Cover Letter

Dear Prof. T. Palpanas and Z. Wu,
Please find enclosed the paper "A Parallel Map-Reduce Algorithm to Efficiently Support Itemset Mining on High Dimensional Data" by Daniele Apiletti, Elena Baralis, Tania Cerquitelli, Paolo Garza, Pietro Michiardi and myself, which we would like to submit to Big Data Research.

Hereby, I declare that this manuscript is the authors original work and has not been published nor it has been submitted simultaneously elsewhere. Furthermore, all authors have checked the manuscript and have agreed to the submission.

Best regards,
Fabio Pulvirenti

Preliminary Workshop Version

In the following we analyze the differences between this paper, submitted to Big Data Research, and our own previous work.

[1] D. Apiletti, E. Baralis, T. Cerquitelli, P. Garza, P. Michiardi and F. Pulvirenti PaMPa-HD: A Parallel MapReduce-Based Frequent Pattern Miner for High-Dimensional Data. *IEEE ICDM Workshop on High Dimensional Data Mining (HDM)*, Atlantic City, NJ, USA, 2015.

A very preliminary version of the PaMPa-HD algorithm was initially introduced in [1]. However, the new journal paper submitted to Big Data Research delivers a set of major contributions which can be divided in two categories: algorithmic and experimental. The former set of additions have been designed to improve the performance of the algorithm in terms of execution time and communication costs. The latter, instead, represents an additional contribution in terms of experiments and comparisons, in order to broadly validate the quality of the proposed approach.

The major algorithmic additions[1] include:

1. An improved Synchronization task is now included in the Job 3 Reducer phase (Section 4), making the overall approach much faster and less I/O intensive. In the previous version, instead, there was an ad-hoc job, burdening the execution of the algorithm with additional I/O overhead.

2. The mappers of Job 3, which run a local Carpenter from the input transposed tables, now directly process the closed itemsets of the previous iteration. This allowed them to store these itemsets and leverage them to improve the impact of the local pruning, which aims to prevent the exploration of useless branches.

3. The impact of the local pruning has been improved and, at the same time, a decrease of the communication costs of the new version of the algorithm has been introduced, by sorting the transposed tables based on their position on the enumeration tree. In the previous implementation, the sorting was not regulated because of Hadoop Distributed File System (HDFS) chunking. Exploring the tree in this way is more similar to a centralized exploration, enhancing the impact of the pruning rule. Additionally, since in the workshop version the tables were sent to the reducers as soon as they were processed (when the max_exp threshold is reached), this modification leads to a decreasing in communication costs because less redundant tables are sent to the reducers.

4. Closed itemsets are not sent to the reducers as soon as they are mined, but only at the end of the process, hence reducing the communication costs because itemsets are firstly pruned locally.

5. In order to improve the execution time, a set of strategies related to the max_exp parameter are presented (Section 5.2). The ratio behind the strategies is to increase the max_exp value along the execution, exploiting the decreasing size of the tables to mine. The benefits of the strategies are related to (i) a simpler initial parameter tuning, and (ii) a performance boost up to almost 20% in some cases.

[1] The source code of PaMPa-HD can be downloaded from https://github.com/fabiopulvi/PaMPa-HD

The other contributions, which have been included to enrich the experiments (Section 5), are briefly described in the following:

1. A whole new extremely-high-dimensional dataset has been introduced, it describes the occupancy rate of different car lanes of San Francisco bay area freeways. The dataset is useful to analyze the limit, in terms of number of transactions of the input dataset, of the high-dimensional design of PaMPa-HD.

2. Two new state-of-the-art distributed algorithms, DistEclat and BigFIM, are used to compare the performance of PaMPa-HD. The algorithms, which have shown to outperform Parallel FP-Growth (the unique distributed approach introduced in the workshop version of the paper) were not available at the time of the original submission.

3. Section 5.2 is completely new, introducing and analyzing the impact of the max_exp strategies.

4. The analysis of the impact of the number of transactions, in Section 5.4 is completely new.

5. A detailed and critical analysis of the communication costs and load balancing of the algorithm has been introduced in Section 5.6.