Smart Routing of Memento Requests

Martin Klein¹
Lyudmila Balakireva¹
Harihar Shankar¹
James Powell¹
Herbert Van de Sompel²

¹Research Library
Los Alamos National Laboratory

²Data Archiving and Networked Services
The Netherlands
Smart Routing of Memento Requests
@mart1nkle1n
IIPC WAC 2018, 11/15/2018, Wellington, NZ
Memento

http://timetravel.mementoweb.org/list/20140809200708/https://www.wellingtonnz.com/

Smart Routing of Memento Requests

@mart1nkle1n

IIPC WAC 2018, 11/15/2018, Wellington, NZ
Discover Wellington

Take a look at the things to do where to stay places to dine and the visitor essentials for Wellington. Plus if you're after business study or work information, we can help there too.

How does this work?
Memento Aggregator (very simplistic view)
Smart Routing of Memento Requests
@mart1nkle1n
IIPC WAC 2018, 11/15/2018, Wellington, NZ
LANL Memento Aggregator - Problem

- As the number of archives grows, sending requests to each archive for every incoming request is not feasible
  - Response times
  - Memento infrastructure load
  - Load on distributed archives
What if…

• We could predict, by merely looking at a URI-R, whether or not to issue a request to a specific archive?
  • A binary classifier per archive

• We could train the classifiers using cached data?

• That would be pretty neat, indeed:
  • Retrain classifiers as web archive collections evolve
  • Not dependent on external data
  • Querying classifiers probably way faster (msec) than polling archives (sec)
We can! Published @ JCDL 2016

- ML models based on simple URI features
  - Character count, n-grams, domain
- Common ML algorithms used per archive
  - Logistic Regression, Multinomial Bayes, SVM
- Optimized for
  - Prediction time, not training time
  - Reduction of false positive rate

**Results:**
- Requests per URI - R: 3.96 vs 11
- Response time:
  - $2.16s$ vs $3.08s$
- Recall:
  - 84.7%

https://doi.org/10.1145/2910896.2910899
Smart Routing of Memento Requests

@mart1nkle1n

IIPC WAC 2018, 11/15/2018, Wellington, NZ
In Production…
Smart Routing of Memento Requests

@mart1nkle1n

IIPC WAC 2018, 11/15/2018, Wellington, NZ
Smart Routing of Memento Requests

@mart1nkle1n

IIPC WAC 2018, 11/15/2018, Wellington, NZ
Populating the Cache

- URI_R
- Datetime
- URI-G
- IA
- LoC
- Arquivo.pt
- Perma.cc
- ...
- 18 others
- Cache
Questions to Ask

• How effective is the cache?
  • What is the hit/miss ratio? Does it vary for different Memento services?
  • Is the cache freshness period appropriate?

• How effective is the ML process?
  • What is the false negative and false positive rate?
  • Do we need to retrain the models? How often?
Evaluation

- Memento Aggregator currently covers
  - 23 web archives
  - 17 with native memento support
  - 6 with by-proxy memento support

- Analysis of log files
  - recorded between July 4th, 2017 and October 17th, 2018
  - > 11m requests in total
  - Approx. 2.6m requests against machine learning process
    - Results in 2.6m lookups to populate cache
      - Used as “truth” to assess ML prediction
Cache Hit/Miss Rate

Memento Cache Performance

- Hit
- Miss
- Stale

- api: 45%
- list: 73%
- memento: 25%
- tg: 82%

Smart Routing of Memento Requests
@mart1nkle1n
IIPC WAC 2018, 11/15/2018, Wellington, NZ
Cache Hit/Miss Rate

Mostly driven by machines humans machines humans

Smart Routing of Memento Requests
@mart1nkle1n
IIPC WAC 2018, 11/15/2018, Wellington, NZ
False Negatives by Number of Archives

Count

Number of Archives

0 1 2 3 4 5 6 7 8 9

1771572 722528 74479 18686 5704 2048 472 188 70 41

Smart Routing of Memento Requests
@mart1nkle1n
IIPC WAC 2018, 11/15/2018, Wellington, NZ
False Negatives by Archive

- York Univ: 322037
- UK Parliament: 1822
- UK Nat Arc: 1
- Stanford Arc: 8604
- PRONI: 1433
- perma.cc: 7560
- Nat Rec Scotland: 0
- Nat Lib Ireland: 5962
- LoC: 36336
- Icelandic Arc: 0
- Croatian Arc: 0
- Canadian Arc: 1036
- British Lib: 5639
- BibAlox.org: 42470
- Bayern Arc: 4061
- Arquivo.pt: 31400
- archive.is: 60351
- Archive-it: 295512

Count
False Positives by Number of Archives

Number of Archives:
- 9: 10487
- 8: 223549
- 7: 271074
- 6: 320979
- 5: 370283
- 4: 377663
- 3: 363014
- 2: 323139
- 1: 152751
- 0: 26703

Count:
- 0 to 50000
- 100000
- 150000
- 200000
- 250000
- 300000
- 350000
False Positives by Archive

<table>
<thead>
<tr>
<th>Archives</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>York Univ</td>
<td>0</td>
</tr>
<tr>
<td>UK Parliament</td>
<td>0</td>
</tr>
<tr>
<td>UK Nat Arc</td>
<td>158990</td>
</tr>
<tr>
<td>Stanford Arc</td>
<td>5224</td>
</tr>
<tr>
<td>PRONI</td>
<td>6976</td>
</tr>
<tr>
<td>perma.cc</td>
<td>397672</td>
</tr>
<tr>
<td>Nat Rec Scotland</td>
<td>0</td>
</tr>
<tr>
<td>Nat Lib Ireland</td>
<td>87365</td>
</tr>
<tr>
<td>LoC</td>
<td>2915</td>
</tr>
<tr>
<td>Icelandic Arc</td>
<td>0</td>
</tr>
<tr>
<td>Croatian Arc</td>
<td>201699</td>
</tr>
<tr>
<td>Canadian Arc</td>
<td>10291</td>
</tr>
<tr>
<td>British Lib</td>
<td>365004</td>
</tr>
<tr>
<td>BibAlex.org</td>
<td>677999</td>
</tr>
<tr>
<td>Bayern Arc</td>
<td>7852</td>
</tr>
<tr>
<td>Arquivo.pt</td>
<td>246579</td>
</tr>
<tr>
<td>archive.is</td>
<td>376835</td>
</tr>
<tr>
<td>Archive-it</td>
<td>22692</td>
</tr>
</tbody>
</table>

Smart Routing of Memento Requests
@mart1nkle1n
IIPC WAC 2018, 11/15/2018, Wellington, NZ
Changes in Archive Holdings

Change in Archive Holdings By Number of Archives

- Number of Archives
  - 214
  - 5171
  - 98611
  - 1439398

- Count
  - 0
  - 200000
  - 400000
  - 600000
  - 800000
  - 1000000
  - 1200000
  - 1400000

Smart Routing of Memento Requests
@mart1nkle1n
IIPC WAC 2018, 11/15/2018, Wellington, NZ
Archives Added

Addition in Archive Holdings by Archive Name

- Internet Arc: 44,129
- BibAlex.org: 30,884
- archive.is: 10,562
- Archive-It: 9,367
- British Lib: 1,608
- Canadian Arc: 2,103
- Icelandic Arc: 802
- LoC: 6,299
- Nat Rec Scotland: 779
- York Univ: 945
Archives Removed

Reduction in Archive Holdings by Archive Name

- York Univ: 15878
- UK Parliament: 2930
- perma.cc: 2432
- Nat Lib Ireland: 5410
- LoC: 2341
- Internet Arc: 27374
- Bayern Arc: 1102
- BibAlex.org: 27065
- Archive-it: 2542
- archive.is: 7997
Takeaways

- Memento Aggregator cache is very effective
  - Especially for human-driven services
- Machine learning process saves!
  - Requests & time while at acceptable recall level
  - FPR: 0.33 (std dev: 0.16)
- Re-training seems necessary, frequency TBD

Optimization

- ML model trained on archival holdings, not usage logs/cache
  - Beneficial for new archives
- Neural network classifier, based on simple URI features, show promising results
Smart Routing of Memento Requests

Martin Klein¹
Lyudmila Balakireva¹
Harihar Shankar¹
James Powell¹
Herbert Van de Sompel²

¹Research Library
Los Alamos National Laboratory

²Data Archiving and Networked Services
The Netherlands