Living Labs for Online Evaluation: From Theory to Practice

Anne Schuth[1(✉)] and Krisztian Balog[2]

[1] University of Amsterdam, Amsterdam, The Netherlands
anne.schuth@uva.nl
[2] University of Stavanger, Stavanger, Norway
krisztian.balog@uis.no

Abstract. Experimental evaluation has always been central to Information Retrieval research. The field is increasingly moving towards *online* evaluation, which involves experimenting with real, unsuspecting users in their natural task environments, a so-called *living lab*. Specifically, with the recent introduction of the Living Labs for IR Evaluation initiative at CLEF and the OpenSearch track at TREC, researchers can now have direct access to such labs. With these benchmarking platforms in place, we believe that online evaluation will be an exciting area to work on in the future. This half-day tutorial aims to provide a comprehensive overview of the underlying theory and complement it with practical guidance.

1 Motivation and Overview

Experimental evaluation has always been a key component in Information Retrieval research. Most commonly, systems are evaluated following the Cranfield methodology [4,22]. Using this approach, systems are evaluated in terms of document relevance for given queries, which is assessed by trained experts. While the Cranfield methodology ensures high internal validity and repeatability of experiments, it has been shown that the users' search success and satisfaction with an IR system are not always accurately reflected by standard IR metrics [29,31]. One reason is that the relevance judges typically do not assess queries and documents that reflect their own information needs, and have to make assumptions about relevance from an assumed user's point of view. Because the true information need can be difficult to assess, this can cause substantial biases [11,30,34]. To address these shortcomings, the field is increasingly moving towards *online* evaluation, which involves experimenting with real, unsuspecting users in their natural task environments. Essentially, the production search engine operates as a "living lab." For a long time, this type of evaluation was only available to those working within organizations that operate a search engine. But this is about to change. For one thing, the need to involve real users is know openly and widely acknowledged in our community (as witnessed, e.g., by the panel discussion at ECIR'15 and the Salton Award keynote lecture of Belkin at SIGIR'15 [2]). For another thing, pioneering efforts to realize the idea of living

N. Ferro et al. (Eds.): ECIR 2016, LNCS 9626, pp. 893–896, 2016.
DOI: 10.1007/978-3-319-30671-1_88

labs in practice are now in place and are available to the community. Specifically the Living Labs for IR Evaluation (LL4IR)[1] initiative runs as a benchmarking campaign at CLEF, but also operates monthly challenges so that people do not have to wait for a yearly evaluation cycle. The most recent initiative is the OpenSearch track at TREC[2], which focuses on *academic literature search*.

Understanding the differences between online and offline evaluation is still a largely unexplored area of research. There is a lot of fundamental research to happen in this space that has not happened yet because of the lack availability of experimental resources to the academic community. With recent developments, we believe that online evaluation will be an exciting area to work on in the future. The motivation for this tutorial is twofold: (1) to raise awareness and promote this form of evaluation (i.e., online evaluation with living labs) in the community, and (2) to help people get started by working through all the steps of the development and deployment process, using the LL4IR evaluation platform.

This half-day tutorial aims to provide a comprehensive overview of the underlying theory and complement it with practical guidance. The tutorial is organized in two 1,5 hours sessions with a break in between. Each session interleaves theoretical, practical, and interactive elements to keep the audience engaged. For the practical parts, we break with the traditional format by using hands-on instructional techniques. We will make use of an online tool, called DataJoy,[3] that proved invaluable in our previous classroom experience. This allows participants to (1) run Python code in a browser window without having to install anything locally, (2) follow the presenter's screen on their own laptop and, (3) at the same time, have their own private copy of the project on a different browser tab.

2 Target Audience and Learning Objectives

The primary target audience are graduate students and lecturers/professors teaching IR classes. Engineers from companies operating search engines might also find the tutorial useful. Our learning objectives include the following topics.

We will start our tutorial with an extensive overview of online evaluation methods. We begin with *A/B Testing* [16], which compares two systems by showing system A to one group of users and system B to another group. A/B testing then tries to infer a difference between the systems from differences in observed behavior. We describe many ways of measuring observed behavior: (1) click through rate (CTR) [14]; (2) dwell time [34]; (3) satisfied clicks [15]; (4) tabbed browsing [13]; (5) abandonment [18,28]; (6) query reformulation [8]; (7) skips [32]; (8) mouse movement [5–7,10,33]; and (9) in-view time [17].

While providing flexibility and control, A/B comparisons typically require a large number of observations. *Interleaved comparison methods* reduce the variance of measurement by presenting users with a result list that combines the

[1] http://living-labs.net.
[2] http://trec-open-search.org/.
[3] http://getdatajoy.com.

rankings of systems A and B. We provide a comprehensive overview of the following interleaving methods: (1) balanced interleave (BI) [14]; (2) team draft interleave (TDI) [21]; (3) document constraints (DC) [9]; (4) probabilistic interleave (PI) [12]; (5) optimized interleave (OI) [20]; (6) team draft multileave (TDM) [27]; and (7) probabilistic multileave (PM) [24].

Next, we discuss a comparison of interleaving and A/B metrics [25]. We then turn to simulating user interactions [26] using *click models* [3]. Finally, we touch on *learning to rank* in two variants: *offline* learning to rank [19] and *online* learning to rank [35], of which the latter requires the aforementioned evaluation methods.

Having provided the necessary theoretical background, we introduce the *living labs for IR* (LL4IR) [1] evaluation platform in depth. We will focus on two specific use-cases [23] from the CLEF lab: product search and web search. the practical sessions, participants will gain hands-on experience with the LL4IR platform [1], which includes: (1) registering and obtaining an API key; (2) getting queries and candidate items; (3) generating and uploading a ranking; and (4) obtaining feedback and outcomes. API documentation and course material are available at http://living-labs.net.

References

1. Balog, K., Kelly, L., Schuth, A.: Head first: living labs for ad-hoc search evaluation. In: CIKM 2014, pp. 1815–1818. ACM Press, New York, USA, November 2014
2. Belkin, N.J.: Salton award lecture: people, interacting with information. In: Proceedings of 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2015, pp. 1–2. ACM (2015)
3. Chuklin, A., Markov, I., de Rijke, M.: Click Models for Web Search. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, San Rafael (2015)
4. Cleverdon, C.W., Keen, M.: Aslib Cranfield research project-factors determining the performance of indexing systems; Volume 2, Test results, National Science Foundation (1966)
5. Diaz, F., White, R., Buscher, G., Liebling, D.: Robust models of mouse movement on dynamic web search results pages. In: CIKM, pp. 1451–1460. ACM Press, October 2013
6. Guo, Q., Agichtein, E.: Understanding "abandoned" ads: towards personalized commercial intent inference via mouse movement analysis. In: SIGIR-IRA (2008)
7. Guo, Q., Agichtein, E.: Towards predicting web searcher gaze position from mouse movements. In: CHI EA, 3601p, April 2010
8. Hassan, A., Shi, X., Craswell, N., Ramsey, B.: Beyond clicks: query reformulation as a predictor of search satisfaction. In: CIKM (2013)
9. He, J., Zhai, C., Li, X.: Evaluation of methods for relative comparison of retrieval systems based on clickthroughs. In: CIKM 2009, ACM (2009)
10. He, Y., Wang, K.: Inferring search behaviors using partially observable markov model with duration (POMD). In: WSDM (2011)
11. Hersh, W., Turpin, A.H., Price, S., Chan, B., Kramer, D., Sacherek, L., Olson, D.: Do batch and user evaluations give the same results? In: SIGIR, pp. 17–24 (2000)
12. Hofmann, K., Whiteson, S., de Rijke, M.: A probabilistic method for inferring preferences from clicks. In: CIKM 2011, ACM (2011)

13. Jeff, H., Thomas, L., Ryen, W.: No search result left behind. In: WSDM, 203p (2012)
14. Joachims, T., Granka, L.A., Pan, B., Hembrooke, H., Radlinski, F., Gay, G.: Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM Trans. Inf. Syst. **25**(2), 7 (2007)
15. Kim, Y., Hassan, A., White, R., Zitouni, I.: Modeling dwell time to predict click-level satisfaction. In: WSDM (2014)
16. Kohavi, R.: Online controlled experiments: introduction, insights, scaling, and humbling statistics. In: Proceedings of UEO 2013 (2013)
17. Lagun, D., Hsieh, C.H., Webster D., Navalpakkam, V.: Towards better measurement of attention and satisfaction in mobile search. In: SIGIR (2014)
18. Li, J., Huffman, S., Tokuda, A.: Good abandonment in mobile and pc internet search. In: SIGIR 2009, pp. 43–50 (2009)
19. Liu, T.-Y.: Learning to Rank for Information Retrieval. Springer, Heidelberg (2011)
20. Radlinski, F., Craswell, N.: Optimized interleaving for online retrieval evaluation. In: WSDM 2013, ACM (2013)
21. Radlinski, F., Kurup, M., Joachims, T.: How does clickthrough data reflect retrieval quality? In: CIKM 2008, ACM (2008)
22. Sanderson, M.: Test collection based evaluation of information retrieval systems. Found. Trends Inf. Retrieval **4**(4), 247–375 (2010)
23. Schuth, A., Balog, K., Kelly, L.: Overview of the living labs for information retrieval evaluation (ll4ir) clef lab. In: Mothe, J., Savoy, J., Kamps, J., Pinel-Sauvagnat, K., Jones, G.J.F., SanJuan, E., Cappellato, L., Ferro, N. (eds.) CLEF 2015. LNCS, pp. 484–496. Springer, Heidelberg (2015)
24. Schuth, A., Bruintjes, R.-J., Büttner, F., van Doorn, J., Groenland, C., Oosterhuis, H., Tran, C.-N., Veeling, B., van der Velde, J., Wechsler, R., Woudenberg, D., de Rijke, M.: Probabilistic multileave for online retrieval evaluation. In: Proceedings of SIGIR (2015)
25. Schuth, A., Hofmann, K., Radlinski, F.: Predicting search satisfaction metrics with interleaved comparisons. In: SIGIR 2015 (2015)
26. Schuth, A., Hofmann, K., Whiteson, S., de Rijke, M.: Lerot: an online learning to rank framework. In: LivingLab 2013, pp. 23–26. ACM Press, November 2013
27. Schuth, A., Sietsma, F., Whiteson, S., Lefortier, D., de Rijke, M.: Multileaved comparisons for fast online evaluation. In: CIKM 2014 (2014)
28. Song, Y., Shi, X., White, R., Hassan, A.: Context-aware web search abandonment prediction. In: SIGIR (2014)
29. Teevan, J., Dumais, S., Horvitz, E.: The potential value of personalizing search. In: SIGIR, pp. 756–757 (2007)
30. Turpin, A., Hersh, W.: Why batch and user evaluations do not give the same results. In: SIGIR, pp. 225–231 (2001)
31. Turpin, A., Scholar, F.: User performance versus precision measures for simple search tasks. In: SIGIR, pp. 11–18 (2006)
32. Wang, K., Walker, T., Zheng, Z.: PSkip: estimating relevance ranking quality from web search clickthrough data. In: KDD, pp. 1355–1364 (2009)
33. Wang, K., Gloy, N., Li, X.: Inferring search behaviors using partially observable Markov (POM) model. In: WSDM (2010)
34. Yilmaz, E., Verma, M., Craswell, N., Radlinski, F., Bailey, P.: Relevance and effort: an analysis of document utility. In: CIKM (2014)
35. Yue, Y., Joachims, T.: Interactively optimizing information retrieval systems as a dueling bandits problem. In: ICML 2009, pp. 1201–1208 (2009)