Examining the Performance of Vertical Fragmentation using FP-MAX Algorithm

Nidhi Thakur
Researcher
LPU, Jalandhar, Delhi G.T.Road
(NH1), Phagwara, Punjab, India-144411

Balwant Ram
Asst. Professor
LPU, Jalandhar, Delhi G.T.Road
(NH-1), Phagwara, Punjab, India-144411

ABSTRACT
Today's business Environment has an increasing need for consistent, scalable, reliable and accessible information which grows steadily.The purpose of this work is to analyse the performance of Vertical Fragmentation on large as well as small database such as educational database, data warehouses, medical databases. Vertical Fragmentation has an important impact in improving the performance of modern applications like document management, multimedia and hypermedia applications. With vertical partitioning, the disk access can be reduced by minimizing the access to irrelevant instance variables when executing the queries. In Present work FP-MAX data mining algorithm is used to extract frequent item set attributes of a large database table. The frequent accessed instance variables are grouped called Vertical fragments. The cost model is applied on Vertical fragments. The results are analyzed for small as well as large datasets in form of Max memory usage, Total time, cost and Frequent itemset count.

Keywords
Database, vertical fragmentation, frequent item set, data mining.

1. INTRODUCTION
Data volumes are increasing at high rate in the commercial world because of increase in number and complexity of transactions. In spite of advancement in computer technology, the data access performance is still a major issue in information system. A distributed database is a collection of data which logically belongs to same system but spread over the network at different sites. It is not necessary that the database should be geographically dispersed/distributed. The primary role of DBMS Design is fragmentation and allocation of fragments of the database [4]. The distribution of data across different sites involves proper fragmentation and allocation decisions. The distributed nature of data involves further new data like data fragmentation, partitioning, replication and allocation of fragments and replicas on different sites.

1.2 Fragmentation
Fragmentation is a design technique to divide a single relation or class of a database into two or more partitions such that their combination provides the original database without any loss of information .This reduces the amount of irrelevant data accessed by the applications of the database, thus reducing the number of disk accesses (Sharma Anshuman2008).The main aim of Fragmentation is to improve Reliability, Performance, Balanced Storage Capacity and Costs, Communication Costs, Security. The following information is used to decide fragmentation.

1.2.1 Vertical Fragmentation
Vertical Fragmentation is a physical database design technique which aims to improve the access performance of transactions. By Vertical fragmentation, the original relation is splitted into set of smaller physical files with each a subset of attributes of original relation. Normally database transactions only require access to subset of attributes. So if we split the relation into subsets and that closely match the user requirements than transactions access time reduces significantly. In this paper, the FP-MAX Vertical fragmentation algorithm, a technique for extracting frequent item sets of attributes is used. Frequent itemsets are then grouped as vertical class fragments. Selected cost model with parameters attribute length and query frequency is applied on fragments. Experimentally results are analyzed on applying FP-MAX algorithm and cost model on large and small datasets.

2. RELATED WORK
In [1] Affiner algorithm is based on the analysis and evaluation of the affinity between each attribute and the combination of high affinity attributes.Affiner algorithm works in five steps First generate the attribute usage matrix(AUM).Secondly ,generating attribute usage frequencies matrix. Then calculating the affinity matrix. After this, sorting the affinity matrix according to their affinity values in descending order. Finally deduce the fragmentation pattern from the ordered list. The vertical fragmentation is tested on the centralized row-oriented DBMS. For validation of the approach, the solution is compared to two existing algorithms Genetic algorithm and Apriori algorithm. Vertical fragmentation with two phase allocation method is used for partitioning a relation vertically and then allocating resultant fragments across different sites [2]. The benefits of distributed database and role of fragmentation are addressed in distributed environment. In distributed database users at a given site able to access data locally or remotely and retains control over the data at their own site [3]. An introduction of Distributed database system in two parts: In first part there is an exploration of distributed database environment and types of fragmentation .In second part Horizontal fragmentation technique of a relation using locality of precedence of its attributes. The proposed fragmentation technique can be applied at both initial as well as later stages of distributed database system for partitioning of relations [4]. Algorithm for Vertical fragmentation and Allocation is proposed in Distributed Object Database Systems Model which consists of simple attributes and simple methods for determining the efficiency and performance of Distributed systems. The main idea of this paper is to introduce a novel technique that considers re-fragmentation of main database, re-allocation of fragments when it is required and update operations on the server

database and its corresponding site fragments .With this approach, time taken for processing is decreased because the data can be accessed from sites databases rather than the server database [6]. Swarm intelligence algorithms are applied to present an algorithm for finding a solution for vertical fragmentation problem. In the proposed algorithm ,the relations are fragmented in such a way so that not only to make transaction processing at each site as much as localized as possible but also to reduce the cost of the operations. This proposed algorithm is based on the Ant collective behavior [7]. A new method to fragment a data warehouse vertically and horizontally based on statistic approach Principal components analysis (PCA)is proposed[8]. The proposed Vertical Partitioning Algorithm using grouping approach starts from the attribute affinity matrix and generates initial groups that are based on the affinity values between attributes. Then initial groups are merged to produce final groups which will represent the fragments [9]. A cost-based approach for horizontal and vertical fragmentation is presented to address the fragmentation problem of complex value databases that covers the common aspects of object-oriented databases, object relational databases and XML [10]. Vertical Fragmentation and allocation are addressed simultaneously in the context to the relational model [11]. Implementation of Heuristic Algorithm that uses an objective function which takes information about the administered dates in a distributed database and evaluate all the scheme of the database vertical fragmentation [12]. A technique is proposed for measuring the performance of object horizontal fragments that are placed at different sites [13]. A technique of Vertical Fragmentation of Views in multidimensional databases is proposed. The main method to extract the useful information from element data in multidimensional database is aggregation [14].

3. PROPOSED WORK

In part of few years Database Technologies goes on hike researches continuously exploring the technique of efficient Vertical Partitioning. We find that:

- FP-MAX Vertical partitioning technique earlier had been tested with a cost model which was based on query frequency parameter but in present work FP-MAX is tested with a cost model having parameters like attribute length and query frequency.
- Cost is computed for large and small datasets by varying the Minimum support value and the parameters value and the performance is analyzed.

3.1 Objectives

The objective of my dissertation is:

- To check the performance of Vertical partitioning on some large as well as small database by applying the selected cost model using specified parameters like attribute length and query frequency.
- The performance is analyzed in form of maximum memory usage, Total time, Frequent itemsets count and cost.

3.2 Research Methodology

- On both large and small datasets, FP-MAX data mining algorithm is applied.
- The frequent item set is generated based on particular minimum support set.
- These frequent item sets are grouped together.
- Vertical Fragments are generated.
- On these vertical fragments, a Cost model is applied using parameters: attribute length, query frequency.
- Performance is evaluated based on the selected Cost model used.
- Results are analyzed.

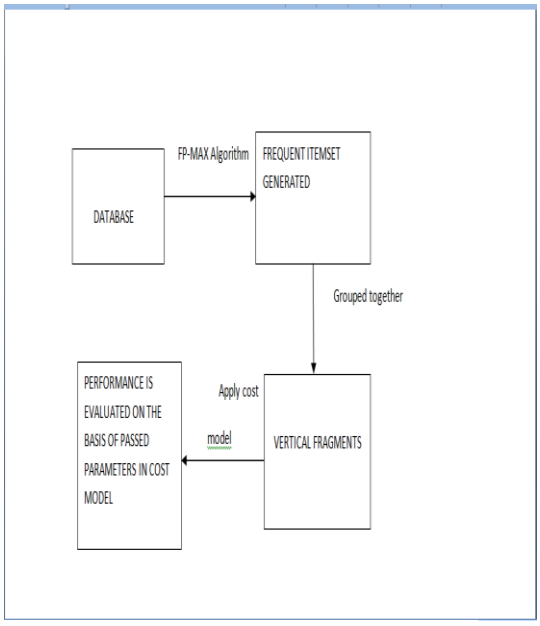

Figure 1: Methodology

4. RESULTS AND DISCUSSIONS

4.1 Case1: Supermarket Dataset

In the figures shown below,for min support values 0.3 to 0.7 the results are obtained in form of max memory usage, total time, cost and frequent itemset count.

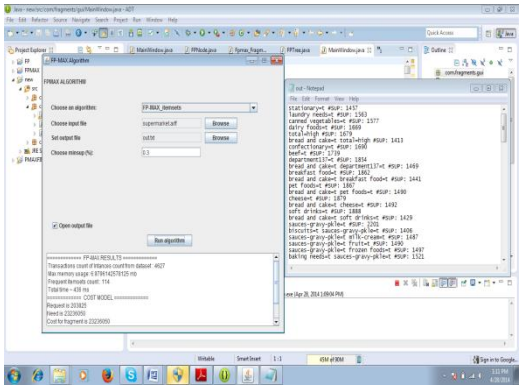

Figure 2: MIN Support 0.3

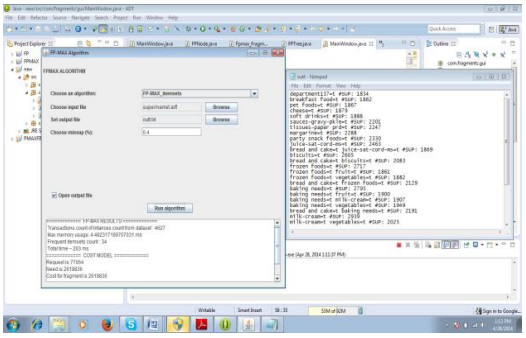

Figure 3: MIN Support 0.4

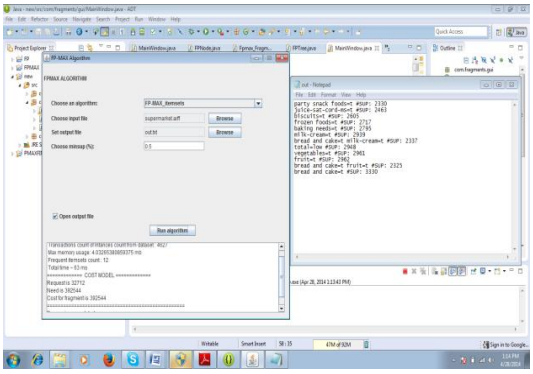

Figure 4: MIN Support 0.5

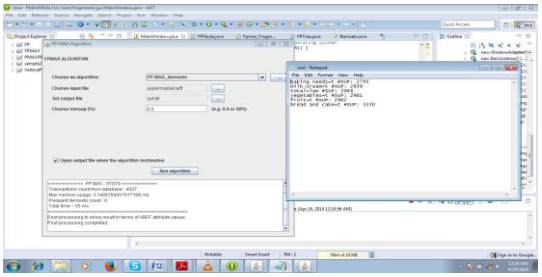

Figure 5: MIN Support 0.6

4.1.1 *Graphical Representation*

The following figures shows the graphical representation of results of supermarket dataset:

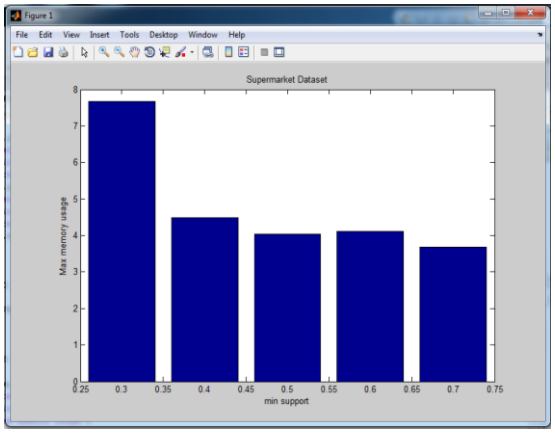

Figure 6: MAX Memory Usage (Supermarket)

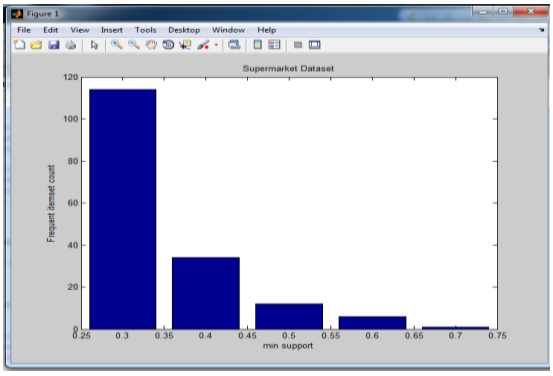

Figure 7: GRAPH- Frequent Itemset Count (Supermarket)

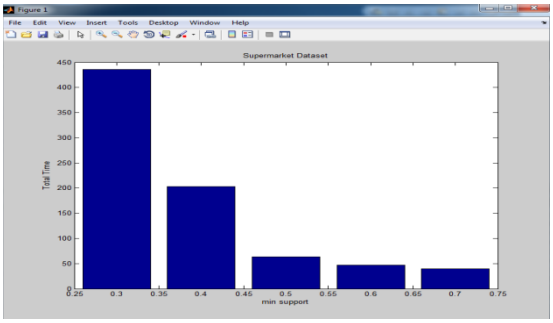

Figure 8: GRAPH Total Time (Supermarket)

4.2 Case 2: Vote Dataset

In the figures shown below,for min support values 0.3 to 0.7 the results are obtained in form of max memory usage, total time, cost and frequent itemset count

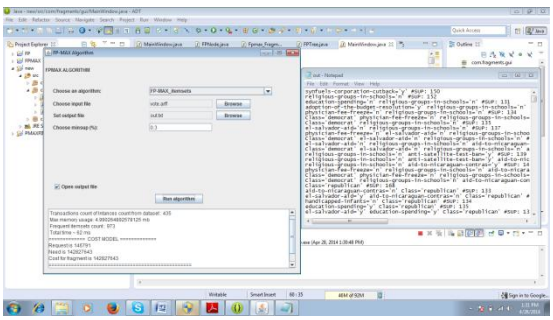

Figure 9: MIN Support 0.3

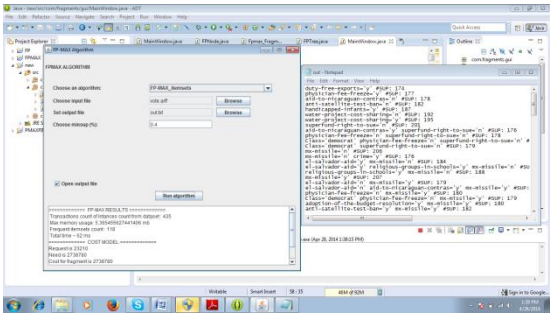

Figure 10: MIN Support 0.4

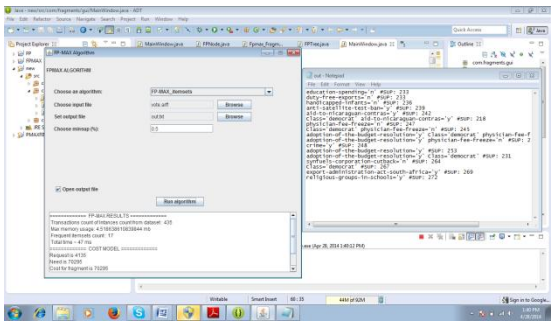

Figure 11: MIN Support 0.5

4.2.1 Graphical Representation

The following figures shows the graphical representation of results of Vote dataset:

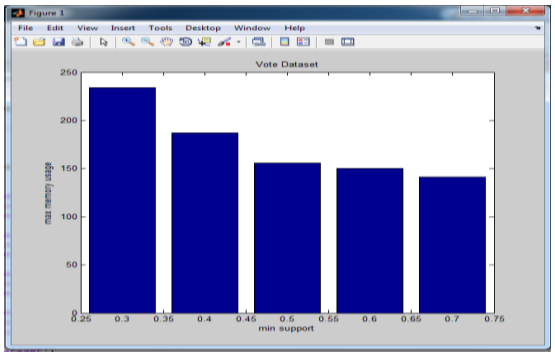

Figure 12: GRAPH- Max Memory Usage (VOTE)

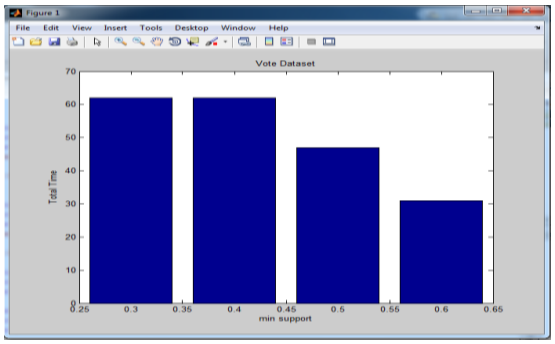

Figure 13: GRAPH- Total Time (Vote)

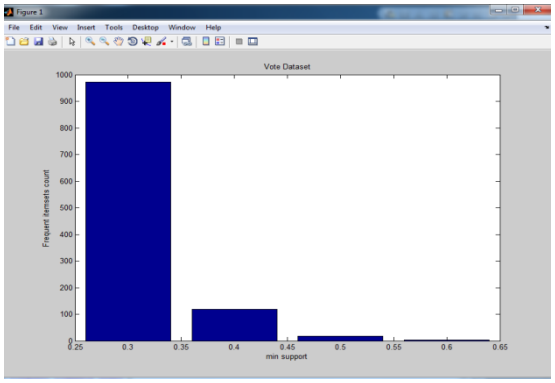

Figure 14: Graph- Frequent Itemset Count (VOTE)

4.3 Case 3: Mushroom Dataset

In the figures shown below,for min support values 0.3 to 0.7 the results are obtained in form of max memory usage, total time, cost and frequent itemset count.

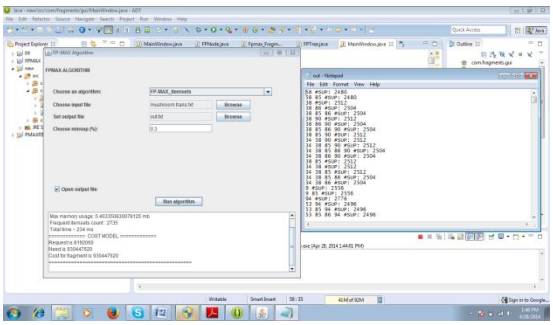

Figure 15: MIN Support 0.3

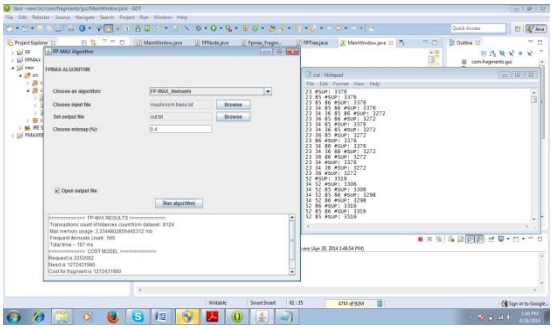

Figure 16: MIN Support 0.4

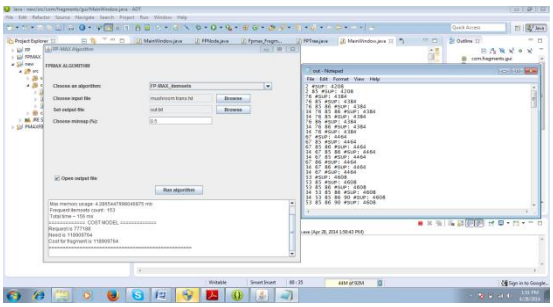

Figure 17: MIN Support 0.5

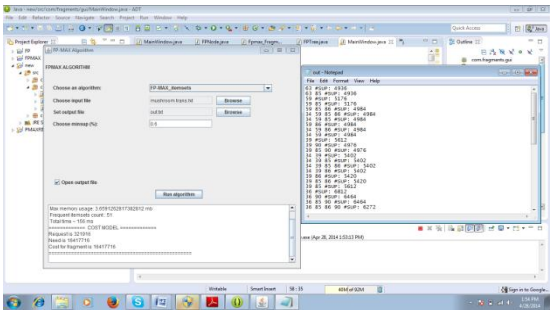

Figure 18: MIN Support 0.6

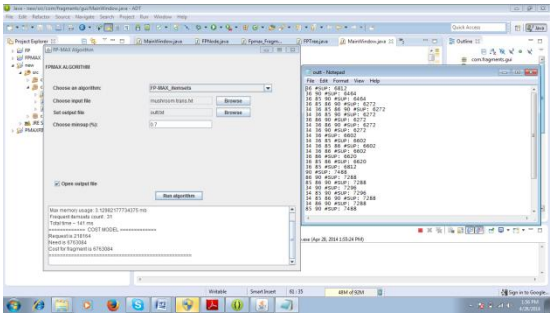

Figure 19: MIN Support 0.7

4.3.1 Graphical Representation

The following figures shows the graphical representation of results of Mushroom Dataset:

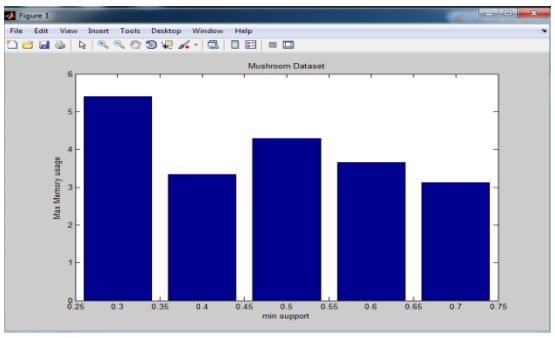

Figure 20: GRAPH- Max Memory Usage

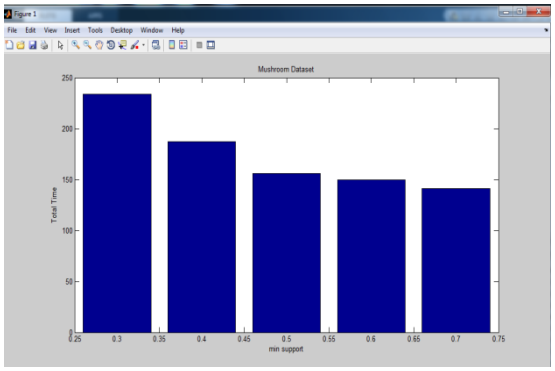

Figure 21: GRAPH- Frequent Itemset Count

4.4 Case 4: Weather Nominal Dataset

In the figures shown below, for min support values 0.3 to 0.7 the results are obtained in form of max memory usage,total time,cost and frequent itemset count.

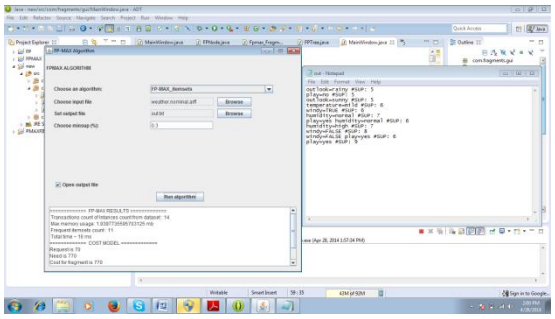

Figure 22: MIN SUPPORT 0.3

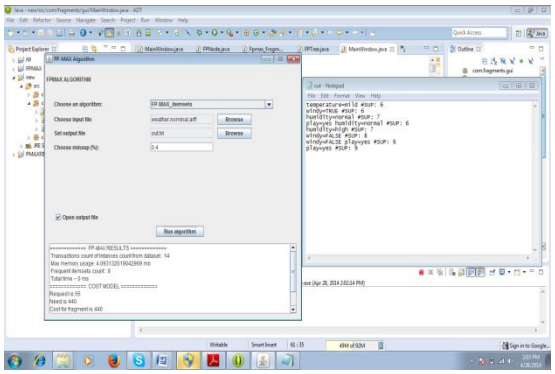

Figure 23: MIN SUPPORT 0.4

4.4.1 Graphical Representation

The following figures shows the graphical representation of results of Weather Nominal Dataset:

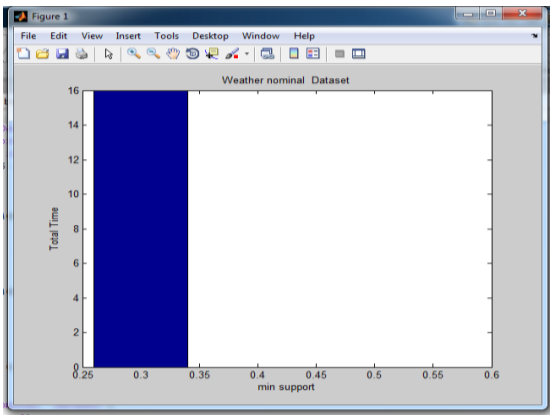

Figure 24: GRAPH- Total Time (Weather Nominal)

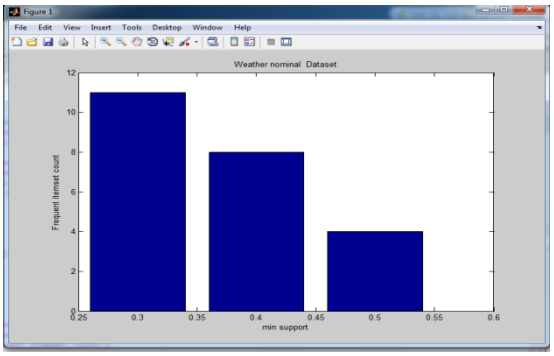

Figure 25: GRAPH- Frequent Memory Usage (Weather Nominal)

5. CONCLUSION

In my present work ,Vertical partitioning of small as well as large datasets like Mushroom dataset, Supermarket dataset, Vote dataset and Weather –Nominal dataset is done using the FP-MAX algorithm and cost model having parameters attribute length and query frequency is applied. From experimental results, for dense datasets Mushroom and Supermarket, a large number of MFI's are mined for higher values of support. For small datasets Vote and Weather-Nominal, MFI's are less i.e., frequent item sets count are less. From these experimental results on real datasets, my conclusions are valid but may not hold for some skewed real datasets.

6. FUTURE SCOPE

In future, further analysis and experiments will be done to understand the impact of skewness of the data. Also, parameters like APL(Average pattern length) and ATL(Average transaction length) can be explored for cost evaluation. My work did not cover horizontal or hybrid fragmentation of datasets which plays prominent role in maintenance overhead. So in future,I plan to extend my work to include horizontal or Hybrid partitioning algorithms to deal with fragmentation and allocation as well.

7. REFERENCES

[1] Adrian Runceanu(2004), "Towards Vertical Fragmentation in distributed databases".

[2] Amer Ali A. and Abdalla Hassan I. (2012) "An

Integrated Design Scheme for Performance Optimization in Distributed Environments “, International Conference on Education and e-Learning Innnovations.

[3] Ankur Bhardwaj et.al,(2012) “Role of Fragmentation in distributed database system” ,International Journal of networking & Parallel Computing Volume 1, Issue 1, September 2012.

[4] Bhuyar P.R. and Gawanda A.D.,(2012) “Distributed Database: Fragmentation and Allocation”, Journal of data Mining and Knowledge Discovery,Volume3.

[5] Bouakkaz M. ,Ouinten Y., Ziani B. (2012) “Vertical Fragmentation of Datawarehouses using FP-MAX Algorithm” ,International Conference on Innovations in Information Technology.

[6] Chaalel,Belbachir,(2013) “An Optimized Vertical Fragmentation Approach”, International Journal of innovative technology and exploring Engineering(IJITEE),Volume-3,Sept 2013.

[7] C.I.Ezeife and Ken Barker(1996), “Vertical fragmentation for advanced object models in a distributed object based system.”

[8] Ezcife.C.I.Zheng Jian “Measuring the performance of database Object Horizontal Fragmentation Schemes ”

[9] Golfarelli Matteo ,Maio Davio and Rizzi Stefano “ Applying Vertical Fragmentation Techniques in Logical Design of Multidimensional Databases”.

[10] Gosta Grahne and Jianfei Zhu(2002), “High performance Mining of maximal frequent itemsets”.

[11] Gupta Surabhi ,Panda Shruti (2012) “Vertical Fragmentation and Re-Fragmentation in Distributed Object Relational Database Systems-(Update Queries Included)” , International Journal of Engineering Research and Development.

[12] Hichame Chaalel and Hafida Belbachir(2013) “ An Optimized vertical fragmentation Approach” ,International Journal of Innovative technology and exploring engineering,Volume 3.

[13] Hui Ma and Markus Kirchberg(2008), “Cost based Fragmentation for distributed complex value databases”.

[14] Jiawei Han et.al,(2000), “Mining frequent patterns without candidate generation.”

[15] Ma Hui ,Scherve Klaus-Dieter and Kirchberg Markus (2006) “ A Heuristic approach to Vertical fragmentation Incorporating Query Information”.

BOOKS

Ozsu. M. Tamer. (2011) “Distributed Database Design”, Principles of Distributed Database Systems, Third Edition.

Sharma Anshuman. (2008) Fundamentals of DBMS, Lakhanpal Publishers, INDIA