Visualizing Digital Collections of Web Archives

Mat Kelly, Michael L. Nelson, Michele C. Weigle
Old Dominion University

Web Archiving Collaboration: New Tools and Models
Columbia University, New York, NY
June 4, 2015

http://ws-dl.cs.odu.edu
@machawk1
Motivation for Thumbnail Summarization

• Change over time - aboutness
Apple.com has > 17k mementos
Many Nearly Identical
(apple.com)
Methods of Summarization

• Including all mementos
  – many redundant thumbnails
  – temporally/spatially/cognitively expensive

• Naively excluding images
  – missing important captures in summary

• Compare image thumbnails
  – temporally expensive for identifying unique thumbnails

Comparing mementos’ markup can identify sufficiently unique mementos
Analyzing Markup

<title>Apple</title>
<meta property="analytics-track" content="Apple - Index/Tab" />
<meta property="analytics-s-channel" content="homepage" />
<meta property="analytics-s-bucket-0" content="appleglobal,applehome" />
<meta property="analytics-s-bucket-1" content="apple{COUNTRY_CODE}global,apple{COUNTRY_CODE}home" />

apple.com at Mar 17, 2008

HTML for memento

SimHash for HTML
SimHash?

HTML snippet for memento

First k characters of markup

Second k characters of markup

63rd k characters of markup

64th k characters of markup

Hash to a character

Hash to a character

Hash to a character

Hash to a character

\[ k = \frac{\text{markup length}}{64} \]
SimHash vs. Other Hashes

• $\text{md5(“aaaaaaaaaa aaaaaaaa”)}$
  $\Rightarrow 12f9cf6998d52dbe773b06f848bb3608$

• $\text{md5(“aaaaaaaaab aaaaaaaa”)}$
  $\Rightarrow e984cee68697eb77577717b532171493$

• $\text{simhash(“aaaaaaaaaa aaaaaaaa”)}$
  $\Rightarrow 8664ee964799c38c156d8f039dae8330$

• $\text{simhash(“aaaaaaaaab aaaaaaaa”)}$
  $\Rightarrow 8664ee964799a48c156d8f039dae8330$
Why SimHash?

• SimHash identifies similarities between documents

• Conventional hashing algorithms are for identifying differences
  – Drastically different output from similar content

• To remove redundancies, we want to detect when temporally adjacent mementos are sufficiently dissimilar
SimHashes for Mementos

HTML of apple.com
March 3, 2008

c39f0abc...b9

HTML of apple.com
March 5, 2008

c39d0abc...c9

HTML of apple.com
April 12, 2008

c39d0abc...b9

HTML of apple.com
October 4, 2008

c770ad1b...b9
# Identifying Similarity by Calculating Hamming Distance

<table>
<thead>
<tr>
<th>HTML of apple.com</th>
<th>March 3, 2008</th>
<th>c39f0abc...b9</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML of apple.com</td>
<td>March 5, 2008</td>
<td>c39d0abc...c9</td>
</tr>
<tr>
<td>HTML of apple.com</td>
<td>April 12, 2008</td>
<td>c39d0abc...b9</td>
</tr>
<tr>
<td>HTML of apple.com</td>
<td>October 4, 2008</td>
<td>c770ad1b...b9</td>
</tr>
</tbody>
</table>

**HAMMING DISTANCE**

- N/A pivot
- March 3, 2008 to March 5, 2008: 1
- March 5, 2008 to April 12, 2008: 2
- October 4, 2008: 7
CRITERIA FOR INCLUSION IN SUMMARY

Temporally Sorted Mementos

Hamming Distance

Threshold

In Summarization
Sliding Hamming Distance

- Selection based on previously selected memento
- Sliding pivot

![Diagram showing criteria for inclusion in summary with Hamming Distance and threshold lines. Points labeled ΔM0, ΔM3, and ΔM6 indicate mementos selected for summarization.]
Project Goals

Develop tools that implement thumbnail summarization for TimeMaps

• **Web Service**
  – Allows anyone to view TimeMap using thumbnail summarization

• **Wayback add-on**
  – Allows any archive using wayback to provide this service to users

• **Embeddable version**
  – Allow web page authors to embed overview of past versions of page on live web page
AlSummarization

• SimHash-based summarization scheme created by Ahmed AlSum
• AlSum + Summarization = AlSummarization

Dr. Nelson’s Homepage

• URI-R: http://www.cs.odu.edu/~mln
• Append onto service URI for summary
Anatomy of the Visualization

Temporally sorted mementos

Memento metadata

Thumbnails for http://www.cs.odu.edu/~mln

3 presentations of the Summary
Additional (optional) Endpoint Parameters

- **Access** – tailors user interface
  - Interactive, Embed, Wayback

- **Strategy** – to use alternative summarization
  - alSummarization, yearly, skipListed, random

- `http://service/?`
  - access=wayback&URI-R=http://www.cs.odu.edu/~mln
  - access=wayback&strategy=random&URI-R=http://www.cs.odu.edu/~mln
Programmatic Flow

User’s Browser

Thumbnails Service

Memento-Compliant Archive
User Requests URI-R Summary

User’s Browser ➔ Thumbnails Service ➔ Memento-Compliant Archive
Service Relays URI-R to Archive

User’s Browser

Service queries archive for all mementos for URI-R
URI-Ms returned to Service

Archive returns TimeMap with URI-Ms to thumbnail service
Service fetches HTML for each Memento
Service generates SimHash for Each Mementos’ HTML

c39f0abc...b9
c39d0abc...c9
c39d0abc...b9
c770ad1b...b9
c770ad1b...b9

Thumbnails Service
Service Calculates Hamming Distance

Mementos in summary selected based on hamming distance
Preliminary UI returned to user

User’s Browser

HTML interface

Thumbnails Service

Templated HTML interface is returned to user with placeholders for thumbnails
Service Generates Thumbnails for Mementos in Summary

Thumbnails Service

Hd()

c39f0abc...b9

2 c39d0abc...c9

1 c39d0abc...b9

7 c770ad1b...b9

0 c770ad1b...b9

PhantomJS
Thumbnails Served to User

Asynchronous polling from HTML pages populates placeholder images once available
Core Implementation

- `node` for thumbnail generation
- `PhantomJS` abstractions preserved for code reuse and extensibility
- Code documented to facilitate extensibility, usage, and fixes

http://github.com/machawk1/ArchiveThumbnails
Initializing the service

User/Service Administrator simply enters:

$ npm install
$ node alSummarization.js

Service responds and is ready for query:

* Local resource (css, js, etc.) server listening on Port 1338...
* Thumbnails service started on Port 15421
Online vs. Offline Generation

• Online Thumbnail Summarization
  – Fetch each mementos’ HTML
  – Calculate SimHashes
  – Calculate Hamming Distance (HD)
  – Select Mementos That Pass HD threshold
  – Generate Thumbnails of Mementos

• Offline Thumbnail Summarization
  – All of the above performed a priori
  – Data potentially updated on access
Adaptive Strategies

- Very large TimeMaps are temporally expensive to generate
- Default behavior:
  
  ```python
  if(timeRequirement == tooLong):
      use(naiveStrategy)
  ```
- User can explicitly override behavior
Other Summarization Strategies

• Random Selection
  – $k$ mementos, uniform selection

• Interval
  – every $m^{th}$ memento, $m = \frac{n}{k}$

• Temporal Interval
  – One memento/year, reverse chronological monthly back-fill

• Temporally Uniform Trimming when $k > 15$
Grid View
AlSummarization vs Random

Dr. Nelson’s Homepage
Random Strategy

Dr. Nelson’s Homepage
AlSummarization Strategy
Grid View

AlSummarization vs Interval

Dr. Nelson’s Homepage
Interval Strategy

Dr. Nelson’s Homepage
AlSummarization Strategy
Grid View

AlSummarization vs Temporal Interval

Dr. Nelson’s Homepage
Temporal Interval Strategy

Dr. Nelson’s Homepage
AlSummarization Strategy
Asynchronous Polling
Server-side SimHash Caching
Four Summarization Strategies
OpenWayback Integration
Service Embedding


- or -

Visualizing Digital Collections of Web Archives

• Codebase:
  – github.com/machawk1/ArchiveThumbnails

• Service URI:
  – http://wsdl-docker.cs.odu.edu:15421
Live Demo