International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952206

Efficient Large Scale Frequent Itemset Mining with Hybrid Partitioning Approach

Priyanka R., Mohammed Ibrahim M., Ranjith Kumar M.
Department of Computer Science and Engineering, Dr. Mahalingam College of Engineering and Technology
Pollachi, Coimbatore, Tamil Nadu, India

ABSTRACT

In today's world, voluminous data are available which are generated from various sources in various forms. Mining or analyzing this large scale data in an efficient way so as to make them useful for the mankind is difficult with the existing approaches. Frequent itemset mining is one such technique used for analyzing in many fields like finance, health care system where the main focus is gathering frequent patterns and grouping them to be meaningful inorder to gather useful insights from the data. Some major applications include customer segmentation in marketing, shopping cart analyses, management relationship, web usage mining, player tracking and so on. Many parallel algorithms, like Dist-Eclat Algorithm, Big FIM algorithm are available to perform large scale Frequent itemset mining. In Dist-Eclat algorithm, datasets are partitioned using Round Robin technique which uses a hybrid partitioning approach, which can improve the overall efficiency of the system. The system works as follows: Initially the data collected are distributed by mapreduce. Then the local frequent k-itmesets are computed using FP-Tree and sent to the map phase. Later the mining results are combined to the center node. Finally, global frequent itemsets are gathered by mapreduce. The proposed system is expected to improve in efficiency by using hybrid partitioning approach in the datasets based on the identification of frequent items.

Keywords : Dist-Eclat Algorithm, Frequent Itemset Mining, Mapreduce, K-Itmesets, Large Data, Data Mining,Frequent Itemset Mining

I. INTRODUCTION

Data mining had been active research area around 20 years. There are many tasks performed in data mining. The huge research efforts have led to a variety of sophisticated and efficient algorithms to find frequent item sets. Industries use the extracted frequent item-sets in decision making or setting policies. For example a retail-sector company is interested to know customer buying habits in particular area to sell out their product. Here, frequent item-set mining helps the company to know customer buying habits. On the other hand, even government of various nations use the frequent item-set technique to extract useful information that further help to provide better services to people.

Frequent item-set mining is the part of frequent pattern mining where frequent pattern represents those sub-sequences and sub-graphs based on the occurrence of items in the given data sets. Traditional data mining tools fails to extract frequent item-sets from large data sets. So, recently developed big data technology can be adopted for processing large scale data.

In Big Data, new approach is required to compute frequent item-sets where data-sets consist of millions

of records. Researchers proposed various approach to deal with Big Data challenges, but all these approaches suffers from synchronization, work load balancing and fault-tolerance problem. To solve this problem there are several parallel algorithms available in data mining that can be implemented using mapreduce. The best known parallel algorithms to find frequent item set from large data sets are Apriori and FP-growth.

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases[3]. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. This has applications in domains such as market basket analysis.

FP-growth is the fastest algorithm, employs a sophisticated and rather complex data structure and thus requires to load the transaction database into main memory[1]. Hence a simpler processing scheme, which maintains efficiency, is desirable. Other lines of improvement include filtering found frequent item sets and association rules, identifying temporal changes in discovered patterns and discovering fault-tolerant or approximate frequent item sets.

Eclat is an acronym for Equivalence class Transformation. Contrarily to algorithms such as Apriori, Eclat uses a depth-first search for discovering frequent itemsets instead of a breadth-first search. The Eclat algorithm uses different techniques like Block partitioning, Round-Robin technique, equal-weight partitioning technique and so on to partition the data sets to the distributed systems. However Eclat algorithm is efficient in finding frequent itemsets, the number of transactions in each distributed system is large. To reduce the number of transactions, a hybrid partitioning technique is proposed in this paper.

1. Block diagram of data mining process

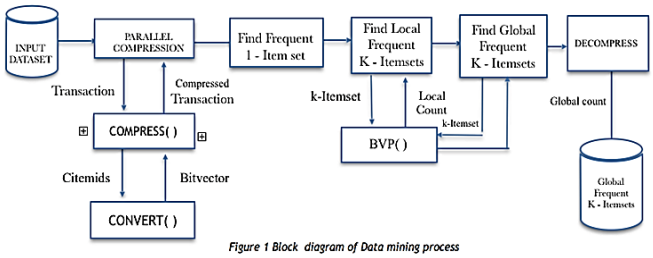


Figure 1 Block diagram of Data mining process

Figure 1, shows the overall block diagram of data mining process. In this process initially the input datasets are parallel compressed in which a transaction of datasets occur to compress the data for further process. Here, the compress and the convert process occurs to complete the transaction.

The first singleton frequent item is found from which the local frequent k-itemsets have to be identified. Bi-vector process takes place to find the local count of the k-itemsets. The local count is used to produce the global count of the k-frequent itemsets. Global k-frequent itemsets are generated by the Bi-vector process using the global count generated in the previous step. Finally the compressed datasets are decompressed to produce the global k-frequent itemsets.

A. Objective

Improving the efficiency of frequent itemset mining algorithm from large scale data.

Demonstrating the frequent itemset mining algorithm in Food Nutrition application in real time scenario.

II.METHODOLOGY

A. Introduction

The existing system shows the implementation of dist-Eclat algorithm using two different partitioning

techniques called block partitioning and Round-Robin partitioning. Eclat algorithm uses a vertical database format for fast support computation. Distributed version of Eclat, partitions the search space more evenly among different processing units. Current techniques distribute the workload by partitioning the transaction database into equally sized sub databases.

B. Preprocessing of input data

T Each item in the dataset is stored together with its cover (also called tidlist) and uses the intersection based approach to compute the support of an itemset less space than apriori if itemsets are small in number. It is suitable for small datasets and requires less time for frequent pattern generation than apriori. Below is the Example of Eclat Algorithm for minimum support = 2.

Initially the database will be in the horizontal data format in which dist-Eclat algorithm is difficult to implement. The database has to be converted into vertical data format in-order to implement the algorithm in the datasets.

The database is converted into vertical data format where each itemsets contains n number of transactions which is represented as tid.

Each singleton itemset contains n number of transactions which is generated in the single phase of the algorithm.

Similarly in the next phase, the itemsets that occur in the pair of transactions is identified for generating frequent itemsets.

TID	List of item IDs
T100	I1,I1,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

a. Database in Horizontal data format

Itemsets	TID_SET
I1	{T100,T400,T500,T700, T800,T900}
I2	{T100,T200,T300,T400, T600,T800,T900}
I3	{T300,T500,T600,T700, T800,T900}
I4	{T200,T400}
I5	{T100,T800}

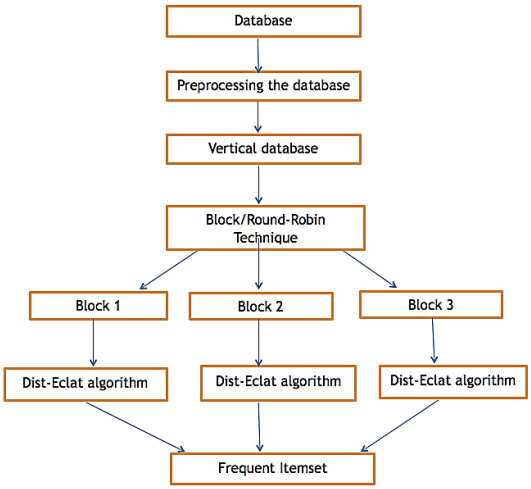
b.Database in Vertical data format

Itemsets	TID_SET
{I1,I2,I3}	{T800,T900}
{I1,I2,I5}	{T100,T800}

c. Frequent 3 itemset generated by intersection of 2-itemset

The datasets are pre-processed and dist-Eclat algorthim is applied using two different partition techniques such as block partitioning and Round Robin techniques. The applied techniques are compared to each other using different minimum threshold to check the number of transactions done in each phase in each client node.

2. Block diagram for existing system



The system consists of three major modules in finding the frequent itemsets. Initially, the database is preprocessed and converted into vertical database. In second module, the datasets are partitioned using Round-Robin or Block partitioning technique. In the third module, dist-Eclat algorithm is implemented to find the frequent itemsets.

C. Preprocessing and partitioning the datasets

3. Preprocessing and partitioning the datasets

The large datasets are processed to remove the unwanted spaces in the horizontal database. The selected database is then preprocessed to get converted into vertical database.

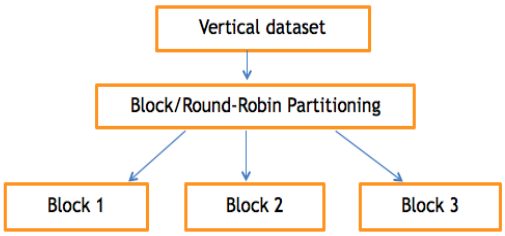


Figure 5 Dividing the datasets and client Communication

D. Dividing the datasets and Client Communication

4. Dividing the datasets and client Communication

In this module, the preprocessed data is divided into

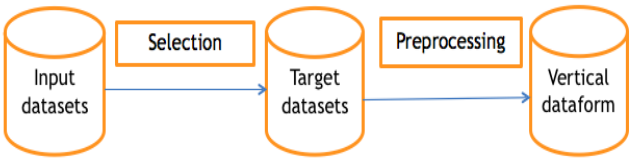


Figure 4 Preprocessing and partitioning the datasets

number of sub datasets from the central dataset as per the desired number of data nodes using Block partitioning or Round-Robin technique. The partitioned data will be then processed by the third module in which dist-Eclat algorithm is implemented to find the frequent itemsets

C. Architecture of Dist-Eclat Algorithm

The datasets are initially converted from horizontal database format to vertical database format and partitioned into equal sub-databases to different mappers using partitioning techniques. Each mapper processes a distinct chunk of the data and produces key-value pairs. In the reduce phase, key-value pairs from different mappers are combined by the framework and fed to reducers as pairs of key and value lists.

Reducers further process those intermediate parts of information and output frequent singleton tid-list. Similarly, in the next part the mappers and the reducers process the singleton tid-list to produce k-Frequent itemset and the prefix tree. The prefix tree used by Eclat can be partitioned into independent groups. Each one of these independent groups can be mined separately on different mappers. The mapreduce phase process the data to produce the frequent itemsets.

However, this approach comes with a large communication cost, i.e., the number of sets to be mined can be very large, moreover, the number of sets that have to be recounted can be very large as well. A possible solution for recounting part, is to mine the sub databases with a lower threshold, hence, decreasing the number of item sets that might have been missed. In this, no extra communication between mappers is necessary and no checking of overlapping mining results is required.

In module 1 the vertical database is divided into equally sized blocks(shards) and distributed to available mappers. Each mapper extracts the frequent singletons from its shard. In the reduce phase, all frequent items are gathered without further processing.

In module 2, P_k the set of frequent itemsets of size k, is generated. First, frequent singletons are distributed across m mappers. Each of the mappers finds the frequent k-sized supersets of the items by running Eclat to level k. Finally, a reducer assigns P_k to a new batch of m mappers. Distribution is done using Round Robin.

In module 3, the prefix tree starting at a prefix is mined from a assigned batch using Eclat. Each mapper can complete this step independently since sub-trees do not require mutual information.

IV. PROPOSED SYSTEM

The input datasets in the vertical database format is partitioned into the distributed system using Hybrid partitioning technique. The partitioned sub-databases are processed individually in the MapReduce phase and dist-Eclat algorithm is implemented to find the frequent itemsets.

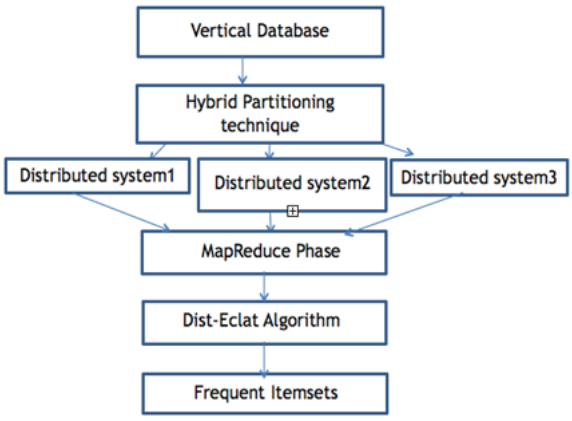


Figure 7 Block diagram for proposed system

5. Block diagram of Proposed system

The block partitioning and the Round Robin technique generates large number of transactions in each distributed system whereas the Hybrid Partitioned technique is expected to reduce the number of transactions in each distributed system. The hybrid partitioning technique is used to distribute the datasets to the mappers based on the number of occurrence of the items in the transaction. Hence, the number of distributed systems is gradually decreased and it is expected that the number of transactions is also reduced in each system.

In the Hybrid partitioning approach, initially the datasets are partitioned using Round-Robin technique and in each block the datasets are again partitioned using the block partitioning technique inorder to reduce the number of transactions in each block.

V. IMPLEMENTATION

A. Dataset

The experiment requires large data sets to analyze the techniques in identifying frequent itemsets. The techniques used to partition the datasets using Dist-Eclat Algorithm are Block Partitioning technique and Round robin technique with different minimum threshold.

B. Dataset statistics

Number of itemsets in the given dataset	2603
Number of transactions in total	541906
Total count of itemsets in the transaction	2934526

d.Dataset Statistics

C. Evaluation Metric

The block partitioning and round robin techniques used in this experiment to partition the datasets gives large number of transactions when different minimum threshold is used to find different frequent itemsets. The number of transactions generated in each system is used as evaluation metric for our proposed system. The proposed hybrid partitioning technique is expected to reduce the number of transactions in the client node. Let $I=\{i_1,i_2,\ldots, i_n\}$ be a set items, a transaction is defined as T=(tid, X) where

tid is the transation identifier and X is a set of items over I. A transaction database D is a set of transactions. Its vertical database D' is a set of pairs that are composed of an item and the set of transactions C_j that contains the item:

D' = {(ij,Cij ={tid | ij € X, (tid, X) € D})}

D. Comparison

Partitioning Technique	No. of transactions in distributed system 1	No. of transactions in distributed system 2	No. of transactions in distributed system 3
Block partitioning	180635	180634	180635
Round-Robin	180635	180635	180635

e.Comparison of transactions using block partitioning and Round-Robin

In this experiment, we compare the number of transactions in round robin and block partitioning technique using dist-Eclat algorithm. The results clearly defines the large number of transactions that occur in each distributed systems using both the techniques in which mapreduce phase is implemented to process the datasets in each distributed system to generate the frequent itemsets.

f.Comparison of transactions with different thresholds

The minimum threshold refers to the minimum support count used for finding the frequent itemsets from the given vertical database. The result shows the gradual decrease in number of transactions when the minimum threshold is increased.

Blocks	Minimum threshold	Threshold 2	Threshold 3	Threshold 4
Block 1		176588	102674	18343
Block 2		153902	123845	22327
Block 3		134452	110374	16923

VI. RESULTS

Large-scale data for frequent itemset mining is gathered and partitioned using the techniques of Round Robin partitioning and Block partitioning. The support which needs to be calculated in the preprocessing is done with intersection-based approach, saving more time and speeding up the process. Thus more transactions are generated than normal.

The comparison with minimum threshold is also enhanced since massive data transactions are generated. The applied techniques are compared with each other using different minimum threshold. System is evaluated using the number of transactions occurred in each partitioning technique. The communication cost incurred could be reduced by the usage of lower threshold.

In Round Robin approach, datasets are partitioned using the technique of dividing the items one by one to each system in a cyclic fashion (also known as cyclic executive). In Block partitioning approach, the datasets are partitioned by dividing total number of itemsets by number of distributed systems or blocks. The system can be extended by using a hybrid approach, which involves block partitioning and Round-Robin within the block. Thus, as a result the number of transactions can be reduced in the client node using this hybrid partitioning method.

VII. ACKNOWLEDGMENT

We are grateful to **Dr. G. Anupriya,** Head of the Department, Computer Science and Engineering, for her direction delivered at all times required. We also thank her for her tireless and meticulous efforts in bringing out this project to its logical conclusion.

Our hearty thanks to our guide **Mrs. V. Priya** Assistant Professor(SS) for her constant support and guidance offered to us during the course of our project by being one among us and all the noble hearts that gave us immense encouragement towards the completion of our project.

VIII. REFERENCES

[1]. Gosta Grahne and Jianfei Zhu," Fast Algorithms for Frequent Itemset Mining Using FP-Trees", IEEE Transactions on Knowledge and Data Engineering, October-2005,USA.

[2]. Go sta Grahne and Jianfei Zhu, "Fast Algorithms for Frequent Itemset Mining Using FP-Trees", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 10, OCTOBER 2005.

[3]. Siddhrajsinh Solanki, Neha Soni," A Survey on Frequent Pattern Mining Methods Apriori, Eclat, FP growth", International Journal of Computer Techniques, 2013, India.

[4]. J. Han, H. Pei and Y. Yin," Mining Frequent Patterns without candidate generation", Conference on the Management of Data, 2014, New York.

[5]. Manjit kaur, Urvashi Grag," ECLAT Algorithm for Frequent Itemsets Generation", International Journal of Computer Systems (ISSN: 2394-1065), Volume 01– Issue 03, December, 2014, India.

[6]. S.N. Patil, "Frequent Itemset Mining for Big Data", International Conference on Green Computing and Internet of Things (ICGCIoT), 2015, India.

[7]. Ferenc Kovács & János Illés," Frequent Itemset Mining on Hadoop", IEEE 9th International Conference on Computational Cybernetics July 8-10, 2015, Tihany, Hungary.

[8]. Savo Tomovic & Predrag Stanišiü," Fast Algorithm for Enumerating Frequent Itemset Pairs in Database of Transactions", 4 th Mediterranean Conference on Embedded Computing MECO – 2015,Budva, Montenegro.

[9]. Zahra Farzanyar, Nick Cercone,"Efficient Mining of Frequent itemsets in Social Network Data based on MapReduce Framework ", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,2015, Toronto, Canada.

[10]. Zhang Xin, Li Kunlun, and Liao Pin," A Depth-First Search Algorithm of Mining Maximal Frequent Itemsets", 7th International Conference on Advanced Computational Intelligence Mount Wuyi, Fujian, March 27-29, 2015,China.

[11]. Tushar M. Chaure and Kavita R. Singh, " Frequent Itemset Mining Techniques - A Technical Review", World Conference on Futuristic Trends in Research and Innovation for Social Welfare (WCFTR'16), 2016, India.

[12]. Zhigang Zhang, Genlin Ji, Mengmeng Tang,"MREclat: an Algorithm for Parallel Mining Frequent Itemsets", International Conference on Advanced Cloud and Big Data, 2016,Nanjing, China.

[13]. Ankit N. Dharsandiya & Mihir R. Patel," A Review on Frequent Itemset Mining Algorithms in Social Network Data", IEEE WiSPNET, 2016 conference, India.

[14]. Dr. Ruchi Agarwal,Sunny Singh & Satvik Vats," Implementation of Improved Algorithm for Frequent Itemset Mining using Hadoop ", International Conference on Computing, Communication and Automation (ICCCA2016), April,2016, Galgotias University, India.

[15]. Savo Tomovic and Predrag Stanisi," Fast Algorithm for Enumerating Frequent Itemset

Pairs in Database of Transactions", 4th Mediterranean Conference on Embedded Computing,2016.

Cite this article as :

Priyanka R., Mohammed Ibrahim M., Ranjith Kumar M., "Efficient Large Scale Frequent Itemset Mining with Hybrid Partitioning Approach", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 845-852, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT1952206
Journal URL : http://ijsrcseit.com/CSEIT1952206