ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Data mining dataset preparation using aggregation – a case study

S. Brintha Rajakumari*, C. Nalini
Dept. of CSE, Bharath University, Chennai.
***Corresponding author: E-Mail: brintha.ramesh@gmail.com**

ABSTRACT

Data mining is popular in research for the past two decades. It is mainly used to analyze the data and give best information from the huge collection of data. The SQL aggregate function used for dataset preparation which gives one column per group of values. We have built alternate method to prepare the data set for mining analysis which reduce the rows but increase the column of the table. The paper analyzed all the study related to the work has done by the combination perform which can used for dataset preparation and proposed new technique for the same.

KEY WORDS: Data Mining, dataset preparation, Horizontal Layout.

1. INTRODUCTION

Data mining is the process of finding useful pattern for the model from the large collection of data. It is generally used in the area of healthcare, fraud analysis, pattern recognition, information retrieval, image analysis and forensic data analysis. Preparing correct data for the data mining also prepare miner to prepare the correct model for the data. Data exploration leads to good decision making for any problem. There are three steps in mining such as preparing data, survey the data, and model the data. In the relational database management system, aggregations of data are used for data analysis.

In the section 2 describe the related analysis within the field of aggregation. In the section 3 presented the comparative analysis of varied analysis works with completely different parameters, section 4 shows the material and method related to data set preparation using SQL aggregate function followed by concluded the paper.

Comparative Study: The table 1 contains list of research works carried in the field of data set preparation using SQL aggregation concept.

Table.1.Comparative study

Title	Authors	Technique	Advantages	Problem	Tools/ Software
Data Set Preprocessing and Transformation in a Database System	C.Ordonez	K-mean Clustering	Overhead is low for large data sets	Overhead is significant for small data sets	SQL
Vertical and Horizontal Percentage Aggregations	Carlos Ordonez, Sasi K. Pitchaimalai	Bayesian Classifiers	Faster than existing OLAP aggregate functions	Two practical issues when computing vertical percentage queries were identified: missing rows and division by zero.	SQL
Bayesian Classifiers Programmed in SQL	Carlos Ordonez, Sasi K. Pitchaimalai	K-means clustering, Naïve Bayes Classifier	Bayesian classifier achieves high classification accuracy. The Bayesian classifier is more accurate than Naïve Bayes and decision trees	Data sets with missing information and subsets of points having significant overlap with each other.	SQL
Constraining and Summarizing Association Rules in Medical Data	Carlos Ordonez, Norberto Ezquerra, Cesar A. Santana	Association Rules	Constraints are shown to significantly reduce the number of discovered rules	Improve running time	C++
On the Efficient Gathering of Sufficient Statistics for Classification from Large SQL Databases	Goetz Graefe, Usama Fayyad, and Surajit Chaudhuri	Greedy Classification Algorithm	Results in significant increase in performance without requiring any changes to the physical layout of the data.	It does not require copies of the data table	SQL
Extended Aggregations for Databases with Referential Integrity Issues	Javier Garcia-Garcia, Carlos Ordonez	Referential Integrity Issues	Improved SQL aggregations to return enhanced answers sets in the presence of referential integrity errors.	Referenced table is assumed to be incomplete.	SQL
Horizontal Aggregations for Building Tabular Data Sets	C. Ordonez	Horizontal Aggregation	Horizontal aggregations tend to produce tables with fewer rows, but with more columns	Query optimization strategies used for vertical aggregations do not work well for horizontal aggregations.	SQL Based

2. MATERIAL AND METHODS

Let assume that there are eight elements in the table T, which are derived from the large number of tables. That means, T is a collection of transaction tables. The table T has one numeric attribute and many discrete attributes. In the table 2, value is the numeric attribute and the columns c1, c2 are discrete attributes.

Table.2.Sample Data

Sl. no.	C1	C2	Value
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

We have used traditional Structure query language aggregation commands such as sum(),avg(),Min(),max() and count() for getting the vertical layout of dataset is in table2. It has five row elements which are derived from the base table 2. And the last column is aggregated by sum() function. The first column will not be used for aggregation. This technique will reduce the row size of the table.

Table.3.Traditional Aggregation

C1	C2	Value
1	X	Null
1	Y	10
2	X	8
2	Y	6
3	X	17

In the paper, the author proposed new horizontal layout of tabular data in an efficient manner and compared with SQL CASE, SPJ and PIVOT method. The Result of the method is in table 4. It has three rows of data with two null entries. In the example table only two unique entries available in the C2 column. Suppose if the table contains more than one then it will increase the size of the column.

Table.4. Horizontal Aggregation

C1	C2X	C2Y
1	Null	10
2	8	6
3	17	Null

TOAC Algorithm: We have proposed new method to reduce the size of the table row which can be used for dataset preparation for mining data analysis. The following steps in the algorithm:

1. Combine all the tables.
2. Identify the aggregate column
3. Find the number of unique element in each column.
4. Find the minimum value attribute from the list.
5. Based on the minimum value column, transpose the element of the aggregated column.

When you apply the proposed method to the existing example, it gives only two rows and three columns of table. Finally, the above methods produce the 2 * 3 or 3* 2 unique element of the table. The resultant table 5 contains 2 * 3 elements from the original table.

Table.5.Result Table

C2	C1, 1	C1, 2	C1, 3
X	Null	8	17
Y	10	6	Null

3. CONCLUSION

We have analyzed about horizontal layout aggregation strategies with related analysis work with number of parameters which can be used for dataset preparation in data mining process. And we have proposed new method to find the solution for the same problem. Our method produces the same result with less number of rows.

REFERENCES

Achudhan M, Prem Jayakumar M, Mathematical modeling and control of an electrically-heated catalyst, International Journal of Applied Engineering Research, 9(23), 2014, 23013.

Brintha Rajakumari S, Nalini C, A Comparative Analysis Of Horizontal Layout Representation Of Data", International Journal of Advance Research In Science And Engineering, 4(1), 2015, 984-990.

Brintha Rajakumari S, Nalini C, An Efficient Compression Method to Reduce the Data Transfer Cost in the Cloud, International Journal of Innovative Research in Computer and Communication Engineering, 3(3), 2014, 2176-2179.

Brintha Rajakumari S, Nalini C, An efficient cost Model for data storage with horizontal layout in the cloud" , Indian Journal of Science and Technology, 7(3), 2014, 45-46.

Brintha Rajakumari S, Nalini C, An efficient Data Mining data Set preparation using aggregation in relational database, Indian Journal of Science and Technology,7(5), 2014, 44-46.

Brintha Rajakumari S, Nalini C, Preparation Of Dataset For Mining Analysis Using Cancer Dataset, Int J Pharm Bio Sci, 6(3), 2015, 883 – 889.

Carlos Ordonez and Zhibo Chen, Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis", IEEE Transactions On Knowledge And Data Engineering, 24(4), 2012.

Cunningham C, Graefe G, and Galindo-Legaria CA, PIVOT and UNPIVOT: Optimization and Execution Strategies in an RDBMS, Proc. 13th Int'l Conf. Very Large Data Bases (VLDB '04), 2004, 998-100.

Galindo-Legaria C and Rosenthal A, Outer Join Simplification and Reordering for Query Optimization, ACM Trans. Database Systems, 22(1), 1997, 43-73.

Gopalakrishnan K, Sundeep Aanand J, Udayakumar R, Electrical properties of doped azopolyester, Middle - East Journal of Scientific Research, 20(11), 2014, 1402-1412.

Gopinath S, Sundararaj M, Elangovan S, Rathakrishnan E, Mixing characteristics of elliptical and rectangular subsonic jets with swirling co-flow, International Journal of Turbo and Jet Engines, 32(1), 2015, 73-83.

Gray J, Bosworth A, Layman A, and Pirahesh H, Data cube: A relational aggregation operator generalizing group-by, cross-tab and subtotal. In *ICDE Conference*, pages 152–159, 1996.

Han J and Kamber M, Data Mining: Concepts and Techniques, first ed. Morgan Kaufmann, 2001.

Ilayaraja K, Ambica A, Spatial distribution of groundwater quality between injambakkam-thiruvanmyiur areas, south east coast of India, Nature Environment and Pollution Technology, 14(4), 2015, 771-776.

Kerana Hanirex D, Kaliyamurthie KP, Kumaravel A, Analysis of improved tdtr algorithm for mining frequent itemsets using dengue virus type 1 dataset: A combined approach, International Journal of Pharma and Bio Sciences, 6(2), 2015, 288-295.

Lingeswaran K, Prasad Karamcheti SS, Gopikrishnan M, Ramu G, Preparation and characterization of chemical bath deposited cds thin film for solar cell, Middle - East Journal of Scientific Research, 20(7), 2014, 812-814.

Ordonez C and Pitchaimalai S, Bayesian Classifiers Programmed in SQL, IEEE Trans. Knowledge and Data Eng., 22(1), 2010, 139-144.

Ordonez C, Data Set Preprocessing and Transformation in a Database System, Intelligent Data Analysis, 15(4), 2011, 613- 631.

Ordonez C, Horizontal Aggregations for Building Tabular Data Sets, Proc. Ninth ACM SIGMOD Workshop Data Mining and Knowledge Discovery (DMKD '04), 2004, 35-42, 2004.

Ordonez C, Integrating K-Means Clustering with a Relational DBMS Using SQL," IEEE Trans. Knowledge and Data Eng., 18(2), 2006, 188-201.

Ordonez C, Statistical Model Computation with UDFs," IEEE Trans. Knowledge and Data Eng., 22(12), 2010, 1752 -1765.

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

Premkumar S, Ramu G, Gunasekaran S, Baskar D, Solar industrial process heating associated with thermal energy storage for feed water heating, Middle - East Journal of Scientific Research, 20(11), 2014, 1686-1688.

Sundar Raj M, Saravanan T, Srinivasan V, Design of silicon-carbide based cascaded multilevel inverter, Middle - East Journal of Scientific Research, 20(12), 2014, 1785-1791.

Thooyamani KP, Khanaa V, Udayakumar R, Application of pattern recognition for farsi license plate recognition, Middle - East Journal of Scientific Research, 18(12), 2013, 1768-1774.

Thooyamani KP, Khanaa V, Udayakumar R, Efficiently measuring denial of service attacks using appropriate metrics, Middle - East Journal of Scientific Research, 20(12), 2014, 2464-2470.

Thooyamani KP, Khanaa V, Udayakumar R, Partial encryption and partial inference control based disclosure in effective cost cloud, Middle - East Journal of Scientific Research, 20(12), 2014, 2456-2459.

Thooyamani KP, Khanaa V, Udayakumar R, Using integrated circuits with low power multi bit flip-flops in different approch, Middle - East Journal of Scientific Research, 20(12), 2014, 2586-2593.

Thooyamani KP, Khanaa V, Udayakumar R, Virtual instrumentation based process of agriculture by automation, Middle - East Journal of Scientific Research, 20(12), 2014, 2604-2612.

Thooyamani KP, Khanaa V, Udayakumar R, Wide area wireless networks-IETF, Middle - East Journal of Scientific Research, 20(12), 2014, 2042-2046.

Udayakumar R, Kaliyamurthie KP, Khanaa, Thooyamani KP, Data mining a boon: Predictive system for university topper women in academia, World Applied Sciences Journal, 29(14), 2014, 86-90.