Association Rule Mining: A Technique for Revolution in Requirement Analysis

Seema Maitrey*, C.K. Jha**

*Department of Computer Science and Engineering,
Krishna Institute of Engineering and Technology, Ghaziabad, India
**Department of Computer Science and Engineering,
Banasthali Vidyapith, Niwai , India.

***Abstract*-** Data mining is the process of extracting interesting, useful and previously unknown information or patterns from large information repositories such as: relational database, data warehouses, XML repository, etc. There are various types of data mining techniques such as association rules, classifications and clustering. Association rule mining is one of the most important and well researched techniques of data mining. Among sets of items in the transaction databases or other data repositories, it seeks interesting correlations, frequent patterns, associations or casual structures. Association Rule Mining is a very potential technique which has the aim to find interesting and useful patterns from the transactional database. It is mainly used in market basket analysis that help to identify patterns of all those items that are purchased together. To denote association with itemsets and their quantities, the Quantitative association mining is used. In this, we partition each item into equi-spaced bins with each bin representing a quantity range. It assumes each bin as a separate bin as we proceed with mining and we also take care to reduce redundancies and rules between different bins of the same item. Here, we make use of Association Rule Mining Technique to create a platform which helps in grouping similar objects together in a transaction process.

***Index Terms*-** *Data mining, Association Rule Mining, Market basket analysis, frequent itemset, Apriori Algorithm*

I. INTRODUCTION

Association Mining is a technique that finds its usage in the market basket analysis.[1] This technique, as can be said in general terms, is used in order to bring together items of the same type. Market basket analysis has also been used to identify the purchase patterns of the Alpha Consumer (people that play a key role in connecting with the concept that lie with a product, then accept that product, and finally validate it for the rest of society). The data collected and analyzed by this type of user has given opportunities to the companies to predict future buying trends and anticipate supply demands.

Data mining plays a role of highly effective tool in the catalog marketing industry. The various tools available in Data mining facilitate in identifying patterns among customers and help to identify the most liable customers to respond [2]. This mining technique is an emerging and promising and has been extensively studied [1][3]. Association rules show attributes value conditions that occur frequently together in a given dataset[4].

The market-basket problem assumes we have some large number of items, e.g., "bread," "butter". Customers select and then fill their market baskets with subset of the items, and we obtain the idea of the people about their buying of items together. This information is used by the marketers to mark the items, and control the way of navigating the store by a typical customer. For Association Rule Mining the terminology goes as follows: Let I = {i1,i2,….in} be a set of items and T be a set of transactions. Each transaction Ti (i = 0,1,…,m) is a set of items such that Ti ⊆ I. An itemset X is a set of items {i1, i2, …,ke}($1 \leq k \leq n$) such that X ⊆ I. An itemset containing k number of items is called k itemset. An association rule is an implication of the form, A=> B, where A ⊂ I, B ⊂ I and A ∩ B=φ. The rule A => B holds in T with support if S% of the transactions in T contain A and B. Similarity rule A=>B holds in T with confidence c if C% of transactions in T support A also support B. For given transaction T, the purpose of association rule mining is to determine all association rules that have supported and confidence greater than the user-specified minimum support *min_sup* and minimum confidence *min_conf.* Quantitative association rule mining refers to association rule forming between frequent items.

Association analysis can be used to improve decision making in a wide variety of applications such as:
telecommunication networks, market and risk management, inventory control, medical analysis, bio-science, survey of data, logistic failure, deception detection in web, customer resource management etc.

II. LITERATURE REVIEW

In data mining, association rule learning is a popular and well researched method to discover interesting relations between variables in huge databases. Association rules were introduced for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. It is based on the concept of strong rules. Market basket analysis rules are employed today in many application areas including Web usage mining, intrusion detection, regular production and biosciences. Association rule learning typically does not consider the order of items either within a transaction or across the transactions. In this way it is opposite to sequence mining [4]. Following the original definition the problem of association rule mining is defined as:

Let I={ $i_1,i_2,\ldots, i_n$} be a set of N binary attributes called items. Let D ={ $t_1,t_2,\ldots, t_m$ } be a set of transactions called the database. Each transaction done in D contains a distinctive transaction ID and includes a subset of the items in I. Rule for association mining can be defined as an inference of the form $X \Rightarrow Y$ where $X,Y \subseteq I$ and $X \cap Y = \phi$. The sets of items (for short itemsets) X and Y are called antecedent (left-hand-side or LHS) and consequences (right-hand-side or RHS) of the rule respectively.

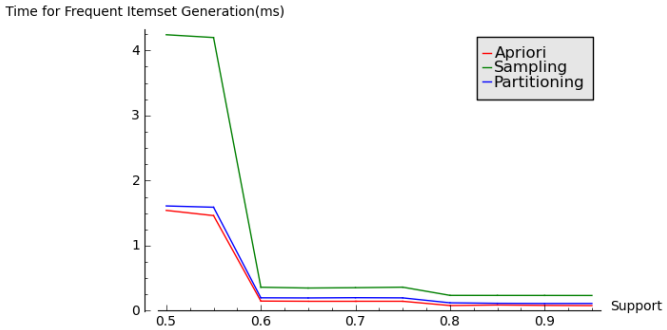


Fig 1: Time for frequent itemset generation vs support

The graph above shows how by using Apriori algorithm time required for frequent itemset generation requires the minimum time. Association rule generation is usually divided into two different steps[5]:

1. Minimum support is applied to find all frequent itemset in a database.
2. These frequent item sets and the minimum confidence constraint are used to form rules.

As it is found that the second step is straightforward, so more attention is required for the first step. It is difficult to find all frequent item sets in a database since it involves searching all possible item combinations. The set of possible item combination is the power set over I and has size 2^n-1 excluding the empty set. In [6] uses an association rule miner to generate high-confidence classification rules (*confidence* >90%) The uses an association rule miner to form rules that can describe individual classes. In [7], the rule discovery programs have been categorized into those that find quantitative rules and those that find qualitative laws.

As we know that the size of the power set grows exponentially in the number of item N in I, so an efficient search is possible using the downward-closure property of support (also called anti-monotonicity). It assure that for a recurrent itemset, all its subsets are also recurrent and thus for an occasional itemset, all its supersets must also be occasional. Efficient algorithms (e.g., Apriori and Eclat) can find all frequent itemsets by exploiting this property. The main costs of apriori's approach have two points: i) the cost of the candidate generation and ii) the cost involved in re-scanning of the database. In the stage of candidate generation, each frequent items at K-1 have to check each other to generate candidate itemset at K. In this step, it requires $O(N^2)$ where N is the number of frequent items at K-1. With the help of hashing technique the cost of generating candidate can be reduced. We can also reduce the cost of re-scanning database, especially x-TB dataset, by using bitmap-based[6] technique. In the following source code, we liberate apriori based approach:

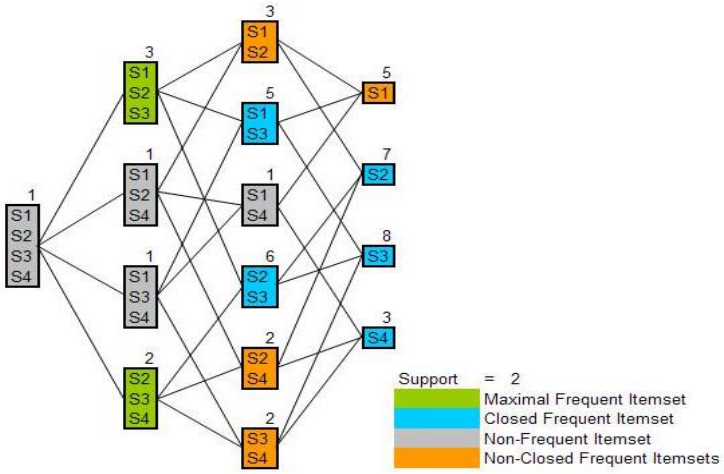


Fig 2: Frequent itemsets[8]

Apriori Algorithm

The first pass of the algorithm simply counts item occurrence to determine the large 1- itemsets. A subsequent pass k,consist of two phases , first the large itemset L_{k-1} found in the (k-1) the pass are used to generate the candidate itemsets C_k using the apriori function , the database is scanned and the support of candidates in C_k is counted.

Algorithm is described below

Apriori(T,c)	
1.	$L_1 \leftarrow$ { large 1 – itemsets};
2.	For (K=2; $L_{k-1} \neq \phi$;K++) do begin
3.	C_k= apriori gen(L_{k-1}); // New Candidates
4.	Forall transaction t ε D do begin
5.	C_t = subset(C_k, t); // Candidate contained in t
6.	Forall candidates c ε C_t do
7.	c.count++;
8.	end
9.	L_k ={ c ε C_k | c.count $\geq$ minsup}
10.	End
11.	Answer = $U_k L_k$;

Fig 3: Apriori algorithm steps

III. PROBLEM FORMULATION

The data of Market basket can be represented in a binary format in which each row correspond to an item. If the item is present in a transaction and possess value one, then it can be treated as a binary variable otherwise zero. The presence of an item in a transaction is often considered more important than its absence thus, an item is treated as an asymmetric binary variable. This representation is perhaps a very simplistic view of real

market basket data because it ignores certain important aspects of the data such as the quantity of items sold or the price paid to purchase them.

Itemset And Support Count

We know that Association rule mining is implemented by organizations to find out association rules among dataset that satisfy the predefined minimum support and confidence. The problem is usually. To make the analysis efficiently, the problem which is to be solved is make to decompose into two subproblems. The first one is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets. The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence.

Let $I=\{i_1,i_2,i_3,\ldots,i_d\}$ be the set of all items in a market basket data and $T=\{t_1, t_2, \ldots, t_N\}$ be the set of all transactions. The transaction t_i which contains a subset of items are selected from association analysis. If an itemset contains k items, it is called a k-itemset. For instance, {bread,butter,milk} is an example of 3-itemset. The null (or empty) set is an itemset that does not contain any items. Mathematically, the support count, $\sum(X)$, for an itemset X can be stated as follows: $\sum(X)=|\{t_i \mid X \subseteq t_i, t_i \in T\}|$, where the symbol |.| denotes number of elements in the set. Association rules are created as: i) by analyzing data for frequent if/then patterns and ii) using the criteria support and confidence to identify the most important relationships. In data mining, association rules are useful for analyzing and predicting customer behavior. The best-known constraints are minimum thresholds on support and confidence. It can be depicted as following:

.

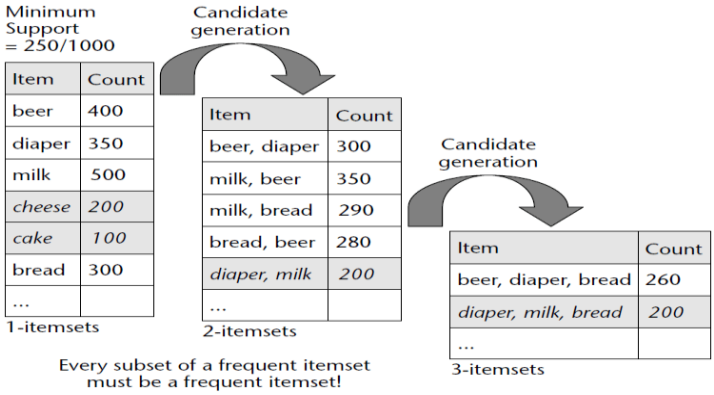


Fig 4: Frequent item set representation[7]

Association Rules find all sets of items (itemsets) that have *support* greater than the minimum support. It then use the large itemsets to create the desired rules having the *confidence* bigger than the minimum confidence. If X and Y were independent then the the *lift* of a rule is the ratio of the observed support to that expected.

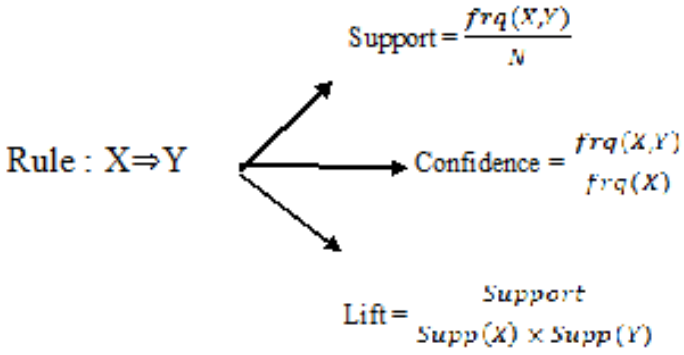


Fig 5: Levels in Association rule[9]

- Support: It is an indication of how frequently the items appear in the database. [10][13]
- Confidence : It determines the number of times the if/then statements have been found to be true.
- It deduce an estimate of the probability P (Y/X) , the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.[7]
- The lift of a rule is defined as lift $(X \Rightarrow Y) = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X) \times \text{Supp}(Y)}$ or the ratio of the observed support to that expected if X and Y were independent. The rule {milk,bread}⇒ {butter} has a lift of $\frac{0.2}{0.4 \times 0.4}$.
- The conviction of a rule is defined as $\text{Conv}(X \Rightarrow Y) = \frac{1-\text{Supp}(Y)}{1-\text{Conf}(X \Rightarrow Y)}$ .The rule{milk,bread}⇒ {butter} has a conviction of $\frac{1-0.4}{1-0.5}$ =1.2 , and can be interpreted as the ratio of the expected frequency that X occurs without Y (i.e., the frequency that the rule makes an inaccurate prediction) if X and Y were independent divided by the observed frequency of incorrect predictions. In the above example, the conviction value of 1.2 shows that the rule {milk,bread}⇒ {butter} would be incorrect 20% more often (1.2 times as often) if the association between X and Y was purely random chance.

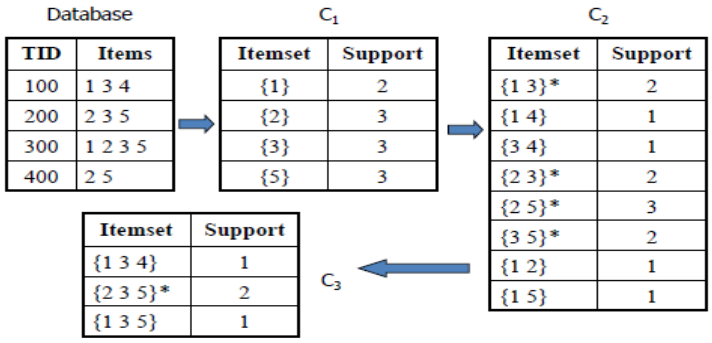


Fig 6: Processing of Association rule

Table 1: representing X→Y dependency

	X	→	Y
1	{A}	→	{B}
2	{A}	→	{C}
3	{A}	→	{B, C}
4	{B}	→	{A}
5	{B}	→	{C}
6	{B}	→	{A, C}
7	{C}	→	{A}
8	{C}	→	{B}
9	{C}	→	{A, B}
10	{A, B}	→	{C}
11	{A, C}	→	{B}
12	{B, C}	→	{A}

An initial step towards improving the performance of association rule mining algorithms is to decouple the support and confidence requirements. From the rule X→Y depends only on the support of its corresponding items, X ∪ Y. For example, the following rules have identical support because they involve items from the same itemset, {Bread, Butter, Milk}[14][15]:

{Bread,Butter}→{Milk},
{Bread, Milk}→{Butter},
{Butter,Milk}→{Bread},
{Bread}→{Butter, Milk},
{Milk}→{Bread,Butter},
{Butter}→{Bread, Milk}

If the itemset is infrequent, then all six candidates rules can be pruned immediately without our having to compute their confidence values. Therefore, a common strategy adopted by many association rule mining algorithms is to decompose the problem into following major subtasks[11]

• Frequent Itemset Generation, whose objective is to find all the itemsets that satisfy the minsup threshold. These itemsets are called frequent itemsets. Let *F* be the set of all frequent itemsets (w.r.t. some minfreq) in data *D*

• Frequent itemset $X \subseteq F$ is maximal if it does not have any frequent supersets

– That is, for all $Y \subseteq X$, $Y \subseteq F$

• Frequent itemset $X \subseteq F$ is closed if it has no superset with the same frequency

– That is, for all $Y \subseteq X$, $supp(Y, D) < supp(X, D)$

• It can't be that $supp(Y, D) > supp(X, D)$ [16].

IV. PROPOSED METHODOLOGY

Simply put our aim is to create Client Server interface similar to the one use in online Transaction System, For this we first use Integrated Development Environment for framing the presentation logic which in turn helps in establishing user interface as well as helps in forming designs for the project.[12] The most important is the database, which is the collection of all the product present and their details which tell about their availability, their cost, company, discount (if offered), product ID, customer requirement. This database is created using MYSQL server. The various tables included in the project are:

- Addproduct
- Admin_registration
- Adminlogin
- Changepassword
- Company_registeration
- Companyprofile
- Consumer
- Product
- Editprofile
- Feedback
- Purchase
- Transpay

This code is written using Java Server Pages (JSP). Various dynamic pages are created using JSP which are useful in the process of searching, carried out when a request is sent by the user to the server.

The code written in JSP cannot run directly on the server so JAVASCRIPT is used which helps to convert JSP code into server side code making it suitable enough to run easily on the server. Thus in the backend TOMCAT server is used for the same.

After the complete code is framed and is ready to run it is to be tested in order to carry out validation.

3.2 Modules

1.Admin
2.Customer
3.Company
4.Suggestions
5.Feedback

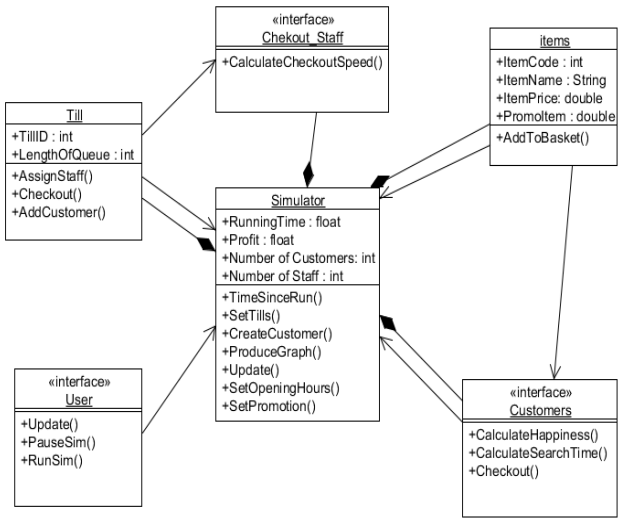


Fig 7: Describing structure of a System

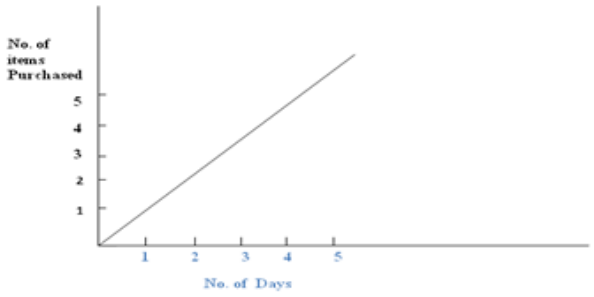
This class diagram illustrate the structure of a system by presenting : the system's classes, their attributes, operations (or methods), and the relationships among the classes. The various classes shown here are:

- Customer
- Items
- Simulator
- Staff
- User

Each of the given class possess their relevant attribute which describe how a particular class is used in an entire process. In the Security Mechanism there are two levels of security. The first level of security is provided by the FRONT END and the second level of security is provided by the database which is being used.

V. EXPERIMENTAL RESULTS

So here we are using Association mining Rule Technique which help us to provide frequent item set and help us to solve market basket problem. According to association mining Rule Technique ,it provide various suggestions to the user related to the product which he want to purchase. It is possible to retrieve the idea of the customer's purchasing behavior by analyzing the number of items purchased versus number of days. This can be shown by the following graph:



:

Fig 8: Number of items purchased vs number of days

Graphs to show requirement change:

Each product has its own requirement time. With time requirement of each product changes for the user. Thus to meet the changing requirement of the customer we use frequent item count and support count which show when a product is needed and when it is not. When an item is presented by 1, then the product is required and when 0, it means the item is not required at that time. This change is shown using the following graphs.

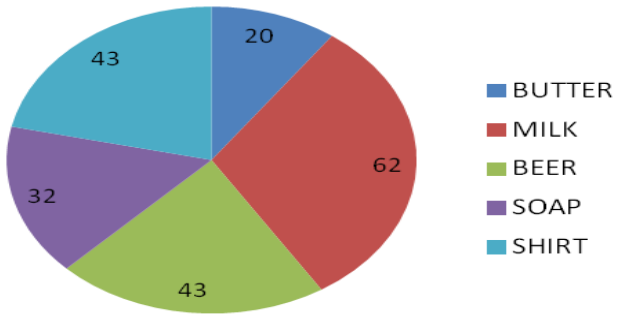


Fig 9 : Initial requirement of each product

Change in requirement of customer needs to be studied in detail. For example, milk might have requirement in one week and in the very next week the requirement of milk might change. This change is shown by the following graph:

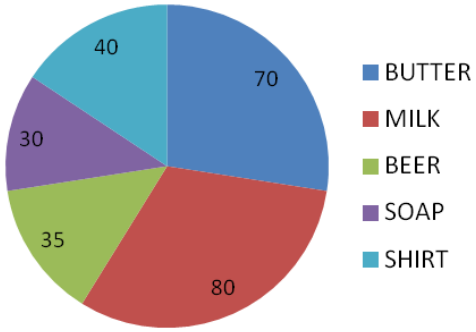


Fig 10: Change in requirement of each product

VI. CONCLUSION

There are several evidence of the success of this mission and there are millions of items listed each day in thousands of diverse categories. Any user may find it listed in the appropriate category, in any configuration from very old and outdated to the most recent greatest machine available.

Association rule mining has been applied to e-learning systems for traditionally association analysis, e.g., the following tasks: automatically guiding the learner's activities and intelligently generate and recommend learning material identifying attributes characterizing patterns of performance disparity between various groups of students .

In the association rule mining area, most of the research efforts went in the first consign to improve the algorithmic performance. In the second place, efforts were made to reduce the output set by allowing the opportunity to express constraints on the desired results. Over the past decade a variety of algorithms that address these issues were developed such as: through the refinement of search strategies, pruning techniques and data structures. While most algorithms focus on the explicit discovery of all rules that satisfy minimal support and confidence constraints for a given dataset.

VII. FUTURE SCOPE

In this paper we are using Association Rule Mining technique for generating suggestions for customer's ease. This technique is widely used for market analysis. When customers are aware of the different products present as an option for them they will buy more. This will increase our market graph and it will help to increase economy of any organization. Future Scope of Association Mining Rule can be e-learning . So by applying this technique we can provide various opportunities to freshers so that they can groom their skills and get better job options for their future. At present every organization is using web technology for their proper functioning, so this web based paper will play an important role from business point of view. Furthermore if this technique would be used, it can be prospective in job searching and any job seeker can register his/her self on the site to carry out their recruitment process turning this technique useful for Job Search. This rule when combined with the effective tools and techniques of web based technology, will become a great assistance in finding out the job opputunities.

REFERENCES

[1] A. Rajak . M. Gupta. "Association Rule Mining: Applications in Various Areas". In *International Conference on Data Management*
[2] H. S. Song, J. K. Kim and S. H. Kim, "Mining the change of customer behavior in an internet shopping mall", Expert Systems with Applications, 2001.
[3] Ali, K., Manganaris, S. and Srikant, R. 1997. "Partial classification using association rules.
[4] R Agrawal, R Srikant . "Fast algorithms for mining association rules". In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, 1994
[5] Agarwal, R., Imielinski, T., and Swami, A. "Mining association rules between sets of items in large databases." In: SIGMOD, 1993, pp. 207-216..
[6] Bayardo, R. J. 1997. Brute-force mining of high confidence classification rules. *KDD-97*, 123- 126.
[7] Sotiris Kotsiantis, Dimitris Kanellopoulos.Association Rules Mining: A Recent Overview . GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82
[8] *U. Ruckert, L. Richter, and S. Kramer. Quantitative association rules based* on half- spaces: An optimization approach. In: icdm, 00:507_510, 2004.
[9] Jiawei, H., Jian, P., Yiwen, Y., and Runying, M. "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree approach*.". In: DMKD, 2004, pp . 53-87.
[10] D. Sujatha and Naveen CH. "Quantitative Association Rule Mining on Weighted Transactional Data" . In *International Journal of Information and Education Technology, Vol. 1, No. 3, August 2011.*
[11] Nan Jiang and Le Gruenwald. "Research issues in data stream association rule mining". In ACM SIGMOD Record , Volume 35, Issue No. 1, Mar. 2006, pages 14-19.
[12] Yu, P., Own, C., Lin, L.: On learning behavior analysis of web based interactive environment. In: Proc. of the Int. Conf. on Implementing Curricular Change in Engineering Education (2001) 1-10.
[13] Rajanish, D., and Ambuj, M. "Fast Frequent Pattern Mining in Real-Time.", In: CSI, 2005, pp. 156-167.
[14] Y. Aumann and Y. Lindell. A statistical theory for quantitative association rules. In: Journal of Intelligent Information Systems, 20:255_283, 2003.
[15] Qiankun Zhao,Sourav S. Bhowmick.Association Rule Mining: A Survey. Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003.
[16] Association rule mining.Winter School on "Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets"pp .298-301.

AUTHORS

First Author – Seema Maitrey, M.Tech(Computer Engineering,and PhD-CSE(Pursuing), KIET,Ghaziabad, seema.maitrey@gmail.com.

Second Author – C.K. Jha,PhD(CSE),Banasthali Vidyapith , Niwai , Rajasthan, India.

Correspondence Author – Seema Maitrey , seema.maitrey@gmail.com, 09720877848