様式 C-19

科学研究費補助金研究成果報告書

平成２３年５月１９日現在

機関番号：６２６１５
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：２０５０００４３
研究課題名（和文）　XML データ統合問題解決のための XQuery の静的解析に基づく書き換え手法の開発
研究課題名（英文）　Developing query rewriting techniques to resolve XML data integration Problems based on a static analysis.
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

研究成果の概要（和文）：融合変換は冗長な中間結果を削除する手法の一つであり、これまで SQL などのデータベース問合せ言語に対する最適化手法として用いられてきたが、本手法を用いた XQuery の最適化は未解決である。その理由は、XQuery は文脈を考慮にいれる必要があり、かつ文書順序の保存が要求されるからである。本研究の成果は、我々の知る限り、文脈と文書順序を取り扱った最初の XQuery 融合変換を開発した。

研究成果の概要（英文）：Fusion is a known technique for eliminating unnecessary intermediate results that are created but then consumed during computation. Being useful for query optimization for many query languages such as SQL, fusion remains as a challenge for XQuery optimization. This is because XQuery has more complicated semantics; it is context sensitive and requires preservation of document order. In this research, we proposed, as far as we are aware, the first XQuery fusion that can deal with both the document order and the context of XQuery expressions. More specifically, we carefully designed a context representation of XQuery expressions based on the Dewey order encoding, developed a context-preserving XQuery fusion for ordered trees by static emulation of the XML store, and proved that our fusion is correct. Our XQuery fusion has been implemented and successfully applied to the multi-step schema mapping, in which fusion is particularly necessary for reducing execution cost of redundant node creations.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	900,000	270,000	1,170,000
２００９年度	700,000	210,000	910,000
２０１０年度	500,000	150,000	650,000
年度			
年度			
総　計	2,100,000	630,000	2,730,000

研究分野：総合領域
科研費の分科・細目：情報学・ソフトウェア
キーワード：アルゴリズム、関数系言語、融合変換、XQuery、データ統合、最適化、XML

１．研究開始当初の背景

異なるスキーマのもとで存在している同じ意味を表わすデータを統合して扱う問題は、データ統合問題と呼ばれ次の二つの点で新たな方向に拡張されている。１）対処となるデータモデルの関係データから XML データへの拡張、２）統合手法は元データ全体をカバーする統合スキーマによる統合から、元データのピア間のスキーママッピングを利用する P2P 手法へと研究対象は移行している。

データ統合問題では閉包性を達成するため、その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。

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。XML ではデータを記述する際、固有のスキーマを用いることで統一的なデータの利用が可能となるが、同じ意味のデータに対して複数のスキーマが存在する。例えば同じメタデータを表わす場合でも RDF、MPEG-7、XTM など様々なスキーマが存在する。XML データにおけるスキーマとはタグの出現の仕方であり、RDF には RDF 用のタグ名とそのタグの出現の仕方が、MPEG-7 にはそれ用のタグ名とそのタグの出現の仕方が規定されている。XML データではタグを含むノードにユニークな ID が割り振られており、このノード ID の順番に基づく順序木または順序付き森をデータモデルとして扱っている。

XML データ統合問題に置ける大きな特徴は、スキーマの変更であるタグの構築に伴い新たに生成された XML ノード ID の取り扱いにある。一般に、データ統合問題における問合せ処理では、スキーマ変更前のデータを検索することで冗長なスキーマ変更のオーバヘッドを回避した問合せ最適化を行う。ところが、XQuery をスキーママッピングに採用した XML データ統合問題においてこの手法をナイーブに適用すると、書換え前後の式が等価にならないという問題が生じる。その理由は以下の通りである。XQuery におけるタグの構築ではデータモデルである木構造を維持するために、構築されたタグとのそのタグの内側に出現する元データのノードに対して、新たなノード ID を割り当てる。尚、この割り当てられたノード ID 上の順序を用いて構築されたタグをルートとする順序木がデータモデルとなる。これにより、スキーマ変更前のデータのノード ID とスキーマ変更後のデータのノード ID が異なってしまい、書換え前の式と、冗長なスキーマ変更を回避した書換え後の式がノード ID については等価ではなくなってしまう。XQuery で頻繁に用いられる経路式（XPath）における軸ステップの意味が「ノード ID についての重複の削除とソート」と XQuery の形式的意味で定義されているため、ノード ID の等価性問題は重要である。

２．研究の目的

本研究では、XQuery に対して上記の問題を解決する融合変換のアルゴリズムを与えることが目的である。

３．研究の方法

本研究では XQuery の式の出現に対して、あるコードを割り当てることでこの問題を解決した。ノード ID 上に定義される順序関係を保存する、ノード ID を定義域とする準同型の値域をこのコードの性質とすることで本問題を解決した。

具体的には、これまで XML データの索引技術に用いられてきた Dewey order encoding をデータではなく、XQuery の式の出現に対して施した。この際、XQuery の for/let 式を扱うために extended Dewey order encoding を導入した。

４．研究成果

本研究の対象とする式は、変数、列式、経路式、for 式、let 式、エレメント構築子とした。W3C による XQuery の形式的意味はその部分集合である XQuery Core 上で定義されている。本研究では経路式を扱っているのに対して、XQuery Coreに経路式は存在せず、for 式に正規化している。本研究で経路式を正規化せずに扱っている理由は次の通りである。(i).経路式を for 式に変換して評価すると非常にコストがかかるが、経路式のまま評価すると線形で評価できることが知られている。(ii).本研究の融合変換の特徴は経路式の部分計算であるため、for 式に変換してしまうと、経路式の正規化により生成された for 式固有の書換え規則を開発する必要が有る。(iii). これまでの XQuery に関する研究でも経路式を扱っている。

本研究の成果は次の四つの項目から構成される。（１）XQuery の式の出現に対して Dewey order encoding を拡張した文脈表現を定義し、（２）これを用いた文脈を保存する XQuery の融合変換アルゴリズムを開発し，（３）その正しさを証明し、（４）プロトタイプ実装を用いた実験を通じてその有効性を確かめた。以下、各項目について概説する。

（１）for/let 式の return 節に出現する式の順番を表わす区切り子`#’を新たに導入した拡張 Dewey order encoding を定義した。

（２）アルゴリズムは三つの要素から構成されている。一つは、全体のアルゴリズム peval であり、式の帰納的定義に従い再帰する。peval がエレメント構築子に対して処理をするときに、文脈を式に出現に割り当てるアルゴリズム dcp が呼ばれる。dcp はエレメント構築子に対して新たな Dewey code を割り当て、そのエレメント構築子の内側に出現する式に対して、再帰的に割り当てを行う。また、peval が経路式に対して処理をするときに、経路式の部分計算をする axis_fusion が呼ばれる。axis_fusion は予め割り振られている拡張 Dewey code を用いて部分計算をする。

（３）アルゴリズムの正しさを証明するた

めに、まず XQuery の形式的定義を Dewey code を用いて再定義した。次に、文書順序は式を評価している最中は重要な役割を果たすが、一旦直列化されると式の評価中に用いられていた文書順序は失われてしまうという性質を利用し、XQuery の形式的定義を用いて、値等価性を定義した。アルゴリズム peval が入力 XQuery 式に対して、値等価性を保持した XQuery 式を出力することが、正しさの証明である。これは次の二つの段階を経て示すことができた。まず、エレメント構築子の内側に出現する式が評価された際に、どのような値を生成するのかを、拡張された Dewey order によって示した。次に、アルゴリズム dcp がこの性質を満たすような Dewey order encoding を式の出現に付与することを示した。また、axis_fusion は新しく導入した拡張 Dewey order のもとで、重複無しソートを実現していることを示した。これによりアルゴリズムの正しさを示すことができた。

（４）OCaml を用いてアルゴリズムのプロトタイプ実装を行った。この実装を用いて多段のスキーママッピングを想定した実験を行った。実験には XQuery エンジンとして Galax バージョン 1.0 を用いた。エンジンのデフォルト最適化オプションは有効にしている。実験に Galax を用いた理由は、エレメント構築の処理において、他のエンジンである Berkeley DB、eXist、MonetDB、QizX、Saxon などはコピーと部分木のマテリアライズをするのに対して、Galax だけはノードのコピーを遅延評価する場合があるからである。この実験ではスキーママッピングのステップ数を 30 までとして実験した。その理由は、これまでの他の研究で使われていたピアの数は 20 数個から 80 数個までで実験しており、ピア同士がスキーママッピングを通じてグラフ構造を形成することから、ピア間の平均的なホップ数を考慮すると最大数は 30 が妥当と考えたからである。この実験を通じて、本研究成果の有効性を確かめることが出来た。

更に、本研究成果は、次の二つの方向においてもその有効性が確かめられた。一つは、データベース問合せにおいては順序は特に重要でない場合がある。その場合 XQuery の unordered modeを使うが、重複の削除は必要であり、この重複の削除をする際に役に立つ。二つ目は、XQuery 式における非決定性の判定に役に立つ。結果の順番が予測不能な問合せ式の検出は、特にデータ統合問題においては重要なことである。

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８件）

① Hiroyuki Kato , Soichiro Hidaka, Zhenjiang Hu, Keisuke Nakano, Yasunori Ishihara, Context -preserving XQuery Fusion, GRACE Technical Report (GRACE-TR-2010-07), 1-21
② Kazuhiro Inaba, Soichiro Hidaka, Zhenjiang Hu, Hiroyuki Kato, Keisuke Nakano, Sound and Complete Validation of Graph Transformations, GRACE Technical Report (GRACE-TR-2010-04), 1-27
③ Soichiro Hidaka, Zhenjiang Hu, Kazuhiro Inaba, Hiroyuki Kato, Kazutaka Matsuda, Keisuke Nakano, Bidirectionalizing Graph Transformations, GRACE Technical Report (GRACE-TR-2010-06), 1-17
④ 加藤弘之、日高宗一郎、胡振江、中野圭介、石原靖哲、順序を考慮にいれた XQuery の融合変換、Web とデータベースに関するフォーラム、査読有、2009
⑤ Soichiro Hidaka, Zhenjiang Hu, Hiroyuki Kato, Keisuke Nakano, A Compositional Approach to Bidirectional Model Transformation, Proc. of ICSE2009, NIER Track, 2009, 査読有
⑥ Hiroyuki Kato , Soichiro Hidaka, Zhenjiang Hu, Yasunori Ishihara, Keisuke Nakano, Rewriting XQuery to Avoid Redundant Expressions based on Static Emulation of XML Store, ACM SIGPLAN Workshop on Programming Languages Techniques for XML,2009, 査読有
⑦ Soichiro Hidaka, ZhenjiangHu, Hiroyuki Kato, Keisuke Nakano, Towards a Compositional Approach to Model Transformation for Software Development, Proc. of the 2009 ACM Symposium on Applied Computing, 2009,468-475, 査読有
⑧ 日高宗一郎、加藤弘之、吉川正俊、書き換えに基づく最適化のための XQuery の相対コストモデル、電子情報通信学会論文誌、査読有、JD91-D、2008、873-888

〔学会発表〕（計 14 件）

① Isao Sasano, Zhenjiang Hu, Kazuhiro Inaba, Hiroyuki Kato, Keisuke Nakano, Toward bidirectionalization of ATL with Ground Tram, 4 th International Conference on Model Transformation (ICMT2011), Zurich, Switzerland, June 27-28, 2011. （発表確定）
② 中野圭介、日高宗一郎、胡振江、加藤弘之、模倣に基づくグラフスキーマを利用したビュー更新可能性判定、第 13 回プログラミングおよびプログラミング言語ワーク

ッショップ、北海道、定山渓ビューホテル、2011年3月9日-11日
③ Hiroyuki Kato , Functional Graph Transformations with Structural Recursion, The 4th DIKU-IST Joint Workshop on Foundations of Software, 東京都、浅草ビューホテル、2011年1月10日-14日
④ Soichiro Hidaka, Zhenjiang Hu, Kazuhiro Inaba, Hiroyuki Kato, Kazutaka Matsuda, Keisuke Nakano, Towards State -based Interface to a Graph Roundtrip Transformation System GRoundTram (poster), 8th Asian Symposium on Programming L anguages and Systems (APLAS2010), Shanghai, China, November 28 - December 1, 2010.
⑤ Hiroyuki Kato , Soichiro Hidaka, Zhenjiang Hu, Keisuke Nakano, Yasunori Ishihara, Context Preserving XQuery Fusion, 8th Asian Symposium on Programming Languages and Systems (APLAS2010), Shanghai, China, November 28 - December 1, 2010.
⑥ Keisuke Nakano, Soichiro Hidaka, Zhenjiang Hu, Kazuhiro Inaba, Hiroyuki Kato, Range Analysis of Graph Transformation for Simulation -based Schema(poster), 8th Asian Symposium on Programming Langua ges and Systems (APLAS2010), Shanghai, China, November 28 - December 1, 2010.
⑦ Hiroyuki Kato, Regular Path Compilation for Graph Updating, 1st PKU-NII International Joint Workshop on Advanced Software Engineering, Beijing, China, October 9-10, 2010.
⑧ Soichiro Hidaka, Zhenjiang Hu, Kazuhiro Inaba, Hiroyuki Kato, Kazutaka Matsuda, Keisuke Nakano, Bidirectionalizing Graph Transformations, 15th ACM SIGPLAN International Conference on Functional Programming (ICFP2010), Baltimore, USA, September 27-29, 2010.
⑨ Hiroyuki Kato , Two Semantics of Updating in GroundTram, 4th International Workshop of Bidirectional Transformation in Architecture-Based Component Composition, March 12-14, 2010, 神奈川、箱根パレスホテル
⑩ Hiroyuki Kato , Soichiro Hidaka, Zhenjiang Hu, Keisuke Nakano, Yasunori Ishihara, An Order -Sensitive XQuery Fusion, The 7th Asian Symposium on Programming Languages and Systems (APLAS2009) (poster), December 14-16, 2009, Seoul, Korea
⑪ Hiroyuki Kato , Towards A Context Preserving Fusion in Optimizing Model Transformations, 3rd International Workshop on Bidirectional Transformation in Architecture -Based Component Composition, November 14-18, 2009, Changsha, China
⑫ Hiroyuki Kato, An XQuery Fusion with Preserving Document Order, 日本ソフトウェア科学会第26回大会、2009年9月16日～18日、島根大学
⑬ Hiroyuki Kato , Rewriting XQuery to Avoid Redundant Expressions based on Static Emulation of XML Store, 6th Asian Workshop on Foundations of Software, April 6-8, 2009, 東京都、学術総合センター
⑭ Hiroyuki Kato, Introduction to UnQL+, GRACE International Meeting on Bidirectional Transfor mations, December 14-18, 2008, 神奈川、湘南国際センター

〔その他〕
ホームページ等
以下のURLではプロトタイプ実装を公開している。
http://www.biglab.org/fusion/

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