Webデータを対象とした包含従属性発見支援のための ランキング手法に関する研究*

弓矢英梨佳 (学籍番号 201021765)
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

DB 分野において，データ一貫性制約は，データ管理，統合，品質維持などに用いる基本的な技術である．しかし，データ一貫性制約は必ずしも明示的に指定されているとは限らない．そのため，これまで既存のデータからデータ一貫性制約を発見することを支援するための研究が多く行われてきた [1][2].

Web データを対象とした包含従属性発見支援に関する研究 [4] では，Web データ中の HTML 要素や XML 要素 (以下，Web ページ要素) 群から包含従属性の必要条件である包含関係が成立する Web ページ要素対（以下，包含対）を効率的に発見する手法を提案している．しかし，大量の Web データを対象とすると一般には膨大な数の包含対が発見される．したがって，ユーザは膨大な数の包含対から包含従属性が成立する包含対を発見しなければならない．

本論文では，膨大な数の包含対から効率よく包含従属性を発見するために包含対のランキング手法を提案する．本論文で提案するランキング手法を適用することで，包含従属性の有無を確認するべき包含対に優先度を付けることが可能となる．これにより，従来よりも Web データに存在する包含従属性を効率良く発見することが可能となる．

本論文の内容は次の通りである．(1) 膨大な数の包含対から包含従属性を効率よく発見するためのランキングの基準として，包含対間のカバー関係を定義する．(2) カバー関係に基づいた効率の良いランキングを実現するために，包含関係が成立する確率を利用したランキング手法を提案する．包含関係が成立する確率は，簡略化したモデルの基では包含対間の確率の値の順序とカバー関係の順序が矛盾しないという特徴がある．(3) 簡略化したモデルではなく，実データの包含対に対して確率の値を計算した時に，包含対間の確率の値の順序とカバー関係の順序がどの程度矛盾するかを示す．

2. ランキング対象の包含対の定義

本論文で扱う包含対の説明をする前に，対象とするデータのモデル化を行う．対象とする Web データは，$(P, elem, words)$ の三つ組である．それぞれ，P $(= \{p_1, p_2, \ldots\})$ は Web ページの集合，$elem(p_k)$ $(= \{e_1, e_2, \ldots\})$ は Web ページ $p_k \in P$ 中の Web ページ要素集合，$words(e_i)$ $(= \{|w_1, w_2, \ldots|\})$ は Web ページ要素 $e_i \in elem(p_k)$ 中の文字列を形態素解析や，N-gram 等で分割した単語の多重集合である．この時，Web ページ要素間の階層構造は次のように表現される．「e_i が e_j の下位要素ならば，$words(e_i)$ は $words(e_j)$ のサブセットである.」

本論文で提案するランキングは，全ての Web ページ要素対 $pairs = \{(e_i, e_j) | e_i, e_j \in E\}$ 中の全ての包含対の集合 $inclusions(pairs) = \{(e_i, e_j) | e_i, e_j \in E, e_i \subseteq e_j\}$ を対象とする．

3. ランキングの基準である包含対間のカバー関係

本論文で提案するランキングは，包含対間の半順序関係である**カバー関係**に基づくものである．このカバー関係は，論文 [3] で提案している包含対の重要度の判定基準を改良し，再定義したものである．

カバー関係の概要は次の通りである．Web ページは階層構造であるため，複数の包含対の内容が重複している場合が多く存在する．このような包含対 α と β $(\alpha, \beta \in inclusions(pairs))$ に対して，α の包含従属性の有無を確認すると，β も同時に確認することができる場合に **α は β をカバーする**といい，$\alpha \geq \beta$ と表記する．この包含対間の関係をカバー関係と呼ぶ．

カバー関係に基づいてランキングすることで，少ない包含対の確認で多くの包含対を同時に確認することができるため，効率よく包含従属性を発見することが可能となる．カバー関係には**演繹カバー関係**と**領域カバー関係**の 2 種類が存在し，次にそれらの定義を示す．

定義 1 2 つの包含対 $\alpha = (e_1, e_4)$ と $\beta = (e_2, e_3)$ の Web ページ要素の階層構造が $words(e_2) \subseteq words(e_1)$ かつ $words(e_4) \subseteq words(e_3)$ である時，またその時に限り，α は β を演繹カバーする $(\alpha \geq_d \beta)$ と言う． □

図 1 は定義 1 の条件である Web ページ要素間の階層構造を図示したものである．この時，$e_2 \subseteq e_3$ は

---

* "A Study on Ranking Methods for the Discovery of Inclusion Dependencies in Web Data" by Erika YUMIYA

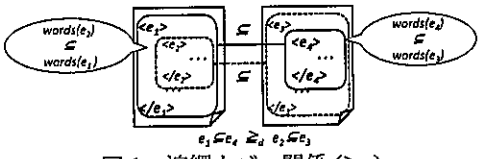

図 1　演繹カバー関係 ($\geq_d$)

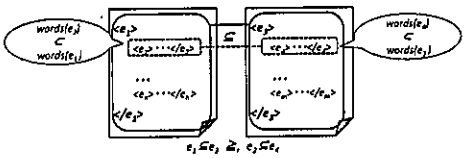

図 2　領域カバー関係 ($\geq_r$)

$e_1 \subseteq e_4$ とそれぞれの階層構造 $e_2 \subseteq e_1, e_4 \subseteq e_3$ から演繹することが可能である．したがって，$e_1 \subseteq e_4$ を確認すれば，そこから演繹できる $e_2 \subseteq e_3$ も同時に確認することが可能であるため，$e_1 \subseteq e_4$ は $e_2 \subseteq e_3$ を演繹カバーしている．

定義 2　2 つの包含対 $\alpha = (e_1, e_3)$ と $\beta = (e_2, e_4)$ の Web ページ要素の階層構造が $words(e_2) \subset words(e_1)$ かつ $words(e_4) \subset words(e_3)$ である時，またその時に限り，α は β を領域カバーする ($\alpha \geq_r \beta$) と言う．□

図 2 は定義 4 の条件を図示したものである．この時，$e_1 \subseteq e_3$ に対して確認を行えば，e_1 と e_3 中にそれぞれ存在する e_2 と e_4 間の包含対 $e_2 \subseteq e_4$ を同時に確認することが可能であるため，$e_1 \subseteq e_3$ は $e_2 \subseteq e_4$ を領域カバーしている．

4.　包含関係が成立する確率を用いたランキング手法

カバー関係に基づいたランキングを実現するためには包含対間のカバー関係を計算し，カバー関係に基づいてトポロジカルソートを行う必要があるため，その計算量は包含対 n 個に対して $O(n^2)$ となる．これに対し，本論文では包含関係が成立する確率を利用した効率的なランキング手法を提案する．

この確率の特徴は，簡略化したモデルの基で確率の値の順序とカバー関係の順序に矛盾が無いことである．すなわち，2 つの包含対 α，β のそれぞれの確率の値を $P(\alpha)$，$P(\beta)$ としたとき，「$\alpha \geq \beta$ ならば $P(\alpha) \leq P(\beta)$」である．したがって，ランキング対象の全ての包含対を確率の値が昇順になるようにソートすればカバー関係に基づいたトポロジカルソートを実現することができる．この確率順に並び替えるランキングは，確率の値を計算し，ソートするだけであるため，その計算量は $O(nlogn)$ と効率的である．

包含関係が成立する確率を厳密に計算することは困難であるため，対象である Web データに対し次の簡略化したモデルを設定する．対象となる Web

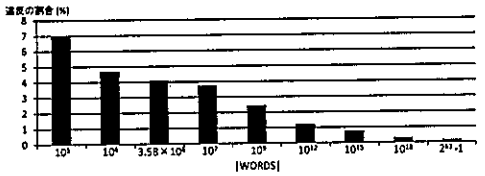

図 3　違反した包含関係の割合 ($inclusions_1$)

ページ群に現れる全ての単語の集合を $WORDS$ とする．$WORDS$ 中の各単語は独立に出現し，その出現確率は一律であるとする．また，$WORDS$ のサイズは $e \in E$ に対して十分大きいものと仮定する ($|WORDS| >> |words(e)|$)．

この時，Web ページ要素 e_1, e_2 が $|words(e_1)| \leq |words(e_2)|$ である時，$e_1 \subseteq e_2$ が成立する確率 $P(e_1 \subseteq e_2)$ は次のようになる．

$$P(e_1 \subseteq e_2) = \frac{|words(e_2)|C_{|words(e_1)|}}{|WORDS|C_{|words(e_1)|}} \quad (1)$$

5.　評価

簡略化したモデル上では矛盾が起こらない包含対間の確率の値の順序とカバー関係の順序が，実データの包含対の集合に対して確率の値を計算した時にどの程度矛盾するかを計測した．実験対象は筑波大学の情報学群と情報学群に所属する各学類の Web サイト中の全ての包含対の集合である．結果は，演繹カバー関係とは矛盾が生じなかったが，領域カバー関係とのみ矛盾が生じた．図 3 は，各 $|WORDS|$ における矛盾が生じた包含対の数の割合である．図 3 から，矛盾した包含対の割合は非常に小さいことがわかる．また，$|WORDS|$ の値を大きくするほど割合が小さくなることがわかった．

6.　まとめ

本論文では，包含従属性の発見支援のための包含対のランキング手法を提案した．具体的には，包含対の集合から効率よく包含従属性の有無を確認するために，ランキングの基準としてカバー関係を利用した．本来ならば，ランキングの実現に $O(n^2)$ かかるところを，$O(nlogn)$ で計算可能な確率を用いたランキングを開発した．ランキングで利用する確率の値は実データに適用した場合でもカバー関係と順序がほぼ無矛盾である事を示した．

文献

[1] Parag Agrawal, Arvind Arasu, Raghav Kaushik. On Indexing Error-Tolerant Set Containment. SIGMOD 2010, 927-938.

[2] Jana Bauckmann, Ulf Leser, Felix Naumann. Efficiently Computing Inclusion Dependencies for Schema Discovery. ICDE Workshops 2006, 2.

[3] 高橋公海，森嶋厚行，松本亜希子，杉本重雄，北川博之. Web コンテンツ管理のための一貫性制約発見手法, 日本データベース学会 Letters, 2008, Vol.7, No. 3, pp. 25-30.

[4] 高橋公海，森嶋厚行，弓矢英梨佳，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．情報処理学会論文誌 TOD，2010, vol.3，No.3, pp. 1-10.