HIROSHIMA MATH. J.
20 (1990), 443–476

A mathematical study on statistical database designs

Yasuyuki KOBAYASHI

(Received September 4, 1989)

1. Introduction

Let be given a finite set U and non-negative integers $f(x)$ for all $x \in U$. Then, by taking the sum of products of them, we have an integer

$$\mathrm{SP}_f(\mathscr{A}) = \textstyle\sum_{A \in \mathscr{A}} \prod_{x \in A} f(x) \tag{1}$$

for each subfamily $\mathscr{A} \subset 2^U - \{\varnothing\}$, especially, for any covering $\mathscr{A}$ of U; and we can consider the following

PROBLEM. For given U, f as above and a covering $\mathscr{B}$ of U, find effectively $\mathscr{A}$ in such coverings $\mathscr{A}'$ that $\mathscr{B}$ is a refinement of $\mathscr{A}'$ so that the function SP_f in (1) takes the minimum value at $\mathscr{A}$ among such coverings $\mathscr{A}'$ (see Definition 2.3).

We call $\mathscr{A}$ in this problem an MSPD for $\langle U, \mathscr{B}, f \rangle$ simply. Of course, an MSPD exists and any MSPD can be found by calculating $\mathrm{SP}_f(\mathscr{A}')$ for all finitely many such $\mathscr{A}'$; but the number of $\mathscr{A}'$ may increase rapidly as $|U|$ increases. ($|X|$ denotes the number of elements in a finite set X.)

Thus, the purpose of this paper is to establish an effective method of finding an MSPD of special type, which is applicable even when $|U|$ may be large.

Our motivation is in the problem on statistical database designs stated in §5. (For databases, cf. Codd [3–5] and Smith-Smith [23], and for statistical databases, cf. Shoshani [22] and several papers in the reference.)

Let R be a given collection of statistical records, that is, a finite subset of the product $D = \Pi_{i=1}^N D_i$ of domains D_i of i-th field. Then, an aggregation function S can be specified by the category fields $X(S)$, the summary fields $Y(S)$ and the summarizing operators g_j over D_j given for each summary field j in $Y(S)$; and S gives us the summary table $S(R)$ corresponding to $X(S)$, $Y(S)$ and g_j's. Moreover, for any finite set $\mathscr{S}$ of aggregation functions, we have

$$\mathrm{NRec}\,(\mathscr{S}) = \textstyle\sum_{S \in \mathscr{S}} |S(R)|\,, \tag{2}$$

the total number of records of $\{S(R)\colon S \in \mathscr{S}\}$. Thus we have the following

PROBLEM. Let R be a given collection of statistical records. Then, for a finite set of summary tables $\{S_0(R)\colon S_0 \in \mathscr{S}_0\}$ to be derived from the database,