On the Reusability of "Living Labs" Test Collections: A Case Study of Real-Time Summarization

Luchen Tan, Gaurav Baruah, and Jimmy Lin
David R. Cheriton School of Computer Science
University of Waterloo, Ontario, Canada
{luchen.tan,gbaruah,jimmylin}@uwaterloo.ca

ABSTRACT

Information retrieval test collections are typically built using data from large-scale evaluations in international forums such as TREC, CLEF, and NTCIR. Previous validation studies on pool-based test collections for *ad hoc* retrieval have examined their reusability to accurately assess the effectiveness of systems that did not participate in the original evaluation. To our knowledge, the reusability of test collections derived from "living labs" evaluations, based on logs of user activity, has not been explored. In this paper, we performed a "leave-one-out" analysis of human judgment data derived from the TREC 2016 Real-Time Summarization Track and show that those judgments do not appear to be reusable. While this finding is limited to one specific evaluation, it does call into question the reusability of test collections built from living labs in general, and at the very least suggests the need for additional work in validating such experimental instruments.

1 INTRODUCTION

Test collections are indispensable experimental tools for information retrieval evaluation and play an important role in advancing the state of the art. Beyond the ability to accurately measure the effectiveness of retrieval techniques, the reusability of test collections is an important and desirable characteristic. Test collections are often constructed via international evaluation forums such as TREC, CLEF, and NTCIR: a *reusable* test collection can be used to assess systems that did not participate in the original evaluation. TREC test collections created in the 1990s are still useful today precisely because they are reusable.

"Living labs" [1, 9, 11, 12] refers to an emerging approach to evaluating information retrieval systems in live settings with real users in natural task environments. Although such evaluation platforms are widespread in industry (e.g., frameworks for running large-scale A/B tests [8]), most academic researchers do not have access to them, and the living labs concept was designed to address this gap. Thus, the product of a living labs evaluation is a record of user actions that are then aggregated to assess the effectiveness of the participating systems.

SIGIR'17, August 7–11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080644

Following standard practices with *ad hoc* test collections constructed via pooling [13], it would be natural to treat these user judgments as part of a reusable test collection—that is, to evaluate systems that did not participate in the original evaluation. However, to our knowledge, the reusability of such test collections has not been explored, and it is unclear if post hoc measurements of new techniques are appropriate or accurate.

This paper explores the reusability of test collections built from living labs evaluations, using the TREC 2016 Real-Time Summarization (RTS) Track as a case study. We show—using a standard "leave-one-out" analysis—that user judgments *cannot* be used to reliably assess systems that did not participate in the original evaluation. Thus, we strongly caution researchers against using the RTS human judgments as a reusable test collection. To our knowledge, we are the first to have examined this issue, and while our finding is limited to a single evaluation, this result calls into question the reusability of living labs data in general. At the very least, more research is needed to validate the appropriateness of reusing data from living labs as evaluation instruments.

2 BACKGROUND AND RELATED WORK

The validation of information retrieval evaluation resources (i.e., meta-evaluations) has a long history that dates back to at least the 1990s. For standard *ad hoc* retrieval test collections built by pooling the results of evaluation participants, researchers have examined the quality of the judgments from a variety of perspectives [13]. There has been a long thread of research focused on reusability [2–4, 14]. The findings are nuanced and reusability is a characteristic of individual test collections, but researchers are generally aware of the pitfalls of reusing relevance judgments from pooled evaluations. One useful technique to study reusability that we adopt in this paper is known as the "leave-one-out" analysis. We evaluate the output of a particular participant by removing its contributions to the pooled judgments—this simulates it never having participated in the original evaluation. We can then assess the impact on the participant's score. This procedure can be repeated for every participant, allowing us to broadly characterize reusability.

Although traditional pool-based *ad hoc* retrieval test collections are generally well-understood evaluation instruments for assessing the quality of retrieval algorithms, researchers often find divergences between system-centered metrics and user-focused metrics. In short, better retrieval algorithms often don't lead to better task outcomes. There is a long line of work exploring this disconnect (a full survey is beyond the scope of this short paper, but see Hersh et al. [7] for a starting point). This realization, in turn, drove researchers to consider more user-focused evaluations. The living labs idea is the latest evolution in this thread of work. In an attempt to

leverage real users in realistic task environments, living labs share similar goals with Evaluation-as-a-Service (EaaS) [6] approaches that attempt to better align evaluation methodologies with user task models and real-world constraints to increase the fidelity of research experiments.

Living labs evaluations are modeled after online experiments in industry such as A/B tests [8] and interleaved evaluations for web search [5]. Of course, academic researchers have difficulty gaining access to such experimental platforms: in this sense, living labs represent an attempt by academic researchers to replicate an evaluation framework that researchers in industry take for granted. One early attempt is described by Said et al. [11], where a company called Plista opened up their news recommender system to academic researchers, who were able to deploy their algorithms in a production setting and receive user feedback. More recent attempts include a living labs deployment at CLEF 2015 [12] and the TREC 2016 Open Search Track [1], both of which explored vertical search using interleaved evaluations, where researchers were invited to submit their ranking algorithms. In this paper, we focus on the TREC 2016 Real-Time Summarization Track [9], a recent living labs evaluation. Unlike pool-based construction of test collections for *ad hoc* retrieval, to our knowledge there has not been reusability studies of data from such evaluations. As far as we are aware, researchers in industry are not concerned with this issue because they have access to large amounts of human editorial judgments (in the case of web search) and the ability to run online experiments on demand, and hence they would not have the need to reuse the results of, for example, previous interleaved ranking experiments.

3 REAL-TIME SUMMARIZATION

The TREC 2016 Real-Time Summarization (RTS) Track tackled prospective information needs against real-time, continuous document streams, exemplified by social media services such as Twitter. Systems are given a number of "interest profiles" representing users' needs (analogous to topics in *ad hoc* retrieval), and their task is to automatically monitor the document stream (Twitter in this case) to keep the user up to date with respect to the interest profiles. The evaluation specifically considers the case where tweets are immediately pushed to users' mobile devices as notifications. At a high level, these push notifications must be relevant, novel, and timely. Here, we provide relevant background, but refer the reader to the track overview [9] for more details.

In order to evaluate push notification systems in a realistic setting, the track defined an official evaluation period (from August 2, 2016 00:00:00 UTC to August 11, 2016 23:59:59 UTC) during which all participants "listened" to the so-called Twitter "spritzer" sample stream. This is putatively a 1% sample of all Twitter public posts, and is available to anyone with a Twitter account. The overall evaluation framework is shown in Figure 1. Before the evaluation period, participants "registered" their systems with the evaluation broker to request unique tokens (via a REST API), which are used in future requests to associate submitted tweets with specific systems. During the evaluation period, whenever a system identified a relevant tweet with respect to an interest profile, the system submitted the tweet id to the evaluation broker (also via a REST API), which recorded the submission time. Each system was allowed to push at

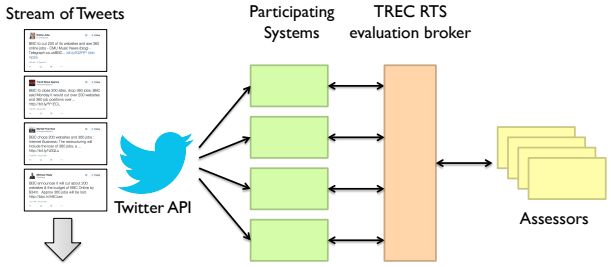

Figure 1: The evaluation setup for push notifications: systems "listen" to the live Twitter sample stream and send results to the evaluation broker, which then delivers push notifications to users.

most ten tweets per interest profile per day; this limit represents an attempt to model user fatigue.

The track organizers recruited a number of users who installed a custom app on their mobile devices. Prior to the beginning of the evaluation period, these users subscribed to interest profiles (i.e., topics) they wished to monitor. During the assessment period, whenever the evaluation broker received a system's submission, the tweet was *immediately* delivered to the mobile devices of users who had subscribed to the particular interest profile and rendered as push notifications—we implemented the temporal interleaving evaluation methodology for prospective information needs proposed by Qian et al. [10]. The user may choose to assess the tweet immediately, or if it arrived at an inopportune time, to ignore it. Either way, the tweet is added to a queue in the app on the user's mobile device, which she can access at any time to examine the queue of accumulated tweets. For each tweet, the user can make one of three judgments with respect to the associated interest profile: *relevant*, if the tweet contains relevant and novel information; *redundant*, if the tweet contains relevant information, but is substantively similar to another tweet that the user had already seen; *not relevant*, if the tweet does not contain relevant information. As the user provides judgments, results are relayed back to the evaluation broker and recorded. These judgments are then aggregated to assess the effectiveness of participating systems.

Our setup has two distinct characteristics: First, judgments happen online as systems generate output, as opposed to traditional batch post-hoc evaluation methodologies, which consider the documents some time (typically, weeks) after they have been generated by the systems. Second, our judgments are *in situ*, in the sense that the users are going about their daily activities (and are thus interrupted by the notifications). This aspect of the design accurately mirrors the intended use of push notification systems. For these two reasons, the RTS track exemplifies a living labs evaluation.

4 EXPERIMENTS

In this paper, we analyzed data from the TREC 2016 RTS Track. In total, 18 groups from around the world participated, submitting a total of 41 systems (runs). Over the evaluation period, these systems pushed a total of 161,726 tweets, or 95,113 unique tweets after de-duplicating within profiles. The organizers recruited 13 users for the study, who collectively provided 12,115 judgments over the assessment period, with a minimum of 28 and a maximum of 3,791 by an individual user. Overall, 122 interest profiles received

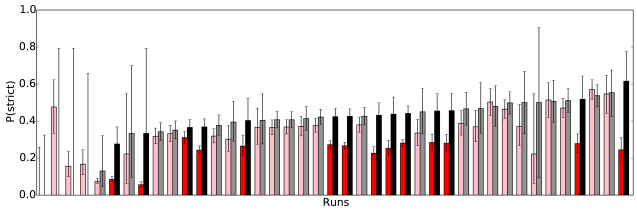

Figure 2: Results of the leave-one-out analysis. Each pair of bars represents precision before and after the leave-one-out procedure. Error bars denote binomial confidence intervals. Darker pairs of bars represent significant differences. Bars are sorted by "after" precision.

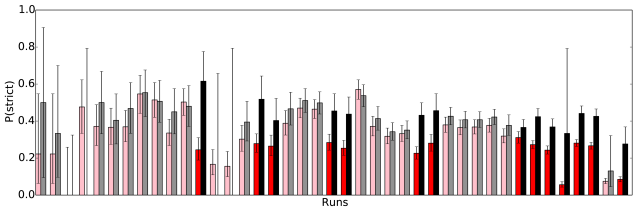

Figure 3: Alternate presentation of the leave-one-out analysis results. The setup is exactly the same as in Figure 2, except that the bars are sorted by increasing push volume.

at least one judgment; 93 received at least 10 judgments; 67 received at least 50 judgments; 44 received at least 100 judgments.

Following the setup of the track, we evaluate systems in terms of "strict" precision, which is simply the fraction of user judgments that were relevant (more precisely, a micro-average across profiles). The metric is "strict" in the sense that redundant judgments are treated as not relevant.

4.1 Leave-One-Out Analysis

We began with a standard leave-one-out analysis at the group level. Since runs originating from each group are likely to be similar (e.g., same underlying algorithm but different parameter settings), a group-level analysis better matches real-world conditions.[1] Specifically, we removed the unique judgments associated with runs from a particular group to simulate what would have happened if the group had not participated in the living labs evaluation, and then evaluated runs from that group with the reduced set of judgments. This procedure was then repeated for all the groups.

In each case, we computed the precision of the runs with the reduced set of judgments (which we call the "after" condition, i.e., after the leave-one-out procedure). This precision can then be compared with the official (i.e., "before") precision using all judgments. If the judgments are reusable, we should not notice significant differences in the score. That is, we would obtain an accurate measurement of precision whether or not the group had participated in the original living labs evaluation.

Results of our leave-one-out analysis are shown in Figure 2. Each pair of bars represents precision before and after the leave-one-out procedure. Error bars denote binomial confidence intervals: systems vary in the volume of tweets they push, and so the confidence intervals tend to be larger for systems that push fewer tweets. All pairs of bars are sorted by the "after" precision, and the darker pairs represent significant differences.

From this analysis, we see that for 14 of 41 runs (approximately one third of the runs), the before and after precision values are significantly different. These differences seem to be in the false positive direction, in the following sense: in trying to use RTS judgments as a reusable test collection to score a run, a researcher might obtain a score (i.e., an "after" score) that is significantly higher than its "true" score (i.e., the "before" score). Thus, there is a danger of over-inflated effectiveness. Interestingly, there doesn't appear to be a case where the "after" condition under-reports the true (i.e., "before") precision.

One reasonable hypothesis might be that push volume (i.e., number of tweets that a system pushes) provides an important feature in determining whether post-hoc assessments are accurate. In Figure 3, the pairs of bars in Figure 2 are resorted by increasing push volume. The results appear to disprove this hypothesis: while high volume systems do show significant "before" vs. "after" differences, systems across the board (in terms of push volume) exhibit significant differences. Note that low volume systems have large confidence intervals, and thus are less likely to exhibit significant differences to begin with.

Taken together, these results suggest that user judgments from the TREC 2016 RTS Track cannot be reused to reliably evaluate systems that did not participate in the original evaluation.

4.2 Temporal Analysis

Since systems pushed tweets at various times during the day, we wondered if there were temporal effects impacting reusability. To explore this question, we performed a series of analyses focused on dropping judgments in different temporal segments. Specifically, we divided each day into four-hour segments and separately dropped all judgments within that particular time window (across all days in the evaluation period). This yielded six different conditions, i.e., dropping judgments from 00:00 to 03:59, from 04:00 to 07:59, etc. All times are in UTC. For each of the segments, we can repeat our before/after analysis, i.e., computing precision based on the full and reduced sets of judgments.

These results are shown in Figure 4, where the pairs of bars are sorted by run id (in other words, arbitrarily), but the position of each pair of bars is consistent from plot to plot. Interestingly, we see significant differences in the first temporal segment (00:00 to 03:59) and the last temporal segment (20:00 to 23:59), but no where else. For the first temporal segment, 12 out of 41 runs exhibit significant differences, and for the final segment, 3 out of 41 runs.

As a sanity check, we repeated the before/after experiments randomly throwing away the same amount of data discarded in each of the temporal segments. For instance, 40% of the judgments are found in the first temporal segment (00:00 to 03:59 UTC), and so as a comparison condition, we randomly discarded 40% of all judgments and repeated our before/after analyses. The results averaged over ten distinct trials are shown in Figure 5. Over ten trials, we did not observe any significant differences in any of the pairs. Thus, we can conclude that the significant differences observed in Figure 4 are *not* due to simply have fewer judgments.

[1] As a detail, the University of Waterloo submitted runs that were quite similar, but using two different group ids. These were collapsed into the same group.

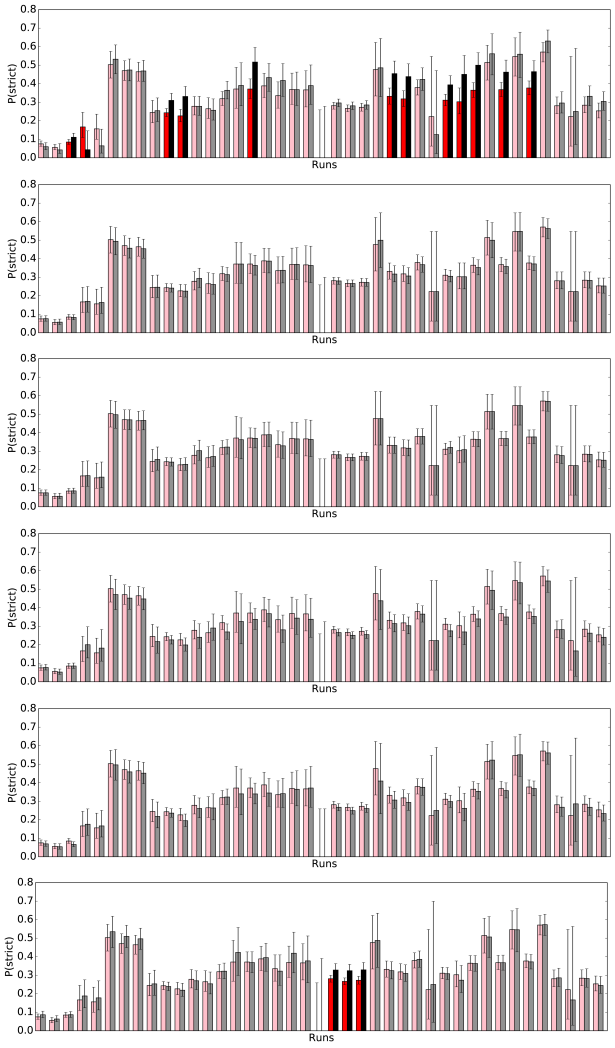

Figure 4: Results of the before/after analysis removing all judgments within a particular four-hour window. Each plot shows the before and after precision of removing each four-hour block (in UTC). Runs are sorted by run id.

As yet, we have no adequate explanation for these findings. UTC 00:00 corresponds to 20:00 local time for the users who participated in the study. While it is true that the RTS broker generally received more judgments during the evening hours throughout the evaluation period (corresponding to the first and last temporal segments), we have confirmed above that the volume of judgments alone does not explain the significant differences we observed. Judgments and system behavior (or both) around this time are somehow "special", in a way that we currently do not understand.

5 CONCLUSIONS

The main contribution of our work is the finding that user judgments from the TREC 2016 RTS Track do not appear to be reusable. That is, they cannot be used to reliably assess the effectiveness of systems that did not participate in the original evaluation. Although our findings are limited to a particular instance of a living labs evaluation, it raises questions about the reusability of such test collections in general. Absent future work to the contrary, experimental results that derive from the reuse of human judgments as part of a living labs must be viewed with suspicion.

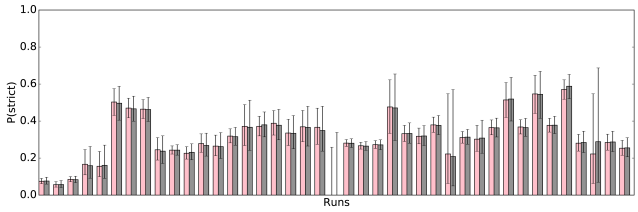

Figure 5: Results of a before/after analysis randomly dropping 40% of the judgments, equal to the amount removed in the first temporal segment. Runs are sorted by run id.

Taken together, these findings unfortunately put information retrieval researchers in somewhat of a quandary: in the quest for more user-centered evaluation techniques that better capture task-level effectiveness, we might have compromised desirable properties of evaluation instruments previously taken for granted. We have identified the problem, which is an important first step, but leave for future work how to actually "fix it".

REFERENCES

[1] Krisztian Balog, Anne Schuth, Peter Dekker, Narges Tavakolpoursaleh, Philipp Schaer, and Po-Yu Chuang. 2016. Overview of the TREC 2016 Open Search Track. In *TREC*.
[2] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2007. Bias and the Limits of Pooling for Large Collections. *Information Retrieval* 10, 6 (2007), 491–508.
[3] Stefan Büttcher, Charles L. A. Clarke, Peter C. K. Yeung, and Ian Soboroff. 2007. Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements. In *SIGIR*. 63–70.
[4] Ben Carterette, Evgeniy Gabrilovich, Vanja Josifovski, and Donald Metzler. 2010. Measuring the Reusability of Test Collections. In *WSDM*. 231–239.
[5] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. 2012. Large-Scale Validation and Analysis of Interleaved Search Evaluation. *ACM TOIS* 30, 1 (2012), Article 6.
[6] Allan Hanbury, Henning Müller, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Ivan Eggel, Tim Gollub, Frank Hopfgartner, Jayashree Kalpathy-Cramer, Noriko Kando, Anastasia Krithara, Jimmy Lin, Simon Mercer, and Martin Potthast. 2015. Evaluation-as-a-Service: Overview and Outlook. *arXiv:1512.07454*.
[7] William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, and Daniel Olson. 2000. Do Batch and User Evaluations Give the Same Results? In *SIGIR*. 17–24.
[8] Ron Kohavi, Randal M. Henne, and Dan Sommerfield. 2007. Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO. In *KDD*. 959–967.
[9] Jimmy Lin, Adam Roegiest, Luchen Tan, Richard McCreadie, Ellen Voorhees, and Fernando Diaz. 2016. Overview of the TREC 2016 Real-Time Summarization Track. In *TREC*.
[10] Xin Qian, Jimmy Lin, and Adam Roegiest. 2016. Interleaved Evaluation for Retrospective Summarization and Prospective Notification on Document Streams. In *SIGIR*. 175–184.
[11] Alan Said, Jimmy Lin, Alejandro Bellogín, and Arjen P. de Vries. 2013. A Month in the Life of a Production News Recommender System. In *CIKM Workshop on Living Labs for Information Retrieval Evaluation*. 7–10.
[12] Anne Schuth, Krisztian Balog, and Liadh Kelly. 2015. Overview of the Living Labs for Information Retrieval Evaluation (LL4IR) CLEF Lab 2015. In *CLEF*.
[13] Ellen M. Voorhees. 2002. The Philosophy of Information Retrieval Evaluation. In *CLEF*. 355–370.
[14] Justin Zobel. 1998. How Reliable Are the Results of Large-Scale Information Retrieval Experiments? In *SIGIR*. 307–314.

Acknowledgments. This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, with additional contributions from the U.S. National Science Foundation under IIS-1218043 and CNS-1405688.