実世界データベース（２）

－生成・概念・表記法－

2J－3

池田　一夫、倉科　周介、大橋　誠
東京都立衛生研究所

１．はじめに

国や地方自治体では、社会・経済の現状把握や政策立案の基礎資料として、各種統計データを収集している。1988年には、統計データ収録管理用データベース（ＤＢ）の設計開発手法が佐藤[1)]により提案されている。

統計データのような一次知識（primary knowledge：外界に存在する対象を直接観測することによって得られる知識）は、調査対象および調査目的・調査結果の編集法や記録法などによって規定される。特に時系列データのように広範な展開規模を持つデータの場合、個別調査はその時点の断片的な情報を与えるにとどまる。したがって、各調査データを総合的に検討することにより得られる知識は、現状の方法に依存する限り生産することができない。

本論文では、この制約を克服するための二次知識（secondary knowledge）の生成法、その結果生成する実世界ＤＢ（real world detabase）およびその表記法について論じる。

２．二次知識の生成法と実世界ＤＢの概念

２－１．一次知識の蒐集

対象とする事物に関する一次知識を網羅的に蒐集する。蒐集した一次知識（主としてデータ）は、記録の原型を保ったまま蓄積する（primary data file）。この段階では、内容の修正などは原則として一切行わない。

２－２．仮想空間の設定

実世界の時空間と同型の座標系を設定する。実世界の事物はすべてこの座標系で固有の位置を占める（real world framework）。さらにこの座標系を基準にして、事物の状態を特徴づける座標系を組み込む。人口現象や疾病現象を対象とする場合は、生物学的座標軸（性別、年代、年齢、疾病類別など）、社会的座標軸（職業、組織、その他各種の社会的条件）等をデータの特性に応じて適切に付加する。

２－３．データの再配置

primary file のデータの正誤を照合調整して、運用のためのデータファイル（revised file）を作成する。これから個別データを抽出して real world framework のしかるべき位置に配置する（real world data file）。この段階では、データの記述単位は一次知識のそれと一致する。

２－４．実世界の復元

real world data fileのデータ構造をもとに、個々のデータを分解・再配置し、データの欠損部分は近傍データとの連続性およびデータ全体の中での整合性の両面に配慮しつつ推計補填して、過去の実世界の全体像を復元する（real world database）。復元に用いるデータは一次データではなく、それに基づいた人工的なもの（artificial data）である。

２－５．実世界の観測

復元された実世界データを基準に、一次知識の生成の場合と同様な観測を行う。すなわち、人工的に構築した世界から実世界を間接観測する。これを二次観測、または知識観測と呼ぶことにする。

Real World Database (2) - Creation, Concept, Notation -
Kazuo IKEDA, Shiusuke KURASHINA, Makoto OHASHI
TOKYO METROPOLITAN RESEARCH LABORATORY OF PUBLIC HEALTH

3．表記法

実世界ＤＢが取り扱うのは統計データである。統計データの基礎となるのは個々のデータそれ自身である。そこで、個々のデータの表記法を基に各種概念の表記法を示す。

ｎ次元空間の各座標軸の名前をA_1，A_2，…… A_nとし、各座標軸の任意の要素をa_1，a_2，…… a_nと表記する。上位カテゴリと下位カテゴリ（都道府県と市町村など）を区別するために、座標軸および要素にレベルの概念を取入れ、それぞれ$A_n(L_n)$，$a_n(L_n)$と表記する。カッコ内の値はレベルを示し、標準状態ではL_nの値は０である。表記を簡略化するために、誤解が生じない限り（０）の表記は省略できるものとする。レベルが下がるほどL_nの値は小さくなり。標準状態より下のレベルでは負の値を持つものとする。

個々のデータはｄ（小文字）、データの集合はＤ（大文字）で表記する。

3－1．一般型

ｎ次元空間の任意の位置における任意のデータの値を

$d_{\mathbf{a}}$　　$\mathbf{a}$：ベクトル［a_1 a_2 …… a_n］　　と表記する。

誤解を生じない限り $\mathbf{a}$ を $a_1a_2\cdots\cdots a_n$ と表記してもよい。

3－2．primary data file のデータおよび座標軸

一次知識であることを表示するための記号 0 を導入し、

データ：　$d^0_{\mathbf{a}}$　　座標軸：　A^0_n　　と表記する。

3－3．real world data fileのデータ

primary data を加工した人工的なものであることを明記する必要がある。real world data を使用する際には、そのデータがどの様な過程を経て生成されたものなのかを式で明示する。形式は次のとおりとする。

座標軸　$A_n \leftarrow A_n^0(L_n)$

$[A_1\ A_2\ \cdots\cdots\ A_n] \leftarrow [A_1^0(L_1)\ \cdots\cdots\ A_n^0(L_n)]$

データは、一般形式で表記する。

3－4．イメージデータ

特定の座標軸A_i，A_k，……A_m上の特定の値a_i，a_k，……a_mにおけるデータのプロフィルを次のように表記する。

$D|_{\mathbf{a}}$　　$\mathbf{a}$：ベクトル［０ ０ ‥ a_i ‥ a_k ‥ a_m ‥ ０ ０］

誤解を生じない限り $\mathbf{a}$ を００‥a_i‥a_k‥a_m‥００または単に $a_ia_k\cdots a_m$ と表記してもよい。

4．表記例

時刻　T　［t_1　t_2　……　t_n］　　世代　G　［g_1　g_2　……　g_n］

空間　S　［s_1　s_2　……　s_n］　　要素　E　［e_1　e_2　……　e_n］

として、各種の表記例を示す（［］内は各々の要素）。

●一般的なデータ：　d_{tsge}　　（t,s,g,e は要素の代表）

●特定時刻・特定空間・特定要素の世代構成：　$D_{|tse}$

●市区町村データを利用して都道府県ごとの合計を計算する。

$$d_{ts(1)ge} = \Sigma_s\, d_{tsge}$$

●さらに世代別合計をとる。

$$d_{ts(1)g(1)e} = \Sigma_g\, d_{ts(1)ge} = \Sigma_g\Sigma_s\, d_{tsge}$$

［参考文献］

１）佐藤英人著、統計データベースの設計と開発、オーム社、1988