ヒューリスティクスを用いたWebコンテンツの包含従属性発見の効率化

弓矢 英梨佳 †　高橋 公海 ‡　森嶋 厚行 ‡　杉本 重雄 ‡　北川 博之 ††

筑波大学 図書館情報専門学群 †　筑波大学大学院 図書館情報メディア研究科 ‡　筑波大学大学院 システム情報工学研究科 ††

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，管理しなくてはならないコンテンツの量が増加している．Webサイトは分散管理されていることが多いが，これらの分散管理されたコンテンツの一貫性の維持は重要な問題である．例えば，大学のWebサイトでは，大学全体と各学部のWebサイトのそれぞれで大学までのアクセス情報や電話番号が掲載されている場合があるが，これらの情報に一貫性がないと，利用者に混乱を引き起こしてしまうなどの問題が生じる可能性がある．このような問題に対処するためにはバックエンドにDBを持つようなシステムを構築することが一般的であるが，現実には分散管理されたWebコンテンツも多く，一貫性の維持は容易ではない．

これに対して我々は，既存のWebサイト群に後付けでコンテンツ一貫性制約を与えることで，分散管理されたWebコンテンツの一貫性の維持を低コストで実現する手法を提案している[1]．図1は，コンテンツ一貫性制約を利用したWebサイト管理手法の概要である．次に具体的な利用手順を示す．(1) まず，Webサイト管理者が，システムに制約を登録する．(2) システムはWebサイトの更新が行われた時などに，登録されている制約に違反が生じていないか確認を行う．(3) 違反が発見された場合，システムはWebサイト管理者に報告を行うか，違反された部分を自動修正する．

本手法を利用すれば，既存Webサイトのコンテンツ一貫性を管理することができる．しかし，システムに登録する制約を手作業で発見することは困難であるため，制約発見の支援が重要となる．

我々は論文[3]において，コンテンツ一貫性制約の一種である包含従属性[2]の発見を支援するために，Webページ中のHTML要素やXML要素（以下総称してWebページ要素）間の包含関係を計算機を用いて効率的に発見するアルゴズムの提案を行った．この手法は，シグネチャを用いたフィルタを利用して厳密な調査をしなければならないWebページ要素の対の数を削減することで，全体としての計算コストを削減する．

しかし，この手法では，フィルタを利用するにしても全てのWebページ要素の対を調査しなければならず，その対の総数が膨大になるという問題がある．

本稿では，ヒューリスティクスを利用して，再現率を多少犠牲にする代わりに，調べなければならないWebページ要素の対の数を削減する問題について議論する．

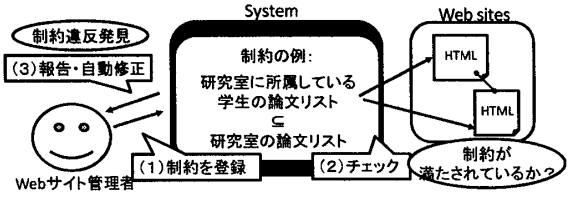

図1　コンテンツ一貫性制約を利用したWebサイト管理

2. 包含率

Webサイトでは間違いや表記の揺れも多いため，厳密な包含関係を計算するのは現実的ではない．したがって，我々はWebページ要素間の包含関係に包含率[3]という概念を導入した．Webページ要素間の関係 $x \subseteq y$ の包含率とは，Webページ要素 x 内の単語集合 X がWebページ要素 y 内の単語集合 Y に含まれている単語の割合 c である．

$$c = \frac{|X \cap Y|}{|X|}$$

我々の問題は，c がある定数以上であるようなWebページ要素の対の発見処理を効率化するという事である．

3. ヒューリスティクスを用いた包含関係発見

本稿では，調べなければならないWebページ要素の対の数を削減するために，Webページ要素間の包含関係の発見過程を次の2段階に分割し，その中の (1) ReductionPhaseにおけるヒューリスティクスの利用について検討する．

(1) Reduction Phase：Webページの集合から，ヒューリスティクスを用いて包含関係があるWebページ要素が含まれていると思われるWebページの対だけを作成し列挙する．

(2) Discovery Phase：(1) で列挙されたWebページの対から得られるWebページ要素の対を全て列挙し，その中から包含率がある定数以上である対を求める．

今回検討したヒューリスティクスは次の通りである．

検討するヒューリスティクス. 異なるWebページに存在するWebページ要素の対 (e_1, e_2) 間に，Webページ要素の包含関係 $e_1 \subseteq e_2$ が存在する時，それらが存在するWebページ間には，それほど長くないリンクのパスが存在する．□

本ヒューリスティクスは一般性のあるものであるが，本稿の4章で示す実験では異なるWebサイト間のWebページにあるWebページ要素の対だけを対象に実験を行うため，異なるWebサイト間のWebページにあるリンク (外部リンクと呼ぶ) を中心にした例を説明する．外部リンクのリンク元から内部リンク (同一サイトのページ間のリンク) をたどる上限を s 回，リンク先からたどる上限を d 回とした時，図2の例では，WebサイトA，WebサイトBのWebページに対して，$s = 1$，$d = 1$ の場合に列挙されるWebページの対を示している．矢印はリンクを示しており，矢印の

Efficient Discovery of Inclusion Dependencies Existing in Web Contents Based on Heuristics

Erika Yumiya† Masami Takahashi‡ Atsuyuki Morishima‡ Shigeo Sugimoto‡ Hiroyuki Kitagawa††

Sch. of Library and Information Science, Univ. of Tsukuba.†

Grad. Sch. of Library, Information and Media Studies, Univ. of Tsukuba.‡

Grad. Sch. of Systems and Information Engineering, Univ. of Tsukuba.††

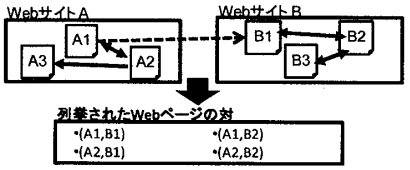

図 2 $s = 1$, $d = 1$ の時に列挙される Web ページの対の例

始点をリンク元，終点をリンク先の Web ページ，点線の矢印を外部リンク，実線の矢印を内部リンクとする．本ヒューリスティクスでは，本来ならば Web サイト A と Web サイト B 間の全ての Web ページの対である $3 \times 3 = 9$ の対から作成される Web ページ要素の対を Discovery Phase に渡すところを，Web ページの 4 つの対だけを列挙して，これらから生成される Web ページ要素の対だけを Discovery Phase に渡すということになる．

4. 予備実験

このヒューリスティクスを適用した影響を調査するために，実際の Web サイトを用いて簡単な予備実験を行った．対象とする Web サイトは，筑波大学の情報学群 (inf)[4]，情報学類 (coins)[5]，知識情報図書館学類 (klis)[6]，情報メディア創成学類 (mast)[7] である．それぞれの Web ページ数は，inf が 60 ページ，coins が 185 ページ，klis が 36 ページ，mast が 39 ページである．実験時には，前章で説明した s と d の値を変化させた．実験の手順は次の通りである．

(1) inf-klis，inf-mast，inf-coins の異なる Web サイト間の Web ページの対の集合 P を求め，さらにこれらの対から，それらの異なるページに存在する Web ページ要素の対の集合 (*BaseSet*) を生成した．次に，*BaseSet* に直接 Discovery Phase を適用し，包含率 $c \geq 0.7$ の Web ページ要素の対の集合 (*CorrectResult*) を生成した．

(2) P に対して Reduction Phase を適用した．今回は，外部リンクのリンク元からたどる上限を $0 \leq s \leq 10$ 回，リンク先からたどる上限を $1 \leq d \leq 10$ 回と変化させた．Reduction Phase で列挙された Web ページの対から，それらの異なるページに存在する Web ページ要素の対の集合 (*ReducedSet*) を生成した．

(3) *ReducedSet* と *CorrectResult* から再現率 (*Recall*)，*ReducedSet* と *BaseSet* から削減率 (*ReductionRate*) を算出した．それぞれの計算式を次に示す．

$$Recall = \frac{|CorrectResult \cap ReducedSet|}{|CorrectResult|}$$
$$ReductionRate = 1 - \frac{|ReducedSet|}{|BaseSet|}$$

結果：実験結果を図 3 に示す．x 軸は再現率，y 軸は削減率である．図を簡潔にするため d の値は一部のみ掲載しているが，全体的に d の値を大きくするほど再現率があがるというグラフになっている．

考察：ある程度高い再現率を維持しつつ，高い削減率を持つ事が出来ることが理想であるが，本予備実験の範囲では，$s \geq 1$，$d \geq 2$ の場合に再現率 80%以上得る場合があることが示された．また，削減率に関しては，$s = 1$，$d = 1$ の場合，67%の削減率を実現できるが，さらにたどると，削減率は 20%程度まで減少する．再現率 80%以上で，再現率と削減率の相乗平均が一番高いのは，$s = 1, d = 2$ の時の結果である．したがって，リンクをたどればたどるほど良い結果が出るとは限らないと言えるため，適切なパラメータの設定が重要な課題となる．今回の予備実験では，適切なパラメータの設定により，85%の再現率を維持しつつ，削減率 41%を得られる場合があることがわかった．

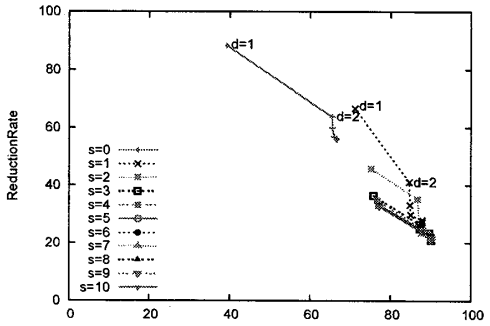

図 3 実験結果

5. まとめと今後の課題

本稿では，ヒューリスティクスを用いて Web コンテンツの包含従属性の発見支援を効率的に 行う問題について議論した．特に，Web ページ間のリンクのパスの距離に注目したヒューリスティクスに関して，実データを用いた予備実験を行いその振る舞いを検証した．その結果，再現率を多少犠牲にしつつも，包含従属性が成立しないと思われる Web ページの対の削減をある程度実現できる可能性があることを示した．

今後の課題は，より効果が高いと考えられるヒューリスティクスの検討，およびより大規模な Web データを対象とした実験などがあげられる．

謝　辞

本研究の一部は科学研究費補助金特定領域研究 (#21013004)，基盤研究 (A)(#21240005)，基盤研究 (B)(#19300081)，若手研究 (B)(#20700076) による．

参考文献

1) 澤菜津美，森嶋厚行，飯田敏成，杉本重雄，北川博之．“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, 2007 年 3 月．
2) Serge Abiteboul, Rechard Hull, Victor Vianu. “Foundations of Databases,” Addison-Wesley Publishing Company, 1995.
3) 高橋公海，森嶋厚行，杉本重雄，北川博之．“ビットシグネチャを用いた Web ページの包含従属性発見の効率化，” WebDB Forum 2009, 2009 年 11 月．
4) 筑波大学情報学群. http://www.inf.tsukuba.ac.jp/
5) 筑波大学情報学群情報科学類. http://www.coins.tsukuba.ac.jp/
6) 筑波大学情報学群知識情報・図書館学類. http://www.klis.tsukuba.ac.jp/
7) 筑波大学情報学群情報メディア創成学類. http://www.mast.tsukuba.ac.jp/