

データベースの観点から見た XML の研究

吉川 正俊

奈良先端科学技術大学院大学

yosikawa@is.aist-nara.ac.jp

1 はじめに

XML は、データおよび文書の表現書式であるため、その格納、検索、変換、出版などを効率良く行うためには、これまでのデータベース分野で研究され蓄積されてきた技術の多くを利用することができる。本稿では、データベースの観点から見た XML の研究について著者の考えをまとめる。また、XML とデータベースに関連する最近の研究動向を概観する。

2 データベースモデルの研究の流れ

XML1.0の標準化やその前身の SGML の発展のためにデータベースコミュニティが果たした役割はきわめて小さいと言わざるを得ない。しかし、XML の出現後その重要性はデータベースコミュニティですぐに認識された。現在では、XML はデータベース分野における非常に重要な研究テーマとなっており、国際会議などで多くの研究成果が発表されている。また、ほとんどの商用データベースシステムで XML を扱えるようになってきている。このように XML に対する迅速な取り組みができた背景には、データベース分野におけるこれまでの長年の研究成果の蓄積があったことを挙げることができる。

これまでのデータベースの研究は、データモデルを一つの軸として進展してきた。その流れを著者の視点から簡単にまとめると次のようになる。

1970年代は関係モデルの提案 [1] とそのシステム実装、理論研究 [2] が盛んに行われた。1980年代始めに商用の関係データベース管理システムが現れるに至り、関係データベースの研究は一つの転機を迎えた。

関係データモデルはデータを二次元の表によって表現するという単純さが成功の理由の一つであるが、一方ではデータ構造があまりに単純であるために、実世界の情報をそのまま忠実に反映するためには不十分である。そこで、1970年代後半から 1980年代にかけては、データの持つ豊富な意味を関係モデルよりも正確に表現するために意味データモデル (semantic data model) と呼ばれるデータモデルが数多く提案された [3] (邦訳は [4])。関係表の値として、集合、リスト、関係など複雑な構造を許す非正規関係データモデルに関する研究もこの一環としてとらえることができる。

意味データモデル自身は実際のシステムとして実装されることはなかった。しかし、その研究の流れを一部引き継ぎ、また、複合オブジェクトモデル (complex object model) の研究と相まって 1980年代半ばからオブジェクト指向データベースに関する研究が盛んになった。これは、Smalltalk-80 などのオブジェクト指向プログラミング言語の影響が大きい。1990年代始めにはその研究成果が商用システムの形で結実した。また、データモデルおよび問合せ言語の標準化も継続的に進められている [5]。

1990年代半ばからは WWW が爆発的に普及し始め、WWW 上のデータをデータベースと

みなし，どのようにモデル化し問い合わせするかという問題が重要な研究テーマとして浮上してきた．このようなデータモデルは，WWW上のデータが通常のDBMSで管理されるデータと異なり，データ構造を必要に応じて自由に変更できる柔軟性を持っていることから，半構造データモデル (semi-structured data model)[6]と呼ばれている．XMLの出現前は研究者がそれぞれ独自の半構造データモデルを定義し議論を展開していたが，現在では，その研究の流れはXMLを標準モデルとすることによって引き継がれている．

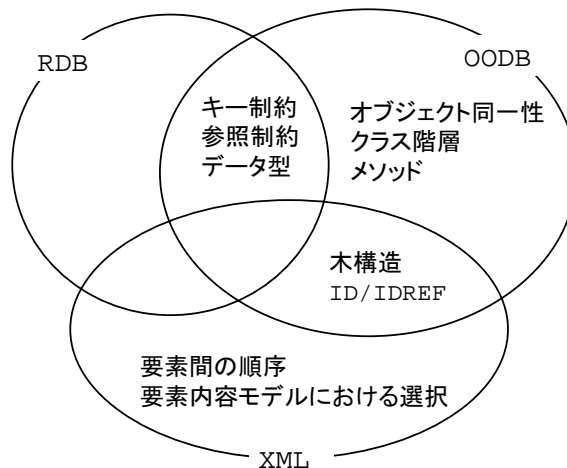


図 1: データモデルの比較

3 データベースモデルとXMLデータモデル

前節で説明したように，データベースの研究の歴史においては様々なデータベースモデルが研究されてきた．研究対象はそれぞれのデータベースモデルの表現能力や問合せ言語など論理的レベルのものから，問合せ最適化，ファイル編成法，並行制御など実装方法に至るまで幅広い．

一方，XML1.0は，このようなデータベースモデルの研究とはほとんど独立に制定された．整形式のXML文書が持つ情報を抽象化して定めているものとしてXML Information Set[7]がある．XML Information Setは，データ構造を木に限定していない．しかし，通常はXML文書のデータ構造は，XPathデータモデルやDOMデータモデルのように木構造でモデル化される．

データベースモデルとXMLのデータモデルの表現能力との比較結果を簡単にまとめたものを図1に示す．データベースモデルとしては現在商用システムが存在する関係データモデルとオブジェクト指向データモデルを対象としている．通常，データベースモデルには，データ構造のみならず（関係モデルにおける関係代数，関係論理などの）問合せモデルも含む．しかし，図1では問合せモデルは考慮していない．

4 XMLデータベースシステム

大量のXML文書を永続的に格納し，検索，更新などの操作を可能とするためのデータベースシステムは，研究用，商用ともに盛んに開発が行われている．この分野はまだ揺籃期であり，多くの可能性が試行されている段階である．

XML文書のデータベースへの格納法は，XML文書を分解しデータベーススキーマに写像する方法とXML文書をそのまま格納する方法の二通りに大別できる．

4.1 XML文書のデータベーススキーマへの写像法

XML文書を分解し，データベーススキーマに写像する法としては次の二つのアプローチを考慮することができる [8] ．

- 構造写像アプローチ

データベーススキーマは，XML文書の論理構造（あるいはDTDが存在する場合はDTD）を表現する．初期の研究では，オブジェクト指向データベースモデルが木構造や参照関係を自然に表現できることを利用

し、基本的に各要素型に対してクラスを定義することにより、SGML 文書をオブジェクト指向データベースで管理する手法(たとえば [9, 10])が提案された。より洗練された手法としては、DTD をグラフで表現し、子要素の数などに応じて適当な部分グラフごとに関係スキーマを生成するもの [11] も提案されている。

構造写像アプローチでは、DTD が存在する場合は DTD ごとに、そうでない場合は、XML 文書ごとにデータベースが定義される。

- モデル写像アプローチ

データベーススキーマは、XML データモデルの構成要素を表現する。このアプローチでは、任意の XML 文書の木構造を格納するために固定したデータベーススキーマを用いる。このアプローチの初期のものとしては、SGML 文書をオブジェクト指向データベースに格納するために、すべてのテキストノードを NODE というクラスで管理する手法 [12] がある。

また、整形式の XML 文書であっても構造を発掘し、それをもとにデータベーススキーマを決定する方法も提案されている [13]。

これら二つのアプローチを比較すると次の様になる。

- 図 1 に示すように、XML のデータモデルには、関係データベースモデルやオブジェクト指向データベースモデルでは表現できない構成要素がある。このことは、これらの構成要素をそのまま自然にデータベーススキーマに写像するような構造写像アプローチは存在しないことを意味する。問題の解決のためには、データベースモデルの表現能力の拡張が必要となる [9, 10]。従って、通常の DBMS は利用できない。

- 構造写像アプローチは、少数の DTD または文書構造に従う大量の XML 文書の管理が必要であり、このような DTD または文書構造が安定している場合には適している。逆に、DTD が不明であるような整形式 XML 文書や DTD が頻繁に更新されるような XML 文書の格納のためにはモデル写像アプローチの方が適している。

4.1.1 関係データベースを用いたモデル写像アプローチ

このアプローチの目的を抽象化すると、任意の順序木を格納する固定した関係データベーススキーマを開発することとなる。文献 [14, 15] では、このアプローチに基づくいくつかの手法を提案し、比較している。これらの手法は、木構造のノードに識別子が付与されていることを仮定している点で制約がある。著者らは、木構造の根から各ノードまでの経路を基本単位とし、経路自身を文字列として関係に格納し、経路とリージョンによって木構造情報を表現する方法 XRel を提案している [8]。関係表は、経路に識別子を与える Path の他に、XML 文書の木構造のノードの種類(要素、XML 属性、テキスト)ごとにそれぞれ Element, Attribute, Text がある。また、XPath の主要な部分から SQL への変換アルゴリズムを開発した。XRel は、格納する XML 文書に何の制約も設けておらず、また、使用する関係データベースにも特に条件を設けていない。性能評価の結果、XPath における子孫ノードを指定する '//' を含む問合せなどを特に高速に処理可能であることを確認している。

4.2 XML 文書をそのまま格納する方法

XML 文書をそのまま格納する方法を採用するデータベースは、ネイティブ XML データベースと呼ばれることもある。関係データベースシ

ステムの場合は，基本的に CLOB (Character Large Object) として格納されることが多い。たとえば，Oracle 9i では，論理的には新たに XML 型が導入されている [16]。XML 型は CLOB データをもとに定義されており，メソッドを用いて検索を行う。このような方法はもともと Waterloo 大学の New OED プロジェクトで開発されたものである [17]。

5 XML データベースのベンチマーク

XML データベースを実現するための多くの手法が提案されているため，それらの性能を定量的に比較するためのベンチマークが開発され始めている。文献 [18] は，XML ベンチマークが備えるべき測定基準を次の様にまとめている。

- (1) Bulk loading
- (2) 再構成 (reconstruction)
Bulk loading の逆。もとの XML 文書の再構成。
- (3) path traversal
冗長性，データ量，断片化の度合いのトレードオフ
- (4) casting
文字列から整数，浮動小数点数，利用者定義型への casting。文字列演算は非常に遅い。
- (5) missing elements
ある種の写像では，多くの NULL 値が生じる。NULL 値をコンパクトに記憶する戦略とは別に，NULL 値(はしばしば興味の対象なので，それ)を問い合わせする効率的な方法も必要。
- (6) ordered access
順序が不要な場合は，それを無視して問合せ処理を高速化するなどの柔軟性が必要。
- (7) references
参照を処理するには，文書に渡ってランダ

ムアクセスを効率的に支援する方法が必要。

- (8) joins
データ中心の応用における，要素や属性の内容に基づく結合
- (9) 大規模な結果の構築
格納されているデータからの大規模な結果の構築
- (10) containment, full-text search

XML 問合せプロセッサの性能評価の枠組みとして XMark [19] が開発されている。このベンチマークは，異なるスケーリングファクタの文書を生成するツールと重要な XML 問合せ処理をカバーする 20 個の XQuery 問い合わせから成る。

このような問合せ処理器以外の面の評価項目としては，次のものがある [18]。

インフラストラクチャ：データベースのフロントエンドとバックエンドが密接に統合されていなければ通信コストがかかり，問合せ処理器の性能が十分に発揮されない。問題点としては以下のものがある。主としてシステムに焦点を当てたベンチマークである XMach-1 [20] では，これらのいくつかを取り上げている。

- アクセスプロトコル ... HTTP, OLE DB, ODMG, ODBC, native APIs など。
- 結果の表示 ... DOM, SAX, 一次元 XML 化など。
- 応答性と完全性 ... 最初のおよび完全な問合せ結果の利用可能性。
- 問合せ言語の表現能力 ... XPath にはない再構成の能力が問合せエンジンが与えられた応用シナリオに適合するかどうかを決める。
- 複数利用者応用シナリオにおけるデータスループット

所有コスト：ソフトウェア自身よりもその所有のコストの方が高価になりつつあるため、次の点が重要である。

- インストール
- スキーマ情報とともに(あるいは、なしに)文書を格納することができるか?
- 入ってきた文書のあるスキーマあるいは他の制約に関して妥当性検査できるか?
- 文書を格納する前にどのような写像およびスキーマ定義が必要か?
- 訓練
- 対話のパラダイム ... 直接対話のためのツールを持つスタンドアロンの文書管理システムか?あるいは、フロントエンド応用のための技術か?
- 細粒度更新機能を提供するか?あるいは文書全体の置き換えに限られているか?

6 XML文書ノードの符号化法

XML文書に対する問合せは、文書構造に基づくものが基本になる。このような問合せを効率良く処理するために、XML文書の木構造におけるノードを符号化する方法が開発されている。それらの方法の中には、問合せのみならず、部分文書単位での更新も考慮したものもある。

文字列としてのXML文書における要素や属性の開始位置と終了位置の対は、レンジ (range) またはリージョン (region) と呼ばれる。レンジは最も単純なノード符号化法である。二つのレンジを比較することにより、それらの先祖-子孫関係や先行-後続関係がわかる。しかし、親子関係を確認するためには二つのレンジの間に他のレンジが存在しないことを確かめる必要がある。

レンジは、単純ではあるが更新に対して弱い。そこで、それを改良する手法が提案されている。著者らは、文書の先頭を起点とする絶対的なレンジではなく、親ノードの先頭位置を起点とする相対的なレンジによってノードを表現する手法を提案した [21]。実際のファイルシステムでは、物理ページに収容できる範囲の部分木における根を起点とすることになる。また、プリオーダ (preorder) と子孫の範囲の対でノードを表現する方法も提案されている [22]。このとき、プリオーダは定義通りではなく、間隔を空けた番号を付与することにより将来のノード挿入に備える。

これらの他にも、XMLの問合せ高速処理、バージョン管理などを目的とするXML木構造の表現法が提案されている [23] [24] [25]。

7 おわりに

2節で説明したように、データベースモデルの長い研究の過程で構造化文書が意識され始めたのは比較的最近のことである。一方、XMLの開発当初は、あくまで構造化「文書」が想定されており、データベースで管理される通常のデータを表現することは考えられていなかった。データベースと文書処理はもともと別々の分野として研究が進められたきたが、XMLを触媒とし今後は融合して発展していくことになるだろう。

XMLの応用の拡散により、XMLデータベースは広範な機器に搭載される基盤システムとなることが予想される。

参考文献

- [1] E. F. Codd. "A Relational Model of Data for Large Shared Data Banks". *Comm. of the ACM*, Vol. 13, No. 6, pp. 377-387, June 1970. Reprinted in M. Stonebraker, *Readings in Database Sys.*, Morgan Kaufmann, San Mateo, CA, 1988.

- [2] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
- [3] R. Hull and R. King. “Semantic Database Modeling: Survey, Applications, and Research Issues”. Vol. 19, No. 3, pp. 201–260, September 1987.
- [4] Richard Hull and Roger King. “意味データベースモデリング：サーベイ、応用、研究課題”. 田中克己 (訳), コンピュータ・サイエンス, bit 別冊, pp. 117–164. 共立出版, 1989 年.
- [5] Rick G. G. Cattell, Douglas K. Barry, Mark Berler, Jeff Eastman, David Jordan, Craig Russell, Olaf Schadow, Torsten Stanienda, and Fernando Velez, editors. *The Object Database Standard: ODMG3.0*. Morgan Kaufmann, 2000.
- [6] 田島敬史. “半構造データのためのデータモデルと操作言語”. 情報処理学会論文誌：データベース, Vol. 40, No. SIG3(TOD1), pp. 152–170, 1998 年 2 月.
- [7] World Wide Web Consortium. “XML Information Set”. <http://www.w3.org/TR/xml-infoset/>, October 2001. W3C Recommendation 24 October 2001.
- [8] Masatoshi Yoshikawa, Toshiyuki Amagasa, Takeyuki Shimura, and Shunsuke Uemura. “XRel: A Path-Based Approach to Storage and Retrieval of XML Documents using Relational Databases”. *ACM Transactions on Internet Technology*, Vol. 1, No. 1, pp. 110–141, August 2001.
- [9] Vassilis Christophides, Serge Abiteboul, Sophie Cluet, and Michel Scholl. “From Structured Documents to Novel Query Facilities”. In *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 313–324, May 1994.
- [10] Serge Abiteboul, Sophie Cluet, Vassilis Christophides, Tova Milo, Guido Moerkotte, and Jérôme Siméon. “Querying Documents in Object Databases”. *International Journal of Digital Libraries*, Vol. 1, No. 1, pp. 5–19, 1997.
- [11] Jayavel Shanmugasundaram, Kristin Tufte, Gang He, Chun Zhang, David J. DeWitt, and Jeffrey F. Naughton. “Relational Databases for Querying XML Documents: Limitations and Opportunities”. In Malcolm P. Atkinson, Maria E. Orłowska, Patrick Valduriez, Stanley B. Zdonik, and Michael L. Brodie, editors, *VLDB’99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK*, pp. 302–314. Morgan Kaufmann, 1999.
- [12] Jian Zhang. “Application of OODB and SGML Techniques in Text Database: An Electronic Dictionary System”. *SIGMOD Record*, Vol. 24, No. 1, pp. 3–8, March 1995.
- [13] Alin Deutsch, Mary F. Fernandez, and Dan Suciu. “Storing Semistructured Data with STORED”. In Alex Delis, Christos Faloutsos, and Shahram Ghandeharizadeh, editors, *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA*, pp. 431–442. ACM Press, 1999.
- [14] Daniela Florescu and Donald Kossmann. “A Performance Evaluation of Alternative Mapping Schemes for Storing XML Data in a Relational Database”. Technical Report 3680, INRIA, May 1999. <http://rodin.inria.fr/dataFiles/FK99.ps>.
- [15] Daniela Florescu and Donald Kossmann. “Storing and Querying XML Data using an RDMBS”. *IEEE Data Engineering Bulletin*, Vol. 22, No. 3, pp. 27–34, September 1999.

- [16] “Oracle Technology Network”. <http://otn.oracle.com/tech/xml/>.
- [17] G. Elizabeth Blake, Mariano P. Consens, Ian J. Davis, Pekka Kilpeläinen, Eila Kuikka, Per-Å. Larson, Tim Snider, and Frank W. Tompa. “Text / Relational Database Management Systems: Overview and Proposed SQL Extensions”. Technical Report CS-95-25, UW Centre for the New OED and Text Research, Department of Computer Science, University of Waterloo, June 1995.
- [18] Albrecht Schmidt, Florian Waas, Martin L. Kersten, Daniela Florescu, Michael J. Carey, Ioana Manolescu, and Ralph Busse. “Why And How To Benchmark XML Databases”. *ACM SIGMOD Record*, Vol. 30, No. 3, pp. 27–32, September 2001.
- [19] A. R. Schmidt, F. Waas, M. L. Kersten, D. Florescu, I. Manolescu, M. J. Carey, and R. Busse. “The XML Benchmark Project”. Technical Report INS-R0103, CWI, Amsterdam, The Netherlands, April 2001.
- [20] Timo Böhme and Erhard Rahm. “XMach-1: A Benchmark for XML Data Management”. In *BTW 2001*, pp. 264–273, 2001.
- [21] Dao Dinh Kha, Masatoshi Yoshikawa, and Shunsuke Uemura. “An XML Indexing Structure with Relative Region Coordinate”. In *Proc. of IEEE 17th International Conference on Data Engineering*, pp. 313–320, April 2001.
- [22] Quanzhong Li and Bongki Moon. “Indexing and Querying XML Data for Regular Path Expressions”. In *Proc. of the 27th International Conference on Very Large Data Bases (VLDB)*, pp. 361–370, Roma, Italy, September 2001.
- [23] Shu-Yao Chien, Vassilis J. Tsotras, and Carlo Zaniolo. “Efficient Management of Multiversion Documents by Object Referencing”. In *Proc. of the 27th International Conference on Very Large Data Bases (VLDB)*, pp. 291–300, Roma, Italy, September 2001.
- [24] Brian Cooper, Neal Sample, Michael J. Franklin, Gisli R. Hjaltason, and Moshe Shadmon. “A Fast Index for Semistructured Data”. In *Proc. of the 27th International Conference on Very Large Data Bases (VLDB)*, pp. 341–350, Roma, Italy, September 2001.
- [25] Amélie Marian, Serge Abiteboul, Gregory Cobena, and Laurent Mignet. “Change-Centric Management of Versions in an XML Warehouse”. In *Proc. of the 27th International Conference on Very Large Data Bases (VLDB)*, pp. 581–590, Roma, Italy, September 2001.