Int J Pharm Bio Sci 2015 July; 6(3): (B) 883 - 889

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

PREPARATION OF DATASET FOR MINING ANALYSIS USING CANCER DATASET

S.BRINTHA RAJAKUMARI[*1] AND DR.C.NALINI[2]

*[*1]Research Scholar, Department of CSE, Bharath University,Chennai, India*
[2]Professor, Department of CSE, Bharath University, Chennai, India.

ABSTRACT

Data preparation is very important for analysis. Incomplete data represent a challenge for achieving a successful data mining process. The data must be prepared and transformed to get the best mineable form. There are many techniques for data preparation that can be used to achieve different data-mining goals. This paper presents the detail description about the types of data, data sources, aggregate functions which are used to prepare the dataset using a new horizontal approach for data mining analysis and experiment with cancer dataset.

KEYWORDS: **Cancer dataset, Dataset Preparation, Data Mining, Horizontal Layout, SPJ.**

S.BRINTHA RAJAKUMARI
Research Scholar, Department of CSE, Bharath University,Chennai, India

***Corresponding author**

INTRODUCTION

Data mining is the discovery of models for data. A model can be one of several things. Modeling can be summarizing the data succinctly and approximately or extracting the most well-known features of the data and ignoring the rest. Building a proper dataset for data mining is a time consuming task. Different methods used for each research discipline to prepare data set for analysis. Aggregation concept is a powerful tool in database design, and consequently, preserving aggregation in database implementation is essential. The aggregation problem becomes especially acute in a database management system (DBMS) since such a system contains a large volume of data that could form aggregates that are more sensitive than their constituent parts. It is the intent of this paper to investigate the aggregation problem in the context of a database [1]. Aggregation is an important concept in database design where composite objects can be modeled during the design of database applications. Therefore, maintaining the aggregation concept in database implementation is essential[2]. Aggregation is a composition relationship, in which a composite object consists of other component objects[3]. This paper will describe a framework to integrate horizontal aggregation with data set preparation. The next four sections will describe each of these phases of the framework in detail, followed by results and conclusion sections. The paper[4] presented the horizontal representation of data used for dataset preparation in data mining analysis and evaluated with the cancer dataset. In this paper[5] introduced three SQL implementations of the popular K-means cluster rule to integrate it with a relative Database management system: an easy translation of K-means computations into SQL. The associate degree optimized version supported improved knowledge organization, economical categorization, adequate statistics and rewritten queries. An incremental version that uses the optimized version as a building block with fast convergence and automated reseeding. In the paper [6] have proposed two aggregate functions to compute percentages. The first function returns one row for each computed percentage and it was called a vertical percentage aggregation. The second function returns each set of percentages adding 100% on the same row in horizontal form and it was called a horizontal percentage aggregation. They are used as a framework to study percentage queries. Two practical issues when computing vertical percentage queries were identified: missing rows and division by zero. The Bayesian classifier could be a basic classification technique. In the paper[7], they targeted on programming Bayesian classifiers in SQL and introduced two classifiers: Naive Bayes and a classifier supported class decomposition victimization K-means clustering and regarded two complementary tasks: model computation and marking an information set. They analyzed the way to remodel equations into economical SQL queries and introduced many question optimizations. They compared the Naive Bayes implementations in SQL and C++: SQL is concerned fourfold slower. Bayesian classifier in SQL achieves high classification accuracy, will with efficiency, analyze massive knowledge sets and has linear quantifiability. Association rules area unit is an information mining technique accustomed discover frequent patterns in a very data set. In the paper [8] association rules area unit employed in the medical domain, wherever knowledge sets area unit usually high dimensional and tiny. For a good type of classification algorithms, scalability of massive databases is often achieved by perceptive that the majority algorithms area unit driven by a group of adequate statistics that area unit considerably smaller than the information within the paper [9]. By counting on a SQL backend to reason the adequate statistics, they leverage the question process system of SQL knowledge bases and avoid the requirement for moving data to the shopper. In the paper[10] introduced a brand new category of mixture functions, referred to as horizontal aggregations. Horizontal aggregations area unit helpful to make knowledge sets in tabular type[11,12]. A

horizontal aggregation returns a set of varieties rather than one number for every group. They projected a straightforward extension to SQL customary mixture functions to reason horizontal aggregations that solely needs specifying sub grouping columns[13, 14]. Data mining or knowledge discovery is the computer-oriented process of digging and analyzing large volumes of data and finally extracting the meaning of the data.

MATERIALS AND METHODS

TYPES OF DATA

The data preparation step covers all activities to construct the final dataset for modeling of the raw data. Tasks include database, table, record, and field selection as well as cleaning, aggregation and transformation of data. The data collected for data set preparation comes from various sources is in figure 1.

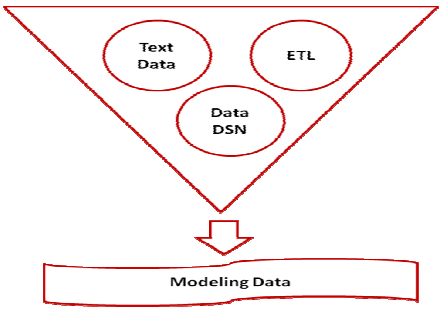

Figure 1
Flow of Data

Data preparation is a time consuming task in data mining. Machine leaning, statistical analysis algorithms assume that the input is given, then it predicts the results. Learning algorithms use the well formed table detail SQL to perform joining different tables, aggregate the operation and filter the details from the database. It is a time consuming process and need to write and maintain a large number of codes. A different type of data to be used for data set preparation is in figure 2.

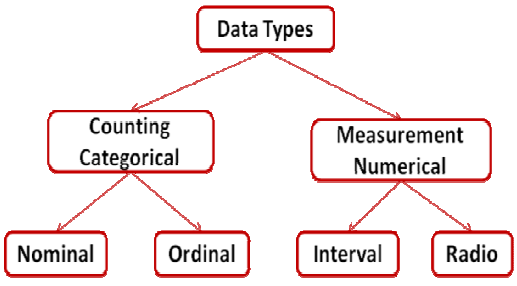

Figure 2
Types of Data

AGGREGATION AND DATA SOURCES

Aggregation concept is a powerful tool in database design, and consequently, preserving aggregation in database implementation is essential. Aggregation problem becomes particularly acute in a Database Management System, since such a system contains a large volume of data that could form aggregates that are more sensitive than their constituent parts. Traditional aggregate functions are listed is in the figure 3. The various data sources and their entities, methods and languages are presented in the table 1.

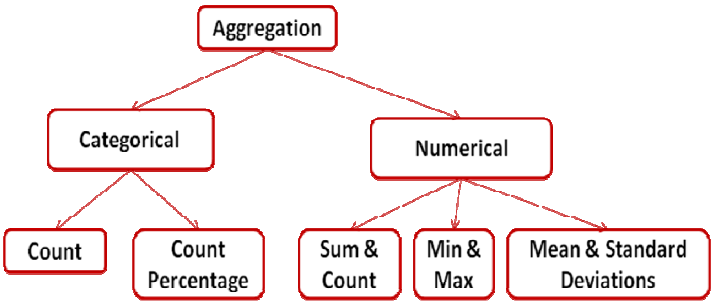

Figure 3
Aggregation Functions for Data Preparation

Table 1
Data Sources

	Text Files	Relational Database	Multi Dimensional Database
Attributes	File	Table	Cube
Entities	Row & Column	Field, Record & Index	Level, Measurement & Dimension
Methods	Read & Write	Insert, Delete , Select & Update	Drill Up & Drill Down
Language	-----	SQL	MDX

EXPERIMENTAL RESULT
SPJ REPRESENTATION AND EVALUATION

Table is a collection of records organized in rows and columns. The definition is in OLAP terms. Let F be a temporary table or view based on a star join query on several tables. In that k is a primary key of an integer attribute and C1, C2 are different dimensions of nominal and numerical attribute. An example showing the input table is on the table 2. And the traditional SQL aggregation sum () result is on the table 3. The horizontal aggregation of query result stored in the table 4. The main goal of horizontal aggregation is to transform the table 3 into table 4 representation. A method, SPJ method,

is used to evaluate a horizontal aggregation, which relies on relational operations. That is, select, project, join and aggregation queries.
In order to evaluate this query, the query optimizer takes three input parameters
The input table F,
The list of grouping columns L1;…. ;Lm ,
The column to aggregate (A).
The Select syntax is as follows
SELECT L1; … ; Lj
FROM F
GROUP BY L1; . . . ; Lj;
Experiment was performed this method using MS SQL Server 2008 and found out the size of the table. Tables 3 and 4 show horizontal layout representation reduced the rows than vertical representation. A data table 2 is presented containing 3 attributes, such as C1, C2 and Values.

Table 2
Example data table

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

Table 3
Vertical aggregation of table

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

Table 3 shows vertical aggregation of similar data and the SQL query is given as,

```
SELECT C1, C2, SUM (Value)
FROM Table 2
GROUP BY C1, C2 ORDER BY C1, C2
```

Table 4
Horizontal layout of data

	C2X	C2Y
1	Null	10
2	8	6
3	17	Null

Table 4 shows horizontal layout of similar data and the SQL query for this is given as,

```
SELECT C1,[X] AS C2X,[Y] AS C2Y FROM
(SELECT C1, C2,Value FROM Table2 AS source table
PIVOT(SUM(Value) FOR C2 IN([X],[Y])) AS pivot table;
```

The sample dataset collected from the http://www.theguardian.com/news/datablog 2011/Dec/07/ cancer-causes-list is in Table 5. The sample data set contains lots of missing data which has to be removed by using data preprocessing using Weka Software before horizontal representation. Experimented with this method using MS SQL Server2008 and find the size of the table. The SQL aggregation function has been applied in the sample data and analyzed the resultant data set. From the tables 2 and 3 shows that horizontal layout representation increases the column of the table than vertical representation. So the resultant table increases the table column in the database.

Table 5
Cancer Data Set

Exposure	Gender	Oral cavity & pharynx	Oeso-phagus	Stomach	Colon-rectum	Liver	Pancreas	Gall-bladder	Larynx	Lung	Meso-thelioma	Melanoma	Breast	Cervix uteri	Corpus uteri
Tobacco	*Male*	69.5	62.6	26.1	6.6	27.3	26.2	-	79	87.3	-	-	-	-	-
Alcohol	*Male*	37.3	25.3	-	15.5	11.4	-	-	27.3	-	-	-	-	-	-
Fruit & veg	*Male*	57.2	46.6	37	-	-	-	-	45.9	8.5	-	-	-	-	-
Meat	*Male*	-	-	-	24.8	-	-	-	-	-	-	-	-	-	-
Fibre	*Male*	-	-	-	10.2	-	-	-	-	-	-	-	-	-	-
Salt	*Male*	-	-	30.9		-	-	-	-	-	-	-	-	-	-
Overweight &	*Male*	-	26.9	-	13.6	-	12.8	19.7	-	-	-	-	-	-	-
Physical exerci	*Male*	-	-	-	3	-	-	-	-	-	-	-	-	-	-
Infections	*Male*	12.3	-	29.2	1.5	19.6	-	-	10.6	-	-	-	-	-	-
Radiation - ion	*Male*	-	2	0.9	1.1	0.6	-	-	-	4.2	-	-	-	-	-
Radiation - UV	*Male*	-	-	-	-	-	-	-	-	-	-	89.8	-	-	-
Occupation	*Male*	0.6	3.3	3	-	0.2	0.02	-	2.9	20.5	97	-	-	-	-
Tobacco	*Females*	54.9	71.3	15.4	9.9	15.3	31	-	79.1	83.6	-	-	-	7.2	-
Alcohol	*Females*	16.9	11.3	-	6.9	5	-	-	12.2	-	-	-	6.4	-	-
Fruit & veg	*Females*	53.6	45.1	33.9	-	-	-	-	43.5	9.3	-	-	-	-	-
Meat	*Females*	-	-	-	16.4	-	-	-	-	-	-	-	-	-	-
Fibre	*Females*	-	-	-	14.6	-	-	-	-	-	-	-	-	-	-
Salt	*Females*	-	-	12.1	-	-	-	-	-	-	-	-	-	-	-
Overweight &	*Females*	-	11.2	-	12.2	-	11.5	17.8	-	-	-	-	8.7	-	33.7
Physical exerci	*Females*	-	-	-	3.6	-	-	-	-	-	-	-	3.4	-	3.8
Post-menopau	*Females*	-	-	-	-	-	-	-	-	-	-	-	3.2	0	1.2
Infections	*Females*	14	-	36	3.1	9.3	-	-	10.6	-	-	-	-	100	-
Radiation - ion	*Females*	-	3.9	1.7	2.2	1.1	-	-	-	5.4	-	-	0.9	-	-
Radiation - UV	*Females*	-	-	-	-	-	-	-	-	-	-	82.4	-	-	-
Occupation	*Females*	0.2	1.1	0.3	-	0.1	-	-	1.6	4.3	82.5		4.6	0.7	-
Reproduction (	*Females*	-	-	-	-	-	-	-	-	-	-	-	3.1	-	-
Tobacco	*Persons*	64.5	65.5	22.2	8.1	23	28.7	-	79	85.6	-	-	-	7.2	-

CONCLUSION

In this paper presented a new approach which reduced the row size in the database using horizontal layout representation and experimented with cancer data set. When we apply this horizontal representation of this dataset will increase the column and reduced the rows in the table. In the future, we planned to implement and analyze the performance of this horizontal layout representation concept using ORACLE and MySQl.

REFERENCES

1. ThomasH.Hinke.,Inference. Aggregation Detection In Database Management System. IEEE,(1988).
2. Johanna Wenny Rahayu and David Taniar. Preserving Aggregation in an Object-Relational DBMS, Springer-Verlag Berlin Heidelberg , 1–10(2002).
3. Rumbaugh, J. et al. Object-Oriented Modelling and Design, Prentice-Hall, (1991).
4. S.Brintha Rajakumari and C.Nalini. An efficient cost Model for data storage with horizontal layout in the cloud, Indian Journal of Science and Technology, 7(S3):45-46,(2014).
5. C. Ordonez. Data Set Preprocessing and Transformation in a Database System, Intelligent Data Analysis,15(4):613-631(2011).
6. C. Ordonez. Vertical and Horizontal Percentage Aggregations, Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '04), 866-871(2004).
7. C. Ordonez and S. Pitchaimalai. Bayesian Classifiers Programmed in SQL, IEEE

Trans. Knowledge and Data Eng., 22(9):139-144(2010).

8. Carlos Ordonez Norberto Ezquerra Cesar A. Santana. Constraining and Summarizing Association Rules in Medical Data, Knowledge and Information Systems (KAIS Journal). 9(3):259-283(2006).
9. G. Graefe, U. Fayyad, and S. Chaudhuri. On the Efficient Gathering of Sufficient Statistics for Classification from Large SQL Databases, KDD-98 Proceedings Proc. ACM. http://www.aaai.org
10. 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): 678-691(2012).
11. S.Brintha Rajakumari and C.Nalini. An efficient Data Mining data Set preparation using aggregation in relational database , Indian Journal of Science and Technology,7(S5):44-46, (2014).
12. S. Brintha Rajakumari, Dr.C.Nalini. A Comparative Analysis Of Horizontal Layout Representation Of Data, International Journal of Advance Research In Science And Engineering, 4 (01): 984-990(2015).
13. S. Brintha Rajakumari, Dr.C.Nalini. A Noval Approach To Prepare Data Set Using Data Stream Mining, Journal of research in computer science, engineering and technology,1:15– 17(2015).
14. Dr.T.Nalini And S.Revathi. An Enhanced Clustering Algorithm Implemented On Biological Data In Data Mining, International Journal of Pharma and Bio Sciences,4(2):1281–1286(2013).