Living Labs for IR Evaluation

LL4IR@CLEF’15

“Give us your ranking, we’ll have it clicked!”

Anne Schuth
University of Amsterdam

Krisztian Balog
University of Stavanger

Liadh Kelly
Trinity College Dublin

http://living-labs.net
@livinglabsnet
News
News

• Funding from EFS ELIAS
  • This meeting
  • Developing the API
• Funding from Microsoft Azure
  • For hosting the API
• Lots of improvements of the API
  • Tracking of errors
• Lots of interest from site that may want to join
  • Several academic search engines?
  • Recipe search?
Introduction
Overview

• **Overall goal**: make information retrieval evaluation more realistic
  • Evaluate retrieval methods in a *live setting* with *real users* in their *natural task environments*

• **Focus**: medium to large sized organizations with fair amount of search volume
  • Typically lack their own R&D department, but would gain much from improved approaches
  • Or, would like to collaborate with academic researchers
Key idea

• Focus on frequent (head) queries
  • Enough traffic on them (both real-time and historical)
  • Ranked result lists can be generated offline

• An API orchestrates all data exchange between live sites and experimental systems

Methodology

Participant

Site

Living Labs API

Researcher

Queries

Documents

click

ranking

query

click

ranking

Queries

Documents
Use cases

- Three Two ad-hoc search tasks

<table>
<thead>
<tr>
<th></th>
<th>Local domain search</th>
<th>Product search</th>
<th>Web search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider</td>
<td>uva.nl</td>
<td>regiojatek.hu</td>
<td>seznam.cz</td>
</tr>
<tr>
<td>Data</td>
<td>raw queries and (generally textual) documents</td>
<td>raw queries and (highly structured) documents</td>
<td>pre-computed document-query features</td>
</tr>
<tr>
<td>Site traffic</td>
<td>relatively low</td>
<td>relatively low (~4K sessions/day)</td>
<td>high</td>
</tr>
<tr>
<td>Info needs</td>
<td>(mostly) navigational</td>
<td>(mostly) transactional</td>
<td>vary</td>
</tr>
</tbody>
</table>
Code: bitbucket.org/living-labs/ll-api

please report issues here!
Living Labs Documentation

- 1. Guide for CLEF Participants
  - 1.1. Schedule
  - 1.2. Key Concepts
  - 1.3. Usage Scenarios [New]
  - 1.4. Implement a Client
    - 1.4.1. Initialize
    - 1.4.2. Obtain Queries
    - 1.4.3. Obtain Doclists
    - 1.4.4. Obtain Feedback and Update Runs
  - 1.5. Running a Client
  - 1.6. Getting Help
  - 1.7. Citation
- 2. API Reference
  - 1. API Reference for Participants
    - 1.1. Query
    - 1.2. Doclist
    - 1.3. Doc
    - 1.4. Run
    - 1.5. Feedback
  - 2. API Reference for Sites
    - 2.1. Query
    - 2.2. Doclist
    - 2.3. Doc
    - 2.4. Ranking
    - 2.5. Feedback
1. Guide for CLEF Participants

Note
This guide is being updated as it is being used. Please tell us what you think is missing. Our contact details are at the bottom of this page.

This guide is meant to be a practical guide to participating in the CLEF Living Lab. Since we deviate significantly from the typical TREC style evaluation setup that most participants are likely to be familiar with, we will focus primarily on those differences.

Participating in the lab involves following these steps:

1. Read the lab description and Key Concepts below. Make sure you’re Getting Help when needed.
2. Sign up:
   1. Register at CLEF.
   2. Register with the lab. You can do this at any moment until the test phase begins.
   3. Sign and send the lab the agreement form. You will receive a link to this form.
   4. Sign up for individual sites (use-cases) you want to obtain data for. You will receive a link by email to do so.
3. Implement your method as a client that can talk to the API. Examples are provided. See Implement a Client below.
4. Run your client:
   1. The client you implement should probably run continuously over several weeks and can potentially constantly update runs.
   2. When the test phase starts, download test queries and submit your test runs. Again, the test phase will last for several weeks but there is no need (nor the possibility) to update runs.
5. Write up your findings. Publication details will become available.
6. Come to and present your work at CLEF 2015 in Toulouse, France in September 2015.

We hope that all steps but 3. and 4. are self-explanatory. Below we detail these two steps in Sections Implement a Client and Running a Client respectively.

1.1. Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Nov, 2014</td>
<td>Training period begins (Note that you can join any time after this date)</td>
</tr>
<tr>
<td>1-15 Apr, 2015</td>
<td>Uploading test runs</td>
</tr>
<tr>
<td>15 Apr, 2015</td>
<td>Testing period begins</td>
</tr>
<tr>
<td>15 May, 2015</td>
<td>Testing period ends</td>
</tr>
<tr>
<td>17 May, 2015</td>
<td>Results released</td>
</tr>
</tbody>
</table>
Dashboard: living-labs.net:5001
Evaluation

- **Train** queries
  - ‘Immediate’ feedback
  - Raw and aggregated feedback
- **Test** queries
  - **No updates** during test period
  - Feedback after test period
  - Only Aggregated feedback
- **Metric**: Team Draft Interleaving
  - Fraction of **wins** against production
F. Radlinski, M. Kurup, and T. Joachims.
How does clickthrough data reflect retrieval quality? In CIKM '08. 2008
• Test periods
  • Last two weeks of every month
• Same set of queries
• Runs will expire
  • This is new behavior
  • Meant to not waste query impressions
Results
Participants

- 39 teams signed up

- Industry:
  904labs, Microsoft, Plista, Yahoo

- Academia:
  au, bw, cz, fr, ie, in, jp, nl, no, uk, us

- 20 teams signed our agreement

- 12 teams submitted runs

- 3 teams submitted 5 runs for test queries
Results
Product Search
Product Search - Inventory

Times. Also, new products may arrive over time that are not available to experimental systems but do get returned by the production system (and hence get clicked by users). We further note that new arrivals are displayed distinctively in the webshop, which may also introduce a bias. Figure 3 displays the number of new arrivals (in green), and the products that became available (blue) or unavailable (red) from the day before. Only products that are candidates for any of the queries (either training or test) are considered. This figure shows absolute numbers. It is apparent that changes do occur, and in particular a great number of new products arrive. (This is actually the least desired type of change, as there is no easy way of dealing with it in our current platform.) An even more revealing statistic would be to measure the ratio of products that were unavailable at a given day, compared to all candidate products that were ever available during the test phase. This is shown in Figure 4. Note that unavailability ratio is specific to a given ranking; the reported numbers are computed for the organizers' baseline. To keep things simple, we use a single value, the average unavailability ratio of all submitted rankings, which is 44%.

If all products were available, the expected probability of winning an interleaved comparison (assuming a randomly clicking user) would be 0.5. However, on average, 44% of the products were actually unavailable. During Round #1, these products were only ever present in the participants ranking (the site's ranking never considered them). And, only after interleaving were these products removed from the resulting interleaved list. We note that this is undesired behavior, as they should have been filtered out before interleaving. The necessary adjustments were made to the implementation for Round #2 of the challenge. As for interpreting the Round #1 results, this means that the chances for products from the participants ranking to be clicked were reduced. We believe that
Product Search - Inventory

• Participants **should** update available products

• Rankings **may** contain stale products

• These products were removed **after** interleaving

  • **Biasing** in favor of production (which never has stale products)

  • Expected interleaving outcome is no longer 0.5

    (we *estimated* it became 0.28)
## Results - Product Search

### Round 1 – Official CLEF Round

<table>
<thead>
<tr>
<th>Teamname</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
</tr>
<tr>
<td>UiS-Mira</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
</tr>
<tr>
<td>UiS-Jern</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
</tr>
<tr>
<td>UiS-UiS</td>
<td>0.2827</td>
<td>54</td>
<td>137</td>
<td>508</td>
<td>699</td>
</tr>
<tr>
<td>Expected</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GESIS</td>
<td>0.2685</td>
<td>40</td>
<td>109</td>
<td>374</td>
<td>523</td>
</tr>
</tbody>
</table>
# Results - Product Search

## Round 2 – June 2015

<table>
<thead>
<tr>
<th>Teamname</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
</tr>
<tr>
<td>Expected</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UiS-Jern</td>
<td>0.4795</td>
<td>82</td>
<td>89</td>
<td>596</td>
<td>767</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
</tr>
<tr>
<td>UiS-Mira</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
</tr>
<tr>
<td>UiS-UiS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
</tr>
</tbody>
</table>
## Results - Product Search

### Round 3 – July 2015

<table>
<thead>
<tr>
<th>Teamname</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRIT</td>
<td>0.4890</td>
<td>89</td>
<td>93</td>
<td>533</td>
<td>715</td>
</tr>
<tr>
<td>UiS-Mira</td>
<td>0.4507</td>
<td>64</td>
<td>78</td>
<td>527</td>
<td>669</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.4430</td>
<td>66</td>
<td>83</td>
<td>498</td>
<td>647</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4134</td>
<td>74</td>
<td>105</td>
<td>513</td>
<td>692</td>
</tr>
<tr>
<td>UiS-Jern</td>
<td>0.3702</td>
<td>67</td>
<td>114</td>
<td>511</td>
<td>692</td>
</tr>
<tr>
<td>UiS-UiS</td>
<td>0.3459</td>
<td>55</td>
<td>104</td>
<td>521</td>
<td>680</td>
</tr>
</tbody>
</table>
## Results - Product Search

### Round 4 – August 2015

<table>
<thead>
<tr>
<th>Teamname</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRIT</td>
<td>0.4654</td>
<td>101</td>
<td>116</td>
<td>767</td>
<td>984</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4292</td>
<td>103</td>
<td>137</td>
<td>804</td>
<td>1044</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.3783</td>
<td>87</td>
<td>143</td>
<td>781</td>
<td>1011</td>
</tr>
</tbody>
</table>
Results - Product Search

Interleaving Outcome

Round 1  |  Round 2  |  Round 3  |  Round 4
Baseline  |  UiS-Mira |  UiS-Jern |  UiS.UiS
GESIS  |  IRIT
Results
Web Search
### Results – Web Search

#### Round 1 – Official CLEF Round

<table>
<thead>
<tr>
<th>Teamname</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploitative Baseline</td>
<td>0.5527</td>
<td>3030</td>
<td>2452</td>
<td>19055</td>
<td>24537</td>
</tr>
<tr>
<td>Expected</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform Baseline</td>
<td>0.2161</td>
<td>430</td>
<td>1560</td>
<td>1346</td>
<td>3336</td>
</tr>
</tbody>
</table>
## Results – Web Search

### Round 2 – June 2015

<table>
<thead>
<tr>
<th>Teamname</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exploitative Baseline</strong></td>
<td>0.6035</td>
<td>3128</td>
<td>2055</td>
<td>18055</td>
<td>23238</td>
</tr>
<tr>
<td><strong>Expected</strong></td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Uniform Baseline</strong></td>
<td>0.2547</td>
<td>435</td>
<td>1273</td>
<td>1053</td>
<td>2761</td>
</tr>
</tbody>
</table>
## Results – Web Search

### Round 3 – July 2015

<table>
<thead>
<tr>
<th>Teamname</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploitative Baseline</td>
<td>0.5203</td>
<td>2161</td>
<td>1992</td>
<td>13206</td>
<td>17359</td>
</tr>
<tr>
<td>Expected</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UvA-LambdaMart</td>
<td>0.2405</td>
<td>2264</td>
<td>7148</td>
<td>7863</td>
<td>17275</td>
</tr>
<tr>
<td>Uniform Baseline</td>
<td>0.2157</td>
<td>313</td>
<td>1138</td>
<td>922</td>
<td>2373*</td>
</tr>
</tbody>
</table>
### Results – Web Search

#### Round 4 – August 2015

<table>
<thead>
<tr>
<th>Teamname</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exploitative Baseline</td>
<td>0.4500</td>
<td>18</td>
<td>22</td>
<td>134</td>
<td>174</td>
</tr>
<tr>
<td>UvA-LambdaMart</td>
<td>0.2059</td>
<td>21</td>
<td>81</td>
<td>89</td>
<td>191</td>
</tr>
</tbody>
</table>
Goals
Goals of this Meeting

• Share findings

• Identify obstacles / problems / confusion

• Establish future directions
Future
Our Future

• We will continue
  • Next year at CLEF?

• New Use Cases
  • Academic Search
  • Recipe Search

• New Task?

• Non-head queries?

• Other metrics?

• Relation between online and offline
  • Write your SIGIR paper
Today
Today’s programme

16:00-16:10 Introduction to the lab

16:10-16:25 Regio use case presentation

16:25-16:40 Seznam use case presentations

16:40-17:25 Lab participants presentations

GESIS, IRIT, UIS (10min each)

17:25-17:35 Questions to participants

17:35-18:00 Discussion session and wrap-up