iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** I **Month of publication:** January 2021

DOI: https://doi.org/10.22214/ijraset.2021.32789

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue I Jan 2021- Available at www.ijraset.com

A Way to Generate Frequent Pattern Mining using Vertical Partitioning of Data with Parallel Computing

Anil Vasoya[1], Dr. Nitin Koli[2]
[1]*Ph.D Scholar, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India*
[2]*Deputy Registrar, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India*

Abstract: ***The performance of association rule algorithms is also evaluated based on time-complexity and accuracy of frequent item set Also, Frequent item set is highly dependent on the user input status such as minimum support. It is difficult to know the meticulous minimum support because these it generate logically incorrect or irrelevant FIS and sometime loose of worthy FIS. These issues can be resolved with the help of Proposed Vertical Approach In this paper, a detailed comparison has been made for the frequent pattern mining with normal approach and vertical approach with proper example. It shows that how can we achieve logically relevant FIS as well as Produces FIS for few categories that are lesser in demand but have higher worth using vertical approach. The Proposed vertical Approach provides a multi-level view of the dataset by clustering w.r.t. to category of the product.***

Keywords: Frequent item set, Association rule Mining, Apriori, Large database , Vertical approach

I. INTRODUCTION

Association rule mining attempt to determine relations among transactions in a transactional database. Ascertaining valuable frequent pattern unknown in database play vital role to increase the business profit in the applications like retail stores, online shopping database etc. ARM can be used to advance decision making in a wide variety of applications. Data mining can abstract important facts such as frequent item set from large data set – nevertheless sometimes it is challenging to achieve all frequent item set if these datasets are fragmented into many clusters when there is a large dataset.

Fast development in technology has resulted increase of great amount of data for any e-commerce organization and therefore now a days fetching needed information from huge dataset has been a big challenge. The discovered interesting patterns from transactional records can help many organizations for decision-making processes like catalog preparations, cross-marketing and to analyze the customer shopping behavior. The basic concept of frequent pattern mining is to find out the patterns of frequent item sets presence in the database is more than a specified minimum support or threshold.

$$\text{Min. Support } (A \Rightarrow B) = P\,(A \cup B)$$

Where A and B is items of transactions. & Min. Support is the % of transactions in which contain both items A and B.

$$\text{Confidence } (A \Rightarrow B) = P\,(B|A)$$

Confidence is ratio of transactions in datasets which containing number of times item A with respect to item B.

Association rule mining is a method for discovered the set of frequent items or interesting pattern based on minimum support and rules based on confidence.

In the method of computing the speed of the occurrences of a remarkable frequent items within the large transactional database and its computational time always could be a big factor of attention for researchers. to get rid of the exponentially more number of candidates, many algorithms, which are most often employed by researchers, try to get less Candidate with user defined min support. while with pruning, finding all frequent item sets from large datasets requires high computation power of CPU and required plenty of memory.

This research paper evaluates the performance of proposed algorithm with Apriori, FP-Growth, ECLAT frequent itemset mining algorithms. This research is limited to frequent itemset mining. By limiting the experimentation to a single implementation of frequent itemset mining this research is able to evaluate the characteristics of the dataset such as number of FIS, transaction, Minimum support affect the performance of all algorithms. So the concept of association rule mining is to discover the sets of items nurture to associate with the others in the database.

The remaining of this paper is organized as: section 3 discuss the literature review in the related field, section 4 emphasize the proposed vertical approach and concern existing normal approach, section 5 gives the experimentation results of the implemented system with example, finally, section 6 will be concluded along with future work.

II. LITERATURE REVIEW

The foremost famous is that the Apriori algorithm which has been brought in 1993 by Agrawal et al. [1] which uses association rule mining[2][3][4] [5][6] Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the identical time. Association rule generation is usually get a divorce into two separate steps: 1. Minimum support (threshold) is applied to go looking out all frequent item-sets during a very database. 2. These frequent item-sets and also the minimum confidence constraints are accustomed form rules. In order to spice up the performance of association rule mining, many researchers tried to distribute dataset into the mining computation over quite one processor or a standard computer have multicore for executing a task[7][8][9]. A method of accelerating the computational speed is by using all core within one computer.

Jun Yang, Haoxiang Huang et al. [10] in their paper, used the Apriori algorithm to mine the frequent access pattern in the web access paths. Aiming at the shortcomings of Apriori algorithm, an improved algorithm named AC-Apriori based on traditional Apriori algorithm is proposed and implemented. The AC-Apriori algorithm combines the Apriori algorithm with the AC automation. Compared with the traditional Apriori algorithm, AC-Apriori algorithm reduces the time scanning database, reduces the runtime and improves the efficiency

Li Zhan, Fusheng Yu et al. [11] proposed a quick algorithm for temporary organization rules. In this paper, the tests in the synthetic dataset and the original data show the best performance of the proposed new algorithm and the efficiency test shows that the proposed new algorithm can obtain the rules of transient integration successfully. The proposed new algorithm can only be acquired temporarily for organization rules on purchase data, but not throughout the time series.

Xiaorong Cheng et al. [12] aimed at the problem of updating the association rules when extending data sets while keeping minimal support unchanged, on the basis of the FUP algorithm, this paper proposes a matrix-based integration of the mining rules algorithm, namely the MBFUP algorithm. The algorithm only scans the data scanner once, which greatly reduces the complexity of the space and the complexity of the algorithm.

LIU Jian-ping et al. [13] has proposed a new type of Pre-FP algorithm based on the concept of large-scale objects. The main idea is that by predicting a "top" and "bottom" support framework, it sets up large-scale forecasts other than the usual and impossible things. This method has reduced the time to delete real transaction data.

Siqing Shan et al.[14] have discovered the problem of continuous updating of large data in their paper. They have proposed a method to reduce data size during the reconstruction process. They have given a description and algorithm for T-tree marketing. The order of things in every T-Tree is different from that in FP-Tree. They have proposed a general continuous algorithm or CIU algorithm in brief. Use of the CIU algorithm, and studied its performance compared to the FP-growth algorithm in big data is seen in this paper.

Anjana Gosain [15] et al. presented an analysis of different association rules proposed by different authors to deal with quantitative data. In this paper, they have examined the rules of the association for various parameters and provided a comparative study of the tabular method. The direct application of organization management to the data may produce a large number of very important rules, so domain knowledge, data features, and application intent to be developed should be considered during the mining process. For future work, they have proposed a framework that involves association rules for data warehouses that exceed the above problems observed by various authors in current findings and implementation.

Jun Tan, Yingyong Bu et al. [16] have proposed Fd-Tree method that does not require scanning of the entire data stream and only scans the revised transactions simultaneously without involving the election generation. Test results in synthetic and real-time data show that the new algorithm surpasses other algorithms not only in terms of speed of algorithms, but also in their memory usage and robustness. They propose, a new approach to the expanding organization that governs mines over data sources. This method is based on the structure of the Fd tree that requires the scanning of the original data and only scans the updated ones simultaneously without involving the candidate generation.

Guoling Liu et al. [17] designed a new and efficient algorithm and proposed parameter C as a stamp. Using the lift, the noise rules are removed from the local branch mines. It avoids I/O transmission and network connectivity. It focuses on the number of resources and the level of support, thereby greatly increasing the accuracy of the global association rule mining. It only looks at all the details of a single database and avoids re-scanning of old data to retrieve new information from incremental addition to dataset. This paper discusses only the positive association rule mining over multiple databases. Future work will consist of negative cases of association mining over multiple databases.

Wenjuan Dong et al. [18] have proposed a new MIIWIU algorithm, the algorithm uses an improved Apriori algorithm and a set of common input elements. The traditional Algorithm Apriori treats each item as identical, ignoring the differences of all objects. The concept of weight is added to the algorithm. Negative association rules exist not only in frequent itemset, bur more exist in infrequent item set. Incremental updating algorithm is important for mining infrequent itemset in dynamic databases.

V.S Tseng et al. [19] propose two algorithms, namely utility pattern growth (UP-Growth) and UP-Growth+, for mining high utility item sets with a set of effective strategies for pruning candidate item sets. Mining high utility item sets from a transactional database refers to the discovery of item sets with high utility like profits. Although a number of relevant algorithms have been proposed, they incur the problem of producing a large number of candidate item sets for high utility item sets. Such a large number of candidate item sets degrades the mining performance in terms of execution time and space requirement.

S. Barua et al. [20] proposed a method for finding patterns based on a statistical test. The proposed model for the null distribution of features, spatial auto-correlation is taken into account and design an algorithm for finding both co-location and segregation patterns. Also it develops two strategies to reduce the computational cost compared to a naïve approach based on simulations of the data distribution, and proposed an approach to reduce the runtime of our algorithm even further by using an approximation of the neighbourhood of features.

III. PROPOSED VERTICAL APPROACH

In proposed vertical approach, category-wise layout is considered. The dataset is split into 21 unique category cluster. For each category, clusters are generated. Each of the Local FIS result is combined to get the Global FIS result.

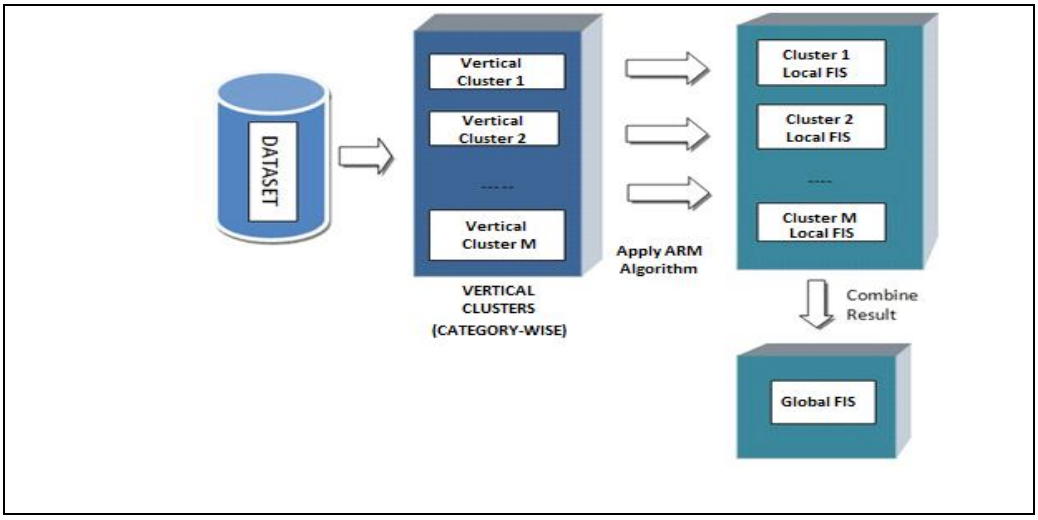


Fig. 4.3: Proposed Vertical Approach

Fig 4.3 shows the visualization of how Proposed Vertical approach works. It shows that how dataset is divided into 'M' clusters and each cluster generates their Local FIS. For FIS generation, the ARM algorithm is applied on each cluster. All Local FISs are combined and the Global FISs are stored in the Database.

A. General Concerns of Existing Normal Approach

In Normal and Horizontal approach, it was observed that due to constant Minimum support across all categories and clusters, it creates two specific issues:

1) *Loss of worthy FIS:* For categories that have lesser number of buyers but have more worth are discarded from Final FIS. Since the number of transactions are comparatively lesser in few categories, the minimum support threshold for that category must be lower.
2) *Inter-Category Issue:* For cases, where minimum support is lower, FIS generated contains irrelevant patterns of items that are practically possible but logically does not have any association. These issues can be resolved with the help of Proposed Vertical Approach. As each level of abstraction must have their own Minimum support threshold, the Proposed vertical Approach provides a multi-level view of the dataset by clustering w.r.t. to category of the product.

IV. EXPERIMENTATION RESULTS

Let us consider a dataset consisting of 50 Transactions (T1-T50) and 24 items (I1-I24).

Table 1: Transactional Dataset to demonstrate Proposed Vertical Approach

Transaction	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9
T1	Charger	Handset	Bread	Oil	Milk	Oven	Mixer		
T2	Handset	Bread	Rice	Tea	Oven				
T3	Battery	Handset	Cover	Bread	Sauce	Tea	Fridge	Mixer	Stove
T4	Handset	Bread	Oil	Sauce	Milk	TV	Oven	Stove	
T5	Battery	Cover	Selfistick	Rice	Tea	Milk	Dal	Fridge	Mixer
T6	Battery	Handset	Selfistick	Oil	Tea	TV	Iron	Oven	
T7	Powerbank	Oil	Tea	TV	Oven				
T8	Rice	Milk	Dal	Oven					
T9	Handset	Oven							
T10	Battery	Cover	Selfistick	Rice	Tea	Dal			
T11	Charger	Handset	Cover	Rice	Sauce	Tea			
T12	Battery	Powerbank	Selfistick	Oil	Sauce	Milk			
T13	Cover	Selfistick	Rice						
T14	Cover	Powerbank	Bread	Tea	Dal				
T15	Powerbank	Selfistick	Sauce	Milk					
T16	Charger	Battery	Cover	Bread	Oil	Tea			
T17	Handset	Cover	Selfistick	Oil	Sauce	Milk			
T18	Charger	Handset	Selfistick	Oil	Tea	Milk			
T19	Battery	Handset	Cover	Rice	Oil	Sauce			
T20	Handset	Cover	Selfistick	Bread	Dal				
T21	Charger	Powerbank	Bread	Sauce	Milk				
T22	Cover	Selfistick	Bread	Sauce	Tea				
T23	Battery	Cover	Selfistick	Rice	Oil	Sauce	Milk		
T24	Handset	Selfistick	Rice	Oil	Tea				
T25	Charger	Cover	Selfistick	Oil	Tea				
T26	Handset	Powerbank	Bread	Sauce	Dal				
T27	Battery	Cover	Selfistick	Oil	Sauce	Tea	Dal		
T28	Battery	Selfistick	Rice	Oil	Tea				
T29	Cover	Bread	Oil	Tea					
T30	Charger	Battery	Selfistick	Sauce					
T31	Charger	Bread	Oil	Dal					
T32	Handset	Cover	Oil	Sauce					
T33	Battery	Selfistick	Bread	Sauce					
T34	Powerbank	Selfistick	Rice	Milk	Dal				
T35	Handset	Selfistick	Bread	Tea					
T36	Charger	Cover	Tea	TV	Mixer	Stove			
T37	Battery	Powerbank	Oil	Sauce	Iron	Oven			
T38	Cover	Selfistick	Milk	Fridge	Oven				
T39	Powerbank	Battery	Bread	Tea	Dal	Fridge	Oven	Stove	
T40	Charger	Cover	Rice	TV	Oven				
T41	Battery	Selfistick	Oil	Sauce	Jeans	T-shirt	Fridge	Iron	
T42	Handset	Cover	Rice	Dal	T-shirt	TV	Iron	Stove	
T43	Cover	Bread	Oil	Tea	Jeans	Shirt	Oven		
T44	Battery	Powerbank	Selfistick	Bread	Oil	Fridge	TV	Oven	
T45	Charger	Battery	Selfistick	Oil	T-shirt	Shirt	Iron	Stove	
T46	Handset	Powerbank	Bread	Milk	Jeans	Fridge	Oven		
T47	Cover	Selfistick	Bread	Sauce	Dress	TV	Stove		
T48	Battery	Cover	Selfistick	Oil	Jeans	Fridge	TV	Oven	
T49	Charger	Cover	Oil	Tea	Dal	TV	Iron	Mixer	Stove
T50	Handset	Selfistick	Bread	Milk	Jeans	Dress	TV		

Solving above dataset using Apriori algorithm for Min. support = 0.2 with Normal approach,

Following were the Global FIS obtained using both the approach,

Table 2: FIS generated in Normal Approach

Normal Approach (FIS)		
S. No	Item1	Item2
1	Charger	
2	Sauce	
3	Tea	
4	Milk	
5	Dal	
6	Battery	
7	TV	
8	Oven	
9	Handset	
10	Cover	
11	Powerbank	
12	Selfistick	
13	Bread	
14	Rice	
15	Oil	
16	Battery	Selfistick
17	Battery	Oil
18	Cover	Tea
19	Cover	Selfistick
20	Cover	Oil
21	Selfistick	Sauce
22	Selfistick	Tea
23	Selfistick	Oil
24	Oil	Tea

As shown in Table 2, As shown in the Table 4.11, number of FIS generated using Normal Approach = 24 and particular frequent item sets such as {Fridge, Oven}, {TV, Oven}, {TV, Stove} are electronics product which are costlier than groceries/accessories but at the same time, they are less purchased by the buyers. In Normal Approach, such FIS are directly discarded due to a uniform Minimum support threshold for all categories

Now, Solving Table 1 dataset using Apriori algorithm for Min. support = 0.2 with Vertical approach,

1) For Proposed Vertical Approach No. of Clusters = No. of Categories,
2) For current dataset No. of Categories = 4 (Accessories, Grocery, Electronics, Clothing)

Category No.	Category Name	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Category-1	MOBILE	Charger	Battery	Handset	Cover	Powerbank	Selfistick	
Category-2	GROCERIES	Bread	Rice	Oil	Sauce	Tea	Milk	Dal
Category-3	CLOTH	Jeans	T-Shirt	Dress	Kurtis	Shirt		
Category-4	APPLIANCES	Fridge	Tv	Iron	Oven	Mixer	Stove	

Now, Table 3 shows the Global FIS obtained using Vertical the approach,

Table 3: FIS generated in Proposed Vertical Approach

Vertical Approach (FIS)		
S. No	Item1	Item2
1	Charger	
2	Sauce	
3	Tea	
4	Milk	
5	Dal	
6	Jeans	
7	T-shirt	
8	Dress	
9	Shirt	
10	Fridge	
11	Battery	
12	TV	
13	Iron	
14	Oven	
15	Mixer	
16	Stove	
17	Handset	
18	Cover	
19	Powerbank	
20	Selfistick	
21	Bread	
22	Rice	
23	Oil	
24	Fridge	Oven
25	Battery	Selfistick
26	TV	Oven
27	TV	Stove
28	Cover	Selfistick
29	Oil	Tea

As shown in the Table 2, number of FIS generated using Normal Approach = 24 whereas in Table 3, number of FIS generated using Proposed Vertical Approach = 29. With vertical approach, we observed that both issues that arise in Normal approach were handled in Vertical approach.

In the Vertical Approach, a reasonable Minimum support threshold is used based on number of transactions within that category, this results in the generation of FIS even from those categories.

Frequent item sets such as {Battery, Oil}, {Cover, Tea}, {Selfistick, Sauce}, etc these item sets, even though, they surpass the Minimum support threshold and does have a logical association.

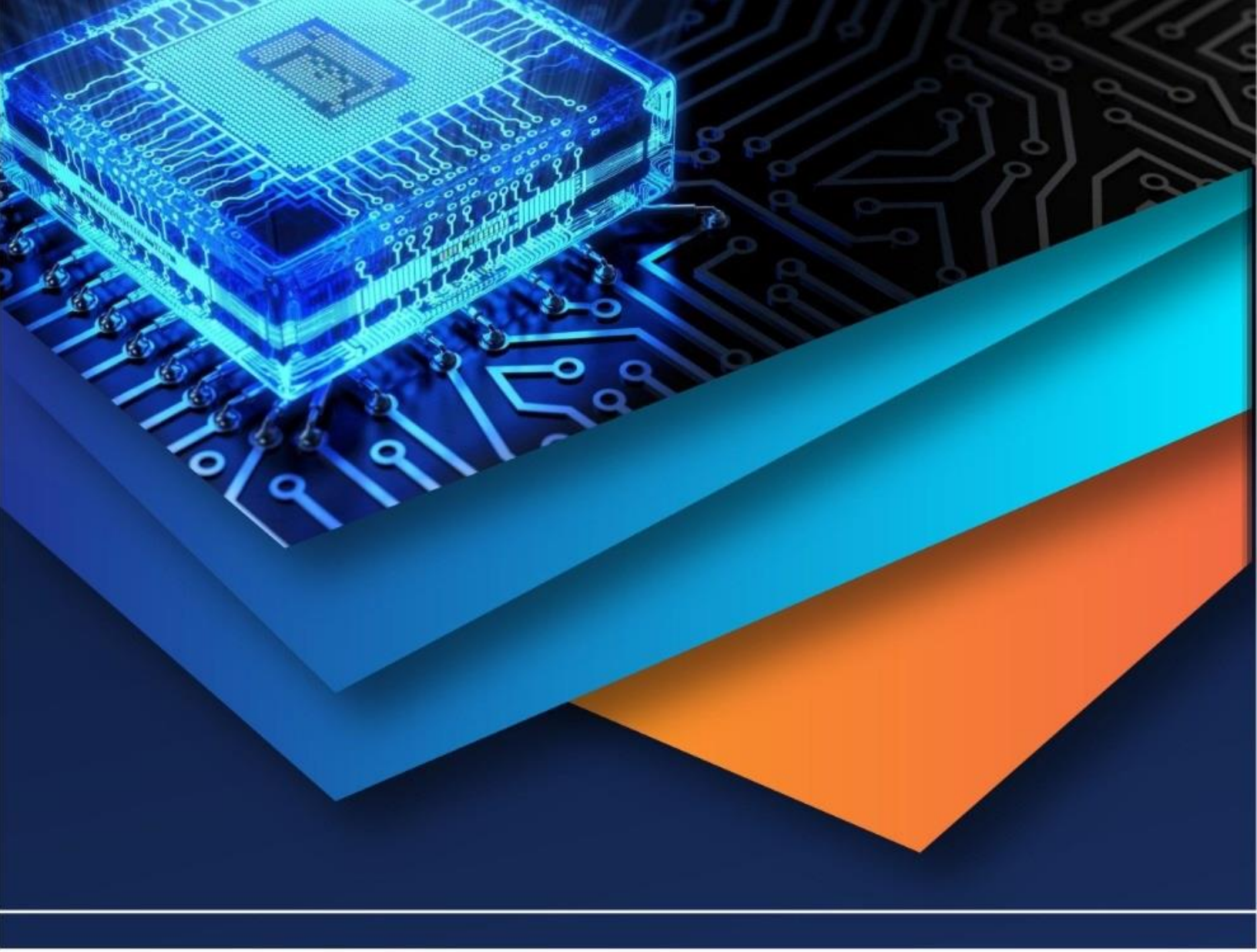
In Normal Approach, such FIS are considered but in the Vertical approach, the inter-category issue is resolved as clusters are formed of similar category items only.

V. CONCLUSION AND FUTURE WORK

It conclude that vertical approach, (A) Resolves Inter-Category issue of irrelevant FIS generation. (B) Generates FIS based on proportionate Minimum support values for clusters. (C)Produces FIS for few categories that are lesser in demand but have higher worth. In future, This paper can be extend to hybrid approach to generate association rule from the generated frequent itemsets in an efficient manner.

REFERENCES

[1] Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (A CM SIGMOD '93), Washington, USA, May 1993.
[2] S. P. Aditya, M. Hemanth, C. K. Lakshmikanth and K. R. Suneetha, "Effective algorithm for frequent pattern mining," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 704-708.
[3] Jian Pei, Jiawei Han, "Mining Frequent patterns without candidate generation," in SIGMOD Proceedings of the 2000 ACM SIGMOD international conference on Management of data, New York, NY, USA, 2000, pp. 1-12.
[4] S. Qianxiang and W. Ping, "Association rules mining based on improved PSO algorithm," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, 2017, pp. 145-149.
[5] C. Yadav, S. Wang and M. Kumar, "An approach to improve apriori algorithm based on association rule mining," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, 2013, pp. 1-9.
[6] Mohammed J. Zaki, "Scalable Algorithms for Association Mining," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, pp. 372-390, 2002.
[7] L.Wang et al., "Efficient Mining of Frequent Item Sets on Large Uncertain Databases, "IEEE Transactions on Knowledge and Data Engineering, vol.24, no. 12, pp. 2170 – 2183,Dec. 2012.
[8] Anil Vasoya, Novel Approach to Improve Apriori Algorithm using Transaction Reduction and Clustering Algorithm, International Journal of Applied Information Systems (IJAIS), pp. 37-44, 2014.
[9] Anil Vasoya, Dr. Nitin Koli,"Mining of association rules on large database using parallel and distributed computing", 7th International Conference on Communication, Computing and Virtualization 2016, Procidia, Computer science, pp.221 – 230,2016
[10] J. Yang, H. Huang and X. Jin, "Mining Web Access Sequence with Improved Apriori Algorithm," 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou, 2017, pp. 780-784.
[11] L. Zhan, F. Yu and H. Zhang, "A fast algorithm for mining temporal association rules based on a new definition," 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, 2017, pp. 1548-1553.
[12] X. Cheng and L. Cao, "Application Research of Improved FUP Algorithm in Network Security Linkage System," 2013 International Conference on Computational and Information Sciences, Shiyang, 2013, pp. 1624-1627.
[13] L. Jian-ping, W. Ying and Y. Fan-ding, "Incremental Mining Alogorithm Pre-FP in Association Rules Based on FP-tree," 2010 First International Conference on Networking and Distributed Computing, Hangzhou, 2010, pp. 199-203.
[14] S. Shan, X. Wang and M. Sui, "Mining Association Rules: A Continuous Incremental Updating Technique," 2010 International Conference on Web Information Systems and Mining, Sanya, 2010, pp. 62-66.
[15] A. Gosain and M. Bhugra, "A comprehensive survey of association rules on quantitative data in data mining," 2013 IEEE Conference on Information & Communication Technologies, Thuckalay, Tamil Nadu, India, 2013, pp. 1003-1008.
[16] Jun Tan, Yingyong Bu and Haiming Zhao, "Incremental maintenance of association rules over data streams," 2010 International Conference on Networking and Digital Society, Wenzhou, 2010, pp. 444-447.
[17] Guoling Liu and Runian Geng, "An efficient algorithm for mining association rules from multiple databases," 2010 2nd International Conference on Computer Engineering and Technology, Chengdu, 2010, pp. V4-349-V4-351.
[18] Wenjuan Dong, H. Jiang, Lei Chen and Guoling Liu, "Incremental updating algorithm for infrequent itemsets on weighted condition," 2010 International Conference on Computer Design and Applications, Qinhuangdao, 2010, pp. V1-36-V1-39.
[19] V. S. Tseng, "Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases," IEEE Transactions on Knowledge and Data Engineering, vol.25 , no. 8, pp. 1772-1786,Aug. 2013.
[20] S. Barua and J. Sander, "Mining Statistically Significant Co-location and Segregation Patterns," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 5, pp. 1185-1199, May 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY