Webコンテンツにおける包含従属性発見支援のためのランキング手法

弓矢英梨佳† 森嶋 厚行†† 杉本 重雄†† 北川 博之†††

† 筑波大学大学院 図書館情報メディア研究科 〒305–8550 茨城県つくば市春日 1–2
†† 筑波大学大学院 図書館情報メディア研究科/知的コミュニティ基盤研究センター 〒305–8550 茨城県つくば市春日 1–2
††† 筑波大学大学院 システム情報工学研究科 〒305–8573 茨城県つくば市天王台 1–1–1
E-mail: †s1021765@u.tsukuba.ac.jp, ††{mori,sugimoto}@slis.tsukuba.ac.jp, †††kitagawa@cs.tsukuba.ac.jp

あらまし DB 分野において，データ一貫性制約はデータ管理，統合，品質維持などの鍵となる技術として広く使われている．本稿では，Web データを対象として，代表的なデータ一貫性制約である包含従属性の発見を支援する手法について議論を行う．特に，包含従属性のヒントとなる包含関係を大量に発見した後で，それらの包含関係を重要と考えられる順にランキングする手法について提案する．本稿で提案するランキング手法は，包含関係が成立する確率を利用する．提案手法の特徴は，計算が簡単であり，かつ複雑なパラメータなどを必要としないことである．本稿では，提案手法が，単純化したモデルのもとではいくつかのヒューリスティクスと順序に相違がないことを証明し，また，実データにおいてもそれらのヒューリスティクスを用いたランキングとほぼ同じ結果を出力することを示す．
キーワード 包含従属性，Web，ランキング

A Ranking Method for the Discovery of Inclusion Dependencies in Web Contents

Erika YUMIYA†, Atsuyuki MORISHIMA††, Shigeo SUGIMOTO††, and Hiroyuki KITAGAWA†††

† Grad. Sch. of Library, Information and Media Studies, Univ. of Tsukuba 1–2 Kasuga, Tsukuba, Ibaraki,Japan 305–8550 Japan
†† Grad. Sch. of Library, Information and Media Studies/Research Center for Knowledge Communities, Univ. of Tsukuba., Univ. of Tsukuba 1–2 Kasuga, Tsukuba, Ibaraki,Japan 305–8550 Japan
††† Grad. Sch. of Sys. and Info. Eng., Univ. of Tsukuba 1–1–1 Tennohdai, Tsukuba, Japan 305–8573
E-mail: †s1021765@u.tsukuba.ac.jp, ††{mori,sugimoto}@slis.tsukuba.ac.jp, †††kitagawa@cs.tsukuba.ac.jp

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，品質維持などの鍵となる技術として広く使われている [1] [2] [3]．特に，関数従属性と包含従属性 [4] は，データベースにおける重要な制約であるとして広く利用されている．包含従属性とは常に包含関係が成立することを保証する制約であり，RDB における外部キー制約が包含従属性の例としてよく知られている．

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 [5]．したがって，既存のデータからデータ一貫性制約の発見を支援する手法がこれまで提案されてきた．これらの手法は，主に RDB を対象として関数従属性や包含従属性の発見を支援するものである．また，XML データに存在する関数従属性の発見を支援する研究もある [6]．

我々は，大量の Web コンテンツを入力として，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**）間の包含従属性の発見を支援する手法について研究を行っている [7] [8] [9] [10] [11] ．Web ページ要素間の包含従属性の応用例としては，図 1 のアプリケーションが考えられる [12]．これは，あらかじめ Web サイト管理者がシステムに包含従属性を入力すると，その後の更新等で包含従属性が満たされなくなった場合に，自動で該当箇所の修正や，Web サイト管理者に報告を行うものである．これにより，バックエンドに DB を配置した Web サイトでなくても Web サイトのデータ一貫性の維持が実現できる．

一般に，包含従属性の発見を支援するためには，包含関係が存在する Web ページ要素の組を発見することが必要条件である．しかし，Web コンテンツには表記の誤りや揺れが多く存在

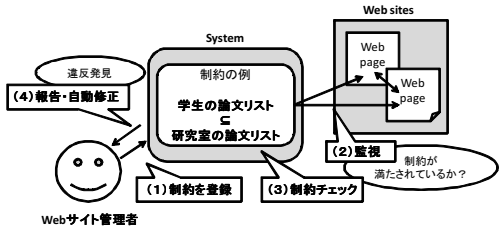

図 1　包含従属性の利用例

するため，それらの存在を前提とした手法が必要である．我々の手法では，後述する**包含率**という概念を導入し，この問題への対処を実現している．しかし，膨大な数の Web ページ要素対からは，膨大な数の包含関係が成立する Web ページ要素対が発見される．したがって，発見された大量の包含関係のランキングが重要であると考えられる．

本稿では，包含従属性の発見支援のために見つけられた包含関係をその**重要度**に応じてランキングする手法について提案する．本手法を適用することで，利用者が，重要な包含従属性をより効率良く発見することを支援できると期待される．

提案手法の特徴は，重要度として確率を用いることにより，アドホックでないランキングを実現する事である．論文 [11] では，包含関係が成立する Web ページ要素対間の階層構造などに基づいて，重要度を比較するヒューリスティクスをいくつか提案しているが，ヒューリスティクスベースの手法では次のような問題がある．(1) あるヒューリスティクスの集合がカバーできる問題の範囲がどの程度なのか判断が難しい．(2) 複数のヒューリスティクスがあるときに，組み合わせるパラメータの決定が困難である．(3) アルゴリズムが複雑になりがちである．

一方，本稿で提案する手法では，包含関係が成立する Web ページ要素対のランキングを実現するために，確率を用いて重要度を計算する．具体的には，包含関係が成立する確率が低いにも関わらず，包含関係が成立している Web ページ要素対は重要度が高い Web ページ要素対であると考える．本提案手法では，重要度を簡単に計算でき，また確率には全順序関係があるため，ヒューリスティクスを組み合わせるためのパラメータ調整などが不要である．本稿では，本提案手法を説明し，本提案手法が，単純化したモデルのもとでは論文 [11] で提案された 3 つのヒューリスティクスと理論的に整合する事を示す．さらに，実験によって，実データを用いた場合においてもヒューリスティクスを用いたランキングとほぼ同一の結果を出すことを示す．

本稿の構成は次の通りである．2 章では関連研究について述べる．3 章では，本稿で扱う Web ページ要素間の包含関係について述べる．4 章では，本稿で提案する包含関係のランキング手法について説明する．5 章は，本稿で提案する包含関係のランキング手法の効果を調査するために行った実験を説明する．6 章はまとめと今後の課題である．

2. 関 連 研 究

情報統合などの文脈において，包含従属性や包含関係の発見を目的とした研究は数多く存在する．我々の知る限り，これらは全て RDB を対象としている．また，本稿の手法では，単に包含関係の発見を行うだけでなく，発見された包含関係の重要度に応じてランキングを行うという点が異なる．

まず，厳密な包含関係の発見を支援する研究としては次のような物がある．論文 [13] では，リレーションのインスタンスが与えられた時に，リレーションの属性間に包含関係が成立するかを判定する効率のよいアルゴリズムを提案している．また，論文 [14] では，ハッシュ値を用いて 2 つのリレーションの属性間に包含関係が成立するかを効率よく発見するアルゴリズムとして，Adaptive Pick-and-Sweep Join (APSJ) と Adaptive Divide-and-Conquer Join (ADCJ) の 2 つを提案している．

次に，厳密性の緩和を考慮した研究としては，次のような物がある．まず，我々の先行研究 [10] では，Web ページ要素間の包含関係を対象として，ビットシグネチャを利用した高速な包含関係の発見手法を提案している．そこでは，厳密性の緩和のために包含率という概念を導入している．論文 [15] では，RDB における外部キー制約を発見する手法を提案している．これは，我々の包含率と同等である Inclusion coefficient を定義し，Inclusion coefficient が 0.9 以上である属性対を包含関係が成立するものとして列挙する．そして，列挙された大量の包含関係から，Randomness という評価基準を用いて外部キー制約が成立すると思われる属性対を発見する．論文 [16] では，レコードマッチングを行うための類似度を算出する際の一般的な概念である Jaccard 係数に方向性を持たせるために改良した Jaccard Containment を用いて類似度を算出し，インデックスを付与する．Jaccard Containment は，論文 [15] で定義している Inclusion Coefficient や，本研究における包含率と同じ式である．

我々の知る限り Web に関する包含関係についての研究は存在しないが，関連する研究として，Web ページやテキストを対象とした類似検索に関する研究が数多く存在する．論文 [17] では，simhash と呼ばれる特殊なハッシュ値を計算することで類似する文書の検索を行う．また，テキスト類似検索の領域では，効率よく Simirality Join を行うための positional filtering [18] という手法が提案されている．類似関係と包含関係は関連する概念ではあるが，一般にはこれらの間には類似ならば包含といった因果関係はなく，同じアルゴリズムで発見することはできない．

3. Web ページ要素対の包含関係

本章では，本稿で扱う Web ページ要素対の包含関係について説明する．Web ページ要素対 (e_i, e_j) に包含関係が成立するとは，Web ページ要素 e_i，e_j のそれぞれに含まれる単語の (多重) 集合間に包含関係がある事である．

具体的には次の通りである．まず，対象となる Web ページの集合 $P = \{p_1, p_2, \ldots\}$ が存在するとする．各 p_i は実際の Web ページそのものでなく，XML データなどへのラッピング結果でもよい．Web ページ p_i の Web ページ要素を $elem(p_i) = \{e_1, e_2, \ldots\}$ と表記する．各 e_i はそれぞれ単語の多

重集合 $W(e_i)$ (以下では断りの無い限り $W(e_i)$ は多重集合であり，それに関する演算は多重集合の演算である) をその要素中に持つ．したがって，同一ページ p において木構造を構成する $elem(p_i)$ 中の Web ページ要素間では「e_i が e_j の下位要素であれば，$W(e_i)\subseteqq W(e_j)$ である」という性質が満たされる．この性質を満たすような単語集合としては，各 Web ページ要素に含まれる文字列を n-gram 等で単純に分割したものや，形態素解析で分割したものが考えられる．我々の議論はその集合の作成方法とは独立しているため，この性質を満たしていればいずれも適用可能である．

この時，我々は，ある Web ページ要素対 e_i, e_j と値 $0 \leqq c \leqq 1$ に関して次が成立するとき，**e_i は e_j に包含率 c で包含される**といい，$e_i \subseteqq_c e_j$ と表記する [10].

$$\frac{|W(e_i) \cap W(e_j)|}{|W(e_i)|} = c \tag{1}$$

$e_i \subseteqq_c e_j$ は，通常の包含関係の一般化になっており，$c = 1$ の時，$W(e_i)\subseteqq W(e_j)$ と同じになる．これにより，誤りや表記の揺れが存在する Web コンテンツの包含関係を扱う事ができる．また，$e_i \subseteqq_{\geqq c} e_j$ を自然な拡張として定義する．すなわち，上式左辺の値が c 以上である時，**e_i は e_j に包含率 c 以上で包含される**といい，$e_i \subseteqq_{\geqq c} e_j$ と表記する．他の不等号も同様である．

4. 包含関係のランキング

我々の最終目的は，対象となる Web ページに含まれる全ての Web ページ要素の集合 $E = \bigcup_{p_i \in P} elem(p_i)$ から構成される全ての Web ページ要素対の集合 $pairs = \{(e_i, e_j) | e_i, e_j \in E\}$ の中から，重要な包含従属性を示唆するような包含関係 $e_i \subseteqq e_j$ が成立する全ての Web ページ要素対を発見する事である．これに対して論文 [10] では，Web コンテンツには誤記や表記の揺れが有ることから，前章で説明したように包含率 c を導入し，上記 $pairs$ と与えられた c に対し，包含率 c 以上の包含関係が成立する全ての Web ページ要素対の集合 $ipairs = \{(e_i, e_j) | e_i, e_j \in E, e_i \subseteqq_{\geqq c} e_j\}$ を漏れなく効率よく計算する手法を提案してきた．

本章では，このようにして計算された $ipairs$ に含まれる Web ページ要素対を，重要度に応じてランキングする手法について説明する．ランキングの際には，発見された $ipairs$ を包含率無しの包含関係を満たす集合 $ipairs = \{(e_i, e_j) | e_i, e_j \in E, e_i \subseteqq e_j\}$ と見なして議論を行う．なぜなら，包含率は誤記や表記の揺れを吸収するために導入したものであり，包含率無しの包含関係の集合の近似として利用しているためである．

本稿の提案手法は確率を用いてランキングを行うが，論文 [11] で提案された，包含関係の重要度を計算する 3 つのヒューリスティクスと順序に相違がないという特徴がある．したがって，まず，それらのヒューリスティクスを説明し，次に，本稿で提案する確率を用いた包含関係のランキング手法を説明する．

4.1 包含関係の重要度を示すためのヒューリスティクス

本節で説明する包含関係の重要度を示すためヒューリスティクスは，2 つの包含関係 α，β を比較した時にどちらが重要であるかを示すものである．今回説明するヒューリスティクスは，演繹可能な包含関係に関するヒューリスティクス H_1，リスト構造を持つ包含関係に関するヒューリスティクス H_2，占有率を用いたヒューリスティクス H_3 の 3 つである．

本稿では，包含関係 α，β $(\alpha \neq \beta)$ においてヒューリスティクス H_i によって包含関係 α の重要度が β 以上であるとみなした時，$\alpha \geqq_{H_i} \beta$ $(i \in [1,3])$ と表記する．次にそれぞれのヒューリスティクスを説明する．

(H_1) 演繹可能な包含関係に関するヒューリスティクス．ヒューリスティクス H_1 は，包含関係 α から包含関係 β が演繹できる時，演繹可能な包含関係 β よりも演繹不可能である包含関係 α の方が重要であるとみなすものである．

例を図 2 に示す．この例では，包含関係 $X_2 \subseteqq Y_1$ は，包含関係 $X_1 \subseteqq Y_2$ と Web ページ要素の階層構造 $X_2 \subseteqq X_1$，$Y_2 \subseteqq Y_1$ から演繹することが可能であるため（$X_2 \subseteqq X_1 \subseteqq Y_2 \subseteqq Y_1 \rightarrow X_2 \subseteqq Y_1$），$X_1 \subseteqq Y_2$ は $X_2 \subseteqq Y_1$ よりも重要と考えられる．

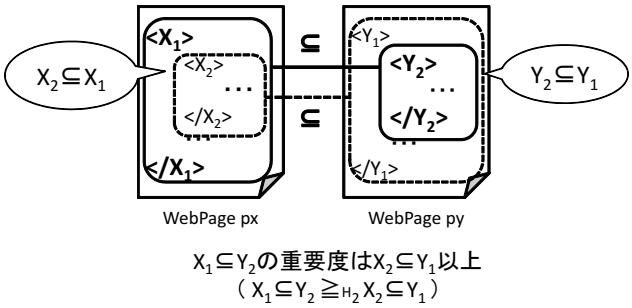

図 2　演繹可能な包含関係に関するヒューリスティクスの適用例

ヒューリスティクス H_1 は次のようにまとめられる．

【ヒューリスティクス H_1】Web ページ要素の階層構造が $X_2 \subseteqq X_1$，$Y_2 \subseteqq Y_1$ であり，かつ 2 つの包含関係 $X_1 \subseteqq Y_2$ と $X_2 \subseteqq Y_1$ が存在すると仮定する．この時，またこの時に限り，2 つの包含関係の重要度は $X_1 \subseteqq Y_2 \geqq_{H_1} X_2 \subseteqq Y_1$ となる．□

(H_2) リスト構造を持つ包含関係に関するヒューリスティクス．ヒューリスティクス H_2 は，包含関係 α で参照される Web ページ要素が，包含関係 β で参照される Web ページ要素をリストの要素として保持している場合，次のようにみなすものである．すなわち，リストの要素間で成立している個々の包含関係 β より，リストを構成する Web ページ要素間の包含関係 α の方が重要であるとみなすものである．

例を図 3 に示す．この例の場合，Web ページ要素 X_a，Y_b はリストを構成する Web ページ要素であり，X_1 と Y_1 はそれぞれのリストの要素となる Web ページ要素である．最終的に重要度でランキングを行い結果を表示すること考慮すると，X_1，Y_1 のようなリストの要素間の包含関係 $X_1 \subseteqq Y_1$ を個々に挙げるより，それらリストの要素を一つにしている Web ページ要素間の包含関係 $X_a \subseteqq Y_b$ を一つ挙げるほうがわかりやすいと考えらえる．したがって，包含関係 $X_a \subseteqq Y_b$ は，包含関係 $X_1 \subseteqq Y_1$ より重要と考える．

ヒューリスティクス H_2 は次のようにまとめられる．

【ヒューリスティクス H_2】Web ページ要素の階層構造が $X_1 \subseteqq X_a$ かつ $Y_1 \subseteqq Y_b$ であり，かつ 2 つの包含関係 $X_a \subseteqq Y_b$ と包含関係

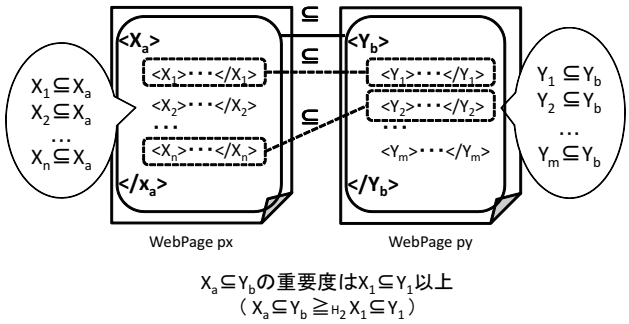

図 3　リスト構造を持つ包含関係に関するヒューリスティクスの適用例

$X_1 \subseteqq Y_1$ が存在すると仮定する．この時，またこの時に限り，2 つの包含関係の重要度は $X_a \subseteqq Y_b \geqq_{H_2} X_1 \subseteqq Y_1$ となる．　□

(H_3) 占有率を用いたヒューリスティクス． ヒューリスティクス H_3 の説明に入る前に，ヒューリスティクス H_3 で使用する**占有率**について説明する．包含関係 $X \subseteqq Y$ の占有率とは．Web ページ要素 Y に含まれる単語集合 $W(Y)$ において，Web ページ要素 X に含まれる単語集合 $W(X)$ がどれくらいの割合を占めているかを示したものである．占有率 $o(X \subseteqq Y)$ を求める式を次に示す．

$$o(X \subseteqq Y) = \frac{|W(X)|}{|W(Y)|} \tag{2}$$

この時，本ヒューリスティクスは，2 つの包含関係において包含されている Web ページ要素の単語数が等しければ，占有率の値が大きい包含関係ほど重要であるとみなすものである．

例を図 4 に示す．包含関係 $X \subseteqq Y_1$ の占有率は $o(X \subseteqq Y_1) = \frac{5}{10} = 0.5$，包含関係 $X \subseteqq Y_2$ の占有率は $o(X \subseteqq Y_2) = \frac{5}{5} = 1.0$ である．したがって，$o(X \subseteqq Y_1) < o(X \subseteqq Y_2)$ より，包含関係 $X \subseteqq Y_2$ は，包含関係 $X \subseteqq Y_1$ より重要であると考えられる．

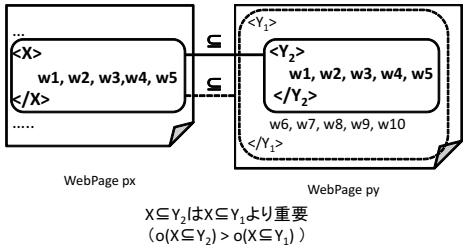

図 4　占有率を用いたヒューリスティクスの適用例

ヒューリスティクス H_3 は次のようにまとめられる．

【ヒューリスティクス H_3】　Web ページ要素の単語数が $|W(X_1)| = |W(X_2)|$ であり，かつ包含関係 $X_1 \subseteqq Y_1$ と $X_2 \subseteqq Y_2$ が存在していると仮定する．この時，それぞれの包含関係の占有率が $o(X_1 \subseteqq Y_1) \leqq o(X_2 \subseteqq Y_2)$ である場合に限り，2 つの包含関係は $X_2 \subseteqq Y_2 \geqq_{H_3} X_1 \subseteqq Y_1$ となる．　□

4.2　確率に基づく包含関係のランキング手法

本稿で提案するランキング手法では，包含関係の重要度の基準として，包含関係が成立する確率を用いる．具体的には，包含関係 α，β に関して，包含関係が成立する確率を $P(\alpha)$，$P(\beta)$ としたとき，$P(\alpha) \leqq P(\beta)$ ならば α の重要度は β 以上とする（$\alpha \geqq_p \beta$ と表記）．そして，包含関係が成立する全ての Web ページ要素対を，確率の値が昇順になるようにソートする．

現実には，包含関係が成立する確率を厳密に計算するのは困難である．したがって，本手法では対象となる Web ページ群に存在する Web ページ要素 X，Y 間の包含関係 $X \subseteqq Y$ の確率を算出するために次のような簡略化したモデルを設定する．すなわち，対象となる Web ページ群に現れる全ての単語の集合を $Words$ とし，各単語は独立に出現してその出現確率は一律であるとする．$Words$ のサイズは十分大きい（$|Words| >> |W(X)|$，　$|Words| >> |W(Y)|$）とする．

この時，包含関係 $X \subseteqq Y$ が成立する確率は次のようになる．すなわち，Web ページ要素 X 中の単語数 $|W(X)|$ が，Web ページ要素 Y 中の単語数 $|W(Y)|$ 以下であるとき（$|W(X)| \leqq |W(Y)|$），包含関係 $X \subseteqq Y$ が成立する確率 $P(X \subseteqq Y)$ は次になる：

$$P(X \subseteqq Y) = \frac{{}_{|W(Y)|}C_{|W(X)|}}{{}_{|Words|}C_{|W(X)|}} \tag{3}$$

4.3　提案手法とヒューリスティクス

本稿で提案した確率に基づいたランキングが，4.1 節で挙げた 3 つのヒューリスティクスの結果と順序に相違がないことを証明する．具体的には，異なる包含関係 α，β ($\alpha \neq \beta$) において $\alpha \geqq_{H_i} \beta$ ならば $P(\alpha) \leqq P(\beta)$ ($\alpha \geqq_p \beta$) であることを証明する．

これらの証明を行う前に，証明で使用する定理を次に示す．

定理 1. 包含関係 $X_1 \subseteqq Y_1$，$X_2 \subseteqq Y_2$ が次の 2 つの条件のいずれかを満たすとき，$P(X_1 \subseteqq Y_1) < P(X_2 \subseteqq Y_2)$ が成立する．(a) $|W(X_1)| > |W(X_2)|$，(b) $|W(X_1)| = |W(X_2)|$ かつ $|W(Y_1)| < |W(Y_2)|$　□

定理 1 の証明は付録に記載する．

この時，確率による順序が，それぞれのヒューリスティクスによる順序と相違がないことを次に示す．

(H_1) 演繹可能な包含関係に関するヒューリスティクス．

定理 2． 包含関係 $X_1 \subseteqq Y_1$，　$X_2 \subseteqq Y_2$ が与えられた時，ヒューリスティクス H_1 による重要度が $X_1 \subseteqq Y_1 \geqq_{H_1} X_2 \subseteqq Y_2$ ならば $P(X_1 \subseteqq Y_1) \leqq P(X_2 \subseteqq Y_2)$ である．　□

証明． ヒューリスティクス H_1 では，包含関係 $X_1 \subseteqq Y_1$ と $X_2 \subseteqq Y_2$ において，$X_2 \subseteqq X_1$ かつ $Y_1 \subseteqq Y_2$ である時に限り $X_1 \subseteqq Y_1 \geqq_{H_1} X_2 \subseteqq Y_2$ が成立する．この時，包含関係 $X_2 \subseteqq X_1$ より，$|W(X_2)| \leqq |W(X_1)|$ である．同様に，包含関係 $Y_1 \subseteqq Y_2$ より，$|W(Y_1)| \leqq |W(Y_2)|$ である．

したがって，$X_1 \subseteqq Y_1 \geqq_{H_1} X_2 \subseteqq Y_2$ は，$|W(X_2)| \leqq |W(X_1)|$ かつ $|W(Y_1)| \leqq |W(Y_2)|$ が成立する十分条件である．定理 1 の条件 (a) より，$|W(X_2)| < |W(X_1)|$ であれば $P(X_1 \subseteqq Y_1) < P(X_2 \subseteqq Y_2)$ が成立する．また，定理 1 の条件 (b) より $|W(X_1)| = |W(X_2)|$ かつ $|W(Y_1)| < |W(Y_2)|$ であれば $P(X_1 \subseteqq Y_1) < P(X_2 \subseteqq Y_2)$ が成立する．さらに，$|W(X_1)| = |W(X_2)|$ かつ $|W(Y_1)| = |W(Y_2)|$ であれば $P(X_1 \subseteqq Y_1) = P(X_2 \subseteqq Y_2)$ が成立するのは自明である．

したがって，包含関係 $X_1 \subseteqq Y_1$ と $X_2 \subseteqq Y_2$ において，ヒューリスティクス H_1 による重要度が $X_1 \subseteqq Y_1 \geq_{H_1} X_2 \subseteqq Y_2$ ならば $P(X_1 \subseteqq Y_1) \leqq P(X_2 \subseteqq Y_2)$ である． □

(H_2) リスト構造を持つ包含関係に関するヒューリスティクス．

定理 3． 包含関係 $X_1 \subseteqq Y_1$， $X_2 \subseteqq Y_2$ が与えられたとき，ヒューリスティクス H_2 による重要度が $X_1 \subseteqq Y_1 \geq_{H_2} X_2 \subseteqq Y_2$ ならば $P(X_1 \subseteqq Y_1) \geqq P(X_2 \subseteqq Y_2)$ である． □

証明． ヒューリスティクス H_2 では，包含関係 $X_1 \subseteqq Y_1$ と $X_2 \subseteqq Y_2$ が与えられた時，$X_2 \subset X_1$ かつ $Y_2 \subset Y_1$ である時に限り，$X_1 \subseteqq Y_1 \geq_{H_2} X_2 \subseteqq Y_2$ が成立する．この時，包含関係 $X_2 \subseteqq X_1$ より，$|W(X_2)| \leqq |W(X_1)|$ である．同様に，包含関係 $Y_2 \subseteqq Y_1$ より，$|W(Y_2)| \leqq |W(Y_1)|$ である．

したがって，$X_1 \subseteqq Y_1 >_{H_2} X_2 \subseteqq Y_2$ は，$|W(X_2)| < |W(X_1)|$ かつ $|W(Y_2)| < |W(Y_1)|$ が成立する十分条件である．定理 1 の条件 (a) より，$|W(X_2)| < |W(X_1)|$ であれば $P(X_1 \subseteqq Y_1) < P(X_2 \subseteqq Y_2)$ が成立する．さらに，$|W(X_1)| = |W(X_2)|$ かつ $|W(Y_1)| = |W(Y_2)|$ であれば $P(X_1 \subseteqq Y_1) = P(X_2 \subseteqq Y_2)$ が成立するのは自明である．

したがって，包含関係 $X_1 \subseteqq Y_1$ と $X_2 \subseteqq Y_2$ において，ヒューリスティクス H_2 による重要度が $X_1 \subseteqq Y_1 \geq_{H_2} X_2 \subseteqq Y_2$ ならば $P(X_1 \subseteqq Y_1) \leqq P(X_2 \subseteqq Y_2)$ である． □

(H_3) 占有率を用いたヒューリスティクス．

定理 4． 包含関係 $X_1 \subseteqq Y_1$， $X_2 \subseteqq Y_2$ が与えられた時， ヒューリスティクス H_3 による重要度が $X_1 \subseteqq Y_1 \geq_{H_3} X_2 \subseteqq Y_2$ ならば $P(X_1 \subseteqq Y_1) \leqq P(X_2 \subseteqq Y_2)$ である． □

証明． ヒューリスティクス H_3 では，包含関係 $X_1 \subseteqq Y_1$ と $X_2 \subseteqq Y_2$ において，$|W(X_1)| = |W(X_2)|$ かつ，占有率が $o(X_1 \subseteqq Y_1) \geq o(X_2 \subseteqq Y_2)$ である時に限り $X_1 \subseteqq Y_1 \geq_{H_3} X_2 \subseteqq Y_2$ が成立する．この時，$|W(X_1)| = |W(X_2)|$ かつ占有率 $o(X_1 \subseteqq Y_1) \geq o(X_2 \subseteqq Y_2)$ であれば，占有率を求める式 (2) より，$|W(Y_1)| \leqq |W(Y_2)|$ が成立する．

したがって，$X_1 \subseteqq Y_1 \geq_{H_3} X_2 \subseteqq Y_2$ は，$|W(X_1)| = |W(X_2)|$ かつ $|W(Y_1)| \leqq |W(Y_2)|$ が成立する十分条件である．定理 1 の条件 (b) より，$|W(X_1)| = |W(X_2)|$ かつ $|W(Y_1)| < |W(Y_2)|$ であれば $P(X_1 \subseteqq Y_1) < P(X_2 \subseteqq Y_2)$ が成立する．さらに，$|W(X_1)| = |W(X_2)|$ かつ $|W(Y_1)| = |W(Y_2)|$ であれば $P(X_1 \subseteqq Y_1) = P(X_2 \subseteqq Y_2)$ が成立するのは自明である．

したがって，包含関係 $X_1 \subseteqq Y_1$ と $X_2 \subseteqq Y_2$ において，ヒューリスティクス H_3 による重要度が $X_1 \subseteqq Y_1 \geq_{H_3} X_2 \subseteqq Y_2$ ならば $P(X_1 \subseteqq Y_1) \leqq P(X_2 \subseteqq Y_2)$ である． □

5. 実　　験

本稿で提案した確率に基づくランキングを実際に実データに適用した時の効果を検証するための実験を行った．

5.1 実験データ

対象とした Web サイトは，筑波大学に所属する情報学群 (以下 inf) [19] ，知識情報・図書館学類 (以下 klis) [20]，情報メディア創成学類 (以下 mast) [21]，情報学類 (以下 coins) [22] の 4 つである．これら各 Web サイトのルートページからクローリングを行い，アクセスが可能であった同一サイト内の全ての Web ページを使用した．各 Web サイトの Web ページ数，Web ページ要素数を表 1 に示す．各 Web ページ要素に含まれる文字列は，形態素解析ツール Sen [23] を利用して単語に分割した．対象となる Web ページ群に含まれる全ての単語の総数は 3,580,217 個であった．

Web サイト	Web ページ数	Web ページ要素数
inf	42	2,065
klis	131	7,924
mast	156	10,142
coins	672	79,778

表 1　実験データ

5.2 実 験 手 順

本実験では，まず，ランキングを行う対象として，同じ Web サイトを含めた全ての Web サイト間の異なる Web ページ間に存在する，包含関係が成立する Web ページ要素対 (以下，包含関係と略記) を抽出した．論文 [24] による実験では，包含率が 0.7 以上の Web ページ要素対を抽出すれば，包含関係を漏れなく抽出可能であることを示している．したがって，今回対象とする包含関係として，包含率が 0.7 以上の包含関係をもつ Web ページ要素対を抽出した．これらの Web ページ要素対の数は 15,086,272 である．Web ページ要素が含む単語数の最大値は 24,970 である．

次に，抽出した包含関係を対象に，本稿で提案する確率を用いたランキング手法と，4 章で挙げた 3 つのヒューリスティクスによるランキングを適用し，その結果を比較した．

提案手法において，$P(X \subseteqq Y)$ の計算で使用する単語集合 $Words$ の単語数は { 100,000, 1,000,000, 3,580,000,10,000,000, 1,000,000,000 } のいずれかに設定した．ここに 3,580,000 が含まれる理由は，実験対象となるデータに含まれる総単語数が実際にその数であるからである．

しかし，我々が仮定したモデルでは，$|Words| >> W(Y) \geq W(X)$ であるとしている．詳細は省略するが，提案する式と実験対象データの統計情報から計算すると，3,580,000 は十分に大きな $|Words|$ の値ではない．したがって，現実のデータの単語数に合わせると，仮定したモデルとは設定が異なることになる．一方，$|Words|$ の数を大きく設定すると，仮定したモデルの設定には近くなるが，現実のデータとは異なる設定になる．

5.3 実験結果 1

まずは，実験の結果，順序に相違がある包含関係の数について説明する．ここで言う順序に相違がある包含関係とは，包含関係 α，β において，ヒューリスティクス H_i による重要度が $\alpha <_{H_i} \beta$ であるにも関わらず，確率の値が $P(\alpha) < P(\beta)$ であった場合の包含関係 α を指す．

ヒューリスティクス H_1 と H_3 に関しては，$|Words|$ の設定によらず，確率を用いたランキングの順序に相違は生じなかった．しかし，ヒューリスティクス H_2 に関しては順序に相違が生じた．図 5 は，ヒューリスィクス H_2 において順序に相違があった包含関係の数を表した累積度数グラフである．横軸は確率を用いたランキングの順位，縦軸は，ヒューリスティクス H_2

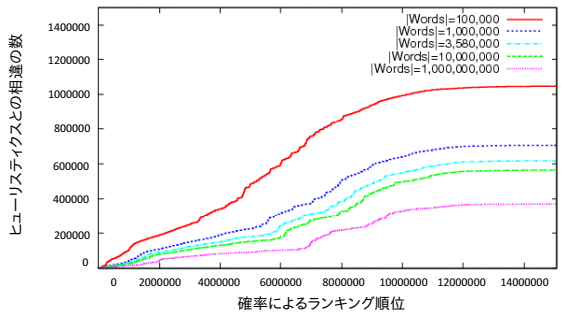

図 5　実験結果 1

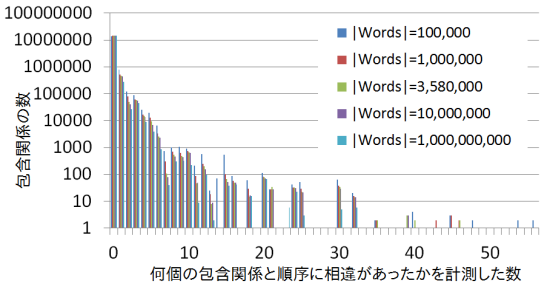

図 6　実験結果 2

と順序に相違があった包含関係の数である．さらに，順序に相違があった包含関係は本実験で対象とした包含関係の総数の何割であるかを算出した．算出結果は，$|Words| = 3,580,000$ の時は約 4.1%，$|Words| = 100,000$ の時は約 6.9%，$|Words| = 1,000,000$ の時は約 4.7%，$|Words| = 10,000,000$ の時は約 3.7%，$|Words| = 1,000,000,000$ の時は約 2.5% である．

このように，ヒューリスティクス H_2 に関して，実データに基づいて $|Words|$ を実データの総単語数に設定するよりも，確率の式を定義する時に仮定した “$Words$ は十分大きい” というモデルをにしたがって設定する方が，良い結果が得られた．ただし，実データに合わせた値を設定しても，順序に相違があった包含関係の数の割合は高々4.1% であるため，次に示す実験結果と合わせると十分に実用的であると言える．

5.4　実験結果 2

次に，順序に相違が生じたヒューリスティクス H_2 に関して，それぞれの包含関係が何個の包含関係と順序に相違があったかを示す (図 6)．例えば，包含関係 α，β，γ において，ヒューリスティクス H_2 による重要度が $\alpha <_{H_2} \beta <_{H_2} \gamma$ であり，確率の値が $P(\alpha) < P(\gamma) < P(\beta)$ であった時，包含関係 α は β と γ の 2 つの包含関係と順序に相違があると計測する．

図 6 の横軸は何個の包含関係と順序に相違があるかを計測した数，縦軸は横軸に対応する包含関係の数を表している．横軸 0 の目盛上の包含関係は，順序に相違がない包含関係である．さらに，1 個以上 10 個以内の包含関係と順序に相違がある包含関係が，順序に相違がある包含関係の総数の何割であるかを算出した．それぞれ，$|Words|$ が $100,000$，$1,000,000$ である時は約 99.8%，$|Words|$ が $3,580,000$，$10,000,000$，$1,000,000,000$ である時は約 99.9% であった．

結果から，ヒューリスティクスと順序に相違がある包含関係うち，相違している数が小さい包含関係がほとんどを占めていることがわかる．したがって，提案する確率を用いたランキング手法を実データに適用した結果は，若干ヒューリスティクスによる結果と一致しない場合があるものの，その違いは小さいと言える．

6.　まとめと今後の課題

本稿では，Web コンテンツにおける包含従属性の発見支援を行うための，包含関係のランキング手法について提案した．本ランキング手法は，包含関係が成立する確率を利用するものであり，次の特徴を持つ．(1) 包含関係が成立する Web ページ要素対の集合に対して，確率を算出し，ソートするだけでランキング結果となる全順序集合を求めることが可能である．(2) 論文 [11] で提案された，包含関係の重要度を示す 3 つのヒューリスティクスに対して，単純化したモデルのもとで理論的に相違がなく，また，実データを用いた実験でもほぼ同じ結果になる．

今後の課題としては，さらに良いランキングの基準の検討がある．具体的には，今回比較対象としたヒューリスティクスと相違しないというだけでなく，Web ページ要素の階層構造などを考慮するなどの工夫が可能ではないかと考えている．

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

文　　献

[1] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Wenyuan Yu. “Towards Certain Fixes with Editing Rules and Master Data”, VLDB2010, 2010, vol.3, No.1, pp.173-184.

[2] D. Lee, M. Mani, F. Chiu, and W.W. Chu. “NeT & CoT: Translating Relational Schemas to XML Schemas using Semantic Constraints ”, CIKM2002, 2002, pp. 282-291.

[3] Michael Karlinger, Millist Vincent and Michael Schrefl. “Inclusion Dependencies in XML: Extending Relational Semantics”. DEXA 2009, pp. 23-37, 2009

[4] Serge Abiteboul, Rechard Hull, Victor Vianu. “Foundations of Databases”, Addison-Wesley Publishing Company”.1995

[5] F. D. Marchi, S. Lopes, and J.-M. Petit. “Unary and n-ary inclusion dependency discovery in relational databases. Journal of Intelligent Information Systems”, Journal of Intelligent Information Systems, 2009, vol.32, No.1, pp.53-73.

[6] Hang Shi, Toshiyuki Amagasa, Hiroyuki Kitagawa, “Fast Detection of Functional Dependencies in XML Data”, XSym 2010, pp.113-127.

[7] 高橋公海，澤菜津美，森嶋厚行，杉本重雄，北川博之. “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），2008, pp.189-190.

[8] 高橋公海，森嶋厚行，松本亜季子，杉本重雄，北川博之. “Web コンテンツ一貫性管理のための制約発見支援”，iDB2008, 2008, pp.127-132.

[9] 高橋公海，森嶋厚行，杉本重雄，北川博之．Web ページを対象とした包含従属性の効率的な発見手法，DEIM2009, 2009.

[10] 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.

[11] 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.

[12] 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”, 2007, DEWS2007
[13] Bauckmann, J., Leser, U. and Naumann F.: Efficiently Computing Inclusion Dependencies for Schema Discovery, InterDB ’06 (ICDE Workshop), 2006.
[14] Melnik, S. and Garcia-Molina, H. “Adaptive Algorithms for Set Containment Joins”, ACM Trans. Database Systems, 2003, Vol.28, No.1, pp.56-99 .
[15] Meihui Zhang, Marios Hadjieleftheriou, Beng Chin Ooi, Cecilia M. Procopiuc, Divesh Srivastava. ”On Multi-Colomn Foreign key Discovery”, 2010, VLDB 2010, vol.3, No 1.
[16] Parag Agrawal, Arvind Arasu, Raghav Kaushik . ”On Indexing Error-Tolerant Set Containment”. SIGMOD 2011, 2011.
[17] Manku, G.S., Jain, A. and Sarma, A.D. “Detecting near-duplicates for web crawling”,WWW2007, 2007, pp.141-150.
[18] Xiao, C., Wang, W., Lin, X. and Yu, J.X. “Efficient Similarity Joins for Near Duplicate Detection”, WWW2008, 2008, pp.131-140.
[19] 筑波大学情報学群, “http://inf.tsukuba.ac.jp”, (参照 2010-07-13)
[20] 筑波大学情報学群　知識情報･図書館学類, “http://klis.tsukuba.ac.jp/”, (参照 2010-07-13)
[21] 筑波大学情報学群　情報メディア創成学類, “http://www.mast.tsukuba.ac.jp/”, (参照 2010-07-13)
[22] 筑波大学情報学群　情報科学類, “http://www.coins.tsukuba.ac.jp/”, (参照 2010-07-13)
[23] 形態素解析ツール Sen, “http://www.mlab.im.dendai.ac.jp/ yamada/ir/MorphologicalAnalyzer/Sen.html”, (参照 2010-05-20)
[24] 高橋公海，澤菜津美，森嶋厚行，杉本重雄，北川博之. “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）, 2008, pp.189-290 日本データベース学会 Letters, 2008, Vol.7, No. 3, pp. 25-30.

付　　録

1. 定理 1 の証明

証明する定理 1 は次の通りである．

定理 1． 包含関係 $X_1 \subseteqq Y_1$，$X_2 \subseteqq Y_2$ が次の 2 つの条件のいずれかを満たすとき，$P(X_1 \subseteqq Y_1) < P(X_2 \subseteqq Y_2)$ が成立する．(a) $|W(X_1)| > |W(X_2)|$，(b) $|W(X_1)| = |W(X_2)|$ かつ $|W(Y_1)| < |W(Y_2)|$ □

ここで，包含関係 $X \subseteqq Y$ の確率 $P(X \subseteqq Y)$ を求める式 (3) は，次の仮定に基づいている．$|Words| >> |W(Y)| \geqq |W(X)|$ である．

この時，定理 1 の証明を次に示す．

証明． 包含関係 $X_1 \subseteqq Y_1$ と $X_2 \subseteqq Y_2$ に関して，それぞれの確率は次の式で表せられる．

$$P(X_1 \subseteqq Y_1) = \frac{|W(Y_1)|C_{|W(X_1)|}}{|Words|C_{|W(X_1)|}} = \frac{|W(Y_1)|! \times (|Words| - |W(X_1)|)!}{(|W(Y_1)| - |W(X_1)|)! \times |Words|!} \quad \text{(A·1)}$$

$$P(X_2 \subseteqq Y_2) = \frac{|W(Y_2)|C_{|W(X_2)|}}{|Words|C_{|W(X_2)|}} = \frac{|W(Y_2)|! \times (|Words| - |W(X_2)|)!}{(|W(Y_2)| - |W(X_2)|)! \times |Words|!} \quad \text{(A·2)}$$

また，式 (A·1)，式 (A·2) より，

$$\frac{P(X_2 \subseteqq Y_2)}{P(X_1 \subseteqq Y_1)} = \frac{|W(Y_2)|! \times (|Words| - |W(X_2)|)! \times (|W(Y_1)| - |W(X_1)|)!}{(|W(Y_2)| - |W(X_2)|)! \times |W(Y_1)|! \times (|Words| - |W(X_1)|)!} \quad \text{(A·3)}$$

である．この時，$P(X_1 \subseteqq Y_1) < P(X_2 \subseteqq Y_2)$ であることを示すため，定理 1 の条件 (a)，(b) がそれぞれ成立する時に，式 (A·3)> 1 であることを証明する．

(a)$|W(X_1)| > |W(X_2)|$ の時．

条件 (a) が成立する時，$|W(Y_1)|$，$|W(Y_2)|$ の大小関係は次の 3 つであることが考えられる．(a-1) $|W(Y_1)| = |W(Y_2)|$，(a-2) $|W(Y_1)| > |W(Y_2)|$，(a-3) $|W(Y_1)| < |W(Y_2)|$．したがって，条件 (a) において条件 (a-1) ，(a-2)，(a-3) がそれぞれ成立する時に $P(X_1 \subseteqq Y_1) < P(X_2 \subseteqq Y_2)$ であることを証明する．

(a-1) $|W(Y_1)| = |W(Y_2)|$ の時． 条件 (a-1) より，$|W(Y_1)| = |W(Y_2)| = |W(Y)|$ とすると，式 (A·3) は次のように変形できる．

$$\frac{P(X_2 \subseteqq Y_2)}{P(X_1 \subseteqq Y_1)} = \frac{(|Words| - |W(X_2)|)!}{(|Words| - |W(X_1)|)!} \times \frac{(|W(Y)| - |W(X_1)|)!}{(|W(Y)| - |W(X_2)|)!} \quad \text{(A·4)}$$

ここで，条件 (a) の $|W(X_1)| > |W(X_2)|$ より，$|W(X_1)| = i + |W(X_2)|$（$i > 0$）と仮定する．この時，式 (A·4) の右辺を $A \times B$ とするとそれぞれ次のようになる．

$$A = \frac{(|Words| - |W(X_2)|)!}{(|Words| - |W(X_1)|)!} = (|Words| - |W(X_1)| + 1) \times (|Words| - |W(X_1)| + 2) \times \ldots \times (|Words| - |W(X_1)| + i) \quad \text{(A·5)}$$

$$B = \frac{(|W(Y)| - |W(X_1)|)!}{(|W(Y)| - |W(X_2)|)!} = \frac{1}{(|W(Y)| - |W(X_1)| + 1) \times (|W(Y)| - |W(X_1)| + 2) \times \ldots \times (|W(Y)| - |W(X_1)| + i)} \quad \text{(A·6)}$$

式 (A·5) と式 (A·6) より，$A \times B$ は次式となる．

$$A \times B = \frac{(|Words| - |W(X_1)| + 1) \times (|Words| - |W(X_1)| + 2) \times \ldots \times (|Words| - |W(X_1)| + i)}{(|W(Y)| - |W(X_1)| + 1) \times (|W(Y)| - |W(X_1)| + 2) \times \ldots \times (|W(Y)| - |W(X_1)| + i)} \quad \text{(A·7)}$$

式 (A·7) > 1 であるためには，任意の自然数 $n \in [1, i]$ において

$$\frac{(|Words| - |W(X_1)| + n)}{(|W(Y)| - |W(X_1)| + n|} > 1$$

であれば良い．仮定 $|Words| >> |W(Y)|$ より，$(|Words| - |W(X_1)| + n) >> (|W(Y)| - |W(X_1) + n|)$ であることは自明である．したがって，式 (A·7) > 1 が成立する．

これにより，条件 (a) $|W(X_1)| > |W(X_2)|$ かつ，(a-1) $|W(Y_1)| = |W(Y_2)|$ の時，式 (A·3)> 1 であるため，$P(X_1 \subseteqq Y_1) < P(X_1 \subseteqq Y_2)$ が成立する．

(a-2) $|W(Y_1)| > |W(Y_2)|$ の時． 条件 (a) より，$|W(X_1)| = n|W(X_2)|$，条件 (a-2) より $|W(Y_1)| = m|W(Y_2)|$ $(n > 1, m > 1)$ とすると，式 (A·3) は次のように変形できる．

$$\frac{P(X_2 \subseteqq Y_2)}{P(X_1 \subseteqq Y_1)} = \frac{|W(Y_2)|! \times (|Words| - |W(X_2)|)! \times (m|W(Y_2)| - n|W(X_2)|)!}{(|W(Y_2)| - |W(X_2)|)! \times (|Words| - n|W(X_2)|)! \times (m|W(Y_2)|)!} \quad \text{(A·8)}$$

この時，式 (A·8) の右辺を $A \times B \times C$ に分解すると，それぞれ次のようになる．

$$A = \frac{|W(Y_2)|!}{(|W(Y_2)| - |W(X_2)|)!} = (|W(Y_2)| - |W(X_2)| + 1) \times (|W(Y_2)| - |W(X_2)| + 1) \times \ldots \times (|W(Y_2)|) \quad \text{(A·9)}$$

$$B = \frac{(|Words| - |W(X_2)|)!}{(|Words| - n|W(X_2)|)!} = (|Words| - n|W(X_2)| + 1) \times (|Words| - n|W(X_2)| + 2) \times \ldots \times (|Words| - |W(X_2)|) \quad \text{(A·10)}$$

$$C = \frac{(m|W(Y_2)| - n|W(X_2)|)!}{m|W(Y_2)|!} = \frac{1}{(m|W(Y_2)| - n|W(X_2)| + 1) \times \ldots \times (m|W(Y_2)| - |W(X_2)|)} \times \frac{1}{(m|W(Y_2)| - |W(X_2)| + 1) \times \ldots \times m|W(Y_2)|} \quad \text{(A·11)}$$

式 (A·9)，式 (A·10)，式 (A·11) より，$A \times B \times C$ は次式となる．

$$A \times B \times C = \frac{(|Words| - n|W(X_2)| + 1) \times \ldots \times (|Words| - |W(X_2)|)}{(m|W(Y_2)| - n|W(X_2)| + 1) \times \ldots \times (m|W(Y_2)| - |W(X_2)|)} \times \frac{(|W(Y_2)| - |W(X_2)| + 1) \times \ldots \times |W(Y_2)|}{(m|W(Y_2)| - |W(X_2)| + 1) \times \ldots \times m|W(Y_2)|} \quad \text{(A·12)}$$

この時，式 (A·12) > 1 を示すため，分母を最大に見積もり，分子を最小に見積もると次のようになる．

$$A \times B \times C > \frac{(|Words| - n|W(X_2)| + 1)^{(n-1)|W(X_2)|}}{(m|W(Y_2)| - |W(X_2))^{(n-1)|W(X_2)|}} \times \frac{(|W(Y_2)| - |W(X_2)| + 1)^{|W(X_2)|}}{m|W(Y_2)|^{|W(X_2)|}} \quad \text{(A·13)}$$

ここで，$(|Words| - n|W(X_2)| + 1) > (m|W(Y_2)| - |W(X_2)|) \times (m|W(Y_2)|)^{\frac{1}{n-1}}$ であれば式 (A·13) > 1 が成立する．仮定 $|Words| >> |W(Y)|$ より，$|Words| > (|Words| - n|W(X_2)| + 1) > (m|W(Y_2)| - |W(X_2)|) \times m|W(Y_2)|^{\frac{1}{n-1}}$ が成立することは自明である．

したがって，式 (A·13) > 1 である．さらに，式 (A·12) $>$ 式 (A·13) であるため，式 (A·12) > 1 である．

これにより，条件 (a) $|W(Y_1)| > |W(Y_2)|$ かつ条件 (a-2) $|W(X_1)| > |W(X_2)|$ の時，式 (A·3) > 1 であるため，$P(X_1 \subseteqq Y_1) < P(X_1 \subseteqq Y_2)$ が成立する．

(a-3) $|W(Y_1)| < |W(Y_2)|$ **の時．** 条件 (a) より，$|W(X_1)| = n|W(X_2)|$，条件 (a-3) より $|W(Y_2)| = m|W(Y_1)|$ $(n > 1, m > 1)$ とすると，式 (A·3) は次のように変形できる．

$$\frac{P(X_2 \subseteqq Y_2)}{P(X_1 \subseteqq Y_1)} = \frac{(m|W(Y_1)|)! \times (|Words| - |W(X_2)|)! \times (|W(Y_1)| - n|W(X_2)|)!}{(m|W(Y_1)| - |W(X_2)|)! \times (|Words| - n|W(X_2)|)! \times (|W(Y_1)|)!} \quad \text{(A·14)}$$

この時，式 (A·14) の右辺を $A \times B \times C$ に分解すると，それぞれ次のようになる．

$$A = \frac{(m|W(Y_1)|)!}{(m|W(Y_1)| - |W(X_2)|)!} = (m|W(Y_1)| - |W(X_2)| + 1) \times (m|W(Y_1)| - |W(X_2)| + 1) \times \ldots \times (m|W(Y_1)|) \quad \text{(A·15)}$$

$$B = \frac{(|Words| - |W(X_2)|)!}{(|Words| - n|W(X_2)|)!} = (|Words| - n|W(X_2)| + 1) \times (|Words| - n|W(X_2)| + 2) \times \ldots \times (|Words| - |W(X_2)|) \quad \text{(A·16)}$$

$$C = \frac{(|W(Y_1)| - n|W(X_2)|)!}{(|W(Y_1)|)!} = \frac{1}{(|W(Y_1)| - n|W(X_2)| + 1) \times \ldots \times (|W(Y_1)| - |W(X_2)|)} \times \frac{1}{(|W(Y_1)| - |W(X_2)| + 1) \times \ldots \times |W(Y_1)|} \quad \text{(A·17)}$$

式 (A·15)，式 (A·16)，式 (A·17) より，$A \times B \times C$ は次式となる．

$$A \times B \times C = \frac{(|Words| - n|W(X_2)| + 1) \times \ldots \times (|Words| - |W(X_2)|)}{(|W(Y_1)| - n|W(X_2)| + 1) \times \ldots \times (|W(Y_1)| - |W(X_2)|)} \times \frac{(m|W(Y_1)| - |W(X_2)| + 1) \times \ldots \times m|W(Y_1)|}{(|W(Y_1)| - |W(X_2)| + 1) \times \ldots \times |W(Y_1)|} \quad \text{(A·18)}$$

この時，式 (A·18) > 1 であるためには，任意の自然数 $i \in [1, (n-1)|W(X_2)|]$ において

$$\frac{(|Words| - n|W(X_2)| + i)}{(|W(Y_1)| - n|W(X_2)| + i)} > 1$$

かつ，任意の自然数 $\mathrm{j} \in [1, |W(X_2)|]$ において

$$\frac{(m|W(Y_1)| - |W(X_2)| + \mathrm{j})|}{(|W(Y_1)| - |W(X_2)| + \mathrm{j})} > 1$$

であれば良い．それぞれ，$(|Words| - n|W(X_2)| + i) > (|W(Y_1)| - n|W(X_2)| + i)$，$(m|W(Y_1)| - |W(X_2)| + \mathrm{j})| > (|W(Y_1)| - |W(X_2)| + \mathrm{j})$ であることは自明である．したがって，式 (A·18) > 1 が成立する．

これにより，条件 (a) $|W(Y_1)| < |W(Y_2)|$ かつ条件 (a-3) $|W(X_1)| > |W(X_2)|$ の時，式 (A·3) > 1 であるため，$P(X_1 \subseteqq Y_1) < P(X_1 \subseteqq Y_2)$ が成立する．したがって，条件 (a) $|W(X_1)| > |W(X_2)|$ の時，$P(X_1 \subseteqq Y_1) < P(X_1 \subseteqq Y_2)$ が成立する．

(b) $|W(X_1)| = |W(X_2)|$ **かつ** $|W(Y_1)| < |W(Y_2)|$ **の時．**
条件 (b) より，$|W(X_1)| = |W(X_2)| = |W(X)|$ とすると，式 (A·3) は次のように変形できる．

$$\frac{P(X_2 \subseteqq Y_2)}{P(X_1 \subseteqq Y_1)} = \frac{|W(Y_2)|! \times (|W(Y_1)| - |W(X)|)!}{|W(Y_1)|! \times (|W(Y_2)| - W(X)|)!} \quad \text{(A·19)}$$

ここで，条件 (b) の $|W(Y_1)| < |W(Y_2)|$ より，$|W(Y_2)| = i + |W(Y_1)|$ $(i > 0)$ と仮定する．この時，式 (A·19) の右辺を $A \times B$ と分解するとそれぞれ次のようになる．

$$A = \frac{|W(Y_2)|!}{|W(Y_1)|!} = \frac{(|W(Y_1)| + i)!}{|W(Y_1)|!} = (|W(Y_1)| + 1) \times (|W(Y_1)| + 2) \times \ldots \times (|W(Y_1)| + i) \quad \text{(A·20)}$$

$$B = \frac{(|W(Y_1)| - |W(X)|)!}{(|W(Y_2)| - |W(X)|)!} = \frac{(|W(Y_1)| - |W(X)|)!}{(|W(Y_1)| + i - |W(X)|)!} = \frac{1}{(|W(Y_1)| - |W(X)| + 1) \times (|W(Y_1)| - |W(X)| + 2) \times \ldots \times (|W(Y_1)| - |W(X)| + i)} \quad \text{(A·21)}$$

式 (A·20) と式 (A·21) より，$A \times B$ は次式となる．

$$A \times B = \frac{(|W(Y_1)| + 1) \times (|W(Y_1)| + 2) \times \ldots \times (|W(Y_1)| + i)}{(|W(Y_1)| - |W(X)| + 1) \times (|W(Y_1)| - |W(X)| + 2) \times \ldots \times (|W(Y_1)| - |W(X)| + i)} \quad \text{(A·22)}$$

式 (A·22) > 1 であるためには，任意の自然数 $n \in [1, i]$ において

$$\frac{(|W(Y_1)| + n)}{(|W(Y_1)| - |W(X)| + n)} > 1$$

であれば良い．$(|W(Y_1)| + n) > (|W(Y_1)| + n - |W(X)|)$ であることは自明である．したがって，式 (A·22) > 1 が成立する．

これにより，条件 (b) $|W(X_1)| = |W(X_2)|$ かつ $|W(Y_1)| < |W(Y_2)|$ の時，式 (A·3) > 1 であるため，$P(X_1 \subseteqq Y_1) < P(X_1 \subseteqq Y_2)$ が成立する．

したがって，包含関係 $X_1 \subseteqq Y_1$，$X_2 \subseteqq Y_2$ が条件 (a) (b) のいずれかを満たす時，$P(X_1 \subseteqq Y_1) < P(X_2 \subseteqq Y_2)$ が成立する． □