The Solitude of Relevant Documents in the Pool

Lipani, A.; Lupu, M.; Kanoulas, E.; Hanbury, A.

DOI
10.1145/2983323.2983891

Publication date
2016

Document Version
Final published version

Published in
CIKM'16

License
Article 25fa Dutch Copyright Act

Link to publication

Citation for published version (APA):
Lipani, A., Lupu, M., Kanoulas, E., & Hanbury, A. (2016). The Solitude of Relevant Documents in the Pool. In *CIKM'16: proceedings of the 2016 ACM Conference on Information and Knowledge Management : October 24-28, 2016, Indianapolis, IN, USA* (pp. 1989-1992). Association for Computing Machinery. https://doi.org/10.1145/2983323.2983891

The Solitude of Relevant Documents in the Pool

Aldo Lipani[1] Mihai Lupu[1] Evangelos Kanoulas[2] Allan Hanbury[1]

[1]Inst. of Software Technology & Interactive Systems, Vienna University of Technology, Vienna, Austria
{lipani, lupu, hanbury}@ifs.tuwien.ac.at

[2]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
e.kanoulas@uva.nl

ABSTRACT

Pool bias is a well understood problem of test-collection based benchmarking in information retrieval. The pooling method itself is designed to identify all relevant documents. In practice, 'all' translates to 'as many as possible given some budgetary constraints' and the problem persists, albeit mitigated. Recently, methods to address this pool bias for previously created test collections have been proposed, for the evaluation measure precision at cut-off ($P@n$). Analyzing previous methods, we make the empirical observation that the distribution of the probability of providing new *relevant* documents to the pool, over the runs, is log-normal (when the pooling strategy is fixed depth at cut-off). We use this observation to calculate a prior probability of providing new relevant documents, which we then use in a pool bias estimator that improves upon previous estimates of precision at cut-off. Through extensive experimental results, covering 15 test collections, we show that the proposed bias correction method is the new state of the art, providing the closest estimates yet when compared to the original pool.

Keywords

Pool bias, $P@n$, test collections, TREC

1. INTRODUCTION

The pool bias refers to an undesirable side effect caused by the use of the *pooling method*. This bias manifests itself by the discounting of Information Retrieval (IR) systems when tested on previously built test collections, due to the retrieval of potentially relevant but non-judged documents [4].

An IR test collection is composed of: a collection of documents, a set of topics, and ideally a complete set of paired relations between topics and documents called relevance assessments, which indicate whether a document is either relevant or non-relevant for a given topic. Soon in the history of IR, with the explosion in size of the document collections and therefore of the number of relevance assessments required, the building of such test collections became impractical; to address this issue, the pooling method was introduced [9].

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

© 2016 ACM. ISBN 978-1-4503-4073-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2983323.2983891

The main advantage of the pooling method is to limit the number of relevance assessments through the use of previously collected search results provided by various retrieval systems. The first introduced and most commonly used pooling strategy is the *fixed-depth at n*. This strategy consists in selecting the top n documents retrieved by each retrieval system, then collecting them in a set, called the pool, which later will be fully judged. Although its main limitation is due the introduced incompleteness of the relevance assessments, as pointed out by Spärck Jones [3], the aim of the pooling method is not to find all the relevant documents but an unbiased sample of them. However, this is not an easy problem because, in the absence of full evaluation, and in the presence of millions of potentially relevant documents, there is no guarantee that a future retrieval method will not retrieve completely different yet relevant documents.

There are three ways to mitigate the pool bias: 1) increasing the depth of the pool, 2) increasing the number of topics, which would reduce the variance and, 3) increasing the number of pooled runs, which should increase the diversity of the pooled documents. Each of them has as effect an increment on the number of judgments to be performed. However, only the first two solutions are directly controllable by the test collection builder, leaving the third to the participation of the IR community in solving the challenge for which the test collection is going to be built.

To tackle the pool bias issue, the IR community has pursued two research paths, one developing pooling strategies aimed at mitigating the pool bias when creating new test collections [2, 8, 7] and the other, developing pool bias estimators aimed at correcting the observed pool bias on existing test collections. Witnessing an explosion in the number of test collections developed each year that use the pooling method with fixed-depth at n pooling strategy and an increasing bias due to an increasing trend in the creation of test collections for niche domain-specific IR, where a required minimal participation is usually not met, in this paper we extend the work done on the latter research path. We develop a new pool bias estimator for the metric Precision at cut-off ($P@n$) based on empirical observations. There are two reasons for working with such a simple metric: there is an increasing demand from practitioners for metrics that make 'sense' [1], and it is a corner stone for other more complex metrics.

2. BACKGROUND

Webber and Park [10] introduced a method to correct the pool bias, which they tested on Rank-Biased-Precision at cut-off but claimed to work also with $P@n$. This method adds to the score of a new run a coefficient equal to the mean difference, indicated with $\delta P@n$, between the score obtained when a run, initially part of the pool

R_p, is pooled and not-pooled:

$$\delta P@n(r_p) = P@n(r_p, Q^{R_p}) - P@n(r_p, Q^{R_p \setminus \{r_p\}}) \quad (1)$$

The correction coefficient for a run ($r_u \notin R_p$) is the expectation:

$$\underset{r_p \in R_p}{\mathrm{E}} [\delta P@n(r_p)] \quad (2)$$

where Q^{R_p} is the set of judged documents created using the set of pooled runs R_p. This approach assumes that any given new run is sampled from the same distribution as the pooled ones. This is of course not always true, because runs are selected based on their performance by human intervention. A limitation of this approach is that it computes a coefficient that is constant and therefore does not depend on the actual status of the biased run. Another limitation of this approach is that the correction is not bounded by the tested run, thereby we may have a score that may exceed the upper limit of the $P@n$ co-domain, which is defined as $[0, 1]$.

Lipani et al. [5] presented a more sophisticated method for correcting $P@n$ that outperforms the Webber approach. It is based on measuring the effect of the new run on the pooled runs in terms of $P@n$, and anti-precision at cut-off ($\overline{P}@n$, the ratio of non relevant document over the cut-off n). The effects are observed via the application of a merging function between the new run and the pooled runs, where the function changes the rank of the documents in the pooled runs based on the shared documents with the new run. After having computed these quantities, an indicator is calculated based on them, and used to trigger or not the correction method. If triggered, the correction is equal to the ratio between the non-judged documents over n times the maximum likelihood estimator for the probability of relevance of the non-judged documents in the run.

3. ANALYSIS

In this section we analyze empirically the assumptions about the Webber approach [10]. The mean in Eq. (2) is valid under the assumption that the distribution of $\delta P@n$, as defined in Eq. (1), is normal. This assumption implies that runs are similar to each other, which is of course not true. To empirically demonstrate the groundlessness of this hypothesis, we observe in Fig. 1, on the top, that the distribution of the $\delta P@n$ is not normally distributed. This makes the estimate for the correction biased for a new run because the distribution of runs is not centered on the calculated mean.

To find a better estimate, we look at the ratio between the number of uniquely identified relevant documents discovered by pooling the run r_p and the number of non-judged documents that the run would have if it had not been pooled. This quantity may be interpreted as the probability of the non-judged documents of the run to be relevant:

$$\begin{aligned} &P(d \in [r_p]_1^n \setminus Q^{R_p \setminus \{r_p\}},\, d \in Q_+^{R_p}) = \\ &= \frac{P@n(r_p, Q^{R_p}) - P@n(r_p, Q^{R_p \setminus \{r_p\}})}{1 - \left(P@n(r_p, Q^{R_p \setminus \{r_p\}}) + \overline{P}@n(r_p, Q^{R_p \setminus \{r_p\}})\right)} = \\ &= \frac{\delta P@n(r_p)}{\bar{k}'@n(r_p)} \end{aligned} \quad (3)$$

where we use the notation introduced by Lipani et al. [5]: $\bar{k}'@n$ is the ratio between the number of non-judged documents and n when r_p is not in the pool; $\overline{P}@n$ is the anti-precision at cut-off, also known in statistics as the false discovery rate, the ratio between the number of non-relevant documents and n; and $Q_+^{R_p}$ is the set of pooled documents that are relevant. We observe empirically that its distribution is log-normal. Indicating with X this distribution and with Y its log-transformation $Y = \ln(X)$, Y is

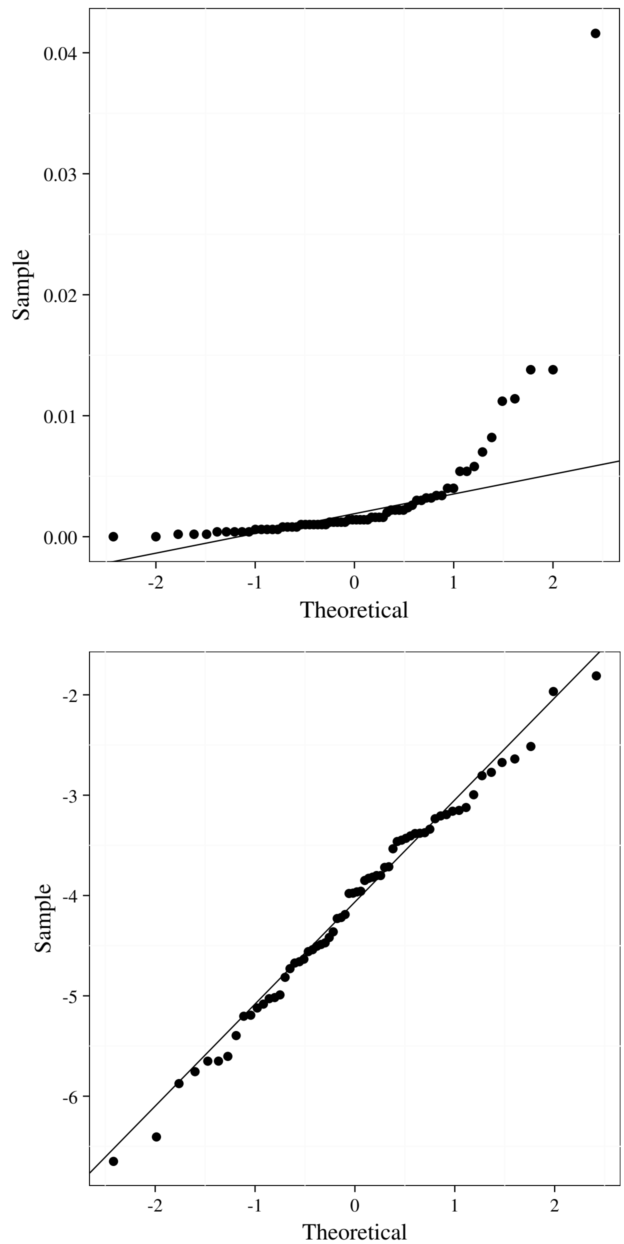

Figure 1: Q-Q Plots of a normal distribution against, on the top, the distribution of δP, as defined in Eq. (1) and, on the bottom, the log transformation of the distribution of the probability of providing new relevant documents to the pool $\delta P/\bar{k}$, as defined in Eq. (3), for the test collection Ad Hoc 8.

normally distributed. In the Q-Q plot in Fig. 1 on the bottom, we observe how the theoretical normal distribution correlates with the sample distribution. To calculate the mean of the new estimate, we compute the mean of the distribution Y and then transform it back to the domain of the distribution X, which leads to the geometric mean of the X:

$$e^{\mathrm{E}[\ln(X)]} = e^{\mathrm{E}[Y]} = \sqrt[|Y|]{\prod_{y \in Y} y} = \mathrm{GM}[X] \quad (4)$$

where GM is the geometric mean. Thereby, the correction coeffi-

Algorithm 1 Revised estimator based on the Webber approach

$r_u \leftarrow$ unpooled run
$R_p \leftarrow$ set of pooled runs
$T \leftarrow$ set of topics
$Q \leftarrow$ qrels on T derived from R_p
$s_{r_u} \leftarrow P@n(r_u, Q)$
$\bar{k}_{r_u} \leftarrow 1 - \left(s_{r_u} + \overline{P}@n(r_u, Q)\right)$
for all $r_p \in R_p$ **do**
  $R'_p \leftarrow R_p \setminus \{r_p\}$
  $Q' \leftarrow$ qrels on T derived from R'_p
  $s'_{r_p} \leftarrow P@n(r_p, Q')$
  $\delta P_{r_p} \leftarrow P@n(r_p, Q) - s'_{r_p}$
  $\bar{k}'_{r_p} \leftarrow 1 - \left(s'_{r_p} + \overline{P}@n(r_p, Q')\right)$
end for
$a \leftarrow \bar{k}_{r_u} \cdot \mathrm{GM}_{r_p \in \{r \in R_p \,:\, \delta P_r \neq 0\}} \left[\delta P_r / \bar{k}'_{r_p}\right]$
return $s_{r_u} + a$

cient for a run ($r_u \notin R_p$) is the following function:

$$\bar{k}@n(r_u) \cdot \underset{r_p \in R_p}{\mathrm{GM}} \left[\delta P@n(r_p)/\bar{k}'@n(r_p)\right] \qquad (5)$$

Comparing Eq. (2) and Eq. (5) we notice that the numerator of this last equation is the same, but with the difference that now, every $\delta P@n$ gets divided by the number of uniquely identified documents provided to the pool, and then multiplied by the same but for r_u, and the arithmetic mean has been substituted by the geometric mean.

However, Eq. (4) is only valid when X is well behaved in the sense of not having zero values. Considering this, in Eq. (5) it mandates the removal of those runs for which δP_{r_p} is zero. This is reasonable because: first, if $\delta P_{r_p} \neq 0$ it can be shown that $\bar{k}'_{r_p} \neq 0$ and consequently the fraction in Eq. (5) is well defined. Second, such zero values bring no information to our estimate of the contribution of the run. In fact, one could generate an unbounded number of runs with $\delta P_{r_p} = 0$.

In Alg. 1 we make use of Eq. (5), filtering out 0 values, and redefining the previous Webber approach. Two benefits of this modification are that no longer every correction is based on a prior probability of the run to find relevant documents among its non-judged ones, and that it is not a constant number as it was before. Also, comparing the algorithm of the Lipani approach [5], we observe here some similarities, such as the use of $\bar{k}'@n$ and of $\overline{P}@n$.

4. EXPERIMENTS & RESULTS

In this section, we present the experimental setup and the material used. It follows the methodology previously used by the already mentioned studies [10, 5, 6]. In the second part, we discuss the results.

4.1 Material & Experimental Set up

To test the effectiveness of the new algorithm, we tested[1] the method on 15 test collections sampled from the TREC evaluation campaign: 7 from the Ad Hoc track, 3 from the Web track, and 1 from each of the Genomics, Robust, Legal, Medical, and Microblog tracks. This sample allowed us to also explore more specific IR domains, where a different behavior may be expected. The experiments consist in running repeated simulations, where a run at each iteration is taken out of the pool with all the runs submitted by the same organization, and a synthetic pool is generated, in a leave-one organization-out fashion. This is done to better simulate the absence of any similar run added to the pool by the same organization. Moreover, to avoid the presence of runs produced by buggy implementations of IR search engines, the bottom 25% of poorly performing runs are filtered out.

To measure the effectiveness of the different approaches, three measures are used: Mean Absolute Error (MAE), System Rank Error (SRE) and System Rank Error with statistical significance (SRE*). MAE measures the mean of the absolute difference of runs' scores when in the pool and when not. SRE measures the difference in ranks between the true ranking and new ranking with using the corrections. Finally, SRE* is similar to SRE, but counts the difference only if statistically significant (paired t-test, $p < 0.05$).

We compared the new approach with the previously mentioned Webber and Lipani approaches, and the baseline, which is the Reduced Pool. The Reduced Pool measures the bias we get if no correction is provided.

4.2 Results

In Tab. 1 the full results are presented. We observe that our approach is top performer in the majority of the cases (45 for MAE, 37 for SRE and 9 for SRE*), with some worst results (15 for MAE, 16 for SRE and, 6 for SRE*) mostly obtained on Ad Hoc 7 and Ad Hoc 8, where the Lipani approach performs better. The Lipani approach, although in general it is not as good as our approach, which has a more stable behavior, collecting fewer worst results (9 for MAE, 6 for SRE and, 1 for SRE*). The Webber approach and the Reduced Pool are the ones that get the majority of the worst results, as also demonstrated in previous work [5].

When our approach is applied to the test collections: Ad Hoc 3, Ad Hoc 5, Ad Hoc 7, Ad Hoc 8, Web 9, Web 2002, Genomics 2005 and Microblog 2011 we can observe that the MAE for the correction applied to $P@5$ gets the worst score (8 of the 15 worst performer cases mentioned above). Here, after an empirical analysis, we have two hypotheses for why this happens, and we claim that it happens for a combination of both. The first hypothesis relies on the observation that for such test collections, the ratio between the number of pooled runs and the number of organizations is much greater than 1. This means that multiple runs from the same organization have been pooled and therefore contribute a very similar set of documents to the pool. Thereby, it nullifies the leave-one run-out approach embedded in our and the Webber methods, as shown in Eq. (2) and (5), looking inside E in the first and GM in the second equation, when the run r_p, originally in the pool Q^{R_p} is removed from the pool ($Q^{R_p \setminus \{r_p\}}$). Note that this leave-one run-out happens inside the algorithm and is different from the leave-one organization-out testing procedure that happens before the algorithm is triggered. However, for Legal 2006 and Medical 2011, although the ratio is also large, we may not observe the same effect due the more shallow pool depth. The second hypothesis is that when we want to count the top number of relevant documents of a run and we have such large number of relevance judgments, there is a high likelihood that the top documents of every run have been already pooled by other runs. This means that, as for the first hypothesis, the effect of the leave-one run-out embedded in the method is nullified. We split these two hypotheses because they have a different nature, although they have the same effect that can be mitigated by the same solution.

In general, these two hypotheses cause a significant error due to the fact that the number of points collected in order to compute an unbiased estimate of the geometric mean is insufficient. When these sets are small, it means that either there are no non-judged documents, or the ones that exist bring no new information in terms

[1] The software is available on the website of the first author.

Table 1: Summary of the results per test collection generated through a leave-one organization-out using the top 75% of the best performing pooled runs. With: $|R|$ number of runs submitted, $|O|$ number of organizations involved, $|R_p|$ number of pooled runs, d depth of the pool and $|T|$ number of topics. Last row counts how many times each approach per metric performed best or worst, tie not included.

Track		P@n	Our			Lipani			Webber			Reduced Pool		
			MAE	SRE	SRE*	MAE	SRE	SRE*	MAE	SRE	SRE*	MAE	SRE	SRE*
Ad Hoc 2	$\|R\|$: 38	5	0.0029	13	1	0.0041	13	1	0.0049	14	1	0.0047	13	1
	$\|O\|$: 22	10	0.0046	12	0	0.0093	18	0	0.0102	16	0	0.0095	18	0
	$\|R_p\|$: 30	20	0.0043	11	0	0.0108	27	0	0.0112	25	0	0.0122	30	0
	d : 100	30	0.0053	7	0	0.0134	30	0	0.0124	19	0	0.0153	32	0
	$\|T\|$: 50	100	0.0069	14	0	0.0226	66	0	0.0163	42	1	0.0307	94	1
Ad Hoc 3	$\|R\|$: 40	5	0.0061	2	0	0.0024	1	0	0.0032	1	0	0.0024	1	0
	$\|O\|$: 22	10	0.0036	3	0	0.0022	2	0	0.0023	0	0	0.0023	2	0
	$\|R_p\|$: 21	20	0.0029	1	0	0.0028	1	0	0.0027	2	0	0.0030	1	0
	d : 200	30	0.0035	4	0	0.0038	7	0	0.0039	8	0	0.0041	7	0
	$\|T\|$: 50	100	0.0047	13	0	0.0065	13	0	0.0060	15	0	0.0085	14	0
Ad Hoc 4	$\|R\|$: 33	5	0.0082	13	0	0.0084	13	0	0.0104	16	0	0.0098	14	0
	$\|O\|$: 19	10	0.0092	12	0	0.0104	14	0	0.0114	15	0	0.0119	16	0
	$\|R_p\|$: 24	20	0.0090	10	0	0.0111	14	0	0.0124	18	0	0.0132	20	0
	d : 100	30	0.0103	10	0	0.0116	13	0	0.0128	17	0	0.0144	20	0
	$\|T\|$: 50	100	0.0127	21	0	0.0128	20	0	0.0144	28	0	0.0218	38	3
Ad Hoc 5	$\|R\|$: 61	5	0.0139	60	1	0.0073	37	3	0.0083	38	3	0.0076	38	3
	$\|O\|$: 21	10	0.0077	43	0	0.0073	40	0	0.0080	42	0	0.0079	45	0
	$\|R_p\|$: 53	20	0.0075	60	0	0.0068	41	0	0.0079	55	0	0.0077	49	0
	d : 100	30	0.0067	59	0	0.0068	41	0	0.0081	57	0	0.0083	54	0
	$\|T\|$: 50	100	0.0069	84	0	0.0053	61	0	0.0084	107	0	0.0102	114	1
Ad Hoc 6	$\|R\|$: 74	5	0.0089	7	0	0.0102	9	5	0.0158	14	5	0.0118	9	5
	$\|O\|$: 29	10	0.0068	4	0	0.0057	6	0	0.0084	6	0	0.0063	6	0
	$\|R_p\|$: 25	20	0.0050	2	0	0.0058	4	0	0.0072	8	0	0.0071	4	0
	d : 100	30	0.0048	3	0	0.0062	1	0	0.0068	3	0	0.0079	2	0
	$\|T\|$: 50	100	0.0051	9	0	0.0068	17	0	0.0073	13	0	0.0116	16	0
Ad Hoc 7	$\|R\|$: 103	5	0.0085	44	5	0.0024	8	2	0.0032	8	2	0.0025	8	2
	$\|O\|$: 42	10	0.0063	47	0	0.0024	7	0	0.0030	7	0	0.0025	7	0
	$\|R_p\|$: 63	20	0.0037	25	0	0.0023	12	0	0.0027	16	0	0.0025	14	0
	d : 100	30	0.0034	28	0	0.0024	28	0	0.0029	26	0	0.0028	29	0
	$\|T\|$: 50	100	0.0035	43	0	0.0028	34	0	0.0035	42	0	0.0043	54	0
Ad Hoc 8	$\|R\|$: 129	5	0.0050	22	2	0.0020	3	0	0.0035	3	0	0.0021	3	0
	$\|O\|$: 41	10	0.0050	29	4	0.0021	3	0	0.0033	3	0	0.0022	3	0
	$\|R_p\|$: 66	20	0.0034	19	2	0.0020	5	0	0.0031	11	0	0.0022	5	0
	d : 100	30	0.0030	15	2	0.0021	8	0	0.0031	15	0	0.0024	10	0
	$\|T\|$: 50	100	0.0024	21	3	0.0022	22	1	0.0032	32	7	0.0036	48	10
Web 9	$\|R\|$: 104	5	0.0030	10	0	0.0024	9	0	0.0023	9	0	0.0022	9	0
	$\|O\|$: 23	10	0.0022	9	0	0.0020	13	0	0.0022	13	0	0.0021	13	0
	$\|R_p\|$: 39	20	0.0019	8	0	0.0035	17	0	0.0036	20	0	0.0039	20	0
	d : 100	30	0.0023	12	0	0.0035	27	0	0.0041	35	0	0.0047	37	0
	$\|T\|$: 50	100	0.0035	23	0	0.0067	67	0	0.0059	66	0	0.0079	82	0
Web 2001	$\|R\|$: 97	5	0.0008	2	0	0.0011	2	0	0.0009	2	0	0.0009	2	0
	$\|O\|$: 29	10	0.0027	11	0	0.0020	11	0	0.0020	11	0	0.0018	11	0
	$\|R_p\|$: 35	20	0.0028	12	0	0.0030	12	0	0.0026	13	0	0.0028	17	0
	d : 100	30	0.0028	37	0	0.0036	44	0	0.0024	28	0	0.0034	34	0
	$\|T\|$: 50	100	0.0039	42	0	0.0082	89	0	0.0037	40	0	0.0070	73	0
Web 2002	$\|R\|$: 69	5	0.0075	88	0	0.0044	52	0	0.0043	53	0	0.0043	53	0
	$\|O\|$: 16	10	0.0038	41	0	0.0047	50	0	0.0043	50	0	0.0052	56	0
	$\|R_p\|$: 60	20	0.0048	78	0	0.0046	74	0	0.0036	49	0	0.0048	73	0
	d : 50	30	0.0041	60	0	0.0055	90	0	0.0035	53	0	0.0054	76	0
	$\|T\|$: 50	100	0.0010	35	0	0.0048	126	0	0.0019	52	0	0.0034	95	0
Genomics 2005	$\|R\|$: 62	5	0.0113	99	0	0.0056	70	0	0.0055	72	0	0.0061	72	0
	$\|O\|$: 32	10	0.0045	71	0	0.0072	122	0	0.0068	118	0	0.0082	129	0
	$\|R_p\|$: 46	20	0.0034	55	0	0.0074	82	0	0.0070	85	0	0.0091	101	0
	d : 60	30	0.0040	69	0	0.0077	90	0	0.0074	103	0	0.0106	148	0
	$\|T\|$: 49	100	0.0041	87	0	0.0052	97	0	0.0054	115	0	0.0094	176	0
Robust 2005	$\|R\|$: 74	5	0.0107	3	0	0.0189	15	2	0.0334	26	4	0.0249	18	4
	$\|O\|$: 17	10	0.0123	6	0	0.0239	14	6	0.0381	25	7	0.0303	19	10
	$\|R_p\|$: 18	20	0.0137	9	2	0.0269	17	6	0.0383	31	11	0.0361	29	11
	d : 55	30	0.0160	8	1	0.0284	18	5	0.0394	33	10	0.0417	30	10
	$\|T\|$: 50	100	0.0121	19	4	0.0154	26	0	0.0210	30	4	0.0300	35	4
Legal 2006	$\|R\|$: 34	5	0.0336	21	2	0.1355	41	3	0.1123	44	4	0.1533	45	4
	$\|O\|$: 8	10	0.0343	27	0	0.1557	22	0	0.1131	44	6	0.1757	47	9
	$\|R_p\|$: 14	20	0.0438	37	3	0.1126	34	7	0.0566	40	6	0.0878	46	10
	d : 10	30	0.0327	39	2	0.0707	26	0	0.0377	41	4	0.0585	44	5
	$\|T\|$: 39	100	0.0107	26	13	0.0215	40	16	0.0113	27	13	0.0176	37	19
Medical 2011	$\|R\|$: 127	5	0.0243	112	0	0.0300	130	0	0.0366	141	0	0.0552	209	0
	$\|O\|$: 29	10	0.0270	118	0	0.0283	113	1	0.0414	157	4	0.0688	263	12
	$\|R_p\|$: 46	20	0.0144	57	0	0.0201	89	0	0.0216	87	0	0.0351	167	0
	d : 10	30	0.0103	78	0	0.0151	82	0	0.0144	88	0	0.0234	117	0
	$\|T\|$: 34	100	0.0035	56	0	0.0063	94	1	0.0042	61	0	0.0069	103	0
Microblog 2011	$\|R\|$: 184	5	0.0211	95	3	0.0112	23	0	0.0120	23	0	0.0115	23	0
	$\|O\|$: 58	10	0.0076	47	0	0.0125	51	0	0.0129	48	0	0.0130	51	0
	$\|R_p\|$: 49	20	0.0087	42	0	0.0139	67	0	0.0137	50	0	0.0150	70	0
	d : 30	30	0.0102	70	0	0.0162	90	0	0.0159	79	0	0.0180	96	0
	$\|T\|$: 49	100	0.0044	53	0	0.0047	58	0	0.0044	48	0	0.0048	58	0
	top performer		45	37	9	18	16	3	7	7	0	2	0	0
	worst performer		15	16	6	9	6	1	12	11	0	37	32	9

of being or not being in the pool. Thereby, the error introduced is bigger than the one we would have observed if we had not corrected the run at all. In fact we can observe that the second best result, but for Genomics 2005, excluding the Lipani approach, is obtained by the Reduced Pool. Therefore to mitigate this effect, it is necessary to introduce a trigger that checks if the number of data points collected are sufficient to compute the estimate and perform the correction, and if not, to fall back to the Reduced Pool error.

5. CONCLUSION

We have presented a new approach, empirically justified, to correct the pool bias for the evaluation measure $P@n$. The method is based on new insights on the fixed-depth at n pooling strategy, namely the observation of a relation between the probability of a run to provide uniquely identified relevant documents to the pool and the log-normal distribution. The proposed method largely outperformed the previous approaches, although for two test collections, Ad Hoc 7 and Ad Hoc 8, where the Lipani estimator works better. We identified two hypotheses that would explain the worst performing cases, both of which tracing their source to situations where, for the metric at hand, the vast majority of runs have all of their documents judged in the reduced pool. A cross comparison between the algorithm of the Lipani estimator and the algorithm presented here shows some similarities which may inspire a new improvement on the correction of the pool bias.

Acknowledgments

This research was partially supported by the Austrian Science Fund (FWF) project number P25905-N23 (ADmIRE), and by the Google Faculty Research Award scheme. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

6. REFERENCES

[1] Charles L. A. Clarke and Mark D. Smucker. Time well spent. In *Proc. of IIiX*, 2014.

[2] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In *Proc. of SIGIR*, 1998.

[3] Karen Sparck Jones. Letter to the editor. *Information Processing & Management*, 39(1), 2003.

[4] Aldo Lipani. Fairness in information retrieval. In *Proc. of SIGIR*, 2016.

[5] Aldo Lipani, Mihai Lupu, and Allan Hanbury. Splitting water: Precision and anti-precision to reduce pool bias. In *Proc. of SIGIR*, 2015.

[6] Aldo Lipani, Mihai Lupu, and Allan Hanbury. The curious incidence of bias corrections in the pool. In *Proc. of ECIR*, 2016.

[7] Aldo Lipani, Guido Zuccon, Mihai Lupu, Bevan Koopman, and Hanbury Allan. The impact of fixed-cost pooling strategies on test collection bias. In *Proc. of ICTIR*, 2016.

[8] Alistair Moffat, William Webber, and Justin Zobel. Strategic system comparisons via targeted relevance judgments. In *Proc. of SIGIR*, 2007.

[9] K. Spärck Jones and C. J. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. *British Library Research and Development Report No. 5266*, 1975.

[10] William Webber and Laurence A. F. Park. Score adjustment for correction of pooling bias. In *Proc. of SIGIR*, 2009.