Improved Data Mining Analysis by Dataset creation using Horizontal Aggregation and B+ Tree

Avisha Wakode, Mrs. D. A. Chaudhari, DYPCOE - Akurdi, Savitribai Phule Pune University

***Abstract*—**Data Mining is one of the emerging field in Research and information retrieval. Data mining tools requires data in the form of data set. Data set preparation is one of the important task in data mining. Data set is collection of data which is stored in relational database where database schema are highly normal- ized. To analyze data efficiency, data mining systems are widely using datasets with columns in horizontal tabular layout. The two main components of sql code is join and aggregation Vertical aggregations have limitations to build data sets because they return one column for aggregated group using group functions. Preparing a data set for data mining analysis is generally the most tedious and time consuming task in a data mining project, which requires many complex SQL queries, joining tables and columns, and aggregating columns. A powerful methods to generate SQL code to return aggregated columns in a horizontal or cross tabular form, returning a set of numbers instead of one number per row is introduced. This new class of methods is called horizontal aggregations. Horizontal aggregations are evaluted using three functions : CASE, SPJ and PIVOT method.Data mining also deals with searching of information. This paper focuses on creation of B+ tree to reduce the time of information search so that efficiency of the system increases.

Keywords*—*Aggregation, PIVOT, SPJ, CASE, Dataset.

*****

I. INTRODUCTION

Data mining is the process of analyzing data from different dimensions,categorizing it and summarizing it into some useful information.Data mining technology automates the process of information search or information retrieval.In a relational database, a lots of effort is required to prepare a data set that can be used as input for a data mining or statistical algorithm. Most algorithms require data set as a input which is in horizontal form, with several records and one variable or dimension per column. There are different models like clustering, classi- fication, regression and PCA. There are different terms used to describe the data set. In data mining the term used to describe data set is point-dimension whereas Statistics literature and machine learning research uses observation variable and in- stance feature. Preparing the useful and appropriate data set for data mining, needs more time. There are two main components of SQL code are join and aggregation.The most well-known aggregation is the aggregating of a column over group of rows. There are many different aggregation functions and operators which includes sum(),count(),avg(),etc in SQL. But all these aggregations have limitations to prepare data sets for data mining purposes. With such drawback in mind, a new class of aggregate methods that aggregates numeric expressions and transpose data to produce a data set with a horizontal layout. Methods belonging to this new class of aggregate functions are called horizontal aggregations. Horizontal aggregations is an extended form of existing vertical aggregations, which return a set of values in a cross tabular form instead of a single value per row. Horizontal aggregations provide different unique features and advantages. Firstly, they provide a template to generate SQL code from a data mining tool. Secondly it minimizes manual work in a data mining project. Horizontal aggregations can be used by a data mining algorithm for data mining analysis. A new class of aggregate functions that can be used to prepare data sets in a horizontal layout or in cross tabular form extending SQL capabilities.There are different ways and methods of information retrieval.But how efficiently information is retrieved is an challenging task.

II. LITERATURE SURVEY

Aggregations plays an vital role in sql code.Aggregation is the grouping the values of multiple rows together to form a single value.There are two types of aggregation techniques which includes vertical aggregation and horizontal aggregation

A. Vertical aggregation

Existing sql aggregations are also called as vertical aggre- gation.Vertical aggregations return single value. The most com- mom vertical aggregation functions includes sum(), count(), avg(), min(), etc.The vertical aggregations are common for numerous programming such as relational algebra.The output that we get from existing aggregations cannot be directly used for data mining algorithm.Data mining algorithm requires data in the form of data set. To get data in the form of data set from the output of existing aggregations require joining tables and columns , aggregating columns and many complex queries.It means that vertical aggregations have limitations to prepare data set.

B. Horizontal aggregation

Horizontal aggregation returns set of values instead of single value.Horizontal aggregations returns the output in the form of horizontal layout or in summarized form.The output that we get from horizontal aggregation can be directly used for data mining.The limitation of vertical aggregation is overcome in horizontal

aggregation.Horizontal aggregation is evaluated using three methods which includes CASE, SPJ and PIVOT method.

C. B+ Tree

There are different ways and methods of information retrieval.But how efficiently information is retrieved is an challenging task. B+tree is a tree data structure that keeps data sorted and allows searches, sequential access, insertions, and deletions in logarithmic time. The B+tree is a generalization of a binary search tree in that a node can have more than two children The B+tree is optimized for systems that read and write large blocks of data.

- C. Ordonez in[2] proposed query evaluation strategy that data in horizontal form require less time. He explained two possible form that is Horizontal form and vertical form.
- In [10] P.Laxmikanth Reddy, C.Rama Krishna, proposed the generation of Horizontal Aggregation in SQL by Using K-Means Clustering algorithm. The system make use of single base table and different derived tables, operations are then perform on the data from several tables. PIVOT operator is use to perform different aggregate operations in this paper.
- In [5] Pradeep Kumar and Dr. R. V. Krishna explained how can sql statements can be used with different combination . These methods can be used to generate the datasets that can be use by many data mining algorithm.
- Durka. C and Kerana Hanirex.D in [4] ,proposed that how the pivot method can help to generate data in horizontal form.
- R.Saravanan, J.Sivapriya in [14] , proposed the use of PIVOT method that helps to generate the output in the form of cross tabular form.
- Jincy Annie V., J. A. M. Rexie in [11], focused on how one can by making the use of sql aggregation can generate data set so that it can be directly used for data mining algorithm.
- Ankita E. Shewale Dr. S. N. Deshmukh[22] gives information about horizontal aggregation using spj method and equivalence of Methods.
- Krupali R. Dhawale and Vani A. Hiremani[18] gives fundamental methods to evaluate horizontal aggregation in sql, where it gives comparison of three horizontal aggregation methods.

III. IMPLEMENTATION DETAILS

As data mining is emerging field in information retrieval.Data mining algorithm requires data in the form of data set. Data sets are stored in relational database which comes from OLTP systems where database schemas are highly normalized To create data in the form of data set relational database requires lots of efforts.Horizontal aggregations can be evaluated using three methods:CASE method,SPJ method and PIVOT method.

1) SPJ method relies on relational operators. In SPJ Method sub query is executed first and after that root query is executed. To implement SPJ method vertical operations can be used. For every column one table is generated and then, the tables generated are joined in order to obtain final output in the form of cross tabular layout Left Outer Join is use in SPJ method, the left outer join is performed in between two tables The both common and uncommon attributes are returned
2) CASE relies on the SQL CASE construct. This task is based on the CASE construction provided by SQL. It deals with many boolean expressions and out of that one of it is returned. Aggregation or Projection is like to this from relational query point of view. In sql CASE programming is done using CASE method.It can be done by using many conditions. In this firstly computations can be directly from input database table and then generated vertical aggregations are stored in some temporary table and this table is used to generate result in horizontal form.
3) PIVOT does the transposition of data.. It transposes the required number of rows into additional new column. Therefore, for evaluating horizontal aggrega- tions pivot operator is used to transfer the data from row into column in it.RDBMS has built in PIVOT operator. This is used for the PIVOT operation. It transposes the fewer of rows into additional new column. Therefore, for evaluating horizontal aggre- gations using pivot method to transfer the data from row into column.

A. System Overview

Architecture of the proposed system is shown in figure 1.

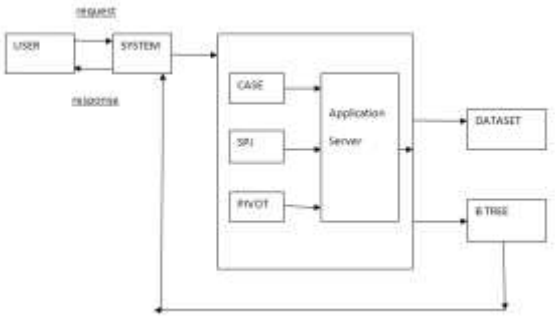

Figure 1. SYSTEM ARCHITECHTURE

Process summary:

1) Designing of database and GUI :- Database designing and creation is done. The GUI is designed which shows the data queries.Module 1 represents GUI design of the project, database design and also login application for user.

2) Estimating of Query :- Query is evaluated using three methods they are CASE, SPJ and PIVOT method.CASE relies on sql case construct. SPJ relies on standard relational operators and PIVOT does the transposition of data.
3) Data generation in the form of Data set :- Using methods of horizontal aggregation data set is created and the output of horizontal aggregation can be directly used for data mining algorithm.
4) Generation of B+ Tree :- To increase the efficiency of system B+ Tree is created on Horizontal aggregation so that information is retrieved in less time.

B. Mathematical Model

INPUT :-

1) Us = {Us1, Us2, Us3, ...} Where 'Us' be the Users.
2) DB = {D1, D2, D3, ...} where DB be the Database.
3) R = {rt1, rtt2, r3, ...} where 'R' be the relations.
4) At = {at1, at2, at3, , ...} where 'At' be the attributes.
5) Entered Queries
Q = {Q1, Q2, Q3, ...}
Where 'Q' be the queries.
6) SYS = {SP J,CASE, P IV OT }

OUTPUT :-

1) Create Dataset = {DS}
- Dataset = Data set are required for data min- ing algorithm.

2) Dataset = {DS}
- DS=SPJ+CASE+PIVOT

$$\prod \text{select}\sigma\text{where}(R1 \times .....R2) \quad (1)$$

IV. EXPECTED RESULTS

The use of horizontal aggregation to prepare a data set is faster technique than existing vertical aggregation.Horizontal aggregation prepares data in the form horizontal or cross tabular form which is used for many data mining algorithms.

Table I. COMPARING QUERY EVALUATION METHOD

n	d	Fv	SPJ	CASE	PIVOT
100K	8	121	87	82	82
	13		85	78	80
	26		106	88	92
1.5M	8	276	231	157	155
	13		418	140	141
	26		1086	161	150
100M	8	350	250	170	170
	13		425	160	160
	26		1090	181	170

Table 1 shows the time in milliseconds.Table 1 shows that keeping the data dimension constant and increase in size of databases the time taken by vertical and horizontal aggregation to give the output.

$$R1(r)\times R2(r) .\times \prod R2.a1.R2.a2\sigma R3.a1 = R2.a2 \quad (2)$$

$$(R1 .\times R2) \cup (R - \prod r1....rn(R1 .\times R2)) \quad (3)$$

$$\prod a1....an\sigma A(R) \quad (4)$$

$$\prod at1.rt1, at2.rt2, ...an.rtn \quad (5)$$

$$\sigma A = at1.r1 \ \wedge \ B = at2.r2.....Z = an.rn \quad (6)$$

$$r1 \times r2..... \times rn \quad (7)$$

C. Operating Environment

a) Software Requirement : Basic software specifications are:

Operating System : Windows XP/7/8
Technology : Dot Net
Front End : Dot net
Database : MySQL
Database Connectivity : MySql server 2008

b) Hardware Requirement : Basic hardware specifications are:

Processor : At Least Pentium Processor
Ram : 64 MB
Hard Disk : 2 GB

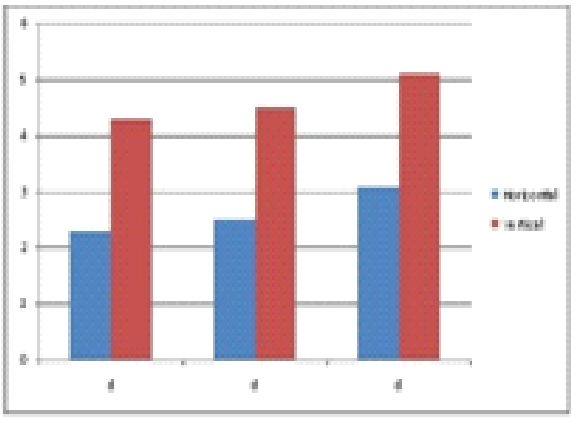

Figure 2. Comparision of Aggregation Technique

Fig 2 shows the comparision between vertical and horizon- tal aggregation n is the size of database. d is the

dimension of the database table.Fv is the table of vertical aggregation.

Figure 3. Data Set CASE method

Fig 3 shows the creation of data set using the CASE method. Fig 4 shows the creation of data set using the SPJ method. Fig 5 shows the creation of data set using the PIVOT method.

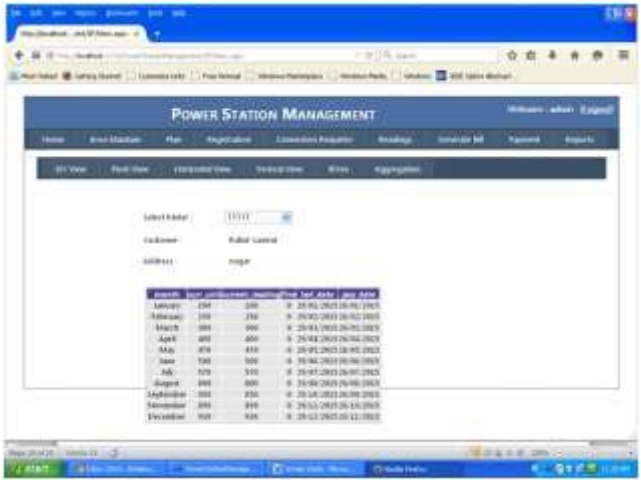

Figure 4. Data Set SPJ method

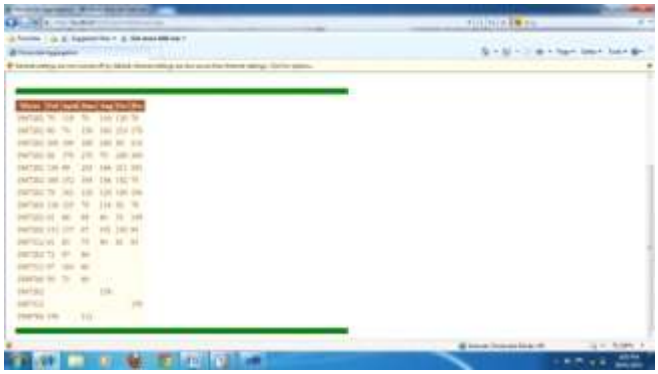

Figure 5. Data Set Pivot Method

V. CONCLUSION

A new class of aggregate methods, called horizontal aggregations which help preparing data sets for data mining analysis is implemented. Specifically, horizontal aggregations are useful to prepare data sets with a horizontal layout or cross tabular form, as commonly required by data mining algorithms and tools. From a query optimization perspective, horizontal aggregations can be evaluated using three methods.

- The first one (SPJ) relies on standard relational oper- ators.
- The second one (CASE) relies on the SQL CASE construct.
- The third (PIVOT) is used for transposition of data.

As data mining also deals with information search the creation of B+ tree makes the system more faster and efficient for data mining. Compare to complexity of Decision tree and B+ Tree, B+ Tree has complexity in logarithmic time so information retrieval is faster.

ACKNOWLEDGMENT

The authors would like to thank the publishers, researchers for making their resources available and teachers for their guidance. We would also thank the college authority for providing the required infrastructure and support. Finally we would like to extend a heart felt gratitude to friends and family members.

REFERENCES

[1] C. Ordonez and Zhibo Chen, “Horizontal Aggregation in SQL to Prepare Data Sets for Data Mining Analysis”, IEEE Trans. 2011.

[2] C. Ordonez, “Vertical and Horizontal Percentage Aggregations,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD’04), pp. 866-871,2004.

[3] B. Susrutha, J. Vamsi Nath, T. Bharath Manohar, I. Shalini, “Horizontal Aggregation in SQL for Data Mining Analysis to Prepare Data Sets”, International Journal of Modern Engineering Research (IJMER) Vol.3, Issue.4, Jul - Aug. 2013 pp-1861-1871 ISSN: 2249-6645.

[4] Durka.C and Kerana Hanirex.D, ” An Efficient Approach for Building Dataset in Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3,pp 1-5, March 2013.

[5] V. Pradeep Kumar, Dr. R. V. Krishnaiah,”Horizontal aggregations in SQL to prepare data sets for data mining analysis”, ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 5 (Nov. - Dec. 2012), PP 36-41

[6] S. Aiswarya, S. Ramadevi ”Multi dimensionalised aggregation in hor- izontal data set using analysis services”,ISSN 2250-2459 , An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013

[7] Y. Chakravarthi, P. Vindhya, ”A Better approach for horizontal aggrega- tions in SQL data sets for data mining analysis”, IJCSMC, Vol. 2, Issue.8, August 2013, pg.230 – 2369

[8] Subbarao Jasti, Dr.D.Vasumathi, ”Creating minimized data set 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)

[9] V. Prashanthi, K. Vinay Kumar, ”Masking the data in horizontal aggrega- tions in sql to prepare data set for data mining analysis”, The International Journal Of Engineering And Science (IJES) Volume 2 ,Issue 10 Pages32-38 ,2013 ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805.

[10] P.Laxmikanth Reddy, C.Rama Krishna, ”A well effective analy- sis of data mining aggregation oriented with horizontal sql”, IJR- RECS/September 2013/Volume-1/Issue-5/724-727.

[11] Jincy Annie V., J. A. M. Rexie, ”Efficient Tabular Dataset Preparations by the Aggregations in SQL”,International Journal of Computer Appli- cations (0975 – 8887) Volume 58– No.15, November 2012

Authors

Avisha Wakode received the B.E. degree in Information Technology from D.Y.Patil college of engg, Akurdi,Pune in 2007. During 2008-2010, she did lecturership in D.Y.Patil Polytechnic Akurdi,Pune. Now she is pursuing Master degree in Computer Engineering from D.Y.Patil Col- lege of engg Akurdi,Pune.

Deepali Chaudhari received the BE degree in Computer Science and Engineering from University of Pune in 2000 and ME in Computer Engineering from University of Pune in 2010 and has 8 years of teaching experience. She is currently working as Assistant professor at D. Y. Patil College of Engineering, Akurdi, Pune.