Head First: Living Labs for Ad-hoc Search Evaluation

Krisztian Balog
University of Stavanger
Norway
krisztian.balog@uis.no

Liadh Kelly
CNGL, School of Computing
Dublin City University
Ireland
lkelly@computing.dcu.ie

Anne Schuth
University of Amsterdam
The Netherlands
anne.schuth@uva.nl

ABSTRACT

The information retrieval (IR) community strives to make evaluation more centered on real users and their needs. The living labs evaluation paradigm, i.e., observing users in their natural task environments, offers great promise in this regard. Yet, progress in an academic setting has been limited. This paper presents the first living labs for the IR community benchmarking campaign initiative, taking as test two use-cases: local domain search on a university website and product search on an e-commerce site. There are many challenges associated with this setting, including incorporating results from experimental search systems into live production systems, and obtaining sufficiently many impressions from relatively low traffic sites. We propose that head queries can be used to generate result lists offline, which are then interleaved with results of the production system for live evaluation. An API is developed to orchestrate the communication between commercial parties and benchmark participants. This campaign acts to progress the living labs for IR evaluation methodology, and offers important insight into the role of living labs in this space.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

Keywords

Evaluation; living labs

1. INTRODUCTION

The Cranfield methodology [9] introduced a way to enable cross-comparable evaluation of information retrieval (IR) systems, using a document collection, queries, and relevance assessments. Since then researchers have strived to make IR evaluations more "realistic," i.e., centered on real users, their needs, and behaviors. Living labs have been proposed as a way for researchers to perform *in situ* evaluations, with real users performing real tasks using real-world applications [13]. This concept has already been used for a number of years as an important instrument for technology development in industrial settings; for example, A/B testing procedures are employed heavily by major web search providers [15]. This form of evaluation, however, is currently only available to those working at the said organizations.

"The basic idea of living labs for IR is that rather than individual research groups independently developing experimental search infrastructures and gathering their own groups of test searchers for IR evaluations, a central and shared experimental environment is developed to facilitate the sharing of resources" [5]. The potential benefits of living labs to the IR community are profound, including the availability of interaction and usage data for researchers and greater knowledge transfer between industry and academia [6]. Progress towards realizing actual living labs, in an academic setting, has nevertheless been limited. Azzopardi and Balog [5] discuss a number of search and recommendation tasks in an online shopping environment and present an idealized architecture based on web services. There are many challenges associated with operationalizing these ideas, including architecture, hosting, maintenance, security, privacy, participant recruiting, and scenarios and tasks for use development [5]. A recent development in this space was the Living Labs for IR Evaluation workshop at CIKM 2013 [6]. A key outcome of this workshop was the need to work towards community-driven living labs benchmarking initiatives.

In this paper we present a living labs for IR evaluation benchmarking platform. We propose that mid-sized organizations that lack their own R&D department are good potential collaborators, as they have the opportunity to gain much improved retrieval approaches. We present two specific use-cases for ad-hoc search: local domain search on a university website and product search on an e-commerce site. These use-cases represent a setting with at least two major challenges: (i) relatively low search volume (especially compared to major web search providers) and (ii) means to facilitate experimentation by "third parties" in live, production systems. We postulate that focusing on *head queries* (i.e., queries most frequently issued) can help overcome these challenges. The choice of head queries is critical because it removes a harsh requirement of providing rankings in real-time for query requests. Instead, experimental search systems (developed by benchmark participants) can generate ranked results lists for these queries offline. These participant rankings can then be used by the live system when head queries are next issued. Finally, feedback is made available to experimental search systems to facilitate improved offline ranking generation. Data exchange between live systems and participants is facilitated by a web-based API.

In summary, the main contributions of this work include the development of evaluation methodology, architecture, specific use-cases, as well as the implementation of the Living Labs API, made

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2598-1/14/11 ...$15.00
http://dx.doi.org/10.1145/2661829.2661962.

available as open source software.[1] An open challenge, with agreements with the use-case organizations in place, is currently being organized.[2] The outcomes of this benchmarking initiative will lead to answers to the following two important research questions: (RQ1) Are system rankings different when using historical clicks from those using online experiments? (RQ2) Are system rankings different when using manual relevance assessments ("expert judgments") from those using online experiments? These answers will provide the research community with concrete insight into the need, or lack thereof, of living labs as an additional tool for IR evaluation.

The remainder of the paper is organized as follows. In §2 we briefly discuss related work. Next, in §3, we introduce our evaluation platform and methodology. We present two particular use-cases in §4. Limitations and directions for future research are discussed in §5. Finally, we conclude in §6.

2. RELATED WORK

The need for more realistic evaluation, involving real users, was reiterated at recent IR workshops [1, 6, 12]. Approaches that attempt to incorporate user behavior into batch-style evaluations can be divided into two main categories. One is to create effectiveness measures that better model user behavior, e.g., [7, 10]. Another approach is to simulate user behavior and then validate these models against actual usage data, e.g., [3, 4]. These ideas have been implemented in a number of community benchmarking efforts, including the TREC Interactive, HARD, and Session tracks, and the INEX Interactive track. While user simulation is a great instrument for fine-tuning systems, it cannot substitute the user. Crowdsourcing, using e.g., Mechanical Turk, enables the sourcing of individuals in the online community to perform various relevance assessment and annotation tasks [2]. However, these individuals do not constitute *real* users performing *real* tasks driven by a *real* information need. Living labs offer this potential.

The living labs notion was first proposed in the information-seeking support space (ISSS) by Kelly et al. [13]: "Such a lab might contain resources and tools for evaluation as well as infrastructure for collaborative studies. It might also function as a point of contact with those interested in participating in ISSS studies." Azzopardi and Balog [5] provided greater insight into what this might be in the IR space: "A living lab would provide a common data repository and evaluation environment giving researchers (in particular from academia) the data required to undertake meaningful and applicable research." Kelly et al. [14] then showed a practical interpretation of this for personal desktop search. However, to date, there have been no attempts at operationalizing a living labs benchmark in the IR space. The nearest to this has been the 2014 CLEF NEWSREEL lab[3] and the Plista contest,[4] addressing the problem of news recommendation. Participants are expected to implement their recommender system as a service that can handle a large number of (recommendation) requests. Their response to a request is shown to a user and resulting clicks are then made available to participants so that they can update their system. One major difference between this and our proposal is the task itself: we are focusing on retrieval as opposed to recommendation. There are also important architectural differences stemming from the nature of our experimental environment; in our setup participants do not get full control over the results shown to the user, they are always interleaved with that of the production system.

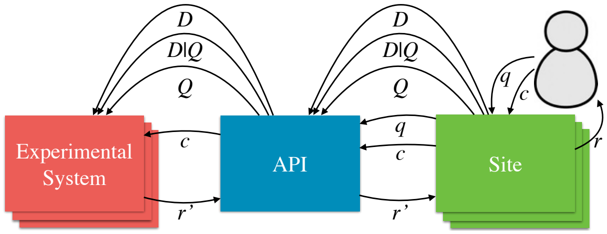

Figure 1: Schematic representation of the Living Labs API.

3. EVALUATION PLATFORM AND METHODOLOGY

The overall objective of our work is to design, implement, and operate an evaluation platform that allows researchers to test retrieval methods in live search environments with real users. A further important desideratum is the availability of usage and interaction data for model training and development. We are particularly interested in developing a solution that caters for medium-sized organizations as participating partners—those with a fair (but not excessive) amount of search volume on their websites, and without their own R&D department. To keep focus, we concentrate on one specific retrieval task: ad-hoc search.

There are certain restrictions to this setting that need to be addressed. First, search volume poses limitations on how much room there is for experimentation. For a fair comparison, systems should be evaluated on the same set of queries and ensured a minimum number of impressions. One of the key design decisions in our approach is to focus exclusively on head queries; this not only makes it easier to plan with and have control over the capacity of experimental resources, but also offers the availability of considerable amounts of historical log data. Moreover, it allows for rapid response time requirements, as rankings can be computed offline, before an actual instance of the query is issued (see §3.2). Second, experimentation takes place in production environments, where it is vital to maintain a certain level of quality of service, both in terms of efficiency (response time) and effectiveness (search relevancy). Organizations are unlikely to give up control of the whole search result page (SERP) and to blindly take ranked lists from experimental search systems. That is why the common online evaluation approach, A/B testing, cannot be used here. Also, given the traffic volume of these websites, it would take a long time to get reliable evaluation results using A/B testing. To overcome this, we propose interleaving experimental search results with those of the production system (see §3.3). Efficiency considerations are addressed head-on by the proposed architecture (see §3.1).

3.1 Architecture

The normal flow of information in a search system is as follows: the end user issues query q on the site, which responds with a ranking r; the clicks c made by the user are recorded by the site. A living lab for IR requires a means of transferring this information (queries, rankings, and feedback) between end users of the site and *experimental search systems*. We propose an architecture in the form of an API for this purpose that encapsulates data storage and access via HTTP calls. All communication between sites and experimental search systems is done via this API (i.e., experimental systems can not directly interact with sites); see Figure 1.

The set of (frequent) queries that are subject to experimentation (see §3.2) is denoted as Q. For each query q in Q, the site makes available (i) a set of candidate documents, (ii) the contents of these documents, and (iii) historical interaction data to the API. Addi-

[1] http://git.living-labs.net
[2] http://living-labs.net/challenge
[3] http://www.clef-newsreel.org
[4] http://contest.plista.com

tionally, sites can also provide collection-level statistics (such as document and collection frequencies of terms). Benchmark participants can obtain this information from the API. Then, for each query q in the fixed set of queries Q, the experimental search system generates a ranking r' that is submitted to the API. When a user issues a query q to the site that is in Q, the site requests a ranking r' from an experimental system through the API. This ranking r' is interleaved (see §3.3) with the production system and presented to the user. The user interactions with the interleaved list are sent back to the API and made available to the participant that contributed r'. Subsequently, this interaction data c can be used to update the ranking r', for one or all queries in Q.

3.2 Queries

The distribution of search queries in web search typically follows a power law [17], where a relatively small set of head queries are frequently posed by many users and there is a long tail of queries that appear in the logs only a few times (often only once). Here, we focus exclusively on head queries for a number of reasons: (i) this allows us to evaluate experimental search systems on the same set of queries, (ii) these queries have a stable volume level, even for mid-sized sites (cf. §4), and (iii) historical click and usage data is available in meaningful quantities. We take a simple measure to filter out "uninteresting" queries: queries for which virtually all historical clicks are associated with a single document are removed.

3.3 Interleaving

Prior work has shown that *interleaving* [8] can produce very reliable comparisons of rankers in online evaluation [16] using much less data than A/B testing. Such interleaved comparison methods take as input two rankings for the same query, and produce as output a combined result list to be shown to the user. The resulting clicks from this user are then interpreted by the interleaving method to decide on a winning ranker. There are several interleaving methods available. *Balanced interleave* (BI) [11] randomly selects a ranker to contribute the first document. Then, alternating, each ranker contributes its next document. A document is added to the interleaving only if it is not yet present there. BI, however, can produce biased results: when two very similar rankers are compared, it can favor one ranker regardless of actual user preferences expressed in clicks. This bias was removed in *team draft* (TD) [16], which makes it the most commonly used interleaving method in industry and, therefore, also our interleaving method of choice.

3.4 Evaluation metrics

We use two forms of online metrics, relative and absolute. The fraction of wins of interleaved comparisons for an experimental system against the production system is a *relative* metric and this is used as our overall evaluation criteria [15]. As shown in [16], relevance can also be inferred from user actions and used to compute *absolute click metrics*. We propose using several standard IR metrics (i.e., nDCG@10, ERR, MAP) as absolute click metrics. These are computed on raw clicks, but also on more reliable (but inherently more sparse) subsets of the clicks: clicks with long dwell times, last clicks, and deepest clicks. Additionally, we compute traditional metrics (again, nDCG@10, ERR, MAP) on assessments from professional human assessors. We refer to these as *offline evaluation metrics*. These offline metrics are merely used for analysis, such that we can answer research question RQ2 (cf. §1).

3.5 Benchmark organization

Evaluation is split into training and test phases. For each of the two use cases described in §4, sets of 50 training and 50 test queries (following the TREC best practice), along with a set of candidate documents to be ranked for each query are provided.

Participants are allowed to partake in the benchmark with a single system. The challenge is to optimally rank the set of candidate documents provided for each query. Participants submit their current ranking to the benchmark API, which makes these rankings available to sites when requested. Note that experimental rankings from the participant are interleaved with the production ranking before presenting results to user (see §3.1). The benchmark operates by giving participants equal number of impressions for queries over both the training and test phases. During the training phases, participants can receive feedback on their own system (such as dwell time, click throughs, and the actual interleaved result list presented to the user), which they can use to refine their system. For the test phase, participants are provided with a two week window in which they can download the training queries and document collections. Once downloaded, participants must submit their rankings for the test queries within 24 hours of downloading the dataset.

We provide a dashboard for monitoring progress of participants. A "leaderboard" is also present with both the relative and absolute online metrics, as described in §3.4. During the training phase, the dashboard will not display offline relevance metrics (metrics that were computed on relevance assessments from human assessors). This is to avoid participants optimizing for this metric instead of learning from user feedback.

4. USE-CASES

We consider ad-hoc search in two different flavors, taking place on the websites of medium-sized organizations: (A) local domain search on the website of the University of Amsterdam[5] and (B) product search in the webshop of a toy retailer operating in Hungary.[6] These organizations have agreed to partake in our challenge and allow for experimentation with head queries.

Queries. Figure 2 shows the historical query counts for head (top 100) queries, for the period 2014-01-01 to 2014-08-20, on a daily basis. The absolute count is the total number of times head queries are issued. Relative click count refers to the fraction of the total query volume that falls on the head queries. There are interesting differences in the characteristics of the two sites; head queries constitute on average 63% of the overall search volume for local domain search, while it is only 25% for product search. There is an order of magnitude difference in absolute terms: $14,500$ vs. $1,500$ queries per day on average for use-cases (A) and (B), respectively. Importantly for us, the plots clearly show that there is a stable volume level on the head queries to experiment with ($9,600$ per day for use-case (A) and 380 per day for use-case (B), on average).

Content. Both sites make available (i) the contents of candidate documents, (ii) relative historical click counts for these documents, and (iii) collection-level term statistics (document and collection frequencies). Use-case (A) represents a rather straightforward document search task; for each (HTML) document, the title and the (cleansed) body are made available. Documents in use-case (B) correspond to products. For each product a fielded representation is provided, including the product's name, description, brand, price (and bonus price, if applicable), product categories, (URLs of) product photos, and date of addition. The product categorization system (a 2-levels deep hierarchy) is also offered.

[5] http://uva.nl

[6] http://regiojatek.hu

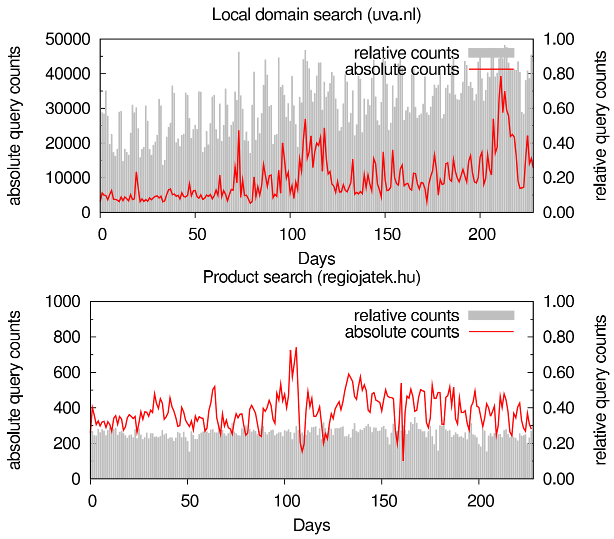

Figure 2: Historical query counts for the challenge use-cases. (Top): local domain search, (Bottom): product search.

Feedback. Feedback includes (i) click throughs, (ii) dwell time, and (iii) the actual interleaved result list presented to the user. Further, for use-case (B), feedback information also includes basket operations and actual purchases made.

5. LIMITATIONS AND FUTURE CONSIDERATIONS

Our proposal is a first of its kind and represents an important step towards making the living labs evaluation paradigm accessible to the wider IR research community. Nevertheless, it is not without limitations. Next, we briefly consider some of these limitations and look at ways in which they could be addressed.

L1) *Head queries only.* While head queries constitute a considerable portion of a site's traffic, they are representative of only one type of request, that is, popular information needs.

L2) *Lack of context.* The search algorithm has no knowledge of the searcher's context, such as location, previous searches, etc. This means that currently there is no room for personalization of results.

L3) *No real-time feedback.* While the proposed API does provide detailed feedback, it is not immediate. Thus, it cannot directly be used in the given search session.

L4) *Limited control.* Experimentation is limited to single searches, where results are interleaved with those of the production system. I.e., there is no control over the entire result list.

L5) *Ultimate measure of success.* Having better search facilities is usually only a means to an end—it is not the ultimate goal. E.g., in the e-commerce case the ultimate measure of success (from the company's perspective) is the profit made on purchases. Evaluation metrics should reflect this overall goal.

L1–L4 could be overcome by a live architecture, in which control is given to benchmark participants over entire sessions, with real-time access to context and feedback. However, it is still a very much open question how to ensure availability, response time, and quality of the experimental methods in production environments. Safety mechanisms are needed for "experiment shutdown" in which case methods can default back to the production system. L5 could be addressed by providing an "utility" score for documents (products); this could already be done with the existing architecture.

6. CONCLUSIONS

Living labs offer the IR community great potential to evaluate their approaches in *live* settings with *real* users. In this paper we presented the first practical methodology and implementation of a living labs for IR benchmarking campaign that is currently being organized, with local domain search and product search as use-cases. Efforts are underway to recruit additional organizations to join our initiative. While significant, this is just the beginning of the practical living labs for IR evaluation story. The results of our research questions, to be answered by this campaign, will yield further light on the living labs for IR paradigm. We expect that this work will pave the way for further progress in this exciting direction for IR evaluation.

Acknowledgments. This research was partially supported by the Evaluating Information Access Systems (ELIAS) ESF Research Networking Programme and the European Community's Seventh Framework Programme (FP7/ 2007-2013) under grant agreement nr 288024 (LiMoSINe project).

References

[1] J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1):2–32, 2012.

[2] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[3] P. Arvola, J. Kekäläinen, and M. Junkkari. Expected reading effort in focused retrieval evaluation. *Inf. Retr.*, 13(5):460–484, 2010.

[4] L. Azzopardi. The economics in interactive information retrieval. In *Proc. of SIGIR '11*, 2011.

[5] L. Azzopardi and K. Balog. Towards a living lab for information retrieval research and development. A proposal for a living lab for product search tasks. In *Proc. of CLEF '11*, 2011.

[6] K. Balog, D. Elsweiler, E. Kanoulas, L. Kelly, and M. D. Smucker. Report on the CIKM workshop on living labs for information retrieval evaluation. *SIGIR Forum*, 48(1), 2014.

[7] B. Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proc. of SIGIR '11*, 2011.

[8] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. Inf. Syst.*, 30(1), 2012.

[9] C. Cleverdon and K. E. Factors Determining the Performance of Indexing Systems (Volume 1: Design; Volume 2: Results). Technical report, Cranfield, UK, 1966.

[10] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[11] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Trans. Inf. Syst.*, 25(2), 2007.

[12] J. Kamps, S. Geva, C. Peters, T. Sakai, A. Trotman, and E. Voorhees. Report on the SIGIR 2009 workshop on the future of IR evaluation. *SIGIR Forum*, 43(2):13–23, 2009.

[13] D. Kelly, S. Dumais, and J. O. Pedersen. Evaluation challenges and directions for information-seeking support systems. *Computer*, 42 (3):60–66, 2009.

[14] L. Kelly, P. Bunbury, and G. J. F. Jones. Evaluating personal information retrieval. In *Proc. of ECIR '12*, 2012.

[15] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

[16] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Proc. of CIKM '08*, 2008.

[17] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *J. Am. Soc. Inf. Sci. Technol.*, 52(3):226–234, 2001.