4J-02

XMLインスタンスからの統一的DTDの作成

鬼沢 友和　安井 浩之　松山 実
武蔵工業大学

1 はじめに

コンピュータネットワーク上において，構造化された文章やデータを管理，交換，配布するためのマークアップ言語としてXML（eXtensible Markup Language）[1]が注目されている．従来のHTMLに比べて文書要素の名前が自由に定義できるなどXML文書を作成する際の自由度が高いという反面，内容が類似するものでも文書の構造は作成者により大きく異なることもある．XML文書をコンピュータで処理する際，文書型定義（Document Type Definition）から文書の構造を判断するため，異なる構造の数だけDTDを定義しなければならない．ここでは無駄に数の多い定義を避けるため，複数のXMLインスタンスから統一化されたDTDを生成する支援ツールの開発を試みている．

2 手作業によるXML化の問題

本来XMLインスタンスはDTDに基づいて作成されるが，XMLには，タグの構造を自由に決定ができるという特徴がある．そのため，既存の文書やデータ（構造化データ）からXML文書に変換する際，先にXMLインスタンスを作成し，インスタンスに合わせてDTDを作成するという場合も多い．しかし，この手順を用いると僅かな違いしかない構造化データに対しても個々に固有のDTDが必要となり効率が悪い（図1）．また，同じ構造化データでも別の人がXML化すると異なった構造のインスタンスになってしまう可能性がある（図2）．このようにDTDもXMLインスタンスも一意には決まらないという問題がある．

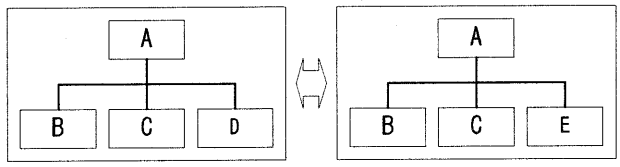

図1. 固有のDTD

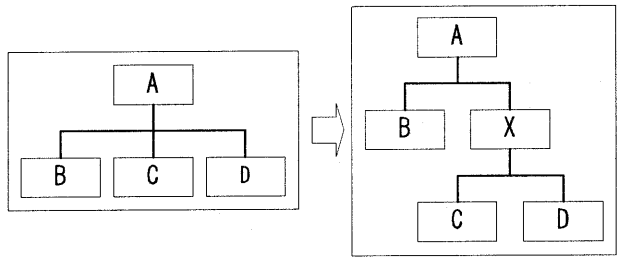

図2. 同じ意味でも異構造

3 支援ツールの概要

ここに報告するXML文書作成支援ツールは実際の構造化データからXML文書化する際にDTDの制約を気にすることなく，インスタンスから作成することを可能にし，既に存在するXML文書に情報を追加したり，まとめるなどの作業を可能な範囲で自動化するものである．その機能を以下に示す．ただし本ツールはXMLエディタの機能を持たないので，テキストエディタやXMLエディタで作成されたXMLインスタンスを別途用意する必要がある．

1)XMLインスタンスからのDTDの自動生成

DTDの無いインスタンスの構造を解析して適合するDTDを自動生成する（図3）．

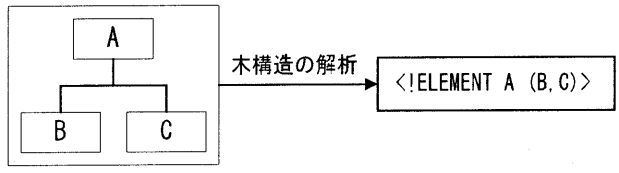

図3. DTDの自動生成-1

Production of unified DTD from XML Instances

Tomokazu Kizawa, Hiroyuki Yasui, Minoru Matsuyama

Musashi Institute of Technology

2)複数のインスタンスからのDTDの自動生成

内容は似ているが構造が異なるインスタンス同士を比較して全てに適合するDTDを生成する．ただし<名前>と<氏名>のように一つにまとめる方が良いとユーザが判断した場合には同じものとして処理する．具体的にはユーザが選択，もしくは木構造の幅や深さを比較してユーザが望むものをコンピュータが選択し，それを基準にその他のインスタンスと比較していく（図4）．

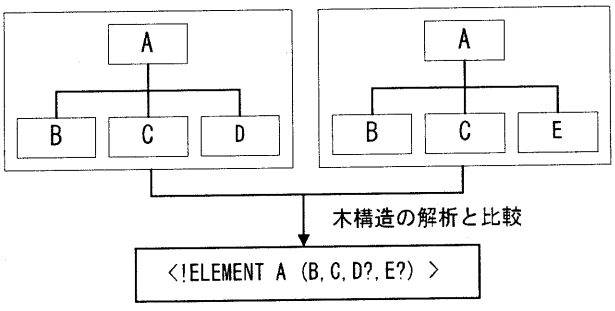

図4. DTDの自動生成-2

3)インスタンスの構造の統一

2)で生成されたDTDに基づいて該当するインスタンス全てを適合するように書き換える．またエンティティの宣言や属性の追加などの作業も行う．これにより対象のインスタンス全てが同じ構造になる．

4 試験的実装

本格的な実装の前段階として試験的実装を行った．

4.1 実装環境

Java2が動作するWindows98上でDOM Level1[2]，SAX version1の各パーサAPIを呼び出すことが可能な環境で行った．

4.2 実装例

2つの構造の異なるXMLインスタンスから共通のDTDを生成し，そのDTDに適合するようにインスタンスを整形した（図6，図7）．

なお，図6の例ではinstance Iは統一DTDに適合しているので書き換えは行われない．

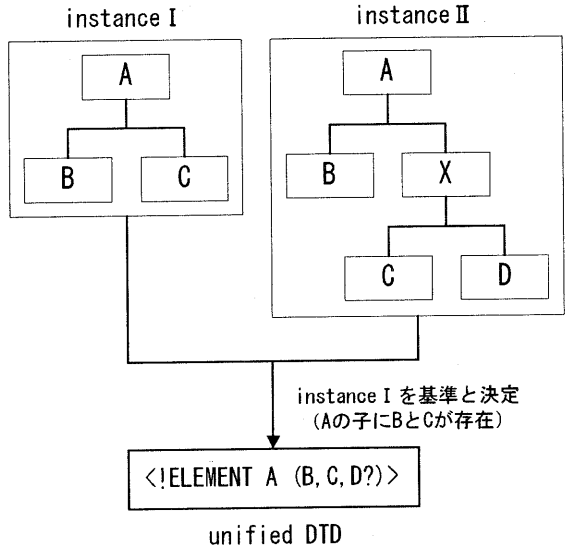

図6. 実装例（DTDの統一）

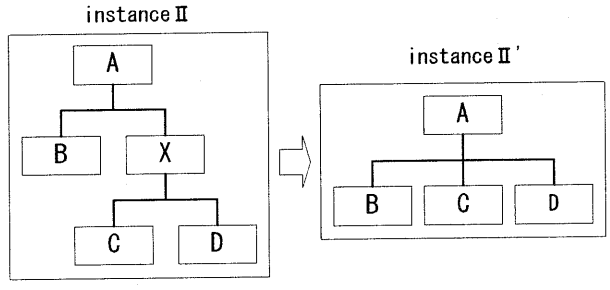

図7. 実装例（インスタンスの変更）

5 おわりに

今回の試験的実装は単純なモデルを対象としたが本来は大量の文書を処理することを目的としているため，最終的な実装段階では膨大なデータに対処できるようなシステムにする必要がある．またユーザから指示を与える場面が多いので，統一の自動化アルゴリズムの検討や，容易な操作を目指す必要がある．

参考文献

[1] W3C, "Extensible Markup Language (XML) 1.0", http://www.w3.org/TR/REC-xml/ , 1998

[2] W3C, "Document Object Model (DOM) Level 1 Specification", http://www.w3.org/TR/REC-DOM-Level-1/ , 1998