Cover Letter

Dear Prof. XXX,
Please find enclosed the paper "Xxx" by Daniele Apiletti, Elena Baralis, Tania Cerquitelli, Paolo Garza, Pietro Michiardi and myself, which we would like to submit to the XXX journal.

Preliminary Workshop Version

In the following we will analyze the differences between this paper, submitted to XXX, and our own previous related 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.

PaMPa-HD algorithm in its first implementation was already introduced in [1]. However, the journal extension submitted to X, delivers a set of additional contribution which can be divided in two categories: algorithmic and experimental contribution. The former set of modifications have been designed to improve the performance of the algorithm with respect to execution time, load balancing and communication cost within the commodity cluster. The latter, instead, represent an additional contribution in terms of experiments and comparisons to validate the quality of the approach.

The set of modifications involving the structure of the algorithm includes:

1. In this new implementation the Synchronization job is included in the Job 3 Reducer phase, as shown in Section 4. In the previous version, instead, there was an ad-hoc job, burdening the execution of the algorithm with addition I/O overhead.

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

3. In order to improve the impact of local pruning and, at the same time, decrease the communication cost of the algorithm, the transposed table that have to be expanded are firstly sorted in a depth-first order. 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 closed to a centralized fashion, 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 carries also a decreasing of communication cost.

4. In a similar way, in the new implementation, all the closed itemsets are not sent to the reducers as soon as they are mined, but only at the end of the process. Even this modification reduces the communication costs.

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 simplification of initial parameter tuning, and, above all, (ii) a performance boost up to the 40% in some cases.

6. To prevent an increasing load unbalancing among the independent tasks of the process, related to the increasing of the max_exp parameter, the synchronization phase is forced after 1 hour of computation. In this way, in the worst case, the resources are not completely exploited for at most 1 hour.

The second set of contributions which have been included to enrich experimental section (Section 5) of the paper include:

1. A whole brand-new very high-dimensional dataset (describing the occupancy rate of different car lanes of San Francisco bay area freeways). The dataset is analyzed in two formats: 100-row format and 440-row format (full dataset). The second format is useful to analyze the limit of the high-dimensional design of PaMPa-HD.

2. Two new state of the art algorithms with which PaMPa-HD has been compared, DistEclat and BigFIM. 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 (correctly working) at the time of the original submission.

3. Section (TO ADD) is completely new, analyzing the communication costs of the algorithms.