



An Empirical Study on Frequent Item Set Mining

Shubham Devrao Mohurle
Computer Engineering
K J C O E M R, Pune, Maharashtra, India
Prof. N. Bogiri
Computer Engineering
K J C O E M R, Pune, Maharashtra, India.

Abstract: In e-commerce or retail business, it is very important to predict the requirements of the customers to improve the business. The easiest and quickest solution for this is to analyze the current market trend to fulfill the customer's requirements. Most of the time these trends are keep changing due to product quality, entry of the new products, endorsements, promotional events, freebies, advertisements and many more reasons. This literally makes impossible to predict the requirements, so one strong and effective method for this is frequent item mining. Apriori is one of the most popular among these frequent item mining algorithms. Because of its horizontal pattern mining it creates more space complexities. So this paper analyses some of the past works on frequent item mining and try to evaluate lesser time and space complexity solution.

Keywords: Frequent Itemset Mining, Apriori Algorithm, Eclat Algorithm.

I. INTRODUCTION

Ever since the advent of humans on this planet, they have been in search of knowledge. At first it was recognizing the dangers that would be fatal for an individual. Such as, predatory animals and birds. They had to recognize the edible fruits and animals that could be eaten for the survival of the species. Human's memory served as the greatest asset in times like these, which would be changed soon enough.

As there was an increase in the amount of information being processed, the humans developed speech and taught their progeny. Therefore, the knowledge was now proliferated with the help of word of mouth. This practice was essential for the survival of our species as knowledge has always been supreme. This was the case until humans were hunter gatherers and required extensive knowledge about what was edible and stay away from poisonous items.

This was extremely difficult to maintain as the agricultural age was ushered and humans settled in pucca houses rather than natural caves and had tools to assist them in farming. This had to be detailed in the form of books. And books were the primary source of knowledge and are still in widespread use due to their flexibility and versatility. This was very instrumental in bringing about the proliferation of knowledge and helped advance the human species forward.

As the technological advancements have occurred significantly with the introduction of silicon as the center of all the technological expansion. Silicon can act as a barrier and allow only for certain values of voltage or current to be able to act like switches which gave rise to the transistor and revolutionized computing forever. The rise of electronics also increased our consumption of data.

Nowadays, we have such an increasing number of smartphones and computers that generate petabytes of data. This data is massive in size and it is very difficult to handle that efficiently. Therefore, there have been some methods that have been designed to manage and extract relevant information from these large sets of data. These large collections of data are called as Big Data very appropriately.

Frequent itemset mining refers to the mining of data elements in a dataset that are frequent of occur with a recurring pattern. This is very critical for the workings of a data mining algorithm. as the frequent itemset mining does not do anything substantial but it is designed in a way to perform certain tasks related to identifying the patterns that exist within a particular dataset.

Frequent itemset mining is also known as association mining because it has the ability to associated patterns with each other and effectively return the set of items occurring with a certain frequency. It is one of the most important parts of the market basket analysis. The market basket analysis utilizes the frequent item set mining to understand the habits of its customers to a larger degree.

The Market Basket Analysis can be really helpful in decoding the various items that are seemingly picked up by the customers. This can provide a greater understanding of the buying habits of people and the kind of items being

bought together. This information is crucial as the data can be used for increasing the efficiency of the store and it can also provide intelligent combos that can be offered by the owner.

This technique is quite useful in making various collections based on what the customers are readily buying together. This provides the customer with a one stop solution for all their needs and decreases the unnecessary navigation that would entail if the items had to be bought separately. This kind of activity increases customer satisfaction and in the long run can benefit the organization greatly.

Frequent itemset mining is extensively used in a lot of applications concerning the derivation of meaningful insight from the data that is generated, such as the spending habits of the customers, their behavior and preferences. This enables the organization of the store to achieve personalized packages and offer discounts on certain objects that are sluggish in their sales.

The Apriori algorithm is one of the most common applications of the concept of frequent itemset mining that is widely applicable to the Market Basket problem. Apriori algorithm thus is suited for application on datasets that are bigger in size as the algorithm excels in managing Big Data very efficiently. One of the ways this is achieved by the Apriori algorithm is that it utilizes the matrix form to organize the data coherently.

The Matrix that is generated by the Apriori algorithm has the columns that accept all the instances that are being observed in the dataset. Whereas, the rows are designed to accept the items that are being encountered by the algorithm. Thus, the binomial coefficient here is expressed as the Apriori algorithm. The algorithm then extracts frequently appearing items in the dataset with the help of the support values.

These support values correspond to the frequent items in the dataset and are therefore aggregated into candidate datasets. The frequency of the candidate datasets is measured against their corresponding support values to determine if it can be accepted or rejected as a frequent itemset or not. The Apriori algorithm is quite a simple algorithm and benefits the area of Frequent itemset mining greatly.

There are a few downsides to the use of Apriori algorithms for mining as they are prone to various bottleneck when the size of the datasets gets increasingly big. This leads to many and long patterns being recognized that affects the performance of the algorithm negatively. These events then increase the computational complexity of the algorithm in general and the algorithm suffers from a decreased computational speed.

What is FP- tree?

FP tree is a very effective application of data mining when the datasets in use are transaction datasets. The FP tree algorithm is modification of the Apriori algorithm as it suffers from the bottleneck that does not allow it to function with bigger datasets containing longer patterns. The FP tree doesn't get affected by the size of the dataset as it utilizes a compact and efficient tree structure instead of the matrix form and this decreases the chances of a data explosion to occur.

FP tree is commonly used algorithm for Data mining procedure as it is very efficient and effectively mines the appropriate datasets when compared to the much slower Apriori algorithm. Because unlike the Apriori algorithm, it utilizes very efficient trees that can process the data faster and in a more efficient manner. But it is not without its demerits as it is significantly more expensive to implement in comparison to the Apriori algorithm which is a lot cheaper.

In this paper, section 2 is dedicated for literature review of past work and Finally Section 3 concludes this paper.

II. LITERATURE REVIEW

P. Nikam [1] introduces the most important process in the procedure for the Mining of Data from large datasets, Frequent Itemset Mining. The researchers explain that the procedure of frequent itemset mining gets severely bottlenecked when popular algorithms encounter excessive amounts of data. These algorithms buckle under the load of the enormous data and are unable to sustain the load. The authors propose a solution to this problem by introducing a modified Apriori algorithm that utilizes parallelization to manage and balance the heavy data load. The technique is named FiDooP as it utilizes the HDFS environment and is demonstrated to be highly effective in comparison to the conventional methods.

Research Gap: The researchers have not utilized the slot configuration in the HDFS framework, addition of which can enable slot ordering and allocation for utilizing the maximum amount of resources.

F. Zhang [2] explores a new concept in the realm of data mining, specifically in the area of Frequent Itemset Mining. The researchers express that most of the algorithms that are designed for the purpose of Frequent itemset mining are very resource intensive and utilize a lot of memory and processing power. Most of the algorithms also do not scale proportionately and that is a big problem. To ameliorate this effect, the authors have presented a unique algorithm that performs the operations parallelly in addition to 3 optimizations. The load balancing is managed by the intelligent system by clustering the frequent itemset mining tasks into smaller tasks.

Research Gap:- The researchers have not extended their approach to allow for optimizations from the FPGA or the GPU. Furthermore, the problem could have been approached dynamically for the dataset characterizations.

S. Haseena [3] suggests that Frequent Pattern mining is one of the most widely discussed methods that are used for the process of Data mining and form an integral part of research. This paper concerns with the mining of extremely large datasets. There a lot of algorithms that are optimized to perform the process of Frequent itemset

mining, and the *modus operandi* remains largely similar across the algorithms. The authors propose an innovative technique for the mining operations that are based on the timestamping information to extract the frequent itemsets in the dataset. The algorithm builds on the FP growth algorithm and utilizes the temporal data for the timestamping information.

Research Gap:- The proposed algorithm extends on the already existing algorithms which are known to bottleneck with large amounts of data and longer patterns.

S. Kawanobe [4] expresses the need for resolving one of the most prevalent problems with mining frequent itemsets, low understandability of the patterns that are generated. To ameliorate this problem, the authors have presented a technique for the categorization of the patterns by using representation learning from different perspectives. The system performs cluster analysis for the identification of the representative and outlier patterns.

Research Gap: - The researchers have not introduced human intervention in the frame work, therefore, a community discovery feature must be added to the system as it will increase the accuracy even further.

Z. Tong [5] states that Data mining is a very powerful concept but is quite limited due to the constraints of computational power. Most of the data mining applications on large amounts of data requires equally large computational power. There is also the problem of data security which is maintained by IDS logs. The analysis of IDS logs of a large database is also really computationally expensive. Therefore, the authors present SnortIDS which solves the problem of security and decreases the computational overhead.

T. Wang [6] explores the realm of data mining and the privacy concerns that are faced by the researchers working on sensitive data. Therefore, to ameliorate this effect of data privacy, the authors present a Local Differential Privacy that allows its users to answer sensitive questions without a breach of their privacy. The frequency of a certain value is estimated by the LDP frequent oracle through an aggregator. Due to a verity of users having various different values as their answers, the padding and supplemental sampling is used to retrieve the values that are frequent and also derive their frequency.

S. Tribhuvan [7] elaborates on the notion of data mining and how it can be beneficial for businesses and help them to be a lot more practical and capable of deriving valuable insight by approximating future trends and customer behavior. As clustering is one of the most significant aspects of association rule mining, the authors observed some inconsistencies in the conventional approach such as data distribution on large clusters, lack of fine load balancing etc. Therefore, to ameliorate these effects, the authors presented a framework based on an enhanced Apriori algorithm that implements MapReduce. the proposed method combats all the shortcomings gracefully and exhibits greater efficiency.

X. Zhao [8] introduces a decision-making system based on the weighted frequent itemset mining technique. The researchers utilized the weighted frequent itemsets by proving their downward closure property and they also implemented the weighted frequent sub-sets by proving their existence property. These techniques help the system get smarter and to be used in various applications. The properties of this type of system can be used to reduce the searching space of WD-FIM increase the time-efficiency.

H. Rao [9] remarks that the introduction of machine learning into businesses and other retailers can enable a far more valuable insight into the business. The authors propose to solve the big data problems with the help of machine learning. to this end, the authors implement a FP-grow Algorithm for mining the frequent itemsets on an E-commerce platform. The technique is used for the classification of user behavior by generating feature through the analysis of the association rules. Naïve Bayesian algorithm is used to achieve clustering for personalized referral services and marketing.

Y. Zhang [10] expresses concerns over the privacy issues and concerns that are faced during the mining operations on large datasets. As there is a lot of sensitive information pertaining to organizations and individuals can be leaked. Therefore, to ameliorate these effects, the authors present an efficient privacy preservation algorithm named PBEK. The proposed algorithm divides the data into small segments. These small fragments are subjected to the BEMASK algorithm individually. The algorithm has been tested to outperform the conventional algorithms.

A. Gassama [11] expresses that frequent itemset mining is one the most fundamental aspects of data mining. As the realm of data mining has been really helping in providing a useful insight in the applications such as market basket analysis and outlier detection. As most of the traditional approaches are not suitable for large datasets. To provide a valuable solution for this problem by introducing a scalable FP-growth algorithm on Apache Spark. The proposed technique can handle large data and provides a much greater efficiency.

Research Gap:- The proposed technique can be improved further by changing the data structure from the Compact FP-tree to a lighter data structure that can reduce the memory consumption even more.

III. CONCLUSION

The frequent item sets are playing vital role in market analysis of the product requirements. Most popular algorithms like Apriori and others are basically working on the horizontal pattern because of this it takes more complexities to handle the big datasets. So this research paper analyzes the all past works on handling frequent item sets for moderate and big data. And come to a conclusion that this process requires more stable algorithms and methodologies to handle the data of moderate and large size.

REFERENCES

- [1] P. Nikam and D. Deshpande, "New approach in Big Data Mining for frequent itemset using MapReduce in HDFS", 3rd International Conference for Convergence in Technology, IEEE, 2018.
- [2] F. Zhang, Y. Zhang, X. Liao and H. Jin, "PNPFI: An Efficient Parallel Frequent Item Sets Mining Algorithm" Proceedings of the 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design, 2018.
- [3] S. Haseena, S. Manoruthra, P. Hemalatha and V. Akshaya, "Mining Frequent Item sets on Large Scale Temporal Data", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology, ICECA 2018.
- [4] S. Kawanobe and T. Ozaki, "Experimental Study of Characterizing Frequent Item sets using Representation Learning", 32nd International Conference on Advanced Information Networking and Applications Workshops, 2018.
- [5] Z. Tong and H. Ying, "Application of Frequent Item Set Mining Algorithm in IDS Based on Hadoop Framework", The 30th Chinese Control and Decision Conference, IEEE, 2018.
- [6] T. Wang, N. Li and S. Jha, "Locally Differentially Private Frequent Itemset Mining", IEEE Symposium on Security and Privacy, 2018.
- [7] S. Tribhuvan, N. Gavai and B. Vasgi, "Frequent Itemset Mining using Improved Apriori Algorithm with MapReduce", Third International Conference on Computing, Communication, Control and Automation, 2017.
- [8] X. Zhao, X. Zhang, P. Wang, S. Chen and Z. Sun, "A Weighted Frequent Itemset Mining Algorithm for Intelligent Decision in Smart Systems", IEEE Access, 2018.
- [9] H. Rao, A. Liu and Z. Zeng, "Research on Personalized Referral Service and Big Data Mining for E-commerce with Machine Learning", 4th International Conference on Computer and Technology Applications, 2018.
- [10] Y. Zhang, T. Wang and S. Wang, "An Efficient Algorithm for Frequent Pattern Mining based on Privacy-preserving", 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 2017.
- [11] A. Gassama, F. Camara2, S. Ndiaye, "S-FPG: A Parallel Version of FP-Growth Algorithm under Apache Spark", 2nd IEEE International Conference on Cloud Computing and Big Data Analysis 2017.
- [12] P. Dong and B. Chen, "new algorithm of maximum frequent itemsets based on FP-tree for mining multiple-level association rules", International Conference on Computer Science and Software Engineering, 2008.