Archiving the Web sites of Athens University of Economics and Business

Michalis Vazirgiannis
Web Archiving

- Introduction
- Context – Motivation
- International Efforts
- Technology & Systems
  - The architecture of a web archiving system
  - Crawling / Parsing Strategies
  - Document Indexing and Search Algorithms
- The AUEB case
  - Architecture / technology
  - Evaluation / Metrics
Why the Archive is so Important

- Archive’s mission: Chronicle the history of the Internet
- Data is in danger of disappearing
- Rapid evolution of the web and changes in the web content
- Hardware systems do not last forever
- Malicious Attacks
- The rise of application-based market is a potential “web killer”
- Critical to capture the web content while it still exists in such a massively public forum
Internet Statistics

- Average Hosting Provider Switching per month: 8.18%
- Web Sites Hacked per Day: 30,000
- 34% of companies fail to test their tape backups, and of those that do, 77% have found tape back-up failures.
- Every week 140,000 hard drives crash in the United States.
Evolution of the Web

Total number of websites (linear scale)

- Hostnames
- Active sites
How Business Owners Can Use the Archive

- Monitor the progress of the top competitors
- Gives a clear view of current trends
- Provides insight into how navigation and page formatting has changed over the years to suit the needs of users
- Validate digital claims
Context–Motivation

- Loss of valuable Information from websites
  - Long-term preservation of the web content
  - Protect the reputation of the institution

- Absence of major web-archiving activities in within Greece

- Archiving the Web sites of Athens University of Economics and Business
  - Hardware and Software system specifications
  - Data analysis
  - Evaluation of the results
International Efforts

The Internet Archive

- Non-profit digital Library
- Founded by Brewster Kahle in 1996
- Collection larger than 10 petabytes
- Uses Heritrix Web Crawler
- Uses PetaBox to store and process information
- A large portion of the collection was provided by Alexa Internet
- Hosts a number of archiving projects
  - Wayback Machine
  - NASA Images Archive
  - Archive-It
  - Open Library
International Efforts

- **The Wayback Machine**
  - Free service provided by The Internet Archive
  - Allows users to view snapshots of archived web pages
  - Since 2001
  - Digital Archive of the World Wide Web
  - 373 billion pages
  - Provides API to access content
  - [https://archive.org](https://archive.org)

- **Archive-it**
  - Subscription service provided by The Internet Archive
  - Allows Institutions & Individuals to create collections of digital content
  - 275 partner organizations
    - University Libraries
    - State Archives, Libraries, Federal Institutions and NGOs
    - Museums and Art Libraries
    - Public Libraries, Cities and Counties
  - [https://archive-it.org/](https://archive-it.org/)
International Efforts

- **Open Library**
  - Goal: One web page for every book ever published
  - Creator: Aaron Swartz
  - Provided by: The Internet Archive
  - Storage Technology: PostgreSQL Database
  - Book information from
    - Library of Congress
    - Amazon.com
    - User contributions

- **International Internet Preservation Consortium**
  - International organization of libraries
  - 48 members in March 2014
  - Goal: Acquire, preserve and make accessible knowledge and Information from the Internet for future generations
  - Supports and sponsors archiving initiatives like the Heritrix and Wayback Project

http://netpreserve.org/
International Efforts

Countries that have made archiving efforts

- Many more individual archiving initiatives
  

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated Total Size</td>
<td>&gt;10 petabytes</td>
</tr>
<tr>
<td>Storage Technology</td>
<td>PetaBox</td>
</tr>
<tr>
<td>Storage Nodes</td>
<td>2500</td>
</tr>
<tr>
<td>Number of Disks</td>
<td>&gt;6000</td>
</tr>
<tr>
<td>Outgoing Bandwidth</td>
<td>6GB/s</td>
</tr>
<tr>
<td>Internal network Bandwidth</td>
<td>100Mb/s</td>
</tr>
<tr>
<td>Front-end Servers Bandwidth</td>
<td>1GB/s</td>
</tr>
</tbody>
</table>

The Internet Archive Hardware Specifications
Technology & Systems Architecture

- **Storage**  ➔ WARC files: A Compressed set of uncorrelated web pages
- **Import**  ➔ Web Crawling Algorithms
- **Index & Search**  ➔ Text Indexing and Retrieval Algorithms
- **Access**  ➔ Interface that integrates functionality and presents it to the end-user

Diagram:
- System Architecture:
  - Search
  - Access
  - Import
  - Index
  - Storage

- Logical View:
  - Aggregate
  - Element
  - Data File
  - Web Crawl
  - Web Page
  - Web Objects: Text, links multimedia files
Web Crawling Strategies

A Web–Crawler’s architecture

- **Selection Policy**
  - states which pages to download

- **Re–visit Policy**
  - states when to check for changes to the pages

- **Politeness Policy**
  - states how to avoid overloading Web sites

- **Parallelization policy**
  - states how to coordinate distributed Web crawlers

- **Crawler frontier**
  - The list of unvisited URLs

- **Page downloader**

- **Web repository**
Web Crawling Strategies

A further look into Selection Policy

- **Breadth First Search Algorithm**
  - Get all links from the starting page and add them to a queue
  - Pick the first link from the queue get all links on the page and add to the queue
  - Repeat until the queue is empty

- **Depth First Search Algorithm**
  - Get the first link not visited from the start page
  - Visit link and get first non-visited link
  - Repeat until there are no unvisited links left
  - Go to first unvisited link in the previous level and repeat steps
Web Crawling Strategies

- **Page Rank Algorithm**
  - Counts citations and backlinks to a given page.
  - Crawls URL with high PageRank first

- **Genetic Algorithm**
  - Based on Evolution Theory
  - Finds the best solution within a specified time frame

- **Naïve Bayes classification Algorithm**
  - Used with structured data
  - Hierarchical website layouts
Document Indexing and Search Algorithms

Text–Indexing

- Text Tokenization
- Language–specific stemming
- Definition of Stop words
- Distributed Index

Text search and Information Retrieval

- Boolean model (BM)
- Vector Space Model (VSM)
  - Tf–Idf weights
  - Cosine–similarity
The AUEB case

System Architecture

- **Heritrix**
  - Crawls the Web and imports content in the System

- **Wayback Machine**
  - Time based document Indexing

- **Apache Solr**
  - Full-Text-Search Feature

- **Open Source Software**
Data Collection

- Heritrix Crawler
  - Crawler designed by the Internet Archive
  - Selection Policy: Uses breadth-first search algorithm by default
  - Open Source Software

- Data storage in the WARC format (ISO 88500 2009)
  - Compressed collections of web pages
  - Stores any type of files and meta-data

- Collects data based on 75 seed URL’s

- Re-visiting Policy: Checks for updates once per Month

- Politeness Policy: Collects data with respect to the Web Server.
  - Retrieves a URL
    - Every 10 seconds from the same Web Server
    - With a time delay of ten times the duration of the last crawl
### Data Collection

<table>
<thead>
<tr>
<th>Line</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WARC/1.0</td>
</tr>
<tr>
<td>2</td>
<td>WARC-Type: request</td>
</tr>
<tr>
<td>4</td>
<td>WARC-Date: 2013-05-02T16:32:28Z</td>
</tr>
<tr>
<td>5</td>
<td>WARC-Concurrent-To: <a href="">urn:uuid:fb3d4fd8-9aa3-4a06-8c08-0d5b22383dbc</a></td>
</tr>
<tr>
<td>6</td>
<td>WARC-Record-ID: <a href="">urn:uuid:4bbf8e9d-b2da-4b9a-a0a5-c5c49495cf57</a></td>
</tr>
<tr>
<td>7</td>
<td>Content-Type: application/http; msgtype=request</td>
</tr>
<tr>
<td>8</td>
<td>Content-Length: 509</td>
</tr>
<tr>
<td>9</td>
<td>GET /index.php/el/newslist/.google-analytics.com/ga.js HTTP/1.0</td>
</tr>
<tr>
<td>10</td>
<td>User-Agent: Mozilla/5.0 (compatible; heritrix/3.1.0 +<a href="http://archive.aueb.gr/">http://archive.aueb.gr/</a>)</td>
</tr>
<tr>
<td>11</td>
<td>Connection: close</td>
</tr>
<tr>
<td>13</td>
<td>Accept: text/html,application/xhtml+xml,application/xml;q=0.9,<em>/</em>;q=0.8</td>
</tr>
<tr>
<td>14</td>
<td>Host: <a href="http://www.dmst.aueb.gr">www.dmst.aueb.gr</a></td>
</tr>
</tbody>
</table>

- Part of a WARC file that crawls through the aueb.gr domain
- Captures all Html text and elements
- Content under aueb.gr can be fully reconstructed
Url based Search

- Creates the index based only on the URL and the day that the URL was Archived
  - Based on the Wayback Machine Software
- Queries must have a time frame parameter

<table>
<thead>
<tr>
<th>Month</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Φεβ 2010</td>
<td>2 pages</td>
</tr>
<tr>
<td>Μαρ 2010</td>
<td>3 pages</td>
</tr>
<tr>
<td>Απρ 2010</td>
<td>0 pages</td>
</tr>
<tr>
<td>Μαΐ 2010</td>
<td>1 page</td>
</tr>
<tr>
<td>Ιουν 2010</td>
<td>2 pages</td>
</tr>
<tr>
<td>Ιούλ 2010</td>
<td>2 pages</td>
</tr>
<tr>
<td>Αυγ 2010</td>
<td>1 page</td>
</tr>
</tbody>
</table>

Keyword based Search

- Full-text search of the archived documents based on the Apache Solr software.
- Uses a combinations of the Boolean model and space vector model for text search
- Documents "approved" by BM are scored by VSM
Evaluation of the results

- ~500,000 URL’s visited every month
- ~500 hosts visited
- The steady numbers indicate the ordinary functionality of the Web crawler.
Evaluation of the results

- Initial Configuration led the crawler into loops
- Several URLs that caused these loops were excluded (e.g. the aueb forums)
- ~50,000 URLs excluded
- ~ Initial crawl is based on ~70 seeds.
Evaluation of the results

- The number of new URIs and bytes crawled since the last crawl
- Heritrix stores only a pointer for entries that have not changed since last crawl
- The number of URIs that have the same hashcode are essentially duplicates
Evaluation of the results

- The system may fail to access a URI due to:
  - Hardware failures
  - Internet connectivity issues
  - Power outage

- The lost data will be archived by future crawls

- In general no information is lost
Data Storage Hardware Specifications

- Archive of ~500,000 URIs with monthly frequency
- Data from the Network:
  - between 27 and 32GB
- Storage of the novel URLs only:
  - Less than 10GB
- Storage in compressed format:
  - between 8 and 15GB
Future Work

- Unify all functionality (Wayback and Solr) into one user-Interface
- Experiment with different metrics and models for Full-text search using Solr
- Collect data through web forms (Deep Web)
References


- Vassilis Plachouras, Chrysostomos Kapetis, Michalis Vazirgiannis, "Archiving the Web sites of Athens University of Economics and Business", in the 19th Greek Academic Library Conference.
