ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Horizontal Aggregation Function Using Multi Class Clustering (MCC) and Weighted (PCA)

Dr. K. Sathesh Kumar[1], P. Sabiya[2], S.Deepika[2]

Assistant Professor, Department of Computer Science and Information Technology, Kalasalingam University, Krishnankoil, Virudhunagar (Dt). India[1]

PG Scholars, Department of Computer Science and Information Technology Kalasalingam University, Krishnankoil, Virudhunagar (Dt). India [2]

Abstract: Data transformation and aggregation is the significant portion in data mining for data analysis and data set preparations. In a relational database environment, building such data set requires joining tables and aggregating columns from different dynamic tables. Several aggregation functions based on the SQL operations have been initiated for multi table aggregation by applying vertical joints. Such previous SQL aggregations are limited since they return a single number static data group. These aggregations worked well in the form of static datasets, but a major effort is still required to build data sets suitable for data mining purposes, where a tabular format is generally required and which need frequent updates. This suggested work proposes a very simple and effective summarization based dynamic join operations over high dimensional dataset. These extents the SQL aggregate functions to produce aggregations in horizontal form, returning a set of numbers instead of single aggregation. The research work also proposes a Multi Class Clustering (MCC) and Weighted PCA method to handle a high dimensional dynamic dataset with summarization technique. In the proposed technique, there are two common data preparation tasks are enlightened which includes transposition/aggregation and transforming categorical attributes into summarized labels. This executes the basic methods to evaluate horizontal aggregations which are named as CASE, SPJ and PIVOT respectively.

Keywords: aggregation, Weighted PCA method, MCC (Multi class clustering)

I. INTRODUCTION

The data has been growing every day because of inventing technology and decision-making is essential to improve the business standards. Analyzing the data in each place is hard to take decision and time-consuming. There is a problem to gathering information from different site and different repositories such as data redundancy, data duplication, data collapse. To defeat this issue, Data mining plays an important role to visualize the data by user defined view. Data Mining (DM) is the process of analytical data from different perspectives and summarized it into useful information. In DM, Data aggregation is a process of grouping the information and expressed in summary form. [14] It purposes are to give particular information about the group of specific variables such as age, profession or income. The information can be used for website [15-18] personalization (commonly used to enhance customer service or e-commerce, sales, personalization is sometime called as one to one marketing). The simplest type of data aggregation is an OLAP (On-line Analytic Processing) in which the marketer uses an on-line reporting mechanism to process the information. This work suggests the aggregation; this is a form of data redundancy, which is computed from other warehouse values. In some per-calculated, average may need to be recomputed as new data and loaded into our data warehouse. Aggregation is mostly used in the component of Business Intelligence [28] (BI) solutions. Aggregation is the person or software search databases, which find the relevant search query data and present data in a summarized format that is used for end user or application [19].

Data aggregation generally works with big data and data marts which will not provide whole information. The issue of data aggregation is occurs when a large amount of data collection on a high security level than individual component of the record. Aggregation is used in dimensional models of the data warehouse to produce dramatic positive effects [30]. Aggregation is a simple summary table that can be grouped by SQL query. The most common use of aggregate is to take dimension and change the granularity of this dimension. Aggregation is referred as pre-calculated summary data. This pre-computing and summarized data are stored in a new aggregated table. When facts are aggregated, it is associated with rolled up dimension. The reason to use aggregation is to increase the performance of the data warehouse in reduction of the number of rows. The aggregation is determined by every possible combination of dimensional granularities. When the request is made by

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

the user the data warehouse should return data from the table with the correct grain [20]. The best way to build the aggregation is to monitor queries and design aggregation to match query patterns [29]. Aggregation data in dimensional model is more complex. To make extra complexity transparent to the user, the functions used to know as aggregate navigation, which is used in query dimensional and fact table. The aggregation navigator is implemented in a range of technology they are as follows: LAP engines, Materialized views, Relational OLAP server and BI application server or query tools [21].

Problem definition

Generally data mining tasks require summarizations that are not readily available from the database. Such requirements typically need computing aggregations at several levels with several segmented datasets. This is because most techniques required by data mining for horizontal and vertical joins, which translates as sums or counts computed with SQL. Unfortunately, such techniques are not hierarchical, which makes the use of separate summary table's necessary.Several other techniques suffer from the high dimensional and dynamic datasets. Those techniques are failed to adapt to such summarized strategies in real world datasets.In a relational database environment with normalized tables, a significant effort is required to prepare a summary data set in order to use it as input for a data mining algorithm. In common all existing data aggregation technique performs single tabular aggregation, where the database should scan every time of aggregation function. When this problem exploits in the high dimensional databases, produces a huge number of delay in operations [22].

II. RELATED WORKS

SQL is widely used in real world applications as it supports interaction with various relational databases. SQL has simple and effective commands to perform operations such as DML, DDL, TCL and DCL. They also support sub queries, joins and aggregations. Optimization of queries is possible by using certain performance tuning activities such as indexing. As part of SQL queries, aggregate functions like MIN, MAX, COUNT, AVG and SUM can be used to get summary of data [1].The aggregate functions provided by SQL generate single row output. They cannot provide data in horizontal layout which is essential for data mining applications. One of the data mining techniques is association rule mining which is widely used in OLAP processing [2]. In [3] an extension to SQL aggregate functions is made for efficient data mining solutions. The result of such aggregate functions can provide data in horizontal layout which is useful for data mining. Clustering algorithm [1] is one of the data mining algorithms that make use of SQL internally in order to perform clustering. SQL extensions provided by them have optimizations for joins but not for resultant groups. For this reason joins can be avoided using PIVOT and CASE constructs. For new class of aggregations in [5] relational algebra is used. Such aggregations are known as horizontal aggregations. In this paper also we focus on the horizontal aggregations such as CASE, PIVOT and SPJ. In this paper also we focus on the horizontal aggregations such as CASE, PIVOT and SPJ. Optimizing joins is also presented in [6] but that is not useful for large queries There was lot of research on aggregations and related optimizations that include cross tabulation [7]. The results contain multiple attribute – value pairs for horizontal aggregations. Transforming data from one format to another format can be done using SQL operations [8]. M. Madhavi and S. Kavitha[10] Experiments with large tables compare the proposed query evaluation methods. CASE method has similar speed to the PIVOT operator and it is much faster than the SPJ method. The CASE and PIVOT methods exhibit linear scalability, whereas the SPJ method does not. Rajesh Reddy Muley, Sravani Achanta and Prof.S.V.Achutha Rao, in [11] explains us the way to use the data mining methods to show the datasets by mining the data from different tables at the same time. Themethods which are suitable for data mining analysis are CASE, SPJ and PIVOT. Coming with CASE they show two possibilities i.e. Vertical view and also the Horizontal View. This paper thus satisfies the main concern i.e. reducing the overload on the databases for retrieval of data. This paper [9] horizontal aggregations can be used as a database method to automatically generate efficient SQL queries with three sets of parameters: grouping columns, sub-grouping columns and aggregated column. The fact that the output horizontal columns are not available when the query is parsed (when the query plan is explored and chosen) makes its evaluation through standard SQL mechanisms infeasible. Our experiments with large tables show our proposed horizontal aggregations evaluated withthe CASE method have similar performance to the built-in PIVOT operator. In a clustering algorithm is explored which makes use of SQL queries internally. It is capable of viewing horizontal layout for further mining operations. SQL extensions to define aggregate functions for association rule mining. Their optimizations have the point of avoiding joins to correspond cell formulas, but are not optimized to achieve limited transposition for each group of effect rows.

III.MATERIALS AND METHODS

A. Proposed method

The proposed system concentrates on the process of horizontal aggregation with automated code using Hybrid techniques; this technique applies the weighted PCA algorithm with existing CASE tool in order to support the high dimensional dynamic dataset. The technique which helps to overcome the existing size and time oriented problems. Attribute linkage is the task of identifying diverse entries that refer to the same entity across

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

different data sources. This helps to aggregate, summarize and link more tables together. Instead of scanning the whole database the proposed system utilizes the summarized aggregation, where temporally segmented portion will be aggregated. So the proposed system applies the temporal aggregation algorithm. The hybrid technique (Multi Class Clustering MCC) has been introduced, which contains the summarization, aggregation and linkage with the considerations of dynamic dataset [23-25].

B. Weighted PCA Aggregation

Principal component analysis (PCA) is a well established technique for dimensionality reduction. The reputation of PCA comes from three important properties.

First, it is the optimal (in terms of mean squared error) linear scheme for compressing a set of high dimensional vectors into a set of lower dimensional vectors and then reconstructing.

Second, the model parameters can be computed directly from the data for example by diagonalizing the sample covariance.

Third, compression and decompression are easy operations to perform given the model parameters they require only matrix multiplications.

Most DBMS provide capabilities to embed code into SQL server. These features include user-defined aggregate functions (UDFs, also called user-defined aggregates) and stored procedures (SPs).

Let X is assumed to be stored on a table with a horizontal layout by default (d dimension columns per row), to enable high dimensional fast data identification. The main limitation of the existing system is the row size limit imposed by the Data base management systems. A possible solution which is explored in this work is to horizontally partition X and joins the partitions for processing. Alternatively, X can be stored with a vertical layout, with one dimension per row for high d and a sparse matrix X (eliminating zeroes and nulls), removing any d limitations. The aggregate user defined aggregate functions for a vertical layout requires clustered storage for dimension values to allow block-based processing[26].

Method (1) is the most portable and easiest to program. This function overcomes the DBMS limitations on a maximum number of columns when X has a horizontal layout.

Method (2), based on a MCC (Multi class clustering), uses a reader () function to scan X table rows, one at a time and calculates n, L, Q. The MCC runs sequentially on a single thread and it can create lists, arrays, or any enumerable object, casting results as relational tables. The MCC approach is more portable compared to an aggregate of existing aggregation approaches.

The weighted PCA implements the above methods together, which will help to support dynamic datasets.

C. Multi Class Clustering (MCC)

The major contribution of the work is the multi class clustering which is the hybrid technique helps to aggregate the dataset in horizontally and as well as vertically [27].

The major aim of the weighted PCA in data aggregation is to reduce the number of variables of interest into a smaller set of components. The weighted PCA analyzes the variance in the attributes and reorganizes it into a new set of components equal to the number of original variables.

The major theme of implementing the weighted PCA in data aggregation is because of two reasons, the first one is the components are independent, that can decrease the time of aggregation if the data is huge. Dimension reduction is the main and strong reason of using weighted PCA for the analysis.

The multi class aggregation query will produce a wide table (t) with n columns, with one group for each unique combination of values V1,…….,Vn and one aggregated value per group (sum(A) in this case). In order to evaluate this query the query optimizer takes three input parameters:

1) The input table t,
2) The list of grouping columns V1,……., Vn,
3) The column to aggregate (MA).

The basic goal of a horizontal aggregation is to reorder (pivot) the aggregated column (MA) by a column subset of V1,……..,Vn

MCC: A Hybrid Technique

The MCC algorithm adopts a strategy consisting in selecting the relevant aggregation techniques of the overall set of conditions.

Step 1: Read dataset from high dimensional table

Read the columns and values from the transaction $T_{N.}$

Step 2: the list of GROUP BY columns L1,…..,Lj,

Step 3: the column to aggregate (M_A),

Set C_a as conditions -Identify base conditions for every attribute or properties

Step 4: the list of transposing columns R1, R2…Rn.

a. Single clustered data set $S_{c.}$
b. If the attribute is already found in the cluster-find the aggregation values.
c. Else if new aggregation process will perform
d. Find next dimensionality
e. Find in next cluster

Step 5: get the last aggregation values.

Step 6: detect the next transaction from the dynamic database

Step 7: perform the summarized aggregation without scanning the whole database T.

Step 8: Return the result F

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

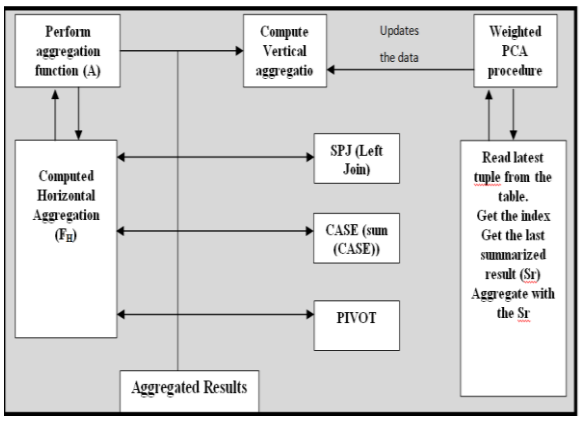

Fig 1 Process of the proposed aggregation using MCC

SYNTAX

This portion of implementation performs the small syntax extension to the SELECT statement, which permit understanding the implementation in a new perspective.
In short this extends the standard SQL aggregate functions with a "transposing" BY clause followed by a list of columns (i.e., R1,…….,Rk), to produce a horizontal set of numbers instead of one number. Our proposed syntax is as follows:

SELECT C1, C2…Cn, sum (C1BY C1+1….Cn)
FROM Table_name
GROUP BY C1 . . . Cn

It represents the subgroup columns C1,……,Cn which is to be a parameter associated to the aggregation itself.

In the context of our work, sum() can be additionally represented as some other SQL aggregation such as

- count()
- min()
- max()
- avg()

The aggregation function must have at least one argument represented by A, followed by a list of columns. The result rows are determined by column C,……Cn in the GROUP BY clause if present.
This permit processing aggregations based on any subset of columns not used in the GROUP BY clause. A horizontal aggregation groups rows and aggregates column values (or expressions) like a vertical aggregation, but returns a set of values (multi- value) for each group.

IV. RESULT AND DISCUSSION

Fig 2 shows the Efficiency comparison based on the Aggregation and our proposed approach MCC (Multi Class Clustering) gives better efficiency than Bayesian classifier.

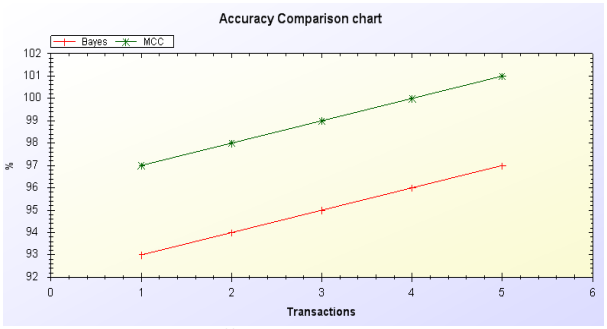

Fig 2 Efficiency comparison chart

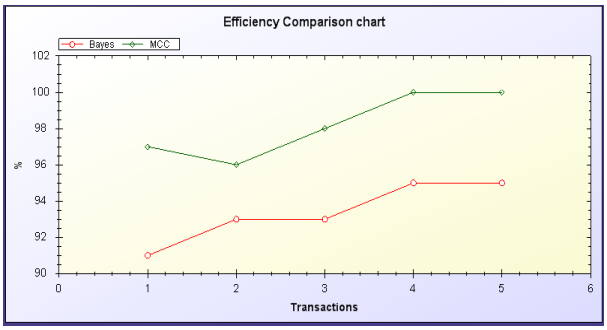

Fig 3 Time comparison chart

Fig. 3 shows the time comparison based on the Aggregation time and our proposed approach MCC (Multi Class Clustering) took less time while compare to Bayesian classifier

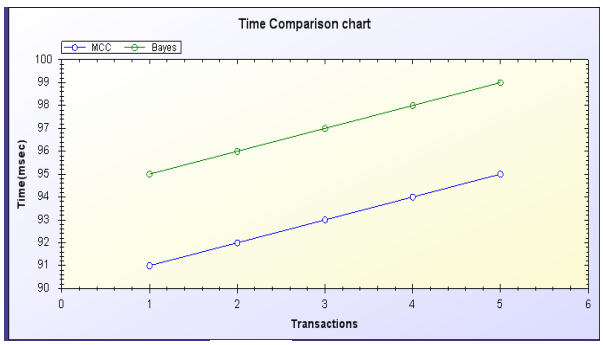

Fig 4 Accuracy comparison chart

Fig 4 shows the Accuracy comparison based on the Aggregation and our proposed approach MCC (Multi Class Clustering) gives more Accuracy than Bayesian classifier.

V. CONCLUSION

The work presented and introduced a new multi class of aggregate functions which is called horizontal aggregations with the innovation of MCC. Horizontal aggregations are useful to build data sets in tabular form when it is huge. A horizontal aggregation returns a set of numbers instead of a single number for each group. This work proposed a simple extension to SQL standard aggregate functions to compute horizontal aggregations

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

that only requires specifying sub grouping columns in a dynamic environment. The system additionally performs the Weighted PCA method in order to reduce the dimensionality in aggregation. This WPCA slightly reduce the time delay when the aggregations take place. The WPCA maintains the summarized results for further aggregation. The experiments and evaluation shows the proposed Multi class aggregation scheme yields best result in the portion of dynamic environment.

ACKNOWLEDGMENT

We thank the Karpagam University for the motivation and encouragement for giving the opportunity to do this research work as successful one.

REFERENCES

[1] C. Ordonez, "Integrating K-Means Clustering with a Relational DBMS Using SQL," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 2, pp. 188-201, Feb. 2006

[2] S. Sarawagi, S. Thomas, and R. Agrawal, "Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '98), pp. 343-354, 1998

[3] H. Wang, C. Zaniolo, and C.R. Luo, "ATLAS: A Small But Complete SQL Extension for Data Mining and Data Streams," Proc. 29th Int'l Conf. Very Large Data Bases (VLDB '03), pp. 1113- 1116, 2003.

[4] Witkowski, S. Bellamkonda, T. Bozkaya, G. Dorman, N. Folkert, A. Gupta, L. Sheng, and S. Subramanian, "Spreadsheets in RDBMS for OLAP," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '03), pp. 52-63, 2003

[5] H. Garcia-Molina, J.D. Ullman, and J. Widom, Database Systems: The Complete Book, first ed. Prentice Hall, 2001

[6] C. Galindo-Legaria and A. Rosenthal, "Outer Join Simplification and Reordering for Query Optimization," ACM Trans. Database Systems, vol. 22, no. 1, pp. 43-73, 1997

[7] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh, "Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross- Tab and Sub-Total," Proc. Int'l Conf. Data Eng., pp. 152-159, 1996

[8] J. Clear, D. Dunn, B. Harvey, M.L. Heytens, and P. Lohman, "Non- Stop SQL/MX Primitives for Knowledge Discovery," Proc. ACM SIGKDD Fifth Int'l Conf. Knowledge Discovery and Data Mining (KDD '99), pp. 425-429, 1999

[9] Carlos Ordonez and Zhibo Chen, "Horizontal Aggregations in SQL toPrepare Data Sets for Data Mining Analysis", IEEE TRANSACTIONSON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 4, APRIL 2012.

[10] M.Madhavi andS. Kavitha," Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis",Mdhavi et al. / IJEA, Vol. 1 Issue 6 ISSN: 2320-0804, PP1-7, 2013

[11] Rajesh Reddy Muley, Sravani Achanta and Prof.S.V.Achutha Rao, "Query Optimization Approach in SQL to prepare Data Sets for Data Mining Analysis",International Journal of Computer Trends and Technology (IJCTT) –volume4Issue8pp 1-5,August 2013.

[12] G. Luo, J.F. Naughton, C.J. Ellmann, and M. Watzke, "Locking Protocols for Materialized Aggregation Join Views," IEEE Trans. Knowledge and Data Eng., vol. 17, no.6, pp. 796-807, June 2005

[13] C. Ordonez. "Integrating K-means clustering with a relational DBMS using SQL," IEEE Transactions on Knowledge and Data Engineering (TKDE), 18(2):188–201, 2006

[14] Susrutha, B., Nath, J. V., Manohar, T. B., & Shalini, I. (2013) .Hortizontal Aggregation in SQL for Data Mining Analysis to Prepare Data Sets. International Journal of Modern Engineering Research. 3(4), pp-1861-1871.

[15] Jagannadh, D., Gayathri, T., & Nagendranadh, M. V. S. S. Horizontal Aggregations for Mining Relational Databases. International Journal of Computer Science and Information Technologies, 3 (2), pp 2012,3483-3487

[16] Manasa, B. K., & Reddy, H. V. Implementing an Efficient Task to Build Data Sets for Datamining Analysis. International Journal of Computer Science and Information Technologies, Vol. 5 (5), 2014,pp 6198-6201

[17] Saravanan, R., Sivapriya, J., & Shahidha, M. Horizontal Aggregations in SQL to Prepare Data Sets Using PIVOT Operator. International Journal of Emerging Technology and Advanced Engineering, 3(12), 2013, pp 477-481.

[18] Saravanan, M., & Mythili, P. A Novel Approach of Horizontal Aggregation in SQL for Data Mining. International Journal of Engineering Trends and Technology (IJETT) – Volume 9 Number 1 - Mar 2014, pp 45-47

[19] Tassa, T. (2014). Secure mining of association rules in horizontally distributed databases. IEEE Transactions on Knowledge and Data Engineering, 26(4), 970-983.

[20] Ordonez, C. (2004, June). Horizontal aggregations for building tabular data sets. In Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery (pp. 35-42). ACM.

[21] Samad, M. A., Rahman, M. R., Zahed, S., & Fattah, M. A. (2013). Creation of Datasets for Data Mining Analysis by Using Horizontal Aggregation in SQL.International Journal of Computer Applications in Engineering Sciences, 3(1), 46.

[22] George, B., & Balaram, A. Dataset Preparation and Indexing for Data Mining Analysis Using Horizontal Aggregations. International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 5, May 2014

[23] Jasti, S., & Vasumathi, D. CREATING MINIMIZED DATA SETS BY USING HORIZONTAL AGGREGATIONS IN SQL FOR DATA MINING ANALYSIS. International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 32-37 (2013)

[24] Dhawale, K. R., & Hiremani, V. A. Fundamental methods to evaluate horizontal aggregation in SQL. International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 10, October 2013, pp 1900-1905

[25] Ordonez, C., & Chen, Z. (2012). Horizontal aggregations in SQL to prepare data sets for data mining analysis. IEEE transactions on knowledge and data engineering, 24(4), 678-691.

[26] Chen, H. (2002). Principal component analysis with missing data and outliers.Electrical and Computer Engineering Department Rutgers University.Sawale, Gaurav J., and S. R. Gupta. "Horizontal Aggregations Based Data Sets for Data Mining Analysis: A Review."

[27] Chaudhari, A. A., & Khanuja, H. K. (2014). Extended SQL Aggregation for Database Transformation. International Journal of Computer Trends and Technology (IJCTT), 18(6), 272-275.

[28] Sumathi, K., Kannan, S., & Nagarajan, K. (2013). A Search Space Reduction Algorithm for Mining Maximal Frequent Itemset. International Journal of Computer Applications, 82(9).

[29] Sumathi, K., Kannan, S., & Nagarajan, K. (2013). Discovering Maximal Frequent Itemset using Association Array and Depth First Search Procedure with Effective Pruning Mechanisms. International Journal of Computer Applications, 76(13).

[30] Sumathi, K., Kannan, S., & Nagarajan, K. (2012). A New MFI Mining Algorithm with effective Pruning Mechanisms. International Journal of Computer Applications, 41(6).

BIOGRAPHY

Dr. K. Sathesh Kumar completed M.C.A., Ph.D. He is presently working as an Assistant Professor in the Department of Computer Science & Information Technology, Kalasalingam

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

University, Krishnankoil, India He has Six years of experience in teaching and research level and also he has published many research Papers in both International and National level Journals and Conferences. His research areas include Data Mining, Image Processing, Computer Networks, Cloud Computing, Software Engineering and Neural Network. He is an editorial board member of the several journals, in India. He has received gold and silver medals in National and International exhibitions for his research products.

P. Sabiya completed B.Sc (IT). She presently perceives the degree of M.Sc (CT) in the Department of Computer Science & Information Technology, Kalasalingam University, Krishnankoil. Her area of interest includes Data Mining.

S. Deepika completed BCA. She presently perceives the degree of M. Sc (CT) in the Department of Computer Science & Information Technology, Kalasalingam University, Krishnankoil. She has published papers in both International and National level conferences. Her area of interest includes Networking and Image Processing.