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

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.

Which is the value of this data? which assets or benefits from it? TO DO

In this scenario, 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 academic and industrial environments. From the academic point of view, 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).

1.1. Data mining

1.1.1. Frequent Itemset Mining

1.2. Big Data and Distributed Frameworks

2. Problem statement

1. Why FIM for Big Data
2. Which are the challenges

3. Related works - Survey

Analysis of the state of the art

4. Frequent Itemset Mining for high dimensional data

PaMPa-HD

5. Applications of Frequent Itemset Mining to distributed frameworks

1. MGI-Cloud
2. Nemico

6. Conclusion