Accessing Web Archives

Web Science Course 2017

Helge Holzmann

05/16/2017

Helge Holzmann (holzmann@L3S.de)
Not today’s topic

http://blog.archive.org/2016/09/19/the-internet-archive-turns-20/

HOW TO BUILD AN ARCHIVE
20 years in the making
What is a Web archive?

- Web archives preserve our history as documented on the Web.
- Consists of all kinds of Web resources:
  - i.e., HTML, images, video, scripts, ...
- ... stored in big files in the standardized WARC format.
Obama Wins a Clear Victory, but Balance of Power Is Unchanged in Washington

Boehner Strikes Conciliatory Tone in Talk of Fiscal Cliff

By JONATHAN WEISMAN and JACKIE CALMES 38 minutes ago

The speaker of the House said Wednesday he is ready to accept a budget deal that raises revenues if it is linked to an overhaul of the tax code and entitlements.

- Boehner's Offer May Bring Sides to Table 5:03 PM ET

Debate Over Size of Federal Government Left Unresolved

By PETER BAKER 3:15 PM ET

After $5 billion, two down presidential primary days, four general election debates and more TV ads than anyone could watch, the two parties essentially fought to a standstill.

- Comments

Favorite

G.O.P. Factions Grapple Over Meaning of Loss

By MICHAEL COOPER 4:59 PM ET

There was no shortage of theories from inside and outside the party about where it fell short and what to do next.

- Comments

Day After Election, a Sharp Loss on Wall Street

By NELSON D. SCHWARTZ 4:50 PM ET

Stocks moved sharply lower in New York, with both major indices down 2.4 percent.

Marriage Gains Cheer Gay Rights Advocates

By ERICK SCHAD 1:30 PM ET

Supporters called votes in Maine and Maryland a turning point. Results in Washington State were still being tallied.

- Comments

MORE ELECTION NEWS
- Californians Back Taxes to Avoid Education Cuts
- Democrats Gain in Senate, Adding to Majority
- Republicans Shore Up Incumbents, Holding House
- Why The Times Was Slower in Calling Election
- Twitter Kills the Fall Whale, One Tweet at a Time

BIRTHDAY GIFT
THANK YOU GIFT
Web archives in Web Science

• Very valuable dataset to study the Web (and the offline world)
  • Access to very diverse knowledge from various disciplines (history, politics, ...)
  • The *whole* Web at your fingertips / processable snapshots
  • Adds a temporal dimension to the Web / captures dynamics

• Access is challenging, questions are among others...
  • How to search / find interesting resources?
  • How to explore temporal snapshots?
  • How to process unstructured / diverse WARCs more efficiently?
  • How to zoom out / study relations?
  • How to investigate underlying structures / hyperlink graphs?
ALEXANDRIA @ L3S

- 5 years ERC Advanced Grant of Prof. Nejdl
  - [www.ALEXANDRIA-project.eu](http://www.ALEXANDRIA-project.eu)
Access from different perspectives

• User centric
  • Search / temporal Information Retrieval
  • Direct access / replay archives

• Data centric
  • (W)ARC and CDX (metadata) datasets
  • Big data processing: Hadoop, Spark, ...
  • Content analysis, historical / evolution studies

• Graph centric
  • Structural view on the dataset
  • Graph algorithms / graph analysis
  • Hyperlink and host graphs, entity / social networks and more
Access from different perspectives

• User centric
  • Search / temporal Information Retrieval
  • Direct access / replay archives

• Data centric
  • (W)ARC and CDX (metadata) datasets
  • Big data processing: Hadoop, Spark, ...
  • Content analysis, historical / evolution studies

• Graph centric
  • Structural view on the dataset
  • Graph algorithms / graph analysis
  • Hyperlink and host graphs, entity / social networks and more
Challenges for IR on Web Archives

• Documents are temporal / consisting of multiple versions
  • Version / snapshot / capture represents are crawl
    • A version may be a duplicate of a previous one
    • Or it may contain slight or drastic changes (might be a completely new page)

• Temporal relevance in addition to textual relevance
  • Temporal relevance is not always encoded in the content
    • Very little changes may be of high importance

• Resource identifiers (i.e., URLs) may change over time
  • A webpage moved to a new URL makes it hard to detect previous versions

• Information needs / query intents are different from traditional IR
  • There is no clear understanding of what is (temporally) relevant
Temporal Archive Search (Tempas)

• **Goal**: find entry points / authority pages in a given time period
  • most *central* URLs of an entity / topic over time
• **Idea**: exploit external information to detect temporal relevance
  • as it is difficult to derive from the documents / contents alone
  • capture temporally relevant keywords / descriptors from external data
• **v1**: based on tags from *Delicious* (tempas.L3S.de/v1)
  • uses temporal frequencies of social bookmarks as proxy for temp. importance
  • biased by Delicious users, only limited available data for 8 years
• **v2**: based on the hyperlink graph of the Web (tempas.L3S.de/v2)
  • uses temp. freq. of emerging in-links to a page as proxy for temp. importance
  • less biased, more data, growing with the Web archive
Tempas v1 (tempas.L3S.de/v1)

Emerging links in $[ta, tb]$:

$$L_{\text{emergence}} = \bigcup_{p \in \mathcal{P}} c \in \{c_t \in p | t_a \leq t \leq t_b\} \setminus \bigcup_{p \in \mathcal{P}} c \in \{c_t \in p | t < t_a\}$$

Relevance of URL $v$ w.r.t. anchor text $a$, based on $freq(v,a)$:

$$freq(v,a) = |\{\text{host}(u) | e = (u,v) \in E \land (e,a) \in L\}|$$

<table>
<thead>
<tr>
<th>Barack Obama</th>
<th>Angela Merkel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>obama</strong> @ [2005, 2006]</td>
<td><strong>merkel</strong> @ [2000, 2004]</td>
</tr>
<tr>
<td><strong>obama</strong> @ [2005, 2007]</td>
<td><strong>angela merkel</strong> @ [2000, 2004]</td>
</tr>
<tr>
<td><strong>obama</strong> @ [2008, 2013]</td>
<td><strong>merkel</strong> @ [2005, 2010]</td>
</tr>
<tr>
<td></td>
<td>1. <a href="http://angela-merkel.de">http://angela-merkel.de</a></td>
</tr>
</tbody>
</table>
Tempas v2 Example Queries (2)

- European Union

  european union @ [1996, 2005]

  european union @ [2005, 2013]

- Wikipedia

  wikipedia @ [2001, 2002]

  wikipedia @ [2003, 2013]

- Creative Commons License

  creative commons license @ [2002, 2003]
  1. http://creativecommons.org/licenses/by-nc-sa/1.0
  2. http://creativecommons.org/licenses/by-nd-nc/1.0

  creative commons license @ [2004, 2006]
  1. http://creativecommons.org/licenses/by-nc-sa/2.0
  2. http://creativecommons.org/licenses/by-nc-nd/2.0

  creative commons license @ [2007, 2013]
  1. http://creativecommons.org/licenses/by/2.5
  2. http://creativecommons.org/licenses/by/3.0
  3. http://creativecommons.org/licenses/by-nc-sa/3.0
Direct Access to Web Archives

• Most common: Wayback Machine
  • Access through URL / timestamp

• Alternative idea: Link temporal references on the live Web
  • here: Link scientific publications to webpages of mentioned software
    • use publication date for temporal links
    • websites represent the corresponding software (version) / serve as surrogates
Selected Related Works

- Marijn Koolen and Jaap Kamps. **The Importance of Anchor Text for Ad Hoc Search Revisited.** ACM Conference on Research and Development in Information Retrieval (SIGIR), 2010

- Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. **Index Maintenance for Time-Travel Text Search.** ACM Conference on Research and Development in Information Retrieval (SIGIR), 2012

- Liudmila Ostroumova Prokhorenkova et al. **Publication Date Prediction through Reverse Engineering of the Web.** ACM International Conference on Web Search and Data Mining (WSDM), 2016

- Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. **Graphs over time: densification laws, shrinking diameters and possible explanations.** ACM International Conference on Knowledge Discovery in Data Mining (KDD), 2005
Access from different perspectives

- **User centric**
  - Search / temporal Information Retrieval
  - Direct access / replay

- **Data centric**
  - (W)ARC and CDX (metadata) datasets
  - Big data processing: Hadoop, Spark, ...
  - Content analysis, historical / evolution studies

- **Graph centric**
  - Structural view on the dataset
  - Graph algorithms / graph analysis
  - Hyperlink and host graphs, entity / social networks and more
German Web Analysis

• The Dawn of Today’s Popular Domains
  • A Study of the Archived German Web over 18 Years

• Analysis purely based on metadata (CDX)

• Emergence of today’s top domains:

  • Intriguing findings
    • Domains grow exponentially, doubling their volume every two years
    • Tomorrow’s newborn URLs will be greater than today’s

German Web Analysis: Age Predictions

- URL age evolution
  - Linear fit with asymptotic error of 8.41%

→ 2020: Average URL age is 23 month (~double the age of 2005)
→ 2038: The Web turns 3 years old (on average)
German Web Analysis: Volume Predictions

• Domain volume evolution
  • Exponential fit with an asymptotic error of 2.07%

→ 2020: ~6 times the number of URLs per domain as in 2014
German Web Analysis: Size Predictions

• URL birth size evolution
  • Linear fit with an asymptotic error of 6.9%

→ 2038: newborn URL will have double the size as of today
Challenge: Web Archives are Big Data

- Processing requires computing clusters
  - i.e., Hadoop, YARN, Spark, ...

- MapReduce or variants
  - Homogeneous data formats
  - Load, transform, aggregate, write
  - Details: https://github.com/helgeho/MapReduceLecture

- Web archive data is heterogeneous, may include text, video, images, ...
  - Common header / metadata format, but various / diverse payloads
  - Requires cleaning, filtering, selection, extraction and finally, processing

Source: Yahoo!
Efficient Processing with ArchiveSpark

• Seamless **two step loading** approach:
  • Filter as much as possible on metadata before touching the archive
  • Enrich records with data from payload instead of mapping / transforming

Example Scenario

• Political scientist wants to analyze sentiments and reactions on the Web from a previous election cycle.

• Five decisions to be taken to narrow down the dataset:
  1. Filter temporally to focus on the election period
  2. Select text documents by filtering on MIME type
  3. Only keep online captures with HTTP status code 200
  4. Choose a captured version, for instance the latest of each page
  5. Look for political signal terms in the content to get rid of unrelated pages

• Finally, extract relevant terms / snippets to analyze sentiments
  • Document lineage, e.g., the title might have more value than the body text
Benchmarks

• Three scenarios, from basic to more sophisticated:
  a) Select one particular URL
  b) Select all pages (MIME type text/html) under a specific domain
  c) Select the latest successful capture (HTTP status 200) in a specific month

• Benchmarks do not include derivations
  • Those are applied on top of all three methods and involve third-party libraries
ArchiveSpark

• **Expressive** and **efficient** Web archives data access / processing
• Joint work with the **Internet Archive**
• **Open source**
  • Fork us on **GitHub**: [https://github.com/helgeho/ArchiveSpark](https://github.com/helgeho/ArchiveSpark)
  • Star, contribute, fix, spread, **get involved!**
  • Easily **extensible**
• **More details in:**
• Marijn Koolen and Jaap Kamps. The Importance of Anchor Text for Ad Hoc Search Revisited. ACM Conference on Research and Development in Information Retrieval (SIGIR), 2010

• Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. Index Maintenance for Time-Travel Text Search. ACM Conference on Research and Development in Information Retrieval (SIGIR), 2012

• Liudmila Ostroumova Prokhorenkova et al. Publication Date Prediction through Reverse Engineering of the Web. ACM International Conference on Web Search and Data Mining (WSDM), 2016

• Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. ACM International Conference on Knowledge Discovery in Data Mining (KDD), 2005
Access from different perspectives

- User centric
  - Search / temporal Information Retrieval
  - Direct access / replay

- Data centric
  - (W)ARC and CDX (metadata) datasets
  - Big data processing: Hadoop, Spark, ...
  - Content analysis, historical / evolution studies

- Graph centric
  - Structural view on the dataset
  - Graph algorithms / graph analysis
  - Hyperlink and host graphs, entity / social networks and more
Graph Challenges of Web Archives

• Different ways to construct / extract (temporal) graphs
  • URLs vs hosts vs ‘temporal merge’ vs snapshots [see Emergence (Tempas v2)]
• Web archives attempt to capture the Web / a subset of the Web
  • However, a Web archive is never complete, graph structures may be broken
Early work: Hyperlink Graph Analysis

- How complete are Web archives / crawls?
  - here: .de 2010 inter-domain out-links vs. availability in .de / Web archive

  ![Graph showing the relationship between number of pages and number of edges.]

- **Question**: How does this impact graph algorithms, such as PageRank?
Selected Related Works

- Marijn Koolen and Jaap Kamps. The Importance of Anchor Text for Ad Hoc Search Revisited. ACM Conference on Research and Development in Information Retrieval (SIGIR), 2010
- Liudmila Ostroumovova Prokhorenkova et al. Publication Date Prediction through Reverse Engineering of the Web. ACM International Conference on Web Search and Data Mining (WSDM), 2016
- Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. ACM International Conference on Knowledge Discovery in Data Mining (KDD), 2005
Synergies: Integrated Data Analysis

**Example**: Compare restaurant menu prices before and after the Euro

Thank you!

- www.L3S.de
- www.ALEXANDRIA-project.eu
- tempas.L3S.de
- github.org/helgeho/ArchiveSpark

Questions?