Living Labs for IR Evaluation

#LL4IR

Overview

Anne Schuth
University of Amsterdam

Krisztian Balog
University of Stavanger

Liadh Kelly
Trinity College Dublin

http://living-labs.net
@livinglabsnet
the lab
use cases
conclusions
Living Labs for IR

- New lab
Living Labs for IR

- **New** lab
- **Realistic IR evaluation**
Living Labs for IR

• **New lab**

• **Realistic IR evaluation**

• Exposing experimental systems to **real users**
Living Labs for IR

• **New** lab

• **Realistic** IR evaluation

• Exposing experimental systems to **real users**

• **Unsuspecting** users
Living Labs for IR

- **New** lab
- **Realistic IR evaluation**
- Exposing experimental systems to **real users**
  - **Unsuspecting** users
  - Users performing **real tasks**
Living Labs for IR

- **New** lab
- **Realistic IR evaluation**
- Exposing experimental systems to **real users**
  - **Unsuspecting** users
  - Users performing **real tasks**
  - Users issuing **real queries**
API
API
API
API
API
API

Living Labs
API

Living Labs API

query

ranking

click

Documents

Queries

Living Labs

Browser
API

Queries → Documents

Researchers

Documents → Queries

Clicks → Ranking

Query → Click

Living Labs
API

Living Labs API

Researcher

Queries → Documents → Ranking

User

Query → Ranking → Click

Query → Documents → Ranking
API

Researcher

Documents

Queries

Living Labs

Documents

Queries

query

ranking

click

query

ranking

Documents
API

Researcher

Documents

Queries

click

Living Labs

click

Documents

Queries

query

ranking

click

ranking

query

Documents
API

Researcher

Living Labs

Documents

Queries

click

ranking

query

Documents

Queries

click

query

ranking

click
API

- **API** (open source) to communicate
API

- **API** (open source) to communicate
  - Queries, documents, runs, clicks, …
API

- **API** (open source) to communicate
  - Queries, documents, runs, clicks, …
- Both **researchers** and **search engines** use API
API

• **API** (open source) to communicate
  • Queries, documents, runs, clicks, …

• Both **researchers** and **search engines** use API
  • **Easy** to connect new search engines
API

- API (open source) to communicate
  - Queries, documents, runs, clicks, …
- Both researchers and search engines use API
  - Easy to connect new search engines
- Fast (for crucial requests)
API

• **API** (open source) to communicate
  • Queries, documents, runs, clicks, …

• Both **researchers** and **search engines** use API
  • **Easy** to connect new search engines

• **Fast** (for crucial requests)

• REST-full, JSON
API

• **API** (open source) to communicate
  • Queries, documents, runs, clicks, ...

• Both **researchers** and **search engines** use API
  • **Easy** to connect new search engines

• **Fast** (for crucial requests)

• REST-full, JSON

• Example clients
API

- **API** (open source) to communicate
  - Queries, documents, runs, clicks, …
- Both **researchers** and **search engines** use API
  - **Easy** to connect new search engines
- **Fast** (for crucial requests)
- REST-full, JSON
- Example clients
  - **Easy** to get started
# Dashboard

## Profile for anneschuth

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teamname</td>
<td>anneschuth</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:anne.schuth@uva.nl">anne.schuth@uva.nl</a></td>
</tr>
<tr>
<td>API key</td>
<td>9F2ECC3BEE4DCFC-********PM</td>
</tr>
<tr>
<td>Creation</td>
<td>2014-06-05 14:51:16.973000</td>
</tr>
<tr>
<td>Verified</td>
<td>Yes</td>
</tr>
<tr>
<td>Type</td>
<td>Participant</td>
</tr>
</tbody>
</table>
API

- **Request**
  
  GET /api/participant/query/(key)

- **Response**
  
  ```json
  {
  "queries": [ {
    "creation_time": "Mon, 10 Nov 2014 17:42:24",
    "qid": "S-q1",
    "qstr": "jaguar",
    "type": "train"
  }, ... ]
  }
  ```
Head Queries

- Evaluate systems on the same set of queries
Head Queries

- Evaluate systems on the same set of queries
- **Stable** volume
Head Queries

- Evaluate systems on the **same set of queries**
- **Stable** volume
- **Historical** click and usage data is available
Head Queries

- Evaluate systems on the **same set of queries**
- **Stable** volume
- **Historical** click and usage data is available
- No (or less) **privacy** issues
Head Queries

- Evaluate systems on the same set of queries
- **Stable** volume
- **Historical** click and usage data is available
- No (or less) **privacy** issues
- Researchers can upload rankings **offline**
Evaluation

- Train queries
Evaluation

- **Train** queries
  - ‘Immediate’ feedback
Evaluation

- **Train** queries
  - ‘Immediate’ feedback
  - Raw and aggregated feedback
Evaluation

- **Train** queries
  - ‘Immediate’ feedback
  - Raw and aggregated feedback
- **Test** queries
Evaluation

• **Train** queries
  • ‘Immediate’ feedback
  • Raw and aggregated feedback
• **Test** queries
  • **No updates** during test period
Evaluation

- **Train** queries
  - ‘Immediate’ feedback
  - Raw and aggregated feedback

- **Test** queries
  - **No updates** during test period
  - Feedback after test period
Evaluation

• **Train** queries
  • ‘Immediate’ feedback
  • Raw and aggregated feedback

• **Test** queries
  • **No updates** during test period
  • Feedback after test period
  • Only Aggregated feedback
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
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
F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM '08. 2008
F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM '08. 2008
Team Draft Interleaving

Production

Reseacher

Researcher is preferred over Production

F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM '08. 2008
Participation

- 39 teams signed up
Participation

• 39 teams signed up

• Industry:
  904labs, Microsoft, Plista, Yahoo
Participation

• 39 teams signed up

• Industry:
  904labs, Microsoft, Plista, Yahoo

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

• 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
Participation

- 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
Participation

• 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
the lab
use cases
conclusions
the lab

use cases

conclusions
Use Cases

Provider

Data

Site traffic

Info needs
<table>
<thead>
<tr>
<th><strong>Use Cases</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Provider</strong></td>
</tr>
<tr>
<td><strong>Data</strong></td>
</tr>
<tr>
<td><strong>Site traffic</strong></td>
</tr>
<tr>
<td><strong>Info needs</strong></td>
</tr>
</tbody>
</table>
### Use Cases

<table>
<thead>
<tr>
<th>Provider</th>
<th>Local domain search</th>
<th>Product search</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td>raw queries and (generally textual) documents</td>
<td>raw queries and (highly structured) documents</td>
</tr>
<tr>
<td><strong>Site traffic</strong></td>
<td>relatively low</td>
<td>relatively low (~4K sessions/day)</td>
</tr>
<tr>
<td><strong>Info needs</strong></td>
<td>(mostly) navigational</td>
<td>(mostly) transactional</td>
</tr>
</tbody>
</table>
## Use Cases

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

<table>
<thead>
<tr>
<th>Provider</th>
<th>Local domain search</th>
<th>Product search</th>
<th>Web search</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td>uva.nl</td>
<td>regiojatek.hu</td>
<td>seznam.cz</td>
</tr>
<tr>
<td><strong>Site traffic</strong></td>
<td>relatively low</td>
<td>relatively low (~4K sessions/day)</td>
<td>high</td>
</tr>
<tr>
<td><strong>Info needs</strong></td>
<td>(mostly) navigational</td>
<td>(mostly) transactional</td>
<td>vary</td>
</tr>
</tbody>
</table>

Not implemented on time
Fig. 2. Screenshot of REGIO, our product search use-case. Those that were unavailable at a given point in time were not displayed to users of the REGIO online store. Further, it may happen (and as we show in Section 4.3 it indeed does happen) during the test period that new products arrive;
Product Search

• Toy store

• Highly structured documents representing products

• Many fields:
  • age_max, age_min, arrived, available, bonus, price, brand, category, category_id, characters, description, etc, …
Product Search - Participation
Product Search - Participation

- 3 teams submitted runs for train queries
Product Search - Participation

- 3 teams submitted runs for train queries
  - UIS
Product Search - Participation

- 3 teams submitted runs for train queries
  - UIS
  - IRIT
Product Search - Participation

- 3 teams submitted runs for train queries
  - UIS
  - IRIT
  - GESIS
Product Search - Participation

• 3 teams submitted runs for train queries
  • UIS
  • IRIT
  • GESIS

• 5 runs submitted for test queries
Product Search - Participation

- 3 teams submitted runs for train queries
  - UIS
  - IRIT
  - GESIS
- 5 runs submitted for test queries
- One baseline
Product Search - Participation

- 3 teams submitted runs for train queries
  - UIS
  - IRIT
  - GESIS
- 5 runs submitted for test queries
- One baseline
  - Sorting by historical clicks
Product Search - Inventory

Fig. 3. Inventory changes during Round #1 of the challenge.

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.

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
Participants should update available products
Product Search - Inventory

• Participants **should** update available products

• Rankings **may** contain stale products
Product Search - Inventory

- Participants **should** update available products
- Rankings **may** contain stale products
- These products were removed **after** interleaving
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)
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
Table 4. Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
<td>0.053</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
<td>0.156</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.2827</td>
<td>54</td>
<td>137</td>
<td>508</td>
<td>699</td>
<td>0.936</td>
</tr>
<tr>
<td>Expected Outcome</td>
<td>0.28</td>
<td></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>
<td>0.785</td>
</tr>
</tbody>
</table>

Table 5. Round #2 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.5. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
<td>0.498</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.4795</td>
<td>82</td>
<td>89</td>
<td>596</td>
<td>767</td>
<td>0.646</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
<td>0.117</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
<td>0.229</td>
</tr>
</tbody>
</table>

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UIS-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for GESIS. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions.
Product Search - Results - #1

Table 4. Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
<td>0.053</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
<td>0.156</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.2827</td>
<td>54</td>
<td>137</td>
<td>508</td>
<td>699</td>
<td>0.936</td>
</tr>
<tr>
<td>Expected Outcome</td>
<td>0.28</td>
<td></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>
<td>0.785</td>
</tr>
</tbody>
</table>

Table 5. Round #2 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.5. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
<td>0.498</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.4795</td>
<td>82</td>
<td>89</td>
<td>596</td>
<td>767</td>
<td>0.646</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
<td>0.229</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
<td>0.117</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UIS-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for GESIS. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions. Due to inventory changes.
## Product Search - Results - #1

### Table 4.
Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
<td>0.053</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
<td>0.156</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.2827</td>
<td>54</td>
<td>137</td>
<td>508</td>
<td>699</td>
<td>0.936</td>
</tr>
<tr>
<td>Expected Outcome</td>
<td>0.28</td>
<td></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>
<td>0.785</td>
</tr>
</tbody>
</table>

### Table 5.
Round #2 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.5. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
<td>0.498</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.4795</td>
<td>82</td>
<td>89</td>
<td>596</td>
<td>767</td>
<td>0.646</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
<td>0.229</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
<td>0.117</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
</tbody>
</table>

randomly clicking user) in this round then was 0.5. In this round no submission outperformed the production system. However, the baseline, still somewhat surprisingly, performed the best albeit not significantly better than the production system. Given that the corrected outcome in Round #1 is only an estimate (that might be biased in favor of the submissions), we regard the Round #2 results a more accurate reflection of system performance.

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UIS-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for GESIS. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions.
### Table 4.
Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASeline</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>UI-S-MIRA</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
<td>0.053</td>
</tr>
<tr>
<td>UI-S-JERN</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
<td>0.156</td>
</tr>
<tr>
<td>UI-S-UIS</td>
<td>0.2827</td>
<td>54</td>
<td>137</td>
<td>508</td>
<td>699</td>
<td>0.936</td>
</tr>
<tr>
<td>Expected Outcome</td>
<td>0.28</td>
<td></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>
<td>0.785</td>
</tr>
</tbody>
</table>

### Table 5.
Round #2 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.5. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASeline</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
<td>0.498</td>
</tr>
<tr>
<td>UI-S-JERN</td>
<td>0.4795</td>
<td>82</td>
<td>89</td>
<td>596</td>
<td>767</td>
<td>0.646</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
<td>0.229</td>
</tr>
<tr>
<td>UI-S-MIRA</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
<td>0.117</td>
</tr>
<tr>
<td>UI-S-UIS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UI-S-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for GESIS. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions.
### Table 4. Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
<td>0.053</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
<td>0.156</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.2827</td>
<td>54</td>
<td>137</td>
<td>508</td>
<td>699</td>
<td>0.936</td>
</tr>
<tr>
<td><strong>Expected Outcome</strong></td>
<td><strong>0.28</strong></td>
<td></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>
<td>0.785</td>
</tr>
</tbody>
</table>

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UIS-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for GESIS. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions.
Product Search - Results - #2

Table 4. Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Expected Outcome</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
<td>0.053</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
<td>0.156</td>
</tr>
<tr>
<td>Expected Outcome</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.2827</td>
<td>54</td>
<td>137</td>
<td>508</td>
<td>699</td>
<td>0.936</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.2685</td>
<td>40</td>
<td>109</td>
<td>374</td>
<td>523</td>
<td>0.785</td>
</tr>
</tbody>
</table>

Table 5. Round #2 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.5. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
<td>0.498</td>
</tr>
<tr>
<td>Expected Outcome</td>
<td>0.5</td>
<td></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>
<td>0.646</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
<td>0.229</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
<td>0.117</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UIS-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for GESIS. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions.
Product Search - Results - #2

**Table 4.** Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>n/a</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
<td>0.053</td>
</tr>
<tr>
<td>n/a</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
<td>0.156</td>
</tr>
<tr>
<td>n/a</td>
<td>0.2827</td>
<td>54</td>
<td>137</td>
<td>508</td>
<td>699</td>
<td>0.936</td>
</tr>
<tr>
<td>Expected Outcome</td>
<td>0.28</td>
<td></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>
<td>0.785</td>
</tr>
</tbody>
</table>

**Table 5.** Round #2 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user is 0.5. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
<td>0.498</td>
</tr>
<tr>
<td>n/a</td>
<td>0.4795</td>
<td>82</td>
<td>89</td>
<td>596</td>
<td>767</td>
<td>0.646</td>
</tr>
<tr>
<td>n/a</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
<td>0.229</td>
</tr>
<tr>
<td>n/a</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
<td>0.117</td>
</tr>
<tr>
<td>n/a</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>n/a</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UIS-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for Gesis. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions.
### Table 4.
Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
<td>0.053</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
<td>0.156</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.2827</td>
<td>54</td>
<td>137</td>
<td>508</td>
<td>699</td>
<td>0.936</td>
</tr>
</tbody>
</table>

**Expected Outcome**

| Expected Outcome | 0.28 |

### Table 5.
Round #2 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.5. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
<td>0.498</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.4795</td>
<td>82</td>
<td>89</td>
<td>596</td>
<td>767</td>
<td>0.646</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
<td>0.229</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
<td>0.117</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UIS-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for GESIS. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions.
## Table 4.

Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.2685</td>
<td>40</td>
<td>109</td>
<td>374</td>
<td>523</td>
<td>0.785</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
<td>0.117</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
</tbody>
</table>

## Table 5.

Round #2 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.5. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
<td>0.498</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
<td>0.229</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.4795</td>
<td>82</td>
<td>89</td>
<td>596</td>
<td>767</td>
<td>0.646</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UIS-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for GESIS. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions.

---

fixed SOLR + click rerank
Table 4. Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th># Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
<td>0.053</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
<td>0.156</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.2827</td>
<td>54</td>
<td>137</td>
<td>508</td>
<td>699</td>
<td>0.936</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.3685</td>
<td>40</td>
<td>109</td>
<td>374</td>
<td>523</td>
<td>0.785</td>
</tr>
</tbody>
</table>

Table 5. Round #2 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.5. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th># Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
<td>0.498</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.4795</td>
<td>82</td>
<td>89</td>
<td>596</td>
<td>767</td>
<td>0.646</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
<td>0.229</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
<td>0.117</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UIS-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for GESIS. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions.
Table 4. Round #1 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.28, see Section 4.3. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.4691</td>
<td>91</td>
<td>103</td>
<td>467</td>
<td>661</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.3413</td>
<td>71</td>
<td>137</td>
<td>517</td>
<td>725</td>
<td>0.053</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.3277</td>
<td>58</td>
<td>119</td>
<td>488</td>
<td>665</td>
<td>0.156</td>
</tr>
<tr>
<td>Expected Outcome</td>
<td>0.28</td>
<td></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>
<td>0.785</td>
</tr>
</tbody>
</table>

Table 5. Round #2 results for the product search use-case ordered by Outcome. The expected outcome under a randomly clicking user for is 0.5. P-values are computed using a binomial test.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>#Wins</th>
<th>#Losses</th>
<th>#Ties</th>
<th>#Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>0.5284</td>
<td>93</td>
<td>83</td>
<td>598</td>
<td>774</td>
<td>0.498</td>
</tr>
<tr>
<td>UIS-JERN</td>
<td>0.4795</td>
<td>82</td>
<td>89</td>
<td>596</td>
<td>767</td>
<td>0.646</td>
</tr>
<tr>
<td>GESIS</td>
<td>0.4520</td>
<td>80</td>
<td>97</td>
<td>639</td>
<td>816</td>
<td>0.229</td>
</tr>
<tr>
<td>UIS-MIRA</td>
<td>0.4389</td>
<td>79</td>
<td>101</td>
<td>577</td>
<td>757</td>
<td>0.117</td>
</tr>
<tr>
<td>UIS-UIS</td>
<td>0.4118</td>
<td>84</td>
<td>120</td>
<td>527</td>
<td>731</td>
<td>0.014</td>
</tr>
<tr>
<td>IRIT</td>
<td>0.3990</td>
<td>79</td>
<td>119</td>
<td>593</td>
<td>791</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Apart from the fact that none of the submission systems outperformed the production system in this round then was 0.5. In this round no submission outperformed the production system. However, the baseline, still somewhat surprisingly, performed the best albeit not significantly better than the production system. Given that the corrected outcome in Round #1 is only an estimate (that might be biased in favor of the submissions), we regard the Round #2 results a more accurate reflection of system performance.

Apart from the fact that none of the submission systems outperformed the production system in this round, their relative performance also changed. Where UIS-JERN came second in Round #1, it won in Round #2, which is more in line with the expectation of the team that submitted this system and the one it overtook. Even more strikingly is the large jump in relative system performance for GESIS. This can most likely be explained by numerous fixes to problems encountered in Round #1. Other than this it is hard to draw any conclusions from these differences as almost all of them have extremely high p-values. We would require more data in order to obtain more stable results before we can draw any conclusions.
Use-case 2: Web Search

The web search use-case is provided by Seznam, a very large web search engine in the Czech Republic. See Figure 5 for a screenshot of the user interface.

Seznam serves almost half the country's search traffic and as such has very high site traffic. Queries are the typical web search queries, and thus are a mixed bag of transactional, navigational and transactional. In contrast to the product search use-case, apart from the scale and the query types, Seznam does not make raw document and query content available, rather features computed for documents and queries. This is much like any learning to rank dataset, such as Letor. Queries and documents are only identified by a unique identifier and for each query, the candidate documents are represented with sparse feature vectors. Seznam provided a total of 557 features. These features were not described in any way. The challenge with this use-case then is a learning to rank challenge.

As described in Section 3, the web search use-case also consists of a training and test phase. For the test phase, there were 97 queries. For the training phase 100 queries were provided. On average, for each query there were about 179 candidate documents. In total, there were 35,322 documents.
Web Search

- Learning to Rank setting
Web Search

• Learning to Rank setting
  • 557 features
Web Search

- Learning to Rank setting
  - 557 features
- >35K documents
Web Search

• Learning to Rank setting
  • 557 features
• >35K documents
• >0.5M impressions
Web Search - Participation
Web Search - Participation

• 6 teams submitted runs for train queries
Web Search - Participation

• 6 teams submitted runs for train queries
• 0 teams submitted runs for test queries
Web Search - Participation

• 6 teams submitted runs for train queries
• 0 teams submitted runs for test queries
  • We report only baselines
The LL4IR CLEF Lab attracted interest from dozens of teams. There were 12 active participants, in particular for the web search use case.

The API infrastructure developed for the LL4IR CLEF Lab offers the potential to host ongoing IR evaluations in a live setting. As such, it is planned that these "challenges" will continue on an ongoing basis post-CLEF, with an expanding number of use-cases as well as refinements to the existing use-cases. Our evaluation challenge is already underway at the time of writing, with some modifications to the initial setup. Moreover, this third round has already attracted more participants, but only 3 teams ended up submitting results for the official evaluation (excluding the baseline systems, provided by the organizers). We found that, while many researchers expressed and showed their interest in the lab, our setup with an API, including the baseline systems, provided by the organizers). We plan to ease this process of adapting to this new evaluation paradigm by providing even more examples and by steering away from a static test collection, which was a hurdle for many.

Web Search - Results

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>Wins</th>
<th>Losses</th>
<th>Ties</th>
<th>Impressions</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exploitative Baseline</strong></td>
<td>0.5527</td>
<td>3030</td>
<td>2452</td>
<td>19055</td>
<td>24537</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td><strong>Expected Outcome</strong></td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Uniform Baseline</strong></td>
<td>0.2161</td>
<td>430</td>
<td>1560</td>
<td>1346</td>
<td>3336</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Submission</th>
<th>Outcome</th>
<th>Wins</th>
<th>Losses</th>
<th>Ties</th>
<th>Impressions</th>
<th>p-value</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>
<td>&lt; 0.01</td>
</tr>
<tr>
<td><strong>Expected Outcome</strong></td>
<td>0.5</td>
<td></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>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>
the lab

use cases

conclusions
the lab
use cases
conclusions
Conclusions

• The first evaluation setup of its kind
Conclusions

• The first evaluation setup of its kind
  • With real users, real task, real interactions
Conclusions

• The first evaluation setup of its kind
  • With real users, real task, real interactions
• Two implemented use cases so far
Conclusions

• The first evaluation setup of its kind
  • With real users, real task, real interactions
• Two implemented use cases so far
  • Web search and Product search
Conclusions

• The first evaluation setup of its kind
  • With real users, real task, real interactions
• Two implemented use cases so far
  • Web search and Product search
• Developed an API (code publicly available)
Conclusions

• The first evaluation setup of its kind
  • With real users, real task, real interactions
• Two implemented use cases so far
  • Web search and Product search
• Developed an API (code publicly available)
• Interest from many teams
Conclusions

• The first evaluation setup of its kind
  • With real users, real task, real interactions

• Two implemented use cases so far
  • Web search and Product search

• Developed an API (code publicly available)

• Interest from many teams
  • Participation from some
Conclusions

• The first evaluation setup of its kind
  • With real users, real task, real interactions
• Two implemented use cases so far
  • Web search and Product search
• Developed an API (code publicly available)
• Interest from many teams
  • Participation from some
• No baselines were beaten, yet
Call

• If you …
  • … own
  • … work at
  • … collaborate with
• … a search engine, please consider joining LL4IR!
Call

• If you do IR research: participate
Call

• If you do IR research: participate

  • it is easy (example code runs out of the box)
Call

• If you do IR research: participate
  • it is easy (example code runs out of the box)
  • we run evaluation periods every 2 weeks
Call

• If you do IR research: participate
  • it is easy (example code runs out of the box)
  • we run evaluation periods every 2 weeks
  • next period in less then a week (plenty of time!)
Call

• If you do IR research: participate
  • it is easy (example code runs out of the box)
  • we run evaluation periods every 2 weeks
  • next period in less then a week (plenty of time!)
• Come to our Lab session Thursday Afternoon
Thank You

#LL4IR

Anne Schuth
University of Amsterdam

Krisztian Balog
University of Stavanger

Liadh Kelly
Trinity College Dublin

Thanks to:
- CLEF
- ESF ELIAS
- COMMIT
- REGIO Jatek
- Seznam

http://living-labs.net
@livinglabsnet