Hyphe, a curation-oriented approach to web crawling for the social sciences

Mathieu Jacomy[1], Paul Girard, Benjamin Ooghe-Tabanou and Tommaso Venturini

Sciences Po Paris médialab

[1]mathieu.jacomy@sciencespo.fr

Abstract

The web is a field of investigation for social sciences, and platform-based studies have long proven their relevance. However the generic web is rarely studied in itself though it contains crucial aspects of the embodiment of social actors: personal blogs, institutional websites, hobby-specific media… We realized that some sociologists see existing web crawlers as "black boxes" unsuitable for research though they are willing to study the broad web. In this paper we present Hyphe, a crawler developed with and for social scientists, with an innovative "curation-oriented" approach. We expose the problems of using web-mining techniques in social science research and how to overcome those by specific features such as step-by-step corpus building and a memory structure allowing researchers to redefine dynamically the granularity of their "web entities".

Introduction

The web has become a legitimate field of study for the social sciences [Ackland 2013]. In qualitative as well as quantitative social sciences, an increasing number of researchers extend their traditional fieldwork towards digital media, as a complement or even as the primary way of collecting data [Rogers 2013]. Social networks like Facebook and Twitter have gained a lot of attention thanks to the possibility of extracting data from their API. These platforms, however, are far from covering all social phenomena on the web. The generic web, although more difficult to harvest and analyze, still represents an important part of actors' presence on the web.

For three years, we worked with sociologists to determine which obstacles prevent them from mining the generic web [Diminescu 2012]. We learned that translating subtle social research questions in terms of web crawling is so difficult that it requires the complete rethinking of how a web crawler works. Our experimentations gave birth to Hyphe, a web crawler using a new approach we call "curation-oriented". It tackles the broad issue of matching research questions with practical uses of a web crawler by allowing scholars to curate their own web corpora.

In this paper we first describe Hyphe and then list a series of issues that we identified as major obstacles to applying web-mining in social sciences, and present the solution as implemented in Hyphe. The first obstacle is the interdisciplinary gap between web mining and sociology. We tackled it by implementing sociological concepts like "corpus" in the crawler's core. The second obstacle is the heterogeneity of the web. We solved it with an innovative memory structure allowing researchers to redefine dynamically the granularity of their "web entities". The third obstacle is the size and structure of the web, and our solution is a step-by-step expansion method.

Description

As all web crawlers, Hyphe allows its user to harvest the content and hyperlinks of a series of web pages. It performs a series of "crawl jobs" demanded by the user. Pages are aggregated as "web entities" (roughly corresponding to websites) in order to reduce the complexity of data. The typical output is a network of web entities to be analyzed through network analysis software such as Gephi [Bastian 2009].

Approach. Hyphe is user-centric. It does nothing unless commanded by the user, and provides a lot of feedbacks for monitoring its processing. We invested a lot of time and energy on interaction design and software engineering to make it usable by people with limited computer skills. Our ambition is to make the whole data management process traceable and understandable by the users, which supports a higher interpretability of the results. We call Hyphe a "curation-oriented crawler" because it uses the concept of "corpus" as a pivot to articulate web-mining constraints with social science methodology, as explained

in the next sections. In particular, we replaced the notion of website with the more flexible notion of "web entity".

Hyphe's demo and source code are available online at: *http://hyphe.medialab.sciences-po.fr*

Structure. Hyphe has a web application, accessible from a HTML5 browser, and a server coded in Python and Java. The application comprises of different screens with lists to monitor data and a live network visualization with a Javascript implementation of the ForceAtlas2 layout algorithm [Jacomy 2014]. The server is composed of: a core, a crawler and a memory structure. The core code in Python Twisted serves the client API and orchestrates the data flows between clients, crawler and memory. We use the Scrapy framework for crawl agents spawning and scheduling (scrapy.org). The downloaded HTML files and related hyperlinks are stored in a queue (a MongoDB database). The queue is consumed by the core which indexes the pages and hyperlinks in the memory structure (a custom Lucene engine). This engine has specific features that support a dynamic aggregation of pages into web entities.

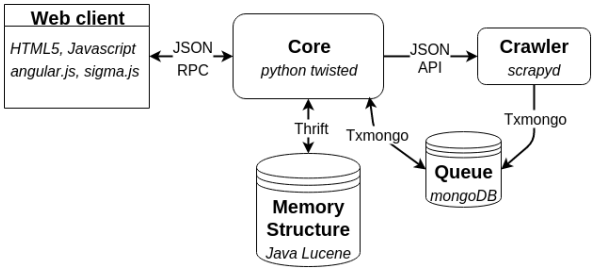

Figure 1: Software architecture of Hyphe

Related works. Considering their usage, web crawlers may be divided in three categories: archiving crawlers, harvesting crawlers and social science crawlers. Archiving crawlers can crawl and index data over time and reconstruct web pages so they can be seen as they were. Heritrix [Mohr 2004], conceived by Internet Archive is the most famous and is used in more research projects than any other crawler, but requires specific computer skills. Harvesting crawlers like UbiCrawler [Boldi 2004] and iRobot [Cai 2008] aim at building data sets with maximal efficiency. They are turned towards industrial needs and focus on performance and scalability to the detriment of methodological quality and ease of use. A few social science crawlers also exist: Issue Crawler [Rogers 2010], SocSciBot [Thelwall 2009] and Voson [Ackland 2006]. These crawlers are comparable to Hyphe in usability, but none of them propose features for corpus curation, and none of them can handle different levels of granularity.

Implementing the concept of corpus

Issue. When considering the web as a field of investigation, social scientists expect their data to have certain methodological properties that make it a *corpus*, something more complex than the simple concept of dataset [Wynne 2005]. It aims at representativeness and has important properties like selection criteria, balance or homogeneity. It can be found everywhere in social sciences, from survey panels in quantitative sociology to document space in archive investigation. When operationalizing their research questions, social scientists use the concept of corpus as a tool to frame the risk of methodological biases and errors. When scholars use a dataset in place of a corpus, they lose the crucial validity metrics that support interpretability.

Solution. Hyphe has corpus-related features that social scientists understand and leverage to monitor and control their methodology. The table below illustrates which methodological properties can help solving which web-mining problems. Their Hyphe implementations are discussed in the following sections.

methodological property	*Hyphe implementation*	*Web-mining issue*
Sampling: How to group web pages?	Dynamic granularity of Web entities	"Website" is not a technical definition [Berners-Lee 2005]
Selection criteria: delineating the field	Step by step expansion method	Topic drift [Chakrabarti 2002]
Representativeness: evaluating how much is missed	Degree-based exhaustiveness	Power law degree distribution [Albert 2002]

Dynamic granularity of "web entities"

Issue. The research objects investigated by social scientists do not necessarily meet the web pages organization (i.e. the URL hierarchy). A person may have several blogs, and a blog several authors (not to mention platform-specific profiles and website nesting). When a social scientist studies an actor's digital traces, he needs to crawl different sources at different levels of granularity.

Solution. We implemented a dynamic memory structure allowing users to redefine the granularity for any domain and at any moment, even after crawling and indexing. Hyphe adds a level of aggregation on top of web pages called "web entities". This aggregation is based upon a morphologic analysis of URLs: URLs are split into stems;

those stems are ordered from genericity to specificity which creates a hierarchy of URL names (see Figure 2). A web entity is defined as a set of references to this tree of URL stems. Web entities are thus a flexible web page aggregation system based on URL hierarchy.

By defining his own web entities, the researcher can group web pages according to his research objects, actors and questions. Though the solution is reasonably simple in practice, its implementation required a totally new indexing engine. Our memory structure stores the web pages network level and the web entity definitions. Then it can dynamically infer the web entities network by aggregating the web pages. This system allows changes in web entities' definition without having to recrawl or reindex.

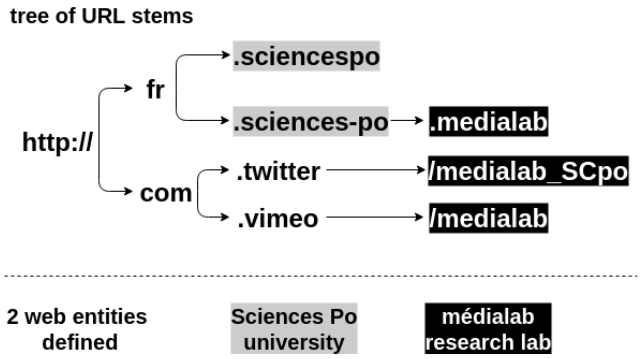

Figure 2: web entities are pointers to the URL name hierarchy.

Examples. We observed that different researchers were able to tune the system in various ways, thus producing different research protocols. Some researchers used this system to merge, into a single web entity, different facets of people's presence on the web: personal blog, Twitter and Facebook accounts… The resulting analysis is then very similar to social network analysis.

Other researchers used web entities to isolate a specific topic across the web. For instance separating controversial articles about "immigration" from the rest of the newspaper websites to differentiate how they are cited and by whom.

Step-by-step expansion method

Issue. Being a "small world" network, the web has both a huge size and small diameter [Broder 2000]. Such structure is produced by the power-law distribution of connectivity. The most notorious websites like Wikipedia or YouTube are hugely cited and impact crawling by capturing the spider and causing a snowball effect where the number of collected pages grows exponentially with the distance to the starting points.

Solution. Hyphe supports a step-by-step expansion method where web entities are curated before being crawled so that hyperconnected websites are not added to the corpus unless necessary to the research, thus keeping topic drift in check. The curation is done through a system of exclusive web entity statuses:

- *DISCOVERED* when the web entity is detected by the crawler, thanks to hyperlinks, but the user has not taken any decision;
- *IN* when the user has explicitly accepted it;
- *OUT* when the user has explicitly rejected it;
- *UNDECIDED* when the user has suspended his decision, often because corpus selection criteria apply ambiguously to it.

Excluding an entity doesn't delete it. We keep a trace of all curation decisions allowing the user to change their mind and the selection criteria to be traceable.

To optimize the curation, the interface dedicated to corpus expansion displays the list of *DISCOVERED* entities from the most to the less cited. The user can then reject (set to *OUT*) the hyperconnected websites while accepting (setting to *IN*) and crawling the relevant ones. This approach allows the user to reject pages with higher degrees of connection and resist the attraction of the hyperconnected nodes and ultimately preventing the snowball effect.

Our curation-oriented approach required a change in the traditional crawling process. Instead of a global page-depth setting akin to IssueCrawler, Hyphe allows each web entity to be crawled at a specific page-depth. The global network properties like diameter are controlled by step-by-step expansion instead of snowball expansion.

Hyphe enables the user to estimate the exhaustiveness of the corpus by exploiting the hyperlink structure of the web. The step-by-step expansion method leads to a systematic analysis of the most cited *DISCOVERED* web entities, up to a certain threshold. Provided that every accepted entity has been properly crawled, that threshold conveys the approximation of exhaustiveness that has been performed in the corpus. The lower the threshold (relatively to the in-degree distribution), the better the approximation of exhaustiveness. It helps researchers manage the tradeoff between corpus quality and time spent on prospection.

Example. The user has a corpus of 200 entities on a given topic, all properly crawled. He accepted or rejected all *DISCOVERED* entities with more than 3 citations (from the corpus). Then Hyphe ensures that every entity cited 4+ times in the corpus has been evaluated. If an important entity is still missing from the corpus, the degree-based exhaustiveness guarantees that:

- either it was poorly cited (and therefore probably not really important);
- or it did not match the selection criteria.

Discussion

Hyphe is dedicated to the generic web, a field richer than specific platforms like Facebook or Twitter exploited in API-driven research. Hyphe is not intended to replace other tools but to unlock new methodologies. Though it is not as efficient as API-driven tools, it can gather data across different platforms, including small and API-less platforms, which is crucial when studying marginal communities or other forms of weak signal.

The curation-oriented approach focuses on quality rather than performance. Industrial crawlers face similar issues but implement opposite tradeoffs because they focus on performance rather than quality. As a major drawback of our approach, curation is time-consuming. Time is a cost that social sciences are more willing to pay in order to ensure reliable, interpretable results.

Hyphe has supported actual research for only a year and most of its results have not been published yet. We believe that a more systematic evaluation of curating web corpora will be possible once a critical number of publications have been reached.

Future work

Hyphe is under constant development and our road map includes tagging features in near future. We are also working on a dedicated browser where users will be able to select and tag web entities while reading their content, which is identified by researchers as a key improvement. We are also experimenting different ways to couple content analysis with links analysis, and many social scientists are looking forward to using such possibilities. We also plan to consolidate user feedback once we reach a critical number of publications, by systematic survey or other means.

Conclusion

In this paper we have presented Hyphe, a curation-oriented web crawler dedicated to social sciences. We listed a series of problems that scholars meet when they use existing crawlers, namely the difficulty to translate their research questions into the technical environment of web-mining; the necessity to control selection criteria; the heterogeneous granularity of actors' presence on the web; the information overflow caused by the web's small-world properties and the evaluation of sampling and exhaustiveness of the data. We detailed our solutions and how they relate to the more general idea of curating a web corpus. We argued that this particular approach leads to a higher level of ground truth and more satisfactory fieldwork for sociologists. We hope that getting more perspectives will allow a better formalization of this approach, both in terms of web-mining and social sciences. We believe that technical features and methodological properties are two sides of a same coin, and that new experiments unlocked by Hyphe may lead to findings in both web studies and sociology.

References

Ackland, R., et al. 2006. *VOSON: A Web services approach for facilitating research into online networks.*

Ackland, R. 2013. *Web social science: Concepts, data and tools for social scientists in the digital age.* Sage.

Albert, R., Barabási, A.-L. 2002. *Statistical mechanics of complex networks.* Reviews of modern physics 74.1: 47

Bastian, M., Heymann, S., Jacomy, M. 2009. *Gephi: an open source software for exploring and manipulating networks.* ICWSM 8: 361-362.

Berners-Lee, T., Fielding, R., Masinter, L. 2005. *Uniform Resource Identifier (URI): Generic Syntax*, STD 66, RFC 3986, DOI 10.17487/RFC3986.

Boldi, P., et al. 2004. *Ubicrawler: A scalable fully distributed web crawler.* Software: Practice and Experience 34.8: 711-726.

Broder A., et al. 2000. *Graph structure in the web*. WWW9 / Computer Networks, 33(1-6):309–320

Cai, R., et al. 2008. *iRobot: An intelligent crawler for Web forums*. Proceedings of the 17th international conference on World Wide Web. ACM.

Chakrabarti, S., et al. 2002. *The structure of broad topics on the web*. Proceedings of the 11th international conference on World Wide Web. ACM.

Diminescu, D. 2012. *E-Diasporas Atlas: Exploration and Cartography of Diasporas on Digital Networks*. Éditions de la maison des sciences de l'homme.

Jacomy, M., et al. 2014. *ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software.* PLoS ONE 9(6): e98679. doi: 10.1371/journal.pone.0098679

Mohr, G., et al. 2004. *Introduction to Heritrix.* 4th International Web Archiving Workshop.

Rogers, R. 2010. *Mapping public Web space with the Issuecrawler.* Digital cognitive technologies: Epistemology and the knowledge economy: 89-99.

Rogers, R. 2013. *Digital Methods*. Cambridge Mass.: MIT Press.

Thelwall, M. 2009. *Introduction to webometrics: Quantitative web research for the social sciences.* Synthesis lectures on information concepts, retrieval, and services 1.1: 1-116.

Wynne, M., 2005. *Arts and Humanities Data Service*. *Developing linguistic corpora: A guide to good practice* (Vol. 92). M. Wynne (Ed.). Oxford: Oxbow Books.