Living Labs for UMAP Evaluation

Liadh Kelly
ADAPT Centre
Trinity College Dublin
Ireland
liadh.kelly@tcd.ie

ABSTRACT

Generating shared task initiatives in the user-modelling, adaptation and personalization (UMAP) space is difficult, especially given individual difference, privacy concerns and the interactive nature of the space. We put forward that the living labs evaluation paradigm, i.e., observing users in their natural task environments, has potential to overcome these difficulties and to allow for comparative evaluation in the UMAP space of research. In particular, the emerging approach to living labs for shared evaluation in other research disciplines, has potential to be adapted to allow for shared evaluation tasks in the UMAP community. Coupled with this, there is the potential to create living labs in different ways for shared UMAP evaluation. In this paper we overview the use to-date of living labs approaches for shared evaluation and set directions for its application in UMAP shared challenges.

Keywords

Evaluation, living labs

1. INTRODUCTION

Developing means to conduct shared evaluation in the user modelling, adaptation and personalization (UMAP) space is inherently difficult. Not least because of privacy concerns, individual differences in behaviours between users of systems and challenges associated with working in interactive scenarios. In this paper we propose the use of a living labs approach as one potential way to overcome these difficulties and to allow for shared task generation in the UMAP domain.

The living labs (i.e. observing users in their natural task environments) evaluation paradigm is already used extensively by industry organisations [7]. More recently researchers in academia have begun to explore its potential application for their evaluations. In 2009 Kelly et al suggested that a living lab could be used as a way for researchers to perform *in situ* evaluations, with real users performing real tasks using real-world applications [5]. Since then researchers have worked towards creation of shared task initiatives using living labs in the information retrieval and recommender system communities [2, 4]. We next describe these initiatives to provide context for our proposal for living labs style shared task evaluations in the UMAP space. We then suggest directions, both stemming from these initiatives and taking a different living labs interpretation, for living labs in the UMAP space.

2. LIVING LABS IN THE INFORMATION RETRIEVAL COMMUNITY

The main goal of the new *living labs for information retrieval evaluation* (LL4IR)[1] challenges, running at CLEF LL4IR 2015-16[2] [8] and at TREC Open Search 2016[3], is to provide a benchmarking platform for researchers to evaluate their ranking systems in a live setting with real users in their natural task environments. The challenges act as proxy between commercial organizations (live environments) and lab participants (experimental systems), facilitate data exchange, and make comparison between the participating systems. Use-cases to-date (and associated commercial systems) for these challenges are product search, web search, and academic search based.

To participate in a challenge, challenge participants take part in a live evaluation process. For this they use a set of provided frequent queries (i.e. queries frequently submitted to the commercial system by it's users). Candidate documents are provided for each query along with historical information associated with the queries. When participants produce their rankings for each query, they upload these to the commercial provider use-case through the provided LL4IR API. The commercial provider then interleaves a given participant's ranked list with their own ranking, and presents the user with the interleaved result list. Participants take turns in having their ranked list interleaved with the commercial providers ranked list. The actions performed by the commercial providers' system users are then made available to the challenge participant (whose ranking was shown) through the API; i.e., the interleaved ranking, resulting clicks, and (aggregated) interleaving outcomes. System performance is scored based on click through rates. Figure 1 shows the LL4IR Living Labs architecture and how the participant interacts with the use-cases through the LL4IR provided API.

3. LIVING LABS IN THE RECOMMENDER SYSTEMS COMMUNITY

The CLEF NEWSREEL lab 2014-16[4] [3, 4], focuses on a recommendation algorithm generation shared challenge using real users interacting with a real commercial system. The use-case for this

[1] http://living-labs.net/
[2] http://living-labs.net/clef-ll4ir-2016/
[3] http://trec-open-search.org/
[4] http://www.clef-newsreel.org/tasks/

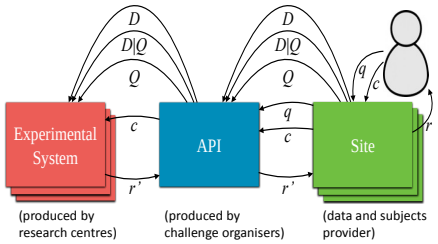

Figure 1: Schematic representation of the Living Labs for Information Retrieval Evaluation (LL4IR) API. Where, Q = frequent queries; (D|Q) = candidate documents for each query; c = user interactions with ranking r' for query q∈Q.

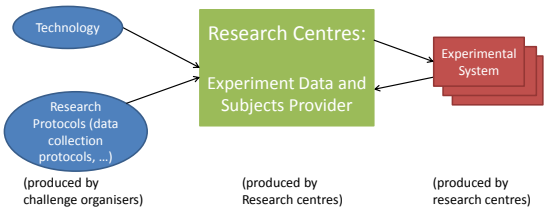

Figure 2: Schematic representation of user centric Living Labs API.

challenge is news recommendation, where the goal is to suggest, within a strict time constraint, news items that a user would click. To participate in this live challenge, participants plug their recommendation algorithm into the CLEF NEWSREEL provided ORP API [3]. The CLEF NEWSREEL API randomly selects a challenge participant's recommendation algorithm for each incoming news recommendation request from live news websites. The selected recommendation algorithm must provide news recommendations for the given user on the given news website in real-time. Click through rate is provided back to the participant's recommender system, and used to measure system performance. Participants have the opportunity to update their live recommender algorithm on an on-going basis. The high-level architecture for this living labs challenge is similar to that of the LL4IR one shown in Figure 1.

4. LIVING LABS FOR UMAP EVALUATION

4.1 API Centred Approach

To-date living labs shared task instantiations (described in Section 2 and Section 3 above) centre around an API (as shown in Figure 1). Challenge participants plug their developed approaches into the challenge provided API. Live commercial systems can then communicate through the API to use, and hence test, challenge participants' algorithms (or techniques), in place of, or in conjunction with, the commercial systems algorithms (or approaches). We believe this general living labs architecture offers promise for UMAP shared evaluation.

4.2 Research-Centre Centred Approach

However, living labs can also be interpreted in different ways. A tool for creating a living lab that centres on research centres providing data and users for shared evaluation is presented in [6] (see Figure 2). Here the focus is on a living lab for evaluation of retrieval techniques for personal desktop collections. In this approach, researchers wishing to evaluate their technologies would participate in a collaborative evaluation effort. Whereby required protocols and technology to gather data for, and to conduct the evaluation, would be distributed to the participating research centres. The retrieval algorithms/techniques developed by each participating research centre would also be distributed to the research centres for evaluation. Individual research centres would then recruit experiment subjects locally, who install and run the provided tool on their personal computer (PC). The tool indexes the items on the PCs. Using the provided protocols and tool, experiment subjects generate queries they would issue on their personal desktop collection (PC collection) and judge the relevance of items on their PC to these queries (i.e. relevance assessment). Participating research centres' IR algorithms can then be evaluated locally on subjects' PCs using the generated index, queries and relevance assessments, with only performance measures returned to investigators thus preserving privacy.

This *research centres* centred living labs approach could be generalized to allow for shared task evaluation in the UMAP space. Whereby evaluation goal specific tools and protocols, and challenge participants algorithms/software are distributed to individual research centres. These are then either used (as shown in Figure 2) to: generate static collections for evaluations as described above; run controlled experiments with the participants' software locally in each research centre; or run the participants' software live, for evaluation purposes, in place of individuals' typical software as they go about their normal activities. Or indeed a hybrid of this *research centres* centred and the earlier *API centred* approach might prove most useful in the UMAP space, depending on the precise scenario to be evaluated.

4.3 Challenges To Address

Realising such living labs requires addressing several challenges associated with living labs architecture and design, hosting, maintenance, security, privacy, participant recruiting, and scenarios and tasks for use development. This is similar to the challenges faced in setting up such living labs in the IR and recommender system spaces, as described in [1]. Lessons can be learned here from the experiences of the IR and recommender systems living labs shared tasks [2, 4].

5. CONCLUSIONS

Living labs is an emerging shared task evaluation paradigm. Living labs hold great promise for conducting realistic evaluation with real users in natural task environments, and importantly they allow for cross comparability across research centres. They have started to be used in IR and recommender system evaluation. We believe that discussion on their applicability and potential use for UMAP evaluation is warranted.

The purpose of this paper is to introduce the living labs shared task notion and to seed discussion on its possible application for UMAP evaluation.

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 *Proc. of CLEF'11*, 2011.

[2] K. Balog, L. Kelly, and A. Schuth. Head first: Living labs for ad-hoc search evaluation. In *CIKM'14*, 2014.
[3] T. Brodt and F. Hopfgartner. Shedding light on a living lab: The clef newsreel open recommendation platform. In *IIiX'14*, 2014.
[4] F. Hopfgartner, B. Kille, A. Lommatzsch, T. Plumbaum, T. Brodt, and T. Heintz. Benchmarking news recommendations in a living lab. In *CLEF'14*, 2014.
[5] D. Kelly, S. Dumais, and J. O. Pedersen. Evaluation challenges and directions for information-seeking support systems. *Computer*, 42(3): 60–66, 2009.
[6] L. Kelly, P. Bunbury, and G. J. F. Jones. Evaluating personal information retrieval. In *Proc. of ECIR '12*, 2012.
[7] 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.
[8] A. Schuth, K. Balog, and L. Kelly. Overview of the living labs for information retrieval evaluation (LL4IR) CLEF Lab 2015. In *CLEF'15*. 2015.