International Journal of Interactive Multimedia and Artificial Intelligence

March 2016, Vol III, Number 6, ISSN: 1989-1660

Data really powers everything that we do.

Jeff Weiner (2015)

Special Issue on Big Data & AI

http://www.ijimai.org

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE
ISSN: 1989-1660–VOL. III, NUMBER 6

Editor's Note

Digital information has redefined the way in which both public and private organizations are faced with the use of data to improve decision making. The importance of Big Data lies in the huge amount of data generated every day, especially following the emergence of online social networks (Facebook, Twitter, Google Plus, etc.) and the exponential growth of devices such as smartphones, smartwatches and other wearables, sensor networks, etc. as well as the possibility of taking into account increasingly updated and more varied information for decision making. [1]

With proper Big Data analysis we can spot trends, get models from historical data for predicting future events or extract patterns from user behaviour, and thus be able to tailor services to the needs of users in a better way.

Using Big Data is becoming widespread among organizations of all kinds. It is a fact that large companies, start-ups, government agencies and non-governmental organizations are gradually being encouraged to use microdata generated by digital devices to operate more efficiently. When these microdata are aggregated they turn into massive amounts of data which require specialized tools and skills to be managed.

The challenge organizations are facing is that the increasing amount of data is too large or too unstructured to be managed and analysed with traditional methods. Think of the data derived from the sequence of clicks from the Web, social media content - tweets, blogs, Facebook wall postings (Facebook alone accounts for more than 1 billion active users generating social interaction content. Google processes on average over 53 thousand search queries per second, making it over 4.6 billion in a single day) - or radio frequency identification systems, which generate up to a thousand times more data than conventional barcode systems (12 million RFID tags – used to capture data and track movement of objects in physical world – had been sold in by 2011. By 2021, it is estimated that that number will have risen to 209 billion. Walmart manages more than 1 million customer transactions per hour). In the World 10.000 payment card transactions are recorded every second. The amount of data transferred over mobile networks increased by 81% to 1.5 Exabyte per month between 2012 and 2014. More than 5 billion people make phone calls, send text messages and surf the Internet with mobile phones. Every day they send 340 million tweets (4.000 per second!). To date they've generated 2.5 trillion bytes of data. However, very little of this information is in the form of rows and columns of traditional databases.

While the increasing amount of data is an undisputable fact, from a social point of view the most relevant issue is the nature and number of problems that the processing of these data is helping to solve. Following Davenport, we could say that although the management of information and data is something that is used in virtually all areas, there are some in which Big Data is used with particular intensity [2]. Among these we include the following: health, politics, finance, security, marketing and business management, the study of social networks, risk analysis, smart cities, human resources, fraud detection, environmental management and education.

Widespread use of Big Data is a result of it having become a problem-solving tool by being able to integrate the traditional tools for managing data with other characteristics of artificial intelligence. Artificial intelligence has been used in several ways to capture and structure Big Data, analysing it to obtain key insights [3].

This special issue is designed with the primary objective of demonstrating what we have just noted: the diversity of fields where big data is used and consequently, how it is gaining increasing importance as a tool for analysis and research. In this sense there are works related to the following topics: digital marketing, optimization of message exchange, sentiment analysis, text analytics, e-learning, financial risk control, forecasting behaviour in video games, energy policy and health.

The first two papers of this special issue are related to Big Data and marketing. Juan Carlos Gonzálvez and Francisco Mochón´s paper, *"Operating an Advertising Programmatic Buying Platform: A Case Study"*, analyses how new technological developments and the possibilities generated by the internet are shaping the online advertising market [4]. More specifically it focuses on a programmatic advertising case study. The origin of the problem is how publishers resort to automated buying and selling when trying to shift unsold inventory. The platform executes, evaluates, manages and optimises display advertising campaigns, all in real-time. The results of this case study show that the platform and discard algorithms incorporated therein are an ideal tool to determine the performance and efficiency of different segments used to promote products. Thanks to Big Data tools and artificial intelligence the platform performs automatically, providing information in a user-friendly, simple manner.

The next paper is also related with marketing and the management of large volumes of data. *"PInCom project: SaaS Big Data platform for ubiquitous marketing using heterogeneous protocols and communication channels"*, written by Juan Manuel Lombardo, Miguel Angel López, Felipe Mirón, and Susana Velasco integrates aspects of cloud computing with the treatment of large volumes of data and the potential and versatility of messaging through various formats and platforms, especially on mobile devices [5]. The importance of this issue arises with the high number of users who have access to these technologies and the information which can be consumed through existing communications networks. The main objective is to offer a cloud service for ubiquitous marketing and loyalty by sending messages using different protocols and communication channels. The platform used is able to handle a lot of traffic and send messages using intelligent routing through standardized protocols while managing information security when it is connected.

"Fine Grain Sentiment Analysis with Semantics in Tweets" presented by the group of researchers of the University of Malaga, consisting of Cristóbal Barba González, José García-Nieto, Ismael Navas-Delgado and José F. Aldana-Montes, is devoted to the study of sentiment analysis [6]. The opinions of Twitter users can be assessed by classifying the sentiment of the tweets as positive or negative [7]. However, tweets can be partially positive and negative at the same time by containing references to different entities within a single tweet. As a result general approaches usually classify these tweets as "neutral". In this paper the authors propose a semantic analysis of tweets using Natural Language Processing to classify the sentiment with regards to the entities mentioned in each tweet. We offer a combination of Big Data tools (under the Apache Hadoop framework) and sentiment analysis using RDF graphs supporting the study of the tweet's lexicon. The work is empirically validated using a sporting event, the 2014 Phillips 66 Big 12 Men's Basketball Championship. The experimental results show a clear correlation between the predicted sentiments with specific events during the championship.

The article by Vidal Alonso and Olga Arranza, *"Big Data & eLearning: A Binomial to the Future of the Knowledge Society"*, focuses on the field of education [8] [9]. The combination of different learning analytical techniques with new paradigms of processing,

DOI: 10.9781/ijimai.2016.360

such as Big Data, will enable relevant information to be accessed by educational authorities and teachers, who in turn will be able to change and optimise current teaching methods. This paper shows a case study where the teacher obtains information about the most popular tools in new learning environments. From this information new strategies of teaching-learning can be set based on student experience so that, looking for greater student participation, the teacher may propose tasks that employ tools preferred by students instead of others less popular. The proposed activities, which involve the use of collaborative tools, stimulate group work, widely regarded as a positive factor in the teaching-learning collaborative process. The use of these tools is highly satisfactory to students, resulting in a more active participation, increased student motivation and a significant improvement in learning.

"Social Network Analysis and Big Data tools applied to the Systemic Risk supervision" written by Mari-Carmen Mochón adopts a different approach to the papers presented so far. It analyses how you could use the huge amount of information generated by financial transactions to combat systemic risk. The importance of this issue is clear and is in line with G20 concerns: the need to strengthen the supervision and control of risk, especially in over- the-counter financial markets [10]. This case study makes a concrete proposal of an analysis methodology. Using Big Data solutions currently applied to Social Network Analysis (SNA), information such as propagation risk can be identified without the need for expensive and demanding technical architectures. This case study exposes how the relations established between the financial market participants could be analysed in order to identify market behaviour and risk of propagation.

The possibilities of using Big Data to address environment and energy problems are discussed in the article by Diego J. Bodas-Sagi and José M. Labeaga, *"Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy"* [11]. This paper analyses the public opinion regarding the Spanish Government's energy policy, using the Global Database of Events, Language, and Tone (GDELT). The aim of the authors is to build sentiment indicators arising from this source of information and, in a final step, evaluate if positive and negative indexes have any effect on the evolution of key market variables such as prices and demand.

"Real-Time Prediction of Gamer Behaviour Using Variable Order Markov and Big Data Technology: A Case of Study", written by Alejandro Baldominos, Esperanza Albacete, Ignacio Marrero and Yago Saez, presents the results found when predicting the behaviour of gamers in commercial videogames datasets. Identifying gaming profiles can help companies to gain a better understanding on how users interact with their games and adapt their products to customers accordingly [12]. Understanding these profiles and learning from them comprise the first step for conditioning the behaviour of gamers to optimize these metrics and to ultimately increase the overall performance of the game. To do so, the company can take an active role so that users from a certain profile can move to other profiles providing better metrics or higher revenues. This paper uses Variable-Order Markov to build a probabilistic model that is able to use the historic behaviour of gamers and to infer what will be their next actions. Being able to predict with accuracy the next user's actions can be of special interest to learn from the behaviour of gamers, to make them more engaged and to reduce churn rate. In order to support a big volume and velocity of data, the system is built on top of the Hadoop ecosystem, using HBase for real-time processing; and the prediction tool is provided as a service and accessible through a RESTful API. The prediction system is evaluated using a case of study with two commercial videogames, attaining promising results with high prediction accuracies.

The use of Big Data in the field of healthcare is experiencing a remarkable growth in such diverse areas as the fight against cancer or the validation of certain drugs [13]. *"Detection of Adverse Reaction to Drugs in Elderly Patients through Predictive Modelling"*, by Rafael San Miguel Carrasco, leverages predictive modelling to uncover new insights related to adverse reaction to drugs in elderly patients. The results of the research show that rigorous analysis of drug interactions and frequent monitoring of patients' adverse reactions to drugs can lower mortality risk.

"Text Analytics: the convergence of Big Data and Artificial Intelligence" by Antonio Moreno and Teófilo Redondo focuses on the study of what is known as text analytics, i.e. the analysis of the text contained in emails, blogs, tweets, forums and other forms of textual communication [14]. Text Analytics has produced useful applications for everyday use. The authors discuss the following three: Lynguo, IBM's Watson, and IPsoft's Amelia.

Francisco Mochón

Juan Carlos Gonzálvez

References

[1] Aldana J, Baldominos A, García JM, Gonzálvez JC, Mochón F, Navas I; (2016). "Introducción al Big Data", *Editorial Garcia-Maroto editores S.L.*

[2] Thomas H. Davenport. (2014). Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. *Harvard Business Review*.

[3] Daniel E. O'Leary. (2013). "Artificial Intelligence and Big Data", *IEEE Intelligent Systems, vol.28, no. 2, pp. 96-99, March-April 2013, doi:10.1109/MIS.2013.39.*

[4] Thomas H. Davenport and Julia Kirby. (2015). Beyond Automation. *Harvard Business Review. June, 2015.*

[5] Viktor Mayer-Schönberger,Kenneth. (2013). CukieBig Data: A Revolution that Will Transform how We Live, Work, and Think. *Houghton Mifflin Harcourt*.

[6] Bing Liu. (2010). Sentiment Analysis and Subjectivity. In Handbook of Natural Language Processing, Second Edition, (*editors: N. Indurkhya and F. J. Damerau*).

[7] F. Mochón y O. Sanjuán. (2014). A First approach to the implicit measurement of happiness in Latin America through the use of social networks. *International Journal of Interactive Multimedia and Artificial Inteligence. March 2014.*

[8] George Siemens and Phil Long. (2011). "Penetrating the Fog: Analytics in Learning and Education" *EDUCAUSE Review, v46 n5 p30-32, 34, 36, 38, 40 Sep-Oct 2011.*

[9] Anthony G. Picciano. (2012). "The Evolution of Big Data and Learning Analytics in American Higher Education." *Journal of Asynchronous Learning Networks, v16 n3 p9-20 Jun 2012.*

[10] Daning Hu, Gerhard Schwabe and Xiao LiEmail (2015). "Systemic risk management and investment analysis with financial network analytics: research opportunities and challenges." *Financial Innovation. June, 2015 1:2.*

[11] Claudia Vitoloa, Yehia Elkhatibb, Dominik Reusserc, Christopher J.A. Macleodd, and Wouter Buytaerta.(2015). Web technologies for environmental Big Data. Environmental Modelling & Software. *Volume 63, January, pages 185–198.*

[12] Steven Rosenbush and Michael Totty. (2013). "How Big Data Is Changing the Whole Equation for Business." *Journal Reports: Leadership. The Wall Street Journal, March 10, 2013*.

[13] Hsinchun Chen, Roger H. L. Chiang and Veda C. Storey. (2012). Business Intelligence and Annalytics: from big data to bog impact. *MIS quarterly, 2012 http://hmchen.shidler.hawaii.edu/Chen_big_data_MISQ_2012.pdf.*

[14] Vishal Gupta and Gurpreet S. Lehal. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging technologies in web intelligence,vol. 1, Nº. 1 August 2009.*

TABLE OF CONTENTS

EDITOR'S NOTE III
OPERATING AN ADVERTISING PROGRAMMATIC BUYING PLATFORM: A CASE STUDY 6
PINCOM PROJECT: SAAS BIG DATA PLATFORM FOR AND COMMUNICATION CHANNELS 16
A FINE GRAIN SENTIMENT ANALYSIS WITH SEMANTICS IN TWEETS 22
BIG DATA & ELEARNING: A BINOMIAL TO THE FUTURE OF THE KNOWLEDGE SOCIETY 29
SOCIAL NETWORK ANALYSIS AND BIG DATA TOOLS APPLIED TO THE SYSTEMIC RISK SUPERVISION 34
USING GDELT DATA TO EVALUATE THE CONFIDENCE ON THE SPANISH GOVERNMENT ENERGY POLICY 38
REAL-TIME PREDICTION OF GAMERS BEHAVIOR USING VARIABLE ORDER MARKOV AND BIG DATA TECHNOLOGY: A CASE OF STUDY 44
DETECTION OF ADVERSE REACTION TO DRUGS IN ELDERLY PATIENTS THROUGH PREDICTIVE MODELING 52
TEXT ANALYTICS: THE CONVERGENCE OF BIG DATA AND ARTIFICIAL INTELLIGENCE 57

OPEN ACCESS JOURNAL

ISSN: 1989-1660

COPYRIGHT NOTICE

Operating an Advertising Programmatic Buying Platform: A Case Study

Juan Carlos Gonzálvez and Francisco Mochón

Zed Worldwide, Universidad Nacional de Educación a Distancia (UNED), Spain

***Abstract* — This paper analyses how new technological developments and the possibilities generated by the internet are shaping the online advertising market. More specifically it focuses on a programmatic advertising case study. The origin of the problem is how publishers resort to automated buying and selling when trying to shift unsold inventory. To carry out our case study, we will use a programmatic online advertising sales platform, which identifies the optimal way of promoting a given product. The platform executes, evaluates, manages and optimizes display advertising campaigns, all in real-time. The empirical analysis carried out in the case study reveals that the platform and its exclusion algorithms are suitable mechanisms for analysing the performance and efficiency of the various segments that might be used to promote products. Thanks to Big Data tools and artificial intelligence the platform performs automatically, providing information in a user-friendly and simple manner.**

***Keywords* — Programmatic Buying, Real Time Bidding, Online Advertising Market, Big Data, Online.**

I. Introduction

The buying and selling of advertising is no different from transactions carried out in any other market. The central issue is one of supply and demand.

Demand is fuelled by advertisers (directly or via their media agencies), who buy advertising space seeking maximum efficacy for their campaigns and optimal cost efficiency. Supply is driven by media or other formats which offer advertising inventories. The goal as always is to achieve maximum performance while maximizing revenues [11].

The development of the online advertising market, which dates back to 1993, has always been linked to the use and evolution of new technology. New technological developments and the new possibilities generated by the internet have always shaped the market and determined the direction of new developments. Business models also change as technology drives new forms of interaction between supply and demand.

This article focuses on the analysis of a case study of programmatic advertising buying and is structured as follows. The second section analyses the evolution of the online advertising market. The third section looks at how publishers resort to automated buying and selling when trying to shift unsold inventory. Sales are executed automatically according to the usual criteria of 'buyers' and 'sellers', leading eventually to a programmatic sale. The fourth section introduces a programmatic sales platform for online advertising. This will then allow us to carry out a case study and to watch the platform pinpoint the optimal way of promoting a given product. The platform executes, evaluates, manages and optimizes display advertising campaigns, all in real-time. The fifth section analyses the platform's architecture and its basic characteristics. The sixth section explains the process followed to obtain the results and summarizes the main achievements of the case study. The seventh section contains the main conclusions and indicates possible future lines of research.

II. The Evolution Of The Online Advertising Market

During the early years, almost all inventory was bought for fixed placements. Clients would pay for a certain quantity of a particular placement within a given format or medium. The amount would usually depend on the length of time the advertising would be displayed and the relevance of the placement to that particular medium, either to its scale or to its placement relative to scale.

Advertisers (and agencies which began to branch out into online media) would purchase placements or print-runs directly from existing formats, known as publishers. Available advertising space, known as the 'inventory', was sold by print-run (the number of times a creative execution would appear in a given placement). Advertisers would buy print-runs by the thousands, in units known as CPM (Cost per Mile). The development and popularization of internet access gave rise to an increase both in the number of publishers, as well as in the volume of content these publishers generated. The surge in supply meant that a large proportion of inventory remained unsold [18].

This unsold inventory, combined with the newest technological capabilities, gave rise to a new business model and a new player in the value chain: Advertising Networks or Ad Networks, which served as agents or brokers, buying unsold inventory from publishers [8] [9] .

Ad Networks make it easier for advertisers to target their campaigns by applying technology to aggregate and segment audiences, packaging and selling advertising accordingly. The technology used by Ad Networks is not particularly sophisticated; in fact it is so negligible as to present a very low barrier to entry into this new segment in the value chain. As a result, the number of Ad Networks soon began to snowball, creating a new problem. The sheer quantity of Ad Networks now marketing unsold advertising inventory in different models or packages soon led to online advertising becoming a highly competitive market, with each actor focused on maximizing their own performance. Buying or selling at the best possible price. The presence of so many players had a negative effect on demand: advertisers could now choose between different Ad Networks to 'buy' the same audience more than once. Furthermore, the market soon became in need of efficiency improvements, culminating in yet another new business model and the appearance of another new player in the value chain: Ad Exchanges.

Ad Exchanges allow both advertisers and publishers to exploit audiences rather than print-runs. Ad Exchanges are able to target audiences via publishers' platforms. Advertisers are then able to select and purchase their target audience. Rather than being booked and purchased directly, audiences are bought via a system of real-time bidding (RTB), in which wins who makes the highest bid [13]. The winning bidder secures the right to position their adverts with the right audience at the right time.

DOI: 10.9781/ijimai.2016.361

The appearance of this new business model did not entail the disappearance of the previous one; advertisers and publishers were now able to choose between buying and selling inventory via Ad Networks or buying and selling audiences via Ad Exchanges (Fig 1).

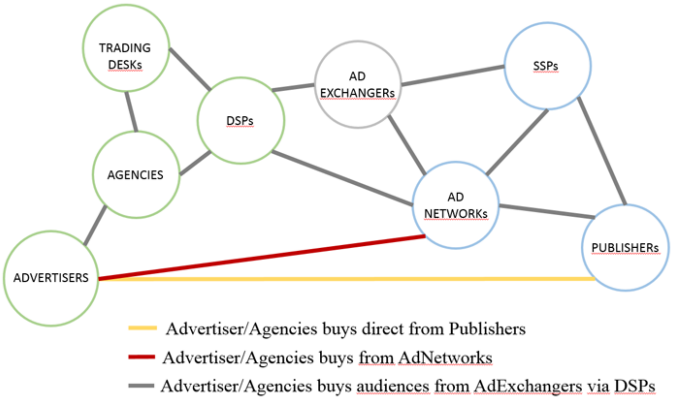


Fig. 1 Online Advertising Business Models.

New opportunities for boosting efficiency and enhancing the buying and selling process began to open up at both ends of the value chain. At the buying end, some advertisers or agencies created their own trading agencies (proprietary agency trading desks), while others joined Demand Side Platforms (DSPs). Both scenarios allow for efficient interaction between advertisers or agencies and Ad Exchanges, via real-time auctions online.

On the selling side, some publishers are able to connect directly with Ad Exchanges, while others connect via Supply Side Platforms (SSPs). The latter process has parallels with the use of DSPs on the purchasing side, in that it allows publishers to interact efficiently with Ad Exchanges, automatically optimizing inventory performance. Nevertheless, the use of Ad Networks remains an effective way of selling inventory.

In addition to the players described above, there are two other crucially important players in the advertising ecosystem: Ad Servers and Data Management Platforms (DMPs). Ad Servers are technological platforms that manage the delivery of creative executions for printing in various formats.

DMPs, as their name suggests, manage data; namely, population segmentation data on groups such as age, gender, location, preferences and so on. This data allows advertisers to select the right audience for each advertising message.

Each and every platform in the advertising ecosystem has been impacted by new developments and technological advances, improving both efficiency and usage rates. This boost has given rise to entirely new forms of interaction, engendering a new purchasing and selling paradigm in the realm of online media [1].

The technological advances responsible for driving this new paradigm have been rolling out over the last five years, precipitating change only when they have coincided efficiently and in concert. Some of these advances include:

- The increase in computational capability: large volumes of complex data need to be processed in milliseconds to allow for real-time decision-making.
- The reduction in storage costs: advertising generates a vast volume of data.
- The application of scientific methods to marketing: marketing professionals make increasing use of Algorithms, Mathematical modelling, Artificial Intelligence or machine learning.
- The increasing speed of data connections.
- Personalization: user recognition mechanisms.
- The use of RTB processes in buying and selling advertising space [17].

Lastly, while the globalization of the advertising market does not in itself qualify as a technological advance, it too has played a role in the birth of the new paradigm. These technologies allow a great deal of rich information to be harvested from advertising activity.

This information can then be applied to boost advertising efficiency, thanks to web analysis and behavioural targeting. As a result, audience acquisition has now become widespread across brands. These brands - no longer exclusively interested in purchasing space - are now also focused on targeting specific audience segments, such as women/men, travel enthusiasts, young people, city dwellers, and so on [4] .

On the whole, the online advertising world presents two main options. The first is search engine advertising: advertisers pay to have their creative executions appear when users type certain keywords into search engines such as Google, Bing or Yahoo. The second is display advertising, in which adverts appear when the user visits certain websites. Display advertising normally takes the form of web banners. This paper will focus on display advertising [12].

III. Online Advertising Today

As internet penetration advances apace across both computers and mobile devices (smartphones, tablets), online advertising supply continues to outstrip demand. Unsold inventory can constitute a very real risk, and it must be addressed carefully; not least because, if handled properly, it opens up new business possibilities.

Publishers resort to automated buying and selling when trying to shift unsold inventory. Sales are executed automatically according to the usual criteria of 'buyers' and 'sellers', leading eventually to a programmatic purchase [2]. In a programmatic purchase, each party to the transaction (purchaser and vendor) entrusts a machine to complete the sale on their behalves. The purchaser is able to refine the profile selection process – a capacity that every brand demands - while the vendor endeavours to optimize their inventory in the most effective way.

The process has been subject to improvements, thanks to the contribution of data by third parties, which allows to verify the suitability of a particular profile (or profiles) to a particular campaign. Before these third parties arrived on the scene, players had to rely on feedback from only one source (usually cookies from user-identification enabled web browsers) to evaluate their impressions. The presence of neutral third party specialists in tagging and identifying audiences (DMPs and third-party data suppliers) has removed this dependency, and the data they offer is now widely used.

Technology has allowed the automatic buying process to go one step further by enabling the purchasing and selling of media to take place in real-time. Until recently, advertisers used to buy website display advertising, advance-booking the number of imprints that they wanted to display to visitors to those sites. In recent years, the use of APIs has become widespread. These allow advertisers to purchase advertising on an impression by impression basis, with prices negotiated through Real-Time Bidding (RTB) via auction [6].

Instead of purchasing inventory directly from publishers, advertisers now enter into the imprint auction system. Using DSPs, they upload their advertising campaigns, target demographics, and the price they are prepared to pay. Publishers put their inventory and audience at the disposal of Ad Exchanges via SSPs. Agents in this ecosystem, including DMPs and Ad Servers, are integrated into the chain via APIs. Technology takes care of everything else.

While programmatic media selling has solved many of the problems

around supply and demand that were typically encountered in the world of online advertising, particularly when it comes to display advertising, this new form of trading presents a new set of challenges and problems to both sides of the value chain. Programmatic media buying has brought with it the automation of media purchasing, however the system requires something more than automation: intelligence [14].

This article will address some of the problems the purchasing side faces within the new paradigm, and will put forward a solution to some of them [12].

When an advertiser decides to execute a campaign there are three key drivers behind the decision making process: creative executions, target segments, and publishers in whose media these creative executions will appear. The goal, as always, is to refine the campaign and maximise its performance.

Audience selection, however, is no trifling matter, and cannot be rushed. Marketers are not able to predict the effect that the advertising campaign will have on consumer behaviour; on how they will respond, and what effect the stimulus will have on consumer motivation. In other words, initial marketing decisions may not be correct or, at the very least, might be in need of some refinement. However, within this new media sales paradigm, every link in the value chain is capable of sharing information in real-time about a campaign's evolution. This offers a great advantage when it comes to campaign optimization: we now have the chance to analyse a campaign's performance in real-time, and make decisions as and when necessary. Within this environment, real-time decision-making is complex, and calls for skilled analysts and in-house business specialists within the workforce. In addition, the volume of information generated within the new paradigm is vast. This compounds the task; rendering it even more complex, specialised and - ultimately - expensive. In spite of the added complexity, current technological advances supply everything we need to be able to carry out this analysis in a profitable way, without labour-intensive and costly need for direct supervision [15].

This article describes a technological platform and processes that must be in place in order to optimize the three key elements that impact on the performance, efficiency and efficacy of an online campaign: creative executions, audience segments and publishers. We can then begin to maximize the ROI on any campaign [16].

IV. Proposal

Taking what has been mentioned in this paper as a reference point, we will now define and implement a programmatic sales platform for online advertising, which we will call PSP. We will carry out a case study, using the platform to promote a given product. This platform executes, evaluates, manages and optimizes display advertising campaigns, all in real-time.

Campaigns can be optimized according to the following variables: user demographics, the media and distribution channel used by the campaign, and the adverts shown.

The platform continually monitors and analyses the performance of programmed campaigns, as part of a cyclical process that focuses on multiple samples (multiple segments, multiple creativities and multiple publishers), in order to highlight which of these samples are the most efficient. Given that campaigns might need to be tested against a range of variables (segments, creativities, publishers), each of which comes with multiple options, the process has the potential to become very costly from an economic perspective. To mitigate this, we need to execute each campaign and its respective options with a small investment, but enough to generate statistically significant results. This will allow us to make an appropriate decision for each case. We will call this process 'prospection'. Prospection, then, involves running a campaign through a platform that applies an automated process to determine which combinations of audience, creativity, or publishers are the best fit for a particular product. The investment is relatively small and allows us to carry out advertising test-runs on the necessary segments. The logic employed for deciding which combinations to rule out is based on the hypothesis test, commonly known as the A/B test.

If we find combinations in which the cost of promoting a product is less than the profit from product sales, we can declare the prospection to have been successful, and can then proceed to promote this product on a large scale.

In the event that a prospection is not successful, PSP will continue updating us on promotion costs according to different combinations of audience/creativity/channels/publishers, in a way that ensures that even though the process may not be profitable on this occasion, any loss is by a small margin. It will do this by trying to refine some of the variables, including improving the product's purchasing price, trying out other creative executions, and so on.

The platform operates according to a four-phase process:

Build

PSP initiates the prospection process for a given product. To do this, the platform links a campaign to the corresponding product, uploads the relevant creativities, and assigns a series of basic parameters according to which the prospecting will be carried out; such as start date or budget.

To build the prospection, PSP will create a specific number (N) of demographically segmented campaigns, factor in a given number (M) of creativities (banners) for each, to be published in a limited number (X) of formats.

To do this, PSP will connect with various DSPs via API, and thence gain access to the inventories of different Ad Exchanges that in turn manage the inventory of a large proportion of publishers.

PSP then launches N campaigns in parallel, and begins to gather information from each and every impression.

Measure

Using the information gathered, the platform begins to calculate all of the metrics necessary for applying statistical logic. To lend greater statistical validity, all calculations are based on the concept of a 'unique impression.' The 'unique impression' concept is comparable to the 'user concept.' When the same user sees the same banner three times, Ad Servers will register three discrete impressions. Conversely, the platform will register a single impression. This decision is more onerous in terms of calculus, given that it precludes the counting of impressions according to how many times they are produced. Instead, a more arduous computation must take place each time an impression is produced. This significantly elevates the complexity of real-time processing. It is worth stressing that every single metric handled by the platform is real. The platform never makes use of statistical estimators for calculable variables.

PSP gathers the following metrics:

- Impressions received per user
- Clicks per user
- Unique impressions
- Unique clicks

These metrics are aggregated:

- By segment
- By advert
- By publisher
- By URL

This task is carried out in what could be considered real-time. Five seconds after an impression is produced, or a click is executed, the platform aggregates the event according to each individual criterion.

Optimise

PSP enhances the information it gathers by applying logic according to the following optimization drivers:

- *Bidding Engine*

  Every five minutes, each segment is analysed according to 'speed.' By 'speed,' we mean the number of impressions per hour that a given campaign is achieving. If the speed doesn't meet the minimum requirements defined at the outset of prospection, we seek to improve it, and vice versa. We control the number of impressions our campaign generates by adjusting our bids in the real-time auction, raising or lowering them as necessary.

- *Fraud Detection Engine*

  One of the problems facing the online advertising industry is the ongoing proliferation of automated systems (bots) that generate impressions and execute clicks on adverts.

  These impressions and 'false' clicks artificially inflate the impression-count, which allows publishers to lay claim to a greater volume of (albeit false) inventory, which they can then sell to Ad Exchanges. Because bot-generated impressions and clicks are clearly unproductive, it is vital that we are able to detect them and, ultimately, eliminate them from our statistical calculations. This is where the Fraud Detection Engine comes in.

  Within the platform's memory, a background program (or daemon) runs an iterative process, reviewing web traffic to seek out patterns that have previously been defined as fraudulent. Some of these patters include detecting the following within a certain period of time:

  - An impression-count that is too high to have been generated by a single user.
  - A click-count that is too high to have been generated by a single user.
  - A click: impression ratio, or Click-through Rate (CTR) that is too high to have been generated by a single user
  - An impression-count that is too high to have been generated by a single IP.
  - A click-count that is too high to have been generated by a single IP.
  - An impression-count that is too high to have been generated by a single IP.
  - Click/impression ratio or Click-through Rate (CTR) that is too high to have been generated by a single IP.
  - A break between impressions that is too short to have originated from a single user

Scale

If prospection for a product proves successful, the platform will increase advertising spend on that particular product for the relevant segments.

V. The Platform: architecture and operation

Based on all of the above, PSP can be described as an exploration tool that optimizes programmatic inventory sales using real-time bidding (RTB) mechanisms [19].

The process begins when an advertiser decides to create a prospection for promoting a product (Fig. 2).

The platform will then configure the DSP and Ad Server in order to be able to purchase the desired inventory [10].

The impression purchasing process takes place via RTB. Each impression purchased comes with a label supplied by the Ad Server. This label is known as an Ad Tag. Ad Tags are responsible for delivering information to the client's web browser in order to serve the advert, track user-related information, and monitor the user's interaction with the advert. Big Data machines then process all of the information gleaned about a user's interaction with the advert. This process takes

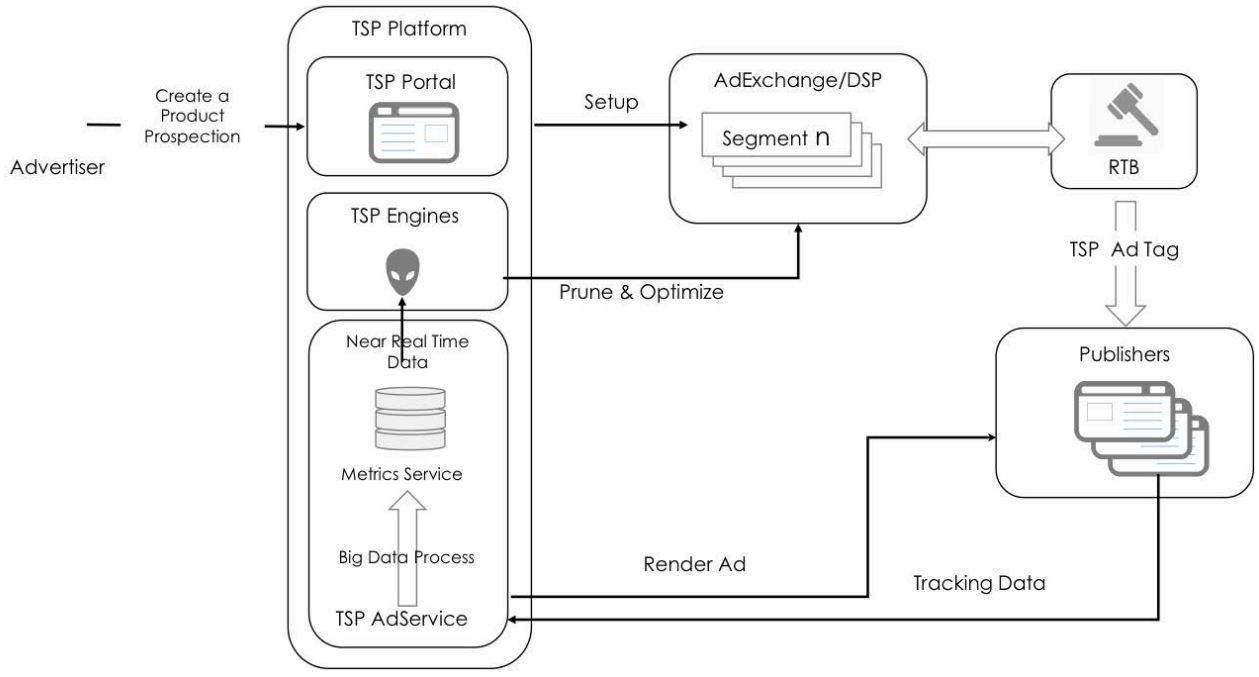


Fig. 2. PSP High Level Architecture

place in Near Real Time (NRT).

Subsequently, the platform's engines then essentially operate in real-time, optimizing inventory purchase, with a view to identifying the segments, adverts and publishers that offer a maximum return on investment (ROI).

Big Data Process

The platform's core is built on Big Data architecture [5]. The rationale for this decision follows on from analysis of the nature and characteristics of the problem at hand. As always, when planning a campaign, we have at our disposal a total of eighty-four possible segmentation combinations; comprised of two gender options, six age ranges and seven HHI (Herfindahl–Hirschman Index) categories. Adding on the various geographical areas (the Designated Market Areas, or DMAs), we would then need to multiply the number of possible combinations by the number of DMAs that exist for a particular country or region. For example, according to Nielsen, the US has 210 DMAs, which would give us 17,640 segmentation combinations. Bearing in mind that digital marketing conventions hold that each individual has between one and one hundred areas of interest, the number of combinations then rockets into millions. In short, the volume of information that needs to be stored and processed is very high [7].

Additionally, given that the platform requires integration into programmatic advertising sales mechanisms, we need to take into account the specific requirements of those mechanisms. The life cycle of a programmatically purchased advertising impression, that is to say, the length of time between a user's arrival at a publisher's page, and the appearance on screen of an advert purchased by auction, is a mere two hundred milliseconds. This means the system has only two hundred milliseconds within which to receive the request, process and analyse the data make a decision and, finally, serve the creative execution. In other words, speed is a key requirement for our platform [3].

In addition, the platform needs to be able to store and process information in a way that satisfies the following requirements:

- Scalability. The platform must be able to execute N prospections simultaneously without requiring any technical adjustments.
- Low latency. Given the fast-paced nature of the RTB market, in which circumstances change from one minute to the next, the platform must be able to analyse events in real-time. Basing a decision on data that are more than five minutes old would, in most cases, be extemporaneous.
- Analytics. While predefined tracking data aggregations do exist (for example the number of impressions and clicks per hour per advert), the platform must also offer a solution that allows data to intersect with the utmost versatility, without needing to pre-process results. This is a fundamental prerequisite for ad hoc analysis – whether seeking to understand user behaviour, find patterns, identify fraud, and so on.

In order to confront these challenges and satisfy these requirements, the platform has been built on Big Data Lamba Architecture. As well as being robust and exact, this architecture allows the platform to combine background (batch) processing with stream processing. While stream processing does not allow for the same level of precision as batch processing, the platform is still able to process and deliver data with very low latency, and continue making decisions practically in real-time.

Flexibility offers further rationale for implementing a Big Data solution; this type of architecture allows virtually infinite scalability. Nevertheless, it is also possible to work with a small and cheap data cluster and incorporate or remove nodes in a straightforward manner as and when necessary. At the same time, technology exists that would allow the platform to process user demographic information, and to bid higher or lower per impression according to the desirability of a given demographic.

Everything described so far takes place in a programmatic manner, in real time, and by auction (with each impression attracting the interest of a number of different advertisers, and the lot going to the highest bidder).

Objectives

The PSP platform's objective is to help advertisers connect with DSPs in order to maximize their campaigns' ROI. In an automated way, the platform is able to create and launch segmented campaigns and, later, to monitor and manage them in real-time without the need for direct supervision. For each product the advertiser promotes, the platform's goal is to identify the following:

- The most appropriate audience
- The optimal channels and formats
- The best creative execution
- The lowest purchase price

Architecture: modules and their interrelatedness

There are three distinct layers to the platform; each layer is charged with a specific function and is capable of interacting and communicating with the other layers, as well as with any external entities upon which service delivery depends.

PSP Portal

The PSP Portal, the platform's management console, takes the form of a website built in ASP.NET MVC and deployed within Microsoft Azure. The site acts as an interface through which the advertiser interacts with the platform. The console permits campaign configuration (setting benchmarks/thresholds, budget, creative executions and so on). This information is then disseminated directly to the relevant DSPs as well as to the Ad Server, removing the need to access these via their respective interfaces.

The console also tracks the campaign's progress and evolution, and logs this information for each campaign.

PSP Core Engines

The PSP Core Engine is the heart of the platform, and the seat of its intelligence. This layer houses the platform's logic and orchestration capabilities, allowing it to co-ordinate the operation of all the elements involved, both in-house and third-party, including the Ad Server, DMPs, DSPs and so on.

This logic layer is composed of a series of processes (or daemons) that continually run algorithms that process data generated by campaigns. These algorithms allow the advertiser to take timely and wise decisions.

This layer is fully hosted in the cloud (Microsoft Azure Cloud Services). Making use of REST APIs from both the Ad Server and from various DSPs, the layer's daemons interface with the former according to decisions taken by the advertiser.

Ad Server

The role of the Ad Server is to make the adverts appear on the screen by supplying the required information to the client (typically the end-user's web browser, delivered via JavaScript).

The Ad Server also supplies the client's monitoring and identification mechanisms, in order to track both the user's identity, and their interaction with any creativity.

A highlight of the Ad Server's features is its ability to generate near-real-time statistics on ongoing campaigns, thereby providing the PSP with the necessary metrics to make decisions.

The Ad Server is configured as a cluster and, like the PSP, is also cloud-hosted (Microsoft Azure Cloud Services). The cluster's machines are balanced through a layer four load balancer, and continually monitored to verify uptime. The system has an elastic response to the platform's workload, so that if the number of requests rises abnormally within a short space of time, more servers can be automatically added to the cluster. In the same way, if the workload is reduced over a sustained period, machines can be eliminated from the cluster in a tidy manner – without missing out on a deal.

Both the Ad Server and the platform are designed in such a way that the time it takes to process a request and serve an impression is always of the shortest possible duration.

The Platform's Features

1) Initial storage of advert-interaction data.

As soon as the tracking information reaches the Ad Server, it is saved in the Event Ingestor, via a Publisher/Subscriber mechanism that allows the same event to be used by different processes at varying paces (Figure 3).

The platform can either utilize a cloud service, such as Azure Event Hubs or Kinesis, or, alternatively, Storm (part of the Hadoop stack). Given the dynamic nature of workload, the use of one of the aforementioned cloud services is highly recommended.

A daemon consumes events as and when they are generated, caching them in a redundant storage system each time the system logs one thousand requests.

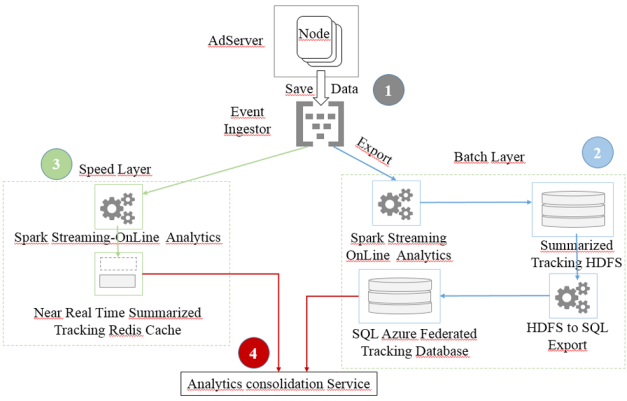


Fig. 3. Platform Features.

2) Batch Layer

Batch processing is tasked with processing all of the information available within a given period of time. This form of information processing delivers accurate aggregated data. The Batch layer consists of an Apache Hadoop cluster distributed by Hortonworks, and all of its processes are orchestrated by Apache Oozie workflows.

The tracking data captured and stored in the previous step is then downloaded in a distributed manner within the cluster's file system. The system proceeds to pre-process, standardize and debug the new logs before finally storing them in HDFS (Hadoop Distributed File System).

In order to improve performance, the file-system is partitioned according to date (via HCatalog). This allows us to easily access and read the logs from a particular time period, without needing to scour through every single log stored in the system.

After the new logs have been stored, the system begins to execute the various processes that are written in Scala on Apache Spark. These processes serve to create aggregated, pre-calculated metrics. Once this data has been calculated, it is inserted into a combined SQL database. This database is also partitioned by date, thereby allowing the data to be accessed and processed rapidly.

3) Speed Layer (stream)

Stream processing allows us to consume the events stored in the Event Ingesters and to process them with very low latency.

The real-time information-processing application is based on Spark Streaming, using the AMQP protocol for its integration. A Redis database stores information for real-time data processing.

4) Consolidation Layer

In order to be able to take automatic decisions for campaign-optimization, or to be able to display campaign-performance reports, the data must be consumed in near-real-time.

To this end, we have rolled out a REST service, which allows us to check and review the filter and aggregation as desired.

Upon receiving a request, the consolidation layer will receive data while carrying out a parallel consultation on the SQL databases previously fed by the batch processing layer. However, if the query refers to data not yet stored in these databases, the service will access the Redis database, fed by the streaming layer, in order to complete the required information.

VI. Testing the Platform: Main Results of the Pilot

In order to test the platform, we will set up a prospection. To illustrate the platform's operating mechanism with maximum clarity, we will simplify our desired objectives and analyse the results that emerge.

Our prospection focuses on the optimization of two variables: the publishers through which our campaign is to be launched via Ad Exchanges, and the distribution channel. The prospection will be composed of fourteen segments: seven oriented towards web navigation via mobile devices (mobile), and seven oriented towards web navigation via PC (web). The prospection will be endowed with a small but sufficient budget, in order to ensure that the sample is statistically significant. We will not carry out any socio-demographic segmentation, and for didactic purposes will always use the same creative executions.

The platform is integrated with the online advertising market's main Ad Exchanges, giving access to a large quantity of publishers. For our test, we will be using seven Ad Exchanges.

The prospection needs to be able to identify which 'Ad Exchange – Distribution Channel (mobile/web)' combination offers the best performance, and we need to be able to do this in a way that rules out - as quickly as possible - those pairings that would render the campaign's performance significantly lower than average. Campaign performance will be measured according to CTR (Click-Through Rate – the ratio of clicks generated to the total number of impressions).

In order to understand how the platform makes decisions, it is important to understand the statistical logic that is applied. The platform makes decisions by applying contrast hypothesis tests to the population, defined by all segments. Given that we have fourteen segments, we will need to compare each segment's CTR with that of each of the other thirteen, in order to evaluate performance. This means that when analysing segment number one, we will compare its CTR with the aggregated CTR of the other thirteen segments.

When examining segment i, CTRi will refer to the segment in question, while CTRri will refer to the aggregated CTR for the remaining thirteen segments, ri. Note that both segment i and the aggregated remaining segments ri constitute samples of our population and, furthermore, that all fourteen segments are discrete. This affords us the statistical independence required to carry out our test.

The problem of how to select the best segments can be tackled in a number of ways. We will outline the most straightforward scheme that has been implemented in the PSP. In order to execute the algorithm that we are about to describe, we first need to establish the significance value of the test (a) as well as that of the statistical significance (β) (the Appendix contains a detailed description of these) which we would like to employ in our study. Or, to put it another way, the degree of confidence in our test, and its statistical power. In this case study we have established the values: a = 0.05, which translates to a confidence level of 95%, and β= 0.2, which equates to a statistical power of 80%.

Once we have established these premises, the algorithm for excluding a particular segment i works in the following way:

1. Calculate (or estimate) CTRi and CTRri and thus estimate the CTRi-CTRri magnitude. We will use this magnitude in our hypotheses.
2. Calculate the statistical Z (see Appendix) and ascertain whether the Z value you have obtained falls below the critical value of Z, determined by α=0.05. To achieve the same goal, you can also identify which p-value corresponds to the Z value obtained, and determine whether it is less than -1.645, this being the corresponding p-value to Z=0.05 in queue analytics. If it does, the null hypothesis is rejected and the opposite case is accepted.

Some of the results obtained for total impressions, clicks and CTR by week, for each segment, are collected in table I. These results were obtained following the application of the exclusion algorithm over the course of a five-week-long prospection. The platform's algorithms use this data as a starting point for carrying out the calculations necessary to make the required decisions. Table 1 includes some of the calculations that are most relevant and essential for making such decisions.

TABLE 1. MAIN RESULTS

		Week 1		Week 2		Week 3		Week 4		Week 5	
	Segment ID	Z	Power	Z	Power	Z	Power	Z	Power	Z	Power
Mobile	1655571	0,3548	10%	0,5065	13%	0,3547	10%	-0,2122	8%	0,7213	18%
	1655572	-1,6707	51%	0,2315	8%	-0,2551	8%	-0,8562	22%	-1,0543	28%
	1655573	7,0197	100%	8,1893	100%	6,9097	100%	5,8508	100%	7,7165	100%
	1655574	0,7197	18%	2,9603	91%	3,3045	95%	5,3317	100%	4,8030	100%
	1655575	-0,7371	18%	-1,7325	53%	-2,2183	72%	-2,7612	87%	-	-
	1655576	0,8309	21%	-1,1253	30%	-1,7253	53%	-2,3646	76%	-3,7834	98%
	1655577	-0,4254	11%	-0,7204	18%	0,5512	14%	-0,3437	10%	-0,7585	19%
Web	1655579	-25,8250	100%	-	-	-	-	-	-	-	-
	1655580	-1,0663	28%	-4,3884	100%	-	-	-	-	-	-
	1655581	4,2081	99%	2,6964	85%	4,5383	100%	5,5031	100%	6,5208	100%
	1655582	-6,5587	100%	-	-	-	-	-	-	-	-
	1655583	-5,4606	100%	-	-	-	-	-	-	-	-
	1655584	-6,9300	100%	-	-	-	-	-	-	-	-
	1655585	-9,7202	100%	-	-	-	-	-	-	-	-

Ruled out Segments
Not ruled out by lack of power

Table 1 shows the values calculated for Z and for statistical power, organised by week and by segment. In the left-hand side of the table, we find the Segment ID column, which shows the identifier that the platform assigns automatically to each segment. We can also see that the data is organised by distribution channel: the first seven lines correspond to campaigns oriented towards the mobile channel, while the final seven correspond to web channel campaigns.

The colour-coding indicates the status of each segment at the end of each week[1]. According to this colour code, there are three possibilities:

- *Red*: segments excluded due to low performance. For example, the segment identified by 1655583 had a Z value of -5.4606 in the first week (way below α=0.05), and a power of 100%, which means that we have power enough to make a decision. Consequently, the segment was excluded. As soon as a segment is excluded due to poor performance, its execution is halted, so that no more impressions are purchased for it.
- *Pink*: segments which produce low performance in comparison with the aggregate, but for which the analysis has not reached sufficient statistical power to be able to reliably declare a verdict of low performance. For example the segment identified by code 1655580, which had a Z value of -1.0663 in the first week (well below α=0.05), nevertheless only has a statistical power of 28% (way below the minimum of 80%). This means that the result cannot be considered conclusive. Therefore, this segment remains active. However, in the second week, the Z value is still below α, but this week its statistical potency is now 100%, which means that the segment is then excluded, and its execution is halted.
- *White*: segments that show similar performance to that of the aggregate. For example the segment identified by code 1655574 had a Z value of 0.7197 in the first week (way above α=0.05). This segment then continues to be executed.

The evolution of calculations across the weeks shows that the platform is able to make conclusive decisions in accordance with the pre-determined significance and power parameters. We can see how the platform excludes those segments whose CTR is significantly lower than that of the aggregate.

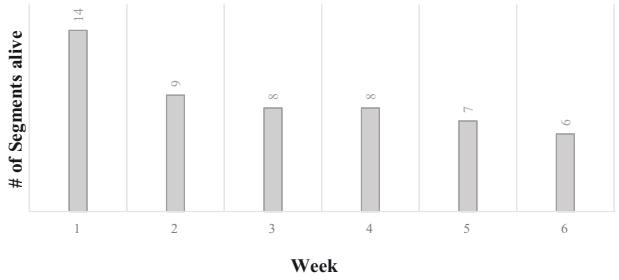


Fig. 4. Segment survival at the beginning of the week.

In figure 4 we can track the survival of different segments on a week-by-week basis, as the algorithm is being executed. At the beginning of week 1, the number of active segments stands at initial fourteen segments. By week 2, the platform has excluded five segments and only nine active segments remain. At the beginning of week 3, eight active segments remain, and the process continues until only six out of the initial fourteen segments have survived by the beginning of week 6.

Of those six active segments that survive past the end of week 5, it is worth highlighting that three (1655573, 1655574 and 1655581) have Z values far above those of the exclusion value a, coupled with high statistical power (100%). One surviving segment (1655571) has high Z values but low power (18%), and two (1655572 and 1655577) have Z values below a, but with power insufficient to lead to exclusion.

The behaviour of the two latter segments is noteworthy. Almost throughout the entire prospection, these two segments maintain consistently low performance, but never attract the statistical power sufficient to lead to their exclusion. What is happening here is either that the segment's CTR values are patchy or uneven, or that they are very close to, or far from, the aggregate CTR. In this situation, we would need to use very large sample sizes if we are to achieve sufficient power to enable us to make a decision. This situation is highly likely to occur, due to the fact that campaigns are subject to price competition within the auction-purchasing model. That is, when the platform goes to an auction in order to purchase impressions for a particular segment, either there is no available inventory or, more

1 In fact, the platform does not make decisions on a weekly basis, but continuously, however this way of introducing information helps us to understand the mechanism of operation thereof.

likely, the auction concludes with an impression purchase price that exceeds the maximum budget established when the campaign was configured. In other words, the platform does not make the purchase. Ultimately, the platform is either unable to purchase impressions, or the traffic it achieves is not of sufficient quality.

Given that the algorithm excludes those segments whose CTR falls below that of the aggregate, the first effect that we can observe is that the prospection's global CTR increases on a weekly basis. Thus, in figure 5, we can see how the prospection's global CTR began at 0.0011 in the first week, concluded at 0.0025 in the fifth week. This is a direct result of the process of excluding lower- performance segments.

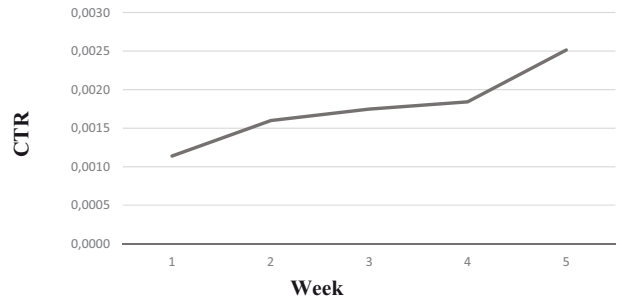


Fig.5. CTR per Week.

A key result of the segment-excluding process is that the mobile communication channel appears to be more appropriate and effective than the web channel when it comes to carrying out campaigns for the product associated with this prospection. In accordance with the data in table 1, we can see that, of the first seven segments that relate to the mobile channel, five survive until the very end of the prospection. Of the web channel segments, only one survives the prospection's full five weeks.

From a business perspective, it is vital that we identify whether the prospection delivers a clear economic benefit. To this end, we will need to evaluate the cost and revenue that the prospection generates. Costs are determined by purchases that the platform makes in the various auctions to which it has access. The information available will permit the establishment either of a global prospection cost for a defined period of time, or of a cost per click.

Revenue will result from conversions; from each and every click that results in a real-life product sale.

In order to gauge the prospection's benefits or performance, we can simply compare total revenue with total cost. Figure 6 shows the evolution of revenue vs. cost throughout the six weeks of the prospection.

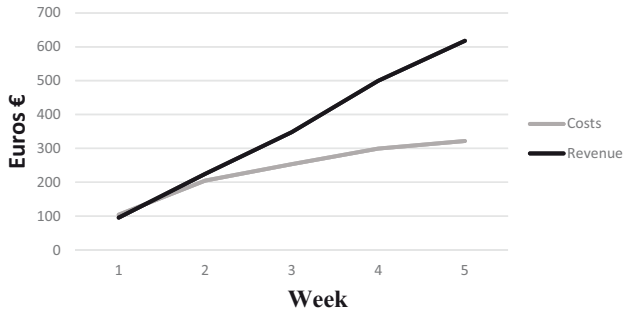


Fig. 6. Costs and Revenue.

As the platform discards segments with a worse outcome the overall campaign CTR grows, which means that the same number of impressions will give us, week after week, a greater number of clicks (figure 5), and logically a greater number of conversions (clicks that ends in a purchase). In short, the selection process is choosing those segments that are more effective.

Costs increase as we continue purchasing impressions for successful segments. In some cases at a higher price due to the bidding process, but as the effectiveness is better, costs grow at a lower rate than revenues do.

As the platform filters segments that are more efficient, the overall number of impression grow; if the number of clicks is higher the number of conversions is higher, and therefore revenues grow at a higher ratio.

VII. Conclusions

This paper has tested the application of Big Data techniques and tools, as well as Artificial Intelligence algorithms, to the online advertising purchase process. Specifically, a platform developed in-house and integrated in real time within the online advertising ecosystem, has been employed. The platform aims to find those segments in a particular campaign with the best performance (the champions), i.e. those that maximize ROI.

This system allows advertisers to run prospections, prior to executing the real campaign, in order to find those champion segments that maximize results. The advantage of running a prospection is that advertisers will find the better segments with a relatively low investment. Finding the champion segments will allow advertisers to spend their money in proven successful segments.

It is worth highlighting that the platform works without the need for direct supervision, and delivers information in a simple, user-friendly manner. This greatly simplifies the workload of personnel engaged in campaign planning, as well as that of staff on the product side. Traditionally, these professionals would have had to access a range of information sources (the interfaces of various Ad Servers or DSPs), with the added complication that this information would burden non-technical personnel with an unwieldy amount of calculation sheets.

Results from the pilot prospection have shown that the platform algorithms are able to determine, rapidly and with enough statistical significance (enough statistical power), which segments are more appropriate for a single campaign, because they are more profitable,

The analysis shows that the platform was able to discard segments with poor CTR. As a result of these decisions taken automatically, the overall CTR of the prospection increased week after week and consequently the overall profitability of the campaign improved significantly.

VIII. Future works

When it comes to possible future investigations, it is work noting that there are a number of ways of improving the platform, among which these are key:

1. Given that one of the determining factors of the platform's statistical analysis is the minimum sample size, it would be advantageous to identify reliable ways of optimizing this process; establishing smaller samples that maintain reliability and still reach the statistical power necessary for making decisions. In this sense, it is worth pointing out the possibility of using methods based on the Bayes theorem, in order to establish minimum sample size.
2. In spite of the fact that CTR can, as we have seen, be an acceptable method of judging a campaign's performance,

from a business perspective it makes more sense to apply an economic performance criterion, one known as performance display. Nevertheless, this selection criterion, while desirable, would significantly complicate both the volume of data and the calculations required.

Future lines of investigation should be guided towards these two paths of action.

Appendix – Applied statistics

The central idea of the platform consists of excluding segments as soon as sufficient statistical evidence supports that CTRi is lower than CTRri.

It is important to note that, while we have been working with CTR (click-through rate, or the relationship between clicks generated and total impressions), this value is a ratio. We could also affirm that this ratio is in fact a probability, given that the fact that a user clicks on an advert can be considered a success, and similarly, the creative execution's total impressions can be viewed as the total number of possible cases. Given that we are comparing the samples' ratios, and are only interested in knowing whether the ratio of one is lower than another, what we are in fact carrying out is a contrast hypothesis test between the ratios of two samples, within only one-tailed test that of the left.

We can present our hypothesis in this way:

Nul hypothesis H_0: CTRi = CTRri

Alternative hypothesis H_1: CTRi < CTRri

Or, alternatively, we could present it thus:

Nul hypothesis H_0: CTRi – CTRri = 0

Alternative hypothesis H_1: CTRi - CTRri < 0

That is to say, we will establish that our nul hypothesis H_0 holds that the sample's CTR (CTRi) is equal to the CTR calculated for the aggregate of the remaining segments (CTRi), and that, therefore, there is no difference between the two. Our alternative hypothesis H_1 maintains that the sample's CTR is lower than the CTR calculated for the aggregate of the remaining segments (CTRri) and that, therefore, there is a difference between the two.

Clear evidence against the null hypothesis and in favour of the alternative hypothesis consists of a CTRi value lower than that of CTRri or, in the same way, a difference that is significantly below zero. The reason we need to employ a contrast hypothesis is precisely because it is difficult to pinpoint the meaning of the term 'significantly,' particularly when sample size is neither fixed nor known, and, furthermore, when there is a significant difference between the sample sizes we are comparing [21]. While it is possible to observe differences in both directions, namely CTRi – CTRri <0 as well as CTRi – CTRri > 0, in reality we are not concerned with whether the segment is significantly more effective than the others, but only with whether it is less effective, in which case it is excluded. Under the same conditions, a one-tailed test offers greater statistical power than a two-tailed test.

The concept of statistical power refers to the reliability of a test when it comes to preventing misguided decisions. Power refers to the probability of rejecting the null hypothesis H_0 when the alternative hypothesis H_1 is true. This is also known as the probability of committing a type II error. [20]

In our scenario, this means that, provided that the sample size is sufficiently representative, the process can be executed sooner and, therefore, a decision can be made sooner. The power of a test depends on the relationship between the sample sizes compared, and the level of significance α established. It is vital that the test has sufficient statistical power to allow us to make decisions with an acceptable margin of error.

We are analysing independent events (the fact that a user clicks does not depend, and does not affect, the behaviour of another user), with discrete variables (a user does or does not click; there no intermediary values), and thus our variable - CTR - will follow a Bernoulli distribution. This allows us to establish variance as: σ=p.q or σ= p (1-p). This amounts to saying σ as the product of the probability of success with that of failure. As stated previously, CTR remains a probability. In this way, we could state that CTR is the probability that a user will click, given that it constitutes the ratio between successes (clicks) and all possible cases (impressions), and is nothing more than CTR. In other words, variance can be expressed thus:

$$\sigma=CTR*(1-CTR) \tag{1}$$

Taking standard deviation into account, we can calculate the standard error as:

$$SE = \sqrt{\frac{\sigma}{n}} = \sqrt{\frac{CTR*(1-CTR)}{n_i}} \tag{2}$$

Given that the sample size of our population n is huge, and in accordance with the central limit theorem, we can reconcile this distribution to a normal distribution provided that the population average and variance have finite values, and as long as $n \cdot p > 5$ and $n \cdot q > 5$ is true. In the preceding equation, n is the population size, p is the probability of success, and q the probability of failure.

This fact will facilitate the process of choosing which statistic to employ when carrying out our calculations. For this purpose, we will use the Z statistic (also known as Z-score), unlike the case of two proportions, calculated as:

$$Z = \frac{P_i - P_{ri}}{SE_{i-ri}} = \frac{|CTR_i - CTR_{ri}|}{\sqrt{\frac{CTR_i*(1-CTR_i)}{n_i} + \frac{CTR_{ri}*(1-CTR_{ri})}{n_{ri}}}} \tag{3}$$

Once we have calculated the Z value, the following formula will tell us the test's power[22]:

$$Power = 1 - \beta = \Phi\left(Z - \Phi^{-1}(1-\alpha)\right) = \Phi(Z - Z_{1-\alpha}) \tag{4}$$

In which Φ() y Φ^{-1}() represent, respectively, the function of normal standard distribution, and the reverse; $Z_{1-\alpha}$ represents up to the (1-α) quantile of Φ(.), that is, the Z value to its left (the area below the curve) is equal to 1-α; β is the type II error, and α is the significance, or type I error.

$$n_{ri_{min}} = \left(\frac{CTR_i(1-CTR_i)}{k} + CTR_{ri}(1-CTR_{ri})\right)\left(\frac{Z_{1-\alpha}+Z_{1-\beta}}{CTR_{ri}-CTR_i}\right)^2 \tag{5}$$

In which k is the relation between the studied segment's sample size and the remaining segments, that is:

$$k = \frac{n_i}{n_{ri}} \text{ and } \quad k = \frac{n_{i_{min}}}{n_{ri_{min}}} \tag{6}$$

Acknowledgment

The platform used in this paper has been developed and deployed by the company Zed Worldwide S.A thanks to whom this article was made possible. We would like to thank, as well those people responsible for the platform, for their recommendations, support and help in the

elaboration to make this paper. Especially to Fernando Perez, Head of Big Data, Alberto Ochoa, Data Scientist, Juan Pablo Rizzo, Platform architecture and William B. Henn Innovation Director.

References

[1] S. Adikari and K. Dutta. (2015) Real Time Bidding in Online Digital Advertisement. In New Horizons in Design Science: Broadening the Research Agenda.10th International Conference, B. Donnellan et al. (Eds.), pages 19-38. Springer.

[2] O. Busch (Editor) (2014). Programmatic Advertising. The Successful Transformation to Automated, Data-Driven Marketing in Real-Time. *Springer Science+Business Media, 2014.* http://www.springer.com/us/book/9783319250212

[3] G. Cormode, and S. Muthukrishnan. (2005). An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms. Volume 55, Issue 1, April*, Pages 58–75.

[4] K. Cox. (2015). Programmatic Helps Brands Make the Most of Micro-Moments. Product Marketing Manager. DoubleClick by Google. https://www.thinkwithgoogle.com/articles/programmatic-helps-brands-make-the-most-of-micromoments.html

[5] L.F. Chiroque , H. Cordobés, A. Fernández Anta, R. A. García Leiva , P. Morere , L. Ornella , F. Pérez and A. Santos. (2015). Empirical Comparison of Graph-based Recommendation Engines for an Apps Ecosystem. *International Journal of Interactive Multimedia and Artificial Intelligence. March, 2015* http://www.ijimai.org/JOURNAL/sites/default/files/journals/IJIMAI20153_2.pdf

[6] U. M. Dholakia. (2015).The Perils of Algorithm-Based Marketing. *Harvard Business Review. June 17*. https://hbr.org/2015/06/the-perils-of-algorithm-based-marketing&cm_sp=Article-_-Links-_-End%20of%20Page%20Recirculation

[7] W. Fan and A. Bife. (2012).Mining big data: current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter*. Volume 14 Issue 2, December, Pages,1-5.

[8] D. Field, P. Zwillenberg, J. Rosenzweig, N. Zuckerman and M. Ruseler. (2015).The Programatic Path to Profit for Publishers. *BCG Boston Consulting Group. July*.

[9] D. Field, O. Rehse, K. Rogers, and P. Zwillenberg. (2013). Efficiency and Effectiveness in Digital Advertising. Cutting Complexity, Adding Value. *BCG.perspectives*. https://www.bcgperspectives.com/content/articles/media_entertainment_marketing_cutting_complexity_adding_value_efficiency_effectiviness_digital_advertising/

[10] A. Gandomi, and M. Haider. (2015).Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management Volume 35, Issue 2, April*, Pages 137–144.

[11] IAB Spain. (2014) Libro blanco de compra programática. IAB Spain, 2014. http://www.iabspain.net/wp-content/uploads/downloads/2014/09/Libro-blanco-de-Compra-Program%C3%A1tica-y-RTB.pdf

[12] P. Minnium. (2014).The Definitive Guide to the Digital Display Ad Ecosystem. Marketing Land. October. http://marketingland.com/digital-simplified-new-advertising-supply-chain-104734

[13] A. Mochón, and Y. Sáez. (2015). Understanding Auctions. Springer Texts in Business and Economics. http://www.springer.com/br/book/9783319088129

[14] J. F. Rayport. (2015). Is Programmatic Advertising the Future of Marketing? *Harvard Business Review. June.* https://hbr.org/2015/06/is-programmatic-advertising-the-future-of-marketing

[15] O. Shani. (2014). Get With The Programmatic: A Primer On Programmatic Advertising. Marketing Land. August. http://marketingland.com/get-programmatic-primer-programmatic-advertising-94502

[16] M. Shields. (2014) CMO Today: Marketers Puzzled by Programmatic Advertising. *The Wall Street Journal. April* http://blogs.wsj.com/cmo/2014/04/01/cmo-today-marketers-puzzled-by-programmatic-advertising/

[17] M. Stange, and B. Funk. (2014) Real-Time Advertising. *Business & Information Systems Engineering, Volume 6, Issue 5*, pages 305-308. http://link.springer.com/article/10.1007%2Fs12599-014-0346-0

[18] P. Zwillenberg, D. Field, M. Kistulinec, N. Rich, K. Rogers, and S. Cohen. (2014). Improving Engagement and Performance in Digital Advertising. Adding Data, Boosting Impact. *.BCG.perspectives. September* 16. https://www.bcgperspectives.com/content/articles/marketing_digital_economy_improving_engagement_performance_digital_display_advertising

[19] A. Woodie. (2015).What's driving the Rise of Real-Time Analytics. *Datanami. September*. http://www.datanami.com/2015/09/15/whats-driving-the-rise-of-real-time-analytics/

[20] Ioannidis JPA (2005) Why Most Published Research Findings Are False. *PLoS Med 2(8): e124.doi:10.1371/journal.pmed.0020124*

[21] Martin Goodson. (2014) Most Winning A/B Test Results are Illusory. *Qubit* http://www.qubit.com/sites/default/files/pdf/mostwinningabtestresultsareillusory_0.pdf

[22] Matthew S. Nelson, Alese Wooditch, Lisa M. Dario (2015). Sample size, effect size, and statistical power: a replication study of Weisburd's paradox. *Journal of Experimental Criminology March 2015, Volume 11, Issue 1*, pp 141-163

Juan Carlos Gonzalvez has a Bachelor Degree in Chemistry from Universidad Autonoma de Madrid and he's got an MBA from IE Business School. He counts on more than eighteen years' experience in Telecom, Technology, Media and Internet sectors. He is currently Chief Innovation Officer at the Zed Worldwide Spanish multinational company, where he is currently working in the following innovation lines: Big Data, Advertising, Mobile Payments and Mobile Financial Services or Security and Privacy among others. He leads several R&D projects, in collaboration with different European universities, funded by various public and semi-public Spanish and European Institutions, under the multiple existing R&D program aids, like Horizon 2020

Francisco Mochón has a PhD in economics from the Autonomous University of Madrid and from Indiana University and is a Fulbright scholar. Currently he is full Professor of Economic Analysis at UNED, Madrid. He has been Advisor to the Ministry of Economy and Finance of Spain, Director General of Financial Policy of the Government of Andalusia, CEO of the research firm ESECA and Chief Financial Officer (CFO) of Telefónica of Spain. He has been the Chairman of the Social Board of the University of Malaga. Currently Prof.dr. Mochón is a member of the advisory committee of U-TAD and member of the advisory committee of the Futures Market of Olive Oil (MFAO). He has published numerous research articles and is the author of more than fifty books on economics, finance and business. Currently his research interests are the Economics of Happiness in the business environment and the Digital Economy. He has been director of the MOOC course "Felicidad y práctica empresarial".

PInCom project: SaaS Big Data Platform for and Communication Channels

Juan Manuel Lombardo, Miguel Ángel López, Felipe Mirón, Susana Velasco, Juan Pablo Sevilla and Juan Mellado

Fundación I+D del Software Libre, Granada, Spain

***Abstract* — The problem of optimization will be addressed in this article, based on the premise that the successful implementation of Big Data solutions requires as a determining factor not only effective -it is assumed- but the efficiency of the responsiveness of management information get the best value offered by the digital and technological environment for gaining knowledge. In adopting Big Data strategies should be identified storage technologies and appropriate extraction to enable professionals and companies from different sectors to realize the full potential of the data. A success story is the solution PInCom: Intelligent-Communications Platform that aims customer loyalty by sending multimedia communications across heterogeneous transmission channels.**

***Keywords* — Optimization, NoSQL, Relational DBMSs, Big Data Analytics.**

I. Overview

In the Fundación I+D del Software Libre (Fidesol), as a research center, we do analysis and experimentation with technological trends to improve the achievement of project goals as well as provide new elements of innovation research organizations and SMEs. We have a research line focused on Big Data and optimization analysis of large data sets to provide our beneficiaries generating value and improving business processes.

Research in "Advanced Analytics of Big Data" [1] is essential to help organizations consolidate technological advantage obtained from the discovery of patterns, trends and useful information to help develop reliable diagnostic and prognostic in many areas.

Companies that take advantage of the potential of data reach a better position to plan, strengthens their capabilities for knowledge discovery and the prediction of behavior in uncertainty scenarios. Knowledge discovery in databases continues extending to almost every field where large amount of data are stored and processed (databases, system logs, activity logs, etc.) [2].

Potential of Big Data was shown in a Mckinsey [3] market study. For example, retail trade, using the full capability offered by the Big Data could increase their profits by 60%.

Also, European governments could save more than € 100 thousand millions in operating efficiency improvements.

According to the EMC Chief Marketing Officer, Jonathan Martin, inside study carried out about digital universe in 2014, conversion of the organizations defined by software companies is supported: "IT must press the reset button and find new ways to optimize storage and take advantage of these masses of data" [4].

We Fidesol create Big Data efficient and quality solutions that respond to the most demanding needs and satisfy the expectations of our customers.

We are introducing PInCom project, an intelligent platform for communication, that shows by using more appropriated advanced strategies and Big Data optimization techniques, that improving of results is possible.

This article is structured as described: Section II shows a summary of PInCom platform, objectives and added value to ICTs area. Under Section III can be found system specifications and an analysis of open source technologies applied for Big Data treatment in this database-focused SaaS platform. Section IV describes system architecture and cloud deployment as well as its benefits. Section V reports the results of the developed pilot to verify and ensure the compliance with the objectives fixed. Final Section VI talks about conclusions extracted from the adoption of this Big Data use.

II. PInCom Project

The team has also developed the "PInCom" project, aligned with a line of research with greater continuity for Fidesol, which is Cloud Computing [5], technological area of interest in our company for its universal and ubiquitous character.

The Cloud is an essential component of any application and, according to analysts, the Cloud services market will grow exponentially in the next few years, along with the Internet of Things industry [6].

Therefore, it is considered that may be of interest to integrate aspects of cloud computing with the treatment of large volumes of data and the potential and versatility of messaging through various formats and platforms, especially on mobile devices, covering another Fidesol major lines, such as R & D applied to mobile technology.

The importance of PInCom project - Intelligent Communication Platform - for Fidesol is based on the quantitative importance of critical mass or target population of the project, ie, the high number of users who have access to those technologies and media that allow access to existing communications networks to receive information of interest. The main objective is to offer a cloud service for ubiquitous marketing and loyalty by sending messages using different protocols and communication channels.

PInCom contribution to ICTs is expressed by:

- The system is scalable and highly available since its inception. This is because the system will be available on a cloud computing platform that allows to increase the computing power of the system as needed.
- It supports sending to several heterogeneous communication services, both sending the same information to all media or, by spreading and diversifying depending on the communication channel and the message you want to provide.
- It is an open platform that easily facilitates the incorporation of new services.

PInCom defines the building and deployment of a high performance Big Data SaaS platform efficiently able to process a large volume of communications. Its cloud architecture provides:

DOI: 10.9781/ijimai.2016.362

- High availability: The system ensures operational continuity, which means that the workload of the system does not affect the number of connections allowed.
- Fault tolerance: The system must be able to continue normal operation after a software/hardware failure.
- Scalability: The hardware features of the system are improved as required, while growing the number of connections or database computing/extension.
- PInCom can dispatch a large number of applications, with low latency, so that it is usable simultaneously by a large number of customers.

The server must meet the objectives of providing a new fully functional platform, able to handle a lot of traffic and send messages by using intelligent routing through the use of standardized protocols, acting as a gateway bursting and managing information security when they connect.

III. Technical Description of the System

The System is a modular application that offers different cloud services:

- User interaction: The PInCom platform provides a web interface that allows complete management of the system after user authentication. On the other hand, allows through a mobile web and native mobile applications, any user with an Internet connection would be able to send communications through the gateway.
- Interconnection with other systems: The system receives requests to send message through standard IT protocols, and transfers these requests to other service providers using their interfaces.
- Intelligent System for optical route detection: PInCom uses heuristics and connections to different nodes in communication networks to select the optimal route for each message, preventing congestion and fault tolerance connection with each supplier.

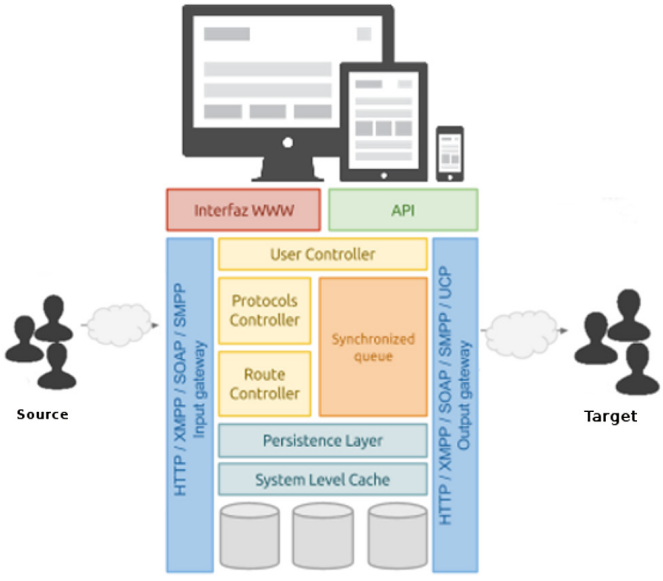


Fig. 1. PInCom's components

PInCom is offered as a modular system that will offer different services deployed in a cloud offering:

- High availability.
- Fault tolerance.
- Scalability.

This will increase the computing requirements as needed.

But to provide this service is not enough to have a conventional clustered database manager considering the huge amount of data to be processed. Due to its specific characteristics (3 Vs) [7], Big Data requires technologies to collect, store and manage very large amount of various data, at high speed in order to allow information optimal processing and get the required results.

A. Analysis of Open Source NoSQL distributed databases

Relational databases have been the choice of many developers and system administrators, opposite to adoption of noSQL-BD, at an early stage. Relational Database Management Systems (RDBMS) are designed for sets of small but frequent transactions of reads and writes, compared with NoSQL databases which are optimized for transactions requiring high load and speed of delivery. The loss of flexibility at run time compared to conventional systems SQL, is offset by significant gains in scalability and performance when it comes to certain data models [8].

The NoSQL BD contains complex information structures in a single record, different information each time, providing polymorphism of data, accessibility and performance, to have the information linked in the same block of data. The ability to process a large amount of data is the ability to run queries type Map-Reduce [9], which can be launched on all nodes at the same time and collect the results to return. In addition, such systems operate in memory and the data dump is done periodically. This allows very fast write operations. Instead the data security is sacrificed, existing the risk of loss and inconsistency that is treated decreasing the time of offset disk dump and validating the writing operations by more than one node. Have to highlight from a list of 150 NoSQL databases: Cassandra [10], with storage capacity for real-time processing and Terracotta [11] that provides organizations through cloud services increased scalability, availability and high real-time performance in their activities and processes.

B. Cassandra and Terracotta systems evaluation results

Researching results (Cassandra and Terracotta) proposed for the new PInCom architecture -whose purpose is to provide the high availability system at all levels and scalability-, are set out below.

Cassandra evaluation

Using a NoSQL database in PInCom and more specifically, Cassandra, arises for two reasons:

- Performance: The performance of the system in the processing of repetitive tasks on the database (query and modify users and communications) may be increased by providing the data cluster of new nodes.
- Stability: The use of a system of distributed databases with automatic data replication and fault tolerant allow PInCom remain stable even when one of the storage nodes collapses.

The main features that make Cassandra an interesting alternative to MySQL according to where more reads and writes are performed are:

- Linear scalability: Performance increases proportional to the number of nodes.
- High Availability: Cassandra's architecture is fully distributed. There are no single points of failure or bottlenecks. All cluster nodes are identical.
- Fault tolerance: To ensure this, the data is automatically replicated. If a cluster node goes down, service is not stopped and housed data inside the node is still available in the cluster.

After testing using Fidesol infrastructure, it was found that: Node-level performance in Cassandra node insertions concerning the volume of data is slightly worse than that in MySQL. This problem can be

solved by distributing data between cluster nodes. This can always be sized so that the nodes do not have an excessive amount of data.

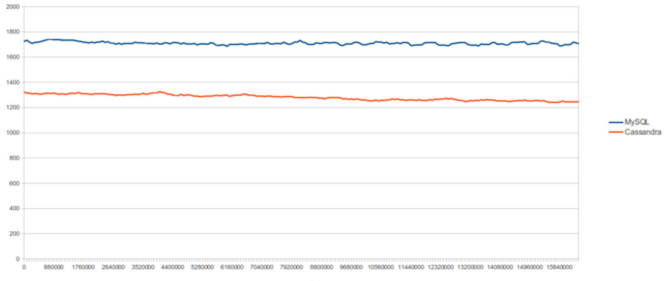


Fig. 2. Cassandra and MySQL queries performance concerning data volume

Read and update operations performance at node level in Cassandra concerning data volume, grows with a lowest rhythm than in MySQL.

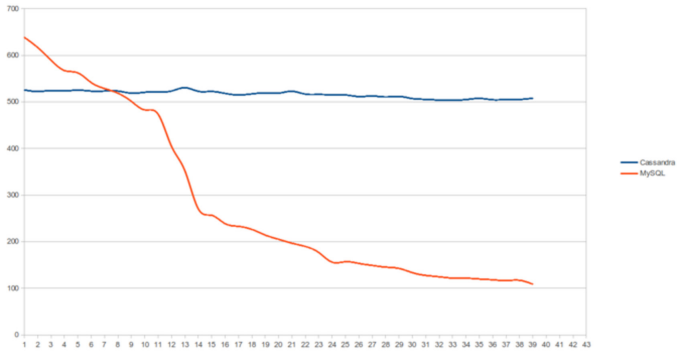


Fig. 3. Cassandra and MySQL queries performance concerning data volume

Cluster-level in Cassandra grows linear, duplicating that in MySQL when using four nodes. This is main advantage for Cassandra, due to cluster performance can be adapted to the load of the system by adding new nodes.

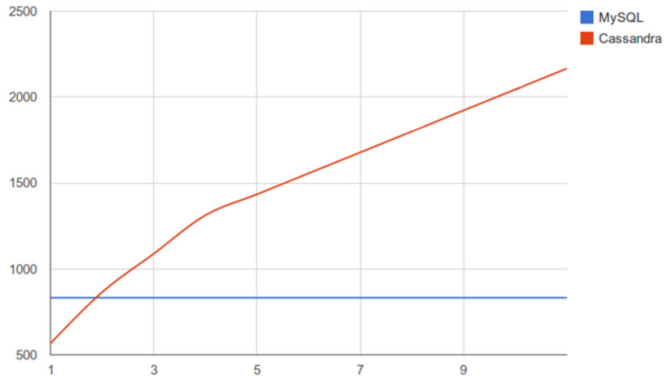


Fig. 4. Cluster-level Cassandra performance concerning number of nodes

This performance is divided by the replication factor (number of replicas of each data stored in the cluster). In tests the value designated for the replication factor has been one.

The stability of a Cassandra node is not higher than a MySQL server when the load limit is maintained. In both cases, the server collapses and, after a given time, the service becomes unavailable. However, since Cassandra offers high service availability and data, this problem at the node level that would lose criticality in a system without high availability.

Terracotta evaluation

Terracotta is proposed as a solution for sharing data between nodes in order to:

- Parallel processing of client requests.
- Establish a mechanism of coordination between processing nodes for distribution of connections to routes, ensuring that for every path, the corresponding connection is established on a single node and the distribution of connections between nodes is consistent, based on some defined criteria (number of connections per node, sum of shipping rates of each node path, etc).

Fidesol testing is focused on the "BigMemory" product in its free version. This version does not have all the functionality offered by the paid version, being limited in cache size allowed, allowed number of customers, etc. However, the basic skills are the same, so it is presented as a viable alternative to test Terracotta as a solution to the sharing of data between processing nodes in PInCom.

Validity is verified, although there is a downside: BigMemory doesn't has the ability to transfer objects partially, so that sharing of data structures of arbitrary size (which will grow accordingly to the increased load on the system) will be a limitation for horizontal scalability.

The following graph shows the evolution of the time required for access to an object in the BigMemory cache depending on the size. BigMemory is found that does not have the partial load capacity of objects, so that the trend is linear. This is a limitation on system scalability, as the growing load of incoming requests to the system will be an increasing load of communication between processing nodes and Terracotta servers.

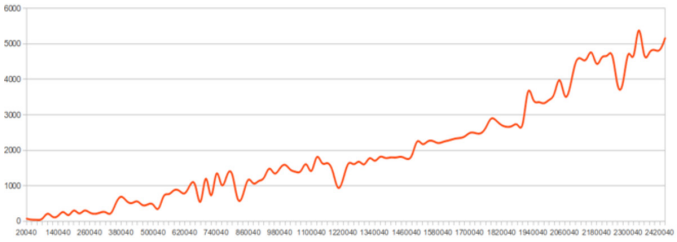


Fig. 5. Time evolution graph to access and object located in BigMemory cache concerning its size.

Therefore, Terracotta BigMemory is not a valid solution. One possible alternative is the use of this software at a lower level and there is not a commercial product for this purpose. Terracotta DSO (Distributed Shared Objects) is the basis of Terracotta products and supports the partial load of certain data structures [12].

This partial data loading allows horizontal-scaling of PInCom processing cluster without overloading the network traffic. However, test results have not been positive: Terracotta DSO can be integrated with Glassfish V2, but not higher, the latter being essential for the deployment of PInCom.

Moreover, Terracotta DSO can be integrated at JVM level. Fidesol testing have not been positive about it, being apparently not feasible such integration in JDK7. In JDK6 it has not been achieved verification successfully.

In addition DSO is no longer available within the packages offered. Because of this, the use of alternative solution for data sharing arises, such as:

- Hazelcast: Apache, supported by benchmarks and success stories.
- Cassandra: it's also considered the possibility of using NoSQL database to house the shared data structure. This would mean a

loss of capacity in terms of timing and consistency that Terracotta and Hazelcast [13] are offering, and must be implemented at the application level in PInCom.

IV. PInCom Architecture

Below can be found PInCom architecture and the benefits of its implementation:

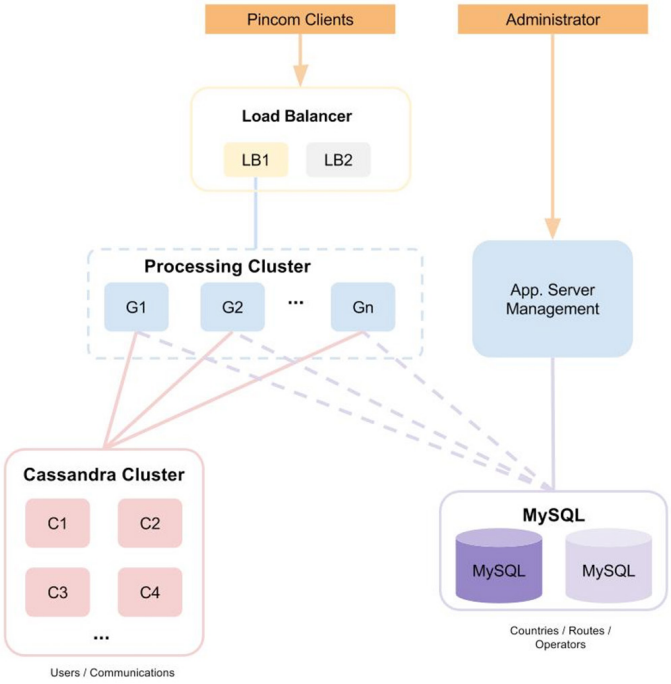


Fig. 6. Proposed architecture

Load Balancer

This machine will be performing balancing tasks toward processing nodes. Balancing will be made at TCP level. Due to task simplicity, it is not necessary a high-performance machine, being enough with the use of a basic machine. However this machine is considered as a single point of failure, so the component goes important for the system high availability.

Processing cluster

All processing nodes (knows as G1 and G2) implement parallel processing requests from clients. Moreover, each connection to a route can only be established in a processing node, so there will be a mechanism for the distribution of routes between nodes. To make this work in a coordinated way, nodes use a series of data structures shared by Cassandra:

- Incoming client-requests queue: This structure contains requests that have been accepted by the system but have not yet been processed.
- Routed queue requests: There is a structure of this type for each system path. Each element inserted into one of these queues contains information of a request made by a client and the list of routing tables that will be used for sending this message.
- Routing table: Processing nodes will use this structure together with a coordination protocol for consistently sharing routes.

Administration server

This machine houses the system management application, where the administrator user can manage the static system information in. Due to its function, it is not required its scalability and so a single machine is dedicated.

MySQL server

MySQL settings will not be different from the proposal so far: two nodes with master-master replication and high availability through heartbeat. This database is used for the following purposes:

- Storage for application management (user management, countries, operators, routes, etc). As the management application has no requirements for scalability, database on which it is support neither does it have.
- Preload of static information for processing nodes. Once this information has been preloaded, processing nodes will only access to Cassandra Cluster for reads and writes.
- Update of static information in processing nodes: When the administration application makes a relevant modification, it will request reload the static information to the processing nodes.

Cassandra cluster

The proposed configuration includes at least 4 nodes and replication factor 2. This will provide high availability and an acceptably high performance.

For the processing cluster to take advantage of data cluster performance, the processing nodes create sets of connections to random nodes of the data cluster and balances the load between these connections. The size of the connection set does not match the data cluster size, as this would compromise the scalability of the system.

It also acts as a synchronization tool between the processing nodes, storing shared structured. The synchronization between the processing nodes and preservation of data consistency is full implemented in these nodes.

V. Pilot Phase

In order to evaluate the PInCom platform functionalities, a pilot phase has been performed to verify the capacity, scalability, performance and efficiency of the system, critical factors to success.

- Verification of the system load for different sizes of processing cluster and data cluster, preserving the scaling ratio between these clusters.
- Evaluation of the system performance with different volumes of requests for different cluster sizes.

A. Project pilot definition

Pilot was built in an iterative way, going from 2 to 12 nodes for processing cluster and Cassandra Cluster, two nodes for MySQL replication and one node for administration service. So we can verify the scalability of the platform and performance that provides this scalability in already mentioned cluster.

The hardware supporting the nodes is as follows:

For processing cluster, servers with 32 Gb of RAM, 8-core, redundant power supplies and RAID 5 have been deployed. In this case RAID 5 has been chosen, for its high performance in reading operations and its high fault tolerance, this ensures that no loss of service if a disk fails.

For Cassandra cluster, servers with 32 Gb of RAM, 8-core, redundant power supplies and RAID 10 have been deployed. We have chosen RAID 10 due to the huge amounts of data that Cassandra writes, as it combines high performance in writing data of RAID 0 and fault tolerance of RAID 0.

For MySQL configuration we have chosen servers with 32 GB of RAM, 8-core, redundant power supplies and RAID 1. We have selected a RAID 1 for it, because storing more static data, high performance is not required and results in cost saving.

Have deployed a server with 32 GB of RAM, 8-core, and redundant power supply and RAID 1 for the management server. Same applies that in MySQL server. It is not required high performance since this server can only be accessed by the system administrator

B. Project pilot development

The following chart shows the maximum load supported by the system, depending on the processing cluster size and data cluster. The number of nodes of both cluster is shown on the horizontal axis -the tests have been performed scaling both cluster at the same time-. In the vertical axis the maximum rate of requests per second that the system can reach is shown.

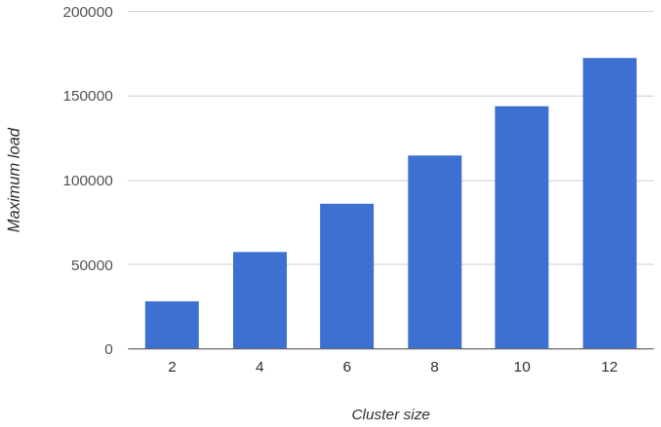


Fig. 7. Maximum load vs. cluster size

The following chart shows the performance of the system to different rates of requests received for different cluster sizes.

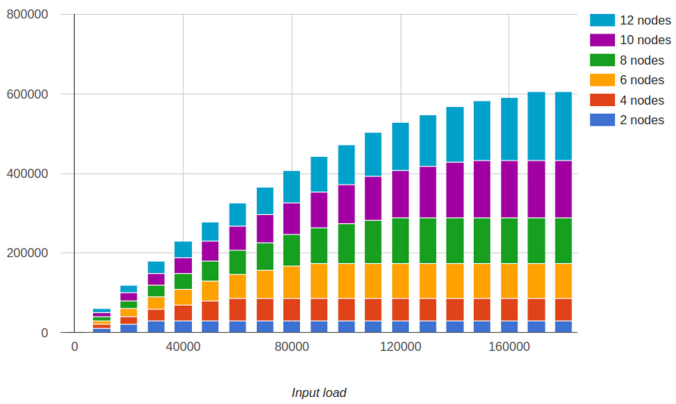


Fig. 8. Throughput vs input load

C. Project pilot results

As shown in the charts, it has achieved the aim of desired scalability and performance since the performance increases almost linearly with the cluster size.

To increase the volume of communications to manage while keeping response level and high-speed of computing capacity, the potential of processing and data clusters should be increased.

In addition, high availability and fault tolerance is managed from computing nodes through copies so that it will able to respond to a fault from database. At the database level with replication is accomplished the same function. A fault in a node will be detectable at runtime and will not compromise the system dramatically.

VI. Conclusion

We have presented the assessment of the information management tools to setting up PInCom, an efficient, available, fault-tolerant and scalable system. Result shows that PInCom is able to deliver a large volume of communications with low latency, so that it can be used by a large number of customers simultaneously. The deployment of multiple computing nodes on an application server allows each node to process different requests in parallel improving overall system speed.

PinCom is fully scalable and system capacity can easily be extended by increasing the number of nodes in the processing cluster, without implying any changes in the code.

Furthermore in system architecture design, has been carried out implementation of non-relational database to manage large volumes of information: The relational databases are not able to handle the size, the speed of delivery nor the diversity of Big Data Formats.

NoSQL databases are the response of scalability and performance that Big Data needs. NoSQL are non-relational, distributed, Open Source and scalable horizontally. This scalability is achieved by increasing the number of nodes also gaining availability. However, not all NoSQL systems have the capabilities required for our communication project.

Before implementing any Big Data technology must study their characteristics and assess whether their qualities correspond with the requirements of the system to be developed.

Also, in the proposed architecture for PInCom it is shown the coexistence between different technologies to obtain the desired configuration. The use of relational database manager is incorporated for the treatment of specific information. Clustered relational database system can solve the problem of inconsistency of own NoSQL system data [14].

NoSQL systems allow management of a large volume of data, but don't offer guarantee of reability that information requires. Its major disadvantage is working with distributed and sometimes replicated data since it gives priority to the availability and immediacy to supply data. It also raises the possibility of creating different versions and inconsistencies [15].

NoSQL systems hardly reaches the consistency of relational databases due to the presence of integrity constraints, concurrency, concurrent updates, etc. Could be possible not has the latest update of the data, which can be somewhat chaotic.

Specialists remarks that NoSQL is an alternative but it does not answer all the data storage issue. It is necessary to keep complex relationships between structured data, and not all NoSQL systems can support it [16].

In the management and analysis of structured and unstructured data based Big Data technologies, NoSQL systems represent a revolution in how they are stored and the data processing is performed.

As Big Data solutions are growing, use cases where relational systems can be attached to the paradigm of this model for optimization and data analysis are discovered.

The key is to choose the most appropriate solutions and integrate them conveniently for the treatment of information very quickly and efficiently in order to get competitive edge shape.

References

[1] L. Columbus, "84% Of Enterprises See Big Data Analytics Changing Their Industries' Competitive Landscapes In The Next Year", 2014. http://

www.forbes.com/sites/louiscolumbus/2014/10/19/84-of-enterprises-see-big-data-analytics-changing-their-industries-competitive-landscapes-in-the-next-year/#1cb8b3863250

[2] J. Campo-Ávila, R. Conejo, F. Triguero, R. Morales-Bueno, "Mining Web-based Educational Systems to Predict Student Learning Achievements", International Journal of Interactive Multimedia and Artificial Intelligence, Vol III, no. 2, España, 2015, pp. 49.

[3] MckInsey Global Institute,"Big Data: The next frontier for innovation, competition, and productivity", 2011. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[4] EMC2, Martin J., "Aprovechamiento del poder del universo digital", 2014, http://spain.emc.com/leadership/digital-universe/index.htm

[5] P.M Timothy, The NIST Definition of Cloud Computing , NIST, 2011. http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

[6] J. Iglesias, G. Amat, "New Challenges on Crossplatform Digital", International Journal of Interactive Multimedia and Artificial Intelligence, Vol III, no. 2, España, 2015, pp 28-29.

[7] IBM, "The four V's of Big Data", 2014. http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

[8] G. Magaña, "El big data y la convergencia con los insigths sociales", 2015, www.bib.uia.mx/tesis/pdf/016079/016079.pdf

[9] IBM, "What is Map Reduce. About MapReduce", 2012. https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/

[10] The Apache Software Foundation, "Cassandra", 2010. http://cassandra.apache.org/

[11] Software AG, Terracotta Enterprise Suite, 2011. http://www.terracotta.org/products/enterprise-suite

[12] O. Letizi, "Terracotta DSO Data Structures Guide", 2011. https://confluence.terracotta.org/display/docs/Home

[13] Hazelcast, "The Leading Open Source In Memory Data Grid: Distribuited computing, simplified" ,2012. http://hazelcast.org/

[14] M.Levi, S. Barjaktarovic, "Proceedings of the XIII International Symposium SymOrg 2012: Innovative management and Business Perfomance", 2012, pp. 979-980.

[15] S. Mei, "Why You Should Never Use MongoDB", 2013. http://www.sarahmei.com/blog/2013/11/11/why-you-should-never-use-mongodb/

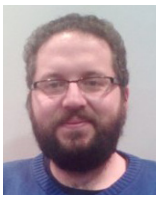
[16] T. Patel, T. Eltaieb, "Relational Database vs NoSQL", Journal of Multidisciplinary Engineering Science and Technology (JMEST), Vol. 2 Issue 4, 2015, pp. 691-695.

Juan Manuel Lombardo, PhD in Computer Science from the Pontifical University of Salamanca, was graduated in Economics and Business Administration in the University of Granada, Spain, Diploma of Advanced Studies (DEA) in Economics from UNED, Research Sufficiency in Business Science from the Complutense University of Madrid and Diploma of Advanced Studies (DEA) in Sociology from the Pontifical University of Salamanca. He is CEO at Fidesol and Professor at Andalusia Business School. Dr. Lombardo is the author of numerous articles and research papers published in journals and books of national and international conferences. Visiting Professor at the Private Technical University of Loja (UTPL Ecuador), The National *University* of the *Northeast* (Argentina), University Francisco José de Caldas (Colombia), Catholic University of Colombia, Catholic University of Ibarra (Ecuador), University of Lisbon (Portugal) and National Engineering University (Peru). Member of the Knowledge Management committee of AEC (Spanish Association for Quality) and the Institute CICTES (Ibero-American Centre on Science, Technology and Society) and trustees of the Fundación Descubre for scientific communication.

Miguel Ángel López, has a degree in Engineering in Technical Engineering in Computer Systems from the University of Almería, graduates in Computer Engineering and Master in Softcomputing and Intelligent Systems from the University of Granada. Currently he is CTO. at Fidesol where performs different roles on the projects. His research focuses on distributed systems, management, integration and analysis of data, robotics, fuzzy logic systems, and the development of virtual reality environments for different purposes.

Felipe Mirón is a member of Technology Transference team, and coordinator of international projects at Fidesol. Born in Granada November 1976, he received his degree in Computer Engineering in 2000. Prior to his coalescence to Fidesol, he worked for Public Administration at Región de Murcia, leading a team to change record processing on Public Social Services. He also joined development team as an Analyst at Onda Cero Radio staff in Madrid, and worked as DBA and developer at Credit Suisse in La Florida (Madrid). He has collaborated with Fidesol team in writing several articles published in "Cuaderno de Avances TIC" 2013.

Susana Velasco, has a Technical Engineer in Computer from the University of Granada. In the past, she worked in manufacturing, financial and service sector enterprises as software engineer and analyst programmer. She has also been R&D engineer and QA engineer at Telefónica I+D and Intecna Soluciones. Currently she is a researcher and QA manager of Fidesol. Her research interests include quality assurance, software quality management systems, Ambient Intelligence (AmI) systems and devices, and new generation of ICT technologies.



Juan Pablo Sevilla, is a student of Computer Engineering in the University of Granada. He currently works as a researcher and software architect for Fidesol in Granada. Previously, he worked as a researcher for the University of Granada. His main fields of work are Java technologies, distributed computing and web applications.

Juan Mellado, has a Technician in Computer Systems Administration and Senior Technician in Automatic Control Systems, with CCNA Certified and LPIC-1 Certified. He has worked since 2001 as a system administrator in different companies like Caton Sistemas Alternativos and Intecna Soluciones. Since 2011 works as a systems administrator at Fidesol.

A Fine Grain Sentiment Analysis with Semantics in Tweets

Cristóbal Barba González, José García-Nieto, Ismael Navas-Delgado, José F. Aldana-Montes

Lenguajes y Ciencias de la Computación, Universidad de Málaga

***Abstract* — Social networking is nowadays a major source of new information in the world. Microblogging sites like Twitter have millions of active users (320 million active users on Twitter on the 30th September 2015) who share their opinions in real time, generating huge amounts of data. These data are, in most cases, available to any network user. The opinions of Twitter users have become something that companies and other organisations study to see whether or not their users like the products or services they offer. One way to assess opinions on Twitter is classifying the sentiment of the tweets as positive or negative. However, this process is usually done at a coarse grain level and the tweets are classified as positive or negative. However, tweets can be partially positive and negative at the same time, referring to different entities. As a result, general approaches usually classify these tweets as "neutral". In this paper, we propose a semantic analysis of tweets, using Natural Language Processing to classify the sentiment with regards to the entities mentioned in each tweet. We offer a combination of Big Data tools (under the Apache Hadoop framework) and sentiment analysis using RDF graphs supporting the study of the tweet's lexicon. This work has been empirically validated using a sporting event, the 2014 Phillips 66 Big 12 Men's Basketball Championship. The experimental results show a clear correlation between the predicted sentiments with specific events during the championship.**

***Keywords* — Microblogging, Big Data, Sentiment Analysis, Apache Hadoop, MapReduce, Twitter, RDF, Named-Entity Recognition, Linked Data.**

I. Introduction

SOCIAL network tools are becoming an important means of social interaction. In this sense, public messages are being analysed to track user opinions on any relevant aspect. Thus, companies are using this kind of analysis to gain insight into the success of new products and services. In this context, Twitter has grown in popularity in recent years and now reports (30th September 2015) that it has a volume of approximately 500 million tweets sent per day and 320 million active users monthly [1]. Therefore, Twitter is a key source of real time opinions of millions of clients or potential clients and so, a valuable source of information for companies.

The term "Big Data" refers to data that cannot be processed or analysed using traditional techniques. Big Data Analytics enables information retrieval from these data. Many software solutions linked to the Apache Hadoop framework [3] are developed to solve problems encountered through the analysis of social media. The problem of analysing tweet streams is a typical example of the use of the Hadoop ecosystem as its technology can provide solutions to make it feasible.

The automatic interpretation of the results of a Big Data analysis, by exposing their semantics and preserving the context of how they have been produced, are some of the challenges to be addressed when trying to add these results to business processes. In this scenario, the concept of *Smart Data* emerges, which could be defined as *(Def. 1)*: *the result of the process of analysis performed to extract relevant information and knowledge from Big Data, including context information and using a standardized format*. By context, we mean all the relevant metadata needed to interpret the analysis of results. This leads to the enforceability of these results thereby facilitating their interpretation, easy integration with other structured data, integration of the Big Data analysis system with the BI systems, and the interconnection (in a standardised way, at a lower cost and with higher accuracy and reliability) of third party algorithms and services.

This has motivated us to propose here, a novel approach for performing sentiment analysis on tweet streams using Big Data technology, although with the aim of obtaining Smart Data. This approach follows the MapReduce programming model for the analysing of tweets by means of an ontology-based [23] text mining method. The analysis is not limited to just calculating the sentiment value of a given tweet, but also the sentiment of entities mentioned in the text. A domain ontology guides the analysis process, providing sentiment values as RDF graphs [23]. The use of RDF enables the publication of the analysis results as Linked Data and their integrated use with other Linked Data repositories.

In this context, there are no benchmarks for measuring the quality of fine grain sentiment analyses. This led us to select a case in which we could relate measured sentiment values with real life events. Thus, if the sentiment analysis correlates these events we have an empirical validation of the proposed solution. In this paper we present the use of sporting events for this validation. The chosen event is the 2014 Phillips 66 Big 12 Men's Basketball Championship [2]. The Big 12 is a set of sporting events, founded in 1994, including sports such as basketball, baseball, and American football. In the 2014 event, 10 universities from around the United States participated: Iowa, Kansas, Oklahoma, Texas and West Virginia. These universities are Baylor University, Iowa State University, Kansas University, Kansas State University, Oklahoma University, Oklahoma State University, Texas Christian University, Texas Tech University and West Virginia University.

The main contributions of this paper are summarised as follows:

- A MapReduce algorithm for the sentiment analysis of tweets that incorporates a semantic layer to improve the text mining, is proposed for the first time.
- A thorough experimentation of our proposal is carried out from three different viewpoints. First, the analysis of the number of tweets and their relation to the match throughout the championship. Second, the analysis of the relationships between sentiment in the tweets and match scores. Third, the use of linear regression to study the relation between number of tweets and sentiment values.

The remainder of this article is organised as follows. The following section presents background concepts. Section 3 reviews the related

DOI: 10.9781/ijimai.2016.363

literature. In Section 4 we present the problem description. Section 5 details our MapReduce approach. In Section 6, experimental results are presented. Section 7 includes a discussion of results, and finally, Section 8 extracts conclusions from this discussion and details future work.

II. Background Concepts

In this section, we describe the different concepts and tools used in this paper, for the sake of a better understanding.

When we think of Big Data Analytics, one of the main frameworks used to address it is Apache Hadoop (Hadoop) [3]. Hadoop is a software framework for the distributed processing of large data sets across clusters of computers using simple programming models. The core of Hadoop consists of a storage component, known as Hadoop Distributed File System (HDFS), and a processing engine called MapReduce.

HDFS is a Java-based file system that provides scalable and reliable distributed data storage. It has been designed to span large clusters of commodity servers. HDFS is a scalable, fault-tolerant, distributed storage system that works closely with a wide variety of concurrent data access applications, usually coordinated by MapReduce.

MapReduce is an increasingly popular, distributed computing framework for large-scale data processing that is amenable to a variety of data intensive tasks. Users specify serial-only computation in terms of a *map* method and a *reduce* method. The underlying implementation automatically parallelises the computation, offers protection against machine failures and efficiently schedules inter-machine communication [4].

The Natural Language Processing (NLP) is used to analyse the tweets. In this approach, we use GATE (General Architecture for Text Engineering [5]). This suite, developed at the University of Sheffield, is used for entity recognition. The GATE developer module is used for a first syntactic analysis that, in our proposal, is complemented with a semantic (ontology guided) analysis of the tweets. The GATE developer contains a component for information extraction (ANNIE) that determines, from an input text, the different terms that compose it. This component divides the text into simple terms such as punctuation marks, numbers or words. During this process, the component identifies certain types of special rules for English, which enables a more effective division (such as contracted forms like "*don't*" or the Saxon genitive). The division into terms helps us to identify the entities described in the domain ontology (populated with synonyms of the different instances).

SentiStrength [6] is a tool used for the identification of the calculation of the sentiment values. SentiStrength lets us perform a quick test, by inputting a phrase that can even include emoticons, acronyms, etc., classifying it on the positive and negative scale. The positive values (sentiment) range between 1 (not positive) and 5 (extremely positive). The negative values (sentiment) range between -1 (not negative) and -5 (extremely negative). This analysis is done in the context of a recognised entity in a tweet. This way, a tweet can deliver several sentiment values for each different entity.

Sentimental force is specific to the context in which the word tends to be used. SentiStrength employs a machine learning algorithm to optimise the emotional power of words in a sentence, which is incremented or decremented by 1 strength point depending on how the accuracy of the ratings increases. Another issue to consider is that the strength of the emotion of a sequence of words can be altered by the words that precede them in the text. SentiStrength includes a list of words that increase or decrease the excitement of a sequence of subsequent words, in a positive or negative sense. Each word in the list increases the strength of emotion by 1 or 2 points (e.g. very or extremely words) or decreases it by 1 point (for example, the word some). The algorithm also has a list of words that reverse the polarity of the subsequent emotion words, including any amplifier preceding word (e.g., "*very happy*" is a positive force for 5 points, but "*not very happy*" has a negative power of 5 points).

SentiStrength employs a spelling correction algorithm that identifies the standard spelling of words that are misspelled by the inclusion of repeated letters. The algorithm also considers the use of repeated letters commonly used to express emotion or energy in the texts, so before correcting them orthographically it increases the emotion words by 1 point, provided there are at least two additional letters (one repeated letter commonly appears in a misspelling).

The use of emoticons is usual in social networks, so the list of forces increases feelings with a list of emoticons (plus or minus 2 points). In addition, any sentence with an exclamation mark is assigned a positive sentiment at least.

III. Literature Overview

In [7] Barbosa et al. propose a 2-step sentiment detection. The first step targets distinguishing subjective tweets from non-subjective or subjective tweets. The second step further classifies the subjective tweets into positive, negative and neutral. This method is called polarity detection. The authors use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labelled tweets for tuning and another 1000 manually labelled tweets for testing.

Argawl et al. [8] have designed a series of models for classifying tweets into positive, negative or neutral sentiments. Their approach proposes two classification tasks: a binary task and a 3-way task. Thus, they use three types of models for the classification: a unigram model, a feature based model and a tree based model and various combinations of the three. For the tree kernel based model, they designed a new tree representation for tweets. The feature based model uses 100 features and the unigram model uses over 10,000 features. They concluded that combining prior polarity of words with their parts-of-speech tags is the most important part of the classification task. They also point out that the tree kernel based model outperformed the other two.

One of the main features of tweets is also their major drawback i.e., their length. Users have to express their thoughts in just 140 characters. Consequently, the messages must be short so people have to use acronyms, emoticons and other characters that convey special meanings and introduce misspelling. Understanding the meanings of these characters is essential for interpreting the sense of the tweets [6]. In this regard, Kumar et al. [9] propose the use of a dictionary-based method and a corpus to find orientation of verbs and adverbs, thus supplying calculated sentiment to the tweets.

Finally, Monchón et al. [25] propose a new study on the measurement of happiness in Latin America, which shows the possibility of measuring the happiness through the use of social networks, and so it is tremendously simple to calculate via objective and empirical means.

In our proposed method, we aim to go one step beyond these previous approaches by incorporating a semantic model to improve the entity identification in the text mining phase, and hence enhance the sentiment analysis.

IV. Problem Description

The problem faced here is the identification of the sentiment at entity level in a set of tweets. Part of the Big 12 competition is used for testing purposes. This data subset is the tweets of the last three competition days for the Big 12 men's basketball championship.

The main challenge in this work is to identify entities of interest found in tweets. This implies the identification of the list of words related to this event, for instance: teams, players, coaches, referee, and matches. Another problem is the number of related tweets that we can filter from Twitter in almost real time using Twitter API [11]. This requires a continuous process of improvement on the filtering keywords used to reach relevant tweets.

In order to address this problem, we have used Hadoop as the parallelization mechanism since the process fits well in the Map-Reduce methodology.

V. Map Reduce Proposed Approach

In this section, we introduce a methodology that combines the distributed computing framework MapReduce with an ontology-based text mining approach to apply fine grain sentiment analyses.

The sentiment analysis uses SentiStrength. The semantic context introduced by the domain ontology enables the early filtering of the tweets based on relevant entities for the analysis tasks. In the use case presented here, many tweets have been discarded because their content has been determined to be irrelevant for the analysis.

The proposed methodology is illustrated in Fig. *1* and includes the following elements:

- The **domain ontology** to describe the analysis context. This ontology includes not only the structure, but also the population of each term with domain knowledge expressed as ontology instances. This knowledge includes names and synonyms for the entities that we aim to recognise in the tweets.
- The tweet **extraction and analysis algorithms**. This process is done using MapReduce as the programming methodology. The search of new tweets is dynamic and can include new search terms as soon as they are detected as relevant during the analysis process.
- The **NLP analysis** algorithm. This algorithm has been developed using GATE and is guided by the domain ontology.
- The **sentiment calculation** algorithm via SentiStrength, later analysed according to the aggregation level needed.

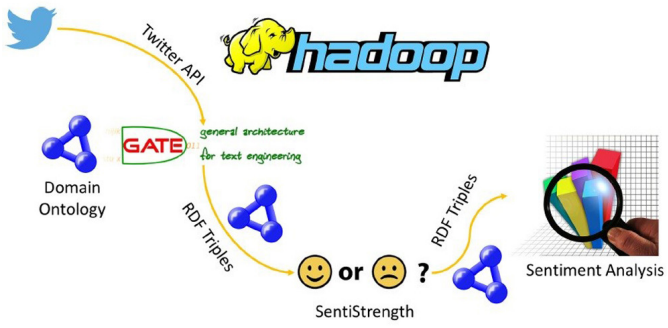


Fig. 1. Methodology for tweet analysis

A. Semantic layer

We have built a full semantic annotation rather than a simple tagging at tweet level. This means that we can identify the location of a term in the text that corresponds to the underlying entity. Ontological proximity disambiguates entities in the text by looking at relationships between entities matched in the tweets. Entities that have a close ontological relationship are deemed to be more likely to be correct. For example, if an analysis of tweets identifies both Texas (university team) and Texas (state) and basketball, then Texas (state) will be disambiguated due to the close ontological relationship with basketball. We have defined an ontology with the concepts of interest for the analysis task. In our use case, these concepts are *Championship, Team, State, City* and *TwitterAccount*. Some examples of the concept properties that relate these concepts with each other are:

- `<http://khaos.uma.es/big data/kb/ont/play>` Property to connect two teams playing a match.
- `<http://khaos.uma.es/big data/kb/ont/cityIn>` Property to indicate relationships between cities and states.
- `<http://khaos.uma.es/big data/kb/ont/locatedIn>` Property to define relations between universities and cities.
- `<http://khaos.uma.es/big data/kb/ont/isAccountOf>` Property to indicate the twitter accounts of the teams.
- `<http://khaos.uma.es/big data/kb/ont/name>` Property to define the names for the entities.

B. Tweet Discovery

The first step of the proposed methodology is a well-known problem, *how to collect the tweets for their analysis*. The extraction of the tweets is done by means of Twitter API [11]. More specifically, in this case we have used Python's API [12]. The discovery component searches for certain keywords on Twitter. The keywords used are intended to obtain the highest number of tweets in order to have as wide as possible coverage, even if this means having to filter out some of them. These keywords are based on *hashtags* of interest and usernames related to the analysis task, but new keywords are detected during the analysis.

In our case, we have chosen as keywords the name of the user accounts of the university teams and hashtags created by team supporters. The championship also has its own hashtag. Examples of an initial set of keywords are *Big12Conference, Big12, Big12Insider, Big12MBB, BaylorMBB, CycloneMBB, TexasMBB, KUHoops, TCUBasketball, OSUMBB, kstatesports, kstate gameday, OU_MBBall, WVUhoops, TechAthletics and TTRaidersSports* [13-21]. We have compiled a wealth of information on approximately 11.5 gigabytes of tweets using these keywords.

Hadoop technologies have been used to deal with the retrieval and analysis of these tweets. The approach uses MapReduce, storing the tweets and the analysis results in HDFS (Hadoop Distributed File System). The following section details the analytical algorithms developed.

C. Ontology-based Tweet Analysis with MapReduce

The algorithms created to analyse the tweets make use of Natural Language Processing and Semantic Web techniques. These techniques take part of the MapReduce model that divides the process into three main functions: the Map function, the Combiner function and the Reduce function.

(1) The Map function retrieves a tweet and analyses its entities' sentiment value. Entities discovered by GATE are annotated as an RDF document. In the test case, GATE searches all the teams listed in the tweet (can be more than one), together with the sentiments associated with them in any individual tweet. SentiStrength calculates the sentiment of an entity based on its context. Hence, a tweet can deliver several sentiment values for each different entity.

The output of the Map function, which is the input of the Combiner function, is the tuple *(key, value)* whose contents are:

- *key*. It is a string formed by the joining of the tweet identification number concatenated with the date and time of the publication of the tweet.
- *value*. It is an RDF triple that is calculated in this Map function.

(2) The Combiner aims to group triples by team. The team associated with each entity is calculated during the execution of the Map function. This function takes care to gather together all positive and negative sentiment values from a particular entity (team in the use case) on a particular date and at a specific time (the date is given as day, month, year, hours, minutes and seconds). The time mark-up of the sentiment values is useful when analysing the results and their correlation with real events. In this case, they could be a score, the elimination of a team, injuries of a player, etc.

The output of the Combiner function is the entry for Reducer function. Therefore, outputs tuples *(key, value)* are:

- *key*. It is the joining of the entity name with the date and time at which the tweet was posted. In the use case, the entity name corresponds to the team name.
- *value*. It is the sum of the positive and negative sentiments calculated from the tweet.

(3) The Reducer collects the values of positive and negative sentiments calculated in the Combiner function and calculates the average of these values for all the tweets using entities (teams in the use case) and dates to order them. Thus, the output of this function is set of tuples *(key, value)* that are constructed as follows:

- *key*. It is the same key as the Combiner function.
- *value*. It is calculated by concatenating the number of tweets, positive sentiment value mean and negative sentiment value mean.

VI. Experimental Results

In this section, we examine the data obtained in the analysis phase. The objective of this analysis is to determine whether the results obtained during the Big 12 Men's Basketball Championship can influence the sentiments of tweets. Therefore, we look to analyse the feelings of the tweets before, during and after matches to check how they change over the course of the match.

The basketball championship began on Wednesday, March 12th, 2014 and ended on Saturday, March 15th, 2014. We started with the capture of tweets to analyse their feelings a few days before the start of the championship and completed the collection one day after it. This was done with the idea of assessing how the championship was affected by the tweets about the teams playing the tournament.

A. Global trends in the championship

In this subsection, we analyse the trends in the number of tweets and their sentiment values throughout the whole championship.

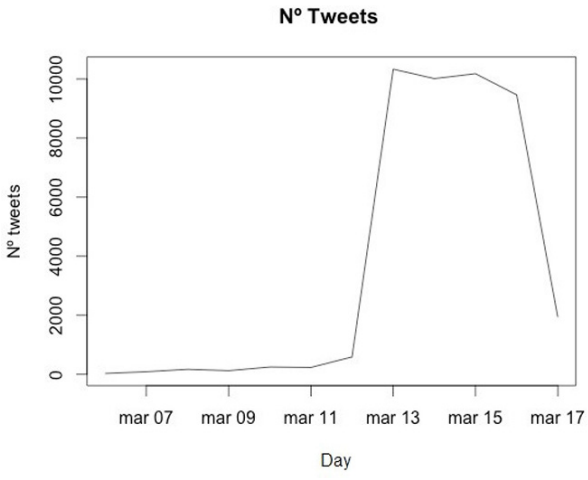


Fig. 2. Total number of tweets mentioning university teams or tournaments

Fig. *2* shows that prior to the start of the championship the number of tweets which mention the university teams or their tournaments were very low. Nevertheless, we note a significant increase in the number of tweets during the championship. Finally, this number falls as soon as the championship ends to a similar number as the initial phase of the championship.

Similarly, the feelings of the tweets, overall, increase their values when the championship starts. This is expected since the mood swings are intensified whenever there is a sporting event. These trends can be observed in Fig. 3 (positive) and Fig. *4* (negative).

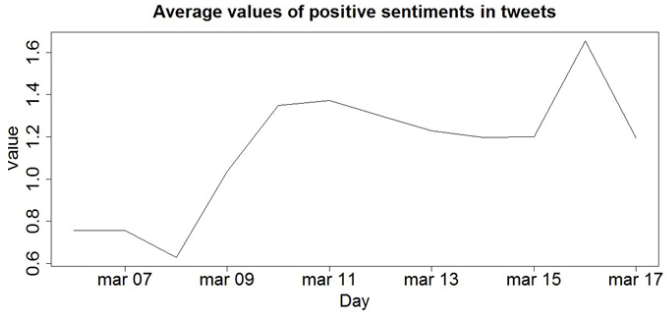


Fig. 3. Positive trends in tweets during the championship

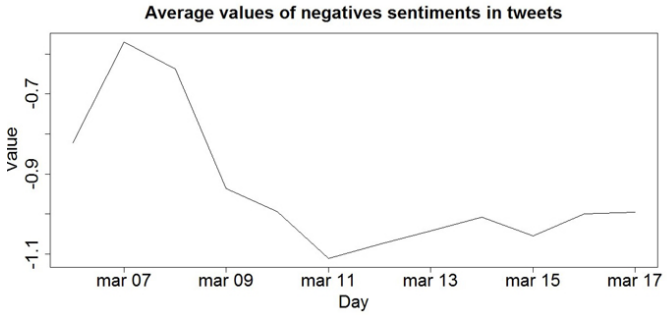


Fig. 4. Negative trends in tweets during the championship

B. Final match

Having viewed the data of the entire championship, we focus on one match specifically so as to compare the changing feelings for the two teams, so we choose the most important match of the championship which is the final.

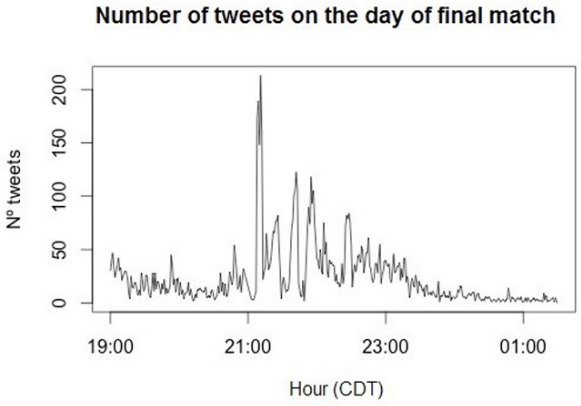


Fig. 5. Evolution in the number tweets in the final match

The final match was played by Baylor University against Iowa State University. Iowa State was the winner. The match began at 20:10 and

ended more or less at 22:10. As shown in Fig. *5*, at the time of the start the game, the number of tweets commenting on something about the championship begins to increase. In the time period between 22:00-22:30, the number of tweets starts descending. The maximum number of tweets is generated in the interval from 21:15 to 21:20. This interval coincides with half-time.

For a better understanding of the trend in the tweets, it should be noted that Baylor University was winning the game until the final minutes. However, Iowa State University, in the final minutes, took the lead and finally won the match.

In Fig. *6* we can observe the trend of positive tweets about Baylor University. A notable aspect is as the outcome becomes less favourable the sentiments in the tweets become less positive.

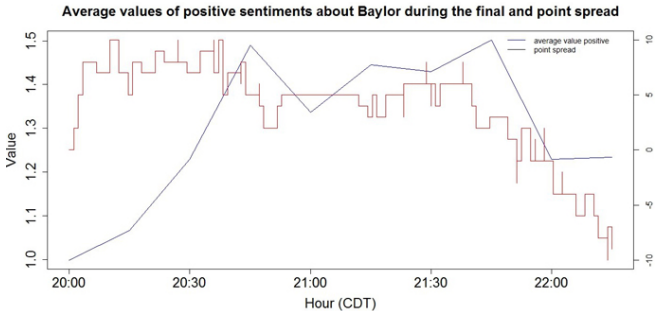


Fig. 6. Baylor University sentiment tweet's trend in the final match

However, we note that in the case of Iowa State University, the trend in tweet sentiments is different from Baylor, when the match is getting closer to the end, the sentiment is mostly positive, as we see in Fig. *7*. This is to be expected as Iowa won the match in the final minutes; therefore, their fans were happier and wrote positive tweets.

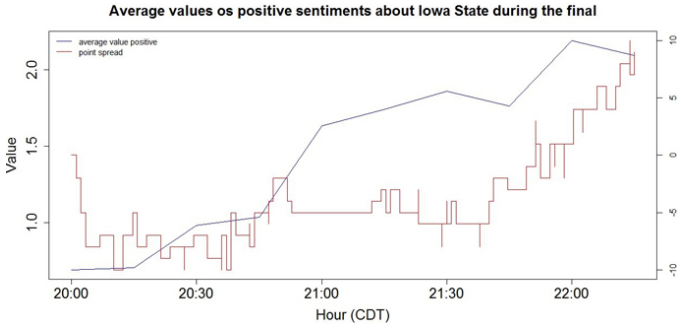


Fig. 7. Iowa State University sentiment tweet's trend in the final match

C. Regression Analysis

This section mathematically analyses the relationship between the number of tweets that are generated, and the positive or negative feelings that they show. This study focuses on the final match.

Linear regression is a mathematical method that models the relationship between pairs of variables. The Pearson product-moment correlation coefficient [24] measures the intensity of this possible relationship between variables. This rate applies when the relationship that may exist between the variables is linear and is calculated as the ratio between the covariance and the product of the standard deviations of both variables (Eq. 1).

$$r \quad \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Eq. 1. Linear regression.

The values can take a correlation coefficient range from -1 (perfect negative correlation) to 1 (perfect positive correlation).

$$-1 \leq r \leq 1$$

As we see in Tables I and II variables which have a better relationship with each other are the positive (+) and negative (-) sentiments. In Table III we note a relationship as a moderate positive sentiment between the number of tweets of Baylor and Iowa State during the final. In Tables IV and V, we can statistically see, as in the case of the Baylor correlation- though not as strong- a negative meaning so that when you increase the point spread against Baylor the number of tweets and sentiment values also decrease. In contrast, we note that in the case of the Iowa State University, when the point spread is higher the number of tweets and positive sentiment are also higher, so the correlation value is positive. However, for Baylor's negative sentiments as well as positive ones, the correlation coefficient is negative.

TABLE I
LINEAR REGRESSION ANALYSIS OF BAYLOR TWEETS. POSITIVE SENTIMENT IS DENOTED WITH (+) AND NEGATIVE SENTIMENT (-)

Baylor	Nº Tweets Baylor	Average (+) Baylor	Average (-) Baylor
Nº Tweets Baylor	N/A	-0.04865	-0.01833
Average (+) Baylor	-0.04865	N/A	0.53622
Average (-) Baylor	-0.01833	0.5363	N/A

TABLE II
LINEAR REGRESSION ANALYSIS OF IOWA TWEETS. POSITIVE SENTIMENT IS DENOTED WITH (+) AND NEGATIVE SENTIMENT (-)

Iowa State	Nº Tweets Iowa State	Average (+) Iowa State	Average (-) Iowa State
Nº Tweets Iowa State	N/A	0.40802	0.28161
Average (+) Iowa State	0.40802	N/A	0.74019
Average (-) Iowa State	0.28161	0.74019	N/A

TABLE III
LINEAR REGRESSION OF BAYLOR IOWA STATE TWEETS. POSITIVE SENTIMENT IS DENOTED WITH (+) AND NEGATIVE SENTIMENT (-)

Baylor/Iowa State	Nº Tweets Iowa State	Average (+) Iowa State	Average (-) Iowa State
Nº Tweets Baylor	0.56581	0.31094	0.225542
Average (+) Baylor	0.11722	0.27056	0.22139
Average (-) Baylor	0.00112	0.15629	0.143035

TABLE IV
LINEAR REGRESSION BAYLOR AND POINT SPREAD

Baylor	Nº Tweets Baylor	Average (+) Baylor	Average (-) Baylor
Point spread	-0.12698	-0.22582	-0.06735

TABLE V
LINEAR REGRESSION IOWA STATE AND POINT SPREAD

Iowa State	Nº Tweets Iowa State	Average (+) Iowa State	Average (-) Iowa State
Point spread	0.42327	0.36465	-0.22899

VII. Discussion

In this paper, a methodology that combines an ontology-based NLP process with a sentiment analysis to produce an accurate analysis of sentiment has been proposed and explained. The philosophy of the MapReduce approach fits perfectly in this type of work, because it allows us to distribute the workload of calculating the sentiment of a set of tweets into a computing cluster. As the calculation of feeling in a tweet is independent of other tweets, distributing the tweets can be done seamlessly without dependencies.

The fine grain of the sentiment calculation enables an analysis at different levels. In this paper, we have shown a use case related to a popular sporting event to show the feasibility of the proposal and the possible analysis that can be performed using the calculated sentiment. However, generating the information extracted and calculating sentiment as ontology instances enables them to be connected with other tools or scenarios:

1. The use of semantics in the analysis process enables the use of context in the discovery of entities in the tweets and the differentiation of the sentiment depending on the entities we are analysing.
2. The use of RDF triples facilitates the publication of the sentiment analysis as SPARQL endpoints, which in turn enables them to be linked with other Linked Data repositories.
3. The fine grain analysis allows a deeper analysis and even the building of data warehouses to operate with different filtering or aggregation functions.

The use case selected has helped test the approach in a real scenario where real events are used to contrast the predicted results with expected sentiment values. In this case, the feelings associated with the tweets correlate to the changes in the sporting event. Thus, the value of this type of analysis has been demonstrated for any context, where one-off events can alter the feelings of social communities and maintain these feelings sometime after the event that produced the changes in sentiment.

This work has focused on the analysis of the sentiment for individual tweets. However, the social interactions between Twitter users could be also of interest in this analysis [24]. Thus, future work will include the consideration of these social interactions to fine tune the analysis results.

VIII. Conclusions

This paper has presented a novel approach for the sentiment analysis of tweets using semantics. This work can be applied in other scenarios with small texts, by defining the analysis context with a domain ontology.

The use of sporting events for testing purposes has been shown to be valid, and so it can be used as a base line for the development of a benchmark for empirical testing of fine grain sentiment analysis tools.

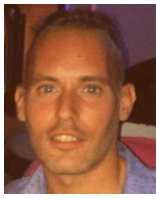
The parallelisation of the problem using MapReduce has shown a good behaviour for the analysis task. However, we are currently looking at ways to improve it. In future work we will be developing approaches, such as Apache Spark or Apache Storm, which are able to deal with real time analysis. These analyses with the support of a parallel platform will lead us to savings in terms of computational effort.

Acknowledgment

This work was supported in part by Grants TIN2014-58304-R (Ministerio de Ciencia e Innovación) and P11-TIC-7529 and P12-TIC-1519 (Plan Andaluz de Investigación, Desarrollo e Innovación). Cristobal Barba is supported by Grant BES-2015-072209 from the Spanish Government.

References

[1] Twitter Statistics, Last Access 28th January 2016. https://about.twitter.com/es/companyB.
[2] Big 12 Men's Basketball 2014 phillips 66 Big 12 Men's Basketball Championship. http://www.big12sports.com/.
[3] Apache Hadoop, 2016 Apache Hadoop http://hadoop.apache.org.
[4] Framework MapReduce, 2008 MapReduce Tutorial
[5] General Architecture for Text Engineering. GATE 2016. https://gate.ac.uk/
[6] SentiStrength . 2013 SentiStrength { sentiment strength detection in short texts sentiment analysis, opinion mining. http://sentistrength.wlv.ac.uk
[7] Luciano Barbosa,Junlan Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. Coling 2010:Poster Volume, Beijing.
[8] S. Agarwal, S. Godbole, D. Punjani y S. Roy. 2007. How much noise is too much:A study in automatic text classication. ICDM, pages 3-12.
[9] Akshi Kumar, Teeja Mary Sebastia. 2012. Sentiment Analysis on Twitter. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012.
[10] Twitter4j. 2016 Api Twitter4j. http://twitter4j.org/en/index.html
[11] Twiter Api, 2016 https://dev.twitter.com/overview/api
[12] Python Api, 2016 https://docs.python.org/2/c-api/
[13] Baylor University Baylor University Texas http://www.baylor.edu.
[14] Iowa State University Iowa State University..
[15] Kansas University Kansas University http://www.ku.edu.
[16] Kansas State University Kansas State University http://www.k-state.edu.
[17] Oklahoma University Oklahoma University https://www.ou.edu.
[18] Oklahoma State University Oklahoma State University http://go.okstate.edu.
[19] Texas Christian University Texas Christian University http://www.tcu.edu.
[20] Texas Tech University Texas Tech University http://www.ttu.edu.
[21] West Virginia University West Virginia University http://www.wvu.edu.
[22] The Pearson product-moment correlation coefficient, 2016
[23] Staab, S., & Studer, R. (2009). Handbook on ontologies. International Handbooks on Information Systems. Springer.
[24] Teófilo Redondo. The Digital Economy: Social Interaction Technologies – an Overview. *International Journal of Interactive Multimedia and Artificial Intelligence*. 3(2): 17-25.
[25] Monchón.F, SanJuan (2014). O. *A First Approach to the Implicit Measurement of Happiness in Latin America Through the Use of Social Networks*. International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 2, Nº 5



Cristóbal Barba-González is a PhD student at the University of Málaga. His research focuses on optimization algorithms (metaheuristics), semantics and analysis of big data.

José Manuel García-Nieto is a PhD and Post-Doctoral Research Assistant at the University of Málaga. His research focuses on optimization algorithms (metaheuristics), semantics, data mining and Big Data.

Ismael Navas-Delgado is PhD and Assistant Professor at the University of Málaga, Spain. His research focuses on the use of Semantics and Big Data technologies in Life Sciences.

José F. Aldana-Montes is Full Professor at the University of Málaga, Spain. He is the main researcher of the Khaos Research Group. His research focuses on management, integration and analysis of data.

Big Data & eLearning: A Binomial to the Future of the Knowledge Society

Vidal Alonso, Olga Arranz

Universidad Pontificia de Salamanca

***Abstract*— There is no doubt that in what refers to the educational area, technology is producing a series of changes that will greatly affect our near future. The increase of students experiences in the new educational systems in distance learning makes possible to have information related to the students 'activities and how these can be dealt with automatic procedures. The implementation of these analytical methods is possible through the use of powerful new technologies such as Data Mining or Big Data. Relevant information is obtained of the use made by the students of the technological tools in a Learning Management System, thus, allowing us to infer a pattern of behavior of the students, to be used in the future.**

***Keywords — Big* Data, Distance learning, Educational technology, Learning analytics.**

I. Introduction

As we all know, technology is having a great impact on people´s lives. If something has marked the progress of advanced societies, over the past few decades, this has been the remorseless development and massive use of technological tools to manage all kinds of tasks [1]. There is no doubt that in what refers to the educational area, technology is producing a series of changes that will greatly affect our near future. The recent emergence of MOOCS (Massively Open Online Courses) is just a sample of the new expectations that are offered to university students.

Another sample is given by the new channels of communication, which are represented by the social networks, which are beginning to be integrated into the educational system by means of lessons via video conferencing or participation in lively debates online. The exchange of information in Facebook or via messages on Twitter, allows a provision of information written according to the new standards [2]. To deal with these circumstances, teachers need to understand the new media available to them and use them creatively.

These educational developments provide the system with a large amount of data coming from the students' activity. The increase of student's experiences in the new educational systems in distance learning makes possible to have information related to the activities of the students and how these can be dealt with automatic procedures.

This trend leads to a role change in the behavior of the different educational agents, who both, teachers and pupils, must conform to the new methods and change their traditional ways of teaching [3]. Academic institutions can´t stay out of this phenomenon and are required to modify its structures and their information systems to meet the student's needs in order to have access to their academic offer.

Is there anything we can do with the vast amount of data provided by students to improve the educational system? Until relatively a short time ago, storage techniques did not permit an exhaustive analysis of the information present in the Learning Management Systems (LMS). Nowadays there are more and more new analytical methods that allow us to deal with the study of these data and infer trends of the use that students make with respect to the tools available in platforms.

In addition, it combines the implementation of these analytical methods with techniques, such as Big Data, that will enable the access of a new system which will provide new information about the students that we have in our classrooms [4]. The implementation of these analytical methods is possible through the use of powerful new technologies such as Data Mining or Big Data that enable the processing of large amounts of information by searching and finding out new knowledge that is present in the data [5]. Big data allow for very exciting changes in the educational field that will revolutionize the way students learn and teachers teach [6].

This is the context where this paper is framed. The use of the information provided by the new analytical learning techniques will provide assistance to the teachers who use the Learning Management System. To achieve this, we will work with a sample of data present in a LMS and we will see the use that the students have made of the available tools.

With this study, relevant information is obtained of the use made by the students of the technological tools which will allow us to infer a pattern of behavior of the students. This pattern may be used by teachers to achieve, through the application of technological teaching strategies, a greater motivation and productivity of students so that they are continually active in their learning process.

II. Big Data Inside Education

A proof of the importance that is reaching the phenomenon of Big Data is its implementation in educational institutions which are beginning to exploit and understand the benefits that it offers them. In the present time, just enterprises and organizations are the ones which have been analyzing these enormous sets of data to better understand their customers by trying to predict market trends, "the educational world is beginning to integrate the data sets available to improve the learning process of the students" [7].

At the beginning, the companies granted almost no value to the data collected in their transactions. When the Big Data Era began, institutions and business organizations became aware of the high potential remaining in the stored data in their files. This fact changes the trend toward the collection and data storage process, making a greater effort to maintain and structure their data repositories, those who are usually disorganized, contain unnecessary details and many times the knowledge locked in them is incomplete, being necessary a purification of them to avoid the generation of uncertainty [8].

The processing of the large amount of existing data in the field of education has been made possible thanks to the development of new Information and Communications Technologies (ICT). This development has led to diverse educational institutions to carry out an

DOI: 10.9781/ijimai.2016.364

analysis of existing data from the interactions of its students, and in this way, draw conclusions that will improve the working environment, generating new educational organizations structures, and what is more important, new learning processes.

In this sense, the NMC Horizon Report, a reference to the global level of the emerging technological trends in education, provides, in its last report of 2014, that "the data analysis shall be adopted, in a meaningful way, in a period of between two and three years, and, in fact, it is already used in some American universities." [9].

This adoption will be supported by the rapid deployment of the virtual learning environments and the MOOC (Massive Open Online Course), where students perform online tasks leaving a significant trail of data on the web. The collection and analysis of the collected data from transactions, that have made the students when interacting with the system, will be used to adapt the content to the students´ needs and thus, to act in the improvement of the education system.

The importance of the impact that Big Data is taking in the education sector is beginning to be reflected in the expectation aroused in a large percentage of teachers and researchers who have placed their hopes in that the analysis provide relevant data and what the use of these data would mean for the educational area.

Teachers must observe the behavior patterns generated and reduce the risk of students who give up, through a more personalized learning process. Therefore, the educational analytical process itself will detect new problems, generating possible corrections to improve the teaching-learning process, or even questioning the effectiveness of the teaching programs that are taught in the educational organization.

On the other hand, students also benefit because, thanks to the analysis of these data, teachers can adapt the learning environments to their needs. This environment adaptation will depend on the creativity of the teacher, who will interpret the patterns from each student and will choose to provide creative solutions that will help the student to learn the skills required.

Higher education has traditionally been inefficient in the use of data, often operating with substantial delays in analyzing readily evident data and feedback. Organizational processes often fail to utilize large amounts of data on effective learning practices, student profiles as well as providing interventions [10].

To analyze this immense amount of information two treatments or processes are beginning to be used increasingly, known as Data Mining and Big Data. Data Mining is also known as KDD process (Knowledge Discovery Databases), which can be described as a process, which allows us to discover hidden information in large volumes of data. In the course of the process it works with data subsets, looking for similar patterns of behavior or predictive models that can be inferred from the processed data. In the educational area it is used in a way that learning processes could incorporate new and relevant knowledge that enables improvements in such processes [11].

Analytics in education must be transformative, altering the existing teaching, learning, and assessment processes, the academic work, and administration tasks. Analytics provides a new model for university leaders to improve teaching and learning processes and will serve as a foundation for changes. But using analytics requires careful thinking about what we need to know [12]. In the same way, academic analytics has the potential to create actionable intelligence to improve teaching, learning, and student success to predict which students are in academic difficulty as well as focusing on specific learning needs [13].

While its use began with economic purposes, their multiple possibilities have allowed us to extend its use to the field of education. The main methods used and their key applications are [14]:

- *Prediction*: Develops a model to infer some aspects of the data. It is used to emulate the behavior of students in the premises of their previous activities and to predict the possible outcomes.
- *Clustering*: Looking for classifying data into groups with the same characteristics providing information of common patterns for students who are in the same group.
- *Relationship Mining*: Finds out relationships among variables. It allows discovering associations of activities that can induce a sequencing of the same nature. It also highlights the most effective pedagogical strategies in the learning process.
- *Visualization:* It allows discovering trends in the use of educational platforms that are outside of the average of students, known as data noise.

However, with the eruption of MOOCS the online information storage is growing in such a way that the processes for managing this information is becoming insufficient, causing a serious problem due to not being able to exploit the data with the necessary guarantees [15]. In order to be able to process such information is necessary to have new methods, being Big Data the last to be applied to the learning area.

This method allows in the present time that organizations can capture and analyze any data, regardless of what type, how much, or how fast it is moving, and makes more informed decisions based on that information. In education, big data allows to understanding how students move through a learning trajectory. This includes gaining insight of how the student accesses the learning activities or measuring optimal practice learning periods [16].

We are very well aware of the fact that there is still a lot to learn about how to work with big data, just like everyone else. But one thing we know for sure is that the traditional ways of working with data will not lead to success in big data analytics [17]. The variety of information sources, the volume of information, latency of processing, even the basic business models are often all different in the big data space. Someone who recommends using the same old tools under these new circumstances is someone who is outside of the data analysis.

III. Big Data and Virtual Learning Environments

According to Castaneda, since the first Virtual Learning Environment was created in 1995, until today, they have made many mixed environments of telematic known tools like Virtual Learning Environments or Virtual Learning Environments (VLE) giving support different teaching and learning modalities available today. The education provided is configured using VLE, and takes shape through the so-called virtual campus or Learning Management System [18].

A lot of LMS are currently being used by companies, schools and universities to assist in their formation processes. You can even say that they have become the essential tool to perform a teaching model eLearning.

The use of these LMS in education eLearning involves generation of a large and complex set of data. These data obtained from the use by many users of the different technological tools in a virtual learning platform can be drawn great benefits to improve eLearning education.

How can one profit in the learning context from Big Data? Everyone interested in eLearning education wants to find the answer to this question. Ambrose pointed to the company in collaboration with IBM Skillsoft how big data can help create a learning experience more personalized and adaptive based on real information about each student. [19]

So, the study on the use of personal data can be applied perfectly to eLearning. It offers us the opportunity to learn more about our students and their behavior patterns in a way not known before. Moreover, we can use this knowledge to develop eLearning courses really geared

to the needs of our students through scenarios that meet their real situations.

Big Data offers us the opportunity to provide students with more efficient courses and more effective on-line learning modules which are attractive and informative. The reasons why large amounts of data can revolutionize the industry of eLearning are [20]:

- **It allows eLearning professionals design more customized eLearning courses.** If you give eLearning professionals the opportunity to learn what works best for their students, in terms of content and delivery, this will allow create more personalized and attractive eLearning courses, thus providing high quality and meaningful learning experience.
- **It provides counseling on effective online strategies.** ELearning big data can give us visions of which eLearning strategies work and which do not.
- **It allows the monitoring of students´ patterns.** With large data eLearning, educators gain the ability to track the students throughout the learning process. This helps them to find out patterns that not only will allow them to learn more about the behavior of each pupil, but also of the group of students as a whole.
- **It enables the possibility to expand our understanding of the process of virtual learning.** It is essential that eLearning professionals get to know how students learn and acquire knowledge. Big Data gives us the opportunity to gain a deeper understanding of the process of eLearning and how students are responding to the eLearning courses. This information can be used to design new learning methods.

Big Data provides teachers with highly relevant information, but we must not forget that it also brings benefits to students. For instance, if one of the benefits is that educators are able to produce better teaching materials to meet their learning patterns, the student will benefit from it as well, that is, if a student is presented with the information in a meaningful way he or she is going to be more motivated in their learning process. [21]

Students and participants in eLearning courses have much to gain from the benefits that the information from Big Data provides us. Next, a case study is presented based on the large amount of data collected using a virtual learning environment, where it is shown how eLearning and Big Data form a binomial that must be considered in the future to improve the knowledge society.

IV. A Case Study: Using Big Data in a Learning Management System

In this practical study it has been taken into consideration the need to evaluate the large amount of data generated from the use of the technological tools used in the teaching / learning environments interaction. These environments of interaction are related to the combined method of learning Known as “Blended Learning”, which Integrates, in a balanced manner, the virtual classroom learning with the learning proposals.

The purpose of this analytical study is to provide information that will allow us to improve the outlook of teachers and students in order to optimize our design in eLearning courses. In addition, the study also aims at guiding the use of the most appropriate virtual educational tools, in order to develop innovative educational strategies according to the patterns obtained from the behavior of student learning.

The study takes, as its starting point, a sample of educational data, educational dataset, from a university that is accessed by the students through a virtual learning environment. The study analyzes the use that the students do of the different technological tools available when they are accessing to the subjects enrolled.

In this study, the educational dataset provides the total number of accesses to the tools that has been collected in the data processing center of the university center. The total number of accesses to each one of the tools is shown in the Fig. 1:

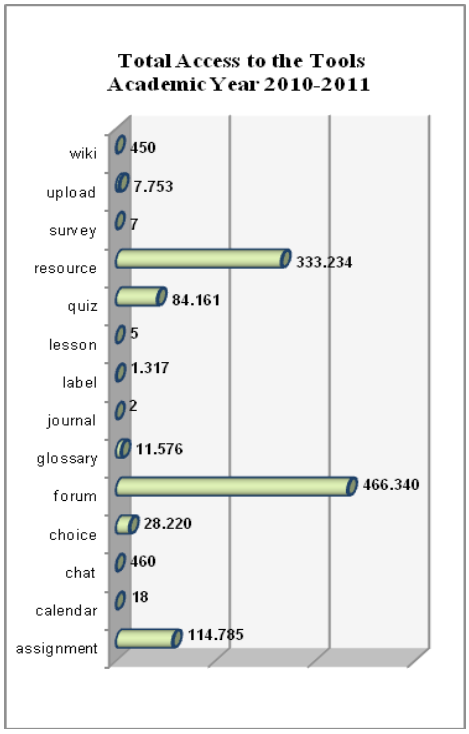


Fig. 1. Total Access to the Tools

As can be seen in Fig. 1 there are 14 different types of technological tools that can be accessed. There are other 2 tools, Course y User, which are not relevant to our study since they are related to the number of users who have accessed as well as to the available virtual courses.

In order to begin this research, it is necessary first to carry out a classification/organization in four different categories (Storage, Collaboration, Communication and Assessment), taking into account the use of the tools and at the same time the types of tools that an VLE should contain. This classification can be observed in Table 1.

TABLE 1. CLASSIFICATION TOOLS GROUPS

Categories	Tools Groups
Storage	Label, Resource, Upload
Collaboration	Forum, Glossary, Wiki
Communication	Calendar, Chat, Journal
Assessment	Assignment, Choice, Lesson, Quiz, Survey

However, it is important to highlight the Collaboration tool group, since it allows us to carry out a collaborative teaching-learning process providing feedback so that we can optimize the learning process as well as an increase in student´s motivation.

Once the group classification tools have been obtained, we can carry out an analysis of the average number of accesses of enrolled students to the different tool groups, always having in mind the final aim of the use of each tool group.

Taking into account the number of enrolled students and the number of total accesses to the different tool groups, we have obtained the

average number of accesses per each student and the type of tool used, getting, in this way the results shown in Fig. 2.

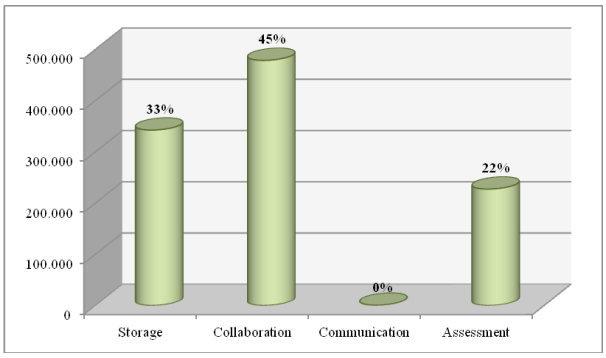


Fig. 2. Students Average Access by Tools Group

From the values shown in Fig. 2, it can be observed that the access to the Communication Tools represents 0% since the Calendar, Chat and Journal tools, which are within the Communication tools, are scarcely used.

On the contrary, Fig. 2 shows that the most used tools are the Collaboration 45% and Storage, 33%, whereas the rest 22% out of the total sample relates to the Assessment tools.

These percentages are aligned with accesses represented in Fig. 1 where it can be observed how the technology tools such as Forum, Resource y Assignment are the ones with a higher volume of accesses. Since they fall into three different categories, the percentage obtained doesn't show any significant differences.

It was also considered of high interest the need to perform a statistical analysis of the dependence or independence between different pairs of tool groups.

The results of this evaluation are shown in Table 2 where the results of asymptotic significance level and Chi-squared statistic obtained from the study of the corresponding media access for student and group sampling tools are presented.

Noted that in the table are represented only half of the values because the relationships between the types of tools are symmetrical blocks.

TABLE 2. REPRESENTATION OF THE VALUES AND SIGNIFICANCE OF THE ASYMPTOTIC CHI-SQUARE BETWEEN PAIRS OF TOOLS, FOR THEIR DEPENDENCE

	STORAGE	COLLABORATION		ASSESSMENT	
		SIG. ASYMPTOTIC	CHI-SQUARE	SIG. ASYMPTOTIC	CHI-SQUARE
STORAGE		0,000	82,96	0,000	148,01
COLLABORATION				0,000	35,08
ASSESSMENT					

In view of the results of the analysis about the asymptotic significance level, we can say that it will always exist dependencies between different tool groups represented. This statement is possible because the significance value in all of them is less than the reference value taken (0,005).

On the other hand, considering the data on chi-square analysis, it can be said that the greater dependence exists between the pair of Storage group with the tools of assessment, since the value of Chi-square is the greatest of all (148.01).

In contrast, less dependence exists between the group with the Collaboration tools and the Assessment ones, since the value of Chi-square is the smallest of all pairs of tools (35,08).

On the contrary, we can state that there is a linear relationship between accesses to the types of Storage tools with those accesses to Collaboration tools, thus, all who access the Storage tools also access the Collaboration tools.

So that, it is possible to extrapolate from data obtained that there is a very high dependence between the group of storage tools and the assessment one.

V. Conclusion

Given the huge amount of data available to the educational area, it is possible to proceed with its processing to obtain sufficient knowledge that allows them to improve their structures. The combination of different learning analytical techniques with the new paradigms of processing, such as Big Data, will enable relevant information to the educational authorities and teachers to change and to optimize the current methods.

These changes can be seen with fear on the part of teachers, who did not know how to deal with the new teaching methods from the pedagogical perspective. To help them, this paper shows a case study where the teacher gets information about which are the most used tools includes in the new learning environments.

From this information you can set new strategies of teaching-learning based on the student's experience. So that, looking for a greater participation by the students, the teacher may propose some tasks where they have to use the tools that are more favorite to him in front of other lesser-used tools.

The proposed activities, that involve the use of collaborative tools, makes the need to work the activity as a group, in line with the new educational circumstances that are supported, mainly, in the teaching-learning collaborative process. The use of these tools is highly satisfactory by part of the students which will result in a more active participation and an increase in the student's own motivation, driving to a learning improvement.

Furthermore, if these collaborative tools are combined with the storage and evaluation tools we shall be creating a teaching strategy that will not be rejected by the student, where he can develop all its intellectual capacity in an enjoyable and satisfactory way.

We have already seen how Big Data can influence eLearning, what benefits can bring and what the ways to use them are. But taking into account how fast this social trend progresses, we wonder if it is as good and beneficial as we are led to believe by the large companies, or, we might end up suffering the collateral consequences that no one has mentioned when describing the advantages of each product.

References

[1] O Arranz et al., *Surviving to the Education Online. Manual to Use Resources of Internet in the Classroom*. Salamanca, Ediciones Demiurgo, 2005.

[2] A. Pentland, "Una Sociedad dirigida por Datos*". Investigación y Ciencia*, pp. 46-51, Jan. 2014.

[3] C. Trevitt, E. Breman, and C. Stocks, "Assessment and Learning: Is it Time to Rethink Student´ Activities and Academic Roles?" *Revista de Investigación Educativa*, num. 30(2), pp. 253-270. 2012

[4] B. Tulasi, "Significance of Big Data and Analytics in Higher Education." *International Journal of Computer Applications.* 68, 14, pp. 21-23. April. 2013.

[5] J. Dean, *Big Data, Data Mining and Machine Learning*. Hoboken, New Jersey: John Wiley & Sons, 2014.

[6] M. V. Rijmenam. (2015, April 29). Four ways Big Data will revoluzione Education. *Datafloq*. [Online] Available: https://datafloq.com/read/big-data-will-revolutionize-learning/206

[7] O. Sánchez. (2014). Learning Analytics, el Big Data en Versión Educativa.

El Blog d'UPCnet. [Online]. Available: http://blog.upcnet.es/learning-analytics-el-big-data-en-version-educativa/

[8] L. Joyanes, *Big Data: Análisis de Grandes Volúmenes de Datos en Organizaciones*. México: Mancorbo, 2014.

[9] L. Johnson, S. Adams, V. Estrada, A. Freeman, *NMC Horizon Report: 2014 K-12 Edition*. Austin, Texas: The New Media Consortium, 2014.

[10] IBM, *Analytics: The Real World Use of Big Data. How Innovative Enterprises Extract Value from Uncertain Data. Executive Report*, IBM Institute for Business Value. 2013.

[11] J. Campo-Avila, R. Conejo, F. Triguero, R. Morales, "Mining Web-based Educational Systems to Predict Student Learning Achievements" *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 3, No. 2, pp. 49-54. 2015.

[12] R. Ferguson. (2012). The State of Learning Analytics in 2012: A Review and Future Challenges. *Technical Report KMI-12-01*, [Online]. Available: http://kmi.open.ac.uk/publications/techreport/kmi-12-01.

[13] P. Campbell, P. Deblois and D. Oblinger , "Academic Analytics: A New Tool for a New Era." *EDUCAUSE Review*, Vol. 42, No. 4, pp. 40-57. 2007.

[14] R. Baker, *Data Mining for Education.* Pittsburg: Carnegie Mellon University. 2009.

[15] F. Jacobi et al.. "Towards a design model for interdisciplinary information systems curriculum development, as exemplified by big data analytics education." In *Proc. of the ECIS 2014, Twenty Second European Conference on Information Systems*, Tel Aviv, June 2014

[16] V. Kellen et al.." Applying Big Data in Higher Education: A Case Study." *The Data Insight & Social BI, Executive Report*, 13 (8). 2013.

[17] D. M. West. (2012). Big Data for Education: Data Mining, Data analytics, and Web Dashboards. *Governance Studies, the Brookings Institution*.[Online]Available: http://www.brookings.edu/research/papers/2012/09/04-education-technology-west

[18] L. Castañeda, "Entre Construir Entornos Virtuales de Enseñanza-Aprendizaje y Enriquecer Entornos Personalizados de Aprendizaje," presented at the EDUTEC2008 Conf., Santiago Compostela, 2008.

[19] J. Ambrose. (2014, Jan 21) Strategy, Corporate Development and Emerging Business, *Skillsoft*, [Online]Available: http://www.skillsoft.com/about/press_room/press_releases/january_21_14_ibm.asp

[20] Ch. Pappas. (2014, Jul 24). Big Data in eLearning: The Future of eLearning Industry, [Online] Available: http://elearningindustry.com/big-data-in-elearning-future-of-elearning-industry

[21] R. Scapin, (2015, Mar 29) Big Data in E-Learning: Looking To the Future, [Online] Available: http://freevideolectures.com/blog/2015/03/big-data-in-e-learning-looking-to-the-future/

Vidal Alonso was born in Luanco, Spain, in 1966. He received the Computer Science Degree in 1992 from the Polytechnic University of Madrid, and the Ph. D. degree, in 2004 from the Pontifical University of Salamanca, Spain. He was a Full Professor of Computer Science at the Pontifical University of Salamanca since 1994. He has occupied the position of Vice rector at his University for five years, until 2015. He also was the Director of the Computer Science School for nine years (2000-2009.) He works in data structures, knowledge discovery and data quality. Dr. Alonso is a member of ALI (Computer Science Spanish Association) and he won the Castilla y Leon Digital Award in 2007.

Olga Arranz was born in Burgos, Spain, in 1973. She received the Computer Science Degree in 2010 from the San Antonio Catholic University of Murcia, and the Ph. D. degree, in 2013 from the Pontifical University of Salamanca, Spain. She was a Full Professor of Education Department at the Pontifical University of Salamanca since 2002. She has occupied the position of Secretary Technical Director at his University for four years, until 2015. She works in elearning processes, knowledge discovery and educational technology.

Social Network Analysis and Big Data tools applied to the Systemic Risk supervision

Mari-Carmen Mochón

Ph program Uned University

***Abstract* — After the financial crisis initiated in 2008, international market supervisors of the G20 agreed to reinforce their systemic risk supervisory duties. For this purpose, several regulatory reporting obligations were imposed to the market participants. As a consequence, millions of trade details are now available to National Competent Authorities on a daily basis. Traditional monitoring tools may not be capable of analyzing such volumes of data and extracting the relevant information, in order to identify the potential risks hidden behind the market. Big Data solutions currently applied to the Social Network Analysis (SNA), can be successfully applied the systemic risk supervision. This case of study proposes how relations established between the financial market participants could be analyzed, in order to identify risk of propagation and market behavior, without the necessity of expensive and demanding technical architectures.**

***Keywords* — Social Network Analysis, Big Data, Financial Markets, OTC Derivatives, Systemic Risk, Supervision, Graphical analysis, EMIR, ESMA, G20, Trade Repository.**

I. Introduction

Following the 2008 financial crisis, the G20 established an international forum for the heads of government of the world's major economies, to reach some consensus about the crash and put in place a high level plan to remediate those causes.

Since that moment, G20 jurisdictions started to draft and issue different regulatory reporting obligations, in order to provide greater transparency to the financial markets and supervisory capability to the national and international supervisors. In Europe several regulations, which imply a reporting obligation, have been implemented or are currently underway; EMIR [1], REMIT [2], SFTR [3] or MIFIR [4].

As a consequence, new market infrastructures appeared; trade repositories (TRs), to collect data from the industry, make sure of the reliability of the data, store it and make it available to regulators. Article 9 in EMIR (European Market Infrastructures Reform) mandates all counterparties to report details of any derivative contract they have concluded, or which they have modified or terminated, to a registered or recognized trade repository under the EMIR reporting requirements. TRs centrally collect and maintain the records of all derivative contracts.

OTC (Over the Counter) derivatives are financial products, which are negotiated bilaterally by two counterparties, and therefore are not registered in an official market. The OTC derivatives market, the financial engineering and the lack of transparency have played a key role in the crisis suffered in 2008.

Each of these regulations imply the daily reporting of billions of messages to be processed and stored in the trade repositories. And even though EMIR was the first regulation to impose a reporting obligation of the derivatives market to trade repositories in February 2014, further reporting regulations appoint TRs as the data centers for the collection of the financial records of different segments of the industry.

Currently in Europe there are several trade repositories operating under EMIR regulation. They collect the information from the whole derivatives industry and make it available to European national competent authorities (NCAs) as defined in article 81 of EMIR.

All market agents are challenged by these regulatory obligations. Market participants have to develop and implement new reporting flows, trade repositories have to create complex and reliable technical architectures to ensure confidentiality and robustness, and at the end of the chain, regulators need to develop the relevant tools to be able to digest and understand the millions of pieces of information they have at their disposal.

If systemic risk supervisors do not achieve their objectives, all the previous work, effort and investments will be of no use. National and international supervisors are facing these mandates with limited human and technical resources, and therefore the achievement of this objective is not always easy.

II. Objective Of The Case Of Study

Regulators whose mandate is to monitor the systemic risk are now facing an adaptation process as hard as the one suffered by market participants. Even if supervisors already analyze an important amount of information, after the 2008 crisis and the G20 commitment towards transparency of the financial markets, millions of new records are under the scope of the regulatory reporting.

This case of study proposes complementary methods to monitor the systemic risk, making usage of the tools that Big Data provides. Big Data technical solutions can cope with the 3 main issues when analyzing and managing data, the 3 Vs; velocity, variety and volume. In the financial industry, the variety of the data is not a technical barrier, as reports are transmitted in quite standardized formats. On the other hand, volume and velocity suppose a big challenge for regulators.

Currently TRs receive hundreds of millions of transactions on a daily basis. National and international authorities are the data consumers, but still need to acquire the capability of extracting the information hidden behind the millions of data records contained in the trade repositories.

Analysis methodologies currently used for Social Network Analysis can be equally applied to the analysis of the relations established between market participants. This could provide valuable information related to the market behavior, tendencies and confidence in the market and clearly identify those participants who hold the highest propagation risk.

Social network analysis (SNA) is the analysis process of mapping and measuring the relationships, connections and exchanges taking place between people, agents, organizations, machines... The nodes in the network are the participants and groups while the links show relationships or communication exchanges between the nodes. SNA provides both a visual and a mathematical analysis of human, or business in this case, relationships.

DOI: 10.9781/ijimai.2016.365

III. Methodologhy

A. Data subject to analysis & hypothesis

This case of study proposes the application of SNA methodologies and Big Data tools, currently used to analyze social networks to study the relations established between financial market participants and its tendencies within the OTC derivatives markets.

This exercise describes how graphical analysis tools could be applied to the OTC derivatives trades executed between European counterparties and reported to a trade repository under EMIR regulation. These reported trade details follow a predefined set of Reporting Technical Standards [5] (RTS) defined by ESMA (European Securities Markets Authority). Based on the reporting technical standards, it can be defined how such analysis should be designed.

Every derivative trade reported to a trade repository must include the reporting fields defined by ESMA in the reporting technical standards. Among the 85 fields subject to the reporting obligation, the following ones are the ones applicable to perform the proposed exercise; Reporting Counterparty ID (RTS field 1.3), ID of the Other Counterparty (RTS field 1.5), Venue of execution (RTS field 2.15).

The fields Reporting Counterparty ID and ID of the Other Counterparty represent a unique code identifying the counterparties to the transaction. The admitted values for this fields are; Legal Entity Identifier (LEI) (20 alphanumerical digits), Interim Entity Identifier (IEI) (20 alphanumerical digits), BIC (11 alphanumerical digits) or a client code (50 alphanumerical digits).

The field Venue of execution will be the datum used to distinguish between the ETD and OTC trades. According to the regulatory technical standards, the field Venue of execution should be informed as follows:

The field is alphanumerical of four characters, and shall contain the values ‘XXXX’ or ‘XOFF’ for bilateral trades (OTC) or in the case of reporting an Exchange Traded Derivative (ETD), the field shall be reported using a Market Identifier Code (MIC) included in the ISO 15022 specification.

The trades reported with “XXXX” or “XOFF” will be marked as OTC, the trades with MIC codes will be classified as ETD.

In order to proceed with the analysis, the following assumptions have been considered:

- The UTI that identifies each trade is truly unique and commonly used by each pair of counterparties.
- All counterparties correctly populate the “venue of execution” field.
- Each regulator has at its disposal the information related to the derivative transactions subject to its supervisory mandate, sent by the TRs.

B. Data collection methodology

The data collection methodology used for this case of study is the “Extract Transform Load and Analyze” process (ETL&A).

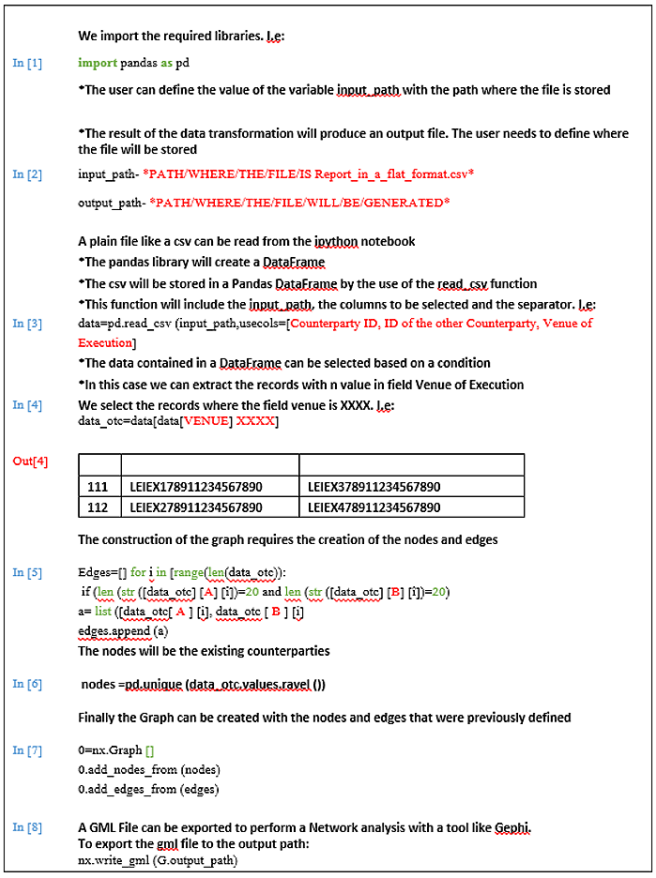
According to article 81 of EMIR, all the NCAs will have access to the information reported to a TR. This proposal assumes that the information to be analysed by a Competent Authority will be contained in a flat file like a CSV.

In order to enhance the user interface, a framework like the Jupyter notebooks [6] is used in order to explore the data contained in the reports. These notebooks are a powerful open source tool that provide a user-friendly interface in the use of different Python libraries [7] like Pandas [8], Numpy [9] or Networxs [10]. Python is an open source programming language that allows quick and flexible integration of systems.

The proposed ETL&A process will be composed of the following phases:

- Data extraction: extract the data from a flat file to a Jupyter notebook.
- Transform: the data is selected and sliced according to the analysis requirements criteria.
  - A new data frame including the; Reporting Counterparty ID, ID of the Other Counterparty and Venue of execution is created.
  - The relevant transformations are applied in order to define the relations established between counterparties of the OTC derivatives market.
  - Networx graph library is used in order to transform the data into a graph file.
- Load: load the graph file obtained into Gephi tool [11], an open source Social Network Analysis tool.
- Analyze: Gephi analysis and visualization tool provides outcomes and calculations that will illustrate the network and related metrics.

The below image represents an example of the script that should be created as part of the ETL process, using Jupiter Network_Analysis.



We import the required libraries. I.e:

In [1]
```
import pandas as pd
```

*The user can define the value of the variable input_path with the path where the file is stored

*The result of the data transformation will produce an output file. The user needs to define where the file will be stored

In [2]
```
input_path- *PATH/WHERE/THE/FILE/IS Report_in_a_flat_format.csv*

output_path- *PATH/WHERE/THE/FILE/WILL/BE/GENERATED*
```

A plain file like a csv can be read from the ipython notebook
*The pandas library will create a DataFrame
*The csv will be stored in a Pandas DataFrame by the use of the read_csv function
*This function will include the input_path, the columns to be selected and the separator. I.e:

In [3]
```
data=pd.read_csv (input_path,usecols=[Counterparty ID, ID of the other Counterparty, Venue of Execution]
```

*The data contained in a DataFrame can be selected based on a condition
*In this case we can extract the records with n value in field Venue of Execution

In [4]
We select the records where the field venue is XXXX. I.e:
```
data_otc=data[data[VENUE] XXXX]
```

Out[4]

111	LEIEX178911234567890	LEIEX378911234567890
112	LEIEX278911234567890	LEIEX478911234567890

The construction of the graph requires the creation of the nodes and edges

In [5]
```
Edges=[] for i in [range(len(data_otc)):
 if (len (str ([data_otc] [A] [i])=20 and len (str ([data_otc] [B] [i])=20)
a= list ([data_otc[ A ] [i], data_otc [ B ] [i]
edges.append (a)
```
The nodes will be the existing counterparties

In [6]
```
nodes =pd.unique (data_otc.values.ravel ())
```

Finally the Graph can be created with the nodes and edges that were previously defined

In [7]
```
0=nx.Graph []
0.add_nodes_from (nodes)
0.add_edges_from (edges)
```

In [8]
A GML File can be exported to perform a Network analysis with a tool like Gephi.
To export the gml file to the output path:
```
nx.write_gml (G.output_path)
```

Fig.1. ETL process script - Source: Own elaboration

C. Tools and programs applied to analyze the data

Gephi is a visualization and exploration software for all kinds of graphs and networks. It is an open-source free tool commonly used in the social networks analysis such as Twitter or Facebook, as provides simple information of complicated networks created by the billions of participants interacting via internet. The software uses a 3D render engine to display large networks in real-time and to speed up the exploration.

Once a graph file is generated and loaded in the Gephi application, the program draws a network that reflects the relations of all the different parties. Additional metrics are generated automatically by the application.

Social Network Analysis can be described as the process of investigating social connections, interactions and structures, by using network and graph theories [12]. The tools used for applying this

technique enhance the visualization of sociograms in which nodes are represented as points or entities and the interactions can be represented by bridges or lines. Such visualization has drawn significant attention in the recent years, because this technique helps researchers to understand, find and predict patterns, and interactions among social actors, i.e., identifying central actors, roles, subgroups or clusters.

The Gephi statistics and metrics framework offer the most common metrics for social network analysis (SNA) and scale-free networks:

- Betweenness Centrality, Closeness, Diameter, Clustering Coefficient, PageRank
- Community detection (Modularity)
- Random generators
- Shortest path

D. *Social Network Analysis metrics*

The following are some of the most relevant metrics used in SNA:

- Degree Centrality

Social network analysts measure the network activity of a determined node by using the concept of degrees, this is the number of direct connections a node has. The nodes holding the most direct connections in the network, are the most active nodes and therefore the 'connectors' or 'hubs' in the network.

- Betweenness Centrality

The location a node has in the network also determines the importance that node has in the network. A node might not have many direct connections, but if it has a good location, close to other important nodes, it may have a powerful 'broker' role in the network. On the other hand this nodes may imply a single point of failure. Supervisors must keep an eye on those nodes with high betweenness, as they has great influence over what flows, and does not, in the network. Location is a key element in network analysis.

- Closeness Centrality

Another important metric is the "closeness centrality" meaning how quickly a node contact any other one. I node may not have many direct and indirect ties, but still have access to all the nodes in the network more quickly than anyone else. These nodes have the shortest paths to all others, meaning that they are close to everyone else. These nodes are in an excellent position to monitor the information flow in the network, as they have the best visibility into what is happening in the network.

- Network Centralization

The relationship between the centralities of all nodes can reveal much about the overall network structure. A very centralized network is dominated by one or a few very central nodes. If these nodes are removed or damaged, the network quickly fragments into unconnected sub-networks. A highly central node can become a single point of failure or financial crash in that market. A network centralized around other set of well-connected hubs can find the turnaround if that hub is disabled or removed. Hubs are nodes with high degree and betweeness centrality.

The less centralized a network is established, the less single points of failure it has. Healthy and robust markets should not be very centralized in order to provide alternatives in the market in case of an agent making default in their financial liabilities.

IV. Results

The graph below is an example of the graphical illustration a regulator could obtain following the proposed methodology to analyze the connections established by the different market participants of a given target sample. Each circle would represent a market participant. Whenever a party is connected to a higher number of other market participants, the circle is represented bigger. This is measured through the "degree". The degree of a node is the number of relation (edge) it has.

Fig.2. Example Network graph result – Source: Gephi 0.9.0

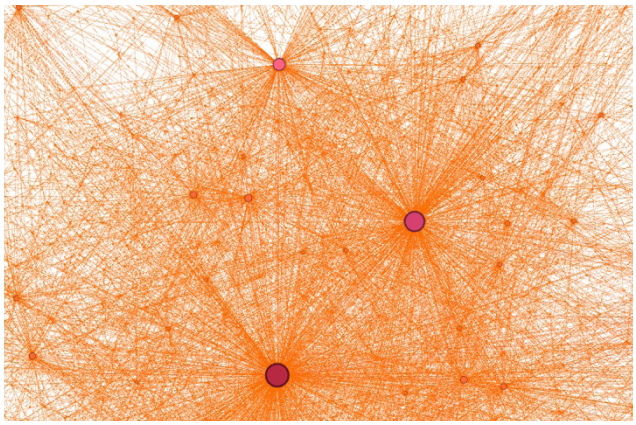


Fig.3. Example Network graph result – Source: Gephi 0.9.0

This graph shows a highly connected network were only a few nodes a significantly bigger that the others. These nodes have the highest number of connections established in the network. Consequently, they would represent a higher propagation risk. Should a credit event occur to any of these nodes, it can be assumed that an important size of the market would be affected.

The layout of the graph distributes the nodes within the graph depending on the closest relations and communities identified.

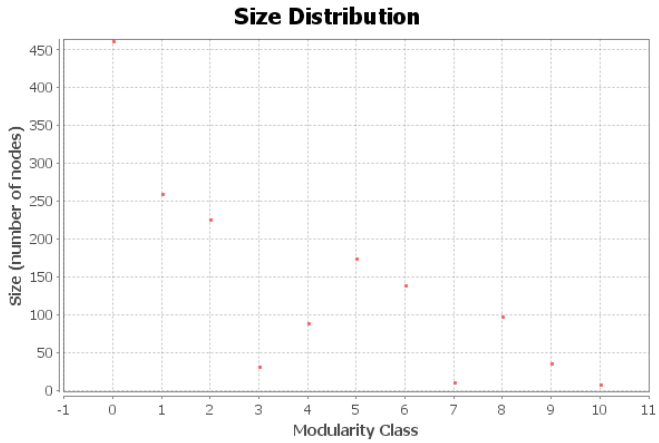


Fig.4. Example of System Modularity representation – Source: Gephi 0.9.0

In addition to the main graph that represents all the relations established in the network, the Gephi application generates the most common metrics for social network analysis and scale-free networks.

For instance, the modularity of the given example is reflected in Fig.5. This system is made up of relatively independent but interlocking communities of members.

The degree of interconnection between counterparties is also identified. A few nodes have a high degree. This would mean that just a few counterparties have the bulk of connections. Presumably, those counterparties or systemic risk concentrators, represent the sell side of the trades, providing financial services to many market participants.

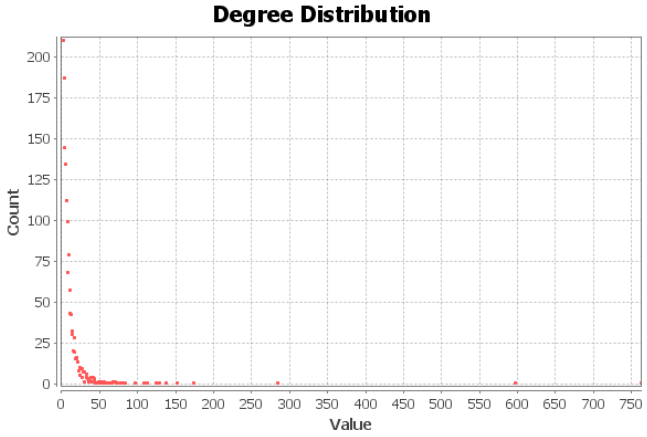


Fig.5. Example System Degree – Source: Gephi 0.9.0

V. Conclusions

Financial markets are clear examples of social structures whose behavior is dictated by external factors. Macroeconomic or microeconomics events, expectations, speculation or risk appetite are some of the factors that compose a market sentiment which is expressed through the commercial relationships established among themselves. As stated in the article "The Digital Economy: Social Interaction Technologies" [13]; "*the daily activities of many businesses are being socialized*". Consequently, these connections and interactions can by analyzed just as any other social network.

The proposed approach may be used to analyze market behavior, tendencies and confidence feelings between counterparties by studying the relations established market participants. Additionally, the identification of the relations established between market participants can disclose information regarding propagation risk factors, and potential cascading failure situations. Through a simple process of analysis, it is possible to identify who are the market participants highly trusted and active in their respective financial fields. These participants concentrate the highest propagation risk and in consequence are systemic risk concentrators.

When many participants establish relations with different counterparties, confidence reigns in the market, and participants are not very sensible to the counterparty credit risk. On the other hand, when the number of relations decrease and only few participants are trusted in the market, it can be interpreted that the market is suspicious and already anticipating potential credit defaults.

Additionally, by analyzing the evolution of the network interaction, it is possible to monitor market feelings and confidence in the market. Competent authorities can monitor market sentiment, by comparing the different snapshots of the market network, by visualizing at a macro level the existent relationships across market counterparties. The analysis tool offers dynamic graph analysis, where users can visualize how a network evolve over time by manipulating the embedded timeline.

It is also important to point out that the tools used in this analysis are free and open-source, therefore no initial cost, license fee or important technical architecture investment is required. Still, these powerful analytic tools enable supervisors to extract relevant information. Regulators with budget constrains can contemplate these analytical tools as an option.

This is a complementary analysis that could be used by regulators in addition to the traditional analysis already performed. Further and more complex studies could be performed with these and other Big Data tools.

References

[1] European Markets Infrastructure Regulation http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32012R0648
[2] Regulation Energy Market Integrity and Transparency https://www.acer-remit.eu/portal/public-documentation
[3] Securities Financing Transaction Regulation http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32015R2365
[4] Markets in Financial Instruments Regulation http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014R0600
[5] ESMA Reporting Technical Standards http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:052:0001:0010:EN:PDF
[6] Jupyter notebook http://jupyter.org/
[7] Python standard library https://docs.python.org/2/library/
[8] Pandas documentation: http://pandas.pydata.org/pandas-docs/version/0.17.1/
[9] Numpy documentation http://docs.scipy.org/doc/
[10] Networxs documentation https://networkx.github.io/documentation/latest/
[11] Gephi documentation https://gephi.org/users/
[12] Otte, Evelien; Rousseau, Ronald (2002). "Social network analysis: a powerful strategy, also for the information sciences"
[13] Teófilo Redondo "The Digital Economy: Social Interaction Technologies". International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 3, Nº 2. 2015

Mari-Carmen Mochón, Institutional Relationship Manager, REGIS-TR. She was born in Sevilla in 1983. In 2006 she was licensed in Business Administration by the University of CUNEF in Madrid and a master title in "Financial Markets and Alternative Investment" in the Instituto BME in 2008. She is Specialized in Regulatory reporting systems/compliance, Institutional and Government Relations, OTC Derivatives products & Financial Markets. She has worked for the Spanish Derivatives Exchange supervision area until 2009. After she worked at the Spanish Central Bank for the European Central Bank, in the creation of a European securities settlement platform. In 2011 she moved to REGIS-TR, for the creation project of a European Trade Repository owned by the Deutsche Börse Group & Bolsas y Mercados Españoles. Currently she is the Institutional Relationship Manger, being responsible for International and European government affairs as well as Market Institutions She has combined her professional career in the financial sector with an academic path; University teacher at Universidad Pontificia de Salamanca (UPSAM), teaching "International Economy" and "Economy of the Europe Union". She has also being post-grade program teacher, in the course "Economic crimes and prevention of financial risk", at the National University for Distance Education (UNED). She has several publications such as; "Teoría de las organizaciones", F. Mochón, M.C. Mochón y M. Sáez, 2015 Alfaomega, Mexico, "Gestión organizacional", F. Mochón, M.C. Mochón y M.Sáez, 2015 Alfaomega, Mexico, "Administración", F. Mochón, MC. Mochón, M. Sáez, 2014. Alfaomega, Mexico and "Economía" in a team of three, 2002. McGraw Hill, Spain. Currently she is finalizing a master degree in Business Intelligence and Big Data in Colegio Universitario Cardenal Cisneros, Madrid, and it is enrolled in the doctoral program at the Faculty of Economics, UNED University.

Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy

Diego J. Bodas-Sagi, José M. Labeaga

Universidad Francisco de Vitoria, Universidad Nacional de Educación a Distancia (UNED), Spain

***Abstract* — The growing demand for affordable, reliable, domestically sourced, and low-carbon electricity is a matter of concern and it is driven by several causes including public policy priorities. Policy objectives and new technologies are changing wholesale market design. The analysis of different aspects of energy markets is increasingly on the agendas of academics, firms' managers or policy makers. Some concerns are global and are related to the evolution of climate change phenomena. Others are regional or national and they strongly appear in countries like Spain with a high dependence on foreign energy sources and high potential of domestic renewable energy sources. We can find a relevant case in Spanish solar energy policy. A series of regulatory reforms since 2010 reduce revenues to existing renewable power generators and they end up the previous system of support to new renewable generation. This policy change has altered the composition of the energy market affecting investment decisions. In this paper, we analyze the public opinion about energy policy of the Spanish Government using the Global Database of Events, Language, and Tone (GDELT). The GDELT Project consists of over a quarter-billion event records in over 300 categories covering the entire world from 1979 to present, along with a massive network diagram connecting every person, organization, location, and theme to this event database. Our aim is to build sentiment indicators arising from this source of information and, in a final step, evaluate if positive and negative indexes have any effect on the evolution of key market variables as prices and demand.**

Keywords — Big Query, GDELT, Public Opinion, Energy, Electricity.

I. Introduction

Public policy plays a critical role in regulating relationship between companies, investors and society. The importance of public policy for long-term investors has grown in recent years, due to [1]:

- Legislative reform of the financial sector in the wake of the global financial crisis.
- Governmental need for investors as a source of long-term growth.
- The increasing impact of environmental, social and governance factors on the ability of investors to deliver long-term returns.

One of the key case for studying the effects of public policy is the energy market. The growing demand for affordable, reliable, domestically sourced, and low-carbon electricity is on the rise. It "is driven in part by evolving public policy priorities, especially reducing the health and environmental impacts of electricity service… Well-designed markets encourage economically efficient solutions, promote innovation and minimize unintended consequences" [2].

Policy objectives and new technologies are changing wholesale market design. A relevant case is the Spanish solar energy policy. A series of regulatory reforms since 2010 reduce revenues to existing renewable power generators and end the previous system of support to new renewable generation. This policy change has caused several claims by various organizations and altered the composition of the energy market. At the end, the Royal Decree of October 2015 strongly affected the solar energy market.

The analysis of the public opinion about specific government measures may be a useful component for some decisions of the agents on the long-term. The public opinion can paly a role on influencing policy makers at several stages of their decision process and this public perception could affect the design of political programs or policy measures. Agents in any market can also be influenced by the positive and negative perceptions about the company. In any case, public representatives can be interested in analyzing the state of the public opinion when adopting measures which can distort the markets.

In this paper, we analyze the public opinion about energy policy of the Spanish Government using the Global Database of Events, Language, and Tone (GDELT). The GDELT Project [3] consists of over a quarter-billion event records in over 300 categories covering the entire world from 1979 to present. Our aim is to build sentiment indicators arising from this source of information and, in a final step, evaluate if positive and negative indexes have any effect on the evolution of key market variables as prices and demand. We do not try to evaluate whether there are causal effects from the sentiment indicators to the market variables but our purpose is to detect the existence of correlation among those variables.

The rest of the paper is structured as follows: First, we provide a brief summary of the GDELT Project and explain the relation with the Big Data paradigm. In section 3, we explain the tools, methods and techniques used in this study. The results obtained are discussed in section 4. The paper ends up providing some policy implications and ideas for future research.

II. The Gdelt Project

The GDELT Project, supported by Google Ideas, share real-time information and metadata with the world. This codified metadata (but not the text of the articles) is then released as an open data stream, updated every 15 minutes, providing a multilingual annotated index of the information. It includes broadcast, print, and online news sources. The project shares a database with trillions data points. Although, data is available as downloadable CSV files, few users have the storing capacity and processing power to download terabytes of data, and effectively query and analyze it. Google's BigQuery platform provides a way to interact with this huge information source. GDELT is a clear example of Big Data, while Google's BigQuery is an example of Infrastructure As a Service (IaaS) technology.

According to [4], Big Data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves

DOI: 10.9781/ijimai.2016.366

too fast, or doesn't fit the structures of our database architectures. To gain value from this data, one must choose an alternative way to process it. Big Data technologies have huge variety of sources, huge volume of information – so much less time is needed to process information thanks to parallel processing and clustering infrastructure.

GDELT maintains the GDELT Event Database, and the GDELT Global Knowledge Graph (GKG). The GKG begins April 1, 2013 and "… attempts to connect every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day" [3]. The data files use Conflict and Mediation Event Observations (CAMEO) [5] coding for recording events. GKG also provides event identification (*EventIDs*) of each event found in the same article as the extracted information, allowing rich contextualization of events.

III. Methodology

In this work, we have used GKG table on Google's BigQuery platform. GKG table provides the "Themes" attribute, the list of all themes found in the document. We want to filter documents related to, at least, one of these two themes: "ENV_SOLAR" (which refers to solar power in general), and "FUELPRICES" (which refers to cost of fuel, energy and heating). The theme attribute is not available for the Event table. At the same time, we have looked for events that refer to Spain at some point using the "Locations" attribute (which contains a list of all locations found in the text). In summary, we are using GKG table to filter information about solar power or cost of fuel, energy and heating and related (in some way) with Spain. Attribute "V2Tone" allows us to analyze the average "tone" of the document as a whole. The score ranges from -100 (extremely negative) to +100 (extremely positive). Common values range between -10 and +10, with 0 indicating neutral. This is calculated as Positive Score minus Negative Score. Positive Score is the percentage of all words in the article that were found to have a positive emotional connotation. Negative Score is the percentage of all words in the article that were found to have a negative emotional connotation. Big Query allows interaction with the whole GDELT dataset using Structure Query Language (SQL). An account in Google Cloud Services and activate Google's Cloud Storage to export and download data is required.

R [6] has been used to analyze and process data. Downloaded data from Google's Cloud Storage have been imported into R. After that, we have done a basic and exploratory analysis of the downloaded data. The analysis shows that there are some references, documents or URL's that are not related to energy policy. For example, some entries refer to scientific news from Canary Institute of Astrophysics ("ENV_SOLAR" theme). For this reason and for a more efficient measurement, we feel that it is necessary to analyze the text of the news and look for references to Spanish Government, council, ministry or ministers. This is a computation intensive task because it requires to deal with HTML tags and extract the text of the document. In the next section, we detail this process and explain different alternatives to improve execution time.

Next and for each theme, we have grouped by day all mentions and calculated the mean tone and typical deviation in tone per day. At this stage, we only have evaluated documents written in Spanish or English because we need to find mentions to Spanish Government (Spanish and English have been the chosen languages to process, other languages will be included in future versions). The results have been placed into context with the policies that the government of Spain has implemented.

Finally, we have applied a Correlated Topics Models (CTM) algorithm [7], based on Latent Dirichlet Allocation (LDA) algorithm [8], on the words contained in the mentions, news or documents written in Spanish. The rest of the dataset has not been included in this part of the study. We have not mixed languages, words with similar meaning but from different languages can be placed on different topics because writing is different. We leave the evaluation of other languages for future research. The "topicmodels" R package [9] allows us to execute LDA and CTM algorithms.

LDA allows to discover topics in large data collections described via topics. It does not require labeled data (unsupervised learning) and uses a stochastic procedure to generate the topic weight vector. LDA represents documents as mixtures of topics that spit out words with certain probabilities. It is a bag-of-words model. For this reason, LDA can be used for document modelling and classification. LDA fails to directly model correlation between the occurrence of topic and, sometimes, the presence of one topic may be correlated with the presence of another (for example: "economic" and "business"). CTM is very similar to LDA except that topic proportions are drawn from a logistic normal distribution rather than a Dirichlet distribution. Applying the "topicmodels" R package to our dataset, we can obtain a list of words for every topic and, also, check the correlation between the topics obtained.

A. Getting the text of the documents

As we have mentioned before, getting the text of the documents is a compute-intensive phase because it requires to deal with HTML tags and extract the text of the document. This task is done in the following steps:

1. First, documents are accessed through its URL.
2. We detect the text language using "textcat" R package [10]. If the document is written in Spanish or English we download the text and confirm that the text contains one of the following words: "gobierno", "government", "council", "ministers", "ministry", "ministro", "ministerio". In other case, we consider that the text does not mention Spanish Government. One should note that Spain location is referenced in the text according to GKG metadata.
3. For all downloaded texts, we clean the text removing HTML tags and stop-words (Spanish and English) in order to improve accuracy and performance. This task reduces the text size. Stop-words refer to the most common words in a language but given our aims they do not add value to the analysis of the topic.

Sequential and parallel execution modes have been tested here. A parallel algorithm, as opposed to a traditional sequential algorithm, is one which can be executed a piece at a time on many different processing p devices or processors, and then put back together again at the end to get the correct result.

The usefulness of this type of parallelization is that once the program structure is known, a few changes must be made to the program to be executed by several processors, and not as a distributed algorithm in which, first, we must establish the optimal communication structure. This usually involves making substantial changes to the program.

Two parallelization schemes have been evaluated. In the first one, we take advantage of multiple cores in one computer. The second parallelization method employs a master-slave architecture [11]. This architecture features a single processor running the main algorithm (master) which delegates the mission of getting the text among a group of processors (slaves). Slaves are responsible for processing URLs and getting the text and communicating results to the central process.

In any case, if we have p processors, the original dataset is divided into p chunks, one processor processes only one of these chunks. In all cases, we have used a computer with Pentium V quad core and 8 GB RAM, managed by Operating System Centos 6.4. The Internet broadband speed is, roughly, 20 Mbps (download speed). Performance are usually measured in terms Speedup (Sp) and Efficiency (Ep):

$$S_p = T_1 / T_p \quad (1)$$

$$E_p = S_p / p \quad (2)$$

where p is the number of processors, T_1 is the execution time of sequential algorithm and T_p is the execution time of the parallel implementation on p processors.

The Simple Network of Workstations (snow) package [12] allows executing parallel code in R. It requires loading the code, loading the snow library, create a snow cluster (or execute in local mode using multicore CPU) and running the code, maintaining this order. Snow library can be used to start new R processes (workers) in our machine. The snow package is a scatter/gather paradigm, which works as follows:

1. The manager partitions the data into chunks and parcels them out to the workers (scatter phase).
2. The workers process their chunks.
3. The manager collects the results from the workers (gather phase) and combines them as appropriate to the application.

Snow can be used with socket connections, Message Passing Interface (MPI), Parallel Virtual Machine (PVM), or NetWorkSpaces [12, 13]. The socket transport does not require any additional packages, and is very portable. We have used socket connections. Snow is a non-shared-memory system example, if we are using a network of workstations, each workstation has its own and independent memory. But, in the multicore and one-computer case, the memory is shared between all the running processes. In both cases, the cost of communications should be kept in mind. The cost of communication is dependent on a variety of features including the programming model semantics, the network topology, data handling and routing, and associated software protocols. Reducing the computation time by adding more processors would only improve marginally the overall execution time as the communication costs remains fixed.

IV. Results

A. Results using data from GDELT GKG

We are analyzing tone in mentions from GKG database. All mentions have some common characteristic, they refer to Spain as location at some point, themes "ENV_SOLAR" or "FUELPRICES" are detected in the text and, the lines contain one of the following words: "gobierno", "government", "council", "ministers", "ministry", "ministro", "ministerio" (we interpret it as the text mentions the Spanish Government). GDELT 2.0 and GKG new version are relativity recent. For this reason, we only have data filling these requisites from February 18th, 2015 to October 28th, 2015. Since is a project in constant update, in order to close our study, we refer here to data obtained in our last interaction with Google's BigQuery on October 28th, 2015.

The following figures show the mean and typical deviation in mentions (tone). All mentions (mentions without filtering words nor languages) and only government mentions are displayed and compared. Blue bars represent mean values in tone, black ones represent error bars. According to Fig. 1 and its histogram in Fig. 3, the average index of the mentions due to fuel and energy prices are negative indicating that the sentiment of news related to the solar energy policies is negative through this period. We must remember that the government introduced in October 2015 what was named as solar tax ("*impuesto al sol*") regulating consumption made by consumers who produce their own energy through photovoltaic systems. The discussions at the media did not began at the time of publishing the Royal Decree on October the 9th, 2015 but several months before as soon as the agents knew government's intention. It is not strange that the sentiment of the agents producing news is negative.

On the other hand, when we include the word government as a control to build the sentiment index, the average as presented in figures 2 and 4 is still more negative. So, the agents (producers, consumers, etc.) clearly express a negative reaction towards the fuel and energy prices and we associate it to the regulations in the energy market referred to these variables. The opinion expressed in other surrounding countries of Europe and also by the authorities of the UE was also negative towards the regulation.

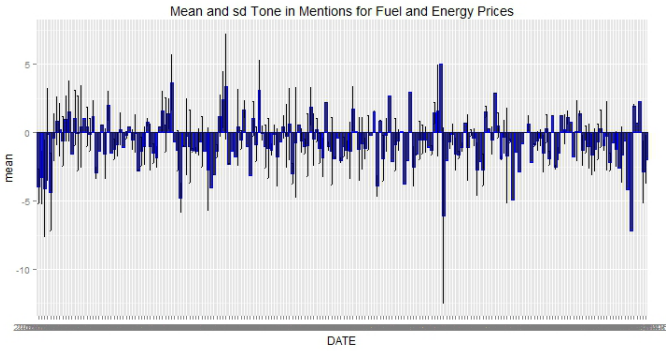


Fig. 1. All mentions for "FUELPRICES" theme in Spain.

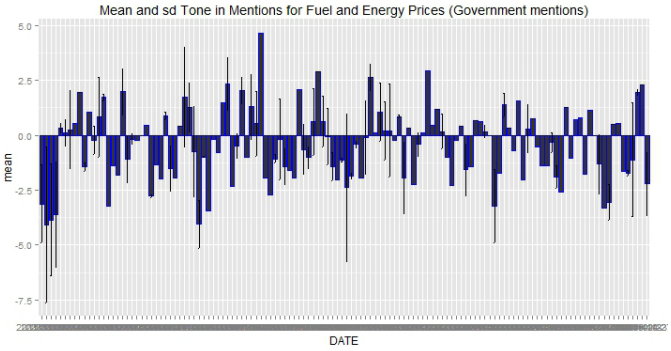


Fig. 2. Mentions for "FUELPRICES" theme in Spain filtering words (government mentions).

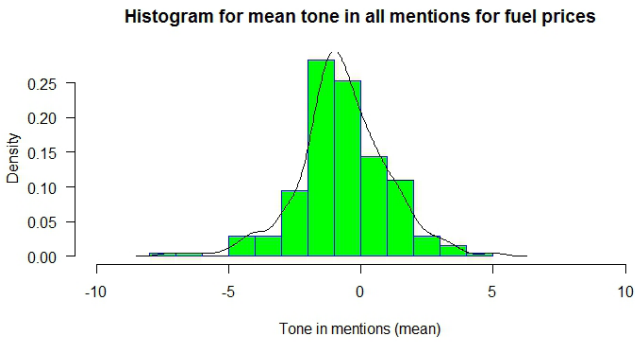


Fig. 3. Histogram - All mentions for "FUELPRICES" theme in Spain.

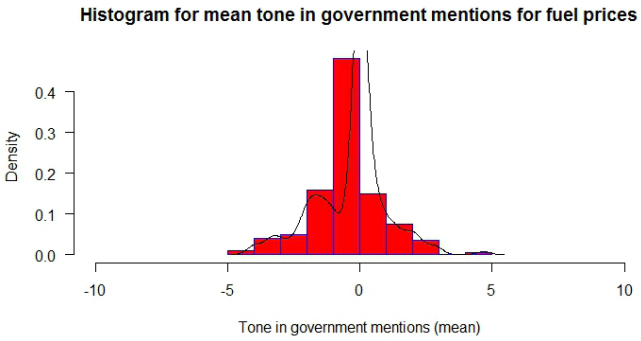


Fig. 4. Histogram - Mentions for "FUELPRICES" theme in Spain filtering words (government mentions).

In order to be able to use these data and conduct some test on them, we first check whether the indexes are normally distributed. figures 5 and 6 present Q-Q plots, which are probability plots, i.e., a graphical method for comparing two probability distributions by plotting their quantiles against each other. Here, we use Q-Q plot to compare data against Normal Distribution with mean and standard deviation according to the sample. Formally, the Shapiro-Wilk test [14] allows us to reject normality. For all data samples mean values are near to zero while typical deviation values are between 0.7 and 1.7. A normal distribution is symmetric about its mean, but this is not the case, and, taking into account the figures, we detect some extreme positive values which are balancing out a more frequent negative values and, for this reason, the mean tone in close to zero.

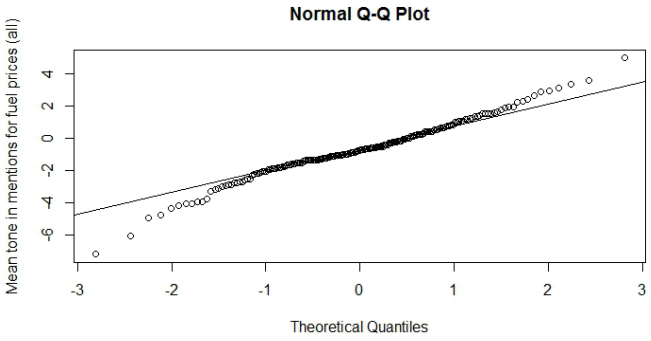


Fig. 5. Q-Q Plot - All mentions for "FUELPRICES" theme in Spain.

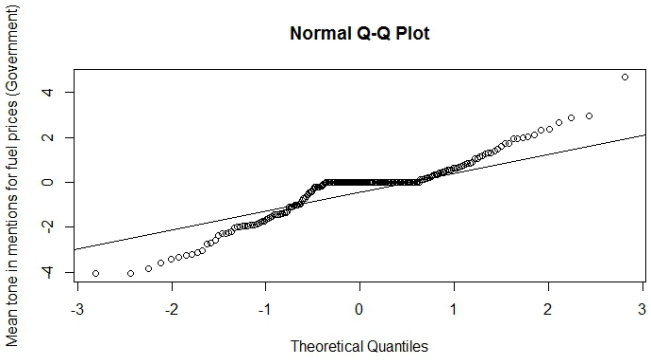


Fig. 6. Q-Q Plot - Mentions for "FUELPRICES" theme in Spain filtering words.

Next, we conduct a similar exercise but filtering in GDELT a different theme than before. So, we include environment and solar ("*ENV_SOLAR*" theme) to the previous exercises and analyze tone in the same way. We can see than the sentiment index does provide some negative tone messages when we do not filter using words related to the government. However, once we filter for words related to the government, the negative tone appears much more clearly.

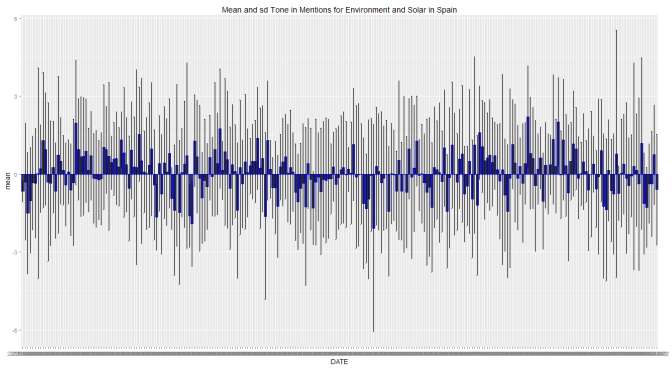


Fig. 7. All mentions for "ENV_SOLAR" theme in Spain.

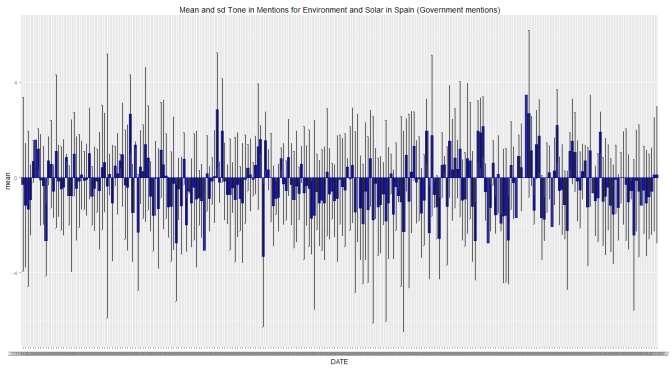


Fig. 8. Mentions for "ENV_SOLAR" theme in Spain filtering words (government mentions).

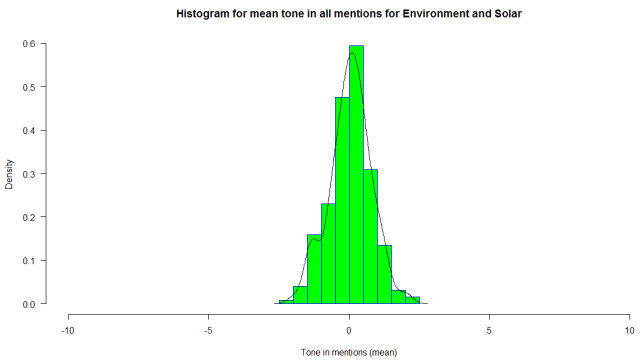


Fig. 9. Histogram – All mentions for "ENV_SOLAR" theme in Spain.

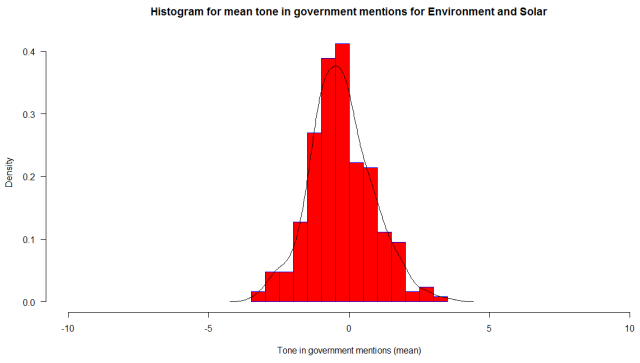


Fig. 10. Histogram – Mentions for "ENV_SOLAR" theme in Spain filtering words (government mentions).

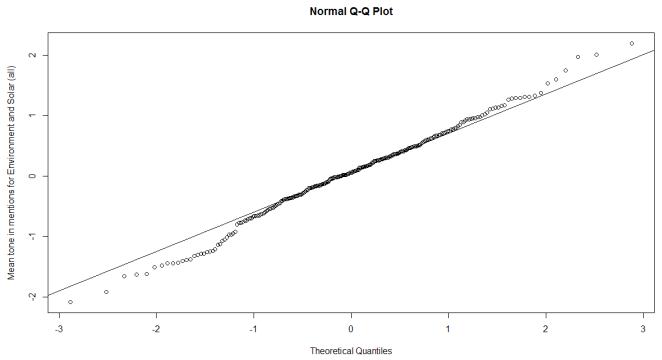


Fig. 11. Q-Q Plot – All mentions for "ENV_SOLAR" theme in Spain.

Despite the graphic, Shapiro-Wilk normality test produces no suspicion about normality. But the mean tone, although is close to zero, is negative.

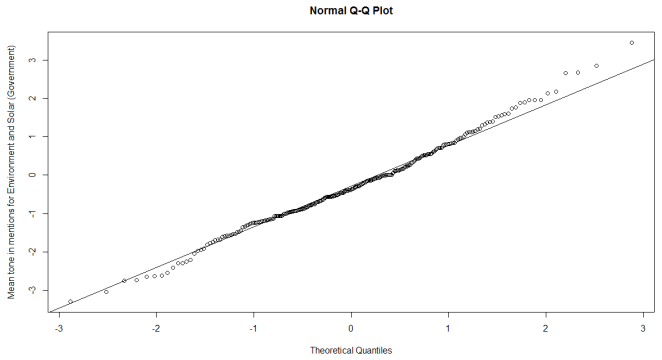


Fig. 12. Q-Q Plot – Mentions for "ENV_SOLAR" theme in Spain filtering words.

B. Correlation analysis: Prices and Demand

We can obtain historical data about electricity prices and demand from OMIE [15]. OMIE manages the electrical market for Spain and Portugal (MIBEL Market). We have used data from February 18th, 2015 to October 28th, 2015 (similar to GDELT data). Our aim is to evaluate whether there is any type of correlation between prices or demand in the MIBEL market and the mean tone of public opinion evaluated thanks to GKG data. For prices and demand, we have taken natural logs. Variables in Fig. 13 must be interpreted in the following way: *MeanALLSolar* (*ENV_SOLAR* theme for Spain) does not include references to Spanish Government, *MeanGovSolar* does include it. Interpretation is similar for *MeanALLFuel* and *MeanGovFuel* (for *FUELPRICES* theme in this case). *LogPrices* and *LogDemand* refer to logarithm mean and daily values for prices and demand (respectively) from MIBEL historical data and the same period.

	MeanALLSolar	MeanGovSolar	MeanALLFuel	MeanGovFuel	LogPrices	LogEnergy
MeanALLSolar	1.0000	0.4132	0.0170	-0.0161	-0.0318	0.0047
MeanGovSolar	0.4132	1.0000	-0.0023	-0.0442	-0.0421	0.0302
MeanALLFuel	0.0170	-0.0023	1.0000	0.4228	-0.0708	-0.0446
MeanGovFuel	-0.0161	-0.0442	0.4228	1.0000	-0.0149	-0.0062
LogPrices	-0.0318	-0.0421	-0.0708	-0.0149	1.0000	0.4234
LogEnergy	0.0047	0.0302	-0.0446	-0.0062	0.4234	1.0000

Fig. 13. Correlations results.

There is some week evidence of correlation between *LogPrices* and mean tone collected by *MeanALLFuel*. A test of the null that the coefficient (-0.0708) is equal to zero gives a normal value of -1.65, which is significantly different from zero at 9.9 percent of significance. The negative coefficient of correlation between *LogPrices* and mean tone collected by *MeanGovSolar* has a p-value of 0.32 for testing the same assumption. Finally, the correlation between *LogPrices* and mean tone collected by *MeanALLFuel* is not significantly different from zero at standard levels. Public opinion could to some extent weakly affect fuel prices and, indirectly, fuel demand in the short-term.

C. CTM results using CTM algorithm

In this subsection we present the results obtained using a CTM algorithm to discover and correlate topics. We must note that we only have applied the algorithm to Spanish texts. The R package "topicmodels" allow us to display the graphs collected in Fig. 14. For the theme named "FUELPRICES" and text written in Spanish, the cluster Group 1 corresponds to HTML tags or other words that have not been properly removed or English words that appear in texts that "textcat" R package has been classified as written in Spanish. On the other hand, clusters named Group 3 and 6 refer to words like (translated from Spanish) "stock exchange", "market", "government", "income", "Europe", "state", "congress"…, "gas", "price", "growth" and other Spanish locations as "Madrid" or "Barcelona" are words contained in the rest of the groups.

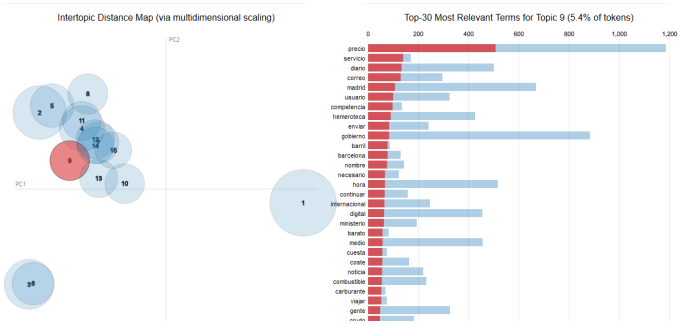


Fig. 14. Using "topicmodels" R package. "FUELPRICES" theme, text written in Spanish.

For "ENV_SOLAR" and text written in Spanish, Group 1 is similar to the last case. Group 5, 12 and 15 refer to words like "solar", "system", "electricity", "law", "change", "tax", and months (written in Spanish).

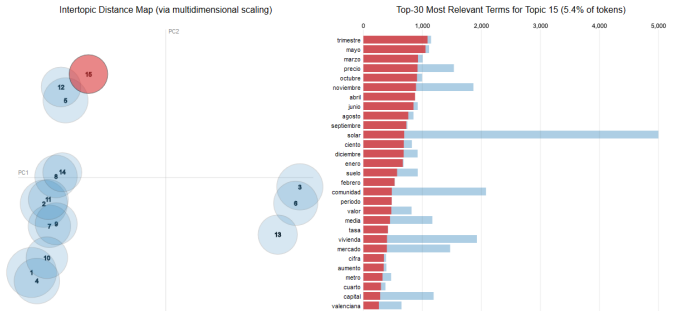


Fig. 15. Using "topicmodels" R package. "ENV_SOLAR" theme, text written in Spanish.

The CTM model allows to classify documents or news in media. We also think that we will be able to use it in the future to do further correlation analysis. Of course our final aim will be to use this technique in future research to do causal analysis from the information collected (the indexes built on it) and the movement of key variables in energy markets.

D. Speedup and Efficiency analysis (getting the text)

As we have explained in the methodology section, getting the text from URLs is a compute-intensive phase. We present two figures for analyzing sequential and parallel execution modes. They summarize the performance in terms of Speedup and Efficiency according to equations (1) and (2). Fig. 16 shows that speedup improves when using

a network of workstations. Although efficiency (in terms of reducing execution time) increases with multicore execution a network of workstations is still preferred:

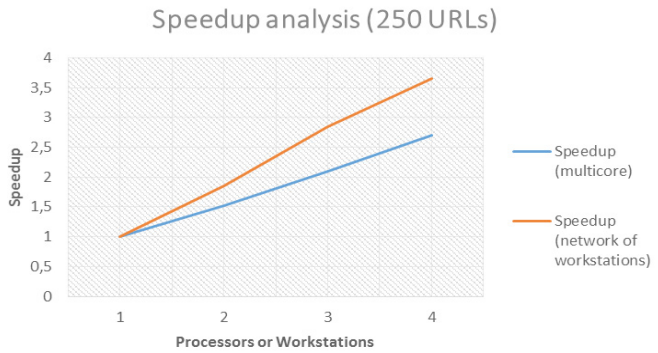


Fig. 16. Speedup comparison.

Our code does not require the dispatch of regular data between processes. Therefore, when we are using a network of workstations, the communication cost is not excessive and efficiency can be maintained at a constant level. However, in the multicore case the computer memory has to be shared and this issue impacts in the efficiency values.

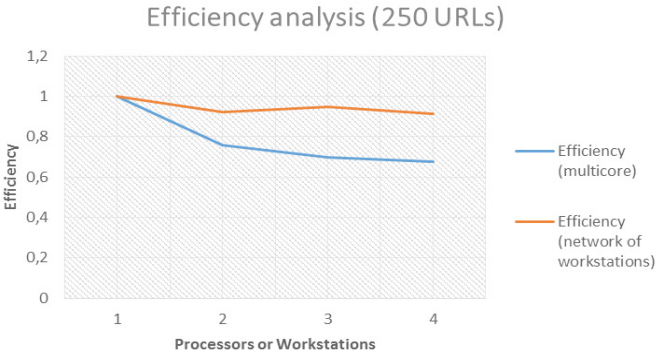


Fig. 17. Efficiency analysis

A multicore execution can be used to reduce execution time. Nevertheless, a network of workstations is preferred.

V. Conclusions And Future Work

In this paper we have used extensive data from several sources to analyze two issues related to energy markets. First, we analyze the public opinion about energy policy of the Spanish government using GDELT. Second, we conduct a correlation analysis between sentiment variables about the public policy and real prices and demand taken from the MIBEL energy market for the same period. Two results are worth emphasizing. On one hand, we detect negative feelings about the solar energy policy introduced by the Spanish government in 2015. On the other, hand, we find weak correlation between the indexes (tone) in mentions from GKG database and average daily log prices of energy. We do not find any correlation to average daily energy demand.

There are many extensions using extensive databases like the one in this paper or similar to follow different research lines in the future. We only quote two possibilities closely related to this exercise. First, we have only taken into account Spanish or English while alternative languages could be important to build sentiment indexes. Second, we have only presented correlation analysis between the indexes and average prices and demand but some formal demand model where to include these indexes as explanatory variables is necessary to accurately measure the potential of these variables to explain the evolution of key energy market variables.

Acknowledgment

We acknowledge very useful comment from an editor of the journal.

References

[1] R. Sullivan, W. Martindale, N. Robins, and Winch H. (2014, November). *Principles for Responsible Investment. Policy Frameworks for Long-Term Responsible Investment: The Case for Investor Engagement in Public Policy*. Available: http://www.unpri.org/wp-content/uploads/PRI_Case-for-Investor-Engagement.pdf

[2] J. Cochran, et al. (2013, October). *Market evolution: Wholesale electricity market design for 21st century power systems*. Technical Report. Contract. Available: http://www.nrel.gov/docs/fy14osti/57477.pdf

[3] The GDELT Project: Watching Our World Unfold. http://gdeltproject.org/

[4] E. Dumbill, *Planning for big data*. O'Reilly Media, Inc. 2012.

[5] D. J. Gerner, R. Abu-Jabr, P. A. Schrodt, and Ö. Yilmaz. "Conflict and Mediation Event Observations (CAMEO): A new event data framework for the analysis of foreign policy interactions." *International Studies Association*, New Orleans (2002).

[6] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available: https://www.R-project.org/.

[7] D. Blei, and J. Lafferty. "Correlated topic models." *Advances in neural information processing systems,* vol. 18, pp. 147. 2006.

[8] D. Blei, Y. Andrew, and M. I. Jordan. "Latent dirichlet allocation." T*he Journal of machine Learning research,* vol 3, pp. 993-1022. 2003.

[9] K. Hornik, and B. Grün. "topicmodels: An R package for fitting topic models." *Journal of Statistical Software,* vol 40.13, pp: 1-30. 2011.

[10] K. Hornik, P. Mair, J. Rauch, W. Geiger, C. Buchta, and I. Feinerer. "The textcat package for n-gram based text categorization in R." *Journal of Statistical Software* vol 52.6, pp. 1-17. 2013.

[11] E. Cantú-Paz, Erick. *Designing Efficient and accurate Parallel Genetic Algorithms*. Ph.D. Dissertation. University of Illinois at Urbana-Champaign, Champaign, IL, USA. Advisor(s) David E. Goldberg. 1998.

[12] L. Tierney, A. J. Rossini, N. Li, H. Sevcikova. (2008). *Snow: simple network of workstations*. R package version 0.3-3, Available: https://cran.r-project.org/web/packages/snow/index.html

[13] N. Matloff, *Parallel Computing for Data Science: With Examples in R, C++ and CUDA*. Chapman and Hall/CRC. 2015.

[14] S. S. Shapiro, and M. B. Wilk. "An analysis of variance test for normality (complete samples)." *Biometrika,* vol. 52, pp. 591-611. 1965.

[15] OMI-Polo Español S.A. (OMIE): Market Results. Available: http://www.omie.es/files/flash/ResultadosMercado.swf

Diego J. Bodas-Sagi is an Associate Professor in the Francisco de Vitoria University (Polytechnic School). He has a PhD from the Complutense University of Madrid in computer science. His research interests include Big Data, Data Science, Computational Economics, modelling and e-Health. Contact address is: Universidad Francisco de Vitoria. Carretera Pozuelo a Majadahonda, Km 1.800, 28223 Pozuelo de Alarcón, Madrid (Spain).

José M. Labeaga is Professor of Economics at the Open University in Madrid and Research Affiliated at UNU-MERIT (Maastricht University) and Economics for Energy. He is Ms. and PhD. in Economics by Universitat Autónoma de Barcelona. He has served for the Spanish Government as General Director of the Institute for Fiscal Studies during the period 2008-2012. His main research interests rely on applied microeconometric models, microsimulation and ex-ante evaluation of programs as well as ex-post or impact evaluation of public policies in several fields as health, energy or taxation. Contact address is: Universidad Nacional de Educación a Distancia. Departamento de Análisis Económico II. C/ Senda del Rey, 11 28040, Madrid (Spain).

Real-Time Prediction of Gamers Behavior Using Variable Order Markov and Big Data Technology: A Case of Study

[1]Alejandro Baldominos, [1]Esperanza Albacete, [2]Ignacio Marrero and [1]Yago Saez

[1]Universidad Carlos III, Madrid, [2]Universidad de Granada, Spain

***Abstract* — This paper presents the results and conclusions found when predicting the behavior of gamers in commercial videogames datasets. In particular, it uses Variable-Order Markov (VOM) to build a probabilistic model that is able to use the historic behavior of gamers and to infer what will be their next actions. Being able to predict with accuracy the next user's actions can be of special interest to learn from the behavior of gamers, to make them more engaged and to reduce churn rate. In order to support a big volume and velocity of data, the system is built on top of the Hadoop ecosystem, using HBase for real-time processing; and the prediction tool is provided as a service (SaaS) and accessible through a RESTful API. The prediction system is evaluated using a case of study with two commercial videogames, attaining promising results with high prediction accuracies.**

***Keywords* — User Behavior, Prediction, Variable-Order Markov, Big Data, Real-Time.**

I. Introduction

WHEN gamers are playing, they are performing a sequence of actions which are determined by the game mechanics and vary from game to game. These actions are generally performed across the whole gaming time, making users switch from one game state to another presenting many similarities across different players. In order to incentivize their engagement, it is common to send messages, events or invitations to the users even when they are not playing. These engagement strategies are aimed to keep gamers interested in the game and active.

If the sequence of actions ends and the user does not play the game during a specific period of time (which depends on the game itself) then these gamers will be referred as churners.

The history of actions of a gamer will depend on how he interacts with the game. While the possible ways of interaction with a game may be unlimited, it is expected that some users will behave in a similar way, thus having similar historical actions records and similar game state transitions. Grouping these users would lead to the appearance of gamer profiles, i.e., groups of users sharing similar patterns of gaming behaviors. Having the ability to correctly identify those behavioral patterns will give videogame companies a chance to accurately predict which the most likely next user action is and to act in consequence.

Identifying gaming profiles can help companies to gain better understanding on how users interact with their games, to adapt their products to the customers and to determine the following metrics:

- The customer engagement: what the degree of commitment of gamers playing the game is.
- The churn rate: a measure for gamers that abandon the game.
- The conversion rate: how much revenue is the company able to earn from a certain profile of gamers, such as those who upgrade their license, make purchases, click on advertising, etc.

Understanding these profiles and learning from them comprise the first step for conditioning the behavior of gamers to optimize these metrics, and to ultimately increase the overall performance of the game. To do so, the company can take an active role so that users from a certain profile can move to other profiles providing better metrics or higher revenues.

In this paper we consider a situation where game events occur in sequences. In such cases, the ability of predicting future events based on current events can be valuable, and in consequence this paper presents a novel approach for predicting the behavior of gamers taking advantage of a Variable-Order Markov Model (VOM) which is built and used for training the model and for forecasting future events. The proposed prediction system will be applied to a case of study using data extracted from two commercial datasets.

In practice, this approach is similar to that of "sales funnel" that can be found in sales process engineering, inasmuch in sales funnels the company wants to know what are the specific sequences or processes that lead to a sale or conversion; and in this paper we want to know what will be the following action performed by a gamer (which can be or not a conversion) given the sequence of actions he/she has carried out.

In addition, it should be noticed that while playing, gamers usually perform a sequence of many actions very fast. This introduces two main challenges: in the first place, storing all of these sequences for hundreds of thousands users will result in a very large dataset; and secondly predictions must be provided in real-time in order to become valuable. In order to face these challenges introduced by the high volume and velocity of data, the proposed user behavior prediction system is designed and tested over a big data platform with high scalability in terms of storage and processing power.

The reminder of this paper is organized as follows. First, relevant papers related to this work are presented and discussed, which also use Markov-based techniques in order to face prediction problems in a variety of domains. Later we provide a formal and sound description of the problem to be tackled in this paper, as well as a proposal for solving this problem. After that, experimental results are shown and discussed. Finally, the paper concludes with a discussion of the contribution in this paper, open problems and future lines of work.

II. State of the Art

The high availability of data in recent years, due to new technologies enabling distributed storage and processing of information, has motivated the study of, among many others, the analysis of users' behavior. This analysis allows to better understand the user profiles in an application and how they will respond to different events. In particular, a subset of these analysis techniques has to do with trying to understand how the user will behave in the future, i.e., predicting his behavior.

Prediction of users' behavior is a topic of extensive research in a wide range of domains. The main reason is that this knowledge can be a useful asset when trying to improve user experience, and in some cases

DOI: 10.9781/ijimai.2016.367

can lead to higher revenues. For instance, a recent work from Yuan, Xu, & Wang [27] have presented a framework that studies the relationship between user behavior and sales performance in e-commerce.

Moreover, a very diverse set of techniques have been developed and applied to face the problem of behavior prediction in very different domains. The most extensively reviewed domain is probably online social networks. In this field, Wang & Deng [24] have recently stated that "being able to predict user behavior in online social networks can be helpful for several areas such as viral marketing and advertisement", and have proposed and evaluated a model for performing this task. A survey of techniques for understanding user behavior in online social networks has been published by Jin et al. [13], which also reviews user behavior in mobile social networks. More interestingly, this work reveals the main advantages of studying user behavior in online social networks for different stakeholders: Internet service providers, OSN service provides and OSN users. These advantages can be extrapolated for domains different than online social networks. Finally, a work from Kim & Cho [16] proposed a system for providing recommendations to mobile phone users by predicting their behavior using Dynamic Bayes networks, which can handle time-series data. Finally, in the field of online social games (which are closer to the present work), Zhu et al. [22] have recently studied theoretically the influence of user behavior in order to determine continuance intention, which is an issue related to churn prediction.

Moreover, techniques derived from Markov models have been extensively used in this kind of prediction problems, as it will be reviewed later. These are probabilistic models which, in essence, should fulfil the Markov property which states that the conditional probability of a future event depends only on the present, and not on past events. This property is usually referred as the model being memory-less, as the probability distribution of events in the future is known even when no information about the past is available.

The specific Markov model fulfilling this property, also known as first-order Markov model, only takes into account the present state to predict the future ones. However, this approach may turn out to be insufficient to carry out the task of behavior prediction. This fact has been shown with certain domains, such as web, as the next action cannot be accurately predicted by only considering the last action performed by the user [8].

More complex models arose from this one, which focuses in introducing memory to the original first-order model, thus considering the last k states in order to predict the future ones. This approach is named the kth-order Markov model. These kind of models evolving from the original concept of Markov models has been successfully used in different domains. For instance, one of the earliest applications of Markov models to the prediction of users' behavior in the web is described by Zukerman et al. [28] and later by Deshpande & Karypis [10], which also uses higher-order Markov models. Other, more recent works regarding the application of such models to the same domain include those of Maheswara-Rao & Valli-Kumari [19] and Madhuri et al. [5], where different variable kth-order Markov models are used in order to face the problem of the big amount of data. Moreover, in the works of Awad & Khalil [3] and Vishwakarma et al. [23], so-called all-kth models are used for similar purposes. Additionally, some works can be found in the fields of social networks, such as those by Lerman & Hogg [17] and Hogg et al. [12] where Markov models are used to predict the behavior of users in Digg and Twitter respectively; e-commerce, such as works by Rendle et al. [20] and Wu et al. [25] where Markov chains are used for recommending items to users (the second one using distributed processing to accommodate the presence of big data); education, such as the work of Marques & Belo [18] where Markov chains are applied for discovering user profiles in e-learning sites in order to customize the learning experience; or even in distributed systems, such as the works from Bolivar et al. where matchmakers based on hidden Markov models are proposed predict the availability of grid resources with the ultimate purpose of decentralized grid scheduling [6, 7].

Beyond the scope of web users' behavior prediction, techniques based on Markov models have also been successfully applied to a wide range of different domains. Some recent works involve using variable-order Markov models for predicting the behavior of drivers in vehicular networks, such as in the work from Xue et al. [26]; predicting the future location of mobile users, as in the work from Katsaros & Manolopoulos [15] or applying Markov-based models to the prediction of the inhabitants' activity in smart ubiquitous homes, as proposed by Kang et al. [14] or Alam et al. [1]. Also, an attempt to build a domain-independent predictor of user intention based on Markov model and considering fixed attributes from users is described by Antwarg et al. [2].

In the domain of games, several works have addressed the issue of predicting the gamers' behavior. For instance, Harrison & Roberts [11] describe a data-driven technique for modelling and predicting players' behavior in a game-independent manner. Moreover, some recent works use Markov-based techniques and sequence analysis to attain this objective, such as those published by Shim et al. [21] where the past performance of players is used to build a model from the observed patterns in users behavior to later allow prediction in MMORPGs (with a specific application to World of Warcraft) or by Dereszynski et al. [9] where hidden Markov models are used to learn models from past experience in order to predict the future behavior of players in real-time strategy games (with an application to StarCraft).

III. Problem Definition

Along this paper, an assumption is established expecting that the interaction of a user with the game can be described as a sequence of events (also referred in this paper as actions, as those events are performed by the user). Formally, let sequence $S = \{e_1, e_2, \dots, e_{t-1}, e_t\}$ be the chronologically sorted sequence of all events performed by a certain user in a certain game. This paper seeks a prediction algorithm A, such that when provided with a subsequence S' of S such that $S' = \{e_i, \dots, e_j\}$ for arbitrary values of i and j fulfilling $0 < i \leq j \leq t$, the algorithm outputs an event prediction: $e'_{j+1} = A(S')$.

If the value of e_{j+1} is known in advance (which happens when e_{j+1}), then the accuracy of this algorithm A can be computed and is defined as follows: for a given sequence $S' = (e\_i, \dots, e\_j)$ it can be stated that A is correct on S' if $A(S') = e_{j+1}$, i.e., if the predicted value e'_{j+1} matches the real value e_{j+1}. When working with more than one subsequences of a sequence S, the accuracy is the average fraction of subsequences in S for which the prediction of the algorithm A is correct (note that for computing this value e_t is excluded from S', as otherwise the value of e_{t+1} would be required to be known a priori).

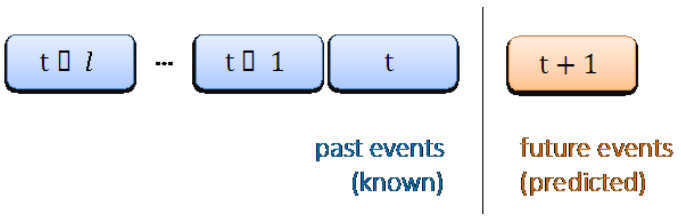


Fig. 1. Conceptual framework for the gaming prediction system. Past behavior events are known, and the system will predict the next event based on the behavior of the community of gamers.

Additionally, A could also be used to predict states of the game under certain circumstances. Let s_t be the state of the game at a time t, where this state is a set of attributes formally defining in an unambiguous way all aspects of the game at a certain point in time. If the state of the game changes only under the gamer's actions, then a new state s_{t+1} can be defined such that $(s_t, e_{t+1}) \rightarrow s_{t+1}$, i.e., the performance of event e_{t+1} under the game state s_t leads to the state s_{t+1}. As far as the game is fully observable (the current state s_t is always known), it is not stochastic, i.e., $\nexists s'_{t+1}$ such that $s_{t+1} \neq s'_{t+1}$ and $(s_t, e_{t+1}) \rightarrow s'_{t+1}$ and the game state is only affected by the gamer's actions (and not by the actions of other gamers or by the system itself), then algorithm A can be used interchangeably to predict either the next action performed by the gamer (e_{t+1}) or the next state of the game (s_{t+1}).

Notice that for games not fulfilling these conditions, then A will not be able to predict the next state of the game, but it will still predict the next user's action, as it depends only on the previous actions and not the states of the game.

This paper seeks an algorithm A providing high prediction accuracy.

IV. Proposal

The system proposed in this paper relies in three different technologies, which are (A) a prediction system for inferring the next action carried out by a gamer given the previous ones; (B) a big data platform for supporting the analysis of huge amounts of information; and (C) a web service for providing the prediction functionality as a service.

This section details how each of these technologies has been applied to the system.

A. Prediction System

The proposed system aims to predict the next action a gamer will carry out given his history of past actions, as depicted in Fig. 1.

In order to provide a prediction, the system will be given a Variable-Order Markov model (VOM) which will store the probabilities that an action occurs just after a particular sequence of actions, and which will be trained from historical data on gamers actions. In order to bound the length of the sequence of actions (as it could be potentially very long), a parameter l is introduced, which determines the maximum number of actions used for training the model, starting from the most recent one and going into the previous ones.

The process for training the model is as follows:

1. The system is introduced a sequence S of length $n+1$ (where $n < t$) containing several actions from a gamer: $S = \{e_{t-n}, \dots, e_{t-1}, e_t\}$.
2. For every different subsequence S' of actions $S' = \{e_i, \dots, e_j\}$, where $i \in [t-n, t-1]$ and $j \in [i, t-1]$ with maximum length l, the conditional probability $p(e_{j+1}|e_j, \dots, e_i)$ is computed.
3. The conditional probabilities computed in step (2) are used to update those already stored in the model. To ease this aggregation, absolute frequencies $n(e_{j+1}|e_j, \dots, e_i)$ may replace probabilities, which would require only additions to update the model.

We will show a very simple example which is not really representative but serves for the purpose of illustrating this process. Let's suppose a gamer performs the next sequence of actions during gameplay in chronological order: *Login*, *StartChallenge*, *KillEnemy*, *KillEnemy*, *KillEnemy*, *CompleteChallenge*. The system is configured to accept a sequence length $l = 2$. As a result, the next frequencies are computed:

$$n(StartChallenge|Login) = 1$$
$$n(KillEnemy|StartChallenge) = 1$$
$$n(KillEnemy|KillEnemy) = 2$$
$$n(CompleteChallenge|KillEnemy) = 1$$
$$n(KillEnemy|StartChallenge, Login) = 1$$
$$n(KillEnemy|KillEnemy, StartChallenge) = 1$$
$$n(KillEnemy|KillEnemy, KillEnemy) = 1$$
$$n(CompleteChallenge|KillEnemy, KillEnemy) = 1$$

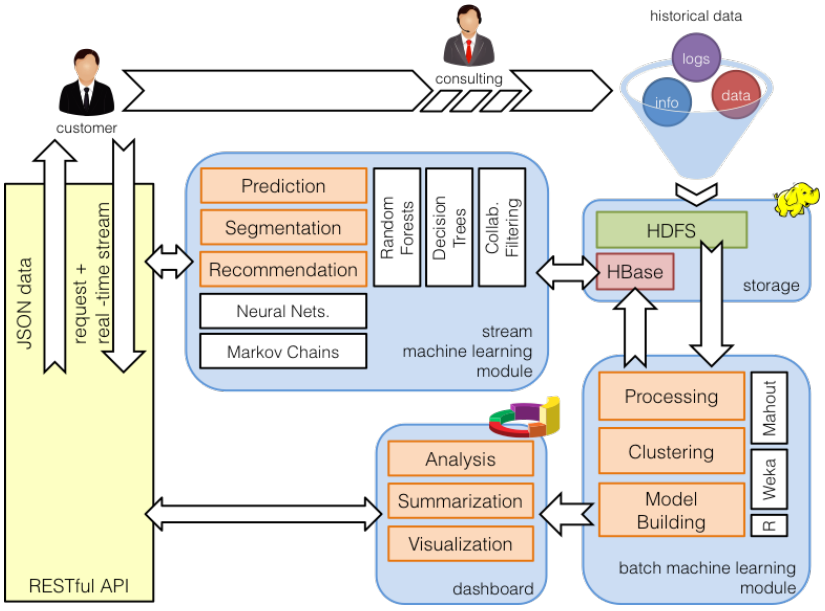


Fig. 2. General framework for online scalable machine learning. This framework comprises different modules which enable storage of Big Data over Apache Hadoop and HBase, as well as batch and streaming machine learning modules; and is operated through a RESTful API.

It is worth noting that, when computing the frequencies, the order of the actions is reversed, as they will be considered in inverse chronological order.

Once the probabilistic model is trained, it can be used for predicting future actions of gamers based on the last $k \leq l$ actions performed by them. In particular, given a known sequence of actions $s = \{e_{t-k}, \dots, e_{t-1}, e_t\}$ the action to be predicted, e_{t+1}, is computed given the next formula, where it should be noted that either relative (f) or absolute (n) frequencies can be used interchangeably:

$$e_{t+1} = \underset{e_{t+1}}{\operatorname{argmax}} f(e_{t+1}|e_t, e_{t-1}, \dots, e_{t-k})$$

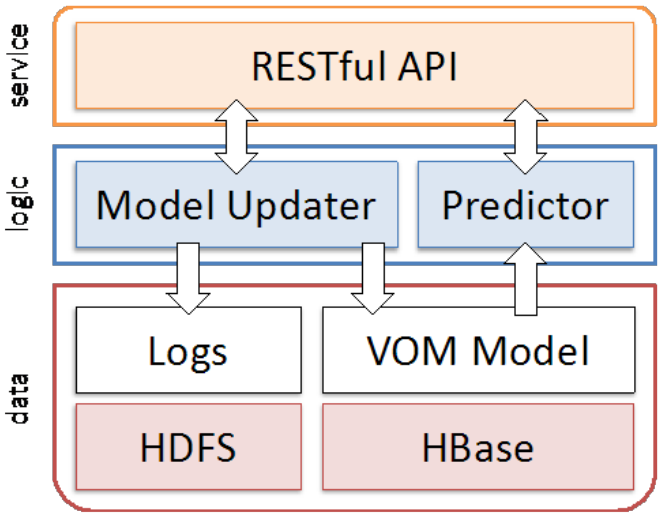


Fig. 3. Architecture of the user behavior prediction system. A web service provides an interface for both updating the probabilistic model stored in HBase (which also stores the incoming data in HDFS) and using the learnt model in order to return a prediction.

It should be noticed that it could be the case where the sequence s is new and, as it never appeared before, the probabilistic model is unaware of it. To provide support for these cases, the previous formula can be generalized as follows:

$$e_{t+1} = \underset{e_{t+1}}{\operatorname{argmax}} f(e_{t+1}|e_t, e_{t-1}, \dots, e_{t-l'})$$

In this formula, $l' \in [1, l]$ and $s' = \{e_{t-l'}, \dots, e_t\}$ is the longest sequence which was already contained in the probabilistic model, i.e., it was also found previously during training. In the literature, this prediction mode is also called prediction by partial match or by longest prefix matching. In the case $s = e_t$ is not contained by the model, then the prediction cannot be performed.

TABLE I
DATA SCHEMA OVER HBASE FOR THE PREDICTION SYSTEM

rk	*cf:predict*	
	$q = e^1_{t+1_1}$	$v: n(e^1_{t+1_1}\|S^1)$
$S^1 = e^1_t, e^1_{t-1}, \dots, e^1_{t-l^1}$	$q = e^1_{t+1_2}$	$v: n(e^1_{t+1_2}\|S^1)$
	…	…
$S^2 = e^2_t, e^2_{t-1}, \dots, e^2_{t-l^2}$	$q = e^2_{t+1_1}$	$v: n(e^2_{t+1_1}\|S^2)$
	…	…
…	…	…

The next example illustrates the prediction process. The system receives the sequence of actions *CompleteChallenge*, *KillEnemy*, which have been carried out by the gamer in chronological order. The system will first try to predict e_{t+1} such that it maximizes the frequency $n(e_{t+1}|KillEnemy, CompleteChallenge)$. However, as that sequence was not already learnt by the model, the system will retry with the longest matching subsequence, which turns out to be *KillEnemy*.

Then, the system will look for the action e_{t+1} that maximizes the frequency $n(e_{t+1}|KillEnemy)$. This action is actually *KillEnemy*, as $n(e_{t+1}|KillEnemy) = 2$ and there are no higher frequencies for this particular sequence. Thus, the system will predict *KillEnemy* as the next action to be performed by the gamer.

This example, however, is extremely simple only serves the purpose of illustrating the proposed algorithm. It can be seen that there are many entries with the same frequency, and that can lead to draws. For instance, if the sequence *KillEnemy, KillEnemy* is considered, then both *KillEnemy* and *CompleteChallenge* are equiprobable predictions. In most cases, when very big datasets are used, draws are very infrequent. However, in the case they do happen and a most likely next event cannot be predicted, then the set of different events is returned.

B. Big Data Technology

When a videogame is popular, it will likely generate huge amounts of data at a high speed, involving the actions performed by its gamers. For this reason, it is important to design the prediction system with scalability in mind, in order to keep processing bounded and to be able to grow in storage capacity and efficiency as it is needed.

To meet these requirements, the prediction system will be built over big data technology. In particular, the Hadoop ecosystem is used following the architecture for big data real-time analysis described in Baldominos et al. [4]. The general framework supporting online scalable machine learning over Big Data is shown in Fig. 2, and the architecture showing how the different components relate to each other is shown in Fig. 3. It can be seen that the web service provides means for both updating the model with new data and providing predictions. The subsystem in charge of updating the model (*model updater* in the figure) also writes the logs to HDFS, so that batch analytics can be performed in later stages.

In order to store the probabilistic model Apache HBase will be used. This tool provides features for storing semi-structured information and is part of the Hadoop ecosystem. Table 1 shows how the model information is structured across the HBase table. As it is shown, the row key (*rk*) comprises the sequence of actions performed by the user, with a maximum length of *l*. There is only one column family (*cf:predict*) which will store the next action to be performed as the qualifier (q) along with the frequency of that action happening after the particular sequence, which is stored as the value (v).

The main advantage of HBase is that it provides indexed row keys, thus enabling efficient query over that field. As the row key in this application stores the subsequences of actions performed by users, we know that the maximum number of queries to be performed for returning a prediction is l, which will typically be a small value. This fact will enable the system to be able to provide real-time predictions.

C. Web Service

The prediction system proposed in this paper is offered as software-as-a-service (SaaS), so that external stakeholders can access it for their own businesses. Moreover, the web service allows selecting one model in each request, so that the system can be reused for different

videogames. The web service is provided through a REST API, which provides the following functionalities:

- Recording a new sequence of actions in the model (`record`), so that the frequencies matrix is updated to incorporate the new sequence. The following parameters must be specified when calling the service:
  - `chainLength`: the maximum chain length (l).
  - `sequence`: the sequence of actions performed by the gamer, in chronological order.
- Predicting the most likely future action (`predict`), for a specified sequence. The next parameter is required for calling the service:
  - `sequence`: the sequence of actions performed by the gamer, in chronological order, with maximum length *l*.
- Returning all the possible next actions along with their probability to occur (`predict_full`), for a specified sequence. This functionality allows the user to get better information about the shortcoming behavior of gamers. The next parameter is required for calling the service:
  - `sequence`: the sequence of actions performed by the gamer, in chronological order, with maximum length *l*.

V. Evaluation

In order to evaluate the proposed prediction system, two different datasets each within the particular domain of social videogames have been considered, as a part of a case of study. The next section provides a description of these datasets. Later, the methodological approach used for conducting the experiments is detailed, and the results obtained are discussed.

A. Datasets Description

Two different commercial games have been used as datasets for evaluating the prediction system. Each of those belong to a different domain, and have a different set of events, so the purpose is to validate that the system is able to generalize well when facing different domains. The commercial names of the games cannot be revealed due to a non-disclosure agreement, but the set of events is described. Although these events could be specified to a higher level of detail, the main purpose for defining these sets was to keep a bounded and reduced number of events while at the same time providing business value from their prediction.

Game #1

It is a social game where users compete to become the best sports director. The set of actions gathered for this game are the next ones:

- `StartSession`, when the user starts a new session.
- `BuyItem`, when the user buys an item in the game.
- `SendNeighbor`, when the user sends an invitation to a friend in the social network to join the game.
- `AcceptNeighbor`, when the user joins the game following an invitation sent by a friend.
- `HelpNeighbor`, when the user helps a neighbor's sport.
- `SendRequest`, when the user sends a social request.
- `AcceptRequest`, when a user accepts a social request sent by a neighbor.
- `UnlockContent`, when the user unlocks content in the game, which can happen after leveling up.
- `OfferAction`, when the user is offered an action by the game.
- `BuyHelp`, when the user buys help in the game (rather than asking neighbors for it).
- `StartTask`, when the user initiates a task associated with a building.
- `TaskCollect`, when the user collects a finished task from a building.
- `UserReferalInfo`, when the user clicks on an advertisement or promotion from a given campaign.

Game #2

It is a social game where users manage an amusement park, and they aim to receive as many visitors as possible. The next actions are logged for this game:

- `StartSession`, when the user starts a new session.
- `BuyItem`, when the user buys an item in the game.
- `SendRequest`, when the user sends a social request.
- `AcceptRequest`, when a user accepts a social request sent by a neighbor.
- `QuestConditionCompleted`, when the user completes a step of a multi-step quest.
- `QuestCompleted`, when the user completes a quest.
- `CollectionCompleted`, when the user completes a collection of items.
- `UserReferalInfo`, when the user clicks on an advertisement or promotion from a given campaign.

B. Experimental Setup

The experiments carried out for the evaluation follow a methodological approach, which is shared between all the datasets used. The purpose is to have the raw data subjected to a preprocessing phase, in order to conform the input format required by the training algorithm to build the probabilistic model. Then, the resulting data is again processed in order to avoid bias in the training or evaluation phases.

For the first preprocessing phase, the raw data is converted to sequences which can be inputted directly to the training algorithm to fed the Variable-Order Markov model. In particular, the original data is contained in the form of logs collecting information about the gamer (including the user identifier), the session (including the session identifier), the action carried out by the gamer, a timestamp and some additional parameters which fall out of the scope of this work. This raw data is converted to a file containing a sequence of actions for each gamer, which can be extracted from the logs just by considering the user ID and the timestamp and which already comply with the format expected by the training algorithm.

After the sequences are produced and before the model is trained, an additional phase takes place in order to guarantee that the evaluation results are not biased. To do so, first the complete set of sequences is randomly shuffled and divided in a training set and a test set, each having 70% and 30% of the original data respectively.

- **No cut**: no cut is performed whatsoever, so the sequence is kept intact and contains all the actions performed by the gamer up from the very beginning to the current moment.
- **Session cut**: the cut is performed randomly, but ensuring that a session is not split, i.e. the cut is performed before a `StartSession` event. The final sequence is kept from the very beginning up to the cut point.
- **Random-fixed cut**: the cut is performed in a random place of the sequence (may break a gaming session) and the final sequence will start at the cut point and have at most length l.

For the experiments, different values of the sequence length (l)

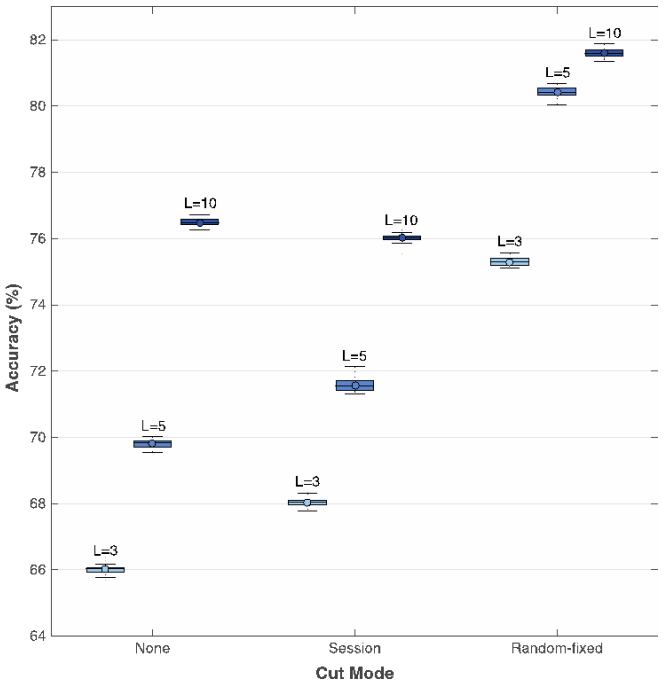


Fig. 4. Boxplot chart for Game #1, showing the distribution of the results of 30 experiments for each cut mode and length value.

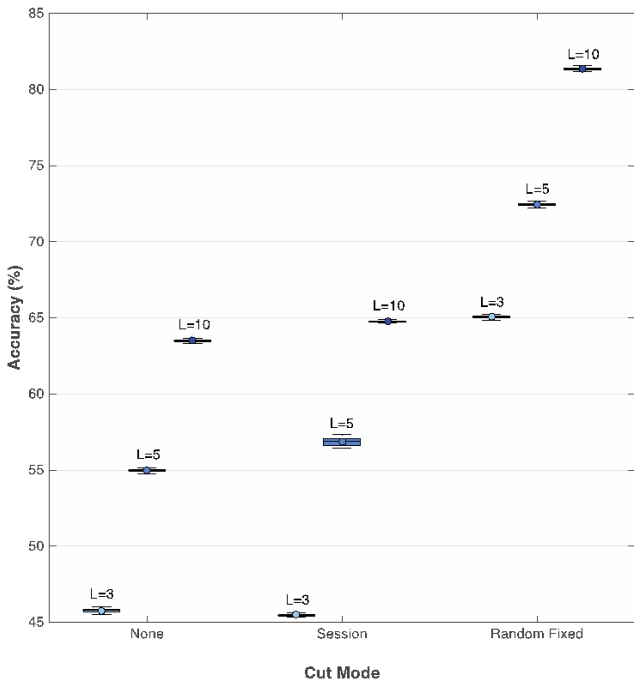


Fig. 5. Boxplot chart for Game #2, showing the distribution of the results of 30 experiments for each cut mode and length value.

TABLE II
PRECISION AND RECALL OBTAINED WHEN PREDICTING EVENTS IN THE GAME #1 DATASET WITH RANDOM-FIXED CUT AND $l = 10$.

Events	***Frequency***	***Precision***	***Recall***
StartSession	02.63%	42.14%	24.49%
BuyItem	06.22%	80.07%	90.75%
SendNeigbor	00.42%	98.60%	98.40%
AcceptNeighbor	00.04%	35.67%	09.75%
HelpNeighbor	01.80%	75.80%	72.19%
SendRequest	06.61%	90.83%	93.01%
AcceptRequest	00.81%	62.41%	24.32%
UnlockContent	02.54%	85.62%	65.00%
OfferAction	00.29%	84.28%	72.96%
BuyHelp	01.64%	69.77%	68.52%
StartTask	40.05%	83.84%	84.80%
TaskCollect	35.79%	63.03%	69.43%
UserReferalInfo	01.17%	64.14%	28.06%

will be tried in order to study how the accuracy changes with the evolution of this parameter, in particular, values of $l \in \{3,5,10\}$ will be tested.

For each particular setup, i.e., each combination of sequence cut mode and length, 30 experiments are executed in order to obtain statistically significant results.

C. Results

The resulting distribution for the prediction system accuracy is described by its boxplot in Fig. 4 for the Game #1 dataset and in Fig. 5 for the Game #2 dataset.

While the results are further discussed in the next section, they clearly show an evidence that for both datasets the same setup achieves a higher accuracy: the one combining a random-fixed cut and a sequence length of $l = 10$. More details about the specific per-event precision and recall, as well as the frequency of occurrence, are shown in table 2 for the Game #1 dataset and in table 3 for the Game #2 dataset.

TABLE III
PRECISION AND RECALL OBTAINED WHEN PREDICTING EVENTS IN THE GAME #2 DATASET WITH RANDOM-FIXED CUT AND $l = 10$.

Events	***Frequency***	***Precision***	***Recall***
StartSession	04.96%	51.96%	15.62%
BuyItem	22.82%	82.89%	84.77%
SendRequest	19.92%	79.38%	65.07%
AcceptRequest	01.40%	50.48%	38.32%
QuestConditionCompleted	37.83%	82.36%	90.07%
QuestCompleted	11.25%	80.60%	88.74%
CollectionCompleted	00.02%	27.14%	02.50%
UserReferalInfo	01.81%	74.19%	42.74%

D. Discussion

As it can be seen in figures 4 and 5, the best results are achieved using $l = 10$ and a random-fixed cut, as this setup outperforms the others for both datasets.

The fact that the accuracy increases with the value of l seems to have a simple explanation: when l is small, then very few past events are considered in order to predict the next ones, and the system will have higher chances to miss the prediction. However, higher values of l does not imply better results in all cases. Moreover, it should be noticed that the time required for training the system grows substantially with the value of l, but the accuracy grows asymptotically.

Regarding the cut mode, results show that higher accuracies are

achieved with the random-fixed cut. In this case, the explanation is not trivial and is domain-dependent. As the datasets used in this paper are social games, it has been observed that the behavior of the users stabilize over time. When users start playing for the first time, they will in most cases perform some random actions in order to get used to the game and explore the scenarios. As time goes by, they stop exploring and start exploiting their resources to achieve higher scores. It should be noticed that the experiments with no cut and with session cut train the probabilistic model using data from the very beginning of the game, while the random-fixed cut may not. This explains why the random-fixed cut achieves higher accuracies, as it has a higher probability of training the model with data extracted when users have already stopped from exploring the game.

Tables 2 and 3 show how precision and recall are compared to the relative frequencies of occurrence of each event in each dataset. With the only exception of the *PayTransaction* event in the Game #1 dataset (an event with an exceptionally small frequency) precision and recall are always higher than the frequency. Also, it can be seen that the system has an outstanding ability to predict some events, such as *SendNeighbor*, even when this event occurs less than 1% of the times. These predictions about future users' behavior could be used to provide a better gaming experience as well as offering players customized offers or experiences to influence their behavior. Moreover, this information can be of great value for the videogame companies, due to the interest of predicting when the users will likely perform some events, such as spending money in in-game purchases (*BuyItem*) or involving other potential users in the game (*SendRequest* and *SendNeighbor* events).

Finally, while it is outside the scope of this evaluation, the time required by the system to provide a prediction is shown in previous work from Baldominos et al. [4]. The experiments were performed in a single-node cluster with 8 Intel Xeon processing cores and 16GB of RAM virtualized over VMWare ESXi 5.0, and Hortonworks HDP 2.1 as the Hadoop distribution, which includes Hadoop 2.4 and HBase 0.98; and JBoss AS 7 as application server. The use of Big Data technology along with a scalable web service allows an average response time of 20.29 ms. when requests are sequential, 98.43 ms. with 10 concurrent requests or 235.04 ms. with 30 concurrent requests. These times are far below one second and enables using this system for real-time prediction as a service. Moreover, the proposed architecture allows to easily scale horizontally just by adding nodes to the cluster, thus being able to perform concurrent reads among these nodes and improving concurrency. Also, the application server could be set in a cluster, and a load balancer could be used for improving performance when many requests are performed at the same time.

VI. Conclusions and Future Work

This paper has presented a novel approach for predicting the future behavior of users of social videogames given their historic behavior. This approach uses variable-order Markov models (VOM) in order to store the frequencies of an event happening after a certain sequence of actions. This model is then able to provide a real time estimation of the next event performed by a gamer.

This prediction system is useful for videogame companies to gain a better understanding on how their users and customers interact with their products, predicting whether certain users will probably make purchases in the shortcoming future or, on the other hand, will abandon the game to become *churners*. This knowledge will enable the company to provide specific experiences and offers customized for each gamer in order to condition his/her behavior; and will provide business insights to discover how their users engage with their products as well as what steps lead them to *convert* into customers.

In a more general way, the system is developed in a general-purpose fashion, so it could be used as a prediction system for any data that can be represented as sequence of events. While this paper has only evaluated the application of this prediction system to the field of videogames, in practice it could be applied to a broad variety of domains, including but not limited to sales processes, web navigation, smartphone usage or even general human behavior analysis.

The system is deployed over Big Data technology, in particular, the Hadoop ecosystem is used in order to scale out horizontally thus being able to tackle the problem of big volume and velocity of data which is inherent to this domain. Data is stored in HBase, thus taking benefit from indexed row keys in order to be able to perform queries on real-time. The prediction system is then provided as a service, which is able to attend 30 concurrent requests in about 200 ms. in a single-node cluster.

In order to evaluate the developed prediction system, experiments have been designed and conducted using two real-world datasets which are in production stages as a part of a case of study. Experiments involved testing the system with different setups of the chain length and the cut mode, this last parameter serving as a way to prevent bias. Results have shown that better results are achieved with a chain length of $l = 10$ and random-fixed cut mode, attaining accuracies of about 82% in both datasets. It makes sense that higher chain lengths lead to better results, as it delves more into the past behavior, being the main disadvantage that the time training the model grows substantially with the value of l. Also, the fact that random-fixed cut provides better results may be explained as the behavior of users in the game stabilize over time, and this cut mode often ignores the behavior at the very beginning of the gaming history of a user.

Per-event results show that precision and recall improve significantly over the frequency of each event, except in very rare cases. This points out that the model is able to learn and extract predictable patterns from data, and thus to provide accurate predictions.

The results from this case of study have not been compared with those resulting for the application of different techniques; as it is the case that other machine learning algorithms are not suitable for providing both training (model update) and prediction in real time. Even when some techniques such as Naive Bayes may be suitable, they must be developed fitting the described Big Data Real-Time architecture, which requires substantial effort. For this reason, this comparative evaluation is left for future work.

Other future lines of work may involve clustering users in order to learn specific models for each cluster or group, so that different types of gamers are detected and prediction models are specialized for each of these. Moreover, it would be interesting in delving in churn prediction, which so far could be supported by the system as long as "*churn*" is defined as an event. Finally, it would be interesting to provide methods not only to predict the next events performed by gamers, but also to condition their behavior so that they will have a higher probability to perform an event which is beneficial to the game company, such as purchasing a certain item. This would require the system to know which are the possible actions to be performed by the game itself in order to provide the more suitable action to achieve a certain behavior in the user, thus providing a more customized gaming experience.

Acknowledgment

This work is part of *Memento Data Analysis* project, co-funded by the Spanish Ministry of Industry, Energy and Tourism with identifier TSI-020601-2012-99 and is supported by the Spanish Ministry of Education, Culture and Sport through FPU fellowship with identifier FPU13/03917.

References

[1] Alam, M. R., Reaz, M. B. I., & Mohd Ali, M. a. (2012). SPEED: An inhabitant activity prediction algorithm for smart homes. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, *42*(4), 985–990.

[2] Antwarg, L., Rokach, L., & Shapira, B. (2012). Attribute-driven hidden markov model trees for intention prediction. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *42*(6), 1103–1119.

[3] Awad, M., & Khalil, I. (2012). Prediction of User's web - browsing behavior : Application of Markov Models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *42*(4), 1131–1142.

[4] Baldominos, A., Albacete, E., Saez, Y., & Isasi, P. (2014). A scalable machine learning online service for big data real-time analysis. *2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD)* (pp. 1–8).

[5] Bindu Madhuri, C., Anand Chandulal, J., Ramya, K., & Phanidra, M. (2011). Analysis of Users' Web Navigation Behavior using GRPA with Variable Length Markov Chains. *International Journal of Data Mining & Knowledge Management Process*, *1*(2), 1–20.

[6] Bolivar, H., Martinez, M., Gonzalez, R., & Sanjuan, O. (2012). A multi-agent matchmaker based on hidden Markov model for decentralized grid scheduling. *Proceedings of the 2012 4th International Conference on Intelligent Networking and Collaborative Systems (INCoS)* (pp. *62*–69).

[7] Bolivar, H., Martinez, M., Gonzalez, R., & Sanjuan, O. (2014). Complexity analysis of a matchmaker based on hidden Markov model for decentralized grid scheduling. *International Journal of Grid and Utility Computing, 5(3)*, 190–197.

[8] Chierichetti, F., Kumar, R., Raghavan, P., & Sarlos, T. (2012). Are web users really Markovian? *Proceedings of the 21st international conference on World Wide Web - WWW '12*, WWW '12 (p. 609). ACM.

[9] Dereszynski, E., Hostetler, J., Fern, A., Dietterich, T., Hoang, T.-T., & Udarbe, M. (2011). Learning Probabilistic Behavior Models in Real-Time Strategy Games. *Proceedings of the Seventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-11)* (pp. 20–25).

[10] Deshpande, M., & Karypis, G. (2004). Selective Markov Models for Predicting Web Page Access. *ACM Transactions on Internet Technology*, *4*(2), 163–184.

[11] Harrison, B., & Roberts, D. L. (2011). Using sequential observations to model and predict player behavior. *Proceedings of the 6th International Conference on Foundations of Digital Games (FDG)* (pp. 91–98).

[12] Hogg, T., Lerman, K., & Smith, L. (2013). Stochastic Models Predict User Behavior in Social Media. *arXiv preprint arXiv:1308.2705*. Retrieved from http://arxiv.org/abs/1308.2705

[13] Jin, L., Chen, Y., Wang, T., Hui, P., & Vasilakos, a V. (2013). Understanding user behavior in online social networks: a survey. *IEEE Communications Magazine*, *51*(9), 144–150.

[14] Kang, W., Shine, D., & Shin, D. (2010). Prediction of state of user's behavior using Hidden Markov Model in ubiquitous home network. *Industrial Engineering and Engineering Management (IEEM), 2010 IEEE International Conference on* (pp. 1752–1756).

[15] Katsaros, D., & Manolopoulos, Y. (2009). Prediction in wireless networks by Markov chains. *IEEE Wireless Communications*, *16*(2), 56–63.

[16] Kim, Y., & Cho, S.-B. (2009). A recommendation agent for mobile phone users using Bayesian behavior prediction. *3rd International Conference on Mobile Ubiquitous Computing, Systems, Services, and Technologies, UBICOMM 2009* (pp. 283–288).

[17] Lerman, K., & Hogg, T. (2012). Using Stochastic Models to Describe and Predict Social Dynamics of Web Users. *ACM Trans. Intell. Syst. Technol.*, *3*(4), 1–33.

[18] Marques, A., & Belo, O. (2010). Discovering student web usage profiles using Markov chains. *9th European Conference on eLearning 2010, ECEL 2010*, *9*(1), 335–342.

[19] Rao, V. V. R. M., & Kumari, V. V. (2011). An Efficient Hybrid Successive Markov Model for Predicting Web User Usage Behavior using Web Usage Mining. *International Journal of Data Engineering*, *1*(5), 43–62.

[20] Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2010). Factorizing personalized Markov chains for next-basket recommendation. *Proceedings of the 19th international conference on World Wide Web (WWW)* (p. 811).

[21] Shim, K. J., Sharan, R., & Srivastava, J. (2010). Player Performance Prediction in Massively Multiplayer Online Role-Playing Games (MMORPGs). *Advances in Knowledge Discovery and Data Mining* (pp. 71–81).

[22] Song-Zhu, D., Jon-Kuo, M., & Hao-Liou, S. (2014). The Study of Continuance Intention for Online Social Games. *Proceedings of the 3rd International Conference on Advanced Applied Informatics* (pp. 230–235).

[23] Vishwakarma, S., Lade, S., Kumar-Suman, M., & Patel, D. (2013). Web User Prediction by Integrating Markov Model with Different Features. *International Journal of Engineering Research and Science & Technology*, *2*(4), 74–83.

[24] Wang, P., & Deng, Q. (2014). User Behavior Prediction: A Combined Model of Topic Level Influence and Contagion Interaction. *Proceedings of the 2014 20th IEEE International Conference on Parallel and Distributed Systems* (pp. 851–852).

[25] Wu, T., He, H., Gu, X., Peng, Y., Zhang, Y., Zhou, Y., & Xu, S. (2013). An intelligent network user behavior analysis system based on collaborative Markov model and distributed data processing. *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 221–228).

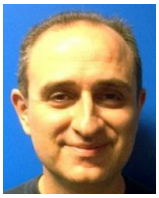
[26] Xue, G., Li, Z., Zhu, H., & Liu, Y. (2009). Traffic-known urban vehicular route prediction based on partial mobility patterns. *Proceedings of the International Conference on Parallel and Distributed Systems (ICPADS)* (pp. 369–375).

[27] Yuan, H., Xu, W., & Wang, M. (2014). Can online user behavior improve the performance of sales prediction in E-commerce? *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2347–2352).

[28] Zukerman, I., Albrecht, D. W., & Nicholson, A. E. (1999). Predicting Users' Requests on the WWW. *Proceedings of the 7th International Conference on User Modeling* (pp. 275–284).

Alejandro Baldominos is Computer Scientist and Engineer since 2012 from Universidad Carlos III de Madrid, and got his Master degree in 2013 from the same university. He is currently working as a researcher in the Evolutionary Computation, Neural Networks and Artificial Intelligence research group (EVANNAI) of the Computer Science Department at Universidad Carlos III de Madrid, where he is currently working in his Ph. D. thesis with a studentship granted by the Spanish Ministry of Education, Culture and Sport. He also works as Professor of the Master in Visual Analytics and Big Data at Universidad Internacional de la Rioja. He has published several conference and journal papers in the fields of context-aware systems, artificial intelligence and big data; and has been involved in several national and European research projects.

Esperanza Albacete received the B.S. in Computer Science from Universidad Carlos III de Madrid in 2009 and the M.Sc. in Computer Science in 2010 in the same university. Currently, she is a Ph. D. student in Computer Science and works as research assistant at the Computer Science Department of Universidad Carlos III de Madrid. Her research interests include Human-Computer Interaction, Ontologies and Advanced Databases Technologies.



Ignacio Marrero is M.Sc. in Astrophysics from Universidad de Granada. He worked in CSIC (most relevant Spanish research center) until 2000, in the areas of experimental data analysis and modelling; and until 2011 worked as a Solution Architect and Manager of Innovation and Research in diverse projects and companies from different sectors. Since 2011 he is working as Project Manager in different companies in the field of Big Data, Business Intelligence, Internet of Things and Advanced Data Analytics. The words "Data Economy & Innovation" clearly summarizes his interests, professional life and overall, his passion; allowing him to be in the front line of Big Data in Spain.

Yago Saez received the degree in computer engineering in 1999. He got his Ph.D. in Computer Science (Software Engineering) from the Universidad Politécnica de Madrid, Spain, in 2005. Since 2007 is vice-head of the Computer Science Department from the Carlos III University of Madrid, where he got a tenure and is associate professor. He belongs to the Evolutionary Computation, Neural Networks and Artificial Intelligence research group (EVANNAI) and member of the IEEE Computational Finance and Economics Technical committee. Nowadays is involved in a European Project in collaboration with public and private companies and he is affiliated with Auctionomics (a US-based World leader auction advisory company). His main research areas encompass the evolutionary computation techniques applied to Computational Economic and Finance and Computational Intelligence to Games fields.

Detection of Adverse Reaction to Drugs in Elderly Patients through Predictive Modeling

Rafael San Miguel Carrasco

Researcher, Universidad Internacional de La Rioja

***Abstract* — Geriatrics Medicine constitutes a clinical research field in which data analytics, particularly predictive modeling, can deliver compelling, reliable and long-lasting benefits, as well as non-intuitive clinical insights and net new knowledge. The research work described in this paper leverages predictive modeling to uncover new insights related to adverse reaction to drugs in elderly patients. The differentiation factor that sets this research exercise apart from traditional clinical research is the fact that it was not designed by formulating a particular hypothesis to be validated. Instead, it was data-centric, with data being mined to discover relationships or correlations among variables. Regression techniques were systematically applied to data through multiple iterations and under different configurations. The obtained results after the process was completed are explained and discussed next.**

Keywords — Geriatrics, Medicine, Data Analytics, Statistical Analysis, Predictive Modeling, Knowledge Management, Adverse reactions, Drugs.

I. Introduction

THE availability of big data and data analytics technologies and data analytics tools in the healthcare sector has not been seen as an advantage until recent times.

Today, managers of healthcare providers notice that data analytics can bring improvements in a broad range of business processes, and can also radically increase the effectiveness of service delivered to patients, while allowing for on-the-go research that can produce net new knowledge not intuitively or easily acquired by traditional research.

This research exercise aimed to generate a predictive model to accurately anticipate occurrence of such a relevant event as mortality (Exitus), among elderly patients admitted in a Geriatric Acute Unit.

While big data is typically associated with a high volume of data, variety (amount of data sources involved in an analysis) and velocity (speed at which data is generated) are also common big data features. This research exercise took advantage of using a dataset including variables from multiple data sources, as opposed to typical single-source, ad-hoc approaches often seen in traditional clinical research.

For this purpose, anonymized clinical records were mined with data analytics software. These records contained details about patients demographics, diagnosis, treatment, physical disability, mental disability, blood tests, admission-related complications and administered drugs. Patients' physical and mental disability were measured using CRF and CRM scales, respectively. These scales have been developed by Hospital Central de la Cruz Roja.

II. State of the Art

Previous research performed by professionals in the fields of interest was reviewed prior to starting the project.

This information allowed to gain an understanding of what other researchers discovered in the past, or are currently investigating on, as a useful reference of how this research must be approached.

Data used in this research work included information on drugs administered to patients. Therefore, it became relevant to understand what knowledge was available about adverse interactions between drugs and clinical variables like mortality, LOS (Length of Stay) and, at a higher level, cost associated with healthcare services delivery.

As Grizzle, F. R. [1] points out, average cost of an error in drugs administration is $977. The total cost is $177,5 billion, of which 70% represents the cost of patients' admissions resulting from these errors.

Matthew G. Whitbeck, R. J. [2] demonstrated that patients with atrial fibrillation suffer from multiple adverse reactions to Digoxin, including a higher mortality rate. The same conclusion is reached by Mate Vamos, J. W. [3], which explains that this adverse interaction is independent from other factors as kidney function, cardiovascular comorbidity or adherence to medications.

Also, Mate Vamos, J. W. [4] confirms that this circumstance is not limited to patients with AF (Atrial Fibrillation), but can also be applied to patients with CHF (Congestive Heart Failure), and suggest that this drug must be used with caution.

Finally, Wooten, J. M. [5] concludes that drug-administration errors have a higher impact on elderly patients. It also states that avoiding polypharmacy, rigorous analysis of drug interactions and frequent monitoring of patients' adverse reactions to drugs can dramatically lower risk.

III. Goals

The goal of this research work was developing a use case in which medical knowledge is extracted from a clinical dataset without a prior hypothesis.

In a broader context, this research work attempts to provide a practical example to shed additional light on finding an answer to the following questions:

1. Can data analytics support traditional clinical research in generating valuable medical knowledge while being less dependent upon intuition?
2. Can data analytics improve current clinical research's efficiency by providing additional insights and shortening deadlines?

IV. Methodology

A. Data sources

Clinical records of patients' admissions to Geriatrics Acute Unit from Jan 1, 2006 to Jul 31, 2015 (N=11.795) were extracted from a clinical dataset in Microsoft Access format. These observations were anonymized and saved in an appropriate format to be used in SAS.

DOI: 10.9781/ijimai.2016.368

In order to make the planned analysis affordable through standard computing resources, the dataset was filtered to extract patients admitted from nursing homes in the first half of 2015 (N=138). Having said that, this use case can scale up to millions of records with no changes in implementation.

The resulting dataset was then combined with data extracted from additional Clinical Information Systems to obtain further clinical variables about each patient. These variables related to drugs-administration, medical tests and consults, and visits to emergency units.

The final number of variables considered was 81 (p=81). The set of observations with inputs from multiple data sources constituted a data lake in which multiple queries can be run.

For the sake of simplicity, this article focuses on insights related to mortality, which is a key clinical variable. However, the same data lake can be used with no changes for any other analysis related to these patients' admissions.

B. Data preparation

Several preparation routines were run against the dataset to facilitate agile and effective mining activities once loaded into SAS.

Particularly:

1. Missing values. The following default values were assigned to empty cells: "Missing", for categorical (discrete) variables, and empty string ("") for numeric variables.
2. Deletion of records. Applied to those records where missing values occurred in key fields for the analysis.
3. Review of minimum and maximum values to detect outliers or erroneous values. These were replaced by the average, minimum or maximum value in the field, depending on each particular case.
4. Removal of irrelevant, unused or redundant variables.
5. Transformation of variables. Admission and discharge dates were replaced by length of stay, and birthdate was converted to age.
6. Replacement of numeric codes with meaningful strings, to allow for faster interpretation of results.
7. Replacement of strings with numeric codes, to fine-tune input variables before executing regression techniques.

C. Data analysis

The following statistical analysis and modeling techniques were used:

1. Calculation of descriptive indicators, to understand each field's structure.
2. Transformation of variables, to increase the degree of linear correlation between available inputs and the target variable.
3. Variable selection, to rapidly discard those inputs that show low predictive capabilities.
4. Logistic regression

D. Toolset

SAS Enteprise Miner[1] was used for this research work. The suggested methodology to perform logistic regression analysis described by Sharma, K. S. [6] was implemented.

E. Limitations

The methodology, tools and data used in this research work is subject to several limitations that are described next. This information will help the reader assess whether obtained results are reliable enough for a particular scenario.

Data sampling

The original dataset was filtered to obtain those patients having been admitted from a set of nursing homes in the first half of 2015. Typically, one year is a more appropriate period for inference techniques to be reliable.

In addition to this, the dataset was subject to bias, given that all patients records belonged to a single hospital. Ideally, these records must have been obtained from multiple hospitals.

Data quality

Data was gathered by healthcare professionals, and input into well-designed clinical Information Systems implementing measures to avoid input errors.

However, the risk of having erroneous data is not fully mitigated. Also, certain variables' values are influenced by the subjective perception of the doctor or nurse.

Accuracy of results

Software used is enterprise-class and commonly used in scientific studies and research. Furthermore, criteria applied to assess statistical significance was based on generally accepted practices.

However, this doesn't imply that they are suitable for other scenarios or use cases beyond the context of this research work.

Seasonality

Selected records covered a period of six months. Therefore, seasonality factors couldn't be accounted for. This might result in biased values. However, the resulting deviation won't likely impact the final results that were obtained.

Geographical factors

As stated previously, the source of data was one hospital in Madrid (Spain).

Therefore, conclusions might not be applicable to other geographies. However, this source of bias is common to most clinical trials.

Methodological errors

Data mining procedures used along this research constitute industry best practices. Nevertheless, other context-related factors might not have been taken into account.

Other limitations

No additional limitations were identified.

In addition, conflicts of interest were not found to apply to the author of this work or any of his collaborators. None of the participants will personally benefit from obtained results.

V. Detailed Procedure

A. Background

Patients mortality is a key clinical variable.

Datasets were mined to discover what variables could accurately predict mortality (Exitus) on a given set of patients.

B. Methodology

The following diagram was designed and run in SAS Enterprise Miner to build the model:

The regression node was configured as follows:

- Two factors interaction: No.
- Polynomials terms: No.

1 http://www.sas.com/en_us/software/analytics/enterprise-miner.html

- Regression type: Logistic regression.
- Link function: Logit.
- Model selection: Stepwise.
- Selection criteria: Validation error.
- Optimization technique: Default.

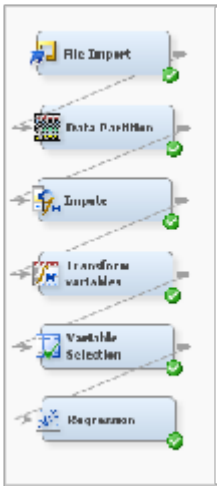


Fig. 1. SAS diagram.

Model fit indicators and relative risk (odds-ratio) values were obtained and analyzed in order to assess reliability and predictive capabilities of the model.

C. Obtained results

The process to build the model was split in several iterations.

In the first iteration, a single variable was found to predict mortality with 100% of accuracy: Place of Exitus. It became obvious that this variable had to be removed from the model.

In the next iteration, Morphine was found to accurately predict mortality. However, since this drug is typically administered to patients when they are about to pass away, the resulting model would offer no predictive capabilities to a doctor. As such, this drug was also removed from the model.

In the third iteration, however, a model containing several meaningful variables was obtained.

These were the following:

1. Digoxin, a drug that has been proved to be associated with higher mortality rates for other populations in previous clinical trials.
2. Number of lab tests requested by the doctor during the admission process.
3. Occurrence of pressure ulcers.

The model was assessed to check whether it was reliable and accurate from a statistical perspective.

Most relevant model fit indicators displayed by SAS are shown in Fig. 2 and discussed next:

- **Global Null Hypothesis**, that tests whether all coefficients in the regression model are zero, was rejected (p-value < 0.0001).
- **Type 3 Analysis of Effects**, that tests whether each individual predictor's coefficient in the model is zero, shows that three of those variables (I_LAB_Number, I_Pharma_Digoxine, TI_GN_Evaluation_Ulceras_p3) exhibit non-zero coefficients when entered in the model (p-values 0.004, 0.0207, 0.0048, respectively).
- **Odds Ratio Estimates**, which are the proportions of observations in the main group (mortality) compared to the control group (no mortality) with respect to each predictor in the model, confirm that the three previous predictors influence patients' mortality (with odds ratio estimates of 1.273, 3.953, 6.280, respectively).

D. Conclusions

The built model turned out to be statistically significant and accurate. Therefore, it would be ready to be implemented in a production environment to predict mortality for a given set of patients.

Furthermore, the confusion matrix depicted below confirms that

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
100.087	63.714	36.3732	3	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
I_LAB_number	1	8.1078	0.0044
I_Pharma_DIGOXINA	1	5.3495	0.0207
TI_G_N_Evaluation_Ulceras_p3	1	7.9718	0.0048

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-2.0981	0.6488	10.46	0.0012		0.123
I_LAB_number	1	0.2414	0.0848	8.11	0.0044	0.7080	1.273
I_Pharma_DIGOXINA	1	1.3746	0.5943	5.35	0.0207	0.5818	3.953
TI_G_N_Evaluation_Ulceras_p3 0	1	0.9187	0.3254	7.97	0.0048		2.506

Odds Ratio Estimates

Effect	Point Estimate
I_LAB_number	1.273
I_Pharma_DIGOXINA	3.953
TI_G_N_Evaluation_Ulceras_p3 0 vs 1	6.280

Fig. 2. Model fit indicators.

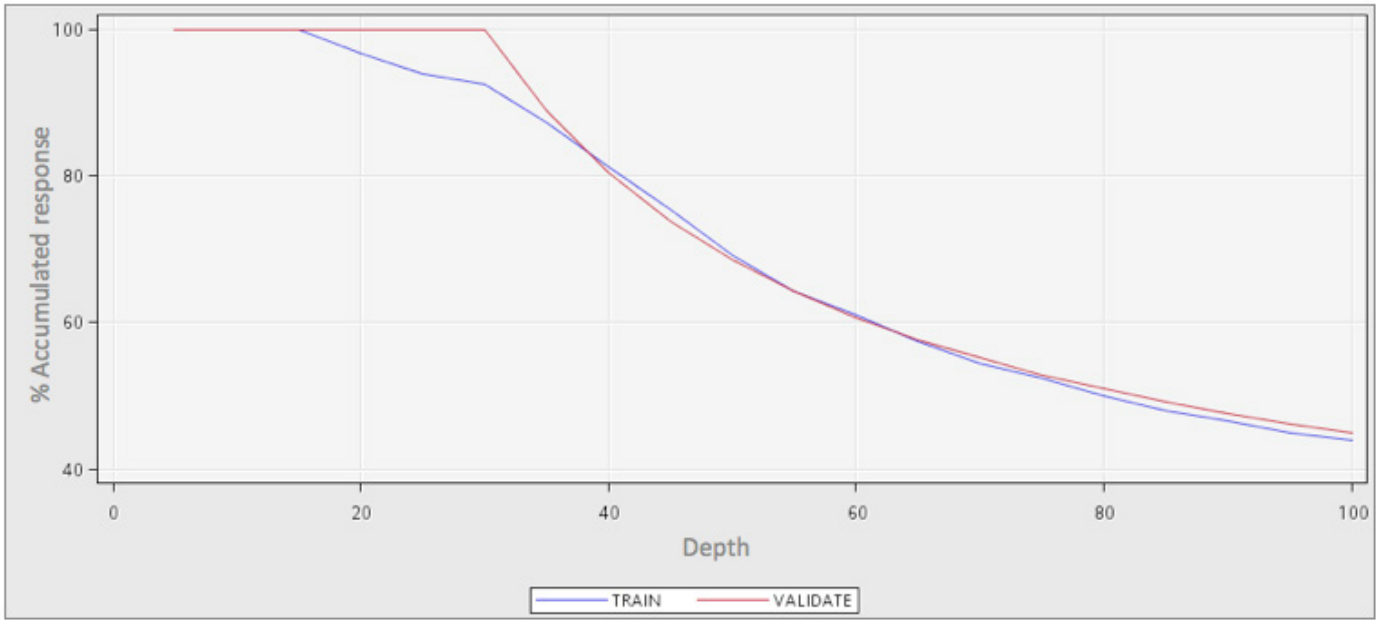


Fig. 3. Percentage of captured response for training and validation data.

the model performed quite well against the train dataset, with no overfitting signals:

Event classification table

Data role=TRAIN Target=B_Exitus_Horus Target label=B_Exitus Horus

False negative	*True negative*	*False positive*	*True positive*
8	*35*	*6*	*24*

Out of 30 cases of mortality in our training data, the model was able to predict 24 of them (80%).

VI. Obtained Results

Digoxin, a regular drug occasionally used in the treatment of various heart conditions, was proved to be linked to patients' mortality. This connection had been demonstrated for other populations but never tested with elderly patients.

Also, other admission-related factors turned out to contribute to increase the likelihood of an elderly patient passing away.

These circumstances were discovered without an initial hypothesis about how available input variables (including drugs administration) could be related to the target variable (mortality).

Beyond actual results, it is also proven that data analytics can uncover non-intuitive or complex relationships in a far more efficient way, with no prior assumptions required to trigger a research exercise, and with multiple potential results brought along the discovery process.

VII. Discussion

Operationalizing a predictive model that is able to anticipate which patients are at a higher risk of mortality allows doctors to adjust their treatments on time and provide further attention to their evolution overtime, hence lowering the mortality rate.

This outcome from a research exercise is highly desirable and fulfills the purpose of enhancing delivery processes in the healthcare environment.

However, one can argue that this result had been proven in a different population by previous studies. What makes this exercise different is the approach taken to get to results: no prior questions were asked and no hypothesis was formulated.

The approach was completely agnostic. As such, the goal was not to uncover an adverse reaction to drugs or to what extent pressure ulcers increases the mortality likelihood. The main purpose was to uncover hidden relationships among clinical variables having an impact on mortality.

VIII. Conclusion

The ultimate goal of this research exercise was to provide a real-life use case in which data analytics added value to traditional clinical research. This added value would translate into insights that were not part of an initial hypothesis, but rather discovered on-the-go while crunching available data.

A doctor's extensive medical knowledge is still limited by his/her professional experience and subjective interpretation of available details. This use case attempts to go beyond this limitation by analyzing data from an agnostic point of view, leveraging all available variables and avoiding prior assumptions that could limit potential results.

Using a data-centric strategy to analyze data, as opposed to generating a custom dataset and focusing on a particular goal, might increase the amount of conclusions derived from the research process and uncover unexpected insights that would have not become a priority otherwise.

The example provided in this article has focused on trying to demonstrate how such data-centric approach would work on a very limited and simplified scenario. Therefore, further research would be highly recommended to gradually prove how far data analytics' contribution could potentially be.

IX. Future Research Paths

Once confirmed that it's feasible to build reliable statistical models to detect adverse reactions to certain drugs and predict mortality, it would be highly suggested to perform additional research in order to go beyond these initial results.

In this research exercise, a limited sample of observations was used. By processing all available patients' data, which can scale up to millions of clinical records in large hospitals, the amount, quality and accuracy of obtained insights would likely be much higher than what was obtained here.

Likewise, in order to realize the benefits of these additional insights, they must be operationalized, that is, made available to doctors through a production-ready Information System implementing a Recommendation Engine (RE).

Lastly, given the inherent computational complexity of such an implementation, RE's performance metrics must be taken into account in the process, as pointed out by Luis F. Chiroque [7].

Acknowledgment

This research was possible thanks to PhD Javier Gómez Pavón, senior doctor, as well as expert and researcher in Geriatrics Medicine.

PhD Beatriz Ares Castro-Conde, senior doctor, also made a contribution to this research.

References

[1] Grizzle, F. R. "Drug-Related Morbidity and Mortality: Updating the Cost-of-Illness Model", J Am Pharm Assoc (Wash). 2001 Mar-Apr;41(2):192-9, 2001.

[2] Matthew G. Whitbeck, R. J. "Increased mortality among patients taking digoxin—analysis from the AFFIRM study", European Heart Journal (2013) 34, 1481–1488 doi:10.1093/eurheartj/ehs348, pp. 1-8, 2013.

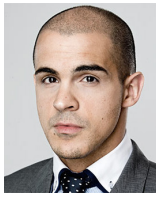
[3] Mate Vamos, J. W. "Increased Mortality Associated With Digoxin in Contemporary Patients With Atrial Fibrillation", Journal of the American College of Cardiology, pp. 1-8, 2014.

[4] Mate Vamos, J. W. "Digoxin-associated mortality: a systematic review and meta-analysis of the literature", European Heart Journal doi:10.1093/eurheartj/ehv143, pp. 2-7, 2015.

[5] Wooten, J. M. "Pharmacotherapy Considerations in Elderly Adults", South Med J. 2012;105(8):437-445 doi: 10.1097/SMJ.0b013e31825fed90, pp. 2-7, 2012.

[6] Sarma, K. S. "Predictive Analytics with SAS Enterprise Miner", SAS, pp. 359-371, 2013.

[7] Luis F. Chiroque. "Empirical Comparison of Graph-based Recommendation Engines for an Apps Ecosystem", IJIMAI , DOI: 10.9781/ijimai.2015.327, pp. 35-36, 2015.



Rafael San Miguel Carrasco has developed his professional career in the Technology industry for the past eleven years. He has taken on roles in the field of research, technology, project management, team leading, business development and middle management. He has worked for multinational firms as Deloitte, Telefónica, Santander and FireEye, engaging on and leading international business initiatives combining technology, management and operations. Rafael works as a Data Scientist at Universidad Internacional de la Rioja (UNIR) in a research initiative in the field of Healthcare Analytics, where big data technologies as SAS, R and Hadoop play a key role.

Text Analytics: the convergence of Big Data and Artificial Intelligence

Antonio Moreno[1], Teófilo Redondo[2]

[1]Universidad Autónoma de Madrid and Instituto de Ingeniería del Conocimiento, Madrid, Spain
[2]ZED Worldwide, Madrid, Spain

***Abstract* — The analysis of the text content in emails, blogs, tweets, forums and other forms of textual communication constitutes what we call text analytics. Text analytics is applicable to most industries: it can help analyze millions of emails; you can analyze customers' comments and questions in forums; you can perform sentiment analysis using text analytics by measuring positive or negative perceptions of a company, brand, or product. Text Analytics has also been called text mining, and is a subcategory of the Natural Language Processing (NLP) field, which is one of the founding branches of Artificial Intelligence, back in the 1950s, when an interest in understanding text originally developed. Currently Text Analytics is often considered as the next step in Big Data analysis. Text Analytics has a number of subdivisions: Information Extraction, Named Entity Recognition, Semantic Web annotated domain's representation, and many more. Several techniques are currently used and some of them have gained a lot of attention, such as Machine Learning, to show a semisupervised enhancement of systems, but they also present a number of limitations which make them not always the only or the best choice. We conclude with current and near future applications of Text Analytics.**

***Keywords* — Big Data Analysis, Information Extraction, Text Analytics**

I. Introduction

Natural Language Processing (NLP) is the practical field of Computational Linguistics, although some authors use the terms almost interchangeably. Sometimes NLP has been considered a subdiscipline of Artificial Intelligence, and more recently it sits at the core of Cognitive Computing, since most cognitive processes are either understood or generated as natural language utterances.

NLP is a very broad topic, and includes a huge amount of subdivisions: Natural Language Understanding, Natural Language Generation, Knowledge Base building, Dialogue Management Systems (and Intelligent Tutor Systems in academic learning systems), Speech Processing, Data Mining – Text Mining – Text Analytics, and so on. We will focus here in this specific article in Text Analytics (TA).

Text Analytics is the most recent name given to Natural Language Understanding, Data and Text Mining. In the last few years a new name has gained popularity, Big Data, to refer mainly to unstructured text (or other information sources), more often in the commercial rather than the academic area, probably because unstructured free text accounts for 80% in a business context, including tweets, blogs, wikis and surveys [1]. In fact there is a lack of academic papers covering this topic, although this may be changing in the near future.

Text Analytics has become an important research area. Text Analytics is the discovery of new, previously unknown information, by automatically extracting information from different written resources.

II. Text Analytics: Concepts and Techniques

Text Analytics is an extension of data mining, that tries to find textual patterns from large non-structured sources, as opposed to data stored in relational databases. Text Analytics, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting non-trivial information and knowledge from unstructured text. Text Analytics is similar to data mining, except that data mining tools are designed to handle structured data from databases, either stored as such or as a result from preprocessing unstructured data. Text Analytics can cover unstructured or semi-structured data sets such as emails, full-text documents and HTML files, blogs, newspaper articles, academic papers, etc. Text Analytics is an interdisciplinary field which draws on information extraction, data mining, machine learning, statistics and computational linguistics.

Text Analytics is gaining prominence in many industries, from marketing to finance, because the process of extracting and analysing large quantities of text can help decision-makers to understand market dynamics, predict outcomes and trends, detect fraud and manage risk.

The multidisciplinary nature of Text Analytics is key to understand the complex integration of different expertise: computer engineers, linguists, experts in Law, BioMedicine or Finance, data scientists, psychologists, causing that the research and development approach is fragmented due to different traditions, methodologies and interests.

A typical text analytics application consists of the following steps and tasks:

Starting with a collection of documents, a text mining tool retrieves a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. The underlying strategy in all the components is to find a pattern (from either a list or a previous process) which matches a rule, and then to apply the rule which annotates the text. Each component performs a particular process on the text, such as: *sentence segmentation* (dividing text into sentences); *tokenization* (words identified by spaces between them); *part-of-speech tagging* (noun, verb, adjective, etc., determined by look-up and relationships among words); *shallow syntactic parsing/ chunking* (dividing the text by noun phrase, verb phrase, subordinate clause, etc.); *named entity recognition (NER)* (the entities in the text such as organizations, people, and places); *dependency analysis* (subordinate clauses, pronominal anaphora [i.e., identifying what a pronoun refers to], etc.).

The resulting process provides "structured" or semi-structured information to be further used (e.g. Knowledge Base building, Ontology enrichment, Machine Learning algorithm validation, Query Indexes for Question & Answer systems).

Some of the techniques that have been developed and can be used in the text mining process are information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, question answering, and deep learning.

DOI: 10.9781/ijimai.2016.369

A. Information Extraction

Information extraction (IE) software identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process usually called pattern matching, typically based on regular expressions. The most popular form of IE is named entity recognition (NER). NER seeks to locate and classify atomic elements in text into predefined categories (usually matching preestablished ontologies). NER techniques extract features such as the names of persons, organizations, locations, temporal or spatial expressions, quantities, monetary values, stock values, percentages, gene or protein names, etc. These are several tools relevant for this task: Apache OpenNLP [2], Stanford Named Entity Recognizer [3] [4], LingPipe [5].

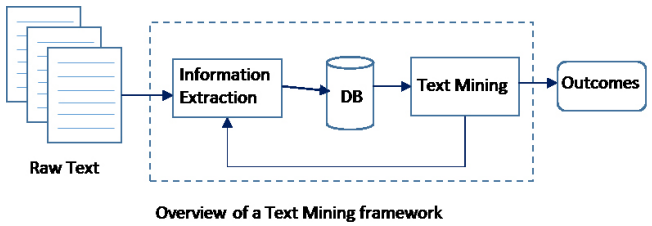


Fig. 1. Overview of a Text Mining Framework

B. Topic Tracking and Detection)

Keywords are a set of significant words in an article that gives a high-level description of its contents to readers. Identifying keywords from a large amount of online news data is very useful in that it can produce a short summary of news articles. As online text documents rapidly increase in size with the growth of WWW, keyword extraction [6] has become the basis of several text mining applications such as search engines, text categorization, summarization, and topic detection. Manual keyword extraction is an extremely difficult and time consuming task; in fact, it is almost impossible to extract keywords manually in case of news articles published in a single day due to their volume.

A topic tracking system works by keeping user profiles and, based on the documents the user views, predicts other documents of interest to the user. Google offers a free topic tracking tool [7] that allows users to choose keywords and notifies them when news relating to those topics becomes available. NER techniques are also used in enhancing topic tracking and detection by matching names, locations or usual terms in a given topic by representing similarities with other documents of similar content [8]. Topic detection is closely related with Classification (see below).

C. Summarization

Text summarization has a long and fruitful tradition in the field of Text Analytics. In a sense text summarization falls also under the category of Natural Language Generation. It helps in figuring out whether or not a lengthy document meets the user's needs and is worth reading for further information. With large texts, text summarization processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning.

One of the strategies most widely used by text summarization tools is sentence extraction. Important sentences from an article are statistically weighted and ranked. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document.

The methods of summarization can be classified in two broad groups:

- shallow analysis, restricted to the syntactic level of representation and try to extract significant parts of the text;
- deeper analysis, assumes a semantics level of representation of the original text (typically using Information Retrieval techniques).

A relatively recent European Union project, ATLAS, has performed an extensive evaluation of text summarization tools [9].

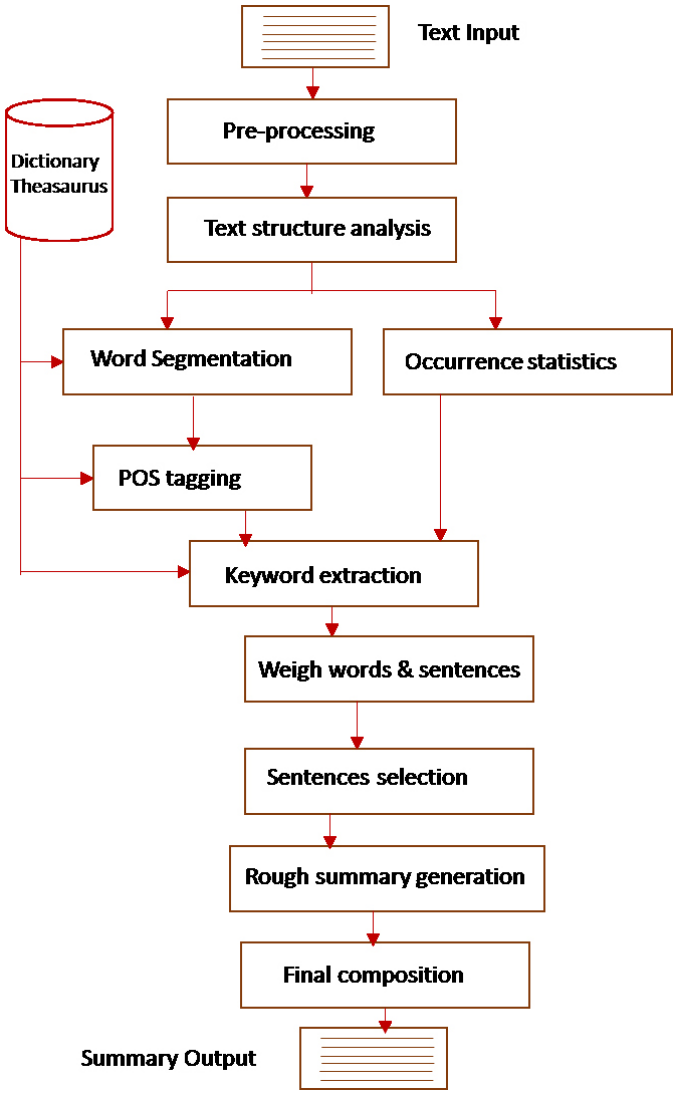


Fig. 2. Text Summarization

A. Categorization or Classification

Categorization involves identifying the main themes of a document by placing the document into a predefined set of topics (either as taxonomies or ontologies). Categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on relationships identified by looking for broad terms, narrower terms, synonyms, and related terms. Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic [10]. Another method is to represent topics as thematic graphs, and using a degree of similarity (or distance from the "reference" graph) to classify documents under a given category [11].

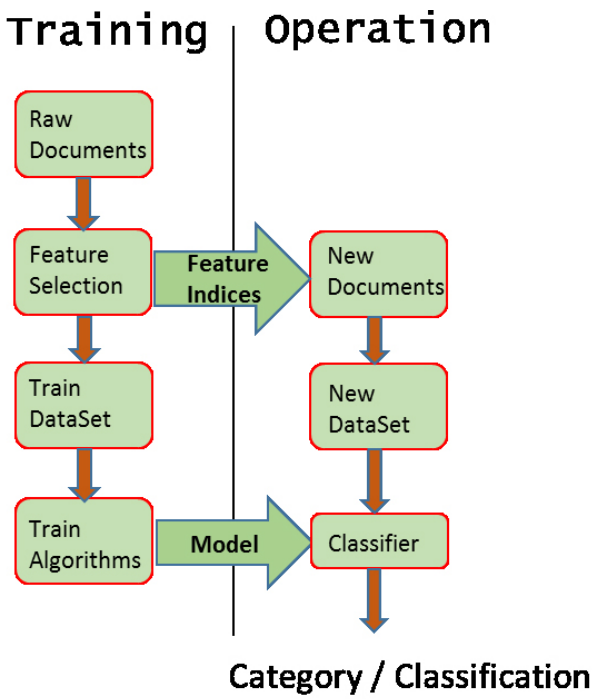


Fig. 3. Text Classification

D. Clustering

Clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered without the use of predefined topics. In other words, while categorization implies supervised (machine) learning in the sense that previous knowledge is used to assign a given document to a given category, clustering is unsupervised learning: there are no previously defined topics or categories. Using clustering, documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results (multiple indexing references). A basic clustering algorithm creates a vector of topics for each document and assigns the document to a given topic cluster.

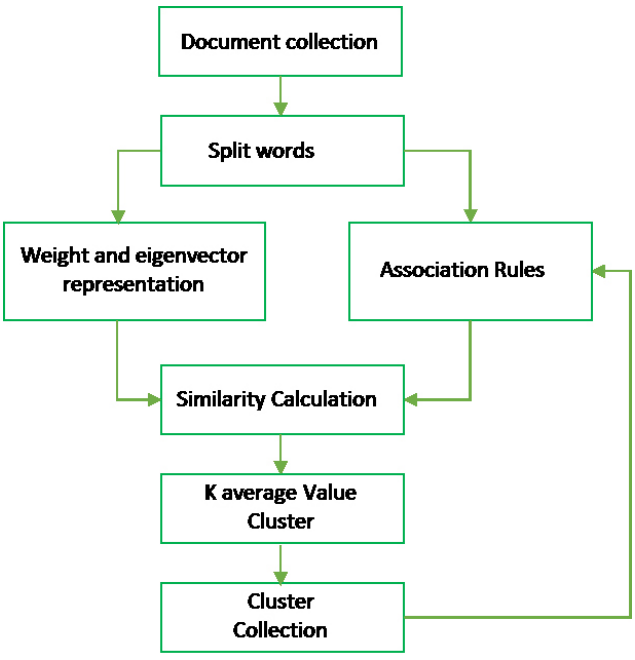


Fig. 4. Document Clustering

Medicine and Legal research papers have been a fertile ground to apply text clustering techniques [12] [13].

E. Concept Linkage

Concept linkage tools connect related documents by identifying their commonly-shared concepts and help users find information that they perhaps would not have found using traditional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable concept in text mining, especially in the biomedical and legal fields where so much research has been done that it is impossible for researchers to read all the material and make associations to other research.

The best known concept linkage tool is C-Link [14] [15]. C-Link is a search tool for finding related and possibly unknown concepts that lie on a path between two known concepts. The tool searches semi-structured information in knowledge repositories based on finding previously unknown concepts that lie between other concepts.

F. Information Visualization

Visual text mining, or information visualization, puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching. Information visualization is useful when a user needs to narrow down a broad range of documents and explore related topics. A common typical example of text information visualization are Tag clouds [16], like those provided by tools such as Wordle [17]. Hearst [18] has written an extensive overview of current (and recent past) tools for text mining visualization, but definitively needs an update with the appearance of new tools in recent years: D3.js [19], Gephi [20], as well as assorted JavaScript-based libraries (Raphäel, jQuery Visualize).

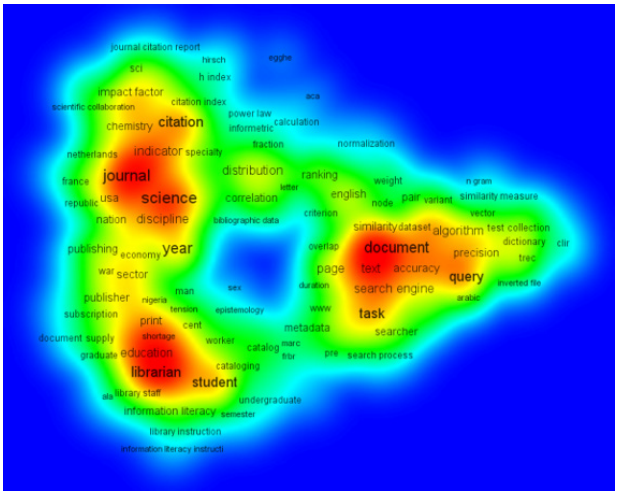


Fig. 5. Text data visualization
(Source: Hearst (2009) “Information Visualization for Text Analysis”)

G. Question Answering

Question answering (Q&A) systems used natural language queries to find the best answer to a given question. Question answering involves a lot of techniques described here, from information extraction for the question topic understanding, question typology and categorization, up to the actual selection and generation of the answer [21] (Hirschman & Gaizauskas, 2001). OpenEphyra [22] was an open-source question answering system, originally derived from Ephyra, which was developed by Nico Schlaefer and has participated in the TREC question answering competition [23]. Unfortunately, OpenEphyra has been discontinued and some alternatives have appeared, such as YodaQA [24], which is general purpose QA system, that is, an “open domain”

general factoid question answering [25].

H. Deep Learning

Deep Learning has been gaining a lot of popularity as of the last two years, and has begun to be experimented for some NLP tasks. Deep Learning is a very broad field and most promising work is moving around Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN).

Neural Networks have a long and prestigious history, and interest within the field of Text Analytics has been revived recently. In a traditional neural network all inputs (and outputs) are independent of each other. The idea behind RNNs is to make use of sequential information (as the words in a sentence). In order to predict the next word in a sentence we must know which words came before it. RNNs perform the same task for every element of a sequence, with the output being dependent on the previous computations. RNNs have a "memory" which captures information about what has been calculated so far. With RNN a given language model can be built, which in turn allows to score arbitrary sentences based on how likely they are to occur in the real world, and later that model allows to generate new text.

CNNs are basically just several layers of convolutions over the input layer to compute the output. Each layer applies different filters, typically hundreds or thousands, and combines their results. These can be used for Sentiment Analysis, Spam Detection or Topic Categorization, but they are of little use with PoS Tagging or Entity Extraction unless additional features are included as filters.

DL4J is a tool for textual analysis using deep learning techniques [26]. It builds on Word2vec, a two-layer neural network that processes text, created by Google [27].

III. Known Problems in Text Analytics

In the context of TA, Big Data is simply a massive volume of written language data. But where does the frontier lie between Big Data and Small Data? There has been a culture-changing fact: while merely 15 years ago a text corpus of 150 million words was considered huge, currently no less than 8.000 million word datasets are available. Not only is it a question simply about size, but also about quality and veracity: data from social media are full of noise and distortion. All datasets have these problems but they are more potentially serious for large datasets because the computer is an intermediary and the human expert do not see them directly, as is the case in small datasets. Therefore, data cleansing processes consume significant efforts and often after the cleansing, the availability of information to train systems is not enough to get reliable predictions, as happened in the Google Flu Trends failed experiment [27].

The reason is that most big datasets are not the output of instruments designed to produce valid and reliable data for analysis, and also because data cleansing is about (mostly subjective) decisions on the relevant design features. Another key issue is the access to the data. In most cases, the academic groups have no access to data from companies such as Google, Twitter or Facebook. For instance, Twitter only makes a very small fraction of its tweets available to the public through its APIs. Additionally, the tweets available do not follow a given pattern (they are an "assorted bunch") so it is difficult to arrive at a conclusion concerning their representativeness. As a consequence, the replication of analyses is almost impossible, since the supporting materials and the underlying technology are not publicly available. Boyd and Crawford [29] go further: limited access to Big Data creates new digital divides, the Big Data rich and the Big Data poor. One needs the means to collect them, and the expertise to analyze them. Interestingly, small but well curated collections of language data (the traditional corpora) offer information that cannot be inferred from big datasets [30].

How to grasp the figurative uses of language, basically irony and metaphor, is also a well-known problem to properly understand text. Essentially, the user's intentions are hidden because the surface meaning is different to the underlying meaning. As a consequence, the words must be interpreted in context and with extra-linguistic knowledge, a fact that being hard on humans, it is even harder for machines. How to translate a given metaphor into another language is extremely difficult. Some estimates calculate that figurative language is about 15-20% of the total content in social media conversations.

IV. Some Use Cases

Text Analytics has produced useful applications for everyday use. Here we just show a sample of these.

A. Lynguo

Social Media Analytics (SMA) consists of gathering data from the social media (Twitter, Facebook, blogs, RSS, websites) dispersed across a myriad of on-line, dynamic, sources; after analyzing automatically the data, these are shown to the client in a graphical environment (dashboard) to help adopting business decisions. Two are the commercial applications: the first one is focused on users' profiles and their network; the other one is targetting the content of the messages. In both cases, the SMA tools support marketing and customer service activities.

Customer profiling is the task of analysing the presence of a brand or the spread of a user's posted content in a social network and extracting information from the users related to gender, age, education, interests, consumer habits, or personality. Typically, the tool provides metrics on influencers and most commented posts. Also, it can identify similarities among users/customers, conversation groups and opinion leaders (Social Intelligence).

Content analytics is the task of analyzing the social media written messages to detect and understand the people's opinions, sentiments, intentions, emotions about a given topic, brand, person. After analyzing millions of comments in almost real-time, these applications help to detect crisis, to measure popularity, reputation and trends, to know the impact of marketing campaigns and customer engagement, or to find out new business opportunities.

Lynguo is an intelligent system developed by the Instituto de Ingeniería del Conocimiento (IIC – http://www.iic.uam.es) that combines both services (the network and the semantics) in a single suite. For instance, Lynguo Observer provides complete metrics on users, comments, trending topics, and their time evolution as well as geolocalization when available. Lynguo Opinion focuses on comments' content: the polarity (positive, neutral or negative) and their distribution through time, and classified by topics and by brands. Lynguo Ideas complements the semantic analysis identifying keywords and concepts associated to brands and people in the social media. This functionality helps to know the distinctive words for a given entity with respect to their competitors based on their exclusive presence or more frequent occurrence of those words in the messages for that entity. Lynguo Alert allows to set up conditions for alerts, for instance, high impact posts or sent by specific users. Finally, Lynguo Report generates personalized reports with graphics from information previously analyzed by other Lynguo modules.

In the next future, two new functionalities will be added: Lynguo Intention and Lynguo Emotions. Analyzing intentions in utterances is an old topic in NLP and AI since the 70s. The goal is to detect what the speaker plans to pursue with his/her speech acts. In the current context of Opinion Mining and Sentiment Analysis [31], intention detection is

focused on determining whether the social media users are expressing a complaint, a question, a wish, a suggestion about a product or a service. The practical application of this knowledge can be useful to the customer services and on-line complaint facilities. Emotions are basically positive or negative, being positive words more frequent, thus carrying less information than negative words [32]. In Lynguo, there are as many as twenty different categories for emotions, from happiness to anger, including intermediate states. The idea is to have a full range of granularity for emotions, and use a different scale for a given task.

There are two main approaches to extraction of opinion polarity in utterances [33]: the statistical or machine learning approach and the lexicon-based approach. The first is the most extended, and basically is a supervised classification task, where the features are learned from annotated instances of texts or sentences. The lexicon-based method involves using a dictionary of words and phrases with polarity values. Then, the program analyzes all relevant polarity words in a given text annotating their scores, and calculates the final aggregation into a final score. The statistical classifiers reach quite a high accuracy in detecting polarity of a text in the domain that they are trained on, but their performance drops drastically when the domain is different. The lexicon-based classifiers operate in a deeper level of analysis, including grammatical and syntactical information for dealing with negation and intensification, for which syntactic parsers are introduced to analyze sentences. Needless to say that Lynguo makes use of this approach.

B. IBM's Watson

IBM has a long history in AI research, and they consider themselves (http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=1569) as one of the founding fathers of AI in the early days in the 1950s. Along the years IBM created a checkers player [34]; a robotic system trained to assemble some parts of an IBM typewriter [35]; Deep Blue, the specialized chess-playing server that beat then World Chess Champion, Garry Kasparov [36]; TD-Gammon, a backgammon playing program using reinforcement learning (RL) with multiple applications [37]; and pioneering work in Neural Network Learning, inspired in biological information processing [38]. More recently work focused on advanced Q&A (question and answer) and Cognitive Computing, of which the SyNAPSE project and Watson are the more well-known examples.

IBM Watson is a technology platform that uses natural language processing and machine learning to reveal insights from large amounts of unstructured data. Watson can extract information from all kinds of text-rich documents in a repository, it can then build patterns of the relationships of the different concepts in the documents, and then can create a Q&A system to query the contents of the repository. Having been trained using supervised learning techniques, Watson uses natural language processing to understand grammar and context by annotating components to extract mentions and relationships, evaluates all possible meanings and determines what is being asked, and presents answers based on the supporting evidence and quality of information provided.

C. IPsoft's Amelia

IPsoft has been one leading proponent of automatization as part of the future of IT, where more and more menial tasks will be performed by expert systems rather than by people. IPsoft launched an initial version of Amelia, as the software platform is known, in September 2014. A new version, Amelia 2.0, has been presented in the last part of 2015, representing a new generation of this artificial intelligence (AI) platform with the promise that near human cognitive capabilities are ever closer [39].

The platform understands the semantics of language and can learn to solve business process queries like a human being would do. It can also solve technical problems by initially learning the same manuals as humans and then learning through experience and by observing the interactions between human agents and customers using semi-supervised machine learning techniques. Amelia 2.0 can complete more tasks and absorb more knowledge now as its core understanding capabilities mature. The latest version has improvements in dialogue management comprehension and emotional engagement [40]:

- *Memory* – Amelia's declarative memory consists of episodic memory and semantic memory. Episodic memory gives the basic cognition of various experiences and events in time in a sequenced autobiographical form. Semantic memory gives a structured record of facts, meanings, concepts and knowledge about the world/domain.
- *Contextual comprehension* – Concepts and ideas in the human brain are semantically linked. Amelia can quickly and reliably retrieve information across a wider and more complex set of knowledge.
- *Emotional responsiveness* – Research shows that a better user experience is directly tied to empathy shown by the agent throughout the interaction. In addition to an increased emotional quotient (EQ), Amelia presents a mood and a personality vector in a 3-dimensional emotional space.

The software is already used for services such as technology helpdesks, contact centres, procurement processing and to advise field engineers, among other business processes.

Both Watson and Amelia were already briefly mentioned in Redondo [41], as the purpose of both systems is to extend a human's capabilities by applying intelligent artificial systems techniques, such as deep learning and interpersonal communication through dialogue management.

V. Examples of TA Applications

We will briefly review two prominent areas of application of Text Analytics, with a large commercial impact: (1) Medical Analytics – classification of articles of medical content, and (2) Legal Analytics – Information extraction from legal texts.

A. Medical Analytics – Classification of articles or medical content

Biomedical text mining or BioNLP presents some unique data types. Their typical texts are abstracts of scientific papers, as well as medical reports. The main task is to classify papers by many different categories, in order to feed a database (like MEDLINE). Other applications include indexing documents by concepts, usually based or related to ontologies (like Unified Medical Language System-UMLS, or SNOMED-CT) or performing "translational research," that is, using basic biological research to inform clinical practice (for instance, automatically extraction of drug-drug interactions, or gene associations with diseases, or mutations in proteins).

The NLP techniques include biomedical entities recognition, pattern recognition, and machine learning for extracting semantic relations between concepts. Biomedical entities recognition consists of recognizing and categorizing entity names in biomedical domains, such as proteins, genes, diseases, drugs, organs and medical specialties. A variety of lexical resources are available in English and other languages (ontologies, terminological databases, nomenclatures), as well as a wide collection of annotated corpora (as GENIA) with semantic and conceptual relations between entities. Despite their availability, no single resource is enough nor comprehensive since new drugs and genes are discovered constantly. This is the main challenge for BioNLP.

There are three approaches for extracting relations between entities:

- Linguistic-based approaches: the idea is to employ parsers to grasp syntactic structures and map them into semantic representations. They are typically based on lexical resources and their main drawbacks are the abundance of synonyms and spelling variations for entities and concepts.
- Pattern-based approaches: these methods make use of a set of patterns for potential relationships, defined by domain experts.
- Machine Learning-based approaches: from annotated texts by human experts, these techniques extract relations in new collections of similar texts. Their main shortcoming is the requirement of computationally expensive training and testing on large amounts of human-tagged data. To extend the extraction system to another type of data or language requires new human effort in annotation.

Friedman et al. [42] presents a survey of the state of the art and prospects in BioNLP, sponsored by the US National Library of Medicine. This report identifies that "the most significant confounding element for clinical NLP is inaccessibility of large scale de-identified clinical corpora, which are needed for training and evaluation."

B. Legal Analytics – Information extraction from legal texts

One area getting a lot of attention about the practicalities of Text Analytics is that concerning the information extraction from texts with legal content. More specifically, litigation data is full of references to judges, lawyers, parties (companies, public organizations, and so on), and patents, gathered from several millions of pages containing all kinds of Intellectual Property (IP) litigation information. This has given rise to the term Legal Analytics, since analytics helps in discovering patterns with meaning hidden in the repositories of data. What it means to lawyers is the combination of insights coming from bottom-up data with top-down authority and experience found in statutes, regulations and court sentences. All this places objective data at the center instead of the so-called anecdotal data.

The underlying problem is that legal textual information is expressed in natural language. While a search can be made for the string *plaintiff*, there are no searches for a string that represents an individual who bears the role of plaintiff. To make language on the Web more meaningful and structured, additional content must be added to the source material, which is where the Semantic Web (semantic roles' tagging) and Natural Language Processing perform their contribution.

We start with an input, the corpus of texts, and then an output, texts annotated with XML tags, JSON tags or other mechanisms. However, getting from a corpus of textual information to annotated output is a demanding task, generically referred to as the knowledge acquisition bottleneck [43]. This task is very demanding on resources (especially manpower with enough expertise to train the systems) and it is also highly knowledge-intensive since whoever is doing the annotation must know what and how to annotate knowledge related to a given domain.

Processing Natural Language (NL) to support such richly annotated documents presents some inherent issues. NL supports all of the following, among other things:

(1) *implicit or presupposed information* – "When did you stop taking drugs?" (presupposes that the person is questioned about taking drugs at some time in the past);

(2) *multiple forms with the same meaning* – Jane Smith, Jane R. Smith, Smith, Attorney Smith… (our NER system must know that these are different ways to refer to the same physical person);

(3) *the same form with different contextually dependent meanings* – An individual referred to as "Jane Smith" in one case decision may not be the individual referred to by the name "Jane Smith" in another case decision; and

(4) *dispersed meanings* – Jane Smith represented Jones Inc. She works for Boies, Schiller and Flexner. To contact her, write to j.smith@bsfllp.com

People grasp naturally relationships between words and phrases, such that if it is true that *Bill used a knife to injure Phil*, then *Bill injured Phil*. Natural Language Processing (NLP) addresses this highly complex problem as an *engineering* problem, decomposing large problems into smaller problems and subdomains (e.g. summarization, information extraction, …) that we have already covered above. In the area of Legal Analytics we are primarily interested in information extraction. Typically a legal analytics system will annotate elements of interest, in order to identify a range of particular pieces of information that would be relevant to legal professionals such as:

- Case citation
- Names of parties
- Roles of parties, meaning *plaintiff* or *defendant*
- Type of court
- Names of judges
- Names of attorneys
- Roles of attorneys, meaning the side they represent (plaintiff or defendant)
- Final decision
- Cases cited
- Nature of the case, meaning using keywords to classify the case in terms of subject (e.g., criminal assault, intellectual property, etc.)

The business implications of Legal Analytics have originated a full branch of textual Big Data applications. Some companies have benefitted from a lucrative market, such as LexisNexis (http://www.lexisnexis.com/en-us/gateway.page), focused on offering predictions on potential medical-malpractice cases to specialized attorneys. Quite recently, LexisNexis has acquired Lex Machina [44], a company that mines mainly litigation data around IP information. Lex Machina originated in the departments of Law and Computer Science of Stanford University. Their Legal Analytics Platform helps to create a well-documented strategy to win cases on Patent, Trademark and Copyright data.

Every day, Lex Machina's crawler extracts data (and indexes documents) from several U.S Law repositories. The crawler automatically captures every docket event and downloads key case documents. It converts the documents by optical character recognition (OCR) to searchable text and stores each one as a PDF file. When the crawler finds a mention of a patent, it fetches information about that patent from the Patents and Trademarks Office (PTO) site. The crawler invokes Lexpressions, a proprietary legal text classification engine. The NLP technology classifies cases and dockets and resolves entity names (using a NER engine). A process of curation of the information extracted is performed by specialized attorneys to ensure high-quality data. The structured text indexer then performs a data cleansing operation to order all the data and stores it for search. Lex Machina's web-based application enables users to run search queries that deliver easy information retrieval of the relevant docket entries and documents.

Most if not all of the work around either medical or legal text analytics has originated from data sets in English (the above-mentioned MEDLINE, LexisNexis, or others like MedScape), with little to no work in other languages. This could be an opportunity to create new text analytics systems in languages other than English. For instance, considering the fact that the legal systems vary quite significantly from

country to country, this is a good field to grow new areas of business in the so-called Digital Economy (see also IJIMAI's last year special issue on this topic).

VI. Future Work

The technologies around text analytics are currently being applied in several industries, for instance, sentiment and opinion analysis in media, finance, healthcare, marketing branding or consumer markets. Insights are extracted not only from the traditional enterprise data sources, but also from online and social media, since more and more the general public has turned out to be the largest generator of text content (just imagine online messaging systems like Whatsapp or Telegram).

The current state of text analytics is very healthy, but there is room for growth in areas such as customer experience, or social listening. This bears good promises for both scientific experimentation and technical innovation alike: Multi-lingual analytics is facilitated by machine learning (ML) and advances in machine translation; customer experience, market research, and consumer insights, and digital analytics and media measurement are enhanced through text analytics; besides the future of deep learning in NLP, long-established language-engineering approaches taxonomies, parsers, lexical and semantic networks, and syntactic-rule systems will continue as bedrocks in the area; emotion analytics, affective states compounded of speech and text as well as images and facial-expression analysis; new forms of supratextual communications like emojis need their own approach to extract semantics and arrive at meaningful analytics; semantic search and knowledge graphs, speech analytics and simultaneous machine translation; and machine-written content, or the capability to compose articles (and email, text messages, summaries, and translations) from text, data, rules, and context, as captured previously in the analytics phase.

VII. Conclusion

Text Analytics, with its long and prestigious history, is an area in constant evolution. It sits at the center of Big Data's Variety vector, that of unstructured information, especially with social communications, where content is generated by millions of users, content not only consisting of images but most of the times textual comments or full blown articles. Information expressed by means of texts involves lots of knowledge about the world and about the entities in this world as well as the interactions among them. That knowledge about the world has already been put to use in order to create the cognitive applications, like IBM's Watson and IPsoft's Amelia, that will interact with human beings expanding their capabilities and helping them perform better. With increased communication, Text Analytics will be expanded and it will be needed to sort out the noise and the irrelevant from the really important information. The future looks more than promising.

References

[1] Xerox Corporation (2015): http://www.xrce.xerox.com/Research-Development/Industry-Expertise/Finance (accessed 26 December 2015)

[2] Apache OpenNLP (2015): http://opennlp.apache.org/ (accessed 19 December 2015)

[3] Stanford Named Entity Recognizer (2015): http://www-nlp.stanford.edu/software/CRF-NER.shtml (accessed 19 December 2015)

[4] J. R. Finkel, T. Grenager, and C. Manning (2005). "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. (online reading: http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf)

[5] LingPipe (2011): http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html (accessed 29 November 2015)

[6] S. Lee and H. Kim (2008). "News Keyword Extraction for Topic Tracking". *Fourth International Conference on Networked Computing and Advanced Information Management*, IEEE.

[7] Google Alerts (2016): http://www.google.com/alerts (accessed 10 January 2016)

[8] W. Xiaowei, J. Longbin, M. Jialin and Jiangyan (2008). "Use of NER Information for Improved Topic Tracking", *Eighth International Conference on Intelligent Systems Design and Applications*, IEEE Computer Society.

[9] ATLAS Project (2013): http://www.atlasproject.eu/atlas/project/task/5.1 (accessed 10 January 2016)

[10] G. Wen, G. Chen, and L. Jiang (2006). "Performing Text Categorization on Manifold". *2006 IEEE International Conference on Systems, Man, and Cybernetics*, Taipei, Taiwan, IEEE.

[11] H. Cordobés, A. Fernández Anta, L.F. Chiroque, F. Pérez García, T. Redondo, A. Santos (2014). "Graph-based Techniques for Topic Classification of Tweets in Spanish". *International Journal of Interactive Multimedia an Artificial Intelligence*.

[12] T. Theodosiou, N. Darzentas, L. Angelis, C.A. Ouzonis (2008). "PuReD-MCL: a graph-based PubMed document clustering methodology". *Bioinformatic*s 24.

[13] Q. Lu, J. G. Conrad, K. Al-Kofahi, W. Keenan (2011). "Legal document clustering with built-in topic segmentation", *Proceedings of the 20th ACM international conference on Information and knowledge management*.

[14] P. Cowling, S. Remde, P. Hartley, W. Stewart, J. Stock-Brooks, T. Woolley (2010), "C-Link Concept Linkage in Knowledge Repositories". *AAAI Spring Symposium Series*.

[15] C-Link (2015): http://www.conceptlinkage.org/ (accessed 10 December 2015)

[16] Y. Hassan-Montero, and V Herrero-Solana (2006). "Improving Tag-Clouds as Visual Information Retrieval Interfaces", *I International Conference on Multidisciplinary Information Sciences and Technologies*, InSciT2006.

[17] Wordle (2014): http://www.wordle.net/ (accessed 20 December 2015)

[18] M. A. Hearst (2009) "Information Visualization for Text Analysis", in *Search User Interfaces*. Cambridge University Press (online reading: http://searchuserinterfaces.com/book/)

[19] D3.js (2016): http://d3js.org/ (accessed 20 January 2016)

[20] Gephi (2016) https://gephi.org/ (accessed 20 January 2016)

[21] L. Hirschman, R. Gaizauskas (2001), "Natural language question answering: the view from here", *Natural Language Engineering 7*. Cambridge University Press. (online reading: http://www.loria.fr/~gardent/applicationsTAL/papers/jnle-qa.pdf)

[22] OpenEphyra (2011): https://mu.lti.cs.cmu.edu/trac/Ephyra/wiki/OpenEphyra (accessed 5 January 2016)

[23] N. Schlaefer, P. Gieselmann, and G. Sautter (2006). "The Ephyra QA system". *2006 Text Retrieval Conference (TREC)*.

[24] YodaQA (2015): http://ailao.eu/yodaqa/ (accessed 5 January 2016)

[25] P. Baudis (2015) "YodaQA: A Modular Question Answering System Pipeline". *POSTER 2015 — 19th International Student Conference on Electrical Engineering*. (online reading: http://ailao.eu/yodaqa/yodaqa-poster2015.pdf)

[26] DL4J (2015): http://deeplearning4j.org/textanalysis.html (accessed 16 December 2015)

[27] Google – Word2vec (2013): http://arxiv.org/pdf/1301.3781.pdf (accessed 20 December 2015)

[28] D. Lazer, R. Kennedy, G. King, and A. Vespignani (2014). "Big data. The parable of Google Flu: traps in big data analysis." *Science*, 343(6176).

[29] D. Boyd, and K. Crawford (2011). "Six Provocations for Big Data". *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. (Available at SSRN: http://ssrn.com/abstract=1926431 or http://dx.doi.org/10.2139/ssrn.1926431)

[30] A. Moreno, and E. Moro (2015). "Big data versus small data: the case of 'gripe' (flu) in Spanish". *Procedia, Social and Behavioral Sciences*, 198.

[31] B. Liu (2012). *Sentiment Analysis and Opinion Mining*. Morgan and Claypool. Chicago.

[32] D. Garcia, A. Garas, and F. Schweitzer (2012). "Positive words carry less information than negative words". *EPJ Data Science*, 1:3. (online reading: http://www.epjdatascience.com/content/1/1/3)

[33] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011).

"Lexicon-based Methods for Sentiment Analysis". *Computational Linguistics*, 37, 2. (online reading: https://www.aclweb.org/anthology/J/J11/J11-2001.pdf)

[34] A. Samuel (1959). "Some Studies in Machine Learning Using the Game of Checkers", *IBM Journal* **3** (3).

[35] C. A. Pratt (1982): "Robotics at IBM", *SIMULATION*.

[36] M. Campbell, A. Hoane, and F. Hsu (2001). "Deep Blue". *Artificial Intelligence,* 134.

[37] G. Tesauro (1995). "Temporal difference learning and TD-Gammon". *Communications of the ACM*, Vol. 38, No. 3.

[38] R. Linsker (1990). "Perceptual Neural Organization: Some Approaches Based on Network Models and Information Theory", *Annual Review of Neuroscience*, Vol. 13.

[39] Karl Flinders (2015). "Amelia, the IPsoft robot". *Computer Weekly* (http://www.computerweekly.com/news/4500254989/Amelia-the-IPsoft-robot-gets-a-makeover

[40] IPsoft (2015) (http://www.ipsoft.com/ipsoft-humanizes-artificial-intelligence-with-the-next-generation-of-its-cognitive-agent-amelia/

[41] T. Redondo (2015). "The Digital Economy: Social Interaction Technologies – an Overview". *International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 3- 2.*

[42] C. Friedman, T. Rindflesch, and M. Corn (2013). "Natural language processing: State of the art and prospect for significant progress, a workshop sponsored by the National Library of Medicine". *Journal of Biomedical Informatics*, 46.

[43] S. Potter (2002). *A Survey of Knowledge Acquisition from Natural Language*, part of the Advanced Knowledge Technologies project, University of Edinburgh.

[44] Lex Machina (2015): https://lexmachina.com/ (accessed 20 December 2015)

Antonio Moreno-Sandoval (BA 1986, MA 1988, PhD 1991, Universidad Autónoma de Madrid, UAM) is Professor of Linguistics and Director of the Computational Linguistics Lab at UAM. He is a former Fulbright postdoc scholar at the Computer Science Dept., New York University (1991-1992) and a former DAAD scholar at Augsburg Universität (1998). His training in Computational Linguistics began as a research assistant in the Eurotra Machine Translation Project (EU FP-2) and then at IBM Scientific Center in Madrid (1989-1990). He was the principal researcher of the Spanish team in the C-ORAL-ROM Project (EU FP-5). He has managed over 15 projects (national, regional-funded) as well as industry contracts. Since 2010 he is Senior Researcher at the Instituto de Ingeniería del Conocimiento (IIC-UAM) in the Social Business Analytics group. Moreno-Sandoval has supervised 9 theses to completion. He is author or co-author of 4 books and over 80 scientific papers.

Teófilo Redondo (BArts -1985, MArts - 1986; Universidad Complutense de Madrid - UCM) is Project Portfolio Coordinator at Zed Worldwide, in the Deparment of Innovation. He was before Technology Architect & Director of Innovation Projects at Universidad Internacional de La Rioja (UNIR). Previously he developed a career at IBM covering several areas like Cloud Computing and Big Data Architectures, Enterprise Solutions Architect (SAP, Oracle Solutions, Dassault Systemes), and as SOA Architect. He started in the IBM Research Division (IBM Scientific Center in Madrid) with several projects on Machine Translation, during which he produced a number of articles on this subject. He was Visiting Scholar at Stanford University (1987). He currently holds an Associate Professorship at UNIR teaching about eLearning in Social Networks as well as Natural Language Processing techniques. He is affiliated with SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) since almost the beginning.

.