Accessing Web Archives from Different Perspectives with Potential Synergies

RESAW/IIPC, London 2017

Helge Holzmann, Thomas Risse

06/15/2017

Helge Holzmann (holzmann@L3S.de)
• 5 years ERC Advanced Grant of Prof. Wolfgang Nejdl  
• www.ALEXANDRIA-project.eu
Not today’s topic
Access from different perspectives

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

- **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
Zooming in Web Archives

Graph-centric view | Data-centric view | User-centric view

[Diagram showing different views]
Access from different perspectives

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

• 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
The Wayback Machine

• Replays Web resources with a temporal dimension
  • Identified by URL and timestamp (crawl time)


http://web.archive.org/web/TIMESTAMP/URL

• Challenges for the user
  1. Find the relevant timestamp
     • At what date / time was the webpage / content of interest online / crawled?
  2. Discover the desired resources
     • What is the URL of the webpage / content of interest at the relevant date / time?
Approach 1: Links from the Web

• Temporal references on the current / live Web

• Semantics of temporal links

  1. webpage@time, e.g., citation at time of visit
  2. entity@event, e.g., president at election

• Examples:

  • Web citation on Wikipedia, specific URL at specific time
    • A news article cited in a Wikipedia article at the time when it was cited
  • Archived surrogates of software in scientific publication at publication time
    • Software websites represent the corresponding software very well
    • Archived sites of mentioned software help to comprehend experiments
Software on the Web

- Analysis based on the hyperlinks on mathematical software pages

Artifacts provided for highly referenced articles

~60% link to some sort of documentation

~30% provide source code
Tempas TimePortal

• Connecting swMATH.org and the Wayback Machine
Software as a First-Class Citizen

• Identified by **software** and **publication**

http://tempas.L3S.de/...?software=866&publication=01415032

• Focus on the software rather than its webpage

• Automatically augmented with software-specific links
  • here: **documentation**, **updates**, **artifacts**

• Meaningful captures rather than random crawl times
Approach 2: Temporal IR in 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 text snippets or 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
Explore more than 284 billion web pages saved over time.


Summary of europa.eu.int

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.
Temporal Archive Search (Tempas.L3S.de)

- **Goal:** find URLs / entry points / authority pages over time
  - most *central* URLs of an entity / topic in a given 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)

[Helge Holzmann, Avishek Anand - “Tempas: Temporal Archive Search Based on Tags”. WWW 2016]
[Helge Holzmann, Wolfgang Nejdl, Avishek Anand - “On the Applicability of Delicious for Temporal Search on Web Archives”. SIGIR 2016]
Emerging links in $[t_a, t_b]$:

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

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

$$freq(v,a) = |\{\text{host}(u) \mid e = (u,v) \in E \land (e,a) \in L\}|$$
Tempas v2 Example Queries (1)

• Barack Obama

  obama @ [2005, 2006]

  obama @ [2005, 2007]

  obama @ [2008, 2013]
  1. http://barackobama.com

• Angela Merkel

  merkel @ [2000, 2004]
  1. http://merkel.de (university bookstore Merkel)

  angela merkel @ [2000, 2004]
  1. http://angela-merkel.de

  merkel @ [2005, 2010]
  1. http://angela-merkel.de

  merkel @ [2010, 2013]
  1. http://angela-merkel.de
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
User View Synergies

• Graph view to identify relevant Web archives
  • Temporal in-links as indicator of relevance
  • **Example**: Software in literature vs. in-links
    • Analysis based on TLD .de (provided by IA)

• Starting point to zoom out to data view
  • Search results as entry points / dataset for data analysis
  • **Future Work**
    • Integration of data analysis capabilities into exploration system, like Tempas
    • Zoom out from user perspective to data analysis
Access from different perspectives

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

- **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
Studying the Web: 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

[Helge Holzmann, Wolfgang Nejdl and Avishek Anand - “The Dawn of Today’s Popular Domains: A Study of the Archived German Web over 18 Years”. JCDL 2016]
• Domain volume evolution
  • Exponential fit with an asymptotic error of 2.07%

\[ f(x) = \text{total} \]

→ 2020: ~6 times the number of URLs per domain as in 2014
Big Data Analysis in Web Archives

• 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!
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
  • Star, contribute, fix, spread, get involved!
  • Easily extensible
• More details in:
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

[Helge Holzmann, Vinay Goel and Avishek Anand - “ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation”. JCDL 2016]
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
Data View Synergies

- Zoom out from user view to process data at scale
  - Search results as entry points / dataset for data analysis
- Graph view to identify entry points
  - Integration of anchor texts as first-level dataset into ArchiveSpark
  - Filter / select relevant records based on links, e.g., Tempas search results
Access from different perspectives

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

- **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
Graphs in Web Archives

• Different ways to construct / extract (temporal) graphs
  • URLs vs hosts vs ‘temporal merge’ vs snapshots [see *L_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
Ongoing Work: Hyperlink Graph Analysis

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

• **Question**: How does this impact graph algorithms, such as PageRank?
Synergies Among Views on Web Archives

Helge Holzmann (holzmann@L3S.de)

source: http://www.okclipart.com/blue-fish-clipart90plakdqvz/
Generic Web Archive Analysis Framework

Graph-centric  Data-centric  User-centric

graph analysis / search  data access  processing / presentation

manual exploration

results

zoom
Example Implementation: DM vs. € Study

- Study of restaurant price when € was introduced
- Steps to be performed
  1. Identify time / keywords of interest
     - restaurant / menu @ the introduction of € (2002)
  2. Find entry points for the study
     - URLs of restaurant and menu pages
  3. Locate suitable documents in the archive
     - WARC records of corresponding URLs
  4. Detect and extract desired information
     - DM and € prices from menus
  5. Aggregate statistics and present results
     - prices on average 23% higher

Example Implementation: DM vs. € Study

Conclusion and Future Work

• Different views on Web archives represent different zoom levels:
  • User view describes the perspective from a user on archived data
  • Data view zooms out to bigger collections, analysis at scale
  • Graph view focuses on relationships among objects / records in archive

• **Synergies** allow for systematic / effective / efficient data analysis

• More research required in **future work**:
  • Web archive graphs not well understood yet
    • What is the impact of incomplete of crawls on a page’s centrality?
    • How do different extraction / construction methods affect the graph properties?
Thank you!

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

Questions?