Proceedings of the Seventh International Workshop on Evaluating Information Access (EVIA 2016),
a Satellite Workshop of the NTCIR-12 Conference, June 7, 2016 Tokyo Japan

An Easter Egg Hunting Approach to Test Collection Building in Dynamic Domains

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

Test collections for offline evaluation remain crucial for information retrieval research and industrial practice, yet the classical Sparck Jones and Van Rijsbergen approach to test collection building based on the pooling of runs on a large collection is expensive and being pushed beyond its limits with the ever increasing size and dynamic nature of the collections. We experiment with a novel approach to reusable test collection building, where we inject judged pages into an existing corpus, and have systems retrieve pages from the extended corpus with the aim to create a reusable test collection. In a metaphorical way, we hide the Easter eggs for systems to retrieve. Our experiments exploit the unique setup of the TREC Contextual Suggestion Track, which allowed both submissions from a fixed corpus (ClueWeb12) as well as from the open web. We conduct an extensive analysis of the reusability of the test collection based on ClueWeb12, and find it too low for reliable offline testing. Then, we detail the expansion with judged pages from the open web, and do extensive analysis on the reusability of the resulting expanded test collection, and observe a dramatic increase in reusability. Our approach offers novel and cost effective ways to build new test collections, and to refresh and update existing test collections. This explores new ways of effective maintenance of offline test collections for dynamic domains such as the web.

1. INTRODUCTION

Evaluation in IR builds on over 50 years of tradition in test collection building, starting from the first large scale experimental evaluations of retrieval effectiveness of various indexing languages for literature at Cranfield [7]. Test collections remain crucial for experimental IR in academia, and for offline evaluation based on editorial judgments in industry. But the test collection approach to IR evaluation is under threat by the fast changing pace of information access, presenting new tasks, new types of data, at an unprecedented scale and intensity. All recent IR research agendas [1, 4, 14] seek ways to embrace these new challenges, while still retaining the advantages of experimental control in the Cranfield/TREC paradigm. One particular challenge is to deal with the dynamic nature of the web and other online sources [19].

This paper is motivated by the TREC Contextual Suggestion track, investigating search techniques for complex information needs that are highly dependent on context and user interests [22]. It offers a personalized venue recommendation task based on a U.S. city as context, and crowdsourced profiles and judgments. The track suffered from the delayed availability of the ClueWeb12 collection, and decided to use no static corpus of documents but accept any page on the web in 2012. In the following years, the track used ClueWeb12 (consisting of 733,019,372 English web pages) but kept on allowing open web results by popular request of the track's participants. This unique setup of the contextual suggestion track leads to two distinct sets of judgments: one set consists of judgments of documents contributed by open web runs, and the other one includes judgments of ClueWeb12 documents provided by ClueWeb12 runs [10].

The open web-based test collection, which includes the majority of the judgments (i.e., 25 out of 31 pooled runs in 2014), is not reusable [12]. In fact, the open nature of the test collection leads to limited overlap between the open web submissions, and reduces the reusability of the test collection. This fact raises two questions: Is the ClueWeb12-based contextual suggestion test collection using a fixed corpus reusable? If not, is it possible to reuse the open web judgments to build a new corpus in order to create a more reusable test collection?

In this paper, our main aim is to study the question: *Can we build a reusable test collection for a dynamic domain by injecting judged documents into a test collection with sparse judgments?* Specifically, we answer following research questions:

1. *How reusable is the ClueWeb12 test collection of the TREC contextual suggestion?*
   (a) *How reusable is the test collection for evaluating non-pooled systems?*
   (b) *What is the fraction of judged documents?*
   (c) *What is the impact of personalization on the fraction of judged documents?*
2. *How to expand a test collection in order to improve its reusability?*
3. *How reusable is the expanded test collection containing judged open web documents?*
   (a) *How reusable is the expanded test collection for ranking systems?*
   (b) *Are retrieval models able to retrieve the judged open web documents?*

In this paper, we first investigate the reusability of the ClueWeb12 test collection of TREC Contextual Suggestion track. Then, we propose a novel approach to expand the ClueWeb12 test collection making use of the open web judgments, and investigate the reusability of the expanded test collection. Our main contribution is a novel approach in building or updating test collections by injecting externally judged documents. This approach can be used to expand test collections having incomplete or imperfect set of judgments [e.g., 2], or update test collections for dynamic domains that have become outdated [e.g., 19]

The rest of this paper is organized as follows. In Section 2, we review some related work on reusable test collection building and reusability tests. Section 3 is devoted to reusability evaluation of the ClueWeb12 test collection. Our proposed test collection building approach is detailed in Section 4, and its reusability is thoroughly evaluated in Section 5. Finally, we present the conclusions and future work in Section 6.

2. RELATED WORK

In this section, we will discuss related work on test collection construction. At TREC, the National Institute of Standard and Technology (NIST) uses the classical Sparck Jones and Van Rijsbergen [21] pooling technique in order to create test collections for the comparative evaluation of retrieval systems. The idea behind pooling is that documents retrieved by a run in ranks deeper than the pool cut-off, is likely retrieved by another run inside the pool. The reusability of the resulting test collection depends on the completeness of the relevance judgments. Therefore, identifying an effective pool depth for building reusable test collections become an important issue. Hence, Zobel [27] studied effects of pool depth on the reusability of test collections, and demonstrated that low pool depth tends to lessen reusability of the test collections.

Within the literature on building a reusable test collection based on the pooling technique, one approach is to sample a more effective set of documents as a pool of documents to be judged. Moffat et al. [15] argued that the importance of all the pooled documents are not the same in building reusable test collections that are able to comparatively rank retrieval systems. They proposed considering of relevance likelihood of documents in creating the pools. Cormack et al. [9] also proposed a move to front pooling approach, which examines documents in order of their relevance likelihood among submissions. In fact, a submission that has more recently retrieved a relevant document is assumed to more likely retrieve another relevant document. Other work focuses on creating a more effective pool by using more diverse pooled runs. To this aim, relevance feedback is used to retrieve a new set of results in order to improve the pool effectiveness [13, 20]. Moreover, in order to build a reusable test collection, Carterette et al. [6] proposed an experimental design, which collects evidence for or against three types of reusability (i.e., within-team, between team and participant comparison) during collecting judgments.

Rather than focusing on pooling itself, the current paper focuses on the problem of how to update an existing test collection with sparse judgments, in case there are new documents with judgments available. Closest in spirit to our work is Soboroff [19], who studied how the GOV2 collection becomes outdated due to the changing Web, looking the effects of pages that disappear and change, and did experiments with simulated re-judging of changed pages. Soboroff also makes the suggestion to judge new pages not included in the original corpus, but did not do any experiments on this, and the current paper addresses this head-on.

There is quite some literature on the reusability of the test collection. Leave out uniques is the standard test for evaluating reusability of test collections in ranking non-pooled systems. Leave-one-run-out is a preliminary version of this test that introduced by Zobel [27] to identify effects of missing relevant documents in evaluating non-pooled systems. Since runs contributed by a same team are similar, leaving all contributions of a team out (i.e. leave-one-team-out [3, 24]) is another reusability test, which is more critical in case teams submitted several similar runs, thereby reducing the number of uniquely retrieved documents in individual runs. Sakai [18] propose take-just-one-team and take-just-three-team experiments to identify effects of missing judgments on a number of evaluation metrics (e.g., AP and bpref).

3. TEST COLLECTION REUSABILITY

This section studies the reusability of the test collection, aiming to answer our first research question: *How reusable is the ClueWeb12 test collection of the TREC contextual suggestion?*

3.1 Experimental Data

In this paper, we use the unique setup of the TREC Contextual Suggestion track. This track allows participants to submit their venue recommendation runs' results based on either open web (in the form of a valid URL) or ClueWeb12 dataset (in the form of a valid ClueWeb12 document ID). In TREC 2014, 31 runs submitted by 17 teams (with 14 teams submitting 2 runs). Among these submissions, 6 runs belong to 3 out of 17 teams who made their submissions based on the ClueWeb12 dataset, and the rest are based on the open web.

In contextual suggestion, a topic consists of a pair of both a context (a North American city) and a profile (consisting the requester's likes and dislikes of venues in another city). For example, a requester's preferences and their ratings of attraction in Chicago, IL are used to recommend venues to visit in the new city of Buffalo, NY. Runs were pooled at depth 5 and in total 299 context-profile pairs, which has 112 judged documents in average, were judged. A short summary of the TREC 2014 contextual suggestion test collection is given in Table 1.

3.2 Leave Out Uniques Analysis

We first look at the question: *How reusable is the test collection for evaluating non-pooled systems?* Specifically, we perform both the leave-one-run-out [27] and leave-one-team-out [3] experiments to see what would have happened if a run had not contributed to the pool of judged documents.

In order to evaluate the test collection reusability in evaluating non-pooled systems, Kendall's τ, which is a standard metric in measuring system rankings correlation, is used. This metric is formulated as follows:

$$\tau = \frac{C - D}{N(N-1)/2},$$

where C is the number of concordant pairs, D is the number

Table 1: TREC 2014 Contextual Suggestion test collection statistics

Subset	#context-profile	# Venues	Depth	avg # judged documents	#Runs	#Teams
Open Web	299	8,441	5	85	25	14
ClueWeb12	299	2,674	5	27	6	3

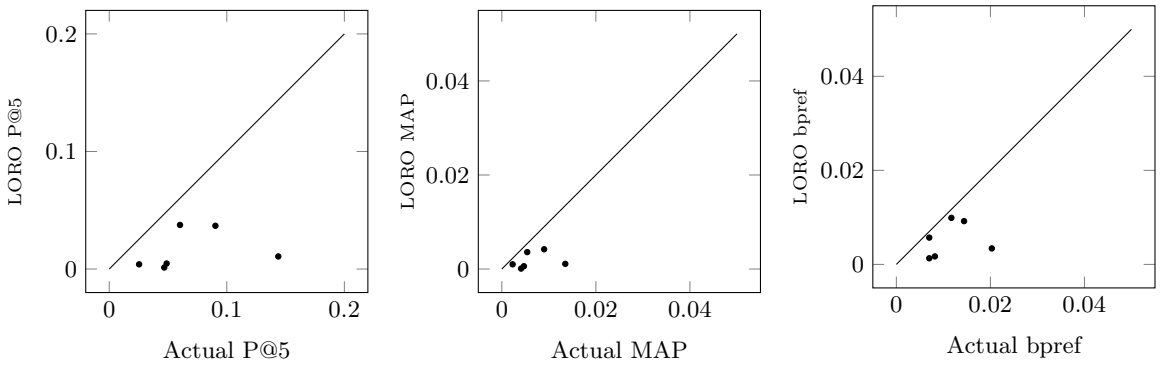

Figure 1: Difference in P@5 (Kendall τ = 0.46, ap corr = 0.11, avg diff = 0.76), MAP (τ = 0.46, ap corr = 0.41, avg diff. = 0.69), and bpref (τ = 0.20, ap corr = 0.02, avg diff = 0.52) based on the leave one run out (LORO) test.

of discordant pairs, and N is the number of systems in the given two rankings [26]. In addition to Kendall's tau, which can be overly optimistic in some conditions [5, 8, 26], AP Correlation Coefficient is used to measure system rankings' correlation more precisely. AP Correlation is formulated as follows:

$$\tau_{AP} = \frac{2}{N-1} \cdot \sum_{i=2}^{n} \left(\frac{C(i)}{i-1}\right) - 1,$$

where $C(i)$ is the number of systems above rank i and correctly ranked [26]. Moreover, the average percentage difference of common IR metrics, before and after leave the out uniques tests, will show the effect of being pooled or not on the systems' absolute scores.

Leave One Run Out In leave-one-run-out (i.e., LORO) experiment, for each pooled run, the run's unique judgments are excluded from the test collection and it is evaluated based on the new test collection in terms of P@5, MAP and bpref.[1] Our main aim in this experiment is finding the correlation of the system ranking in the case that they are pooled and judged in the test collection building process with the one ranked based on the assumption that the systems are not pooled.

As it is shown in Figure 1, leave-one-run-out system ranking's correlation with the actual system ranking is lower than the usual scores reported on reusable test collections in previous research. Specifically, Kendall's τ of the LORO experiment based on the MAP metric is 0.46, which is much lower than 0.9 that is the threshold usually considered as the correlation of two effectively equivalent rankings [23].

[1]The track uses P@5 as main measure, and also supplies MRR and a modified time-based gain (TBG) measure. As we are dealing with sparse judgments, we opt to include MAP which is know to be more stable, and bpref which is designed to be stable under incomplete judgments. Experiments (not reported) confirm that MRR is very unstable and that TBG resembles the P@5 results.

According to Figure 1, even rank correlation based on bpref metric, which works better than precision based metrics for evaluating systems based on incomplete test collections, is not acceptable. Moreover, difference between actual P@5, MAP and bpref and the ones based on LORO test indicates that scores of systems are considerably underestimated by excluding their unique judgments from the test collection. In particular, the mean of the percentage differences of MAP is 0.69, which is not reusable in comparison to the scores reported as reusable test collections (e.g., from 0.5 to 2.2 [3, 25]).

Leave One Team Out In order to study the reusability problem of the ClueWeb12 Contextual Suggestion test collection more precisely, we study a more realistic leave out uniques experiments. According to the observation made in [12], open web contextual suggestion runs submitted by each team is based on a similar or a same data collection. Therefore, leave-one-team-out (i.e., LOTO) is a better indicator of the test collection reusability in evaluating a new non-pooled run, which might use a completely different collection than the ones used by the pooled runs. According to this experiment, leaving one team's judgments out has a dramatic effect on runs' evaluation. Specifically, MAP score of 3 out of 6 runs is 0 after leaving their teams' judgments out of the test collection. In fact, P@5, MAP and bpref scores are dropping to zero or almost zero in LOTO test. Therefore, we do not plot the LOTO test to save space.

We will study the causes of the lack of reusability in the rest of the section, starting with the fraction of judgments in the runs.

3.3 Fraction of Judged Pages

We now look at the question: *What is the fraction of judged documents?* We want to find out if the ClueWeb12 contextual suggestion test collection has enough judgments for venues suggested in ranks beyond the pooling depth. To this aim, we have analyzed overlap@N [12] as the fraction of the top-N suggestions that is judged for the given set of

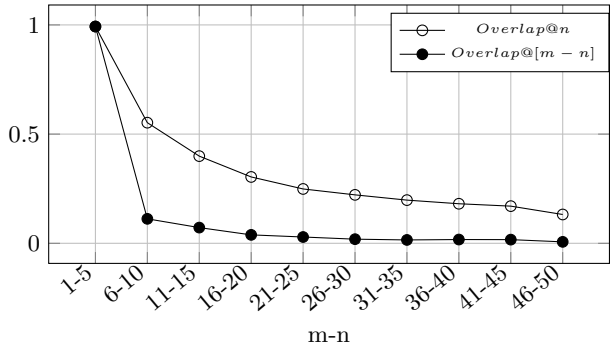

Figure 2: Overlap@N versus Overlap@N over rank intervals.

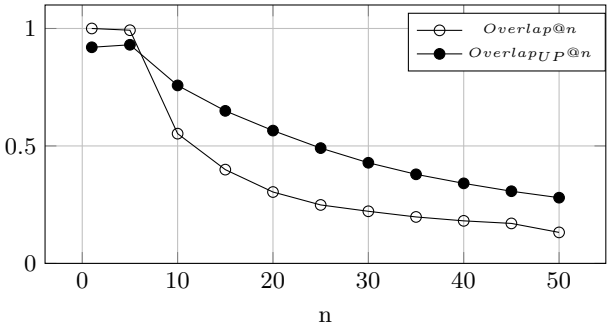

Figure 3: Effect of personalization: Overlap@N versus Overlap$_{UP}$@N based on the non-personalized test collection.

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.

According to Figure 2, the personalized test collection overlap is dropping fast after the pool cut-off, rather than declining gracefully. This observation indicates that the overlap between pooled runs is relatively low, and consequently the test collection is incomplete in terms of recall. The Overlap@N scores at lower ranks are almost completely a result from the top 5 pooled documents guaranteed to be judged, as Overlap@10 is not much higher than 0.5. Figure 2 also shows overlap at rank intervals, which shows this even more clearly: below the pool depth, the overlap at the next interval is almost zero. The low fraction of judged pages beyond the pool depth explains the low reusablity observed above.

3.4 Impact of Personalization

In this part, we answer the question: *What is the impact of personalization on the fraction of judged documents?* In the case of the TREC Contextual Suggestion's open test collection, [12] founf that personalization and shallow pool depth affected the test collection's reusability. In the case of the TREC Contextual Suggestion's ClueWeb12 test collection studied here, we may expect a similar effect of personalization. We will analyze now to what extent the personalization affects the fraction of judged documents, and hence the reusability of the resulting test collection, by "depersonalizing" the official test collection to see whether the non-personalized test collection has enough judgments for reliably ranking systems. We have used the Borda count fusion over profiles to build non-personalized runs based on the pooled personalized runs. For the evaluation purpose, any suggestion, which judged as a relevant suggestion for the given city and one of the judged profiles, is counted as relevant suggestion for the given city.

Figure 3 demonstrates that personalization has a considerable effect on the overlap, as it is causing greater diversity between the runs as well as spreading the evaluation effort thin over each profile—leading to a low pool depth.

To summarize, in this section we investigated the reusable of the TREC contextual suggestion track's ClueWeb12-based test collection. The outcome is rather negative: the system rank correlation in the leave out uniques test is below 50%, with MAP and bpref scores close to zero; the fraction of judged documents after the pooling depth plummets down; and the combination of shallow pools over personalized runs aggravates the problem considerably.

4. EXPANDING TEST COLLECTIONS

In this section, we answer the question: *How to expand a test collection in order to improve its reusability?*

4.1 Injecting Judged Documents

Our approach is rather straightforward: in case a fixed test collection becomes outdated and systems return documents not included in the outdated corpus, we simply judge the new documents, and merge them into an expanded test collection. We metaphorically hide the new documents in the old collection as Easter eggs for systems to retrieve as in an Easter egg hunt.

So assume we have a test collection based on a fixed corpus, which is not reusable. This test collection is formulated as follows:

$$TC_f = \{(t,d,r)|t:T, d:D_f, r:R_f\},$$

where t is a topic from the judged topics set (i.e., T), d is a document belongs to the fixed corpus, and r is a relevance judgment from judgments given for the fixed corpus (i.e., R_f). Moreover, consider that we have set of new pages for the same problem and a same topic set, of which some or all are judged. This second set of judged documents has a similar formulation:

$$TC_s = \{(t,d,r)|t:T, d:D_s, r:R_s\},$$

where D_s is a set of documents from the secondary collection and R_s is a set of judgment for some documents of the second corpus (which could be an open collection like the web).

In order to use the second test collection for expanding the test collection, for each document $d_1 \in D_s$, the document is injected to the fixed collection (i.e., D_f), and relevance judgments of document d_1 (i.e., $\{(t,d,r)|t:T, d == d_1, r:R_s\}$) are added to the fixed test collection (i.e., D_f). Finally, each judgment in the new test collection is an instance of the following set:

$$TC_e = \{(t,d,r)|t:T, d:D_f \cup D_s, r:R_f \cup R_s\},$$

where d is a document judged in either the fixed test collection (i.e., D_f) or the secondary test collection (i.e., D_s), and r is a relevance judgment based on either the relevance judgments set created for the fixed collection (i.e., R_f) or the secondary relevance judgments set (i.e., R_s).

4.2 Expanded Contextual Suggestion Test Collection

The unique setup of TREC Contextual Suggestion track, which is discussed in Section 3, allows us to test our approach on this test collection. To this aim, we inject the judged open contextual suggestions into a fixed contextual suggestion collection (i.e., ClueWeb12 touristic sub collection, which is provided by the TREC organizers). To be specific, the ClueWeb12 sub collection contains 176,970 documents focusing on the touristic domain, and there are 7,434 judged open web documents as candidates to be merged into this collection.

The expansion of the test collection consists of two steps: First, we determine which open web pages are also included in ClueWeb12, based on the mapping of [11]. We retain the copy of the page in ClueWeb12, as these pages tend to describe venues and still describe the same entity, although an alternative is crawl the pages and update them. The qrels are expanded with the judgments for this page. Second, for remaining open web pages, we have either fetched rest of the web pages from the web or used the touristic aggregators' websites' (e.g., Yelp) API to gather the judged web pages' textual content. These judged documents are added to the collection, and the qrels are expanded with the judgments for this page. The new qrels are substantially richer. To be specific, the contextual suggestion ClueWeb12 test collection has 8,043 judgments including 682 relevant judgments, and we add 25,407 open web judgments including 9,738 relevant judgments into that.

To summarize, in this section we investigated an approach to update or expand a test collection with a secondary set of judged pages. The general approach is to simply "hide" the judged pages in the original collection, with the goal of systems to retrieve the relevant pages amongst the rest of the collection. We applied the approach to the case of the TREC contextual suggestion track, merging the large set of judged open web pages into the ClueWeb12 based collection, leading to an updated test collection with a far greater number of judged documents.

5. REUSABILITY OF THE EXPANDED TEST COLLECTION

In this section, we look at the question: *How reusable is the expanded test collection containing judged open web documents?*

5.1 Leave Out Uniques

We evaluate reusability of the test collection by discussing the correctness of the non-pooled system ranking based on the expanded test collection. Specifically, we look at the following research question: *How reusable is the expanded test collection for ranking systems?*

In this experiment, we would like to test whether the expanded test collection is effective enough in ranking high quality runs higher than the low quality ones or not. To this aim, two groups of personalized runs are built to retrieve suggestions relevant to the given city name and profile. One of them is a group of runs based on personalized query expansion using a group of defined touristic categories (i.e., LM JM BQ, LM JM, LM two-stage, LM JM2 and LM Dir.). The other one is based on retrieving relevant suggestions to the given city name, and then ranking suggestions based on similarity of suggestions to the given profile (i.e., Okapi, tfidf, Okapi2 and tfidf2). A short summary of these runs is given in Table 2. We know that the second group of runs might miss some suggestions relevant to the given profile in case the city name is not mentioned explicitly in their contents. For example, some of the relevant suggestions might include the name of another (nearby) city rather than the city name of the context. We expect lower rank for the second group of runs in comparison to the first more effective runs.

Table 2: Personalized non-pooled runs and their descriptions. In these runs, personalization is done based on users' positive profiles

Ranker	Description
LM JM BQ	Language modeling, default JM smoothing (i.e., $\lambda = 0.4$), Boolean personalization
LM JM	Language modeling, default JM smoothing (i.e., $\lambda = 0.4$)
LM two-stage	Language modeling, default two-stage smoothing (i.e., $\mu = 2{,}500$ and $\lambda = 0.4$)
LM JM2	Language modeling, JM smoothing and $\lambda = 0.001$
LM Dir.	Language modeling, default Dirichlet smoothing (i.e., $\mu = 2{,}500$)
Okapi	Okapi, default parameters (i.e., $k_1 = 1.2$, $b = 0.75$ and $k_3 = 7$)
tfidf	tf.idf, default parameters (i.e., $k_1 = 1.2$ and $b = 0.75$)
Okapi2	Okapi, $k_1 = 0.001$, $b = 0.001$ and $k_3 = 0.001$
tfidf2	tf.idf, $k_1 = 0.001$ and $b = 0.001$

As it is shown in Table 3 (top), the expanded test collection is able to discriminate these two groups of runs, and also rank relatively reasonable within each group of runs. On the other hand, Table 3 (bottom) indicates system ranking of the same runs based on the official TREC test collection, indicating that the official test collection is not able to rank systems in a logical order. In order to test reusability of the test collection, leave out uniques test is done using LORO test. According to Figure 4, the actual system ranking is exactly same as the LORO system ranking, and they have the highest rank correlation in terms of Kendall's τ and AP correlation. Specifically, Kendall's τ and AP correlation of this test is 1, which presents the strongest possible evidence for the reusability of the expanded test collection for ranking non-pooled personalized systems.

5.2 Retrieving Judged Documents

In order to evaluate effectiveness of the expanded test collection, we study the research question: *Are retrieval models able to retrieve the judged open web documents?* Recall that

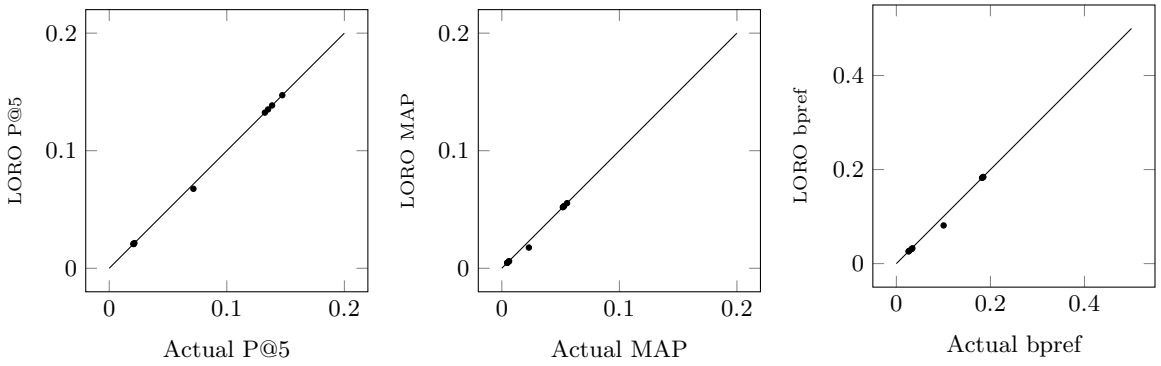

Figure 4: Difference in P@5 (Kendall τ = 1.00, ap corr = 1.00, avg diff = 0.01), MAP (τ = 1.00, ap corr = 1.00, avg diff = 0.03), and bpref (τ = 1.0, ap corr = 1.0, avg diff = 0.03) based on the leave one run out (LORO) test on the expanded test collection.

Table 3: Personalized, non-pooled system ranking based on MAP and their overlap

Run	P@5	MAP (%)	bpref	Overlap@50 (%)
Expanded test collection				
LM JM BQ	14.72	05.55	18.49	31.57
LM JM	13.85	05.29	18.35	31.45
LM two-stage	13.51	05.25	18.44	31.49
LM JM2	13.24	05.19	18.23	31.43
LM Dir.	7.16	2.30	10.05	27.28
okapi	2.14	0.62	3.41	17.69
tfidf	2.07	0.58	3.24	17.44
okapi2	2.07	0.46	2.71	16.28
tfidf2	2.07	0.46	2.59	16.08
Official test collection				
LM Dir.	2.94	0.41	2.15	11.49
okapi	1.87	0.26	1.61	14.39
tfidf	1.74	0.26	1.59	14.30
okapi2	2.01	0.24	1.50	13.97
tfidf2	2.01	0.24	1.46	13.78
LM JM2	0.40	0.07	0.98	3.81
LM JM BQ	0.33	0.06	0.83	3.33
LM JM	0.40	0.06	0.83	3.31
LM two-stage	0.40	0.06	0.80	3.19

in this section we use a new set of runs on the expanded test collection, based on the touristic subset of ClueWeb and the Open Web runs, making these results not directly comparable to those in Section 3.

Figure 5 shows Overlap@N of the non-pooled runs with expanded test collection as well as the official contextual suggestion test collection. This experiment indicates that the injecting judged documents approach has a considerable impact on the personalized test collection fraction of judgments. In particular, Overlap@50 is improved from 0.14 to 0.26, which is 85% improvement in the fraction of judgments.

As discussed in Section 3, depersonalization of the contextual suggestion has a great impact on the fraction of judged documents. Table 4 (top) indicates that injecting judged documents in a fixed corpus has a great impact on the non-personalized test collection fraction of judgments. In addition, same as personalized expanded test collection, the system ranking based on the non-personalized expanded test collection is reasonable. However, according to Table 4 (bottom), system ranking of the same runs based on the official TREC test collection shows that the official test collection overlap is poor and it is not able to rank non-personalized systems in a logical order.

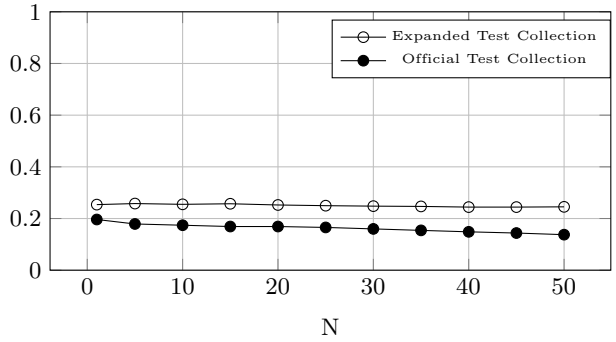

Figure 5: Overlap@N of non-pooled runs: official test collection versus expanded test collection for personalized runs.

To summarize, in this section we investigated the reusability of the expanded contextual suggestion test collection. The result is positive: we determined the reusability by doing a leave out uniques analysis, leading to perfect system rank agreement over a set of nine systems. In order to explain the ranking stability we looked at whether systems are indeed retrieving the inserted judged pages, and found that a fair and stable fraction of judged documents is retrieved, more than doubling the fraction of judged documents, and that this fraction is gently decreasing of the ranking. The effect of personalization remains large, and de-personalized versions of the qrels ignoring the profile lead to substantially higher fractions of retrieved judged documents. This gives strong support to the test collection expansion approach proposed in this paper.

6. CONCLUSIONS

In this paper, we investigated the challenges of expanding or updating a test collection in a dynamic domain. We experimented with a novel approach to reusable test collec-

Table 4: Non-personalized, non-pooled system ranking based on MAP and their overlap

Run	P@5	MAP (%)	bpref	Overlap@50 (%)
Expanded test collection				
LM JM BQ	48.94	15.30	22.91	87.10
LM two-stage	50.21	15.27	22.85	87.14
LM JM	49.36	15.21	22.84	87.14
LM JM2	49.36	15.14	22.81	87.19
LM Dir.	26.81	05.82	12.74	66.38
okapi	4.68	0.91	3.87	36.42
tfidf	3.83	0.86	3.80	36.29
okapi2	4.68	0.71	3.30	33.36
tfidf2	5.11	0.70	3.35	33.95
Official test collection				
LM Dir.	11.49	0.79	2.99	30.08
okapi	4.68	0.47	2.41	30.68
tfidf	3.83	0.44	2.38	30.38
okapi2	4.68	0.39	2.12	28.72
tfidf2	5.11	0.38	2.17	29.14
LM JM BQ	3.83	0.18	1.20	9.74
LM JM	3.83	0.18	1.23	9.87
LM JM2	3.83	0.18	1.26	10.04
LM two-stage	3.83	0.17	1.19	9.65

tion building, where we inject judged pages into an existing corpus, and have systems retrieve pages from the extended corpus with the aim to create a reusable test collection. In a way, we metaphorically hide the Easter eggs for systems to retrieve. The approach was motivated by, and applied to, the TREC Contextual Suggestion Track offering a personalized venue recommendation task, which allowed both submissions from a fixed corpus (ClueWeb12) as well as from the open web.

Our main research question was: *Can we build a reusable test collection for a dynamic domain by injecting judged documents into a test collection with sparse judgments?* Specifically, we answer following research questions: Our first research question was: *How reusable is the ClueWeb12 test collection of the TREC contextual suggestion?* The outcome is rather negative: the system rank correlation in the leave out uniques test is below 50%, with MAP and bpref scores close to zero; the fraction of judged documents after the pooling depth plummets down; and the combination of shallow pools over personalized runs aggravates the problem considerably. Our second research question was: *How to expand a test collection in order to improve its reusability?* Our approach is to simply "hide" the judged pages in the original collection, with the goal of systems to retrieve the relevant pages amongst the rest of the collection. The above scenario is a common case in all dynamic domains, such as online services on the web. We applied it to the case of the TREC contextual suggestion track, merging the large set of judged open web pages into the ClueWeb12 based collection, leading to an updated test collection with a far greater number of judged documents. Our third research question was: *How reusable is the expanded test collection containing judged open web documents?* The result is positive: we determined the reusability by doing a leave out uniques analysis, leading to perfect system rank agreement over a set of nine systems. We found that a fair and stable fraction of judged documents is retrieved, more than doubling the fraction of judged documents, and that this fraction is gently decreasing over the ranking. The effect of personalization remains large, and de-personalized versions of the qrels ignoring the profile lead to substantially higher fractions of retrieved judged documents.

Our general conclusion is that our proposed approach to update or expand a test collection offers novel and cost effective ways to build new test collections, and to refresh and update existing test collections. This offers new ways of effective maintenance of test collections for offline evaluation in dynamic domains such as the web. There are some open questions to address in future work. How general can the approach be applied? The case of the TREC contextual suggestion track had a unique configuration with both a fixed offline test collection and judged results from the open web, which greatly facilitated the experiments of this paper. The general case underlying the approach is dynamic data, such as almost all web data, and the track setup even models this with a crawled web collection from 2012 in combination with live web results from 2014. How to avoid bias in the set of pages to add? Clearly, if the set of additional pages is based on the unjudged pages retrieved by a single system, this will introduce bias toward this system and preclude a fair comparison to other systems. Our experimental data started with very sparse judgments (ClueWeb12) in combination with a considerable higher number of added pages and judgments (open web), how much is the impact in case the initial test collection was more complete? Web data is highly dynamic, with considerable numbers of new pages appearing in the index continuously making offline test collection age fast [17]. This leads to many high ranked but unjudged pages creating an obvious need to update the offline tests, and ways to reuse old judgments are of obvious value. How sensitive is the approach to the quality of the judgments on the inserted pages? Clearly adding just any labeled data may have some risks, as the judgments may be noisy or made under very different task assumptions, or even give opportunities for spamming [16]. We assume the new and old judgments are created in a similar way, typically by trusted editorial judges or through crowdsourcing platforms as used in this paper.

Acknowledgments

This research is funded in part by the European Community's FP7 (project meSch, grant # 600851).

REFERENCES

[1] J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. *SIGIR Forum*, 46(1):2–32, 2012.

[2] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, 2004.

[3] 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.

[4] J. Callan, J. Allan, C. L. A. Clarke, S. Dumais, D. A. Evans, M. Sanderson, and C. Zhai. Meeting of the MINDS: An information retrieval research agenda. *SIGIR Forum*, 41(2):25–34, 2007.

[5] B. Carterette. On rank correlation and the distance between rankings. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 436–443, 2009.

[6] B. Carterette, E. Kanoulas, V. Pavlu, and H. Fang. Reusable test collections through experimental design. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 547–554, 2010.

[7] C. W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield UK, 1962.

[8] G. V. Cormack and T. R. Lynam. Power and bias of subset pooling strategies. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 837–838, 2007.

[9] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 282–289, 1998.

[10] A. Dean-Hall, C. L. Clarke, J. Kamps, P. Thomas, and E. M. Voorhees. Overview of the TREC 2014 contextual suggestion track. In *Proceeding of Text REtrieval Conference (TREC)*, 2014.

[11] S. H. Hashemi and J. Kamps. Venue recommendation and web search based on anchor text. In *23rd Text REtrieval Conference (TREC)*, 2014.

[12] S. H. Hashemi, C. L. Clarke, A. Dean-Hall, J. Kamps, and J. Kiseleva. On the reusability of open test collections. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.

[13] G. K. Jayasinghe, W. Webber, M. Sanderson, and J. S. Culpepper. Improving test collection pools with machine learning. In *Proceedings of the 2014 Australasian Document Computing Symposium*, 2014.

[14] J. Kamps, S. Geva, C. Peters, T. Sakai, A. Trotman, and E. Voorhees. Report on the sigir 2009 workshop on the future of IR evaluation. *SIGIR Forum*, 43(2): 13–23, 2009.

[15] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 375–382, 2007.

[16] M. P. O'Mahony, N. J. Hurley, and G. C. M. Silvestre. Recommender systems: Attack types and strategies. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 1*, AAAI'05, pages 334–339, 2005.

[17] K. Radinsky, K. M. Svore, S. T. Dumais, M. Shokouhi, J. Teevan, A. Bocharov, and E. Horvitz. Behavioral dynamics on the web: Learning, modeling, and prediction. *ACM TOIS*, 31(3):16:1–16:37, 2013.

[18] T. Sakai. Comparing metrics across TREC and NTCIR: The robustness to system bias. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 581–590, 2008.

[19] I. Soboroff. Dynamic test collections: Measuring search effectiveness on the live web. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 276–283, 2006.

[20] I. Soboroff and S. Robertson. Building a filtering test collection for trec 2002. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 243–250, 2003.

[21] K. Sparck Jones and C. J. Van Rijsbergen. Report on the need for and provision of an 'ideal' information retrieval test collection. Technical report, University Computer Laboratory, Cambridge, 1975.

[22] TREC. Contextual suggestion track. https://sites.google.com/site/treccontext/, 2015.

[23] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, 2001.

[24] E. M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF*, pages 355–370, 2002.

[25] E. M. Voorhees, J. Lin, and M. Efron. On run diversity in evaluation as a service. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962, 2014.

[26] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 587–594, 2008.

[27] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, 1998.