INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Área Departamental de Engenharia de Electrónica e Telecomunicações e de Computadores

Automatic Identification of Not Suitable For Work images

DANIEL PESCA RODRIGUES BICHO

Master Degree

Relatório preliminar para obtenção do Grau de Mestre
em Engenharia Informática e de Computadores

Orientadores : PhD Artur Ferreira
PhD Nuno Datia

January, 2018

INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Área Departamental de Engenharia de Electrónica e Telecomunicações e de Computadores

Automatic Identification of Not Suitable For Work images

DANIEL PESCA RODRIGUES BICHO

Master Degree

Relatório preliminar para obtenção do Grau de Mestre
em Engenharia Informática e de Computadores

Orientadores :	PhD Artur Ferreira
	PhD Nuno Datia

January, 2018

Contents

List of Figures **vii**

List of Tables **ix**

1 Not Suitable for Work Images **1**

1.1 The Problem 3

1.2 Project Proposal 4

1.2.1 Planned Activities 5

1.3 Arquivo.pt Data Characterization 5

1.4 Infrastructure of Arquivo.pt 7

1.4.1 Crawler System 8

1.4.2 Replay System 8

1.4.3 Indexing System 8

1.4.4 Search System 9

1.5 State of the Art 9

List of Figures

1.1 Example of Arquivo.pt Image Search. 2

1.2 Example of Arquivo.pt Image Search problematic content. 3

1.3 Workflow of the classification system. 4

1.4 Planned activities. 5

List of Tables

1.1 Measure of the number of resources mime-types collected on Arquivo.pt last broad domain crawl (AWP24). 7

1.2 Mime-types distribution for image type contents. 7

1

Not Suitable for Work Images

Web Archiving [**?**] is a research field about how to collect portions of the World Wide Web to ensure that information is preserved in an archive for future researchers, historians and the general public.

The most common way used by Web Archives to collect this information for preservation means is through web crawlers as Heritrix [**?**], specialized at archiving the web that automates the process of harvesting web pages and preserving their contents. These contents include any resource type, such as Hyper Text Markup Language (HTML), style sheets, javascript, images, videos and metadata about the preserved resources such as access times, resource mime-types and content length.

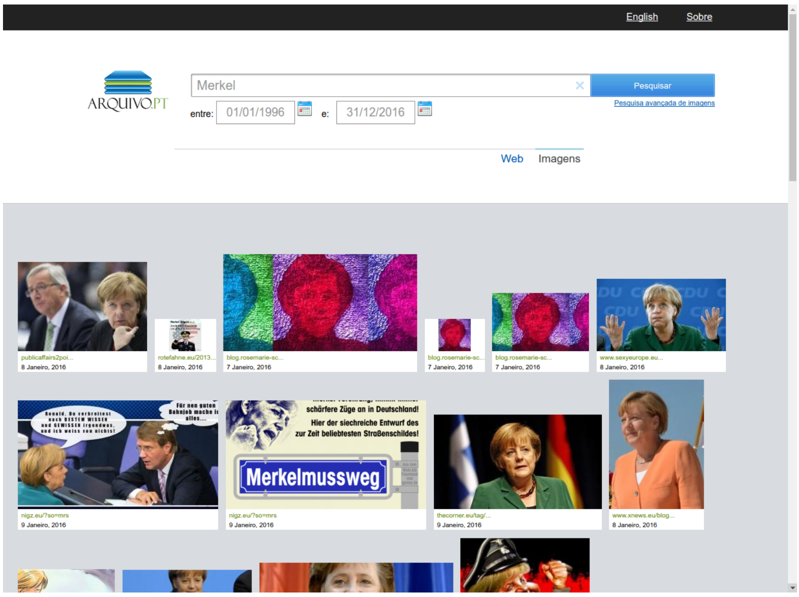
Choosing what to be preserved is a hard challenge, since there isn't enough storage space to preserve everything and the amount of data that is available on the web is permanently growing.

There are several Web Archiving initiatives worldwide that try to preserve the Web. Some Web Archives have more narrow scopes, preserving just some very specific kind of pages like institutional web pages (European Commission Historical Archives[1]), others try to preserve the entire national top-level domain (UK Web Archive[2]), and others the entire web (Internet Archive)[3].

[1] http://ec.europa.eu/historical_archives/index_en.htm

[2] https://www.webarchive.org.uk/ukwa/

[3] https://archive.org/



Merkel

entre: 01/01/1996 e: 31/12/2016

publicaffairs2poi... 8 Janeiro, 2016

rotefahne.eu/2013... 8 Janeiro, 2016

blog.rosemarie-sc... 7 Janeiro, 2016

blog.rosemarie-sc... 7 Janeiro, 2016

blog.rosemarie-sc... 7 Janeiro, 2016

www.sexyeurope.eu... 8 Janeiro, 2016

nigz.eu/?so=mrs 9 Janeiro, 2016

nigz.eu/?so=mrs 9 Janeiro, 2016

thecorner.eu/tag/... 9 Janeiro, 2016

www.xnews.eu/blog... 8 Janeiro, 2016

Figure 1.1: Example of Arquivo.pt Image Search.

Arquivo.pt[4] is one of those initiatives and it preserves the Portuguese *.pt* top-level domain and all web pages that publish Portuguese related important information. It also acts as a research infrastructure making its contents available in open access to researchers and to the general public.

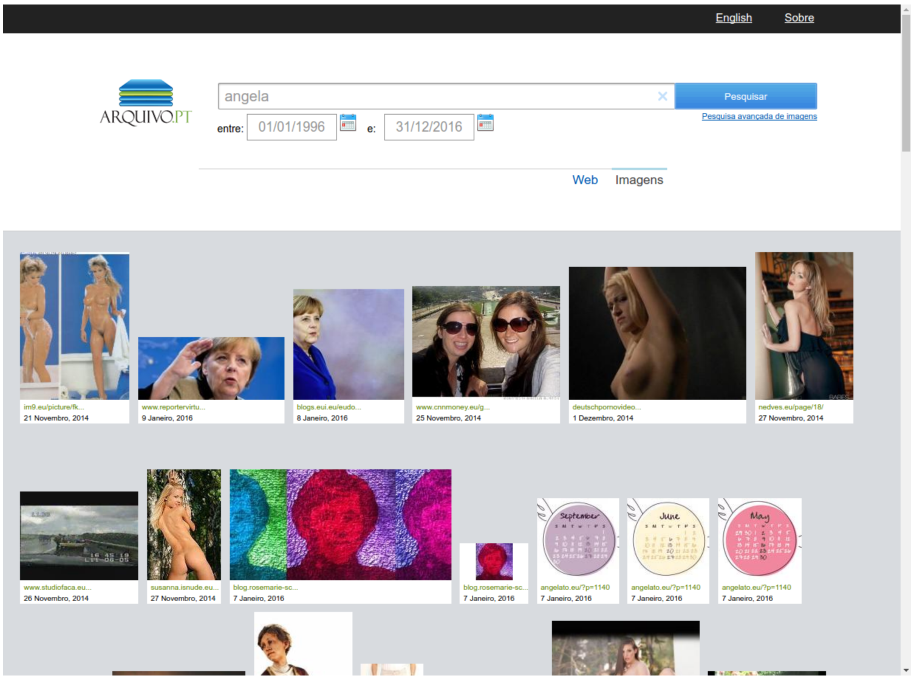
It is important to make available and discoverable this unique and historically valuable information. Without the tools for users to find the desired information, the usefulness of Web Archives is hampered.

To accomplish this, Arquivo.pt provides a full-text search system to all its preserved data. There are significant studies and efforts to improve the Web Archiving Information Retrieval (WAIR) capabilities. For instance, Miguel Costa tries to improve the information search on Web Archives exploring the temporal information that is intrinsic to them [**?**].

Based on this pursuit to provide better searching capabilities to this important information, Arquivo.pt is developing an Image Search service. This service enables image retrieval capabilities to Arquivo.pt preserved contents, presenting an interface in which users can perform queries and the service will try to retrieve images related to the user query.

Figure 1.1 presents the prototype of the Image Search service. A text box is available for users to fill with query terms like *Merkel* and the service displays images related to the user query. On this specific use case, the service retrieves images

[4] http://arquivo.pt



angela

entre: 01/01/1996 e: 31/12/2016

Figure 1.2: Example of Arquivo.pt Image Search problematic content.

about *Angela Merkel*. Each image is clickable, and when a click is performed, the service presents the Archived Web Page from where the images were found.

The Problem

There are huge amounts of visual content on the Web, some part of this visual content is Not Suitable For Work (NSFW) for most users, because it contains offensive or explicit images of violence or pornography. The access to these types of content is particularly critical for children.

The Image Search service retrieves images based on the filename, anchor text and the surrounding text of an image presented on a web page. Therefore, because of the nature of the web, there are no guarantees that the images retrieved to answer a search query will not retrieve an offensive image, even with an apparently not problematic query. For instance, a website that got hacked for web spam can show this behavior.

An example of this problem using the Image Search service is shown in Figure 1.2 where an apparently not offensive query term *angela* was fulfilled on the system and the retrieved results contain content that can be considered offensive to the users.

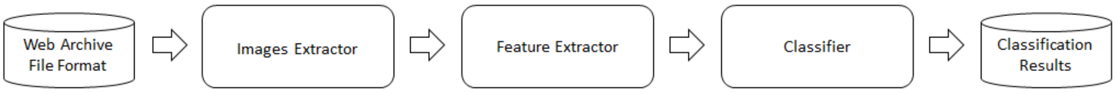


Figure 1.3: Workflow of the classification system.

There are several types of NSFW content, but the type that Arquivo.pt is trying to identify and filter out is pornography.

Detection of NSFW content on the preserved web pages from Internet is challenging because of the scale (billions of images) and diversity (small to very large images, graphic, color images, etc.) of image content.

Project Proposal

The application that will be developed needs to automatically identify this kind of contents, classifying the contents has NSFW or Suitable for Work (SFW). For instance, providing a score from 0.0 to 1.0. Close to 0.0 means SFW content and close to 1.0 NSFW content. These scores can then be used by the Image Search service to filter out the content, for instance providing an option of 'Safe Search' or 'Not Safe Search' to the users searching at Arquivo.pt. This is similar to how Google and other Search Engines work. Another approach is, for example, to blur the thumbnails of this type of content.

Providing image content analysis capabilities to Arquivo.pt would also enable extraction of other information from the images, providing better tools and retrieve results based on this extra information.

The application must be able to perform classification of the images contents in a reasonably time frame. The number of contents available at Arquivo.pt for each collection can be very large. Today, one broad crawl to the top-level domain *.pt* can collect a total amount of 13 Terabytes of compressed information.

The proposed system needs to be modular enough to be used for other use cases by switching the underlying model, for instance, get other type of information from the images. A Web Service interface for real time classification should also be implemented.

An example of a possible workflow is presented at Figure 1.3.

Planned Activities

The work will be split in 4 main activities (Figure 1.4): 1. Datasets building. 2. Solution development - Classification System. 3. Solution development - Scaling Issues. 4. Report writing.

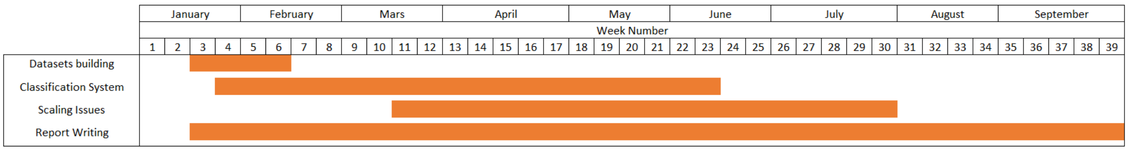


Figure 1.4: Planned activities.

At the first activity will be build a dataset with labeled data from Arquivo.pt. This dataset will be used as ground-truth to measure and compare the possible solutions and fine-tune them with Arquivo.pt data. The main work on this activity will be to label by hand the maximum of images from Arquivo.pt as NSFW or SFW category. The dataset will be a text file, where each row will be composed by an URL to the image resource and the class label.

The development of the solution will start at the next 2 activities. Several solutions will be tested with a small dataset at first to gather insights earlier. Then the solution will keep being developed and consolidated with bigger datasets and possible problems and limitations identified resolved. After the initial development stage, concerns about the scalability of the solution and how to solve them will be analyzed and handled. On this stage the concerns will be to test if the proposed solution can classify in a reasonable time frame the Arquivo.pt data. The questions that this stage will try to address are how much time it would take to classify all the data with the current solution? How can we speed up the process? Is it trading accuracy in exchange of speed plausible? How to serve the models?

The writing of the report activity will take place during the entire project duration. New contents will be written and incremented to the report side by side with the solution development. The last months the efforts will focus at finishing the report.

Arquivo.pt Data Characterization

As of 17 December 2017, Arquivo.pt has a total amount of 4 533 859 612 preserved web files, gathered in 2 160 318 Archive file format (ARC) and Web ARChive file format (WARC[5]) compressed files fulfilling 222.4 Terabytes of disk space storage.

The ARC/WARC file formats specify a method for combining multiple digital resources into an aggregate archival file together with metadata information. Each WARC file is a concatenation of one or more WARC records, each WARC record consisting of a header with metadata information and followed by a content block with the resourced downloaded, such as images, text documents or any type of resource found on the Web.

The information preserved by Arquivo.pt is gathered automatically using web crawlers. This web crawlers ran within a specific scope, and all the information they collect is kept for preservation.

There are four different web crawlers scopes:

- Broad Domain Web Crawlers - (top level domain *.pt* crawl and suggested websites).
- Daily Web Crawls to specific targets - (daily crawl of Portuguese news websites).
- Special Crawls (thematic crawl jobs, like Portuguese elections of *.eu* domain crawl).
- High Quality Web Crawls (specific websites crawl with high quality system and fewer restrictions).

Each crawl has different configurations. For instance, the broad domain crawlers don't download files from the web with more than 10 MB, but the special crawls and high-quality crawls don't have these limitations. Also, they are configured to accept all mime-types found on the Web. For this reason, the type of data that can be found at Arquivo.pt can be very widespread and heterogeneous.

Table 1.1 reports the top 10 mime-types found during a 2017 *.pt* top-level domain crawl. The presented mime-types are reported by the web servers using the HTTP meta information when each resource is collected. It's important to note that sometimes the reported mime-type is wrong regarding the real resource type that the webserver is providing, although these situations are not very common.

Table 1.2 presents the top 8 mime-types regarding image contents that were found during the same crawl. The most common image type is the *jpeg* with 75% of the

[5] ISO 28500:2009 - https://www.iso.org/standard/44717.html

Table 1.1: Measure of the number of resources mime-types collected on Arquivo.pt last broad domain crawl (AWP24).

% Amount	Number of URLs	mime-types
79.60	237966251	text/html
10.03	29997689	image/jpeg
2.45	7328179	image/png
0.77	2305834	application/pdf
0.76	2271770	application/javascript
0.72	2145481	text/xml
0.62	1869902	application/rss+xml
0.62	1861165	application/json
0.60	1820236	image/gif
0.58	1783630	text/css
3.25	1783630	all others

Table 1.2: Mime-types distribution for image type contents.

% Amount	Number of URLs	mime-types
75.22	29997689	image/jpeg
18.37	7328179	image/png
4.56	1820236	image/gif
0.98	389839	image/svg+xml
0.37	147849	image/jpg
0.34	135720	image/x-icon
0.09	38577	image/pjpeg
0.06	22717	image/bmp

total images followed by *png* with 18%. Although only 4% of the images are *gif*, this type can present an extra challenge to classify because the images have an animation.

Because the origin of the images comes from the web, from multiple different websites, the images can have many sizes, different resolutions and any type that is allowed on the web.

Infrastructure of Arquivo.pt

Arquivo.pt infrastructure is composed by 4 key systems to provide its functionality.

Crawler System

The crawler system job is to harvest the web collecting and storing its contents. It uses crawlers such as Heritrix to perform this task. The crawler is configured with a list of starting URLs (seeds) and with a specific scope and budget, a example of configuration is for instance to crawl only URLs from *.pt* top level domain and to download a maximum of 10 000 URLs per host, so it not run forever. It then starts to crawl those seeds, discovering more URLs to crawl and preserve them. Other crawler used by the system is the Brozzler[6], a crawler that uses Chromium instances to render the web pages instead of only fetching the resources representation. These crawlers can discover and preserve more content with higher quality with the downside that it demands more hardware resources, the process is also slower. All contents that are fetched by those crawlers are stored in ARC and WARC file formats to be used by the other systems.

Replay System

The replay system is responsible to reproduce the preserved contents page back to the user trying to provide the a similar experience and navigability as the original web page. The software used for this kind of task is commonly named Wayback Machine. There several implementations of Wayback Machines, for instance OpenWayback[7] written in Java. Arquivo.pt use for this purpose the PyWB[8] a the Wayback Machine written in python. These systems use special indexes named CDX [**?**] that maps the URLs and their position within the ARC/WARC files.

Indexing System

This system is responsible for all the indexing working so that the contents preserved can be search and reproducible. It is composed by a Hadoop[9] Cluster that process the Terabytes of ARC/WARC files stored by the Crawler system and produce Lucene[10] indexes to be used by the full-text Search System and CDX indexes to be used by the Replay System.

[6] https://github.com/internetarchive/brozzler
[7] https://github.com/iipc/openwayback
[8] https://github.com/ikreymer/pywb
[9] https://hadoop.apache.org/
[10] https://lucene.apache.org/

Search System

The Search System is responsible to provide for the users a way to find information at Arquivo.pt. Traditionally, Web Archives only allow to search for a web page through the URL. Arquivo.pt provides more search capabilities allowing users to search through query terms. The system will display preserved web pages that its contents are related with the submitted terms. This system uses a modified NutchWAX[11] implementation and it uses the Lucene indexes generated by the Indexing System.

State of the Art

There are currently available several methods to address this problem, using different techniques to analyze image contents and to detect if it they have Not Suitable for Work content.

Traditional methods use skin-detection algorithms to identify regions of interest and then analyze features of these skin regions to decide whether they are pornographic or not. An implementation of this methodology is the POESIA filter[12], an open source implementation of a skin-color-based filter.

The performance of those methods relies on the accuracy of the skin detection algorithm and the extracted features, usually handcrafted.

Other techniques that showed good image classification results was through *bag-of-visual-words* models (BoVW). These techniques extract from an image a set of visual features represented as words, similar to the *bag-of-words* to document classification, building a vocabulary vector with the number of occurrences of these visual words representing local images features. A classifier that uses this representation is then trained to classify the image content as pornographic or not [**?**].

Recently, Deep Learning has showed state of the art results in almost all tasks of image recognition, more specifically, the Convolution Neural Networks (CNN) has been used widely on image recognition tasks [**?**]. New CNN architectures have continuously been published with improved accuracy of the standard ImageNet [13] classification challenge.

[11] http://archive-access.sourceforge.net/projects/nutchwax/

[12] http://web.archive.org/web/20051030090955/http://www.poesia-filter.org:80/

[13] http://www.image-net.org/

The application of deep learning combines both features extraction and classification, so there is less involvement of a designer in terms of selecting the features or the classifier.

The draw side of these techniques is the amount of training data and infrastructure that they require. Nowadays improved training datasets are available, such as ImageNet. Pre-trained models have been published openly to be used by people without the need to train a Neural Network from scratch. An example of this kind of initiatives is Yahoo! release of an open source model to identify NSFW images, specifically, pornographic images[14].

Open NSFW is a Deep Learning solution released by Yahoo! for classification of NSFW images. The model is published openly to be used for classification by developers, but the images dataset that Yahoo! applied for training is not available, due to the nature of its contents.

They have trained a Residual Network architecture model [**?**] with ImageNet 1000 classes dataset first. Then they replaced the last neural network layer, a 1000 node fully-connected (FC) layer, with a 2 node FC layer and fine-tuned the model to the specific task of NSFW classification using their NSFW images dataset. This model can be reused and fine-tuned for each user specific use case and dataset.

The model is published using Caffe[15] [**?**], an open source deep learning framework developed by Berkeley AI Research[16].

Inspired by Open NSFW other models are being published openly to solve this problem but focusing in architectures that can keep a good accuracy with less hardware resources. An example is NsfwSqueezenet[17] that uses a SqueezeNet [**?**] architecture that is designed to use less parameters and memory, so it can be used with more common hardware.

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

[15] http://caffe.berkeleyvision.org/

[16] http://bair.berkeley.edu/

[17] https://github.com/TechnikEmpire/NsfwSqueezenet