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 (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Link to publication

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
The Solitude of Relevant Documents in the Pool

Aldo Lipani1   Mihai Lupu1   Evangelos Kanoulas1   Allan Hanbury1
1Inst. of Software Technology & Interactive Systems, Vienna University of Technology, Vienna, Austria {lipani, lupu, hanbury}@ifs.tuwien.ac.at
2Informatics 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].

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...
relevant: as the probability of the non-judged documents of the run to be uniquely identified relevant documents discovered by pooling the 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 \( \delta \bar{P} \bar{n} \), 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 \( \delta \bar{P} \bar{n} \) is not 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 \( X \):

\[ e^{E[\ln(X)]} = e^{E[Y]} = \sqrt[n]{\prod_{y \in Y} y} = GM[X] \]  

where \( GM \) is the geometric mean. Thereby, the correction coeffi-
of the run. In fact, one could generate an unbounded number of
consequently, the fraction in Eq. (5) is well defined. Second, such
because: first, if \( \delta P \) can be shown that
\( \delta P \) gets divided by the number of uniquely identified docu-
ments 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 \( \delta P \) is zero. This is reasonable
because: first, if \( \delta P \neq 0 \) it can be shown that \( \delta P \neq 0 \) 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 \( \delta P = 0 \).
In Alg. 1 we make use of Eq. (5), filtering out 0 values, and redefin-
ing the previous Webber approach. Two benefits of this modification
are that no longer every correction is based on a priori probabili-
ty 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 \( k \otimes n \) and of \( P \otimes n \).

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 spe-
cific 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 or-
ganization. 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 dif-
ference in ranks between the true ranking and new ranking with us-
ing 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 Re-
duced 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
approach is top performer in the majority of the cases (45 for MAE,
37 for SRE and, 9 for SRE\(^*\)). MAE measures the mean of the absolute difference of
runs, 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|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, ties not included.

<table>
<thead>
<tr>
<th>Track</th>
<th>F1@10</th>
<th>Our MRE</th>
<th>Lipman MRE</th>
<th>Webber MRE</th>
<th>Reduced Pool MRE</th>
<th>Li 1988</th>
<th>Lipman 1988</th>
<th>Webber 1988</th>
<th>Reduced Pool 1988</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad Hoc 1</td>
<td>0.026</td>
<td>0.026</td>
<td>0.032</td>
<td>0.032</td>
<td>0.042</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>Ad Hoc 2</td>
<td>0.026</td>
<td>0.026</td>
<td>0.032</td>
<td>0.032</td>
<td>0.042</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>Ad Hoc 3</td>
<td>0.026</td>
<td>0.026</td>
<td>0.032</td>
<td>0.032</td>
<td>0.042</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>Ad Hoc 4</td>
<td>0.026</td>
<td>0.026</td>
<td>0.032</td>
<td>0.032</td>
<td>0.042</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>Ad Hoc 5</td>
<td>0.026</td>
<td>0.026</td>
<td>0.032</td>
<td>0.032</td>
<td>0.042</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>Ad Hoc 6</td>
<td>0.026</td>
<td>0.026</td>
<td>0.032</td>
<td>0.032</td>
<td>0.042</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
</tr>
</tbody>
</table>

5. CONCLUSION

We have presented new insights, empirically justified, to correct the pool bias for the evaluation measure F1@10. 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 Lipman estimator works better. We identified two hypotheses that would explain why this may happen, 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 Lipman 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