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　２９　年　　６　月　１９　日現在

機関番号：　１２１０２
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０１２６

研究課題名（和文）スキーマ更新に応じたXSLTスタイルシート修正アルゴリズムの開発

研究課題名（英文）Construction of An Algorithm for Correcting XSLT Stylesheets According to Schema Updates

研究代表者

鈴木　伸崇（SUZUKI, Nobutaka）

筑波大学・図書館情報メディア系・准教授

研究者番号：６０３０５７７９

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究では，スキーマ更新に応じてXSLTスタイルシートを修正する問題について考察した．XSLTスタイルシートのサブセットとして，UTT，UTTpat，UTTpat,sel という３つの木変換機のクラスを定義した．まず，UTTpat と UTTpat,sel に対して，本問題はNP困難であることを示した．次に，UTTに対して，スキーマ更新により影響を受けるXSLTスタイルシートの規則を求める多項式時間アルゴリズムを作成し，これを用いて本問題を解くためのアルゴリズムを開発した．本アルゴリズムをJavaで実装して評価実験を行い，本アルゴリズムの有効性を確認した．

研究成果の概要（英文）：In this study, the problem of correcting XSLT stylesheets according to schema updates is considered. As subsets of XSLT stylesheets, three classes of tree transducers UTT, UTTpat, and UTTpat,sel are defined. Firstly, it is shown that the problem becomes NP-hard for UTTpat and UTTpat,sel. Secondly, a polynomial-time algorithm for detecting XSLT rules affected by schema updates is constructed. Based on the algorithm, an algorithm for solving the problem is constructed. The algorithm is implemented in Java and experiments are conducted, which verifies the effectiveness of the algorithm.

研究分野：　情報学

キーワード：　XML　XSLT　木変換機

１．研究開始当初の背景

XML データを継続的に蓄積・管理して使用する場合，格納すべきデータの構造をスキーマで定義しておき，それに妥当なデータを作成・格納することが一般的である．また，時間の経過と共に利用者の要求や使用状況が変化するため，格納すべきデータの構造や種類が変化し，それに応じてスキーマ定義が更新されることも多い．スキーマが更新されると，それまで使用していた XSLT スタイルシートが正常に動作しなくなるため，スキーマの更新内容に合わせて適切に修正する必要が生じる．しかし，スキーマは近年複雑化しサイズも増大しているため，スキーマの更新内容と XSLT スタイルシートとの依存関係を把握して適切な修正を行うのは容易でない．このような状況から，スキーマの更新内容に応じて XSLT スタイルシートを適切に自動修正する効率のよいアルゴリズムが求められている．

これまで，XML のスキーマ更新に関する研究は数多く行われているが，スキーマに対する更新操作や更新に伴う XML データの修正・管理に関するものが中心である．例えば，Kwietniewski や報告者らは，スキーマ更新に応じて XML データを自動的に変換する手法を提案している．Oliveira らは，スキーマ更新によって生じる問題を分類し，それらの問題を検出する手法を提案している．しかし，報告者の知る限り，スキーマ更新に応じた XSLT スタイルシートの自動修正に関する研究は存在しない．一方，申請者らは，スキーマ更新に応じて XPath 式を自動修正するアルゴリズムを提案している．本研究で扱う XSLT は XPath を部分的に含んでいるが，中心的な機能は（XPath には備わっていない）データ変換であり，この機能は複数の規則が互いに関連して実現されるものである．したがって，このアルゴリズムを XSLT スタイルシートの修正に適用することは困難であり，スキーマ更新に応じて適切に XSLT スタイルシートを修正するためには新たなアルゴリズムの開発が必要である．

２．研究の目的

本研究の目的は，スキーマ更新に応じて XSLT スタイルシートを適切に修正する効率のよいアルゴリズムを開発することである．ただし，XSLT スタイルシートは複雑な機能を有しており，本問題は計算困難なことが見込まれることを考慮して，以下を本研究の目的とする．

(1) 本問題が計算困難であることを形式的に証明する．さらに，本問題が多項式時間可解となるための十分条件を明らかにする．
(2) (1)で得られた十分条件の下で動作する，本問題を解くための多項式時間アルゴリズムを求める．
(3) 得られたアルゴリズムを計算機上に実装し，アルゴリズムが所望の効率で動作するか，木変換機に対する修正が実用上有用であるかなどを明らかにする．

３．研究の方法

本研究では，XML データをラベル付き順序木として考える．また，スキーマとして最も広く用いられている DTD を対象とし，DTD に対する更新操作として以下の６種を定める．

- 要素の追加
- 要素の削除
- 要素のネスト
- 要素のネスト解除
- 演算子の追加
- 演算子の削除

また，XSLT を下降型木変換機(top-down tree transducer, 以下単に木変換機)として形式化する．

DTD の更新に応じて XSLT スタイルシートを適切に修正するためには，DTD の更新によって XSLT スタイルシートのどの規則が影響を受けるかを特定する必要がある．そこで，上記の XML および更新操作の形式化に基づいて，まず「DTD の更新によって影響を受ける XSLT スタイルシートの規則」を形式的に定義する．この定義に基づいて，「XSLT スタイルシートの規則 r に対して，r が DTD の更新によって影響を受けるか否か」の計算複雑さについて考察し，この問題の計算困難性の証明，および，この問題が多項式時間で判定できるための十分条件を求める．そして，この十分条件の下で，DTD の更新によって影響を受ける規則を修正する効率の良いアルゴリズムを開発する．さらに，ここで得られたアルゴリズムを Java により計算機上に実装し，アルゴリズムが所望の効率で動作するか，XSLT の規則に対する修正が実用上有用であるかなどについて評価実験を通じて明らかにする．

４．研究成果

（１）アルゴリズムの開発にあたっては，対象となる XSLT スタイルシートや，DTD の更新によって影響を受ける規則などの概念を形式的に定義する必要がある．まず，XSLT スタイルシートについて，その計算能力はチューリング完全であることがわかっており，このままでは DTD の更新によって影響を受ける規則の検出や修正は困難である．そこで本研究では，XSLT スタイルシートのサブセットとして，XSLT の主要な機能を備えた木変換機を定義して用いることとした．より具体的には，以下の３つの木変換機を定義した．

- UTT
- $\mathrm{UTT}^{\mathrm{pat}}$
- $\mathrm{UTT}^{\mathrm{pat,sel}}$

ここで，UTT は，一般的な下降型ランク無し木変換機である．$\mathrm{UTT}^{\mathrm{pat}}$ は，XSLT スタイルシートにおける template 要素のパターン

(pattern)属性に長さ 2 以上の XPath ロケーションパスを記述できる機能を加えたものである（言い換えると，パターンとして記述できるロケーションパスの長さを 1 に限定したものが UTT である）．また，$UTT^{pat,sel}$ は，UTT^{pat} に apply-templates 要素の select 属性の記述を許したものである．

次に，DTD の更新によって影響を受ける規則を形式的に定義した．D を DTD，u を D に対する更新操作列，e を D の要素，r 木変換機の規則とする．また，u(D)は D に u を適用して得られる更新後の DTD を表す．以下のいずれかが成り立つ場合，e において r が u によって影響を受けるという．

A) D において r は e に適用可能であり，u(D)において r は e に適用不可能である
B) D において r は e に適用不可能であり，u(D)において r は e に適用可能である

（2）以上の定義に基づいて，DTD の更新によってどの規則が影響を受けるかを特定する問題の計算複雑さについて考察した．DTD D におけるどの内容モデルに対しても，同じ要素が複数回出現しない場合，D は duplicate-free であるという．得られた結果は以下の通りである．

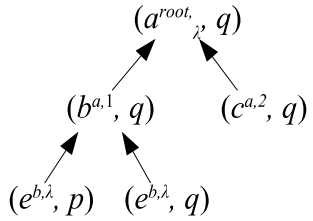
① UTT^{sel} について，DTD の高さを 1 に限定しても本問題は NP 困難である
② UTT^{sel} について，DTD を duplicate-free なものに限定しても本問題は NP 困難である
③ UTT^{pat} について，DTD を duplicate-free なものに限定しても本問題は NP 困難である

①および②は 3SAT 問題からの帰着，③は 3DNF-tautology 問題からの帰着により証明したものである．これらの結果は，UTT^{pat} および UTT^{sel} においては，DTD が高さ 1 や duplicate-free なものに限定されていたとしても，DTD の更新によって影響を受ける規則の発見と修正候補の提示を効率良く行うことが困難であることを示している．

（3）UTT に対して，DTD の更新によって XSLT スタイルシートのどの規則が影響を受けるかを特定するための多項式時間アルゴリズム(FindAffectedRules)を開発した．このアルゴリズムは，DTD D，D に対する更新操作列 u，および UTT に属する木変換機 Tr に対して，以下の手順で DTD の更新によって影響を受ける規則を発見する．

① まず，D と Tr から依存グラフ G を作成する
② 次に，u(D)と Tr から依存グラフ G'を作成する
③ ①と②で得られた G と G'の差分を抽出し，その差分から各要素において影響を受ける規則を求める

ここで，依存グラフとは，DTD の要素と木変換機の状態の組をノードとするグラフであり，これを用いることで各要素に適用可能な規則を効率的に得ることができる．以下は依存グラフの例である．



例えば，(a^{root},q) は，ルート要素 a に対して適用可能な規則があり，適用時の状態が q であることを表している．

次に，上記のアルゴリズム FindAffectedRules に基づいて，DTD の更新によって影響を受ける規則に対する修正候補を提示するアルゴリズムを開発した．このアルゴリズムは，以下のようにして修正候補を求めている．

① アルゴリズム FindAffectedRules を用いて，DTD の更新によって影響を受ける木変換機の規則とその規則が適用される要素をすべて求める
② ①で得られた，規則 r とそれが適用される要素 e の組それぞれに対して，以下を行う
   i. r と e が（1）の A)に該当する場合，r が e に適用可能となるように r を修正し，それを修正候補とする
   ii. r と e が（1）の B)に該当する場合，r が e に適用されなくなるように r を修正し，それを修正候補とする
③ ②で得られた修正候補を提示する

（4）以上で得られた，DTD の更新によって影響を受ける規則に対する修正候補を提示するアルゴリズムを Java により実装し，評価実験を行なった．評価実験に用いた DTD は以下の 2 種である．

- MSRMEDOC（バージョン 2.2.1 および 2.2.2）
- The NLM Journal Publishing Tag Set Library（バージョン 2.3 および 3.0）

また，DTD の更新操作として，2 つのバージョン間の差分を抽出したものを用いた．XSLT スタイルシートは，MSRMEDOC に関しては，人手で作成したものを用いた．また，The NLM Journal Publishing Tag Set Library に関しては，NISO Journal Article Tag Suite (JATS) version 1.0 に基づいて作成した．

このようにして得られた DTD，DTD の更新操作，および XSLT スタイルシートを本アルゴリズムへの入力として用い，本アルゴリズムに関する評価を行なった．まず，本アルゴリズムの出力した修正候補の妥当性について評価を行なった．評価の方法は以下のとおりである．

① 被験者（それぞれの DTD に対して 2 名ずつ）に対して，更新前と更新後の DTD，および XSLT スタイルシートを提示し，その内容から DTD の更新によって影響

を受けた規則を検出してもらう
② 本アルゴリズムを用いて, XSLT スタイルシートの規則に対する修正候補を出力する
③ ②の結果を被験者に見せて, それらが妥当なものか否かを判断してもらう

まず，MSRMEDOC に関して得られた結果は以下の通りである.

被験者	結果	所要時間
1	34/37 (92%)	30 分
2	37/37 (100%)	35 分

ここで，被験者 1 の「結果」列の「34/37」は，アルゴリズムにより修正された規則の総数 37 のうち，被験者により修正が妥当と判断されたものが 34 であることを表している. なお，本アルゴリズムの実行時間は平均 1.5 秒であった．次に，The NLM Journal Publishing Tag Set Library に関して得られた結果は以下の通りである.

被験者	結果	所要時間
1	28/28 (100%)	41 分
2	28/28 (100%)	33 分

また，本アルゴリズムの平均実行時間は 37 秒であった.

以上の結果から，まず，本アルゴリズムが出力した修正候補の妥当性に関して，MSRMEDOC および The NLM Journal Publishing Tag Set Library いずれの場合も，被験者から概ね妥当であるとの評価が得られており，本アルゴリズムは適切な修正候補を出力することが可能であると言える．次に，本アルゴリズムの動作効率について，MSRMEDOC の場合は概ね所望の効率で実行できている．一方，The NLM Journal Publishing Tag Set Library の場合は，やや実行時間が長くなる傾向にある. これは, The NLM Journal Publishing Tag Set Library の DTD は各要素に対する内容モデルのサイズが大きく，計算の手間がより大きくなるためであると考えられる．これらの結果から，アルゴリズムの動作効率に関しては，DTD が複雑な場合において効率の改善を図る余地があると考えられるものの，修正候補の品質に関しては概ね良好であり，当初の目的を概ね達成していると考えられる.

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３件）

① Y. Wu and N. Suzuki, Detecting XSLT Rules Affected by DTD Updates, DBSJ Journal, 査読有, vol.16, 8p.,（採録決定）
② Y. Wu and N. Suzuki, An Algorithm for Correcting XSLT Rules According to DTD Updates, Proceedings of the 4th International Workshop on Document Changes: Modeling, Detection, Storage and Visualization (DChanges 2016), 査読有, 2016, 8p.
DOI: 10.1145/2993585.2993588
③ Y. Wu and N. Suzuki, Detecting XSLT Rules Affected by Schema Evolution, Proceedings of the 15th ACM SIGWEB International Symposium on Document Engineering (DocEng 2015), 査読有, 2015, 143-146
DOI: 10.1145/2682571.2797086

〔学会発表〕（計４件）

① Y. Wu, An Algorithm for Correcting XSLT Rules According to DTD Updates, the 4th International Workshop on Document Changes: Modeling, Detection, Storage and Visualization (DChanges 2016), 2016 年 9 月 13 日, Vienna (Austria)
② Y. Wu, Detecting XSLT Rules Affected by Schema Evolution, the 15th ACM SIGWEB International Symposium on Document Engineering (DocEng 2015), 2015 年 9 月 11 日, Lausanne (Switzerland)
③ Y. Wu, An Algorithm for Detecting XSLT Rules Affected by Schema Updates, 第 161 回 DBS・第 119 回 IFAT 合同研究発表会, 2015 年 8 月 5 日, 東大寺総合文化センター（奈良県奈良市）
④ 呉揚，スキーマ進化によって影響を受ける XSLT 規則の検出手法, 第 7 回データ工学と情報マネジメントに関するフォーラム(DEIM2015), 2015 年 3 月 2 日,（磐梯熱海ホテル華の湯（福島県郡山市）

６．研究組織
(1)研究代表者

鈴木　伸崇（SUZUKI, Nobutaka）
筑波大学・図書館情報メディア系・准教授
研究者番号：60305779