Diachronic Mining of Web Archives

Facing the online history of diasporas

Stakeholders

TELECOM ParisTech

INSTITUT Mines-Télécom

Author

Quentin Lobbé
Supervised by Pierre Senellart and Dana Diminescu
Télécom ParisTech
Université Paris Saclay
Paris, France

Extending The E-Diasporas Atlas Project

An investigation at the frontier of computer science and social sciences

- **E-Diasporas Atlas :** Pioneer research program in the sociological analysis of on-line activities associated with migrant populations, the e-Diasporas Atlas project revealed diasporic communities and collectives that organize first and foremost on the Web as networks of migrant websites connected to each other through hypertext links.
- **A corpus of web archives** : Crawled and captured during the past 6 years, 70 Terabytes, 10,000 migrant websites distributed and aggregated along 30 diasporic networks (Moroccan, Tunisian, Egyptian, etc.)
- We want to understand the different kinds of **evolutions** experienced by each online migrant community (from extension to reduction, from partition to reunification, etc.) and confront them to their historical and socio-political contexts. The inspiration of our research is the argument that the structure and content of web archives can be **permeable to the effects of shocks** and external events, political and social mobilizations, catastrophes, etc.

Using Automatic Web Archive Enrichment

Through a scalable platform for data processing and extraction

- The archives are uploaded into a **Hadoop Distributed File System** (HDFS). Then a **Spark** pipeline ingests the HDFS where a library extracts fields from the HTML. Finally we index all these fields in a **Solr** search engine in order to make them searchable
- **Content Extraction** : Based on CMS & ad hoc rules, extracting links, users, comments, etc.
- **Event Detection** : Using both variations in topic modeling and text similarity with a set of keywords
- **Quality and Consistency** : Defining archive quality, measuring probability of archive inconsistency based on a given query
- **Querying** : Building a new formalism for querying archives through time, link, page, site, etc.

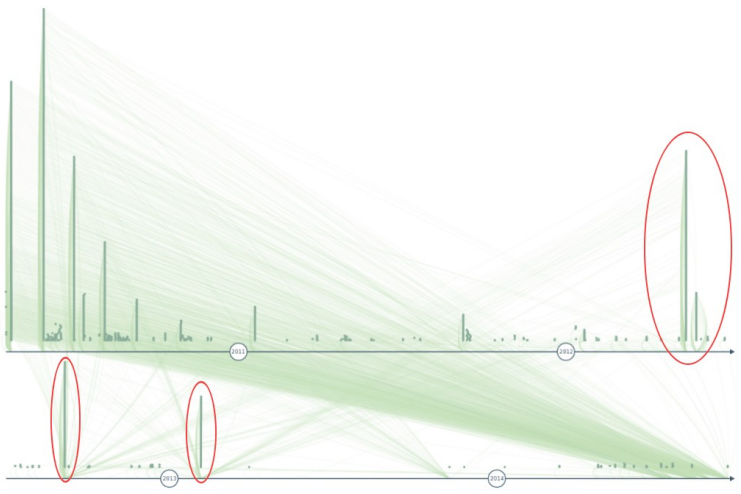

Fig 1 : Out-links distribution of lailalalami.com between 2010 and 2015

To Establish The Premises Of Web Archaeology

Following Tukey: the well known Exploratory Data Analysis

- This is a fundamentally **iterative** process that is deliberately part of a logic of **observation**, discovery and astonishment. EDA's process can be reduce as a loop where each new tool or proposition of **visualization** leads to new research questions, to the definitions of new indicators that will guide the **exploration** and stabilize the ways to represent and see the nature of our set of data and of the contained informations.
- How can we understand the web through its transitions and mutations, from web 1.0 to 2.0, from the web of blogs to the social web, from laptop to smartphone screens, etc ?
- What was the role of web forums during the 2011's Arab Spring ?
- What kind of temporality can we detect on the Web ?
- Is it possible to follow slow trends or topics ?
- Does the web have its own micro-history ?
- Can we study the content of a web page and the gesture behind that content ?
- Can we follow a user through time from a simple commenter status to an influencer status inside a community ?

Contact : quentin.lobbe@telecom-paristech.fr