Exploiting the Diversity, Mass and Speed of Territorial Data by TELCO Operator for Better User Services

Fabrizio Antonelli
Telecom Italia
antonelli.fabrizio@telecomitalia.it

Antonino Casella
Telecom Italia
casella.antonino@telecomitalia.it

Cristiana Chitic
Telecom Italia
chitic.cristiana@telecomitalia.it

Roberto Larcher
Telecom Italia
larcher.roberto@telecomitalia.it

Giovanni Torrisi
Telecom Italia
torrisi.giovanni@telecomitalia.it

1. INTRODUCTION

The Semantic and Knowledge Innovation Lab (SKIL) is an innovation lab opened by TELECOM Italia as a result of its partnership with the European Institute of Technology (EIT). The lab is located in Trento which hosts one of the 6 Information and Communication Technologies Lab of the EIT and two reference research groups: University of Trento and the Fondazione Bruno Kessler (FBK) center. The aim of the lab is to exploit the potential of big data and open data by creating a platform to share data among academia, industrial and institutional partners. It is our belief that such raw data gives a new image of the territory/community from which the data was collected and by analyzing it new trends, enriched information can be extracted. The results of the analysis can turn into revenue and profit not only for the companies but also for the communities.

2. DATA USAGE

The huge amount of data in our world is generating value in many domains such as public sector, retail, telecommunications, manufacturing, etc. For instance, the impact of public open data on the European Union economy is estimated to add a 40 billion € value annually. This boost comes as a result of cutting down on public expenditure and increased tax income from applications based on open data.

Enterprises are generating an increasing volume of information. They realized that this data could be a source of revenue and an important factor of production, alongside labor and capital, not just a storage problem or a side effect of their activities. Companies and organizations can obtain additional by making information available at higher frequency allowing the potential of big data to be materialized. Sophisticated analytics run on top of a variety of data sets (such data may come from inside or outside the company) can improve decision-making and the development of the next generation of products and services. Although the idea that big data drives economic growth is more and more spread among private sector, there are several issues that have to be addressed by the businesses in order to capture the full potential of big data. One of these issues is the lack of people with deep analytical skills to ask questions about the data and interpret the results. Moreover, organizations need not only to put the right talent and technologies in place but also create the workflows to optimize the use of big data. Most of the time, companies find hard integrating information from multiple sources especially from third parties. That is, the variety component of big data becomes a challenge for the companies.

Not only the private sector is unlocking economic value from data, but also governments and public authorities across the world are leveraging open data for economic benefits. The public authorities produce tones of data under the form of documents and statistics every year. Many governments have given support and incentives to public sector in establishing open data portals. Such governmental approaches enable an increase in business activity by allowing the creation of new firms, new products and services and by supplementing existing products. Also, the open data portals are leading to additional public sector revenues: namely charging from data and tax income from commercial usage of open data. However, very few public authorities are taking the right measures in achieving benefits from publishing their data: sharing comprehensive data that has both significant breadth and granularity, it is regularly updated, easy to use, enhanced search capabilities. Another factor that affects the success of open data programs is user participation. This participation is directly proportional to the amount of apps creation, discussion forums and blogs.

Many of the current big data initiatives are prone to fail due to lack of integrating data from heterogeneous sources into a flexible platform. The data will produce more revenue if combined with data from third party sources. Another overlooked factor is the enrichment of the data with semantics that in short terms seems unimportant but in long term it could create a common understanding of the data, derive hidden interpretations and questions about data, and additionally a sharable data format.

A simple example emphasizing this idea is the true story case of a small restaurant owner in Trento, Italy, which was interested in estimating the right amount of expecting customers to decide the required daily staff and tables. A wrong

Proceedings of the VLDB Endowment, Vol. 6, No. 11

estimation could have cost him money and resources. The solution he found given his situation was to call the customer service of the highway company Autostrada A22 and ask the number of barrier crossings at the Trento exit the same period of the previous year. In this way he was able to roughly estimate the number of people entering the city and the potential customers. This experiment shows that the same data in different contexts and under custom tailored analysis captures value for a variety of businesses.

3. TOLD PROJECT

Telecom Italia is one of the biggest telecommunication companies in Italy that due to the nature of its business produces large datasets every year. Any time a person is making a phone call (landline or mobile) a set of records (referred to as call-detail records) is stored: the phone numbers of the caller and the called person, the date and time of the call, the duration of the call, the cellular radio towers and the cell sites (land areas that are under the coverage of a cellular radio tower) from and where the call has been initiated and terminated as well as the cell sites that were crossed while the call has lasted, the respective zip codes and dispatcher centers in the case of landline calls, etc. Similar information is stored in the case of SMS exchanges. Also the information regarding the data traffic that is handled through the mobile phones is recorded: the client id and phone number, the connection plan that the client is using, the duration time and the volume of data downloaded per connection, if any, the cell sites entered/exited in the process etc. Traditionally individual departments of Telecom analyze these internal datasets targeting certain business interests of the company: improve the cellular network, produce statistics and billing papers for its own customers, etc.

Telecom Italia realized that investing in research and innovation groups that leverage data-driven strategies to innovate and capture value from massive amount of batch and up-to-real-time information is going to be a boost for the company itself. Moreover, as in the case of the restaurant owner in Trento mentioned before, integrating the telecommunication data (i.e the anonymized call-detail records) in conjunction with other territorial and private business sets of data drives growth by improving the decision-making process and the quality of its services and leads to the development of new products and services.

As a consequence, our lab and more specifically one of our projects called TOLD (Trentino Open Living Data) aims at the development of use-cases where our telecom data is analyzed with other data coming from external providers. One such external provider is the local energy supplier that provides data about the daily readings of the consumption meters and solar panels, the spatial information of the distribution lines and meters. Another example is the highway company (Autostrada A22) providing information about tollbooths, cars and highway segments. A third example is the public administration data: civil, financial, meteorological and geographical data published by the territorial or national open portals.

The reason the lab was founded in Trento is because the local government is extensively investing in services for its citizens and our findings can serve as pilots for other territories in Italy.

We have various scenarios that we are working on. One of them is the provenance of the tourists: based on the phone calls made or received in a specific area (city, points of interest etc.) we can infer the provenance of the tourists in that area. Since the call records used in the estimation process are anonymized, we tried different statistics that will measure as accurate as possible the interaction between the different areas through the phone call traffic. We built another use-case to validate the statistics on the touristic flow computed in the previous scenario using only telecom data. More specifically, we analyze the daily energy consumption of the distribution lines that are powering the ski lifts of the territory and the daily phone traffic encountered by the cell sites over these areas. The results show a high correlation between the two different datasets. Another scenario is to provide a fairly good estimation of the energy consumption in the touristic areas by monitoring the phone call traffic of the cellular coverage over the different highway segments in conjunction with the dataset provided by the highway A22. Later on, this use-case will be extended to the adjacent roads that are crossing the touristic areas of the territory. This will provide insights on how people are moving over the territory once they exit the highway and how this movement affects the energy consumption. All the use-cases described so far are making use of open data: geographical maps of the territory, meteorological data and information regarding the opening and closing of the ski lifts.

The technologies we are using are mainly of-the-shelf technologies chosen based on the rapidity to implement the use-cases and derive fast the results of our analysis. We set up a Hadoop cluster on our cloud infrastructure (Nuvola Italiana). In this way, we are able to effectively and efficiently store large amounts of data. The data is vertically partitioned by day and topology of the data records, for efficiency using the DFS-Datastore library. The Cascalog is used to write our map reduce jobs and do the extraction, transformation and a first aggregation of the data. For more advanced analysis algorithms we are using RHadoop. To deal with the heterogeneity of the datasets coming from different sources we use some semantics [1] and spatial information using technologies provided by Oracle. We are currently investigating the capabilities that the Pivotal tools can provide to our project. In terms of visualizing the geo-location information and the results of the data analysis, we build our interfaces using Google Visualization API and experimenting aside with D3.js and ggmap-ggplot packages of R. The work of this project is in collaboration with the Data and Knowledge Management group [1] of the Fondazione Bruno Kessler and the Data Management group [2] at the University of Trento.

4. REFERENCES

[1] O. Hassanzadeh, A. Kementsietsidis, and Y. Velegrakis. Data management issues on the semantic web. In *ICDE*, pages 1204–1206, 2012.

[2] T. Palpanas and Y. Velegrakis. dbtrento: the data and information management group at the university of trento. *SIGMOD Record*, 41(3):28–33, 2012.

[1] https://dkm.fbk.eu