Alma Mater Studiorum · Università di Bologna

Physics and Astronomy Department
PhD Thesis in Applied Physics

Implementation and optimization of algorithms in Biological Big Data Analytics

Supervisor:

Prof. Daniel Remondini

Correlator:

Prof. Gastone Castellani
Prof. Armando Bazzani

Presented by:

Nico Curti

Session 2019/2020

Chapter 1

Biological Big Data - CHIMeRA project

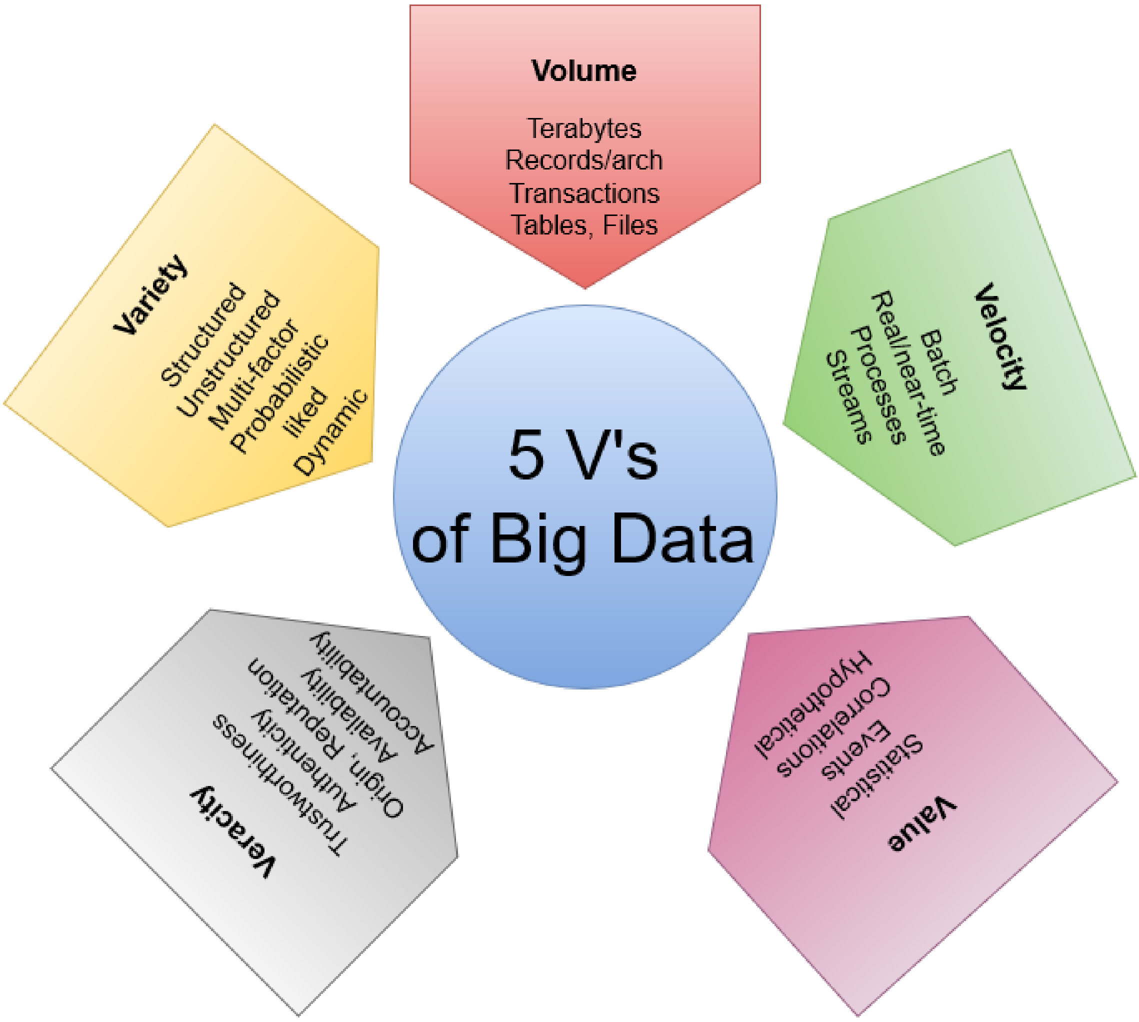

Figure 1.1: Big Data 5 V's

Every second a large quantity of data are produced and shared through Internet and Web-pages. Data are collected by social networks, messages, video streaming and images. Everyone, in fact, can easily create new data sources and share or put them in Internet pages. The growth of this data is not limited to multimedia data but it involves many different fields. This is one of the most important feature of the contemporary time, the so-called Big Data era: this huge volume of data has created a new field in data processing which is called Big Data Analytics that nowadays positioned among top ten strategic technologies (Gartner Research, 2012).

It is still difficult to provide a definition of what exactly are the Big Data and we can find many slight different nomenclatures and categories which aim to formulate it. Moreover, Big Data does not define a particular data type but more than we normally think sources can be labeled as it. The *International Journal of Computer Applications* defined them as "[· · ·] a collection of large and complex datasets that cannot be processed and analyzed using traditional computing techniques. It is not a single technique or a tool, rather it involves many areas of business and technology". This definition certainly involves many aspects of Big Data processing but it does not provide any definition about their nature. Moreover, it is easy to identify them as "big" and thus difficult to analyze, but they are all around us every day and just using the Internet connection every smart-phone or laptop can extract and visualize our web queries so could be not properly correct to define them in this way. However it is certainly sure that the standard computing techniques have to be reviewed to face on this vast amount of data and a even more important attention has to be payed on the algorithm implementations.

While a global definition of them is evidently difficult we can however describe them using some of their "essential" features. One of the most common and used set of labels is given by the so called 5 V's of Big Data: volume, velocity, variety, veracity and value. Despite the first twos are quite obvious (the Big Data are certainly *big* in volume and they are produced very *fast*), the remaining three need a particular attention. Moreover, we have already treated problems about the volume of data (ref. Chapter**??**) and the need of very fast processing and algorithm optimizations (ref. Chapter**??**). Now in this chapter we want to focus on the remaining three characteristics of Big Data Analytics.

As pre-announced there many different sources able to provide data and this feature describes the extreme heterogeneity and variety of them. We can however broadly classify this variety into three global classes: *structured data*, *semi-structured data* and *unstructured*. A dataset is *structured* if we can easily manage the informations in it or, in other words, if it is described using the standard formats of data and thus can it can be "quearable". On the other side we have the completely *unstructured* dataset in which the data are disorganized and we need one or multiple pre-processing steps before use them or we have to develop different techniques to handle them. The intermediate step is given by the *semi-structured* data in which only a part of them could be handle with standard techniques or can be easily brought to their structured version. The data organization has been a crucial task on this work of thesis and we will return on this topic in the next sections.

The fourth essential characteristic of Big Data is the *veracity* of them due to data inconsistency and incompleteness. The data are shared very fast using Internet and there could be found some ambiguities and/or deceptions between different data sources. If we want to merge and aggregate different kind of informations we have also to face on this kind of problems. The final task of every Big Data Analytic application is, in fact, to process these large quantities of data and obtain a unique answer which can not vary in relation to the portion of data or dataset used.

The last and probably most important feature is certainly their *value*: it is certainly good to have access to large amount of data but unless we can turn them into value they are useless. In this vast amount of data only a small part of them can be considered as informative and it is always harder to extract this core of informations from them. Moreover, we have to take into account also the difficulties about the management of these data and their more or less complex structure. However, also in this case, it is hard to generalize this property on all the amount of data contained in Internet: every day we see a large quantity of useless information on the web and it is hard to figure out that some of them can be useful for research applications. A key role is played by the *questions* which we ask: for every data source there is always an appropriated question which can be answer using it and vice versa. In this way also the seemingly useless datasets can acquire importance for an appropriated research project.

In this chapter we will discuss about one of the latest project developed during my PhD and which is still in work in progress: the CHIMera (*Complex Human Interactions in MEdical Records and Atlases*) project. The project is founded on Big Data applications and it is accidentally born as separated branch of the INFN FiloBlu project (ref. next sections and Appendix E for further informations about the project) which financed my last PhD year. CHIMeRA aims to create a unified database of bio-medical records using Natural Language processing techniques. The final purpose is to merge multiple data sources available on-line into a single network structure which highlights the relevant interactions between bio-medical informations, i.e starting from diseases to the biological agents involved into their causes and consequences. The realization of the first version of CHIMeRA required a lot of time and the development of novel pipelines of data processing. The project does not still achieve its conclusions but in this chapter we will cross through the main key points which allowed its construction.

1.1 The CHIMeRA project

The increasing availability of large-scale biomedical literature under the form of public on-line databases, has opened the door to a whole new understanding of multi-level associations between genomics, protein interactions and metabolic pathways for human diseases via network approaches. Many structures and resources aiming at such type of analyses have been built, with the purpose of disentangling the complex relationships between various aspects of the human system relating to diseases [3, 1]. All these data come from different kind of studies performed by independent research groups who want to prove their theory about a particular aspect of biological agent interactions. Modern biological analyses perform very capillary studies on biological agents: in this way we can easily study the deeper relationships between them but we completely ignoring what are around them. This approach is certainly extremely efficient for the detection of the minimal causal agents of the problem but it tends to loose the global and complex[1] environment and prospective in which the process occurs. This is also the main problem of complex systems, i.e a system composed of multiple components with a mutual interactions between them. The study of an individual aspect, in fact, could give us only a partial overview of the system but we have to take in account the interactions between the multiple components for a global description.

The network structures are acquiring even more importance on this kind of studies. Complex System and System Biology researches have proposed multiple models about the dynamical and evolutionary interactions of the human system agents aiming to study the hidden relationships between them using network models. A network model, in fact, is able to highlight and quantify the non-trivial correlations between the system components. The main problem of this kind of approach is certainly the increasing dimensionality of the involved data: a network model could be described via its adjacency matrix, i.e a matrix $(N \times N)$ in which each row/column identifies an agent/node of the underlying problem and each entry (i, j) quantifies the importance of the interaction between the agent/node i and the agent/node j. In real data applications we can often reasonably assume that a wide amount of the matrix entries are null, i.e the interaction between the involved agent is

[1] From a physical point-of-view.

quite sparse, and thus we can used the important properties related to the sparse matrices to manage our network. However, when the amount of data increase also the management of a such sparse matrix could be difficult. More efficient solutions are provided by the modern Database formats and languages (e.g MySQL, SQLite, InfluxDB, $\cdots$) which store all these informations into a binary format and they allow to submit queries to extract the desired portion of data. A global visualization of these huge amount of data is, in fact, without practical-sense and none valuable informations can be extracted from the global representation of the system. The most important feature of network model is, in fact, the definition of a hierarchy of the interactions: the relationship between two nodes is given by the amount of connections which link them or, in other words, by their path. Starting from a node its nearest neighbors are given by the set of nodes connected to it: re-iterating this concept we can explore all the network structure[2]. In this way we can study the interactions of each node at different precision orders and causalities.

In light of these considerations we started to develop the CHIMeRA project (*Complex Human Interactions in MEdical Records and Atlases*) in which we aim to merge the state-of-art studies and databases about biomedical researches into a unified network structure. A key role on our network structure is played by the diseases: the major part of biomedical researches are focused on causes and consequences of a given disease and thus the corresponding databases involve the interactions between them and other biological factors. The diseases are also the most bigger manifestations of biological malfunctions and a large part of biological researches are financed on their study looking for their fine grain causes. Thus a disease could be a valid “bridge” between multiple data sources: in each database we can find the associations between a disease and a series of multiple biological aspects related to it which can be merged together using disease nodes.

The crucial point of this project was, in fact, the merging of different kind of informations provided by multiple distinct data structured. As told above, the major part of researches have focused on a partial aspect of the problem and they provide an independent result from the others, reducing the possibility of interactions between the outputs. Moreover, a lot of time is always spent for the creation of a practical visualization of the results using web pages and on-line services which drastically affect the real usage of these informations when we want to combine multiple sources. The CHIMeRA project started from these independent sources and it aims to maximize their overlap and thus the communications between them.

A final attention has to be payed about the format of these data: in physics we are friendly with numerical data but in these context we have to work with words. The above told databases include only the research outputs and thus the performed interpretations of numerical data studied. For example, if a numerically significant correlation was found between a disease and a gene we would found an association between them into a database and thus we can model it as a link. The only information available into this database is a link between two words, the disease name and the gene name, without any numeric value. While numbers have a unique representation (the number 2 is always 2) we can use multiple periphrases, i.e set of words, to identify the same concept. The biomedical community, in fact, has not yet provided a unified standard for disease identification or, at least, it has not yet provided a rigid standard as for other kind of data as genes or SNPs. So, if the diseases could be an efficient way to link together multiple data sources since all the studies prove correlations between them and other biomedical informations, they have the payback of an extreme variability in their nomenclature. The CHIMeRA project tried to overcome this issue using a Natural Language Processing (NLP) approach.

In the next sections we will discuss about the multiple steps which bring us to the formulation of our unified CHIMeRA database. We will start from the preliminary studies

[2] We assume that our network structure has not isolated nodes and undirected connections.

performed into the INFN FiloBlu project which allow the creation of the SymptomsNet structure, i.e a "small" network based on Italian words which connects diseases to related symptoms. Then we will briefly introduce the most common NLP techniques, also used into the CHIMeRA pipeline and finally we will show the main developed features of CHIMeRA.

1.2 SymptomsNet

The INFN FiloBlu project was developed by the collaboration between the Physics Department of the University of Bologna and the INFN group of the Sapienza University of Rome. The project aims to implement a NLP pipeline to process messages with a medical theme. The critical issue of this project is about the message language: the project was financed by the Lazio region and it was developed to work at the Sant'Andrea Hospital of Rome so all the project involves Italian words. This constrain drastically affect the data availability which are very hard to find on-line. We tried to overcome this issue with a synthetic generator of phrases. To this purpose, a large set of medical words have to be provided and a valid interactions between symptoms and diseases can be useful to create reasonably phrases. The development of the NLP pipeline goes out from the scopes of this thesis (see Appendix E for further informations about our contributions about this task) but our first contribution to the INFN FiloBlu project concerned the realization of an Italian disease ontology. If we have a sufficient disease ontology we can use it to train our synthetic generator with reasonable phrases.

The English is becoming the predominant language in the research community and it is really hard to find (enough) data in other languages: everyone who wants to share his data and informations via Internet have to provide them in English if he wants to increase its visibility and availability. The Italian constrain posed by the project drastically limits the data sources and no public database was found. We would stress that as database we consider a public available set of structured data which can be downloaded and easily used.

Surfing on Internet many web pages can be found about diseases and their interactions with symptoms and causes, the so-called *on-line doctor*[3] (or Medical Services) pages. An on-line doctor is querable Internet service which allow to provide an auto-diagnosis of the user based on the information provided. The validity of the informations inside these tools is only partially guaranteed by the service provider and thus it can not be considered as a scientific method for medical diagnosis. However, the amount of informations collected by these applications is very interesting and it can be used to simulate reasonable medical queries, needed by our projects. Also in this case is important to notice that despite the availability of these public informations the data are structured according to the web page needs and moreover there is not an immediate download of the raw data.

So, how can we obtain these useful informations and re-organize them into a structured data format? The answer is given by the web-scraping techniques. With the term web-scraping we identify the wide set of algorithms aim to extract the information from a website, or more in general from the Internet. All the Internet pages are intrinsically pieces of codes written in different programming languages (HTML, PhP, ·). The major part of websites are written in HTML, an extreme verbose language, with more or less JavaScript supports. Write a code can be compared to an art form and the way chosen to reach the desired result is left to the programmer: in these way we do not have a rigid standard (excepted by the programming language constrains) and under each website underlies a potential completely different ensemble of code lines. Thus, the realization of a web-scraper

[3] Famous English applications are SteadyMD, MDLIVE, Sherpaa, LiveHealth Online and so on. Each service provides slight different informations and the choice of the best one vary according to the user needs.

poses several issues to the programmer who has to find the underlying patterns inside the web page to get the information stored.

A web-scraping algorithm is made by a series of multiple step which has to be performed without automatically. First of all, the algorithm has to recognized the unique website structure which can be summarized as the parsing of the underlying HTML code. Inside the large amount of code lines[4] are stored the set of informations useful for our application. So, the algorithm should be able to detect the relevant and interesting part and filter them. At this step we can easily reorganize the informations into a usable data format and save them.

There are multiple way in which all these steps could be performed and multiple open source libraries provide user friendly interfaces for the creation of own web-scraper. The most common one (and also used in our applications) is the BeautifulSoup [2] Python package. This package provides a very powerful Python library designed to navigate and read website source codes. The integration of this library with other pre- and post-processing techniques allow the extraction of the desired informations from websites and moreover their reorganization into a structured and usable data format.

Using the Italian version of public *on-line doctor* websites we can obtained the needed informations. We applied a set custom web scraping pipelines[5] to the different web pages to extract the medical informations and we mainly focused on sites which highlight the relationship between symptoms and diseases. As discussed above the Italian data sources are quite fewer than the English ones so only three web pages were involved into our analysis: My PersonalTrainer, SaniHelp and Sapere.it. All these three sites provide an organized series of tables which associate a disease to the corresponding symptoms. These databases are certainly not robust from a scientific point-of-view and their vulnerability is shown also by a non-rigid labeling of the two classes: in multiple cases we can find a disease as symptom of an other one and in many cases there is not a perfect agreement between the three data sources.

1.3 CHIMeRA query

1.4 Data extraction - Web scraping

[4] Very large if we consider a pure HTML web page.

[5] We would stress that the extremely variety of the websites requires an equally varied set of web-scraping algorithms. Thus for each web page taken into account an appropriated web-scraper was developed.

Bibliography

[1] C. A. e. a. Hidalgo. A dynamic network approach for the study of human phenotypes. *PLOS Computational Biology*, 5(4):1–11, 04 2009.

[2] L. Richardson. Beautiful soup documentation. *April*, 2007.

[3] X. Zhou and et al. Human symptoms–disease network. *Nature Communications*, 5.