Technical Report - Portuguese Web Archive Image Search

André Mourão and Fernando Melo

Abstract The popular sentence *a picture paints a thousand words* illustrates the information richness an image can provide.

There are billions of images available on the Web. Such graphic and other complex resources require special search capabilities.

In the late 90's, Altavista released the first major text to image search engine on the Web, followed by Google in 2001. Searching for images is a prominent need for users in the live Web. Thus, finding images using text in Web Archives is an important task, as it adds a temporal perspective on how images in the web change.

User studies show that a significant number of users use Web Archives to find past images for a specific subjects or events. In 2016, the Internet Archive has released an image search portal for Animated GIF images harvested from the Geocities website. In 2017 the Royal Danish Library developed a new Wayback software, with image search capabilities, and in 2018 Arquivo.pt has launched an image search beta service, allowing temporal image searches (e.g. to search for images related with the *Olympic games* between 1996 and 2010).

In this report, we will describe how to build a temporal image search system for a Web archive, namely the workflow and technical implementation required to extract, classify, index and rank millions of Web images collected through the years.

The report ends with a brief discussion on possible research applications and future work on image search for Web archives, such as scene classification and automatic object and color detection.

André Mourão and Fernando Melo
FCCN , Avenida do Brasil 101, e-mail: andre.mourao@fccn.pt

1 Introduction

Web archives periodically harvest the web for preservation purposes. However, little advantage is taken from this vast wealth of information if there are no efficient search mechanisms.

Different content type (e.g. web page text, images, videos) requires specialized processing techniques to enable the information to be accessed and searched efficiently.

Fishkin research at the Inbounder[1] shows that images are amongst the most searched forms of content on the Web. Google Image Search has a count of 22,6% of the searches done on the Internet in the USA amongst the major search engines; 15 times more than YouTube; almost 10 times more than Amazon, 10 times more than Bing, and almost 20 times more than Facebook[2].

However, search on a web archive has characteristics which distinguish it from a conventional search engine. A web archive search system must be able to process large amounts of data and deal with its temporal features. It should easily enable users to efficiently find and access images compiled over the years.

Costa (2014) user study on information needs of web archive users has identified image search as a significantly mentioned users' need, that was not supported at the time of the thesis publication.

This report describes Arquivo.pt Image search platform. Section 2 gives brief overview of the recent advancements on image search systems for web archives. Section 3 describes Arquivo.pt image search, including the Web UI, API and the indexing process. Section 4 shows possible directions for the future of web archive image search.

2 Image search in web archives

2.0.1 Gifcities

In 2009 Yahoo announced it was closing down the United States version of the Geocites (hosting service where users could build their own websites), containing more than 38 million web pages. The Internet Archive has made a special effort to preserve Geocities web pages[3].

On the 20th anniversary of Internet Archive in 2016, the project *GifCities: The GeoCities Animated GIF Search Engine* was released. The Internet Archive team has created a text to image search engine for archived animated GIF images from the website geocities.com.

[1] https://theinbounder.com/videos/search-marketing-s-evolution-2018-and-beyond.html

[2] https://www.stateofdigital.com/google-images-and-visual-search/

[3] https://archive.org/details/geocities&tab=about

Gifcities is available at https://gifcities.org. Figure 1 shows a static capture of Gifcities animated image search results for the query *cat*.

In order to build *Gifcities* search engine, 4,5000,000 animated GIF images (1,600,000 unique images) were extracted from archived Web resources. Both the directory path and the filenames text were used to build the text to image search engine. When searching in the *Gifcities* portal (e.g. searching for *cat*) a list of animated GIFs related with the search is shown. Each of these animated GIF images links to an archived Web page available at the Wayback machine (i.e. a web.archive.org archived URL) which embeds the image.

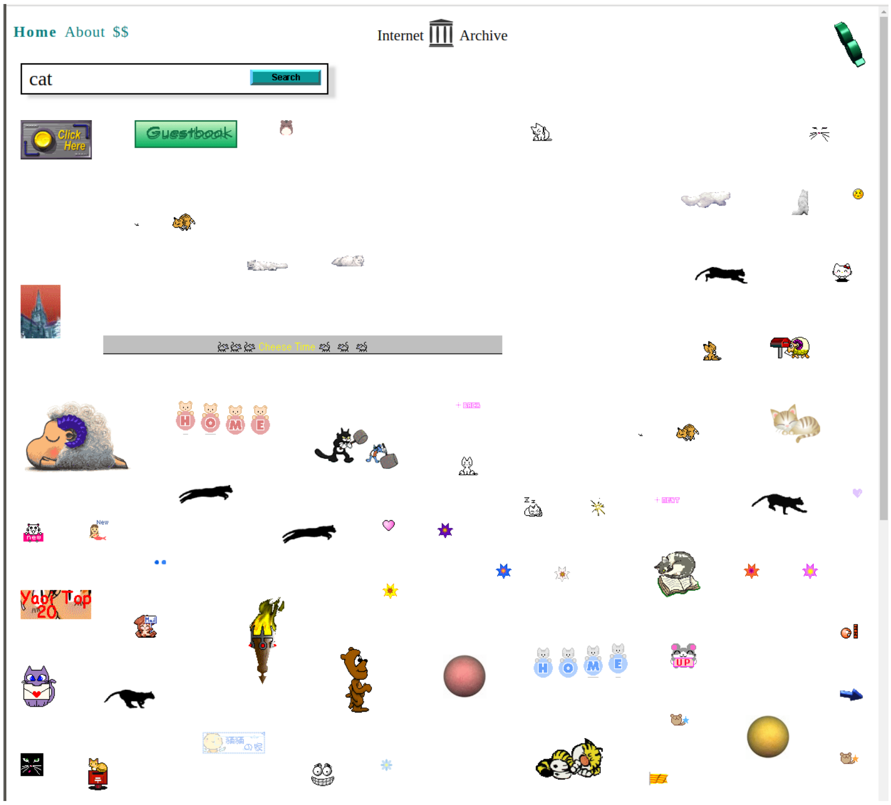


Fig. 1 A static frame of the image search results for the query *cat* in the Gifcities web portal.

2.0.2 SolrWayback

In 2017, the Royal Danish Library[4] released the first version of SolrWayback[5], a web application for browsing historical harvested ARC/WARC files, similar to the Internet Archive Wayback Machine[6]. The software is open-source and is avaliable at Github[7]. Instead of using the traditional capture indexes (CDX) to replay web archived resources, SolrWayback relies on Apache Solr server(s). ARC/WARC files have been indexed using the British Library WARC-Indexer software[8].

One of the main novelties of this Wayback software, when comparing with OpenWayback [9] or pywb[10], is that it ships with image search capabilities. Figure 2 shows the image search results within a Web archive for a single term query (*obama*) using SolrWayback. Bellow each image there is an option to *Search for image*, in order to find all occurrences of the image within the Web archive. Besides the traditional text to image search, SolrWayback also enables to search for images by uploading an image, i.e. you upload an image, and discover if that particular image has been harvested, and from which domains. Another interesting SolrWayback feature is that it allows to search by location for images that contain (EXIF)[11] location information. Figure 3 shows a location search for all images within a 50 km radius centered on the Royal Danish Library.

3 Arquivo.pt Image Search System

In 2018, Arquivo.pt launched a text to image search system for its entire Web archive, comprising a total of 23 million unique images from 1996 to 2017. Inspired by Google Image Search[12], the goal is to provides a familiar experience when searching for images from the past web.

table 1 shows the latest set of detailed statistics for the service. The service is available at arquivo.pt/images.jsp.

The user inputs a textual query and a set of image results are displayed in a grid, as illustrated in Figure 4. Similarly to Gifcities, each image search result links to an archived Web page that embeds the image. When clicking on an image result, an image viewer is displayed (Figure 5), showing a larger image together with details

[4] http://www.kb.dk/

[5] https://github.com/netarchivesuite/solrwayback

[6] https://archive.org/web/

[7] https://github.com/netarchivesuite/solrwayback

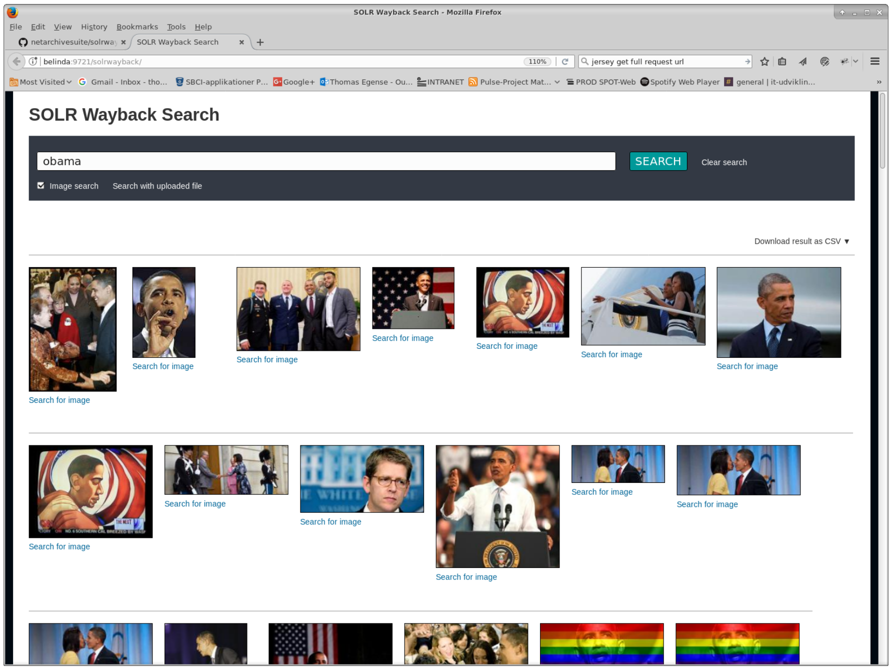
[8] https://github.com/ukwa/webarchive-discovery/

[9] https://github.com/iipc/openwayback

[10] https://github.com/webrecorder/pywb

[11] https://www.exif.org/Exif2-2.PDF

[12] https://images.google.pt/

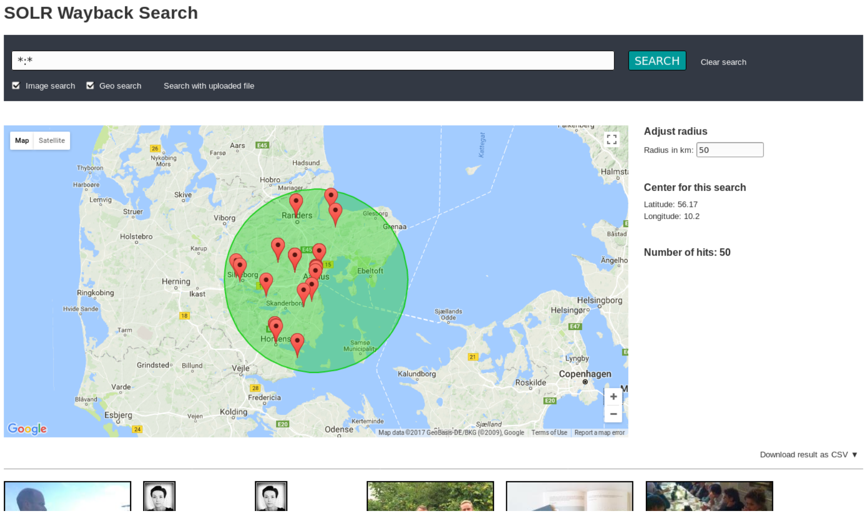


SOLR Wayback Search

obama

☑ Image search

Fig. 2 Image search results for the query *Obama* in a web archive using SolrWayback.



SOLR Wayback Search

:

☑ Image search ☑ Geo search

Adjust radius

Radius in km: 50

Center for this search

Latitude: 56.17
Longitude: 10.2

Number of hits: 50

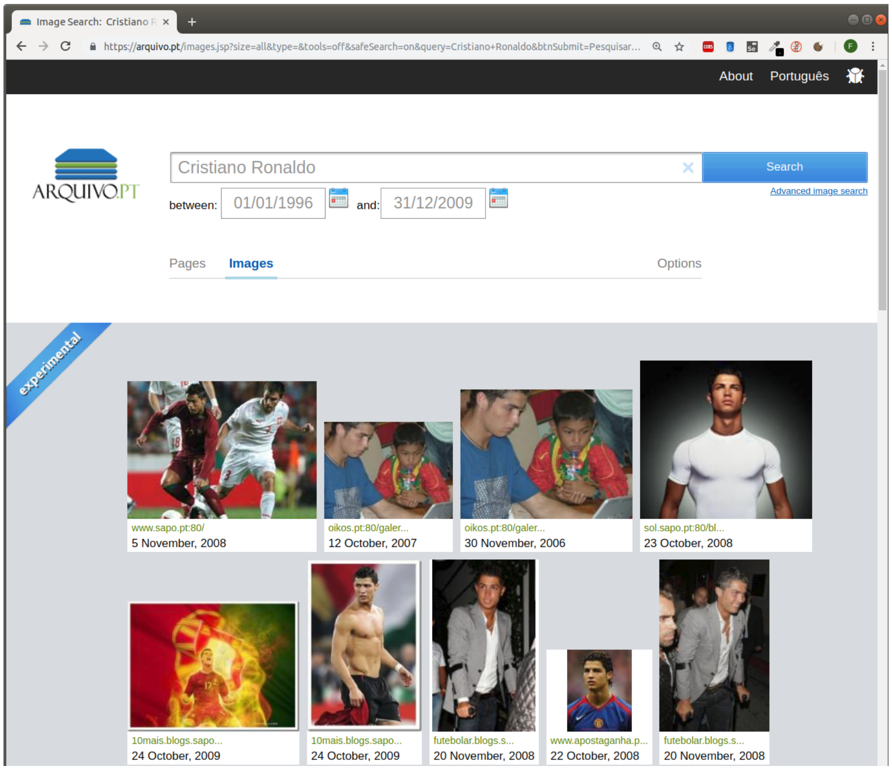
Fig. 3 SolrWayback geosearch for images within a 50km radius centered on the Royal Danish Library.

about the current image and the web page where it was found. The source code of Arquivo.pt image search system is open-source and is freely available at Github[13].

[13] https://github.com/arquivo/ImageSearchIndexing

Table 1 General statistics for the image search system

Indexed images	23,589,395
Crawl/Collection count	88
(W)ARCS	3,414,742 (336.47 TB)
Total collected files	6,086,768,283
Start of crawl	01/01/1996
End of daily crawls	31/12/2018
End of special crawls	14/11/2019
Oldest image date	15/04/1994
Newest image date	14/11/2019



https://arquivo.pt/images.jsp?size=all&type=&tools=off&safeSearch=on&query=Cristiano+Ronaldo&btnSubmit=Pesquisar...

Cristiano Ronaldo

between: 01/01/1996 and: 31/12/2009

www.sapo.pt:80/
5 November, 2008

oikos.pt:80/galer...
12 October, 2007

oikos.pt:80/galer...
30 November, 2006

sol.sapo.pt:80/bl...
23 October, 2008

10mais.blogs.sapo...
24 October, 2009

10mais.blogs.sapo...
24 October, 2009

futebolar.blogs.s...
20 November, 2008

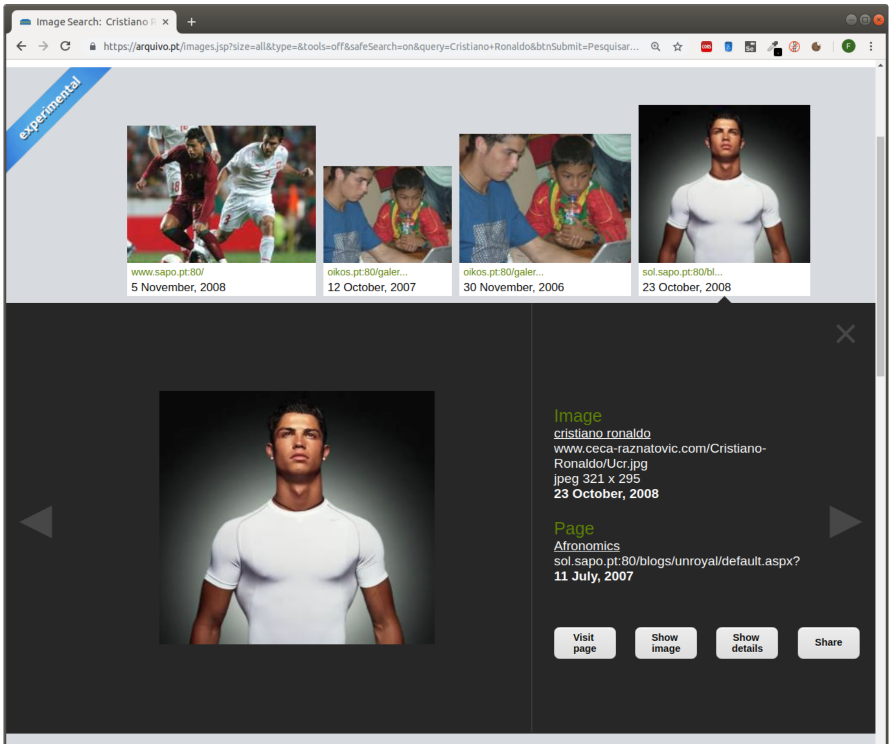
www.apostaganha.p...
22 October, 2008

futebolar.blogs.s...
20 November, 2008

Fig. 4 Image search results for the query *Cristiano Ronaldo* between the dates *01/01/1996* and *31/12/2009* in Arquivo.pt web archive.

3.1 Arquivo.pt Image Search Filters

Arquivo.pt's image search system supports multiple filters, to help the user fine tune the displayed search results:



https://arquivo.pt/images.jsp?size=all&type=&tools=off&safeSearch=on&query=Cristiano+Ronaldo&btnSubmit=Pesquisar...

www.sapo.pt:80/
5 November, 2008

oikos.pt:80/galer...
12 October, 2007

oikos.pt:80/galer...
30 November, 2006

sol.sapo.pt:80/bl...
23 October, 2008

Image
cristiano ronaldo
www.ceca-raznatovic.com/Cristiano-Ronaldo/Ucr.jpg
jpeg 321 x 295
23 October, 2008

Page
Afronomics
sol.sapo.pt:80/blogs/unroyal/default.aspx?
11 July, 2007

Fig. 5 Arquivo.pt image viewer. When clicking on an image result, the image is expanded, showing more details the image and the page that contains the it

- Time filter (e.g. find images of *Cristiano Ronaldo* between 1996 and 2000). See Figure 4. This is a key filter for users that are looking for images from the past, as it allows focusing on specific time periods;
- Filter by domain (e.g. find all images related with the term *goal* collected from the website portugalnews.com). See Figure 6;
- Filter by mimetype (e.g. search for images related with the term *Homer Simpson* in the *GIF* mimetype) See Figure 7;
- Filter by image size (e.g. search for *large* images of *Lisbon*); See Figure 8;
- Filter by collection (e.g. find only images which belong to a specific collection[14]. A complete list of Arquivo.PT collections is available here https://docs.google.com/spreadsheets/d/1SjijGAMXgUcwBaH

[14] Collections are set of related web content, captured together. The relation may be temporal (e.g. AWP30 collection: periodic crawl of the PT domain performed on April 2019) or a part of an thematic crawl (e.g. BlogsSapo2018 collection: deep crawl of all sapo.pt blogs)

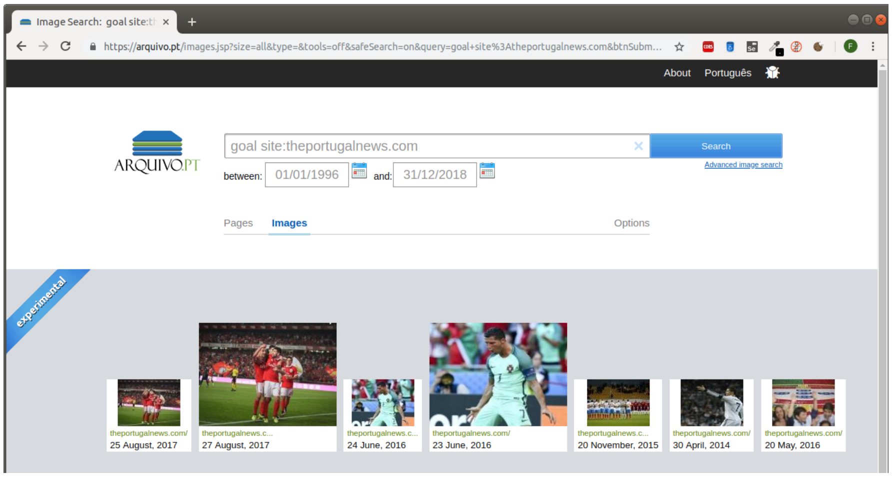


Image Search: goal site:th

https://arquivo.pt/images.jsp?size=all&type=&tools=off&safeSearch=on&query=goal+site%3Atheportugalnews.com&btnSubm...

About Português

goal site:theportugalnews.com

Search

Advanced image search

between: 01/01/1996 and: 31/12/2018

Pages Images Options

experimental

theportugalnews.com/	theportugalnews.c...	theportugalnews.c...	theportugalnews.com/	theportugalnews.c...	theportugalnews.com/	theportugalnews.com/
25 August, 2017	27 August, 2017	24 June, 2016	23 June, 2016	20 November, 2015	30 April, 2014	20 May, 2016

Fig. 6 Arquivo.pt image search by domain. Searching for images related with the term *goal* within the domain *theportugalnews.com*

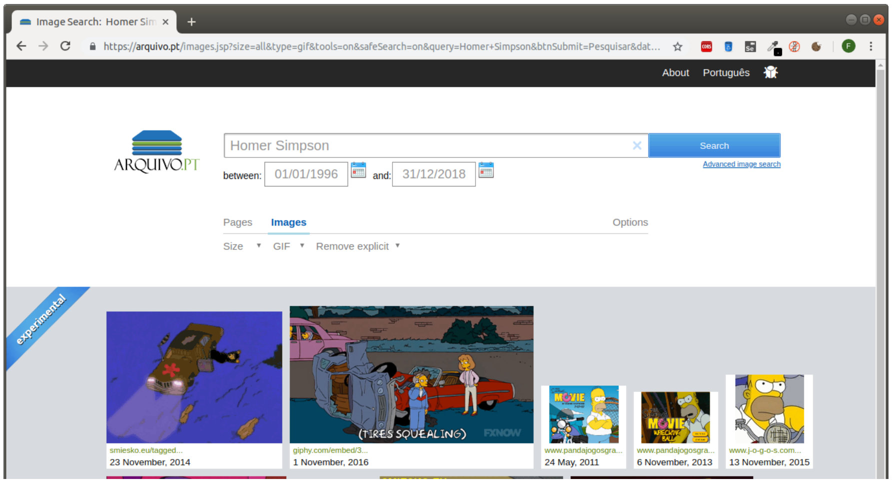


Image Search: Homer Sim

https://arquivo.pt/images.jsp?size=all&type=gif&tools=on&safeSearch=on&query=Homer+Simpson&btnSubmit=Pesquisar&dat...

About Português

Homer Simpson

Search

Advanced image search

between: 01/01/1996 and: 31/12/2018

Pages Images Options

Size ▾ GIF ▾ Remove explicit ▾

experimental

smiesko.eu/tagged...	giphy.com/embed/3...	www.pandajogosgra...	www.pandajogosgra...	www.j-o-g-o-s.com...
23 November, 2014	1 November, 2016	24 May, 2011	6 November, 2013	13 November, 2015

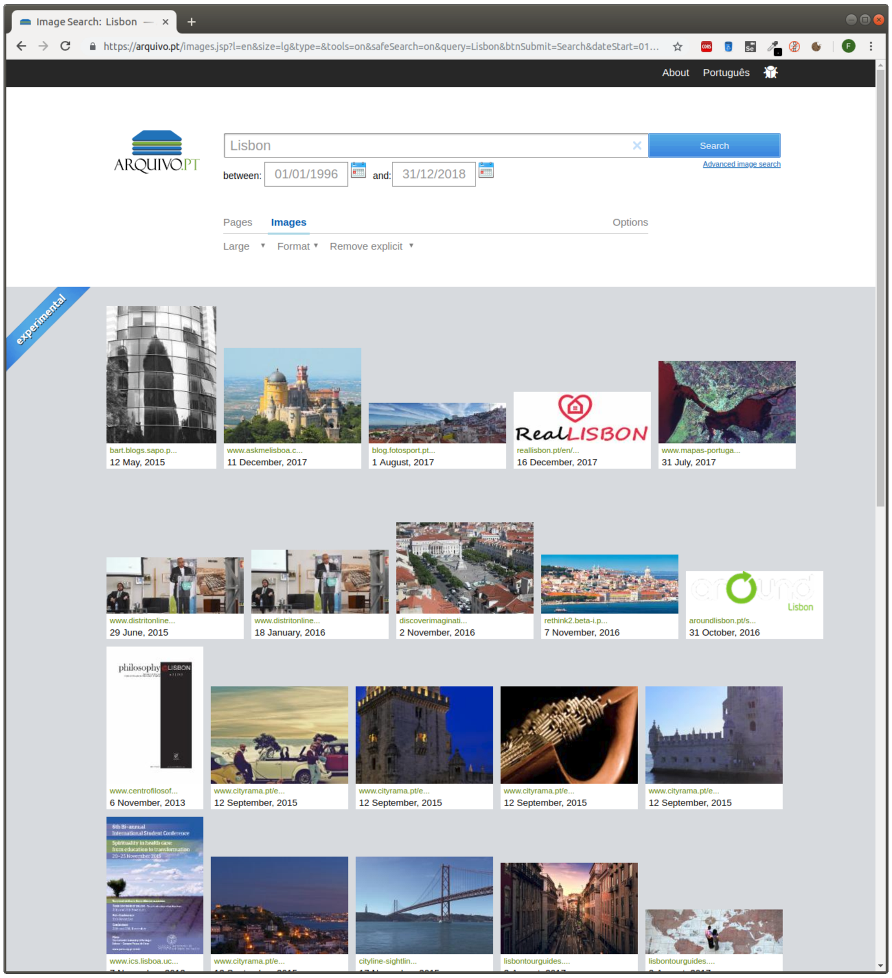
Fig. 7 Arquivo.pt image search restricting mimetype. Searching for images related with *Homer Simpson* in the *GIF* image format

3.2 Arquivo.pt Image Indexing Workflow

In order to build a fast image search system for a large Web archive, we created image specific indexes.

Our image indexing system consists of three main steps:

1. Creation of image indexes from ARC/WARC files;
2. Image classification;



https://arquivo.pt/images.jsp?l=en&size=lg&type=&tools=on&safeSearch=on&query=Lisbon&btnSubmit=Search&dateStart=01…

Lisbon

between: 01/01/1996 and: 31/12/2018

Large ▾ Format ▾ Remove explicit ▾

Fig. 8 Arquivo.pt image search restricting image size. Searching for *large* images related with the term *Lisbon*.

3. Solr indexing.

3.2.1 Creation of image indexes from ARC/WARC files.

Creating an index for image search using data captured for archival purposes poses a specific set of challenges, including separating images from the remaining web content, finding the page where images were found and finding which metadata is associated with the image itself.

Arquivo.pt is using a cluster of servers running Apache Hadoop[15] (Lam, 2010), an open-source software library that allows distributed processing of large data, to process hundreds of thousands of ARC/WARC files per collection. To store image metadata, Arquivo.pt chose a MongoDB sharded cluster of servers. MongoDB is a NoSQL (Cattell, 2011) document database, scalable and flexible, which allows to store, and retrieve large volume of data. Each document is a JSON-like object, allowing fields to vary as well as the data structure.

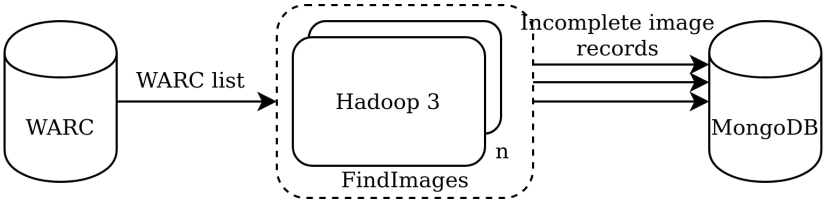


Fig. 9 Data flow for Image Extraction in (W)ARC files

In order to create image indexes two main jobs are run for each collection of ARC/WARC files:

1. Image Extraction.
   All images are extracted from the ARC/WARC files in a given collection; image records containing image attributes, such as, *image width*, *image height*, *image timestamp*, *image mime type*, together with a generated *image thumbnail in base64* are stored in a MongoDB sharded database, fig. 9.
2. HTML <img> tags extraction.
   Figure 10 shows the flow of data for this job. <*img*> tags are extracted from the HTML records in the ARC/WARC files within the collection. Then, the image record is compared with the image record found in the Image Extraction stage (stored in MongoDB). If no image matches the URL in the found <*img*>, it is ignored, as this means that the original image was not crawled, and thus, cannot be added to the index. For each matching tag, the following information is extracted and added to MongoDB:

   - *imgSrcTokens* - The keywords of the image src attribute. The image src attribute identifies an image by a URL, which often includes the filename of the image;
   - *imgTitle* - Image title attribute; it is used to provide additional information about the image;
   - *imgAlt* - Image alt attribute; it provides alternative information about an image if a user cannot view it;
   - *pageTitle* - Page title attribute; it is used to provide additional information about an HTML page;
   - *pageURLTokens* - The keywords of the URL of the HTML page that contains the image.

[15] https://hadoop.apache.org/

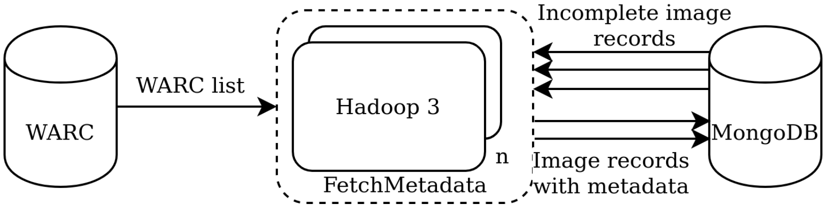


Fig. 10 Data flow for <img> tag extraction in (W)ARC files

In addition to the removal of non-crawled images, Arquivo.pt performs two additional filtering steps:

- discard all images with less than 50 pixels of width or less than 50 pixels height, as they are mostly decorative HTML documents from early web design standards (e.g. vertical and horizontal lines, rounded table corners);
- discard image duplicates per collection of ARC/WARC files. Each image has a unique fingerprint (called digest) and for each image digest, only one image index document is created. If multiple websites (or the same website captured at different times) link to an image with the same image digest (SHA-256 file hash), Arquivo.pt only stores the record with the older image timestamp per digest and per collection.

The image indexes from Item 2 are exported from the MongoDB database to JSON files where each line contains an image index.

Output

MongoDB database with image records. It follows a simplified example of a single image index obtained after the creation of image indexes from ARC/WARC files.

```
{
  "pageURL": "http://www.celticgold.eu/b_en/gold-university/h
ow-bank-guarantys-work.html",
  "imgWidth": 299,
  "imgHeight": 168,
  "pageTstamp": 20141122085132,
  "pageTitle": "How Bank Guarantys work",
  "imgThumbnailBase64": "/9j/4AAQSkZJRgABAg...RRUgf/2Q==",
  "pageProtocol": "http",
  "pageHost": "www.celticgold.eu",
  "imgTstamp": 20141122085145,
  "imgTitle": "Angela Merkel",
  "imgSrc": "http://www.celticgold.eu/media/wysiwyg/CGNews26t
hFeb2012_page6_image9.jpg",
  "imgMimeType": "image/jpeg",
  "imgDigest": "29bae2624f71c1b473...a41f3a068d17e8e17aeeb375",
  "imgAlt": "Angela Merkel",
}
```

3.2.2 Image classification.

Arquivo.pt is a public service that does broad domain crawls to the .pt domain, towards preserving as much information that may be of interest for the Portuguese community as possible. In that context, and due to the graphic nature that some images provide, Arquivo.pt enables a default filter in order to exclude pornographic images from the search results, automatically classified as Not Safe for Work (NSFW). This filter can be disabled by the user through the image search interface.

Arquivo.pt has developed a custom NSFW image classifier based on OpenNSFW[16] (Sanzgiri et al., 2018; Gangwar et al., 2017) - a Deep Neural Network (DNN) solution (He et al., 2016; Simonyan and Zisserman, 2014; He and Sun, 2014) released by Yahoo! for classification of NSFW images.

Arquivo.pt is using 2 servers equipped with Nvidia P40 Graphics Processing Units, in order to process and classify hundreds of thousands of images per collection.

Figure 11 shows a simplified representation of how the data goes from MongoDB to the Solr, including NSFW classification. All the images extracted from Section 3.2.1 are exported from MongoDB and classified with Arquivo.pt NSFW classifier. The result from this step is the addition of a field with the name *safe* for each image index. This value varies from 0.000 to 1.000 for each image. The closer this value is to 1.000, the more *safe* the image is according to the automatic classifier.

In the future, more image classifiers may be added in this stage. For example, a scene recognition classifier to extract text categories from images (Zhou et al., 2018), or an image caption sentence generator (Vinyals et al., 2015).

Output

A safe field is added to each image index from Section 3.2.1.

```
{
    ...
    "safe": 0.997,
}
```

[16] https://github.com/yahoo/open_nsfw

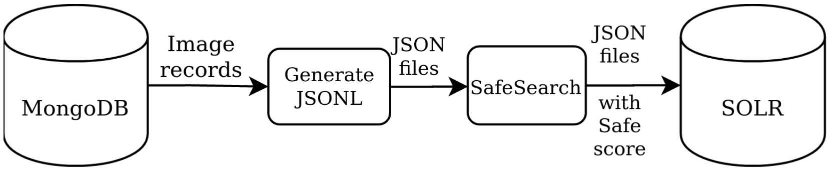


Fig. 11 SafeSearch classification and SOLR index creation

3.2.3 Solr indexing.

Solr[17] is an open source enterprise search platform which enables distributed full-text search. It is written in Java and was built on top of Apache Lucene project and is developed and maintained by the Apache Foundation.

Arquivo.pt image search system uses Solr to enable text to image search. In this step, the JSON image indexes from Section 3.2.2 are indexed in Solr, which converts and stores these image records in Apache Lucene indexes.

Arquivo.pt currently has 2 independent servers with Apache Solr software installed. In a near future Arquivo.pt is going to change to a SolrCloud[18] architecture which enables to set up a cluster of Solr servers that combines fault tolerance and high availability. Each server consumes about 4 GB of RAM to enable text to image search.

After Solr indexing is complete, all the images that were indexed are automatically searchable in arquivo.pt/images.jsp.

Output

Lucene indexes - the underlying structure that Solr uses for its powerful text-search capabilities. Once Solr indexing is completed images are searchable in Arquivo.pt.

3.3 Ranking image search

Arquivo.pt image search engine finds images based on the user input text, and retrieves the top results, i.e. the images with higher scores.

Each image index contains text attributes extracted from the <*img*> HTML tags (e.g. alt text, tokenized URL), together with text attributes extracted from the HTML page that contains the image (e.g. original HTML page title).

Thousands of images are returned for certain queries. It is thus of extreme importance to provide the best ranking possible, in order to show to the user the most relevant images from these large sets of results.

[17] https://lucene.apache.org/solr/

[18] https://lucene.apache.org/solr/guide/6_6/solrcloud.html

Arquivo.pt is currently ranking the image search results based on the following fields, described in detail in section 1: *imgSrcTokens*, *imgTitle*, *imgAlt*, *pageTitle* and *pageURLTokens*.

The Arquivo.pt image search system uses BM25 (Robertson and Zaragoza, 2009) ranking function for each field. Multiplicative boosts are then given to each field according to their importance. The image search ranking for a single term query is calculated according to Equation (1).

$$\begin{aligned} originalScore = {} & 4 \times imgTitleBM25 + 3 \times imgAltBM25 + \\ & 2 \times imgSrcTokensBM25 + pageTitleBM25 + \\ & pageURLTokensBM25 \end{aligned} \tag{1}$$

In Equation (1), *imgTitleBM25, imgAltBM25, imgSrcTokensBM25, pageTitleBM25 and pageURLTokensBM25* correspond to the BM25 score of the query term for the ranking fields. As one can observe, the most important field for the image search ranking is the image title, followed by the image alt text.

Additional ranking scores are given to phrases, according to the following Solr query fields[19]:

- *Phrase fields (pf)* - boosts the score of an image index when all the terms of a query exist in a given ranking field in close proximity.
- *Phrase slop (ps)* - specifies the distance the indexed search terms can have in the document and still influence relevancy. The amount of slop, i.e. the distance between indexed search terms, is defined by the *ps* field and affects the phrase fields (pf).
  E.g. if *ps=1* and the input query is *UEFA 2004*, a ranking field containing the text *UEFA Euro 2004*, is considered as a phrase match, because *ps=1* allows up to 1 word of slop between query terms.
- *Phrase bi-gram fields (pf2)* - similar to *pf* field, but it breaks the input down into word bi-grams (Brown et al., 1992).
- *Phrase slop 2 (ps2)* - similar to *ps* field, the amount of slop applied to the *pf2* field.
- *Phrase tri-gram fields (pf3)* - similar to *pf,pf2* fields, but it breaks the input down into word tri-grams.
- *Phrase slop 3 (ps3)* - similar to *ps,ps2* fields, the amount of slop applied to the *pf3* field

Equations (2) to (4) show Arquivo.pt image search ranking boosts for phrase fields (pf), phrase bi-gram fields (pf2) and phrase tri-gram fields (pf3), according to their respective phrase slops (*ps1, ps2, and ps3*).

[19] https://lucene.apache.org/solr/guide/6_6/the-extended-dismax-query-parser.html#TheExtendedDisMaxQueryParser-Theps2Parameter

$$\begin{aligned} ps1 &= 1 \\ pf1 &= 4000 \times imgTitleBM25 + 3000 \times imgAltBM25+ \\ &\quad 2000 \times imgSrcTokensBM25 + 1000 \times pageTitleBM25+ \\ &\quad 1000 \times pageURLTokensBM25 \end{aligned} \tag{2}$$

$$\begin{aligned} ps2 &= 2 \\ pf2 &= 400 \times imgTitleBM25 + 300 \times imgAltBM25+ \\ &\quad 200 \times imgSrcTokensBM25 + 100 \times pageTitleBM25+ \\ &\quad 100 \times pageURLTokensBM25 \end{aligned} \tag{3}$$

$$\begin{aligned} ps3 &= 3 \\ pf3 &= 40 \times imgTitleBM25 + 30 \times imgAltBM25+ \\ &\quad 20 \times imgSrcTokensBM25 + 10 \times pageTitleBM25+ \\ &\quad 10 \times pageURLTokensBM25 \end{aligned} \tag{4}$$

The rationale between choosing this boosting structure is that it ensures that documents that have all query terms close together, but still ensuring partial query matches are still part of the search results. This process matches users expectation and feedback, as most queries consist on a person or institution name (Costa, 2014), and showed good results on our empirical evaluations.

The final document score is as follows:

$$finalScore = \ originalScore + pf1 + pf2 + pf3 \tag{5}$$

For each query, image indexes are ranked, and the image indexes with higher scores are shown to the user.

3.4 Arquivo.pt Image search API for researchers

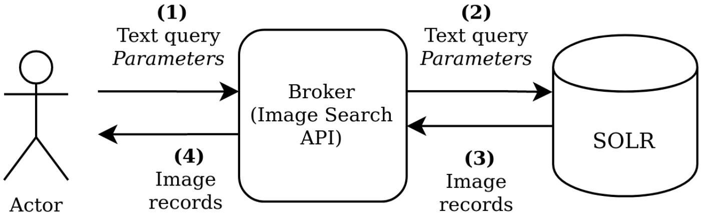


Fig. 12 Image Search API data flow

Arquivo.pt developed and open image search API, so that third-party software developers can integrate the Arquivo.pt image search results in their applications, Figure 12.

The API is documented on GitHub[20], and the endpoint is located at arquivo.pt/imagesearch.

The response is given in the JSON file format, and is composed by a response header, which contains basic information such as the links to the next and the previous page of results, the total number of images found for the query, and by a list of response items (responseItems) containing the highest scored image indexes.

The main search parameter of this API is the *q* parameter (query). For example, one can search for images from the past preserved by arquivo.pt and related with the keyword *european* by entering the url arquivo.pt/imagesearch?q=european. As with the Web UI, the API also allows filtering by time, domain, collection and type of image (gif, jpg, png, ...). Filtering by collection is specially important, as it allows to find documents captured under special requests (e.g. images from a defunct blogging platform or Portuguese research institution pages). An example of a response from the Image search API is shown below.

Arquivo.pt Image search API JSON response

```
{
  "serviceName": "Arquivo.pt - image search service.",
  "linkToService": "https://arquivo.pt/images.jsp",
  "linkToDocumentation": "https://github.com/arquivo/
pwa-technologies/wiki/ImageSearch-API-v1-(beta)",
  "linkToMoreFields": "https://arquivo.pt/imagesearch?
q=european&prettyPrint=true&more=imgThumbnailBase64,
imgSrcURLDigest,imgDigest,pageProtocol,pageHost,pageImages,
safe",
  "nextPage": "https://arquivo.pt/imagesearch?q=european
&prettyPrint=true&offset=50",
  "previousPage": "https://arquivo.pt/imagesearch?
q=european&prettyPrint=true&offset=0",
  "totalItems": 87642,
  "numberOfResponseItems": 50,
  "offset": 0,
  "responseItems": [
  ...
   {
      "imgAlt": "|-eu-flag.gif 170x113px",
      "imgTstamp": 20160314101513,
      "imgWidth": 170,
      "imgTitle": "European Union - European Community",
      "pageTstamp": 20160314114224,
      "pageTitle": "LSIWC",
```

[20] https://github.com/arquivo/pwa-technologies/wiki/ImageSearch-API-v1-(beta)

```
      "pageURL": "http://www.kki.lv/index.php?lang=en&id=88
&izmers=3",
      "collection": "EAWP10",
      "imgMimeType": "image/gif",
      "imgSrc": "http://www.kki.lv/galerija/lielas
/eu-flag.gif",
      "imgHeight": 113,
      "imgLinkToArchive": "https://arquivo.pt/wayback
/20160314101513/http://www.kki.lv/galerija/lielas
/eu-flag.gif",
      "pageLinkToArchive": "https://arquivo.pt/wayback
/20160314114224/http://www.kki.lv/index.php?lang=en&id=88
&izmers=3"
    }
  ...
}
```

Each response item contains information such as the image width (*imgWidth*) and height (*imgHeight*), the timestamp when the image was crawled from the Web (*imgTstamp*), how to obtain the image (*imgLinkToArchive*), and how to obtain the page that contains the image (*pageLinkToArchive*).

For more information consult our APIs page[21].

4 The Future of Image Search on Web Archiving

Despite the recent advancements on image search systems for Web archiving described throughout this report, there are still many opportunities for further enhancements.

With the current evolution of hardware such as Graphic Processing Units (GPUs), and image classification techniques, one can build better image search models for Web archiving in a near future.

Research possibilities to improve Web archive image search systems include but are not limited to: extracting categories from images using scene recognition classifiers (Zhou et al., 2018), or even generating image captions (Vinyals et al., 2015); retrieving similar images (Wang et al., 2014); and retrieving images by dominant color(s) (Wang and Hua, 2011), shapes and textures.

Another research possibility for image search systems is the development of mobile user interfaces, justified by a growing percentage of users that access Web archives using tablets and mobile phones.

[21] https://arquivo.pt/apis

Finally, the development of standards and APIs is fundamental to increase interoperability, facilitating the emergence of new applications and research based on images from the past.

References

Brown PF, deSouza PV, Mercer RL, Pietra VJD, Lai JC (1992) Class-based n-gram models of natural language. Comput Linguist 18(4):467–479

Cattell R (2011) Scalable sql and nosql data stores. SIGMOD Rec 39(4):12–27, DOI 10.1145/1978915.1978919, URL https://doi.org/10.1145/1978915.1978919

Costa M (2014) Information search in web archives. PhD thesis, Universidade de Lisboa

Gangwar A, Fidalgo E, Alegre E, González-Castro V (2017) Pornography and child sexual abuse detection in image and video: A comparative evaluation. In: 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017), pp 37–42, DOI 10.1049/ic.2017.0046

He K, Sun J (2014) Convolutional neural networks at constrained time cost. 1412.1710

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778, DOI 10.1109/CVPR.2016.90

Lam C (2010) Hadoop in Action, 1st edn. Manning Publications Co., Greenwich, CT, USA

Robertson S, Zaragoza H (2009) The probabilistic relevance framework: Bm25 and beyond. Found Trends Inf Retr 3(4):333–389, DOI 10.1561/1500000019, URL https://doi.org/10.1561/1500000019

Sanzgiri A, Austin D, Sankaran K, Woodard R, Lissack A, Seljan S (2018) Classifying sensitive content in online advertisements with deep learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp 434–441, DOI 10.1109/DSAA.2018.00056

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. 1409.1556

Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Wang J, Hua XS (2011) Interactive image search by color map. ACM Trans Intell Syst Technol 3(1), DOI 10.1145/2036264.2036276, URL https://doi.org/10.1145/2036264.2036276

Wang J, Song Y, Leung T, Rosenberg C, Wang J, Philbin J, Chen B, Wu Y (2014) Learning fine-grained image similarity with deep ranking. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, USA, CVPR '14, p 1386–1393, DOI 10.1109/CVPR.2014.180, URL https://doi.org/10.1109/CVPR.2014.180

Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2018) Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(6):1452–1464, DOI 10.1109/TPAMI.2017.2723009