Research directions for harvesting cross-sectorial correlations towards improved policy making

A. Drosou[1], N. Dimitriou[1], N. Sarris[2], A. Konstantinidis[3], Dimitrios Tzovaras[1]

[1] *Centre for Research & Technology Hellas;* [2] *Athens Technology Center*
[3] *Imperial College London*
{drosou,nikdim,tzovaras}@iti.gr, n.sarris@atc.gr; a.konstantinidis16@imperial.ac.uk

Abstract

The current paper focuses on the emerging problem of the management of and the processing required for the massive amounts of interdisciplinary data produced nowadays. As it becomes apparent that "the truth and the useful information is drowned in a sea of irrelevance due to the vast amount of information available"[1], there are strong indications that seemingly irrelevant co-occurrences of events and subtle links between them may form pieces of the same puzzle that complement each other towards the revealment of predictive or explanatory indicators for many sectors of the modern economy. To this direction, modern technologies like data mining, data and visual analytics, artificial intelligence, etc. can be of significant value, if offering a comprehensive communication of potentially useful information to the appropriate stakeholders and/or policy-makers. In this context, a multi-purpose platform for data analytics is briefly exhibited in order to demonstrate the potential of such approaches to policy making.

Keywords— data stream analysis, data visualization & analytics, data integration, information retrieval, cross-sectorial data correlations, finance, news media, journalism.

1 Introduction

It is recently becoming a common place that the world is becoming quite a noisy place; not only in terms of sound levels (acoustics)1 but most importantly in terms of uncertain (incl. expression and content) information production (computer science). To this end, the term Big Data (a.k.a. the 4V's[2]) has been introduced to describe the large volumes of data produced constantly at a high velocity from a large variety of sources. Its origins lay in the massive digitization wave, growing big during the information age, which in turn led to the production, collection, processing and storage of an enormous volume of interdisciplinary data that still grows exponentially, while parallelly grows the uncertainty they carry (see Figure 1)

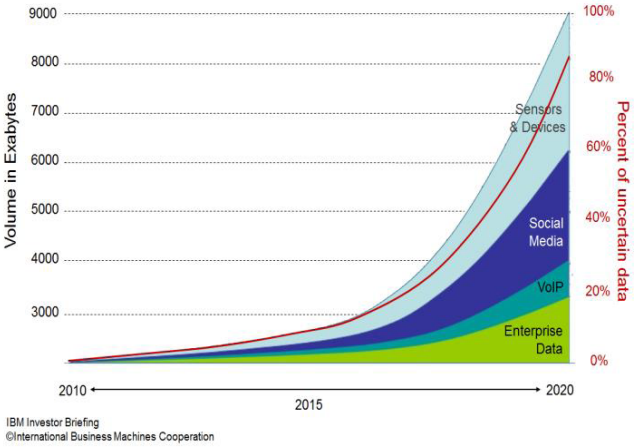


Figure 1: Volume of available data is exponentially increasing over time and so does also the uncertainty of them. *(source: IBM Investor Briefing - International Machines Corporation)*

The advance of social, event and news media, as well as the emergence of the Internet of Things (IoT) has led to the generation of enormous amounts of streaming and diverse types of data, ranging from textual and cognitive ones (e.g. quotes in tweets, elaborated opinions in news articles, etc.) to arithmetic and behavioural ones (e.g. stock market indices, geolocational routes in traffic, BGP routing in communications, etc.). In turn, this outburst has given rise to a

[1] Aldous Huxley, "Brave New World" (1932)

[2] The Big Data paradigm invariably suffers from the so-called curse of dimensionality, but it also harbours intrinsic advantages – namely, it provides multiple sources of redundancy, through both the sheer volume of data and the complex mutual information and statistical dependencies between the data/information channels

surge in the so-called *datafication*, i.e. the ability to quantitatively encode many aspects of the world that have never been quantified before (i.e. with traditional analysis of the pre-Big Data era). Data have nowadays infiltrated into every critical infrastructure of society and have become an essential asset of every sector and function of the global economy.

In particular, the real-time and high-frequency nature of data is also important in order to accurately implement the so-called *nowcasting* (i.e. the ability to estimate metrics - contraction of "now" and "forecasting"), which previously could only be done retrospectively. Thus, considerable power to prediction is added, while similarly, the high frequency of data allows hypothesis testing in near real-time and to a level never before possible. Even further, the interdependencies and correlations among data originating from multiple sources or describing multiple sectors can be exploited, allowing a deeper understanding of the true knowledge behind the bulk of data and leading to more accurate cause analysis, predictions and in the long term contributing to improved policy planning for corporate and governmental institutions.

2 Big Data exploitation for real-world policy making

Striking examples highlighting the need for mining significant cross-sectorial correlations can be easily identified in several domains of the modern economy, and they all can be addressed by the concept presented in Figure 2.

The sectors illustrated in Figure 2 could be particularized by the triangular relationship between journalism (incl. social and mass media), government, and the finance sector. News stories or rumors spreading rapidly through the social media have a large and immediate impact on financial indices and in the long-term affect policy makers [1]. In this concept and as Prof. R. Parker explains in his discussion paper at Harvard University [3], there is a symbiotic relation between journalism and economics, two of the major sectors/industries of modern consumer societies that are majorally driven from the latter, but also have a great impact on them. By projecting his view on the modern society, it can be noticed that this relation is becoming even more solid. In this respect, it has become a common place among key stakeholders [4] in this field that Big Data and the related analysis technologies have the potential to help companies grasp the social pulse and the social impact of their activities, early detect fraudulent events and/or evidence of systemic risks (e.g. Lehman Brothers case, false news spreading [5],

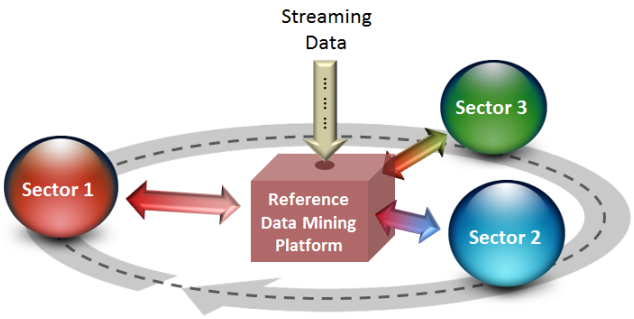


Figure 2: Cross-sectorial functionality of an open-architecture Platform. All available data are collected and processed so as to detect possible correlations and to evaluate them with respect to their significance (e.g. entropy). The data released by each side/sector are filtered via dedicated interfaces. The open architecture highlights the need for input modularity from different sectors under a common data exchange concept for improved decision making & mutually beneficial data sharing for x-sectorial correlation detection.

etc.) with catastrophic social and economic impact and be more efficient in long-term policy and strategy planning.

Alternatively, another cross-sectorial and triangular relationship can be formed by the transport research (e.g. behaviourally common traffic/travelling patterns), urban planning[20] (incl. real estate) and resilience management[21] that have also been proven mutually beneficial. However, the key innovation would be to catch the subtle public behaviours under certain normal and/or extreme conditions and to offer the policy makers the incentives to act pro-actively in emergent situation, to support the sustainability (e.g. CO2 emissions, temperature, etc.) of the transport system, to increase the productivity of the local labour class, to carefully have control on the real estate balance, to avoid the ghettoization of certain areas, etc.

2.1 Expectations from Big Data

Indeed, in a highly competitive environment, the need to take advantage of the opportunities offered by Big Data analysis techniques is now prominent, in order to provide high quality products and services:

- **new product development and optimized service offerings**, by gaining a deeper understanding of the market and the society factors that influence it most, as well as the ability to more accurately predict future events and trends,
- **cost and time reduction**, by being able to exploit the

constant stream of produced data and predict future events faster, thus saving resources,

- **smart decision making**, by being able to assess the data validity and importance more accurately.

Yet, nowadays, the concept of Big Data can even go beyond simple data collection and encompasses new technologies (e.g. deep learning, etc.), new skills (e.g. cloud computing, etc.) and new analysis techniques (e.g. root cause analysis) that have only become available in recent years. In addition, the exploitation of cross-sectorial correlations can provide new insights and significantly improve policy and decision making.

2.2 Cornerstones for Big Data analysis

More concretely, venues of scientific research that can be followed and ultimately integrated in order achieve multi-faceted and cross-sectorial Big Data analysis, can be identified:

- **Data integration:** In order to exchange data between sectors a common data model is needed that will allow the flawless exchange of information while respecting intellectual property rights (IPRs). To this end, a hierarchical data abstraction can be deployed with multiple layers [7, 8] with each one corresponding to different levels of granularity that will be available to users according to their requests and permission rights. Regarding data management, a NoSQL [6] framework on a distributed system is very promising solution, ensuring scalability, responsiveness and reliability.

- **Data mining from multi-modal and multi-dimensional data:** Multi-modal content analysis can combine traditional text processing techniques [9, 10] with language-agnostic processing algorithms. To this end, continuous vector representations and skip-gram architectures [11, 12] can be examined under the deep learning paradigm of artificial intelligence [14, 13] in order to capture non-linear dependencies and reveal concealed trends/correlations of heterogeneous data sources.

- **Signal processing:**In this category there is great potential in simulating/predicting the behaviour of stakeholders and estimating the fluctuation of related metric (e.g economic factors) by taking into account sentiment induced by external factors (e.g. news media. This can be achieved by following the principles of Catastrophe Theory [15] and fusing multifaceted data under a generic optimization framework such as the Alternating Direction Method of Multipliers (ADMM) [16] suitable for a distributed deployment. In order to deal with the high dimensionality of cross-sectorial data, expressive representations can be used based on tensors and structured graph models [17] in order to find common factors in diverse data such as financial indices and news snippets. Taking into account the validity of a data source as estimated by sentiment analysis and text processing algorithm erroneous or inaccurate data can be properly handled.

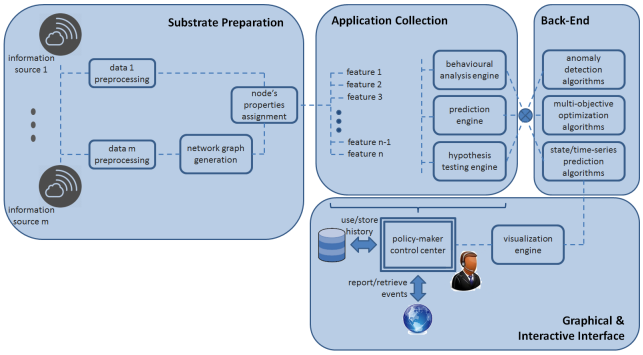


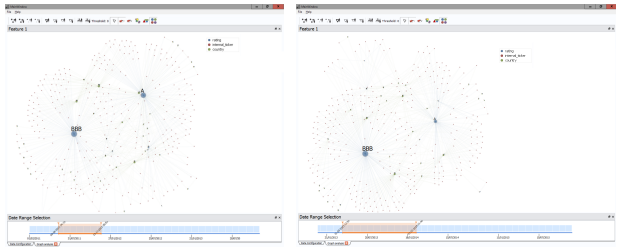
Figure 3: The conceptual architecture of the multi-purpose Graph Analytics Platform.

- **Visual (data) analytics:** To accommodate the interactive analysis of heterogeneous data sources visualization technologies [18, 19] can be used to provide insight on correlations among data from multiple sources and the facilitation of root cause analysis and hypothesis testing by end users.

3 GAP - The potential of data mining for policy making

The Graph Analytics Platform (GAP) is a powerful that offers a rich toolkit for and is built upon the concept of top-down data mining approach (i.e. data minimization) aiming at the detection of data correlations. Its conceptual architecture is illustrated in Figure 3.

The key feature of this platform is that it is data-type agnostic and it manages to model any input of multi-dimensional and interdisciplinary data into efficient graph structures. Once this is done, GAP manages to perform several sophisticated processes like multi-attribute clustering, hypothesis tests, sentiment analysis, behavioural analysis facilitating thus, highly accurate root cause analysis tasks either predictive for future states or the past records.



(a) (b)

Figure 4: Subfigures (a) & (b) show different placements of the countries around the available ratings at different time instances, as shown in the timeline in the bottom. The relevant position of a country to the rating nodes shows is analogous to the accumulated rating of its corresponding stocks.

Possible application for effective and imminent policy making, as derived by the GAP platform are shown in the following subsections.

3.1 A clustering example for the country-based analysis of ratings

The securities of every company in the stock market are accompanied by a rating index. Provided that each company is also categorized based on the country of its origin, an useful index for making predictions in country-related stock market (e.g. currencies) is the average rating of a country among the different rating levels. In this context, GAP can reveal such information and present it in different levels of abstractions (see Figure 4), helping thus the analyst or the policy maker to extract the appropriate conclusions and proceed to the most profitable placements or take countermeasures to prevent negative impacts.

3.2 A social media example for early detection of emerging trends

The analysis that is illustrated in Figure 5 concerns the Twitter related dataset as it was collected the days before and after the terroristic attack in Paris on the 13^{th} of November 2015. The graph structure in Figure 5 reveals the following information: On the right side of the image two uni-modal graphs can be seen, each one illustrating the nodes around the city they have been sent from and about the city they have been talking about, respectively. The large network graph in the middle of Figure 5 has been the results of the aforementioned uni-modal graphs in a large multi-modal one. The clusters that are close to each other inidcate these cities between which a large amount of information is exchanged. For instance on can notice that there are a lot of tweets from New York about Paris and also quite a few from Ottawa, Toronto & Singapore talking about New York. The wordclouds can give an overview to the analyst about what the emerging trends in these regions are. Similarly the sentiment bar shows the corresponding sentiment expressed by each cluster selected.

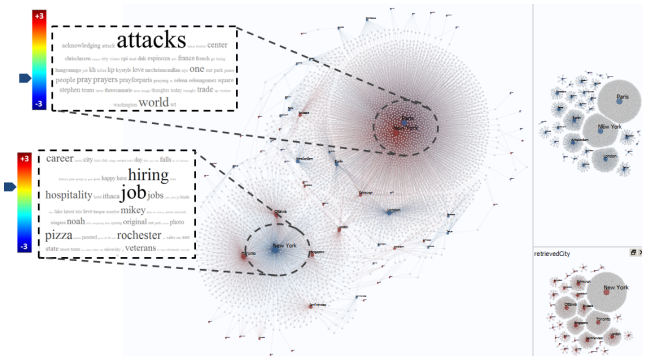


Figure 5: k-partite Graph visualization of all the tweets (i.e. nodes) that were posted between the 10^{th} and the 16^{th} of November 2015. The utilized features were the name of the city the tweet retrieved from and the name of the city contained within its text.

As the reader can notice in Figure 5, the most frequent word found in the tweet and re-tweeted messages is the word 'attacks', which directly links to the terroristic attack that occurred on that day. Similarly, the tweets about New York mainly concern the spreading out of some new vacancies and 'job'/'hiring' opportunities in the area and this can also be justified by the fact that the discussion is being held within the neighbouring cities. Apparently the tweets regarding the terroristic attack have a significantly lower sentiment than the ones about hiring opportunities.

Acknowledgements

This work has been partially supported by the European Commission through project H2020-DRS-2014-653460-RESOLUTE funded by the "H2020-EU.3.7.-Secure societies" program. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the European Commission.

References

[1] https://www.tradeking.com/investing/stock-investment-news

[2] http://www.sharemarketschool.com/how-does-news-affect-stock-prices/

[3] R. Parker, "Journalism & Economics: The Tangled Webs of Profession, Narrative, and Responsibility in a Modern Democracy", Press Politics, Press Policy, The Joan Shorenstein Center, Harvard University, 1997.

[4] http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[5] http://www.reuters.com/article/net-us-usa-whitehouse-ap-idUSBRE93M12Y20130423

[6] R. Cattell, "Scalable SQL and NoSQL data stores", ACM SIGMOD Record, vol. 39, no. 4 pp. 12–27, 2011.

[7] D. Ganesan, B. Greenstein, D. Perelyubskiy, D. Estrin and J. Heidemann, "An evaluation of multi-resolution storage for sensor networks", In Proceedings of the 1st international conference on Embedded networked sensor systems, ACM, pp. 89–102, 2003.

[8] M. Buevich, A. Wright, R. Sargent and A. Rowe, "Respawn: A distributed multi-resolution time-series datastore", In IEEE 34^{th} Real-Time Systems Symposium (RTSS), pp. 288–297, 2013.

[9] E. Lloret and M. Palomar, "Text summarisation in progress: a literature review", Artificial Intelligence Review, vol. 37, no. 1, pp. 1–41, 2012.

[10] A. Passos, V. Kumar and A. McCallum, "Lexicon infused phrase embeddings for named entity resolution", CoRR, vol. abs/1404.5367, 2014.

[11] T. Mikolov, Kai Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space" CoRR, vol. abs/1301.3781, 2013.

[12] https://code.google.com/archive/p/word2vec/

[13] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y.Ng and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank", Proceedings of the conference on empirical methods in natural language processing (EMNLP), vol. 1631, 2013.

[14] Y. Kim, "Convolutional neural networks for sentence classification", Proceedings of the conference on empirical methods in natural language processing (EMNLP), pp. 1746–1751, 2014.

[15] P. A. I. Hartelman, "Stochastic catastrophe theory", Dissertatie reeks, Faculteit Psychologie, Universiteit can Amsterdam, 1997.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, "Distributed optimisation and statistical learning via the alternating direction method of multipliers", Foundations and Trends in Machine Learning, vol. 3, no. 1, pp. 1–122, 2010.

[17] A. Cichocki, D. Mandic, A-H. Phan, C. Caiafa, G. Zhou, Q. Zhao and L. De Lathauweret,"Tensor decompositions for signal processing applications", IEEE Signal Processing Magazine, vol. 32, no. 2, pp. 145–163, 2015.

[18] D. Keim, G. Andrienko, J.D. Fekete, C., J. Kohlhammer and G. Melancon, "Visual analytics: Definition, process, and challenges" Springer Berlin Heidelberg, pp. 154–175, 2008.

[19] J.S. Yi, Y. ah Kang, J.T. Stasko and J.A. Jacko, "Toward a deeper understanding of the role of interaction in information visualization", IEEE Trans. Vis. Comput. Graphics, vol. 13, no. 6, pp. 1224-1231, 2007.

[20] J. Xu, D. Deng, U. Demiryurek, C. Shahabi and M. v. d. Schaar, "Mining the Situation: Spatiotemporal Traffic Prediction With Big Data" in IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 4, pp. 702–715, 2015.

[21] C. Giovanna, M. Giuseppe, P. Antonio, R. Corrado, R. Francesco and V. Antonino, "Transport models and intelligent transportation system to support urban evacuation planning process", in IET Intelligent Transport Systems, vol. 10, no. 4, pp. 279–286, 2016.