


PaloAnalytics project concept, scope and outcomes: an opportunity for culture

Vassilis Pouloupoulos¹[0000-0003-1707-3153], Manolis Wallace¹[0000-0002-4629-5946], Costas Vassilakis²[0000-0001-9940-1821], and George Lepouras²[0000-0001-6094-3308] *

¹  Knowledge and Uncertainty Research Laboratory
University of the Peloponnese, Tripolis, Greece 221 31
{wallace, vacilos}@uop.gr
<http://gav.uop.gr>

² University of the Peloponnese
Tripolis, Greece 221 31
{costas, gl}@uop.gr

Abstract. This paper describes the national funded project entitled PaloAnalytics, which develops an innovative platform that allows companies and organizations, that operate in several countries, to monitor and analyze, in depth, the markets' interest to their products and successfully plan their marketing and communication strategy, with data and insights collected from all the local media, and focuses on its application to cultural spaces and museums. In this notion, we examine the effect that this project can have in cultural spaces or companies related to arts and culture. PaloAnalytics platform allows organizations to investigate the impact of their products on consumers across different countries and this is achieved with the analysis of content from sites, blogs, social networks and open data. This implies that cultural organization can benefit by adopting the implemented services, so that the can recognize and analyze their audience, their online marketing campaigns as well as examine the impact of their messages and the spread of their messages on the Internet. In this paper, we briefly describe the project and discuss on the impact on cultural related organizations.

Keywords: big data, data monitoring, trending topics, influencers, info graphics, data visualization, deep learning

1 Introduction

The data that is generated daily in the world of the internet is vast. The amount of information is such that it is impossible for companies and organizations to fetch, analyze and learn from all the data produced. In this scope, PaloAnalytics is a project that aims to perform the procedures of collecting, analyzing and

* Cultural Informatics 2019, June 9, 2019, Larnaca, Cyprus. Copyright held by the authors.

extracting useful information from different sources of the internet, web pages, news portals, open sources, and social media. The procedure of collecting and analyzing information from diverse sources is not something new, and has attracted research during the last 20 years [10]. It resides to the area of Data Mining [12], [11] and it focuses on Big Data analysis, which is based on multiple custom Data Warehouses [13]. In this notion, we present a project that intends to employ resources from all these sectors in order to produce its final results; in depth analysis of social media and web data in order to support organizations and companies.

Market research has proven that companies and organizations are in strong need for an holistic market monitoring and analysis service in several countries and not solely the country of their origin; or at least they are convinced that they can perform much better if they have such a tool. Besides, the competition of such companies and organizations is usually international. Furthermore, it clear that when analyzing data in an international environment each local information can easily affect the whole organization, but it is usually difficult to become a part of the organization's international policy. In general, it seems possible that data can be collected and analyzed in some extent locally but is usually not transferred as knowledge to the international level. In fact, for such organizations it would be extremely useful to utilize a unique language for all the data analyzed and it seems that the English language is acceptable and consistent. A number of tools have been developed including Mention³ and Brandwatch⁴ in order to collect and analyze data internationally but they have some major disadvantages. They focus mainly on social media and target experienced users, while in parallel they do not provide translations of reports from local languages in a universal language. Furthermore, they do not offer a homogeneous overall picture for all the countries that are of interest for a business.

In this ground, we introduce PaloAnalytics project, which intends to focus on the basic challenges that organizations face and includes the ability to have a universal monitoring tool, with links and interconnections between data collected and analyzed from a number of different sources and different languages. In this way the project will be the ideal solution for international companies (or companies willing to become international) and companies that their international competition affect their local business. An ideal solution, through which the organization will be able to get information out of large sets of data.

The proposed design and implementation, introduces a series of software modules that will

- analyze multilingual content posted on news sites, social networks and open data
- extract knowledge and information about products and companies, including product characteristics
- analyze sources, their influence and trends

³ <https://mention.com> - Mention: Scour the web, social media, and more for powerful market insights

⁴ <https://www.brandwatch.com/> - Brandwatch: Know what your customers think

- help in assessing the image of the business and its products as well as its competitors
- visualize the knowledge in order to easily understand the analyzed information

These procedures describe how this project can be used by any type of company. Research has shown that cultural spaces and organizations have started to take seriously the world of the Internet and the Social Media. Consequently, they find it attractive to spread their messages through these mediums, as it is expected to reach a larger and global audience, they can make serious debates and conversations, and, generally, have an alternative active role in order to challenge the mass culture. Having the aforementioned as a base, it is evident that the project can help all these organizations have a holistic presence in social media and the internet; a presence that can be expected to be international.

The rest of the paper is structured as follows: Section 2 presents the methodology of the project, while section 3 discusses the system architecture. In section ?? a detailed description of each component is presented, providing more emphasis on the Trending Topics software module and its results. Section 4 defines the expected outcomes of the proposed system and the final section presents a discussion on the project.

2 Methodology of the project

Due to the large number of different modules, the high complexity of their implementation and the importance in precision of their algorithmic procedures, an advanced methodology is employed. As such, the Rational Unified Process is used. It is a software engineering procedure that ensures producing high quality software and achieving end user needs within a specific timetable and cost. Two cycles of project evolution are followed, one that leads to the basic implementation and is longer, while a shorter one will be done in order to perform refinements. Both of the cycles will go through the same steps of development. During the first cycle the implementation will be ensured, while the second cycle will focus on the quality of the outcomes. The cycle phases include:

- Inception Phase
- Elaboration Phase
- Construction Phase
- Transition Phase

Figure 1 depicts the cycle of system design, implementation and integration.

During the inception phase a general description of the key requirements of the project is done; key points and the basic constraints are defined and the system use cases are defined in brief. An initial business case including the business framework, the success criteria and financial forecasting is the ones that lead to the project plan and to a draft business model. While analyzing the

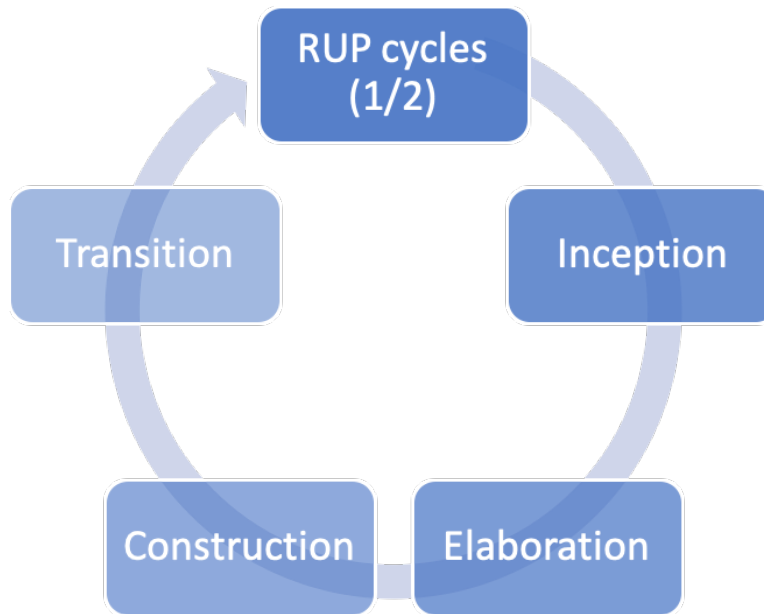


Fig. 1: Rational Unified Process

information, during the processing phase, the use case model was completed, and the final requirements were recorded. The architecture reached its final form and the project's development plan was finalized. Currently the project is under the first construction phase, where modules are implemented and starting to be integrate into the PaloAnalytics platform. Upon completion of the first phase of implementations an overall system functionality, performance and usability test will be done.

The development of the platform follows a bottom-up approach, based on the proposed architecture as presented in figure 2), starting from data collection that will directly lead to data aggregation services which will be used individually. On the produced data, multilingual content analysis' services are employed, while in parallel, at this stage, business intelligence extracting solutions are applied. The availability of the proposed services will be both on Web and Mobile application enabling increased penetration into the business community. Each service is built supporting endpoint integration in order to be available for use as an individual component even for third party systems, external to PaloAnalytics platform.

This will develop a complete development stack, that is based on multilingual content from news sites, open data sources and social media. The services of this stack are expected to attract third-party businesses companies, public bodies and researchers who will develop new management modes of business data from the sources incorporated by PaloAnalytics platform and will set up new business models on them, multiplying the benefits for the companies and organizations

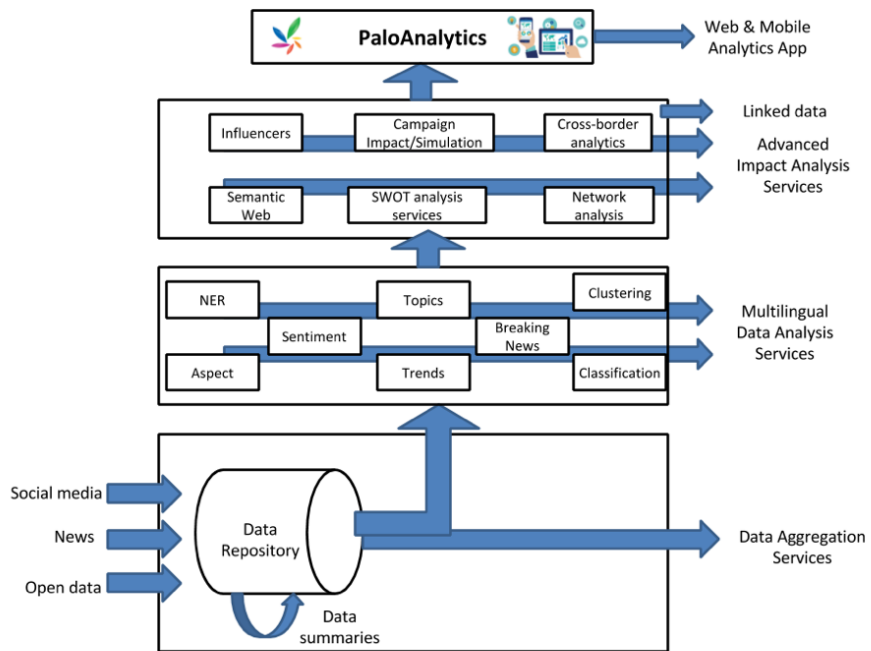


Fig. 2: Proposed architecture

while maximizing the influence of the proposed solutions for the scientific and business community.

3 Architecture

According to the architecture presented in figure 2 the proposed system is divided into several components and modules enabling in this way individual design and integration. The system, though, can be separated into four major components:

- data entrance point / data storage
- deep data analysis
- semantics and metadata analysis
- point of presentation

Each of the major components consists of a number of modules in order to successfully achieve its scope. Furthermore, each component will offer services for direct data extraction and usage by third party systems.

3.1 Entry point

The entry point of the system is the component that is responsible for collecting and storing data from the several different sources (social media, news and open data). The data storage is built enabling several interfaces to be connected in order to fetch and store data. In general, it follows a hybrid scheme including both an SQL and a noSQL database.

The system acts as a data warehouse, including modules for data extraction, data transformation as well as data loading. The extraction of data is done from several different sources including news websites, blogging platforms, social media - focusing on text based ones - and open data sources. The data collected is transformed in order to formulate similar objects with specific unified structure. The unified structure of each unique object includes a unique identifier, title, body, source, timestamp and author.

The aforementioned is the main object of the system and described the main form of data collected. A number of metadata and objects analyzing in depth each object is used including detailed information about the source, the author, accompanying multimedia and more. Figure 3 presents a generic schema of the database infrastructure that is used in order to support the essential for the system storage.

The data collected are stored on both an SQL-like storage environment as well as a noSQL environment. The hybrid scheme will help for storing elements for fast access in the noSQL nodes and collection of all the collected data in an SQL based structure for better interconnection between them and permanent storage of data with historic metadata [7]. Furthermore, a time-series database is used in order to keep track of the records that are stored in the database, including information about the source or the author. The latter is extremely useful when defining the rate of update for each source or the frequency of posting for authors and their relation to period and time.

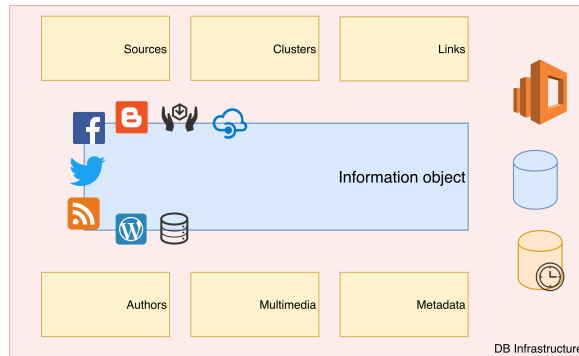


Fig. 3: Generic database scheme

3.2 System Core

The system core contains all the key elements and services of the system. It consists the basis upon which the complete system is designed and implemented. Each of the modules formulating the system core can act as an autonomous system providing endpoints for independent usage. These endpoints can also be used by the system internally in order to perform the physical interconnection between the different services.

The system core includes the following elements:

- **Named Entity Recognition (NER)**, which is a module for recognizing entities in bunches of text. A machine learning mechanism based on OpenNLP⁵, a set of language features and a set of annotated documents for finding candidate NERs, enhanced by the use of dictionaries is used [5].
- **Breaking news detection**, which is a component for recognizing important news topics. This is usually based on the number of similar articles produced in a period of time, but it should be considered that not all news topics are increased in numbers in the same manner. As such, machine learning algorithms are employed that are able to recognize breaking news based on the growth rate in time [6].
- **Clustering**, which is responsible for finding interconnections between the different entities. It should be noted that the objects collected derive from several different sources and the scope of this module is to create physical interconnections between objects having identical meaning. According to the definition of the object (without any attached metadata) the main scope of the clustering procedure is to interconnect conceptually two objects. Furthermore, as the system intends to operate regardless of the language of origin, the interconnection of the object should be language agnostic.

⁵ OpenNLP: a machine learning based toolkit for the processing of natural language text. <https://opennlp.apache.org/>

- **Classification**, which is a module for automatic categorization of objects to predefined categories. As the categories of the system are predefined, due to the fact that Palo is used as a news aggregation service, the categorization is done in several primal categories. The current mechanism will be enhanced in order to enable multilevel categorization including two different levels [4].
- **Sentiment Analysis**, which is responsible for extracting the polarity of the objects. A machine learning algorithm will be employed in order to replace a currently used algorithm based on the bag of words method [1].
- **Summarization**, which is responsible for extracting summaries out of the clusters of objects. As the clustering procedure evolves in time, the summarization procedure must adapt to changes that are done to the size of the cluster in time.
- **Trending topics detection and enrichment**, which is responsible for analyzing social media and open sources in order to detect topics that are trending and enrich them accordingly in order to detect their trends to other countries and languages.

3.3 High level analysis

The high level data analysis of the system includes a number of components that combine the outcomes of the deep data analysis and they include:

- Discovering social media influencers [2], [8]
- Applying cross-border analytics [3]
- Performing network analysis
- Exploring semantic means of the web [9]
- Simulating web and social media campaigns and measuring their impact

3.4 Frontend

The system frontend consists of both web and mobile applications that utilize the data collected and analyzed in order to present reports, visualize data and make it easy to explore the combined information.

The web and mobile applications will have a public part that will make parts of the collected available to public. This is a news aggregation service including rich media format of data as well as interconnection of information and multilingual content. The same is for the mobile application which can be formulated in order to enhance portability and usability of the presented content.

4 Expected outcomes and opportunities for cultural organizations

The design and development of the proposed system consists of a new and innovative product for the international market, which is expected to be the attraction

for many companies and organizations primarily organizations that operate internationally. The absence of specialized competitive products in this field offers a significant advantage and allows it to be a leading player in the Greek market, which is the country of origin, and to penetrate the emerging and demanding international market of high-volume data analysis technology by providing innovative services and products.

All the aforementioned, is expected to provide a new dynamic in the field of application development in the referred emerging sectors. This is achieved by using state-of-the-art technologies and methodologies together with the extensive knowledge in the field by the partnership. At the same time, within the framework of the proposed project, the know-how acquired in the areas of large volume analysis is fully exploited, thereby enhancing the company's policy towards the increased use of cutting-edge technologies, as well as the partnerships' research background. Finally, we should consider the valuable know-how acquired by all the participating bodies during the implementation of the proposed project through the two research organizations, which will be done by the research and development in order to achieve the desired objectives. The know-how to be transferred will improve all the organizations' and especially the company's scientific potential by increasing its knowledge and expertise and consequently the company's capabilities for future support as well as developing new applications and undertaking new research projects in the context of its activities.

Focusing on cultural related organizations it is possible to find opportunities that these venues never had. It is important to note that cultural spaces have recognized the important role of technology and online synchronous and asynchronous communication, and are willing to utilize modern and edge-cutting technological features in order both to enhance the experience of the visitors as well as attract a broader audience. In this scope, it is extremely difficult for people related to arts and culture perform an advanced step towards analyzing the impact of their presence and marketing procedures on the internet. Palo-Analytics project can play the role of the companion when it comes to their online presence. The project can help recognize the supporters and fans, can measure the impact of the online marketing strategy, can keep a record of other spaces' impact or connection and can help towards the improvement of the online presence.

5 Discussion

We presented the project PaloAnalytics, which is reaching its first year of undergoing. During this period the first crucial steps have been made, including the definition of the system use-cases, the formulation of the system architecture, the set-up of the system infrastructure, as well as the design and initiation of the first system components. Furthermore, the business-case is completed and the implementation of the first sub-systems is almost finalized. The infrastructure of the system is set-up and the means of integration are defined. An

interesting feature of the project is the participation of two research laboratories from two different institutions in Greece, which will join their research teams to produce the results of the project. In order to achieve the objectives of the project, cutting-edge technology and algorithms are used, which means that the participants will join forces towards the research.

Despite the fact that the actual outcome of the project is minimal compared to the algorithmic procedures that lead to it, a number of related research fields will be explored during the design and implementation of the components. First of all, data mining algorithms will be researched in order to produce the optimal solution for fetching data. Furthermore, the infrastructure that stores the data is the basis of the system and as such its design and integration is part of a research and development procedure. On the other hand, a number of algorithms and techniques including deep machine learning will be investigated in order to achieve procedures listing: clustering of data (including text objects deriving from social media), summarization of clusters, named entity recognition, sentiment analysis, aspect mining and breaking news definition. Furthermore, apart from the core algorithms, a number of “high level” procedures are required in order to achieve the complete set of project scopes. These include influencers mining, semantic web, network analysis, campaign impact, swot analysis and more, which are based on the metadata that accompany the information collected and processed.

It should be noted, that all the aforementioned are not just part of a research procedure; meaning that the research should not stand on the feasibility and soundness of the results. The system is a production based environment targeting large business and organizations, which can even test and formulate the procedures and the use-case scenarios. It lies on the ground of applied research and it is expected that all the implemented solutions will be able to endure large volumes of data, users and demanding procedures.

As far as the role that the system can play for cultural organizations it is clearly defined as an important one. Specifically we defined the system as a valuable companion that can totally alter the procedures of online marketing strategies and social media interactions. The system can be used to examine the behavior of the users towards exhibitions and presentations as well as towards individual cultural objects. The project can be the beginning of a new era in cultural informatics, acting as a novel pioneer procedure, that can involve edge cutting technologies directly on the relation of the organizations and visitors introducing a new way of mass culture.

Acknowledgment

This research has been cofinanced by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH CREATE INNOVATE (project code: T1EDK-03470)

References

1. Castellano, G., Kessous, L., Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction* (pp. 92-103). Springer, Berlin, Heidelberg.
2. Caridakis, G., Karpouzis, K., Wallace, M., Kessous, L., Amir, N. (2010). Multimodal users affective state analysis in naturalistic interaction. *Journal on Multimodal User Interfaces*, 3(1-2), 49-66.
3. Vlachostergiou, A., Caridakis, G., Kollias, S. (2014). Investigating context awareness of affective computing systems: a critical approach. *Procedia Computer Science*, 39, 91-98.
4. Varlamis, I., Tsirakis, N., Pouloupoulos, V., Tsantilas, P. (2014, October). An automatic wrapper generation process for large scale crawling of news websites. In *Proceedings of the 18th Panhellenic Conference on Informatics* (pp. 1-6). ACM.
5. Makrynioti, N., Grivas, A., Sardianos, C., Tsirakis, N., Varlamis, I., Vassalos, V., Pouloupoulos, V. Tsantilas, P. (2017). PaloPro: a platform for knowledge extraction from big social data and the news. *International Journal of Big Data Intelligence*, 4(1), 3-22.
6. Varlamis, I., Hilliard, D. F. (2017). Finding influential sources and breaking news in news media using graph analysis techniques. *International Journal of Web Engineering and Technology*, 12(2), 143-164.
7. Tsirakis, N., Pouloupoulos, V., Tsantilas, P., Varlamis, I. (2017). Large scale opinion mining for social, news and blog data. *Journal of Systems and Software*, 127, 237-248.
8. Margaritis, D., Vassilakis, C., Georgiadis, P. (2018). Query personalization using social network information and collaborative filtering techniques. *Future Generation Computer Systems*, 78, 440-450.
9. Bampatzia, S., Bravo-Quezada, O. G., Antoniou, A., Nores, M. L., Wallace, M., Lepouras, G., Vassilakis, C. (2016, September). The use of semantics in the CrossCult H2020 project. In *Semantic Keyword-based Search on Structured Data Sources* (pp. 190-195). Springer, Cham.
10. Levy, A., Rajaraman, A., Ordille, J. (1996). Querying heterogeneous information sources using source descriptions. Stanford InfoLab.
11. Hand, D. J. (2006). *Data Mining*. Encyclopedia of Environmetrics, 2.
12. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*.
13. McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., Barton, D. (2012). Big data: the management revolution. *Harvard business review*, 90(10), 60-68.
14. Ghoshal, A., Swietojanski, P., & Renals, S. (2013, May). Multilingual training of deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7319-7323). IEEE.
15. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., ... and Sainath, T. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29.
16. Violos, I., Tserpes, K., Varlamis, I., Varvarigou, T. (2018). Text classification using the n-gram graph representation model over high frequency data stream. *Frontiers in Applied Mathematics and Statistics*, section Mathematics of Computation and Data Science Journal. doi: 10.3389/fams.2018.00041