関係表の垂直表現を利用した河川情報データの統合

王 毅† 天笠 俊之†,‡ 北川 博之†,‡

†筑波大学大学院システム情報工学研究科 ‡筑波大学計算科学研究センター

1 はじめに

ICT 技術を活用して計算資源，情報資源，人的資源の交流を促進させ，これまでは成しえなかった科学的な発見を目指す「e-サイエンス」が注目されている．これまでは，研究組織や研究者個人に閉じて生成，蓄積されていた科学データが，ネットワークを通じて広く公開され，計算機によって膨大な量のデータを処理することが可能になったことから，科学分野において新しい発見が期待されている．

科学分野の一例として河川情報に目を向けると，我が国はもともと災害を受けやすい地形・気象条件下にあり，毎年各地で台風や集中豪雨による水害や水不足が発生し，深刻な問題となっている．また，河川環境の改善のため，雨量，流量，水質等の水文水質の観測情報を収集することが重要となっている．国土交通省では，全国の主要河川において雨量及び水位，流量，水質等の観測を実施しており，その一部はインターネット上に公開されている [2].

各観測所で観測されたデータは，観測所や各自治体などに独自に蓄積されている．その記録媒体は，古いものは紙媒体，年代を追うに従い，フロッピーディスク，光磁気ディスクなどまちまちである．また，データフォーマット等も異なっており，それらの情報へ統一的なアクセスは実現できていない．図 1 は河川データの一部である．

今後災害の防止や軽減化のため，あるいは人間の社会経済活動が水文循環与える影響を評価するため，過去の水文データへのアクセスの重要性は今後ますます増すものと思われる．本論文では，このような異なるスキーマを持つ科学データの統合のため，関係表の垂直表現 [1] を利用する．関係表の垂直表現は，多数の属性からなる関係タプルを，属性–値ペアに分解して格納する手法のことである．本来は，膨大な属性を持つ e-コマースデータの格納を意図して提案された手法であるが，異なるスキーマを持つ関係データを固定したスキーマで格納できることから，情報統合の一手段として利用する．

Integrating Heterogeneous Water Information using Vertical Representation of Relations

†Graduate School of Systems and Information Engineering
‡Center for Computational Sciences
University of Tsukuba
1-1-1 Tennodai, Tsukuba 305-8573, JAPAN
wang@kde.cs.tsukuba.ac.jp,
{amagasa, kitagawa}@cs.tsukuba.ac.jp

2 問題定義

図 1 に示したような河川データを統合しようとする際，次のような問題がある．

- データによってスキーマが異なる．
- データによってに属性名が異なる．例えば，「堤内地盤高」と「地盤高」，「距離標」と「累加距離」はそれぞれ同じ概念を指す属性であるが，異なる属性名が与えられており，このままでは統合することは難しい．
- 計測データには一般に，ノイズや欠損値を含むことが多い．例えば，前掲の例では，属性「地点名」には大量の欠損値が含まれる．欠損値を大量に含む関係表をそのまま格納することは，格納効率の面から問題である．

この問題を解決するために、関係表の垂直表現手法を利用する．

3 関係表の垂直表現

Agrawal 等は，数百から数千の属性を持つ関係表を効率よく格納，検索するための手法として，関係表の

水平表 H1

Oid	距離標	堤内地盤高	現況堤防高
1	0.2	23.774	26.764
2	0.4	23.887	27.377
3	0.6	24.431	28.691
4	0.8	25.327	29.209

水平表 H2

Oid	地点名	累加距離	地盤高
1	テイナイジバシ	−46.5	23.770
2	⊥	−39.6	23.770
3	⊥	−36.5	21.860
4	⊥	−35	21.660

図 1: 河川データの例.

垂直表 V1

Oid	Key	Val
1	距離標	0.2
1	堤内地盤高	23.774
1	現況堤防高	26.764
2	距離標	0.4
2	堤内地盤高	23.887
2	現況堤防高	27.377
3	距離標	0.6
3	堤内地盤高	24.431
3	現況堤防高	28.691
4	距離標	0.8
4	堤内地盤高	25.327
4	現況堤防高	29.209

垂直表 V2

Oid	Key	Val
1	地点名	テイナイジバン
1	累加距離	−46.5
1	地盤高	23.770
2	累加距離	−39.6
2	地盤高	23.770
3	累加距離	−36.5
3	地盤高	21.860
4	累加距離	−35
4	地盤高	21.660

図 2: 垂直表の例.

垂直表現を提案している [1]. 本節では，本研究の基礎となるこの手法の概要を説明する.

e-コマースなどに代表される応用においては，あるオブジェクトに関連する属性は数百から数千にも及ぶことがある. 既存の関係データベースシステムでは，そのように多くの属性は想定しておらず，特に，属性数はタプルのサイズに直接影響するため，効率的な処理はできない. ただ，このようなデータは多くの場合疎であり，属性のほとんどが空値 (NULL) である場合が多い. この特徴に着目し，この手法では，関係表を，タプル ID (OID)，キー，値の三つ組みに分解し，固定したスキーマに格納する. これを関係表の垂直表現 (vertical representation) と呼ぶ，一方，従来の関係表は水平表現 (horizontal representation) と呼ばれる.

3.1 関係表の垂直・水平表現の変換

通常の関係表 (H) から垂直表現 (V) への変換は，関係代数演算で行うことができる.

$$\begin{aligned}\mho^n(V) \;=\; & [\cup_{i=1}^{k}\pi_{Oid},\text{`}A_i\text{'},A_i(\sigma_{A_i\neq\text{`}\perp\text{'}}(H))]\cup \\ & [\cup_{i=1}^{k}\pi_{Oid},\text{`}A_i\text{'},A_i(\sigma_{\wedge_{i=1}^{k}A_i=\text{`}\perp\text{'}}(H))]\end{aligned}$$

逆に，垂直表現から水平表現への変換は次の式で表わされる. なお，'⊐⋈' は左外部結合を表す.

$$\begin{aligned}\Omega^n(V) \;=\; & [\pi_{Oid}(V)]\sqsupset\bowtie \\ & [\sqsupset\bowtie_{i=1}^{k}\pi_{Oid,Val}(\sigma_{Key=\text{'}A_i\text{'}}(V))]\end{aligned}$$

具体的な例を示すと，図 1 の水平表 H1, H2 を垂直表に変換するとそれぞれ図 2 の垂直表 V1, V2 になる.

4 提案手法

垂直表を用いると，第 2 で指摘した問題点のうち，スキーマの不一致の問題と欠損値の問題は解消される. 任意の関係表は，全て TID, キー，属性値の三つ組に分解されるので，あらゆる関係表を固定したスキーマに格納することができる. また，空値に関しては，垂直表に変換する際に削除される. ただし，属性名については何らかの手立てが必要である.

Tid	Oid	距離標	地盤高	現況堤防高	地点名
1	1	0.2	23.774	26.764	⊥
1	2	0.4	23.887	27.377	⊥
1	3	0.6	24.431	28.691	⊥
1	4	0.8	25.327	29.209	⊥
2	1	−46.5	23.770	⊥	テイナイジバン
2	2	−39.6	23.770	⊥	⊥
2	3	−36.5	21.860	⊥	⊥
2	4	−35	21.660	⊥	⊥

図 3: 関係表の統合例.

本研究では，上で述べた垂直表から水平表への変換処理に，オントロジによるドメイン知識を導入することを提案する. 一般的な概念や用語については，既存のオントロジを利用することが考えられる. 専門性の高いもの，特定の応用に依存する概念・用語については，処理しようとするデータについて，あらかじめ知識を整理して，専用のオントロジを構築しておく. 垂直表から水平表に変換する際，各属性の検索条件を，オントロジによって拡張すれば，類似する概念を統一することができる. 具体的には，次のようになる：

$$\begin{aligned}\Omega^n(V) = & [\pi_{Tid,Oid}(V)]\sqsupset\bowtie \\ & [\sqsupset\bowtie_{i=1}^{k}\pi_{Tid,Oid,Val}(\sigma_{Key\in\{\text{`概念同義の項目'}\}}(V))]\end{aligned}$$

この処理によって，二つの関係表を統合した例が図 3 である. なお，タプルがどの関係表から来たものかわかるように，関係表の ID (TID) が加わっている. この表をビューとして定義しておけば，利用者は統合した表が存在するものとして問合せを行うことができる.

5 まとめ

この論文は異なるスキーマを持つ河川データを統合するために，関係表の垂直表現とオントロジを組み合わせた手法を提案した. 異なるスキーマを持つデータを固定したスキーマで格納するため，関係表の垂直表現を用いるとともに，異なる属性名の統合のために，ドメイン知識を表現したオントロジを利用する. 今後は，提案手法の実装と実データによる評価を行う予定である.

謝辞

本研究の一部は科学研究費補助金萌芽研究 (#18650018) によるものである. ここに記して謝意を表す.

参考文献

[1] Rakesh Agrawal, Amit Somani, and Yirong Xu, Storage and Querying of E-Commerce Data. *Proc. VLDB 2001*, pp. 149–158, 2001.

[2] 国土交通省，水文水質データベース. `http://www1.river.go.jp/`