The Science in the Media Monitor (SMM) system

Federico Neresini
University of Padua, Italy
Observa Science in Society
federico.neresini@unipd.it

Andrea Lorenzet
University of Padua, Italy
Observa Science in Society
andrea.lorenzet@unipd.it

Abstract

In this talk we present the features of Science in the Media Monitoring (SMM). a system for automatically tracking science and technology issues in the digital media. The system's architecture is able to be adapted to different languages and topics, and to collect different sources of information (like online newspapers, blogs and others). At the moment our database is collecting contents from mainstream Italian online newspapers starting from 2008 (with a consistent and coherent database from 2010), and from 2013 of several Italian blogs and from three International English and French newspapers, but it can be expanded for new and other sources, like for example Twitter and social media. Texts collected are parsed and cleaned from html, automatically tagged by a classifier which works through a thesaurus of weighted keywords and stored in the database, that is thus able to make automatic distinctions of relevant and not relevant contents in relation to the selected topics (for example contents regarding science and technology, but also more specific topics like food safety, nanotechnology and others). The systems has a user interface for Boolean search in its database, with charts and word clouds, that give results based on the metadata, like the length of each entry, the source, date, link and other basic

information useful for analysis. Results can be exported by the user in txt and excel formats.

Introduction

The project here presented regards the development of a software infrastructure for the automatic analysis of on-line mass media coverage of science and technology issues. The system, called Science in the Media Monitor (SMM) regards at the moment the analysis of the Italian media, but its architecture is projected in order to allow the collection and the analysis of large quantities of textual data from any web source, and can thus be adapted to the analysis of media coverage of science and technology issues in different type of media, languages, countries, and cultural contexts.

The system constitutes the output of a multidisciplinary project at Observa Science in Society, featuring close collaboration between experts in the field of Public Communication of Science and Technology, statisticians and computer scientists working on database construction, information retrieval, search engine optimization, and on the construction of a series of dedicated indicators for assessing and studying the presence of S&T topics and issues in the media.

The architecture of the software is based on two main elements: a database system with an acquisition function that collects data from the RSS of selected content sources (at the moment online newspapers and blogs) and stores them in a relational open source-based database, and an interface/search engine module that indexes texts according to a set of pre-defined metadata, with a function of tagging regarding the relevance for science and technology topics, and a GUI interface for data consultation, download and analysis.

The overall goal of the project is to develop a methodology and a data structure in order to analyze and monitor media coverage of science and technology issues and topics within the perspective of big data: that is to allow content analysis of mass media coverage in order to build a sound and reliable structure for conducting PCST analyses of the presence, the tone, and discourses regarding science and technology in web contents.

As already specified, the project is also aimed at the definition of a series of indicators for mapping the cultural authority of science and technology in the public

sphere (Bauer, Shukla, and Allum 2012, Bauer and Bucchi 2008, Neresini and Bucchi 2012). This approach is part of a wider framework for the mapping of "cultural authority of science", that is to map public opinion processes regarding science and technology within a normative RRI research framework (Responsible Research and Innovation), in order to feed both scientific understanding and policy/regulatory action on controversial subjects.

Within this approach, the public sphere is considered a three-fold arena within which public opinion on science and technology issues is being formed, made up of regulation, mass media, and everyday conversations (Bauer and Gaskell 2001). This can be particularly interesting for the analysis of public controversies, that is public debates on science and technology that allow us to discuss uncertainty and institutionalization as a key feature of the interaction among heterogeneous social actors in the public sphere, as it was clearly exemplified in the case of the biotechnology movement:

> "The public sphere modelled as a tripartite arena allows us to characterize the changing symbolic environment of the biotechnology movement and, at the same time, the potentials for organizational and institutional learning that may result from public resistance."
> (Gaskell and Bauer 2001, p.16)

Within this frame the research work regards the definition of indicators determining the "cultural authority of science" in different cultural contexts, thus avoiding ethnocentrism in their definition. The CASI system covers three levels of science culture: individual attitudes, media attention, and public engagement opportunities.

The SMM project was devised in order to act at the level of media attention on science and technology issues, to store information on a database and make it available with a speed and easiness never known before to scholars working on PCST issues and media analysis of Science and Technology.

In this paper we describe the technical infrastructure of the system that has been devised by the Observa SMM team and two basic indicators that can be used to track in

quantitative terms the presence of science and technology in the media: salience and visibility, leaving aside the description of more complex and theoretically informed indicators for reaching these goals in relation to several different dimensions.

The SMM technical infrastructure methodology has been designed in order to comply with the new field of big data analysis, and thus according to a new set of principles for advanced and breakthrough automated media and content analysis of texts.

The first principle regards the use of the full data set for the analysis, instead of using a sample of data created ad hoc. Thanks to the availability of large amount of text and the possibility to store them in databases, we now can abandon the logic of sampling data and periods, and get a more broad view of phenomena thanks to the availability of these large datasets for content analysis, within the frame of big data analysis (Mayer-Schönberger, and Cukier 2013). The objective of the SMM system is to collect and make available for PCST and CASI research data and information regarding mass media and science and technology coverage.

The second principle regards the use of new measures, and in particular to basis the analysis on the logic of correlation, that is to establish connections within variables by using large amount of data and a "scalable" approach: that is to develop a use of indicators that can act in a comparative research design.

Software architecture, metadata, and the application of classifiers

The architecture of the SMM software is organized as described in **figure 1**. HTML documents are retrieved by the system through RSSs feeds thanks to a procedure of crawling and scraping. Documents' text is cleaned and inserted in an xml file together with a series of standard metadata that comprises the following elements:

- Identifier (the ID of the entry, automatically assigned to each document entry)
- URL (the URL at which the document was retrieved)
- Document hash (a string of automatically generated alphanumeric characters that identifies the document and that is generated on the basis of the URL)
- Chosen date (the date of the download according to the selected time zone)
- Chosen title (the title of the article/blog post)
- Document length (length of the document)

- Keywords (the set of keywords detected among the list provided)
- Multipliers (the set of generic keywords acting as multipliers in the scheme)
- In homepage (a selector that tells if the article has been published or not in the home page)
- Document collector (the subset of database from which data come from)
- Source (the newspaper/blog from which the document is retrieved by the system)
- Source URL (the URL of the source)
- Source set (a criteria to cluster different sources in one coherent subset: example "core set newspapers", mainly used for comparative purposes)
- Source set type (the type of source, being them newspapers, blogs, social media, and so on…)
- Assessment (a binary classification to determine if the article is considered or not relevant according to classification criteria)
- Assessment type (an optional criteria that describes if the article was automatically or manually tagged)
- Document score (the score assigned from the system to the document according to the relevance criteria)

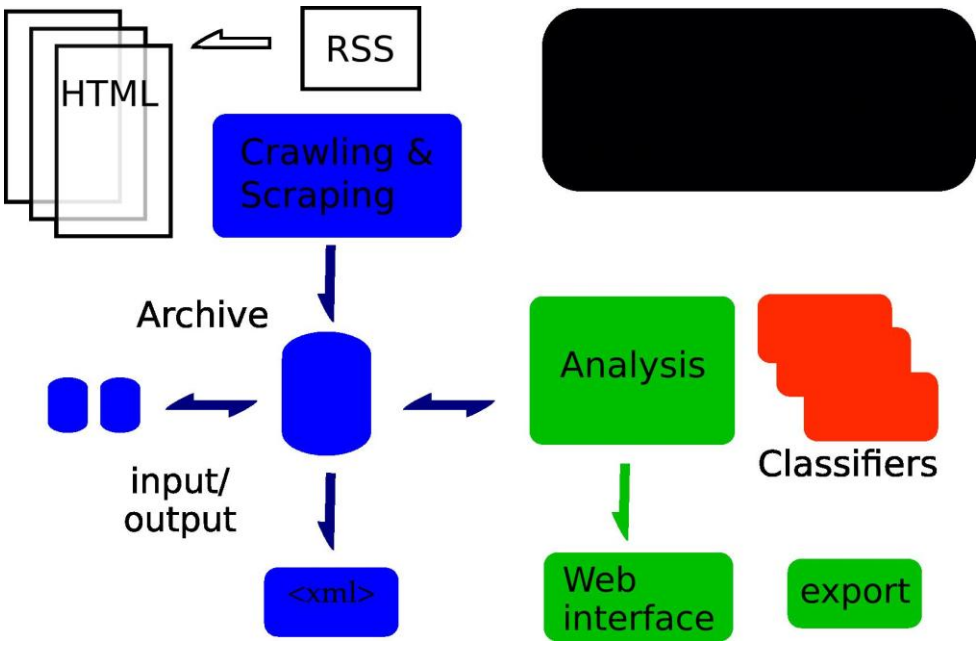

Figure 1: the software architecture of the SMM system

The system is then made up of a Web interface allowing users to analyze data. The search system allows Boolean search within full text, the specification of parameters according to some metadata (such as time, date, score, and so on), and file download, that is the export of an excel file with the information regarding the dataset, and/or the full text export of the texts of the articles.

The other element of the system regards the tagging of the articles according to a relevance criteria automatically generated by the system, that is the potential to apply to the corpus a classifier, made of a set of keywords combined in an algorithm able to select articles that are relevant for a given topic. At the moment two classifiers are implemented for the analysis of Italian media; the first selects articles that are relevant for the "science and technology" topic, while the second selects articles relevant for the topic of "food safety".

The system of tagging is based, as said, on the presence in the text of keywords, grouped in two classes: "keywords" and "multipliers". Each keyword and multiplier has a given weight and the system calculates a score for each item according to a series of rules: if only multipliers are present in the article and no keyword is contained in it, the score of the article remains zero, and this is because the words that are inserted in the multipliers' list are the more generic in the semantic domain of that field, while keywords refer to core semantic field of the topic.

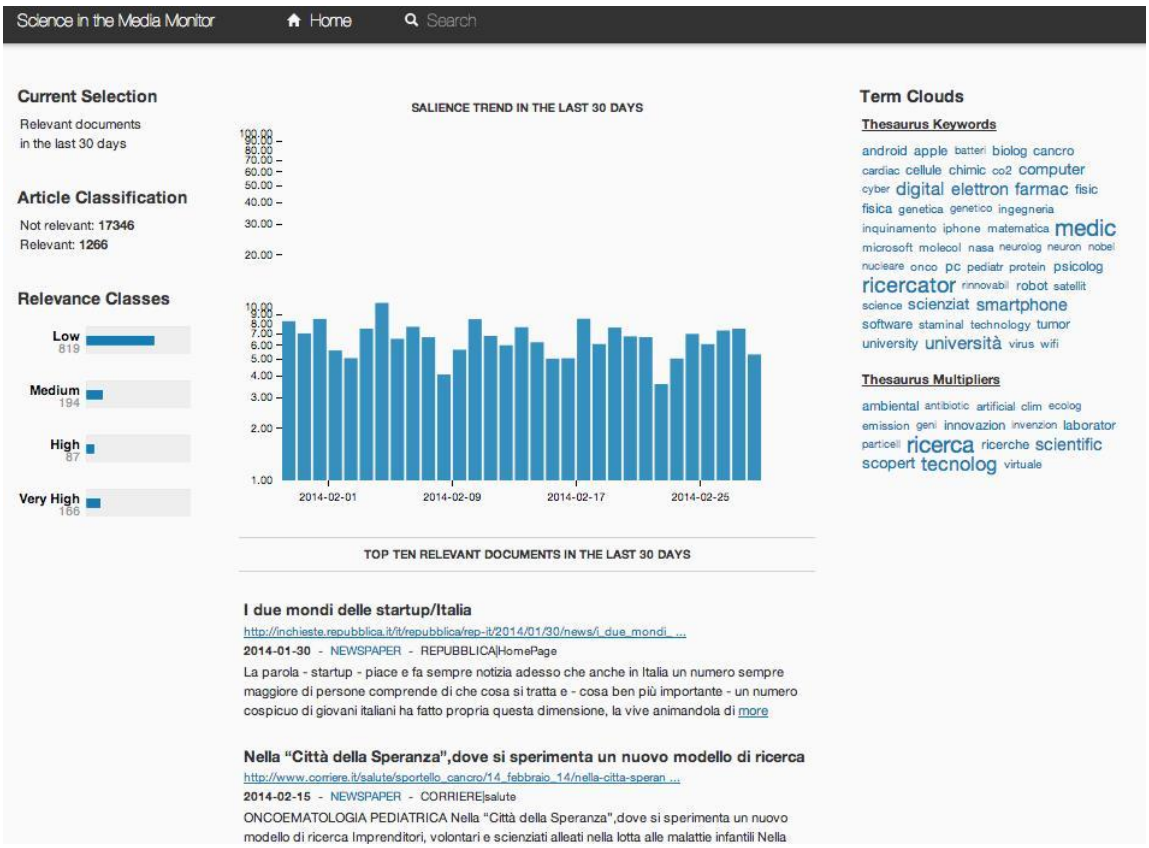

Figure 2: The home interface of the SMM system

On the basis of this system of tagging, the database is weighted for each selected classifier in the search interface of the SMM, which is made of a publicly available homepage displaying the latest articles and trends (see Figure 2) in news coverage of science in the Italian newspapers, while the "search and download" interface is accessible only by researchers with a login and password.

Articles are classified along a multiple relevance criteria, and thus assigned to one of four classes (low, medium, high, and very high S&T content); the home and interface features also a trend graph according to the relevance of articles during the latest period, and a tag cloud reporting on the most detected words among both keywords and multipliers.

Some data about the actual system and two indicators for the quantitative detection of S&T in the press.

The SMM collects on daily bases at the moment two types of sources from the Italian web: 8 newspapers among the most popular ones (Avvenire, Corriere della Sera, il Giornale, la Repubblica, La Stampa, Il Sole 24 Ore, Il Mattino di Napoli, Il Messaggero di Roma), and 500 blogs with ~1000 posts per day. At the date of the 1st March 2014 the system indexed ~700000 online newspaper articles, and has a consisted database starting from 1st January 2010.

Moreover, a clone of the full SMM regards a sample created within the collaboration of the SMM team for the MACAS project (Mapping the Cultural Authority of Science), coordinated by Martin Bauer of the London School of Economics and Petra Pansegrau from Bielefeld University, that covers a sample of articles for each year in the period 1992-2013 on two Italian newspapers, La Repubblica and Il Corriere della Sera, generated with the methodology of the constructed week.

Moreover, although not indexed yet, the system is collecting data from English language newspapers, and more exactly from the website of the New York Times, the Guardian, the Daily Mirror, the Telegraph and the Times of India. At the moment the SMM English version has retrieved 548309 articles, scraped 528712, retaining after a dedupping procedure 256103 articles. Finally, a French version of the system is being tested on French online newspapers (including news from Le Figaro, Le Monde, Liberation, Les Echos, Parisien and Lacroix).

Among the measures that the team has identified to study the presence of S&T coverage in the press, two are particularly relevant and constitute a first way of understanding the potentiality of the system.

The first measure is *salience* of S&T content. This measure is calculated as the percentage of relevant S&T articles on the total of articles for a given period. This is a measure of the share of S&T in the press. On the total database starting from 2010, salience has a value of 7.1%).

The second indicator is *visibility*; this measure can be calculated either as the percentage of the articles published in home pages on the total of relevant articles (in this case the value on all the database of the SMM system from 2010 is 28.7%) or as the

percentage of articles in home page on the total articles in the database (in this case the value for the SMM system from 2010 is 2%).

These indicators are only two of a series of other tools that can be used in order to analyze the presence of science and technology in the media, on which the research team of the SMM is working on; besides that the team is implementing automated topic detection and topic modeling for detecting science and technology contents in the media.

Bibiliography

Bauer, M. W., & Bucchi, M. (2008). Journalism, science and society: Science communication between news and public relations. Routledge.

Bauer, M., Shukla, R., & Allum, N. (2011). The Culture of Science: How does the Public relate to Science across the Globe, London Routledge.

Bucchi, M., and Neresini F., (2012), Monitoring Science in the Public Sphere: The Case of Italy in N. Allum, M. Bauer, R. Shukla (eds.), "The Culture of Science", New York: Routledge, 2011, p. 449-462.

Gaskell, G., and Bauer M. (2001) (eds.), Biotechnology 1996-2000: the years of controversy, London, Science Museum Press.

Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.