EVALUATING THE TEMPORAL COHERENCE OF ARCHIVED PAGES

SCOTT G. AINSWORTH
MICHAEL L. NELSON
OLD DOMINION UNIVERSITY

HERBERT VAN DE SOMPEL
LOS ALAMOS NATIONAL LABORATORY
HE WENT TO VIEW AN ARCHIVED PAGE. YOU WON’T BELIEVE WHAT HE SAW NEXT...

SCOTT G. AINSWORTH
MICHAEL L. NELSON
OLD DOMINION UNIVERSITY

HERBERT VAN DE SOMPEL
LOS ALAMOS NATIONAL LABORATORY
CONTENTS

- Motivation
- Composite Mementos
- Coherence Framework
- Temporal Coherence
- Future Work
- Conclusion
RESEARCH TO DATE

- How to crawl a site to maximize coherence
  - Ben Saad et al., JCDL 2011, TPDL 2011

- Detecting, visualizing temporal defects
  - Spaniol et al., WICOW 2009, IWAW 2009, VLDB 2009
RESEARCH TO DATE: @WEBSCIDL

- How much of the web is archived?
  - Ainsworth et al., JCDL 2011

- Are the archives stable?
  - Brunelle et al., JCDL 2013

- Temporal drift while browsing in an archive?
  - Ainsworth et al., JCDL 2013

- Are the missing resources important?
  - Brunelle et al., JCDL 2014

- Are the present resources correct?

(now 404, but that's a different story…)

4/27/15 Scott G. Ainsworth • Michael L. Nelson • Herbert Van de Sampel
NOT ALL 2004-12-09T19:09:26

-15 hours

Varina, Iowa
Local Time: 1:09 PM CST
Set My Timezone

Current Conditions
Updated: 12:55 PM CST on December 09, 2004
Observed at Storm Lake, Iowa (History)
Elevation: 1486 ft / 453 m

- Light Drizzle
- Wind: 6 mph / 8 km/h from
- Temperature: 41 °F / 5 °C
- Humidity: 100%

5-Day Forecast for ZIP Code 50593
Thu Fri Sat Sun Mon
- Chance of Rain: Mostly Clear
- Temperature: 43° | 29° 36° | 29° 31° | 15° 31° | 15°

+9 hours
CLEAR OR CLOUDY?

Lake, Iowa (History)  453 m
5 °C
3 °C
5 °C

5-Day Forecast
Fri

Chance of Rain
Mostly Cl

Alternate Computer Forecast: from Pollen.com

+9 months

Meet Locals

tripad
QUESTIONS

• How prevalent is temporal incoherence?
• Can temporal coherence be improved by using multiple archives?
• Can temporal coherence be improved by introducing memento selection heuristics?
CONTENTS

☑ Motivation
☐ Composite Memento
☐ Coherence Framework
☐ Temporal Coherence
☐ Future Work
☐ Conclusion
COMPOSITE MEMENTO

PRESENTATION

STRUCTURE
CONTENTS

☑ Motivation
☑ Composite Memento
☐ Coherence Framework
☐ Temporal Coherence
☐ Future Work
☐ Conclusion
COHERENCE STATES

• Prima Facie Coherent
  Evidence that the memento **existed** in its archived state when the root was acquired.

• Prima Facie Violative
  Evidence … **did not exist** …

• Possibly Coherent
  Evidence … **might have existed** …

• Probably Violative
  Evidence … **probably did not exist** …
CONSIDER THIS PAGE...

<html>
<img src="foo.jpeg">
</html>
WITH THESE RESPONSE HEADERS

HTTP/1.1 200 OK
Server: Tengine/2.0.3
Date: Mon, 27 Apr 2015 22:03:32 GMT
Content-Type: image/jpeg
Content-Length: 15632
Connection: keep-alive
Memento-Datetime: Tue, 07 Feb 2006 00:58:23 GMT
Link: <Memento links deleted...>
X-Archive-Orig-server: Apache/1.3.26 (Unix) ApacheJServ/1.1.2 PHP/4.3.4
X-Archive-Orig-etag: "4978-3d10-3e4d822e"
X-Archive-Orig-content-length: 15632
X-Archive-Orig-accept-ranges: bytes
X-Archive-Orig-date: Tue, 07 Feb 2006 00:58:20 GMT
X-Archive-Orig-content-type: image/jpeg
X-Archive-Orig-last-modified: Fri, 14 Feb 2003 23:56:30 GMT
X-Archive-Orig-connection: close
<other headers deleted>
PRIMA FACIE COHERENT

Bracket Pattern:
Memento-Datetime + Last-Modified

(Yes, Last-Modified is sometimes wrong, but many of those cases can be detected)
Equal Pattern: simultaneous capture
(with an optionally tunable “bubble of simultaneity”)

\[ T_0 = T_{i,1} \]
PRIMA FACIE VIOLATIVE

Closest memento created and acquired after the root was acquired
POSSIBLY COHERENT

Closest (or only) memento captured before the root
PROBABLY VIOLATIVE

Closest (or only) memento captured after the root but no Last-Modified (possibly indicating a dynamically generated representations) (for both PC & PV, you could do content comparison if there are 2 mementos that straddle the root page)
TEMPORAL COHERENCE

Old Dominion University

College of Sciences

Computer Science

Featured Student

Upcoming Events

Computer Science Colloquium, Friday May 14 at 10:00AM
Speaker: Dr. Allen Fisher from Carnegie Mellon [More]

CS Dept News

Jessica Crouch will be joining the computer science faculty as an
assistant professor in the 2015 fall semester.

Featured Alum

James S. Cheng
BS in CS
Class of 1982

About

Welcome from the Chair
Mission Statement
Student Goals
Bylaws
Organization
Facilities
FAQ
TEMPORAL COHERENCE

+9 days

+7 months

+18 days

+18 days

+2.1 years

2005-05-14 01:36:08

4/27/15 Scott G. Ainsworth • Michael L. Nelson • Herbert Van de Sampel
## Embedded Resources

<table>
<thead>
<tr>
<th>Resource</th>
<th>Memento-Datetime</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>mm_menu.js</code></td>
<td>2005-05-23 02:39:12</td>
<td>9.0 d</td>
</tr>
<tr>
<td><code>style.css</code></td>
<td>2005-05-23 02:39:39</td>
<td>9.0 d</td>
</tr>
<tr>
<td><code>ddmenu_ddown.js</code></td>
<td>2005-05-23 02:39:43</td>
<td>9.0 d</td>
</tr>
<tr>
<td><code>university.js</code></td>
<td>2005-05-23 02:39:56</td>
<td>9.0 d</td>
</tr>
<tr>
<td><code>rmenu_1st_about.png</code></td>
<td>2005-05-23 21:21:45</td>
<td>18.8 d</td>
</tr>
<tr>
<td><code>rmenu_1st_featured_alumni.png</code></td>
<td>2005-05-23 21:21:45</td>
<td>18.8 d</td>
</tr>
<tr>
<td><code>shadow-bl.gif</code></td>
<td>2005-05-23 14:07:29</td>
<td>18.5 d</td>
</tr>
<tr>
<td><code>rmenu_bottom_229.gif</code></td>
<td>2005-05-23 14:55:53</td>
<td>18.6 d</td>
</tr>
<tr>
<td><code>shadow-br.gif</code></td>
<td>2005-05-23 14:55:57</td>
<td>18.6 d</td>
</tr>
<tr>
<td><code>gfx-btn-go-dblue.gif</code></td>
<td>2005-05-24 02:36:07</td>
<td>2.1 years</td>
</tr>
<tr>
<td><code>shadow-tr.gif</code></td>
<td>2005-05-24 02:35:17</td>
<td>2.1 years</td>
</tr>
<tr>
<td><code>header-right1.gif</code></td>
<td>2005-05-24 02:35:17</td>
<td>2.1 years</td>
</tr>
</tbody>
</table>

### Embedded Resources

- Embedded Resources: 26
- Mean Delta: 125.9 days
- Standard Deviation: 207.7 days
- Minimum Delta: 9.0 days
- Maximum Delta: 2.1 years

---

4/27/15

Scott G. Ainsworth • Michael L. Nelson • Herbert Van de Sampel
REPRESENTING COHERENCE

Memento-Datetime

2006-11-10
2006-12-06
2007-01-01

-1d 0 1d

Mementos by Delta

Root
REPRESENTING COHERENCE

Memento-Datetime

2006-11-10
2006-12-06
2007-01-01

Mementos by Delta

-1d
0
1d

Prima Facie Coherent

Root

Last-Modified
REPRESENTING COHERENCE

[Diagram showing mementos by delta with timestamps and labels for prima facie coherent and probably coherent mementos.]

Prima Facie Coherent
Probably Coherent
Root
REPRESENTING COHERENCE

Mementos by Delta

-1d 0 1d

2006-11-10
2006-12-06
2007-01-01

Prima Facie Coherent
Probably Coherent
Probably Violative

Root-Modified
Last-Modified

23 days
3.6 days
REPRESENTING COHERENCE

Prima Facie Coherent  Prima Vacie Violative  Root
Probably Coherent  Probably Violative

Last-Modified

23 days
3.6 days
THE FULL CHART

Mementos by Delta

-3y  -1y  0  1y  2y  3y  4y  5y  6y


Prima Facie Coherent  Prima Vacie Violative  rURI-M
Probably Coherent  Probably Violative

2005-03-10
EXPERIMENT: DATA SET

• 4,000 sample URI-Rs (data set from JCDL 2011)
• Single and Multiple Archives
• Two Heuristics:
  • Minimum distance (current default Wayback behavior)
    • choose closest Memento-Datetime
  • Bracket (proposed here)
    • use combination of Memento-Datetime + Last-Modified

• Download all TimeMaps
• Download all root mementos
• Download all embedded resources
EXPERIMENT: SAMPLING

• For each root URI-R TimeMap, choose a single memento per month

• Extract embedded URI-Rs
• Download TimeMaps for embedded URI-Rs
• Download heuristically best URI-Ms

• Repeat recursively
### ROOT URI-R STATISTICS

#### Archival Data

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root URI-Rs archived</td>
<td>2,756</td>
<td>68.9%</td>
</tr>
<tr>
<td>In multiple archives</td>
<td>1,180</td>
<td>29.5%</td>
</tr>
<tr>
<td>Mean archives per URI-R</td>
<td>1.58</td>
<td></td>
</tr>
<tr>
<td>Mean mementos per URI-R</td>
<td>124.57</td>
<td></td>
</tr>
</tbody>
</table>

#### URI-M Status

<table>
<thead>
<tr>
<th>Status</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 OK</td>
<td>82,425</td>
<td>93.6%</td>
</tr>
<tr>
<td>503 Service Unavailable</td>
<td>4,444</td>
<td>5.0%</td>
</tr>
<tr>
<td>404 Not found</td>
<td>583</td>
<td>0.7%</td>
</tr>
<tr>
<td>403 Forbidden</td>
<td>388</td>
<td>0.4%</td>
</tr>
<tr>
<td>Others</td>
<td>214</td>
<td>0.3%</td>
</tr>
</tbody>
</table>
## EMBEDDED URI-R STATISTICS

### Archival Data

<table>
<thead>
<tr>
<th>Embedded URI-Rs</th>
<th>1,623,127</th>
</tr>
</thead>
<tbody>
<tr>
<td>per root URI-M</td>
<td>19.7</td>
</tr>
<tr>
<td>Embedded URI-Ms available</td>
<td>1,332,993 • 93.6%</td>
</tr>
<tr>
<td>per root URI-M</td>
<td>15.1</td>
</tr>
</tbody>
</table>

### URI-M Failure Reasons

<table>
<thead>
<tr>
<th>Reason</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not archived</td>
<td>312,641 • 83.9%</td>
</tr>
<tr>
<td>404 Not found</td>
<td>44,852 • 12.0%</td>
</tr>
<tr>
<td>403 Forbidden</td>
<td>6,116 • 1.6%</td>
</tr>
<tr>
<td>503 Service Unavailable</td>
<td>5,442 • 1.5%</td>
</tr>
<tr>
<td>Others</td>
<td>3,508 • 0.9%</td>
</tr>
</tbody>
</table>
## COMPOSITE MEMENTO (ROOT)
### COMPLETENESS & COHERENCE

### Completeness (and Missing)

<table>
<thead>
<tr>
<th>Description</th>
<th>MinDist Single</th>
<th>MinDist Multi</th>
<th>Bracket Single</th>
<th>Bracket Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Complete</td>
<td>76.1%</td>
<td>80.2%</td>
<td>76.2%</td>
<td>80.3%</td>
</tr>
<tr>
<td>Mean Missing</td>
<td>23.9%</td>
<td>19.8%</td>
<td>23.8%</td>
<td>19.7%</td>
</tr>
</tbody>
</table>

### Coherence

**Multiple archives: +completeness, -coherence?**

<table>
<thead>
<tr>
<th>Description</th>
<th>MinDist Single</th>
<th>MinDist Multi</th>
<th>Bracket Single</th>
<th>Bracket Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Prima Facie Coherent</td>
<td>41.0%</td>
<td>40.9%</td>
<td>54.7%</td>
<td>54.6%</td>
</tr>
<tr>
<td>Mean Possibly Coherent</td>
<td>27.3%</td>
<td>28.7%</td>
<td>12.8%</td>
<td>14.2%</td>
</tr>
<tr>
<td>Mean Probably Violative</td>
<td>2.5%</td>
<td>5.3%</td>
<td>2.5%</td>
<td>5.3%</td>
</tr>
<tr>
<td>Mean Prima Facie Violative</td>
<td>5.3%</td>
<td>5.3%</td>
<td>6.2%</td>
<td>6.2%</td>
</tr>
</tbody>
</table>

---

*At least 5% of pages can be shown to have temporal violations!*
## EMBEDDED MEMENTO COHERENCE

<table>
<thead>
<tr>
<th>Description</th>
<th>MinDist Single</th>
<th>MinDist Multi</th>
<th>Bracket Single</th>
<th>Bracket Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prima Facie Coherent</td>
<td>622,565</td>
<td>621,447</td>
<td>864,736</td>
<td>859,625</td>
</tr>
<tr>
<td>Possibly Coherent</td>
<td>497,405</td>
<td>466,046</td>
<td>244,104</td>
<td>215,585</td>
</tr>
<tr>
<td>Probably Violative</td>
<td>104,376</td>
<td>53,734</td>
<td>104,339</td>
<td>53,694</td>
</tr>
<tr>
<td>Prima Facie Violative</td>
<td>100,760</td>
<td>103,662</td>
<td>114,062</td>
<td>117,469</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>1,325,106</strong></td>
<td><strong>1,244,889</strong></td>
<td><strong>1,327,241</strong></td>
<td><strong>1,246,373</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
<th>MinDist Single</th>
<th>MinDist Multi</th>
<th>Bracket Single</th>
<th>Bracket Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prima Facie Coherent</td>
<td>47.0%</td>
<td>49.9%</td>
<td>65.2%</td>
<td>69.0%</td>
</tr>
<tr>
<td>Possibly Coherent</td>
<td>37.5%</td>
<td>37.4%</td>
<td>18.4%</td>
<td>17.3%</td>
</tr>
<tr>
<td>Probably Violative</td>
<td>7.9%</td>
<td>4.3%</td>
<td>7.9%</td>
<td>4.3%</td>
</tr>
<tr>
<td>Prima Facie Violative</td>
<td><strong>7.6%</strong></td>
<td><strong>8.3%</strong></td>
<td><strong>8.6%</strong></td>
<td><strong>9.4%</strong></td>
</tr>
</tbody>
</table>

**At least 7% of embedded resources are used violatively!**
CONTENTS

☑ Motivation
☑ Related work
☑ Preliminary work
☑ Temporal Coherence
☐ Future work
☐ Conclusion
MINOR OR MAJOR VIOLATIONS?

• This is a temporal violation. But is it meaningful?

• How to judge?
  • Most archives transform HTML
    • Not all archives support export of original file
  • How to measure similarity on binary files?
    • early results: very few cases of equivalent binaries
How to convey coherence & contributing archive?

How to scale to > 100 embedded mementos?
POLICIES & HEURISTICS

• Tradeoffs:
  • Fast: minimize distance
  • Accurate: maximize coherence
  • Complete: query all (not just top k) archives in order to maximize completeness
CONTENTS

☑ Motivation
☑ Composite Memento
☑ Coherence Framework
☑ Temporal Coherence
☑ Future Work
☐ Conclusion
CONCLUSION

• Defined four classes of temporal coherence for describing relationship between root & embedded mementos
  • Prima Facie {Coherent|Violative}
  • Possibly Coherent / Probably Violative

• Determine classes using a combination of HTTP metadata, primarily Memento-Datetime & Last-Modified

• At least 5% of IA pages have 1 or more temporal violations

• Using multiple archives increases completeness, but with a possible loss of coherence

• Determining semantic impact of violations and UI issues (status, policy choices) are areas of future research