Archival HTTP Redirection Retrieval Policies

Ahmed AlSum, Michael L. Nelson
Old Dominion University
Norfolk VA, USA
{aalsum,mln}@cs.odu.edu

Robert Sanderson, Herbert Van de Sompel
Los Alamos National Laboratory
Los Alamos NM, USA
{rsanderson,herbertv}@lanl.gov

Temporal Web Analytics Workshop 2013, Rio De Janeiro
Agenda

- Introduction
- Abstract Model
- Experiment And Results
- Retrieval Policies
Memento Terminology


URI-T, TM  TimeMap

```
1 <http://mementoproy.cs.oue.edu/aggr/timemap/link/http://www.amazon.com>; rel="self"; type="application/link-format",
2 <http://mementoproy.cs.oue.edu/aggr/timegate/http://www.amazon.com>; rel="timegate",
3 <http://www.amazon.com>; rel="original",
```
Live Redirect


% curl -I http://bit.ly/r9kIfC
HTTP/1.1 301 Moved
....
Location: http://www.cs.odu.edu/
...
Live Redirect

Archived Redirect

$R_1$

www.draculathemusical.co.uk

redirects

www.dracula-uk.com/index.html

http://api.wayback.archive.org/memento/20020212194020/http://www.draculathemusical.co.uk/

Archived redirects

Abstract Model
Abstract Model

- TimeMap for $R$

\[ TM(R) = \{M_1, M_2, ... M_n\}; \text{where } M_i = M(R) \text{ at } t_i \]
URI Stability

• URI’s stability is a count of the change in HTTP responses across time (200, 3xx, or 4xx) and the number of different URIs in the “Location” for 3xx status code.

High Stability = 1

No Stability = 0

\[
\text{Stability}(R) = 1 - \frac{\sum_{M \in TM} \text{Change}(M_i, M_{i-1})}{|TM|}
\]

\[
\text{Change}(M_i, M_{i-1}) = \begin{cases} 
1 & \text{Status}(M_i) \neq \text{Status}(M_{i-1}) \\
0 & \text{or Location}(M_i) \neq \text{Location}(M_{i-1}) \\
& \text{Otherwise}
\end{cases}
\]
Timemap Redirection Categories

- **Category 1**

All Mementos have **200 HTTP status code**
Timemap Redirection Categories

- Category 2

All Mementos have redirection to the same URI.
Timemap Redirection Categories

- Category 3

All Mementos have redirection to different URIs.
Timemap Redirection Categories

- Category 4

Mementos have different HTTP status code.
Timemap Redirection Categories

All Mementos have 200 HTTP status code

Stability = 1

All Mementos have redirection to the same URI.

Stability ≈ 0

All Mementos have redirection to different URIs.

Mementos have different HTTP status code.

Stability lim t→∞ = ?
URI Reliability

\[
R_{\text{URI}} = \frac{\# \text{Mementos end 200}}{|TM|}
\]

[Diagram showing the relationship between different nodes labeled with mementos and responses (200, 404, 3xx), with arrows indicating the movement through time stamps (TM). The diagram illustrates the concept of URI reliability with a formula that calculates the reliability based on the number of mementos ending with 200 and the total number of time stamps.]
HTTP Redirection Relationship between URI-R & URI-M

<table>
<thead>
<tr>
<th>Web Archive URI-M</th>
<th>Live Web URI – R</th>
</tr>
</thead>
<tbody>
<tr>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>Case 1</td>
<td>Case 1</td>
</tr>
<tr>
<td>5</td>
<td>3,4</td>
</tr>
<tr>
<td>OK</td>
<td>Redirection</td>
</tr>
<tr>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>Redirection</td>
<td>2</td>
</tr>
</tbody>
</table>

Case 1

Case 2

Case 3

Case 4

Case 5
Experiment & Results
## Experiment

- **Dataset:** 10,000 sample URIs from [Dmoz Open Directory Project](http://dmoz.org).
- **Dataset doesn’t have bit.ly nor doi.**
- **Experiment focused on the root page (no embedded resources)**

### URIs Live HTTP status code

<table>
<thead>
<tr>
<th>HTTP Status/Code</th>
<th>(10,000 URI-R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OK (200)</td>
<td>82.83%</td>
</tr>
<tr>
<td>Redirection (3xx)</td>
<td>14.71%</td>
</tr>
<tr>
<td>Redirection (301)</td>
<td>8.4%</td>
</tr>
<tr>
<td>Redirection (302)</td>
<td>6.1%</td>
</tr>
<tr>
<td>Redirection (others)</td>
<td>0.2%</td>
</tr>
<tr>
<td>Not-Found (4xx)</td>
<td>1.18%</td>
</tr>
<tr>
<td>Others</td>
<td>1.28%</td>
</tr>
</tbody>
</table>

### Memento HTTP status code

<table>
<thead>
<tr>
<th>HTTP Status/Code</th>
<th>(894,717 URI-M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OK (200)</td>
<td>93.46%</td>
</tr>
<tr>
<td>Redirection (3xx)</td>
<td>5.69%</td>
</tr>
<tr>
<td>Not-Found (4xx)</td>
<td>0.26%</td>
</tr>
<tr>
<td>Others</td>
<td>0.59%</td>
</tr>
</tbody>
</table>
Relationship between $TM(R)$ and $TM(\overline{R})$

<table>
<thead>
<tr>
<th>URI – $\overline{R}$ TimeMap Cases</th>
<th>#TM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>251</td>
<td>3</td>
</tr>
<tr>
<td>161</td>
<td>4</td>
</tr>
<tr>
<td>80</td>
<td>5</td>
</tr>
<tr>
<td>366</td>
<td>6</td>
</tr>
<tr>
<td>98</td>
<td>7</td>
</tr>
</tbody>
</table>

- Time span
- Number of Mementos

19%

65%
URI Stability

<table>
<thead>
<tr>
<th>TimeMap Category</th>
<th>Percentage</th>
<th>Stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Mementos have OK</td>
<td>52%</td>
<td>1</td>
</tr>
<tr>
<td>Mementos have mix status code</td>
<td>36%</td>
<td>0.91</td>
</tr>
<tr>
<td>All Mementos have Redirection</td>
<td>0.92%</td>
<td>0.85</td>
</tr>
<tr>
<td>Redirection to the same URI</td>
<td>0.62%</td>
<td></td>
</tr>
<tr>
<td>Redirection to different URIs</td>
<td>0.30%</td>
<td></td>
</tr>
<tr>
<td>URI has no Mementos at all</td>
<td>10.97%</td>
<td>0</td>
</tr>
</tbody>
</table>

Stability in semi-log scale

Stability for $|TM(R)| < 300$
URI Reliability

- 23% of the mementos did not lead to a successful memento at the end.
HTTP Redirection Relationship between URI-R & URI-M

<table>
<thead>
<tr>
<th>Web Archive URI-M</th>
<th>Live Web URI – R</th>
</tr>
</thead>
<tbody>
<tr>
<td>OK</td>
<td>OK Case 1 5</td>
</tr>
<tr>
<td>Redirection</td>
<td>2 3,4</td>
</tr>
</tbody>
</table>

Case 1 80.8%
Case 2 2.74%
Case 3 1.34%
Case 4 1.33%
Case 5 13.7%
RETRIEVAL POLICIES

ARCHIVED HTTP REDIRECTION RETRIEVAL POLICIES
Current Wayback Machine Policy

• **Live Redirect**: Wayback Machine ignores the live redirects. Use $R$ instead of $\bar{R}$.

• **Archived Redirect**: Wayback Machine follows the redirection.
Policy one: URI-R with HTTP redirection

- **Scope**: Selection between $R \rightarrow \overline{R}$ on the live web.
- **Algorithm**:

  1. Retrieve the memento $M$ for $R$.
  2. If `Status(M) = 200`:
     - **Yes**: Stop
  3. If `Status(M) = 3xx`:
     - **Yes**: Go to Policy 2
     - **No**: Stop
  4. If `Status(M) = 4xx` & $R$ has $\overline{R}$:
     - **Yes**: Use $\overline{R}$ instead of $R$
     - **No**: Stop


Policy one: URI-R with HTTP redirection

- **Evaluation:**
  - Policy scope has: 1471 URIs (that have live redirection)
  - 77 out of 1471 have no mementos at all
  - 17 out of 77 have been retrieved mementos based on live redirection

- **Implementation**

<table>
<thead>
<tr>
<th>Tool</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA Wayback Machine</td>
<td>For bit.ly URIs only</td>
</tr>
<tr>
<td>MementoFox</td>
<td>v 0.9.6+</td>
</tr>
<tr>
<td>mcurl</td>
<td>v 1.0</td>
</tr>
</tbody>
</table>
Policy two: URI-M with HTTP redirection

- **Scope:** Selection between $M \rightarrow \overline{M}$ in web archive.
- **Example:**

- **Algorithm:**

```
M → M̅
```

```
Extract original from M̅
http://www.cnn.com/
```

```
Repeat content-negotiation in datetime for original(M̅)
Accept-Datetime: Sun, 13 May 2006
http://www.cnn.com/
```
Policy two: URI-M with HTTP redirection

- Evaluation:
  - Policy scope: 2980 TimeMap (that showed HTTP redirection status code in at least one memento)
  - Success criteria: Using policy two contributed to the original TimeMap
  - Success percentage: 58% of the cases
Conclusion

• Quantitative study with 10,000 URIs.
• 48% were not fully stable through time.
• 27% were not perfectly reliable through time.
• New archival retrieval policy:
  - Policy one: successfully retrieved mementos for 17 out of 77
  - Policy two: Expanded the timemap for 58% of cases.

aalsum@cs.odu.edu
@aalsum