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

Krisztian Balog[1], David Elsweiler[2], Evangelos Kanoulas[3], Liadh Kelly[4], Mark D. Smucker[5]
[1] University of Stavanger, Norway, *krisztian.balog@uis.no*
[2] University of Regensburg, Germany, *david@elsweiler.co.uk*
[3] Google, Switzerland, *ekanoulas@gmail.com*
[4] Dublin City University, Ireland, *lkelly@computing.dcu.ie*
[5] University of Waterloo, Canada, *mark.smucker@uwaterloo.ca*

Summary

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 which involve and integrate users within the research process. The basic idea of living labs is that a central and shared environment resource is used rather than individual research groups having to develop their own experimental environment and their own individual set of experiment subjects.

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 experiment purposes; cross-comparability across research centres; and 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. Progress towards realising actual living labs has nevertheless been limited. There are many challenges to be overcome before the benefits associated with living labs for IR can be realised, 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 organisations, 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 programme committee. Two short papers, two position papers and three demo papers were accepted for presentation at the workshop and inclusion in the workshop proceedings (available at http://dl.acm.org/citation.cfm?id=2513150). In addition to the regular paper presentations, the programme included an invited talk by Georg Buscher (Microsoft Bing).

The workshop was successful in bringing people from research communities as well as from industry together, and offered a highly interactive environment with lively discussions throughout the whole day. A final discussion session wrapped up the event with the objective to identify and formulate specific action items for future research and development.

Description of the scientific content of and discussions at the event

Keynote

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.

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 ignore 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 what checks 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.).

Presented Papers

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 utilises 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 benefit 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 system’s users, offering the chance to evaluate their algorithms in situ. The infrastructure provides a strong model of how a living-lab can be realised 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 month’s 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 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 and combining scores 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 generalisability 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, 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 condusive to building 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 partularly conducive to trust. All of these issues raise challenges in terms of recruitment and retention in a living-labs 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.

Presented Demos

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, that 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.

The second demo titled "FindiLike: A Preference Driven Entity Search Engine for Evaluating Entity Retrieval and Opinion Summarization" by Kavita Ganesan, ChengXiang Zhai was the one that 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 currently the site is 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 environment 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 is 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 has been already implemented, while researchers can add to this arsenal through the provided framework by implementing a set of functions described.

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 approach 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), e.g. 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, underserved by commericial 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.com (which runs as a CLEF Lab in 2014 [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.

Assessment of the results and impact of the event on the future directions of the field

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 categorise 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.

Annexes

Programme of the meeting

09:00-09:30 Opening
09:30-10:30 Keynote
10:30-11:00 Break
11:00-12:00 Paper Session (15min x 4)
12:00-12:30 Brainstorming topics / breakout ideas
12:30-13:30 Lunch
13:30-14:00 Demo Session (10min x 3)
14:00-15:00 Discussion (runs into break)
15:00-15:30 Break
15:30-16:30 Discussion
16:30-17:00 Wrap-up / closing

Full list of speakers and participants

Name	**City, Country**	**convenor**	**speaker**	**participant**
Dr. Alistair Moffat	Melbourne, (AU)			X
Dr. Catherine Smith	Kent, OH, (US)		X	X
Dr. Christian Beutenmüller	Leipzig, (DE)		X	X
Dr. David Elsweiler	Regensburg, (DE)	X		
Dr. Evangelos Kanoulas	Zürich, (CH)	X		
Dr. Georg Buscher	Redmond, WA, (US)		X	X
Dr. Henry Feild	Beverly, MA, (US)		X	X
Dr. Jimmy Lin	Maryland, MD, (US)		X	X
Dr. Krisztian Balog	Stavanger, (NO)	X	X	X
Dr. Liadh Kelly	Dublin, (IE)	X		
Dr. Mark Smucker	Waterloo, (CA)	X	X	X
Dr. Trotman Andrew	Dunedin, (NZ)			X
Mr. Anne Schuth	Amsterdam, (NL)		X	X
Mr. Damien Lefortier	Moscow, (RU)			X
Mr. Eugene Kharitonov	Moscow, (RU)			X
Mr. Jielong Zhou	Beijing, (CN)			X
Mr. Ke Zhou	Glasgow, (UK)			X
Mr. Matthieu Denoual	Brussels, (BE)			X
Mr. Tao Peng	Beijing, (CN)			X
Mrs. Kavita Ganesan	Champaign, IL, (US)		X	X
Mrs. Melanie Imhof	Winterthur, (SZ)		X	X