NORMASEARCH: a Big Data application for financial services.

Ylenia Maruccia[1] and **Giovanni Pansini**[1] and **Gloria Polimeno**[1] and **Felice Vitulano**[1]

Abstract. In the recent years banking and financial markets are trying to learn how Big Data can help to transform their processes and organizations, improving customer intelligence, reducing risks, and meeting regulatory objectives. The collection and the analysis of new legislations, understanding if they are introducing new aspects with potential impacts on different fields, could be the basis of a system able to give support in the strategic decision making process and to evaluate the potential impacts on both management and strategic activities. Here we want to present *NormaSearch*, a Big Data application developed by Exprivia, an international leading company in Italy in the process consulting, technology services and information technology solutions. NormaSearch is able to analyse specifical information taken from the web, both in a structured and unstructured form, and its application in the financial fields.

1 Introduction

The last years have seen a continuous increase of data generated in many fields, from science to social life, passing through industries, which we refer to as "Big Data". Each actor, e.g. individual, administration, organization or business, is a producer of new forms of data, both in structured or unstructured form: they can be personal data, conversations on social network, medical data, meteorological information, shared photos, and so on.

The challenge of both public and private companies is to easily manage these huge amount of data with improved technologies, different from the traditional ones, and extract knowledge from all these hidden information.

The main characteristics of Big Data are that they are too big, move too fast and do not fit the structures of traditional database architectures, so new technologies are necessaries to manage them. Moreover, one of the main difficulties, as aforementioned, is the format in which all the information are generated: in particular, they can be in a structured, unstructured or semi-structured form. So new forms of databases, programming languages and hardware architectures are used to either store Big Data or to transform it from unstructured or semi-structured format into a well-structured one, with consequences in many fields of application.

According to [1], Big Data help to better listen to customers, understand their ways of using services and hence the offer, simplifying also the decision making process. To this aim, an important role is played by those applications that tailor the information based on the needs of the customers. Technologies such as Recommender System are now used by many brands with the aim of suggesting products or services which a user may be interested in.

[1] Exprivia Spa, Bari - Italy, email:{ylenia.maruccia, giovanni.pansini,gloria.polimeno, felice.vitulano}@exprivia.it

In these years banking and financial markets firms are continuing to learn how Big Data can help to transform their processes and organizations. In particular, for banks, Big Data initiatives predominately still revolve around improving customer intelligence, reducing risk, and meeting regulatory objectives.

Machine learning techniques, for example, can be applied within the fraud and risk sectors, improving models and allowing acceleration towards more real-time analysis and alerting. Finding new legislations, understanding what are the differences with the existing ones and/or if new aspects have been introduced, can be a very important challenge in this field of application, with the purpose of evaluating the potential impacts on both management and strategic activities and of giving support in taking those strategic decisions that could minimise potential costs.

Here we want to present our solution, called NormaSearch, developed in Exprivia to manage the data generated from different sources and coming mainly in an unstructured format, at the aim of adapting it in the banking system. In Section 2 it will be described the scenarios in which NormaSearch could be applied, while in Sections 3 and 4 it will be presented the application with its component.

2 Scenarios

The financial crisis and the speculative use of the derivative instruments has placed the reform of the derivative markets "Over The Counter" among the priorities of the legislature in terms of standard negotiation procedures, as well as more stringent rules pertaining to the capitalization of financial intermediaries:

- In terms of rules designed to standardise the trading of OTC derivatives, it has been promulgated different regulations, such as *EMIR/DOTT Frank Act* (European Market Infrastructure Regulation), that revived the role of the Central Counter-Parties (CCPs), with the aim of increasing transparency and reduce both the counterpart risk and the operational one (see Fig. 1).
- To ensure the soundness of the banking system, the Basel agreements require the banks of the leading world countries some limits about their operational activities, especially regarding the amount of assets which they have to equip themselves for their clients' protection, thus allowing the capitalization of banks (and, consequently, liquidity guarantees), to guarantee the operations - collection, financing an investment - put in place with customers.

Therefore, it can be deduces as today a Financial Intermediary is called to observe the dictates imposed by the regulations in the area of interest, involving adjustments to the operational processes and/or IT architectures, in compliance with regulations.

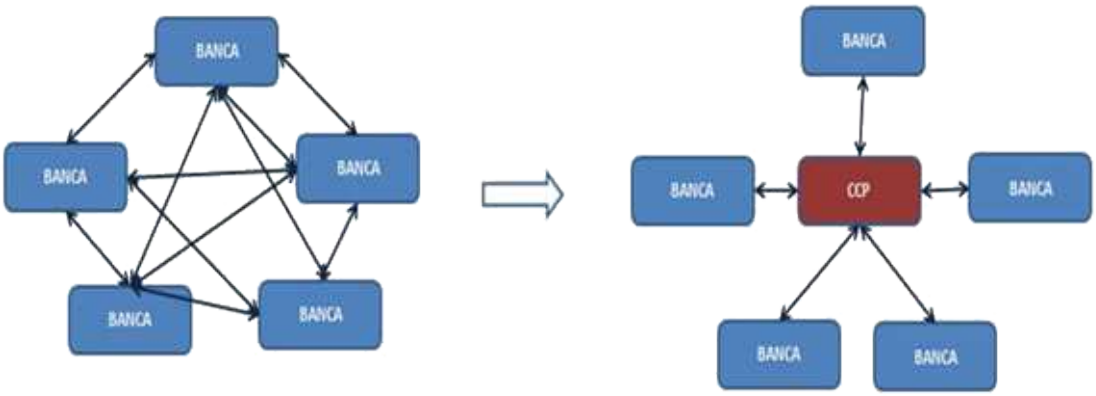

Figure 1. The role of the Central Counter-Parties.

The granularity and, at the same time, the complexity of these regulations, necessitate a constant attention and monitoring of them, in order to anticipate future changes, or integrations, or evolutions.

In this context it engages the idea of providing a machine learning tool which, through the analysis of the newly introduced legislation (or in the approval process) and/or the changes in the requirements previously promulgated (detectable by special certificates internet sites), may provide guidance on the bank process involved and, therefore, indicate with almost predictive function the impacts on the IT application, in terms of changes or new implementations, in support of the above processes.

To this aim, we developed a Big Data application that is able to analyse specifical information, both in a structured and unstructured form, taken from the web. This application, called *NormaSearch*, is described in the next section.

3 NormaSearch Functionalities

As aforementioned, NormaSearch is a Big Data application that allows to browse the web, search and analyse specifical information on the bases of given rules that are defined by the user. The application has two distinct sections, one of Administration and one of Fruition: the first allows to train to recognise and classify the information of its own interest through a series of examples (weakly supervising training), while the second allows users the analysis of the sites collected independently by the system. Once trained, the application allows to analyse web pages and documents in the sites, news group, blogs, forums and so on, according to a process specifically designed for linguistic and conceptual analysis of online content.

This process allows to achieve an optimum precision to coverage ratio in the search, as well as to limit in an important way the amount of downloaded web pages and, consequently, the hardware resource consumption.

The application operates their own research in an incremental way: by doing so, the pages are presented to the user only in the case in which they have never previously been recognised or, in the case where the content is changed. In details, the application is therefore able to:

- Refer autonomously a set of sites, blogs, forums and so on, looking for info about a set of concepts of interests identified by the machine training activity by examples (weakly supervised training); the system is able to consult a set of predetermined sites (authoritative sites) or even the whole www.
- Identify, in every web page retrieved, the individual portions of text (HTML page section) in which are expressed the sophisticated concepts, by associating a percentage indicator of relevance to such concepts with each section identified.
- Automatically classify and organize web sites and pages that belong to them according to a predetermined conceptual taxonomy or derivable during the training phase machine.
- Filter, as needed, specific types of web sites that tend to generate noise, such as for example search engines based on search engine spamming techniques.
- Identify only new content found on each new consultation.
- Present the results through a simple web interface or as a report directly downloadable from the interface; reports can also be sent from the application via e-mail.
- Independently identify potentially authoritative sites and recognise inactivity of authoritative sites.

4 NormaSearch Architecture

In Fig. 2 it is shown the architecture of this application.

It is made up of two main components, a client and a server ones, both described below.

NormaSearch Client. It is specialised on the interaction with the user and the transmission of user requests to the server component. It is structured into two main parts:

- A fruition console (in Fig. 2, Retr.UI). Here the user can manage documents and decide which of them have to be processed, or could be useful for the launch of new experimental projects on specific themes, or dismissed.
- An administration console (in Fig. 2, Admin. Ui). Here the user can define the security rules, the loader dedicated to the web monitoring and to the retrieval of documents of interest, the definition of categories and subcategories of the safety rules through which the conceptual framework of the rule itself is defined in terms of topics and the organization of them, the training of the security rules and the related conceptual categories. Moreover, here it can be also specified new projects where where the user can insert additional or parallel categories to the security rules established

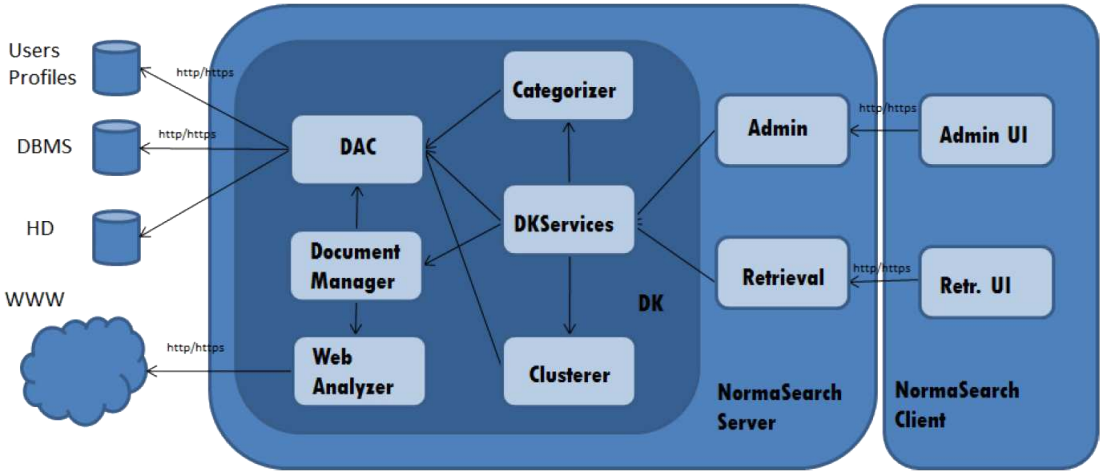

Figure 2. Norma Search Architecture.

above, in order to look "cool stuff" that can be used as a starting point to expand the research in the field of banking regulation or to add/modify existing safety rules.

NormaSearch Server. It is specialized on the receipt of the users' requests and the sorting of the same to the server components, these ones suitable of taking charge of specific requests. As NormaSearch Client, it has two components: the administration component and the fruition one, designed to manage requests from the administration and fruition consoles, respectively, and send them to the specified server components.

At the moment NormaSearch is in productive use by an Italian bank. Moreover, in NormaSearch Server there is an important software component, called Big Knowledge [4] , developed by Exprivia.

Big Knowledge is able to manage both structured and unstructured data and, as it can be seen in Fig. 2, it is made up by six main components:

- **DAC** The Data Access Component is the centralized component to access data, either in a DBMS or, in the BigData cases Solr.
- **Document Manager** It is the component through which BK is able to convert the document provided in textual form, and clean it from useless portion of text (e.g. html banners).
- **Information Extraction Manager** It annotates the document, extracting relevant information like Named Entity Recognition by the usage of Finite State Automata, and elements inside custom gazeteers.
- **Clusterer** It is dedicated to the extraction of conceptual groups (clusters). These ones are automatically extracted using advanced techniques of NLP (Natural Language Processing) based on Latent Semantic Analysis[2] (LSA) and a Markov clustering algorithm[3]. The generated clusters are crucial for the training tuning of the system.
- **Categorizer** It is intended for the automatic classification and organization of the document according to a conceptual taxonomy expressed as a set of clusters and constructed manually or semi-automatically by describing the category with a small text.
- **Geo Recognizer** Using a gazeteers of places kept from Geonames[2], it annotates the documents gathering: nations, regions, cities, airport, port and generically geographic points of interest. A kml export of single of multiple document, is provided when a WMS[3] compliant system is integrated.

5 Conclusion

Big Data are changing the industrial world and, for this reason, all kind of companies need to be capable of managing this huge amount of data and to extract useful information from them. This great interest in Big Data is present also in the financial services, which can obtain important information from the analysis of both structured and unstructured data. It is also important to have these information in an useful time, in order to prevent losses and to predict important event before they happen. In this paper we discussed a Big Data solution in financial service field developed by Exprivia. It is called NormaSearch and it aims at predicting the impact of a legislation change, or the introducing of a new one. After a brief introduction about Big data and machine learning technologies, we described NormaSearch and its components and how it works. It can be analysed the new introduced legislation and also provided a guidance on the bank process involved. It can be predicted, in particular, the impacts on the IT application in support of that bank process. Moreover, in this paper it has been described an important Big Data solution that is present in NormaSearch. It is called Big Knowledge and it is composed by six components that speak together in order to manage all the documents in input and extract important information that can be then classified. This paper showed how a Big Data solution can be useful in financial field and can predict important information in an useful time in the strategic decision making process.

REFERENCES

[1] C.L.P. Chen and C.Y. Zhang, 'Data-intensive applications, challenges, techniques and technologies: A survey on Big Data', *Informatics and Computer Science Intelligent Systems Applications*, **275**.

[2] http://www.geonames.org/
[3] http://www.opengeospatial.org/standards/wms

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science*.
[3] Stijn van Dongen, *Graph Clustering by Flow Simulation*, PhD thesis, University of Utrecht, 2010.
[4] F. Vitulano, M. Cammisa, and Y.Maruccia, 'Unleashing big data power for sea emergency control', in *Proceedings Tethys 2015*.