International Journal of Engineering Research
Volume No.3, Issue No.7, PP : 446-448
ISSN:2319-6890)(online),2347-5013(print)
01 July 2014

Preparation of Datasets for Data Mining Analysis Using Horizontal Aggregation

Vidya M.Bodhe, Prof.Jyoti Mankar
K.K.W.I.E.E.R, Nashik, University of Pune, Pune, Maharashtra India
vidya.jambhulkar@gmail.com, aim_jyoti@yahoo.co.in

Abstract- Data aggregation is a process in which information is gathered and expressed in a summary form, and which is used for purposes such as statistical analysis. A common aggregation purpose is to get more information about particular groups based on specific variables such as age, name, phone number, address, profession, or income. Most data mining algorithms takes as input data set with a horizontal layout. Significant effort is required to prepare summary data set in a relational database with normalized tables. For preparing data sets suitable for data mining analysis, we have to write complex SQL queries, operation of joining tables and column aggregation. Horizontal aggregation can be performing by using operator, it can easily be implemented inside a query processor, much like a select, project and join. PIVOT operator on tabular data that exchange rows, enable data transformations useful in data modeling, data analysis, and data presentation. Two main ingredients in SQL code are joins and aggregations Standard aggregation returns one column per aggregated group and produce table with a vertical layout and Standard aggregations are hard to interpret when grouping attributes have high cardinalities. All these are limitations of standard aggregation. Because of these limitations, standard aggregation is not much suitable for preparation of data set for data mining analysis. Horizontal aggregation is a simple method which generates SQL code to return aggregated columns in a horizontal tabular layout and returns set of numbers instead of one number per row. This project is useful for building a suitable dataset for data mining analysis using horizontal aggregations in SQL. Two fundamental methods are used to evaluate horizontal aggregations: SPJ and Left Outer Join. This project will evaluate horizontal aggregation using Left outer join method.

Index Terms – **Data Mining, Data set, Horizontal aggregation, Left outer join, SPJ, Standard aggregation**

Introduction

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is widely used domain for extracting trends or patterns from historical data. However, the databases used by enterprises can't be directly used for data mining. RDBMS has become a standard for storing and retrieving large amount of data. This data is permanently stored and retrieved through front end applications. The applications can use SQL to interact with relational databases. Preparing databases needs identification of relevant data and then normalizing the tables.Generally, data sets that are stored in a relational database or a data warehouse come from On-Line Transaction Processing (OLTP) systems in which database schemas are highly normalized. But data mining, statistical or machine learning algorithms generally require aggregated data in summarized form. Suitable data set building for data mining purposes is a time- consuming task. This task requires writing long SQL statements or customizing SQL Code if it is automatically generated by some tool. There are two main ingredients in such SQL code: joins and aggregations. Many aggregations functions and operators are exists in SQL. But all these aggregations have limitations to build data sets for data mining purposes. Standard aggregations are hard to interpret when there are many result rows, especially when grouping attributes have high cardinalities

2 EXISISTING SYSTEMS

Datasets are prepared for data mining analysis using standard aggregation functions. Most data mining algorithm requires input a datasets with a horizontal layout with several records and one variable or dimensions per column. But data set prepared using standard aggregation produce dataset in vertical tabular layout. And converting vertical data set into summarized form requires writing long SQL statements or customizing SQL code if it is generated by some tool. Significant effort is required for computing aggregations using available functions and clauses in SQL to convert data set into cross tabular form suitable for data mining analysis.

Let F be a table having a simple primary key K represented by an integer, p discrete attributes and one numeric attribute: F(K,D1, . . . , Dp, A). Using standard aggregation functions datasets are prepared from table F shown in table 1 and result is shown in table 2.

Table 1: Input table, F

K	D1	D2	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

ISSN:2319-6890)(online),2347-5013(print)
01 July 2014

Table 2: Vertical table, F_V

D1	D2	A
1	X	NULL
1	Y	10
2	X	8
2	Y	6
3	X	17

3. PROPOSED SYSTEM

The proposed method focuses on minimizing effort and time that is spent in preparing and cleaning a data set for data mining algorithms in data mining project. A big part of this effort involves deriving metrics and coding categorical attributes from the data set in question and storing them in a tabular (observation, record) form for analysis so that they can be used by a data mining algorithm. A new class of aggregations that have similar behaviour to SQL standard aggregations, but which produce tables with a horizontal layout as shown in table 3. Horizontal aggregations just require a small syntax extension to aggregate functions called in a SELECT statement. Alternatively, horizontal aggregations can be used to generate SQL code from a data mining tool to build data sets for data mining analsis. Proposed syntax is as follows.

SELECT (L1… Lj), H (A BY R1…Rk)
FROM F
GROUP BY (L1… Lj);

Dataset can be prepared using two methods SPJ and Left outer join as shown in fig.1

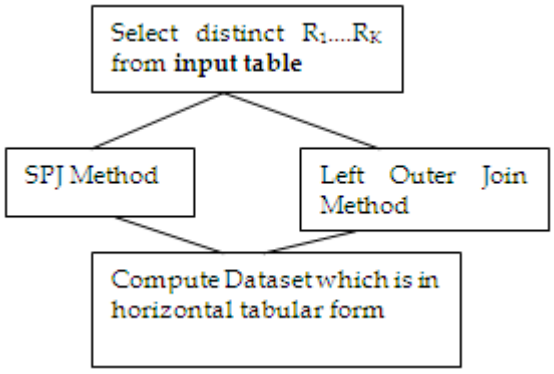

Figure 1: System architecture

Table 3: Horizontal table, F_H

D1	D2X	D2Y
1	Null	10
2	8	6
3	17	Null

4. EXPERIMENTS AND RESULTS

In order to compare the performance of the proposed system, the system is checked with dataset generated by TPC-H generator having input table lineitem with 700 records, |F|=700 and following parameters as shown in table 4.

Table 4: Summary of Grouping Columns from TPC-H TableLineitem (N=700).

L1(grouping column)	R1(transposing column)	n (answerset size)	d (no.of dimensions)
suppkey	linestatus	50	2
Suppkey	Weekday	50	7
Suppkey	Month	50	12
Suppkey	Brand	50	24
partkey	linestatus	100	2
Partkey	Weekaday	100	7
Partkey	Month	100	12
partkey	Brand	100	24
orderkey	linestatus	200	2
Orderkey	Weekaday	200	7
Orderkey	Month	200	12
orderkey	Brand	200	24

Table 5: Query Optimization (N =700) Times in Seconds

n	d	SPJ	LOJ
50	2	2294	1562
	7	3793	3635
	12	6084	5475
	24	11141	9796
100	2	2621	1920
	7	4463	3839
	12	6551	5039
	24	11105	9706
200	2	1950	1325
	7	5083	3447
	12	6582	5585
	24	11842	10794

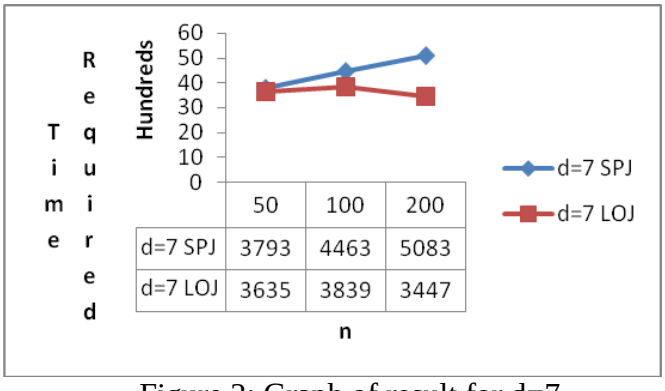

Figure 2: Graph of result for d=7

ISSN:2319-6890)(online),2347-5013(print)
01 July 2014

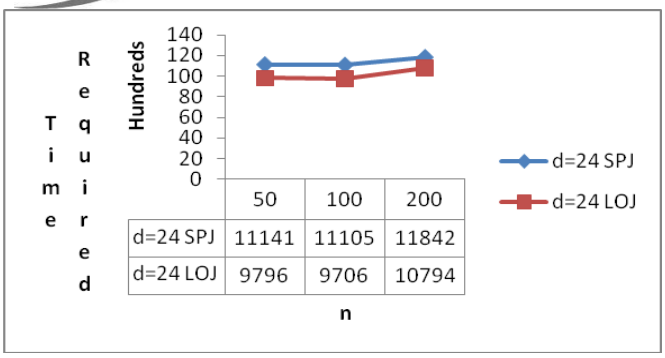

Figure 3: Graph of result for d=24

Figure 2 and 3 shows the graph for result for input table size 700 and answerset size 50,100 and 200 and for dimension d=7 and d=24.

5. Conclusion

We are introduced a new class of extended aggregate functions, called a horizontal aggregations which are help to preparing datasets for OLAP cube exploration and data mining. In particularly, horizontal aggregations are useful to create data sets with a horizontal layout. Mainly a horizontal aggregation returns a set of numbers instead of one number per each group. For a query optimization perspective, we are proposed two fundamental query evaluation methods. The first method is SPJ. It relies on standard relational operators. Second method is Left Outer Join. A SPJ method is important from a theoretical point of view because it is based on select, project and joins queries. Left Outer Join method takes less time for preparing dataset than SPJ method.

5 References

i. *C. Ordonez, "Data Set Preprocessing And Transformation In A Database System," Intelligent Data Analysis, Vol. 15, No. 4, Pp. 613-631, 2011.*

ii. *C. Ordonez, "Statistical Model Computation With Udfs," Ieee Trans. Knowledge And Data Eng., Vol. 22, No. 12, Pp. 1752 -1765,Dec. 2010.*

iii. *C. Ordonez And S. Pitchaimalai, "Bayesian Classifiers Programmed In Sql," Ieee Trans. Knowledge And Data Eng., Vol. 22, No. 1 ,Pp. 139-144, Jan. 2010.*

iv. *J. Han And M. Kamber, Data Mining: Concepts And Techniques, First Ed. Morgan Kaufmann, 2001.*

v. *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.*

vi. *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.*

vii. *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.*

viii. *A. 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.*

ix. *H. Garcia-Molina, J.D. Ullman, And J. Widom, Database Systems: The Complete Book, First Ed. Prentice Hall, 2001.*

x. *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.*

xi. *G. Bhargava, P. Goel, And B.R. Iyer, "Hypergraph Based Reordering Of Outer Join Queries With Complex Predicates," Proc. Acm Sigmod Int'l Conf. Management Of Data (Sigmod '95), Pp. 304-315, 1995.*

xii. *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.*

xiii. *G. Graefe, U. Fayyad, And S. Chaudhuri, "On The Efficient Gathering Of Sufficient Statistics For Classification From Large Sql Databases," Proc. Acm Conf. Knowledge Discovery And Data Mining (Kdd '98), Pp. 204-208, 1998.*

xiv. *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.*

xv. *C. Cunningham, G. Graefe, And C.A. Galindo-Legaria, "Pivot And Unpivot: Optimization And Execution Strategies In An Rdbms," Proc. 13th Int'l Conf. Very Large Data Bases (Vldb '04), Pp. 998-1009, 2004.*

xvi. *C. Ordonez, "Horizontal Aggregations For Building Tabular Data Sets," Proc. Ninth Acm Sigmod Workshop Data Mining And Knowledge Discovery (Dmkd '04), Pp. 35-42, 2004.*

xvii. *Carlos Ordonez And Zhibo Chen, "Horizontal Aggregations In Sql To Prepare Data Sets For Data Mining Analysis", Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 4, April 2012.*