差分によって記述されたXMLデータの格納検索方式

西村　雄介 (410095)

北陸先端科学技術大学院大学 情報科学研究科

2006 年 2 月 9 日

キーワード: XML，差分記述，XPath，問合せ処理.

現在，XML フォーマットによるデータは様々な分野で用いられており，システム間のデータ交換用として使用されるだけでなく，用途によってはシステム内のデータフォーマットとしてもその利用が期待されている．とくに，従来の関係データベースなどによる定型的なデータ項目から成り立つデータではなく，データ項目が不定型なデータや，文書が混在するようなデータについて，XML はより有用である．

一方で，XML フォーマットが利用されている様々な応用の中には，扱うデータの中に類似性の高いデータが多く含まれるようなものがある．これらの類似したデータをそれぞれ独立した XML 文書とするのではなく，類似したデータへの参照と異なる部分（差分）を記述するだけで，それぞれ独立した文書を記述しているのと同じ扱いになるような機構を提案する．これにより，データ量を圧縮できるだけでなく，共通部分を一括に更新することができ，また，類似データ間における違いが明確になり可読性が向上する．

本研究では，このような差分を用いた表現を実現するための枠組みと，その枠組みで表現されたデータを管理するための機構を提案する．本論文では特に，これらのデータ群を経路ベースアプローチを用いて関係データベースに格納し，XPath による問合せを効率良く実行するための手法を提案する．

本研究では，差分は「参照対象の子孫要素に含まれるある要素への上書き指示」とする．この参照と上書きの指示の記述を XML で記述し，XML のルート要素を含む任意の要素として記述することができるものとした．

参照と差分の記述の要素をどのように格納するかについては，いくつかの方式が考えられる．まず，参照や差分を記述した XML データを格納する一方で，参照や差分を全て解釈した完全な XML データも同時に格納する方法が考えられる．この方式の場合，問合せは通常の XML 文書と同様に扱うことができる．しかし参照する要素が多くの要素を子孫要素に保持している場合には，データ量が増加する．また，データの更新があった場合には，その箇所を参照しているような要素の全てを更新しなければならないため，更新時の負荷が高くなる．

また，逆に差分の記述を解釈せずに格納し，問合せ実行時に参照関係を解釈しながら問合せの評価を行う方法もある．この方法は先述の方式とは全く逆の長所と欠点を持つことになる．つまり，更新は通常の XML 文章と同様に行うことができ，また，差分を記述したデータ以外のデータを一切生成しないため，データ量を少なくすることができる．しかし一方で，問合せは要素間の参照関係をたどりながら行わなければならず，負荷が高くなる．また，XPath による問い合わせが行われた場合に，考えられる参照関係や上書きのパターンが多いため，検索を行う SQL が複雑化する．さらに，参照先の要素の子孫要素の中にさらに参照の指示がある場合など，再帰的な処理が必要な場合，すべてのパターンをあげて検索することは事実上不可能である．以上の点から，この方式は実用的であるとはいえない．

そこで，本研究では，これら二つの方式に加えて，参照と差分の関係を格納する索引を生成する中間的な方法も併せて提案する．この手法では，文書の参照関係を記述した木構造の索引を構築する．各ノードは元の文書の要素とリンクするための値と，その要素に到達するまでのパス式が記述されている．この索引は，参照と差分の関係のみを格納しているため，それ以外の要素が変更された場合でも大きな影響を与えない．また，参照関係はすでに解釈され，存在しうるパターンは絞られており，XPath による問い合わせを SQL に変換しても，非常にシンプルなものになる．先ほど問題としてあげた再帰構造についても，すでに解釈し，パス式を生成しているため，検索時に意識する必要はない．よって，検索処理についても実用的な速度で可能になることが期待できる．

本研究では，前述の三つの手法について，関係スキーマを示し，検索における具体的な SQL を示した．また，そのスキーマを用いて，実験を行い，データ量，検索速度，更新速度という観点で検証を行った．この実験により，少なくとも単純な XPath における問い合わせおよび単純な参照関係については，各手法について上記で述べたような性質を持っていることを確かめることができた．

今後の課題として，より複雑な XPath 問い合わせや，より大規模な文書群に対する問い合わせ，複雑な再帰関係を持った文書に対する問い合わせについて検証を行う必要があること，があげられる．