Monitoring Issues in Digital Public Space - From Data Collection to Issue Mapping

Constance de Quatrebarbes, Antoine Mazieres, Jean-Philippe Cointet

Monitoring Issues in Digital Public Space - From Data Collection to Issue Mapping

Constance de Quatrebarbes, Antoine Mazières and Jean-Philippe Cointet

Cortext-LISIS, Champs-sur-Marne. March 12, 2015.

This article objectives are twofold. Its aim is both to introduce the principles of CrawText: an online data collection tool, and to identify some perspectives on the possible use of those digital traces for the analysis of public issues. CrawText specificity is to enable the construction of controlled corpora focusing on a given topic. Besides it allows to track the evolution of digital territories along time, paving the way toward the monitoring of actors and content dynamics in those public spaces. We will briefly illustrate those analytical perspectives with a case study on current discussions about the Climate Change Conference COP 21 to be held in Paris next Fall.

"We apperceive through our sieves as much as we sieve through our appreciation. We appersieve if you will."[1]

[1] Paul Kockelman. The anthropology of an equation. *Journal of Ethnographic Theory*, 2013

Introduction

Cortext Lab[2] develops tools for collecting, gathering and analyzing data in various contexts, targeting social science scholars at large. Principal motivation is to facilitate data acquisition and data analysis tasks for enabling any user to quickly relate a research questions with various analysis and visualization offered by data analysis algorithms.

[2] `http://cortext.net/`

CrawText has been designed[3] to tackle the raw material constituted by traditionnal web, specifically targeted to monitor, analyze and visualize public issues in digital space.

[3] CrawText is available to download on github under MIT license `http://github.com/cortext/crawtext/`

Mining and analyzing public issue

The traditional web: forums, blogs,traditional websites, online newspapers, etc. constitute a massive raw material to investigate the way a public issues are discussed.

But mining the web to collect data on a specific issue often confronts the researcher with three big technical challenges:

- first the profusion of available informations on the web challenges the traditional methods to collect, store and structure the ressources,
- then, the variety of ressources types and sources (such as blogs, websites, social networks, documents, videos, audio platforms) call for the development of specific information processing tools,

- finally, the essentially dynamic nature of the web requires an *a priori* approach which allows to monitor how digital space change in time.

Taking into account this three obstacles, we have developed CrawText, an open software for:

1. crawling the web in a controlled way, that is only retrieving pertinent webpages regarding a specific issue,
2. archiving the crawl process and its results on a daily/weekly/monthly basis.
3. allowing to extract textual features and mapping relationship between ressource producers, to analyze the evolution of opinions, frames, and actors circulation in the public space and their potential reorganization throught time.

Crawtext: a web crawler for issue mapping

CrawText works as follows. First, the user defines a query representing the issue to be investigated[4]. The choice of the correct query is crucial not to exclude certain actors in the data collection phase. Typically, a political issue could be named differently by two strongly differentiated opinions on the problem. For instance, the recent reform of medical care in the United States is usually named "Obamacare" by its advocates and "Affordable Care Act" by its critics, referring or not to its charismatic carrier, or to its simple legal status. Users can combine or exclude subsets, through the usage of Boolean keywords such as "AND", "OR", "NOT" in order to build the most coherent and balanced territory. The query also allows for orthographical variations through the use of few regular expression operators such as "*" and "?"[5].

Once the query is defined, *Crawtext* initiates its walk from seeds obtained either from a controlled url list defined by the user or from a traditional search engine (Bing). It is also possible to mix manually curated lists with search engines results. Typically the seeds are composed a few tens of urls that will be used as initial webpages for the crawl to start.

Crawtext mimics and systematizes the behavior of a surfer "googling" a query and then clicking here and there on the hyperlinks that she/he considers the most interesting. CrawText works the same way except that it will explore every possible path, as far as it considers every webpage as "pertinent". The crawl extension process is summarized on the right[6]. A scheme (see figure 1) also visually illustrates

[4] Example of possible queries:

```
IAmCharlie AND IAmNotCharlie
nuclear waste
GMO OR organic food NOT insects
(pesticide? OR DDT OR phytosanit*) AND bee*
```

[5] Operators AND OR NOT * ? () " " are shared operators between BING search engine (that may be use for initiating the crawl) and internal matching system

[6]

```
bootstrap:
  url_queue = get seeds
while url_queue is not empty:
  for url in url_queue:
    crawl url
    if query matches url content:
      store url
      add new cited urls to url_queue
```

this snowball process. It can be described as the sequential exploration of new circles. Starting from the first circle populated with seeds, a second circle gathering every pertinent (that is matching the user query) webpages cited by the seeds is built. The third circle is only made of urls that were cited (though an hyperlink) by webpages of this second circle, etc. The total number of circles is called the total depth of the crawl and can possibly be limited by the user if he/she is only interested in the most visible part of the web.

CrawText monitors a web territory on a specific question enlightening analysts with the link structure between information sources or opinions being expressed by any actor or group of actors. In order to make those different level of analysis possible, the following information is being stored for each pertinent url (including seeds provided by search engines): domain, title, textual content, cited links, depth at which the webpage was found and the crawl date.

This last information is capital as CrawText can be automatically launched on a day/week/month regular basis. Whenever a new crawl starts, CrawText simply adds the new trending seeds and update the existing url queue. It checks the content and hyperlinks of existing and new urls in the database and updates them if needed. This features really paves the way toward a dynamical analysis of online public issue.

Crawl Process Analysis

The crawling process itself may provide interesting insights for the understanding of a digital territory structure and the state of the underlying problem connecting those urls. Beyond the size of the corpus which simply echoes the public attention around an issue, the number of steps necessary to collect the entire corpus (i.e. *depth*) may provide an interesting information about how connected the different actors debating the issue are. The analysis of the top seeds provided by Bing search engine and more importantly their stability in time may be of prime importance to track not only the visibility of each actor but also their evolution.

Beyond those simple measures, corpus hyperlinks enable to build a directed graph (either at the url or domain level), which structure could bring informative elements at a global scale: diameter, degree distribution, modularity, or derived metrics. For instance [Shwed and Bearman, 2010][7] proposed to use a normalized modularity score based on the publication citation network to measure the degree of consensus reached in scientific community confronted to disputed questions. At a local level, scores of betweenness centrality, degree, or any structural measure at the node level could help characterize actor

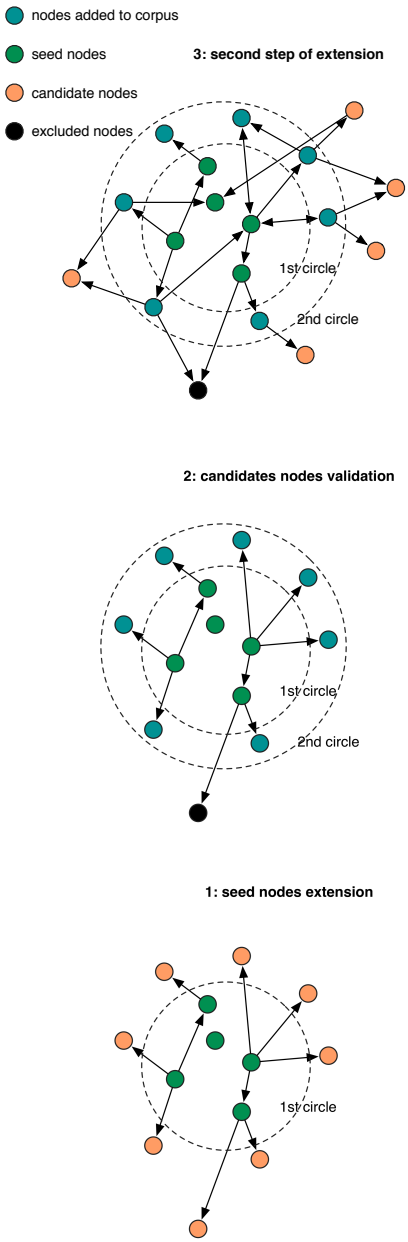

Figure 1: Crawtext snowball process, from down to top

[7] Uri Shwed and Peter S Bearman. The temporal structure of scientific consensus formation. *American Sociological Review*, 75(6):817–840, 2010

strategic positions regarding the issue at stake.

Most issues are characterized by a potentially evolving set of identifiable arguments, references or concepts that frame the issues. The textual content found in each webpage is then a prime material to unveil the public issue "key components". It would be interesting to describe the vocabulary used by actors 1) At an individual level, to frame the problem and 2) from a dynamical point of view, to observe how dominant framing may change in time[Snow et al., 1986][8]. Natural Language Processing provides tools to automatically extract from raw textual content most pertinent terms or phrases[Castellví et al., 2001][9] which are *specifically* being used by certain actors. Other tools such as sentiment analysis detection[Boullier and Lohard, 2012][10] may provide some supplementary information on the emotional involvement of actors and their optimisitic/neutral/pessimistic approach of the problem. At a more general level, considering every webpage as part of a global textual corpus will enable the analyst to draw a semantic network from which main clusters constituting the different frames can be derived. In the next section we show some preliminary results on an actual dataset to illustrate a typical visualization that can be expected from such a workflow.

[8] David A Snow, E Burke Rochford Jr, Steven K Worden, and Robert D Benford. Frame alignment processes, micromobilization, and movement participation. *American sociological review*, pages 464–481, 1986

[9] M Teresa Cabré Castellví, Rosa Estopa Bagot, and Jordi Vivaldi Palatresi. Automatic term detection: A review of current systems. *Recent advances in computational terminology*, 2:53–88, 2001

[10] Dominique Boullier and Audrey Lohard. Opinion mining et sentiment analysis. 2012

Empirical Investigation

France will host the 21st Conference of the Parties to the *United Nations Framework Convention on Climate Change* (COP21) in November and December 2015. Following COP20 at Lima, it will be a very high stake conference as its objective is to produce an international binding climate agreement enabling to limit global warming to below 2 degrees Celsius. Civil society (NGOs, ecologists, social movements) is already mobilized to try to influence the Conference agenda, already creating a rich online conversation about the upcomming event.

Launching Crawtext on the query "COP21 OR COP 21" and stopping the process after two steps we collected 1 566 unique urls which constitute the most visible part of the online informational ecosystem around the event. We then use the CorText Manager[11] to analyse the raw content extracted from html pages. We applied a term extraction algorithm[12] to identify more than 150 key phrases that characterize the public discourses around the COP21. The algorithm uses Python linguistic processing library that we had to configure for a given unique language. This is the reason why we end up with a list of terms essentially made of english phrases such as: "*land degradation and drought*", "*vulnerable countries and populations*", "*food*", etc.

Given this list, a matrix of co-occurrences compiles every joint appearance of two terms in the same document (or more precisely,

[11] `http://manager.cortext.net`

[12] `http://docs.cortext.net/lexical-extraction/`

a co-occurrence between two words is counted when they appear in two sentences which are "close" enough in the original text). From this matrix we derive a *semantic network* which features terms as nodes, and semantic distance as edges strength [Weeds et al., 2004][13]. Louvain community detection algorithm [Blondel et al., 2008][14] is then run to extract the most cohesive subgraphs in the network that form pertinent clusters of terms. The network is visualized figure 2, each emerging topic (cluster) being colored differently and the nodes size scaling with term frequencies. We also tagged those topic with the sources that were their main "contributors" *i.e.* specifically using the vocabulary enclosed in the clusters.

It is beyond the scope of the article to provide a fine-grained description of the various arrangements emerging from this network or to extend the quantitative analysis as our objective was rather to provide a "proof-of-concept" of the possibilities opened by the monitoring and analysis of online issues. But as a matter of introduction of a potential analysis, Figure 2 shows the common terms shared by the datasets and organize it into a graph. On the topic of COP21 organisation,the different actors use a different semantic that enlight their major preoccupation on the coming debate and their future position. We can see that on the blue community focused on the figure 2 they have a very concrete way of expressing their wishes for their conference. It means that the community composed by researchers (Versaille's University), NGO such as ATTAC, and a big french fair on ecological solution have a common discourse and concrete wishes for this Conference. They use a very different lexical and semantic discourse such as *measures*, *best practice* and that this community is linked switching to more governemental concepts such as *agreement government business*. On the other side of the graph the encyclopedic world is expressing itself using general concept of knowledge in which the conference take place such as *law and policy* or *history*, *natural ressources management issues*.

[13] Julie Weeds, David Weir, and Diana McCarthy. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1015. Association for Computational Linguistics, 2004

[14] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008

Conclusion and future work

The general framework that we introduced starting from online data collection to data analysis only proposes some open perspectives on the opportunities opened by such a workflow to explore the dynamics of online public issues. No doubt that the multiplication of case study will help to identify most convincing methods and measures to describe and visualize these kind of dynamics. Once this "toolbox" will be ready, we can expect that some generalizing principles could be drawn from the systematic comparison of several case studies leading to a conceptual model of public issues dynamics.

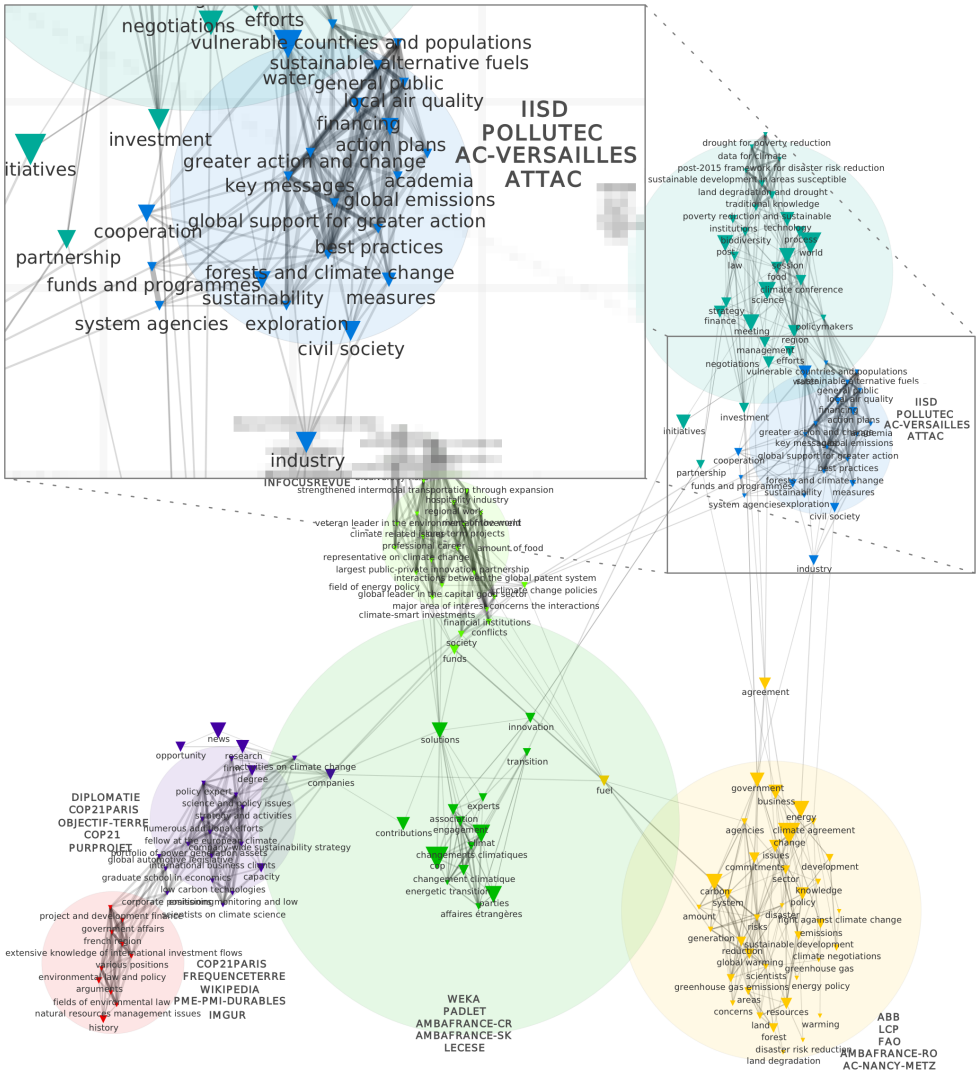

Figure 2: Semantic map: terms are linked when they cooccurence profiles are close, cohesive subgraphs form clusters which are then tagged with most specific sources sharing the same vocabulary (capital letters). One of the cluster is highlighted which terms essentially captures vocabulary around the consequences of climate change for "vulnerable" countries and population regarding "forest", "water", "energy", etc. We observe that this topic gather quite heterogeneous sources among its top contributors (a NGO (Attac), a non-profit non-governmental research organization (IISD), the Versailles education academy (AC-Versailles) which launched a regional project around COP21, and a fair organizer (POLLUTEC) on environmental technologies

The framework presented here allows the researcher to easily flow in various combinations of queries, corpuses and analysis to test, validate, refute, explore hypothesis on controversies and their traces left on the web.

As CrawText allows use to track the change regarding the topic of COP21 we would like to survey the major changes on the frame and the play of actor regarding the debate before, during and after the conference. We would also borrow the biais of langage by using a language detection algorithm to be able to monitor also the evolution of the debate in the different language areas and characterize the main topic evolution from the organization to the conclusion of this conference.