

# A Novel MapReduce-based Algorithm for Association Rules Mining

**Bin Wu**

JiangXi University of Science and Technology

**Yimin Mao**

JiangXi University of Science and Technology

**Deborah Simon Mwakapesa**

JiangXi University of Science and Technology

**Yaser Ahangari Nanehkaran**

JiangXi University of Science and Technology

**Qianhu Deng**

JiangXi University of Science and Technology

**Jianbing Yi**

JiangXi University of Science and Technology

**Xueyu Huang (✉ [lycmym@163.com](mailto:lycmym@163.com))**

JiangXi University of Science and Technology

---

## Research Article

**Keywords:** CanTree, data compression, information entropy, MapReduce, association rules

**DOI:** <https://doi.org/10.21203/rs.3.rs-388532/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

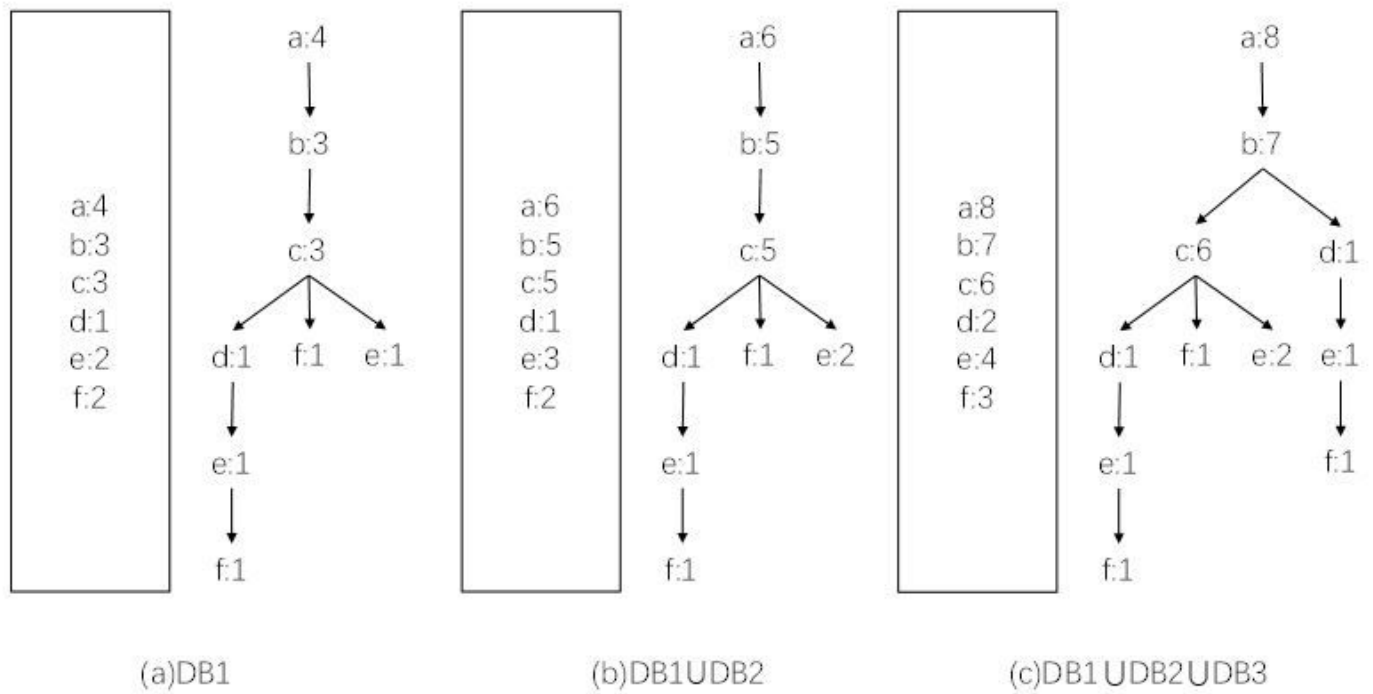
# Abstract

AR (Association rule) is considered to be one of the models for data mining. With the growth of datasets, conventional association rules are not suitable for big data mining, which has aroused a large number of scholars' interest in algorithm innovation. This study aims to design an optimization parallel association rules mining algorithm based on MapReduce, named as PMRARIM-IEG algorithm, to deal with problems such as the excessive space occupied by the CanTree (CanTreeCanonical order Tree), the inability to dynamically set the support threshold, and the time-consuming data transmission in the Map and Reduce phases. Firstly, a structure called SIM-IE (similar items merging based on information entropy) strategy is adopted for reducing the space occupation of the CanTree effectively. Then, a DST-GA (dynamic support threshold obtaining using genetic algorithm) is proposed to obtain the relatively optimal dynamic support threshold in the big data environment. Finally, in the process of MapReduce parallel, a LZO (Lempel-Ziv-Oberhumer) data compression strategy is used to compress the output data of the Map stage, which improves the speed of the data transmission. We compared the PMRARIM-IEG algorithm with other algorithms on five datasets, including Wikipedia, LiveJournal, com-amazon, kosarak, and webdocs. The experimental results obtained demonstrate that the proposed algorithm, PMRARIM-IEG, not only reduces the space and time complexity, but also obtains a well-performing speed-up ratio in a big data environment.

# Full Text

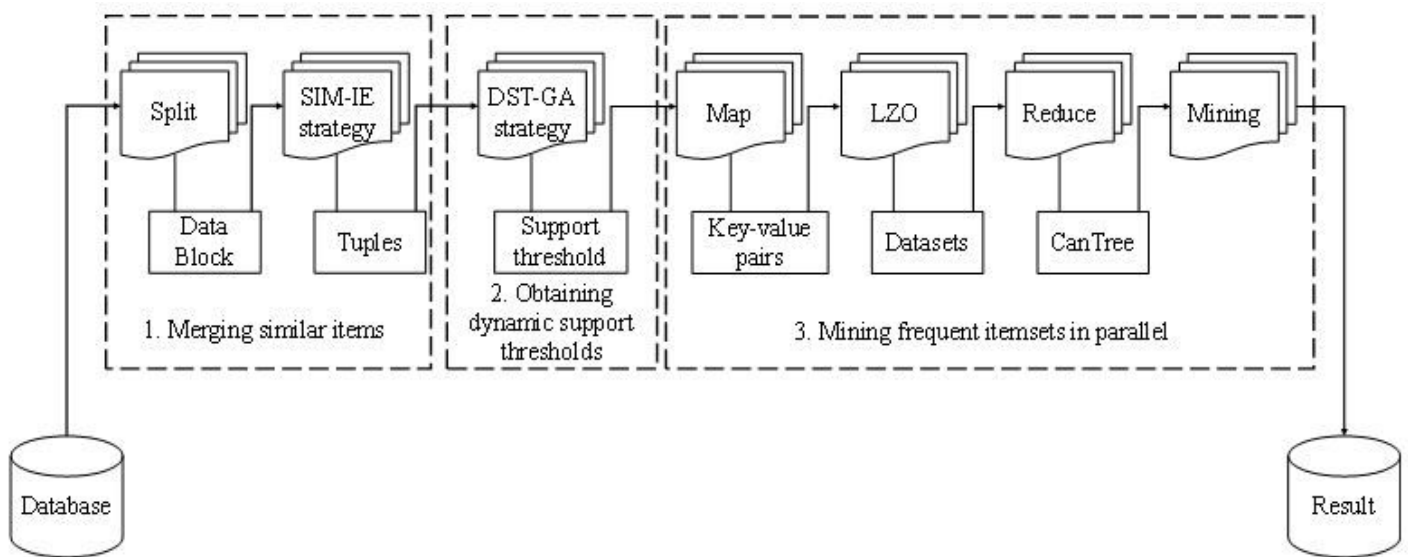
This preprint is available for [download as a PDF](#).

# Figures



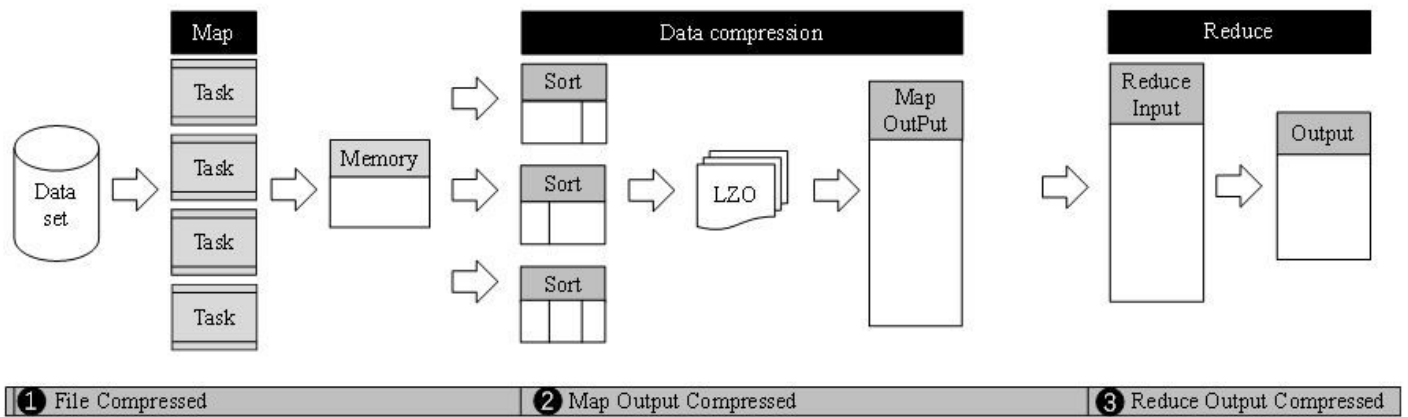
**Figure 1**

CanTree Constructed Process



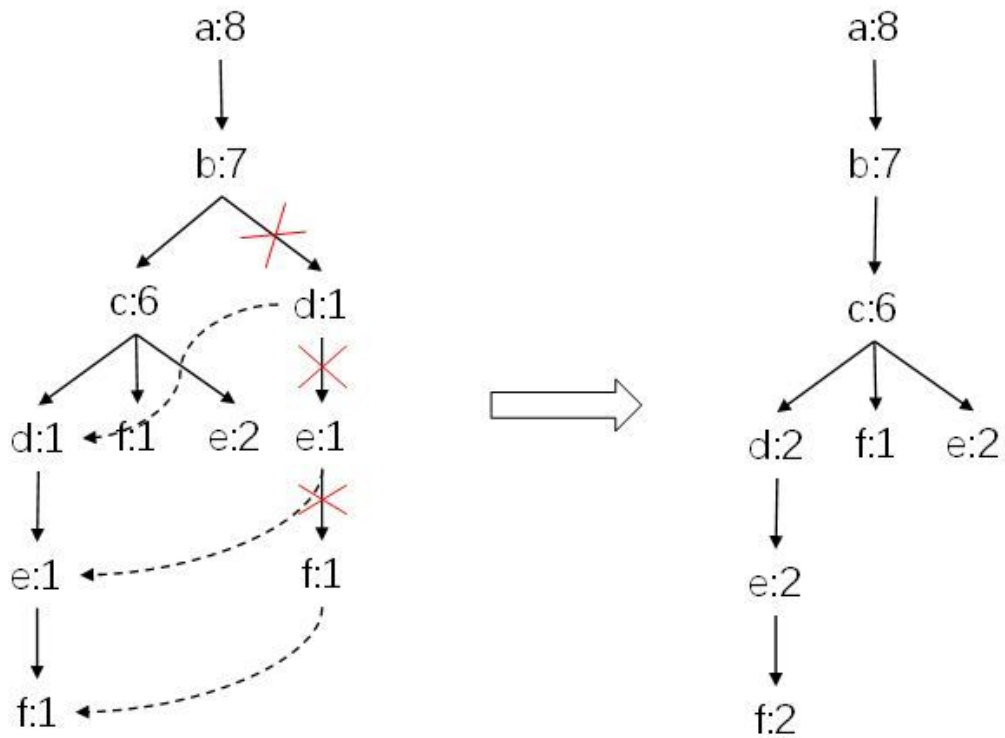
**Figure 2**

PMRARIM-IEG algorithm structure



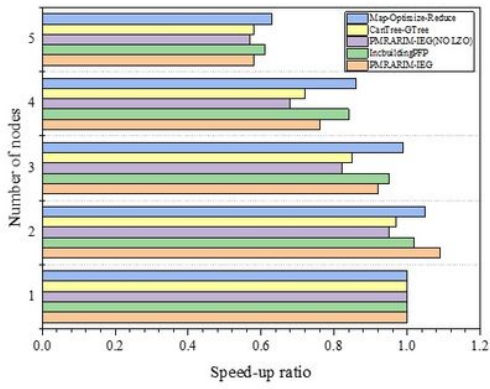
**Figure 3**

Data Compression processing in MapReduce

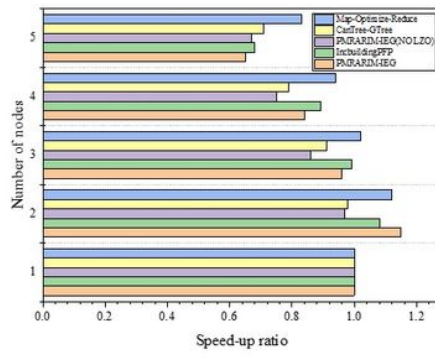


**Figure 4**

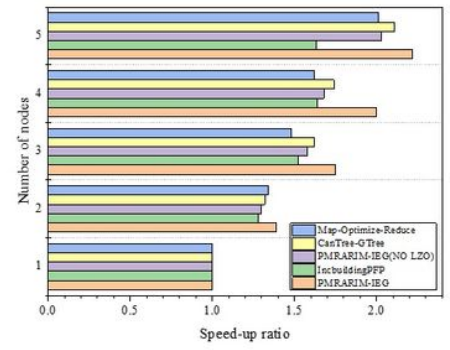
Optimization of CanTree structure



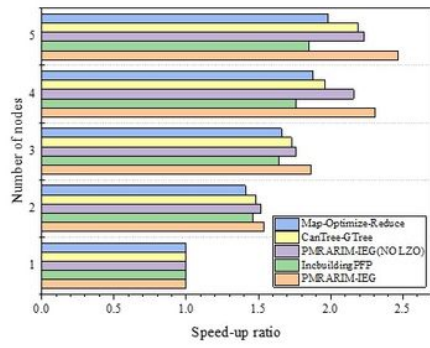
(a) Speed-up ratios for each algorithm on com-amazon



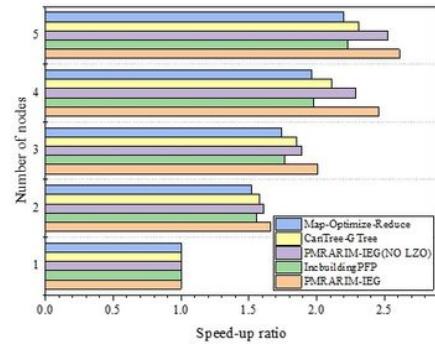
(b) Speed-up ratios for each algorithm on kosarak



(c) Speed-up ratios for each algorithm on Wikipedia



(d) Speed-up ratios for each algorithm on LiveJournal



(e) Speed-up ratios for each algorithm on webdocs

## Figure 7

Speed-up ratios for each algorithm