Test Collection Building and Maintenance 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

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. 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 agenda's 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.

We experiment with a novel approach to reusable test collection building in dynaimc domains, 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. Our experiments exploit the unique setup of the TREC Contextual Suggestion Track, which offering a personalized venue recommendation task starting from a U.S. city as context, and exploiting crowdsourced profiles and judgments. The track 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.

In this study, 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, or update test collections for dynamic domains that have become outdated. This explores new ways of effective maintenance of offline test collections for dynamic domains such as the web.

2. REUSABILITY OF THE EXPANDED TEST COLLECTION

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?*

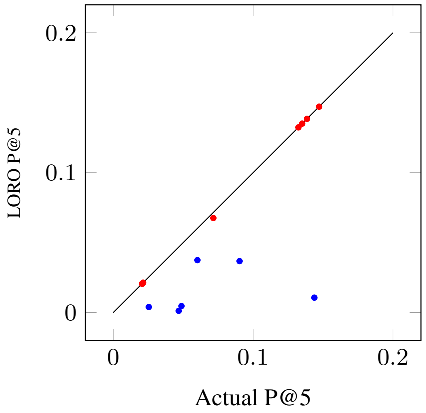

Figure 1: Difference in P@5 (Kendall τ = 1.00, ap corr = 1.00, avg diff = 0.01) based on the leave one run out (LORO) test on the expanded test collection (red dots) and difference in P@5 (Kendall τ = 0.46, ap corr = 0.11, avg diff = 0.76) based on the leave one run out (LORO) test on the official TREC Contextual Suggestion test collection (blue dots).

In order to test reusability of the test collection, we build nine different personalized contextual suggestion runs. Then, leave out uniques test is done using leave-one-run-out (LORO) test. According to Figure 1, the actual system ranking of the nine built runs 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. As it is shown in Figure 1, the Easter egg hunting approach effectively improves reusability of TREC Contextual Suggestion track test collections based on the ClueWeb12 corpus, which is not reusable same as the TREC Contextual Suggestion track test collection based on open web [1–3].

References

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

[2] 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*, pages 827–830, 2015.

[3] S. H. Hashemi, C. L. Clarke, A. Dean-Hall, J. Kamps, and J. Kiseleva. An easter egg hunting approach to test collection building in dynamic domains. In *Proceedings of NTCIR-EVIA 2016*, pages 1–8, 2016.

[*] This is an extended abstract of Hashemi et al. [3].