情報の三元分解と再合成
——次世代データモデルの開発——

遠山　元道

■ 研究のねらい

データベースシステムは、情報のセントラルリポジトリとして、多くの応用に情報を供給する役割を担う。1970 年に E.F. Codd によって開発された関係データモデルは、平坦な表として表わされる関係の集合として必要な情報をすべて表現する。その上で関係代数、関係論理、SQL などのデータ操作言語によってこれらを応用に必要な形に加工することによって、関係の集合に閉じた世界で上記の要求を満たしてきた。

しかしながら WWW のホームページなど、我々が日常的に目にする際の多くの情報は平坦な表とは異なる多彩なレイアウトに基づくプレゼンテーションとなっている。データベースから得られる関係は、必要な情報を全て持ちながら、最終製品としての情報出版物を得るためにはさらに加工が必要である。

本研究では、データベースから得られるホームページ、印刷物、表計算ワークシートなど様々な情報出版物を、「データ」、「構造」、「メディア抽象」の３つの直交する要素に分解できるという発想に基づき、Trinity データモデルを提唱する。データをコンテンツとメディアの２元に分解する見方が一般的だが、Trinity モデルの「構造」はデータに属するか、メディアに属するか認識が曖昧なため、情報やプログラムの再利用を妨げてきた。Trinity モデルはその構成をクリアにすることによって、データの共有、再利用などデータベース本来の役割を大きく拡大するとともに、三要素の合成のために提案する SuperSQL 言語によって情報出版ソフトウェアの開発を著しく容易にする。

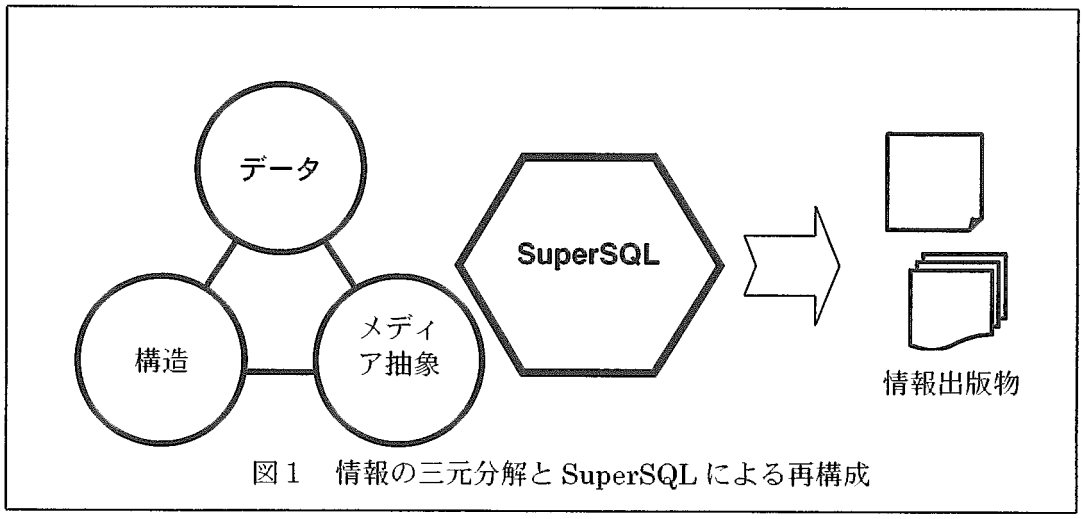

図１　情報の三元分解と SuperSQL による再構成

■ 研究内容

1）モデルの構成

データ： 30年余の実績から、関係データモデルがあらゆるジャンルのデータを過不足なく表現可能であることが証明されていると考えられる。関係データモデルの優れた点は、整備された正規化手続きによって情報の部品化を行い、逆に結合、射影、選択によってこれらを合成したり一部を取り出したりできることにある。Trinityモデルでは、データベースから必要な「材料」を得るのに関係データモデルの処理系をそのまま仮定しているので、全体としては関係データモデルに対して上位互換性を持つ。

構造： あまりに漠然とした命名になるが、ここで言う構造とは出版物における情報のグループ化とレイアウトを指している。関係データモデルではタプルの集合（平坦な表に相当）以上の構造を許さない。このため、1対多の対応関係を持つデータ項目を含む場合には多くの重複が生じる。成績（科目名、学籍番号、評点）の3項関係では、同一科目を10名が履修していればその科目名は表の中に10回繰り返して現れる（図2a）。我々が目にする出版物では、科目名は用紙の情報に一度だけ現れるような表現（図2b）が普通だが、関係データモデルではこのような構造を扱えない。

科目名	学籍番号	評点
数学1	1234	A
数学1	1256	B
数学1	1278	A
数学1	1298	C
英語1	1234	B

図2a 平坦な表

科目名　数学1

学籍番号	評点
1234	A
1256	B
1278	A
1298	C
1234	B

図2b グループ化

このため、SuperSQLでは2項演算子としての結合子と、グルーピングを表わす反復子を使ってグルーピングと項目間のレイアウトを表現する能力を持たせる。これらの演算子を総称してレイアウト演算子と呼んでいる。図2の例では、

[科目名 ！ [学籍番号, 評点]!]!

のようにする。コンマ（ , ）、感嘆符（ ！ ）、パーセント（%）、シャープ（#）がそれぞれ第1次元から第4次元の結合子を表わす。また、角括弧対にこれらの結合子の一つをつなげたものをその次元の反復子としている。平面媒体では2次元で十分だが、第3次元結合子はWWWにおけるハイパーリンク、第4次元は動画の編集に用いる。ターゲットリストにこれらの演算子による拡張を加えたものを、TFE（Target Form Expression）と呼ぶことにする。

メディア抽象： Trinityモデルが従来のデータベースの守備範囲を超えてさまざまな応用メディア

をモデル化の対象とする上で、キーポイントとなるアイディアは第 3 要素のメディア抽象である。メディア抽象とは、構造化したデータ群から特定のメディアの出版物への写像を定義するものである。対象となるメディアにはさまざまな構成子（コンストラクタ）がある。これらを使い分けるために、構造化の段階で次元の区別という概念を導入した。1 次元の接続子は、直感的にはオペランドを左右につなげる構成子に写像する。HTML を対象メディアとする場合を例にとれば、これは〈TABLE〉タグを用いて左右のオペランドを内容とする 1 行 2 列の表を生成する。あまり複雑でない場合には、このようにテンプレートとしてメディア抽象を与えることができる。一方、3 次元バーチャルリアリティ空間の生成をターゲットとするような複雑な場合は、メディア抽象にレイアウトに関する多くの手続きや知識を持たせることになる。

2）SuperSQL

本研究の主題は Trinity モデルを確立し、世の中に存在する多くの情報出版物をその主要三要素に直交分解可能であることを示し、かつそれぞれの要素の再利用や多重利用のメリットを検証することにある。しかし、三元分解しただけでは情報のブラックホールになってしまう。データベースに格納した情報から情報出版物を生成する役割を持つのが SuperSQL 質問文である（図 1）。SuperSQL 質問文は、通常の SQL 質問文の SELECT 句に代わり、対象メディアの指定と構造化を指示する TFE をもつ GENERATE 句を持たせることによって、三要素の統合を指示する。

GENERATE 〈メディア指定〉〈TFE〉
FROM 〈関係の並び〉
WHERE 〈選択、結合条件〉

SuperSQL 質問文の評価は、通常の SQL と同様に FROM 句と WHERE 句から必要な情報を全て得る（Trinity モデルの第 1 要素）。次に、GENERATE 句に与えられた TFE の入れ子の形に応じてタプル集合を入れ子関係に構造化する。最後に、TFE で指示した次元に基づいてメディア抽象の写像を選択し、これによって構造化されたデータを目的のメディアに変換する。

通常の SQL では、WHERE 句の後に、GROUP BY 句、HAVING 句、ORDER BY 句などを付加し、ある種の出力の構造化を行なうことができる。しかし、例えば ORDER BY は 1 通りしか指定できないため、その表現力は SuperSQL において反復子を 1 組だけしか使わないよう制限したものに等しく、極めて貧弱である。

■ 研究成果

1）木構造スキーマの情報容量について

情報を表現する能力という最も根本的な部分について考察を行なった。同一のデータベースから 2 つの SuperSQL 文によって生成される出版物を比較する。このとき、FROM 句と WHERE 句が等しく、かつ GENERATE 句に現れる属性の並びが同一ならば同じ情報内容が結果として得られるだろうか？極端なケースを考えれば自明だが、この等価性は成立しない。前出の例では、どの学生がどの科目でどの成績をとったかがすべて表現されている。しかし、次のような TFE に基づいて得られる結果は何を意味するだろうか？

GENERATE HTML [科目]!, [学籍番号]!, [評点]!

この場合、第1列には科目一覧、第2列には全学生の学籍番号一覧、第三列には評点の種類（A, B, C, D, * など）からなるページが作られる。ところが、これからは学生、科目間の関係は失われ、当然個々の成績も分からない。

ある TFE に反復子を一組加える操作と、その上で冗長な反復子を除去する操作を適用することによって得られる TFE では、後者の出力からの情報量は前者と等しいかもしくはより少なくなることが示せる。この「反復子の導入」操作に基づき、反復子一つだけからなる TFE を最大元、各属性を単独に反復子でくくる TFE を最小限とする束構造が得られることを示すことができる。

最初の設問に対して、TFE の形によって得られる結果の情報量は異なり、同一の物を集めた同値類と、それぞれの間の情報量の包含関係に基づく半順序が定義できることが分かった。この結果の直接的な応用として、帳票や WEB ページのデザインによって、データベースから得られた情報を損失無く表現できる場合、損失を伴って表現する場合、本来データベースに無かった情報があたかも存在するように見せる誇大表現などを厳密に定義し、議論することができるようになる。同様に、木構造であらわされる XML のスキーマの情報容量と、これから生成する Web ページの情報容量との比較なども実用的には重要な話題となる。

2）多メディア生成（XML 他）

SuperSQL の GENERATE 句で指定できるメディアとして、従来から LaTeX, HTML, EXCEL などを試作していた。これを増やすためには対象となるメディアのメディア抽象を与えれば良い。現状ではメディア抽象はプログラミングによって与えている。これらに加え、新たな生成対象として XML、O2（関係データベースからオブジェクト指向データベースへの情報移転の実現）、VRML（三次元仮想空間の生成）などを試作した。従来の SuperSQL 処理系では、生成対象のコンストラクタは固定であり、メディア抽象はこれを生成するコードジェネレータとして開発する構成になっている。一方、データの流通もしくは保存の媒体として、XML はますますその重要性を高めているが、XML の特徴の一つとして利用者が任意のタグ名をもつ XML 文書を自由に作成することができる。このため、SuperSQL によってデータベースから XML 文書を生成するという重要な応用において、出力するタグをどこでどのように指定するかが問題となる。これについて、SuperSQL の質問文中にアドホックに記入する方法と、メディア抽象に組み入れる方法とが考えられる。前者の利点は、使用したいタグの種類が多い場合、TFE の次元数以上に自由に質問文に書くことができることである。一方、メディア抽象に定義すれば質問文は非常にシンプルになる。XML の実現例ともいえる HTML の生成では定義を固定する方式でよい結果が得られている。実際には XML 生成においては、制約を無くすために質問文記入方式のインプリメントを行ない、メディア抽象に定義したものを質問文変換で導入することを試みた。

3）応用指向メディア抽象

従来の SuperSQL では生成対象の汎用性を重視し、HTML, VRML, XML など、汎用のメディアをターゲットとしていた。しかしながら、たとえばデータベースのコンテンツを利用して仮想美術館を生成することを考えてみよう。SuperSQL を用いることにより、出展する絵画の選択や配置を自由にコントロールできるというメリットがある反面、単純な直方体に絵やタイトルを貼り付けて並べただけでは美術館らしい臨場感は得られない。臨場感のためには壁や照明などを念入りにモデリングし、

作りこむ必要がある。この際用いられるメディアはVRMLであったとしても、美術館に特化したものは他の応用に転用することはできない。このように、品質と適用範囲にはトレードオフが存在することは明らかである。

そこで、SuperSQLの生成対象として、実在の世田谷美術館の内装を基本とし、仮想美術館を生成するシステムを試作した。GENERATE句にはメディア指定として、VRMLではなく、Museumと書く。この経験から、SuperSQLの生成対象メディアとして応用により近い、特化したテンプレートを用いることの必要性を、この応用を通じて確認した。

WWWページなど、汎用メディアであるHTMLを表現媒体とする場合でも、例えば特定の新聞社のページなど、高度のデザイン性が求められる場合には、応用寄りにメディア抽象をカスタマイズすることが有効となるが、デザインの変更などへの柔軟性は著しく制約されてしまう。

4）質問分割と動的生成による生成プロセスの最適化

Trinityモデルにおいて情報出版物を三元分解することの大きなメリットの一つとして、それぞれの側面において独立に変換を定義することできる点が挙げられる。質問分割では、一つのSuperSQL質問文を2つ以上の副質問に分解することによって、データベース全体ではなく、利用者が必要とする部分だけを生成対象とすることができる。WWWなどで、分割した質問文をハイパーリンクのクリックに同期させた動的呼び出しに組み合わせれば、対象メディアの動的生成が実現できる。これにより、生成のリフレッシュレートとスループットのバランスを取る上で、極めて抽象度の高いレベルでの設計の最適化を実現することができる。在庫に基づいてWWWで通信販売を行なうサイトの設計を例に取ると、商品カテゴリの目次ページと各商品のページを一つの質問文で一括生成すると極めておおきな負荷となってしまう。目次ページと下位のページの生成を分割し、動的呼び出しをすることにより、目次ページで選択された商品に関する最新の在庫状況を表示するページを、商品単位で生成することができる。このような等価性に基づく設計変換は、プログラム変換などでは実現しえないと考えられる。

■ 今後の展開

この三年間は、基本的なアイディアを中心に置き、データモデルとしての汎用性を確認するためにいわば外向的に考えを発展させてきた。しかし、データモデルは広くその価値が認知され、多くの研究やソフトウェア開発の基礎となってこそその価値がある。今後の方向としては、まずSuperSQLの利便性をアピールするために、たとえばフリーのDBMSとして発展し、その安定性から広く業務にも用いられているPostgreSQLの機能拡張としていままで開発した処理系をコントリビュートすることを考えている。その上で、本研究の主題である情報の三元分解の意義がより多くの利用者、ソフトウェア技術者、研究者に理解されるよう努めようと考えている。

■ 参考文献

- 赤堀正剛、有澤達也、遠山元道、SuperSQLによる関係データベースとXMLデータの統合利用、情報処理学会論文誌 Vol.42, No.SIG8, 2001.7

- Yoko Maeda, Motomichi Toyama, ACTIVIEW: Adaptive Data Presentation Using SuperSQL, Proc. 27th International Conference on Very Large Databases (VLDB2001), 2001.9
- Shiro Udoguchi, Tadashi Iijima, Motomichi Toyama, Application of SuperSQL Query Language for the Migration from a Relational to an OO Database, Proc. International Database Engineering and Applications Symposium (IDEAS2000), IEEE, 207-218, 2000.9
- 鵜戸口志郎、飯島正、遠山元道、データベース出版技術を用いた関係データベースからオブジェクト指向データベースへのデータ移行、情報処理学会 アドバンストデータベースシンポジウム 98, 63-70, 1998.12
- 赤堀正剛、遠山元道、SuperSQLによる印刷の高品質化と対話型レイアウト入力の実現, 電子情報通信学会 データ工学ワークショップ 99
- 多崎央、遠山元道、SuperSQL によるデータマイグレーションツールの拡張、情報処理学会 夏のデータベースワークショップ 1999, DBWS99, 1999.7
- 有澤 達也、遠山 元道、応用データ自動生成のための変換定義言語、情報処理学会 夏のデータベースワークショップ 1999, DBWS99, 1999.7
- 多崎 央、遠山 元道、OODB/ORDB 間のデータマイグレーションシステム、 電子情報通信学会 データ工学ワークショップ 2000, 2000.3
- 高橋悠子、河村嘉之、遠山元道、芸術情報のデジタルアーカイビングにおけるXMLの利用、情報処理学会 夏のデータベースワークショップ 2000, DBWS2000, 2000.7
- 團琢海、遠山元道、データベース出版における物理的レイアウト制約の反映、情報処理学会 夏のデータベースワークショップ 2000, DBWS2000, 2000.7
- 赤堀正剛、田中隆一、遠山元道、SuperSQL による XML データドキュメントの自動生成、情報処理学会 夏のデータベースワークショップ 2000, DBWS2000, 2000.7
- 赤堀正剛、有澤達也、遠山元道、SuperSQL による関係データベースと XML データの統合利用、情報処理学会 アドバンストデータベースシンポジウム 2000, 2000.12
- 有澤達也、遠山元道、SuperSQL 処理系におけるグルーピング操作の効率的な実装、電子情報通信学会 データ工学ワークショップ 2001, 2001.3
- 前田葉子、遠山元道、Actiview:SuperSQL を利用した適応型表示ビューの実現、電子情報通信学会 データ工学ワークショップ 2001, 2001.3
- 多田光伸、遠山元道、SuperSQL によるインタラクティブプレゼンテーションの自動生成、情報処理学会 夏のデータベースワークショップ 2001, DBWS2001, 2001.7
- 高畑理、藤沼健太郎、石橋玲、遠山元道、Magic Mirror Mailing: 個人情報データベースを利用する柔軟なメイル配送システム、情報処理学会 夏のデータベースワークショップ 2001, DBWS2001, 2001.7