STATISTICAL RELATIONAL MODEL

Sakti P. Ghosh
Computer Science
IBM Almaden Research Center
San Jose, California.

Abstract

We will outline some of the important problems of extending the relational model for processing statistical data bases. We achieve this by augmenting to relational algebra few more numerical operations without altering the relational algebra. This algebra is referred to as statistical relational algebra. Applications of this extended relational model for solving some common statistical problems are discussed in the paper.

1. Introduction

In the last seven years many researchers in databases and statistics have been exploring and contributing in the area of statistical database management. The biannual workshop on statistical and scientific databases have contributed to this advancement
also (Wong (1982a), Hammond & McCarthy (1983). Many of the research have concentrated on inventing new data models and data languages for processing statistical databases (Sue (1983), Shoshani (1982), Wong (1982), Ghosh (1984). Recently attempts have also been made to invent new algebra for statistical databases (Fortunato, Rafanelli, Ricci & Sebastio (1986)). Some work has also been done in the area of statistical metadata (Ghosh (1985), but no attempt has been made upto now to extend the relational model to handle statistical processing also. The relational data base model was invented by Codd (1970) to provide a simple representation of logical structured data using relational algebra. Many attempts have been made to extend the relational algebra for other data models (including applications like: office data, distributed data). These have heavily emphasized only the logical/communication aspects of data. Most of the data models have claimed that they are suitable for handling statistical data, which is true but they are very inefficient.

The goal of this paper is to extend the Codd's relational algebra by augmenting it with commands which perform statistical processing, without altering the relational structure of the data. Thus the extended relational model will have the capability of simultaneously performing logical structural and statistical manipulation of data when it is executing a command. The syntax of a query language for performing basic statistical processing within the relational DBMS environment and some proposed extensions of SQL (1983) are discussed. This language extensions can also be included in INGRES (Stonebraker (1986). This extension shifts many of the tasks of the statistical application programmers to the DBMS, and thus increasing efficiency of statistical applications.

One important feature of relational algebra is that a relational operation can be executed on a table (or two tables for joins) and the result is also a table. On a relational table it is possible to define separate types of algebra on columns (domains) and rows (tuples). We shall define the operations of relational algebra on the domains (attributes) and numerical algebra on the rows or tuples of the relation. This type of algebra is most advantageous for many statistical database management problems. Thus it is possible to define the relational operations of projection, selection, ordering, union, intersection,