# Frogenstein: the SAPO (scientific automatic press observer) system and beyond

Carlos Vogt
State University of Campinas (Unicamp)
Virtual University of the State of São Paulo (Univesp)
cvogt@uol.com.br

Ana Paula Morales
State University of Campinas (Unicamp)
Virtual University of the State of São Paulo (Univesp)
ana.morales@univesp.br

Daniel Carnelossi
Virtual University of the State of São Paulo (Univesp)
daniel.carnelossi@univesp.br

Angelo Grossi
Virtual University of the State of São Paulo (Univesp)
angelo.grossi@univesp.br

## Abstract

Labjor/Unicamp has developed in recent years a computer system known as SAPO (Scientific Automatic Press Observer), which collects, selects, organizes and measures, in an automated fashion, content related to scientific topics published in non-specialized online media. Besides performing searches and retrieving articles containing certain keywords and/or published in a set period of time, the system also produces indicators of the presence of S&T in online media. Articles extracted from the vehicle are analyzed and classified as of scientific content (S&T) or not. The automatic classification of articles as S&T and not S&T introduced by SAPO has inspired other groups to create similar systems in different languages. The existing classifiers are based on a set of keywords (thesaurus) related to science and technology (S&T), but the method has been subject of debate among the collaborative research groups. In this

article we will briefly introduce SAPO and present a work in progress on enhancing the classification method based on text mining.

## Introduction

Relevant information found in text documents can be – and has been increasingly – identified, systematized and used to support a wide range of studies using text mining practices. These practices are based on the organization of databases and procedures for indexing and classifying information, involving increasingly sophisticated computer systems, and resulting in denser and more qualified evaluations. The tools already developed for such purposes within the ambits of various studies (such as language, semiotics, public opinion, sociology and anthropology) may also be applied in the analysis of print, radio and television media (BAUER & GASKELL, 2002).

Outside the academic context, there is also interest from companies, institutions, government agencies and newspaper editors in measuring their visibility in the media, assessing the impact of policies in the press, monitoring how the effects on public readers evolve over time etc[1]. To meet these diverse interests, the Laboratory for Advanced Studies in Journalism at the State University of Campinas (Labjor/Unicamp) in Brazil has embarked on efforts to develop a system of collection, indexation and measurement of the presence and impact of issues related to Science and Technology (S&T) in the online media, called SAPO, the English acronym for Scientific Automatic Press Observer. This is a computer system based on a database integrated with quantitative indicators, measured automatically. Its goal is to evaluate the presence of scientific themes in Brazilian online media, allowing for studies related to the public understanding of science and technology, such as: i) evaluation and measurement of trends in the coverage of different scientific themes, ii) analysis of media coverage of new cases, iii) study of the temporal evolution of a longitudinal coverage of a news item and classic topics; iv) study of the public's perception and response, and v) correlation between the type of coverage of a particular topic and other variables.

[1] "Public and private institutions may need to analyze the impact and repercussion of their press releases, or of their public stance. Newspaper editors and administrators may require quantitative tools to compare their editorial policies with that of other newspapers. The relative weight given to different sorts of news and content, the process throughout the year of the type and quality of published material, may be important data. Professional journalists themselves may be interested in a qualitative and quantitative analysis which compares the coverage which they accorded with that of their colleagues on similar topics or events." (VOGT, C. *et al.*, 2006)

The hypothesis that guided these efforts is that the studies based on text mining of the online media, using the SAPO system, are able to monitor and measure the presence of scientific topics in the media and thus reflect the public understanding of science from a perspective different and complementary to the traditional one, based on surveys. The system has incorporated several improvements since its inception and the team is currently working on developing a dynamic thesaurus based on machine learning, as presented in this article.

## The SAPO system

In order to perform the automated functions of collection, selection, organization and measurement of the content published in online vehicles, SAPO automatically indexes the content of selected websites and classifies the articles as scientific content (S&T) or not (VOGT et al., 20011).

The system currently consists of a set of routines prepared to scan and subsequently index the whole content of the portal Estadao.com.br (EOL), which congregates the major vehicles of the Estado Group, including the nationwide newspaper *O Estado de S. Paulo* (one of the most important in terms of print run and circulation in Brazil)[2].

The classification system (which we will discuss in the following section with more details) is capable of automatically selecting those texts that deal with issues related to S&T in general, scientific-technological and innovation policies, biomedicine and the environment.

The content captured, indexed and analyzed by the SAPO system may be exploited by means of a user-friendly interface, principally in two ways: keyword and/or period search and quantitative indicators.

An intelligent search engine allows several queries on the structured database generated with the collected content: by subject (keywords), by author, by period (day, week, month, year).

Based on the data collected and classified, SAPO also makes it possible to generate quantitative indicators and graphs for specific periods of time. These indicators are useful tools to aid research on online media, particularly with regard to the public

[2] In January 2003, the portal surpassed one million monthly visitors, consolidating its leadership position in queries in real time journalism vehicles in Brazil.

understanding of science from the perspective of the offer of scientific articles in the media. The indicators generated are:

- Indicator of mass (M): number of S&T articles published in vehicle, over a specific period. The temporal analysis of this indicator makes it possible observe the seasonality and to highlight moments of peaks on the coverage of certain scientific subjects.
- Indicator of frequency (f): relative quantity of S&T articles over the total articles published in the vehicle during the selected period. This indicator points to the degree of scientific content of the vehicle relative to its total content, revealing peaks on specific days characterized by the presence of sections intensively "populated" by S&T issues and highlights media cases.
- Indicator of density (d): relative space of S&T articles, that is to say, the percentage of words of these articles over the total number of words in the vehicle during the period under analysis. The density indicator was inspired by the old procedure for measuring the proportion of subjects in print newspapers, made with the use of ruler and calculating the area occupied by topic of interest, in square centimeters.
- Indicator of depth (A): relative weight of the S&T articles compared to the "average" of the vehicle. This indicator combines the indicators of density and frequency, dividing them (A = d / f).

## Mapping S&T in media content: the classification methodology

The current method used by SAPO for classification of the articles as of scientific content (S&T) or not is based on a set of keywords (thesaurus), where each keyword has a specific score, according to their attributed weight. The article under analysis is scanned and each keyword found in the text adds to the value of its score, so that the article's score is the sum of the points of the keywords present (counted only once, without considering repetitions). The score determines whether the text is of scientific content, may be of scientific content (intermediate zone), or is not of scientific content, according to the score range where it stands.

SAPO´s thesaurus-based classification method has been tested and proven reliable both in its initial development phase (VOGT et al., 2006) and more recently (VOGT, 2013). The system´s classification results have been compared to a manual evaluation carried out by human coders and it has been verified a high level of reliability of the system.

Discussions with academic and technical partners, however, have raised some ideas on how to improve the method and develop a finer grained classification system that may help the production of more detailed science culture indicators. One of the main issues raised relate to the comparability of data produced by the system (or similar systems) for different countries or languages. For the data to be compared, the different thesauri need to have the same structure and founding concept. In other words, the idea is that thesauri have a basic structure or guidelines in order to be applied for any language or country. A standard model for the thesauri should be valuable also regarding another important aspect related to further development of the systems: its use not only for classifying the media content as S&T or not, but also for scanning sub-areas, such as environment and astronomy, e.g.

The definition of the boundaries of S&T is a complex issue also to be considered. The common sense understanding of what is S&T may change over periods of time and/or in different contexts, as well as the semantics of scientific vocabulary may vary over time. As an example, many words that have once been used only in specific scientific environment or publications, such as “DNA”, have now a whole different (and broad, in this case) use.

**A possible solution for content classification: an automatic and dynamic thesaurus**

Existing classifiers are based on a predefined list of keywords (thesaurus) and as exposed the method may pose some reliability and validity issues. A possible solution in this direction would be applying machine learning techniques in order to improve the system.

The SAPO team has explored the application of text mining techniques in order to obtain a dynamic and automatically built thesaurus. Text mining is a process that uses techniques of extraction and analysis of data from texts, phrases or words. The technique involves computer algorithms that process explicit or implicit information, which could not be obtained by traditional forms of consultation (MORAIS & AMBRÓSIO, 2007). The application of the technique allows, for example, qualitative and quantitative analysis of large volumes of textual data.

Data mining is a process of identifying potentially useful patterns available in data (FAYYAD, 1996). This is not exactly a quest, but the analysis of documents which may lead to the production of knowledge. In the case of text mining, the process usually follows these steps: selection of documents, semantic (functionality) or statistics (frequency) analysis, data preparation, indexing and normalization, calculating the relevance of terms selection of terms and post processing (MORAIS & AMBRÓSIO, 2007, pg 7).

The identification step can be simple, identifying the terms present in the text; or composite, identifying the occurrence pattern between terms, which may be interesting to the analysis of content related to S&T. Other important steps in this sense are the relevance calculation, in which scores are assigned to words considered most relevant; and the selection of terms, which may be based on their scores or their syntactic position in relation to the text.

Several techniques are currently used for the discovery of patterns of occurrence of terms or for the generation of knowledge from texts. The Latent Semantic Analysis (LSA) is a method for extracting and representing the semantic meaning of words in context, based on the co-occurrence of words in a text.

For this kind of analysis to work properly, it is necessary that the collection of documents to be analyzed by the system is very well known: "The first step in this process is the selection of a collection of texts about certain subjects. This collection serves as the basis for indexing, ie, generation of a semantic vector, as the terms that semantically represent the documents" (MORAIS & AMBRÓSIO, 2007, pg 22; free translation).

Based on this principle, the text mining of a database composed exclusively of previously known scientific articles seems a reliable manner to obtain patterns of occurrence of relevant words in the context of S&T. The result of this type of analysis – meaning the selection of relevant terms of a given set of scientific papers – could therefore set up a new and automatic thesaurus, that could also be dynamic in the sense that it could be rebuilt from time to time, following possible changes in scientific vocabulary.

Furthermore, the technique could be applied to different scientific papers database – in different languages or groups of specific sub-areas of knowledge (such as the environment, astronomy etc.). Still, it allows the influence of the researchers in the construction of the thesaurus to be significantly reduced.

## Reconstructing SAPO: an experiment

In order to test the hypothesis that by using text mining techniques it would be possible to generate a S&T thesaurus in an automated and dynamic fashion, we have decided to reconstruct the SAPO system with software compounds appropriate for this purpose. We have chosen to experiment with an application suite from Apache Software Foundation (ASF).

The SAPO system has been reconstructed following a modular design that is consistent with big data analysis and entirely developed using free tools available in open source. Even though the focus of this experiment was enhance the method of the classification step, we had to change the some other software that compose the system (responsible for collecting and indexing the contents, e.g.) in order to the whole process to function properly. Thus, this new version of SAPO system came to be formed by the aggregation of different software, and hence the metaphorical and alliterative pun with Frankenstein and the meaning of SAPO in Portuguese.

The collection of computer programs chosen is capable of capturing, processing and indexing data present in predefined websites, creating structured data files that can be subject of text mining techniques. By running the software suite on online repositories of scientific articles one could be able to generate a reliable S&T thesaurus. Following this purpose, we are working on the SciELO (Scientific Electronic Library Online) platform – an electronic library covering a selected collection of Brazilian scientific journals in diverse areas.[3]

The text mining of the papers featured in the SciELO platform – in general or in specific areas of knowledge – will bring as a result a structured data file that contains the words and groups of words considered most relevant in that context. By using this new thesaurus, the software suite would then be able to perform the classification of a distinct database (texts) captured and indexed from another platform such as Estadão.com.br.

The experiment described has been carried out in parallel to the prior SAPO system. The classification results obtained from this new approach – based on the text mining of the scientific articles basis of SciELO – will be compared to the already validated classification results of the previous system – based on a set of pre-defined words built in an intuitive fashion by researchers.

[3] Brazil accounted for 195,106 original or review articles and other documents in SciELO 's database in 2012.

## Conclusions

Preliminary search tests using SAPO conducted to date indicate that the system provides extremely useful assistance to studies of online media, with consistent indicators and results. The system, which is under development, offers the research community daily information on media coverage of S&T, making it possible to see not only how often such topics greet the reader, but also the manner in which the reader encounters them in newspapers.

SAPO has inspired the development of similar systems and the exchange of ideas among the group has led to the experimentation of new methodologies based on machine learning for improving the functionalities of the software. The ideas presented in this paper relate to the enhancement of the method of classification of media content as of scientific or not. The proposed technique would allow the standardization of thesauri for different systems, languages and countries. Also, it may be the solution for overcoming the temporal issue regarding the changes on common sense understanding of S&T and on the scientific vocabulary.

## References

BAUER, M.W.; GASKELL, G. (2002) Pesquisa qualitativa com texto, imagem and som. Um manual prático [Qualitative Research in Text, Image and Sound. A Practical Manual]. Vozes , Petrópolis, RJ.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From data mining to knowledge discovery in databases. Ai Magazine, 17:37–54, 1996; In AMBRÓSIO, A.P.L. & MORAIS, E.A.M. Mineração de Textos. Relatório Técnico. Universidade Federal de Goiás, 2007.

VOGT, C.A. et al. (2006) SAPO (Science Automatic Press Observer): Construindo um barômetro da ciência e tecnologia na mídia [Constructing a Barometer of Science and Technology in the Media]. In: Cultura científica: desafios. EDUSP FAPESP, São Paulo, p.85-130.

VOGT, C.A.; CASTELFRANCHI, Y.; RIGHETTI, S.; EVANGELISTA, R.A.; MORALES, A.P.; GOUVEIA, F. (2011) Building a science news media barometer SAPO. In: Bauer, M.; Shukla, R.; Allum, N.. (Org.). The culture of science - how the public relates to science across the globe. 1st ed. New York/London: Routledge, p. 400-413.

VOGT, C.A.; GOUVEIA, F.; MORALES, A.P.; DAHER, F.; PISARUK, F. (2013) Scientific Automatic Press Observer (SAPO): sistema automatic de geração de indicadores de Cultura Científica e de monitoramento de temas científicos na mídia. IX Congreso Iberoamericno de Indicadores de Ciencia y Tecnologia, Bogotá. Available at: http://congreso2013.ricyt.org/files/mesas/2fPercepcion/SAPO.pdf