M-009

XML 文書に適合する RELAX NG スキーマの自動生成実験

An experiment on the automatic generation of RELAX NG schema from XML documents

稲葉 健治† 野口 健一郎‡
Kenji Inaba Kenichiro Noguchi

1．まえがき

XML 文書には、その構成規則である文書型の定義を持たないものもある。そのような XML 文書に対して、文書型定義を自動生成できれば、文書を修正したときなどに、その文書が構成規則を満たした妥当なものかどうかのチェックが可能になる。また自動生成は、人間が生成するときに入り込みやすいミスを防ぐことも期待できる。

文書型定義としてシンプルな構造を持つ RELAX NG スキーマを用い、それを自動生成するシステムを実験した。XML 文書を入力し、その構造を解析して、RELAX NG スキーマを生成するようにした。さらに、同じスキーマに従うはずの文書を複数入力することにより、RELAX NG スキーマを次第に洗練していくようにした。

2．システム概要

複数個の XML 文書を入力し、入力した XML 文書すべてに対し適合する文書型を生成*する。文書型には XML Schema と同等の機能を持ちながらもよりシンプルな構造の RELAX NG を採用する。生成の流れを図 1 に示す。

* すなわち、入力の XML 文書は生成された文書型からみて「妥当な XML 文書」である。

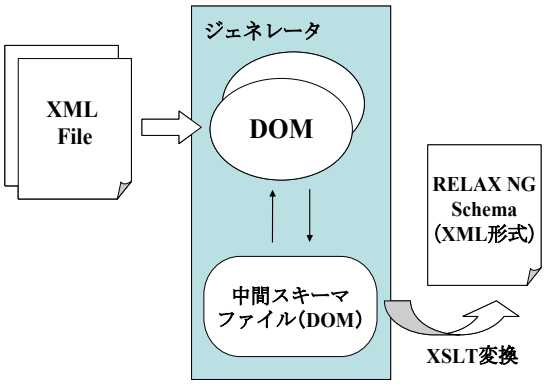


図 1　システム構成図

3．研究課題

（1）文書型の帰納、生成方式
（2）人間が介在する方法

4．解決策

4.1　ジェネレータにおける内部表現

内部表現には XML 文書を動的に扱える DOM を利用した。

† 神奈川大学理学部情報科学科（現在　（株）日立 INS ソフトウェア）
‡ 神奈川大学理学部情報科学科

さらに RELAX NG を生で扱うのではなく、DOM でアクセスし易やすいような中間スキーマファイルに対して文書型を構築し、最後に XSLT を用いて完全な RELAX NG スキーマに変換するようにした。

4.2　要素型の生成

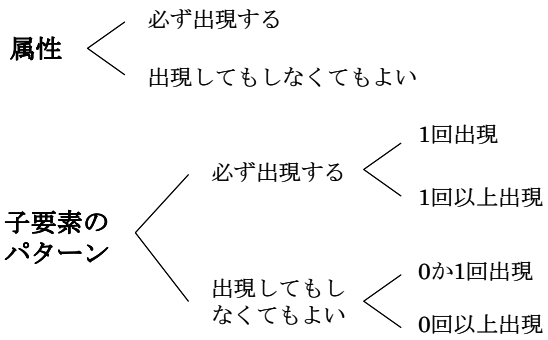
ある要素が初めて出現したとき、その要素が持つ属性群と子要素のパターンを以下のようした。

```
<define name="要素名">
  <element name="要素名">
    <attribute name="属性名" />
    <ref name="defineの参照名" />
      …
  </element>
</define>
```

RELAX NG 構文の define 要素、element 要素、attribute 要素、ref 要素（define の参照）を用いて定義していくようにした。

4.3　文書型の洗練

同じ要素が 2 回目以降出現したときは、一度定義した文書の型を修正していく方法を取る。属性、子要素パターンは以下のように分類した。



分類したら、中間スキーマの段階では use 属性を用いて "require", "optional", "oneOrMore", "zeroOrMore"で区別するようにした。

4.4　RELAX NG の出力

XSLT 変換においては次の変換を行うようにした。

1. ルート要素が複数定義されている場合は<choice>の追加
2. 名前空間処理時の ns, combine 属性の追加
3. 0 か 1 回出現時の<optional>の追加
4. 1 回以上出現時の<oneOrMore>の追加

5. 0 回以上出現時の<zeroOrMore>の追加
6. テキストデータを持つ場合
   6.1 string 型の指定
   6.2 混合要素を示す<mixed>の追加
7. 空要素は<empty />の追加

4.5 人間が介在する方法

自動生成において、生成が難しい部分はあらかじめユーザが自身で定義した define 要素を読み込ませることによって、そちらを最優先したスキーマの生成も可能にした。

５．生成例

例として、図 2 の XML 文書 2 つを実際に読み込ませた。

```
<?xml version="1.0" encoding="Shift_JIS"?>
<書籍 xmlns="http://bookstore.com">
  <グループ name="暗号">
    <書籍名 管理No="333-2121">暗号技術</書籍名>
  </グループ>
  <グループ name="XML">
    <書籍名 管理No="333-2125">XML入門</書籍名>
    <書籍名 管理No="333-2126">XML応用</書籍名>
  </グループ>
</書籍>
```

```
<?xml version="1.0" encoding="Shift_JIS"?>
<書籍 xmlns="http://bookstore.com">
  <グループ name="Java">
    <販売エリア>3F</販売エリア>
    <書籍名 管理No="333-2190">Java言語</書籍名>
  </グループ>
</書籍>
```

図２　２つの入力 XML 文書

結果として得られた RELAX NG スキーマを図 3 に示す。スキーマと XML 文書を妥当性検証器 Jing にかけてみたところ、妥当であるという結果が得られた。

６．考察

（１）実用性について

シンプルな構造の XML 文書ならばユーザが求めるようなスキーマを得ることができた。千行を越す膨大な XML 文書でも妥当性を保つことができた。また、複数の XML 文書を読み込ませる際、ルート要素が異なっているものでも対応することができる。

（２）より高度な文書型について

複雑な構造になると、本来求めたいと思うスキーマを得ることが難しい。その例として属性値や要素が持つテキストデータの列挙、要素出現における choice などが挙げられる。現在の生成方式では、ルート要素以外では、本来 choice となるべきところが optional の連続で表現される。

（３）データ型について

RELAX NG スキーマは外部で定義されたデータ型を使用することができる。本実験プログラムは XML Schema Part 2 の string のみに対応している。この適用範囲を広げることにより、実用性を向上させることは課題である。

```
<?xml version="1.0" encoding="Shift_JIS"?>
<grammar datatypeLibrary="http://www.w3.org/2001/XMLSchema-datatypes"
xmlns="http://relaxng.org/ns/structure/1.0">
  <start>
    <ref name="書籍"/>
  </start>
  <define name="書籍">
    <element name="書籍" ns="http://bookstore.com">
      <oneOrMore>
        <ref name="グループ"/>
      </oneOrMore>
    </element>
  </define>
  <define name="グループ">
    <element name="グループ" ns="http://bookstore.com">
      <attribute name="name">
        <data type="string"/>
      </attribute>
      <optional>
        <ref name="販売エリア"/>
      </optional>
      <oneOrMore>
        <ref name="書籍名"/>
      </oneOrMore>
    </element>
  </define>
  <define name="書籍名">
    <element name="書籍名" ns="http://bookstore.com">
      <attribute name="管理No">
        <data type="string"/>
      </attribute>
      <data type="string"/>
    </element>
  </define>
  <define name="販売エリア">
    <element name="販売エリア" ns="http://bookstore.com">
      <data type="string"/>
    </element>
  </define>
</grammar>
```

図３　出力 RELAX NG スキーマ

７．今後の課題

（１）XML Schema Part 2 のデータ型の適用範囲拡大
（２）列挙やリスト型への対応
（３）choice や interleave への対応
（４）Relaxer との連動

参考文献

[1] RELAX NG Specification
http://www.oasis-open.org/committees/relax-ng/spec.html
[2] RELAX NG Tutorial
http://www.oasis-open.org/committees/relax-ng/tutorial.html
[3] 屋内恭輔 安陪隆明 著：「XML スキーマ書法」
毎日コミュニケーションズ (2003)