# ネットワーク上でのXML問い合わせ集合の最適化

福井 佳紀 (210078)

北陸先端科学技術大学院大学 情報科学研究科

2004年2月13日

**キーワード**： XML, XPath, Query Processing, Distributed Database, Optimization.

今日，XMLフォーマットは頻繁に利用されるようになり，インターネット上でのデータ交換やデータ発信の標準ともいわれるようになった．その結果，XMLデータはネットワーク上に散在するようになり，ネットワーク上のXMLデータを効率的に問い合わせるための実現方法が求められている．上述のようなXMLデータの情報サービスシステムでは，1999年にW3Cの勧告となったXPathという問い合わせ言語が一般的に用いられている．XPathは，XMLデータ中の特定のノード集合をパス式によって選択することができる非常にシンプルな問い合わせ言語である．XMLデータは通常，ラベル付き木で表現され，データ中のエレメントのうち，問い合わせのパス式にマッチするエレメントを根とする部分木の集合が返される．

ネットワーク上のXMLデータベースに対して，クライアントが複数のXPathによる問い合わせを行う場合を考えると，返送される解集合には次の場合に，冗長性が含まれている可能性がある．第一に，ある解集合に含まれている，あるエレメントと，別の解集合に含まれているあるエレメントが全く同一のものであり，重複している場合がある．第二に，ある解集合中のあるエレメントが，別の解集合中のあるエレメントの部分木となっている場合がある．この二つ場合にXPathの問い合わせの解集合に冗長性が発生し得る．さらにいえば，複数のXPathではなく，単一のXPath問い合わせを発行する場合にも自己冗長性が発生する場合がある．これは，その問い合わせの解に含まれるあるエレメントが，同じ解に含まれる別のエレメントの部分木になっている場合があるためである．サーバがこれらの冗長性をもった解をそのままクライアントに送信すると，ネットワーク上に同じデータが何度も流されることになり，通信コストの上では最適とはいえない．そこで，本研究では，ネットワーク上でXPath問い合わせを実行する場合に生じる通信コストを最適化するための手法を提案する．本研究では，サーバに手を加えられない場合と，サーバに手を加えられる場合の二種類の場合を想定し，それぞれに対して研究を行っている．

まず，サーバに手を加えられないシステムを考える．ネットワーク上のXMLデータベースがサーバとしてクライアントからのXPathの問い合わせを待ち受けており，受け取った問い合わせを評価し，解集合をクライアントに返すというシステムを前提に考えている．

そのため，クライアントが自由にサーバを変更することができないので，計算コストを最適化するためには，問い合わせ内容を変更するしかない．我々は，上述のような解の冗長性による通信コストの増大を防ぐために，XPath 問い合わせの集合を与えられた場合，それらの問い合わせ全てに答えることができるサイズ最小のビューを求め，これをサーバからクライアントに送信し，クライアント側でこのビューから，オリジナルの問い合わせによって得られるはずであった解集合を生成する方法をこれまでに提案した．これまでに提案した手法では，通信コストの最適化に特化されており，サーバの計算コストは増大してしまう場合がある．これは，サイズ最小のビューに変換された問い合わせは，オリジナルのものと比べて複雑な計算を必要とするのが原因である．例えば，サイズ最小のビューでは，問い合わせ集合間のすべての共通部分を取り除くために，可能な限り解を分割して取り出せるような問い合わせに変換し，解中に共通部分が発生しないようにする．これによって，否定を求める演算や集合の共通部分を取る演算の数が増え，計算が複雑になる．また，自己冗長性を取り除く演算には，根からみて一番浅い場所にあるエレメントを取り出す必要があるため，さらに複雑な処理が必要となる．そこで，本論文では，サイズ最小のビューに変換された問い合わせ集合の中で頻繁に現れるパターンの部分について，サーバが保有する XML データ中の統計情報を利用することによって，より簡単な問い合わせに変換し，計算コストを軽減する手法を提案する．

一方，サーバ側に手を加えられる場合では，上述の手法に加えて，さらにクライアント側での計算コストも減らすことが可能である．一般的なデータベース問い合わせシステムでは，サーバが返信する解データをそのまま所望する解としてクライアントが利用することが可能である．しかし，上述のサイズ最小のビューに変換する手法では，クライアントが，受け取ったサイズ最小の解からオリジナルの問い合わせで得られるはずであった解集合を取り出す追加処理が必要となる．クライアントが携帯電話や PDA といった，組込み機器を利用しており，マシン性能がある程度制限されている環境にあれば，この解を取り出す追加の計算は好ましくない．そこで，これを解決するために二つの手法を提案する．まず，はじめに，クライアント側での計算コストを軽減するために，解集合を簡単な問い合わせで取り出せるように，サーバがデータを加工する手法を提案する．次に，XML 問い合わせにおいて，大きな負荷がかかるパーズ処理に着目し，この負荷を取り除くため，取り出すべき解の位置のバイトオフセットをサーバが送信することによって，クライアントがパーズ処理を行わずに解を取り出せる手法を提案する．実験の結果，後者の方がより計算コストが改善された．

上述のように，ネットワーク上の XML データベースに対して問い合わせを行うときに生じる，通信コストと計算コストの問題に焦点を当て，様々な効率化手法を提案した．そして，XML データを関係の形にエンコーディングをしたものを，関係データベースに格納し，SQL で問い合わせを行うシステムと，XML データをメモリ上に展開し，インメモリ上で XPath 問い合わせ処理を行うシステムを用意し，評価実験と検証を行った結果，これらの手法の有用性を示した．