Web Archives in Web Science
Web Science Course 2018

Helge Holzmann

May 8, 2018

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

a. 25 Petabytes Archived
b. 435 Billion Web Captures
c. 4 Million eBooks
d. 2 Million Tweets

e. 3 Million Hours of TV

http://blog.archive.org/2016/09/19/the-internet-archive-turns-20/
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 efficiently?
  • How to tackle it from different views / zoom out / study relations?
  • How to investigate underlying structures / hyperlink graphs?

08/05/2018 Helge Holzmann (holzmann@L3S.de)
http://nytimes.com

Saved 49,342 times between November 12, 1996 and February 12, 2017.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.
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

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

By MICHAEL COOPER 4:50 PM ET

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

Day After Election, a Sharp Loss on Wall Street

By NELSON D. SCHWARTZ 3:20 PM ET

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

Debate Over Size of Federal Government Left Unresolved

By PETER BAKER 3:15 PM ET

After $6 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.

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
ALEXANDRIYA @ L3S

• 5 years ERC Advanced Grant of Prof. Nejdl
• www.ALEXANDRIYA-project.eu
Working with Web Archives

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

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

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

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

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

- **Graph-centric View**
  - 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?
Temporal Search 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
**Temporal Archive Search (Tempas)**

- **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
  - capture temporal keywords / descriptors from external data (surrogates)
- **v1**: based on *tags* from Delicious ([tempas.L3S.de/v1](http://tempas.L3S.de/v1))
  - uses temporal frequencies of social bookmarks as proxy for temporal 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](http://tempas.L3S.de/v2))
  - less biased, more data, growing with the Web archive
Tempas v1 (tempas.L3S.de/v1)

UI with support for temporal queries and temporal results

[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]
Tempas v2 (tempas.L3S.de/v2)

- From tags to anchor texts
  - Delicious is biased
  - Available data is limited
  - Tags do not allow for natural keyword search
- Anchor texts serve well for the task of temporal navigational queries
Temporal Retrieval Model

• Based on temporal hyperlink graphs with yearly granularity
  • only emerging links in time period \([t_a, t_b]\) are considered

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

• Relevance is determined by hosts linking to a URL \(v\) w.r.t. anchor text \(a\)

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

\[
\text{rel}(v, a) = \log\left(\frac{\text{freq}(v, a)}{\max_{v \in V, a \in A} \text{freq}(v, a)} \cdot \gamma\right)
\]
Index Construction and Retrieval

• Based on the **German Web archive** (TLD .de) from 1996 to 2013
  • > 2 billion distinct archived pages linking to 26,443,384,902 URLs (not only .de)
  • 319,574,156 URLs indexed after cleaning infrequent and malformed links

• Anchor texts classified by **relevance class per year**: $\varphi = relc(v, a) = \lceil rel(v, a) \rceil$
  • Weighted by maximum frequency in a year: $\text{boost}(\varphi) = \varphi \cdot \log(\max_{v \in V, a \in A} \text{freq}(v, a))$

• Indexed into **Elastic Search**
  • One full-text index per year and relevance class
  • Scored by tf-idf with boost
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
Alternative: Direct Links

• 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

~60% link to some sort of documentation

~30% provide source code

Artifacts provided for highly referenced articles
Temporal Software Links

- 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
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 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
Working with Web Archives

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

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

• Graph-centric View
  • 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

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

→ 2020: ~6 times the number of URLs per domain of 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 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
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
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]
Core Components of ArchiveSpark

- Data Specifications (**DataSpecs**) and Enrich Functions (**EnrichFuncs**)
Simple and Expressive Interface

• Based on **Spark**, powered by **Scala**
  • This does not mean you have to learn a new programming language!
  • The interface is rather **declarative** and writing scripts for ArchiveSpark does not require deep knowledge about Spark / Scala

• **Simple data accessors** are included
  • Provide simplified access to the underlying data model

• **Easy extraction / enrichment mechanisms**
  • Customizable and extensible by advanced users

```scala
val query = MhlSearchOptions(query = "polio", collections = MhlCollections.Statemedicalsocietyjournals)
val rdd = ArchiveSpark.load(MhlSearchSpec(query))
val enriched = rdd.enrich(Entities)
enriched.saveAsJson("enriched.json.gz")
```
Implicit Lineage Documentation

- Nested JSON output encodes lineage of applied enrichments

```json
"title": {
  "text": {
    "-_": "Libyan Revolution | Alexander Higgins Blog",
    "entities": {
      "persons": [
        "Alexander",
        "Higgins"
      ]
    }
  }
}
```
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:
  • *Helge Holzmann, Vinay Goel, Avishek Anand.*
    *ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation.*
    *In Proceedings of JCDL, Newark, New Jersey, USA, 2016.*
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
Working with Web Archives

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

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

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

08/05/2018

Helge Holzmann (holzmann@L3S.de)
Graph Challenges of Web Archives

- Different ways to construct / extract (temporal) graphs
  - URLs vs hosts vs 'temporal merge' vs snapshots
- Web archives attempt to capture the Web / a subset of the Web
  - However, a Web archive is never complete, graph structures may be broken
Challenge: Incompleteness

• **Hypothesis:** Incomplete graphs (e.g., crawls, Web archives, ...) cause deviations in random walk algorithms, such as PageRank

• **Consequence:** Rankings corresponding to PageRank differ from the (unavailable) complete / original graph

• **Example:**
  • Women@NASA (women.nasa.gov) is among top 300 .gov pages
  • A big majority of its PageRank comes from nasa.gov
  • Missing this link (e.g., nasa.gov down / removes link) would result in women.nasa.gov to fall into oblivion
  • By missing 50% of the graph, average displacement of 66,216 ranks among the originally top 30% (leading to Kendall’s Tau of 0.55)
Observations on Incomplete Crawls

- 3 Web archives + 1 social graph
  - Slightly different trends due to different structures (more or less scale-free)
- Kendall’s Tau measures correlation between ranking on complete graph vs. incomplete crawl
  - Block fraction describes of much the crawler missed
- HAK Measure is an attempt to estimate the deviation, purely on the crawl
  - work in progress, but seems possible
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 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
Synergies: Integrated Data Analysis

• Combining **user-**, **data-**, **graph-centric** views in data analysis

[H. Holzmann and T. Risse - “Accessing Web Archives from Different Perspectives with Potential Synergies”. Web Archiving Week (RESAW/IIPC) 2017. ]
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

Thank you!

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

Questions?