RTL
A Relation and Table Language for statistical databases

Lotfi Lakhal[1,3], Rosine Cicchetti[2,3]

and

Serge Miranda[3]

1- University of Tunis, ENSI, 2049, Ariana, TUNISIE.
2- CERAM, Sophia Antipolis BP 120, 06561 Valbonne cedex, FRANCE.
3- BAOU Project, University of Nice-CNRS, LISAN, Bat 3, Rue A. Einstein,
06560 Sophia-Antipolis, Valbonne, FRANCE.

Abstract

In this paper, we present a multidimentional data language-RTL (Relation and Table Language) allowing statisticians to handle micro and macro statistical databases. We first describe our data model for statistical applications, based upon a twofold structure : relation and Complex Statistical Table (CST). Then, we present RTL, that encompasses relational algebra with aggregate operator, transformation operators, in order to achieve conversion between relations and CSTs, and specific CST-manipulation operators, allowing to modify CST organization (and making this data-structure dynamic) or to aggregate macro-data.

Key-Words

Aggregate functions ; Complex Statistical Table ; Macro-data ; Micro-data ; Relational algebra ; Statistical databases ; Statistical Table.

0 Introduction

They are two broad kinds of Statistical DataBases (SDBs), micro and macro-SDBs [WONG84], [RAFA86]. Micro-SDBs contain records of individual entities or event, for instance, census raw-data, seismic events, mortality data of individual persons. This kind of statistical database can be modelized by relational structure. Macro-SDBs contain only "summary-data sets" achieved by statistical mathematical operations on micro-SDBs. Our attention is devoted to study the use of these two classes of SDBs, by proposing a multidimensional data language allowing statisticians to handle micro and summary data.

Summary-data sets are aggregated-data sets according to a particular combination of criteria [JOHN81], [CHAN81]. They are characterized by two kinds of attributes : **summary attributes**, also called arithmetical or quantitative variables and **category attributes** that permit to classify values of summary attributes. Category attributes frequently have an alphanumerical basic type and have generally (but not mandatory) small domain size [TURN79]. Summary attributes can only be of numerical type since their values are achieved by applying aggregative functions (Sum, Avg, Count, Min, ...) on quantitative measures of a studied micro-data sub-set [KLUG81, 82], [SHOS85], [OZSO85a].
For instance, in a macro-SDB concerning "student population", the summary-data set S1 permits to count the Nice students by sex, study year and discipline. S1 contains the category attributes SEX defined in the value set {W, M}, DISCIPLINE defined by {Computer science, Mathematics, Physics}, STUDY_YEAR defined by {1, 2}, and the summary attribute NB (number of students) defined by the integer set. They are several levels of complexity in the semantical contents of summary-data sets. Some, as S1 in the previous example, have a single semantic, others can represent several semantics, through different summary attributes, classified according to values of the same or different category attributes. For instance, the summary-data set S2 permits to count the Nice students in graduate studies, by sex, discipline and study year (first semantic) and gives mean of student scholarship by sex and study year (second semantic). It is very important to have a suitable logical model in order to represent summary-data sets, and to avoid processing statistical units that are micro data.

"Statistical tables" have been used for a long time, by statisticians, to present achieved results (output forms), as in statistical packages (P-STAT [PSTAT81] with the command TABLES or SPSS [NIE75] with BREAKDOWN and CROSSTABS). They are easy to comprehend (since they facilitate interpretation of statistical macro-data) and deal with ergonomically (because of data rational organization). Recently, such tables have inspired some research contributions with summary-data modelization. The "statistical table" representation is seen as a logical structure with operators [GHOSH84], [OZSO85a,b], but often limited (to simple tables) as in Rafanelli's approach [RAFA86]. Some of the most significant contributions, in this field, have been proposed by Ghosh [GHOS84] and Ozsoyoglu [OZSO84], [OZSO85a, b]. These approaches, that we examine in [LAKH89a], inspire us with a multidimensional data structure called "Complex Statistical Table" (that can be seen as an extension of usual "Statistical Table") and suitable operators for macro-SDBs.

A Statistical Table (ST) may modelize summary-data sets with a simple semantics. ST involves scheme (intension) and values (extension). The ST scheme is presented as an array, in which category-attribute names appear, in the row and column headings, and the unique summary-attribute name appears in the cell. ST row (or column) are actually organized according to category-attribute scheme defined as an ordered set of category-attribute names.
For instance, in the statistical macro-SDB "Student population", let's consider the ST P12 , that counts the Nice students in graduate studies by sex, discipline and study year. P12, whose scheme is given in figure 1, has a unique summary attribute : NB12.

P12		STUDY_YEAR
SEX	DISCIPLINE	NB12

figure 1: P12 schemes

ST extension is a table of summary-attribute values and category-attribute-values. The row and column organization correspond to the cross product of row or column category-attribute values.

P12				STUDY_YEAR	
				1	2
SEX	W	DISCIPLINE	Computer sciences	45	42
	W		Mathematics	40	38
	W		Physics	37	37
	M	DISCIPLINE	Computer sciences	58	55
	M		Mathematics	40	40
	M		Physics	35	30

Figure 2 : P12 extension

A Complex Statistical Table (CST), may represent summary-data sets with a complex semantics. CST is the composition, according to rows or columns, of N STs (N ≥ 1), which just makes a concatenation of their row or column category-attribute scheme. The definition of the CST row (and column) organization is a category-attribute multi-scheme (ordered set of schemes).
For instance let's consider the CST S1 composed of two STs, whose schemes are illustrated in figure 3. S1 scheme is given in figure 4.

P11		STUDY_YEAR
SEX	CLASS	NB11

P12		STUDY_YEAR
SEX	DISCIPLINE	NB12

Figure 3: P11, P12 schemes

S1 P11 P12		STUDY_YEAR
SEX	CLASS	NB11
SEX	DISCIPLINE	NB12

Figure 4 : S1 scheme