Average Rank-Biased Overlap between independent rankings Revealing average benchmarks: An Empirical Investigation

Mark Dragnev

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

Background

Rankings are prevalent in daily life, from sports to search engine results, and are often incomplete, top-weighted, and indefinite. Existing measures, such as Bar-Ilan's ρ [1], Buckley's AnchorMAP [2], and Kendall's τ [3], address (non-)conjoint and (un)weighted rankings but have limitations. Yilmaz et al. [5] improved these with the top-weighted τ_{AP}. Rank-Biased Overlap [4] tackles non-conjoint and indefinite rankings by accommodating differences in length and assigning proper weight to ranks. However, the lack of a reference value for RBO complicates interpreting results. Understanding the factors influencing RBO, like the p-parameter, ranking length, and degree of conjointness, helps establish a benchmark for expected RBO between independent rankings. This benchmark is crucial for assessing the significance of observed scores.

Research Questions

The question that this research strives to answer is *What is the average Rank-Biased Overlap between independent rankings?* In order to answer this, the following sub-questions have to be answered:

- What is the expected RBO between independent rankings when the p-parameter changes?
- What is the expected RBO between independent rankings when their prefix length changes?
- What is the expected RBO between independent rankings when their degree of conjointness changes?

In more detail, the behavior of the expected RBO between independent rankings will be investigated as a function of these variables. Subsequently, insights will be derived to determine, or at least approximate, a reference value based on the observed patterns across different variable configurations. This work only focuses on the assumption that two rankings are of equal size, and that they do not contain any ties.

Example Reference Values for different P and N

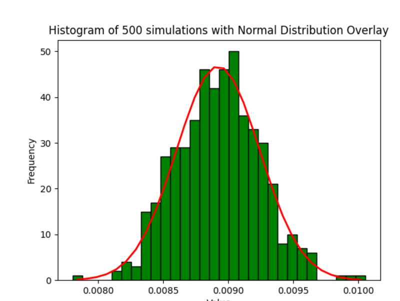
Table 1 gives some Reference Values for different P, N, degree of conjointness between domains, and their sizes. Every simulation used $10,000$ iterations, and is tested 500 times, in order to provide a *mean* value, *standard deviation* and *coefficient of variation*.

P	N	Conj	D1 D2	E[RBO] *mean*	E[RBO] *sd*	E[RBO] *cv*
0.8	5	1	1000	*0.003364*	*7.968e-05*	*0.0237*
	10			*0.004455*	*7.375e-05*	*0.0166*
	15			*0.00482*	*7.013e-05*	*0.0145*
	20			*0.00495*	*6.506e-05*	*0.0131*
	30			*0.005030*	*0.00022*	*0.044*
0.9	5	1	1000	*0.004155*	*0.00026*	*0.064*
	15			*0.007969*	*0.00025*	*0.0314*
	20			*0.008782*	*0.00019*	*0.0226*
	40			*0.009839*	*0.00019*	*0.0198*
	100			*0.010006*	*0.00019*	*0.0198*
0.95	10	1	1000	*0.008025*	*0.00026*	*0.0327*
	20			*0.012823*	*0.00020*	*0.0162*
	40			*0.017427*	*0.00017*	*0.0102*
	100			*0.019876*	*0.00021*	*0.0106*
0.99	10	1	1000	*0.009533*	*0.00029*	*0.0306*
	20			*0.018280*	*0.00027*	*0.0151*
	40			*0.033102*	*0.00023*	*0.0072*
	50			*0.039467*	*0.00023*	*0.0058*
	100			*0.063377*	*0.0002*	*0.0031*
	200			*0.086579*	*0.0002*	*0.0023*
	350			*0.097012*	*0.00017*	*0.0018*

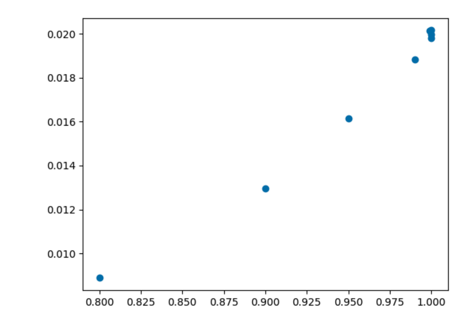
Table 1. Different E[RBO] when p and N (prefix size) change. At some point, it converges around a value, when N is large enough.

Trend of E[RBO] when the p-parameter changes

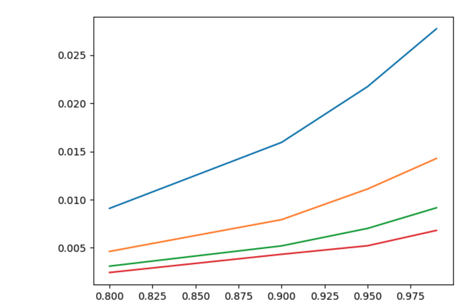
Number of iterations in a single simulation is $10,000$ and still gives room for error. Figure 1a presents a simulation for specific parameters of $P, N, conjointness, Domain_1$ and $Domain_2$. Coefficient of variation is 3.5%. In figure 1b, as p increases, the assumption that the agreement continues indefinitely becomes more significant, reducing the impact of top-ranked discrepancies and leading to higher overall scores. Figure 1c presents the trend when domains gets larger, leading to $E[RBO] \to 0$.



(a) Different values for E[RBO]. $p = 0.8, N = 10, D1 = D2 = 500$.



(b) Growth of E[RBO] when p is increasing. $N = 10, D1 = D2 = 500$.



(c) Growth of E[RBO] when p is increasing. $N = 15, D1 = D2 = [500, 1000, 1500, 2000]$.

Trend of E[RBO] when prefix size changes

$E[RBO]$ increases as more elements are sampled from the same domain, with p determining the growth rate. Lower p values lead to faster convergence, which occurs at larger sizes with higher p. Convergence starts when a prefix reaches a significant weight threshold like 99%.

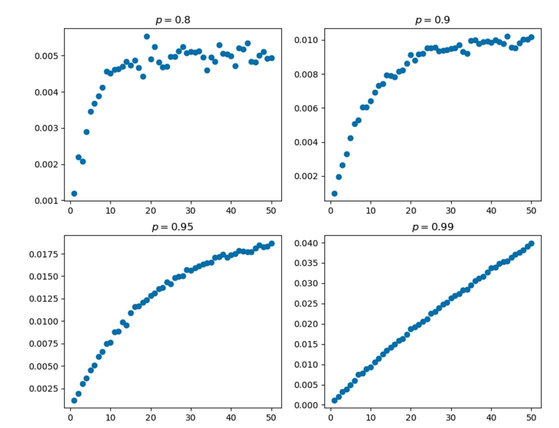
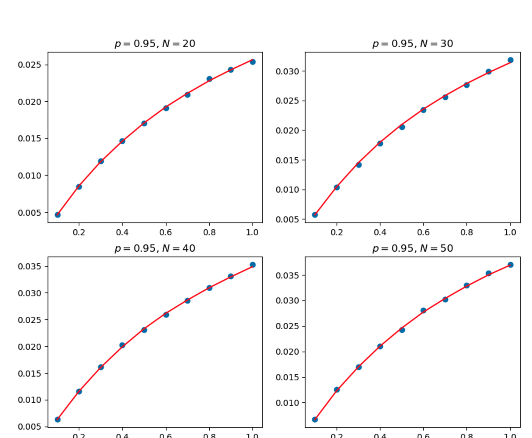


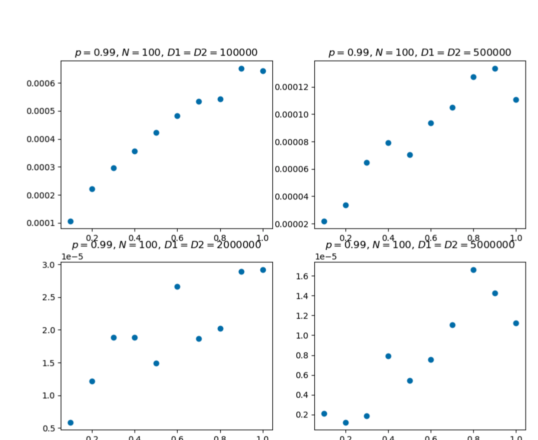
Figure 2. Four different plots for $p = [0.8, 0.9, 0.95, 0.99]$, when prefix size changes. $D1 = D2 = 1000$.

Trend of E[RBO] when conjointness changes

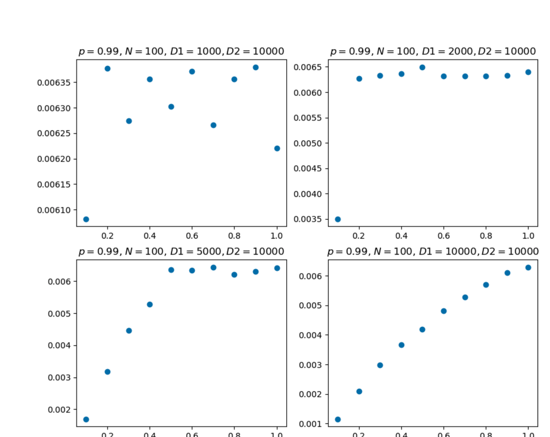
Figure 3a provides different trends. As the conjointness *tends* to 1, the growth decelerates. The growth is similar for larger domains (Figure 3b). And for different sizes of the domains, $E[RBO]$ converges when $conjointness = min(D1, D2)/max(D1, D2)$. If conjointness is fixed, the smaller domain doesn't influence $E[RBO]$.



(a) Trend of E[RBO] for different degrees of conjointness. $p = 0.95, N = [20, 30, 40, 50], D1 = D2 = 500$.



(b) Trend of E[RBO] for different degrees of conjointness. $p = 0.99, N = 100$.



(c) Trend of E[RBO] for different degrees of conjointness. $p = 0.99, N = 100$.

Rank-Biased Overlap

For any $i \in \mathbb{N}$, S_i and T_i represent the elements at position i. *Rank-Biased Overlap* [4] between S and T is defined as the infinite and weighted sum of the agreements at all depths:

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

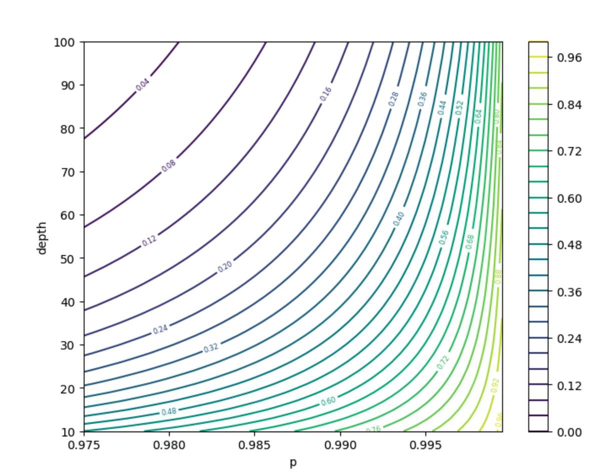
The weight of a prefix is defined as

$$W_{RBO}(1 : d) = 1 - p^{d-1} + \frac{1-p}{p} \times d \times (\ln \frac{1}{1-p} - \sum_{i=1}^{d-1} \frac{p^i}{i}) \tag{2}$$

And finally, RBO_{EXT} is defined as

$$RBO_{EXT}(S, T, p, k) = \frac{X_k}{k} \times p^k + \frac{1-p}{p} \sum_{d=1}^{k} \frac{X_d}{d} \times p^d \tag{3}$$

Weights of prefixes



(a) A contour map of different weights of tails based on N and P

	P=0.8	P=0.9	P=0.95	P=0.99
N=5	0.860864	0.671989	0.476300	0.168775
N=10	0.969034	0.855585	0.672422	0.274809
N=15	0.992234	0.931032	0.784015	0.356887
N=20	0.997931	0.965613	0.853407	0.424174
N=30	0.999838	0.990792	0.928893	0.529912
N=40	0.999986	0.997383	0.964006	0.610101
N=50	0.999999	0.999229	0.981277	0.673131
N=100	1.000000	0.999998	0.999119	0.851864
N=150	1.000000	1.000000	0.999951	0.927287
N=200	1.000000	1.000000	0.999997	0.962768
N=300	1.000000	1.000000	1.000000	0.989501
N=400	1.000000	1.000000	1.000000	0.996861
N=500	1.000000	1.000000	1.000000	0.999027

(b) Weights of prefixes for different sizes and different p, chosen for evaluation.

References

[1] Judit Bar-Ilan.
Comparing rankings of search results on the web.
Information Processing Management, 41(6):1511–1519, 2005.
Special Issue on Infometrics.

[2] Chris Buckley.
Topic prediction based on comparative retrieval rankings.
In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 506–507, New York, NY, USA, 2004. Association for Computing Machinery.

[3] Maurice George Kendall.
Rank correlation methods.
1948.

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

[5] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson.
A new rank correlation coefficient for information retrieval.
In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 587–594, New York, NY, USA, 2008. Association for Computing Machinery.