On the reusability of open test collections

Hashemi, S.H.; Clarke, C.L.A.; Dean-Hall, A.; Kamps, J.; Kiseleva, J.

DOI
10.1145/2766462.2767788

Publication date
2015

Document Version
Final published version

Published in
SIGIR 2015

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

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 426, 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)
ABSTRACT
Creating test collections for modern search tasks is increasingly more challenging due to the growing scale and dynamic nature of content, and need for richer contextualization of the statements of request. To address these issues, the TREC Contextual Suggestion Track explored an open test collection, where participants were allowed to submit any web page as a result for a personalized venue recommendation task. This prompts the question on the reusability of the resulting test collection: How does the open nature affect the pooling process? Can participants reliably evaluate variant runs with the resulting qrels? Can other teams evaluate new runs reliably? In short, does the set of pooled and judged documents effectively produce a post hoc test collection? Our main findings are the following: First, while there is a strongly significant rank correlation, the effect of pooling is notable and results in underestimation of performance, implying the evaluation of non-pooled systems should be done with great care. Second, we extensively analyze impacts of open corpus on the fraction of judged documents, explaining how low recall affects the reusability, and how the personalization and low pooling depth aggravate that problem. Third, we outline a potential solution by deriving a fixed corpus from open web submissions.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Query formulation, Search process, Selection process
General Terms: Algorithms, Measurement, Experimentation

1. INTRODUCTION
Controlled test collections remain crucial for evaluation and tuning of retrieval systems, both for offline testing in industry and for public benchmarks in academia. The TREC Contextual Suggestion Track experimented with an open test collection, where participants were allowed to submit any web page result for a personalized venue recommendation task. This option proved exceedingly popular amongst participants of the track, e.g., in 2014 the track received 25 permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI http://dx.doi.org/10.1145/2766462.2767788

open web submissions against 6 runs based on ClueWeb12. We focus here exclusively on the open web submissions, and investigate the reusability of the resulting open web test collection. There are at least three factors that may impede the reusability of the resulting test collection. First, the open nature may result in little to no overlap between the submissions, frustrating the pooling effect and limiting its evaluation power. Second, the track includes personalization of results to a specific user profile, hence a “topic” consists of the main statement of request (in this case a North American city) and a profile of the requester. Third, the resulting pooling depth over submissions per topic (i.e., a unique context and profile pair) are limited to rank 5. It is well known that low pool depth affects reusability [7]. The key factor in case of sparse judgments is the presence or absence of pooling bias [1].

In this paper, our main aim is to study the question: How reusable are open test collections? Specifically, we answer the following research questions:

1. How does the open nature affect the evaluation of non-pooled systems?
   (a) What is the effect of leave out uniques on the score and ranking over all systems?
   (b) What is the effect of leave out uniques on the score and ranking over top ranked systems?

2. How does the open nature affect the fraction of judged documents?
   (a) What is the fraction of judged documents over ranks?
   (b) What is the effect of personalization on the fraction of judged documents?

2. EXPERIMENTAL DATA
The TREC Contextual Suggestion Track asks participants to submit venue recommendations (in the form of a valid URL). We give some statistics of the open web submissions in 2014. There were a total of 25 submissions by 14 teams (with 11 teams submitting 2 runs). A topic consists of a pair of both a context (a North American city) and a profile of the requester (consisting of likes and dislikes of venues in another city). For example, to recommend venues to visit in the unknown city of Buffalo, NY, based on a profile with ratings of attractions in Chicago, IL. Runs were pooled at depth 5 and in total 299 context-profile pairs were judged, with an average of 28.2 unique judged venues per pair, hence 8,441 judgments in total. Details of the run and their P@5 scores are shown later in Table 2.
In a leave-one-run-out (LORO) experiment, we exclude a run had not contributed to the pool of judged documents. We also measure the effect on the runs’ scores and ranking over all systems? 

Moreover, difference of actual P@5, MAP and bpref and the ones based on LORO test is shown in Figure 1. As it is shown in this figure, average difference of MAP is 0.36 which is much higher than the ones reported for reusable test collections (e.g., from 0.5 to 2.2 [4]). Figure 1 shows that bpref is a more reliable metric in comparison to (mean average) precision.

Leave One Team Out.

The LORO experiment can be biased in case teams’ submit closely related runs containing many mutual venues. In reality, a non-pooled system might use completely different collection than the ones used by the pooled runs. Hence, we also conduct a leave-one-team-out (LOTO) experiment. Figure 2 demonstrates the same pattern as observed above for the LORO experiment, with somewhat lower rank correlations, and larger differences in scores. Again, bpref remains the most stable of the three measures.

3. IMPACT ON REUSABILITY

This section studies the reusability of the test collection, aiming to answer our first research question: How does the open nature affect the evaluation of non-pooled systems?

3.1 Leave Out Uniques Analysis

We first look at the question: What is the effect of leave out uniques on the score and ranking over all systems? Specifically, we perform both the leave-one-run-out [4] and leave-one-team-out [1] experiments to see what would have happened if a run had not contributed to the pool of judged documents. We also measure the effect on the runs’ scores as well as their system ranking—as the main goal of a test collection is to determine the system ranking rather than absolute scores. We also reusability of the test collection is evaluated based on the mentioned metrics. The Kendall’s τ of this experiment based on P@5, MAP and bpref metrics are much lower than 0.9 that is the threshold usually considered as the correlation of two effectively equivalent rankings [1].

Moreover, difference of actual P@5, MAP and bpref and the ones based on LORO test is shown in Figure 1. As it is shown in this figure, average difference of MAP is 0.36 which is much higher than the ones reported for reusable test collections (e.g., from 0.5 to 2.2 [4]). Figure 1 shows that bpref is a more reliable metric in comparison to (mean average) precision.
### 3.2 Top Ranked Systems

The leave out uniques experiments give a clear call to caution on the reuse of the open web judgments, but we observe in the scatter plots that the top ranked runs seem to fare slightly better. Hence, we look at the question: What is the effect of leave out uniques on the score and ranking over top ranked systems? We look both at Kendall’s τ and the τsig, which only consider significant inversions [3]. We also look at bias, which is the fraction of all significant pairs that are significant inversions [3]. We use a paired Student’s t-test with α = 0.05 to find significant inversions (i.e., p < α). Table 1 reports the more critical LOTO test. Over all runs, we see that τsig is somewhat better than τ but still low enough to be very careful with using the resulting test collection for evaluating non-pooled runs. Over the top ranked systems (based on P@5 as reported in Table 2), the bias, τ and τsig correlations are substantially better.

In this section we looked at the leave out uniques analysis for the open test collection in both leave run and leave team out experiments. The outcome is mixed at best, while there is a strongly significant rank correlation, the effect of pooling is notable, and results in underestimation of score and hence affects the ranking. Although we observe a somewhat more reliable evaluation of the better scoring systems, this means that the judgments should be used with caution, and evaluating non-pooled systems requires great care.

### 4. Impact on Judged Documents

This section studies in more detail the factors contributing to the observed low reusability, trying to answer our second research question: How does the open nature affect the fraction of judged documents?

#### 4.1 Fraction of Judged Documents

We first look at the question: What is the fraction of judged documents over ranks? We define Overlap@N as the fraction of the top – N suggestions that is judged for the given set of topics:

$$\text{Overlap@N}(C, P) = \frac{1}{N} \sum_{(c,p) \in (C,P)} \frac{\#\text{Judged}(c,p)}{N},$$

where #Judged(c,p) corresponds to the count of judged suggestions for the given context and profile pair (c,p) in the top-N suggestions, and (C,P) is a set of judged context and profile pair. Table 2 shows the overlap@N of runs submitted to the contextual suggestion track in 2014. We see a significant drop after the pooling cut-off at rank 5, signaling that the recall base may be incomplete and the overlap between the different runs is relatively low. Clearly the lack of a fixed collection will have contributed to this.

In order to investigate the relation of the fraction of judged pages with the pooled runs’ effectiveness, we plot Overlap@N vs. P@5 (i.e. the main official metric in this track) in Figure 3. Points in the graph represent pooled runs. Arguably, evaluating the best runs reliably is more important than separating the blatant failures. As it is shown in Figure 3 runs having higher P@5 usually have higher Overlap@N. This explains why for the evaluation is more reliable for the better performing runs. This figure also shows that runs that are outliers in terms of low fractions of judged documents. These two runs did usually provide fewer than 5 venues for the given topics.

<table>
<thead>
<tr>
<th>Run</th>
<th>Overlap@N (%)</th>
<th>P@5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BJUTa</td>
<td>100.00</td>
<td>61.43</td>
</tr>
<tr>
<td></td>
<td>59.36</td>
<td>32.38</td>
</tr>
<tr>
<td></td>
<td>20.65</td>
<td>20.23</td>
</tr>
<tr>
<td></td>
<td>100.00</td>
<td>59.43</td>
</tr>
<tr>
<td></td>
<td>59.43</td>
<td>28.86</td>
</tr>
<tr>
<td></td>
<td>16.19</td>
<td>14.25</td>
</tr>
<tr>
<td></td>
<td>97.08</td>
<td>57.52</td>
</tr>
<tr>
<td></td>
<td>31.47</td>
<td>16.19</td>
</tr>
<tr>
<td></td>
<td>95.33</td>
<td>55.93</td>
</tr>
<tr>
<td></td>
<td>27.62</td>
<td>17.35</td>
</tr>
<tr>
<td></td>
<td>10.99</td>
<td>49.36</td>
</tr>
<tr>
<td></td>
<td>22.81</td>
<td>42.21</td>
</tr>
<tr>
<td></td>
<td>44.88</td>
<td>0.290</td>
</tr>
<tr>
<td></td>
<td>0.480</td>
<td>0.154</td>
</tr>
<tr>
<td></td>
<td>0.393</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>0.111</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Table 1: Reusability in top of the ranking**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Depth</th>
<th>P@5 MAP bpref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendall τ</td>
<td>✓</td>
<td>0.800 0.800 0.000</td>
</tr>
<tr>
<td>Kendall τsig</td>
<td>✓</td>
<td>1.000 0.777 1.000</td>
</tr>
<tr>
<td>Bias</td>
<td>✓</td>
<td>0.000 0.111 0.000</td>
</tr>
</tbody>
</table>

**Table 2: Overlap@N and P@5 of each pooled open web run based on the official TREC judgments**

**Figure 3: Overlap@N versus P@5 for open web runs.**

829
5. CONCLUSIONS

We have studied reusability of the TREC 2014 Contextual Suggestion open test collection in terms of the reusability of the judgments to evaluate non-pooled runs and in terms of fraction of judged venues. We analyzed the effectiveness of the pool for building a reusable test collection. Experimental results of leave out uniques (i.e., run or a team) tests based on various metrics, including Kendall’s $\tau$, AP correlation and average difference, showed that the test collection should be used with extreme care: non-pooled systems tend to be underestimated. However, for the high quality runs (i.e., top-5 of the ranking), the test collection performs somewhat better and had the highest correlation with the official ranking in terms of the $\tau$ based on significant inversions. Our empirical investigation has also shown that using an open collection tends to produce a diverse pool and consequently less fraction of judged venues at ranks deeper than the pool cut-off (e.g., only 16% overlap at ranks between 6 and 10). In addition, we looked at the role of personalization and low pooling depth, and showed that the lenient profile-ignorant fractions of judged page leads to considerable larger fractions of judged documents.

Our general observation is that the open collection leads to significantly lower recall, and low fraction of judged results, over individual runs. There are several ways in which this could be addressed. First, it is still an open question on whether we can derive a post hoc corpus and test collection from the open web submissions, by constructing a corpus based on the combined retrieved pages, and use this to evaluate runs over the combined set. We have done an initial analysis of this approach showing promising results. Second, the organizers of the TREC 2015 contextual suggestion track aim to collect open web results as a pre-task in early 2015, and use these submissions to construct a fixed open web collection shared to all track participants. The results of this paper give support to the creation of a fixed collection of open web results, and suggest that this will substantially increase the reusability of the benchmark for non-pooled runs in follow up experiments.

Acknowledgments This research is party funded by the European Community’s FP7 (project meSch, grant # 600851).

References