Volume 4, Issue 4, April 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Frequent Pattern Analysis in Horizontal Layout Using Apriori Algorithm

M.Karunya
PG Scholar, Dept. of CSE ,
Sona College of Technology,
Salem – 636005, TN, India

R.Reena Devi
Asst. Professor, Dept. of CSE,
Sona College of Technology,
Salem-636005,TN. India

Abstract— Data presentation is important phase in analysis. So data is collected and represented in the form of vertical or horizontal dataset. To prepare a dataset which is horizontally aggregated from various databases three major techniques were used such as SPJ, PIVOT and CASE, which make pre-processing easier through query tool. In this paper with horizontal aggregation, the analysis process is combined and pre-processed using apriori algorithm.

Keywords–Dataset preparation, Horizontal aggregation, Structured query language, Apriori algorithm

I. INTRODUCTION

Knowledge Discovery is the major task of data mining. The challenge is to model a suitable Data. Centralized database Management is achieved through data warehousing, mainly to improve efficiency while extracting and on data preparation. Dataset for analysis can be get by large queries, joins and aggregating several databases.

Task-Relevant data are summarized into smaller set of data and this summarization is achieved by aggregation [8].Preprocessing the raw data improves the quality of data preparation. Because more than half time spent on preparation phase. At once data is prepared, the mining is performed.

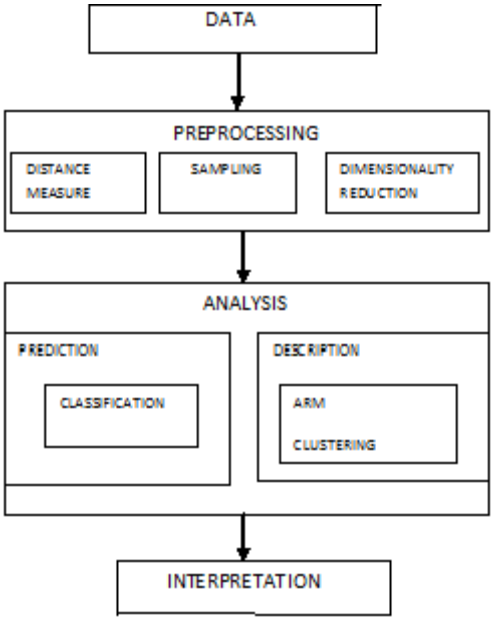

Fig. 1. Pre-processing and Analysis of data

Observation incompleteness, Erroneous in measures and inference information that is missing produce data uncertainty and managing uncertain data is an critical task [5].Statistical models which are multidimensional are computed outside the DBMS by exporting large data sets[5]. The above two considered problems can be solved by using User Defined Functions. Our paper makes use of user defined functions, which process and analyze the data by complex and domain specific algorithms [4].

UDF's are coded in any language and can be called by an "SELECT" statement. The UDF is advantageous because of following reasons; there is no need to internal DBMS code. They can be used as like any other SQL functions. The Code Exploits flexibility and speed of any language.UDF can reduce disk I/O, runtime in because it can work in main memory. But they cannot call another UDF. The UDF is mainly used to pre-process the dataset inside DBMS[4].

Preparation of dataset is an time consuming task, So functions which produce in horizontal layout are introduced. by automating the SQL query writing and extending their capabilities. These functions used aggregate functions and operators in SQL and produce a Data Set in summarized form. The three functions such as PIVOT, CASE AND SPJ are produced. Pivot operator exists in DBMS, CASE is an programming construct and SPJ depends on standard relational operators. Horizontal aggregations aggregate numeric expressions and transpose the resultant data set. It also reduces the manual work in data preparation phase [1].

II. RELATED WORK

Dataset is the vital property to discover knowledge from a database. Therefore considerable time and effort spends on the preparation of the dataset. One of the major requirements in data mining is summarization of data. Hence data is extracted or collected for the purpose of data mining requirements.

The obtained data is processed with three steps like data pre-processing, analysis and result interpretation. Real life or raw data must be pre-processed in order to use it in the analysis step. Hence raw data is susceptible to noise, inconsistent and missing. This affects the quality of data; so as to improve efficiency the data must be pre-processed before mining process [4].This pre-processing deals with the preparation and transformation of initial dataset. So pre-processing requires data cleaning, integration, Transformation and reduction.

Such pre-processing is done inside DBMS. The basic operations are introduced of DBMS in knowledge discovery without changing the syntactical model of the programming language. The basic requirement is a data mining model and operations performed in that model. Such operations must have ability to define, populate, predict and browse for reporting and visualizing applications [5].

So OLEDB is developed to represent data mining model in SQL. The statements such as create, delete, insert into and select can be done in this models. This method of mining is advantageous, because it avoids excess data movement, extraction and copying which results a hike in performance [6].

Traditionally the mining was done outside the DBMS, but this optimization is created by the deployment of data mining model and by eliminating the interface which process of SQL query writing makes easy.

SQL queries can be expressed as user defined functions and scans the database. Most of the pre-processing functions in the DBMS is by using the UDF. Stored procedure is slower than UDF. So it is possible for passing entire mining algorithm as SQL functions. Association rules are formed for every transactions, A->B where A and B are set of items. Association rule can be divided into two problems as finding frequent item sets and using the item sets to generate rules [9].

The main process is to integrate mining with RDBMS architecture, where the performance is based on loose coupling and stored procedure approach [9].The chosen ARM algorithm has multiple passes over data sets. The K passes over all item sets having K items is called the K item sets. They have multiple roles such as discovering patterns in databases, extracting knowledge from software engineering metrics, text mining etc.

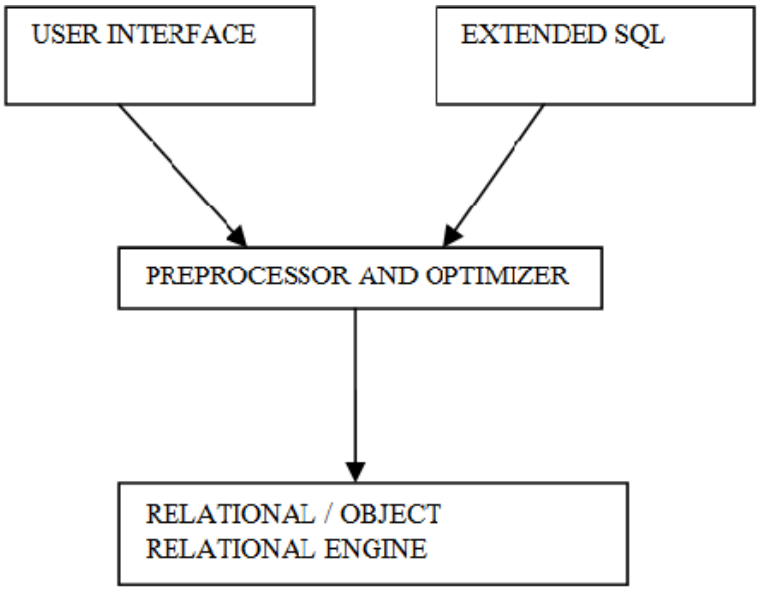

Fig. 2. SQL Architecture with Data Mining

By taking this architecture as an advantage, The preparation of data set is pre-processed using three functions, such as PIVOT, CASE and SPJ. These functions return aggregation results in cross-tabular layout. They combine the operations of transposition and aggregation together [1]. These functions group named as Horizontal aggregations, where the main objective of these horizontal aggregations is to reduce manual work by automating the functions by using a tool. Interpretation of standard aggregations is hard and they may result in various rows. Here UDF is used by defining a template for these functions inside the data mining tool.

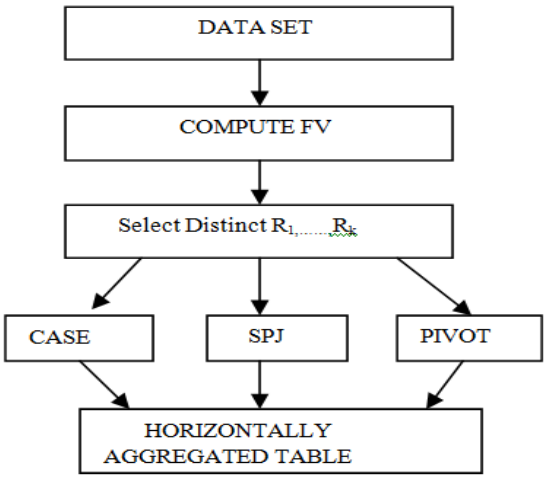

Fig. 3. Horizontal Aggregations

Unusual patterns of data are found by using data analysis. They categorize, extract, and contrast the category with one another. The data needed to be formulate, extract, visualize and then for analysing.

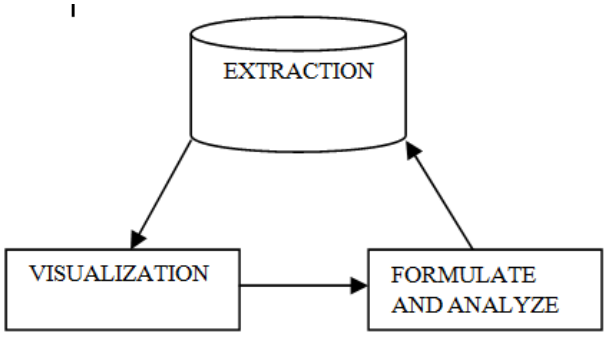

Fig. 4.Extract-Visualize-Analyze group

Some of the visualization tools display data in N-dimensional space [10].This paper introduces pre-processing the dataset preparation in a cross tabular form and the analysis process using an Association Rule Mining Algorithm.

III. PROPOSED WORK

The current work optimizes SQL extended for mining operation to run complex, long queries based on the mining semantics.

A. System model

A new class of aggregations known as Horizontal Aggregations is introduced to produce result in horizontal tables. A small SQL extension is needed for tool to process task relevant data. We are extending this scope by analysing the cross tabular data set produced by SQL language. The SQL is used for analysing process and Fig 1 explains the System model of proposed system.

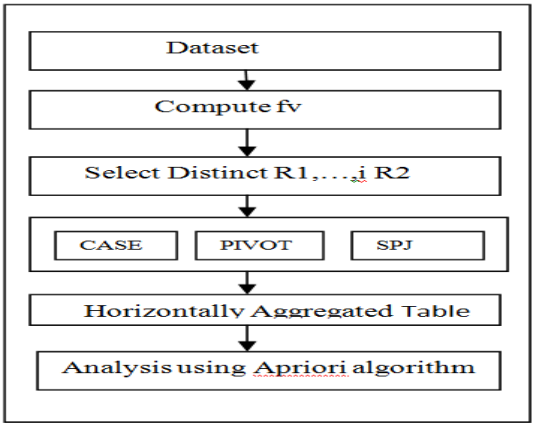

Fig. 5. Pre-processing analysis process

B. Template Generation

Generating templates from the given databases includes the method of expanding templates to generate new templates. To accomplish atomic templates, the templates with only one entity node are generated and then they are taken as the basis of template generation. Then the expansion rule is applied to generate new templates. This generation is called Query template model.

C. Fv Computation

Vertical layout is computed as F_V to form F_H, The vertical table computation is included for optimisation purposes. Instead of computing the result set directly from F, the Temporary table F_v is computed by grouping from F. Hence F_v is the compressed version of F.

D. Analysis of dataset

Apriori algorithm is chosen to pre-process the analysis process. The task is to identify all the frequent item sets from the set of candidate item sets.

To generate candidate item sets three basic approach is used,

- K-way Join
- Two way Group by
- Query sub Query

This paper implements Query sub Query approach for candidate Item set Generation.

E. Query Sub Query

Generating intermediate sub-Queries for support counting Sub Queries are denoted as Q_i. Sub Query Q_i is used to select items from Sub Query Q_{i-1} and relations C_i and input table F [7].

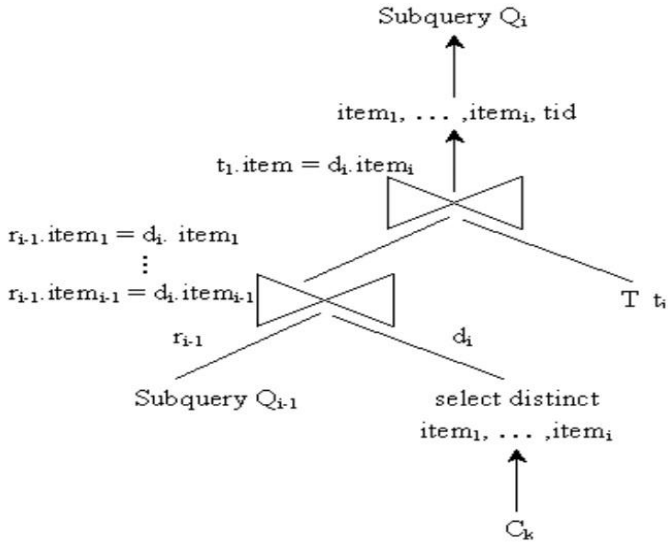

Fig. 6.Candidate generation

D. Frequent item set Generation

Stored procedures and UDF's are used to generate frequent item set. SQL-OR construct are used for better representation of data. This is UDF architecture based and UDF's are used to allocate memory for candidate item sets. With that for each pass of the mining algorithm, a collection of UDF is used for generating candidate item set. The data table is scanned sequentially for support counting and UDF is provided for each tuple to update the count in memory.

IV. CONCLUSION

We extended the horizontal Aggregation by pre-processing the analysis process. The main advantage of this process is automating analysis and dataset preparation using a tool. This reduces manual work while preparing a data set. Precomputing cube will accelerate on all the methods of horizontal Aggregations. Optimizing workload is a problem. High dimensionality produced by horizontal Aggregations is another challenging problem.

REFERENCES

[1] Carlos Ordenez,Zhibo chen "Horizontal Aggregations to prepare data sets for Data mining Analysis " , In Transactions and Knowledge Discovery ,IEEE 2011.

[2] Amir Netz,Surajit Chaudhri Usama Fayyad,Jeff Bernhardt,"Intergrating Data Mining with Databases:OLEDB for Data Mining" Microsoft Cooperation

[3] Goetz Graefe, Surajit Chaudhri Usama Fayyad,"On the Efficient Gathering of Sufficient Statistics for Classfication from Large SQL Databases",In Copyright 1998,Microsoft Cooperation.

[4] Carlos Ordenez,"Building Statistical Models and Scoring with UDFs",In SIGMOD Conference,ACM 2007

[5] Thanh T.L.Tran,Yanlei Diao,Charles Sutton,Anna Liu,"Supporting User-Defined Functions on Uncertain Data",In International Conference on very Large Databases,VLDB endowment Vol 6 No 6.

[6] Amir Netz,Surajit Chaudhuri,Jeff Bernhardt,Usama Fayyad,"Integration of Data Mining and relational Databases",Proceedings of the 26 th International Conference on Very Large Databases,Egypt 2000.

[7] Pratyush mishra,"Performance evaluation and analysis of Sql based approaches for Association rule mining",Dec 2002.

[8] Jiawei Han and Micheline Kamber *"Data Mining: Concepts and Techniques", Second Edition*

[9] Sunita Sarawagi Shiby Thomas * Rakesh Agrawal," Integrating Association Rule Mining with Relational DatabaseSystems:Alternatives And implications",In SIGMOD '98 Seattle, WA, USA.

[10] Jim Gray,Adam Bosworth,Andrew Layman,Hamid Pirahesh,"Data cube:A relational Aggregation operator,Generalizing Group By,Cross tab,and Sub totals", Technical Report, 5 February 1995.

[11] Ronnie Alves and Orlando Belo," Integrating Pattern Growth Mining on SQL-Server RDBMS", University of Minho, Department of Informatics.

[12] Arthur.A.Shaw, N.P. Gopalan," Frequent Pattern Mining of Trajectory Coordinates using Apriori Algorithm",In International Journal of Computer Applications (0975 – 8887),Volume 22– No.9, May 2011.

