Searching images in a web archive

André Mourão\textsuperscript{1,2}, Daniel Gomes\textsuperscript{2}
Arquivo.pt\textsuperscript{1,2}, NovaLincs\textsuperscript{1}
a.mourao@campus.fct.unl.pt

DSAA 2023: Thessaloniki, Greece
October 11th 2023
Images of Lopez in the green dress were downloaded from the Grammy website 642,917 times in just 24 hours after the event.

But back in 2000, search results were still just a list of blue links. When the Search team realized they weren’t able to directly surface the results that people wanted—a picture of Jennifer in the dress—they were inspired to create Google Images.

https://blog.google/products/search/18-years-after-google-images-versace-jungle-print-dress-back/
Does image search matter today?

Data provided by Jumpshot, assembled by Rand Fishkin from SparkToro

sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

Arquivo.pt Google Analytics
Does image search matter today?

~1 in 5 web searches is for images

~1 in 5 Arquivo.pt searches is for images

[sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/]

Arquivo.pt Google Analytics
Arquivo.pt Image Search
80% of web content is not available in its original form after only one year.

Archived image indexing challenges

- Metadata extraction for web content over decades
- Multiple versions of the same image over time
  - Captured more than once
  - Shows up on multiple pages
- Processing and indexing large amounts of data for real-time search

(W)ARC sizes  520 TB
Total collected files  8,500 million
Total collected images  2,408 million
Oldest image date  1994/04/15
Newest image date  2020/11/14
Finding images in pages results

- <img> tag attributes
- <a> tag attributes
- Inline CSS background image
- Inline base64 images
- Images set by JS
- <figure>, <picture>

<table>
<thead>
<tr>
<th>Reference Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;img&gt;</td>
<td>90.6%</td>
</tr>
<tr>
<td>&lt;a&gt;</td>
<td>8.7%</td>
</tr>
<tr>
<td>css</td>
<td>0.7%</td>
</tr>
<tr>
<td>imgAlt or imgTitle</td>
<td>49%</td>
</tr>
<tr>
<td>URL only</td>
<td>51%</td>
</tr>
</tbody>
</table>

Image caption extraction

First parent with text

- Default method
- Works well for images in boxes or reasonably structured pages
Image caption extraction

First parent with text
- Default method
- Works well for images in boxes or reasonably structured pages

Previous and next node text
- Used if the first parent with text is at the level of the page with more siblings
- List of images as in a blog

49% -> 99% images with specific metadata
Indexing architecture

The collection becomes available for search in Arquivo.pt
Map Reduce: Extract images and metadata
Dealing with duplicate information at scale

● The amount of data produced by this step is huge!

● But most of this information is duplicate
  ○ Images and pages that were crawled at different times but have not changed
  ○ References to the images that have the same caption/metadata
Deduplication selected solution

- We arrived at three deduplication scenarios:
  
  a. every page-image pair is a document
  
  b. the oldest page that references the image is the canonical document
  
  c. **oldest page information and image specific information from all pages**
     - keep reference to oldest page
     - Add all new image specific information (title, alt, caption) to the document
     - replace oldest page reference if a new oldest document shows up
## Impact of deduplication

<table>
<thead>
<tr>
<th>Number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a</strong> 1,862 million image-page pair documents</td>
</tr>
<tr>
<td><strong>b</strong> 971 million before deduplication across collections</td>
</tr>
<tr>
<td><strong>c</strong> 584 million documents, containing information from all 1,862 million image-page pairs</td>
</tr>
</tbody>
</table>

How will we index these **584 million** documents?
Planning SolrCloud resource allocation

- Expected index size: ~720GB
- SolrCloud servers:
  - 8 servers, 4 per branch
    - 512GB: p87, p91 (20/40 cores/threads)
    - 256GB: p82, p83 (12/24 c/t), p93, p94, p98, p99 (20/40 c/t)
  - 2560GB total, 1280GB per branch
- No SSD, only HDD, but we have more RAM than indexed data
How we configured SolrCloud?

- **solr1_1**
  - shard1_1
  - 20 GB
  - 18M documents

- **solr2_1**
  - shard2_1
  - 20 GB
  - 18M documents

- **solr3_1**
  - shard3_1
  - 20 GB
  - 18M documents

- **solr4_1**
  - shard4_1
  - 20 GB
  - 18M documents

- **solr1_8**
  - shard1_8
  - 20 GB
  - 18M documents

- **solr2_8**
  - shard2_8
  - 20 GB
  - 18M documents

- **solr3_8**
  - shard3_8
  - 20 GB
  - 18M documents

- **solr4_8**
  - shard4_8
  - 20 GB
  - 18M documents

- **solr_0**
  - No shards, will only take requests and aggregate results from other instances
Realistic query performance test

<table>
<thead>
<tr>
<th># requests</th>
<th>Avg.</th>
<th>Med.</th>
<th>P_{95%}</th>
<th>P_{99%}</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>115 ms</td>
<td>74 ms</td>
<td>235 ms</td>
<td>769 ms</td>
<td>8 q/sec</td>
</tr>
<tr>
<td>3</td>
<td>120 ms</td>
<td>76 ms</td>
<td>259 ms</td>
<td>872 ms</td>
<td>24 q/sec</td>
</tr>
<tr>
<td>5</td>
<td>136 ms</td>
<td>85 ms</td>
<td>304 ms</td>
<td>1059 ms</td>
<td>36 q/sec</td>
</tr>
<tr>
<td>10</td>
<td>211 ms</td>
<td>128 ms</td>
<td>501 ms</td>
<td>1718 ms</td>
<td>46 q/sec</td>
</tr>
<tr>
<td>25</td>
<td>489 ms</td>
<td>266 ms</td>
<td>1297 ms</td>
<td>4334 ms</td>
<td>50 q/sec</td>
</tr>
<tr>
<td>50</td>
<td>970 ms</td>
<td>593 ms</td>
<td>2694 ms</td>
<td>6699 ms</td>
<td>50 q/sec</td>
</tr>
</tbody>
</table>

- Random pairs of Portuguese words
- Warmup the index using 50 queries
- Query for 5 minutes and parse the results
Summary

● More images processed
  ○ 22M -> 1,862M images processed, 584M after deduplication

● More metadata per image
  ○ 99%+ have image metadata (imgAlt, imgTitle, imgCaption)

● Improved NSFW image processing
  ○ 7x faster processing (40 -> 280 images per second)

● Improved processing and indexing architecture
Contributions

- Fast image metadata and caption extraction for web images from all over the history of the web
- Impact of the deduplication for web archived data
- Detailed technical system architecture for a live system in the scale of the hundreds of millions
Arquivo.pt

- **Arquivo.pt** makes **8,000+ million pages** and **1,800+ million images** available for visualization and search:
  - Archived web pages -> **Text Search API/Memento/CDX Server**
  - Text and metadata search -> **Text Search API**
  - Image search -> **Image Search API**
- Available to the general public without registration
- Open Source
- [https://github.com/arquivo/pwa-technologies/wiki/APIs](https://github.com/arquivo/pwa-technologies/wiki/APIs)