Revealing Historical Events out of Web Archives

Quentin Lobbé

Revealing Historical Events out of Web Archives

Quentin Lobbé[0000−0003−2691−5615]

LTCI, Télécom ParisTech, Université Paris Saclay & Inria, Paris, France
quentin.lobbe@telecom-paristech.fr

Abstract. As the living Web expands, worldwide volumes of Web archives constantly increase, making difficult to identify relevant archived contents. Here we propose an application for detecting historical events out of a corpus of Web archives and based on an entity called *Web Fragment*: a semantic and syntactic subset of a given Web page. The Web fragment has the particularity to be indexed by its edition date instead of its archiving date. We apply our framework on an archived Moroccan forum and witness how it reacted to the Arab Spring at the end of 2010.

Keywords: Web archives, event detection, online migrants collectives

1. Introduction

Since the creation of the Web in the early 90's [2], the loss of the digital content that constitutes the Web itself has been considered a major issue [6]. Whereas related works mainly focus on upstream Web archive acquisition [8], we choose here to perform the exploration of an existing corpus. But apart from the online portal of the WayBack Machine[1], the majority of corpora of Web archives only allows local consultation points, with no remote access or API. In this paper, we first introduce the usage of a new entity called *Web fragment* to guide researchers through Web archives at retrieval time (Section 2). As we think that most explorers of Web archives pursue the discovery of events of some sort, we describe an application called *Web Archive Explorer* (WAE) for detecting historical events (Section 3). We finally use the WAE to understand how the Moroccan forum *yabiladi.com* reacted to the Arab Spring at the end of 2010 (Section 4).

2. Setup

An online migrants collective. As input data for WAE we use the Moroccan section of the e-Diasporas Atlas [3]. The Atlas revealed diasporic communities organized as networks of migrant Web sites connected to each other through hypertext links. But facing the partial or total disappearance of some of the observed Web sites, it was decided to start archiving them. Thus, the corpus was archived from March 2010 to September 2014, covering 254 Web sites[2]. In section 4, we will focus on the Moroccan forum *yabiladi.com*: an old established

[1] https://archive.org/Web/

[2] Publicly available at http://maps.e-diasporas.fr/index.php?focus=map&map=5&section=5

Web site, representing a set of 2.8 million archived Web pages.

Web fragment. A Web fragment is a coherent set of textual, audiovisual or animated contents extracted from a Web page and understandable on its own. It can be a meaningful object like a post inside a forum, a news article, or a comment, and it has the particularity to be indexed by its edition date (the time when it was written). We assume that an original *edition date* will always be more historically accurate than the *download date* of the parent archived Web page. In practice, a Web fragment is the result of the aglomerative clustering of some of the HTML nodes that constitute a given Web page. To extract them, we extend the boilerplate method from [7] and use a combination of vision-based [1] and tag-based scraping strategies [5].

Event detection model. Following our logic of archive exploration, we don't want to detect specific events with expert knowledge, so we avoid patterns and clustering methods. We instead use a threshold-based heuristic [4] within a sliding time frame of one week. We define an event as a detected outlier in the temporal distribution of a set of Web fragments that matches a given keyword. We try explaining the events by finding semantic correlation between the text content of the Web fragments (splited in bigrams) and a set of Morrocan news titles. As *yabiladi.com* is a combination between a news provider and a forum, we choose to construct an index of potential events using the titles and the edition dates extracted from its news section. To sum up, a historical event is the semantic encounter between a well-dated news title and a burst of web fragments.

3. Architecture

We now introduce the components of WAE[3]. We refer to Figure 1 as an illustration of its architecture : **(1)** Our data set is recorded under the Digital Archive File Format (DAFF) formalism that separates the metadata (URL, download date, etc.) from the archived data contents (original HTML content). Our Moroccan corpus results in a 30GB metadata DAFF file and a 300GB data DAFF file. **(2)** The ArchiveMiner component grabs the files using a Java extractor which uploads them into a Hadoop Distributed File System (HDFS). Then a distributed Spark[4] pipeline ingests the HDFS and groups the metadata by time-stable versions and joins them to the data contents. A set of filters based on download dates or domain names are then applied. **(3)** We enrich the original corpus by adding qualitative informations such as the main language (French, Moroccan, Spanish, etc.) or category (forum, blog, media, etc.) of each Web site. **(4)** The FragmentsExtractor component divides each archived Web page into Web fragments (Section 2). Every edition date is translated from a natural language format into a normalized date format. Additionally, the component extracts the

[3] Open source and available at https://github.com/lobbeque/archive-miner and https://github.com/lobbeque/peastee

[4] See http://hadoop.apache.org/, http://spark.apache.org/ and http://lucene.apache.org/solr/

text content, author and title of each Web fragment and joins them with the information inherited from the parent Web pages (URL, download date, etc). **(5)** The ArchiveSearch component indexes the Web fragments into a Solr search engine. A lemmatizer is then applied to increase the accuracy of the full-text facilities. Custom requestHandlers are built to allow different time query strategies. **(6)** The WAE provides two different inverted indexes (Section 2): first the Web fragments extracted from the forum section of *yabiladi.com* and then the events extracted from its news section. **(7)** The ArchiveViz component provides an interface to request the archives by writing a set of keywords and choosing the granularity of the query : Web pages or Web fragments. The results are displayed as a list of documents, illustrated with histograms and a bigrams viewer linked to the events detection system.

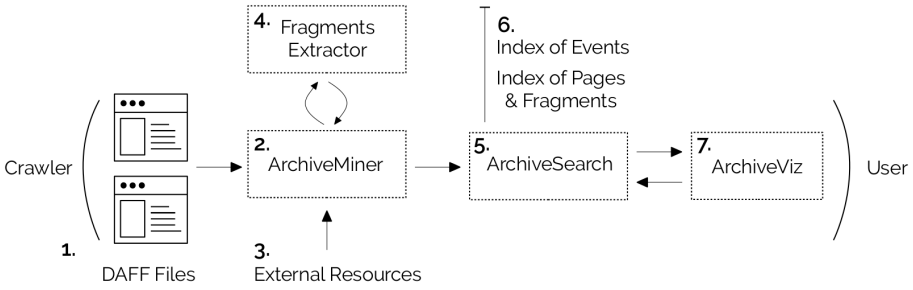

Figure 1: Architecture of the Web Archive Explorer (WAE)

4. Demonstration scenario

We now describe[5] a set of use cases where WAE helps to reveal historical events : **(1)** The user first tries to query the Wayback Machine. But, as it is built on top of a search-by-URL system, the keywords *morroco king* do not match the real content of the archived Web pages. They can only match strict URLs or HTML titles. **(2)** With our system, the user request for *roi* (meaning king in french) and selects *pages* for granularity. The user has to pick up a range of dates to filter the archives. The top ten resulting archived pages are ordered by default Lucene similarity. **(3)** The user now specializes his query by focusing on one of the main author contributions. The 5 most prolific authors are displayed in the facets section of the GUI. **(4)** One can use the first histogram to see the number of matching Web pages by week. Below, there is a line chart displaying the ratio of matching bigrams by weeks. There, the user can follow the evolution of the word *king* in the corpus. The event detection system does not find any

[5] See the accompanying video https://youtu.be/snW4O-usyTM for a peek at the GUI

matching event because the user chose to use Web pages as a scale. But pages are timestamped by download date without regard for any historical accuracy. **(5)** Now the user switches to the Web fragment level, enters the same query and witnesses that she has the possibility to study Web fragments written up to 2003. The event detection system now understands that around the late 2004 an event concerning the king may have focused the conversations on *yabiladi.com*. The system identifies it as an official visit of the Morrocan king to Mexico in November 2004. **(6)** The WAE supports multiple queries (using comma as a separator) for comparison purpose. It displays a coloured line in the n-gram viewer for each query and a union of the resulting fragments in the list below. The user can clearly see a growing percentage of the phrase *Ben Ali* (the former Tunisian president) during the late 2010 that we may correlate to the beginning of the Arab Spring in Tunisia. This assumption is reinforced by a triggered event about the destitution of Ben Ali in January 2011. (7) Finally, more seasonal keywords can be entered in the search box such as the muslim month of fasting *ramadan*. Here the user observes a temporal pattern in the archives that can be explained by the cultural specificity of our Moroccan corpus.

5. CONCLUSION

In this paper, we proposed an application to reveal historical events. We introduced a new entity called *Web fragment* to guide researchers through an exploration of Web archives at retrieval time. We described the architecture of our application and, as a demonstration, we witness how the online community of the Moroccan forum *yabiladi.com* reacted to the Arab Spring at the end of 2010. In the future, we will feed our application with more diverse sets of Web archives (social media streams, blogging platforms, etc.) and work in close collaboration with sociologists and historians to investigate multidisciplinary research questions based on Web archive analysis.

References

1. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: Vips: a vision-based page segmentation algorithm (2003)
2. CERN: The document that officially put the world wide web into the public domain (1993), http://cds.cern.ch/record/1164399
3. Diminescu, D.: e-Diasporas Atlas. Explorations and Cartography of Diasporas on Digital Networks. Ed. de la Maison des Sciences de l'Homme, Paris (2012)
4. Fung, G.P.C., Yu, J.X., Yu, P.S., Lu, H.: Parameter free bursty events detection in text streams. In: Proceedings of the 31st international conference on Very large data bases. pp. 181–192. VLDB Endowment (2005)
5. Jatowt, A., Kawai, Y., Tanaka, K.: Detecting age of page content. In: Proceedings of the 9th annual ACM international workshop on Web information and data management. pp. 137–144. ACM (2007)
6. Kahle, B.: Preserving the internet. Scientific American pp. 276, 82–83 (Mar 1997)
7. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. pp. 441–450. WSDM '10, ACM, New York, NY, USA (2010)
8. Masanès, J.: Web Archiving. Springer, New York (2006)