Scalable Frequent Itemset Mining algorithms for Big Data

Fabio Pulvirenti

Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy
Email: name.surname@polito.it

Abstract

Keywords:

1. Introduction

probably to rethink In the last years, we have literally been overwhelmed with data. We have witnessed, at the same moment, very strong advances in the domain of data generation, data collection and data storage. Just think about the new social applications which gathers information about every possible aspects of the users. From the voluntary data (tweets, comments, pictures) to data extracted with less straightforward techniques (cookies, pointer tracking, machine learning algorithms applied to photo repositories,...). What about the data generated by the wearable devices, or by the car black-boxes installed by car insurances on the customers' cars? The advances related to data generation and collection came together with the possibility of storing data which we would have trashed in the past. The reason behind this new trend about gathering as much data as one can is related to the new value that is given to such data. Everybody are collecting

data because it is useful. And if it is not clear how can be exploited now, probably it will be useful in the future. Lying hidden in all this raw data is potentially useful knowledge, which is rarely exploited.

The value of these data is directly correlated to the knowledge which can be extracted from it. It is very related to the use cases. Therefore, for example, it is possible to think about companies which, through the analysis of huge amount of customer attributes, are able to develop predictive models which target customers. Another example could be related to the incredible amount of data collected by sensors in the automotive domain. The possible exploitation of this information are several: from self-driving car algorithms training to predictive component fixing. Finally, many efforts are nowadays spent in pre-crime projects. By means of big data and prediction models, crimes are predicted and customized counter-measures are adopted.

In a scenario characterized by this huge amount of valuable data, the interest towards Data mining, which is a branch of computer science which extracts useful and effective knowledge from data, has risen. The trend is noticeable in both industrial and academic environments. Companies are interested in the strategic benefits that big data could deliver, even directly. In [**?**], the authors present a study to illustrate that larger data indeed can be more valuable assets for predictive analytics. The deduction is that institutions with larger collections of data and, of course, the skill to take advantage of them, can obtain a competitive advantage over institutions without. On the other hand, from the academic point of view, the design of big data algorithms represent a very stimulant challenge. The application of traditional data mining techniques to such large collection of data is very

challenging. As the amount of data increases, the proportion of it that people is able to interpret decreases (cit. Data mining: practical Machine learning tools and techniques). For this reason, there is a concrete need of a new generation of scalable tools which, often, need to be redesigned from scratches to cope with such an extreme environments.

In this dissertation, we focus on one of the most popular data mining technique, frequent itemset mining. Frequent itemset mining is an exploratory data analysis method used to discover frequent co-occurrence among the items of a transactional dataset (attribute-value pairs). Frequent itemsets are very useful for data summarization and correlation analysis. **ADD MORE ON FIM AND Association rules and rule-based classifier and raccomandation**. Frequent itemset extraction is a very challenging problem in the big data domain. The reason is related to the nature of the problem which requires a full knowledge of the input data. In the last years several scalable techniques have been introduced. All of them relies on different search space exploration strategy and this leads to different performances related to the use case.

Thesis statement: *This dissertation is an effort to thoroughly analyze the current scalable frequent itemset mining tools and, eventually, try to fill in the discovered gap.*

In the final part of this Chapter, we resume this dissertation plan highlighting our research contribution.

2. Dissertation plan and research contribution

The main contribution of this dissertation is to deeply examine the current state of the art of frequent itemset mining algorithms and its usage. Therefore, after discovering the environment lacks and issues, try to enrich it with new solutions and algorithms. This target is achieved through three main steps, useful to cluster together and label the research contributions behind them:

1. A deep analysis of the most reliable frequent itemset mining tools for big data
2. The introduction of a new scalable frequent itemset mining algorithm
3. The contribution of frequent itemsets to big data mining frameworks for the extraction of misleadig generalized itemsets

The remainder part of this section will briefly introduce each phase in order to deliver a clear idea of the structure of the dissertation work.

2.1. Frequent Itemset Mining for Big Data: an experimental analysis

Itemset mining is a well-known exploratory data mining technique used to discover interesting correlations hidden in a data collection. Since it supports different targeted analyses, it is profitably exploited in a wide range of different domains, ranging from retail store informations to network traffic or biological repositories. As already mentioned, with the increasing amount of generated data, different distributed and scalable algorithms have been developed. They have been developed exploiting the computational advantages of distributed computing platforms, such as Apache Hadoop and Apache Spark. However, depending on the use case, it is not easy to select the best

fitting algorithm. Several features affects this choice, such as data cardinality or data distribution. Therefore, the algorithm selection often relies on analyst expertise. For this reason, the delivered analysis will examine both theoretically (survey bigdap) and experimentally (survey itemset) some state-of-the-art implementations of frequent itemset mining algorithms. The ratio is to guide the analyst in selecting the most suitable approach based on the use case and the outline lesson learned. The review takes into account also some aspects typical of distributed environment, such as communication costs and load balancing. Many real and synthetic datasets have been considered in the comparison.

The takeaways of the review is that no algorithm is universally superior and performances are heavily skewed by the use cases and the relative input data. However, it is very clear that all of the algorithms focus on being able to deal with a huge number of transactions. None of them has been designed to cope with a huge number of attributes, i.e. high-dimensional data. As shown in the next subsection, we have tried to fill in this gap.

2.2. A Parallel Map-Reduce Algorithm to Efficiently Support Itemset Mining on High Dimensional Data

In today's world, many scientific applications such as bioinformatics and networking, are continuously generated large volumes of data. Since each monitored event is usually characterized by a variety of features, high-dimensional datasets have been continuously generated. Frequent itemset is one of the technique used to extract value from these complex collections, discovering hidden and non-trivial correlations among data. Thanks to the spread of distributed and parallel frameworks, the development of scalable approaches

able to deal with the so called Big Data has been extended to frequent itemset mining. Unfortunately, as mentioned in the previous Subsection (and clearly shown in Chapter **??**), most of the current algorithms are designed to cope with low-dimensional datasets, delivering poor performances in those use cases characterized by high-dimensional data. This work introduces PaMPa-HD, a MapReduce-based frequent closed itemset mining algorithm for high dimensional datasets. An efficient solution has been pro- posed to parallelize and speed up the mining process. Furthermore, different strategies have been proposed to easily tune-up the algorithm parameters. The experimental results, performed on real-life high-dimensional use cases, show the efficiency of the proposed approach in terms of execution time, load balancing and robustness to memory issues.

2.3. Big Data Mining frameworks and Misleading Generalized Itemsets

Data analysis is very large family of processes; frequent itemset mining represents just one of the steps required to deal with data. Along with other data mining algorithms, they represent just the knowledge extraction and exploration step of the whole process, which is strongly composed of many data preparation phases. The availability of distributed and parallel platforms has allowed the design of big data mining systems. These systems, with a design which is parallel starting from the very first data preparation steps, are able to deal with the so called data revolution.

In these environments, distibuted frequent itemset mining is just one of the possible 'modules' of the framework. It can be replaced by other data mining analyses or used to support further data mining processes. The latter is the case of 'misleading generalized itemset', a particular type of itemsets

obtained from frequent itemsets and a taxonomy of the input data. In this dissertation will be analyzed two real life use cases. The first is related to smart cities while the second will analyze network traffic logs.

2.4. Dissertation Plan

This dissertation is organized in the following way. Chapter **??** introduce the background related to frequent itemset mining and the distributed platforms involved. Most of all, it will deepen the problem statement and explain the challenges related to scalable implementations of new frequent itemset mining algorithms. In Chapter **??** a thorough review of the most affirmed solutions will be introduced. The performance of the best-in-class implementation will be evaluated through the utilization of synthetic and real datasets, evidencing the current limitation of the academic state of the art. Then, in Chapter **??** an innovative distributed algorithm will be presented and evaluated, demonstrating its effectiveness in the context of high-dimensional pattern mining. In Chapter **??** and **??** a big data mining framework will be introduced and, respectively, exploited to obtain a special type of frequent itemsets. Finally, Chapter **??** summarizes the main results we achieved and provide some future possible work directions.

3. Frequent Itemset Mining

1. preliminaries and details
2. Why FIM for Big Data
3. Which are the challenges

4. Related works - Survey

Analysis of the state of the art

5. Frequent Itemset Mining for high dimensional data

PaMPa-HD

6. Applications of Frequent Itemset Mining to distributed frameworks

1. MGI-Cloud
2. Nemico

7. Conclusion