Living Ranking: from online to real-time information retrieval evaluation

Lamjed Ben Jabeur*, Laure Soulier**, Paul Mousset*, and Lynda Tamine*

* IRIT, Université de Toulouse, CNRS, UPS, 118 Route Narbonne, Toulouse, France
** Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 75005 Paris
{jabeur,mousset,tamine}@irit.fr,laure.soulier@lip6.fr

Abstract. The Living Labs for Information Retrieval Evaluation (LL4IR) initiative have provided a novel framework for evaluating retrieval models that involve real users. In this position paper, we propose an extension to the LL4IR framework that enables to evaluate real-time IR.
Keywords: Information retrieval, Evaluation, LL4IR, Living Labs

1 Introduction

Most of the frameworks used today for evaluating IR models, such as TREC tracks, are faced to the limitation of the consideration of the user within the effectiveness measurement. To tackle this gap, two main user-centered initiatives, namely the LL4IR [2] and the NewsREEL [1] benchmarks running for CLEF, have been launched. Both initiatives borrow the concepts of *"living labs"* for online evaluation. While NewsReel is specific to recommendation systems, LL4IR addresses experimental evaluation in the context of IR. The LL4IR proposes an evaluation framework relying on real use cases in which real users evaluate, for a set of predefined queries, online rankings produced offline by participants. The main issues of such approach are: *(1)* the restriction of predefined queries among the frequent ones which may not reflect for instance time-sensitive information needs; *(2)* the discard of user context such as search session; and the most important one *(3)* the dismissing of real-time updates, namely freshly created or disappeared documents. Accordingly, the LL4IR framework does not allow to provide a full living lab methodology regarding time constraints or real-time search tasks (e.g., microblog search and news search).

We propose here to enhance the LL4IR framework with a real-time ranking component, called "Living Ranking". The latter enables to deliver a real-time ranking for fresh queries while maintaining the simplicity of LL4IR framework. In particular, participants need to provide, instead of offline rankings, algorithms that will be executed in real time. Unlike NewsReel framework [1], the proposed extension does not require any infrastructure to be deployed by participants.

2 Extending LL4IR with Living Ranking

The "Living Ranking" is an extension component pluggable to LL4IR framework, as illustrated in Figure reffig:ll4R-extention. In contrast to offline rankings that must be provided in advance through LL4IR participant's API, the new component stands as a new source that provides API with rankings generated on-the-fly for each online submitted query while maintaining the initial framework overflow.

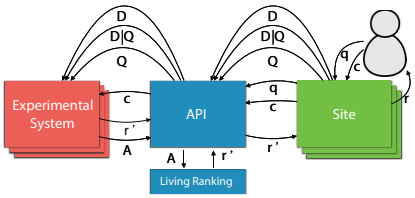

Fig. 1. LL4IR Framework extension. "A" illustrates participant's algorithm.

To do so, participants must provide a ranking algorithm which may be executed online via the Living Ranking component. Ranking algorithms provided by participants may respect a standard interface with well-defined input and output formats. For instance, the format of the rankings issued from the Living Ranking component could be structured as the one currently required to participants. We outline that this architecture allows to restrict the visibility of real-time submitted queries and eventually of documents, which avoid bias in the algorithm design and gives more credibility to evaluation results. However, this component might be resource-consuming. One solution could be to execute ranking algorithms on demand, for instance when changes occurred on the result set. Such on-demand strategy may balance between efficiency and effectiveness.

The integration of the Living Ranking component within the LL4IR framework suggests some changes or brings further enhancements detailed below:

- *Framework architecture:* Living Ranking should offer a flexible interface so participants can easily implement their algorithm without requiring complex infrastructure for all tiers. We suggest implementing algorithms in sandbox-based scripts (i.e., JavaScript) that support online execution under strict constraints.

- *Challenge Organization:* Since test queries and produced rankings may not be visible, we suggest to introduce a *debugging phase* with simulated queries, standing before uploading ranking algorithms. This would help participants to validate the effectiveness and efficiency of their algorithms.

- *Evaluation Metric:* Living Ranking components allow to produce additional evaluation metrics in terms of algorithm computation resources (e.g., execution time and used memory). Although this type of metric is not commonly used in IR, we think that such metrics are relevant for evaluating real-time IR models.

3 Conclusion

We propose in this paper to extend the LL4IR framework with a Living Ranking component in the aim of providing an evaluation framework for real-time ranking. This approach may add some technical complexity. We are also aware of the additional efforts to be deployed for benchmark organization but we believe that the proposed extension would open LL4IR to other retrieval tasks that attract a lot of interest in IR community, namely real-time search tasks.

References

[1] B. Kille, A. Lommatzsch, R. Turrin, A., M. Larson, T. Brodt, J. Seiler, and F. Hopfgartner. Stream-based recommendations: Online and offline evaluation as a service. In *CLEF 2015*, 2015.

[2] A. Schuth, K. Balog, and L. Kelly. Overview of the living labs for information retrieval evaluation (ll4ir) clef lab 2015. In *CLEF 2015*, 2015.