Towards a Feature-Rich Data Set for Personalised Access to Long-Tail Content

Christin Seifert, Jörg Schlötterer, Michael Granitzer
University of Passau, Germany
IAR@ACM SAC, Salamanca, 2015-04-16
Motivation

Motivation

Motivation

Unique Visitors (%)

Rank of the Site

Digital libraries, museums, archives

Motivation

Unique Visitors (%) vs. Rank of the Site

User-content distance

long tail resources

Digital libraries, museums, archives

Motivation

How to reduce this distance?

Motivation

Bring content to users.
Motivation

Bring content to users.

• “contextualised, personalised, automatic search”
• Just-in-time retrieval (JITR)
• optimisation needs ground truth data
• existing data sets not suitable
Motivation

Bring content to users.

• “contextualised, personalised, automatic search”
• Just-in-time retrieval (JITR)
• optimisation needs ground truth data
• existing data sets not suitable
Motivation

Bring content to users.

- “contextualised, personalised, automatic search”
- Just-in-time retrieval (JITR)
- optimisation needs ground truth data
- existing data sets not suitable
Motivation

Bring content to users.

- “contextualised, personalised, automatic search”
- Just-in-time retrieval (JITR)
- optimisation needs ground truth data
- existing data sets not suitable
Motivation

Bring content to users.

- “contextualised, personalised, automatic search”
- Just-in-time retrieval (JITR)
- optimisation needs ground truth data
- existing data sets not suitable
Motivation

Bring content to users.

- “contextualised, personalised, automatic search”
- Just-in-time retrieval (JITR)
- optimisation needs ground truth data
- existing data sets not suitable
Motivation

Bring content to users.

• “contextualised, personalised, automatic search”
• Just-in-time retrieval (JITR)
• optimisation needs ground truth data
• existing data sets not suitable
Motivation

Bring content to users.

- “contextualised, personalised, automatic search”
- Just-in-time retrieval (JITR)
- optimisation needs ground truth data
- existing data sets not suitable

long tail: specialised, diverse, unrelated, low usage per item
Approach
Overview

Ground truth data set for just-in-time retrieval of long tail content
Overview

What to collect?

Ground truth data set for just-in-time retrieval of long tail content

acquisition planning
Overview

What to collect?
- Acquisition planning

What to acquire?
- Acquisition procedure

Ground truth data set for just-in-time retrieval of long tail content
Overview

Ground truth data set for just-in-time retrieval of long tail content

- **What to collect?**
  - acquisition planning

- **How to acquire?**
  - acquisition procedure

- **What’s in there?**
  - data set statistics
Overview

Ground truth data set for just-in-time retrieval of long tail content

What to collect?
- acquisition planning

How to acquire?
- acquisition procedure

What’s in there?
- data set statistics

What for? (example)
- experiment
Acquisition Planning


---

Acquisition Planning

What to collect?

- Search engine viewed as black box (no direct access to search engine, federation of search, privacy aspects)

## Acquisition Planning

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse clicks (+target)</td>
<td>I</td>
</tr>
<tr>
<td>textual input</td>
<td>I</td>
</tr>
<tr>
<td>browsing history</td>
<td>I</td>
</tr>
<tr>
<td>browser profile (plugins, ...)</td>
<td>I</td>
</tr>
</tbody>
</table>

### User and Task

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>predefined task name</td>
<td>E</td>
</tr>
<tr>
<td>custom task name</td>
<td>E</td>
</tr>
<tr>
<td>task start-time</td>
<td>E</td>
</tr>
<tr>
<td>task end-time</td>
<td>E</td>
</tr>
<tr>
<td>level of expertise</td>
<td>E</td>
</tr>
<tr>
<td>topics relevant to task</td>
<td>E</td>
</tr>
<tr>
<td>demographic data</td>
<td>E</td>
</tr>
</tbody>
</table>

### Resources

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>recommendations desirable?</td>
<td>E</td>
</tr>
<tr>
<td>queries</td>
<td>E</td>
</tr>
<tr>
<td>preferred language of results</td>
<td>E</td>
</tr>
<tr>
<td>rating of results (binary)</td>
<td>E</td>
</tr>
<tr>
<td>overall helpfulness of results</td>
<td>E</td>
</tr>
</tbody>
</table>

---

Table 1: Overview of collected features. E refers to data collected by the user's extension interface (UI) and UI refers to user input (UI).
### Acquisition Planning

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse clicks (+target)</td>
<td>I</td>
</tr>
<tr>
<td>textual input</td>
<td>I</td>
</tr>
<tr>
<td>browsing history</td>
<td>I</td>
</tr>
<tr>
<td>browser profile (plugins, ...)</td>
<td>I</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User and Task</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>predefined task name</td>
<td>E</td>
</tr>
<tr>
<td>custom task name</td>
<td>E</td>
</tr>
<tr>
<td>task start-time</td>
<td>E</td>
</tr>
<tr>
<td>task end-time</td>
<td>E</td>
</tr>
<tr>
<td>level of expertise</td>
<td>E</td>
</tr>
<tr>
<td>topics relevant to task</td>
<td>E</td>
</tr>
<tr>
<td>demographic data</td>
<td>E</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resources</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>recommendations desirable?</td>
<td>E</td>
</tr>
<tr>
<td>queries</td>
<td>E</td>
</tr>
<tr>
<td>preferred language of results</td>
<td>E</td>
</tr>
<tr>
<td>rating of results (binary)</td>
<td>E</td>
</tr>
<tr>
<td>overall helpfulness of results</td>
<td>E</td>
</tr>
</tbody>
</table>

---

**Personalisation as post-processing**
Acquisition Planning

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse clicks (+target)</td>
<td>UI</td>
</tr>
<tr>
<td>textual input</td>
<td>UI</td>
</tr>
<tr>
<td>browsing history</td>
<td>UI</td>
</tr>
<tr>
<td>browser profile (plugins, ...)</td>
<td>UI</td>
</tr>
</tbody>
</table>

**User and Task**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>predefined task name</td>
<td>E UI</td>
</tr>
<tr>
<td>custom task name</td>
<td>E UI</td>
</tr>
<tr>
<td>task start-time</td>
<td>E UI</td>
</tr>
<tr>
<td>task end-time</td>
<td>E UI</td>
</tr>
<tr>
<td>level of expertise</td>
<td>E UI</td>
</tr>
<tr>
<td>topics relevant to task</td>
<td>E UI</td>
</tr>
<tr>
<td>demographic data</td>
<td>E Q</td>
</tr>
</tbody>
</table>

**Resources**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>recommendations desirable?</td>
<td>E UI</td>
</tr>
<tr>
<td>queries</td>
<td>E UI</td>
</tr>
<tr>
<td>preferred language of results</td>
<td>E UI</td>
</tr>
<tr>
<td>rating of results (binary)</td>
<td>E UI</td>
</tr>
<tr>
<td>overall helpfulness of results</td>
<td>E Q</td>
</tr>
</tbody>
</table>
Acquisition Prototype

How to acquire?

extension + screencasts @ http://purl.org/eexcess/datasets/umlt
Data Set Statistics

<table>
<thead>
<tr>
<th>users</th>
<th>tasks</th>
<th>anno</th>
<th>queries</th>
<th>views</th>
<th>ratings</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>217</td>
<td>1332</td>
<td>4562</td>
<td>3267</td>
<td>15043</td>
<td>10252</td>
</tr>
</tbody>
</table>

5. DATA SET ANALYSIS

In this section we provide a qualitative and quantitative analysis of the data set we collected. First, we provide a general overview and then we provide detailed analyses for predefined tasks (section 5.1), custom tasks (section 5.2) and the relation between the selected text and the query issued for the selection (section 5.3).

Table 2 provides an overall summary of the collected data. Column “views” refers to the number of times users opened the detail page of results. We count each single view and do not distinguish between different results or one result that has been opened multiple times. Column “time” accumulates the duration (in minutes) of all performed tasks, “anno” counts the total number of users’ annotations of web pages with any of the retrieved results. Interestingly, the amount of ratings is more than three times larger than the amount of result views. It is obvious that a lot of results have been rated solely by the short summary as provided in the result list. A potential reason for this is that users assessed the quality of additional information provided in the detailed view as below average in the questionnaire (2.75 on a 1-5 scale).

5.1 Predefined Tasks

This section provides an analysis of the collected data from a task-centric perspective, aggregated over all users. Table 3 presents statistics for the predefined tasks. Tasks prefixed with “A” are content consumption tasks and tasks prefixed with “B” refer to content creation tasks. For each task, annotations were made on one single web page, either within a web site or within one Wiki page. The total number of annotations varies considerably for each task, ranging from 55 (T.B2, wiki entry about an important person) to 112 (T.A8, English Wikipedia article about the Berlin Wall). However, the annotations per user and task do not vary significantly (mean = 11.5, SD = 1.3).

The average task duration for predefined tasks was about 74 minutes per user, with a minimum of 46 minutes (task A5.de) and a maximum of 121 minutes (task B1). Users needed about twice as much time for content creation tasks (112 minutes on average) than for content consumption tasks (65 minutes on average). This tendency is also reflected in the number of clicks amounting to 63 per task per user for content consumption and 224 for content creation tasks.

We collected 8,091 positive and 6,826 negative ratings, amounting to 14,917 ratings. This amount is explained by the assignment asking participants to rate at least 10 results for each search. On average, users rated equally positive and negative, we found no significant difference between the number of positive and negative ratings (Shapiro-Wilks test for normality, W = 0.9305, p = 0.2777 for negative, W = 0.9608, p = 0.7063 for positive ratings, paired T-Test at confidence level = 0.05: t = 1.5687, d f = 1 4, p = 0.139).

The column “pages visited” lists the number of visits beyond the predetermined page (i.e., the page to annotate or the page on which a wiki entry had to be created) and result page visits.

5.2 Custom Tasks

Apart from altogether 114 different predefined task executions, users performed 103 freely chosen tasks, and indicated that for 76 of them they would have liked recommendations. After grouping obviously equal tasks, a set of 18 distinct tasks remained. For example, “watch online video”, “watch-
Data Set Statistics

<table>
<thead>
<tr>
<th>users</th>
<th>tasks</th>
<th>anno</th>
<th>queries</th>
<th>views</th>
<th>ratings</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>217</td>
<td>1332</td>
<td>4562</td>
<td>3267</td>
<td>15043</td>
<td>10252</td>
</tr>
</tbody>
</table>

approx. 170 hours
### Data Set Statistics

<table>
<thead>
<tr>
<th>users</th>
<th>tasks</th>
<th>anno</th>
<th>queries</th>
<th>views</th>
<th>ratings</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>217</td>
<td>1332</td>
<td>4562</td>
<td>3267</td>
<td>15043</td>
<td>10252</td>
</tr>
</tbody>
</table>

approx. 170 hours

<table>
<thead>
<tr>
<th>task id</th>
<th>webpage topic</th>
<th>language</th>
<th>annotations -</th>
<th>+</th>
<th>time [min]</th>
<th>clicks</th>
<th>queries issued</th>
<th>results viewed</th>
<th>pages visited</th>
<th>expertise [0-10]</th>
<th>users</th>
</tr>
</thead>
<tbody>
<tr>
<td>T.A1</td>
<td>wikipedia: Munich</td>
<td>de</td>
<td>98</td>
<td>554 645</td>
<td>586</td>
<td>412</td>
<td>347</td>
<td>387</td>
<td>8</td>
<td>3.4</td>
<td>8</td>
</tr>
<tr>
<td>T.A1</td>
<td>wikipedia: Munich</td>
<td>en</td>
<td>81</td>
<td>487 481</td>
<td>430</td>
<td>381</td>
<td>315</td>
<td>230</td>
<td>11</td>
<td>3.2</td>
<td>8</td>
</tr>
<tr>
<td>T.A2</td>
<td>wikipedia: World War I</td>
<td>de</td>
<td>97</td>
<td>625 414</td>
<td>467</td>
<td>354</td>
<td>311</td>
<td>165</td>
<td>32</td>
<td>3.4</td>
<td>8</td>
</tr>
<tr>
<td>T.A2</td>
<td>wikipedia: World War I</td>
<td>en</td>
<td>96</td>
<td>588 463</td>
<td>461</td>
<td>373</td>
<td>219</td>
<td>147</td>
<td>12</td>
<td>4.0</td>
<td>8</td>
</tr>
<tr>
<td>T.A3</td>
<td>warfarehistorian.blogspot.de: warefare</td>
<td>en</td>
<td>108</td>
<td>468 846</td>
<td>705</td>
<td>991</td>
<td>596</td>
<td>294</td>
<td>25</td>
<td>2.1</td>
<td>8</td>
</tr>
<tr>
<td>T.A4</td>
<td>wikipedia: Women in Workforce</td>
<td>en</td>
<td>92</td>
<td>557 606</td>
<td>586</td>
<td>559</td>
<td>373</td>
<td>176</td>
<td>54</td>
<td>2.6</td>
<td>8</td>
</tr>
<tr>
<td>T.A5</td>
<td>habsburger.net: Franz III &amp; Napoleon</td>
<td>de</td>
<td>86</td>
<td>681 295</td>
<td>371</td>
<td>519</td>
<td>271</td>
<td>217</td>
<td>4</td>
<td>3.5</td>
<td>8</td>
</tr>
<tr>
<td>T.A6</td>
<td>wikipedia: Pope</td>
<td>de</td>
<td>91</td>
<td>662 448</td>
<td>467</td>
<td>497</td>
<td>356</td>
<td>229</td>
<td>18</td>
<td>3.4</td>
<td>8</td>
</tr>
<tr>
<td>T.A6</td>
<td>wikipedia: Pope</td>
<td>en</td>
<td>92</td>
<td>552 687</td>
<td>601</td>
<td>587</td>
<td>302</td>
<td>298</td>
<td>14</td>
<td>2.9</td>
<td>8</td>
</tr>
<tr>
<td>T.A7</td>
<td>britannica.com: French Revolution</td>
<td>en</td>
<td>92</td>
<td>588 554</td>
<td>461</td>
<td>495</td>
<td>238</td>
<td>228</td>
<td>35</td>
<td>4.0</td>
<td>8</td>
</tr>
<tr>
<td>T.A8</td>
<td>wikipedia: Berlin Wall</td>
<td>de</td>
<td>99</td>
<td>761 290</td>
<td>483</td>
<td>502</td>
<td>283</td>
<td>322</td>
<td>11</td>
<td>6.2</td>
<td>8</td>
</tr>
<tr>
<td>T.A8</td>
<td>wikipedia: Berlin Wall</td>
<td>en</td>
<td>112</td>
<td>702 524</td>
<td>610</td>
<td>423</td>
<td>291</td>
<td>261</td>
<td>12</td>
<td>5.9</td>
<td>8</td>
</tr>
<tr>
<td>T.B1</td>
<td>sights of a city</td>
<td>de</td>
<td>60</td>
<td>325 180</td>
<td>729</td>
<td>1612</td>
<td>144</td>
<td>97</td>
<td>142</td>
<td>6.0</td>
<td>6</td>
</tr>
<tr>
<td>T.B2</td>
<td>important person</td>
<td>de</td>
<td>55</td>
<td>294 211</td>
<td>692</td>
<td>1322</td>
<td>187</td>
<td>108</td>
<td>73</td>
<td>6.4</td>
<td>6</td>
</tr>
<tr>
<td>T.B3</td>
<td>historic event</td>
<td>de</td>
<td>63</td>
<td>247 182</td>
<td>609</td>
<td>1101</td>
<td>279</td>
<td>90</td>
<td>199</td>
<td>5.4</td>
<td>6</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
<td>1322</td>
<td>6826 8091</td>
<td>8266</td>
<td>10128</td>
<td>4512</td>
<td>3249</td>
<td>650</td>
<td>Ø 4.2</td>
<td>8 (max)</td>
</tr>
</tbody>
</table>
## Data Set Statistics

<table>
<thead>
<tr>
<th>users</th>
<th>tasks</th>
<th>anno</th>
<th>queries</th>
<th>views</th>
<th>ratings</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>217</td>
<td>1332</td>
<td>4562</td>
<td>3267</td>
<td>15043</td>
<td>10252</td>
</tr>
</tbody>
</table>

**approx. 170 hours**

<table>
<thead>
<tr>
<th>task id</th>
<th>webpage topic</th>
<th>language</th>
<th>annotations</th>
<th>ratings</th>
<th>time [min]</th>
<th>clicks</th>
<th>queries issued</th>
<th>results viewed</th>
<th>pages visited</th>
<th>expertise [0-10]</th>
<th>users</th>
</tr>
</thead>
<tbody>
<tr>
<td>T.A1</td>
<td>wikipedia: Munich</td>
<td>de</td>
<td>98</td>
<td>554</td>
<td>645</td>
<td>586</td>
<td>412</td>
<td>347</td>
<td>387</td>
<td>8</td>
<td>3.4</td>
</tr>
<tr>
<td>T.A1</td>
<td>wikipedia: Munich</td>
<td>en</td>
<td>81</td>
<td>487</td>
<td>481</td>
<td>430</td>
<td>381</td>
<td>315</td>
<td>230</td>
<td>11</td>
<td>3.2</td>
</tr>
<tr>
<td>T.A2</td>
<td>wikipedia: World War I</td>
<td>de</td>
<td>97</td>
<td>625</td>
<td>414</td>
<td>467</td>
<td>354</td>
<td>311</td>
<td>165</td>
<td>32</td>
<td>3.4</td>
</tr>
<tr>
<td>T.A2</td>
<td>wikipedia: World War I</td>
<td>en</td>
<td>96</td>
<td>588</td>
<td>463</td>
<td>461</td>
<td>373</td>
<td>219</td>
<td>147</td>
<td>12</td>
<td>4.0</td>
</tr>
<tr>
<td>T.A3</td>
<td>warfarehistorian.blogspot.de: warefare</td>
<td>en</td>
<td>108</td>
<td>468</td>
<td>846</td>
<td>705</td>
<td>991</td>
<td>596</td>
<td>294</td>
<td>25</td>
<td>2.1</td>
</tr>
<tr>
<td>T.A4</td>
<td>wikipedia: Women in Workforce</td>
<td>en</td>
<td>92</td>
<td>557</td>
<td>606</td>
<td>586</td>
<td>559</td>
<td>373</td>
<td>176</td>
<td>54</td>
<td>2.6</td>
</tr>
<tr>
<td>T.A5</td>
<td>habsburger.net: Franz III &amp; Napoleon</td>
<td>de</td>
<td>86</td>
<td>681</td>
<td>295</td>
<td>371</td>
<td>519</td>
<td>271</td>
<td>217</td>
<td>4</td>
<td>3.5</td>
</tr>
<tr>
<td>T.A6</td>
<td>wikipedia: Pope</td>
<td>de</td>
<td>91</td>
<td>662</td>
<td>448</td>
<td>467</td>
<td>497</td>
<td>356</td>
<td>229</td>
<td>18</td>
<td>3.4</td>
</tr>
<tr>
<td>T.A6</td>
<td>wikipedia: Pope</td>
<td>en</td>
<td>92</td>
<td>552</td>
<td>687</td>
<td>601</td>
<td>587</td>
<td>302</td>
<td>298</td>
<td>14</td>
<td>2.9</td>
</tr>
<tr>
<td>T.A7</td>
<td>britannica.com: French Revolution</td>
<td>en</td>
<td>92</td>
<td>588</td>
<td>554</td>
<td>461</td>
<td>495</td>
<td>238</td>
<td>228</td>
<td>35</td>
<td>4.0</td>
</tr>
<tr>
<td>T.A8</td>
<td>wikipedia: Berlin Wall</td>
<td>de</td>
<td>99</td>
<td>761</td>
<td>290</td>
<td>483</td>
<td>502</td>
<td>283</td>
<td>322</td>
<td>11</td>
<td>6.2</td>
</tr>
<tr>
<td>T.A8</td>
<td>wikipedia: Berlin Wall</td>
<td>en</td>
<td>112</td>
<td>702</td>
<td>524</td>
<td>610</td>
<td>423</td>
<td>291</td>
<td>261</td>
<td>12</td>
<td>5.9</td>
</tr>
<tr>
<td>T.B1</td>
<td>sights of a city</td>
<td>de</td>
<td>60</td>
<td>325</td>
<td>180</td>
<td>729</td>
<td>1612</td>
<td>144</td>
<td>97</td>
<td>142</td>
<td>6.0</td>
</tr>
<tr>
<td>T.B2</td>
<td>important person</td>
<td>de</td>
<td>55</td>
<td>294</td>
<td>211</td>
<td>692</td>
<td>1322</td>
<td>187</td>
<td>108</td>
<td>73</td>
<td>6.4</td>
</tr>
<tr>
<td>T.B3</td>
<td>historic event</td>
<td>de</td>
<td>63</td>
<td>247</td>
<td>182</td>
<td>609</td>
<td>1101</td>
<td>279</td>
<td>90</td>
<td>199</td>
<td>5.4</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
<td>1322</td>
<td>6826</td>
<td>8091</td>
<td>8266</td>
<td>10128</td>
<td>4512</td>
<td>3249</td>
<td>650</td>
<td>8 (max)</td>
</tr>
</tbody>
</table>

---

### Ratings
- **Positive ratings:**
- **Negative ratings:**

### Task Distribution
- **Predefined tasks (section 5.1):**
- **Custom tasks (section 5.2):**
- **Content consumption tasks (excluding predefined tasks):**

### Time Duration
- **Minimum:** 46 minutes
- **Maximum:** 121 minutes
- **Average:** 74 minutes

### Additional Information
- **Annotations:**
- **Queries:**
- **Views:**
- **Ratings:**

---

### Data Quality
- **Data Cleaning:**
- **Feature Engineering:**

---

### Analysis
- **Task Reoccurrence:**
- **User Familiarity:**
- **Recommendation Performance:**

---

### Conclusion
- **Overall Implications:**
- **Future Directions:**
### Data Set Statistics

<table>
<thead>
<tr>
<th>webpage topic</th>
<th>language</th>
<th>annotations</th>
<th>ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>wikipedia: Munich</td>
<td>de</td>
<td>98</td>
<td>554</td>
</tr>
<tr>
<td>wikipedia: Munich</td>
<td>en</td>
<td>81</td>
<td>487</td>
</tr>
<tr>
<td>wikipedia: World War I</td>
<td>de</td>
<td>97</td>
<td>625</td>
</tr>
<tr>
<td>wikipedia: World War I</td>
<td>en</td>
<td>96</td>
<td>588</td>
</tr>
<tr>
<td>warfarehistorian.blogspot.de: warefare</td>
<td>en</td>
<td>108</td>
<td>468</td>
</tr>
<tr>
<td>wikipedia: Women in Workforce</td>
<td>en</td>
<td>92</td>
<td>557</td>
</tr>
<tr>
<td>habsburger.net: Franz III &amp; Napoleon</td>
<td>de</td>
<td>86</td>
<td>681</td>
</tr>
<tr>
<td>wikipedia: Pope</td>
<td>de</td>
<td>91</td>
<td>662</td>
</tr>
<tr>
<td>wikipedia: Pope</td>
<td>en</td>
<td>92</td>
<td>552</td>
</tr>
<tr>
<td>britannica.com: French Revolution</td>
<td>en</td>
<td>92</td>
<td>588</td>
</tr>
<tr>
<td>wikipedia: Berlin Wall</td>
<td>de</td>
<td>99</td>
<td>761</td>
</tr>
<tr>
<td>wikipedia: Berlin Wall</td>
<td>en</td>
<td>112</td>
<td>702</td>
</tr>
<tr>
<td>sights of a city</td>
<td>de</td>
<td>60</td>
<td>325</td>
</tr>
<tr>
<td>important person</td>
<td>de</td>
<td>55</td>
<td>294</td>
</tr>
<tr>
<td>historic event</td>
<td>de</td>
<td>63</td>
<td>247</td>
</tr>
</tbody>
</table>
## Data Set Statistics

<table>
<thead>
<tr>
<th>webpage topic</th>
<th>language</th>
<th>annotations</th>
<th>ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>wikipedia: Munich</td>
<td>de</td>
<td>98</td>
<td>554</td>
</tr>
<tr>
<td>wikipedia: Munich</td>
<td>en</td>
<td>81</td>
<td>487</td>
</tr>
<tr>
<td>wikipedia: World War I</td>
<td>de</td>
<td>97</td>
<td>625</td>
</tr>
<tr>
<td>wikipedia: World War I</td>
<td>en</td>
<td>96</td>
<td>588</td>
</tr>
<tr>
<td>warfarehistorian.blogspot.de: warefare</td>
<td>en</td>
<td>108</td>
<td>468</td>
</tr>
<tr>
<td>wikipedia: Women in Workforce</td>
<td>en</td>
<td>92</td>
<td>557</td>
</tr>
<tr>
<td>habsburger.net: Franz III &amp; Napoleon</td>
<td>de</td>
<td>86</td>
<td>681</td>
</tr>
<tr>
<td>wikipedia: Pope</td>
<td>de</td>
<td>91</td>
<td>662</td>
</tr>
<tr>
<td>wikipedia: Pope</td>
<td>en</td>
<td>92</td>
<td>552</td>
</tr>
<tr>
<td>britannica.com: French Revolution</td>
<td>en</td>
<td>92</td>
<td>588</td>
</tr>
<tr>
<td>wikipedia: Berlin Wall</td>
<td>de</td>
<td>99</td>
<td>761</td>
</tr>
<tr>
<td>wikipedia: Berlin Wall</td>
<td>en</td>
<td>112</td>
<td>702</td>
</tr>
<tr>
<td>sights of a city</td>
<td>de</td>
<td>60</td>
<td>325</td>
</tr>
<tr>
<td>important person</td>
<td>de</td>
<td>55</td>
<td>294</td>
</tr>
<tr>
<td>historic event</td>
<td>de</td>
<td>63</td>
<td>247</td>
</tr>
</tbody>
</table>
### Data Set Statistics

<table>
<thead>
<tr>
<th>webpage topic</th>
<th>language</th>
<th>annotations</th>
<th>ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>wikipedia: Munich</td>
<td>de</td>
<td>98</td>
<td>554 645</td>
</tr>
<tr>
<td>wikipedia: Munich</td>
<td>en</td>
<td>81</td>
<td>487 481</td>
</tr>
<tr>
<td>wikipedia: World War I</td>
<td>de</td>
<td>97</td>
<td>625 414</td>
</tr>
<tr>
<td>wikipedia: World War I</td>
<td>en</td>
<td>96</td>
<td>588 463</td>
</tr>
<tr>
<td>warfarehistorian.blogspot.de: warefare</td>
<td>en</td>
<td>108</td>
<td>468 846</td>
</tr>
<tr>
<td>wikipedia: Women in Workforce</td>
<td>en</td>
<td>92</td>
<td>557 606</td>
</tr>
<tr>
<td>habsburger.net: Franz III &amp; Napoleon</td>
<td>de</td>
<td>86</td>
<td>681 295</td>
</tr>
<tr>
<td>wikipedia: Pope</td>
<td>de</td>
<td>91</td>
<td>662 448</td>
</tr>
<tr>
<td>wikipedia: Pope</td>
<td>en</td>
<td>92</td>
<td>552 687</td>
</tr>
<tr>
<td>britannica.com: French Revolution</td>
<td>en</td>
<td>92</td>
<td>588 554</td>
</tr>
<tr>
<td>wikipedia: Berlin Wall</td>
<td>de</td>
<td>99</td>
<td>761 290</td>
</tr>
<tr>
<td>wikipedia: Berlin Wall</td>
<td>en</td>
<td>112</td>
<td>702 524</td>
</tr>
<tr>
<td>sights of a city</td>
<td>de</td>
<td>60</td>
<td>325 180</td>
</tr>
<tr>
<td>important person</td>
<td>de</td>
<td>55</td>
<td>294 211</td>
</tr>
<tr>
<td>historic event</td>
<td>de</td>
<td>63</td>
<td>247 182</td>
</tr>
</tbody>
</table>

Users rated equally positive and negative (T-test @0.05, p=0.139)
Data Set Statistics

- freely chosen tasks

What’s in there?
Automatic Query Generation

Can we predict manual queries from a text selection?

Automatic Query Generation

Can we predict manual queries from a text selection?

“.. The gathering of troops around Paris and the dismissal of Necker provoked insurrection in the capital. On July 14, 1789, the Parisian crowd seized the Bastille, a symbol of royal tyranny. Again the king had to yield; visiting Paris, he showed his recognition of the sovereignty of the people by wearing the tricolour cockade...” [1]

Can we predict manual queries from a text selection?

“.. The gathering of troops around Paris and the dismissal of Necker provoked insurrection in the capital. On July 14, 1789, the Parisian crowd seized the Bastille, a symbol of royal tyranny. Again the king had to yield; visiting Paris, he showed his recognition of the sovereignty of the people by wearing the tricolour cockade...” [1]

storming Bastille 1789

Automatic Query Generation

Can we predict manual queries from a text selection?

“.. The gathering of troops around Paris and the dismissal of Necker provoked insurrection in the capital. **On July 14, 1789, the Parisian crowd seized the Bastille, a symbol of royal tyranny.** Again the king had to yield; visiting Paris, he showed his recognition of the sovereignty of the people by wearing the tricolour cockade...” [1]

storming Bastille 1789

• 2499 text selection - query pairs
• features: term identity, case-sensitivity?, POS-tag type, stop word?, + features of preceding and succeeding term
• 29 feature combinations, CRF, cross-validation

Automatic Query Generation

<table>
<thead>
<tr>
<th></th>
<th>feature set</th>
<th>trivial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i, c, t</td>
<td>i, t</td>
</tr>
<tr>
<td>users mean</td>
<td>76</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>tasks mean</td>
<td>82</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>10-fold mean</td>
<td>89</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

- i - the identity of a term, i.e. the term itself
- c - whether the term begins with upper- or lowercase
- t - POS tag

- Manual queries can be predicted with high accuracy from text selection
- Feature set \{term identity, case-sensitivity, and POS tag\} is best
- Model better transferable across tasks than across users
### Automatic Query Generation

#### Table 4: Accuracies [%] for query prediction from selected text

<table>
<thead>
<tr>
<th></th>
<th>feature set</th>
<th>trivial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i, c, t$</td>
<td>$i, t$</td>
</tr>
<tr>
<td>users mean</td>
<td>76</td>
<td>77</td>
</tr>
<tr>
<td>SD</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>tasks mean</td>
<td>82</td>
<td>83</td>
</tr>
<tr>
<td>SD</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>10-fold mean</td>
<td>89</td>
<td>88</td>
</tr>
<tr>
<td>SD</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

- $i$ - the identity of a term, i.e. the term itself
- $c$ - whether the term begins with upper- or lowercase
- $t$ - POS tag

» Manual queries can be predicted with high accuracy from text selection
» Feature set \{term identity, case-sensitivity, and POS tag\} is best
» Model better transferable across tasks than across users
Automatic Query Generation

<table>
<thead>
<tr>
<th>feature set</th>
<th>users mean</th>
<th>tasks mean</th>
<th>10-fold mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i, c, t</td>
<td>i, t</td>
<td>c, t</td>
</tr>
<tr>
<td>i - the identity of a term, i.e. the term itself</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c - whether the term begins with upper- or lowercase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t - POS tag</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trivial rejector</td>
<td>76</td>
<td>82</td>
<td>89</td>
</tr>
<tr>
<td>trivial acceptor</td>
<td>77</td>
<td>83</td>
<td>88</td>
</tr>
<tr>
<td>SD</td>
<td>15</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>SD</td>
<td>15</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>SD</td>
<td>18</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>SD</td>
<td>35</td>
<td>35</td>
<td>2</td>
</tr>
<tr>
<td>SD</td>
<td>35</td>
<td>35</td>
<td>2</td>
</tr>
</tbody>
</table>

» Manual queries can be predicted with high accuracy from text selection
» Feature set \{\text{term identity, case-sensitivity, and POS tag}\} is best
» Model better transferable across tasks than across users
Summary and Outlook
Summary

- Feature analysis for ground truth data acquisition based on literature
- Prototype for data collection (backend europeana.eu) available
- Data set with 170 hours of user interactions available
- Experiment on query prediction from user selection

GitHub: http://purl.org/eexcess/datasets/umlt
Ada Lovelace

From Wikipedia, the free encyclopedia

Augusta Ada King, Countess of Lovelace (10 December 1815 – 27 November 1852), born Augusta Ada Byron and now commonly known as Ada Lovelace, was an English mathematician and writer chiefly known for her work on Charles Babbage's early mechanical general-purpose computer, the Analytical Engine. Her notes on the engine include what is recognised as the first algorithm intended to be carried out by a machine. Because of this, she is regarded as the first computer programmer.[1][2][3]

Lovelace was born 10 December 1815 as the only child of the poet Lord Byron and his wife Anne Isabella Byron.[4] All Byron's other children were born out of wedlock to other women.[5] Byron separated from his wife a month after Ada was born and left England forever four months later, eventually dying of disease in the Greek War of Independence when Ada was eight years old. Ada's mother remained bitter towards Lord Byron and promoted Ada's interest in mathematics and logic in an effort to prevent her from developing
Ada Lovelace

From Wikipedia, the free encyclopedia

Augusta Ada King, Countess of Lovelace (10 December 1815 – 27 November 1852), born Augusta Ada Byron and now commonly known as Ada Lovelace, was an English mathematician and writer chiefly known for her work on Charles Babbage’s early mechanical general-purpose computer, the Analytical Engine. Her notes on the engine include what is recognised as the first algorithm intended to be carried out by a machine. Because of this, she is regarded as the first computer programmer.[1][2]

Lovelace was born 10 December 1815 as the only child of the poet Lord Byron and his wife Anne Isabella Byron.[4] All Byron’s other children were born out of wedlock to other women.[5] Byron separated from his wife a month after Ada was born and left England forever four months later, eventually dying of disease in the Greek War of Independence when Ada was eight years old. Ada’s mother remained bitter towards Lord Byron and promoted Ada’s interest in mathematics and logic in an effort to prevent her from developing...
Ada Lovelace was born 10 December 1815 as the only child of the poet Lord Byron and his wife Anne Isabella Byron.[4] All Byron's other children were born out of wedlock to other women.[5] Byron separated from his wife a month after Ada was born and left England forever four months later, eventually dying of disease in the Greek War of Independence when Ada was eight years old. Ada's mother remained bitter towards Lord Byron and promoted Ada's interest in mathematics and logic in an effort to prevent her from developing.
Ada Lovelace

Ada, Countess of Lovelace

*Ada King, Countess of Lovelace* (10 December 1815 – 27 November 1852), born *Augusta Ada Byron* and now commonly known as *Ada Lovelace*, was an English mathematician and writer chiefly known for her work on *Charles Babbage’s* early mechanical general-purpose computer, the *Analytical Engine*. Her notes on the engine include what is recognised as the first *algorithm* intended to be carried out by a machine. Because of this, she is regarded as the first computer programmer.

Loovelace was born 10 December 1815 as the only child of the poet *Lord Byron* and his wife *Anne Isabella Byron*. All Byron's other children were born out of wedlock to other women. Byron separated from his wife a month after Ada was born and left England forever four months later, eventually dying of disease in the *Greek War of Independence* when Ada was eight years old. Ada's mother remained bitter towards Lord Byron and promoted Ada's interest in mathematics and logic in an effort to turn her from an intellectual.
Outlook

Ada Lovelace

Ada Lovelace, Countess of Lovelace

From Wikipedia, the free encyclopedia

Ada Lovelace, Countess of Lovelace, 1840

Ada, Countess of Lovelace

Augosta Ada King, Countess of Lovelace (10 December 1815 – 27 November 1852), born Augusta Ada Byron and now commonly known as Ada Lovelace, was an English mathematician and writer chiefly known for her work on Charles Babbage's early mechanical general-purpose computer, the Analytical Engine. Her notes on the engine include what is recognised as the first algorithm intended to be carried out by a machine. Because of this, she is regarded as the first computer programmer.1[3][9]

Lovelace was born 10 December 1815 as the only child of the poet Lord Byron and his wife Anne Isabella Byron.4 All Byron's other children were born out of wedlock to other women.5 Byron separated from his wife a month after Ada was born and left England forever four months later, eventually dying of disease in the Greek War of Independence when Ada was eight years old. Ada's mother remained bitter towards Lord Byron and promoted Ada's interest in mathematics and logic in an attempt to prevent her from becoming a poet...

Results for a selection

Results for a paragraph

Results for a page
Questions?

christin.seifert@uni-passau.de

http://eexcess.eu

http://purl.org/eexcess/datasets/umlt
Appendix
Motivation

How do users find long tail content?
Motivation

How do users find long tail content?

Use major search engines
Motivation

How do users find long tail content?

Use major search engines

Not optimised for long-tail, privacy issues
Motivation

How do users find long tail content?

Use major search engines

*Not optimised for long-tail, privacy issues*

Use dedicated portals
Motivation

How do users find long tail content?

Use major search engines

Not optimised for long-tail, privacy issues

Use dedicated portals

Need to know of their existence
Motivation

How do users find long tail content?

Use major search engines

Not optimised for long-tail, privacy issues

Use dedicated portals

Need to know of their existence

Don’t know of existence of portals
Motivation

How do users find long tail content?

Use major search engines

Not optimised for long-tail, privacy issues

Use dedicated portals

Need to know of their existence

Don’t know of existence of portals

Not solvable
Motivation

How do users find long tail content?

- Use major search engines
  - Not optimised for long-tail, privacy issues
- Use dedicated portals
  - Need to know of their existence
- Don’t know of existence of portals
  - Not solvable

User finds content. Limited success.
Germany is a large country located in central Europe, covering an area of 357,021 square kilometres (137,846 sq mi) with a largely temperate seasonal climate. Its capital and largest city is Berlin. The seat of government is Berlin and Bonn. Germany is a major economic and political power and traditionally a leader in many cultural, theoretical and technical fields.

With 80.7 million inhabitants, Germany is the most populous member state in the European Union. After the United States, it is also the second most popular migration destination in the world. Germany has the world's fourth-largest economy by nominal GDP and the fifth-largest by PPP. As a global leader in several industrial and technological sectors, it is both the world's third-largest exporter and third-largest importer of goods. It is a developed country with a very high standard of living, featuring comprehensive social security that includes the world's oldest universal health care system. Known for its rich cultural and political history, Germany has been the home of many influential philosophers, artists, musicians, cineasts, entrepreneurs, scientists and inventors.

Germany was a founding member of the European Communities in 1957, which became the European Union in 1993. It is
As a global leader in several industrial and technological sectors, it is both the world's third-largest exporter and third-largest importer of goods. It is a developed country with a very high standard of living, featuring comprehensive social security that includes the world's oldest universal health care system. Known for its rich cultural and political history, Germany has been the home of many influential philosophers, artists, musicians, cinemasts, entrepreneurs, scientists and inventors. Germany was a founding member of the European Communities in 1957, which became the European Union in 1993. It is
As a global leader in several industrial and technological sectors, it is both the world's third-largest exporter and third-largest importer of goods. It is a developed country featuring comprehensive social security that includes the world's oldest universal health care system. Known for its rich cultural and political history, Germany has been the home of many influential philosophers, artists, musicians, cinemastas, entrepreneurs, scientists and inventors. Germany was a founding member of the European Communities in 1957, which became the European Union in 1993. It is
As a global leader in several industrial and technological sectors, it is both the world's third-largest exporter and third-largest importer of goods. It is a developed country featuring comprehensive social security that includes the world's oldest universal health care system. Known for its rich cultural and political history, Germany has been the home of many influential philosophers, artists, musicians, cineasts, entrepreneurs, scientists and inventors. Germany was a founding member of the European Communities in 1957, which became the European Union in 1993. It is
<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Exports</th>
<th>Date of information</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>World</td>
<td>$17,779,000,000,000[2]</td>
<td>2011 est.</td>
</tr>
<tr>
<td>1</td>
<td>China</td>
<td>$2,210,000,000,000</td>
<td>2013</td>
</tr>
<tr>
<td>—</td>
<td>European Union</td>
<td>$2,173,000,000,000</td>
<td>2012 est.</td>
</tr>
<tr>
<td>2</td>
<td>United States</td>
<td>$1,575,000,000,000</td>
<td>2013 est.</td>
</tr>
<tr>
<td>3</td>
<td>Germany</td>
<td>$1,493,000,000,000</td>
<td>2013 est.</td>
</tr>
<tr>
<td>4</td>
<td>United Kingdom</td>
<td>$813,200,000,000</td>
<td>2013 est.</td>
</tr>
<tr>
<td>5</td>
<td>Japan</td>
<td>$697,000,000,000</td>
<td>2013 est.</td>
</tr>
<tr>
<td>6</td>
<td>France</td>
<td>$578,600,000,000</td>
<td>2013 est.</td>
</tr>
<tr>
<td>7</td>
<td>Netherlands</td>
<td>$576,900,000,000</td>
<td>2013 est.</td>
</tr>
<tr>
<td>Rank</td>
<td>Country</td>
<td>Exports</td>
<td>Date of information</td>
</tr>
<tr>
<td>------</td>
<td>-----------------</td>
<td>-----------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>1</td>
<td>China</td>
<td>$2,210,000,000,000</td>
<td>2013</td>
</tr>
<tr>
<td>2</td>
<td>European Union</td>
<td>$2,173,000,000,000</td>
<td>2012 est.</td>
</tr>
<tr>
<td>3</td>
<td>United States</td>
<td>$1,575,000,000,000</td>
<td>2013 est.</td>
</tr>
<tr>
<td>3</td>
<td>Germany</td>
<td>$1,493,000,000,000</td>
<td>2013 est.</td>
</tr>
<tr>
<td>7</td>
<td>Netherlands</td>
<td>$576,900,000,000</td>
<td>2013 est.</td>
</tr>
</tbody>
</table>
Schleppende deutsche Exportentwicklung: zyklisch oder strukturell?

Published in: Wirtschaftsdienst : Zeitschrift für Wirtschaftspolitik ; 95 (2015)
Author: Peters, Heiko ; Schneider, Stefan
Availability: Full text

Deutschland (Subject) ... Bundesrepublik Deutschland (Subject) ...
Why to reduce the user-content distance?
Why to reduce the user-content distance?

Long-tail resources in context:
- Discover new information
- Verify facts
- Enrich existing information
User Context Detection

• Translate user context to information need
• Example — browser extension
User Context Detection

- Translate user context to information need
- Example — browser extension

Find out what users need

1. Results for a page
2. Results for a paragraph
3. Results for a selection

User context

Search Backend

Personalised Results