データベース・システム 41 － 1
（１９８４．６．５）

Statistical Database Design Method

R. Hotaka
University of Tsukuba
Sakura, Niihari, Ibaraki 305 Japan

ABSTRACT

The design procedure is roughly divided into 2 steps. First, if there are many ways to represent the same statistical fact, more elemental or fundamental representations are obtained by decomposing the given statistical files. This is accomplished by purifying summary attributes and orthogonalizing category attributes. Next, decomposed files are synthesized by horizontal composition and vertical composition.

1 Statistical Data Model

We call a triplet (AES, FS, VS) a statistical observation where AES is a set of atomic events, VS a set of numerical values and FS a set of functions from AES to VS, i.e.,

$$FS \subset VS^{AES}.$$

where A^B denotes the set of all functions defined on B and maps into A. A subset $E \subset AES$ is called an event. Given a statistical observation, we can define various statistical files on it. We do not give the detailed explanation of statistical files, but readers can grasp the image of it by examining the example 1 given below.

First, we introduce a category set CS. We define a mapping S from CS into 2^{AES}. E.g., if $x \in CS$, $S(x) \subset AES$. A member of CS is called a category.

A set $C \subset CS$ is called a classification hierarchy if it satisfies the following:

(1) There is a special category w, called the whole category such that
$w \in C$ and $S(w) = AES$

(2) For every x in C, $S(x) \neq \emptyset$

(3) If $x_1, x_2 \in C$, then $S(x_1) \cap S(x_2) = \emptyset$ or $S(x_1) \subset S(x_2)$ or $S(x_2) \subset S(x_1)$

This work was supported in part by the grant (No. 58115004) from the Ministry of Education of Japan.

(4) If $x_1, x_2 \in C$ and $S(x_1) = S(x_2)$ then $x_1 = x_2$

A classification hierarchy is called partition-type if it satisfies the following:

(5) For every x_1, $x_2 \in C$ such that $S(x_1) \supsetneq S(x_2)$ there exist $x_3,\ldots,x_m \in C$ such that $S(x_1) = U_{i=2}^{m} S(x_i)$ and $S(x_p) \cap S(x_q) = \emptyset$ for $2 \leq p < q \leq m$.

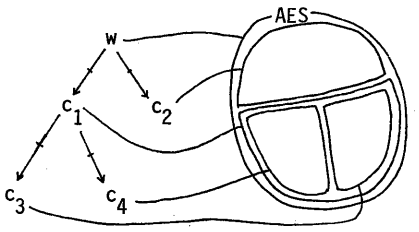

Fig. 1 A classification hierarchy

Above figure 1 shows a classification hierarchy which is also partition-type.

Given m classification hierarchies $C_i (i=1,\ldots,m)$ and n functions $f_j \in FS (j=1,\ldots,n)$, we define a statistical file $SF(C_1,\ldots,Cm; f_1,\ldots,f_n;$ AES, FS, VS) as the set of tuples

$$\{(x_1,\ldots,x_m; B_1(E),\ldots,B_n(E)) \mid x_1 \in C_1, x_2 \in C_2,\ldots, x_m \in C_m, E = \bigcap_{i=1}^{m} S(x_i) \}$$

where $B_j(X)$ is generally defined for any event $X \subset$ AES by

(6) $B_j(X) = \sum \{f_j(e) \mid e \in X \}$

B_j is called a summary attribute.

The above file may be interpreted as a relation which has the relational schema

$$SF(A_1,\ldots,A_m \mid B_1,\ldots,B_n).$$

$A_i (i=1,\ldots,m)$ will be called a category attribute. We placed '|' to distinguish category attributes and summary attributes.

Example1

Atomic events (Employee)	Descriptions of events		Values of functions			
	CITY	SEX	Count	Income	No-of-Male-emp	No-of-Female-emp
e_1	Tokyo	Male	1	1	1	0
e_2	Osaka	Male	1	4	1	0
e_3	Osaka	Female	1	8	0	1
e_4	Tokyo	Male	1	3	1	1
e_5	Tokyo	Female	1	5	0	0

Statistical Observation

AES = $\{e_1, e_2, e_3, e_4, e_5\}$
FS = {Count, Income, No-of-Male-emp, No-of-Female-emp}
VS = { x | x is a real number }

A Statistical File

C_1 ={w, Tokyo, Osaka}
C_2 = {w, Male, Female}

where S(w) = AES

S(Tokyo) = { e | CITY(e) = Tokyo }
S(Osaka) = { e | CITY(e) = Osaka }
S(Male) = { e | Sex(e) = Male }
S(Female) = { e | Sex(e) = Female }

$f_1(e)$ = Count(e)
$f_2(e)$ = Income(e)
$f_3(e)$ = No-of-Male- emp(e)
$f_4(e)$ = No-of-Female-emp(e)

$SF(C_1; f_1, f_2, f_3; AES, FS, VS)$

={(w; 5, 22, 3),
  (Tokyo; 3, 10, 2),
  (Osaka; 2, 12, 1)}

Relational schema of the above statistical file

= SF(CITY | Count, Income, No-of-Male-emp)

(We deliberately omitted No-of-Female-emp.)

2 Summary attribute

Definition (pure summary attribute)

A summary attribute B is called pure iff there does not exists an event $D \subsetneq AES$ such that

$B(E) = B(D \cap E)$ for every event of AES.

According to this definition, the purity of an attribute depends on a particular statistical observations. This is not favorable in general, but since the improvement is straightforward but rather cumbersome, we leave it as the future task.

Assumption (extraction of pure summary attribute from a summary attribute)

For each summary attribute B, there exists an event D and pure summary attribute V such that

$B(E) = V(D \cap E)$ for every event E.

D is called the a subtype event and V the pure summary attribute of the summary attribute B.

Example 2

In the above example 1, attribute B_3 is not pure since

$$B_3(E) = B_3(\{e_1, e_2, e_4\} \cap E)$$

for every $E \subset AES$. Attributes B_1 and B_2 are pure since for every i(i=1,2), we cannot find $D \subsetneq AES$ such that

$$B_i(E) = B_i(D \cap E)$$

for every E.

3 Orthogonality of a category attribute

Sometimes a category x can be represented using two categories y, z by the following equation.

(1) $S(x) = S(y) \cap S(z)$

This suggests the possibility of replacing one classification hierarchy by others.

Definition (orthogonality)

m partition-type classification hierarchies $\{C_i | i=1,\ldots,m\}$ are orthogonal iff for each category $x_i \in C_i$ $(i=1,\ldots,m)$, $\bigcap_{i=1}^{m} S(x_i) \neq \emptyset$.

Theorem 1 Let classification hierarchies $C_1, C_2, \ldots, C_m$ be partition-type and orthogonal. If $x_i, y_i \in C_i$ $(i=1,\ldots m)$ and $\bigcap_{i=1}^{m} S(x_i) = \bigcap_{i=1}^{m} S(y_i)$, then $x_i = y_i$ for every i $(i=1,\ldots,m)$.

Proof. The proof is similar to that of the stronger theorem 2 below. Hence ommitted.

C_i corresponds to an axis and x_i to the coordinate value of vector space. Ordinarily, the coordinate value depends on the selection of axes. But, in our case, we can show that the coordinate value is independent of the other axes. Formally, we have:

Theorem 2 Suppose $\{C, C_1\}$ and $\{C, C_2\}$ are two sets of orthogonal classification hierarchies. If $x_1, x_2 \in C$, $y_1 \in C_1$ and $y_2 \in C_2$ such that $S(x_1) \cap S(y_1) = S(x_2) \cap S(y_2)$ then $S(x_1) = S(x_2)$.

Proof. Suppose $S(x_1) \neq S(x_2)$. Then without loss of generality, we can assume $S(x_1) \supsetneq S(x_2)$. Choose $x_3, \ldots, x_m \in C$ such that

$$S(x_1) = \bigcup_{i=2}^{m} S(x_i) \quad (m \geq 3) \text{ and } S(x_p) \cap S(x_q) = \emptyset \ (2 \leq p < q \leq m)$$

Then,

$$S(x_1) \cap S(y_1) - S(x_2) \cap S(y_2)$$
$$= S(x_1) \cap S(y_1) - S(x_1) \cap S(x_2) \cap S(y_1) \cap S(y_2)$$

$\supset (S(x_2) \cap S(y_1)) \cup (S(x_3) \cap S(y_1)) - S(x_1) \cap S(x_2) \cap S(y_1) \cap S(y_2)$

$\supset S(x_3) \cap S(y_1) - S(x_1) \cap S(x_2) \cap S(y_1) \cap S(y_2)$

$= S(x_3) \cap S(y_1) - S(x_1) \cap S(x_2) \cap S(x_3) \cap S(y_1) \cap S(y_2)$

$= S(x_3) \cap S(y_1)$ (because $S(x_2) \cap S(x_3) = \emptyset$)

$\neq \emptyset$ (by the orthogonality of $\{C, C_1\}$)

hence the contradiction.

Example 3

$\{C_1, C_2\}$ in example 1 are orthogonal.

Usually, category attributes of a statistical file is selected such that domains of each of the category attributes are orthogonal. Note, however, in a particular statistical observation, those domains may happen to fail to be orthogonal. But in design stage, one performs the purification or orthogonalization by considering only the theoretical possibility or tendency in the long run. One does not care what the actual statistical observation really is.

4 Canonical representation of statistical files

4.1 Design procedures

Starting from existing statistical files, we follow the following procedures to get the canonical representations of statistical files.

First, similar summary attributes are collected to form a group (section 4.2). For each group, purification (section 4.3) and orthogonalization (section 4.4) are performed. After these processes, original statistical files are decomposed into atomic units of statistical information. Next step is to synthesize these information to obtain more bigger units. This is achieved by horizontal composition (section 4.5) and vertical composition (section 4.6).

4.2 Grouping of summary attributes

Consider a statistical file schema

$$F(A_1,\ldots,A_m \mid B_1,\ldots,B_n).$$

Let D_j and V_j be the subtype event and pure summary attribute of B_j respectively. The above file is equivalent to the following n files each of which has a simple attribute.

$$F_1(A_1,\ldots,A_m \mid B_1)$$
$$\vdots$$
$$F_n(A_1,\ldots,A_m \mid B_n) \quad (1)$$

F_i's are grouped into one group if the pure summary attributes V_i's are the same. F_i's are thus grouped into r groups G_p (p=1,...,r). G_p has r_p files

$$F_{i_1}(A_1,\ldots,A_m \mid B_{i_1}),$$
$$\vdots$$
$$F_{i_{r_p}}(A_1,\ldots,A_m \mid B_{i_{r_p}}), \quad (2)$$

which have the same pure summary attribute V_p.

Example 4

In example 1, B_3(defined by No-of-Male-emp) and B_4(defined by No-of-Female-emp) have the same pure summary attribute B_1(defined by No-of-emp).

4.3 Purification

Consider the group G_p (p=1,...,r_p). Let D_j be the subtype event of B_j and x_j the category such that

$$D_j = S(x_j) \quad (j=1,\ldots,n).$$

Assumption For any p,q (p,q=$i_1,\ldots,i_{r_p}$),

$$D_p \cap D_q = \emptyset \text{ or } D_p \subset D_q \text{ or } D_q \subset D_p.$$

Create a minimal calassification hierarchy C_p' such that

$$x_j \in C_p' \quad (j=1,\ldots,n)$$

and a category attribute A_p' whose domain is C_p'. The set of r_p statistical files (2) is equivalent to the purified statistical file

$$G_p(A_1,\ldots,A_m,A_p' \mid V_p).$$

Example 5

$$F(\text{CITY} \mid B_3, B_4)$$

is equivalent to

$$G(\text{CITY}, \text{SEX} \mid B_1)$$

where the domain of CITY = C_1 and the domain of SEX = C_2.

4.4 Orthogonalization

If a domain of an attribute A can be represented by s partition-type orthogonal classification hierarchies $C_p(p=1,\ldots,s)$, A is replaced by s attributes $A'_1,\ldots,A'_s$ which have domains $C'_1,\ldots,C'_s$ respectively. If partition-type orthogonal classification hierarchies are selected, then the decomposition is unique. But the selection of classification hierarchies itself may not be unique. These orthogonal decomposition is repeated as far as there exists a domain of a category attribute that can be decomposed orthogonally.

4.5 Horizontal composition

After purification and orthogonalization, we have many statistical files each of which has a single summary attribute. By the following rules, these statistical files will be horizontally composed to make bigger statistical files.

Two files

$$F_1(A_1,\ldots,A_{m-1},A'_m \mid B)$$

and

$$F_2(A_1,\ldots,A_{m-1},A''_m \mid B)$$

will be composed to form

$$F(A_1,\ldots,A_m \mid B)$$

when we can define a category attribute A_m and its domain(classification hierarchy) which contains both domains of A'_m and A''_m.

4.6 Vertical composition

We still have many statistical files each of which has a single summary attribute. Suppose

$$F_1(A_1,\ldots,A_m \mid B_1)$$

$$\vdots$$

$$F_k(A_1,\ldots,A_m \mid B_k)$$

be k statistical files. We vertically compose these files into one statistical file

$$F(A_1,\ldots,A_m \mid B_1,\ldots,B_k)$$

After the every possible vertical composition, we finally get the canonical representation of original statistical files.

5 Conclusion

Statistical database design procedures have been proposed to obtain the canonical representations of statistical files starting from existing ones. A statistical data model has been rigorously presented following Chan and Shoshani's SUBJECT model[(2)].

New concepts, purification and orthogonalization are introduced to assure unique statistical file decomposition. File composition procedures are somewhat arbitrary, but following the same procedures makes it possible for various designers to reach the same statistical file design.

We have assumed the existence of the same whole category w throughout the statistical files. In practice, however, this cannot be assumed. Extentions are future tasks.

As for the summarizing operator, only summing operator Σ was considered. The extension is also a future task.

Acknowledgement

The author is very grateful for Dr. H. Sato for his kind comments and introduction to statistical database problems.

References

(1) R. Hotaka: Name and Meta-object-oriented Data Model, Database symposium, Dec. 1983, pp.7-21.

(2) P. Chan and A. Shoshani: SUBJECT: A Directory driven System for Organizing and Accessing large Statistical Databases, VLDB 1981, pp.553-563.

(3) W. Kent: Choices in Practical Data Design, VLDB, 1982, pp.165-180.