L_037

XML 文書に適合する RELAX NG スキーマの自動生成実験 －スキーマ仕様のサポート範囲の拡大－

An experiment on the automatic generation of RELAX NG schema from XML documents - Extending the support of schema features -

小松崎 裕介† 野口 健一郎‡
Yusuke Komatsuzaki Kenichiro Noguchi

1．まえがき

現在、XML 文書を用いて動作するアプリケーションが増え続けている。受け渡された XML 文書の受け入れの可否を判断するために、その構造を定義した文書型があることが望ましい。しかし、XML 文書に対応した文書型が必ずしも作られているとは限らない。そこで、XML 文書をもとにして文書型を自動生成する研究を行ってきた[1]。しかし、複雑な XML 文書になるとユーザが本来求めるような文書型を得ることが難しかった。本研究では、属性値のデータ型のサポート、要素の出現順序への対応、要素と属性の名前空間への対応などの改良を行い、より高度な文書型を生成できるようにした。

2．システム概要

入力した XML 文書から適合する文書型の生成を行う。生成の流れを図 1 に示す。複数の XML 文書を読み込んだ場合にもそれら全てに適合する文書型を生成する事が出来る。

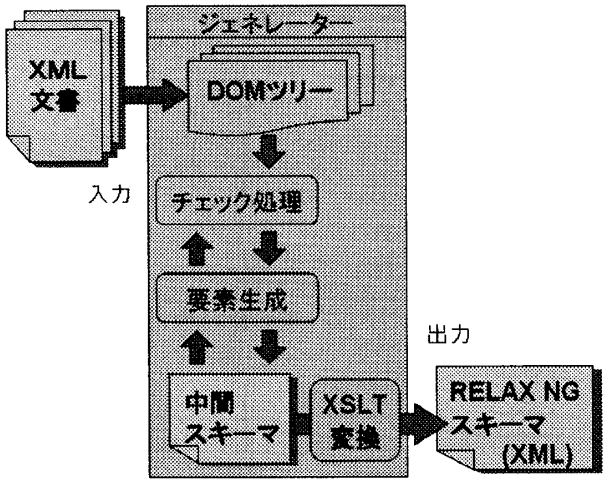

図 1 システム構成図

スキーマ言語には XML 用の簡潔なスキーマ言語である RELAX NG を採用した。内部の処理では XML 文書のパーサとして DOM を使用し、XML 文書は内部では DOM ツリーで表される。また、生成したスキーマの内部表現を RELAX NG スキーマの形式に変換するのに XML 文書用のスタイル変換言語である XSLT を使用した。

† 神奈川大学理学部情報科学科（現在 テクノバン株式会社）

‡ 神奈川大学理学部情報科学科

3. 研究課題

(1) RELAX NG スキーマ仕様のサポート範囲の拡大
(2) 要素の出現順序の判別方法

4. 解決策

4.1 サポート範囲の拡大

これまでの研究[1]ではサポートされていなかった、属性値のデータ型、要素の出現順序への対応、要素と属性の名前空間に対応させた。

(1) 要素型の生成方法

ある要素が始めて出現したときに、RELAX NG 構文の define 要素、element 要素、attribute 要素、data 要素、ref 要素(子要素定義)を使って定義した。生成結果は図 2 のようになる。

```
<define name="要素名">
  <element name="要素名">
    <attribute name="属性名">
      <data type="属性のデータ型">
    </attribute>
    <ref name="参照する要素名"/>
    …
  </element>
</define>
```

図 2 要素の生成例

(2) 属性値のデータ型

RELAX NG では XML Schema Part2 で定義されているデータ型を利用することができる。属性値のデータ型を入力 XML 文書から次の 4 つに判別した。

integer	整数型
double	倍精度浮動小数点数型
date	日付(ISO8601 形式(yyyy-MM-dd))
string	文字列(上記以外)

(3) 要素の出現順序への対応

同名要素が複数ある場合のその子要素の出現順序を調べ、中間スキーマ内で属性を用い同順、順不同の定義をするようにした。

(4) 名前空間への対応

要素や属性に名前空間が指定されている場合、ns 属性で指定するようにした。

4.2 要素の出現順序の判別方法

一度定義した要素と同名の要素を定義する際には、その

子要素の要素名と出現順序を比較、判定によりその親要素に"group"、"interleave"の印をつけた。最終的なスキーマを得るとき、この印を頼りにこれらのタグを生成する。それぞれの意味は次の通りである。

group	子要素の出現順序が全て同順の場合
interleave	子要素の出現順序が順不同の場合
(なし)	特に指定しない場合

例えば図3のように、①要素名とその出現順序を既に中間スキーマに定義済みの要素型と比較、②順不同と判断し、生成したorder要素にinterleaveの印をつける、といった操作を行っている。

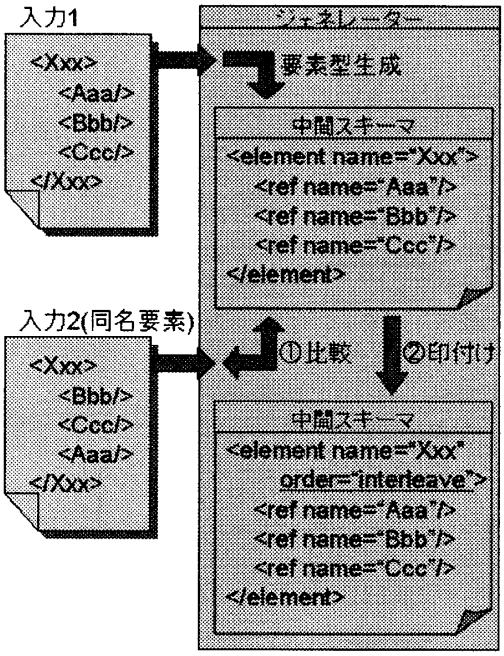

図3 interleave適用例

5. 生成例

次の2つのXML文書を入力する。

```
<?xml version="1.0" encoding="Shift_JIS" ?>
<書籍 xmlns:b="http://book.com"
      xmlns:m="http://magazine.com">
 <グループ No="1000">
 <雑誌名 m:管理No="1248">月刊JAVA</雑誌名>
 <書籍名 b:管理No="1322">JAVA言語入門</書籍名>
 </グループ>
 <グループ No="1010">
  <書籍名 b:管理No="1323">C言語入門</書籍名>
  <雑誌名 m:管理No="1249">月刊C言語</雑誌名>
 </グループ>
</書籍>
```

```
<?xml version="1.0" encoding="Shift_JIS" ?>
<書籍>
 <グループ No="4000">
  <書籍名>JAVA言語入門</書籍名>
  <書籍名>JAVA言語応用</書籍名>
 </グループ>
 <グループ No="4444">
  <雑誌名>週刊C言語</雑誌名>
  <雑誌名>週刊JAVA</雑誌名>
 </グループ>
</書籍>
```

結果として次のスキーマが得られる（一部のみ示す）。

```
・・・
<define name="グループ">
  <element name="グループ">
    <attribute name="No">
      <data type="integer"/>
    </attribute>
    <interleave>
      <zeroOrMore>
        <ref name="雑誌名"/>
      </zeroOrMore>
      <zeroOrMore>
        <ref name="書籍名"/>
      </zeroOrMore>
    </interleave>
  </element>
</define>
<define name="雑誌名">
  <element name="雑誌名">
    <optional>
      <attribute name="管理No"
                 ns="http://magazine.com">
        <data type="integer"/>
      </attribute>
    </optional>
    <data type="string"/>
  </element>
</define>
・・・
```

6. 考察

(1) 生成の範囲について

本研究では要素の出現順序、属性値のデータ型についてサポート範囲を拡大したが、要素の選択や列挙等に未対応である。その為、複雑な構造のXML文書では適合はするが、ユーザが望むようなスキーマを生成することが難しい。

(2) 要素の出現順序について

子要素の出現順序によって<group>、<interleave>を生成している。しかしすでにinterleaveと判定された要素順が出現した後にinterleaveとならない要素順がきた時に、要素の出現回数が0回でも許可するように変換している。これでは<interleave>のタグの意味が薄れてしまう。

解決策として新たに定義を生成して要素順を定義する方法も試みたが、シンプルなXML文書の場合には無駄な定義の多いスキーマ文書となってしまった。選択や列挙を用いることによって解決すると思われる。

(3) 関係ないXML文書の除外方法について

本研究では新たなXML文書を読み込むたびにユーザに入力の可否を問い合わせている。しかし入力文書の構造から対象外の文書かどうかを判断できることが望ましい。

7. 今後の課題

(1) 列挙やリスト型への対応

(2) 対象外とするXML文書の判定方法

参考文献

[1] 稲葉健治・野口健一郎：「XML文書に適合するRELAX NGスキーマの自動生成実験」FIT（情報科学技術フォーラム）2004.

[2] RELAX NG Specification
http://www.oasis-open.org/committees/relax-ng/spec.html

[3] RELAX NG Tutorial
http://www.oasis-open.org/committees/relax-ng/tutorial.html