

機関番号：12102

研究種目：若手研究 (B)

研究期間：2008～2010

課題番号：20700077

研究課題名 (和文) スキーマに適用される編集操作列から推測されるXML変換アルゴリズムの開発

研究課題名 (英文) Constructing an XML Transformation Algorithm Inferred from Update Script to Schema

研究代表者

鈴木 伸崇 (SUZUKI NOBUTAKA)

筑波大学・大学院図書館情報メディア研究科・准教授

研究者番号：60305779

研究成果の概要 (和文)：

スキーマとそれに関して妥当なXMLデータが蓄積・管理されている状況を考える。

スキーマが更新された場合、XMLデータは通常妥当でなくなるため、スキーマの更新に応じてXMLデータを妥当なものに変換する必要が生じる。

しかし、このような変換をすべて手作業で行うのは非現実的である。

そこで本研究では、スキーマに適用される編集操作列からXML変換アルゴリズムを推測する問題について考察し、その推測が一般には計算困難であることを示し、ある十分条件の下でK最適編集操作列を求める擬多項式時間アルゴリズムを構成した。

研究成果の概要 (英文)：

Suppose that we maintain a schema and XML documents valid against the schema. An XML document usually becomes invalid when its schema is updated, thus we have to transform the XML document according to the schema update. However, it is impractical to achieve such a transformation by hand. Therefore, in this research we consider the problem of inferring an XML transformation algorithm from an update script to a schema, and showed that such an inference is in general intractable and constructed a pseudo-polynomial time algorithm for finding K optimum update scripts to a schema under a sufficient condition.

交付決定額

(金額単位：円)

	直接経費	間接経費	合計
2008年度	500,000	150,000	650,000
2009年度	1,200,000	360,000	1,560,000
2010年度	600,000	180,000	780,000
年度			
年度			
総計	2,300,000	690,000	2,990,000

研究分野：情報科学

科研費の分科・細目：情報学 メディア情報学・データベース

キーワード：XML, スキーマ進化, XML変換

## 1. 研究開始当初の背景

XML の普及に伴い、近年、大量の XML データが生成・蓄積されている。

ある程度まとまった量の XML データを蓄積・管理する場合、予め格納すべきデータの構造をスキーマとして定義しておき、それに沿った構造の XML データを格納するのが一般的である。

ここで、スキーマとそれに関して妥当な XML データが蓄積・管理されている状況を考える。

この状況でスキーマが更新された場合、XML データは通常妥当でなくなるため、スキーマの更新に応じて XML データを妥当なものに変換する必要が生じる。

しかし、データ量が膨大な場合や、更新操作が複雑な場合など、このような変換をすべて手作業で行うのは非現実的である。

したがって、スキーマの更新に応じて XML データを適切に変換できるアルゴリズムが構成できれば有用であると考えられる。

ここで、更新の際にスキーマに適用される編集操作列は、適切な変換を行う上での有力な手がかりと考えられる。

そこで本研究では、スキーマに適用される編集操作列から XML 変換アルゴリズムを推測する問題について考察を行った。

## 2. 研究の目的

上記 1. で述べたように、スキーマに適用される編集操作列はユーザのデータ変換に関する意図を知る上での有力な手がかりとなると考えられる。

しかし、スキーマの定義や編集操作列の内容によっては、追加・削除すべき要素が一意に定まらないことがある。

以下、これを曖昧な変換と呼ぶことにする。曖昧な変換は、ユーザの意図に沿わない要素の追加や削除を引き起こす恐れがあり、望ましくないと考えられる。

以上の考察から、本研究では、以下の(1)～(4)を研究目的とすることとした。

- (1) スキーマに適用される編集操作列から推測される効率のよい XML 変換アルゴリズムの開発
- (2) (1)の XML 変換アルゴリズムが曖昧であるか否かを決定する効率のよいアルゴリズムの開発
- (3) スキーマに適用される編集操作列と XML データに対して、その XML データに対する最適な(元の XML データの構造が最も保存される)編集操作列を求める効率の良いアルゴリズムの開発
- (4) 上記アルゴリズムに関する評価実験の実施

## 3. 研究の方法

XML で使用されているスキーマにはいくつか種類があり、表現力にも違いがある。

上記 2. で挙げた問題は計算複雑さが相当高いと予想されるため、スキーマとしてまず表現力はあまり高くないものの記述が簡潔で最も普及している DTD について考察する。

そして、DTD より表現力の高いスキーマ(正規木文法等)に対しても効率の良いアルゴリズムが構成できる見込みが得られた場合、そのようなスキーマに対しても考察を行うものとした。

スキーマに適用される編集操作として、スキーマを更新するための操作(例えば、内容モデルにおけるラベルや演算子の追加・削除、文法の生成規則の追加・削除など)を形式的に定義した。

これらの定義に基づいて、以下の手順で研究を行うものとした。

- (1) まず、スキーマ  $G$  とそれに対する編集操作列  $s$  から推測される XML 変換アルゴリズムを開発する。

このアルゴリズムは、 $G$  に関して妥当な任意の XML データを「 $s$  の内容に沿って」 $s(G)$  に関して妥当なものに変換する。

ここで、 $s(G)$  は  $G$  に  $s$  を適用して得

られるスキーマを表す。

なお、このアルゴリズムは曖昧な変換を許すアルゴリズムであり、 $G$ と $s$ の内容によっては曖昧な変換を行うことがある。

- (2) 次に、(1)で定義した、 $G$ と $s$ から推測されるXML変換アルゴリズムが曖昧か否かを決定する多項式時間アルゴリズムを開発する。

ただし、これが困難な場合、 $G$ と $s$ から推測されるXML変換アルゴリズムが曖昧か否かを効率よく決定可能な十分条件を求め、その条件の下で動作する多項式時間アルゴリズムを構成する。

- (3) 更に、スキーマに適用される編集操作列とXMLデータに対して、そのXMLデータに対する最適な編集操作列を求める効率のよいアルゴリズムを開発する。

ただし、これが困難な場合、スキーマに適用される編集操作列とXMLデータに対して、そのXMLデータに対する最適な編集操作列が効率よく得られる十分条件を求め、その条件下で動作する多項式時間アルゴリズムを構成する。

- (4) これらアルゴリズムを計算機上に実装し、動作効率等に関する評価実験を行う。

#### 4. 研究成果

上記3.で示した研究の方法に沿って研究を行った。

これまでに得られている主な結果は以下の通りである。

- (1) DTDとそれに適用される編集操作列から推測される、多項式時間で動作するXML変換アルゴリズムを構成した。
- (2) 上記(1)のアルゴリズムが曖昧か否か(XML変換が一意に決まるか否か)を決定する問題がPSPACE困難であることを示した。
- (3) 上記(1)のアルゴリズムが曖昧か否かを効率よく決定可能であるための十分条件を求め、その十分条件の下で曖昧さ判定

を行う多項式時間アルゴリズムを構成した。

- (4) DTDに適用される編集操作列とXMLデータに対して、そのXMLデータに対する最適な(元のXMLデータの構造が最も保存される)編集操作列を求める問題がNP困難であることを示した。
- (5) 編集操作列の長さを1に限定した場合における、 $K$ 最適編集操作列(コストが小さいものから順に $K$ 個の編集操作列)を求める擬多項式時間アルゴリズムを構成した

更に、上記で得られたアルゴリズムをRubyを用いて実装し、動作効率や変換の適切さなどに関する評価実験を行った。

この評価実験を行うためには、スキーマとそれに関して妥当なXMLデータを相当数用意する必要がある。

このため、XML Data Repository等インターネット上で入手可能なスキーマとそれに関して妥当なXMLデータを収集して用いた。

これらのスキーマやXMLデータを実装したアルゴリズムに適用し、動作効率や変換の適切さに関して評価したところ、概ね期待通りの結果が得られたが、一部の場において必ずしも適切な変換が得られないことが示唆されている。

このため、これらの点について今後更なる検討を行う予定である。

#### 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3件)

N. Suzuki, "An Algorithm for Inferring K Optimum Transformations of XML Document from Update Script to DTD", IEICE Transactions on Information and Systems, vol. E93-D, no. 8, pp. 2198-2212, Aug. 2010 (査読有り)

機関リポジトリ URL:

<http://hdl.handle.net/2241/107914>

N. Suzuki and Y. Fukushima, "An XML Transformation Algorithm Inferred from an Update Script between DTDs," IEICE Transactions on Information and Systems, vol.E92-D, no.4, pp.594-607, Apr. 2009 (査読有り)

N. Suzuki, "On Inferring K Optimum Transformations of XML Document from Update Script to DTD", Proceedings of the 14th International Conference on Management of Data (COMAD 2008), pp.210-221, 2008 (査読有り)

## 6. 研究組織

### (1) 研究代表者

鈴木 伸崇 (SUZUKI NOBUTAKA)

筑波大学・大学院図書館情報メディア研究科・准教授

研究者番号：60305779