社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143 (38)
2007／7／3

XMLの整形出力問合せ言語PPXにおけるイレギュラーXMLデータの自動フォーマット方式の提案

金 哲† 遠山 元道‡

†慶應義塾大学 大学院 理工学研究科 開放環境科学専攻 〒223-0061 横浜市港北区日吉 3-14-1
‡慶應義塾大学 理工学部 情報工学科 〒223-0061 横浜市港北区日吉 3-14-1
E-mail: †tetsu@db.ics.keio.ac.jp, ‡toyama@ics.keio.ac.jp

あらまし: 既存のXMLデータを変換する方法は汎用プログラミング言語を利用するもの，スタイルシート言語を利用するもの，変換言語を利用するものと問合せ言語を利用するもの，独自なフォーマット言語によるものなどがある．本論文ではフォーマット言語の一つであるPPX(Pretty Printer for XML) 問合せ言語においてイレギュラーXMLデータに対してフォーマットを行う自動フォーマット方式を提案する．提案する自動フォーマット方式はレイアウト自動決定ルールに基づいてイレギュラーXMLデータに対してレイアウトする方法を自動決定して自動フォーマットを行う．実験では提案した自動フォーマット方式を人工的なデータ，実世界のデータに対して適用し，数行の記述量でイレギュラーXMLデータのレイアウトができることを示した．

キーワード: XML，イレギュラーXMLデータ，PPX，TFE

An Automatic Formatting Method for Irregular XML data by Pretty Printer for XML

Zhe JIN† Motomichi TOYAMA‡

†School of Science for OPEN and Environmental Systems, Faculty of Science and Technology, Keio University.
‡Department of Information and Computer Science, Faculty of Science and Technology, Keio University.
E-mail : †tetsu@db.ics.keio.ac.jp, ‡toyama@ics.keio.ac.jp

Abstract: Many existing methods are used to converting the XML instance into another all kinds of XML instance, such as the programming language, the Style Sheet language, the conversion language, the query language and the Format language and so on. In this paper, we propose an automatic formatting method to layout irregular XML data by the PPX(Pretty Printer for XML) query language that is the format language. The Automatic Formatting Method generates layout method to layout irregular XML data based on layout decision rule. The automatic formatting method is tested in the experiment which contains artificial data and real data. The results shows that the automatic formatting method can work correctly and efficiently.

Keyword: XML, Irregular XML data, PPX, TFE

1 まえがき

XML(eXtensible Markup Language)で記述された大量のデータの存在はXMLデータからHTMLへの変換技術を重要な研究課題にしている。既存の方法は汎用プログラミング言語によるもの，スタイルシート言語によるもの，変換言語によるもの，問合せ言語によるもの，独自なフォーマット言語であるTFE(Target Form Expression)の応用によるものなどがある．TFEを応用するPPX(Pretty Printer for XML)[1, 2]はXMLデータをHTMLへの変換を問合せ言語である．例えば，Q1はGENERATE句のレイアウト表現式でレギュラーXMLデータに対して実際に処理を行う変数をTFEのレイアウト指定演算子と組合わせて完全指定してレイアウトを行う固定フォーマット方式を使う．

```
Q1: GENERATE html
[$j/year/text() ,
 [ {$j/id/text() ,
    $j/title/text()@{width=400}} !
      [$l/name/text() !
       $l/univ/name/text()@{width=500}
], ]! ]!
FOR $i in db2('db.TEXT')/papers
FOR $j in $i/paper
FOR $l in $j//author
```

Q1で示したように固定フォーマット方式はXSLT，JAVAなどの方法などは違い，簡単な記述でXMLデータをHTMLにレイアウトができる．しかし，イレギュラーXMLデータに対して固定フォーマットを行うことができない．また，Q1から見るとレイアウト表現式で変数とTFEのレイアウト指定演算子をすべて完全に指定してレイアウトを行うので記述量が多くになる．

本論文では固定フォーマット方式より簡単なフォーマット方式でイレギュラーXMLデータもレイアウトができる自動フォーマット方式を提案する．本方式はレイアウト自動決定ルールの利用によってイレギュラーXMLデータに対して自動フォーマットを行ない、表示する。Q2はGENERATE句のレイアウト表現式で&ルール演算子(レイアウト自動決定ルール)を利用してpapersノードが持つ全てのイレギュラーXMLデータをレイアウトを行う自動フォーマット方式を使う．

```
Q2: GENERATE html [ & ( $i ) ],
    FOR $i in db2('paper.TEXT')/papers
```

Q2で示したように自動フォーマット方式は一般ユーザはもちろん，高度なプログラミング・スキルを有するユーザにとってもXMLデータ構造などを直接意識することなくレイアウトの洗練に集中するだけでXMLデータの変換作業を容易に行うようにする．

本論文の構成は2章で関連研究を述べる．3章で自動フォーマット方式を述べる．4章ではレイアウト自動決定ルールを述べる．5章では評価を述べ，最後にまとめと今後の課題を述べる．

2 関連研究

XMLデータをHTMLに変換する既存の方法は次のとおりである．

2.1 汎用プログラミング言語によるもの

JAVA，PERL，PHP，C++など汎用プログラミング言語を利用する場合，XMLパーサ(DOM)によってXMLデータをメモリ上に要素のツリーを作成し，これに基いて新しいドキュメントオブジェクトを作ってから得られたXMLデータをHTMLに変換する．

2.2 変換言語によるもの

変換言語であるインフォテリア社のiXSLT [3]はXMLデータをXSLT規格に準拠して変換を行うXSLTプロセッサであり，変換元となるXMLデータとXSLT規格に則ったXSLTスタイルシートを読み込み，XSLTスタイルシートに記述されたルールに従って別のXMLもしくはHTMLやCSV等の構造化データへ変換を行う．

2.3 スタイルシート言語によるもの

スタイルシート言語であるXSLはXMLデータからデータを取り出し，XMLデータの見ばえに関する情報を与えて表示するための仕様である．XSLのうち，データを取り出すためのXSLT，表示のための物理構造を記述するためのXSL(FO)とXML

インスタンスに表示情報を付加するための CSS がある．

2.4 問合せ言語によるもの

問合せ言語である Oracle 社の XSQL [4] は XSQL サーブレットが JDBC 接続でデータベースに問い合わせをし，データベースからデータを XML データとして取得し，その結果を XSQL サーブレットに返す．XSQL サーブレットがデータベースから出力された XML データを XSL で表記されたスタイルシートによって HTML などに変換する．

2.5 独自なフォーマット言語にもの

TFE はターゲットリストの拡張であり，結合演算子，反復演算子，装飾演算子などのレイアウト指定演算子を持つ一種の式である．TFE を利用する SuperSQL [5] は関係データベースの出力結果を構造化し，HTML 以外に XML などへの出力を可能とする SQL の拡張言語である．しかし，SQL を基づいた SuperSQL は XML データを HTML にレイアウトすることはできない．TFE を応用する PPX 問合せ言語は XML データを HTML にレイアウトすることができる．

3 自動フォーマット方式

本研究では XML データをレギュラー XML データとイレギュラー XML データに分ける．PPX 問合せ言語はレギュラー XML データに対して固定フォーマット方式を利用してレイアウトを行ない，イレギュラー XML データに対して自動フォーマット方式を利用してレイアウトを行う．本論文では自動フォーマット方式を論じる．

3.1 PPX 問合せ言語

PPX は XQuery に基づいて探索された XML データを SuperSQL の TFE を拡張したものによって XML データを HTML へのレイアウトを行う問合せ言語である．本問合せ構文は GENERATE 句，FOR 句，WHERE 句，Order By 句などがある．

3.2 システム構成

本システムは図 1 で示したように構文解析部，リスト構造生成部，レイアウト編集部からなる．PPX

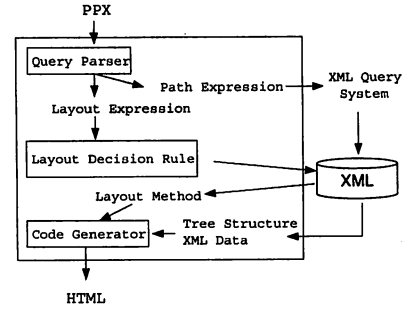

図 1: システムの流れ

問合せの問合せ文は構文解析部によってレイアウト表現式とパス表現式に分ける．パス表現式は実際に XML データの探索を行い，フラットなリスト構造の XML データはリスト構造生成部でレイアウト表現式に従って再構造化を行い，最後にレイアウト編集部に渡され様々な表構造を持つ HTML に変換される．

4 レイアウト自動決定ルール

レイアウト自動決定ルール (&ルール演算子) はレイアウトされる XML データから DTD を抽出し，様々な DTD パス表現式集合を持つ連結式が生成する．生成された連結式集合が持つ DTD パス表現式集合は XML データに対してパス表現式マッチングによってレイアウトされる．

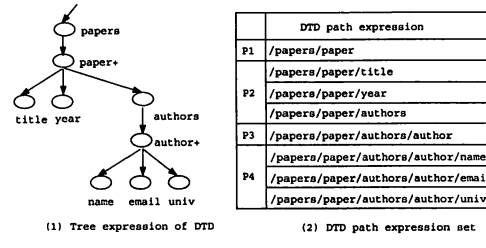

図 2: DTD と DTD パス表現式集合

4.1 DTDを抽出

XMLデータからDTDを抽出する方法[6]は既存の方法を利用する．1章のQ2は抽出したDTD(図2(1))からDTDパス表現式集合(図2(2))が生成，レイアウトされるXMLデータからXMLデータパス表現式集合などが生成する．

その後，DTDパス表現式にマッチするXMLデータパス表現式が持つ反復出現する部分XML(サブ木)数を統計する．また，反復出現するノードが持つXMLデータ数も統計する．図3はDTDパス表現式にマッチするXMLデータパス表現式が持つ部分XMLであり，一部を示した．

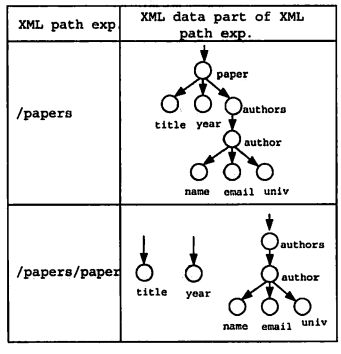

図3: XMLパス表現式が持つ部分XML

4.2 連結式

連結式はDTDパス表現式集合(P(d))，DTD演算子(operator)，部分XML数(m)，XMLデータ数(n)で構成され，以下のように表わせる．

連結式 $= \{(P(d), operator, m, n) | L, d, m, n \in I\}$

本連結式は以下のような連結方法を持つ．

連結式 ∈{縦連結，横連結，縦横連結，横縦連結，横横連結，縦縦連結 }

次は連結式について簡単に説明する．

(1) “＋”(“＊”)などの反復演算子を持つ図2(2)のP(1)のDTDパス表現式によって構成された連結式

連結式 ={(P(1), +, m, n)}

は縦連結と横連結などの連結方法を持つ．さらに，縦横連結，横縦連結，横横連結，縦縦連結などの連結方法によって反復出現する部分XML数(反復出現するXMLデータ数)を何等分折り返して連結することもできる．

(2) “？”などの反復出現しない演算子を持つ図2(2)のP(2)のDTDパス表現式によって構成された連結式

連結式 ={(P(2), , m, n)}

は縦連結と横連結などの連結方法を持つ．

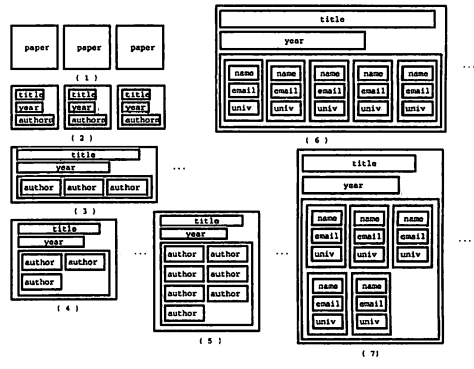

図4: レイアウト生成

次は連結式が持つDTDパス表現式集合にマッチするXMLデータは連結式集合によってレイアウトされる．例えば，図2(2)のP(1)にマッチする最初の部分は1章のQ2のGenerate句の最も外側の縦反復演算子を参照して横方向に反復しながら並べる(図4(1))．また，図2(2)のP(2)にマッチする部分は前の部分が横方向に並べ過ぎを避ける為に縦連結(図4(2))が行う．その後，図2(2)のP(3)にマッチする部分は横方向に反復しながら並べる(図4(3))際に反復出現する部分XML数(XMLデータ数)の量によって何等分ずっつ折り返して縦連結(図4(4))か，或いは何等分ずっつ折り返して横連結(図4(5))することによって全体が横方向に並べ過ぎるのを避ける．最後，図2(2)のP(4)にマッチするXMLデータは縦連結(図4(6))が行う際に反復出現するXMLデータ数を何等分折り返して縦連結(図4(7))することによって全体が横方向に並べ過ぎるのを避ける．ここで何等分するため，部分XML数，XMLデータ数，文字サイズなどを参照するが詳しく説明は省略する．

4.3 パス表現式マッチング

連結式集合はトップダウンアプローチ，ボトムアップアプローチ，ハイブリッドアプローチなどのパス表現式マッチング方法を利用してXMLデータに対してレイアウトを行う．

4.3.1 トップダウンアプローチ

トップダウンアプローチはパス表現式集合が持つレイアウト方法を利用してXMLデータのルートノードから葉ノードの方向に接頭語パスマッチングを行いながらレイアウトする．ここでマッチではないパス表現式が持つXMLデータはXListで表示する．例えば，図5(1)のXMLデータは図4のP1，P2，P3，P4などパス表現式集合順番に接頭語マッチングによってレイアウトされる．

4.3.2 ボトムアップアプローチ

ボトムアップアプローチはパス表現式集合が持つレイアウト方法を利用してXMLデータの葉ノードからルートノードの方向に接尾語パスマッチングを行いながらレイアウトする．ここでマッチではないパス表現式が持つXMLデータはXListで表示する．例えば，図5(2)のXMLデータは図4のP4，P3，P2，P1などパス表現式集合順番に接尾語パスマッチングによってレイアウトされる．

4.3.3 ハイブリッドアプローチ

ハイブリッドアプローチは図5(3)で示したようにトップダウンアプローチとボトムアップアプローチを組合わせてインスタンスXMLデータに対してそれぞれマッチングを行ないながらレイアウトを行う．例えば，図5(3)のXMLデータは図4のP1，P2，P3，P4などパス表現式集合順番に接頭語マッチングによってレイアウトされた続いて図4のP4，P3などパス表現式集合順番に接尾語パスマッチングによってレイアウトされる．

5 評価

本研究で提案した自動フォーマット方式はDTD演算子だけ参照によるレイアウト自動決定ルールを使う．スペースのことで詳しい結果は示してないがDBLP，SIGMOD Recordなど実際のXMLデータ，

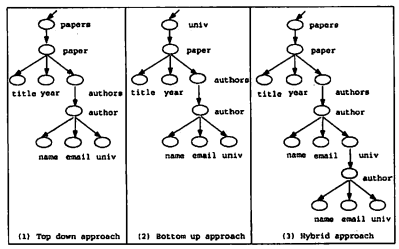

図5: パス表現式マッチング方法

そして人工XMLデータからHTMLへの変換実験を行い，有効性を確かめた．また，固定フォーマット方法，従来の方法と比較を行った．

5.1 自動フォーマット方式と固定フォーマット方式の比較

固定フォーマット方式はレギュラーXMLデータに対してレイアウトを行ない，自動フォーマット方式はレギュラーXMLデータとイレギュラーXMLデータに対してレイアウトを行う．このようなXMLデータをレイアウトする為に，固定フォーマット方式はGENERATE句のレイアウト表現式でレギュラーXMLデータに対して実際に処理を行う変数をTFEのレイアウト指定演算子と組合わせて完全に指定する必要がある．しかし，自動フォーマット方式はレイアウトを行うXMLデータ部分をレイアウト自動決定ルール演算子で指定すれば結構である．また，固定フォーマット方式はDTDに適合しないXMLデータは見なければ指定ができない．しかし，自動フォーマット方式はXMLデータを詳しく見なくても簡単に指定できる．さらに，XMLデータを探査する為に固定フォーマット方式はテキストノードまで指す完全なパス表現式を利用しなければならない．しかし，自動フォーマット方式はテキストノードまで指してない不完全なパス表現式の利用も可能である．探索されたXMLデータに対して固定フォーマット方式はレイアウト表現式でTFE演算子を詳しく指定できるので元のXMLデータ構造とまったく違う破壊的な構造変換ができる．しかし，自動フォーマット方法はレイアウト表現式でTFE演算子を詳しく指定できないので生成したデータ構造変換は単純である．さらに，表構造変換のために固

定フォーマット方式に基づいてレイアウト自動決定ルール演算子を何ヶ所も利用しなければならない．

5.2 従来方法との比較

PPX 問合せ言語は XSLT，SuperSQL とは次のような相違点がある．

5.2.1 XML データ探索

SQL に基づいた SuperSQL は関係データベースからデータを探索することができるが XML データを探索することができない．パス表現式に基づいた PPX，XSLT などは構造化された XML データ，フラットなリスト構造を持つ XML データを探索することができる．

5.2.2 XML データ構造変換

PPX，XSLT などは XML データから特定の部分を抽出し，別の XML データ構造に整形することができる．実際にプログラミング言語である XSLT はパターンマッチングとルール間の競合解消を基本とする変形プロセスの理解が要求され，作成することは必ずしも容易でない．PPX は探索されたフラットなリスト構造の XML データをレイアウト表現式で変数と拡張 TFE のレイアウト指定演算子を組合わせて完全指定は XML データを様々な構造を持つ再構造化が行うことができる．

5.2.3 表構造を持つ HTML への変換

SuperSQL と PPX はともに TFE を利用してデータを様々な表構造を持つ HTML へのレイアウトができる．XSLT は XML データを HTML に変換のためのルールを記述していくことが必要であり，多様な表構造を持つ HTML へのレイアウトするためにスタイルを書き直さなければならない．XSLT で変換した XML データを CSS を使ってレイアウトしてブラウザに表示する．PPX はレイアウト表現式で結合演算子を利用して XML データを HTML に変換する．ここでは同じ構造を持つ XML データに対して結合演算子の位置変換だけで様々な表構造を持つ HTML へのレイアウトが生成する．さらに，装飾演算子の利用によって HTML 表に生成される際に XML データの装飾もできる．

6 終わりに

本論文では PPX 問合せ言語においてイレギュラー XML データをフォーマットする自動フォーマット方式を提案した．本フォーマット方式は DTD 演算子だけの参照によるレイアウト自動決定ルールはイレギュラー XML データに対してレイアウトする方法を自動生成し，フォーマットをが行う．また，少ない記述量で XML データに対してレイアウトができることから初心者や一般ユーザも簡単に利用できることを確認した．

今後の研究課題として自動フォーマットを行うトップダウンアプローチ，ボトムアップアプローチ，そしてハイブリッドアプローチなどを実装し，最適化が必要である．

参考文献

[1] 金哲，慎祥揆，有澤達也，遠山元道: XML データの整形出力処理系，信学技報，vol.103, no.190，pp.37-42，Jul. 2003.

[2] 金哲，慎祥揆，有澤達也，遠山元道: XML データの整形出力処理系における DTD の利用，信学技報，vol.105，no.172，pp.187-192，Jul. 2005.

[3] Lionel Villard, Nabil Layaida: iXSLT: An Incremental XSLT Transformation Processor for XML Document Manipulation, WWW2002，pp.7 − 11，Mai 2002.

[4] M. Kifer, W. Kim, Y. Sagiv: Querying object-oriented databases, Proc. ACM SIGMOD，pp.393 − 402，1992.

[5] M. Toyama: SuperSQL: An Extended SQL for Database Publishing and Presentation, Proc. ACM SIGMOD，pp.584-586，1998.

[6] Jong-Seok Jung, Dong-Ik Oh, Yong-Hae Kong, Jong-Keun Ahn: Extracting Information from XML Documents by Reverse Generating a DTD, EurAsia-ICT 2002: 314-321.