Rank-Biased Precision Reloaded: Reproducibility and Generalization

Nicola Ferro and Gianmaria Silvello

Department of Information Engineering, University of Padua, Italy
{ferro,silvello}@dei.unipd.it

Abstract. In this work we reproduce the experiments presented in the paper entitled "Rank-Biased Precision for Measurement of Retrieval Effectiveness". This paper introduced a new effectiveness measure – *Rank-Biased Precision (RBP)* – which has become a reference point in the IR experimental evaluation panorama.

We will show that the experiments presented in the original RBP paper are repeatable and we discuss points of strength and limitations of the approach taken by the authors. We also present a generalization of the results by adopting four experimental collections and different analysis methodologies.

1 Introduction

In this paper we aim to reproduce the experiments presented in the paper by A. Moffat and J. Zobel entitled "Rank-Biased Precision for Measurement of Retrieval Effectiveness" published in the ACM Transaction on Information System in 2008 [12]. This work presents an effectiveness measure which had quite an impact on the *Information Retrieval (IR)* experimental evaluation field and also inspired the development of many other measures. Indeed, *Rank-Biased Precision (RBP)* is built around a user model where the browsing model, the document utiliy model and the utility accumulation model are explict [4]; it does not depend on the recall base, which is a quantity difficult to estimate and actually unknown to real users; finally, it matches well with real users, being well correlated with observed click behaviour in system logs [5,21] and allowing to learn models which capture a good share of actual users' way of behaving [11].

The core of RBP resides in its user model, which is defined starting from the observation that a user has no desire of examining every item in a ranking list. The idea is that a user always starts from the first document in the list and then she/he progresses from a document to the other with a probability p, called the *persistence parameter*, and, conversely, ends her/his examination of the list with probability $1 - p$. This assumption allows for the definition of user models representing both patient and impatient users by varying p.

Given a run of d documents, RBP is defined as:

$$\mathrm{RBP} = (1-p)\sum_{i=0}^{d} r_i \cdot p^{i-1}$$

A. Hanbury et al. (Eds.): ECIR 2015, LNCS 9022, pp. 768–780, 2015.

where r_i is the relevance grade of the document at rank i. Since RBP is defined for binary relevance, r_i can assume only two values: 0 or 1.

At the time of writing the RBP paper counts 80 citations on the *Association for Computing Machinery (ACM)* digital library[1] and more than 170 on Google scholar[2], several other works about effectiveness measures rely or are inspired by the user model of RBP and exploited it to define new effectiveness measures.

To repeat the experiments we rely on an open source and publicly available software library called MATTERS (MATlab Toolkit for Evaluation of information Retrieval Systems)[3] implemented in MATLAB[4]. The use of MATLAB allows us to exploit a widely-tested and robust to numerical approximations implementations of the statistical methods needed for analysing the measures such as Kendall's τ correlation measure [10], Student's t test [8] or the Wilcoxon signed rank test [18]. All the data and the scripts we used for reproducing the experiments in [12] are available at the URL: `http://matters.dei.unipd.it/`.

We take reproducibility also as the possibility of both generalizing the original experimental outcomes to other experimental collections in order to confirm them on a wider range of datasets, and validating the experimental hypotheses by means of additional analysis methods. The former led us to repeat the experiments on four different test collections from *Text REtrieval Conference (TREC)* and *Conference and Labs of the Evaluation Forum (CLEF)*; the latter led has to assess the robustness of RBP to shallow pools by using stratified random sampling techniques. We will show how this extended analysis on the one hand allows us to point out additional aspects of RBP and on the other hand provides a solid basis for future uses of this measure.

The paper is organized as follows: Section 2 describes the experiments conducted in the RBP original paper and details the aspects concerning their reproducibility; Section 3 reports about the extended experiments we conducted on RBP and in Section 4 we draw some conclusions.

2 Reproducibility

The experiments in [12] are based on the TREC-05, 1996, Ad-Hoc collection [16] composed of 61 runs (30 automatic and 31 manual runs), 50 topics with binary relevance judgments (i.e. relevant and not-relevant documents), and about 530,000 documents. The authors conducted three main experiments to explore how RBP behaves with shallow pools, also varying the persistence parameter $p = \{0.5, 0.8, 0.95\}$. RBP has been compared against P@10, P@R (precision at the recall-base), and *Average Precision (AP)* [3]; *Normalized Discounted Cumulated Gain (nDCG)* [9], and *Reciprocal Rank (RR)* [17], by considering two pool depths 100 (the original depth of TREC-05) and 10.

[1] `http://dl.acm.org/citation.cfm?id=1416952`

[2] `http://scholar.google.com/`

[3] `http://matters.dei.unipd.it/`

[4] `http://www.mathworks.com/`

The original pool was calculated by taking the union of the first 100 documents of each run submitted to TREC-05 and then assessing the resulting set of documents, whereas the pool at depth 10 was calculated by exploiting the original assessments but applying them to a reduced set composed of the union of the first 10 documents of each run; all the documents not belonging to this set are considered as not-relevant. From the reproducibility point-of-view, this downsampling technique has the advantage of being deterministic not involving any randomization technique in the downsampling of the pools.

The experiments to be reproduced can be divided into three parts:

1. Kendall's τ correlation coefficients calculated from the systems ordering generated by pair of metrics using TREC-05 runs and by considering two pool depths. With respect to the original paper we aim to reproduce Figure 2 on page 9, Figure 4 on page 16 and Table 3 on page 23.
2. Upper and lower bounds for RBP as the p parameter is varied and increasing number of documents (from 1 to 100) are considered. With respect to the original paper we aim to reproduce Figure 5 on page 19.
3. t test and Wilcoxon test for determining the rate at which different effectiveness metrics allow significant distinctions to be made between systems. With respect to the original paper we aim to reproduce Table 4 on page 24.

The TREC-05 data needed to reproduce the paper is released by *National Institute of Standards and Technology (NIST)* and available on the TREC website[5]; it is composed of the original pool with depth 100 and the set of 61 runs submitted to the campaign. When it comes to reproducing some experiments using this kind of data, the first consideration that has to be made regards how to import the run files; indeed, in the TREC format of a run there is the following:

```
<topic-id> Q0 <document-id> <rank> <score> <run-id>
```

where: `topic-id` is a string specifying the identifier of a topic, `Q0` is a constant unused field which can be discarded during the import, `document-id` is a string specifying the identifier of a document, `rank` is an integer specifying the rank of a document for a topic, `score` is a decimal numeric value specifying the score of a document for a topic and `run-id` is a string specifying the identifier of the run. Track guidelines ask participants to rank the output of their systems by increasing value of `rank` and decreasing value of `score`.

The standard software library adopted by TREC for analysing the runs is `trec_eval`[6]. When importing runs, `trec_eval` may modify the actual ordering of the items in the file since it sorts items in descending order of `score`[7] and descending lexicographical order of `document-id`, when `score`s are tied; note that the `rank` value is not considered. We call this *`trec_eval` ordering*.

[5] http://trec.nist.gov/

[6] http://trec.nist.gov/trec_eval/

[7] Note that `trec_eval` also casts the scores of the runs to single precision (float) values while often they contain more decimal values than those supported by single precision numbers. So two `score` values may appear as tied if regarded as single precision value whereas they would have not if regarded as double precision values.

Note that the *trec_eval ordering* represents a cleaning of the data for those runs which have not complied with the track guidelines as far as ordering of the items is concerned but it may modify also correctly behaving runs, if two items have the same `score` but different `rank`, since in this case `trec_eval` reorders them in descending lexicographical order of `document-id` which may be different from the ordering by increasing `rank`.

RBP is not part of the standard `trec_eval` and the paper under exam does not explicitly say whether the authors have extended `trec_eval` to plug-in also RBP or whether they relied on some other script for carrying out the experiments. In the latter case, if one does not deeply know the internals of `trec_eval`, when importing the run files, the original ordering of the items may be kept as granted, under the assumption that the files are well-formed and complying with the guidelines since they have been accepted and then released as official submissions to TREC. We call this latter case *original ordering*.

This aspect has an impact, though small, on the reproducibility of the experiments; indeed, by considering all the documents of all the runs for TREC-05, the *trec_eval ordering* swaps about 2.79% of documents with respect to the *original ordering*; the impact of the swaps on the calculation of the metrics is narrowed down by the fact that most of the swaps (89.21% of the total) are between not-relevant documents, 6.56% are between equally relevant documents while only 4.23% are between relevant and not-relevant documents, thus producing a measurable effect on the metrics calculation.

Table 1 is the reproduction of Table 3 on page 23 of the original RBP paper; we report the Kendall's Tau correlations calculated from the system rankings generated by pairs of metrics by using both the *trec_eval ordering* and the *original ordering* in order to understand which one was most likely used in [12] and to show the differences between the two orderings. We report in bold the numbers which are at least 1% different than those in the table of the reference paper. As we may see for the *trec_eval ordering* only two numbers are at least 1% different from the ones in the paper, whereas there are more differences for the *original ordering*, especially for the correlations with P@R which seems to be more sensitive to small changes in the order of documents with respect to the other metrics. The differences between the two orderings are small, but in the case of the correlation between P@R with depth 100 and RBP.95 with depth 10, if we consider the *original ordering* the correlation is above the 0.9 threshold value [14], whereas with the *trec_eval ordering* – as well as in the reference paper – it is below this threshold. Another significant difference can be identified in the correlation between P@R with depth 100 and RBP.95 with depth 100; indeed, with the *trec_eval ordering* the difference is very close to the threshold value (i.e. 0.895), whereas with the*trec_eval ordering* it goes down to 0.850.

The correlation values obtained with the *trec_eval ordering* are closer to the ones in the reference paper even if they present small differences probably due to numeric approximations and two values present a difference greater than 1%. From this analysis we can assume that the reference paper adopted the

Table 1. Kendall's Tau correlations calculated from the system orderings generated by metric pairs with TREC-05 by using the *trec_eval ordering* and the *original ordering*. Numbers in bold are those which are at least 1% different from the correlations in [12].

treceval ordering					
		depth 100			
Metric	depth	RR	P@10	P@R	AP
RR	10	0.997	0.842	0.748	0.732
P@10	10	0.840	1.000	0.861	0.845
P@R	100	0.746	0.861	1.000	0.908
RBP.5	10	0.926	0.858	0.764	0.755
RBP.8	10	0.888	0.930	**0.819**	0.809
RBP.95	10	0.778	0.882	0.877	0.896
RBP.95	100	0.793	0.916	0.895	**0.859**
nDCG	100	0.765	0.831	0.877	0.915

original ordering					
		depth 100			
Metric	depth	RR	P@10	P@R	AP
RR	10	0.997	0.841	0.747	0.730
P@10	10	0.840	1.000	0.860	0.844
P@R	100	**0.769**	0.861	1.000	0.907
RBP.5	10	0.924	0.858	**0.776**	0.755
RBP.8	10	0.889	0.929	**0.828**	0.809
RBP.95	10	0.779	0.880	**0.905**	0.894
RBP.95	100	0.792	0.913	**0.850**	**0.859**
nDCG	100	0.763	0.829	**0.886**	0.913

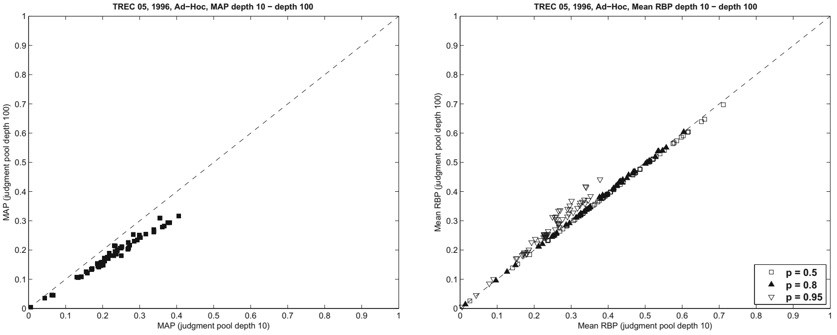
trec_eval ordering, thus in the following we conduct all the other experiments by assuming this ordering for importing the runs.

Another small issue with the reproduction of this experiment is that in the original paper there are no details about the parameters – i.e. weighting schema and log base – used for calculating nDCG; we tested several weighting schema and log bases and we obtained the same number as those in the reference paper by assigning weight 0 to not-relevant documents, 1 to relevant ones and by using log base 2.[8]

Figure 2 and 4 of the original paper regard similar aspects to those presented above in the comment to Table 1 and they concern the correlation between *Mean Average Precision (MAP)* values calculated on the TREC-05 Ad-Hoc collection considering pool depth 100 and pool depth 10 which we show in Figure 1a and the correlation between mean RBP values with p set at 0.5, 0.8 and 0.95 as reported in Figure 1b.

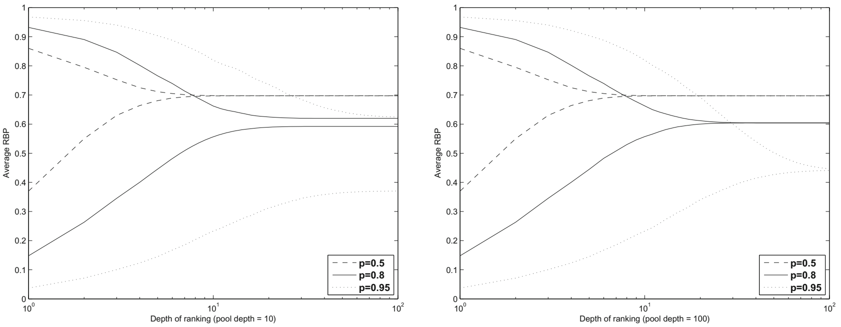
As we can see these two figures are qualitatively equal to those in the original paper and thus these experiments can be considered as reproducible. The main difference regards the choice of the axes which in the reference paper are in the range $[0, 0.4]$ for MAP and $[0, 0.6]$ for mean RBP, whereas we report the graph with axes in the range $[0, 1]$, which is the actual full-scale for both measures. In this way, we can see some MAP values which are above 0.4, showing that MAP calculated with shallow pools tends to overestimate the good runs more than the bad ones. Also for mean RBP we can see some values above the 0.6 limit reported in the original paper; these points show that RBP with $p = 0.5$ with the depth 10 pool tends to overestimate good runs a little more than the bad ones even though these points are also very close to the bisector.

[8] Note that the log base might have guessed by the fact that, on page 21 of the paper, when presenting DCG the authors report that [9] suggested the use of $b = 2$, and employed that value in their examples and experiments.



(a) Mean average precision. **(b)** Mean rank-biased precision.

Fig. 1. Correlation between MAP and mean RBP at pool depth 10 and 100



(a) Pool depth 10 **(b)** Pool depth 100

Fig. 2. Upper and lower bounds of RBP as p is varied and increasing number of documents are considered in the ranking for the "ETHme1" run

The second set of experiments in [12] we aim to reproduce regards upper and lower bounds of RBP evaluated at depth 10 and depth 100. In the usual TREC evaluation setting some documents of a run are assessed (either relevant or not relevant in the binary case), but most of them are left unjudged and normally considered as not-relevant when it comes to calculating effectiveness measures. In [12] it is stated that with this assumption "quoted effectiveness rates might be expected to be pessimistic" and thus represent a lower bound of the measurement; thus, RBP values calculated with this assumption are considered the lower bounds of the measure. They proposed a method to compute a *residual* that captures the unknown component (determined by the unjudged documents) of RBP. Basically, the residual is calculated on a item-by-item basis by summing the weight that the documents would have had if they were relevant; the upper bound is defined by the sum of RBP (i.e. the lower bound) and the residual.

The goal of this experiment is to show that lower and upper bounds stabilize as the depth of the evaluation is increased, even if for higher values of p and

Table 2. Significant differences between systems; the total number of system pairs is 1830 and numbers in bold are at least 1% different from [12]

	Wilcoxon		t test	
Metric	99%	95%	99%	95%
RR	1030	763	1000	752
P@10	**1153**	904	1150	915
P@R	1211	994	1142	931
AP	1260	1077	1164	969
RBP.5	1077	**845**	1052	812
RBP.8	1163	921	1167	918
RBP.95	1232	1009	1209	987
nDCG	1289	**1104**	1267	**1089**

shallow pools they do not converge. This experiment is summarized in Figure 5 on page 19 of the original paper which reports upper and lower bounds of RBP (with p varying from 0.5 to 0.95) for a given run. In the original paper there is no indication about which run has been used in this experiment; as a consequence to reproduce the experiment we had to calculate upper and lower bounds for all the runs and then proceed by inspection of the plots to determine the run used in the original paper. We determined that the used run is named “ETHme1”.

In Figure 2 we present a replica of the figure reported in the original paper where we can see that the upper and lower bound for RBP.5 with the original pool converge before rank 100, whereas for RBP.8 and RBP.95 they converge later on; for the measures calculated with pool depth 10 only RBP.5 converges before rank 100. In this case the original experiment is not easily reproducible because the name of the chosen run was not reported; the same problem prevents the possibility of replicating the plot of Figure 6 on page 20 of the original paper, where the upper and lower bounds of “two systems” are shown: there is no indication about which system pair among the 1830 possible pairs in in TREC-05 have been chosen.

The last experiment to be reproduced regards the t test and the Wilcoxon signed rank test for determining the significant differences between retrieval models according to different measures. In Table 2 we report the values we obtained that have to be compared to those in Table 4 on page 24 of the reference paper. We reported in bold the numbers presenting a difference higher than 1% from the original ones; as we may see there are three major differences for the Wilcoxon test and only one for the t test. We highlight that for the Wilcoxon test 94% of the values are different from the original paper even though the differences are very small (less than 1%); on the other hand, for the t test the 31% of the values we obtained are different from those in the original paper.

Table 3. Features of the adopted experimental collections

Collection	CLEF 2003	TREC 13	CLEF 2009	TREC 21
Year	2003	2004	2009	2012
Track	Ad-Hoc	Robust	TEL	Web
# Documents	1M	528K	2.1M	1B
# Topics	50	250	50	50
# Runs	52	110	43	27
Run Length	1,000	1,000	1,000	10,000
Relevance Degrees	2	3	2	4
Pool Depth	60	100 and 125	60	30 and 25
Languages	EN, FR, DE, ES	EN	DE, EL, FR, IT, ZH	EN

3 Generalization

The experiments conducted in [12] are all based on TREC-5, but these results have not been proven in a wider environment by using different experimental collections (e.g. collections with more runs, more topics, higher and lower original pool depths) or using different pool sampling techniques. Indeed, to the best of our knowledge, the only one other systematic analysis of RBP on different experimental collections is the one by [13], even if it does not concern the original RBP as defined in the reference paper under examination but its extension to multi-graded relevance judgements.

In this section we aim to investigate three main aspects regarding RBP:

- stability to pool downsampling at depth 10 by using two CLEF and two TREC collections;
- the robustness of RBP to downsampled pools (with different reduction rates) according to the stratified random sampling method [2];
- the behavior of RBP upper and lower bound in the average case presenting confidence intervals.

In the following we consider four public experimental collections, whose characteristics are reported in Table 3: (i) CLEF 2003, Multilingual-4, Ad-Hoc Track [1]; (ii) TREC 13, 2004, Robust Track [15]; (iii) CLEF 2009, bilingual X2EN, *The European Library (TEL)* Track [7]; and, (iv) TREC 21, 2012, Web Track [6].

As we can see these collections have different interesting characteristics which allow us to test the behaviour of RBP in a wider range of settings. CLEF 2003 has been used for evaluating multilingual systems with 50 topics and the corpus of one million documents in four different languages; TREC-13 has a high number of runs, topics (i.e. 250) and pool depth (i.e. 125 for 50 topics and 100 for the other 200); CLEF 2009 presents a corpus of documents composed by short bibliographic records and not newspaper articles as in the other CLEF collections and has been used to evaluate bilingual systems working on topics in English and documents in five different languages; and TREC-21 presents a huge multilingual Web corpus, topics are created from the logs of a commercial search engine and it

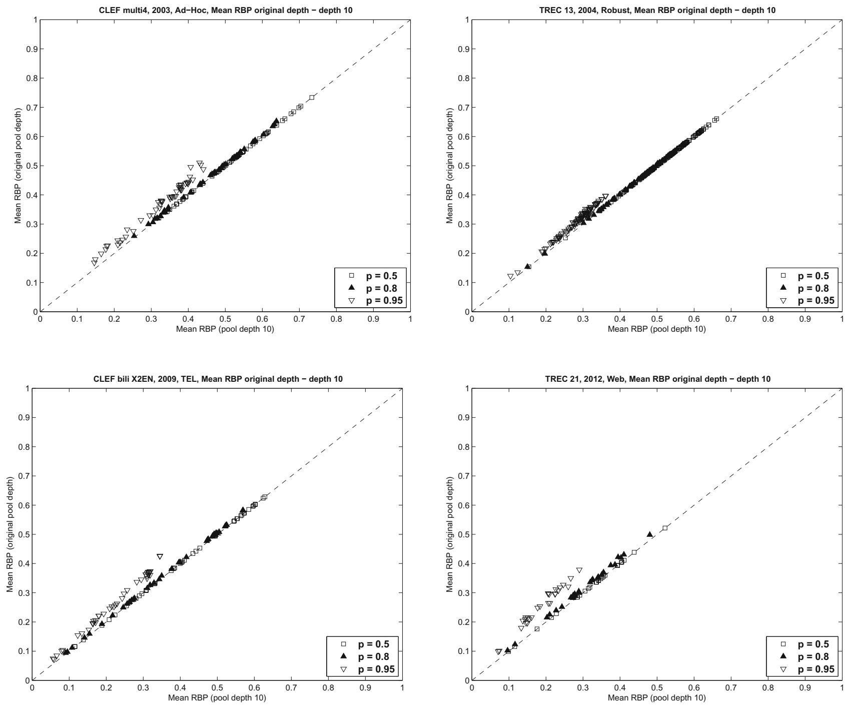


Fig. 3. Robustness of RBP to pool downsampling using different collections

allows us to evaluate up-to-date IR systems working on a Web scale, furthermore 25 topics were judged to depth 30 and 25 to depth 20 [6].

In Figure 3 we can see the correlation between RBP (with the three usual values of $p = \{0.5, 0.8, 0.95\}$) calculated with the original pool depth and with pool depth 10 across the four selected test collections. The results presented in [12] with TREC-05 are confirmed for all the tested collections showing that RBP.5 and RBP.8 are robust to pool downsampling, whereas RBP.95 tends to underestimate the effectiveness of the runs when calculated using pool depth 10; this effect is more evident with TREC-21 where also RBP.8 values are slightly above the bisector.

The *stratified random sampling* of the pools allows us to investigate the behavior of RBP as the relevance judgment sets become less complete following the methodology presented in [2]: Starting from the original pool (100% of the relevance judgments) for each topic we select a list of relevant documents in random order and a list of not-relevant documents in random order; then, we create alternative pools by taking $\{90, 70, 50, 30, 10\}\%$ of the original pool. For a target pool which is $P\%$ as large as the original pool, we select $X = P \times R$ relevant documents and $Y = P \times N$ not-relevant documents or each topic where R is the number of relevant documents in the original pool and N is the number of judged not-relevant documents in the original pool. We use 1 as the minimum number of

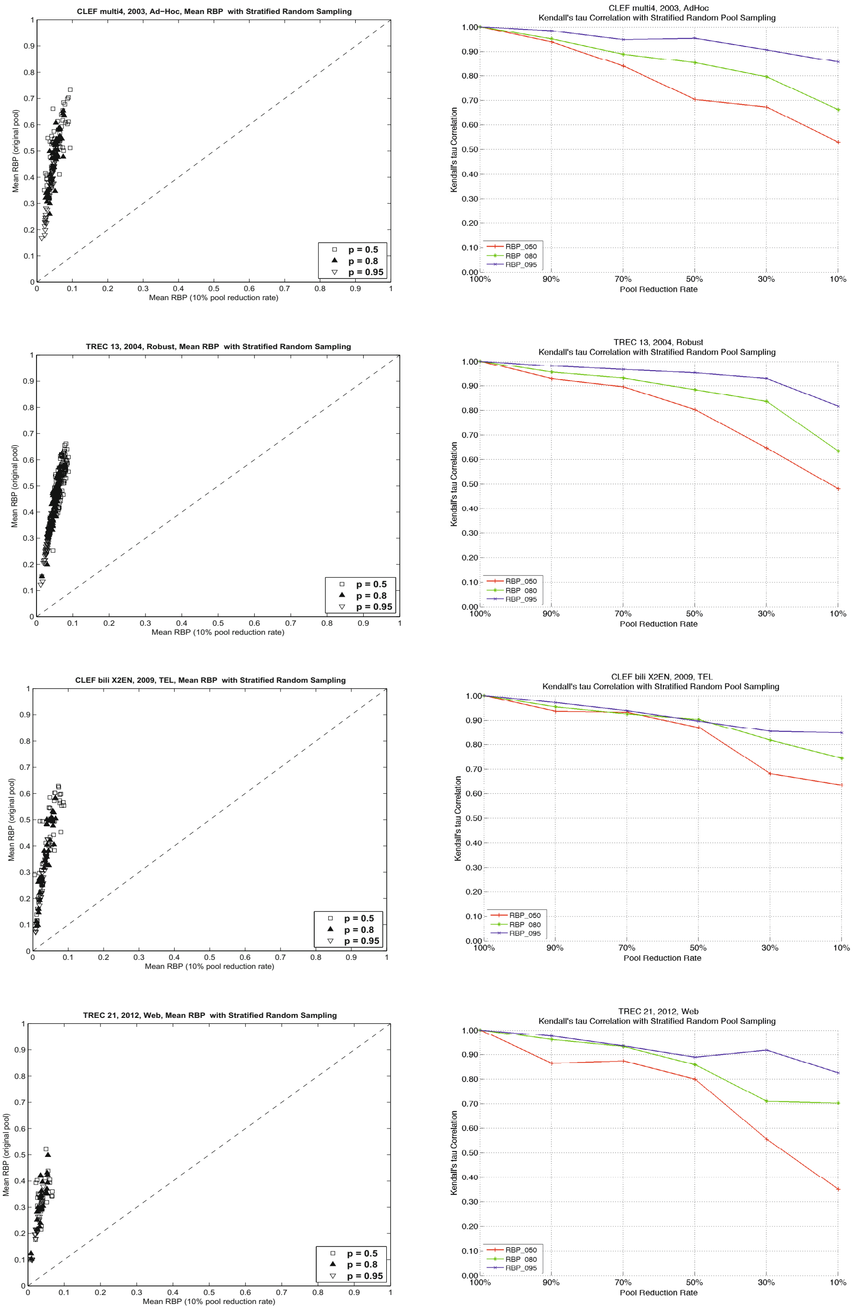


Fig. 4. On the left-hand side are the Kendall's τ between the original pool and the 10% downsampled pool (that can be compared with those in Figure 3 adopting a pool downsampled to depth 10) and on the right-hand side there is the change in Kendall's τ as judgment sets are downsampled for the CLEF and TREC collections.

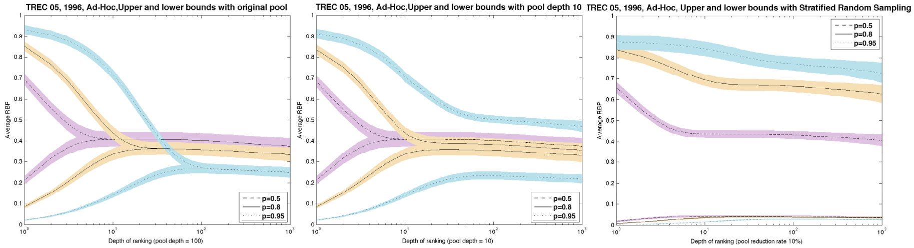


Fig. 5. Upper and lower bounds of average RBP as p is varied and the number of documents in the ranking increases from 1 to 1000

relevant documents and 10 as the minimum number of judged not-relevant documents per topic. Since we take random subsets of a pool that is assumed to be fair, the reduced pools are also unbiased with respect to systems; this methodology is equivalent to perform uniform random sampling of the pool [19], which is desirable to infer statistical properties. This methodology allows us to further explore the robustness of RBP to pool downsampling; it must be underlined that for each pool sample, relevant documents are selected at random and thus the results here reported are not exactly reproducible even if the conclusions emerging from this test do not change from sample to sample. Figure 4 shows how RBP behaves as the pool is downsampled with the stratified random sampling techniques for the TREC collections.

The plots on the left-hand side of the figures show the correlation of RBP values calculated with the original pool versus RBP calculated with a pool at a 10% reduction rate. We can see that RBP calculated with the pool at a 10% reduction rate highly underestimates the effectiveness of the runs and highly reduces the interval of values it assumes – i.e. most of the values are in the $[0, 0.1]$ interval. From these plots it is not possible to see a significant difference between RBP.5, RBP.8 and RBP.95. The plots on the right-hand side show the robustness of RBP at different reduction rates: the higher the curves the more stable the measure. As we can see for all TREC collections show the same ordering between RBP.5, RBP.8 and RBP.95, where RBP.95 is always more robust than the other two. This result contradicts the previous one (see Figure 3) where RBP.95 is the less robust measure. The results obtained with the stratified random sampling allow us to say that RBP with different p values calculated with a pool reduction rate of 10% seriously narrows down the interval of effectiveness values a run can achieve; on the other hand, we see that RBP.95 always has a Kendall's τ correlation between the original and the 10% downsampled pool in the $[0.8, 0.9]$ interval.

Lastly, in Figure 5 we present a generalization of Figure 2 which reports RBP upper and lower bounds calculated by averaging over all the runs of the TREC-05 collection instead of choosing a specific run as representative of the whole collection. We also reported the confidence interval of the measures and we show how the bounds behave up to rank 1000 (i.e. the maximum length of the runs); furthermore, we show how the bounds behave when RBP is calculated by adopting a pool with 10% reduction rate determined with the stratified random sampling technique. We can see that with the original pool as well as with pool downsampled to depth 10 the results are consistent with those reported in the RBP original paper for RBP.5 and RBP.95, whereas it shows that RBP.8 tends to converge between rank 10 and 100. However, upper and lower bounds of RBP calculated with 10% pool reduction rate never converge for all the considered values of p showing a high impact of unjudged documents on RBP values. The very same trends emerge for the RBP bounds calculated with the other collections we presented above; we do not report the plots for space reasons.

4 Conclusions

In this paper we discussed the experiments conducted in [12] where the RBP measure was presented and described for the first time. We have shown that most of the experiments presented in the original RBP paper are reproducible, even though there are precautions that should be taken with presenting experiments about experimental evaluation in IR. These include: (i) explicitly describing the choices made about document ordering – e.g. explaining if the `trec_eval` document ordering is applied or not; (ii) explicitly reporting the name or id of the systems used for the experiments – e.g. the "ETHme1" run in Figure 2 – or specifying which subset of systems has been selected from the whole collection; (iii) reporting all the parameters used for calculating a measure – e.g. weighting schema and log base for nDCG. It must be highlighted that the experiments were reproducible because they were originally conducted on publicly available and shared datasets such as the TREC-05 Ad-Hoc collection.

From the reproducibility point of view, the presentation of the results by means of tables would be preferable to only using plots, because they allow for a thorough verification of the results; graphs and plots are useful for understanding the results from a qualitative perspective, but they always should be accompanied by the numerical data on which they rely (they can be presented also in an appendix of the paper or made available online).

The generalization part of this work shows that the results presented in the original RBP paper are verifiable also with other public and shared experimental collections. On the other hand, we show that the use of different analysis methodologies (e.g. a different pool downsampling technique) could lead to different conclusions that must be taken into account in order to employ RBP for experimental evaluation. As we have seen by using pool downsampling RBP.5 is the most robust measure, but it is the less robust by using the stratified random sampling method; we reach the same conclusion by considering RBP bounds that, with a 10% pool reduction rate, do not converge for any p value up to rank $1,000$.

References

1. Braschler, M.: CLEF 2003 – Overview of Results. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 44–63. Springer, Heidelberg (2004)

2. Buckley, C., Voorhees, E.M.: Retrieval Evaluation with Incomplete Information. In: Proc. 27th Ann. Int. ACM Conference on Research and Development in IR (SIGIR 2004), pp. 25–32. ACM Press, USA (2004)
3. Buckley, C., Voorhees, E.M.: Retrieval System Evaluation. In: TREC. Experiment and Evaluation in Information Retrieval, pp. 53–78. MIT Press, Cambridge (2005)
4. Carterette, B.A.: System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In: Proc. 34th Ann. Int. ACM Conference on Research and Development in IR (SIGIR 2011), pp. 903–912. ACM Press, USA (2011)
5. Chapelle, O., Metzler, D., Zhang, Y., Grinspan, P.: Expected Reciprocal Rank for Graded Relevance. In: Proc. 18th Int. Conference on Information and Knowledge Management (CIKM 2009), pp. 621–630. ACM Press, USA (2009)
6. Clarke, C.L.A., Craswell, N., Voorhees, H.: Overview of the TREC 2012 Web Track. In: The Twenty-First Text REtrieval Conference Proceedings (TREC 2012), NIST, SP 500-298, USA, pp. 1–8 (2013)
7. Ferro, N., Peters, C.: CLEF 2009 Ad Hoc Track Overview: TEL and Persian Tasks. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 13–35. Springer, Heidelberg (2010)
8. Gosset, W.S.: The Probable Error of a Mean. Biometrika (1), 1–25 (1908)
9. Järvelin, K., Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems (TOIS) 20(4), 422–446 (2002)
10. Kendall, M.G.: Rank correlation methods. Griffin, Oxford, England (1948)
11. Moffat, A., Thomas, P., Scholer, F.: Users Versus Models: What Observation Tells Us About Effectiveness Metrics. In: Proc. 22h Int. Conference on Information and Knowledge Management (CIKM 2013), pp. 659–668. ACM Press (2013)
12. Moffat, A., Zobel, J.: Rank-Biased Precision for Measurement of Retrieval Effectiveness. ACM Transactions on Information Systems 27(1), 1–27 (2008)
13. Sakai, T., Kando, N.: On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments. Inf. Retrieval 11(5), 447–470 (2008)
14. Voorhees, E.: Evaluation by Highly Relevant Documents. In: Proc. 24th Ann. Int. ACM Conference on Research and Development in IR (SIGIR 2001), pp. 74–82. ACM Press, USA (2001)
15. Voorhees, E.M.: Overview of the TREC 2004 Robust Track. In: The 13th Text REtrieval Conference Proceedings (TREC 2004), USA, pp. 500–261 (2004)
16. Voorhees, E.M., Harman, D.K.: Overview of the Fifth Text REtrieval Conference (TREC-5). In: The 5th Text REtrieval Conference (TREC-5), NIST, SP 500-238, pp. 1–28 (1996)
17. Voorhees, E.M., Tice, D.M.: The TREC-8 Question Answering Track Evaluation. In: The 8th Text REtrieval Conference (TREC-8), NIST, SP 500-246, USA, pp. 83–105 (1999)
18. Wilcoxon, F.: Individual Comparisons by Ranking Methods. Biometrics Bulletin 1(6), 80–83 (1945)
19. Yilmaz, E., Aslam, J.A.: Estimating Average Precision when Judgments are Incomplete. Knowledge and Information Systems 16(2), 173–211 (2008)
20. Yilmaz, E., Shokouhi, M., Craswell, N., Robertson, S.: Expected Browsing Utility for Web Search Evaluation. In: Proc. 19th Int. Conference on Information and Knowledge Management (CIKM 2010), pp. 1561–1565. ACM Press, USA (2010)
21. Zhang, Y., Park, L., Moffat, A.: Click-based evidence for decaying weight distributions in search effectiveness metrics. Inf. Retrieval 13(1), 46–69 (2010)