Indian Journal of Science and Technology, *Vol 7(S5), 44–46, June 2014*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

An Efficient Data Mining Dataset Preparation using Aggregation in Relational Database

S. Brintha Rajakumari* **and C. Nalini**

Department of CSE, B. I. S. T. (Bharath University), Chennai, India;
brintha.ramesh@gmail.com, nalinicha2002@gmail.com

Abstract

To prepare the data set from relational database management system for data mining is very difficult and time consuming task. These prepared data can be used as input in data mining analysis. But traditional structured query language aggregate function returns the records in one column per aggregated group. This paper presents the horizontal representation of data used for dataset preparation in data mining analysis and reduce memory space when evaluated with the cancer dataset.

Keywords: Aggregation, Data Mining

1. Introduction

Data mining is the discovery of models for data. A model can be one of several things. Modelling can be summarizing the data succinctly and approximately, or extracting the most prominent 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. This paper presented the horizontal representation of data used for dataset preparation in data mining analysis and evaluated with the cancer dataset.

2. Related Works

Aggregation is an important concept in database design where composite objects can be modelled during the design of database applications. Therefore, maintaining the aggregation concept in database implementation is essential[2]. Aggregation is a composition (part-of) relationship, in which a composite object ("whole") consists of other component objects ("parts")[3].

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]. Since large-scale aggregation queries typically are used to get a "big picture" of a data set, a more attractive approach is to perform online aggregation, in which progressively refined running estimate of the final aggregate values are continuously displayed to the user. The estimated proximity of a running estimate to the final result is indicated by means of an associated confidence interval. An online aggregation system must be optimized to provide more useful information quickly, rather than to minimize the time to query completion[4].

3. Problem Definition

A cancer data set presented in Table 1 will be used for aggregation using SQL queries. In the table 1, the first column is used as primary key and the remaining three columns contain the information about exposure,

*Author for correspondence

Table 1. The oral cavity and Pharynx cancer data set

S.No	Exposure	Gender	Oral cavity and pharynx
1	Tobacco	Male	69.5
2	Alcohol	Male	37.3
3	Fruit and vegetables	Male	57.2
4	Meat	Male	0
5	Fibre	Male	0
6	Salt	Male	0
7	Overweight and obesity	Male	0
8	Physical exercise	Male	0
9	Infections	Male	12.3
10	Radiation - ionising	Male	0
11	Radiation - UV	Male	0
12	Occupation	Male	0.6
13	Tobacco	Female	54.9
14	Alcohol	Female	16.9
15	Fruit & vegetables	Female	53.6
16	Meat	Female	0
17	Fibre	Female	0
18	Salt	Female	0
19	Overweight & obesity	Female	0
20	Physical exercise	Female	0
21	Post-menopausal hormones	Female	0
22	Infections	Female	14
23	Radiation - ionising	Female	0
24	Radiation - UV	Female	0
25	Occupation	Female	0.2
26	Reproduction (breast feeding)	Female	0

gender and percentage of oral cavity and pharynx. From that table, one of the columns passed to standard SQL aggregations. The primary key column will not be used for aggregations. Normally SQL aggregation returns results in a vertical layout. The percentage of cavity which is a non key field will be used for analysis. Firstly we apply the SQL aggregation function to the main table that gives a vertical layout of information with rows lesser than the original table. Again, we reduce the size of the table by using transformation functions pivot which gives the result in horizontal layout.

4. Analysis of Cancer Data Set

Cancer of the mouth or the oral cavity and the oropharynx is referred to as oral cancer. The sample dataset collected from the http://www.theguardian.com/news/datablog 2011/Dec/07/ cancer-causes-list is in Table 1. The SQL aggregation function has been applied in the sample data and the resultant data set is in Table 2.

The oral cavity and pharynx cancer table has 26 records of four attribute which contains a number of records, risk factor exposure, gender and percentage of oral cavity and pharynx cancer. The risk factor percentage will be calculated based on tobacco, Alcohol, Fruit and Vegetables, Meat, Fibre, salt, overweight and obesity, physical exercise, infections, radiation in ionizing and UV, occupation, Post-menopausal hormones and reproduction (breast feeding).

The statistical survey on oral cancers reveals that more men are affected by the disease than women. Experimented with this method using MS SQL Server2008 and find the size of the table. In the tables 1 and 2 shows that horizontal layout representation record size is lesser than vertical representation. So the resultant table occupies less memory space in the database.

Table 2. Data after horizontal layout

S.No	Exposure	Male Oral cavity and pharynx	Female Oral cavity and pharynx
1	Tobacco	69.5	54.9
2	Alcohol	37.3	16.9
3	Fruit and vegetables	57.2	53.6
4	Meat	0	0
5	Fibre	0	0
6	Salt	0	0
7	Overweight and obesity	0	0
8	Physical exercise	0	0
9	Infections	12.3	14
10	Radiation – ionizing	0	0
11	Radiation – UV	0	0
12	Occupation	0.6	0.2
13	Post-menopausal hormones	0	Null
14	Reproduction (breast feeding)	0	Null

5. Conclusion

In this paper presented a new approach which reduced the size of storage space in the database using horizontal layout representation and experimented with cancer data set to 27 records. In the future, a large data set can be worked out.

6. References

1. Thomas H. Hinke.,Inference Aggregation Detection In Database Management Systems,IEEE,1988.
2. Johanna Wenny Rahayu and David Taniar Preserving Aggregation in an Object-Relational DBMS, Springer-Verlag Berlin Heidelberg , pp. 1–10, 2002.
3. Rumbaugh, J. et al, Object-Oriented Modelling and Design, Prentice-Hall, 1991.
4. Peter J. Haas Joseph M. Hellerstein, Ripple Joins for Online Aggregation, ACM 1999.