IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

A Novel Way to Compute High Utility Itemsets

S. J. Vivekanandan[1], R. Reni Hena Helan[2], Harish K[3], Sriharish S[4], Vetrivel R.[5]
Assistant Professor, Dhanalakshmi College of Engineering, Chennai, India[1,2]
Students, Dhanalakshmi College of Engineering, Chennai, India[3,4,5]

Abstract: *One of the key fact-finding areas in data mining is association rule mining. For estimating association rules, there are numerous algorithms. These methods use confidence and support values, which are insufficient for real-time applications. High utility itemset is a new and rapidly expanding area of data mining that is highly beneficial in real-time applications. The user receives the important or more expensive stuff. Using the Novel utility-frequent apriori algorithm, we compute high utility itemsets in this study. This algorithm offers a great opportunity to alter the utility range in order to target the profit and obtain more valuable things.*

Keywords: Association rule mining, Support and Confidence, High Utility Itemsets, Range of utility

I. INTRODUCTION

A large amount of data is now being produced daily. To handle the data, an effective technique is therefore required. Data mining is an important strategy. The process of extracting new patterns from a body of data is known as data mining. Many other methods exist, including text mining, categorization, spatial mining, web mining, association rule mining, etc. One of the significant and rapidly expanding subfields of data mining is association rule mining. To estimate an association rule, numerous approaches are available. The Apriori algorithm and the Frequent Pattern tree method are well-known algorithms. For real-time applications, these algorithms' use of support value and confidence value is insufficient. They require products that are more profitable given the current state of business.

One of the fastest growing research areas among scientists is association rules mining (ARM). The aim of ARM is to find the strong association rules or the relationship between the itemsets in a large amount of data. The concepts of "what goes with what" and "the buying of one thing when we purchase another product" can be grasped in a single sentence. GH, where G stands for antecedent and H for consequent, is the symbol for association rules. In other words, if G happens, H might also happen. Market basket research mostly uses association rule mining to pinpoint consumer purchasing patterns.

Let I ={i1, i2, i3,..., in} represent a group of objects. Every transaction in the collection Tran DB is likewise a set of objects (t1, t2, t3,..., tn). Support and confidence can be used to gauge association rules. The number of times an item appears in the transaction as a whole is called support. Support (i1) is calculated as the ratio of the total number of transactions to the number of occurrences of i1 in the database.

Support (i1i2) is calculated as the ratio of the number of times i1i2 appear together in the database to the total number of transactions. Confidence is the likelihood that A will happen if B also happens, where A and B are two distinct item sets. Confidence (i1i2) is equal to the support-to-confidence ratio (i1:i2) (i1). When an item or group of items does not fall below the minimal support value, it is referred to as a frequent item or set of frequent items. When a product or product sets are smaller than the minimal support value, they are referred to as infrequent products or infrequent product sets. Finding powerful association rules is a process called association mining [15] [16]. It only need two steps to finish. In step 1, "Generating Frequent Itemsets," we compute all itemsets larger than or equal to the support value. Association Rule Generation comes next. Pull out all the rules that are greater than the confidence value based on the first step. Strong association regulations are what they are referred to as.

In general, there are two types of association rule mining algorithms: those that use a candidate generation strategy and those that do not. The popular algorithm for the first category is the apriori algorithm, and the popular algorithm for the second category is the frequent pattern tree algorithm. A well-known approach for determining the association rules is the apriori algorithm [15]. It operates via apriori properties. Property 1 states that all of an itemset's supersets are infrequent itemsets if the itemset itself is infrequent. Property 2 states that all subsets of a frequent itemset are also

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

frequent itemsets. The database can support boolean transactions, i.e. It only has one or zero entries. An unpurchased item or itemset is represented by the number 0. 1 indicates the acquisition of an item or collection of things. These algorithms serve as the foundation for a lot of the research in these areas

II. LITERATURE SURVEY

A lot of studies have been done in the area of association rule mining. Mr. Rakesh and others have presented an efficient algorithm that generates all significant association rules among items in the database. The authors proposed Apriori property which identifies frequent itemsets in a database [4], [5]. The authors have used sales data from a large retailing company and tried to find the associations between products by taking the minimum support=1% and minimum confidence=50%. They also assured the effectiveness of the estimation and pruning techniques by measuring accuracy

Authors Abdulsalam, Hambali and others used Apriori algorithm for market basket analysis. The authors tried to represent the sales pattern of a supermarket by representing six (6) distinct products across thirty (30) unique transactions. The authors assumed minimum support is 50% and tried to find the itemsets which are frequent by using Apriori algorithm in JAVA programming language [6].

Dhanabhakyam and Punithavalli highlighted Classification Dependent Predictive Association Rules(CPAR), Associative Classification, Classification Association Rule Mining(CARM), Distributed Apriori Association Rule, Six Sigma Technique and Apriori algorithm in their paper [1]. They marked out the advantages and disadvantages of the methods and tried to drag a conclusion that which method is better. According to authors among all the methods Apriori algorithm is found to be better for association but it has many difficulties. For this the authors proposed to combine fuzzy logic with Apriori algorithm which will return better result.

Again the authors Liu and Guan have used FP Growth in their paper which can solve the disadvantages of Apriori [2]. According to the authors, FP Growth constructs an FP tree which has highly compressed information. The authors have taken five transactions and generated the FP tree to discover the relationship between transactionsAuthorsJiangtaoQiu and others have tried to build a model of customers purchase behavior in the e-commerce context, which known as customer purchase prediction model(COREL) [7]. This model mainly has two stages. At first a candidate production collection is built by discovering associations among products and by this it predicts customers' motivations. The second stage is used to determine the most purchased candidate products based on customers' preferences. Authors have gained customers' information and product reviews from "Jingdong". The outcome of their paper showed that customers' preference plays a great role in purchasing decisions

Authors Kaura and Kanga proposed an approach to identify the changing trends of market data using association rule mining [8]. They at first described various techniques of data mining and then tried to describe why market basket analysis is important. They have used extended bakery datasets and tried to detect outliers. Also the authors suggested to extend this method to other areas.

III. PROPOSED METHODOLOGY

In this section, we discuss our proposed algorithm called the Novel Utility-Frequent Apriori algorithm. In a real-time scenario, frequent itemsets will not give significance to the user because it does not have utility value. On the other hand, items with only utility value do not make any guarantee that it occurs frequently. Therefore, we need a method that adopts frequent itemsets and utility value based itemsets. To overcome this issue, we proposed the Novel Utility-Frequent Apriori algorithm.

This proposed approach will be useful in E-Commerce to make more profit, in the Medical field to discover new diseases, and Banking sector to discover fraud activities. In day-to-day E-Commerce applications, few products are frequent sale but non profitable, few products are nonfrequent sale but profitable. So this approach will be helpful in categorization of itemsets based on sales and profit. In the medical field, to diagnose a disease, a list of predefined symptoms would be used. If we use traditional apriori, it will neglect the rare symptoms. But our proposed approach will take all rare symptoms which cause any new disease. Similarly, in the banking sector, abnormal transaction can be neglected by traditional apriori but our approach will take all transactions into the different categories of banking transactions..

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

3.1 Utility-Frequent Apriori model

The diagrammatic representation of utility-frequent apriori algorithm is Utility-Frequent Apriori Model. It is depicted in Fig. 1. min_sup, min_util, and IU & EU of the items are the input to the Utility-Frequent Apriori Model. In UtilityFrequentApriori Model, it executes the Novel UF-Apriori algorithm and categorizes the itemsets as HPHF, HPRF, LPHF, and LPRF.

3.2 Novel utility-frequent Apriori algorithm

Novel Utility-Frequent Apriori Algorithm – Categorizes the itemsets with support and utility

Input: Transaction Database (TB) with different Items (I_i), Profit table, min_sup and min_util

Output: Different categories of Itemsets with utility and support

Steps:

1. Initialize four empty sets such as HPHF, HPRF, LPHF, and LPRF
2. Scan the TB and for all $I_i \in$ TB do
   - 2a. Compute its support value i.e. Support (I_i)
   - 2b. Compute its utility value i.e. Utility (I_i)
3. For each transaction $T_i \in$ TB do
   - 3a. If (Ti is repeated) neglects it
   - 3b. Else, generate a combination of all $I_i \in T_i$
4. For each item $I_i \in$ combination do
   a. If (Utility(I_i) >= min_util) and (Support (I_i) >= min_sup) then add into HPHF
   b. If (Utility(I_i) >= min_util) and (Support (I_i) < min_sup) then add into HPRF
   c. If (Utility(I_i) < min_util) and (Support(I_i) >= min_sup) then add into LPHF
   d. If (Utility(I_i) < min_util) and (Support(I_i) < min_sup) then add into LPRF
5. Repeat steps 3 and 4, until there are no more transactions T_i.
6. Return HPHF, HPRF, LPHF and LPRF
7. Stop

3.3 Explanation of novel utility-frequent Apriori algorithm

Step 1 HPHF, HPRF, LPHF, and LPRF are initialized to null set i.e., empty set. Step 2 reads the database and checks the available item. For every item, it computes its utility value and support value. Step 3 takes each transaction from the database and generates the combination of all the available itemsets in each transaction. Step 4 takes each item from the combination, calculates its utility value and support value. Assigns itemsets into their categories, if itemsets were not present in their category already. Repeatsteps 3 and 4 until there is no transaction in the database. Step 6 displays all the categories of itemsets and stops the process.

Table 3 has six columns and five rows (i.e. transactions). The first column indicates the transaction identifier (TID) i.e. every transaction must have a unique identification. The second, third, fourth, fifth, and sixth column indicates the items A, B, C, D, and E respectively, and their purchased quantity in the transaction. Table 4 has two rows. The first row represents the different items available in the transaction and the second row represents the gain of the item. The external utility of an item, the transaction utility, and the total utility of the database are all evaluated through profit value.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

Table 3 Transaction table

Transaction ID	A	B	C	D	E
T101	0	0	18	0	1
T102	0	6	0	1	1
T103	5	0	4	0	2
T104	2	3	1	1	1
T105	0	0	4	0	3

Table 4 Profit table

A	B	C	D	E
2	11	4	7	5

By using Tables 3 and 4, we have simply discussed our algorithm. We have assumed min_sup = 30% and min_util = 100. Step 1 HPHF, HPRF, LPHF, and LPRF are initialized to null set i.e. HPHF = {}, HPRF = {}, LPHF = {} and LPRF = {}. In step 2, Support (A) = 40%, Support (B) = 40%, Support (C) = 80%, Support(D) = 40%, Support (E) = 100%, Utility (A) = 14, Utility (B) = 99, Utility (C) = 108, Utility (D) = 14 and Utility (E) = 40. In step 3, T101 contains only C and E so its combinations are (C, E, CE) and Support (C) = 80%, Utility (C) = 108, Support (E) = 100%, Utility (E) = 40, Utility (CE) = 143, Support (CE) = 80%. So HPHF = {C, CE} and LPHF = {E}.

T102 contains B, D and E so its combinations are {B, D, E, BD, BE, DE, BDE} and Support (B) = 40%, Utility (B) = 99, Support (D) = 40%, Utility (D) = 14, Support (E) = 100%, Utility (E) = 40, Support (BD) = 40%, Utility (BD) = 113, Support (BE) = 40%, Utility (BE) = 109, Support (DE) = 40%, Utility (DE) = 24, Support (BDE) = 40%, Utility (BDE) = 123. So, HPHF = {C, CE, BD, BE, BDE} and LPHF = {B, D, E, DE}.

T103 contains A, C and E so its combinations are {A, C, E, AC, AE, CE} and Support (A) = 40%, Utility (A) = 14, Support (C) = 80%, Utility (C) = 108, Support (E) = 100%, Utility (E) = 40, Support (AC) = 40%, Utility (AC) = 34, Support (AE) = 40%, Utility (AE) = 29, Support (CE) = 80%, Utility (CE) = 148, Support (ACE) = 40%, Utility (ACE) = 49. So HPHF = {C, CE, BD, BE, BDE} and LPHF = {A, B, D, E, AC, AE, DE, ACE}.

T104 contains A, B, C, D, E so its combinations are {A, B, C, D, E, AB, AC, AD, AE, BC, BD, BE, CD, CE, DE, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE ABCDE} and Support(AB) = 20%, Utility (AB) = 37, Support (AD) = 20%, Utility (AD) = 11, Support (BC) = 20%, Utility(BC) = 37, Support (CD) = 20%, Utility (CD) = 11, Support (ABC) = 20%, Utility (ABC) = 41, Support (ABD) = 20%, Utility (ABC) = 44, Support (ABE) = 20%, Utility (ABE) = 42, Support (ACD) = 20%, Utility (ACD) = 15, Support (ADE) = 20%, Utility (ADE) = 16, Support (BCD) = 20%, Utility (BCD) = 44, Support (BCE) = 20%, Utility (BCE) = 42, Support (CDE) = 20%, Utility (CDE) = 16, Support (ABCD) = 20%, Utility (ABCD) = 48, Support (ABCE) = 20%, Utility (ABCE) = 46, Support (ABDE) = 20%, Utility (ABDE) = 49, Support (ACDE) = 20%, Utility (ACDE) = 20, Support (BCDE) = 20%, Utility (BCDE) = 48, Support (ABCDE) = 20%, Utility (ABCDE) = 53, so HPHF = {C, CE, BD, BE, BDE}, LPHF = {A, B, D, E, AC, AE, DE, ACE}, LPLF = { AB, AD, BC, CD, ABC, ABD, ABE, ACD, ADE, BCD, BCE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE}.

T105 is a repeated transaction, so we can neglect this transaction. In the next step, it displays the output as different categories of Itemset i.e., HPHF = {C, CE, BD, BE, BDE}, LPHF = {A, B, D, E, AC, AE, DE, ACE}, LPRF = {AB, AD, BC, CD, ABC, ABD, ABE, ACD, ADE, BCD, BCE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE}.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

IV. SYSTEM DESIGN

The system's architectural layout and user interface are presented in this part.

Architectural Design

Fig. 1 shows the basic system components and their interactions.

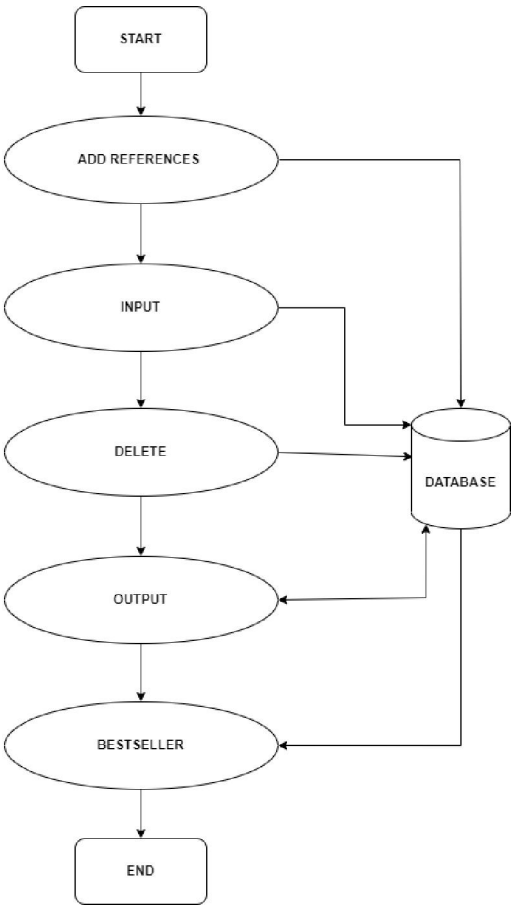

Fig 1. System Architecture

- Data Input Module: This module is responsible for receiving the transactional database as an input in the form of a CSV or TXT file.
- Preprocessing Module: This module is responsible for preprocessing the input data by removing duplicates, transforming the transactional format to a vertical format, and converting the item names to unique integer values.
- Candidate Generation Module: This module is responsible for generating the candidate itemsets using a hybrid approach that combines the strengths of Apriori and FP-Growth algorithms.
- Utility Calculation Module: This module is responsible for calculating the utility of each candidate itemset using a bottom-up approach that computes the utility of each itemset based on the utilities of its sub-itemsets.
- Pruning Module: This module is responsible for pruning the candidate itemsets that do not meet the minimum utility threshold or are redundant, based on utility-based pruning and monotonicity pruning.
- Itemset Selection Module: This module is responsible for selecting the high utility itemsets that satisfy the minimum utility threshold using a threshold-based approach.
- Output Module: This module is responsible for outputting the high utility itemsets in the form of a CSV or TXT file.
- User Interface: This module is responsible for providing a graphical user interface (GUI) for the user to interact with the system, select the input file, set the minimum utility threshold, and view the output file

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

V. IMPLEMENTATION

- **Data Preparation:** You will need a dataset to mine high utility itemsets. You can use any public dataset or prepare your own dataset. The dataset should be in a transactional format, where each row represents a transaction and each column represents an item.
- **Utility Measure:** You need to define a utility measure to calculate the utility of each itemset. The utility measure depends on the application domain. For example, in a retail store, the utility measure can be the profit earned from selling a particular item or a set of items.
- **Candidate Generation:** You need to generate a set of candidate itemsets that satisfy the minimum utility threshold. You can use any method to generate candidate itemsets, such as Apriori or FP-Growth. However, to improve the efficiency of mining, you can use a hybrid approach that combines Apriori and FP-Growth.
- **Utility Calculation:** You need to calculate the utility of each candidate itemset. To do this, you can use a bottom-up approach that starts from the individual items and combines them to form larger itemsets.
- **Pruning:** You need to prune the candidate itemsets that do not satisfy the minimum utility threshold or are redundant. You can use several pruning techniques, such as the utility-based pruning and the monotonicity pruning.
- **Itemset Selection:** You need to select the high utility itemsets that satisfy the minimum utility threshold. You can use any method to select high utility itemsets, such as the top-k approach or the threshold-based approach.

VI. RESULTS AND CONCLUSION

Since the substantial advancements in computer era, mining of big data to gain useful knowledge has been a hot topic studied from several aspects. One of the important fields of research is ARM which aims to uncover the subtle relations between the entries of huge bulk data so that some meaningful rules of associations can be generated. The obtained association rules can be exploited to identify which instances correlate in certain dimensions. ARM techniques have been used in many different areas ranging from retail industry to healthcare and diagnosis of illnesses. In this work, a comprehensive literature review on the existing algorithms of ARM is conducted with a special focus on the performance and application areas of the algorithms. Many algorithms have been proposed to discover association rules since the beginning of research in this area. The developed algorithms are, in general, classified into three main classes: (1) based on frequent itemsets, (2) based on sequential pattern, and (3) based on structures pattern. This classification mainly groups the algorithms based on the given structure of the dataset. Thus, the structure of the dataset subject to ARM study essentially determines the algorithm to be employed. Within each of these three classes, each algorithm provides superiority in certain aspects in comparison with each other. The above discussed algorithms were developed to improve the accuracy and decrease the complexity, and execution time. However, they do not always succeed to optimize all these objectives simultaneously

REFERENCES

[1]. S J Vivekanandan, G Gunasekaran. A Study on Utility-Frequent Itemset Mining. Journal of Seybold Report. Volume 15 Issue 9 2020; 3573-3581.

[2]. N. Alhusaini, S. Karmoshi, A. Hawbani, L. Jing, A. Alhusaini and Y. Al-sharabi, "LUIM: New Low-Utility Itemset Mining Framework," in IEEE Access, vol. 7,pp. 100535-100551, 2019, doi: 10.1109/ACCESS.2019.2929082.

[3]. Q. Yang, Q. Fu, C. Wang and J. Yang, "A Matrix-Based Apriori Algorithm Improvement," 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), Guangzhou, China, 2018, pp. 824-828, doi: 10.1109/DSC.2018.00132.

[4]. S J Vivekanandan, G Gunasekaran. An Improvisation Apriori Algorithm Applied in Medical Transcation. Journal of Green Engineering. Volume 10 Issue 10 October 2020; 8574-8586.

[5]. Sun, Ln. An improved apriori algorithm based on support weight matrix for data mining in transaction database. J Ambient Intell Human Comput 11, 495–501 (2020). https://doi.org/10.1007/s12652-019-01222-4.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, May 2023

[6]. Wang, C., Zheng, X. Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint. Evol. Intel. 13, 39–49 (2020). https://doi.org/10.1007/s12065-019-00234-5

[7]. G.K.Gupta, "Text Book: Introduction to Data Mining with Case Studies" 3rd edition, pp. 91-151 PHI Learning Pvt. Ltd. 2019

[8]. Mohammed Al-Maolegi, Bassam Arkok, "An Improved Apriori Algorithm for Association Rules", IJNLC vol. 3, No.1, February 2014.

[9]. P. S. Sandhu, D. S. Dhaliwal, S. N. Panda and A. Bisht, "An Improvement in Apriori Algorithm Using Profit and Quantity," 2010 Second International Conference on Computer and Network Technology, Bangkok, Thailand, 2010, pp. 3-7, doi: 10.1109/ICCNT.2010.46.

[10]. D.Narmada, G.Naveen, S.Geetha, " An efficient approach to prune Mined association rules in large databases", IJCSI, Vol.8, Issues 1, jan 2011.

[11]. L. Wu, K. Gong, Y. He, X. Ge and J. Cui, "A Study of Improving Apriori Algorithm," 2010 2nd International Workshop on Intelligent Systems and Applications, Wuhan, China, 2010, pp. 1-4, doi: 10.1109/IWISA.2010.5473450.

[12]. L. Ji, B. Zhang and J. Li, "A New Improvement on Apriori Algorithm," 2006 International Conference on Computational Intelligence and Security, Guangzhou, China, 2006, pp. 840-844, doi: 10.1109/ICCIAS.2006.294255.

[13]. Wang K, Zhou Q, Yeung, "Mining Customer value: From Association Rules to Direct marketing", Data Mining and Knowledge Discovery, Vol. 11, pp. 57-79 2005.

[14]. J.-S. Park, M.-S Chen, and P.S. Yu, " An Effective Hash based algorithm for mining association rules", Proc. of ACM-SIGMOD, pp. 175-186, May 1995.

[15]. Agrawal R, Srikant R, " Fast algorithms for mining association rules", Proceedings of 20th International Conf. on Very large Databases, Santiago, Chile, pp 487-499, 1994.

[16]. Agrawal R, Imielinski T., Swami A, "Mining association rules between set of items in large databases", Proceedings of the ACM SIGMOD Intl. Conf. on Management of Data, Washington, D.C.. may 1993, pp 207-216.

[17]. S J Vivekanandan, G Gunasekaran. A Survey on Association Rules Mining. Asian Resonance. VOL.- 8, ISSUE-1, pp. 1 – 4, January (Part-1) 2019. K. Liu, G. Motta, and T. Ma, "XYZ indoor navigation with augmented reality: A research in progress," Proceedings - 2016 IEEE International Conference on Services Computing, SCC 2016, pp. 299–306, 2016.

[18]. S J Vivekanandan, G Gunasekaran. A Survey on Utility Mining. Asian Resonance. VOL.- 8, ISSUE-1, pp. 5 – 9, January (Part-1) 2019.

[19]. S. Shankar, N. Babu, T. Purusothaman and S. Jayanthi, "A Fast Algorithm forMining High Utility Itemsets," 2009 IEEE International Advance Computing Conference, Patiala,India,2009,pp. 1459-1464, doi:10.1109/IADCC.2009.4809232..