Preparing Data Sets For Data Mining Using CASE, PIVOT And SPJ

I.Lakshmi Kantha Reddy[#1], M.Samba sivudu[*2]

[#]*Student & Department of CSE, MRCET, JNTUH, Hyderabad, India*

[#]*Associate Professor & Department of CSE, MRCET, JNTUH, Hyderabad, India*

Abstract--Data mining plays an important role in real time applications for extracting business intelligence from business data and make expert decisions. Datasets are used in order to mine data for the purpose of discovering knowledge from data. However, preparing datasets manually is a tedious task. The reason behind it is that it involves aggregation of relations and other complex operations. Another important reason for the difficultly is the fact that SQL aggregations do not provide datasets. Instead they can give only single value results that are not suitable for data mining. Data in horizontal layout is required for data mining purposes. For this reason, in this paper we focus on the horizontal aggregations that can produce datasets. Towards it we build three constructs that can be used along with SQL queries to produce datasets automatically. The novel aggregations include SPJ, CASE and PIVOT constructs. We built a prototype for making experiments and the results revealed that the proposed aggregations are able to produce datasets required.

Index Terms– SQL, aggregations, horizontal aggregations

I. INTRODUCTION

RDBMS is widely used database management system for storing valuable business data of organizations. Generally regular data is stored in RDBMS. For data mining purposes, this data has to be transformed and moved to data warehouses. The data in warehouses is suitable for data mining. However, preparing data sets for data mining is very important for the successful mining of data. Data mining can't be done directly with regular databases. For preparing datasets SQL aggregate functions can be used but they produce single row output. Therefore it is not suitable for data mining purposes. What is expected for data mining purposes is the preparation of data with horizontal layout. The data in horizontal layout is suitable for data mining. Vertical aggregations such as SUM(), MAX(), MIN(), AVG() and COUNT() can't produce horizontal aggregations. Therefore they can't be used directly [1]. In case of statistical algorithms vertical aggregations are useful [2], [3]. Many data mining techniques like classification, clustering, regression, and so on need data in horizontal layout for data mining [4].

To overcome this problem in this paper we built new constructs like CASE, SPJ and PIVOT. These constructs are used to produce horizontal aggregations that give rise to data in horizontal layout. These constructs internally use SQL commands along with some business logic as required. Thus the generated datasets can be used for data mining in case of OLAP applications. We also built a prototype application that can demonstrate the usefulness of the constructs developed in this paper. They are able to produce the datasets that can be used in data mining. The underlying logic is built in stored procedures which are pre-compiled objects that work faster than normal queries. The proposed application has web based and user-friendly interface. The remainder of this paper is organized as follows. Section II reviews literature. Section III describes horizontal aggregations. Section IV presents experimental results while section V concludes the paper.

II. RELATED WORK

SQL has been around for many decades which are the de facto standard to interact with relational databases. In all kinds of applications of all platforms this is commonly used language. The commands of SQL are generally categorized into DCL, TCL, DML and DCL. The aggregation functions supported by SQL are widely used to generate summary of data. In the process they can also be used for producing outputs but they are not suitable for data mining purposes. The reason behind is that they produce single row output. The format expected by data miming applications is horizontal layout [5]. Data mining techniques like association rule mining [6] are used to mine data. However, they operate on datasets to produce patterns from the data [7]. In this paper we focus on producing new aggregate functions that make use of SQL aggregations internally to produce datasets that can be used in data mining. The new aggregations produced by us are CASE, SPJ, and

PIVOT. In [5] SQL queries which are used in clustering algorithms. In [8] spreadsheet like operations are proposed through SQL. They also provided optimizations for joins and other SQL commands. New class of aggregations such as PIVOT, SPJ, and CASE are produced in this paper which is based on the traditional relational algebra [9], [10]. Tree based plans are traditionally used in optimizing queries [11]. Lot of research went on aggregations. The literature has cross and cube tabulations as well [12]. Unpivoted relational tables are presented in [13]. Transformations can be made from normal aggregations to horizontal aggregations [14]. The operations like TRANSPOSE and Unpivot are similar. Number of operations is less in case of transpose when compared it with PIVOT. Inverse relationship exists between them. Using them vertical aggregations can be produced that are widely used in decision trees in case of data mining domain. Relational databases support both of the operations [15]. In [16] and [17] also horizontal aggregations are worked out but they have limitations. One of the limitations is that they produce data that can't be directly used for data mining purposes. In this paper we produced new operators such as CASE, SPJ and PIVOT.

III. HORIZONTAL AGGREGATIONS

These are the operations that make use of aggregate functions to generate data in horizontal layout. However, it can't be done using normal SQL aggregations. Towards it in this paper we built new constructs namely CASE, SPJ and PIVOT that can be used with SQL commands in order to produce output with horizontal layout.

K	D_1	D_2	A
1	3	X	9
2	2	Y	6
3	1	Y	10
4	1	Y	0
5	2	X	1
6	1	X	null
7	3	X	8
8	2	X	7

D_1	D_2	A
1	X	null
1	Y	10
2	X	8
2	Y	6
3	X	17

D_1	D_2X	D_2Y
1	null	10
2	8	6
3	17	null

Fig. 1 – Input table (a), traditional vertical aggregation (b), and horizontal aggregation (c)

As seen in fig. 1, sample data is given in input table. Vertical aggregation result is presented in (b). In fact the result generated by SUM function of SQL is presented in (b). Horizontal aggregation results are presented in (c).

Steps Used in All Methods

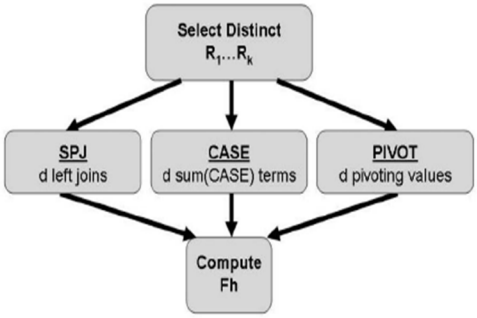

Fig. 2 shows steps on all methods based on input table

As seen in fig. 2, for all aggregations such as PIVOT, CASE and SPJ certain steps are carried out. However, the first step of all operations starts with SELECT query. Then based on the operation other activities are performed for computing horizontal aggregations.

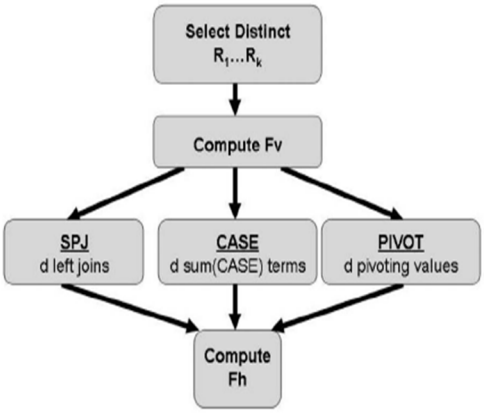

Fig. 3 shows steps on all methods based on table containing results of vertical aggregations

As seen in fig. 2, for all aggregations such as PIVOT, CASE and SPJ certain steps are carried out. However, the first step of all operations starts with SELECT query. Then based on the operation other activities are performed for computing horizontal aggregations.

SPJ Method

Vertical operations are used in SPJ method. For every column one table is generated in this model. Afterwards, the tables generated are joined in order to obtain final horizontal aggregations. The procedure followed is as given in [18].

PIVOT Method

RDBMS has built in PIVOT operation. This is used by the PIVOT operation we proposed in this paper. This construct can provide transpositions. Therefore for evaluating horizontal aggregations it can be used.

```
SELECT DISTINCT R1; . . .;Rk
FROM FV ;
INSERT INTO FH
SELECT L1,..,Lj
,sum(CASE WHEN R1 ¼ v11 and .. and Rk ¼ vk1
THEN A ELSE null END)
SELECT DISTINCT R1
FROM F; /* produces v1; . . . ; vd */
SELECT
L1; L2; . . . ; Lj
,v1; v2; . . . ; vd
INTO FH
FROM (
SELECT L1; L2; . . . ; Lj;R1;A
FROM F) Ft
PIVOT(
V ðAÞ FOR R1 in (v1; v2; . . . ; vd)
) AS P;
```

Listing 1 – Shows optimized instructions for PIVOT construct

As seen in listing 1, the queries have been optimized by choosing only the columns that are required by horizontal aggregations.

CASE Method

The CASE operation has many Boolean expressions to evaluate multiple conditions. It is also present in SQL. When all expressions are evaluated, the resultant value is returned. Many conditions with functions such as AND, OR are used internally in order to achieve it. Two strategies are used in this case. The first one is that computations are directly done on the given table while the second one makes vertical aggregations and the results are kept in an intermediary table. The procedure is as explored in [18].

IV. EXPERIMENTAL EVALUATION

We built a prototype web application that demonstrates the proof of concept. The environment used to build the application includes a PC with 4GB RAM, core 2 dual processor running Windows XP operating system. The web application is built using Java platform. Servlets and JSP technologies are used to provide web interface. JDBC is used to interact with relational databases. Java is the main programming language for building constructs. SPJ results are shown in figure 4.

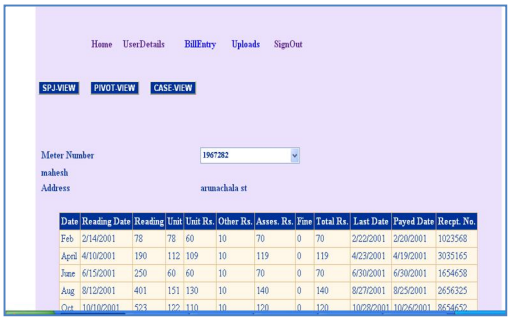

Fig. 4 – Results of Spj aggregation

As seen in fig. 4, SPJ operation's results are presented in horizontal layout. This kind of data can be used further for data mining operations.

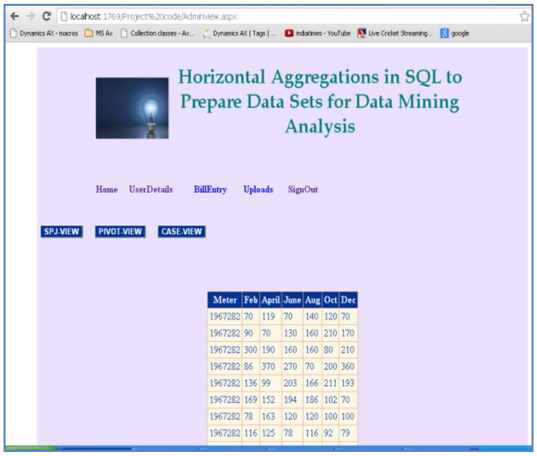

Fig. 5 – Result of Pivoting Aggregation

As seen in fig. 5, PIVOT operation's results are presented in horizontal layout. This kind of data can be used further for data mining operations.

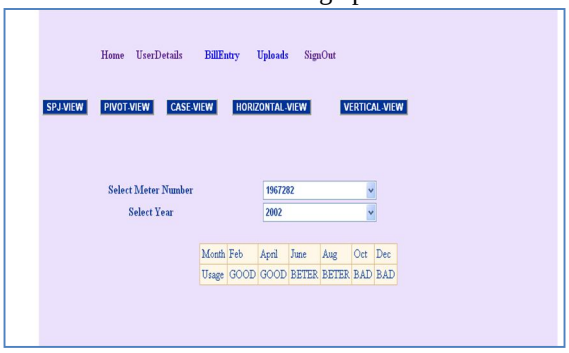

Fig. 6 – Result of CASE Aggregation

As seen in fig. 6, CASE operation's results are presented in horizontal layout. This kind of data can be used further for data mining operations.

V. CONCLUSIONS

In this paper we studied the preparation of datasets for data mining purposes. As vertical aggregations

provided by SQL such as SUM, MAX, MIN, AVG, and COUNT are unable to provide results in horizontal layout, in this paper, we built horizontal aggregations such as PIVOT, CASE and SPJ that can produced desired results that can be used for data mining operations. Our new aggregations are the programming constructs that make use of SQL aggregations internally along with certain business logic. Moreover our constructs are pre-compiled objects that work faster than normal aggregations. We built a prototype application that demonstrates the proof of concept. The empirical results revealed that the proposed aggregations are very useful in producing datasets with horizontal layout that can be used in data mining operations.

References

[1] C. Ordonez, "Data Set Preprocessing and Transformation in aDatabase System," Intelligent Data Analysis, vol. 15, no. 4, pp. 613-631, 2011.

[2] C. Ordonez and S. Pitchaimalai, "Bayesian Classifiers Programmedin SQL," IEEE Trans. Knowledge and Data Eng., vol. 22,no. 1, pp. 139-144, Jan. 2010.

[3] C. Ordonez, "Statistical Model Computation with UDFs," IEEETrans. Knowledge and Data Eng., vol. 22, no. 12, pp. 1752-1765, Dec.2010.

[4] J. Han and M. Kamber, Data Mining: Concepts and Techniques, firsted. Morgan Kaufmann, 2001.

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

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

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

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

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

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

[11] G. Bhargava, P. Goel, and B.R. Iyer, "Hypergraph BasedReorderings of Outer Join Queries with Complex Predicates,"Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '95),pp. 304-315, 1995.

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

[13] G. Graefe, U. Fayyad, and S. Chaudhuri, "On the EfficientGathering of Sufficient Statistics for Classification from LargeSQL Databases," Proc. ACM Conf. Knowledge Discovery and DataMining (KDD '98), pp. 204-208, 1998.

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

[15] C. Cunningham, G. Graefe, and C.A. Galindo-Legaria, "PIVOTand UNPIVOT: Optimization and Execution Strategies in anRDBMS," Proc. 13th Int'l Conf. Very Large Data Bases (VLDB '04),pp. 998-1009, 2004.

[16] C. Ordonez, "Horizontal Aggregations for Building Tabular DataSets," Proc. Ninth ACM SIGMOD Workshop Data Mining andKnowledge Discovery (DMKD '04), pp. 35-42, 2004.

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

[18] Carlos Ordonez and Zhibo Chen, "Horizontal Aggregations in SQL to PrepareData Sets for Data Mining Analysis", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 4, APRIL 2012.

Authors

Lakshmi kantha Reddy.I, he is pursuing M.Tech (CSE) in MRCET, Hyderabad, AP, INDIA. He has received M.C.A. His main research interest includes data mining, Databases and DWH.

Samba Sivudu.M He is currently with the Department of Computer Science and Engineering, MRCET, Andhra Pradesh,India. He has received M.Tech Degree from JNTU. His main research interest includes data mining, Databases and DWH.