Traffic Analytics for Linked Data Publishers

Luca Costabello, Pierre-Yves Vandenbussche, Gofran Shukair
Fujitsu Ireland

Corine Deliot, Neil Wilson
British Library
The Problem: Measuring Traffic on RDF Datasets

- Linked Data publishers have limited awareness of how datasets are accessed by visitors.
- No tool to mine Linked Data servers access logs
- Why is this such a big deal?
  - Justify investment in Linked Data IT infrastructure
  - Cost control
  - Identify abuses
  - Interpret access peaks
Challenges

- Which traffic metrics?
  - Adapt conventional web analytics metrics
  - Define Linked-Data specific extensions

- How to extract and compute such metrics?
  - Which data sources? (client tracking? server access logs mining? both?)
  - Need to support dual data access protocol (HTTP operations + SPARQL)
  - How to filter noise? (i.e. robots, search engines crawlers)
  - How to detect client sessions? (no client tracking, dual data access protocol)
  - How to detect SPARQL activity peaks?
Filling a Gap in Prior Art

- Existing tools do not include Linked Data-specific metrics:

- Linked data-specific metrics, but no platform [Moller et al, WebScience 2010]

Learning from Linked Open Data Usage: Patterns & Metrics

Knud Möller, Michael Hausenblas, Richard Cyganiak, Siegfried Handschuh
Digital Enterprise Research Institute (DERI), National University of Ireland, Galway
IDAI Business Park, Lower Dangan, Galway, Ireland
knud.moeller@deri.org, michael.hausenblas@deri.org, richard.cyganiak@deri.org, siegfried.handschuh@deri.org

Gunnar AAastrand Grimnes
German Research Center for Artificial Intelligence (DFKI) GmbH, Knowledge Management Group
D-67663 Kaiserslautern, Germany
grimmno@dfki.uni-kl.de
Our Contribution / Agenda

- Traffic analytics platform for LD servers
- Metrics
- Metrics Extraction
  - Visitor Sessions
  - Heavy/Light SPARQL Queries
- Results & British Library Trial Insight
• Traffic analytics platform for LD servers
• Metrics
• Metrics Extraction
  • Visitor Sessions
  • Heavy/Light SPARQL Queries
• Results & British Library Trial Insight
• Traffic Analytics Platform for LD Servers
• **Metrics**
• Metrics Extraction
  • Visitor Sessions
  • Heavy/Light SPARQL Queries
• Results & British Library Trial Insight
## Content Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances*</td>
<td>How many times RDF instances have been requested (HTTP requests and URIs in SPARQL queries)</td>
</tr>
<tr>
<td>Classes*</td>
<td>The count of URIs used as RDFS/OWL classes in SPARQL queries. URIs must be objects of <code>rdf:type</code>.</td>
</tr>
<tr>
<td>Properties*</td>
<td>The count of URIs used as predicates in SPARQL queries.</td>
</tr>
<tr>
<td>Graphs*</td>
<td>The count of URIs used as graphs in SPARQL queries (FROM/FROM NAMED, USING/USING NAMED, GRAPH).</td>
</tr>
</tbody>
</table>

* Linked Data-specific
**Protocol Metrics**

<table>
<thead>
<tr>
<th>Data Access Protocol*</th>
<th>The count of HTTP lookups and SPARQL queries, over a specific time frame.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPARQL Query type*</td>
<td>The count of SPARQL verbs over a specific time frame (e.g. how many SELECT, ASK, DESCRIBE, and CONSTRUCT).</td>
</tr>
<tr>
<td>Light/Heavy SPARQL*</td>
<td>Indicates the number of light and heavy SPARQL queries (see Section 4.2 for details).</td>
</tr>
<tr>
<td>HTTP Methods</td>
<td>Indicates the count of HTTP verbs (e.g. GET, POST, HEAD).</td>
</tr>
<tr>
<td>303 Patterns*</td>
<td>The count of HTTP 303 patterns found in logs (303 URIs).</td>
</tr>
<tr>
<td>Misses</td>
<td>Keeping track of HTTP 404s is useful to understand whether visitors are looking for resources which are not currently included in the dataset.</td>
</tr>
<tr>
<td>Malformed queries*</td>
<td>The count of HTTP 400s shows how many malformed HTTP requests have been issued. We also show how many SPARQL queries contain syntax errors.</td>
</tr>
<tr>
<td>Other client-side errors</td>
<td>Useful, for example, to detect whether visitors attempt to access forbidden RDF resources (HTTP 403).</td>
</tr>
<tr>
<td>Server-side errors</td>
<td>The count of HTTP 5xxs is important to estimate whether error-triggereing SPARQL queries have been repeatedly issued to a triplestore.</td>
</tr>
</tbody>
</table>

* Linked Data-specific
<table>
<thead>
<tr>
<th><strong>Location</strong></th>
<th>Continent, country, subdivision, and city of origin of a visitor.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network Provider</strong></td>
<td>The name of the company or institute associated to a visitor’s IP address, as returned by WHOIS lookups. If WHOIS data is not available, we provide the visitor’s host network.</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>The preferred language requested by a visitor. Such information is extracted from the Accept-Language HTTP header (for HTTP lookups) and retrieved from SPARQL xsd language-tagged string literals (e.g. @en) or FILTER lang()s (e.g. lang(?1) = &quot;en&quot;).</td>
</tr>
<tr>
<td><strong>User Agent</strong></td>
<td>The user agent string provided by a visitor.</td>
</tr>
<tr>
<td><strong>User Agent Type</strong></td>
<td>To provide a clearer estimate of which clients are used to access a dataset, we group user agent strings into Software Libraries (e.g. Jena, Python SPARQL-client, etc.), Desktop Browsers, Mobile Browsers, and Others.</td>
</tr>
<tr>
<td><strong>New vs Returning</strong></td>
<td>New visitors versus visitors that have performed at least one visit before.</td>
</tr>
<tr>
<td><strong>External Referrer</strong></td>
<td>HTTP Referrer: headers. When dereferencing an RDF resource, the HTTP request might contain this optional header field that specifies the URI from which the request has been issued.</td>
</tr>
<tr>
<td><strong>Sessions Count</strong></td>
<td>The count of all visitors sessions.</td>
</tr>
<tr>
<td><strong>Session Size</strong></td>
<td>The number of requests sent by a visitor during a session (requests might be a mix of HTTP lookups and SPARQL queries).</td>
</tr>
<tr>
<td><strong>Session Depth</strong></td>
<td>The number of distinct RDF resources (graphs, classes, properties, instances) requested by a visitor during a session.</td>
</tr>
<tr>
<td><strong>Session Duration</strong></td>
<td>The duration of a session.</td>
</tr>
<tr>
<td><strong>Bounce Rate</strong></td>
<td>Indicates the percentage of sessions that contain only one resource request (whether this is an HTTP lookup or a SPARQL query).</td>
</tr>
</tbody>
</table>
• Traffic Analytics Platform for LD Servers
• Metrics
• **Metrics Extraction**
  • Visitor Sessions
  • Heavy/Light SPARQL Queries
• Results & British Library Trial Insight
Visitor Session Detection

- **Session**: sequence of requests issued with no significant interruptions by a uniquely identified visitor. Expires after a *period of inactivity*.

- We use the **HAC** variant by [Murray et al. 2006, Mehrzadi et al. 2012]
  - Unsupervised, gap-based session boundary detection
  - Traditional web logs analysis
  - **Benefit**: *visitor-specific* temporal cut-off

- Two-step procedure:
  - Set visitor-specific session cut-off as time interval that significantly increases the variance.
  - Group HTTP/SPARQL requests into sessions according to the cut-off
• Traffic Analytics Platform for LD Servers
• Metrics
• **Metrics Extraction**
  • Visitor Sessions
  • **Heavy/Light SPARQL Queries**
• Results & British Library Trial Insight
Heavy/Light SPARQL Queries Binary Classifier

- **Rough** estimate of *heavy* and *light* queries with **supervised binary classification**.
- **Heavy SPARQL Query**: if it requires considerable computational and memory resources.

```sql
SELECT ?x
WHERE
{?x a foaf:Person.}

SELECT ?u ?c ?l
WHERE {{
    ?u rdfs:comment ?c .
    FILTER(regex(str(?c), '([^\\\\\W]fr', 'i'))) UNION {
    OPTIONAL {?u skos:prefLabel ?l.}}}}
```
### Heavy/Light SPARQL Queries Binary Classifier

- **Feature vectors:** SPARQL 1.1 syntactic features *only:*
  
<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>● UNIONs count</td>
<td>● Triple patterns count</td>
<td>● SELECT * count</td>
<td>● FILTERs expressions count</td>
<td>● Join types count $(bub, bbb, uuu, ubu, ubb, bbu, uub, buu)^10</td>
</tr>
<tr>
<td>● OPTIONALs count</td>
<td></td>
<td></td>
<td>● FILTERs strstarts count</td>
<td></td>
</tr>
<tr>
<td>● Joins count</td>
<td></td>
<td></td>
<td>● FILTERs contains count</td>
<td></td>
</tr>
<tr>
<td>● FILTERs count</td>
<td></td>
<td></td>
<td>● FILTERs in count</td>
<td></td>
</tr>
<tr>
<td>● FILTERs regex count</td>
<td></td>
<td></td>
<td>● FILTERs or count</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>● FILTERs lang count</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>● Presence of ORDER BY</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>● Projections count</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>● Projection variables count</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Traffic Analytics Platform for LD Servers
• Metrics
• Metrics Extraction
  • Visitor Sessions
  • Heavy/Light SPARQL Queries
• Results & British Library Trial Insight
Datasets

- British National Bibliography access logs
  - bnb.data.bl.uk (access logs are not public)
  - 13 months
  - ~ 10M HTTP requests/month

- DBpedia 3.9 access logs
  - USEWOD 2015 Dataset
Visitor Session Detection: Results

- How well do we detect the beginning of a new session?

- Dataset
  - British National Bibliography access logs (3 consecutive days)
  - ~16k HTTP/SPARQL requests
    - 32% Desktop browsers (115 visitors)
    - 68% Software libraries (10 visitors)
  - Manually annotated records
    - 1=session_start / 0=internal

- Baseline: fixed-length cut-offs
- HAC outperforms fixed-length cut-offs

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAC</td>
<td>0.91</td>
<td>0.72</td>
<td>0.80</td>
</tr>
<tr>
<td>15 m</td>
<td>0.58</td>
<td>0.99</td>
<td>0.73</td>
</tr>
<tr>
<td>30 m</td>
<td>0.72</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>60 m</td>
<td>0.87</td>
<td>0.67</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Heavy/Light SPARQL: Experiment Protocol

- Random distinct queries from DBpedia 3.9 access logs
- Run the queries multiple times on local clone of DBpedia
- Kept ~3.7k queries with low variance (3.1k light, 600 heavy)
- Cut-off threshold: 100ms

![SPARQL Queries by Execution Time](image)

- Naïve Bayes and SVM
- Grid search & randomized search w/ 10-fold CV
# Heavy/Light SPARQL: Results

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>F1_{avg}</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S V C_{all}$</td>
<td>light</td>
<td>0.92</td>
<td>0.99</td>
<td>0.95</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>Heavy</td>
<td>0.87</td>
<td>0.54</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>$N B_{all}$</td>
<td>light</td>
<td>0.94</td>
<td>0.88</td>
<td>0.91</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Heavy</td>
<td>0.49</td>
<td>0.67</td>
<td>0.56</td>
<td></td>
</tr>
</tbody>
</table>

**Precision-Recall Curve**

- $A U C_{S V C_{all}} = 0.67$
- $A U C_{S V C_{2}} = 0.44$
- $A U C_{S V C_{3}} = 0.30$
- $A U C_{S V C_{4}} = 0.52$
- $A U C_{N B} = 0.53$
- $A U C_{S V C_{1}} = 0.28$
Some Insights on BL Traffic Logs

- Genuine calls account for 0.6% of total traffic!
- +30% of HTTP/SPARQL traffic over the observed 13 months
- Sharp increase in requests from Software Libraries (95x)
- SPARQL accounts for 29% of traffic
- 6% of heavy SPARQL queries
- 37 days have unusual traffic spikes
- Bounce rate: 48%
- Software Libraries have bigger, deeper, and longer sessions.
Summary

- We relieve publishers from manual and time-consuming access log mining

- Support Linked Data-specific metrics
  - Break down traffic by RDF content
  - Capture SPARQL insights
  - Properly interpret 303 patterns

- Reconstruction of Linked Data visitors sessions

- Heavy/light SPARQL classifier w/ SPARQL syntax + supervised learning

- Revealed hidden insights on 13 months of access logs of the British Library
Future Work

- Statistics on noise (i.e. web crawlers)
- Heavy/light classifier
  - Feature set refinements
  - Does it generalize to other datasets?
- Enhance session detection with content-based heuristics
  - Relatedness of subsequent SPARQL queries
  - Structure and type of requested RDF entities

[Graphs and charts depicting various data metrics]

innovation.ie.fujitsu.com/kedi