Report on the CIKM Workshop on Living Labs for Information Retrieval Evaluation

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

David Elsweiler
University of Regensburg
Germany
david@elsweiler.co.uk

Evangelos Kanoulas
Google Inc.
Switzerland
ekanoulas@gmail.com

Liadh Kelly
Dublin City University
Ireland
lkelly@computing.dcu.ie

Mark D. Smucker
University of Waterloo
Canada
msmucker@uwaterloo.ca

Abstract

Evaluation is a central aspect of information retrieval (IR) research. In the past few years, a new evaluation methodology known as living labs has been proposed as a way for researchers to be able to perform in-situ evaluation. The first CIKM workshop on Living Labs for IR evaluation (LL'13) was held on 1st November 2013 in San Francisco, USA. The workshop consisted of an industrial keynote, four oral paper presentations, three demo presentations, and a discussion session. This report presents an overview of the scope and contents of the workshop and outlines the major outcomes.

1 Introduction

In the past few years the information retrieval (IR) community has been exploring ways to move further away from the Cranfield-style evaluation paradigm, and make evaluations more "realistic" (more centered on real users, their needs and behaviours). As part of this drive, living labs which involve and integrate users in the research process have been proposed. 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. These would, not only, enable the capture of real interaction and usage data, but also provide a context for testing and evaluating IR models, methods and systems. Kelly et al. [3] outlined what this might be for information-seeking support systems (ISSS):

> *A living laboratory on the Web that brings researchers and searchers together is needed to facilitate ISSS (Information-Seeking Support System) evaluation. 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 [1] elaborated further on the benefits of living labs for IR:

> *Living labs are seen as a way to bridge the data divide within the research community, because currently interaction data is often only available to those working within organizations that provide real world IR applications. 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. More generally though, a living lab has been presented not just as a platform for collaborative research, but also as a platform where users co-create the product, application or service (i.e., users are not just subjects of observation, but also part of the creation). Essentially, the users explore emerging ideas and scenarios in-situ, the evaluation process is then fed back into the design of the product to further enhance their user experience.*

In general, living labs would offer huge benefits to the community, such as: availability of, potentially larger, cohorts of real users and their behaviours, e.g., querying behaviours, for experimental purposes; cross-comparability across research centres; greater knowledge transfer between industry and academia, when industry partners are involved. The need for this methodology is further amplified by the increased reliance of IR approaches on proprietary data; living labs are a way to bridge the data divide between academia and industry [2]. Progress towards realizing actual living labs has nevertheless been limited. The most notable contribution being that of Azzopardi and Balog [1], where a possible architecture for product search tasks in an e-commerce setting is presented. However, their idea has not been operationalized yet. There are many challenges to be overcome before the benefits associated with living labs for IR can be realized, including challenges associated with living labs architecture and design, hosting, maintenance, security, privacy, participant recruiting, and scenarios and tasks for use development.

The aim of the CIKM workshop on Living Labs for Information Retrieval Evaluation was to further develop the living labs for information retrieval evaluation paradigm and formulate practical next steps for post-workshop progression. Issues include implementation options, how to make it attractive to commercial organizations, alternatives when commercial providers will not get involved, coping with data privacy issues, and tasks and usage scenarios.

Papers submitted to the workshop were reviewed by an international program committee. Two short papers, two position papers, and three demo papers were accepted for presentation at the workshop and inclusion in the workshop proceedings. Some of the presentations made at the workshop are available on the workshop website: `http://ll2013.dcu.ie/`. The workshop proceedings are available at: `http://dl.acm.org/citation.cfm?id=2513150`.

What follows is our personal interpretation of the workshop activities, including the presented papers and the various discussion sessions. We conclude by summarizing the main points raised during the workshop, the main achievements during the day and the open points for future investigation.

2 Keynote – Georg Buscher, Microsoft Bing

The keynote talk of the workshop was given by Georg Buscher, senior researcher at Bing, and was entitled *IR Evaluation: Perspectives From Within a Living Lab*. In the past, during his PhD studies, Buscher performed several lab studies with users. Now, he leads the online metrics team at Bing, where he gets to experiment with millions of users. This puts him in a unique position where he has the expertise and perspective, both in academic and in industrial contexts.

Lab studies provide a more realistic setting than using offline judgments (i.e., the traditional TREC setup) while still allowing for controlled experiments. Nevertheless, lab studies are still artificial, given that users are observed in a lab, outside their natural environment. Moreover, lab studies are costly and do not scale. Living labs, on the other hand, offer a perfectly realistic setting; most users are not even aware that they are the subject of experimentation (laboratory guinea pigs). Importantly, information needs are not only real, but are also representative. Living labs scale very well and make it possible to perform evaluations on millions of users. Despite the attractive opportunities living labs offer, there are a number of challenges involved.

First, experiments in a living lab must not be destructive and need to meet a minimum quality bar. There are procedures to ensure this: (i) running offline evaluation, on representative query sets, before starting online experimentation, (ii) piping real historic traffic through the experimentation system, to check for both back-end and front-end errors, and (iii) alerting early experiment shutdown, if metrics do not stay within certain bounds.

Second, complex systems can produce unexpected side effects. Search result pages, in Bing, are composed by a layered stack of modules, where changes in modules lower down in the stack may have upstream effects. This means that a small degradation lower down the stack might be amplified into a large degradation on the whole page. Therefore, it is vital to understand whether and what side effects happened, to be able to adequately interpret the eventual experiment log data. In effect, many living laboratory experiments can fail to be controlled. The experimenter may attempt to change one variable, but in fact, the experimenter changes many variables. Herein lies the trade-off between traditional lab studies and living labs. While traditional lab studies are controlled, they lack full reality, and while living labs are fully real, they may lack control.

Third, online experimentation requires different metrics than those used in offline evaluation; there is no ground truth data anymore, only user interactions. There are different types of online metrics with different applicability: (i) *feature-specific metrics* (e.g., for result ranking, result snippet generation, query auto-completion, etc.) target specific features with built-in assumptions about what good/successful interactions look like, while ignoring other (important) aspects of the overall search experience; an improvement in a feature-specific metric can regress a metric on a higher level; (ii) *user utility metrics* are specific to the service (i.e., the search engine) but are mostly oblivious to page composition/features; the basic assumption is that clicks are "good" and more effort (time or queries) is "bad," but there are exceptions (for example, satisfaction is not observable when the user abandons the page because her information need has been answered by a rich snippet); (iii) *retention metrics* generally applicable; they do not make service-specific assumptions and are not subject to inherent metric trade-offs; on the flip side, they are extremely insensitive.

Finally, real-world data is messy and may contain strange user interactions.

Buscher concluded his talk with advice for conducting experiments in a living lab: (i)

focus on very specific and well-defined problems/scenarios and be aware of possible unwanted side-effects; (ii) work out guidelines for checks that a feature has to pass before online experimentation; (iii) specify and agree on well-defined metrics that capture all/most aspects of the feature change; and (iv) data cleaning has to be handled and is done best by the commercial system if sufficient methods are available there (in conjunction with anonymization, etc.).

3 Presented Papers

The next session consisted of four short paper presentations, each of which described a different perspective on what living labs are and how they can be achieved. Each paper was given a slot of 10 minutes followed by 5 minutes for audience questions. The presented work featured diverse viewpoints and aspects with respect to evaluation and thus provided an excellent platform for discussions later in the workshop.

The first paper *A Private Living Lab for Requirements Based Evaluation* by Christian Beutenmüller, Stefan Bordag, and Ramin Assadollahi was presented via a pre-recorded video by Stefan Bordag. The work described attempts to evaluate a framework that facilitates the integration and sharing of information across multiple apps on a mobile device (PTPT), which can be used, for example, to generate user specific recommendations. The evaluation approach utilizes use cases and personas to establish a simulated evaluation to avoid compromising the privacy of real users. Paid testers assume virtual personas and evaluate items with respect to what the authors refer to as evaluation points snap shots of the data available to device at particular time points. The presentation discussed the costs of the approach, both in terms of creating datasets with paid testers and in the limitations in terms of validity. The method was presented as a complementary alternative to other evaluation approaches and represents a move towards some of the benefits of a living lab approach.

The next paper presented was *A Month in the Life of a Production News Recommender System* by Alan Said, Jimmy Lin, Alejandro Bellogin, and Arjen de Vries and was presented by Jimmy Lin. This work was closer to a more traditional definition of a living lab setup, describing an infrastructure for a real life news article recommender system, Plista, whereby external researchers and practitioners can connect their recommendation algorithms to the Plista infrastructure as part of a competition and deliver recommendations in real time to the systems users, offering the chance to evaluate their algorithms in situ. The infrastructure provides a strong model of how a living lab can be realized in practice. Systems from different groups are periodically requested to provide recommendations, but the interaction and performance data is open to all participants. Analyses of one months worth of interaction data with the system were presented, which highlighted several trends in news recommendation and showed that in situ evaluation is sensitive to factors not related to the recommendation itself. For example, such as natural temporal variation in user behaviour and biases in click-throughs for particular types and sources of articles. These analyses show that great care must be taken when interpreting the results of living lab evaluations.

The third paper to be presented was *(An) Evaluation for Operational IR Applications - Generalizability and Automation* by Melanie Imhof, Martin Braschler, Preben Hansen, and Stefan Rietberger. Melanie Imhof presented the work. This work presents a framework for "black box" appraisal and evaluation of IR systems based on a number of individual tests that, when taken together, provide a strong evaluation of the complete system. The evaluation framework is motivated by explaining the shortcomings of the more traditional

Cranfield approach, particularly its lack of focus on the users of the system, and framing it as a single part of a greater set of tests. Here various different aspects including the user interface, the underlying IR engine and data layers are evaluated and scores combined via a weighted average. In an evaluation of the approach the authors found that the score for this approach correlated with user experience measures. The presentation discussed the generalizability of the approach to different domains and the automation of the approach, which added nicely to the living labs discussion.

The final paper in the session, presented by Catherine Smith, broadened the focus somewhat by dealing with *Factors Affecting Conditions of Trust in Participant Recruitment and Retention*. This position statement built on the work of Nissenbaum [4], who proposed conditions associated with the formation of trust online. Smith discussed what these conditions could mean in terms of acquiring and retaining participants for a living lab situation. The first condition relates to the reputation of the trustee (researcher(s)) and their history, which could be influenced by the reputation of the institution in which they work, but also if the individual researcher(s) are known personally to the participants. The desired property of a large and diverse user population makes personal relationships unlikely (and undesirable). A further condition conducive to building a trust relationship is the existence of reciprocity in the relationship between truster and trustee. Smith argues that because the lab assumes no risk comparable to that taken by the participant, there is no mutuality, which makes recruitment a challenge. She further argues that while offering a monetary or other kind of reward can engender reciprocity, this is not particularly conducive to trust. All of these issues raise challenges in terms of recruitment and retention in a living lab setting and these must be addressed in order to achieve the benefits such evaluations offer. One suggestion Smith made was to offer contributors innovative new tools, methods, and systems, which may produce greater reciprocity among some populations and engender higher trust and increased rates of participation.

4 Presented Demos

There were three demos presented during the workshop. The first two touch on the problem of sharing a user pool along with infrastructure resources to conduct in situ experiments for a variety of search tasks ranging from ad-hoc search, entity ranking, and summarization through A/B testing and interleaving, while the latter considers the problem of limited pool of users in academic environments from a different perspective exploring the use of simulations for interleaving experiments and allowing sharing the appropriate infrastructure.

In details, the first demo titled *Using CrowdLogger for In Situ Information Retrieval System Evaluation* by Henry A. Feild and James Allan demonstrated an open-source browser extension for Firefox and Google Chrome, which can be used as an in situ evaluation platform. CrowdLogger serves as a client-side platform that tracks certain user interactions with web pages. Interactions include queries, result sets, clicks, page loads among others. The data is stored locally at the client side hence users have full privacy control over it. Users can inspect their activity logs, remove data from them, and upload them to the CrowdLogger server. A privacy API is used to provide control mechanisms regarding the privacy of the data such as client-side encryption and server-side decryption. CrowdLogger supports study modules developed by researchers for in situ experiments. The developed modules are distributed through CrowdLogger and users can choose to participate in the study by downloading and

installing the module. CrowdLogger provides the necessary API for the researcher to set up experiments that can use the history of user activities and/or live data. There were two major challenges/directions identified, the first was about saving engineering effort from researchers to build study modules by providing a module builder to automatically build code for common patterns, and the second was about increasing the pool of users by providing CrowdLogger as a desktop application that communicates with light-way browser extensions so that it is easier to extend the system for more browsing platforms.

The second demo titled *FindiLike: A Preference Driven Entity Search Engine for Evaluating Entity Retrieval and Opinion Summarization* by Kavita Ganesan and ChengXiang Zhai was the demo which received the best demo award. FindiLike is a preference-driven search engine that finds entities of interest based on preferences set by the user. Preferences may be structured (e.g., price) or unstructured (e.g., a hotel being clean). FindiLike explores a large set of online reviews about the entities of interest and matches these with the user preferences. Abstractive summarization is used to generate option summaries. In terms of the theme of the workshop an extension to the system was presented that allows the in situ evaluation of retrieval systems for the tasks of opinion-based entity ranking and summarization. Regarding the former, any search algorithm can be used to rank entities based on preferences; interleaved results can be shown to the users allowing the use of any interleaving algorithm that has been proposed in the literature. Regarding the evaluation of abstractive summarization algorithms, the current algorithms display sentences that summarize certain aspects of interest of the entities described in the online reviews. Sentences are clickable so that users can explore the underlying reviews summarized by them. Different algorithms can be implemented and sentences coming from the baseline and experimental algorithm can be randomly mixed. Clicks can again be used as a proxy of summary quality. New algorithms could be uploaded through an interface provided by FindiLike. FindiLike is already live, being used for the ranking of hotels and can be found at `eval.findilike.com`. Since January there has been about 1000 unique visits to the site. A couple of challenges were identified; first given the small amount of traffic which the site is currently receiving, new algorithms should be of good quality. Peer reviewing of the algorithms to be uploaded was suggested as a potential solution. The second challenge is about the efficiency of the uploaded algorithms with a potential solution being a threshold on the response time in the live system. A more general solution to all these issues could be an automatic allocation of opportunities to compete a baseline to multiple new algorithms; details of how such an evaluation could be performed are to be studied.

The last demo titled *Lerot: an Online Learning to Rank Framework* by Anne Schuth, Katja Hofmann, Shimon Whiteson, and Maarten de Rijke views the problem of limited user interaction data in academic environments from a different perspective providing a framework to simulate these data and perform interleaving experiments. The demonstrated framework allows the implementation of different models to simulate user clicks and the implementation of different interleaving methods. Combining the two one can simulate clicks over an interleaved ranking of two competing algorithms. This simulation framework can be used to learn a ranker. In each step of learning a ranker is perturbed and the two competing algorithms are the original ranker and the perturbed version of it, so weights can be learned based on the click behaviour of the users. A large number of algorithms have already been implemented, while researchers can add to this arsenal through the provided framework by implementing a set of described functions.

5 Discussion Session

Participants discussed how to make living labs a reality. There are two main possibilities for realizing living labs: (i) using an existing site or service and (ii) building something new together, as a community.

The advantage of (i) is that it would provide an immediate starting point for research and development. Two approaches were discussed for using existing sites and services. The first involves the creation of results in advance that are interleaved for users when a given query is entered. The interaction logs for this query would then be shared with the contributor of results. The second approach is some sort of API that makes requests of participants to provide results on the fly to a system and then also provides interaction data. Challenges are to find a site or service where there is enough traffic and the components to be researched are of interest to sufficiently many people. Sharing potentially sensitive data (such as search and usage logs) raises additional difficulties. Using CiteSeer was discussed as one option, but it is suspected that queries would primarily consist of paper titles, which would not be very interesting.

Option (ii), like building local domain search for universities, would have the advantage that it would lower the barrier to entry (by sharing indices, code, etc.). On the other hand, there is no short-term incentive for people to contribute. Also, it would mean running production IR systems; something that academics are not necessarily prepared to do. The idea here is that local domain search is important, under-served by commercial interests, and a challenging problem that may be within the scale do-able by researchers unable to work at the web-scale. Experimenting with university-wide search engines was discussed as a possibility that could combine the benefits of both (i) and (ii). It comprises components and data sources that are typical to most universities (news, study guide, staff homepages, etc.); therefore, data would not need to leave the walls of the organization. At the same time, all could benefit from a shared set of source code.

The news recommendation challenge from Plista (which runs as a CLEF Lab in 2014, see `http://www.clef-newsreel.org`) provides a working example for living labs. Although this is a real-time task, users' expectations towards response time are likely to be different for search than for recommendation. As a possible remedy, one workshop organizer suggested the idea of focusing on head queries; for these, rankings could be generated offline and then interleaved with the baseline search results.

Another idea was to create a plug-in for a search engine such as Lucene that would enable people to have a standard set of online metrics.

Finally, there are ethical issues involved with living labs, including if and how to ask permissions from users. Currently, there are no set guidelines that are universally accepted.

6 Conclusions

Overall, the workshop was an engaging, enjoyable event, which shed further light on the living lab for IR paradigm and avenues for progression. In particular, the challenges associated with generating living labs in the research community were further highlighted, and the benefits to be obtained by industry involvement exemplified. Initial exciting steps are now being made in the use of living labs for evaluation, some of which were showcased at the workshop. The notion of what constitutes a living lab within our community and multiple takes on this were highlighted. As a next step the community now needs to clearly categorize the

types of living labs possible for use in IR evaluation, and focus on targeted progression steps within these categories. Further individual developments, followed by (or indeed potentially coupled with) initial community driven initiatives, such as low barrier approaches in shared initiatives, should see living lab evaluation approaches mature over the coming years.

7 Acknowledgements

We are grateful for the financial support received from the ESF Research Networking Programme "Evaluating Information Access Systems" (ELIAS). We would also like to thank CIKM for hosting the workshop. Thanks also go to the program committee (Leif Azzopardi, Glasgow University; Ben Carterette, University of Delaware; Yi Chen, Dublin City University; Charles Clarke, University of Waterloo; Carsten Eickhoff, Delft University of Technology; Nicola Ferro, Department of Information Engineering - University of Padua; Morgan Harvey, University of Lugano (USI); Claudia Hauff, Delft University of Technology; Gareth Jones, Dublin City University; Noriko Kando, National Institute of Informatics; Diane Kelly, University of North Carolina; Séamus Lawless, Trinity College Dublin; Henning Müller, HES-SO; Ian Ruthven, University of Strathclyde; Falk Scholer, RMIT University; Alan Smeaton, Dublin City University; Paul Thomas, CSIRO; Ellen Voorhees, NIST), paper authors and workshop attendees, without whom the workshop would not have been the success it was.

References

[1] 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 *CLEF 2011: Conference on Multilingual and Multimodal Information Access Evaluation*, pages 26–37, 2011.

[2] J. Callan and A. Moffat. Panel on use of proprietary data. *SIGIR Forum*, 46(2):10–18, Dec. 2012.

[3] D. Kelly, S. Dumais, and J. O. Pedersen. Evaluation Challenges and Directions for Info. Seeking Support Systems. *Computer*, 42(3):60–66, 2009.

[4] H. Nissenbaum. Securing trust online: Wisdom or oxymoron? *Boston University Law Review*, 81:101–131, 2001.