A survey of web archive search architectures

Miguel Costa, Daniel Gomes
(Portuguese Web Archive@FCCN)
Francisco Couto, Mário J. Silva
(University of Lisbon)
The Internet Archive was founded in 1996

Web-archived page of the first Web Archive
Web archiving has been growing

- 77 web archiving initiatives
- 282 billion web-archived files
Web archives must be searchable

• Users demand “Google-like” search
  –Searchable means at least full-text search

• Unsearchable=Useless
How to enable web archive search?

Our pursued answer since 2001...
Research on web archiving (2001)

- A web archive of online publications
- Project between the University of Lisbon and the National Library of Portugal

Publications are changing from the traditional formats, like paper magazines, to digital media, such as online news feeds. In addition everyone with a connected computer is now a potential publisher. This is increasing the number of new publications, making the management of deposits more complex. The archival of publications has a significant role in preserving the historical past. Publications in traditional media have been archived since ancient times. However, archiving publications on the Internet with techniques designed to allow their long term preservation is a non trivial task. The tools available for building digital publications weren't designed with preservation in mind and so don't meet most of the requirements involved.

Research Team:
- Mário Gaspar da Silva (coordinator)
- Ana Paula Afonso
- António Ferreira
- Daniel Gomes
- João Campos
- Norman Noronha

Digital Deposit
Research on search engines (2001)

- Portuguese-web search engine
Portuguese Web Archive = Web search + Web archiving

2008

Search and access pages of the past

See or rediscover pages that are no longer available online.

Search examples:
- Expo 98
- Euro 2004
- Elections in 2004
- University of Lisbon

US elections

6 January, 2007 - other dates

US Elections The Embassy Consular Service: Portugal's What direction for America after the mid-term elections have become clear. many people are speculating on what will happen.

http://www.american-embassy.pt/OpEd

US Embassy Lisbon, Portugal - Government & Other Links

18 April, 2004 - other dates

Primary and Caucus System in U.S. Elections, Foreign Press Center Briefing with Curtis Gans, Director of the... ASKED QUESTIONS FAQs about U.S. elections, FAQs About Voting System Standards, ELECTION PROCESS

U.S. Foreign Policy U.S. Government

http://www.american-embassy.pt/elections/04.html

Die Zeit - Politik: The US Elections: What Europeans expect...

23 January, 2006 - other dates

Die Zeit - Politik: The US Elections: What Europeans expect ZEIT.DE + POLITIK + us-wahl us-wahl The US Elections: What Europeans expect Von Constanze Steinmei”ler, Timothy Garton Ash has called it, the most important American election in living memory. Even to those of us who don't think that the... http://www.zeit.de/2004/4/europoeis_us_wahl
Survey about web archiving initiatives (2011)

- URL search: 89%
- Meta-data search: 79%
- Full-text search: 67% (28 initiatives)
  - The knowledge is out there
For this survey of web archive search architectures

- Identified prevalent search architectures
- Compared main features based on:
  - Available publications (still few)
  - Our experience
Portuguese Web Archive

• Based on NutchWAX
  – Archive-access tools are widely used to support search

• Full-text search over 1.2B docs at archive.pt
Search workflow

• For large collections, indexes must be **partitioned** across several machines
Time, document partitioning (PWA)

• Advantages
  – Selects *time partitions* according to query timespan
  – Progressive degradation
    • All computers have all terms
    • One partition fails, remaining respond to query term
  – No index rebuilding required to add new collections

• Disadvantages
  – High workload: all *document partitions* within timespan must be scanned for each query
  – Centralized data center approach

![Diagram of document partitioning with time intervals and document identifiers]
Everlast

- P2P architecture
- Different types of nodes
  - Crawlers
  - Version directories
  - Indexes
- Low cost nodes
- Full-text search
- “Unlimited” scalability
- Tested in laboratory
Term, time partitioning (Everlast)

- **Advantages**
  - Robustness of decentralized architecture
  - Lower workload: only one *term partition* is contacted for each query

- **Disadvantages**
  - Index updates to add new collections
  - *Term partition* unreachable may prevent response to query term
    - Redundancy required
  - Latency due to network
Wayback Machine (URL search)

- Doesn’t use inverted indexes
  - Flat sorted files of URLs
- URL partitioning
- Advantages
  - High throughput with millions of queries daily
  - Easy to manage: “no PhD required”
- Disadvantages
  - High communication workload because all queries are broadcasted to all index partitions
  - Limited search features
Overall comparison

<table>
<thead>
<tr>
<th>Search requirement</th>
<th>Wayback Machine</th>
<th>Portuguese Web Archive</th>
<th>Everlast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage and workload scalability</td>
<td>High</td>
<td>High</td>
<td>Very High</td>
</tr>
<tr>
<td>Service reliability</td>
<td>High</td>
<td>High</td>
<td>Medium</td>
</tr>
<tr>
<td>Time-aware indexing</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Performance of response times and</td>
<td>High</td>
<td>Very High</td>
<td>Medium</td>
</tr>
<tr>
<td>throughput</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- None is the best, just different.
- Our objective was to improve documentation about web archive search
Food-for-thought
Problem for **existing** web archives

• Users don’t know where to search for past web content
  —“Page unavailable” means lost forever

• Dissemination of web archive services is expensive
It would be nice to have a single portal for cross-web archive search but...

- Web-archived data is spread
- Search architectures are different
- Search technologies are different
- Interoperability is required
How to design a cross-web archive search architecture?
Our proposal: Web archive metasearch based on OpenSearch

- OpenSearch is a widely supported and simple technology
  - Most web archives use NutchWAX and it supports OpenSearch
- Portal would be simple and cheap to implement
  - Extremely useful to web users
  - Increase visibility of web archiving initiatives
- Easily combines live-web with past-web search results
Successfully tested by Computer Science students

- Web applications that gather information about politicians from several sources: Wikipedia, Youtube, Twitter, Portuguese Web Archive
Web archive search easily integrated on web browsers
Required research to cross-web archive search

• Cross-web archive ranking algorithms
  –How to rank search results?

• User interface design
  –How to adequately present results from different sources?
Conclusions

• Web archives must support full-text search
• Web archive search architectures are different but search interoperability should be a requirement
• OpenSearch has potential to quickly enable cross-web archive search
  – What do you think?
Contact us whenever you like
Thanks.

www.archive.pt
daniel.gomes@fccn.pt