高速性と信頼性を両立したコンテンツ管理向けネイティブXMLデータベース

Native XML Database for Contents Management Offering High Performance and Reliability

服部　雅一　■HATTORI Masakazu
野々村　克彦　■NONOMURA Katsuhiko
金輪　拓也　■KANAWA Takuya

XML（eXtensible Markup Language）は，インターネット上の交換･蓄積のフォーマットとして標準化された新たなデータ記述言語として期待されている。それとともに，XMLデータベースも様々な管理方式が提案されている。東芝は，XMLの持つ柔軟性や拡張性を生かすため，スキーマレスでXMLを格納するネイティブXMLデータベースを開発した。

このデータベースは，性能を出すためだけのXML検索エンジンではなく，信頼性と性能を両立したものである。マルチメディア管理やドキュメント管理など，コンテンツ系アプリケーションから異なる環境での情報交換を可能とするEAI（Enterprise Application Integration）などのネット系アプリケーションまで，様々なドキュメント系アプリケーションへ適用が可能である。

Extensible Markup Language (XML) is emerging as a major new standard for representing data on the World Wide Web. Several XML database systems have been proposed for storing XML documents in different data models.

Toshiba has developed a native XML database, for which XML documents are not required to have an associated schema or document type definition. The database is not a retrieval-oriented XML engine to achieve high-speed performance; rather, it provides the optimal balance of speed, size, functionality, and reliability. It can therefore be applied to various document-centric applications from document management to network system applications.

1　まえがき

XMLは，データの意味を表すタグを目的に応じて定義することが可能な国際標準のデータ記述言語であり，1998年にXML1.0としてW3C（World Wide Web Consortium）によって標準化された。XMLの具体的な記述例を**図1**に示す。

XMLは，HTML（HyperText Markup Language）と比較して次のような特長を持っている。

⑴　コンテンツとコンテンツの表現方法が分離されている。
⑵　タグの定義が可能である。
⑶　柔軟なリンク機能を持っている。

近年，企業や行政などで扱われる文書をはじめとしてXMLで記述することも一般的になってきている。身の回りにあるデータは，その構造化の度合いから以下の三つに分類できる。

⑴　構造的なデータ（売上げデータや従業員データなど）
⑵　構造があいまいなデータ，いわゆる半構造データ（議事録や設計データなど）
⑶　非構造データ（メモや通常のWWW（World Wide Web）ページなど）

XMLは半構造データモデル（Semistructured Data Model）に属する言語として，これらの異なる三つのデータを同一の枠組みで表現することができる。そのため，コンテンツのXML化は急激に進んでいる。同時に，大量のXMLデータを適切に管理･処理する仕掛けが必要になっている。

東芝では，XMLの持つ特性を生かすため，XMLネイティブ（XML文書を変換せずに格納できる）な高性能データベースを開発した。

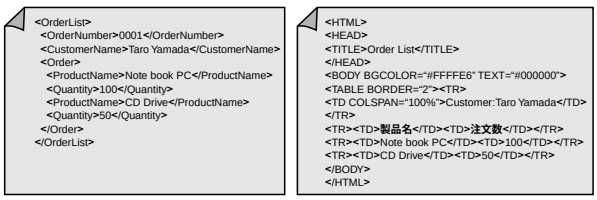

図1．XMLの記述例 ― 発注データの例である。左側はXML，右側はHTMLである。
Example of XML data

2　XMLデータベース

2.1　位置づけ

従来，コンテンツを管理するときに二つのタイプのデータベースが想像できる。一つは，テキストなど非構造データを管理する全文検索エンジンである。もう一つは，テーブルにより構造データを管理するRDB（Relational DataBase）であ

る。先にも述べたようにXMLは，非構造データと構造データを取り持つ半構造データモデルに属する。当然の帰結として，XMLデータベースも，全文検索エンジン的な要求とRDB的な要求が合わさって存在している(**図2**)。

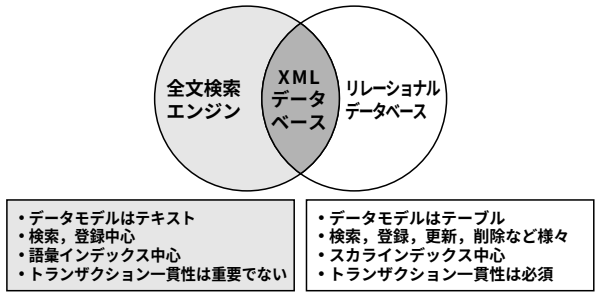

図2．XMLデータベースの位置づけ―全文検索エンジン的な要求とRDB的な要求が存在する。
Positioning of XML database

2.2 格納という観点から見た分類

XMLデータの管理には，いくつかの方式が提案されている(**図3**)。

単純な方式としては，XMLデータをそのままテキストファイルとして管理する方式である。これでは，データ数やサイズが大きくなると格納効率が悪くなったり，XMLの特性を生かした問合せが困難になる。次に考えられるのが，RDBにXMLデータを管理させるものである。更に，構造データを管理するために開発されたOODB(Object Oriented DataBase)で管理する方式もある。

基幹系などで広くRDBが使われているが，その拡張としてXML対応RDBが製品化されている。RDBは，データをフラットなテーブル形式に格納するため，XMLデータのような階層構造をテーブルに対応づける複雑なマッピングが必要となる。このマッピングのため，テーブルに関する事前の構造(スキーマ)設計を十分に行わないと，パフォーマンスが低下してしまう問題が発生する。これに対応するため，マッピングの自動化，テキスト(CLOB：Character Large OBjects)形式での格納，OODB的な技術の導入，などの方式が組み込まれている。

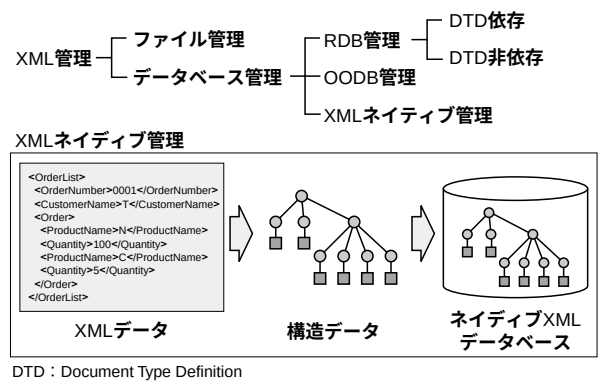

図3．XMLデータの管理方式―半構造であるXMLデータを管理するにはいくつかの方法が提案されている。ネイティブXMLデータベースは，XMLデータを特別なマッピング処理することなしに格納する。
Approaches to management of XML data

そもそもXMLデータを管理するときに，ユーザーはどのような機能や特長がポイントになるのだろうか。あるアンケートによると，“XMLデータを扱うときの高速性”や“XML階層構造をそのまま管理できること”などの回答が上位にきていた。そのような要求に応えるために提案されている方式が，“ネイティブXMLデータベース”である。

ネイティブXMLデータベースは，多種多様な階層構造を持つXMLデータを特別なマッピング処理することなしに格納する。そのため，格納や取得時に余計な処理が存在しない。また，コストのかかる事前のスキーマ設計が不要になり，ビジネス環境の変化により必要に応じてXMLデータの構造を自由に変更することが可能である。

このような特長は，マルチメディア管理やドキュメント管理などのコンテンツ系アプリケーションを中心に，異なる環境での情報交換を可能とするEAIなどのネット系アプリケーションまで，非定型・半構造データを対象とした適用分野で効果を発揮する。

2.3 問合せ言語

XMLデータが効率良く格納されるとともに，それらのデータを取り出す手段が必要になる。格納されたデータを取り出す手段として，問合せ言語がある。

RDBの世界では，SQL(Structured Query Language)がある。W3Cにおいて，XMLの問合せ言語に関する標準化活動が進んでおり，XQuery(XML Query)がワーキングドラフト段階で公開されている[(1)]。XQueryの問合せ例を**図4**に示す。

XQueryは，XMLデータをデータベースのように扱うための言語である。そのため条件に合致するデータ集合の取出

```
FOR $book IN document("bib.xml")//book
LET $a:= $book/author
WHERE contains($book/publisher,"Addison-Wesley")
RETURN
   <book> {
      $book/title,
      <count>Number of authors: {count($a)}</count>
   }</book>
```

図4．問合せの例―出版社がAddison-Wesleyである本のタイトルと著者の数の一覧を表示する。
Example of query data

しや集計・分析を行うための手段が提供されている。また，XMLデータは親子や兄弟などの要素が組み合わさった階層構造を持つため，その階層構造をたどったりする手段（XPath）が提供されている。更に，自由自在に構造を変換した結果をXMLデータとして出力できるのも特長である。

3 当社XMLデータベースの概要

当社は，XMLの持つ柔軟性や拡張性を生かしたXMLデータベースを開発した。以下にその概要を述べる。

3.1 目標

ネイティブXMLデータベースは，以下の目標のもとで研究・開発を行った[(2),(3)]。

⑴ XMLの適用分野の特性上，XMLデータ構造を厳密に定義しておくことは困難である。そこで事前のスキーマ設計をなくしたい。

⑵ 大規模なXMLデータでも，問合せや格納が高速でなければならない。

⑶ 語彙（ごい）や構造を条件にしてXMLデータを検索できる“検索エンジン”が既に製品化されているが，“検索エンジン”なのでトランザクション一貫性，同時実行性，リカバリなどのRDB並みの信頼性が保証されていない。これでは，ユーザーがアプリケーション開発やサービス提供を続けていくうえで大きな障害となる。そのため，検索エンジン並みの高速性とデータベース並みの信頼性を両立させなければならない。

⑷ 格納されたXMLデータを少ないプログラミングコストで検索・加工したい。ユーザー側のアプリケーションの開発コストも抑えたい。

3.2 特長

⑴ ネイティブXMLデータベース　DTDなどのスキーマなしで，XMLデータが持つ階層構造を維持したままストレージに格納している。XMLデータ取得時にも，その階層構造をたどることで原本のXMLデータを得ることができる。

⑵ 高性能　独自の問合せ最適化技術とインデキシング技術を新規開発することにより，数百万規模のドキュメントに対する通常検索（条件付き語彙検索）を数秒以内で応答することができる。

⑶ データベース管理システム　トランザクション一貫性，リカバリ機能，同時実行制御，API（Application Programming Interface），などのデータベース管理システムとして必須の機能を備えている。

トランザクション一貫性とは，同時に複数トランザクションが存在しても，各トランザクションの処理は完全に成功するか失敗するかのいずれかで中途半端に実行されないことを保証する。

リカバリ機能とは，マシン障害や基本ソフトウェア（OS）ダウンなどのトラブルが原因でデータベースがダウンしたとしても，正しい状態を完全復元する機能のことである。

同時実行制御とは，複数トランザクションが同時に同じデータを利用（参照・更新）する場合の不具合や競合を避けるという機能である。

⑷ XQueryのサポート　最新仕様であるXQueryを問合せ言語としてサポートしている。DOM（Document Object Model）に基づくプログラミングをしなくとも，多種多様なコンテンツを作成することが可能である。

システムの構成を**図5**に示す。

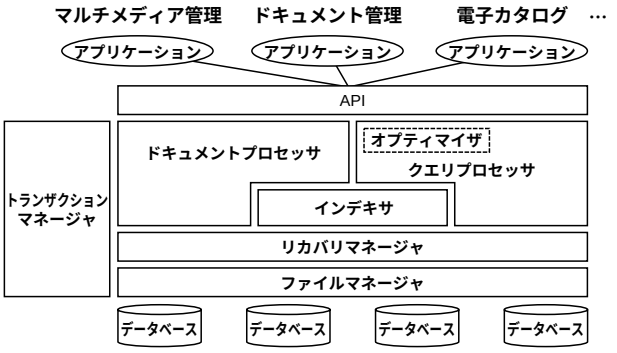

図5．システムの構成— XQueryを処理するQP（Query Processor），トランザクションを管理するTM（Transaction Manager）などいくつかの機能ブロックから構成されている。

System configuration of XML database

3.3 論理的な格納モデル

論理的な管理方法として，ツリー方式とコレクション方式がある。このデータベースでは，スペース，フォルダ，ドキュメントから成るツリー方式を採用している。ツリー方式には，次のメリットがある。

⑴ 複数のXMLデータに対する横断検索が可能

⑵ パスをベースに検索範囲の絞込みが可能

⑶ マニュアルのような階層的ドキュメント構築が可能

3.4 高性能

問合せ処理では，データベース上に存在する大量データ（多くはディスクにある）にアクセスする必要がある。そのため，複雑な問合せ要求では，処理時間が長くなる可能性がある。

一般に問合せ処理では，問合せ解析→問合せプラン生成→問合せプラン最適化→問合せプラン実行，という流れを踏む。問合せプラン最適化では，ファイルアクセスの最小化や結合演算コストの最小化などの観点で問合せプランの最適化が行われる。

RDBにおいてもSQLの問合せ最適化技術が長年研究・開発されてきたが，XMLにその技術を援用するには次のようないくつかの問題がある。

⑴ データモデルを階層構造に変更する必要がある。

⑵ 階層の部分構造への照合検索が必要である。同一構造を持つ要素の発生位置は一定でなく，入れ子が発生する可能性がある。そのため，正規表現のようなあいまいパスでの問合せが必要である。

⑶ 要素の順番を考慮しなければならない。

⑷ 複雑な語彙条件を含んだ検索が多い。

このデータベースでは，このような問題に対して以下のような方式で対処している。

まず，最適化技術については，独自の制約充足的な問合せ最適化技術を開発した。データの階層構造に対応した問合せグラフを設計し，問合せグラフを制約充足問題ととらえ，インデックス情報や構造情報から見積もられるコストデータによって最小の問合せプランを生成できる。柔軟性も高く，複雑なAND/ORを含んだ問合せも最適化することができる。

更に，弱スキーマの自動抽出技術を開発した。これは，格納されたXMLデータから緩やかな構造情報（弱スキーマ）を自動抽出するものである。格納されたXMLデータの構造をあらかじめ知っていれば，照合検索に対してアクセスすべきXMLデータベース上の範囲を最小限に絞り込むことができる。弱スキーマの自動抽出は，スキーマ自動生成に相当するものである。この弱スキーマの自動抽出により，従来比で検索速度が100倍近く高速化されている。

インデキシングについては，XMLデータ構造とXMLテキストに関するインデックスを自動生成する。構造インデックスと語彙インデックスの2種類のインデックスである。語彙インデックスは，N-Gramベースの語彙切り出しを行っており，独自の圧縮技術により格納されている。

3.5 リカバリ方式と同時実行性

独自のマルチバージョン管理を採用している。トランザクションによりデータ更新が発生すると，そのオリジナルデータに関するコピーを作る。システムダウンなどノード障害時には，オリジナルを復活させることでリカバリを行う。またコピーバージョンを適切に管理することで，問合せトランザクションと更新トランザクション間の競合を最小限に抑えている。

4 適用領域

XML関連のコンテンツは増大傾向にあり，このような特長を持つネイティブXMLデータベースを大容量の文書管理・コンテンツ管理アプリケーションへの適用を図っている。また，XMLはこれまでテキストなど非構造データでは困難であった分析や加工が容易になる。XMLベースのナレッジマネジメントでは，組織内で保有するノウハウや業務データを共有するだけでなく，より高度な知識創造の環境を促進することにもなる。現在，ナレッジマネジメントのミドルウェアとして適用を進めている。

更に，企業や行政だけでなく，家庭で楽しんでいるデジタルコンテンツもXMLで記述されることが多くなった（UPnP A/V（Universal Plug and Play Audio/Video），BML（Broadcast Markup Language）など）。家庭内で取り扱うコンテンツを管理するためのデータベースとして適用を検討しており，軽量化を中心に改良を進めている。

5 あとがき

XMLは，インターネット上の交換・蓄積のフォーマットなどに適用され，IT（情報技術）の基盤技術として定着しつつある。米国調査会社IDCの2002年9月の調査によると，XMLデータベースを含めたXMLサーバ市場は，2005年までに35億ドルに達するとの予測をしている。

しかし，現実にコンテンツ系アプリケーションでXMLデータを直接取り扱うユーザーから“スキーマ変更に追従できない”，“パフォーマンスが悪い”などの問題が指摘されている。このような問題に対して，当社ではネイティブXMLデータベースを研究・開発してきた。従来のデータベースと比較し，柔軟性や拡張性という観点でメリットを享受でき，前記のような問題を解消できるよう，更なる技術開拓に取り組んでいく。

文 献

⑴ W3C. XQuery 1.0: An XML Query Language. <http://www.w3.org/TR/xquery>, (accessed 2003-11-26).

⑵ 宮部安男，ほか．次世代ネットビジネスを支えるXML技術特集．東芝レビュー．**56**，11，2002，p1－34.

⑶ 服部雅一,ほか. 情報処理学会論文誌：データベース．**43**，SIG12（TOD16），2003，15p.

服部 雅一 HATTORI Masakazu

研究開発センター 知識メディアラボラトリー主任研究員。XMLデータベース及びナレッジマネジメントの研究・開発に従事。情報処理学会，人工知能学会会員。

Knowledge Media Lab.

野々村 克彦 NONOMURA Katsuhiko

研究開発センター 知識メディアラボラトリー研究主務。XMLデータベース及びナレッジマネジメントの研究・開発に従事。

Knowledge Media Lab.

金輪 拓也 KANAWA Takuya

研究開発センター 知識メディアラボラトリー。XMLデータベース及びナレッジマネジメントの研究・開発に従事。

Knowledge Media Lab.