International Journal of Scientific Research in Computer Science and Engineering
Vol.6, Issue.2, pp. 10-14, April (2018)

Research Paper

E-ISSN: 2320-7639

Different Approaches for Frequent Itemset Mining

P.V. Nikam[1*], D.S. Deshpande[2]

[1]Department of CSE, Jawaharlal Nehru Engineering College, Aurangabad, India
[2]Department of CSE, Jawaharlal Nehru Engineering College, Aurangabad, India

[*]*Corresponding Author: pallavinikam19@gmail.com, Tel.: +91-7798653564*

Available online at: www.isroset.org

Received: 07/Mar/2018, Revised: 15/Mar/2018, Accepted: 31/Mar/2018, Online: 30/Apr/ 2018

***Abstract*—** Data mining is the retrieval of hidden analytical information from huge databases, is a controlling new technology with great possible to help organizations as well as research focus on the mainly essential information in their data warehouses. Data mining tools forecast future development and performances, allowing businesses to create proactive, idea for decision making systems. Frequent Itemset Mining (FIM) is one of the traditional data mining problems in mainly of the data mining approaches. It requires very huge computations and input and output traffic capacity. Also resources like single processor's memory and CPU are very limited, which degrades the presentation of algorithm. In this research work system proposed one such distributed approach which will run on Hadoop cluster – one of the recent most popular distributed frameworks which basically focus on parallel processing. The proposed framework takes into account extends characteristics of the Apriori algorithm related to the frequent itemset invention and throughout a block-based partitioning uses a dynamic workload management. The algorithm greatly improves the performance and gets high scalability compared to the existing approaches. Proposed algorithm is implemented and tested on large scale datasets distributed system on heterogeneous cluster.

***Keywords*—** Frequent Itemset, Apriori, FP Growth, Modified Apriori.

I. INTRODUCTION

In a data mining process, it is a tedious task to find the frequent patterns that are extracted from the Data Warehouse. Patterns are the set of rules behind the Mining frequent pattern of data, which comprises of Frequent Item Set Mining, Frequent Sequence Mining, Frequent Tree Mining, and Frequent Graph Mining. From that pattern, extracting knowledge is tedious. To solve this problem, so many algorithms have come up. Now days, the data sizes are in about Peta bytes. Hence it is hard to precede the problem whenever a worst case occurs. Selecting the suitable algorithm is a biggest task ever. In order to extract the accurate knowledge, the data should not be damaged. Each algorithm faces the trouble with the data collection and processing in the different view. Every algorithm is not suitable for the all kind of itemsets. This paper presents few latest algorithms used in various scenarios.

Rest of the paper is organized as follows; Section II presents related work done in frequent itemset mining. In Section III, we present our proposed algorithm. Section IV presents results and discussion and conclusion is drawn in Section V with future direction.

II. RELATED WORK

Finding the association rules in large databases play a key role in data mining. Kalli Srinivasa Nageswara [1] has considered the prior researches and present working status in order to restore the gaps between them with present known information. There were two problems regarding this context: identifying all frequent item sets and to generate constraints from them. Here, first problem, as it takes more processing time, was computationally costly. Consequently, many algorithms were proposed to solve this problem. Their current study considers such algorithms and the related issues.

An efficient tree structure for mining high utility itemsets was presented in [2]. At first, the authors have developed a utility frequent-pattern tree structure, an extended tree structure for storing crucial information about utility itemsets. Then, the pattern growth methodology was utilized for mining the complete set of utility patterns. Improved high utility itemsets mining efficiency was achieved using two major concepts: 1) Compressing a large database into a smaller data structure as well as the utility FP-tree avoids repeated database scans, 2) The pattern growth method utilized in the proposed FP-tree-based utility mining avoids

the costly generation of a large number of candidate sets and thereby reduces the search space dramatically. Experimental analysis was carried out on tree structure mining concept using different real time datasets. The performance evaluation results have demonstrated the efficiency of the proposed approach in mining high utility itemsets.

To discover the relationships among the attributes in a database, association rules are the most important tool used. In [3] the authors have discussed that the existing Association Rule mining algorithms were applied on binary attributes or discrete attributes, in case of discrete attributes there was a loss of information and these algorithms take too much computer time to compute all the frequent itemsets. By using Genetic Algorithm (GA), it is possible to improve the generation of Frequent Itemset for numeric attributes. The major advantage of using GA in the discovery of frequent itemsets is that they perform global search and its time complexity was less compared to other algorithms as the genetic algorithm was based on the greedy approach. The main aim of their paper is to find all the frequent itemsets from given data sets using genetic algorithm.

Association Rule Mining (ARM) is used to find all frequent itemsets and to build rules based of frequent itemsets. But a frequent itemset only reproduces the statistical correlation between items, and it does not reflect the semantic importance of the items. To overcome this limitation, Kannimuthu et al. [4] have utilized a utility based itemset mining approach. Utility-based data mining is a broad topic that covers all aspects of economic utility in data mining. It takes in predictive and descriptive methods for data mining. High utility itemset mining is a research area of utility based descriptive data mining, aimed at finding itemsets that contribute most to the total utility. The well known faster and simpler algorithm for mining high utility itemsets from large transaction databases is Fast Utility Mining (FUM). In this proposed system, they made a significant improvement in FUM algorithm to make the system faster than FUM. The algorithm was evaluated by applying it to IBM synthetic database. Experimental results have shown that the proposed algorithm was effective on the databases tested.

An efficient approach based on weight factor and utility for effectual mining of significant association rules was proposed by Parvinder S. Sandhu et al. [5]. Initially, the approach has utilized traditional Apriori algorithm to generate a set of association rules from a database. The proposed approach exploits the anti-monotone property of the Apriori algorithm, which states that for a k-itemset to be frequent all (k-1) subsets of this itemset also have to be frequent. Subsequently, the set of association rules mined were subjected to weightage (W-gain) and utility (U-gain) constraints, and for every association rule mined, a combined utility weighted score (UW-Score) was computed. Ultimately, they have determined a subset of valuable association rules based on the UW-Score computed. The experimental results have demonstrated the effectiveness of the proposed approach in generating high utility association rules that can be lucratively applied for business development

An enhanced association rule mining algorithm to mine the frequent patterns was proposed by Venu Madhav Kuthadi [6]. The algorithm utilized weightage validation in the conventional association rule mining algorithms to validate the utility and its consistency in the mined association rules. The utility is validated by the integrated calculation of the cost/price efficiency of the itemsets and its frequency. The consistency validation is performed at every defined number of windows using the probability distribution function, assuming that the weights are normally distributed. Hence, validated and the obtained rules are frequent and utility efficient and their interestingness are distributed throughout the entire time period. The algorithm was implemented and the resultant rules were compared against the rules that can be obtained from conventional mining algorithms.

III. PROPOSED SYSTEM

The figure 1 shows the proposed system architecture, with traditional system, which can be transformed into new system that can work with in big data environment. The figure [1] show the overall procedure of our system**.**

System first input the different data set for processing. The proposed system has used three different algorithms (Apriori, FP growth and hybrid algorithm) which are used to evaluate the system results. Each algorithm provides the different results with same dataset. The system can able to process the transaction database.

Algorithm

Proposed Hybrid Algorithm for finding frequent itemset

Input: Dataset D, Support generation denominator De, minreqitems mk;
Output: generate T item set
Step 1: for all (T in DBi) do
Step 2: items [] ←split(T)
Step 3: for all (item in items []) do
Step 4: if the item is available in FLits
Step 5: Add a[] ← item
Step 6: end for
Step 7: add all a to ArrayList <items name, count> All items.
Step 8: Generate the support base on support=(T.count/De)
Step 9: for(k in ArrayList)
Step 10:if(k.count>=support)
Step 11: FreqItems(k);
Step 12: end for
Step 13: apply step 9 to 12 when reach mk

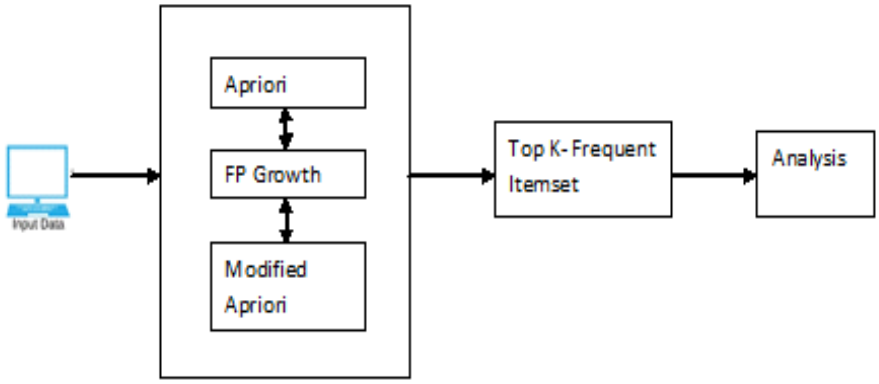

Figure 1: Proposed System architecture

IV. RESULTS AND DISCUSSION

We used synthetic data resemble market basket data with short frequent patterns. The experimental results of this framework are set to the minimum support threshold (or, proportionally, bigger data sizes) than having even yet been considered. These upgrade same at no execution cost, as prove by the way that our implementation achieves the performance compare to other methods in less time.

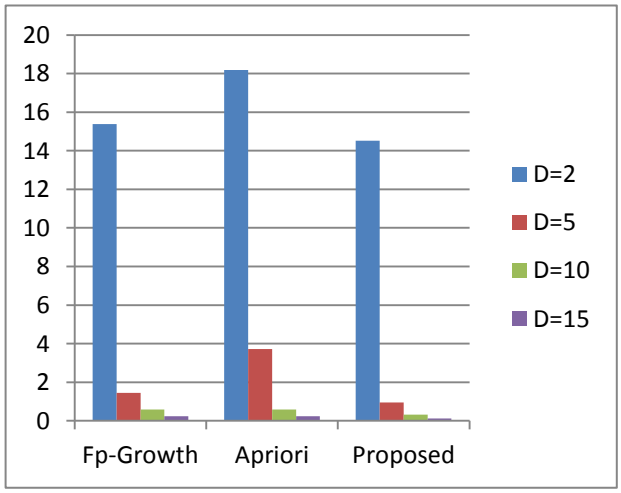

Figure 2(a). Comparative analysis of different algorithms execution time with 100 instances

Before this work, it has been accepted that the execution of Apriori is slow. Through these examinations we have demonstrated that at increased support our implementation produces the same results with improved performance. When threshold value decreases, other methods occupy larger memory as well as increased time. While comparing other methods limits our execution creates an indistinguishable outcomes with a similar execution from our implementation.

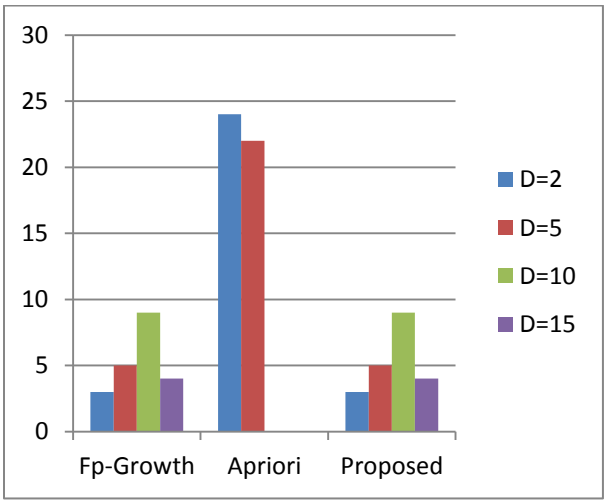

Figure 2(b). Comparative analysis of different algorithms K-items extracted with 100 instances

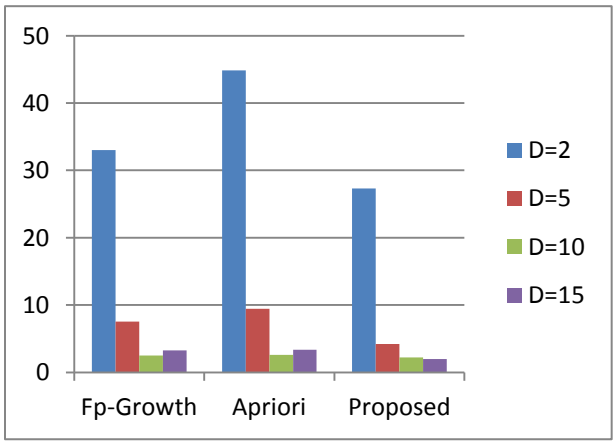

Figure 3(a). Comparative analysis of different algorithms execution time with 400 instances

The performance of this system against the different frequent itemset mining methods is as shown in fig 2 and fig 3. In existing system implementation tree structure is used to store the results. There is need to choose the best programming language or algorithm to achieve the best results within less time and less memory. Accurate results are also depends on the proper selection of algorithms. Therefore we are taking Modified Apriori algorithm, it can be best algorithm to give the accurate results as compare to existing systems. Proposed system algorithm shows the faster execution even for large database on hadoop. We could create our own vast dataset against which to likewise run tests, however the cost for doing as such is negligible. Here we performed two experiments. In experiment 1, we took 100 instances with different support denominators. The figure 2 (a) shows the execution time in seconds required to run algorithms, while figure 2 (b) shows total extracted top-k item set count. In experiment 2, we took 400 instances with different support denominators. The figure 3 (a) shows the execution time in seconds for different algorithms and figure 3 (b) shows the total extracted top-k item set count.

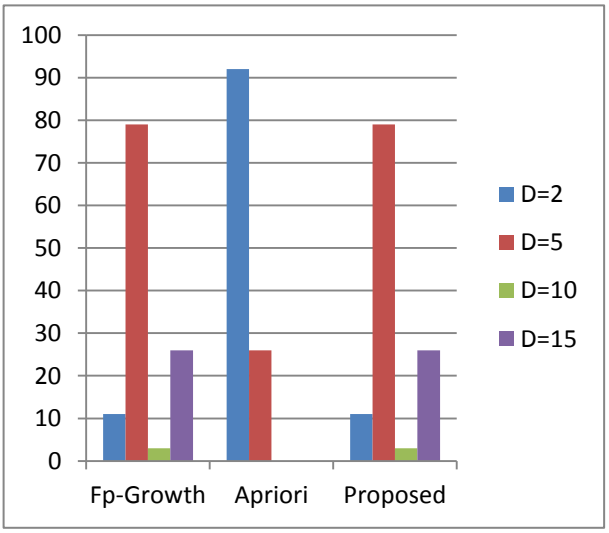

Figure 3(b). Comparative analysis of different algorithms K-items extracted with 400 instance

V. CONCLUSION

Frequent Item set Mining has attracted plenty of attention but much less attention has been given to mining Infrequent Item sets. The mining of frequent item set is an important research area in the field of data mining. The association rules are formed using the frequent item set mined. Many different methods have been proposed for mining the frequent item sets. The three different algorithms result shows the efficiency of each algorithm with different datasets and experimental results shows the each algorithm efficiency base on time as well as accuracy. The system also show the different results when we given a different support value. Hence the system can be applicable for all kind of small business data mining process. In the different observations it is reflected how the system can provide different accuracy levels with heterogeneous datasets. The modified algorithm also improves the overall execution as well as data extraction accuracy for same.

VI. FUTURE WORK

To implement the system in multi cloud with distributed environment, and work with different synthetic dataset like semi structured as well as unstructured datasets. The recommendation system can also work in big data environments like climate analysis, stock market prediction etc.

REFERENCES

[1] K. S. N. Prasad, S. Ramakrishna, "*Frequent Pattern Mining and Current State of the Art*", International Journal of Computer Applications, Vol. 26, No. 7, pp. 33-39, 2011.

[2] C. Saravanabhavan, R. M. S. Parvathi, "*Utility FP-Tree: An Efficient Approach to Mine Weighted Utility Itemsets*", European Journal of Scientific Research, Vol. 50 No. 4, pp. 466-480, 2011.

[3] R. V. Prakash, S. Govardhan, S. S. V. N. Sarma, "*Mining Frequent Itemsets from Large Data Sets using Genetic Algorithms*", IJCA-Artificial Intelligence Techniques - Novel Approaches & Practical Applications, No. 4, Vol. 7, pp. 38-43, 2011.

[4] S. Kannimuthu, K. Premalatha, S. Shankar, "*iFUM - Improved Fast Utility Mining*", International Journal of Computer Applications, Volume 27– No.11, pp. 32-36, 2011.

[5] P. S. Sandhu, D. S. Dhaliwal, S. N. Panda, "*Mining utility-oriented association rules: An efficient approach based on profit and quantity*", International Journal of the Physical Sciences Vol. 6, No. 2, pp. 301-307, 2011.

[6] V. M. Kuthadi, "*A New Data Stream Mining Algorithm for Interestingness-Rich Association Rules*", Journal of Computer Information Systems, pp. 14-27, 2013.

[7] R. Agrawal, T. Imielinski, A. Swami, "*Mining association rules between sets of items in large databases,*" ACM SIGMOD Rec., vol. 22, no. 2, pp. 207–216, 1993.

[8] M. Y. Lin., P. Y. Lee, S. C. Hsueh "*Apriori based frequent itemset mining algorithms on Mapreduce*," ICUIMC'12, 6th international conference on Information Management and communication,Malaysia, article no. 76, Feb 2012.

[9] L. Zhou, Z. Zhong, J. Chang, "*Ballanced Parallel FP-growth with Mapreduce,*" Information Computing and telecommunications, 5713090, Nov. 2010.

[10] Y. J. Tsay, T. J. Hsu, J. R. Yu "*FIUT: A new method for mining frequent itemsets*," Inf. Sci., vol. 179, no. 11, pp. 1724–1737, May 2009.

[11] S. Moens, E. Aksehirli, B. Goathals, "Frequent Itemset Mining for Big Data," IEEE conference in Big Data, Silicon Valley, USA, 978-1-4799-1293-3, Dec. 2013.

[12] M. Riondato, J. A. DeBrabant, R. Fonseca, and E. Upfal, "*PARMA:A parallel randomized algorithm for approximate association rules mining in MapReduce*," Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.,Maui, HI, USA, pp. 85–94,2012.

[13] Goethals B. "*Survey on frequent pattern mining*," Univ. of Helsinki, 19:840-52.

[14] S. Jarkad, J. E. Nalavade, "*Approach for Big Data Mining in Hadoop Framework : An Overview*," IJIRSET, Jan 2017.
[15] Y. Xun, J. Zhang, and X. Qin, "*FiDoop: Parallel Mining of Frequent Itemsets Using Mapreduce*," IEEE Trans. Systems, Man, and Cybernetics, vol.46, no.3, Mar. 2016.
[16] C. V. Suneel, K. Prasanna, M. R. Kumar, "Frequent Data Partitioning using Parallel Mining Itemset and Mapreduce", IJSRCSEIT, Vol.2,Issue 4,2017, pg.641-644.

Authors Profile

Pallavi V. Nikam has completed her Bachelor of Engineering degree and currently pursuing Masters of Engineering in Computer Science & Engineering from Jawaharlal Nehru Engineering College under BAMU University from Aurangabad, Maharashtra, India.

Dr. Deepa S. Deshpande is currently the Associate Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Engineering College, Aurangabad. She obtained her Bachelor of Engineering Degree and her Masters of Technology degree in Computer Science. She has also completed her Ph.D. Her research interests include data mining, data science. She has a 21 years of working experience in teaching.

