Creating Topical Collections: Web Archives vs. Live Web

Martin Klein
@mart1nkle1n

Research Library
Los Alamos National Laboratory
Background - Live Web

- Live web is dynamic, lives in a “perpetual now”
- Subject to link rot and content drift (= reference rot)*
- Significant platform/source for news publication/consumption

- Pew Research Center survey from August 2017:
  - 43% often get news online
  - 50% often get news from TV
  - 38% and 57% in early 2016

*See:
https://doi.org/10.1371/journal.pone.0115253
https://doi.org/10.1371/journal.pone.0167475
Background – Collection Building from the Live Web

- Often orchestrated by subject matter experts, archivists, special collection librarians, technicians
- Potentially with guidance from institutional collection policy
- Results in a list of seeds (URIs, social media accounts, etc)
- Utilization of crawling services such as Archive-It, Social Feed Manager

- Relevance of seeds assessed by humans
- Time passed since event is a concern because:
  - Stories evolve
  - Reference rot
  - API restrictions
Background – Archived Web

• Web archives are an invaluable resource for researchers, historians, journalists, etc.
• Often broad in scope, large in scale, covering different temporal intervals
• Makes discovery, access, and analysis difficult
  • In particular, for topic-specific resources
Access to the Archived Web

Memento allows to access many web archives, simultaneously!

http://timetravel.mementoweb.org/
http://mementoweb.org/guide/rfc/
<Intermezzo>
Web Crawling

Seed

Child 1
- Child 1.1
- Child 1.2

Child 2
- Child 2.1

Child 3
- Child 3.1
- Child 3.2
- Child 3.2
Web Crawling

Seed

Child 1
Child 1.1
Child 1.2

Child 2
Child 2.1

Child 3
Child 3.1
Child 3.2
Child 3.2
Web Crawling

Seed

Child 1
- Child 1.1
- Child 1.2

Child 2
- Child 2.1

Child 3
- Child 3.1
- Child 3.2
Web Crawling

Seed

Child 1
- Child 1.1
- Child 1.2

Child 2
- Child 2.1

Child 3
- Child 3.1
- Child 3.2

Focused Crawling

- Seed
  - Child 1
    - Child 1.1
    - Child 1.2
  - Child 2
    - Child 2.1
  - Child 3
    - Child 3.1
    - Child 3.2
    - Child 3.2

- Crawled and relevant
- Crawled and not relevant
- Not crawled
Focused Crawling

Seed

Child 1
  - Child 1.1
  - Child 1.2

Child 2
  - Child 2.1

Child 3
  - Child 3.1
  - Child 3.2

Crawled and relevant
Crawled and not relevant
Not crawled
Focused Crawling

- Seed
  - Child 1
    - Child 1.1
    - Child 1.2
  - Child 2
    - Child 2.1
  - Child 3
    - Child 3.1
    - Child 3.2
    - Child 3.3

- Crawled and relevant
- Crawled and not relevant
- Not crawled
</Intermezzo>
Inspiration from Previous Work

Extracting Event-Centric Document Collections from Large Scale Web Archives

Gerhard Gossen, Elena Demidova, and Thomas Risse
L3S Research Center, Leibniz Universität Hannover
{gossen, demidova, risse}@l3s.de

Abstract. Web archives are typically very broad in scope and extremely large in scale. This makes data analysis appear daunting, especially for non-computer scientists. These collections constitute an increasingly important source for researchers in the social sciences, the historical sciences and journalists interested in studying past events. However, there are currently no access methods that help users to efficiently access information, in particular about specific events, beyond the retrieval of individual disconnected documents. Therefore we propose a novel method to efficiently extract event-centric document collections from large scale Web archives. This method relies on a specialized focused extraction algorithm. Our experiments on the German Web archive (covering a time period of 19 years) demonstrate that our method enables efficient and effective extraction of event-centric collections for different event types.

https://doi.org/10.1007/978-3-319-67008-9_10
Previous Work - Setup

• Extract event-centric document collections via focused crawling of an archive
• Archive = web pages from .de top-level domain, captured by the Internet Archive until 2013 (30TB, 4b captures, 1b URIs)
• Identified 28 topics, likely covered in archive
• Text of topics’ Wikipedia page used for content relevance evaluation
• Crawled page datetime used for temporal relevance evaluation
• Overall relevance = content relevance + temporal relevance
• Wikipedia page outlinks used as seeds for focused crawl
Previous Work – Results

Costa Concordia grounding

![Graph showing relevance (accumulated) vs. downloaded URLs for Costa Concordia grounding]

German federal election 2009

![Graph showing relevance (accumulated) vs. downloaded URLs for German federal election 2009]
Our Questions

• Can we create high-quality topical collections by focused crawling online-available web archives?
• What is the effect of including multiple archives in the crawl?
• How do collections created from the archived web compare to those created from the live web?
• How does the amount of time passed since the event affect the quality of the collection?
Our Experiment

- Topics limited to terror attacks and mass shootings in the U.S.
- From different times in the past
- Focused crawl of:
  - 22 archives, simultaneously, via Memento infrastructure
  - the live web
- Take content and temporal relevance into account
  - Equally weighted: \( R = (0.5 \times CR) + (0.5 \times TR) \)
- Use events’ Wikipedia page as input for focused crawler
1. Content of Wikipedia page + random 60% of page’s references
   - Generate topic vector (TF-IDF of 1grams + 2grams)

2. Content of remaining 40% of Wikipedia page’s outlinks
   - Generate topic vector (TF-IDF of 1grams + 2grams)
   - Compute cosine similarity value between vectors 1 and 2
   - Run 10 times
   - Take average similarity value as content threshold
Temporal Relevance

- Define temporal interval for which crawled pages are considered relevant
- Event date extracted from Wikipedia event page
- Change point determined from graph of proportional Wikipedia page edits per day
Change Point Detection

![Graph showing change point detection with edit dates and percentage]
Datetime Extraction

- Extract datetime from pages via:
  - URI
  - Meta tags
    <meta property="article:published" itemprop="datePublished" content="2017-12-09T10:14:50-05:00" />
  - ODU’s Carbondate tool
    http://carbondate.cs.odu.edu/
  - Memento datetime
  - X-Header
Crawls

- Use version of Wikipedia page that was live at change point
- Possible crawl stop conditions:
  - Total number of documents crawled
  - Accumulated size of crawled documents
  - Time elapsed since crawl started
  - Crawl x levels deep
  - No more relevant documents left

- Our pick:
  - Crawl relevant documents
  - 5 levels deep
  - with priority queue
Collections Crawled (in late November)

• New York City, October 31\textsuperscript{st} 2017
• Las Vegas, October 1\textsuperscript{st} 2017

• Orlando, June 12\textsuperscript{th} 2016
• San Bernadino, December 2\textsuperscript{nd} 2015

• Tucson, January 8\textsuperscript{th} 2011
• Binghampton, April 3\textsuperscript{rd} 2009
NYC, 10/31/2017 – URIs per Level

Archived Crawl

Live Crawl

Levels

Levels

All URIs

Relevant URIs

All URIs

Relevant URIs
Intermezzo – Focused Crawling

- Seed
  - Child 1
    - Child 1.1
    - Child 1.2
  - Child 2
    - Child 2.1
  - Child 3
    - Child 3.1
    - Child 3.2
    - Child 3.3

Crawled and relevant
Crawled and not relevant
Not crawled

Creating Topical Collections: Web Archives vs. Live Web
@mart1nkle1n
CNI Fall Meeting 2017, 12/11/2017, Washington, DC
NYC, 10/31/2017 – Relevance over URIs

Relevant Documents

All Crawled Documents

Accumulated Relevance

Documents

Accumulated Relevance

Documents

Archived

Live

Archived

Live
NYC, 10/31/2017 – Relevance over Crawl Time

Relevant Documents

All Crawled Documents

Accumulated Relevance

Accumulated Relevance

Time in Seconds

Time in Seconds

Archived

Live

Archived

Live

0 50000 100000 150000 200000

0 100 200 300 400 500 600

0 500 1000 1500

0 50000 100000 150000 200000

0 50000 100000 150000 200000
NYC, 10/31/2017 – Web Archive Distribution

- web.archive.org
- wayback.archive-it.org
- archive.is
- perma-archives.org
- webarchive.nationalarchives.gov.uk

Legend:
- Blue: All Mementos
- Red: Relevant Mementos

Graph showing the distribution of mementos across various archival sites.
Creating Topical Collections: Web Archives vs. Live Web

Binghampton, April 3rd 2009 – URIs per Level

Archived Crawl

Live Crawl

Levels

All URIs

Relevant URIs

Levels

All URIs

Relevant URIs
Binghampton, April 3rd 2009 – Relevance over URIs

Relevant Documents

All Crawled Documents

Accumulated Relevance

Archived
Live

Documents

Accumulated Relevance

Archived
Live

Documents
Binghampton, April 3rd 2009 – Relevance over Crawl Time

Relevant Documents

All Crawled Documents
San Bernadino, December 2\textsuperscript{nd} 2015 – URIs per Level

**Archived Crawl**

- **All URIs**
- **Relevant URIs**

**Live Crawl**

- **All URIs**
- **Relevant URIs**
San Bernadino, December 2\textsuperscript{nd} 2015 – Relevance over URIs

Relevant Documents

![Graph of Relevant Documents]

All Crawled Documents

![Graph of All Crawled Documents]
San Bernadino, December 2\textsuperscript{nd} 2015 – Relevance over Crawl Time

Relevant Documents

All Crawled Documents

Accumulated Relevance vs. Time

Archived vs. Live

Accumulated Relevance vs. Time

Archived vs. Live
San Bernadino, December 2\textsuperscript{nd} 2015 – Web Archive Distribution

![Graph showing distribution of mementos and relevant mementos for various domains.](image-url)

- **All Mementos**
- **Relevant Mementos**

---

Creating Topical Collections: Web Archives vs. Live Web

@mart1nkle1n

CNI Fall Meeting 2017, 12/11/2017, Washington, DC
Take-Aways

• Web archives are good resources to build topical collections of web resources
• Utilizing multiple web archives is beneficial for the collection
• Crawling web archives is much slower than the live web
• Collections about very recent events benefit more from the live web than the archived web
  but
• Collections about events from the distant past benefit more from archives than the live web
  but
• Collections about less recent events can (still) benefit from the live web and (already) from the archived web
Where to go next

- Forgive one level of “irrelevance”
- Compare with manually curated collections (from AIT)
- Diversify to international topics and beyond shootings
- Investigate questions of optimal start and end time of crawls
"Two days after 26 people were massacred in a Texas church, the incident — one of the worst mass shootings in American history — had nearly vanished from the major cable news networks." @teamtrace thetrace.org/rounds/texas-c ...

https://web.archive.org/web/20171206181955/https://twitter.com/TVNewsArchive/status/938466726190096384
Creating Topical Collections: Web Archives vs. Live Web

Martin Klein
@mart1nkle1n

Research Library
Los Alamos National Laboratory