On the Reusability of Personalized Test Collections

Seyyed Hadi Hashemi
University of Amsterdam
Amsterdam, The Netherlands
hashemi@uva.nl

Jaap Kamps
University of Amsterdam
Amsterdam, The Netherlands
kamps@uva.nl

ABSTRACT

Test collections for offline evaluation remain crucial for information retrieval research and industrial practice, yet reusability of test collections is under threat by different factors such as dynamic nature of data collections and new trends in building retrieval systems. Specifically, building reusable test collections that last over years is a very challenging problem as retrieval approaches change considerably per year based on new trends among Information Retrieval researchers. We experiment with a novel temporal reusability test to evaluate reusability of test collections over a year based on leaving mutual topics in experiment, in which we borrow some judged topics from previous years and include them in the new set of topics to be used in the current year. In fact, we experiment whether a new set of retrieval systems can be evaluated and comparatively ranked based on an old test collection. Our experiments is done based on two sets of runs from Text REtrieval Conference (TREC) 2015 and 2016 Contextual Suggestion Track, which is a personalized venue recommendation task. Our experiments show that the TREC 2015 test collection is not temporally reusable. The test collection should be used with extreme care based on early precision metrics and slightly less care based on NDCG, bpref and MAP metrics. Our approach offers a very precise experiment to test temporal reusability of test collections over a year, and it is very effective to be used in tracks running a setup similar to their previous years.

KEYWORDS

Test Collection Building, Reusability, Contextual Suggestion, Personalization

ACM Reference format:
Seyyed Hadi Hashemi and Jaap Kamps. 2017. On the Reusability of Personalized Test Collections. In *Proceedings of UMAP'17 Adjunct, July 09-12, 2017, Bratislava, Slovakia, ,* 5 pages.
DOI: **10.1145/3099023.3099044**

1 INTRODUCTION

Test collection building is one of the most popular evaluation activities in Information Retrieval since more than 50 years ago, starting from the first large scale experimental evaluations of retrieval effectiveness of various indexing languages for literature at Cranfield [1].

UMAP'17 Adjunct, July 09-12, 2017, Bratislava, Slovakia

DOI: **10.1145/3099023.3099044**

Test collections have been extensively used in both academia and industry for offline evaluation purposes. However, due to the amount of human efforts needed to achieve a desirable pooling depth in the TREC-Style test collection building based on pooling approach, building a high quality and reusable test collections for personalization evaluation is very challenging and difficult [2–4, 6].

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 [5, 7]. TREC contextual suggestion track offers a personalized venue recommendation task, in which contextual suggestion systems have to provide suggestions based on a context (e.g., U.S. city) and a user profile (e.g., the user preferences). Creating a test collection for the contextual suggestion is different from the traditional non-personalized test collections, as we have to collect judgments for each user separately.

In the TREC contextual suggestion track, different users might have different preferences in a same city, which leads to different judgments for different user profiles. It basically adds another dimension to the complexity of building reusable test collections. Therefore, a TREC-style test collection building for contextual suggestion based on deep pool depth is very difficult and unrealistic as it needs a lot of human judgments. Using a shallow pool-depth in the test collection building process, makes it very challenging to create a reusable personalized test collection for comparative evaluation of both pooled and non-pooled personalization systems.

Moreover, personalization approaches seem to change considerably every year. For example, the emergence of new trends in using retrieval models (e.g., neural retrieval models in last few years) leads to variety of personalization systems retrieving very different suggestions for a same information need over years. Therefore, even if a test collection is reusable at the time of its creation, it might not be reusable over time.

One way to test reusability of test collections for comparative evaluations over years would be the comparison of system rankings created based on a set of runs implemented a year after the test collection creation time with a ground-truth system ranking created based on the new set of runs but using a new test collection. We investigate on this approach in the rest of this paper.

In this paper, our main aim is to study the question: *How reusable could be a test collection over a year?* Specifically, we answer the following research questions:

(1) *How are retrieved documents changed over a year in TREC Contextual Suggestion Track?*
(2) *Can we evaluate TREC 2016 contextual suggestion submitted runs based on TREC 2015 contextual suggestion test collection?*

We first investigate the mutuality of retrieved documents by TREC Contextual Suggestion submissions over a year. Then, we

propose a novel experiment to test temporal reusability of test collections over a year, which is our main contribution.

The rest of this paper is organized as follows. In Section 2, we give a short summary of the TREC Contextual Suggestion Track. Section 3 is devoted to investigation on system agreement in retrieving mutual documents for same requests. Our proposed temporal reusability test is detailed in Section 4. Finally, we present the conclusions and future work in Section 5.

2 TREC CONTEXTUAL SUGGESTION TRACK

In this section, a short overview of the TREC Contextual Suggestion track is given and we discuss how we ran TREC 2016 contextual suggestion track to address research questions of this paper.

In TREC 2016, the track followed the setup of TREC 2015, which facilitates testing temporal reusability of the personalized test collection over a year. The track has two phases, namely, phase 1 and phase 2. In both phase 1 and phase 2 tasks, participants were asked to develop a system providing relevant suggestions to a specific person based on their given profile and context. The contextual suggestion track organizers provide a set of profiles, a set of contexts and a set of example suggestions (URLs of pages corresponding to POIs in a given context) as input of the task. Each profile corresponds to a user's preferences in example suggestions of another context or city, their gender and age. Moreover, the target city (i.e., the target location), a trip type, a trip duration, a type of group the person is travelling with, and a season the trip will occur in are considered as context.

Profiles correspond to the stated preferences of real individuals, who either are recruited by crowdsourcing or are editorial judges. These assessors first judged example attractions in seed locations, later returning to judge suggestions provided by the phase 1 participants for new contexts. At seed location judgement phase or each return, assessors were able to ask for suggestions relevant to a context that was chosen by them at that point.

As output of the phase 1 task, participants were required to provide a ranked list of 50 suggestions for each context and profile pair. Each suggestion was expected to be relevant to the given profile and the context. As output of the phase 2 task, participants were expected to rerank the given suggestion candidates with respect to the user's profile and context. In this study, we are interested in personalized retrieval task rather than reranking. Therefore, In the rest of this paper, we will detail the test collection created for the TREC 2016 phase 1 task (similar to TREC 2015 contextual suggestion live task) and TREC 2015 test collection reuability over a year.

The TREC contextual suggestion track use a collection of URLs corresponding to POIs in each context, see the examples in Table 1. The track has also released the TREC Contextual Suggestion Web corpus. The TREC CS web corpus is a web crawl of the suggestions' URLs available at the TREC contextual suggestion collection, and includes attraction web pages of 272 different North American cities. For more information about the TREC Contextual Suggestion Web corpus refer to the TREC 2016 contextual suggestion overview paper [5].

2.1 TREC Contextual Suggestion Test Collection

In TREC 2016 contextual suggestion track, the track organizers released 438 requests, in which 211 of them were borrowed from TREC 2015 contextual suggestion track requests. Including the released requests of TREC 2015 in the requests of TREC 2016 helps us to investigate contextual suggestion systems variations over a year. However, as the judgments of the 211 borrowed requests from 2015 were already available since 2015, we have not used those judgments as official test collection to rank TREC 2016 submissions.

In order to rank TREC 2016 submissions, 61 new high-quality set of requests has been used as the official TREC 2016 requests and the corresponding judgments have been used as the official TREC 2016 contextual suggestion test collection. The official TREC 2016 test collection is used as a ground truth for system ranking in this paper.

Relevance judgments of the suggestions retrieved by the participants for the given set of requests were collected through crowdsourcing and editorial judges. They were asked to rate suggestions in a graded 5 point scale judgments as follows:

(1) 4: Strongly interested
(2) 3: Interested
(3) 2: Neither interested or uninterested
(4) 1: Uninterested
(5) 0: Strongly uninterested
(6) -1: Not loaded or no rating given

However, in the test collection, we have shifted the raw assessors' 5 point scale judgments with -2, making the judgments in the range -3 to 2, and making a score of 1.0 or higher correspond to a "interested" or "strongly interested" judgment. Therefore, the trec_eval [1] can be used to evaluate contextual suggestion runs based on all the common IR measures, included graded measures like NDCG.

2.2 TREC 2016 Submissions

In TREC 2016 contextual suggestion track, 8 different organizations participated in the phase 1 and submitted 15 different runs. We will use these 15 runs to test reusability of the TREC 2015 test collection over a year. These runs provide suggestions for both the 211 borrowed requests from TREC 2015 and the official 61 high-quality requests judged in 2016.

3 SYSTEMS AGREEMENT OVER A YEAR

This section studies the contextual suggestion systems agreement in providing suggestions relevant to a context and profile pair, aiming to answer our first research question: *How are retrieved documents changed over a year in TREC Contextual Suggestion Track?*

In reality, we could create a reusable test collection if we have an acceptable variety of submissions, avoid pooling bias in favor of some of the submissions and pool deep enough. In this way, the test collection can last long over years, and it might be possible to evaluate new sets of runs created years later than the test collection creation time. However, it is a very difficult task for a personalization problem.

[1] http://trec.nist.gov/trec_eval/

Table 1: TREC Contextual Suggestion track collection example.

Attraction ID	City ID	URL	Title
TRECCS-00000005-418	418	http://www.greatfallsmt.net/people_offices/park_rec/gibson.php	"Gibson Park"
TRECCS-00000006-418	418	http://www.mackenzieriverpizza.com	"MacKenzie River Pizza Co"
TRECCS-00000007-418	418	http://www.bostons.com	"Bostons Restaurant Sports Bar"
TRECCS-00000008-418	418	http://pink.victoriassecret.com	"Victorias Secret PINK"

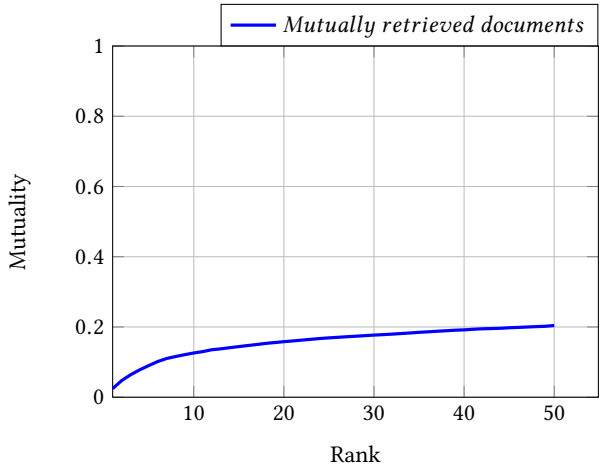

Figure 1: Mutually retrieved documents in both TREC 2015 and 2016 contextual suggestion submissions.

As it is shown in Figure 1, just 9% of the top-5 retrieved documents for the mutual 211 requests available in both TREC 2015 and 2016 are mutual between TREC 2015 and 2016 contextual suggestion submissions. Specifically, 91 % of the top-5 retrieved documents for the 211 context and user profile pairs by the TREC 2016 contextual suggestion submissions are exclusively retrieved by them and they are not retrieved as top-5 documents in TREC 2015.

According to Figure 1, the mutuality score of the top-N retrieved documents by TREC 2016 submissions with the top-N retrieved documents by TREC 2015 submissions increases as we go deeper in the ranking. Specifically, at rank 50, the mutuality is the highest among the possible documents ranks by having 20 % of mutuality. This experiment potentially means that although the official TREC contextual suggestion track metrics are based on early precision, they might not be the best metrics for evaluating non-pooled systems in contextual suggestion. In fact, more stable metrics for evaluating retrieval systems based on incomplete test collections such as bpref metric is a potentially better metric to use. In the next section, we will investigate on this assumption by doing leave-mutual-topics-in temporal reuasability experiment.

4 LEAVE MUTUAL TOPICS IN

We now look at the question: *Can we evaluate TREC 2016 contextual suggestion submitted runs based on TREC 2015 contextual suggestion test collection?* In order to test the reusability over a year for the TREC 2015 contextual suggestion test collection, we did the leave-mutual-topics-in (LMTI) test in TREC 2016 contextual suggestion, in which we asked participants to retrieve suggestions related to the 211 requests already have been used in 2015. Participants were not aware of using the 211 borrowed requests in the new set of TREC 2016 requests.

In TREC 2016 test collection building phase, we did not pool retrieved documents for the 211 borrowed requests. In fact, we wanted to try evaluating TREC 2016 submitted contextual suggestion systems based on TREC 2015 contextual suggestion test collection. This is the main idea behind the LMTI experiment. Specifically, if the LMTI system ranking have an acceptable correlation with the official system ranking, then we can conclude that the test collection is reusable and last over a year.

In order to test the reusability over a year, we calculate the LMTI system ranking correlation with the official system ranking as the ground-truth based on 2 groups of evaluation metrics. The first group includes NDCG@5, P@5 and MRR, which are early precision based metrics. The second group is based on relatively more stable evaluation metrics in incomplete test collections including bpref, MAP and NDCG.

As it is shown in Figure 2, the system ranking correlation of the LMTI system ranking and the official system ranking based on kendall's τ is much lower than 0.9, the threshold usually considered as two effectively equivalent rankings in leave uniques out tests [8]. This is observed for all the early precision metrics of the first group of evaluation metrics. Therefore, reusability of test collection based on NDCG@5, P@5 and MRR is not approved based on the LMTI reusability test over a year.

Moreover, we test the reuability over a year based on the second group of IR evaluation metrics, which are more stable than the first group in system ranking based on incomplete test collections. Figure 3 indicates that the reusability of the TREC 2015 test collection over a year is higher based on the second group of metrics in comparison to the first group, which are early precision based metrics.

As it is shown in Figure 3, the system ranking correlation of the LMTI system ranking and the official system ranking based on kendall's τ is 0.68, 0.68 and 0.64 using NDCG, bpref and MAP, respectively. These system ranking correlations based on kendall's τ is lower than 0.9, the threshold usually considered as two effectively equivalent rankings in leave uniques out tests. According to this experiment, although the TREC 2015 test collection reusablity over a year based on NDCG, bpref and MAP metrics is higher than the reusability based on early precision metrics, the test collection should be used with some cares based on even stabler metrics.

5 CONCLUSIONS

In this paper, we investigated the contextual suggestion systems agreement in ranking attractions for same requests over a year

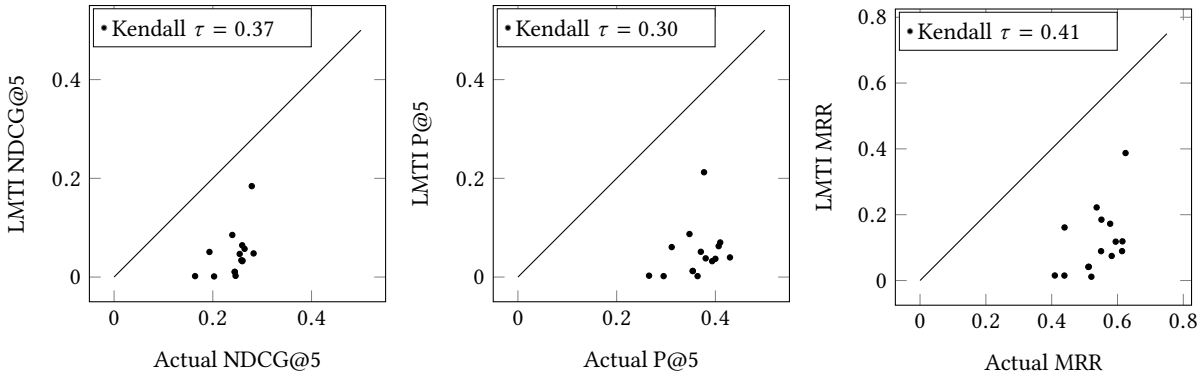

Figure 2: Leave Mutual Topics In (LMTI) temporal reusability test of the TREC 2015 contextual suggestion test collection over a year for ranking TREC 2016 contextual suggestion submissions based on NDCG@5, P@5 and MRR metrics.

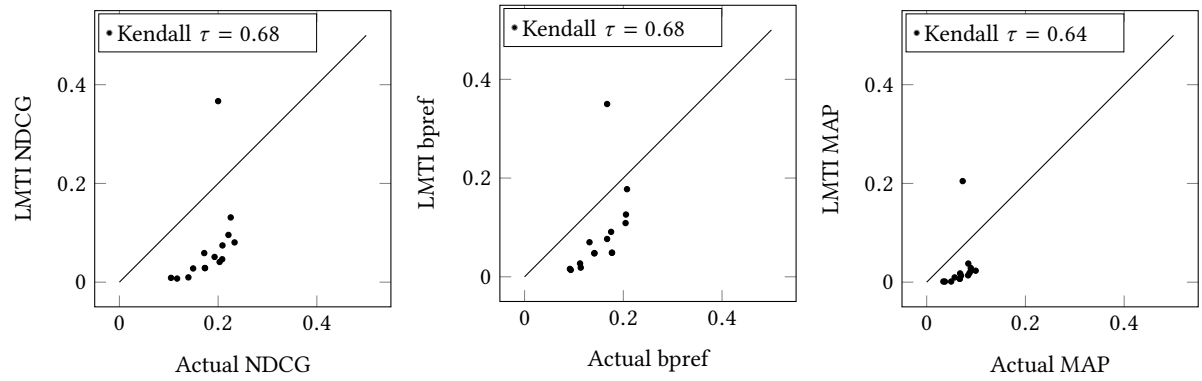

Figure 3: Leave Mutual Topics In (LMTI) temporal reusability test of the TREC 2015 contextual suggestion test collection over a year for ranking TREC 2016 contextual suggestion submissions based on NDCG, bpref and MAP metrics.

and its effect on reusability of test collections. We observed that contextual suggestion systems change significantly over a year and retrieve very different suggestions for a same user in a same context. This observation shows the difficulty of the reusability problem for TREC contextual suggestion, which is a highly personalized and contextualized problem. We also observed that contextual suggestion systems retrieve a higher percentage of mutually retrieved documents over a year in deeper ranks in comparison to the shallower ranks. This leaded to higher reusability of the test collection based on more stable and recall-based metrics in comparison to early precision based metrics. Moreover, our experiment based on the novel leave mutual topics in (LMTI) temporal reusability test indicates that the TREC 2015 contextual suggestion test collection is more reusable based on NDCG, bpref and MAP metrics in comparison to the used early precision based metrics, namely, NDCG@5, P@5 and MRR. The TREC 2015 test collection have not passed the LMTI temporal reusability tests based on all the tested evaluation metrics, and the overall conclusion of the temporal reusability of the TREC 2015 test collection is that the test collection should be used by extreme care for evaluations based on early precision metrics and it can be used with slightly less care for evaluations based on NDCG, bpref and MAP metrics. Our proposed LMTI temporal reusability test is very precise and effective reusability experiment, however, our proposed approach is not applicable for those tracks that are not supposed to run again or tracks that significantly change their setup over years. As a future work, we will work on a temporal reusablity test using simulations that is very useful for measuring temporal reusability of test collections without running the track for one more year or even before creating test collections.

ACKNOWLEDGMENTS

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

REFERENCES

[1] C. W. Cleverdon. 1962. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems.* Technical Report. College of Aeronautics, Cranfield UK.

[2] Seyyed Hadi Hashemi, Charles L.A. Clarke, Adriel Dean-Hall, Jaap Kamps, and Julia Kiseleva. 2015. On the Reusability of Open Test Collections. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[3] Seyyed Hadi Hashemi, Charles L.A. Clarke, Adriel Dean-Hall, Jaap Kamps, and Julia Kiseleva. 2016. An Easter Egg Hunting Approach to Test Collection Building in Dynamic Domains. In *Proceedings of NTCIR-EVIA 2016*. 1–8.

[4] Seyyed Hadi Hashemi, Charles L. A. Clarke, Adriel Dean-Hall, Jaap Kamps, and Julia Kiseleva. 2016. Test Collection Building and Maintenance in Dynamic Domains. In *15th Dutch-Belgian Information Retrieval Workshop (DIR)*.

[5] Seyyed Hadi Hashemi, Charles L. A. Clarke, Jaap Kamps, Julia Kiseleva, and Ellen M. Voorhees. 2016. Overview of the TREC 2016 Contextual Suggestion Track. In *Proceeding of Text REtrieval Conference (TREC)*.

[6] Seyyed Hadi Hashemi and Jaap Kamps. 2014. Venue Recommendation and Web Search Based on Anchor Text. In *23rd Text REtrieval Conference (TREC)*.

[7] TREC 2016. Contextual Suggestion Track. **https://sites.google.com/site/treccontext/**. (2016).

[8] Ellen M. Voorhees. 2001. Evaluation by Highly Relevant Documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 74–82.