ISSN (Online): 2455-7838
SJIF Impact Factor (2017): 5.705

EPRA International Journal of

Research & Development

(IJRD)

Monthly Peer Reviewed & Indexed
International Online Journal

Volume: 3, Issue:11,November 2018

Published By :
EPRA Journals

SJIF Impact Factor: 5.705 Volume: 3 | Issue: 11 | November | 2018 ISSN: 2455-7838(Online)

EPRA International Journal of Research and Development (IJRD)

FREQUENT ITEMSET MINING USING HADOOP

Archana Nikam[1]

[1]Pimpri Chinchwad College of Engineering, Nigdi Pradhikaran, Akurdi, Pune

Sandhya Waghere[2]

[2]Pimpri Chinchwad College of Engineering, Nigdi Pradhikaran, Akurdi, Pune

ABSTRACT

*Searching frequent item-sets in large size heterogeneous databases in minimal time is considered as one of the most important data mining problem. As a solution of this problem, various algorithms have been proposed to speed up execution. Most of the recent proposed algorithms focused on parallelizing the workload using large number of machine in distributed computational environment like Map Reduce framework. A few of them are actually capable to determine the appropriate number of required computing computers, considering workload balancing and execution efficiency. But internally not capable to determine exact number of required iteration for any large size datasets in advance to find out the frequent item-set based on iterative sampling. In this paper, we propose an improved and compact algorithm (ICA) for finding frequent item-set in minimal time, using distributed computational environment. It is also capable of determining the exact number of internal iteration required for any large size datasets whether data is in structured or unstructured format.**[2]***

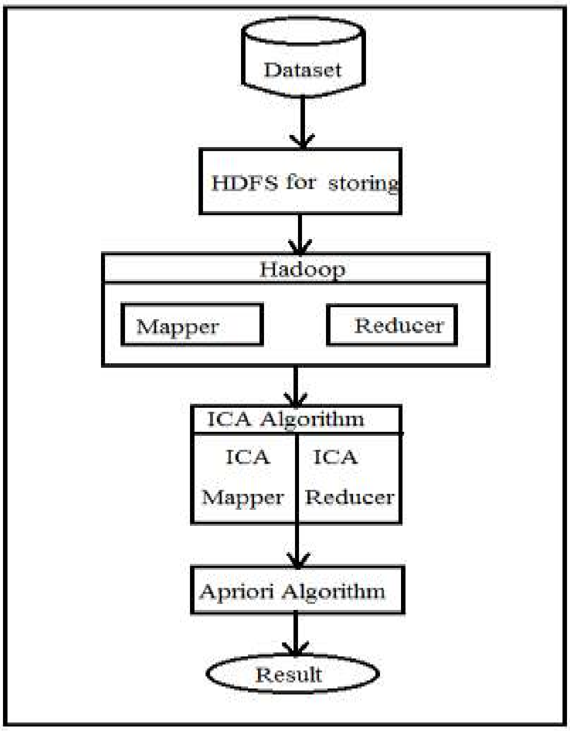
KEY WORDS: *Frequent itemset mining, ICA,mapreduce.*

1. INTRODUCTION

Frequent item-set mining is one of the most important techniques to find out frequent item-sets in data mining. 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 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 which are occurred many times frequently in a given data sets.

Traditional data mining tools fails to extract frequent item-sets when the size of transactional database is too large to compute. In Big Data era, we need a new approach 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 overcome this problem MapReduce model come into existence, originally proposed by Google.[2]

2.SYSTEM ARCHITECTURE



3. ALGORITHM S/TECHNIQUE

3.1 Hadoop MapReduce Programming model

Hadoop Framework is open source framework.it is suitable for large data, and it is scalable.it provide fault tolerance capability Hadoop provide map reduce programming model this map reduce model consist of 4 stages:

Input Split
Mapping
Shuffling
Reducing

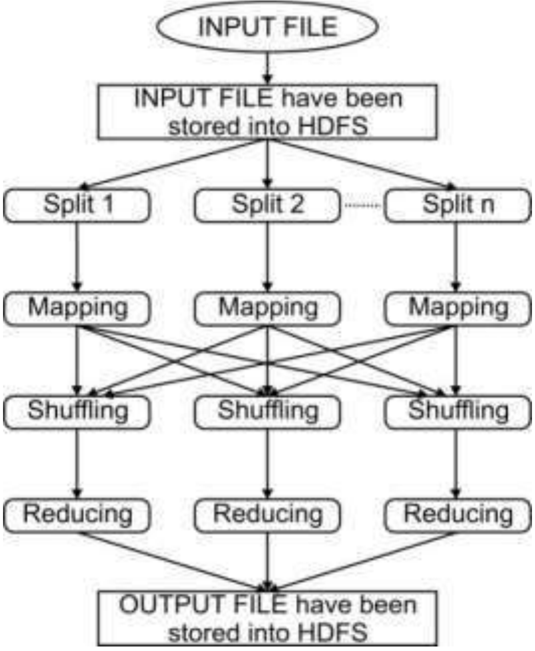


Figure 1: Hadoop Mapreduce Programming Model[2]

3.2 Improved And Compact Algorithm

It convert whole datasets into manageable data chucks (sample data sets) and send them to others nodes in distributed computational environment . Each and every nodes consisting of Mapper receive the data chunks or sample data in terms of key value. Here Mapper reads the input in accordance with local minimum support and assigned value after tokenization to the keyword that occurs first time. Mapper function generates the output in terms of key-value on each and every node. These Mapper output is passed to combiner function where shuffle sort is used to shuffle all the key-values. A new key-value form is generated after combining the values and these values is further aggregated by reducer function. Reducer also checks the occurring frequency of each and individual item-sets to the global minimum support values. [2]

3.3 Apriori Algorithm

Apriori is the classic algorithm for finding frequent item sets. It was proposed by Agrawal and Srikant. Apriori uses bottom-up approach to find frequent item sets. Frequent subsets are added one at a time (join). This process is known as candidate generation. The item sets are checked against the minimum support. If the item set's support count is less than the minimum support then the particular item set will be removed and this process is known as pruning. Apriori uses an efficient search methodology called downward-closure property of support also called anti- monotonicity. This is also known as Apriori property which states that the subsets of the frequent item set is also frequent. Likewise for infrequent subsets also.[1]

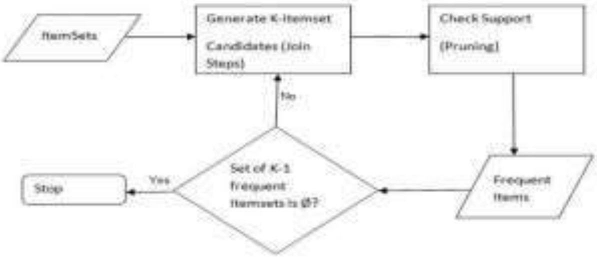

ItemSets
Generate K-itemset Candidates (Join Steps)
Check Support (Pruning)
No
Stop
Yes
Set of K-1 frequent itemsets is Ø?
Frequent Items

Figure 2: Apriori Algorithm

Table1.Comparative study Of Algorithms

Algorithms	Efficiency	Quality	Scalability	Relevance	Accuracy
Apriori	Less	Low	No	Yes	Less
Eclat	Less	High	No	Yes	Less
FP Growth	Less	High	No	Yes	Less
ICA	High	Highest	Yes	Yes	More

.

4.APPLICATION

1. Financial Data Analysis
2. Retail Industry
3. Telecommunication Industry
4. Biological Data Analysis
5. Intrusion Detection
6. customer credit policy analysis

5.CONCLUSION

We have validated and demonstrated the effectiveness of the proposed approach on transactional data-set under Hadoop architecture.

We use is an apriori algorithm for frequent item set mining and association rule learning over transactional databases. 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.

6.REFERENCES

1. *Dr Ramah Sivakumar,J.G.R.Sathiaseelan,"A Performance Based EmpiricalStudy Of Frequent ItemSet Mining Algorithm "IEEE,2017.*
2. *Dr.Ruchi Agarwal,Sunny Singh,Satvik Vats,"Implementation of an Improved Algorithm for Frequent Itemset Mining using Hadoop", International Conference on Computing, Communication and Automation (ICCCA2016) .*
3. *J. Xie, et al., "Improving mapreduce performance through data placement in heterogeneous hadoop clusters," in Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on, 2010, pp. 1-9.*
4. *A. Thusoo, et al., "Hive-a petabyte scale data warehouse using hadoop," in Data Engineering (ICDE), 2010 IEEE 26th International Conference on, 2010, pp. 996-1005*
5. *Singh and S. Agrawal, "A review of research on MapReduce scheduling algorithms inHadoop," in 2015 International Conference on Computing,Communication & Automation (ICCCA), 2015, pp. 637-642.*