Measures for assessing the data freshness in Open Data portals

Sebastian Neumaier and Jürgen Umbrich
Vienna University of Economics and Business
Problem/Challenge

- How up-to-date are resources in Open Data portals?

- Required information for such a metric:
  - Change history of documents in portals

- Challenge:
  1. Collect available change history
  2. Estimated next change time to assess up-to-dateness

- Two scenarios:
  - Portal provider: wants to add freshness measure to metadata
  - Data consumer: updating of application, DB, etc.
Open Data Portals

- Single point of access
- Local and external resources
- Meta data
  - Title
  - Modification date
  - ...

- Typical software:
Sources of change information in OD portals

- **Push-based history:**
  - Data provider push change information to portal
    - If *local*, by uploading new version
    - If *external*, by updating a specific metadata field
      
      ```
      last_modified: "2013-09-25T00:00:00"
      ```

- **Pull-based history:**
  - **Age sampling:**
    - Access to latest change time of a resource (i.e., last-modified timestamp in *HTTP Header*)
      
      ```
      Last-Modified: Mon, 04 Nov 2013 13:00:08 GMT
      ETag: "21096456b7f7d72268dc99b3bf082565"
      ```

  - **Comparison sampling:**
    - Detect changes by monitoring and comparing the resources
Open Data Portal Watch

http://data.wu.ac.at/portalwatch/

- Periodically monitoring over 260 Open Data portals
- Metadata quality assessment
  - Uniform handling of metadata (using DCAT mapping)
- Evolution tracking & archiving
  - Meta data
  - Data
Available change information

- **CKAN**: age- and comparison-sampling required
- **Socrata & OpenDataSoft**: push-based possible

<table>
<thead>
<tr>
<th></th>
<th>CKAN</th>
<th>Socrata</th>
<th>OpenDataSoft</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLs</td>
<td>4,049,851</td>
<td>181,548</td>
<td>8757</td>
</tr>
<tr>
<td>distinct</td>
<td>2,116,940</td>
<td>165,966</td>
<td>8757</td>
</tr>
</tbody>
</table>
Estimation of next updates

- Evaluating three change estimation heuristics:
  - **Poisson process**
    - Cho and Garcia-Molina (2003) propose Poisson process model to estimate updates in the context of Web sites
  - **Markov chain approach**
    - Umbrich et al. (2015) use Markov chains to schedule next crawl times for URLs based on previous observed changes
  - **Empirical distribution**
    - Build empirical distribution of changes based on intervals
Estimation of next updates (cont’d)

- **Age sampling**
  - **Poisson distribution**
    - $X/T$ ($= \frac{\text{number of changes}}{\text{monitoring period}}$) as estimator for Poisson parameter
    - Compute next change time by considering $p$-quantiles
  - **Empirical distribution**
    - Use intervals between the observed last-modified times
    - $p$-quantiles of empirical distribution
Estimation of next updates (cont’d)

- **Comparison sampling**
  - Only binary information/states available:
    
    \[
    \begin{array}{c|c|c|c|c|c|c|c|c}
    & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \\
    \hline
    \hline
    0 & 3 & 3 & 6 \\
    1 & 1 & 2 & 3 \\
    \end{array}
    \]

- **Markov chain approach**
  - Probability of next change based on previous state, e.g.:

    \[
    P(1|0) = \frac{3}{6}
    \]
Evaluation Summary

- Controlled environment:
  - Evaluation using revision histories of Wikipedia articles
    - 1562 randomly Wiki articles with >3 years history and >30 revisions
    - Wiki change history does not follow Poisson distribution

- Different confidence values:
  - For fixed $p$, we report the ratio of successfully predicted updates

- Conclusion:
  - Markov chain approach best for comparison-based sampling
  - Empirical distribution best for push-based and age-based sampling

<table>
<thead>
<tr>
<th>Estimator</th>
<th>All</th>
<th>Regular</th>
<th>Irregular</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 0.7$ $S = 10d$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C_{EmpDist}$</td>
<td>0.59</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>$C_{ChoNaive}$</td>
<td>0.67</td>
<td>0.67</td>
<td>0.63</td>
</tr>
<tr>
<td>$C_{ChoImpr}$</td>
<td>0.66</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>$C_{UmbMarkov}$</td>
<td>0.51</td>
<td>0.76</td>
<td>0.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimator</th>
<th>All</th>
<th>Regular</th>
<th>Irregular</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 0.7$ $S = 50d$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C_{EmpDist}$</td>
<td>0.54</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>$C_{ChoNaive}$</td>
<td>0.65</td>
<td>0.36</td>
<td>0.63</td>
</tr>
<tr>
<td>$C_{ChoImpr}$</td>
<td>0.27</td>
<td>0.31</td>
<td>0.47</td>
</tr>
<tr>
<td>$C_{UmbMarkov}$</td>
<td>0.58</td>
<td>0.59</td>
<td>0.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimator</th>
<th>All</th>
<th>Regular</th>
<th>Irregular</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 0.9$ $S = 10d$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C_{EmpDist}$</td>
<td>0.81</td>
<td>0.87</td>
<td>0.80</td>
</tr>
<tr>
<td>$C_{ChoNaive}$</td>
<td>0.71</td>
<td>0.70</td>
<td>0.67</td>
</tr>
<tr>
<td>$C_{ChoImpr}$</td>
<td>0.57</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>$C_{UmbMarkov}$</td>
<td>0.88</td>
<td>0.94</td>
<td>0.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimator</th>
<th>All</th>
<th>Regular</th>
<th>Irregular</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 0.9$ $S = 10d$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C_{EmpDist}$</td>
<td>0.81</td>
<td>0.87</td>
<td>0.80</td>
</tr>
<tr>
<td>$C_{ChoNaive}$</td>
<td>0.71</td>
<td>0.70</td>
<td>0.67</td>
</tr>
<tr>
<td>$C_{ChoImpr}$</td>
<td>0.57</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>$C_{UmbMarkov}$</td>
<td>0.88</td>
<td>0.94</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Thank you for your attention

- **Goal**
  - Data Freshness estimation in Open Data

- **Challenge**
  - Collecting change history (push vs pull)

- **Approach**
  - Estimators for different scenarios
  - Empirical evaluation

Sebastian Neumaier
WU Vienna, Institute for Information Business
email: sebastian.neumaier@wu.ac.at
url: https://sebneumaier.wordpress.com/
twitter: @sebneum