Web ページを対象とした包含従属性発見支援手法とその評価*

高橋公海 (学籍番号 200821663)
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

近年，Web サイトを通じた情報発信が広く普及し，コンテンツの量も増加している．Web の特徴の一つは分散管理であるが，一方で，その特徴がコンテンツの一貫性維持を困難とする一因となっている．例えば，大学の研究室の Web サイトでは，各構成員が自分のホームページ上で研究論文リストを公開することが多いが，これらの論文リストの間には矛盾が多く見られる等といった問題がある．一般に，コンテンツの一貫性を維持するためには，バックエンドに DB システムを配置し，DB に格納されているデータから Web ページを作成するアプローチがとられる．しかし，筑波大学の Web サイトを対象とした予備調査 [1] では，バックエンドに DB 等をもたずに手作業で管理されている Web サイトも数多く存在することが分かっている．また，ある 2 つの Web サイトが同じ内容を含むにも関わらず，管理者が別であるために統一的に管理されていないということもよく見られることである．例えば，ある学科の Web サイトの入試説明会情報に変更があった際に，その学科に属する学部の Web サイトでも同様の変更を行わなければならないが，このような場合に確実な変更を保証することは多大な労力を要する．

この問題に対し，DB 等をバックエンドに持たない Web コンテンツの一貫性管理支援を目的としたシステムの研究開発を行ってきた [1][2][3][4]．本システムは，Web コンテンツ間に成立すべき制約 (例えば，研究室の構成員である学生の Web サイト上に掲載されている論文リストは，研究室の Web サイトに掲載されている論文リストのサブセットである，等) を与えることにより，既存の Web コンテンツに対して**後付けで**コンテンツの一貫性管理が行えるようにするものである．

図 1 は本システムの仕組みを表したものである．利用手順は次の通りである．

(1) 利用者がコンテンツ一貫性制約を登録する．
(2) システムは定期的，もしくは更新が行われた際などに Web サイトのチェックを行う．
(3) 先に登録しておいた制約と照らし合わせ，制約が破られていないかどうかを調べる．
(4) 制約違反を発見した場合は，Web サイト管理者に報告，またはコンテンツの一貫性が保たれるよう自動的に修正するといった対応を行う．

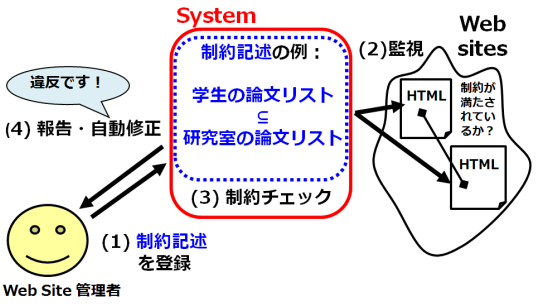

図 1　コンテンツ一貫性制約を用いた Web サイト管理

本システムを利用することにより，既存の Web サイトを DB をバックエンドにした Web サイトに再構築しなくとも，コンテンツの一貫性管理が可能になる．しかし，本システムを利用する際には，既存の Web コンテンツからそこに存在する制約を発見し，システムに与えなくてはならない．膨大な Web コンテンツを対象とした場合には手作業で一つずつ制約を発見することは現実的ではなく，既存の Web コンテンツに対する制約発見支援が必須である．

本研究では，Web コンテンツ間の制約として，Web ページの HTML 要素や XML 要素間 (以降では総称して **Web ページ要素**と呼ぶ) の**包含従属性** [5] に焦点を当て，システムを用いてその発見を支援するための手法を提案する．

2. Web ページを対象とした包含従属性発見支援

包含従属性の発見を支援するためには，その根拠となる包含関係が Web コンテンツ間で成立している事を発見することが有効である．しかし，伝統的に包含従属性が議論されてきた RDB と異なり，Web コンテンツにおいては意図せぬ間違いなどによって完全な包含関係が成立しているとは限らない．そこで，本研究では次の 2 つを行う．(1) 包含関係に**包含率** という概念を導入する．具体的には，

* “A Method to Find Inclusion Dependencies in Web Pages and Its Evaluation” by Masami TAKAHASHI

Web ページ要素 x に対応する単語集合 X の要素のうち割合 c が Web ページ要素 y に対応する単語集合 Y に含まれるとき，$X \subseteq_c Y$ (X は Y に包含率 c で含まれる) と表記する．なお，X が Y に包含率 c 以上で含まれるとき，$x \subseteq_{\geq c} y$ と表記する．(2) 包含率 c が与えられたとき，既存の Web コンテンツに存在する包含率 c 以上で成立する包含関係をもれなく発見する手法を提案する．

本研究では，包含率 c が一定以上の Web ページ要素の組合せを効率よく発見するための重要な要素技術として，次の 3 点に関して研究を行った．

2.1 包含率を導入した包含関係の発見アルゴリズム

全ての Web ページ要素組がどれだけの包含率で成立するかを，1 度のディスクスキャンで計算する *Single-Pass* アルゴリズムを提案した．これは，厳密な包含関係を発見するための J.Bauckmann ら [6] のアルゴリズムを拡張し，包含率を導入したものである．

2.2 包含関係の厳密な計算を行う候補を減らすためのフィルタリング手法

包含関係発見処理を効率化するフィルタリング手法の研究を行った．一般に，対象となる全 Web ページ要素組の集合 $pairs$ のサイズは膨大なものになるため，全ての可能な組 $(x, y) \in pairs$ に対して $x \subseteq_{\geq c} y$ を確認することは多大な労力を要する．そこで本研究では，c が与えられたとき，低コストであらかじめ可能性の無い組を除去することができるフィルタ $filter(x, y, c)$ を提案した．これは，$x \subseteq_{\geq c} y$ が成立する可能性が無いときのみ偽を返す述語である．すなわち，$pairs' = \{(x, y) | (x, y) \in pairs \wedge filter(x, y, c)\}$ としたとき，$|pairs'| < |pairs|$，かつ，$pairs'$ 中の要素の組と包含率 c に対する結果が，元の $pairs$ を用いた結果と変わらないようなフィルタである．具体的には，$filter(x, y, c)$ を実現する手法として，interval-based filter (以下，*i-filter*)，bit-signature-based filter (以下，*b-filter*) という 2 つの手法を開発した．

i-filter では，Web ページ要素 x, y に対応する単語集合 X, Y をあらかじめソートした単語列 $L(x), L(y)$ の範囲を利用してフィルタリングを行う．$L(x), L(y)$ の共通な範囲が $c \cdot |L(x)|$ 以上なければ $x \subseteq_{\geq c} y$ が成立する可能性もないと判断できる．

b-filter は，各 Web ページ要素 x, y に対してビットシグネチャ $b(x), b(y)$ をあらかじめ求めておき，$b(x)$, $b(y)$, c を用いて $filter(x, y, c)$ の判定を行う方法である．

実験では，コンテンツが分散管理されている実際の Web サイトを対象とし，*i-filter* と *b-filter* がどれだけ組合せの数を減らすことが可能であるか，比較を行った．実験の結果，*i-filter* では 15% 前後，*b-filter* では 85% 前後組合せ数を削減することができた．特に *b-filter* はフィルタとして効果が高いと考えられる．

2.3 発見した包含関係のスコアリング手法

発見した包含関係のスコアリングを行うための重要度スコアを導入し，その計算を行うルールを定義した．重要度スコアを導入した動機は次の通りである．すなわち，ページに含まれる全ての Web ページ要素の組合せの包含関係を計算する場合，自明な包含関係とそうでない包含関係が混在し，価値の高い包含関係が埋もれてしまうからである．定義したルールを用いて，出力される各包含関係に対し重要度スコア ($0 \leq s \leq 1$) を計算し，自明でない包含関係の発見を支援する．

3. まとめ

本研究では，Web コンテンツで成立する包含従属性の発見を支援するために，次の 3 つの要素技術の開発を行った．(1) 包含率を導入した包含関係の発見アルゴリズム．(2) 包含関係の厳密な計算を行う候補を減らすためのフィルタリング手法．(3) 発見した包含関係のスコアリング手法．これらの要素技術に関して実際の Web ページを用いた実験を行い，本アプローチの実現可能性を示した．

文献

[1] 澤菜津美 他．コンテンツ一貫性制約を用いた Web サイト管理手法の提案．DEWS2007，7 pages, 2007.

[2] Natsumi Sawa. et al. Wraplet: Wrapping Your Web Contents with a Lightweight Language. Proc. IEEE SITIS' 2007.

[3] 澤菜津美 他．情報統合利用を目的とした HTML ページのラッピング支援．DEWS2008, 2008.

[4] 高橋公海 他. Web コンテンツ一貫性管理支援ツールの開発. 第 70 回情報処理学会全国大会講演論文集 (第 5 分冊), pp. 189-190, 2008.

[5] Serge Abiteboul. et al. Foundations of Databases. Addison-Wesley 1995.

[6] J. Bauckmann. et al. Efficiently Computing Inclusion Dependencies for Schema Discovery. InterDB'06 (ICDE Workshop), 2006.