Strategies for Efficiently Keeping Local Caches of Linked Open Data Sources Up-To-Date

Renata Dividino, Thomas Gottron, and Ansgar Scherp
ISWC - ESE 2015
Linked Data Evolution Experiment

Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, Aidan Hogan: Observing Linked Data Dynamics. ESWC 2013
Effects on Indices and Caches

Impact on the accuracy of indices!

Thomas Gottron, Christian Gottron: Perplexity of Index Models over Evolving Linked Data. ESWC 2014
Problem: Updates of Indices and Caches

How to keep data “fresh”, i.e. identical to data in the LOD sources?
Solution: Scheduling Update Strategies

Update strategies are need to keep caches up-to-date, i.e. synchronized to the LOD sources.
Keeping Caches of the LOD Sources Up-To-Date

HTTP Response

HEADER

Last-Modified: Tue, 15 Nov 1994 12:45:26 GMT

CONTENT

Renata Queiroz Dividino, André Kramer, Thomas Gottron: An Investigation of HTTP Header Information for Detecting Changes of Linked Open Data Sources. ESWC (Satellite Events) 2014

Evaluation: Existing Scheduling Strategies for the WWW
Evaluation: Scheduling Update Strategies

Prioritized by:

1. PageRank [Page et al., 1999, Boldi et al., 2004, Cho et al., 2007]
2. Site Size [Baeza-Yates et al., 2005]
3. Age or Last Visited [Cho et al., 2000]
5. Change Rate [Olston et al., 2008, Ntoulas et al., 2004, Dividino et al., 2013]
6. Change History Information [Dividino et al., 2014]
Page Rank

- It captures the popularity/importance of the LOD source.

LOD Source Size

- Data from the biggest/smallest LOD sources should be updated first.

Age or Last Visited

- The time elapsed from the last update (the difference between query time and last update time).
Change Ratio

- It captures the quantity change.
- How much of data items in the cache that are out-to-date.
Change Rate

- Comparison of two RDF data sets
  - \( X \): Set of triple statements
  - \( \Delta \): Distance function e.g. Jaccard, Dice, Cousine

\[
\Delta_{\text{Jaccard}}(X_1, X_2) = 1 - \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|}
\]
Dynamics

- It uses history information
- It quantifies the *evolution* of a dataset over a period of time via change rate

\[
\theta (X_{t_j}) - \theta (X_{t_i}) = \int_{t_i}^{t_j} c(X_t) \, dt
\]
Evaluation: Data

- 154 snapshots (approx. 3 years)
- 590 data sources (PLD)

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Average Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia.org</td>
<td>3,406,364.5</td>
</tr>
<tr>
<td>edgarwrap.ontologycentral.com</td>
<td>982,631.0</td>
</tr>
<tr>
<td>dbtune.org</td>
<td>864,107.6</td>
</tr>
<tr>
<td>dbtropes.org</td>
<td>787,299.9</td>
</tr>
<tr>
<td>data.linkedct.org</td>
<td>498,986.3</td>
</tr>
<tr>
<td>aims.fao.org</td>
<td>416,708.9</td>
</tr>
<tr>
<td><a href="http://www.legislation.gov.uk">www.legislation.gov.uk</a></td>
<td>399,601.6</td>
</tr>
<tr>
<td>kent.zpr.fer.hr</td>
<td>387,034.8</td>
</tr>
<tr>
<td>identi.ca</td>
<td>278,316.2</td>
</tr>
<tr>
<td>webenemasuno.linkeddata.es</td>
<td>250,557.9</td>
</tr>
</tbody>
</table>

We use data from the Dynamic Linked Data Observatory

Weekly snapshots, 14M triples
Evaluation: Single Step Update

Which strategy is the most appropriated one to keep the cache up-to-date?
Evaluation: Iterative Updates

Simulates a LOD search engine continuously updating its caches
Evaluation: Metrics

- Precision: portion of cached data that are up-to-dated
- Recall: portion of data in the LOD cloud that is identical to the cached data
Results: Single Step Update

![Graph showing precision vs bandwidth for different update strategies.]

- **BiggestFirst**
- **SmallestFirst**
- **PageRank**
- **ChangeRate-J**
- **ChangeRate-D**
- **ChangeRatio**
- **Dynamics-J**
- **Dynamics-D**

Precision is shown on the Y-axis, and bandwidth is on the X-axis.

- At 100% bandwidth, *BiggestFirst* shows the highest precision, followed by *SmallestFirst*.
- At lower bandwidths, the precision for *PageRank* is consistently higher than the others.

**Key points**:
- *BiggestFirst* updates the most important data first, leading to higher precision initially.
- *PageRank* prioritizes updates based on importance scores, maintaining high precision across bandwidths.
- Other strategies may not perform as well at lower bandwidths.
Results: Single Step Update
Experiments - Iterative Updates

![Graph showing iterative updates over time with exponentially increasing cache sizes.](image)

Each iteration increases the cache size by 60% to 95%, starting from 100% and ending at 5%, 40%, 15%, and 5% cache sizes. The graphs show the precision over iterations for various methods (BiggestFirst, SmallestFirst, PageRank, ChangeRate, ChangeRate-D, ChangeRatio, Dynamics, Dynamics-D, and Age).
Experiments - Iterative Updates
Efficiently Keeping Local Caches of Linked Open Data Sources Up-To-Date

Experiments - Iterative Updates

![Graph showing iterative updates over time with different percentages](image)

- $t_i$: Initial time
- $t_f$: Final time
- 60%: 40% of updates
- 95%: 5% of updates
- 100%: Complete update
- 5%, 15%, 40%: Different iteration percentages

Graphs illustrating recall over iterations with different update percentages.
Experiments - Iterative Updates

The diagram illustrates the efficiency of keeping local caches of linked open data sources up-to-date. It shows iterative updates over time, with different cache states at each iteration, indicated by the percentages (60%, 95%, 40%, 5%, 40%, 15%, 40%, 5%). The recall values decrease over iterations, with labels for different strategies like BiggestFirst, SmallestFirst, PageRank, ChangeRate-J, ChangeRate-D, ChangeRatio, Dynamics-J, Dynamics-D, and Age. The graph shows a decline in recall from 0.96 to around 0.88 over 4 iterations.
Conclusion

- Evaluation: apply existing strategies for maintaining indices of Web documents to LOD sources
- Goal: keep a cache as 'fresh' as possible despite of limitations
- Data: weekly snapshots – weekly changes
- Most appropriate strategies: capture the change behaviour (dynamics) over time
- Specially for low-relative-bandwidth (strong limitations)
- Not consider strategies dealing with topic/query/user-centred coverage.

Future work

- Impact of updates when combining strategies
- Analysis of the influence of sources that go offline
Strategies for Efficiently Keeping Local Caches of Linked Open Data Sources Up-To-Date

THANKS!

Renata Dividino, Thomas Gottron, and Ansgar Scherp

dividino@uni-koblenz.de  gottron@uni-koblenz.de  asc@informatik.uni-kiel.de

References

- Baeza-Yates, Ricardo; Castillo, Carlos; Marin, Mauricio und Rodriguez, Andrea (2005), *Crawling a country: better strategies than breadth-first for web page ordering*, in: WWW ’05: Special interest tracks and posters
- Alexandros Ntoulas, Junghoo Cho, Christopher Olston: *What's new on the web?: the evolution of the web from a search engine perspective*. WWW 2004
- Renata Queiroz Dividino, Thomas Gottron, Ansgar Scherp, Gerd Gröner: *From Changes to Dynamics: Dynamics Analysis of Linked Open Data Sources*. PROFILES@ESWC 2014
- Renata Queiroz Dividino, André Kramer, Thomas Gottron: *An Investigation of HTTP Header Information for Detecting Changes of Linked Open Data Sources*. ESWC (Satellite Events) 2014: 199-203
- Renata Queiroz Dividino, Ansgar Scherp, Gerd Gröner, Thomas Grotton: *Change-a-LOD: Does the Schema on the Linked Data Cloud Change or Not?* COLD 2013