Research Networking Programmes

Science Meeting – Scientific Report

Proposal Title: *Living Labs for IR Challenge Workshop*

Application Reference N°: *5520*

1) Summary (up to one page)

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. Living labs would offer huge benefits to the community, including the availability of usage and interaction data for experimental purposes and greater knowledge transfer between industry and academia. Progress towards realising actual living labs has nevertheless been limited. There are many challenges to be overcome, including architecture and design, hosting, maintenance, security, privacy, participant recruiting, and scenarios and tasks for use development.

Having a community benchmarking platform with shared tasks would be a key catalyst in enabling people to make progress in this area. This is exactly what we are trying to set up in the form of a challenge, with the ultimate goal of turning it into a CLEF, TREC, or NTCIR track in the future. For this challenge we have sourced involvement from two use-cases: 1) the e-commerce domain with the involvement of a medium-size online retailer and 2) a search engine behind a university website. Challenge participants are able to access products/documents, usage and query log data, and trading logs (for the e-commerce use-case) through an API and test their approaches in a live setting.

Organizing this challenge is a huge challenge in itself. There are many potential pitfalls caused by the fact that outputs of experimental methods will need to be fed into live production systems. Further, making more data (usage and trade logs) available through the API provides more room for creative approaches, but, at the same time, raises privacy and security concerns. Our main goal for the LLC workshop was to test and further develop the challenge platform in the form of a "hackathon." A beta version of the challenge platform (including the API, documentation, and a "scoreboard") was implemented and prepared for the event.

While our attempt is a first of its kind for IR, there exist similar initiatives for news recommendation: the Plista contest and the CLEF 2014 NewsReel lab. There are some important differences between their task and ours; yet, the setups share some architectural similarities. The workshop, therefore, featured two invited talks from two of the organizers behind these efforts.

2) Description of the scientific content of and discussions at the event (up to four pages)

Invited talks

Setting Up a Living Lab for Information Access Research

Frank Hopfgartner (Technische Universität Berlin, Germany)

The first keynote provided an overview of current progress and drives in creating living labs in other domains. Key for these drives is creation of a realistic setting where users are not restricted by closed lab conditions. Hopfgartner then provided insights on his experiences organizing the CLEF NewsReel Challenge. This challenge uses a living lab evaluation approach in the recommendation space, where participants are invited to evaluate news recommendation techniques in real-time by providing news recommendations to actual users that visit commercial news portals to satisfy their information needs. The challenge consisted of two tasks. Task 1 is an offline evaluation, where participants must predict what articles the user will click on. Task 2 is an online evaluation consisting of a number of pre-defined evaluation periods. This challenge uses multiple publishers (sport, news, etc), which the industrial partner (plista) involved in the challenge provides recommendations for. In this challenge search queries are not provided, nor are users identified. Instead queries with pseudo ID for geo-region source of query are provided. In describing these tasks and their set-up Hopfgartner provided the audience with insights into how to create a living lab evaluation campaign, his experiences, challenges encountered, and solutions adopted. The importance of constantly monitoring the system was emphasized, and the need to regularly contact participants to ensure they are not experiencing issues with using the system.

Slides from the keynote are available at:
http://living-labs.net/wp-content/uploads/2014/06/LLC-Hopfgartner.pdf

Open Recommendation Platform for Researchers & Developers

Torben Brodt (plista GmbH)

The second keynote provided insights into the plista industry partner of the CLEF NewReel living labs recommendation challenge. Plista is a data-driven content and advertising platform. Details on how the plista system was built and how it acts in the background were provided. Brodt then described the research and development being conducted by a team at plista into recommendation algorithm development and how researchers can get involved with plista. Details on how the CLEF NewsReel challenge operates from the plista end were provided. In this challenge, when a participant submits results to plista, these results are first validated by plista and then shown directly to the users (i.e., neither changed nor interleaved with other system results). They discussed challenges in creating a living lab evaluation campaign and points for consideration when creating a campaign. For example, they reiterated the need to give participants as much as possible before the challenge. They also highlighted that response time might be an issue, and hence that it might be necessary to give a virtual server to participants. Also highlighted was that there may be a start-up problem for participants integrating into a real application, and hence that it may be beneficial to give participants a run time environment and to provide them with a baseline system.

Slides from the keynote are available at:
http://living-labs.net/wp-content/uploads/2014/06/LLC-Torben.pdf

Challenge overview

Next in the programme, the Living Labs Challenge was introduced. First, the two main use-case organizations and application scenarios were presented.

- REGIO Játék (REGIO, www.regiojatek.hu) is the largest (offline) toy retailer in Hungary with currently over 30 stores; their webshop is among the top 5 in Hungary. The company is working on strengthening their online presence; improving the quality of product search on their website is directed towards this larger goal.
- The University of Amsterdam (UvA, www.uva.nl) offers search functionality on their website. Currently, the search system is an out of the box search engine and the university is looking for ways to improve upon the system. They initiated collaborations with information retrieval researchers and agreed to participate in the living labs challenge. Part of the search traffic to the search box on the front page will soon be flowing through the living labs challenge API.
- Additionally, Seznam, the largest search engine in the Czech Republic, also expressed interest in joining our living labs initiative with a third use-case: web search. One main difference compared to the first two use-cases, apart from the scale, is that Seznam would not make raw document and query content available, but features computed for documents and queries.

This was followed by a description of the challenge API, given by Anne Schuth.
The API documentation is available at: http://doc.living-labs.net

Breakout group sessions

Participants were divided into 3 breakout groups: 1) evaluation, led by Liadh Kelly; 2) use-cases and participation, led by Krisztian Balog; and 3) API and technicalities, led by Anne Schuth. Next, we briefly report on the groups' activities and outcomes.

Breakout group #1: Evaluation

This group considered the way in which evaluation of systems developed by participants of the living labs challenge should be conducted. The group began by exploring the use-cases' websites, from which several useful insights were discovered:

- On the UvA website, many documents returned in response to searchers are password protected. The group questioned the impact of this and whether the content of such sites could be provided to challenge participants. This possibility is being explored with UvA.
- In the REGIO use-case, shoppers might select their product for purchase based on the visual appeal of the product (depicted in image accompanying product). It had been planned to provide textual information only associated with products to challenge participants. However, we are now exploring the possibility of also providing images.
- For some of the queries provided in the challenge different categories of users will find different documents to be relevant. For example, for the query 'VPN' the information need of staff and student would be quite different. While this is part of the challenge, we are now looking into the possibility of also providing some user information to challenge participants.

The need to be careful about the type of data we release in the challenge was discussed. The possibility of releasing the data (logs etc.) after the lab challenge, for use by the research community, was also discussed. This could be good for traction, but the possibility to do so needs to be discussed with the commercial and university data owners.

The group then moved on to consider an evaluation approach and evaluation metrics for the challenge. They introduced the idea of having a training and test phase. For this the 100 queries per challenge would be split into two sets of 50. A split of 6 weeks training and 2 weeks testing was proposed for the REGIO use-case, and a split of 2 weeks training and 6 weeks testing for the UvA use-case. This could operate by providing for example a two week window to request the test queries, and once downloaded the team would have 24 hours to upload their final system for the test phase.

During the test phase results presented via a dashboard would show the performance of competing systems, with perhaps daily updates. These results would be based on clickthrough information for the UvA use-case. The possibility of using a graded relevance assessment for the REGIO use-case where user actions of purchase, placing item in basket, dwell time, click through data would be used as implicit relevance indicators was discussed. Using business focused evaluation metrics was also considered for the REGIO use-case, where the amount of profit on purchases would be used as a performance metric. For the REGIO use-case we will provide an overall score of this nature. Similar information would be provided to individual teams on their system performance, on a query by query basis, during the training phase.

Breakout group #2: Participation

This group considered the challenge API from the participants' perspective, with a primary focus on the product search use-case.

The generally little amount of textual material associated with products (only name, description, and name of categories) was noted, and a number of additional pieces of information that make the product search task interesting were identified.

(1) *Historical click information for queries.* Products that received clicks along with relative click frequencies.

(2) *Collection statistics*. The dictionary of terms with collection and document frequencies.

(3) *Product taxonomy*. The categories that the product is assigned to are already available, but the category system itself is also of interest.

(4) *Product photos*. Arguably, in an e-commerce setting, product photos might have a significant influence on which items get clicked.

(5) *Date/time when the product first became available*. Users might be more interested in "new arrivals."

(6) *Historical click information for products*. Queries along with relative click counts that led to a given product.

(7) *Sales margins*. From the vendor's viewpoint, the ultimate metric is not relevance, but the profit made on the purchases.

Note that (1) and (2) are, in fact, not specific to the product search use-case. (2) -- (4) could be obtained by crawling the website, therefore they do not qualify as sensitive information; (3) was already implemented on site during the event. (1), (6), and (7) represent sensitive data; we were already given permission to make (1) available; (6) and (7) are currently being discussed with REGIO.

More technical requests concerning API calls were also articulated (e.g., the need for identifying the use-case in the HTTP requests); some of these were accommodated by the technical breakout group. Finally, parts where the documentation needs to be extended were identified.

Breakout group #3: Technicalities

This group looked into technicalities regarding the implementation of the challenge API.

- Initially, every member of the breakout group installed the API locally on their own machines. This brought up numerous compatibility issues and clarity issues with the installation manual. All these issues were fixed on the spot.
- Furthermore, scaling up with the number of participants and sites talking to the API simultaneously were brought to light. Bottlenecks have been identified and we have started addressing these.
- Members of the breakout group also started implementing new sites and participants that can communicate with the API.

3) Assessment of the results and impact of the event on the future directions of the field (up to two pages)

The target outcome of the workshop was the operationalization of living labs in production environments, for product search and site search. Our specific aims were:

1. to implement initial ideas and deploy them using the platform,
2. to perform break testing of the platform,
3. to identify whether any of the shared data may be misused (i.e., sensitivity and privacy testing),
4. to collect and prioritize additional features requests.

Our assessment of the outcomes, with respect to the goals originally set out for the event, are as follows.

(1) Implement initial ideas and deploy them using the platform.

All participants (in breakout groups #2 and #3) managed to connect to the challenge API. The time available proved too short to implement and deploy new ideas, however, participants did get a glimpse of how the platform works using the simulation feature of the API.

(2) Perform break testing of the platform.

A number of bottlenecks were identified as an increasing number of participants and sites started talking to the API simultaneously. Some of these issues were fixed on the spot, while others were collected on a todo list and are currently being addressed.

(3) Identify whether any of the shared data may be misused (i.e., sensitivity and privacy testing).

Sensitivity and privacy issues were thoroughly discussed in breakout group #1. A number of issues also popped up in relation to additional feature requests, concerning the product search use-case. In general, information that could potentially be crawled from the web (even if involving manual effort) is not considered sensitive. Historical queries and clicks are taken to be sensitive data; the possibility of releasing them is currently being discussed with the use-case organizations and is likely to be allowed subject to the agreement of challenge participants to organizational terms and conditions. These terms and conditions should include, among others, that (i) information may not be distributed, (ii) copies of resulting publications must be sent to the challenge organizers, (iii) participating websites must not be referenced by their name, but as use-cases, (iv) organizations or challenge organizers can request data or part of the data to be deleted.

(4) Collect and prioritize additional features requests.
A number of additional feature requests were collected. Some of these were implemented as part of the hackathon. Others have been identified as essential for the challenge and are currently being implemented. Finally, some have been added to a wishlist for future development.

Overall, the workshop was an engaging, enjoyable event, and was successful in bringing people from research communities (information retrieval and recommendation) together. Our initiative also raised interest among commercial parties (as evidenced by the participation of Seznam) as well as evaluation efforts in other sub-disciplines of information retrieval (including multimedia evaluation, http://www.multimediaeval.org/, whose representatives were also present). In total, we estimate that dozens of European researchers in search engine evaluation methodology stand to benefit from these outcomes.

As a follow-up to the workshop, a proposal for running the challenge as a lab at CLEF has recently been submitted. Our platform acts as a solid, practical starting point for progressing the living labs for IR evaluation paradigm, by offering the first practical implementation of this evaluation methodology. This is intended as a first step for further progress. We see several opportunities for growth of this years tasks in subsequent years.

Firstly, while we currently propose a single training and test phase, we envision a process where training and testing phases follow continuously, essentially running the challenge every few months. This would provide academics with a continuous possibility to test their new ideas on real users. CLEF would then still serve as the yearly venue to present findings.

Secondly, we would like to extend to additional tasks and uses-cases. With the API that has been implemented, incorporating new use-cases is straightforward.

Thirdly, the current setup caters for one-off queries for which results are pre-computed offline. Ultimately, we would like participants to have control over entire sessions and allow for the personalization of search results. Achieving this (along with the appropriate safety mechanisms) will require a number of incremental steps.

There are also other ways that living labs might be created, these possibilities will also be considered and progressed in subsequent years.

4) Annexes 4a) and 4b): Programme of the meeting and full list of speakers and participants

4a) Programme of the meeting

09:30-10:00	**Opening**
10:00-12:00	**Invited talks** by CLEF NEWREEL organisers (TU Berlin and Plista Gmbh) followed by an interactive Q&A session
12:00-13:00	Lunch
13:00-16:00	**Hackathon**
16:00-17:00	**Wrap-up and conclusions**
17:00-19:00	Dinner

4b) Full list of speakers and participants

Dr. Krisztian Balog	Stavanger, (NO)	Organizer
Dr. Lamjed Ben Jabeur	Toulouse Cedex 9, (FR)	Participant
Mr. Richard Berendsen	Amsterdam, (NL)	Participant
Mr. Torben Brodt	Berlin, (DE)	Speaker
Dr. Claudia Hauff	2600 GA Delft, (NL)	Participant
Mr. Tjeerd Hes	Amsterdam, (NL)	Participant
Dr. Frand Hopfgartner	Berlin, (DE)	Speaker
Dr. Liadh Kelly	Dublin, (IE)	Organizer
Mr. Gebre Kirstos	Amsterdam, (NL)	Participant
Dr. Martha Larson	Delft, (NL)	Participant
Dr. Ilya Markov	Amsterdam, (NL)	Participant
Mr. David Maxwell	Glasgow, (UK)	Participant
Mr. Tomáš Páleníček	Praha 5, (CZ)	Participant
Mr. Ridho Reinanda	Amsterdam, (NL)	Participant
Dr. Alan Said	2600 GA, Delft, (NL)	Participant
Mr. Anne Schuth	Amsterdam, (NL)	Organizer
Mr. Isaac Sijaranamual	Amsterdam, (NL)	Participant
Dr. Gianmaria Silvello	Padova, (IT)	Participant
Mr. Jaspreet Singh	Tampere, (FI)	Participant
Dr. Manos Tsagkias	Amsterdam, (NL)	Participant