Zhe-jun KUANG, Hang ZHOU, Dong-dai ZHOU, Jin-peng ZHOU, Kun YANG, 2019. A non-group parallel frequent pattern mining algorithm based on conditional patterns. *Frontiers of Information Technology & Electronic Engineering*, 20(9):1234-1245. https://doi.org/10.1631/FITEE.1800467

A non-group parallel frequent pattern mining algorithm based on conditional patterns

Key words: Frequent pattern mining; Parallel algorithm; Conditional pattern bases; MapReduce; Big data

Corresponding author: Dong-dai ZHOU
E-mail: ddzhou@nenu.edu.cn
ORCID: http://orcid.org/0000-0001-5053-2935

Motivation

1. With the limitations in computing space and performance, the association of frequent items in large data mining requires both extensive time and effort, particularly when the datasets become increasingly larger.

2. Because inner-relationship data are redundant, storage of these data will significantly increase the space usage in comparison with the original dataset.

Main idea

1. We proposed a non-group PFP (NG-PFP) mining algorithm that cancels the grouping mode and reduces data redundancy between sub-tasks.

2. A non-group algorithm is designed and implemented.

3. All the proposed algorithms are validated by simulations and tests.

Method

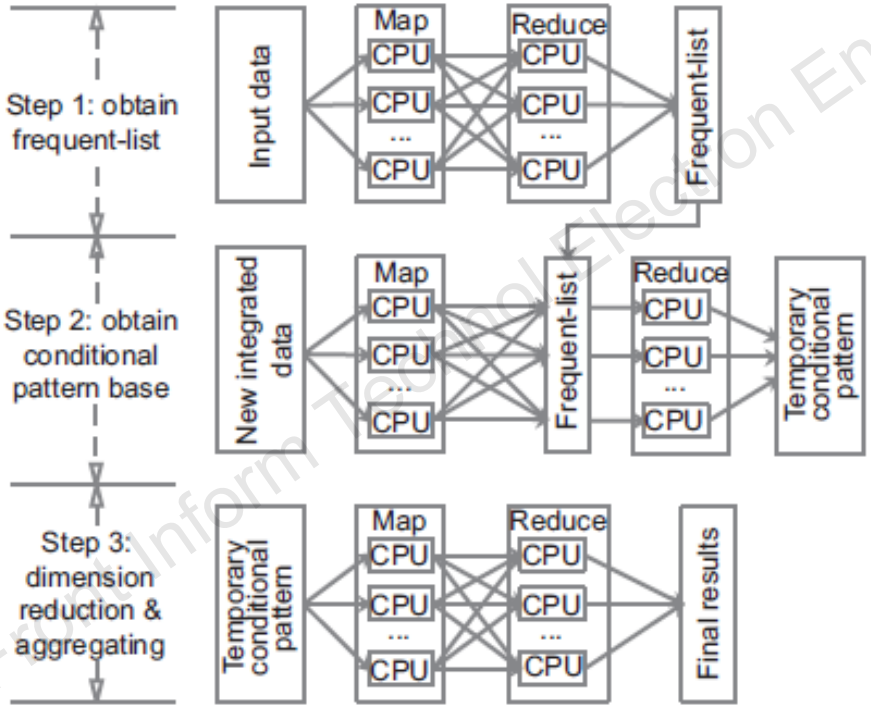

Fig. 4 Process of the non-group parallel frequent pattern (NG-PFP) algorithm

Method (Cont'd)

1. The global F-list is computed, and transaction data sharding is performed to obtain the frequent computation of all itemsets via MapReduce. The F-list obtained is presented in descending order.

2. The conditional pattern bases of all items are obtained, after which the local FP-tree is formed on the basis of the global F-list. Then the conditional pattern bases of all items are identified using the bottom-up method.

Method (Cont'd)

3. The full set of frequent items is obtained, and the infrequent items in conditional pattern bases are eliminated as per the minimum threshold. Using the FP-growth approach, the full set of frequent items is obtained on the basis of the conditional pattern bases of infrequent items as a minimum threshold.

Major results

Test results

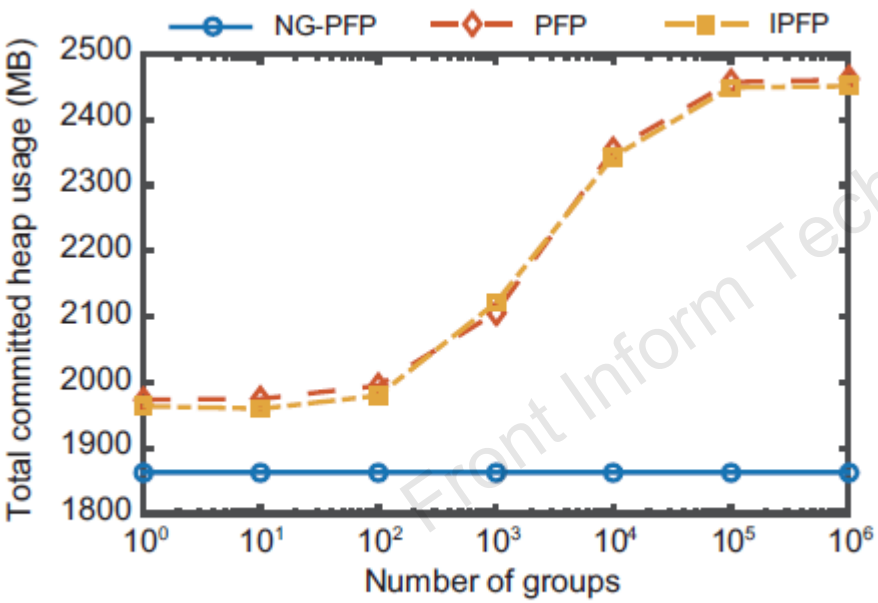

Fig. 9 Computing space occupancy rate

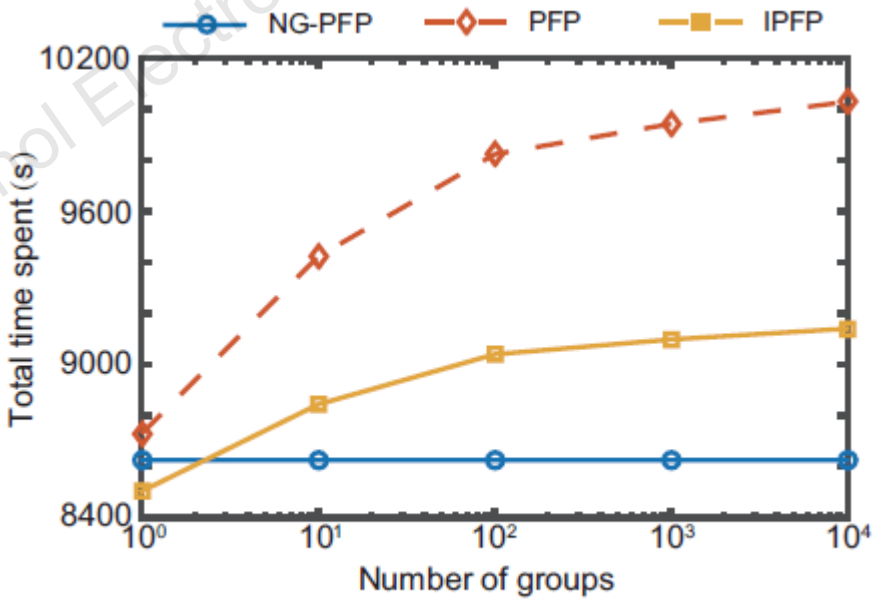

Fig. 10 Total time spent by groups

Major results (Cont'd)

Test results

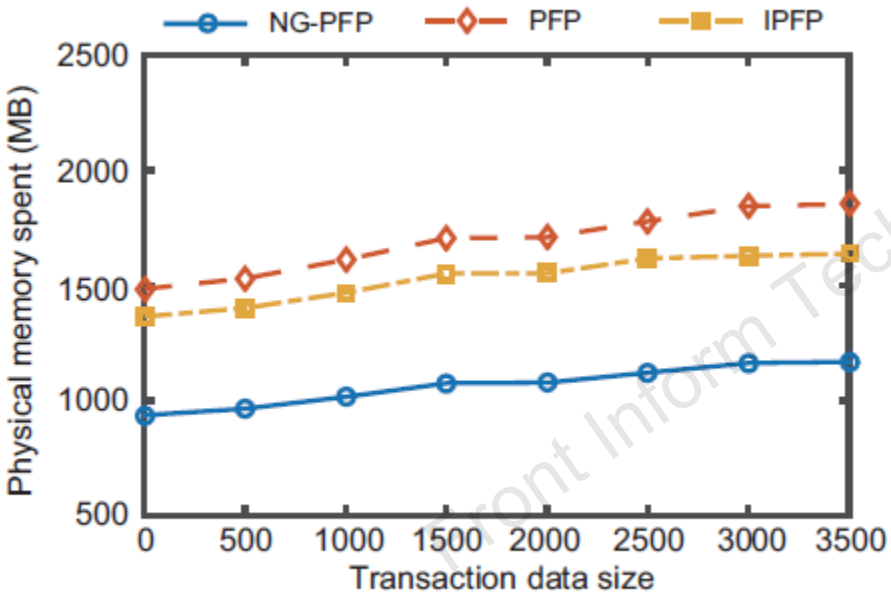

Fig. 11 Physical memory spent by data

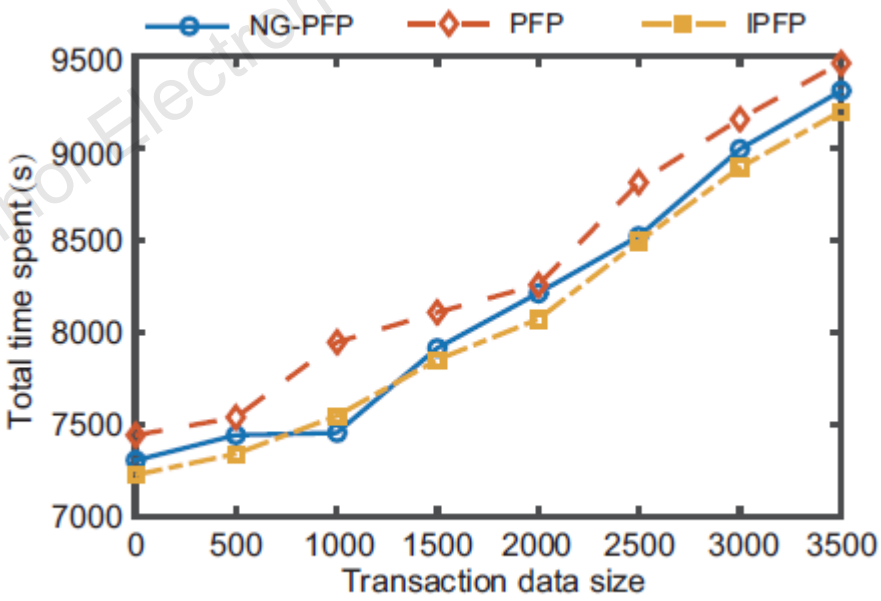

Fig. 12 Total time spent by data

Conclusions

1. A non-grouping PFP algorithm has been designed.

2. The proposed novel approach can cancel the grouping mode and reduce data redundancy between sub-tasks.

3. Through simulations and tests, the strategy and the current cost of time/computing space are thoroughly analyzed and verified.