Arquivo.pt Query Dataset for Research

Pedro Gomes (pedro.gomes@fccn.pt)
Daniel Gomes (daniel.gomes@fccn.pt)
Arquivo.pt (contacto@arquivo.pt)

Last review: 26 April, 2022

Introduction

Previous work with Arquivo.pt query logs

- Initial work (2009-2014)
- Recent work (2021)
  - First analysis in June 2021
  - Second analysis in December 2021

Additional references related to Arquivo.pt access logs

Future work

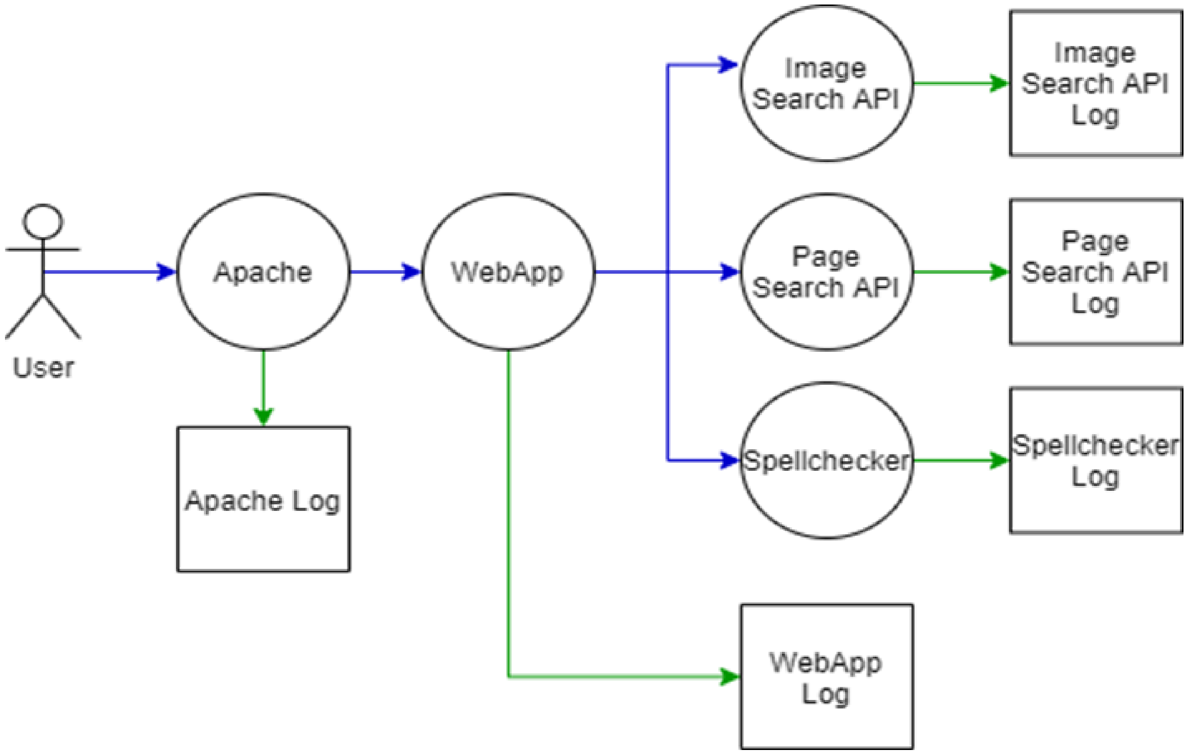
Introduction

Arquivo.pt preserves millions of files collected from the web since 1996 and provides public search services over this information. It contains information in several languages.

Periodically it collects and stores information published on the web. Then, it processes the collected data to make it searchable, providing a “Google-like” service that enables searching pages and images from the past web (English user interface available at www.archive.pt).

The objective of this document is to provide a summary of references about previous work, query workflows and structure of the corresponding query dataset produced by Arquivo.pt, to enable external researchers to study these data sets.

In general, the flow of interactions between the users and the Arquivo.pt search services is the following:



Initial work on Arquivo.pt query dataset (2009-2014)

The first analysis was done during the PhD of Miguel Costa in Java, which is now obsolete. The work done is in the repository PwaLogsMiner. The following related publications were published:

- Miguel Costa, Mário J. Silva, Towards Information Retrieval Evaluation over Web Archives, SIGIR 2009 Workshop on the Future of Information Retrieval Evaluation, Boston, U.S.A, July 2009 (poster, bibtex)
- Miguel Costa, Mário J. Silva, Characterizing Search Behavior in Web Archives, Temporal Web Analytics Workshop 2011, Hyderabad, India, March 2011 (presentation, bibtex)
- Miguel Costa, Mário J. Silva, Understanding the Information Needs of Web Archive Users,10th International Web Archiving Workshop, Vienna, Austria, September 2010 (presentation, bibtex)
- Miguel Costa, Mário J. Silva, A Search Log Analysis of a Portuguese Web Search Engine,INFORUM – Simpósio de Informática, Braga, Portugal, September 2010 (presentation, bibtex)
- Miguel Costa, Mário J. Silva, Evaluating Web Archive Search Systems, Web Information Systems Engineering 2012, Cyprus, November 2012 (presentation, bibtex).
- Miguel Costa, Information Search in Web Archives, PhD thesis, Universidade de Lisboa, December 2014 (video, ppt, bibtex).

Recent work on Arquivo.pt query dataset (2021-2022)

First analysis in June 2021

The folder First Analysis - Pedro Gomes (Jun 2021) contains the work done by Pedro Gomes in June 2021:

- Web Archive Search Evaluation Metrics is the report in which we defined the metrics used to set the long and short-term goals to decide which new products and features should be released to the user. Most of this report is based on research papers.
- Arquivo.pt search query dataset analysis - first report is the report in which we had first contact with the data to understand which metrics are possible to implement and which metrics need further development.
- Logs.pptx is the presentation made internally to the team, where we discuss which metrics have the most potential. The PowerPoint is available in this folder, while the video (in Portuguese) of the presentation and discussion is in the following private folder available under request:
    - \\paris\AWP\videos\2021\LogAnalyzer

Second analysis in December 2021

The folder Second Analysis - Pedro Gomes (Dec 2021) contains the work done by Pedro Gomes in December 2021. Based on the research work done in June, two scripts were made to process the information, which are present in the LogAnalyzer repository. The Generate_General_Logs.py script processes information from different logs (e.g., Apache

Log and Page Search API Log) and aggregates the information into a single dataset (i.e., a CSV file). The DataAnalitics.py script processes the CSV file and generates a set of charts.

Finally, the report An Analysis on a Query Log from a Arquivo.pt Search Engine was generated, which is the document that describes generally the characteristics of the Arquivo.pt search engine with simple characterization statistics of the Query Log Dataset from June 1 to August 31, 2021.

In addition to these reports and scripts, other resources were organized as support:

- Query Logs - Sample, which is a document (google sheet) containing a sample of the query logs dataset. The dataset was generated Generate_General_Logs.py
- Log_Page_Search_Log4j_Sample.txt, which is a document containing a sample of the Page Search Log (using Log4j) (e.g., check the user 109.49.178.21).
- Log_Apache_Sample.txt, which is a document containing a sample of the Apache Log (e.g., check the user 109.49.178.21).
- Logs in CSV format (in UTF-8): be careful when opening because some readers such as Microsoft Excel may use the wrong charset and damage the content for instance of column L “QUERY”.
    - See How to set character encoding when opening a CSV file in Excel? - Super User
    - Generated a Logs file in XLSX format

Additional references related to Arquivo.pt access logs

- AWstats system (private access available on request)
- Google Analytics (private access available on request)
- Dataset to facilitate research in Learning to Rank for Web Archive Information Retrieval
- Web Archive Information Retrieval video

FAQs

1. Do the dataset queries logs all come from the same Logs file?
    - No. Arquivo.pt is a service based on the concept of microservices. Thus, there are several microservices that talk to each other to return certain information to the user. So, each microservice has its own logging system. However, the Arquivo.pt team has created the Generate_General_Logs.py script which aggregates the information from different logs into a single dataset. More specifically, there are 4 logs that are used to generate the query log dataset:
        i. Apache Log
        ii. Webapp Log
        iii. Image Search API Log
        iv. Page Search API Log

2. In the ArquivoPT.csv (inside LogsArquivoPT.7z) that you sent us, what is the PAGE argument? Is it the index of the URL clicked in the results given for one query ? Then if PAGE is 0 does that mean that no result link has been clicked or has the first result been clicked ?"
   - So, First, there are 4 arguments related to each other: PAGE, MAXITEMS, POSITION and ABSOLUTE_POSITION.
   - Whenever there is a click on a result the argument POSITION records the position of that click (on the page that the user is viewing). So, the position will always be a value between 1 and MAXITEMS. MAXITEMS is the number of results per page because users can choose to view 10 results per page or 24 for example. Second, the user can click on the first POSITION on the fifth PAGE. Therefore, the POSITION will be 1 and the PAGE will be 5. So, to be easier I create a new column ABSOLUTE_POSITION, which is the clicked position within the results returned by the API with the formula "= PAGE * MAXITEMS + POSITION".
   - You can check the example from line 89 (query "rtp play nutri ventures", IP "213.190.219.72", and timestamp "1622550233").
   - PAGE = 1, MAXITEMS = 24, POSITION = 17, ABSOLUTE_POSITION = 41. (ABSOLUTE_POSITION = PAGE * MAXITEMS + POSITION).
   - The user clicked on the second page (with 24 results per page) in position 17, which is the fortieth first position (41).
   - There is no aggregation of interactions between the user and Arquivo.pt. For instance, in the lines 66, 67 and 68 (IP "2001:818:e73e:7500:dc54:f713:c1d5:6488", query "apartadox"), for the same query we have different clicks. This information is useful to know if the user makes more than one click per query.
   - To answer your question, there is no click when the POSITION argument is empty.
3. I tried for each portuguese query to get the clicked url with this PAGE argument (if it indicates the clicked url) with the help of arquivo API but sadly I think my query list was too long because my code just stopped working after getting a bit more than 1000 urls. Do you have any tip ? I could also be doing something wrong.
   - This will not work. Because we are incrementally increasing the number of pages available at Arquivo.pt. That is, the results that the user analyzed in 2021 for the query "HELLO" will be different 2022. However, the dataset that I gave you is missing one column, which will be useful to you to get this information. However, you can do the same analysis for the images. You can use the argument IMAGE_SEARCH_RESULTS to see the results return to the user in that moment.
4. Flavie's study resulted in only 1.18% of the queries were URL searches:
   - This statement is not correct because only regular queries were considered in the dataset. In other words, when the user searches for a query, only the searches made to the endpoints: https://arquivo.pt/page/search and https://arquivo.pt/image/search will be taken into account for this study. When the user searches directly for a URL in the search box, our webapp has a built-in regex that detects URLs and searches directly in the endpoint https://arquivo.pt/wayback/cdx?url=.

Future work

- Publish the datasets generated by the analysis on dados.gov.pt (Arquivo.pt account).