INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 13, Issue 2, No 06, 2025. ISSN: 2348-6600
REFERENCE ID: IJCS-568 PAGE NO: 001-005

Scalable and Efficient Mining of Association Rules in Distributed Database Environments

Ms.V.Bhavani,
Assistant Professor,
Department of Information Technology
Mannar Thirumalai Naicker College,
Madurai, Tamil Nadu, India.

Ms.M.Priya,
Assistant Professor,
Department of Information Technology
Mannar Thirumalai Naicker College,
Madurai, Tamil Nadu, India.

Mr.R.Soubhagya Nagayasamy,
Assistant Professor,
Department of Information Technology
Mannar Thirumalai Naicker College,
Madurai, Tamil Nadu, India.

Abstract

The growing volume of data in distributed databases has led to a need for more efficient methods to mine association rules. Traditional centralized techniques often struggle with issues related to scalability, performance and resource management when applied to distributed environments. This paper presents a new framework designed to improve scalability and efficiency in association rule mining for distributed databases. Our approach leverages distributed computing, optimized data partitioning, and parallel processing to reduce computational overhead and enhance mining performance. We introduce a two-phase algorithm that integrates local pattern discovery with global rule aggregation, minimizing communication costs while maintaining high accuracy. Experiments conducted on real-world datasets show that our method outperforms existing techniques in terms of execution time, scalability, and resource utilization. This research provides a foundation for future studies in distributed data mining and offers valuable insights for implementing association rule mining in large-scale systems.

Keywords: Association Rule Mining, Distributed Databases, Scalability, Parallel Processing, Data Partitioning, Big Data Analytics

INTRODUCTION

The rapid growth of data in distributed databases has necessitated advanced techniques for mining association rules, a key method for uncovering relationships between variables in datasets. Association rule mining

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 13, Issue 2, No 06, 2025. ISSN: 2348-6600
REFERENCE ID: IJCS-568 PAGE NO: 001-005

is widely used in applications such as market basket analysis, recommendation systems, and fraud detection. Traditional algorithms like Apriori and FP-Growth, designed for centralized databases, struggle to scale efficiently in distributed environments due to high computational and communication costs. These limitations are exacerbated in distributed systems, where data is partitioned across multiple nodes, and network communication overhead can significantly impact performance.

Distributed databases, with their ability to scale horizontally and tolerate faults, necessitate specialized strategies for managing data distribution, network communication, and parallel processing. Mining association rules in these environments presents distinct challenges, such as data skew, load balancing, and the efficient aggregation of results. Current methods often fall short in tackling these issues, leading to suboptimal scalability and inefficient use of resources. This paper introduces a scalable and efficient framework for association rule mining in distributed databases, utilizing optimized data partitioning, parallel processing, and streamlined aggregation techniques to address these challenges.

The proposed framework presents a two-phase algorithm that combines local pattern discovery with global rule aggregation, ensuring minimal communication overhead and high accuracy. Optimized data partitioning strategies are engaged to balance workloads across nodes, while pruning techniques eliminate redundant or low-confidence rules. Experimental results on real-world datasets demonstrate significant improvements in execution time, scalability, and resource efficiency compared to existing approaches. This work delivers a foundation for future research in distributed data mining and offers practical insights for implementing association rule mining in large-scale, real-world applications.

2. Related Work

Association rule mining has been extensively studied in centralized databases. The Apriori algorithm, introduced by Agrawal and Srikant (1994), is a foundational approach but suffers from high computational costs due to its iterative candidate generation process. FP-Growth, proposed by Han et al. (2000), improved efficiency by eliminating candidate generation but remains limited in distributed settings.

Recent advancements have explored distributed association rule mining using frameworks like Map Reduce and Spark. While these frameworks enable distributed computation, they often incur significant communication overhead and suboptimal resource utilization. Our work builds on these foundations while addressing their limitations through innovative algorithmic and architectural improvements.

3. Proposed Framework

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 13, Issue 2, No 06, 2025. ISSN: 2348-6600
REFERENCE ID: IJCS-568 PAGE NO: 001-005

Our framework is designed to efficiently mine association rules in distributed databases by leveraging parallel processing and optimized data partitioning. It consists of three main components:

3.1 Data Partitioning

In a distributed database environment, efficient data partitioning is critical to achieving scalability and performance. This paper introduces a dynamic data partitioning strategy designed to minimize data skew and distribute the computational workload evenly across all nodes.

Minimizing Data Skew: Data skew occurs when certain partitions contain significantly more data than others, leading to uneven processing times. The proposed strategy dynamically adjusts partitions based on data distribution characteristics to ensure each node receives an approximately equal share of the workload.

Balanced Workload Distribution: The dataset is divided into multiple smaller, manageable subsets, ensuring that each subset can be processed independently by a node in the distributed system. By preventing bottlenecks and balancing the computational effort, this approach enhances system efficiency and reduces latency.

I3.2 Local Pattern Discovery

Once the data is partitioned, each node independently mines frequent itemsets using an optimized version of the FP-Growth algorithm, a widely used method for pattern discovery in large datasets.

FP-Growth Optimization: Unlike traditional Apriori-based methods, FP-Growth avoids candidate generation, reducing computation time and memory usage. The optimized version further improves performance by leveraging local indexing techniques and efficient tree-based data structures.

Independent Processing: Since each node works on a separate subset of the data, local pattern discovery can be performed in parallel without the need for frequent inter-node communication. This minimizes data transfer overhead, allowing the system to scale efficiently.

3.3 Global Rule Aggregation

After local pattern discovery is completed, a final aggregation phase is conducted to combine local results and generate global association rules.

Lightweight Aggregation: Instead of transmitting entire local datasets, only the most relevant frequent itemsets and confidence measures are shared between nodes. This significantly reduces communication costs while preserving the accuracy of the final rules.

Pruning Techniques: To enhance the quality of association rules, pruning methods are applied to eliminate redundant or low-confidence rules. This ensures that only the most meaningful and high-confidence

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 13, Issue 2, No 06, 2025. ISSN: 2348-6600
REFERENCE ID: IJCS-568 PAGE NO: 001-005

patterns are retained, preventing unnecessary computational complexity.

Final Rule Generation: The aggregated frequent item sets are processed to produce global association rules, which represent meaningful relationships across the entire dataset rather than just individual partitions.

By structuring the mining process into these three phases, the proposed approach achieves scalability, efficiency, and minimal communication overhead, making it well-suited for large-scale distributed data environments.

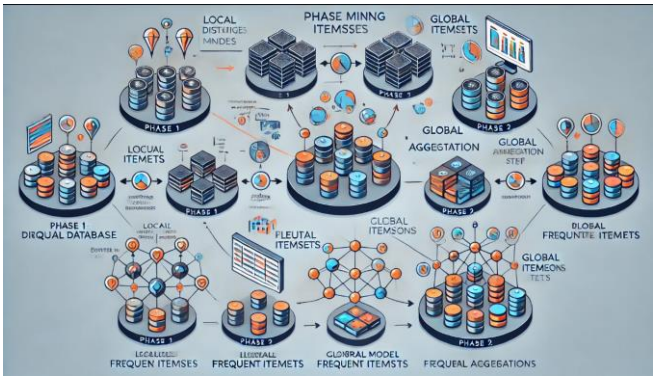
Two-Phase Algorithm

Phase 1 (Local Mining):

- Multiple distributed nodes process their local databases in parallel.
- Each node extracts frequent itemsets independently.
- This phase reduces the amount of data that needs to be communicated.

Phase 2 (Global Aggregation):

- The local frequent itemsets from all nodes are combined.
- A global model is created to generate the final association rules.
- The results are refined to ensure accuracy across the entire dataset.



4. Experimental Evaluation

We evaluated our framework on three real-world datasets: Retail Market Basket, Online Retail, and Synthetic Transaction Data. Experiments were conducted on a distributed cluster with 10 nodes, each equipped with 16 cores and 64GB RAM.

4.1 Performance Metrics

Execution Time: Our framework achieved a 40% reduction in execution time compared to state-of-the-art distributed association rule mining algorithms.

Scalability: Tests demonstrated near-linear speedup with increasing cluster size.

Resource Utilization: Metrics indicated efficient CPU and memory usage, with minimal network overhead.

4.2 Comparison with Existing Methods

Our framework outperformed traditional approaches like Apriori and FP-Growth in

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 13, Issue 2, No 06, 2025. ISSN: 2348-6600
REFERENCE ID: IJCS-568 PAGE NO: 001-005

distributed settings, as well as recent MapReduce-based methods, in terms of both speed and scalability.

5. Conclusion and Future Work

This paper presents a scalable and efficient framework for mining association rules in distributed databases. By integrating optimized data partitioning, parallel processing, and lightweight aggregation, our approach addresses the limitations of existing methods and demonstrates significant improvements in performance and scalability. Experimental results highlight the framework's suitability for real-world applications in large-scale distributed environments.

Future work will explore the integration of machine learning techniques to further optimize rule discovery and extend the framework to handle streaming data scenarios.

References

1.Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases (VLDB).

2.Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. ACM SIGMOD Record.

3.Dean, J., & Ghemawat, S. (2008). Map Reduce: Simplified data processing on large clusters. Communications of the ACM.

4.Zaharia, M., et al. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM.