On the Reusability of Open Test Collections

Seyyed Hadi Hashemi[1] Charles L.A. Clarke[2] Adriel Dean-Hall[2] Jaap Kamps[1] Julia Kiseleva[3]

[1]University of Amsterdam, Amsterdam, The Netherlands
[2]University of Waterloo, Waterloo, Canada
[3]Eindhoven University of Technology, Eindhoven, The Netherlands

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 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.

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.

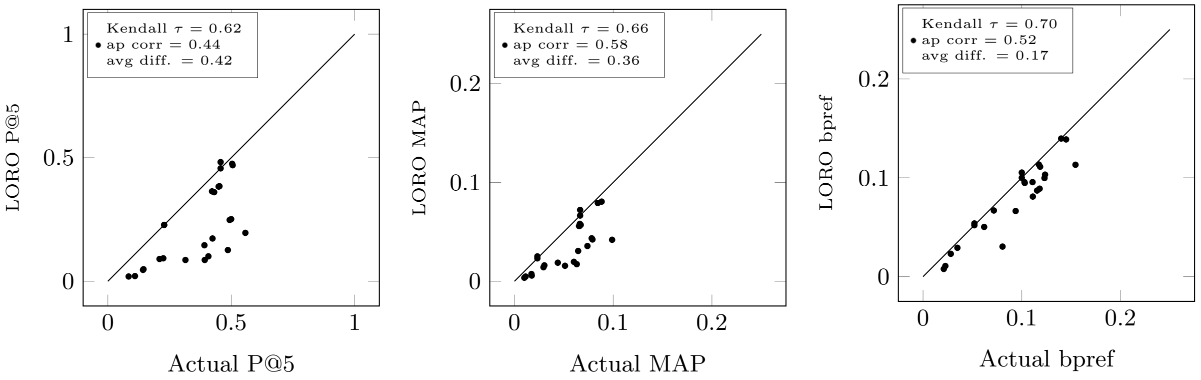

Figure 1: Difference in P@5, MAP, and bpref based on the leave one run out (LORO) test.

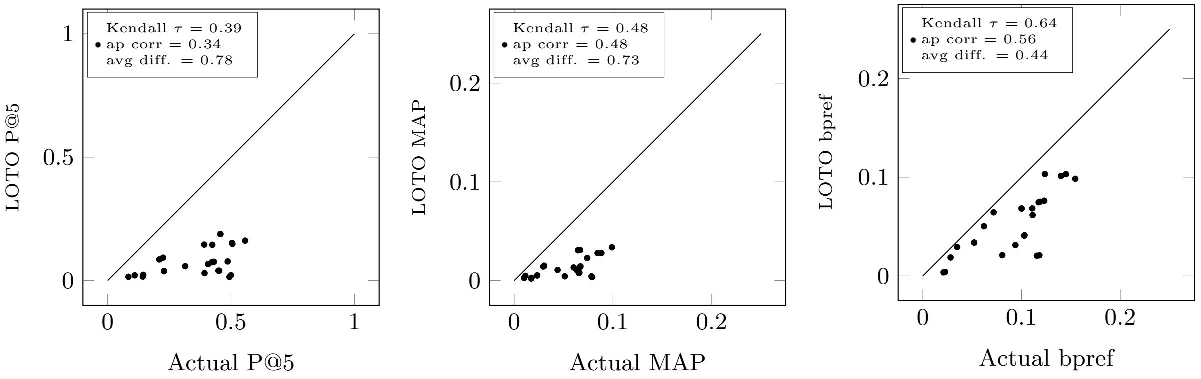

Figure 2: Difference in P@5, MAP, and bpref based on the leave one team out (LOTO) test.

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 [7] 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. The standard system rank correlation measure in IR research is Kendall's τ (i.e. $\tau = \frac{C-D}{N(N-1)/2}$), where C is the number of concordant pairs, D is the number of discordant pairs, and N is the number of systems in the given two rankings [6]. However, there are a number of researches studied that the Kendall's τ is not promising in some conditions [2, 3, 6]. In order to more precisely measure the test collection reusability, we also use AP Correlation Coefficient (i.e., $\tau_{AP} = \frac{2}{N-1} \cdot \sum_{i=2}^{n} (\frac{C(i)}{i-1}) - 1)$, where $C(i)$ is the number of systems above rank i and correctly ranked [6].

Leave One Run Out.

In a leave-one-run-out (LORO) experiment, we exclude a pooled run's unique judgments from the test collection, and evaluate the run based on the new test collection in terms of P@5, MAP, or bpref metrics. This test is done for all of the pooled runs—hence for each run we obtain the score as if it had not been pooled and judged. Then, the ranking correlation of the official ranking of runs with the new one is estimated. In Figure 1, 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 [4].

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 [1, 5]). 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.

Table 1: Reusability in top of the ranking

Metric	Depth 5	 All	P@5	MAP	bpref
Kendall τ	✓		0.800	0.800	0.000
Kendall τ_{sig}	✓		1.000	0.777	1.000
Bias	✓		0.000	0.111	0.000
Kendall τ		✓	0.393	0.480	0.646
Kendall τ_{sig}		✓	0.418	0.572	0.691
Bias		✓	0.290	0.213	0.154

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 τ_{sig}, 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 $\alpha = 0.05$ is used to find significant inversions (i.e., $p < \alpha$). Table 1 reports the more critical LOTO test. Over all runs, we see that τ_{sig} 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 τ_{sig} 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:

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

where $\#Judged@N(\langle c, p\rangle)$ corresponds to the count of judged suggestions for the given context and profile pair $\langle c, p\rangle$ in the top-N suggestions, and $\langle C, P\rangle$ 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.

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

Run	Overlap@N (%) N=5	 N=10	 N=25	 N=50	P@5 (%)
BJUTa	100.00	61.43	32.38	20.65	50.57
BJUTb	100.00	60.26	32.10	20.23	50.37
BUPT_PRIS_01	44.88	23.24	09.36	04.68	14.45
BUPT_PRIS_02	47.02	25.21	10.27	05.13	14.25
cat	99.93	59.06	31.90	19.83	20.94
choqrun	97.85	57.52	31.47	16.19	22.47
dixlticmu	100.00	59.49	32.33	21.46	39.13
gw1	97.99	51.97	24.98	14.21	10.99
lda	100.00	53.57	24.73	14.31	08.43
RAMARUN2	100.00	57.99	27.78	15.53	49.97
run_DwD	99.53	61.00	35.30	24.68	31.44
run_FDwD	99.59	79.79	37.61	23.68	42.41
RUN1	99.93	58.56	28.58	16.00	49.36
simpleScore	100.00	58.82	28.60	16.25	44.88
simpleScoreImp	100.00	59.43	28.86	16.34	45.22
tueNet	99.86	52.64	23.90	14.33	22.81
tueRforest	99.86	52.64	23.90	14.33	22.81
UDInfoCS2014_1	100.00	57.45	28.64	17.35	40.74
UDInfoCS2014_2	100.00	59.36	31.41	19.83	55.72
uogTrBunSumF	100.00	55.75	22.38	11.20	48.63
uogTrCsLtrF	100.00	55.75	27.30	16.40	39.26
waterlooA	99.79	64.28	31.10	19.67	42.21
waterlooB	99.79	59.53	30.50	19.36	43.08
webis_1	98.59	56.75	27.50	15.16	45.69
webis_2	98.59	56.75	27.50	15.16	45.69
Average	95.33	55.93	27.62	16.48	36.06

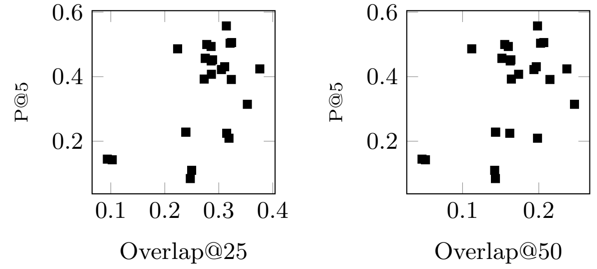

Figure 3: Overlap@N versus P@5 for open web runs.

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 two 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.

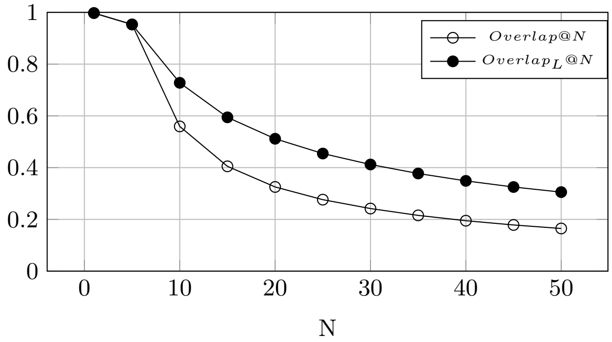

Figure 4: Effect of lenient judgments on Overlap@N.

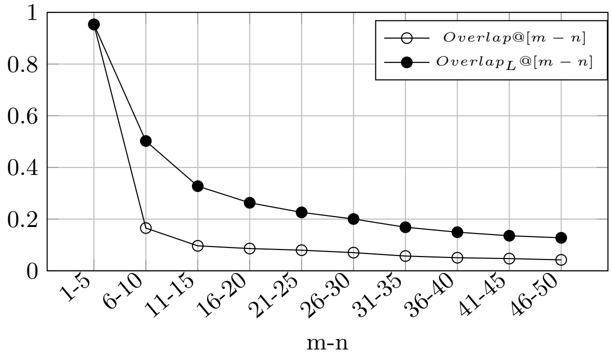

Figure 5: Overlap@N over rank intervals.

4.2 Impact of Personalization

We now look at the question: *What is the effect of personalization on the fraction of judged documents?* Specifically, we exploit the fact that contexts (i.e., cities) are judged for multiple profiles of the same (and other) submissions: in the case that the relevance of a venue to the given context is not judged for the given profile, judgments made for other profiles will be used. We define *Lenient Overlap* (i.e., Overlap$_L$@N) that is an instance of Overlap@N, in which $\#Judged@N$ is calculated by ignoring profile assumption. The results are shown in Figure 4, which shows that ignoring the exact profile substantially improves the fraction of judged pages.

To highlight the number of judged pages after the pooling depth, we show the same data in an interval level analysis in Figure 5. Obviously, for pooled runs, the Overlap@5 is guaranteed to be 1, making Overlap@10 guaranteed to be at least 0.5, etc. This shows the drop in fraction of judged paged for the personalized runs in an even more dramatic way. The lenient profile-ignorant overlap measure however remains more stable over the intervals. This signals that the relatively low fractions of judged pages can be attributed for some part to the low pool depth and personalization, rather than the open nature of the test collection.

This section looked at the fraction of judged pages in the open web submissions. The outcome clearly show the low recall: after the pooling depth the fraction plummets down, explaining the relatively low reusability of the open web judgments. We looked in the relative contribution of the open nature of the collection and the personalization and pool depth, which suggested that the latter play a major role in explaining the low fraction of overlap.

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 τ, 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 τ 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

[1] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Information retrieval*, 10(6):491–508, 2007.

[2] B. Carterette. On rank correlation and the distance between rankings. In *SIGIR*, pages 436–443, 2009.

[3] G. V. Cormack and T. R. Lynam. Power and bias of subset pooling strategies. In *SIGIR*, pages 837–838, 2007.

[4] E. M. Voorhees. Evaluation by highly relevant documents. In *SIGIR*, SIGIR '01, pages 74–82. ACM, 2001.

[5] E. M. Voorhees, J. Lin, and M. Efron. On run diversity in evaluation as a service. In *SIGIR*, pages 959–962, 2014.

[6] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR*, pages 587–594, 2008.

[7] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *SIGIR*, pages 307–314, 1998.