様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１４　日現在

機関番号：　６２６１５
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３３００９７

研究課題名（和文）データ相互運用問題解決のためのスキーママッピングを用いたXQueryの書換え手法

研究課題名（英文）Rewriting XQuery using Schema Mappings to Solve Data Interoperability Problems

研究代表者
加藤　弘之（KATO, Hiroyuki）

国立情報学研究所・コンテンツ科学研究系・助教

研究者番号：１０３２１５８０

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：同じ意味を表す異なるデータに統一的にアクセスする問題である，データ相互運用問題では，データ交換のための国際標準であるXMLデータとその問合せ言語であるXQueryが使われている．
XML/XQueryの実用上の普及を妨げている問題が「順序に関する厳しい制限」であると国際会議でも指摘されている．特に，典型的なXQueryの使用法の一つである，指定された構造を有するデータの抽出である，twig問合せでは，この「順序に関する厳しい制限」によって問合せの効率が悪いことが知られている．
本研究では，このtwig問合せを効率的に実行するためのXQueryの書き換え技術を開発し，実験によりその有効性を示した．

研究成果の概要（英文）：In the data interoperability problem, the goal of which is to provide a uniform access to different data structure having the same meaning data, XML/XQuery are used because XML has been developed for data exchanging format and XQuery is a query language of XML. The reason why XML/XQuery ended up failing in prattle is in their "heavily ordered" feature as pointed out in FADS2017, an international workshop on Failed Aspirations in Database Systems.
In this research, we developed a method for rewriting XQuery to optimize a twig query, which is a typical use of XQuery. Also, the experimental results shows efficiency of the method.

研究分野：　データベースプログラミング言語

キーワード：　XQuery　静的解析　最適化

１．研究開始当初の背景

「データ相互運用問題(Data Interoperability Problems)」とは，異なるスキーマのもとに存在している同じ意味を表すデータを統一的に扱う問題であり，二つのアプローチとして「データ統合」と「データ交換」がある[1]. これら二つの大きな違いは，データ交換はデータを重複して持つのに対して，データ統合ではデータは重複して持たずに問合せの書き換えを通じて回答するデータを検索する．またこれら二つに共通する点は，スキーママッピングを用いて異なるスキーマの違いを記述する点にある．本研究では主にデータ統合に基づくアプローチを採用する．尚，データ統合では閉包性を達成するために，使われているデータモデルの問合せ言語を用いてスキーママッピングを記述する．

データ相互運用問題は，近年次の二つの点で新たな方向に拡張されている．1)対象となるデータモデルは関係データから XML データへと拡張．2)統合手法は，データ全体をカバーするスキーマによる統合からピア間のスキーママッピングを利用する P2P 手法へと移行している．

Web 上での情報交換フォーマットとして開発された XML は，木または森をデータモデルとして採用し，様々なデータを記述するのに用いられている．また，XML データを対象とする問合せ言語として当初開発された XQuery は，様々なユーティリティを兼ね備えた関数型言語としての完成度の高さと，その記述能力の高さから，問合せとしてだけでなく，様々なアプリケーション開発言語としても用いられている．実際，アマゾンなどの IT 企業で実際の業務に使われている．

[1] Balder ten Cate, Phokin G. Kolaitis, Wang-Chiew Tan, Schema Mappings and Data Examples, Tutorial, In EDBT 2013.

２．研究の目的

本研究の目的は，XML データ統合問題において，より現実的な問題を取り扱うために，大規模な実用に耐えるような最適化機構を開発することである．具体的には，XML データに対する問合せの典型例の一つとして，広く知られている「twig 問合せ」の効率処理に取り組む. twig 問合せとは与えられた木のパターンを満たす部分木の抽出であり，特に順序木を考慮した場合の非効率さは広く知られている．本研究では，実用上よく使われているスキーマ情報を使う場合と，セキュリティやプライバシなどの理由によりスキーマ情報が手に入らない場合とで，それぞれ twig 問合せの効率化を目標とした.

３．研究の方法

(1)．スキーマ情報を使わない場合とスキーマ情報を使う場合とで，実際に使われるデータの性質および，スキーマのクラスを整理し，それぞれの性質に応じて，研究を進めることにした.

(2)．調査の結果，データ統合においてよく使われているスキーマのクラスが Nested-Relational DTD であることがわかり，このスキーマを使って，静的解析による最適化の開発を目指すことにした.

(3)．一般的な解法の一つである生成検査法(generate-and-test approach)を，問合せの構成に適用することで，系統的な書き換え手法が確立できた.

(4)．本研究に興味を示している欧州の XQuery エンジン企業 BaseX 社から，問合せ処理解析ツールの提供を受け，このツールを使う事で，最適化の効果をより詳細に分析した.

４．研究成果

(1)．効率的な静的ストアの設計と目標とすべき問合せのクラスについての調査を行った．本研究の最大の特徴は静的解析に基づく問合せの最適化である．一般にストアは実行時に値を格納し参照するために用いられるが，本研究では，静的解析から実行時に取りうるエレメント名とその文書順序を格納し参照する枠組みとして，「静的ストア」を用意することで，問合せの最適化に利用している．スキーマ情報から得られる等価問合せを用いた場合，静的ストアにはエレメント構築子が含まれるため，冗長なエレメント構築が存在してしまい最適化を阻害してしまう．そこで，エレメント構築子を用いずに変数を使うことで，等価問合せを用いた場合と同じ効果でより効率的な問合せを生成できることがわかった.

(2)．本研究で目標とすべき二つの問合せのクラス Tree-Free XQuery と DDO-Free XQuery を定義した．これら二つのクラスは比較不能であるが, Tree-Free かつ DDO-Free とすることでより最適な問合せへと変換できることがわかった．但し，水平軸を対象とした DDO-Free XQuery への変換は，特に複数の同じエレメントが同じレベルに存在するとき，難しいことがわかった．更に，DDO-Free XQuery への変換で構造に関する条件を伴う if 式が生成されるが, スキーマ情報を使うことでこの構造に関する条件を満たすかどうかを静的に判断できる場合があることがわかった.

(3)．スキーマ情報を使わない場合，生成検査法の生成段階では, descendant-or-self 軸を使うことで，順序に基づいた全ての値を参照できる問合せをスケルトン問合せとして

用意する．次に，検査段階として，与えられた木のパターンがこれらの値を満たすかどうかの書き換えをすることで，twig 問合せを効率的に処理することに成功した．また，この手法は，ストリーミングデータにも適用可能なことがわかった．

(4)．スキーマ情報を使う場合，生成段階では，順序に基づく全ての木のノードを生成する問合せをスケルトン問合せとして用意する．スキーマを Nested-Relational DTD のクラスに制限することで，child 軸だけを使った細粒度の問合せとなる．次の検査段階では，与えられた木のパターンを適切な位置に配置することで，効率的問合せへと変換できることがわかった．特に，入力問合せ中の条件を適切に抽出し，スケルトン問合せの適切な位置に配置する必要がある．この条件抽出の為の変換は等価変換である必要はなく，modulo DDO のもとでの変換で十分であることを証明した．この変換のために１１個の変換規則を定義し，その正しさを証明した．

(5)．上記(4)．で得られた問合せの最適化技術を実装し，定量的に評価するために，二つの XQuery エンジン，BaseX と SAXON を用いて，実験を行った．この実験によりおよそ 2000 倍速くなることが確かめられた．また，書き換え前の問合せではメモリ消費量が大きく，大規模データでは実行自体ができないものに対しても，書き替え後の問合せでは，メモリ消費量が抑えられ実行可能となった．

(6)．XML/XQuery が実用的に普及しない原因の一つが，「順序に関する厳しい制限(heavily ordered)」であると，2017 年の VLDB 併設ワークショップ(Failed Aspiration in Database Systems(FADS@VLDB)で指摘された．この「順序に関する厳しい制限」とは，XML を順序木として扱う際の木のノードに関するソートと重複の削除(Distinct Document Order，DDO)のことであり，XQuery の意味に組み込まれている．本研究課題の一つは，与えられた XQuery 問合せを DDO 処理のない DDO-Free XQuery に変換する手法の開発である．入力 XQuery を DDO-Free XQuery に書き換える手法において，long-distance 軸，parent 軸，self 軸の削除によって列式が導入されてしまい，変数に束縛される木のレベルが一定とならなくなってしまうという問題があった．この問題は，書き換えステップの最初の段階で，スキーマ情報から取得できる木の最大の高さを用いて書き換えた後で対応することで解決することができた．

(7)．Nested-Relational DTD から identity 問合せが導出できる事を示し，そこから XQuery 式の抽象表現が定義できる事を示した．この抽象表現は，XQuery 式を評価した結果の木の構造を適切に表現しており，軸計算に利用できるだけでなく，provenance 情報としても活用できる可能性があることがわかった．Provenance 情報とは，そのデータの来歴情報のことであり，データの信頼性の評価に利用できることがわかっており，近年注目されている．この provenance 情報を用いることで，データベースの分野で長年の課題として取り組まれてきた「answering query using materialized views」に寄与できることがわかった．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

(1) Hiroyuki Kato, Soichiro Hidaka, Zhenjiang Hu, Keisuke Nakano and Yosunori Ishihara, Context-preserving XQuery fusion, Mathematical Structures in Computer Science, 査読有, Vol.25, 2015, pp.916-941

〔学会発表〕（計 6 件）

(1) 加藤弘之，石原靖哲，Torsten Grust, DDO-Free XQuery, The 16th International Symposium on Database Programming Languages (DBPL 2017), 2017

(2) 朴柱英，吉川正俊，加藤弘之，Cell-based Provenance for Scientific Data, ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL2017), 2017

(3) 朴柱英，吉川正俊，加藤弘之，セル単位データ来歴--データ引用に向けて，日本データベース学会，2017

(4) 杉村憲司，石原靖哲，加藤弘之，藤原融，制限された DTD クラスのもとでの XQuery 式の型推論，電子情報通信学会ソフトウェアサイエンス研究会，2016

(5) 石原靖哲，加藤弘之，Torsten Grust，ノードの並べ替え処理と重複削除処理の回避による XQuery 最適化，電子情報通信学会ソフトウェアサイエンス研究会，2016

(6) 鬼塚真，加藤弘之，日高宗一郎，中野圭介，胡振江，Optimization for iterative queries on MapReduce, 40th International Conference on Very Large Data Base (VLDB 2014), 2014

６．研究組織
(1)研究代表者
加藤 弘之（KATO, Hiroyuki）

国立情報学研究所・
コンテンツ科学研究系・助教
研究者番号：10321580