D-020

# 属性の動的追加を考慮したデータキューブの差分構築の一方式

## An Incremental Datacube Construction Scheme Incorporating Dynamic Addition of Attributes

千葉 陽介† 都司 達夫† 樋口 健†
Yosuke Chiba Tatsuo Tsuji Ken Higuchi

## 1. まえがき

データキューブは多次元データのオンライン分析において不可欠な概念である．一方で，業務の拡大や多様化に応じて，システム運用時に新たに必要な属性が追加されるなどのデータベース定義の動的な変更を必要とする機会が増えている．しかし，属性追加に対応するためには，一般にデータキューブの再構築を行う必要があり，多大なコストを伴う．本研究室では，経歴・パターン法と呼ぶ多次元データセットのエンコード方式 [1]を提案している．本研究では， 新たな属性の動的追加を考慮した多次元データに対するデータキューブの差分構築の一方式を提案する．

## 2. 経歴パターン法の概要

経歴・パターン法は拡張可能な論理配列空間の座標値をエンコードする方法であり，任意の次元の座標を経歴値とパターンの組で表現する(図 1)．論理配列の拡張では，拡張直前の配列と同じ大きさの部分配列が拡張次元方向に付加される．拡張した順序を部分配列の経歴値と呼ぶ．境界ベクトルは座標の各次元添字の表現に必要なビット数である．座標はそれが属する部分配列の経歴値 h と境界ベクトルを用いて，各次元添字をビット列として結合したビットパターン p の対<h，p>としてエンコードされる．p のビット長は h の値と等しくなるという性質がある．添字のビットサイズは固定長ではなく，配列拡張に応じて漸増する可変長であり，コンパクトに実装でき，また，値の照合はシフトとマスク命令のみで行われ，高速な検索を保証し得る[1]．任意のデータ型の属性値のタプル集合を扱うために，属性値を添字に変換するデータ構造(例えば $B^+$木)および，添字から対応する属性値に変換するデータ構造(1 次元配列)を次元毎に必要とする．

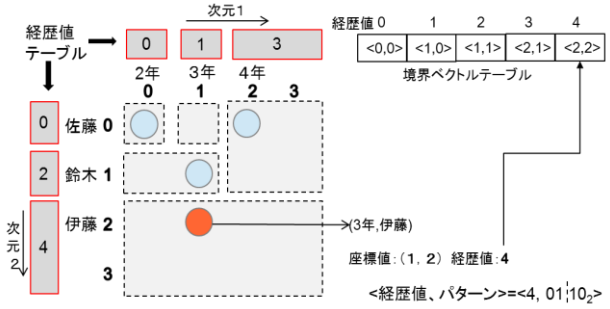



図 1. 経歴パターン法によるエンコード例

## 3. 属性の動的追加[2]

経歴パターン法における新たな属性の追加は，拡張可能配列の次元を動的に 1 つ増加させることで対応できる．n 次元拡張可能配列では，新たな属性の追加要求後，実際に n+1 次元に新たな属性値が登録された際に次元拡張が起こる(図 2)．このとき，次元拡張前の n 次元タプルの拡張次元 n+1 の属性値は NULL 値に対応させる．図 2 ではこの NULL 値は n+1 次元添字 0 に対応させている．これにより，従来のエンコード/デコードのアルゴリズムは，ほぼそのまま利用できる．境界ベクトルテーブルには配列の次元数の情報が増設され，これによりタプルのデコード時に境界ベクトルの次元数を知ることができる．

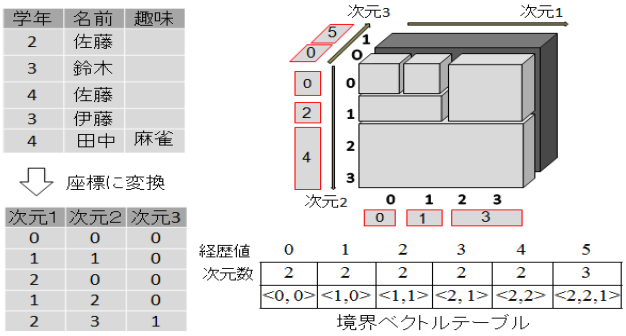



図 2. 図 1 に対する属性の追加と次元拡張

動的追加された新たな属性に対して既存のタプルに NULL 値を対応させる上述の場合以外にも既存タプルには NULL 値属性を含み得る．つづいて，このような NULL 値属性に対応するためのタプルのエンコード法を示す．タプル中に各属性値が NULL 値であればビット 0，既定義であればビット 1 として，ビット列 B を構成する．B を当該タプルの未定義パターンと呼ぶ．図 3 の 3 番目のタプルの場合，B=1101 となる．

タプルのエンコードは n 次元タプル t の 1 次元目に次元パターンを挿入した n+1 次元タプル t'に対して行われる．境界ベクトルを用いて t'内の未定義パターン B を求め，既定義の次元のみをエンコードし，<h,p> の対を得る．p には未定義属性値分のサイズは含まないので，p のビット長は実際には経歴値 h 以下である．図 3 にこの場合のエンコード例を示す．B は他の属性値と同様にエンコードされるので，その添字の最大は未定義パターンの種類の数 m であり，新たな未定義パターンの出現順に 1~m の値が割当てられるが p 中の B の添字ビットサイズは可変である．

| 未定義パターン | 学年 | 名前 | 趣味 | クラブ |
|---|---|---|---|---|
| 11 | 2 | 佐藤 | | |
| 111 | 3 | 鈴木 | パソコン | |
| 1101 | 4 | 佐藤 | | 陸上部 |

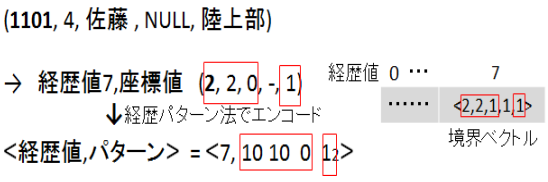
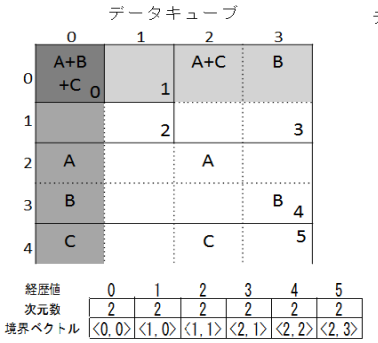



図 3. NULL 値属性を含むタプルのエンコード

†福井大学 工学研究科, Graduate School of Eng., Univ. of Fukui.

## 4. データキューブの構築

T を n 次元多次元データセットとして，集約対象となる T の fact data の属性を含む属性の部分集合からなるタプルデータセットを T のベーステーブルといい，BT とする．BT の属性集合のすべての部分集合に対する BT の射影演算により得られるタプル集合について fact data の集約値を計算した多次元データセットを BT のデータキューブという． ここでは，BT を経歴･パターン法でエンコードすることで拡張可能とし，データキューブの差分構築を実現する．図 4 に{(2,2,A), (3,3,B), (2,4,C)}の 3 タプルを BT とするデータキューブを示す．各次元の添字 0 は集約キューボイドを，また，添字 1 は以後の属性の動的追加に対応するために NULL 値に対応づける．図 4 左図の各部分配列の右下隅の数字は部分配列の経歴値を表す．




データキューブのエンコード

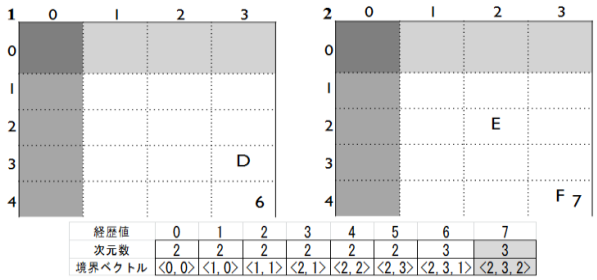
| 経歴値 | パターン | 集約値 |
|---|---|---|
| 4 | 10 10 | A |
| 4 | 11 11 | B |
| 5 | 10 100 | C |
| 0 | | A+B+C |
| 4 | 00 10 | A |
| 4 | 00 11 | B |
| 5 | 00 100 | C |
| 3 | 10 0 | A+C |
| 3 | 11 0 | B |

**図 4. 3 タプルを BT とするデータキューブ**

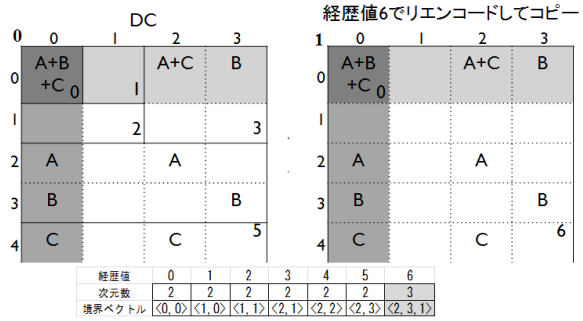
## 5. 属性の動的追加とデータキューブの差分構築

BT による 4. のデータキューブを以後オリジナルデータキューブ（以下 DC）という．DC 構築後に，BT に挿入されるタプル集合(以下ΔBT)について，そのデータキューブ(以下ΔDC)である差分データキューブを構築する．ΔBT を格納中に新たな属性が動的に追加されたとして，ΔBT を格納後ΔDC を構築する．その後，DC をΔDC で集約更新することでデータキューブを差分構築する．図 3 に見るように，一般に，属性の追加時には既存タプルに多くの NULL 値が現れる．3.で示した方法はこのような NULL 値の記憶コストを抑制する．以下では，上記の差分構築のアルゴリズムを図 4 の DC に基づいて，説明する．

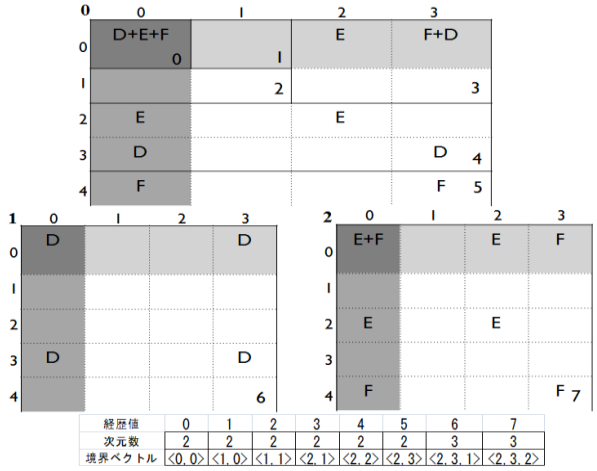
**(a)** 差分のタプル集合ΔBT を格納する．エンコード用の各種データ構造は DC のものを共有することに注意されたい．次の例では，ΔBT={ (3,3,NULL,D), (2,2,2,E), (3,4,2,F)}．図の左上隅の数字は追加次元の添字を表す．



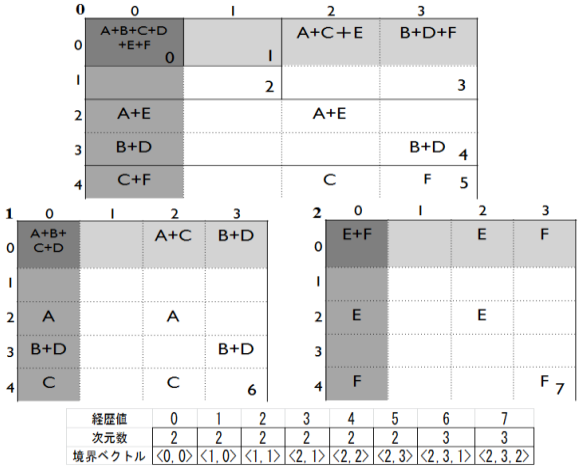

**(b)** ΔBT 格納中に属性が追加された時は DC を追加次元方向に 1 つ拡張する．追加次元の添字を 1（未定義属性値）として，添字 1 の領域（例ではこの領域の経歴値は 6）に DC をリエンコードしてコピーする．




**(c)** ΔBT 格納後，ΔBT について，ΔDC を構築．




**(d)** (b) の DC とそのコピーに(c)のΔDC を集約．この時点でΔBT によるデータキューブの差分構築が完了．




## 6. むすび

経歴･パターン法を使った，データキューブの差分構築方式を提案した．今後，本方式を実装･評価する予定である．

### 参考文献

[1] M.Makino, T.Tsuji, H.Higuchi, ” History-Pattern Implementation for Large-Scale Dynamic Multidimensional Datasets and Its Evaluations”, Proc. of DASFAA (2), 275-291(2015).
[2] 千葉，北嶋，都司，樋口，”属性の動的追加を考慮したタプルデータセットの実装方式 “，情処学会全国大会講演論文集，1N-062015(1),611-612(2015).