
Kolici, V. [et al.] (2014) Scalability, memory issues and challenges in mining large data sets. *2014 International Conference on Intelligent Networking and Collaborative Systems: IEEE INCoS 2014: 10–12 September 2014, University of Salerno, Salerno, Italy: proceedings.* [S.l.]: IEEE, 2014. Pp. 268-273 Doi: http://dx.doi.org/10.1109/INCoS.2014.50.

Kolici, V. [et al.] (2014) Scalability, memory issues and challenges in mining large data sets. *2014 International Conference on Intelligent Networking and Collaborative Systems: IEEE INCoS 2014: 10–12 September 2014, University of Salerno, Salerno, Italy: proceedings.* [S.l.]: IEEE, 2014. Pp. 268-273 Doi: http://dx.doi.org/10.1109/INCoS.2014.50.

Scalability, Memory Issues and Challenges in Mining Large Data Sets

Vladi Kolici
Polytechnic University of Tirana
Tirana, Albania
Email: vkolici@fti.edu.al

Fatos Xhafa
Universitat Politècnica de Catalunya
Barcelona, Spain
EMail: fatos@lsi.upc.edu

Leonard Barolli
Fukuoka Institute of Technology
Fukuoka, Japan
EMail: barolli@fit.ac.jp

Algenti Lala
Polytechnic University of Tirana
Tirana, Albania
Email: alala@fti.edu.al

***Abstract*—Data mining is an active field of research and development aiming to automatically extract "knowledge" from analyzing data sets. Knowledge can be defined in different ways such as discovering (structured, frequent, approximate, etc.) patterns in data, grouping/clustering/bi-clustering data according to one or more criteria, finding association rules, etc. Such knowledge is then fed-back to decision support systems enabling end-users (actors) to make more informed decisions, which in economic terms could lead to advantages as compared to traditional decision support systems. It should be noted however, that data mining algorithms and frameworks have been proposed prior to the "Big Data" explosion. While data mining algorithms have considered efficiency and computational complexity as an important requirement, they did not take into account features of Big Data such as very large size, velocity with which data is generated, variety, etc. On the other hand, these features are indeed posing issues and challenges to data mining algorithms and frameworks. In this paper we analyse some of the issues in mining large data sets such as scalability and in-memory needs. We also show some computational results pointing out to such issues.**

Keywords: Data Mining, Large Data Sets, Distributed Data Mining, Hadoop, Map Reduce, Scalability, Memory.

I. Introduction

Data mining is about automatically (or semi-automatically) extracting knowledge from the data. The "knowledge" to be found/discovered from the data can be defined in different ways such as finding (structured, frequent, approximate, etc.) patterns in data [3], [6], association rules [1], [2], [8], grouping/clustering/bi-clustering data according to one or more criteria and many more [5]. In any case, such knowledge should be meaningful and useful to problem solving and should shed light on the phenomena (e.g. a hypothesis) under study. In most cases, the knowledge is fed-back to decision support systems enabling end-users (system actors) to make more informed decisions, which could lead to advantages as compared to traditional decision support systems. Nowadays, many business intelligence applications rely on data mining as powerful and sophisticated techniques to discover not only straightforward information but even complex patterns, user behaviour, rare events, etc.

The data sources to be mined can be diverse, while most data mining techniques has been applied to databases as well as to data warehousing, given that the data stored in data warehouses are easier and more efficient to mine than mining data spread across multiple databases, hosted on different physical networks.

Data mining is a rather new research and development field, although its root go back to a few decades ago as statistical analysis can be seen / is part of mining techniques. A plethora of data mining techniques, tools and frameworks have been developed and are in production to date. However, with the fast development of Internet, data mining techniques and frameworks are being challenged! What has changed? The answer is the target of the mining: the data! The changes to data sources have occurred in a number of ways:

- *Scale*: databases storing hundreds of millions of records (e.g. online transactions of big companies such as Amazon).
- *Variety*: data stored in databases is richer comprising text, images, videos, transactions, contextual data, etc.
- *Data acquisition*: computer-based acquisition of data is dramatically changing the scale of the data and its variety. Indeed, computerised accounting records about transactions, mobile data, monitoring systems, etc. are continuous data sources.
- *Data observations*: within the data users are interested to handle (1) many cases of interest (e.g. looking at transactions, products and costumers) and (2) online observations (e.g. spotting online failed transactions, fraudulent transactions, "rare" events, i.e. finding needles in haystacks, etc.)

The above data features, commonly known nowadays as "big data" are posing many issues and challenges to

data mining techniques. Indeed, while data mining techniques were designed with the efficiency and computational complexity criteria in mind [7], *reliability* was not often considered. However, handling with hundred of millions of records, very large graphs from social networks of hundreds of millions of users, monitoring urban traffic of millions of cars every day, etc., require to ensure the mining process is reliable, that is, it scales with the *current* and *future* size of the data and that the process is reliable due very demanding needs for memory. In this paper we analyse such issues and challenges and bring some computational results of an empirical study with real data of a Virtual Campus to point out to them. In all, successful data mining very much depends on making the right assumptions and correctly designing the study of the data mining.

The rest of the paper is organized as follows. In Section II we briefly overview some concepts and features of Big Data. We discuss scalability and memory issues in mining large data sets in Section III, where we also refer to online data mining requirements. Then, in Section IV, we briefly discuss how such issues are being addressed by a new generation of distributed data mining frameworks running on Hadoop. We end the paper in Section V with some conclusions and remarks for future work.

II. Large Data Sets –Big data

Large Data Sets (also referred to as Big Data) refer to datasets whose sizes are beyond the ability of typical software tools to capture, store, manage, and analyse them. At present there is data being generated everywhere: from traditional domains of eScience (simulation data can take up easily to Terabytes), data repositories and Internet archives, server logs, data gathered from World Sensing (raw data ranging into the Terabytes scale), mobile data sets (estimated at about 600 Petabytes passing through mobile devices *per* month), exhaust data, enterprise transaction data (estimated global enterprises can have hundreds of millions of transactions per day!

What does really make the data "BIG"? The answer comes from the definition of the main Vs:

1) ***V**olume*: The size of the data (especially when combined from different sources) is large/very large for conventional computing systems.
2) ***V**elocity*: There is a continuous data growth. As an example, every year business data almost doubles; in fact, 90% of this vast amount of data has created over the last three years. Thus the continuous data growth is actually ITs biggest challenge!
3) ***V**ariety*: This feature refer to data heterogeneity, as there are heterogeneous data sources, data could come in structured/unstructured and can have various formats (text, multimedia,...)
4) ***V**eracity*: Data is checked against potential biases, noise, missing values, errors, etc. This feature usually requires pre-processing of data (data cleaning). The veracity is also related to*data validity*.
5) ***V**olatility* defined as *data life time*: how long will data be considered valid (for the worth of analysis) and how long should it be stored. As an example sensing data from a smart city application has a shorter lifetime as compared to sensing data of patients; the later however may have shorter lifetime than patient records.
6) ***V**alue*: The value is defined in terms of data quality, that is, useful "knowledge" that can be extracted from data, how can it support business processes, decision making, etc.

There is a growing number of examples of Big Data scenarios arising in every field of science and human activity. We briefly describe next a few of them.

Big data from businesses: Companies do have "mountains of data"! Data sets are considered as a new asset for enterprises (additionally to their labour force and capital) and are an important factor of production. For example, Shell, is already collecting up to a Petabyte of geological data *per* well using its advanced seismic monitoring sensors, and plans to use the sensors on 10,000 wells. Likewise, banking and administration sectors are using online systems to improve operational efficiency, reduce costs, reduce fraud and errors. Processing of this data has led to business analytics, and extended form of business intelligence.

Big Data and Genomics: Genomics aims to study of the complete genetic material (*genome*) of organisms (sequencing, mapping, and analyzing a wide range of RNA and DNA codes). Just the human genome has about 20.000-25.000 genes x 3 million base pairs leading to 100 Gigabytes of data. Processing of this data has led to development of next-generation genomics through reduced sequencing time and costs and is thought to be the basis of personalizing the Healthcare and development of a predictive medicine, that is, understanding each individual genome.

Big Data and Internet of Things: The Internet of Things is spreading out everywhere and thus Internet of Things Companies may be suppliers of Big Data and analytical software that can help extract meaningful information from the enormous flows of data coming from many large scale applications (smart grids, smart cities, smart buidlings, "smart world"...)

Big Data and Health Care: Big Data is expected to make a revolution in healthcare, where data sets to be stored and analysed go far beyond traditional patient records. For example, there is a growing interest in continuous patient monitoring at hospitals, care centres or even at home, leading to Big Data. In fact, remote patient monitoring is seen by large as potential way to reduce the burden to premature or long time patient institutionalization (patients in hospital or care centres) due to caring of elderly, and growing population of patients that require long term care (e.g. patients of dementia).

Big Data, Virtual Campuses & MOOCs: Virtual Campuses are a widespread form of online learning and teaching. For instance, the Open University of Catalonia (Barcelona, Spain) accounts for more than 50.000 students and all academic and administrative activity is performed online! Log data files recording students activity can range from 15 to 20 Gb daily. It is of course of interest to know whats going on in the Campus (in virtual classrooms) in order to improve the online academic activity and learning outcomes as well as to improve the usage of resources, efficiency and security of the Virtual Campus, etc. Log data files are in this context considered as an important source of information. This has led to definition of learning analytics (also referred to as educational data mining). One recent example is that of MOOCs (Massive Open Online Courses), which unlike traditional online courses, can admit an "unlimited" number of students (there are some examples of courses exceeding 100.000 registrants). Learning analytics here can be useful to students monitoring, developing personalized MOOCs, designing personalized learning paths, etc.

III. Issues in mining large data sets

Most data mining algorithms and frameworks were developed before large data sets and Big Data come into play. Therefore, although efficiency and computational complexity were concerns [2], [3], [8] in designing data mining algorithms, they did not take into account features of Big Data such as very large size, velocity with which data is generated, variety, etc., requiring scalable and reliable data mining to extract knowledge from amounts of data of unprecedented size.

Data-mining for large, complex data sets should therefore address the many Vs of Big Data described in Section II.

A. Volume of data

The large data sets do not fit into main memory for processing, besides the requirement for efficient access to data as data transfer could be infeasible. Recent approaches advocate the use of Data-as-a Service (DaaS) [9] in the Cloud to provide fast and transparent access to data independently of the persistent data layer and enable interactive exploration of large data sets. Altogether, current data mining algorithms may critically run short to address scalability for processing large data sets.

B. Variety of data

With the variety of the data the amount of features to be taken into account increases considerably. Consider for example contextual information. Context is inherently complex and the amount of features could be large to fully capture the features of various contexts. This has implications for both processing, memory and storage.

C. Response time

While for computing analytics and historical processing of data (in batch mode), the processing time can arguably be large, each time more end-users need faster responses, as this may have an immediate impact on the decision making. Indeed, finding patters in the data about failed transactions and their context or fraudulent transactions require response in almost real time. In some case even, end users can trade accuracy for time, that is, knowing less (e.g. approximate patterns) but much faster.

The good news is that, for some data mining algorithms, the above requirements can be satisfied by implementing distributed data mining versions and thus can cope with Big Data however issues and challenges of scalability, I/O operations and memory needs arise. The bad news is that many data mining techniques are inherently sequential and thus parallel and distributed computing can't help much.

D. Scalability issues

We used real data from the Virtual Campus of Open University of Catalonia, which offers distance education through the Internet in different languages. About 50,000 students, lectures and tutors from everywhere participate in some of the 30 official degrees and other PhD and post-graduate programs resulting in more than 600 official courses. The campus is completely virtualised, made up of individual and community areas (e.g. personal electronic mailbox, virtual classrooms, digital library, on-line bars, virtual administration offices, etc.) through which users are continuously browsing in order to fully satisfy their learning, teaching, administrative and social needs. All users' requests are processed by a collection of Apache web servers as well as database servers and other secondary applications, all of which are providing service to the whole community and thus satisfying a large number of users. For load balance purposes, all HTTP traffic is smartly distributed among the different Apache web servers available and each web server stores in a log file each user request received and the information generated from processing it. Once a day (namely, at 01:00 a.m.), all web servers in a daily rotation merge their logs producing a single very large log file containing the whole user interaction with the campus performed in the last 24 hours. The size of the log files keeps growing, due, on the one hand, to the increase in the number of users of the virtual campus, and on the other, to the variety of event information kept in the log data files. Currently a typical daily log file size may be up to 15-20 GB (about more than 50% increase in log data files size of 5 years before).

We used RDLab as distributed infrastructure [1] (see Fig. 1) for processing and analyzing the data.

We processed the data contained in the log files to find navigation patterns of online users, specifically chains

[1] http://rdlab.lsi.upc.edu/index.php/en/

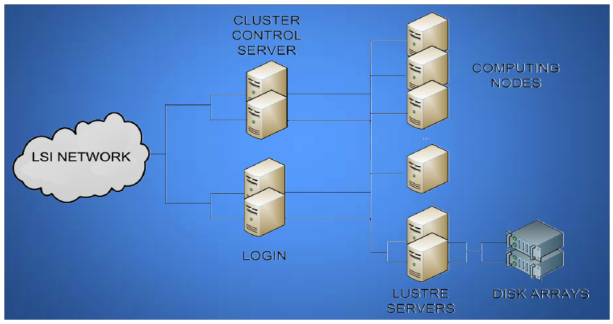

Figure 1. RDLab Cluster

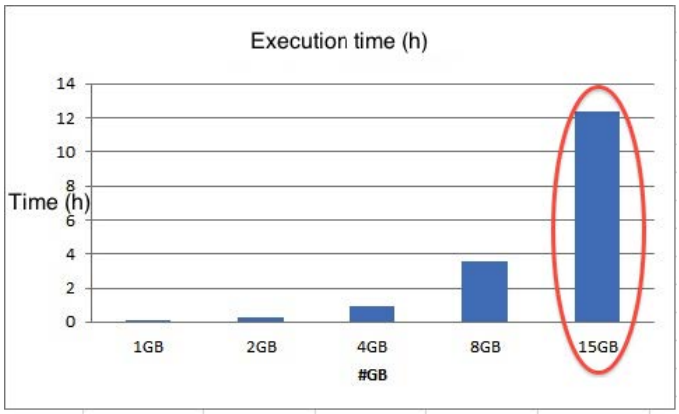

Figure 2. Execution time of log files of different sizes

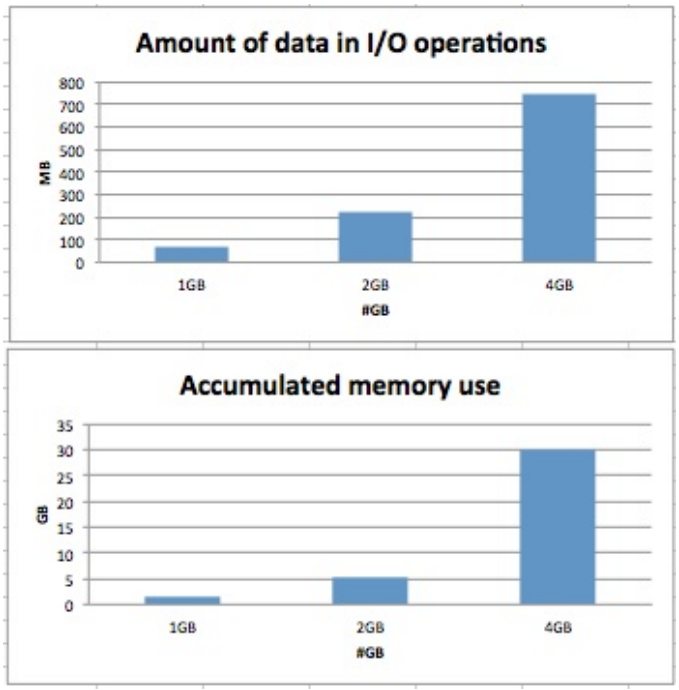

Figure 3. Amount of data in I/O operations (top) and Accumulated memory (bottom).

of access to resources in the Virtual Campus. We show the graphical representation of execution time according to different data size (in Gb) in Fig. 2.

As can be seen from Fig. 2, processing more than 8 Gb requires hours of processing time (for the case of 15 Gb, the processing in a single machine needed 12h of execution (CPU) time). This amount of time is obviously prohibitive for most real life application needs.

E. Memory issues

Another issue that arises when mining large data sets such as the log files of the Virtual Campus is the amount of data in I/O operations and the accumulated memory. For the case of k-Means algorithms, we observed an fast increase in both cases even for rather small log data files (up to 4 Gb; see Fig. III-E).

As can be seen from III-E, the amount of data in I/O operations grows at a rate of 4 times when log data size is doubled. In the case of accumulated memory an exponential growth was observed. As a matter of fact, processing of log files of more than 4Gb ran short of the memory, showing critical reliability.

The memory issue requires carefully analysing three possible options while mining large data sets, namely:

- *In-memory:* data is loaded completely into memory. This option provides the fastest access to data yet the amount of data that can fit into memory has clear limitations.
- *In-database:* the data access is not as fast as in-memory case and depends on the database server, the fact that database can be distributed, etc. The advantage is that large amounts of data can be stored.
- *In-Hadoop:* data access is fast, though depending on Hadoop cluster and HDF / HBase configuration.The advantage is that large amounts of data can be stored.

Some studies (with RapidMiner[2] software) have shown that in-Hadoop option stands in the "middle" of in-memory and in-database options regarding the data access, while can provide *unlimited* amount of data storage.

F. Issues from online mining of data streams

There is an increasing interest in *online* data mining. Indeed, many Internet-based applications generate data streams that could be of interest to mine, especially for patterns discovery. For example, online banking applications might be interested to detect in real time failed transactions, or monitoring in real time flights information globally, mining streaming of tweets in Twitter, etc. Differently from mining large data sets already stored in databases, data warehouses or (distributed) file systems, online data mining has to cope with the incoming flow of data.

[2]http://rapidminer.com/

The issues here however, as in the case of Big Data, are, on the one hand the *rate* at which data is generated and, on the other, the data variety (structured/unstructured, low/high volume, etc.) that can appear along the data stream. Window-based sampling, chain sampling techniques have been proposed to deal with the incoming flow of data. Memory issues arise if large window time are to be considered. Some empirical evidence was obtained by processing data streams from FlightRadar[3] (detailes are omitted here).

IV. Data mining frameworks running on Hadoop

After the MapReduce and Hadoop platform, a new generation of distributed data mining frameworks is emerging. Some of them, like Weka which is prior to Hadoop, are realising distributed versions that can run on Hadoop, while others, like Mahout, are developed anew to be fully distributed.

In this section, we briefly review some features of distributed Weka, Mahaout and R package.

A. Distributed Weka

Weka comprises a large collection of data mining algorithms, most of which are batch-based and operate on data held in main memory. The latest versions of Weka offer new packages for distributed Weka, namely *distributedWekaBase*, which provides base map and reduce tasks that are not tied to any specific distributed platform, and *distributedWekaHadoop*, which provides Hadoop-specific wrappers and jobs for these base tasks. Of course, not all algorithms can be distributed (especially a critical problem is for training models from large datasets).

B. Mahout

The Apache Mahout project aims to address scalability issues of data mining and machine learning algorithms. Thus, algorithms for clustering, classification and collaborative filtering are implemented on top of scalable, distributed systems. The limitation is that to date it only supports three use cases, namely, *recommendation*, *clustering* and *classification*.

C. R for Data mining and Hadoop

R is free software designed for statistical computing, which includes also packages for data mining. There is also a version of R running on Hadoop aiming to take advantage of the computing capacities offered by Hadoop platform (one of them is released under the name of RHadoop).

[3]http://www.flightradar24.com/

V. Conclusions

In this paper we have analysed some issues and challenges arising in the field of mining Big Data. As large data sources are proliferating everywhere, there is a clear need of data mining algorithms and frameworks to deal with large data sets. The features of the Big Data, also known as the many Vs, pose several issues and challenges to the current state of the art in data mining. We have analysed particularly the scalability, I/O operations and accumulated memory issues. By using real data of a Virtual Campus, we have obtained computational results pointing out to such issues.

Data mining for Big Data is in its infancy and there is much research work to be done to successfully address the identified issues. On the one hand, there is the direction of developing new versions of existing data mining software such as Weka, R and Rapid Miner to distributed versions running on Hadoop clusters. In this same line, new frameworks are emerging designed with the aimed to be fully distributed, such as Mahout. The number of such frameworks is clearly very limited to date. On the other hand, and perhaps most interesting direction is that of designing data mining algorithms anew to take into account features of Big Data. Finally, there are similar issues in mining data streams, which are even more challenging due to the high rate of data generated by online applications (e.g. users tweets in Twitter) due to the fast increase in the number of online users.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proceedings 1993 ACM-SIGMOD International Conference on Management of Data (SIGMOD93), Washington, DC, pp. 207216, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proceedings of 1994 International Conference on Very Large Data Bases (VLDB94), Santiago, Chile, pp. 487499, 1994.

[3] R.J. Bayardo. Efficiently mining long patterns from databases. In Proceedings 1998 ACM-SIGMOD International Conference Management of Data (SIGMOD98), Seattle, WA, pp. 8593, 1998.

[4] S. Caballé, F. Xhafa: Distributed-based massive processing of activity logs for efficient user modeling in a Virtual Campus. Cluster Computing 16(4): 829-844 (2013)

[5] J. Han, M. Kamber. Data Mining -Concepts and Techniques. Morgan Kaufmann Publishers, 2000.

[6] J. Han, J. Pei, Y. Yin, and R. Mao. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Min. Knowl. Discov. 8, 1 (2004), 53-87.

[7] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. In Proceedings 1998 International Conference on Very Large Data Bases (VLDB98), New York, NY, pp. 594605, 1998.

[8] Savasere, A., Omiecinski, E., and Navathe, S. 1995. An efficient algorithm for mining association rules in large databases. In Proceedings 1995 International Conference on Very Large Data Bases (VLDB95), Zurich, Switzerland, pp. 432443.

[9] O. Terzo, P. Ruiu, . Bucci, F. Xhafa: Data as a Service (DaaS) for Sharing and Processing of Large Data Collections in the Cloud. CISIS 2013: 475-480, IEEE CPS

[10] F. Xhafa, A. Lopez Martinez, S. Caballé, V. Kolici, L. Barolli: Mining Navigation Patterns in a Virtual Campus. EIDWT 2012: 181-189, IEEE CPS.

[11] F. Xhafa, C. Paniagua, L. Barolli, S. Caballé: Using Grid services to parallelize IBM's Generic Log Adapter. Journal of Systems and Software 84(1): 55-62 (2011)