Archiving news on the Web through RSS flows. A new tool for studying international events

Marta Severo, Laurent Beauguitte, Hugues Pecout

Archiving news on the Web through RSS flows. A new tool for studying international events

M. Severo (Université Lille 3), L. Beauguitte (CNRS/UMR IDEES), H. Pecout (CNRS/GIS-CIST)

Introduction

Media news is more and more used in academic research as data for social sciences' studies. News allows detecting and monitoring events from social movements to natural catastrophes. In the last decades, several scholars have worked on the definition and identification of media events (Dayan and Katz, 1992; McCombs and Shaw, 1972). Among them, some investigated cross-national media coverage of different types of events (Galtung and Ruge, 1965; Koopmaas and Vliegenthart, 2011) and focused on mechanisms that may explain diffusion of media attention. One of the main issues related to this type of studies is that media data can be retrieved only in commercial databases such as DowJones Factiva. The use of these databases is not only expensive, but it also raises several issues. Yet, recently, with the emergence of the Web 2.0, media (especially newspapers) publish news directly on the Web, mainly free of charge, and often they provide free push services such as RSS feeds to get real-time information to the reader. Our hypothesis is that newspapers' RSS flows can be an alternative source of information for media studies. This paper will be organized in three parts: firstly, we focus on the problems raised by traditional media databases; secondly, we present the features and the expected advantages of using RSS feeds; and finally, we present the results of the ANR Corpus Geomedia project (2013-2016, http://geomedia.hypotheses.org). In this project, we build a database storing RSS flows associated with articles published in one hundred newspapers in different parts of the World in order to extract two types of information, flows among countries and international events. We briefly present the methodology used for building this archive, its main features and possible uses of RSS data for studying international events with a multi-dimensional viewpoint.

1. Media data from identifying of international events: the commercial databases

In the last decades, several scholars have worked on the definition and identification of media events. From the definition of so-called 'news values' (Galtung and Ruge, 1965) and of the media agenda setting (McCombs and Shaw, 1972) to the concept of media event defined by Dayan and Katz (1992), studies focused on the different factors and mechanisms that affect the media attention and may explain media coverage of specific events. Among them, some scholars investigated cross-national media coverage of different types of events, from human events, such as social movements (Herkenrath and Knoll, 2011), to natural events such as earthquakes and tsunamis (Koopmaas and Vliegenthart, 2011). Such a research object is intriguing for its interdisciplinary and multi-dimensional nature. As underlined by Koopmaas and Vliegenthart (2011), studying diffusion of news imposes to question the relationship between media and extra-media data and the interactions among the spatial, temporal and thematic dimensions of events.

Studies on media coverage of events are usually based on the analysis of broadcast traditional media such as press and television. In this article we focus on studies based on the press. As known (Earl *et al.*, 2004), the use of newspaper data for studying events such as collective action may raise several critics concerning the data collection and the selection and description bias related to articles' content (McCarthy and McPhail, 1996), yet they can be very useful for research on social phenomena because they may allow studying some aspects of these events that scholars could hardly approach with other methods. We can distinguish more qualitative studies that focus on the description of events provided by specific media and other more quantitative studies that seek to detect general patterns to the identification of events. These two types of studies are partly unsatisfactory. The first type of studies is usually based on manual analysis of a corpus of media discourse (usually press articles). These analyses are very time-consuming and therefore are applied on very limited data samples in space and in time (one country, one event and a limited period). Conversely, quantitative analyses rely on large sets of data obtained through the access to commercial database for analysing phenomena on a broader temporal and geographical perspective (longer periods, several media).

In both cases, one of the more complicated issues of the research about media coverage concerns how to retrieve media data. Usually, they can be retrieved only in commercial databases such as DowJones Factiva, LexisNexis or Europresse. The use of these databases is not only expensive, but it also raises several technical and methodological problems. From the technical viewpoint, the extraction of data is very limited (usually less than 100 items simultaneously) for copyright and commercial reasons and consequently the creation of large samples is quite time-consuming. From the methodological viewpoint, these databases are usually not very accurate and transparent concerning the presence of the sources (gaps in the database are not clearly signaled) and the attribution of keywords to the articles. For these reasons, the necessity of finding other kinds of media data has emerged as a priority in the research related to international media events.

2. The RSS feeds

Using the RSS feeds provided by the online version of worldwide newspapers appears as an alternative source in studies about media coverage of international events. RSS feeds are files (in XML) updated regularly by websites that give concise information about the publication of new content on a website. A feed consists of several items and each item is constructed according to standards (RSS 1.0, RSS 2.0 and Atom) and is characterized by a number of mandatory fields (date, title, description, hyperlink). For newspapers, each item corresponds to an article and consists of a title, which generally coincides with the title of the article, a summary, which can be written by the journalist or match the beginning of the article, and the hyperlink to the newspaper article. An item usually consists of only two or three lines of text.

RSS are supposed to have three great advantages: they are freely accessible, so they may be archived and tagged without limits; they have a quite homogenous structure, so they are easily comparable; they are generally provided as the news is ready and they can therefore be suitable for a real-time analysis. However, RSS are still little studied. Until now the majority of studies have investigated the technical aspects of this format of syndication. They concern particularly the different standards (Hammersley 2005; Hammond *et al*., 2004). Few studies focused on specific contexts of use such as education (Duffy and Bruns, 2006) or science publishing (Hammond *et al*., 2004). Recently, certain studies started to investigate the informational value of RSS. For example, Marty *et al*. (2012) investigate the informational plurality on the Web by analysing the contents of hundreds of websites' RSS feeds.

If the literature is reduced, we can notice that several scholars have recently based their empirical projects on RSS feeds without really questioning and justifying that choice. In France, the ANR has funded several projects in this direction: the Transmedia Observatory (www.otmedia.fr/) that aimed to trace media events on all broadcast media (web, press, radio and television); the project ChronoLines (http://chronolines.fr/) that aimed to model "Event-based Chronologies" by treating news; and the project Webfluence (http://webfluence.csregistry.org/) aimed to study, model and rebuild information flow dynamics by differentiating the flows produced by the established media (traditional media and pure players) and those produced by different communities of interest in the French blogosphere. Similarly, the European Union funded the Glocal Project (www.glocal-project.eu) to identify and model the local and global events from different media data corpora, and the platform EMM NewsBrief (http://emm.newsbrief.eu/) that collects news in real-time from multiple sources in 43 languages and identify in real time the most important stories (top stories).

Finally, there is an important issue to keep in mind. In 2013 Google Reader, the main RSS aggregator, was closed and with this fact, many announced the death of RSS. Yet, they are still alive and diffused. Even if a lot of people today privilege other tools to be informed like the Facebook wall or push news on the mobile phones, main newspapers still guarantee this format.

3. The Geomedia project

We build a database storing RSS feeds associated with articles published in one hundred newspapers in different parts of the World and we extract two types of information: flows among countries and international events.

One of the main aims of the Geomedia corpus project is to identify flows among countries (which are spaces of interest according to the media localization?) and international events (can we distinguish between inter-national, regional and global events in our corpus?). Due to the size of the gathered corpus, automatic procedures are required. A first step was to identify elements of information we were able to identify in our items. With a low margin of errors, identifying country's names was possible. Conversely, except for worldwide known leaders (Obama, the pope, Putin), identifying persons was not possible. Our corpus being composed of 100 flows from 40 different countries, each newspaper gives variable levels of precision regarding quoted persons. Creating a dictionary of main political leaders on a World scale is a daunting task, as it would need permanent updates. Identifying countries was easier and thematically more relevant: it allows controlling differential hierarchies regarding places quoted according to newspapers locations. The only issue was to create a complete dictionary to tag properly all items without creating false positive identification. Our objective is to get a margin of error below of 5%. The main difficulties were caused by the plural ways to qualify the United States, and, more marginally, the way a given newspaper names its proper country.

Regarding events, its lexical spectrum needs to be as reduced as possible to be identified in our corpus. For instance, extracting all items regarding Ebola is easy as there is no ambiguity regarding terms used in the four languages of the corpus. Reversely, extracting all items regarding terrorism for instance would be more tricky as many terms are used to qualify the phenomena, new actors and words emerge on an unpredictable basis (*Charlie-Hebdo* suddenly became synonymous of a terrorist attack on a world scale on January 2015), and some used words are polysemous (an attack in an item could be a heart attack striking a personality as well as a terrorism attack) which would cause many false positive.

Quantitative treatments already launched on the corpus regard three main directions:

1. time persistence of given events (Ebola, Syrian war – Giraud *et al*., 2013 -, Ferguson riot, Wukan protest – Severo *et al*., 2012);
2. hierarchy of places amongst newspapers and time;
3. co-occurrences of places in RSS items.

The first direction tries to model news flows from a thematic and a spatial perspective and aims to investigate the following questions: when and how a given event spread across newspapers on an international scale? Is it possible to identify patterns and trends of diffusion? Which are the barriers (cultural, linguistic, political) to these diffusion processes? When does media saturation appear?

Hierarchy amongst places in international news is a well-known topic of investigation and we would like to check if, despite changes in information production due to the Internet, rules proposed nearly fifty years ago by Galtung and Ruge remain valid on a world scale.

Examining states co-occurrences was, to our full knowledge, barely examined in previous studies. However, our first tests (Beauguitte and Severo, 2014) seem to indicate that several teachings regarding world structure could be brought by such investigation. For instance, if main powers are often quoted alone in international news (an American event often becomes a world event), least developed countries are never quoted alone, except in case of major crisis.

Conclusion

Two main aspects need further research: a technical one, a thematic one. From the technical perspective, at least two main issues appear when using RSS feeds for news archiving: (1) limited coverage of the feeds – as feeds only contain materials selected by their editors and not the entire news content, and (2) heterogeneity of the feed content and structure.

From a thematic perspective, tests initiated in the Geomedia project – events detection, co-occurrence flows, States hierarchy - are just a starting base as our database will be made available to researchers that will be able to test their own hypotheses. The format of the deliverable is not defined yet, even if the choice of a machine-readble format (XLM, JSON etc.) appears as a relevant option. A R package encapsulating data gathered, dictionaries (places, given events and topics) and analysis programs is another option currently studied.

References

Beauguitte, L. & Severo, M., 2014, *Do International News Reflect World Structure? A Network Approach*, Paper presented at the First EUSN, Barcelona, http://geomedia.hypotheses.org/176

Dayan D. & Katz E., 1992, *Media Events: The Live Broadcasting of History*, Cambridge, Harvard University Press.

Duffy, P.D. & Bruns, A., 2006, The Use of Blogs, Wikis and RSS in Education: A Conversation of Possibilities, In *Online Learning and Teaching Conference 2006*, 26 Sep. 2006, Brisbane.

Galtung, J. & Ruge, H.M., 1965, The structure of foreign news, *Journal of Peace Research*, 2(1): 64-91.

Giraud, T. *et al*., 2013, Identification of international media events by spatial and temporal aggregation of newspapers rss flows. Application to the case of the Syrian Civil War between May 2011 and December 2012, *Procceding ECTQG 2013*, Paris.

Glance, N. S., Hurst, M., & Tomokiyo, T., 2004, BlogPulse: Automated trend discovery for weblogs. In *Proceedings of the 13th international WWW conference* – workshop on weblogging ecosystem – aggregation, analysis and dynamics.

Grasland C., Giraud T. & Severo M., 2012, Un capteur géomédiatique d'événements internationaux in *Fonder les Sciences du Territoire*, Karthala, Paris, p. 184-190.

Grossnickle J., Board T. & Vrian Pickens, M. B., 2005, *RSS - Crossing into the Mainstream*, http://www.egoboss.com/pdfs/Yahoo_RSS_whitePaper.pdf

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A., 2004, Information diffusion through blogspace. In *Proceedings of the 13th international WWW conference,* 491–501.

Hammersley, B., 2005, *Developing feeds with RSS and Atom,* Sebastopol, CA: O'Reilly.

Hammond, T., Hannay, T., & Lund, B., 2004, The role of RSS in science publishing: Syndication and annotation on the web, *D-Lib Magazine*, 10(12), http://www.dlib.org/dlib/december04/hammond/12hammond.html.

Herkenrath, M., & Knoll, A., 2011, Protest events in international press coverage: An empirical critique of cross-national conflict databases, *International Journal of Comparative Sociology*, 52(3): 163-180.

Koopmans, R. & Vliegenthart R., Media Attention as the Outcome of a Diffusion Process—A Theoretical Framework and Cross-National Evidence on Earthquake Coverage Ruud and Rens, *European Sociological Review*, 27(5): 636-653.

Marty E. *et al*., 2010, Variété et distribution des sujets d'actualité sur Internet. Une analyse quantitative de l'information en ligne, *Mots. Les langages du politique*, 93: 107-126.
Marty, E. *et al*., 2012. Diversité et concentration de l'information sur le web. *Réseaux*, 176(6):27-72.
McCombs, M.E. & Shaw, D.L., 1972, The Agenda-Seting Function of Mass Media, *The Public Opinion Quarterly*, 36(2): 176-187.
Severo M., Giraud T. & Douay N., 2012, The Wukan's protests: just-in-time identification of international media events, in *Proceeding of Workshop Just-in-Time Sociology, SocInfo international conference*, 4-6 december, Lausanne, Switzerland.