数理解析研究所講究録
第593巻 1986年 197-215

統計データベース管理システム

大阪大学　工学部　通信工学教室
打浪　清一　(Seiichi UCHINAMI)

[I] 統計データベースシステムの特徴

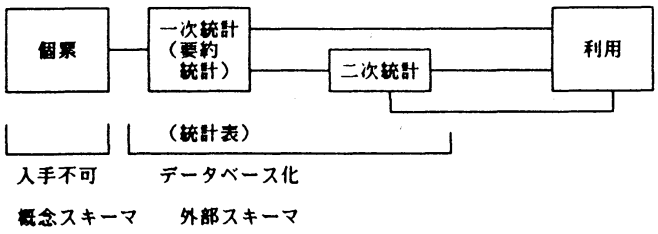
統計データベースにおいては，個票データはプライバシ保護のために公開されないことが多い。そのため，各種観点から集計された表が多量に出来，これをうまく管理する必要が生じる。また利用者の見たい表が値一属性変換等により多くの形をとる。これらの管理が可能な概念スキーマの設計法と，その上での外部スキーマ管理について述べる。また統計データベースシステムに於いては，単一データをアクセスするのではなく，ある属性の値が同じものを検索し，その集計を取るというように，グループデータの検索を行うことが多い。この様な用途に適した新しいファイル構成法を述べる。なおファイル構成法の詳細は次の『統計データベースに於ける一連検索用ファイル編成法』（近藤他）に述べる。

1．統計データベースシステムの利用法

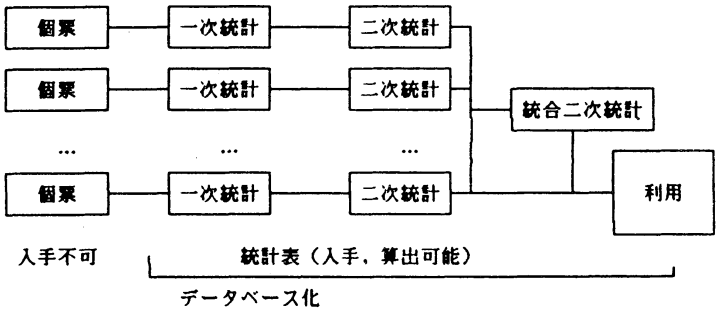
統計データベースシステムにおいては次の２つの利用レベルがある。

第一レベル　図１(a) に示すように一種類の調査原票に基づく要約統計データを蓄積し，これを組合わせ或いは統合して利用者の要求する統計表を作成する。ここでは個票から概念スキーマを作成する。蓄積された要約統計は外部スキーマの集合と考える。

第二レベル　図1(b)に示すように異種の個票から作成された異種の統計表やそれらの二次統計表を組合わせて利用者の要求する統計表を算出する。このレベルでは個票の集合から概念スキーマを構成する。また蓄積あるいは算出する統計表は外部スキーマ集合を構成する。



(a) 第一レベルの統計データベース利用



(b) 第二レベルの統計データベース利用

図1 統計データベース利用のレベル

2．統計データの特徴

経済統計等の統計データの特徴としては，次のものが挙げられる。

(1) 統計データは本質的に多次元性を備えたカテゴリカルデータである。

(2) 各統計表に表れる属性はカテゴリ属性とサマリ属性に２分され，カテゴリ属性は一般に階層構造をなす。

カテゴリ属性：調査，測定されたデータを分類する属性　例；地域，性別，産業分類

サマリ　属性：カテゴライズされた集合の特性値で，統計量を値として持つ属性

例；人口，世帯数，面積

(3) 一つの統計表に表われるカテゴリ属性は多くても３つまでである。それ以上増えると複雑となり，平面的なテーブルでは表現しにくい。

(4) 統計量は複数のカテゴリ属性と，一つのサマリ属性が決まると一意的に決まる。

(5) 一般に表側にカテゴリ属性，表頭にサマリ属性を表示するが，表の外形を統一し，見易くする為にカテゴリ属性の一部を表頭に用いることが多い。

(6) 異種のカテゴリ属性が統計表に直和型で表せられている時は統計表を分離出来る。

(7) 統計データベースシステムの検索処理に於いては，一連性検索がよく用いられる。このため，一連性検索に適したファイル構成をするべきである。

[11] 統計データベースシステムの概念スキーマ設計論

統計データは，その利用目的に合わせ，表の縦軸と横軸を交換したり，個々の値と合計を同じ欄に記録したりしており，このような利用上の表形式をもとに，データベースシステムの貯える全体像である概念スキーマを設計すると，見通しがわるく，最適なものが設計出来ない恐れがる。そこでこれを効果的に行なえる方法を提案した。

1．概念スキーマ

各種調査の結果については，各観点からの集計結果の要約統計が表になって出版されておりこれをＤＢ化していることが多い。しかしこの場合に個票を出発点として概念スキーマを設計すべきである。また属性一値変換操作により，同一対象を表した統計が何通りもの異なる表の形態をとることが多い。これらを統一的に扱うための概念スキーマの設計法として次にあげる方法を提案する。

概念スキーマは位相情報空間モデルに基づく多次元位相情報空間として定義する。位相情報空間は３レベルの空間生成文法により定義する。第一レベルは部分空間の合成法を指定し，直積・，連結和＃，直和｜がある。第二レベルは各部分空間を構成する各座標軸を定義し，第三レベルで各座標軸に位相を割り当てる。

[定義　1]　空間構成文法

全体空間がどのような部分空間から構成されるのかを規定する。

$$G_i = < V_{Ni}, V_{Ti}, P_i, S_i > \qquad i = 1$$

以下　第二，第三レベルも同じ形の文法として，定義される。（ｉ＝２，３）

$$V_{Ni} \cap V_{Ti} = \phi, \quad V_{Ni} \cup V_{Ti} = V, \quad (i = 1, 3)$$

生成規則は次の形をしている。

$$P1 : A \rightarrow \omega, \quad V_{N1} \ni A, \quad L^{\triangle}(V1, D) \ni \omega$$

$$L^{\triangle}(V, D) = \{ (\alpha\, \delta)^{*}\, \beta \mid V \ni \alpha, \beta, \quad D \ni \delta \}$$

$$D = \{ \cdot, \#, | \}$$

Ｄは空間構成子で，直積・，直和｜，連結和＃がある。

直和とは同一座標軸上の互いに素な区間の和を表し，連結和とは異なる座標軸を並べたものを意味する。直積は互いに素な座標軸をもってより高い次元の空間を構成することを意味する。

［定義　2］　空間軸規定文法

各部分空間を構成する座標軸を規定する。G_2 は定義１で示された形であって，生成規則は

$$P_2 : A \rightarrow \omega, \quad V_{N2} \ni A, \quad V_{T2} \ni a, b$$

$$\{a, a, aa, aba^{-1}b^{-1}, aa^{-1}, aba^{-1}\} \ni \omega$$

・

aは巡回軸で始端と終端が同一視される。

［定義　3］　空間位相規定文法

このレベルでは，生成規則が２種類あり，各軸の位相を規定するP_{3T}と，利用時にその一部をピックアップして指し示すP_{3A}とがある。

P_{3A} ：A→ω，　{各種の位相空間} ∋ ω

P_{3T} ：A→ω，　{位相空間の領域} ∋ ω

位相としては，数軸，順序軸，その他種々の位相がとられる。

［定義　4］　統計データベース空間生成文法

$$G_I = < G_1, G_2, G_3 >$$

ここで，G_i は，上述の文法であって，三者の間には

$$S_1 = S, \quad S_2 = V_{T1}, \quad S_3 = V_{T2}$$

なる関係がある。

［定義５］　座標軸組　C

Ｎ個の属性を持つ統計データベースは，Ｎ次元統計空間に写像される。このＮ個の属性の組を座標軸組といい，次のように記す。

C＝（属性１，属性２，属性３，…　，属性ｉ，　　，属性ｎ）

［定義６］　第一レベルの統計データベース空間

各第ｉ属性が，第k_i　レベルの抽象化レベルまである統計調査データベースを次のよう

に表す。

$$F_d = (k_1, k_2, k_3, \cdots, k_n)$$

［定義７］　統計データベースファイル

各第ｉ属性を，第m_i　レベルで集計を行って得られる統計表を次のように表す。

$$F = (m_1, m_2, \cdots, m_n), \quad (\text{ここで } m_i \leqq k_i)$$

統計データベース空間F_d　から計算される統計データベースファイルの全集合を $\mathbf{F}$ と表す。

［定義８］　表の選択

第ｉ属性の第m_i　レベルに於いて，その範囲をＶに制限したものを次式で表す。

$$F = (m_1, m_2, m_3, \cdots, m_i[V], \cdots, m_n)$$

ここで［］内には制約条件を記述する。もし制約条件が等間隔なサンプリングの場合は，ＦＯＲＴＲＡＮのＤＯループのパラメータの様に，［初期値，終了値，増分幅］と記述する。増分幅が１の場合はこれを省略し，［初期値，終了値］と記述してよい。

［定義９］　抽象化（導出）

統計表Ｇを幾つかの欄にわたって集計し，統計表Ｆが算出出来る時，ＦはＧから抽象化（導出）可能であるという。

［定理　１］　単一属性抽象化（導出）

F_1 と F_2 を $\mathbf{F}$ に属する統計表とし，（１），（２）で表されるとする。

$$F_1 = (q_{11}, q_{12}, q_{13}, \cdots, q_{1n}) \qquad (1)$$

$$F_2 = (q_{21}, q_{22}, q_{23}, \cdots, q_{1n}) \qquad (2)$$

ここで　　第ｐ属性のみ抽象度が異なり，他は同一レベルであるとする。

$$F_1 \xrightarrow{p} F_2 \quad <==> \quad q_{1i} = q_{2i} \ \text{for all i except p} \ \text{and} \ q_{1p} > q_{2p}$$

第ｐ属性は，F_1 の方がF_2 属性より抽象度が低いとする。このとき，統計表F_1 の第ｐ属性を集計することにより，統計表F_2 を導出することが可能である。これを

$F_1 \xrightarrow{p} F_2$ と記し単一属性pの抽象化という。また$D_p\{F_1\}=F_2$とも略記する。

［定理　２］　抽象化（導出）の複合演算

単一属性抽象化（導出）に於いては，次の規則が成立する。

可換律　　$D_p D_q \{F_1\} = D_q D_p \{F_1\}$

結合律　　$(D_p D_q) D_r \{F_1\} = D_p (D_q D_r) \{F_1\}$

（証明）

可換律は　$F_1 \xrightarrow{p} F_2 \xrightarrow{q} F_3 \iff F_1 \xrightarrow{q} F_2 \xrightarrow{p} F_3$　と書ける。

ここで表F_1のレコードを属性pとqにのみ着目し，r（a，b）と書く。すると

$$\sum_a \sum_b r(a,b) = \sum_b \sum_a r(a,b)$$　となる。

即ち，属性pで１つのカテゴリになるものを先ず集計し，ついで属性qで１つのカテゴリになるものを集計したものは，その集計順序を変えて，先ず属性qで１つのカテゴリになるものを集計し，次いで属性pで集計したものと等しくなる。これは２次元表の集計に於いて縦横の順でも，横縦の順でも答は同じになることを言っている。

結合律は

$$(F_1 \xrightarrow{p} F_2 \xrightarrow{q} F_3) \xrightarrow{r} F_4 \iff F_1 \xrightarrow{p} (F_2 \xrightarrow{q} F_3 \xrightarrow{r} F_4)$$

となる。これは属性p，q，rに着目すれば，３次元表をどの順で集計するかという問題となり，２次元の場合（可換律）の拡張となっている。よって同様に証明できる。

［定理　３］　多属性抽象化

F_1，F_2を（１），（２）で表せられるファイルとする。

$$F_1 \xrightarrow{} F_2 \iff \text{for all } i \quad q_{1i} \geqq q_{2i} .$$

上記の条件が成立したら，多属性の抽象化が可能である。

（証明）　単一属性抽象化を繰り返すことにより，多属性抽象化が可能である。

［定義　１０］　抽象化（導出）の距離

定理３に示した抽象化に於いて，表F_1とF_2の導出距離は次式で定義される。

$$D(F_1, F_2) = \sum_{i=1}^{n} (q_{1i} - q_{2i})$$

［定理　4］　ファイル抽象化（導出）

F_1，F_2 が式（1），（2）で示され $\mathbf{F}$ に属するとする。

このとき

$$F_1 \rightarrow F_2 \quad <==> \quad \text{for all } i,\ q_{1i} \geqq q_{2i} \qquad (1 \leqq i \leqq n).$$

即ち，上述の条件が満たされたときファイル F_2 はファイル F_1 から抽象化可能である。

（証明）上述の条件が満たされているので，各属性はすべて F_1 から抽象化（導出）できる。そこで第一属性から順に抽象化を行ってゆけば求める表が得られる。

［定理　5］　空間構成子は次の律を満たす。

結合律　　$(A \# B) \# C = A \# (B \# C)$

　　$(A \mid B) \mid C = A \mid (B \mid C)$

分配律　　$(A \# B) \cdot C = A \cdot C \# B \cdot C$

　　$C \cdot (A \# B) = C \cdot A \# C \cdot B$

　　$(A \mid B) \cdot C = A \cdot C \mid B \cdot C$

　　$C \cdot (A \mid B) = C \cdot A \mid C \cdot B$

（証明）連結和に於いては，２つの被演算項は共通な属性を持たない。そこで集合的に要素を列挙していることになる。この列挙においては，順序は無関係なので結合律が成り立つ。直和においては，被演算項は同じ座標軸上にある。そしてこの軸上のトポロジーは空間生成文法で規定するので，利用時には，これを参照して用いればよい。そうすると単一座標軸上の区間を列挙すればよくなり，この列挙においては演算順序は意味が無い。よって直和は結合律が成立する。

直積との分配律に関しては，直積が直交する新しい軸を追加するのみなので，連結和，直和の演算の定義からすぐ分かる。

２．概念スキーマの設計法

設計は次のステップに従って行う。

(1) 第１レベルの処理

第一レベルでは，調査に用いられた個票（調査原票のフォーマット）から概念スキーマを表す生成規則へ，次の手順に従って変換する。

(i) 異なった個体型は異なった部分空間を割り当てる。

(ii)同一の個体型に関係する属性で，互いに素なものは一つの部分空間内の互いに素な座標軸（カテゴリ）を構成する。

例えば，国勢調査に於いて，人口が年度，性別，年齢別，地域別で，昼間人口，常住人口が集計されているとすると，

人口：＝時期・性別・年齢・地域・＜統計種別＞

と書かれる。ここで，時期，性別，年齢，地域はカテゴリ属性で，＜統計種別＞はサマリ属性である。

(iii) 階層的な関係にある属性は，生成規則に於ける適用順にその階層を対応させ，同一軸上に表現することにより，生成規則を定める。

例えば，アジアは，日本，中国，韓国，‥からなっているが，これは，

アジア：＝日本｜中国｜‥

と記述される。このとき左辺は右辺のより上位カテゴリになっている。

(iv)生成規則は分配律などを用いて統合化される。

例えば，

事業所数：＝産業・地域・＜統計種別＞

従業員数：＝産業・地域・＜統計種別＞

のとき

事業所数　&　従業員数：＝産業・地域・＜統計種別；事業所数　# 従業員数＞

となる。

人口統計の場合の生成規則例を図２に示す。またカテゴリ階層の例を図３に示す。

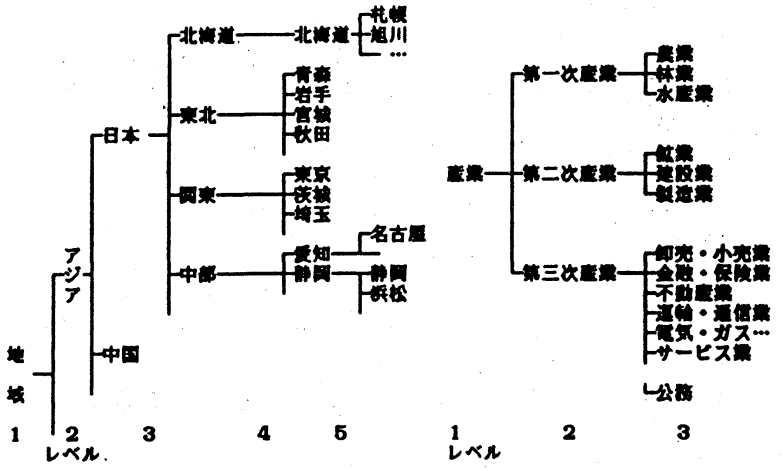
```
人口統計：＝時期・性別・年齢・地域・統計種別                    全空間      P1
時期：＝時刻｜期間                                            時刻1       P1
統計種別：＝常住人口｜昼間人口｜…                              種別1       P1

時刻   ：＝年                                                 時刻2       P2
年     ：＝半年                                               時刻3       P2
半年   ：＝四半期                                             時刻4       P2
四半期 ：＝月                                                 時刻5       P2
月     ：＝旬                                                 時刻6       P2
旬     ：＝日                                                 時刻7       P2
日     ：＝時                                                 時刻8       P2
時     ：＝分                                                 時刻9       P2
分     ：＝秒                                                 時刻10      P2
         …
地域   ：＝大陸                                               地域1       P2
大陸   ：＝国                                                 地域2       P2
国     ：＝地方                                               地域3       P2

年齢層 ：＝乳児｜幼児｜児童｜青年｜成年｜壮年｜熟年｜老年      年齢1       P3T
乳児   ：＝0｜1                                               年齢2       P3T
幼児   ：＝2｜3｜4｜5                                         年齢2       P3T
児童   ：＝6｜7｜…｜12                                       年齢2       P3T
老年   ：＝60｜61｜…                                         年齢2       P3T
性別   ：＝男｜女                                           性別1       P3T
地域   ：＝アジア｜北米｜南米｜オセアニア｜欧州｜アフリカ    地域1       P3T
アジア ：＝日本｜中国｜韓国｜タイ｜インド｜インドネシア｜…  地域2       P3T
日本   ：＝北海道｜東北｜関東｜中部｜関西｜中国｜四国｜九州  地域3       P3T
東北   ：＝青森｜岩手｜宮城｜…                               地域4       P3T
関西   ：＝大阪｜京都｜奈良｜兵庫｜…                         地域4       P3T
九州   ：＝福岡｜長崎｜…                                     地域4       P3T

性別   ：＝男                                               性別1       P3A
         …
日本   ：＝関西                                             地域3       P3A
関西   ：＝大阪                                             地域4       P3A
大阪   ：＝吹田                                             地域5       P3A
```

図2　生成規則例



(a)　地域カテゴリ階層　　(b)　産業カテゴリ階層

図3　カテゴリ階層別

(2) 第2レベルの処理

第二レベルは概念スキーマフリー原則に基づいて行う。

各種調査統計データベースを統合する。即ち，各個表に基づく統計表群を第一レベルで個々のデータベースとして構成し，これらを統合することにより行う。

(i) 各データベース間で，スキーマ中に現れる仕様群中で，同一の仕様または等価な仕様を捜しこれを統合する。スキーマ・フリーシステムでは，データとともにスキーマ

を投入する際この操作を行うので，後で行うのではない。これは属性項目の統合である。

(ii)フイールドの統合が(i) ステップで行なえると，次は生成規則レベルでの統合を行う。共通属性を沢山含む生成規則を統合して一つにまとめれないか調べ，まとめれるものをまとめ，生成規則でそれを定義する。

[III] 統計データベースシステム構成法

本システムは次のような構成をしている。

(1) 概念スキーマフリーである。

(2) 投入データは詳細な外部スキーマと共に入力される。各属性を記述するための仕様，仕様の組とそのスキーマに関する情報を記述したスキーマがデータと共に入力されＤＤ／Ｄに記録・管理される。

(3) 概念スキーマはその時点迄に投入された外部スキーマの統合されたものとして陰的（Implicit）に決定される。

(4) 概念スキーマは多次元位相空間として定義される。属性と値の変換可能なものは次元を上げた形で定義される。

(5) 内部スキーマはアクセスの高速化のため，逆に低次元に圧縮する。

(6) データ統合のためＤＤ／Ｄを完備する。これはメタＤＢでもある。

(7) ファイルの重ね合わせのためにもＤＤ／Ｄを完備する。

(8) 位相空間モデルに従い，類似レコードは自動的に近傍に集まる機構を持たす。これによりアクセスの高速化が期待出来る。

[IV] 統計データベース用ファイル構成法

1．ファイル構成

概念スキーマフリーＤＢＭＳは次のようなファイル構成している。全体のファイル構成を図４に示す。

(1) 仕様ファイル；各属性のメタ情報を記録する。この内容により，異なるファイル中にあっても，同一個体を記述したものを見出して統合することが可能となる。

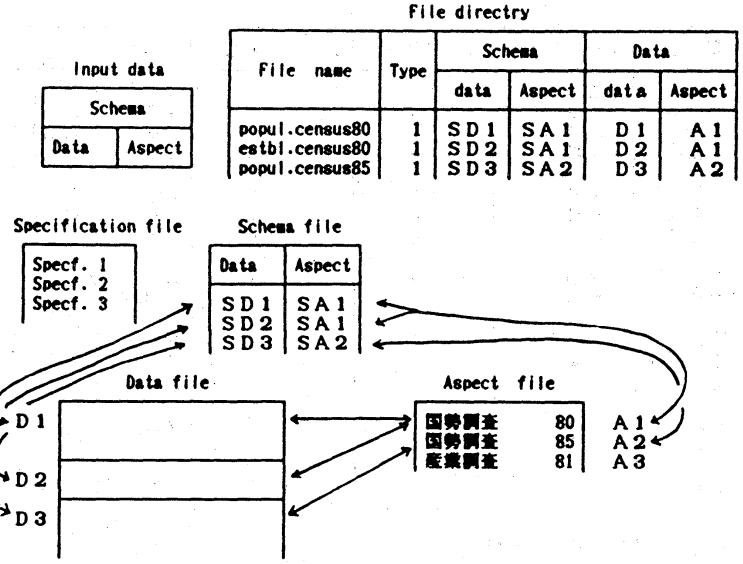


図4　スキーマ・フリー統計データベースシステムのファイル構成概略図

(2) 視点ファイル；データが採取された環境，条件を記録する。

(3) 外部スキーマファイル；同じ分類基準で取られた一つの統計書に相当する。これには記述に使われている仕様組や統合の為のメタ情報が記録される。

(4) ディレクトリ；各ファイルの所在を記録する。

(5) データファイル；生の統計データ，あるいはそれから抽象されたデータが記録される。多次元クラスター分割ファイル構成をとり，類似したデータは自動的に近傍に格納される。また共通属性が多いファイルは共有化あるいは近傍に配置される。

このファイルとしては，次の方式の中からふさわしいものを選ぶ。

(a) Clustered Indexed File

(b) 静的分割ファイル

(c) 適応分割ファイル

このファイル構成については以下に概説する。詳しくは次の発表参照のこと。

(6) データディクショナリ；フィールドの明細は仕様ファイルに記述され，このファイルにはファイルレベルの情報や，利用者，利用プログラムとの関連情報を記録する。

(7) 属性カテゴリ転置索引；属性名から，それが統計表のどのカテゴリの何番目の階層かを求める為の転置索引である。またその座標軸のどの辺りなのかを表した転置索引である。図5にその例を示す。

属性名	カテゴリ名	カテゴリID	カテゴリ階層	位相
青森市	地域	4	5	2次元 1.1.2.1.1
青森県	地域	4	4	2次元 1.1.2.1
秋田市	地域	4	5	2次元 1.1.2.2.1
秋田県	地域	4	4	2次元 1.1.2.2
…	…	..	..	…
運輸通信	産業	6	2	集合 3.3
大阪府	地域	4	4	2次元 1.1.6.1
大阪市	地域	4	5	2次元 1.1.6.1.1
男	性別	2	1	順序 1
女	性別	2	1	順序 2
関東	地域	4	2	2次元 1.1.3
金融保険	産業	6	2	集合 3.1
建設業	産業	6	2	集合 2.1
鉱業	産業	6	2	集合 1.2
公務	産業	6	2	集合 3.6
日本	地域	4	2	2次元 1.1

図5　属性カテゴリ転置索引

(8) 統計表抽象度表；データベースに蓄えている統計表が，夫々の属性がどの抽象化レベルで集計されたものかを表す表である。図6にその例を示す。

Table	カテゴリ 1	カテゴリ 2	カテゴリ 3	…	カテゴリ n
表 1	2	3	0		1
表 2	3	2	4		3
表 3	0	4	2	…	2
表 4	3	3	4		0

図6　統計表抽象度表

(9) バケット分割転置索引ファイル；データを各属性の属性値から高速に捜す為の索引である。バケット分割転置索引ファイルの例を図7に示す。

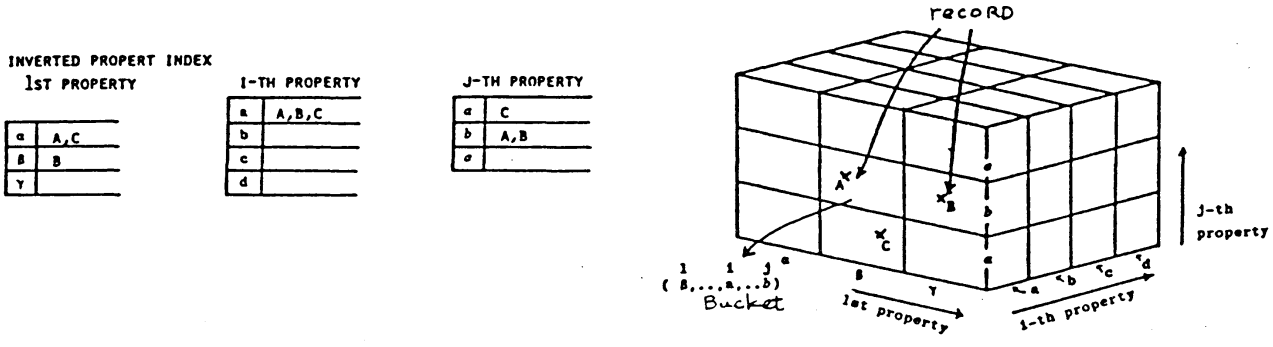


図7　バケット分割転置索引ファイル

(10)統合用基準軸データファイル；統合する際の標準となるデータで，例えば，年度，緯度，経度のメッシュなどがある。

２．統計データベース用ファイル構成法

統計データの検索処理に於いては，統計表の縦軸あるいは横軸に沿っての集計を行いたいということが多い。このような検索に適したファイル構成法に一連性検索ファイルと呼ばれる構成法がある。これは一度ディスクをアクセスして欲しいデータを読むと，同じ条件で検索されるデータがまとまって入っており，あちこちと分散したデータを捜しながら処理する必要のないファイル構成法である。一連検索特性を持つ，統計用データベースに向く新しいファイル構成法を述べる。　なおこのファイル構成法は，統計データのみならず，一連性検索特性を要求される応用に対して同様に有効である。

なおここで議論するのは，前節の(4) のデータファイルの持ち方に関してである。

その設計方針は，(1) データの大域的および局所的クラスタリングを行う，(2) 検索に対するアクセス域を限定（前処理として，射影，選択を行ってしまい，結果をファイル化）する，(3) データ圧縮を行う，にある。

クラスタリング法に単純分割法，静的分割法と適応分割法の３種類がある。

A）単純分割法

空間の各座標軸を幾つかの区間に分割し，その直積として，生成文法で生成された全体空間を超立方体に分割する方法である。各超立方体をバケットと呼び，この各バケットに入ったデータ群をまとめて格納する。レコード内のフィールドの分割格納は行わない。

バケットへのアクセスを高速にする為に，属性値ーバケット転置索引を準備する。

このファイル構成をBucket resolved clustered file with inverted index と呼ぶ。

B）静的分割法　　属性数（位相空間の次元数：カテゴリ数）をｎとするとき，

各属性A_i を有限のk_i 個に分割する。位相空間は $k_1 \times k_2 \times k_3 \times \cdots \times k_n$ 個の直積部分空間に分割される。データはこの各部分空間毎にクラスタ化される。

ファイルは次の様に分けて持つ。

(1) 項目ファイル：項目についての記述および部分領域ファイルへのポインタを持つ。

仕様ファイルの拡張となっている。

(2) 部分領域索引ファイル：属性A_i の属性値がどの範囲のものがどのブロックに格納さ

れているかを示す索引ファイルである。属性カテゴリ転置索引の拡張となっており，

（属性名，属性領域名）→ブロック番号　なる写像を行う。

(3) ブロックファイル：ブロック項目値ファイルの位置を示すポインタファイルである。

（属性名，ブロック番号）→指定属性ブロック項目値ファイルアドレス　写像を行う。

(4) ブロック項目値ファイル：ブロックに対応するレコードの各１つの属性に関する属性値をＩＤ順に持つファイルである。

結局この方法では，一つのレコードは，１つのブロック項目値ファイルに格納されるのではなく，各属性のブロック項目値ファイルにまたがって入る。そのため，１つのレコードの全属性をアクセスするためには，属性数個のブロック項目値ファイルにアクセスしなければならない。しかしその代わり或る特定属性の値のみに興味がある場合，うまくゆけば一つのブロック項目値ファイルのアクセスですんでしまう。　データ圧縮に関しては，１つのブロックの中に入る値がほぼ近いものばかりなので，属性値とその個数の組で記述し，圧縮する方法，ずれ（これは有効桁がずっと少ない）で記述する方法などをとる。

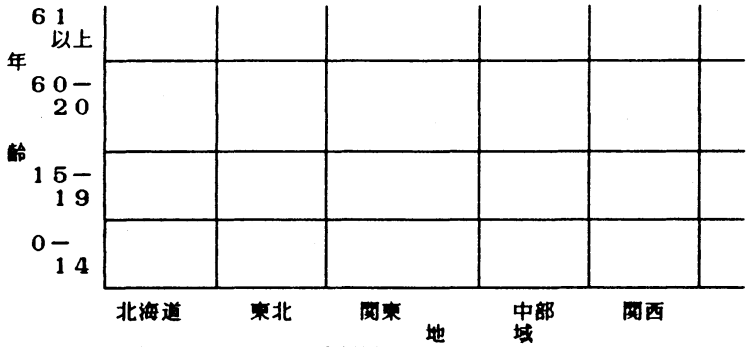
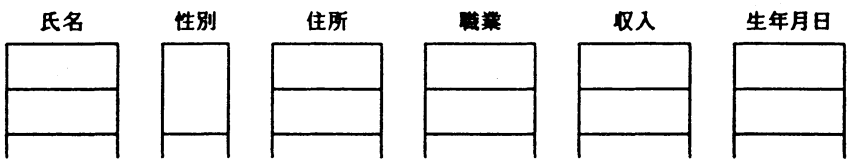


図８　ブロック分割例

氏名	性別	住所	職業	収入	生年月日

(a)　元のファイル



(b)　各属性毎に分割したファイル
各箱が１回のＤＡＳＤのアクセス単位となる。

図９　ブロック項目値ファイル

統計データの一連性検索向きと提案されているトランスポーズドファイルと比較すると

(a) 条件項目が存在すると，トランスポーズファイルより効率の良い分割法が存在する。

(b) 条件項目の数が増加するに従いアクセスすべきブロック数が少なくてすむ。

(c) 分割数を多くしすぎると，多くのブロックを必要とし，記憶域の利用率の低下を引き起こす。

(d) データの属性値の分布にむらがあると，均等な分割では各部分空間に入る個数の分散が大きくなり効率の低下を引き起こす。

ことがいえる。

一つの調査から得られた統計報告書を入手し，これをもとにデータベース化をはかるときは，一括設計が可能なので，この方法を用いるのが有効であろう。

C）適応型分割法　　静的分割法では，各座標軸（属性）毎でデータが均等になるように分割するが，部分空間は各座標軸の直積で定義されるため，その中にはいるデータの個数は必ずしも均等とはならない。この個数が均等なほうが効率が良くなるので，各部分空間に入るデータ数が大体等しくなるように，隣接するブロックをまとめる方法である。なおまとめるに際して凹多角形となるようなまとめ方はしない。

次々と統計表を収集しながらデータベース化してゆくときは，この方法が適している。

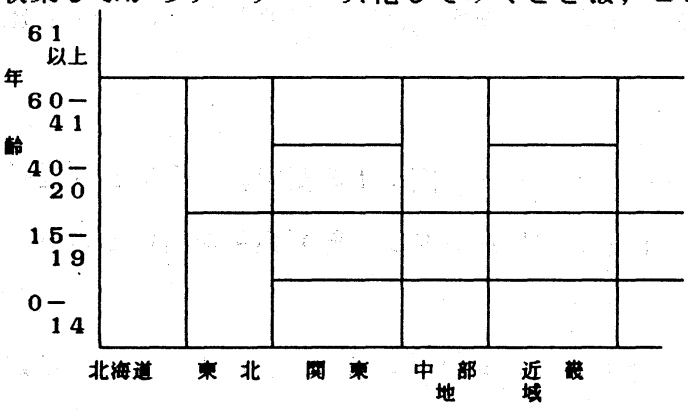


図10　　適応型ブロック分割例

［V］統計データベースシステムの利用

1．統計用ＤＢＭＳの構築

昨年開発してきた，概念スキーマフリーＤＢＭＳを拡張し上述の機能を備えたものに設計を拡張し，核部分を試作している。データのファイリングは次のように行う。もし個票

データが入手可能なら，これを一次データとして持ち，よく使う抽象化レベルで集計をとった二次ファイルを持つ。個票データがプライバシの関係で持てないときは，これから集計された各種の表を管理することになる。これらの表は概念スキーマに対し，各種属性毎の集計レベルで求め，そのレベルの組でその抽象度を符号として付ける。各表はこの符号表によりアクセスされる。図４に示す人口統計に関しては，（時期，性別，年齢，地域，統計種別）の５つ組で表示され，例えば，（２，１，２，４，１）は，年単位の統計で，性別，年齢は１才きざみ，地域は都道府県別の統計表を表す。

２．外部スキーマ管理

データの検索　　利用者の要求する統計表のフォーマットを呈示すると表に含まれる各属性名から逆引索引を引いて，どの属性の第何レベルかを，夫々について求める。要求する表の属性名が既定義の統計表生成文法に含まれない時は，それを含むような形に生成規則を追加することにより，目的とする抽象化レベルを作り出すことも出来る。要求表のレベル表現が求まると，ＤＤに持つファイルの符号表を検索し，要求組と同一のものがあればそれが答えとなり，無い場合は，各項目のレベルが等しいかそれより大きい組を捜し，そのうち要求組に最も近い組を求め，その表をもとに集計することにより要求表を作成する。値一属性変換があつてもこの操作により，標準形ともいうべき概念スキーマを仲介として処理出来る。

その各ステップは次のようになる。図１１の要求表を例にとり説明する。

(1) 表中の各属性に対して，属性ーカテゴリ転置索引を引いて，カテゴリ名とそのレベルを求める。

例えば，『鉱業』はカテゴリ名が『産業』，レベルが２，『青森』はカテゴリ名が『地域』，レベルが４である。

サマリ属性に関しては《》で囲み，カテゴリ属性に関しては＜＞で囲む。

(2) もし該当カテゴリが見当たらなければ，シソーラスをひいてもう一度これを行う。それでもみつからなければ，導出出来ない。

(3) 表は生成規則に書き換えられる。

各カテゴリとそのレベルは　＜カテゴリ名：レベル＞　のように書かれる。もし総計も含まれているなら　＜Ｔ　＋　カテゴリ名：レベル＞　と書かれる。

生成規則の左辺は表の名前となる。

右辺は構成軸の直積として定義される。もし表が単一の表でなく，幾つかの表を一つにまとめたものの時は，素な表の連結和として表す。表が全体の一部を抽出したものの時は直和として表す。

要求表：＝＜1940年から1985年迄５年刻み　＃　地域　４レベル　＞

・＜Ｔ　＋　産業　２レベル＞・《事業所　＃　従業員》

となる。

(4) もし右辺が複合表の場合は，これを展開する。

例では，Ａ＝＜1940年から1985年迄５年刻み＞，Ｂ＝＜地域　４レベル＞，

Ｃ＝＜Ｔ　＋　産業　２レベル＞，Ｄ＝《事業所》，Ｅ＝《従業員》，Ｚ＝要求表

とすると，

Ｚ：＝（Ａ　＃　Ｂ）・Ｃ・（Ｄ　＃　Ｅ）

＝ＡＣＤ　＃　ＡＣＥ　＃　ＢＣＤ　＃　ＢＣＥ

となる。

(5) 各表を属性組表現に変換する。

例では，属性組　Ｔｃを

Ｔｃ＝＜時刻，産業，年齢，地域，種別＞

としたとき，

ＢＣＥ　は　（１：［１９８５］，２，０，４，［１］）

ＡＣＤ　は　（１：［１９４０，１９８５，５］，２，０，０，［２］）

となる。

(6) 夫々の属性組表現について，統計表抽象化レベル索引を引き，どの統計表から算出すればよいかを求める。もし複数通の候補があがった時は，最も距離の短いものを選ぶ。

(7) もし要求表が複合表である場合は複数の統計表から算出することになるが冗長な表は落とす。

	総数		農林水産		鉱業		建設業		製造業		…		.. ..
	事業所数	従業員数	事	従	事	従	事	従	事	従	事	従	
1940													
1945													
1950													
…													
1985													
北海道													
青森													
岩手													
宮城													
秋田													
…													

図１１　要求表

表側　年度　→　カテゴリ　時期
北海道→　カテゴリ　地域　レベル　３　or　４　他より４と判定
青森　→　地域　〃　４
岩手　→　地域　〃　４

表頭　総数　カテゴリ（産業）レベル　１
農林水産　産業　レベル　２
鉱業　〃　〃　２
建設業　〃　〃　２
製造業　〃　〃　２

要求表：＝＜５年毎＃都道府県＞・＜総計＃産業第２レベル＞・＜事業所＃従業員＞
地域第４レベル

３．データのアクセス

利用者の要求より，上述の方法で，属性（カテゴリ）転置索引を引きながら要求表を生成規則に書きなおす。次に要求表を算出するのに必要な，属性は何と何かを列挙する。もし要求データの範囲に制限があるなら，部分領域索引ファイルを引き，どのブロックをアクセスすればよいかを知る。

単純分割法の場合は，属性ーバケット転置ファイルを経由してデータをアクセスする。

静的分割法あるいは，適応分割法の場合は，ブロックファイルを経由して，必要な属性に関する，該当する属性範囲のデータが入っているブロック項目値ファイル群をアクセスする。その際，必要属性の各ブロック（例えば，住所ファイルと生年月日ファイル）を同時に読み出し，前からｉ番目の住所と年齢を調べ，これをもとに第ｉレコードの内容を知り，集計を行ってゆく。この場合だと，住所ファイルと年齢ファイルしか読まないので，不用な属性値を読み込まず，効率の良いアクセスが出来る。

何れの方法に於いても，クラスタ化されているので，各属性に関しても，必要な属性値の範囲の入っている所しか読み込まないので，この観点からも効率の良いアクセスとなっている。

４．おわりに

概念スキーマを個票をもとに生成規則で定義し，各集計段階をレベルの組で記録することにより，種々バラエティのある統計表を統一的に管理することが可能となった。また一連性検索に適した新しいファイル構成法を述べた。

文献

１．打浪，手塚：”統計用ＤＢＭＳの構成に関する一考察”
電子通信学会総合全国大会　５９年度

２．近藤，打浪，手塚：”学術情報，統計用データベースシステムの構成に関する一考察”，情報処理学会全国大会，１９８４．

３．打浪，手塚：”統計ＤＢの概念スキーマとViewサポートについて”，電子通信学会総合全国大会　６０年度

４．打浪：”統計データベースシステムにおける概念スキーマとビューサポートについて”，情報システム研究会資料，１９８５．

５．近藤，打浪，手塚：”統計データベースにおける一連検索様ファイル編成法について”，電子通信学会オートマトンと言語研究会資料，１９８６．２．

６．総務庁：”統計年鑑”１９８５．

７．Chan P.,A.Shoshani：”SUBJECT:A Directry driven system for Organizing and Accessing Large Statistical Databases”, 7th VLDB, １９８１．