Discovery of Hidden Relationship in a Large Data Itemsets through Apriori Algorithm of Association Analysis with UML

Narander Kumar
Department of Computer Science,
B. B. A. University, Lucknow.

Vishal Verma
Department of Computer Science,
B. B. A. University, Lucknow.

Vipin Saxena
Department of Computer Science,
B. B. A. University, Lucknow.

ABSTRACT
An association rule is a method to find out the frequent hidden relationship from a large amount of datasets in a database. Association analysis into existing database technology is very useful for indexing and query processing capabilities of database system and developing efficient and scalable mining algorithms as well as handling user specified or domain specific constraints and post processing the extracted patterns.

In the present work, a methodology known as association analysis is presented which is very useful for discovery of interesting relationship hidden in large dataset, and an algorithm for generation of frequent data item set known as Apriori algorithm is used and validated the relations through Unified Modeling Language (UML). Authors used the lattice structure and also discussed the various association rules for the frequent data itemset which is found by Apriori algorithm. The different strategies in generation and traversal are breadth first and depth first search traversal. These techniques provide different tradeoff in terms of the input and output memory and computational time requirements. The entire concept is implemented by considering a real case study of Vehicle Insurance Policy system (VIPS) in context of Indian scenario.

Keywords
Association rule, Frequent data item sets, Apriori, Lattice structure, VIPS.

1. INTRODUCTION
The study of relationships between the different data items in a database through association rules for any sales transactions is important. Trailing of the buying patterns of consumer behavior through such rules is a very important before sale transactions. The problem of association rule generation has recently gained considerable prominence in the data mining community because of the capability of its being used as an important tool for knowledge discovery. Consequently, there has been a forge of research activity in the recent years surrounding this problem [1].

The problem of finding association rules is the first innovative idea has been given by Agrawal, Imielinski, and Swami [2]. This problem is concerned to find the relationships between different items in a database which contains the customer transactions. Such information can be used in different purposes which generates the revenue because the consumer's behavior of purchasing can be extrapolated from one another [3].

In the association rule problem, which described above may be a minimum level of support and confidence. There are mainly two phase procedures behind most of the association rule algorithms. In the first phase, all frequent data itemsets are found. If any data item is satisfying the user-concerned minimum support requirement that data item is known as frequent. In the second phase, for the generation of all the association rules (according to particular problem) for a problem then we use these frequent data item sets as described in the first phase. Frequent data item sets satisfy the user concern minimum confidence [4].

The formulation of association rules initially, considered that the effort towards the research has been devoted on it. There are different types of algorithms to generate the frequent data item sets proposed by various researchers [1, 2-8]. Different methodologies of association rules like generalized association rules, quantitative association rules as well as multilevel association rules have also been studied in [9-11].

In the present work a methodology known as association analysis which is useful for discovery of interesting relationship hidden in large dataset is described and an algorithm for generation of frequent data itemset by the technique known as Apriori algorithm which is presented in this paper. The lattice structure shows the various association rules for the frequent data itemset which is found by Apriori algorithm. The different strategies in generation and traversal such as breadth first and depth first search traversal are also demonstrated. These techniques provide different tradeoff in terms of the input and output memory and computational time requirements. The relations of associations rule are also described. For modeling purpose, the concept of Unified Modeling Language is used for a real case study of Vehicle Insurance Policy System.

2. RELATED WORK
From the literature, it is observed that a lot of work is done on the association rules for the relational database management system but limited work is available on the object-oriented database system. Let us first describe some of the important research contributions towards the association rules. In [12], algorithms for efficiently generating bases for association rules are described. A basis is a set of non-redundant rules from which all association rules can be derived, and it captures all, the useful information. Moreover, the size is significantly reduced as compared with the set of all possible rules because redundant and unimportant rules are discarded. New approach has two advantages on one hand, the user is provided with a smaller set of resulting rules, easy to handle, and improved quality information. On the other hand, execution times are reduced as compared with the discovering of all association rules. In the above paper, a case study of chess dataset is demonstrated as an experiment.

In a research paper [13], authors have proposed a new approach in finding association rules. This approach has the concept of rough set theory and Bayesian network classification to generate association rules and provides a way

for decision maker to get more information to generate association rules than traditional approach. The new approach for reviving association rules has the ability to handle the certainty in the classifying of the process so that they can reduce information loss and enhance the results of data mining. The algorithm can simulate the value of probability which is based on continuous data set.

Parvinder et al. [14] have proposed an approach that exploits the anti-monotone property of the Apriori algorithm that states that for a k-item set to be frequent all (k-1) subsets of this item set have also to be frequent. Afterword, the set of association rules mined that are subjected to Weightage (W-gain) and Utility (U-gain) constraints, for every association rule mined, a combined Utility Weighted score (UW-Score) is computed. At last they determine a subset of valuable association rules based on the UW-Score computed. Their results demonstrate the effectiveness of the proposed approach in generating high utility association rules that can be lucratively applied for business development.

Prasannal and Seetha [15] have discussed about an algorithm, which emphasized that mining the data from the cloud using sector/sphere framework which generates the association rules. Since cloud computing has very large processing data sets over the related clusters in which processing can be done through the given right programming model. They discussed about Apriori algorithm to mine the association rules. They define the cloud as an infrastructure, which provides the resources and services on the internet. This research work provides a review about the design and implementation of sector storage cloud and sphere compute cloud. They considered the distributed file system as a sector and parallel in-storage processing of data as sphere. Jiang and Gruenwald [16] have discussed some issues i. e. there exists emerging applications of data streams that require association rule mining such as network traffic monitoring and web click streams analysis. Different form of data in traditional static databases, data streams arrive continuously in high speed with huge amount and changing data distribution. From the above discussion, raises new issues that are need to be considered when developing association rule mining techniques for stream data.

Aggarwal [17] has proposed criteria which emphasis the importance of the actual correlation of items with one another with an algorithm which has a good computational efficiency and as well as maintain statistical robustness. Association rules in data sets which have varying density or even negative association rules have designed by the author. Temporal association rules are studied in [18] with an innovative approach and considered the time constraints and evaluated the performance of the proposed methods. Dunkel and Soparkar [19] have considered three important aspects; namely representation, organization and access of the data. When input and output costs are considered then the access of the data may significantly affect the performance. Authors also made comparison between column-wise and row-wise approaches of data access of Apriori association algorithm and found that counting in the Apriori algorithm with data accessed in the column-wise manner is better by reducing the degree to which the data and counters need to be repeatedly brought into the memory. Zhu et, al. [20] have developed methodology on the principle of the maximum entropy for studying the privacy consequences of data mining results. Aggarwal and Yu [21] have applied a graph-theoretic search algorithm which is proportional to the size of the output, on the stored pre-processed data in such a way that online processing may be done or in other words one can say that online mining. The algorithm is capable to find the rules with specific items in antecedent or consequent and also support for discovering association rules for large data item sets. The association rules help in the reduction of irrelevant noise in the data mining process. Cai et al. [22] have proposed a technique, where items have been given weights, which may be useful for special promotions on some products or maximize the revenue of different or particular items. Authors also discussed weighted association rules with weights and presented an algorithm to solve the problem of down-words closure property which support measure in the un-weighted case, no longer exist. The research work available in [23], has considered the problem of analyzing market-basket data and they presented some contributions for finding large item sets based on sampling. The idea of item reordering helps to improve the low level efficiency of the algorithm. The other important related references on data mining are [24-25].

3. ANALYSIS OF ASSOCIATION RULES

Many of the business enterprise gathered large quantities of data taken from routine work eg. huge amounts of customer purchase data items from counters of grocery stores considered as dataitems. In this reference, classical example of a data mining problem is market based basket analysis. Collect the information on how many items are purchased by the customers. The hope is, by finding out what products are frequently purchased jointly (that are associated with each other), that are able to optimize the marketing of the products for e.g. the layout of the store, by better targeting certain groups of customers. A famous example, the discovery of that people who buy diapers also frequently buy beers [25]. For the current work, let us consider the an example of Vehicle Insurance Policy System in Indian scenario which consists of items namely taken as Motor_Bike, Model,Color, Policy, Premium_Slave and Discount offered by the company. These are represented in the following Table 1.

Table 1: Items Sets for Vehicle Insurance Policy System

Sl.No.	ITEMS NAME
1	{Motor_Bike,Policy}
2	{Motor_Bike,Policy,Premium_Slave,Discount}
3	{color,Model,Policy,Premium_Slave,Discount}
4	{Motor_Bike,Policy,Premium_Slave,Discount}
5	{Color,Motor_Bike,Policy,Premium_Slave,Discount}

An item can be treated as a binary variable whose value is one if the item is present in the transaction and zero if not present in the transaction, but this representation ignores the price and quantity of the item. It does not have complete detail of the item. Binary representation of Vehicle Insurance Policy System is shown below in Table 2. The itemset is the set of items that occur jointly in transactions (e.g. the items bought) as represented in Table1.

Table 2: Binary Representation of VIPS

TID	Motor _Bike	Policy	Premium_ Slave	Disc ount	Color	Model
1	1	1	0	0	0	0
2	1	1	1	1	0	0
3	0	1	1	1	1	1
4	1	1	1	1	0	0
5	1	1	1	1	1	0

Let us define the support which is the number of transactions in which hold the association rule, i.e. in which all items of the rule occur.

Support = probability that body and head occur in transaction.

The other term confidence is the probability that in case the head of the rule (the condition) is satisfied as well as the satisfied the body of the rule (the conclusion). This indicates to which degree the rule is true, in the cases where the rule is applicable.

On the basis of support and confidence, let us define the association rules which are given below:

$A \Rightarrow B$ [s, c], where s=support and c= confidence
A, B item sets ($A, B \subseteq I$), where I is also a set;
$A \cap B$ empty;

support s = probability that a transaction contains $A \cup B = P(A \cup B)$;

confidence c = conditional probability that a transaction having A also contains B = P(B|A);

confidence = probability when if body occurs also head occurs;
Association rule is an implication expression of the form $X \rightarrow Y$ where X and Y are disjoint itemsets, i. e. $X \cap Y = \phi$. The strength of an association rule can be measured in terms of its support and confidence.

4 GENERATION OF FREQUENT ITEMSET IN THE APRIORI ALGORITHM

Apriori is the first association rule mining algorithm. Let us consider itemset of Table 1 with support count as described below:

(a) Candidate 1-itemset: Every item is taken as a candidate 1-item set. In this, if the item is appeared less than 3 then that item is discarded. For eg. {Color} and {Model} are discarded because count of Table 3 candidate 1-itemset is less than 3.

Table 3: Candidate 1-Itemset

ITEM	Count
Motor_Bike	4
Policy	5
Premium_Slave	4
Discount	4
Color	2
Model	1

Item set removed because of low support count

(b) Candidate 2-itemset: In the candidate 2-itemset we take collectively the combination of two itemsets. The Aprioiri principle ensures that all the superset of the infrequent must be infrequent. In this, there are three frequent 1-itemset, the number of candidate 2-itemset is

.

Table 3: Candidate 2-Itemset

Itemset	Count
Motor _Bike, Policy	4
Motor _Bike, Premium_Slave	2
Motor _Bike, Discount	2
Policy, Premium_Slave	4
Policy,Discount	4

Item set removed because of low support count.

(c) Candidate 3-itemset: Again according to the algorithm we take combination of transaction id 1 and 2 from the candidate2 itemset by counting their support .The itemset is discarded which is less than three minimum supports.

Table 4: Candidate 3-Itemset

Itemset	Count
Motor_Bike,Policy,Premium_Slave	3
Motor_Bike,Policy,Discount	3

With the Apriori principle we only need to keep candidate three itemsets whose subsets are frequent**.** The designed association rules are given below:

1. {Motor_Bike}⟶{Policy}
2. {Motor_Bike, Policy}⟶{Premium Slave}
3. {Motor_Bike,Policy,Premium_Slave}⟶{Discount}

4.1 Lattice structure: A lattice structure can be used to count the list of all possible itemsets in lattice structure if data sets that contain k itemsets. That it can contain 2^k-1 frequent itemsets excluding null sets. According to this the search space is very large for exploring the itemsets so that computational complexity is also very high. Hence there are various methods to reduce this complexity. Apriori principle is the effective method to remove the candidate itemsets without counting their support values.

According to the Apriori principle "if an itemset is frequent then all of its subset must also be frequent."

Inversely - "if an itemset is frequent then all its superset must be infrequent too".

For eg. If any transaction that contain {a, b, c} must also contain subset {a} {b}, {c}, {a, b} {a, c} & {b, c}

Let us consider a, b, c, d, e and f represented Motor_Bike, Policy, Premium_Slave, Discount, Color and Model. The lattice structure containing above by using Apriori algorithm is shown below in fig .1.
We observe that Apriori algorithm find frequent itemset & then scan the database to find the frequent itemset. The lattice structure shows the basic concept in association rule mining that an itemset can be frequent only if all the subsets of the itemset are frequent. For the itemset abc to be frequent must be frequent which in turn required that ab, ac, bc be frequent which is finally required abc to be frequent. We can say that Apriori algorithm work bottom up one level at a time by seeing at lattice structure.

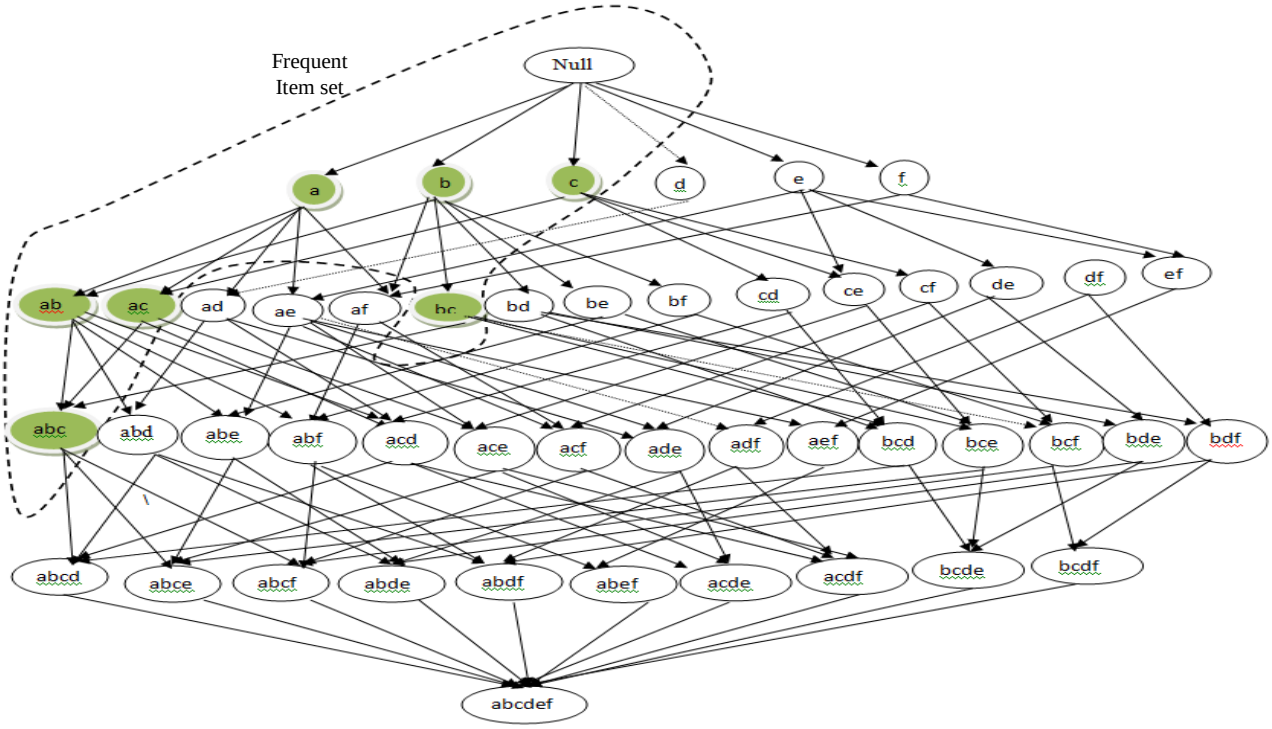


Figure1: Lattice Structure using the Apriori Algorithm

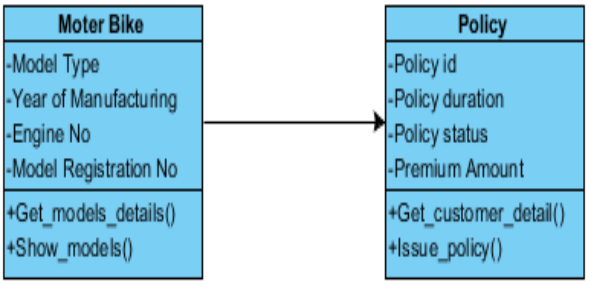
If compute the frequent 1- itemsets candidate 2-itemset then the frequent 2-itemsets. The number of scans of the transaction is equal to the maximum number of the itemset in a candidate database itemset. In this lattice structure we take each item from the Table of transaction level 0. At level 1 we combine the first item with all the other items by taking the joint operation & find the frequent itemset by counting the support count & generate candidate itemset similarly at the next higher level we generate candidate 2 itemset &so on.

Hence in conclusion we say that by forming the lattice structure we scan the database & find the all possible number of itemset as well as the frequent itemset and show the graphical view of our database.

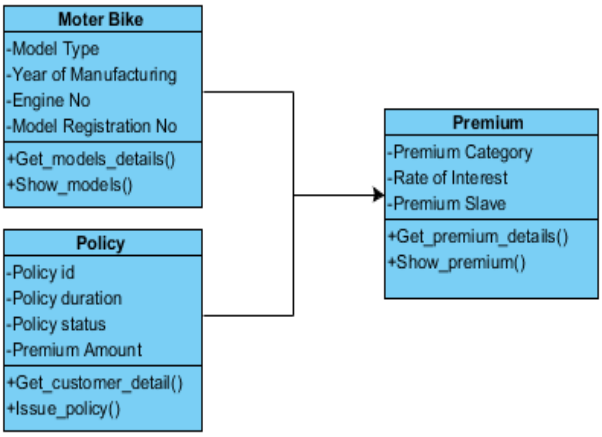
5. VALIDATION OF ASSOCIATION RULES THROUGH UML

Association analysis is useful for discovering of interesting relationships hidden in huge data sets. The uncovered relationships can be represented in the form of association rules [25]. From the Table, the association rules have been extracted from the data set as given above. It is observed that a strong relationship exists between the sale of Motor_Bike and Policy because there are many customers who buy Motor Bike and has to take Policy also as the insurance is compulsory. Agencies of motor bike can use this type of rules to help them and to identify new opportunities for cross selling their products to the customers. These rules are demonstrated by the use of UML. It is a well known and strong tool to represent diagrammatic modeling language which is used to model the software research problems. It is a very popular for analysis of the any problem. Modeling plays an important role for any system and contributes the understanding of the source inputs and outputs.

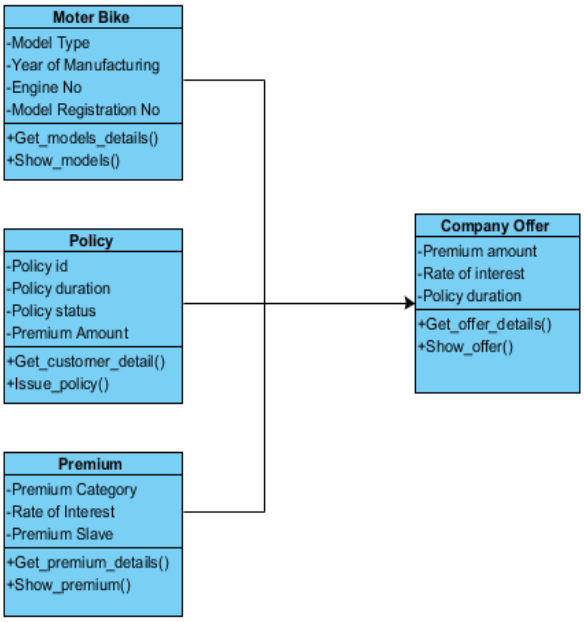
UML diagram shows the representation of the research problem. The UML class models for the aforesaid association rules are described below in figure 2. Vehicle Insurance Policy System (VIPS) contains four major classes along with attributes as shown below. In the UML model, Motor Bike class associates with Policy class; this means if any customer purchases a motor bike then has to take an insurance policy as represented in the figure 2(a). A customer who purchased the motor bike and took policy then has to pay the premium amount of policy as shown in figure 2(b). From the above association rules, if a customer has three things namely Motor_Bike, Policy and Premium amount then company provides some attractive offers according to premium slave to the customer for enhancing the numbers of customers as shown in figure 2(c).



(a)



(b)



(c)

Figure 2: UML Class Diagram of Vehicle Insurance Policy System

6. CONCLUDING REMARKS

From the above it is concluded that there are two key issues that need to be addressed when applying association analysis to Vehicle Insurance Policy System data. First discovering patterns from a large transaction datasets can be expensive. Second, some of the discovered patterns are potentially spurious because they may happen simply by chance. The proposed paper mainly emphasis around these two issues. First is to explain the basic concepts of association analysis and the algorithms used to efficiently mine such patterns of datasets as discussed in the work. Second, it deals with the issues of evaluating the discovered patterns in order to prevent the generation of spurious results. For future direction, the association analysis is also applicable to other application domains such as bio-informatics, medical diagnosis, web mining, scientific data analysis, etc.

7. REFERENCES

[1] Aggarwal, C.C. and Yu, Philip S., "Mining Large Itemsets for Association Rules", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, pp 1-9, 1998.

[2] Agrawal, R. C., Imielinski T. and Swami A., "Mining Association Rules between Sets of Items in Very Large Databases." Proceedings of the ACM SIGMOD Conference on Management of Data, pages 207-216, 1993.

[3] Agarwal, R. C., Aggarwal C. C., Prasad V. V. V., and Crestana V., "A Tree Projection Algorithm For Generation of Large Itemsets for Association Rules." IBM Research Report, RC 21341.

[4] Agrawal R., Mannila H., Srikant R., Toivonen H. and Verkamo A. I., "Fast Discovery of Association Rules." Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Chapter 12, pages 307-328. Proceedings of the 20th International Conference on Very Large Data Bases, pages 478-499, 1994.

[5] Bayardo R. J., "Efficiently Mining Long Patterns from Databases." Proceedings of the ACM SIGMOD, pages 85-93, 1998.

[6] Lin D. and Kedem Z. M., "Pincer-Search: A New Algorithm for Discovering the Maximum Fre-quent Itemset." EDBT Conference Proceedings, pages 105-119, 1998.

[7] Savasere A., Omiecinski E. and Navathe S. B., "An ancient Algorithm for Mining Association Rules in Large Databases." Proceedings of the 21st International Conference on Very Large Databases, 1995.

[8] Toivonen H., "Sampling Large Databases for Association Rules". Proceedings of the 22nd In-ternational Conference on Very Large Databases, Bombay, India, September 1996.

[9] Han J. and Fu Y., "Discovery of Multi-level Association Rules From Large Databases." Proceedings of the International Conference on Very Large Databases, pages 420-431, Zurich, Switzerland, September 1995.

[10] Srikant R. and Agrawal R., "Mining Generalized Association Rules." Proceedings of the 21st International Conference on Very Large Data Bases, pages 407-419, 1995.

[11] Srikant R. and Agrawal R., "Mining quantitative association rules in large relational tables", Proceedings of the ACM SIGMOD Conference on Management of Data, pages 1-12, 1996.

[12] Ramaraj, E., Gokulakrishnan, R. and Rameshkumar, K. "Information Quality Improvement through association rule mining algorithms DFCI, DFAPRIORI-CLOSE, EARA, PBAARA, SBAARA." Journal of Theoretical and Applied Information Technology, pp 948-960, © 2005 – 2008.

[13] Venkateswara Rao Vedula and Thatavarti, S. "Binary Association Rule Mining Using Bayesian Network" International Conference on Information and Network Technology, vol.4, pp 171-176 © (2011) IACSIT Press, Singapore.

[14] Sandhu, P.S., Dhaliwal, D.S. and Panda, S.N. "Mining Utility-Oriented Association Rules: An Efficient Approach Based on Profit and Quantity," International Journal of the Physical Sciences Vol. 6(2), pp. 301-307,

18 January, 2011. Available online at http://www.academicjournals.org/IJPS ISSN 1992 - 1950 ©2011 Academic Journals.

[15] Prasanna1, K. and Seetha, M. "Association Rule Mining Algorithms for High Dimensional Data", International Journal of Advances in Engineering & Technology, ISSN: 2231-1963, pp. 443-454, Jan 2012.

[16] Jiang, N. and Gruenwald, Le "Research Issues in Data Stream Association Rule Mining", SIGMOD Record, Vol. 35, No. 1, pp 14-19, Mar. 2006.

[17] Aggarwal, C. C., "Mining \Associations with the Collective Strength Approach" IEEE Transactions on Knowledge and Data Engineering, , Volume: 13, Issue: 6, pp-863-873, Nov/Dec 2001.

[18] HuiNing, Haifeng Yuan and Shugang, Chen, "Temporal Association Rules in Mining Method," First International Multi-Symposiums on Computer and Computational Sciences, Volume: 2, pp739-742 (IMSCCS '06). 20-24 June 2006.

[19] Dunkel, B. and Soparkar, N. , "Data Organization and Access for Efficient Data Mining", 15th International Conference on Data Engineering, Proceedings, pp-522 – 529, 23-26 Mar 1999.

[20] Zutao, Zhu, Guan, Wang, and Wenliang, Du, "Deriving Private Information from Association Rule Mining Results"

[21] Aggarwal, C.C. and Yu, P.S., "A New Approach to Online Generation of Association Rules", IEEE Transactions on Knowledge and Data Engineering Volume 13, Issue: 4, pp 527 - 540 Jul/Aug 2001.

[22] Cai, C.H., Fu, A.W.C., Cheng, C.H. and Kwong, W.W., "Mining Association Rules with Weighted Items", International Proceedings. IDEAS'98 Database Engineering and Applications Symposium, pp-68–77, 1998.

[23] Brin, Sergey, Motwani, Rajeev, Ullman, Jeffrey D. and Tsur Shalom, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", International Conference proceedings ACM SIGMOD, pp- 255-264, May 13-15, 1997.

[24] Walter A. Kosters, Elena Marchiori, and A. J. Oerlemans, "Mining Clusters with Association Rules", Proceedings of Symposium on Advances in Intelligent Data Analysis, ISBN:3-540-66332-0 IN 1999.

[25] Tan, Pang-Ning, Steinbach, Michael and Kumar, Vipin, "Introduction To Data Mining", Pearson Education, ISBN 978-81-317-1472-0, Fourth Edition, 2009.

8. AUTHOR'S PROFILE

Dr. Narander Kumar received his Post Graduate Degree and Ph. D. in CS & IT, from the Department of Computer Science and Information Technology, Faculty of Engineering and Technology, M. J. P. Rohilkhand University, Bareilly, Uttar Pradesh, INDIA in 2002 and 2009, respectively. His current research interest includes Quality of Service (QoS), Software Engineering, Computer Networks, Resource Management Mechanism, in the networks for Multimedia Applications, Performance Evaluation. Presently he is working as Assistant Professor, in the Department of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, INDIA.

Vishal Verma is a research scholar in Department of Comupter Science, Babashaheb BhimRao Ambedkar University, and Lucknow, India. Earlier he got his Master of Computer Application (MCA) from the above University and presently he is working on Data Mining Applications through UML.

Vipin Saxena is a Professor and Head, Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India. He got his M.Phil. Degree in Computer Application in 1992 & Ph.D. Degree work on Scientific Computing from University of Roorkee (renamed as IndianInstitute of Technology, Roorkee, India) in 1997. He has more than 17 years of teaching experience and 20 years of research experience in the field of Scientific Computing & Software Engineering. He has published more than hundred International and National research papers and authored four books in the Computer Science field. Dr. Saxena is a life time member of Indian Science Congress.