Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Generating Frequent Pattern Mining From Big Datasets Using Hybrid Apriori And Genetic Algorithm

[1]B.Bazeer Ahamed , [2]D.Yuvaraj

[1]Department of IT University of Technology and Applied Sciecnes Al Musannah Sultanate of Oman

[2]Department of Computer Science Cihan University – Duhok Kurdistan Region, Iraq

Abstract— A major research field in data mining is the frequent pattern of mining. It has attracted the attention of many researchers since its launch. The generation and collection of data diagonally improve all areas exponentially in scale. Data mining is the use of advanced analytics to determine new information in the method of patterns, trends, and correlations in vast volumes of data. Information retrieval and decision-making involve a flexible and effective way of processing and extracting from Big Data relevant information. One of the most widely used methods for extracting useful knowledge from data is Regular Itemset Mining. However, when this approach is applied to Big Data, the combined outburst of the candidate object sets has become a problem. Recent advances in parallel programming have created excellent tools to address this problem. However, these tools have their technical disadvantages, such as unbiased sharing of data and intercom costs. In our work, we examine the applicability of Frequent Mining Items in the Map-Reduce framework. This method is optimized for use in extremely large datasets. Our approach is similar to FP-growing but uses a different data structure based on algebraic topology.

Keywords—Frequent Pattern Mining, Frequent Item Sets, Hybrid Apriori,Big data

1. INTRODUCTION

Data mining included a broad understanding of the intrigue within the IT industry and civilization as untouched at the display, not to mention the widespread accessibility of massive amounts of information, which is also an imminent requirement to spiral this information into practical data and facts. During this Big Data age, all the people, in one way or a different throw in producing data. It could be prearranged, which is created by several business-related solicitations and naturally stored as records and fields with a definite structure. Semi-structured data generated by feed sensors, stock market feedback and transactions and network traffic could also be security-related data. Semi-structured data typically have metadata showing their configuration. We could also have unstructured Big Data, which is naturally produced by the community employing social networks

and this includes different kinds of data like images, video and audio. In addition to this array of data structures, Big-Data can be created in a huge amount at a fast speed without any clear indication of the actuality of the data. With these attributes, the data has exceeded the capacity to be discarded and used by a number of ordinary frameworks. [1]. The cost of the information shall be realized from the primary to the last experience, taking into account the deliberation that the value of a few focal points of the information shall be taken into account by a very rapid rain check. Regularly more, achievement has become dependent on how quickly and efficiently the petabytes of data they collect can be converted into actionable information [2].

A.FREQUENT ITEMSET EXTRACTION

Frequent itemsets are sequences that exists frequently in data-set. Discovering frequent itemsets performs a crucial role in extracting correlations and several further motivating connections between data. Consider a sample, a list of items, like paste and brush that seem repeatedly self-possessed within transaction data is again frequent item-set. Consider a situation such as starting with the purchase of a computer, at that point the data card, and finally the write drive, and in the event that this design occurs as often as possible in the past information set, at that point the design may be a frequent item set.[15]. Substructures may include dissimilar essential methods, like sub-graphs, subtrees, that are pooled with itemsets. Extraction of Frequent itemsets examines for repetitive interactions in the assumed data set. An analysis of the market basket is the initial use of the extracting of frequent items to obtain association rules. Big Data Analytics will show businesses the way to discover hidden features of this information and use these designs to determine the plausibility of events approaching. Data Analysis may well be reminiscent of what has been reused to summarize what has been shown. It can be extrapolative, using a range of measurable, displaying, information mining, as well as machine-learning methods to monitor current and old information, thus empowering businesses to create expectations about what is inescapable[16]. There is a promising form of data review called Prescriptive Analytics that proposes one or more courses of activity and shows the likely outcome of each choice.

A few algorithms are available for viewing mining designs and the issue was that finding a competent calculation and extracting information from database data could be a troublesome errand. Knowledge extraction from the information set is the assignment of an understanding of the space issue and an extrication of the designs of the use. This stimulated the search for an appropriate calculation to mine a visitor design from huge datasets [17]. Gigantic use work has been done to compare the results of the proposed calculations with those of the previous algorithm..

Generating every frequent itemset is typically extremely large, which does not require all applications. The subset that is needed by these applications regularly includes particularly a minute number of itemsets. Thus more time remains spent in considering all unwanted frequent itemsets to extract frequent itemsets.[3]. In addition to this, memory is also wasted in storing all unimportant frequent itemsets. So constraints can be introduced to remove these unimportant itemsets. Consequently, here it is a necessity to recommend a competent method to discover frequent itemsets by using Big Dataset utilizing restraints. Virtually altogether Frequent Item-Set extraction methods, frequent 1-itemsets are produced to determine support value (occurrences) of every item present in

the complete data set. The main goals of this work to extract cumulative frequent itemsets from multiple files and designing the suitable Map-Reduce computing model for parallel FP Tree called Big FM.

II.LITERATURE SURVEY

Frequent patterns are itemsets, sub-series, which appear in an information set that satisfies the client with the least back esteem demonstrated by the event, not below the desired client boundary. This method is applied to trade data of a large trading company also the effectiveness of the method has proved. Agarwal et al projected two strategies Apriori and Apriori Tid to shed light on the issue of extricating visit designs, and then combined these two algorithms. Depict the comparative analysis and classification of accessible data mining strategies. Advances of competent methods for pulling out numerous types of information at manifold generalization levels are available. DB-Miner has established a framework for the extraction of social information at Simon-Fraser College for the extraction of compound types of rules at various concept levels, such as attribute-based rules, discriminating rules, association extraction rules, classification rules, etc.

Babi et.al [4] , suggested FP-growth, for extracting an entire set of frequent-itemsets, the proficiency of mining is accomplished with three methods: a huge data set is trampled into an extremely shortened abundant reduced data structure that evades expensive, recurrent data set scans, “FP-tree" established mining espouses a pattern portion growth process to dodge the expensive production of an enormous number of candidate sets. Cheng-Yue Chang et al., explore a new method called Segmented Progressive Filter that segment data set into sub data sets in such a manner that items in every sub data set would have whichever joint begin-time or the joint end-time. For every sub-data set, SPF gradually strainers candidate 2-itemsets by way of growing purifying limits whichever forward or backwards in time. Here mentioned attribute licenses SPF of acceptance the scan reduction procedure through creating candidate k-itemsets (k >2) of candidate (two) 2-itemsets.
Bova et al,[5] selected CATs tree extended the idea of FP- Tree in the process of improving the storage compression and FP mining without generating the candidate itemsets and allow single pass over the data set and insertion and deletion of transactions in the tree can be done very efficiently. Choi et al.[6] suggested two efficient algorithms to identify sequential patterns from d-dimensional sequence data. First algorithm Apriori-MD algorithm is an updated Apriori method to extract sequential patterns from a multidimensional data set. It uses tree structures termed candidate tree similar to a hash tree. Another method is prefixed span algorithm which is an updated version of the prefix span algorithm. The classical Apriori and Apriori Tid algorithms exist used to construct the frequent item-set, the main disadvantage is that it consumes more time for scanning the dataset. To avoid this, Gangavarapu, T., & Patil [7] generates a highly efficient Apriori Tid algorithm, for reducing the scanning time.

Sornalakshmi et.al.[14] improves the Apriori method, by using the division operation, governance strategy and introduction of Hash technology, scanning time of data set is reduced. The Apriori method is the utmost recognized algorithm for extraction of frequent itemsets and numerous executions of the Apriori algorithm have been informed and assessed.

Hamdad et al.,[8] suggested an Apriori algorithm toward mining spatial, temporal patterns. Apriori like algorithm has been generated based on candidate generation and a comparing function. Apriori like algorithms in the case of approximate search generates a large number of sequences. Khotimah et al.,[9] describe Rough Apriori Algorithm which is the combination of Apriori Algorithm and Rough Set theory for mining frequent patterns.

Liu et al, [10]proposed an enhanced Apriori method entails less memory decreases the occurrence of computer Input and Output the efficiency has enhanced. Mining progression can be exhibited as a poison binomial distribution by M. Malek and H. Kadima et al. [11] that could proficiently and correctly determine frequent itemsets in a large unclear dataset. The probabilistic algorithm discovers association rules and frequent itemsets by using the perception of repetitive medians to compute the thinning out in the transaction set for every itemset. Repetitive medians are used to compute the highest number of collective transactions for any two itemsets so that it reduces the scanning of database and generation of unsuccessful candidate sets.

III.FPTREE USING MAPREDUCE- BIGFM

We propose an algorithm that contains two novel techniques for extracting frequent itemsets in concurrently on top of Map-Reduce frame-work, in which frequency limits could be little. Present a strategy for the time being, which could be a cross-breed strategy that fundamentally employs the FP Tree-dependent approach to retrieve frequent item-sets for size 'k' and then changes through Eclatas as soon as the expected data sets are stored in memory. Initially, extracting for k-Frequent-Itemsets could previously impartible [Neysiani et.al] [12]. Second, on the whole of the mappers necessitate the entire dataset placed it on memory to extract subtrees [13]. The following figure 1 explains the functionality of map-reduce.

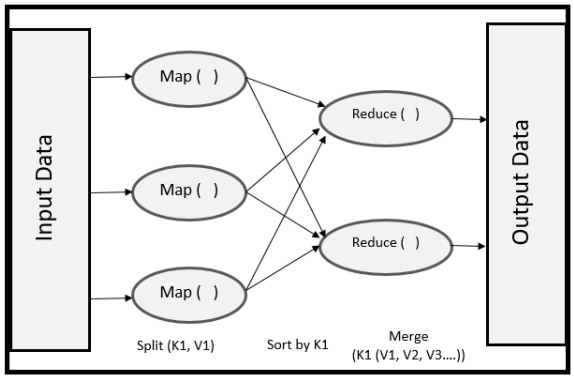

Figure.1 Functionality of Map Reduce

1. 1.. It could be accomplished by acclimatizing Word-Counting predicament concerning document records, i.e., specific mapper gains apart from the report, including testimony some item-sets in support of these we desire to find the support (count). Each reducer combines all nearby regularities and declarations with all inclusive visits[17].
2. Pronouncement of prospective Extensions: After calculating prefixes, subsequently, step is calculating permissible augmentations, i.e., fetching tid-lists concerning (k+1)-FIs. This could be done in terms of how word-counting is performed.
3. Subtree-Mining: To end with, the mappers' exertion upon entity prefixes clusters. Prefix clusters describe a restricted data set that placed into memory[18]. The extraction fraction after that make use of different sets to extract restricted data set for frequent itemsets utilizing dfs that is depth-first search. The iterative preparation shall continue until a set of k Visit Things has arrived, which are adequate for the diminutive[23][24]. To reduce the network passage, we preset the extracted items using a condensed tire string illustration for each batch of items[25].

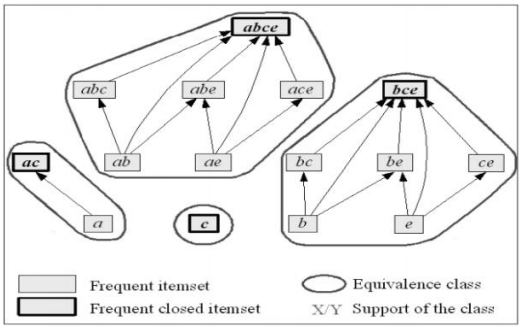

Figure 2. Frequent Equivalence Classes

IV EXPERIMENTAL RESULTS AND ANALYSIS

The 3 parallel Association Rule Mining Algorithms for extraction of Frequent Item-Sets and Association-Rules from Big Data Sets. The Algorithms are Big FM, Parallel Apriori using Genetic Algorithm for Optimization and Partition based Apriori algorithms. These algorithms, when implemented, performed efficiently when compared to corresponding conventional FP Growth, Apriori algorithms that extract Frequent Item - Sets and Association - Rules of traditional data sets. Here we would like to compare the three Parallel Association Rule Mining algorithms BigFM, Parallel hybrid Apriori using Genetic Algorithm and Partition based Apriori algorithm and find how these algorithms generate the association rules and perform execution efficiency based on parameters time complexity and space complexity

A. BIGFM Vs PARALLEL HYBRID APRIORI

This section observed the execution time of algorithms “BigFM and hybrid Apriori for different minimum support value. We analyzed the execution time by changing the least back values and by increasing the number of mappers. A small Hadoop-2.6.0 cluster is introduced with five hubs; all are running Ubuntu 14.04. One hub is assigned as Title Hub and four other hubs serve as Information Hubs. Title Hub is designed with 4 centers and 4 GB of memory running in a virtualized

environment on a window host. [19]. Two data hubs run on an isolated physical machine, each with 4 centers and 2 GB of memory. The other two Information Hubs with 4 centers and 4 GB of memory run in a virtualized environment on the same window. All calculations are performed in Java and the Outline Diminish 2.0 library is used[20]. The Tests were performed on both engineered datasets and real-time.

Algorithm-1 BigFPTree ():

Input: Big Data Set of Transactions, Support
Output: Frequent Patterns
Begin
Construct Header Table ()
Find One Frequent ItemSets ()
Process Transaction Sets ()
End
Algorithm Construct Header Table ()
Begin Scan Dataset
Count support of each item
Construct Header Table (TH) by using Items and their support
(Header Table Consists of 3 fields name, support and link)
Link refers to all nodes of an item on Tree
End
Algorithm Find One Frequent Itemsets ()
Begin
While (TH Not empty)
Do
Remove Items with support less than min support from TH
Sort the TH based on support in descending order
Done
Algorithm Process Item ()
Begin
Insert Data Item Q into Base Item (BI) In TH,
Q.link includes all Nodes in Tree T from this Item is Q
Scan complete Items beginning Node Ni
i= 1 to k to the root of Tree T
Create Sub Header Table (SHT) with items and support
While (SHT Not empty) Do
Remove items from SHT with support less than
min support
Sort SHT on support in decreasing order Done

End

Finally it is observed the execution time of algorithms for different minimum support values in the datasets above[21][22]. As the size of the chunk, i.e. the number of input lines decreases, the number of splits increases. Comparison of Parallel Hybrid Apriori and Parallel BigFM Algorithms over single-node Hadoop Cluster is shown in the following table.1 and represented in chart diagram (Figure.3).

TABLE.1PARALLEL HYBRID APRIORI VS PARALLEL BIGFM ALGORITHMS OVER SINGLE-NODE HADOOP

Number of Transactions	Parallel BigFM Algorithm (Time in secs)	Parallel Hybrid Apriori Algorithm (Time in secs)
20000	78.350	33.599
50000	216.492	216.492
100000	603.050	384.879
125000	661.511	626.736
150000	3971.819	3392.179
200000	6368.553	6227.533

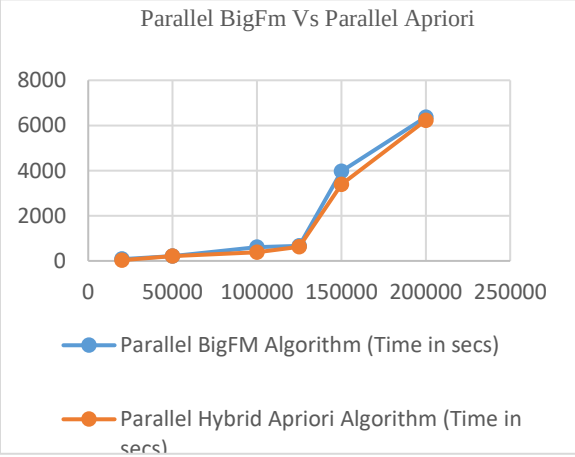

Figure.3 Parallel BigFm Vs Parallel Apriori Algorithms over single-node Hadoop

Table.2 shows the BigFM and Apriori data sets for Huge Information.That is BigFM algorithm is compared to Apriori and generating results. It explains how BigFm provides efficient time complexity than Apriori in stipulations of the number of association rules generated Vs time consumed. This is verified and validated for the number of transactions with a different set of data

records with different execution runs. The following table represents the Results of BigFM and Apriori on Big Data sets.

TABLE.2 BIGFM AND APRIORI ON BIG DATA SETS RESULTS

Status	No.of New Transactions	No.of Transactions	BigFm	Apriori
Iteration -1	10	10	0.51	0.17
Iteration -2	10	20	0.53	0.42
Iteration -3	50	70	0.91	1.88
Iteration -4	200	270	1.13	3.61
Iteration -5	400	670	1.30	5.00
Iteration -6	600	1270	2.67	6.98
Iteration -7	1200	2470	5.28	14.98
Average frequent run times			1.9	5.482
Decrease percentage				63.95 %

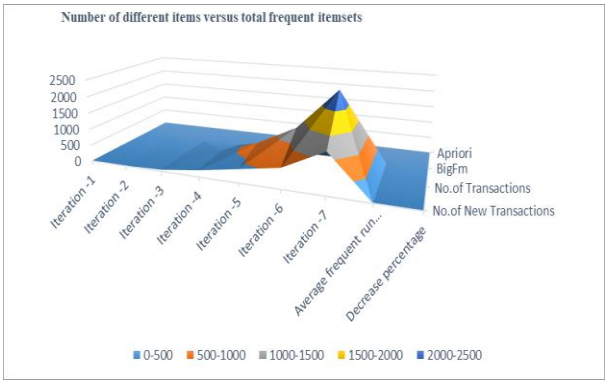

Figure.4. Total Number of different items versus total frequent itemsets

CONCLUSION

Frequent Item-set extraction is the mainly comprehensively practical method to get back constructive in turn from input data. Moreover, after this method is affected by Big Data, the combined explosion of the candidate items must turn out to be a confrontation. Latest advances in the domain of concurrent programming have shown exceptional devices to overcome the difficulty. Nonetheless, these devices encompass their technological disadvantages. We propose an algorithm that contains two novel strategies for the extrication of frequent items and, at the same time, a beat of Map-Reduce frame-work, in which recurrence limits may be small. Introduce a second method, which is a hybrid method which primarily employs FP Tree dependent approach to retrieve frequent item-sets regarding size k and subsequently changes through Eclatas soon as anticipated data sets placed into memory. we focus on generating frequent patterns and association rules of Big Data sets relating FP Tree method. BigFM method is introduced and compared with various methods and considered the results of the comparisons.

REFERENCES

[1] Ahamed, B. B., & Hariharan, S. (2012, December). State of the art process in query processing ranking system. In 2012 Fourth International Conference on Advanced Computing (ICoAC) (pp. 1-5). IEEE.

[2] Ahamed, B. B., & Ramkumar, T. (2015). Deduce user search progression with a feedback session. Advances in Systems Science and Applications, 15(4), 366-383.

[3] Ahamed, R. M. S. Najimaldeen and Y. Duraisamy, "Enhancement Framework of Semantic Query Expansion Using Mapped Ontology," 2020 International Conference on Computer Science and Software Engineering (CSASE), Duhok, Iraq, 2020, pp. 56-60, doi: 10.1109/CSASE48920.2020.9142093.

[4] Babi, C., Rao, M. V., & Rao, V. V. (2018). Efficient Association Rule Mining for Retrieving Frequent Itemsets in Big Data Sets. Current Journal of Applied Science and Technology, 1-14.

[5] Bova, V., Shcheglov, S., & Leshchanov, D. (2019, September). Modified Approach to Problems of Associative Rules Processing based on Genetic Search. In 2019 International Russian Automation Conference (RusAutoCon) (pp. 1-5). IEEE.

[6] Choi, S. Y., & Chung, K. (2019). Knowledge process of health big data using MapReduce-based associative mining. Personal and Ubiquitous Computing, 1-11.

[7] Gangavarapu, T., & Patil, N. (2019). A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets. Applied Soft Computing, 81, 105538.

[8] Hamdad, L., Ournani, Z., Benatchba, K., & Bendjoudi, A. (2020). Two-level parallel CPU/GPU-based genetic algorithm for association rule mining. International Journal of Computational Science and Engineering, 22(2-3), 335-345.

[9] Khotimah, B. K., Miswanto, M., & Suprajitno, H. (2020). Optimization of feature selection using a genetic algorithm in naïve Bayes classification for incomplete data. Int. J. Intell. Eng. Syst, 13(1), 334-343.

[10] Liu, X., Niu, X., & Fournier-Viger, P. (2020). Fast Top-K association rule mining using rule generation property pruning. Applied Intelligence, 1-17.

[11] M. Malek and H. Kadima, 2013. "Searching frequent itemsets by clustering data: Towards a parallel approach using mapreduce", In Proc. WISE 2011 and 2012 Workshops, Springer Berlin Heidelberg, 251–258

[12] Neysiani, B. S., Soltani, N., Mofidi, R., & Nadimi-Shahraki, M. H. (2019). Improve performance of association rule-based collaborative filtering recommendation systems using genetic algorithm. Int. J. Inf Technol. Comput. Sci, 2, 48-55.

[13] Patel, J., & Shah, P. (2019, April). Hiding Sensitive Association Rules Using Modified Genetic Algorithm: Subtitle as needed (paper subtitle). In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 30-34). IEEE.

[14] Sornalakshmi, M., Balamurali, S., Venkatesulu, M., Navaneetha Krishnan, M., Ramasamy, L. K., Kadry, S., ... & Muthu, B. A. (2020). Hybrid method for mining rules based on enhanced Apriori algorithm with sequential minimal optimization in healthcare industry. Neural Computing and Applications, 1-14.

[15] SuryaNarayana, G., Kolli, K., Ansari, M. D., & Gunjan, V. K. A Traditional Analysis for Efficient Data Mining with Integrated Association Mining into Regression Techniques. In ICCCE 2020 (pp. 1393-1404). Springer, Singapore.

[16] Yifan Chen, Xiang Zhao, Xuemin Lin, and Yang Wang, 2015. "Towards frequent subgraph mining on single large uncertain graphs", In 2015 IEEE International Conference on Data Mining (ICDM 2015), Atlantic City, NJ, USA, 41-50

[17] Ahamed, B. B., & Yuvaraj, D. (2019, October). Dynamic Secure Power Management System in Mobile Wireless Sensor Network. In International Conference on Intelligent Computing & Optimization (pp. 549-558). Springer, Cham.

[18] Sivaram, M., Yuvaraj, D., Porkodi, V., & Manikandan, V. (1941). Emergent news event detection from facebook using clustering. Journal of Advanced Research in Dynamical and Control Systems, Pages, 1947, 2018.

[19] Ahamed, B. B., & Yuvaraj, D. (2018, October). Framework for faction of data in social network using link-based mining process. In International Conference on Intelligent Computing & Optimization (pp. 300-309). Springer, Cham.

[20] Nithya, S., Sundara Vadivel, P., Yuvaraj, D., & Sivaram, M. (2018). Intelligent based IoT smart city on a traffic control system using raspberry Pi and robust waste management. Journal of Advanced Research in Dynamical and Control Systems, Pages, 765-770.

[21] Sivaram, M., Yuvaraj, D., Porkodi, V., & Manikandan, V. (1941). Emergent news event detection from Facebook using clustering. Journal of Advanced Research in Dynamical and Control Systems, Pages, 1947, 2018.

[22] Ahamed, B. B., & Yuvaraj, D. (2018, October). Framework for a faction of data in social network using link-based mining process. In International Conference on Intelligent Computing & Optimization (pp. 300-309). Springer, Cham.

[23] Yuvaraj, D., Sivaram, M., Ahamed, A. M. U., & Nageswari, S. (2019, October). An Efficient Lion Optimization Based Cluster Formation and Energy Management in WSN Based IoT. In International Conference on Intelligent Computing & Optimization (pp. 591-607). Springer, Cham.

[24] Karthikeyan, B., Raj, M. A., Yuvaraj, D., & Sundar, K. J. A. (2020). A Hybrid Approach for Video Steganography by Stretching the Secret Data. In Inventive Communication and Computational Technologies (pp. 1081-1087). Springer, Singapore.

[25] Venkateasan, R., Yuvaraj, D.,(2018). Predicting Students' Academic Drop Out and Failures Using Data Mining Techniques, International Journal of advance Science and Technology,28(2),182-193.