明示的な制約記述による Web コンテンツ一貫性管理支援*

澤菜津美 (学籍番号 200621318)
研究指導教員: 森嶋厚行
副研究指導教員: 杉本重雄

1. はじめに

現在，Web サイトの構築手法としては，大きく分けて，次の 2 つの手法がある．**手法 A**:各ページのコンテンツ直接作成，**手法 B**:ページとは別の情報源 (DB 等) からオンデマンドで動的にページを作成するシステムを構築．一般的に，Web サイトに含まれるコンテンツは互いに関連していることが多い．例えば，大学の研究室の Web サイトでは，各人のページに研究室の名前，住所，電話番号が含まれており，これらは一致するはずである．また，研究室メンバの発表論文の一覧は，研究室の発表論文一覧のサブセットであることが一般的である．このような関連を表す制約を，本研究ではコンテンツ一貫性制約と呼ぶ．Web サイトのコンテンツの変更があった場合には，これらのコンテンツ一貫性制約が保持されるように更新されることが望ましい．しかし，Web サイトの規模が大きくなるにつれ，手法 A ではコンテンツ一貫性の維持が困難になる．したがって，ある程度大規模な Web サイトは手法 B で構築される．

問題は，(1) ページ数の少ない Web サイトでは手法 A の方が構築コストが小さいこと，および，(2) サイトの規模が大きくなってしまうと，手法 A で構築されたサイトから手法 B への移行コストが大きいこと，である．したがって，あらかじめ大規模サイトであることが分かっており，かつ，構築のためのリソースがあり，かつ，コンテンツ一貫性が維持されないことが致命的な問題になる (ビジネス関係など) 場合のみ，最初から手法 B で構築される事が多い．それ以外の場合では，手法 B の採用を決断するケースは少なくなってしまう．

以上の技術的背景と問題点を踏まえ，本研究では，直接ページ作成のようなバックエンドの情報源によってコンテンツの管理が行われないような場合にでも，Web コンテンツの一貫性維持を容易に行うための仕組みを提案する．

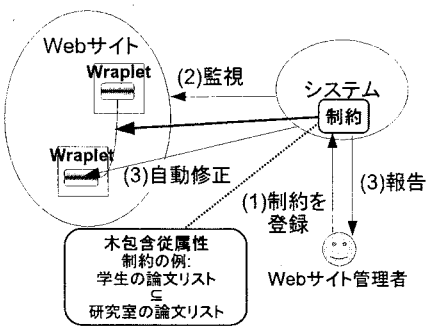

図 1　コンテンツ一貫性制約を用いた Web サイト管理手法

2. コンテンツ一貫性制約を利用した Web サイト管理手法

図 1 はコンテンツ一貫性制約を用いた Web サイト管理の仕組みを表したものである．以下にその手順を述べる．まず，利用者がコンテンツ一貫性制約を登録する (図 1 (1))．登録は 2 ステップである．まず，ステップ 1 として，利用者が監視したい HTML ページに対し，XML 形式で書かれた構造データ (structured data) を抽出するためのラッピング規則を記述する．ラッピング規則の記述には，本研究で提案する要素技術であるラッピング言語 Wraplet を用いる．Wraplet によるラッピング規則 (Wraplet 式と呼ぶ) の記述は，利用者自身が行っても良いが，本研究で提案する Wraplet 式生成支援手法を利用することもできる．次に，ステップ 2 として，抽出した構造データ間の制約を記述する．制約の記述には，本研究で提案するコンテンツ一貫性制約の一種である木包含従属性で表す事ができる．木包含従属性によって，例えば，「学生の論文リストは研究室の論文リストのサブセット」という制約を記述することができる．制約が登録されると，システムは定期的もしくは Web サイトの更新が行われた際などに Web サイトのチェックを行い，先に発見しておいた制約と照らし合わせて，制約が破られていないかどうか調べる (図 1 (2))．その際，もし制約違反を発見したら，Web サイト管理者に報告もしくは自動修正を行う (図 1 (3))．

* "Support of the Web Content Integrity Management with Explicit Constraint Descriptions" by Natsumi SAWA

(a) 果物在庫リスト

```
<ul>
  <li>りんご,10</li>
  <li>みかん,20</li>
  <li>桃,30</li>
</ul>
```

(b) Wraplet 式でパースした結果

```
<在庫>
    <果物><名前>りんご</名前><数量>10</数量></果物>
    <果物><名前>みかん</名前><数量>20</数量></果物>
    <果物><名前>桃</名前><数量>30</数量></果物>
</在庫>
```

図 2　Wraplet 式の適用例

3. 本アプローチ実現のための要素技術

本研究では，本アプローチを実現するための重要な要素技術として，次の 3 点に関して具体的に研究を行った．

3.1 木包含従属性

コンテンツ一貫性制約の一種として木包含従属性 (tree inclusion dependencies) を定義した．木包含従属性は，URL u と v から Wraplet を利用して抽出した Web コンテンツを表す木構造 t_u と t_v 間において，木に含まれるノードの集合間の包含関係を表すためのものである．

3.2 ラッピング言語 Wraplet

HTML で記述された Web コンテンツから構造データを抽出するためのラッピング言語 Wraplet を提案した．既存のラッピング手法 [1][2] と異なる特徴としては，簡易な構文やライブラリの工夫を行っている点，および，HTML に組み込んで利用可能な点がある．例えば，図 2(a) のような果物在庫を表現する HTML ページがあり，ラッピングした結果の XML データが図 2(b) であるとする．この変換を行うための Wraplet 式は次のようになる．

```
在庫:/
  {果物:#li/
    [名前:_val(#anytext)\,, 数量:_val(#num)]}
```

Wraplet 式は，XML データの要素名であるラベル (在庫，果物など) と XML 要素に対応する HTML データの範囲を指定するパターン (`#li`, `_val(#anytext)\`, など) を持つ．また，入れ子の式で，XML データの木構造を指定する．

3.3 Wraplet 式生成支援手法

Wraplet 式によって HTML データを XML に変更するラッパの生成支援手法の研究を行った．本研究では，HTML ラッパ構築の問題を，XML データからの DTD 生成の問題の一般化としてモデル化し，ラッパ選択の基準として，DTD の適切性の評価尺度である MDL コスト [3] と，ラッピングの具体性尺度として本研究で提案する S-Value の相乗平均を利用したスコア付け手法を提案した．実験では，提案したスコア付け手法により，人が適切であると考えたラッピング (図 3 の X) が最も高いスコアを獲得した．これは，提案手法が適切であることを示している．

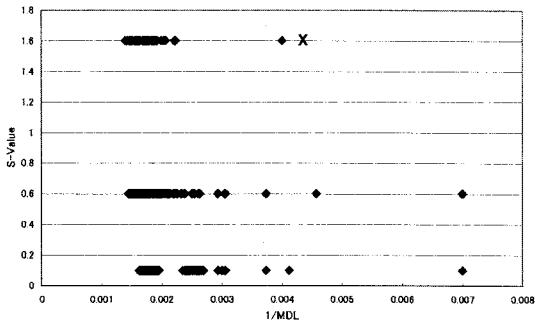

図 3　スコア分布

4. まとめ

本研究では，Web サイトの HTML データから XML 形式で書かれた構造データを抽出し，コンテンツ一貫性制約を明示的に指定および保持することにより Web サイト管理を行う手法を提案した．特に，本アプローチを実現するための重要な要素技術に関して研究を行い，コンテンツ一貫性制約の維持を行うための技術基盤を構築した．これにより，本アプローチの実現可能性を示した．

文献

[1] L. Liu, C. Pu, and W. Han. XWRAP: An XML-enabled wrapper construction system for web information sources. International Conference on Data Engineering (ICDE), pp. 611-621, 2000.

[2] Arvind Arasu, Hector Garcia-Molina. Extracting Structured Data from Web Pages. ACM SIGMOD International Conference on Management of Data, pp.337-348, 2003.

[3] Minos Garofalakis, Aristides Gionis, Rajeev Rastogi, S. Seshadri and Kyuseok Shim. XTRACT: A System for Extracting Document Type Descriptors from XML Documents. Proc. of the 2000 ACM SIGMOD international conference on Management of data, pp. 165-176, 2000.