# Novel Most Frequent Pattern Mining Approach Using Distributed Computing Environment

Parag M. Moteria
*PhD Scholar,*
*School of Computer Science,*
*R K University, Rajkot.*

Dr. Y. R. Ghodasara
*Associate Professor*

## Abstract

Frequent patterns are frequent data set in transactional data set, play an essential role in mining associations, correlations and many other interesting relationships among data that leads knowledge discovery and helps in many business decision making processes [1]. Data mining is a very basic operational technique in knowledge discovery and decision making processes. Frequent pattern mining techniques have become necessary for massive amount datasets in distributed data mining approach using distributed computing environment. This paper discuss novel approach for efficient and scalable distributed algorithm for most frequent itemsets generation on Boolean types of single dimensional and single level data mining using distributed computing environments in transactional dataset.



## *1. ITRODUCTION*

Data mining is the process of finding interesting trends or patterns in large datasets to steer decision about future activities. Knowledge discovery in databases and data mining helps to extract useful information from raw data. Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases or transactional dataset, such as association rules, correlations, sequences, episodes, classifiers, clusters. Frequent pattern mining is one of the most important and well researched techniques of data mining. Association rules can be useful for decisions concerning product pricing, promotions, store layout and many others [2]. Thus, frequent pattern mining has become an important data mining task and a focused theme in data mining research [3]. Our novel most frequent pattern mining approach using distributed computing environment is data mining where computations are spread over many independent nodes with central transactional dataset. This paper describes theoretical approach to mine most frequent pattern itemset without using user threshold in distributed computing environment.

## *2. MOST FREQUENT PATTERN MINING (MFPM) APPROACH*

The proposed novel approach to design efficient and scalable MFPM in distributed computing environment using transactional dataset is as under [4].

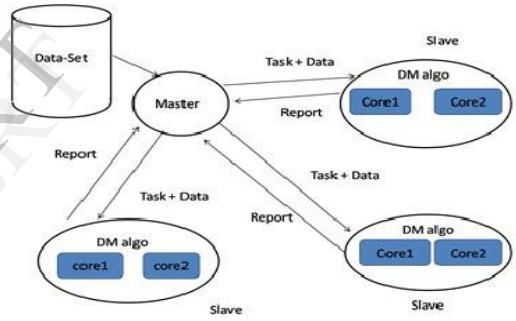


**Figure 1**

Assume that, we have one server (master) say S and n numbers of nodes (slaves) say $N_i$ (where i=2, 3, … , n. Here, n equals to total number of different items in itemset. Each itemset consists unique items per transaction.

Consider following transactional dataset in lexicographic order [1]:

**Table 1**

| TID | List of ITEM IDs |
|---|---|
| T100 | I1, I2, 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 |

Step 1:
Transactional dataset resides into server.

Step 2:
Build cardinality table of each itemset by server and store maximum number of different items in itemset in variable say n.

| Table 2 | |
|---|---|
| Cardinality | List of ITEM IDs |
| 3 | I1, I2, I5 |
| 2 | I2, I4 |
| 2 | I2, I3 |
| 3 | I1, I2, I4 |
| 2 | I1, I3 |
| 2 | I2, I3 |
| 2 | I1, I3 |
| 5 | I1, I2, I3, I5 |
| 2 | I1, I2, I3 |

Step3:
Server sends maximum number of items in itemset to each node $N_i$. Each node $N_i$ generates flags say $f_p$ (where p = 1, 2, … , ${}^nC_i$).
For example,
total number of different items in itemset n=5
Cardinality of itemset=4,
Node $N_4$ generates $f_p$ (Here, p=1, 2, … , ${}^5C_4 = 5$)

Step 4:
Server reads cardinality from table2, depending upon cardinality corresponding itemset send to node $N_i$.

Step 5:
Node $N_i$ scans itemset, flag $f_p$ sets on and code this combination of itemset say $comb_p$ (where p = 1, 2, … , ${}^nC_i$). Each flag and combination of itemsets are predetermined. If all predetermined flags set on, stop message raised by node $N_i$ to server followed with computed result by node $N_i$.
For Example,
Node $N_4$ scan itemset like {I1, I2, I3, I4} set $f_1$ on and generate code $comb_1$ that represents {I1, I2, I3, I4}.

Step 6:
When end of last itemset reach, appropriate message send by server to each node $N_i$, except nodes those have been submitted their result.

Step 7:
Now, each node $N_i$ compute intersection operation on $comb_p$ that corresponding flags set on. Result sends to server.

Step 8:
Compute intersection on results submitted my each node gives final result.

Case i
If final result is NULL, then make union each result. It generates most frequent itemset in transactional dataset.

Case ii)
Omit resultant set with cardinality with one and compute intersection on remaining results send by each node $N_i$. It generates most frequent itemset in transactional dataset.

Case iii)
Otherwise we get final result as most frequent itemset.

As per above steps,
$N_2$ generates result – {I1, I2, I3}
$N_3$ generates result – {I1, I2}
$N_4$ generates result – {I1, I2, I3, I5}
$N_5$ generates result – no itemset with cardinality five

Final result equals to {I1, I2}, which is most frequent itemset in transactional dataset.

## *3. Conclusion*

Our novel approach for most frequent pattern mining using distributed computing environment may build efficient and scalable distributed mining approach to enhance strength to discover knowledge. It may help to determine most frequent pattern item in transactional dataset. This novel approach is developed with theoretical background. Hence, implementation is needed as our future work.

## References

[1] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques - Third Edition, ELSEVIER Morgan Kaufman Publisher, July 6, 2011

[2] D. N. Goswami, Anshu Chaturvedi, C. S. Raghuvanshi, "Frequent Pattern Mining Using Record Filter Approach",International Journal of Computer Science, Vol. 7, Issue 4, No 7, July 2010, pp 38-43

[3] Jiawei Han, Hong Cheng, Dong Xin, Xifeng Yan, “Frequent pattern mining: current status and future directions”, Springer Science+Business Media, LLC 2007, pp 55-86

[4] Anjan K Koundinya,Srinath N K,K A K Sharma, Kiran Kumar, Madhu M N and Kiran U Shanbag, "Map/Reduce Design And Implementation Of Apriorialgorithm For Handling Voluminous Data-Sets", ACIJ, Vol.3, No.6, November 2012, pp 29-39

[5] Lamine M. Aouad, Nhien-An Le-Khac and Tahar M. Kechadi, "Distributed Frequent Itemsets Mining in Heterogeneous Platforms", Journal of Engineering, Computing and Architecture, Vol. 1, Issue 2, 2007

[6] Bagrudeen Bazeer Ahamed and Shanmugasundaram Hariharan, "A Survey On Distributed Data Mining Process Via Grid", International Journal of Database Theory and Application, Vol. 4, No. 3, September 2011, pp 77-90

[7] Goswami D.N., Chaturvedi Anshu., Raghuvanshi C.S.,"An Algorithm for Frequent Pattern Mining Based On Apriori", IJCSE, Vol. 02, No. 04, 2010, pp 942-947

[8] Sunil Joshi, R S Jadon and R C Jain, "A Frame Work for Frequent Pattern Mining Using Dynamic Function", IJCSI, Vol. 8, Issue 3, No. 1, May 2011, pp 141-147

[9] Sumithra, R.; Paul, S.; , "Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery," Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on , vol., no., 29-31 July 2010, pp 1-5