JADT' 18

PROCEEDINGS OF THE 14TH INTERNATIONAL CONFERENCE ON STATISTICAL ANALYSIS OF TEXTUAL DATA

JADT’ 18

PROCEEDINGS OF THE 14TH INTERNATIONAL CONFERENCE ON STATISTICAL ANALYSIS OF TEXTUAL DATA

(Rome, 12-15 June 2018)

Vol. I

UniversItalia
2018

ISBN 978-88-3293-137-2

Program Committee

Ramón Álvarez Esteban: Univ. of León, E
Valérie Beaudouin: Telecom ParisTech, F
Mónica Bécue: Poly. Univ. of Catalunya, E
Sergio Bolasco: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
François Daoust, UQÀM, Montreal, CDN
Anne Dister, FUSL, Bruxelles / UCL, Louvain, B
Jules Duchastel: UQÀM, Montreal, CDN
Serge Fleury: Univ. Paris 3, F
Cédrick Fairon: UCL, Louvain, B
Luca Giuliano: Sapienza Univ. of Rome, I
Serge Heiden, ENS, Lyon, F
Domenica Fioredistella Iezzi, Univ. of Tor Vergata, I
Margareta Kastberg, Univ. of Franche Comté, F
Ludovic Lebart: CNRS / ENST, Paris, F
Jean-Marc Leblanc: Univ. of Créteil, F

Alain Lelu: Univ. of Franche Comté, F
Dominique Longrée, Univ. of Liège, B
Véronique Magri: Univ. of Nice Sophia-Antipolis, F
Pascal Marchand: Univ. of Toulouse, F
William Martinez: Univ. of Lisboa, P
Damon Mayaffre: CNRS, Nice, F
Sylvie Mellet: CNRS, Nice, F
Michelangelo Misuraca: Univ. of Calabria, I
Denis Monière: Univ. of Montréal, CDN
Bénédicte Pincemin: CNRS, Lyon, F
Céline Poudat: Univ. of Nice Sophia-Antipolis, F
Pierre Retinaud: Univ. of Tolouse, F
André Salem: Univ. Paris 3, F
Monique Slodzian: Inalco, F
Arjuna Tuzzi: Univ. of Padua, I
Mathieu Valette: Inalco, F

Organising Committee

Domenica Fioredistella Iezzi: Univ. of Tor Vergata, I
Sergio Bolasco: Sapienza Univ. of Rome, I
Livia Celardo: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
Francesca della Ratta: ISTAT, I
Fiorenza Deriu: Sapienza Univ. of Rome, I
Francesca Dolcetti: Sapienza Univ. of Rome, I

Andrea Fronzetti Colladon: Univ. of Tor Vergata, I
Francesca Greco: Sapienza Univ. of Rome, I
Isabella Mingo: Sapienza Univ. of Rome, I
Michelangelo Misuraca: Univ. of Calabria, I
Arjuna Tuzzi: Univ. of Padua, I
Maurizio Vichi: Sapienza Univ. of Rome, I
Francesco Zarelli: ISTAT, I

Local Organisation

Francesco Alò, Giulia Giacco,
Paolo Meoli, Vittorio Palermo, Viola Talucci

Table of contents

Abstracts

DOMINIO: A Modular and Scalable Tool for the Open Source Intelligence

Francesca Greco[1], Dario Maschietti[2], Alessandro Polli[3]
[1] La Sapienza University of Rome, Prisma S.r.l. – francesca.greco@uniroma1.it
[2] Prisma S.r.l – d.maschietti@prismaprogetti.it
[3] La Sapienza University of Roma – alessandro.polli@uniroma1.it

Abstract

Prisma has developed an innovative technology for the Open Source Intelligence (OSINT) which aims to provide a solution for those processes of knowledge management, which require the intervention of a human operator, unaided by information technology (IT) support, in one or more stages of the procedure. Such intervention involves a considerable waste of time and resources that could be reduced through the use of an IT tool, partially or totally automating entire stages of the procedure. DOMINIO is a platform that implements tools for automatic online information aggregation, its analysis, the possible alignment with traditional databases and the representation through infographic and georeferencing tools, in order to generate a report. This paper describes the platform architecture, the main algorithms used in the analysis stage of the contents and possible directions of development.

Abstract

Prisma ha sviluppato una tecnologia innovativa finalizzata all'Open Source Intelligence (OSINT) che intende fornire risposta alle necessità di knowledge management, che richiedono l'intervento di un operatore umano, non assistito da supporti di information technology (IT), in una o più fasi della procedura. Tale intervento comporta un notevole dispendio di tempo e risorse che potrebbe essere ridotto attraverso l'utilizzo di uno strumento di IT, automatizzando parzialmente o totalmente intere fasi della procedura. DOMINIO è una piattaforma che implementa strumenti per l'aggregazione automatica di informazioni on line, la loro analisi, l'eventuale allineamento con banche dati di tipo tradizionale, la rappresentazione attraverso tool di infografica e georeferenziazione, allo scopo di generare una reportistica. Il presente contributo descrive l'architettura della piattaforma, i principali algoritmi adottati nella fase di analisi dei contenuti e le possibili direzioni di sviluppo.

Keywords: knowledge management, Open Source Intelligence tool, Information Technology,

1. Introduction

There is a close link between data management and knowledge on the one hand, and knowledge and innovation on the other. The growing mass of unstructured information from disparate channels (search engines, RSS feeds, social networks) and traditional databases entails the need to drastically simplify the preparation, analysis and reporting stages required to structure the information. In fact, only a structured information translates into knowledge. Knowledge, in turn, is a major driver of innovation and, properly managed, it translates into a competitive advantage. The idea at the basis of the tool OSINT (Open Source Intelligence) stems from the needs expressed by analysts – mainly involved in the field of sentiment analysis and opinion mining industry. However, this idea is enough comprehensive to encompass all those activities of knowledge management, similar to the former, which require intervention by a human operator, unaided by IT support (Information Technology), in one or more stages of the procedure, the intervention of which involves a great deal of time and resources. Although in high-end solutions machine learning systems are starting to spread, the available technology is still characterized by significant limitations, especially in the presence of unstructured information. In particular, with regard to supervised machine learning systems, intervention is required by an operator in the initial stages of the procedure and, in general, with reference to any automated system applied to the analysis of a text, it is still impossible to identify complex cognitive functions (for example, irony). Of course, these problems are immanent in many fields of OSINT, and they also affect the stage of reporting, which requires a direct involvement of the analyst, unaided by IT. So, the availability of an IT tool that minimizes human operator intervention − partially or totally automates entire stages of the procedure − would result in substantial advantages, like time savings, increased productivity and the resulting increased efficiency in the allocation of human and financial resources.

Prisma has developed an innovative technology of OSINT, which aims to fix the problems briefly described above. The platform implements tools for automatic aggregation of the online information, their analysis, the alignment with traditional databases, the representation through infographic and georeferencing tools, aimed to automate also the phase of elaboration of the final report.

This paper will describe the architecture of the platform, the main analysis modules and the possible directions of development.

2. Platform Architecture

DOMINIO is an OSINT (Open Source Intelligence) platform that automatically aggregates information from online and traditional databases, analyses it and generates reports on a user-defined subject. The platform collects information by querying several channels: search engines (Google, Yahoo, Bing), social networks (Facebook, Twitter, Google+), RSS feeds, blogs (Blogger, Wordpress, Tumblr), traditional databases. The goal of DOMINIO is to build a structured set of contents, as broad as possible, and to carry out a wide range of qualitative and quantitative analysis. DOMINIO stores these contents within a non-relational database (DB) (MongoDB, 2018; Morphia, 2018), classifying the various documents by channel of origin (Twitter, Facebook, RSS, etc.) to ensure the homogeneity of the collections.

Among the options, the DOMINIO user can make queries on-demand or in a continuous mode. The on-demand option carries out an asynchronous search, while the continuous mode option enables to aggregate periodically data and to track a subject over an extended time span. The DOMINIO’s architecture allows the user to switch from one mode to another; the availability of two searching modes allows overcoming the trade-off between accuracy of analysis and speed of processing.

With regard to one or more subjects selected by the operator, DOMINIO performs synchronous or asynchronous research on a set of Internet channels, such as search engines (Google, Yahoo, Bing), social networks (Facebook, Twitter, Google+), RSS feeds, blogs (Blogger, Wordpress, Tumblr). The user can also extend the search to the Deep Web, through specific search engines, such as Torch or Grams.

Moreover, to meet specific information needs, DOMINION can match these search results with the information achievable from the traditional databases to support many types of analysis (brand reputation, country risk assessment, opinion polls, cyber security, etc.), considerably increasing the operability and flexibility of the tool.

Among the traditional databases already available, DOMINIO includes:

- IHS Jane's (2018), which provides updates on military and political situation, terrorist acts, civil wars, transportation system, for most of the countries in the world;
- Bureau Van Dijk (2018), which collects firms data on ratings, shareholdings, equity investments and M&A;
- MIG (a geographic information database drawn up by one of the authors).

In addition, for specific information purposes, DOMINIO is open for interfacing with Enterprise Resource Planning databases (like SAP, Oracle, etc.) through market tools (Business Object, Quick View).

The search results are recalled by the analyst, who operates from a CMS (Content Management System) application to manage the structured set of content and conduct a wide range of qualitative and quantitative analyses (from simple summary statistics to sophisticated multivariate analyses and text and opinion mining techniques).

The statistical methods implemented on DOMINIO are chosen by the Prisma research team according to a set of criteria that privileges the suitability of one algorithm to automate entire stages of the procedure, in accordance with the original design idea. Moreover, the modular architecture of DOMINIO, described briefly below, allows a quick integration of the latest analysis tools and innovative methodologies produced in the academic field.

Once the stage of content analysis is completed, the CMS application generates a micro-site containing the results (geo-referenced maps, summary statistics, multivariate analysis results, textual and semantic analysis of sentiment analysis, etc.). After selecting a graphic layout for the final report, the analyst has only to write notes and final remarks.

The possibility of including features generating automatic and/or auto-completion comments, customizable by the user, is also being studied. Once the last stage is completed, the report is ready for online publication or traditional diffusion in pdf format, or linked to external services.

From an architectural point of view, DOMINIO is designed following the most modern criteria of modular software design, with the parallel development of the platform’s modules. In short, in order to ensure a greater fault tolerance and high safety standards, the system is divided into three independent logical units (cfr. Figure 1):

- DOMINIO Engine Unit (MEU), which implements the features of 1) scraping information from the sources mentioned above (web, social networks, RSS feeds, traditional databases); 2) storage of results on MEDB database; 3) qualitative and quantitative analysis;
- DOMINIO RESTurl Unit (MRESTU), which receives requests from the MCMS unit, verifies the consistency and forwards the request to the unit ME. Upon receiving the response, it implements the request by adding additional fields (username, token, etc.) and returns them to the MCMS client. The MRESTU unit contains the database (MRESTDB) for user profiling;
- DOMINIO Content Management System Unit (MCMSU), which manages the stage concerning the reporting and archiving of reports according to pre-logical criteria (organization by topic, chronologically, for templates, etc.).

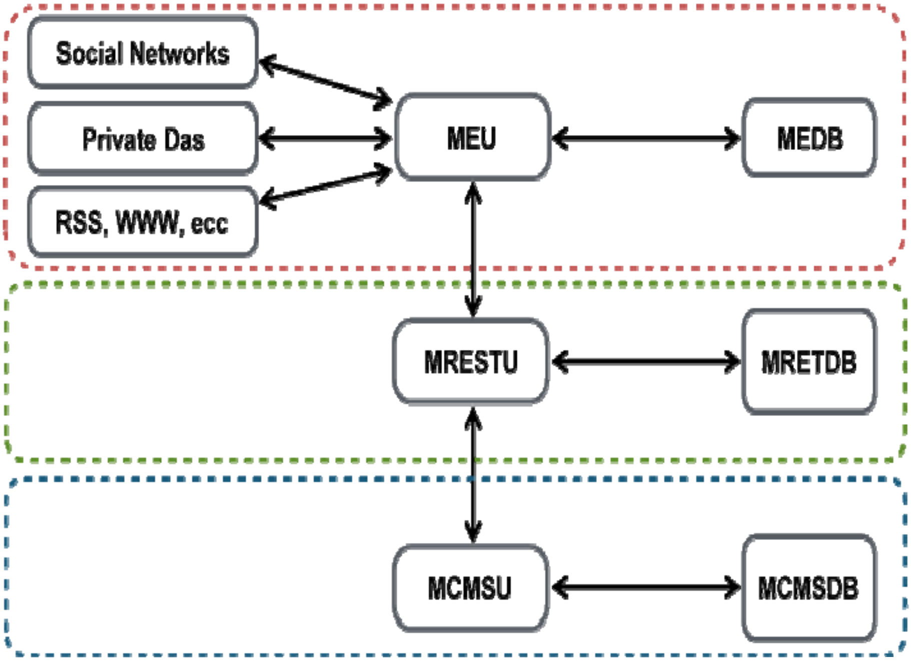

Figure 1 - DOMINIO General Overview

3. Main analysis modules

3.1. Country Threat Assessment

The Country Threat Assessment module supports the Company Intelligence and Security analyst in the country's risk assessment process. Through a responsive type interface, it aggregates information from major global industry databases (eg, IHS Jane's) giving an assessment of external and internal risk and that due to political and socio-economic factors and potential outbreaks or revolutionary movements for 192 different countries. Country Threat Assessment is integrated with intelligence information updated weekly on each country. Through an automatic report, data is aggregated into a single file by optimizing timing of risk assessment and providing a solid foundation for any further detailed analysis. DOMINIO offers the possibility of making a full or partial information download, and the generation of an automatic report, thus optimizing any drafting processes.

3.2. Due Diligence

The Due Diligence module supports the Economic Intelligence analyst in the process of business valuation in relation to suppliers, partners and customers. Among the sectors analysed in the module are included assessments of profitability and financial performance as well as creditworthiness. Through a simple and intuitive interface, the module aggregates information from leading industry databases and returns an economic, financial and credit risk profile on hundreds of millions of businesses around the world. The Due Diligence Module also allows an assessment of individuals, through the analysis of individuals exposed politically, returning an automatic report that integrates the main aspects of each business and its economic risk analysis.

3.3. Open Source Intelligence

On completion of the aggregation of large amounts of data from major social networks (Facebook, Twitter, Youtube) and the main Italian newspapers based on predetermined keywords analyst, a statistical representation of the main trending topic is returned and an output of structured data for subsequent multivariate analysis is generated. Furthermore, the module allows the geo-referencing of content, highlighting even at geographic levels useful signs for the analyst. As for each of DOMINIO's modules, it is possible to generate automatic reporting.

3.4. Geographic Information Module

This is a module that analyses the information inferable from a dataset of basic statistical information and related indicators, with reference to a multitude of subjects, 9 of which are in a current stage of development. The basic statistical information, refers to the division of the Italian territory into provinces, covering a time period between 1995 and the latest available year, which for some subject areas is ongoing or, more frequently, the previous year to the current one. The dataset will be supportive to a wide range of applications - from forecasting and scenario analysis, counterfactual analysis to spatial analysis.

3.5. Text Mining Module

On completion of the automatic analysis of textual data using statistical methods (Lebart et Salem, 1994; Feldman et Sanger, 2006; Bolasco, 2013), in order to extract structured information, the main statistical methods of analysis of textual data implemented in DOMINIO are: factor analysis (correspondence analysis, multiple correspondence analysis); cluster analysis (k-mean, bisetting k-mean, fuzzy clustering, etc.); network analysis; Markov analysis; pattern recognition.

For example, during the French presidential campaign of 2017 we analysed the sentiment about migration, that was one of the most debated theme. We performed an Emotional Text Mining (Greco et al., 2017) in order to explore

the emotional content of the Twitter messages concerning migration written in French in the last two weeks before the first round of the presidential election in 2017. The aim was to analyse the opinions, feelings and shared comments, classifying the contents and the sentiments. We retrieved the messanges from the Twitter repository collecting a sample of over une hundred thousand tweets The large size corpus of 2.154.194 tokens (TTR = 0,01; Hapax percentage = 40,4) underwent a multivariate analysis based on a bisecting *k*-means algorithm (Savaresi et Boley, 2004) to classify the text, and a correspondence analysis (Lebart et Salem, 1994) to detect the latent dimensions setting the cluster per keywords matrix. The advantage connected with this approach is to interpret the factorial space according to words polarization, thus identifying the emotional categories that generate migration representations, and to facilitate the interpretation of clusters, exploring their relationship within the symbolic space (Greco, 2016).
The results interpretation allowed for the detection of seven representations of migrants that corresponded to three different sentiments: positive (42%), negative for the community (45%), and negative for migrants (13%). We considered as negative the representation of migrants as squatters, invaders, terrorists, trafficking slaves and migration victims, and positive the sport heroes and the EU solidarity target. Among the negative clusters, we distinguished negativity according to the direction of the action: squatters, terrorists and invaders are negative for the community and trafficking slaves and migration victims are negatives for migrants themselves (see Greco et al., 2017). Moreover, It was possible to highlight the connection between the real life events and the tweets production. While the terrorist attack three days before the first round of voting in the centre of Paris had slightly modified the production of messages, the candidates' interviews had a higher impact. This suggests that the medialization was more important than the terrorist attack in the production of messages (see Greco et al., 2017).

4. Conclusion

The innovative aspect that characterizes DOMINIO is the ability to aggregate data of different types and from different channels of information, automatically, simply and transparently. Moreover, its structure allows for the integration of the latest analytical tools and innovative methodologies produced in academia. By means of an automated reporting system, the analyst is supported in the assessment of risk and the collection of information in the geopolitical and economic field and from open sources. The set of modules allows the analyst to generate knowledge from an ever-growing amount of data by optimizing the processes of assessment and risk reduction.

References

Bolasco S. (2013). *L'analisi automatica dei testi: Fare ricerca con il text mining*. Carocci.

Bureau von Dijk (2018). *A Moody's Analytics Company*. Bureau von Dijk, https://www.bvdinfo.com/it-it/home

Feldman R. and Sanger J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.

Greco F., Maschietti D. and Polli A. (2017). Emotional text mining of social networks: The French pre-electoral sentiment on migration. *RIEDS,* 71(2): 125:36.

IHS Jane's (2018). *Jane's Information Group*. IHS Jane's, http://www.janes.com

Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod

MongoDB (2018). *MongoDB for GIANT ideas.* MongoDB, https://www.mongodb.com

Morphia (2018). *The Java Object Document Mapper for MongoDB*. MongoDB, https://mongodb.github.io/morphia/

Savaresi S.M. and Boley D.L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis,* 8(4): 345-362.