D-008

素材管理データベースの XML 部分更新によるパフォーマンス向上

XML partial update method to improve performance of news data management system

橋本 光治　瀬田 賢治　渡部 敏雄　北川 亘 †
Kohji Hashimoto　Kenji Seta　Toshio Watabe　Wataru Kitagawa

1. はじめに

通信社、新聞社、メディア、プロバイダーでニュースを交換するフォーマットとして NewsML[i]が業界標準に定着し、各社がその採用を進めている[1]。IBM は、初期から NewsML に対応したソリューションを提案し、先進的なシステムを数社の新聞社で既に構築してきた。筆者らのチームは、その NewsML ソリューションのコンセプトデザインから参画し、共通コンポーネントの開発を通して IBM の NewsML ソリューションを支えてきた。

現在、大手新聞社様において NewsML による素材管理データベースを構築し、「コンテンツ流通」「データ交換の標準化」を実現してサービス・インをした。2次開発において、標準を維持しながら、より良いパフォーマンスを達成するアプローチに取り組んだ。筆者らは更新時に常に NewsML の XML 全体がサブシステム間を往来することにオーバーヘッドがあることを指摘した。その問題を解決するためにサブシステム間のインターフェースを見直し、部分更新ができる効率的な NewsML データベースの仕組みを提案した。現在、2次システムに向けてその開発を進めている。本稿では、XML でデータ交換を標準化したデータベースのパフォーマンス上の問題点とその解決策を述べる。

2. 素材管理データベースとその機能

IBM の NewsML ソリューションである NewsML データベース(図1）はつぎのコンポーネントから構成される。

- 共通インターフェース
- NewsML プロセッサー
- NewsML ストア
- NewsML リポジトリ IBM DB2 UDB/ IBM DB2 Content Manager

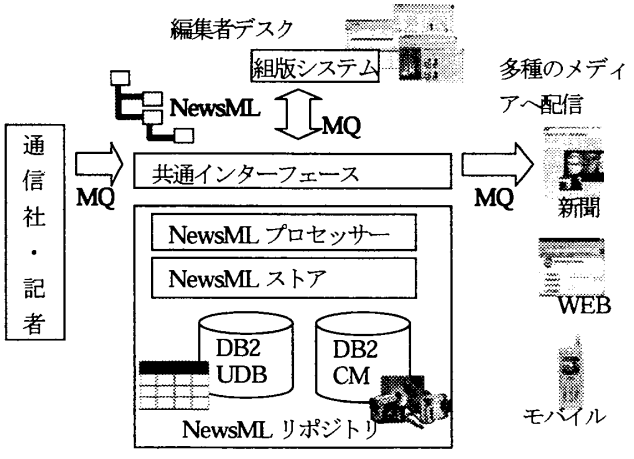

図 1　NewsML データベース

- 共通インターフェースは通信に WebSphere MQ、データ交換のフォーマットに NewsML を用いている。
- NewsML プロセッサーは、NewsML の XML 文書から NewsML 要素のオブジェクト・ツリーを生成し、NewsML の構造を操作する基本機能を提供する。
- NewsML ストアは、NewsML のオブジェクト・ツリーを NewsML リポジトリに保管し、XQuery で検索する機能を実現する。
- NewsML リポジトリは、NewsML の検索に利用される要素を IBM DB2 UDB のテーブルにマッピングし、NewsML から外部参照しているイメージ・オブジェクトを IBM DB2 Content Manager に保管する。検索に用いられない要素はテーブルにマッピングせず、NewsML の XML 文書全体を DB2 UDB の LOB[ii]列に保管する。

3. コンテンツモデル

ＩBMの素材管理データベースは、入稿・制作から配信先の複数のメディアまでワンソース・マルチユースで共有し、効率的でスピーディーな紙面制作を実現する。サブシステム間でデータを共有したマルチユースを実現するために、NewsML で来る素材コンテンツをメタデータで管理する。そのメタデータのデータモデルを定義することは非常に重要である[2]。大手新聞社様の素材管理データベース・システムで管理するのは、素材に限らず、素材の集合を表す素材セットと、組版システムによって編集され付加されるレイアウト情報を持った組・大組（面）パッケージ(図2)と呼ばれるデータ構造とがある。前述のセットやパッケージは NewsML 文書の参照関係によって集合を関連付けられる[3]。このように NewsML データベースで管理するすべてのデータが NewsML によって表現され、その NewsML がサブシステム間の情報交換する単位となる。

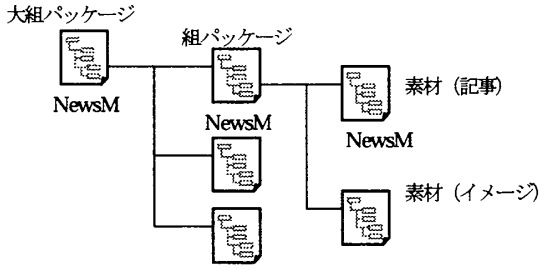

図 2　紙面を構成するデータモデル

4. パフォーマンスを妨げる要因

共通インターフェースのデータの粒度を NewsML 単位にしたことで、単一項目の更新時にも NewsML 全体がサブシステム間を往来する。紙面を作成する編集部の作業においては、更新が非常に頻繁で新聞制作システムの性格上、その更新はピーク時間に集中する。素材のある項目の更新時においても、XML 全体がデータとして流れ、受け取ったサブシステムでは、そのごく一部のデータを更新するために、全体の XML をパースする工程がオーバーヘッドとなっている。

† 日本アイ・ビー・エム株式会社　大和ソフトウェア開発研究所

4.1 部分更新をかけたいケース

制作系システムのユースケースを分析すると、つぎのようなケースが部分更新によってデータ交換の効率化とともに、素材管理データベースのパフォーマンス改善ができると考えられる。

①リンクの更新

大組パッケージをはじめとする参照関係[3]では、参照先のポインターが変わることで参照元の変更が発生する。この場合にNewsMLの意味的な内容は同じでも、参照元のNewsML全体がデータ交換されていた。ここに参照のみの部分更新が有効である。

②ワーキング中の素材や組みデータ(組・大組パッケージ)

素材や組みデータが編集デスクによって変更されている間は、随時変更がかかり、その都度全体更新が行なわれている。これはデータ交換にNewsMLのXML全体を必要としないため、部分更新に置き換えることができる。

5. 部分更新機能

サブシステムからNewsMLデータベースにXMLの部分更新を実装するにあたり、本システムで工夫した点を述べる。

5.1 XPathとNewsML要素

部分更新をするには、NewsML文書の特定と文書内の要素の特定、そして置き換える部分要素が必要となる。NewsMLは文書をユニークに特定するPublicIdentifierというIDを持ち、これを利用する。文書内の要素の特定は、XPathを用いる。XPathは汎用的であるため、部分更新に必要となる次の仕様に限定した。

i. 絶対パス

相対パスはカレントのオブジェクトがどの位置にいるか、双方で同期を維持することが複雑になる。部分更新をかけるサブシステムは、その時点でNewsML文書の全体を保持しており、ルートからの絶対パス指定が可能である。複雑性を排除するためにも絶対パスのみの採用とした。

ii. 各階層は要素を一意に特定できる条件を持つ

各階層は上位から要素を一意に特定する。下層から特定する機能は使用不可とした。上位層から下位層に向かって曖昧性を排除して要素を特定することが最も効率が良い。

iii. 要素を特定は出現順番と属性値の単純等価比較

上記i, ii, iiiの制限したXPathの仕様を実装することで、簡略化した非常に軽いXPathのパースエンジンを実装した。

XPathで対象となる要素をNewsMLリポジトリから特定し、更新する要素は、XMLの一部の要素から生成された部分オブジェクト・ツリーをそのNewsMLリポジトリに挿入、置換、削除の基本操作を実現する。

5.2 複合コマンドとトランザクション

サブシステム間はMQを介したリクエスト・リプライのメッセージ通信をする。そのため部分更新の1回毎にサブシステム間のメッセージが往復するのでは、オーバーヘッドが大きい。共通インターフェースは複合コマンドに対応し、一度のMQメッセージ通信で、複数の部分更新を実現する。(図3)

しかし、複合コマンドに対応しようとすると整合性の維持が難しくなった。一貫した整合性を保つために複合コマンドの1回を1トランザクションとして処理する。つまり、複合コマンドの部分更新でエラーが発生した場合は、複合コマンド全体をロールバックすることで、NewsML文書全体のデータ整合性を確保する。

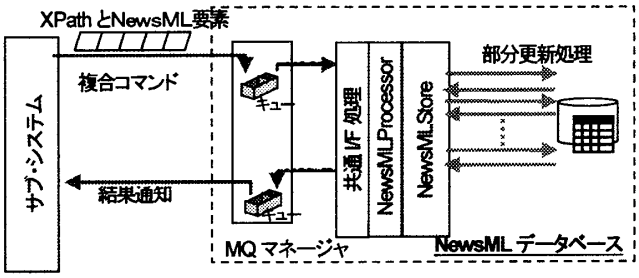

図3 複合コマンドによる部分更新処理

5.3 LOB更新を一括化

NewsMLは拡張が自由なためテーブルにマッピングされない項目も存在し、NewsML文書全体をLOB列に保管する。そのため部分更新時にもLOB列を更新する。部分更新の該当するテーブルの更新に比べ、XML文書を更新し、そのXMLをLOBへ更新するコストが高い。そのため部分更新では、XML文書をメモリーに保持し、複合コマンドではメモリー内で更新かけ、最後のコミット処理時に、LOBへの更新をする効率化を図っている。

テスト機によるベンチマークでは、全体更新よりも部分更新は更新の所要時間で1/7になった。

6. まとめ

XMLでモデル化されたデータをデータベースに格納する際は、そのデータ交換の粒度を考慮することが非常に重要である。本稿では、部分更新における効率的な部分抽出をXPathで実現した。また、整合性と効率を両立するために複合コマンドを1トランザクションにまとめ、LOB更新を一括化することが有効であることを確認した。

7. 今後の課題

同じXML文書に対して部分更新も数多く実施すれば一度の全体更新の方が効率で上回るケースが考えられる。今後、全体更新と部分更新のベンチマーク評価をしていき、部分更新の有効となる範囲を分析調査していく。

謝辞

NewsMLソリューションの提案者である日本アイ・ビー・エム　大和ソフトウェア開発研究の石井宏和氏、多大なご助言を頂いている同研究所 高倉伸氏に感謝します。

参考文献

[1] 日本新聞協会「NewsMLレベル1解説書(第1.0.3版)」2003.7.2
[2] 渡部敏雄「NewsMLのシステム間共通インターフェースへの適用」第4回情報科学技術フォーラム（FIT2005）2005.9
[3] 北川亘「マルチリファレンスを適用したNewsMLの高速アクセス手法の提案」第4回情報科学技術フォーラム（FIT2005）2005.9

---

i 1998年ロイター通信社がニュース素材の配信フォーマットとして提唱。NewsMLは日本新聞協会で規定しているNSK NewsMLレベル1に準拠している

ii ラージ・オブジェクト