2004－AL－95　(1)
2004／5／21

高さ制約変数を持つ順序木パターン言語の正データからの多項式時間帰納推論可能性について

鈴木祐介[1]　正代隆義[2]　松本哲志[3]　宮原哲浩[4]

[1] 九州大学大学院システム情報科学府　[2] 九州大学大学院システム情報科学研究院
[3] 東海大学理学部情報数理学科　[4] 広島市立大学情報科学部
e-mail: {y-suzuki,shoudai}@i.kyushu-u.ac.jp, matumoto@ss.u-tokai.ac.jp, miyahara@its.hiroshima-cu.ac.jp

概要

近年，Web 文書のような木構造データが増大しており，木構造データからの情報抽出がより重要になっている．これら木構造データに共通する特徴的なパターンを表現するために，本論文では順序項木を用いる．順序項木とは内部に構造的変数を持つ順序木パターンで，その変数には任意の順序木が代入可能である．本論文では高さ制約変数という新しい変数を導入する．(i,j)-高さ制約変数には，変数に対応する頂点間のパスの長さが少なくとも i であり，高さが高々 j であるような順序木しか代入することはできない．本論文では，高さ制約変数を持ち，変数のみからなるチェーンを持たない順序項木のクラスが多項式時間帰納推論可能であることを示す．

Polynomial Time Inductive Inference of Ordered Tree Languages with Height-Constrained Variables from Positive Data

Yusuke Suzuki[1]　Takayoshi Shoudai[1]　Satoshi Matsumoto[2] and Tetsuhiro Miyahara[3]

[1] Department of Informatics, Kyushu University
[2] Department of Mathematical Sciences, Tokai University, Hiratsuka
[3] Faculty of Information Sciences, Hiroshima City University
e-mail: {y-suzuki,shoudai}@i.kyushu-u.ac.jp, matumoto@ss.u-tokai.ac.jp, miyahara@its.hiroshima-cu.ac.jp

Abstract

Due to the rapid growth of tree structured data such as Web documents, efficient learning from tree structured data becomes more and more important. In order to represent structural features common to such tree structured data, we propose an ordered term tree, which is a rooted tree pattern with ordered children and structured variables. A usual variable can be replaced with an arbitrary tree. In this paper, we introduce a new kind of variable, called height-constrained variables. An (i,j)-height-constrained variable can be replaced with any tree such that the minimum length of the corresponding path of the tree is i and the maximum height of the tree is j. Let $\mathcal{OTT}^h$ be the set of all ordered term trees with (i,j)-height-constrained variables for any i and j $(1 \leq i \leq j)$ without any variable-chain. We show that the class $\mathcal{OTT}^h$ is polynomial time inductively inferable from positive data.

1 はじめに

近年，情報技術の急速な発達により，Web 文書のような半構造データが増大している．そのため，これら半構造データからの情報抽出がより重要になっている．HTML/XML のような Web 文書は，明確な構造を持たないため，半構造データと呼ばれる．半構造データから有益な情報を抽出するためには，半構造データに共通するパターンの抽出が必要である．データマイニングや知識発見の分野では，多くの研究者が機械学習手法を用いて，これらのデータの解析を行っている．Object Exchange Model[1] に基づき，本論文では半構造データを木構造データとして扱う．木構造データは順序付けられた子を持つ根付き木として表され，各辺はラベル付けされている．木構造データに共通する木構造パターンの表現方法として，本論文では順序項木を用いる．順序項木は内部構造変数を持つ順序木パターンであり，変数には任意の順序木が代入可能である [10, 11, 12, 13].

従来の変数には任意の順序木を代入できるため，その表現能力はかなり大きいと言える．項木を共通パターンとして用いる情報抽出の対象である半構造データとして，検索サイトの検索結果などを考えると，１つの検索結果の大きさには，ある程度の制限がある．この検索結果を木構造データとみなすと，木の高さが制限されているとみなすことができる．また，半構造データは，同じ種類のデータの重複やデータの欠落があり，明確な構造を持たないが，木構造データとみなすと，幅の自由度は大きいが，高さはある程度制限されているとみなすことができる．よって本論文では，構造的特徴をより表現するために，高さ制約変数という新しい変数を導入する．(i,j)-高さ制約変数には，変数に対応する頂点間のパスの長さが少なくとも i であり，高さが高々j であるような順序木しか代入することはできない．

Λ を辺ラベル集合し，少なくとも 2 つの要素を持つものとする．$\mathcal{OTT}^h_\Lambda$ を (i,j)-高さ制約変数を持ち，変数のみからなるチェーンが現れない全ての順序項木の集合とする．本論文では，$\mathcal{OTT}^h_\Lambda$ の正データからの多項式時間帰納推論可能性を考察する．

従来の順序項木は任意の順序木を代入できる構造的変数を持っている点で，他の木構造パターン [4, 6] と異なる．我々は [10, 11, 12, 13] において，高さ制約変数と持たない基本的な順序項木のクラスについてその学習可能性を考察した．また [7] において，順序項木を共通パターンとして用いた半構造データからのデータマイニング手法について提案した．また，順序項木を用いた HTML 文書からの情報抽出の実装も行った [2].

2 順序項木と高さ制約変数

本論文では，2 ポート変数のみを持つ順序項木を取り扱うものとする．多重ポートを持つ一般的な順序項木の定義は [12] で与えられている．集合 S に対して，S の要素数を $|S|$ と表す.

定義 1 (順序項木) $T = (V_T, E_T)$ を順序付けられた子を持つ根付き木とする．ここで V_T は頂点集合であり，E_T は辺集合とする．順序付けられた子を持つ根付き木を順序木と呼ぶ．E_g と H_g を，$E_g \cup H_g = E_T$, $E_g \cap H_g = \emptyset$ を満たすような E_T の分割とする．また $V_g = V_T$ とする．このとき 3 つ組み $g = (V_g, E_g, H_g)$ を**順序項木** (ordered term tree) または単に**項木** (term tree) と呼ぶ．V_g, E_g, H_g の要素をそれぞれ頂点，辺，変数とよぶ.

順序項木 g とその頂点 v_1, v_i に対して，v_1 から v_i へのパスとは，いかなる $1 \leq j < i$ である j に対しても，v_j と v_{j+1} から構成される辺または変数が存在するような頂点の列 $v_1, v_2, \ldots, v_i$ である．v と v' から構成される辺または変数が存在し，v が根から v' のパス上にあるとき，v を v' の親といい，v' を v の子という．順序項木 g の根とは，親を持たない頂点のことであり，**葉**とは，子を持たない頂点のことである．順序項木 g の高さとは，g の根から葉までの最大のパスの長さであり，また，g の頂点 v_g に対して，頂点 v_g の高さとは，v_g から葉までのパスの長さのことである．

v が v' の親であるような変数 $\{v, v'\} \in H_g$ を $[v, v']$ と表す．このとき，v を変数 $[v, v']$ の**親ポート** (parent port) といい，v' を変数 $[v, v']$ の**子ポート** (child port) という．順序項木 g に対して，g の各内点 u の全ての子はそれぞれ順序付けられており，u の 2 つの子 u', u'' において，$u' <^g_u u''$ は u' が u'' より u の子の順序において小さいことを表す．各辺及び，各変数はそれぞれ辺ラベル集合 Λ と変数ラベル集合 X の要素によってラベル付けされる．順序項木 g は H_g の全ての変数が互いに異なる X に属する変数ラベルを持つとき，**線形** (linear) または**正則** (regular) であるという．本論文

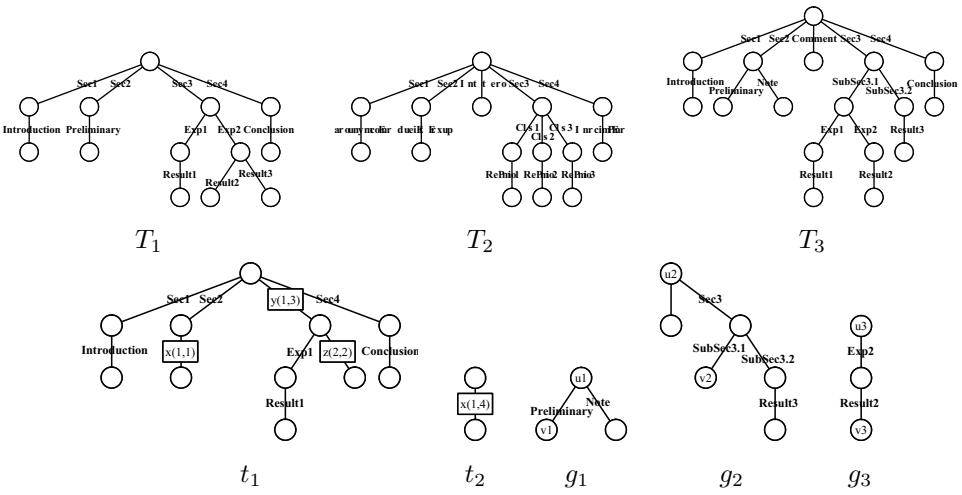

図 1: 順序項木 t_1,t_2，順序木 T_1,T_2,T_3，枠の中の $x(i,j)$ は変数ラベル x を持つ，(i,j)-高さ制約変数であることを表す.

では線形な項木のみを取り扱うので特に断らない限り，項木は全て線形であるとする.

定義 2 (高さ制約変数) $X^{\mathcal{H}}$ を変数ラベル X の無限の大きさを持つ部分集合とする．2 つの数 $1 \leq i \leq j$ に対して，各 $X^{\mathcal{H},(i,j)}$ を $X^{\mathcal{H}}$ の部分集合とする．ここで $X^{\mathcal{H}} = \bigcup_{1 \leq i \leq j} X^{\mathcal{H},(i,j)}$ かつ，$(i,j) \neq (i',j')$ に対して，$X^{\mathcal{H},(i,j)} \cap X^{\mathcal{H},(i',j')} = \emptyset$ が成り立つものとする．2 つの数 $1 \leq i \leq j$ に対して，$X^{\mathcal{H},(i,j)}$ の要素の変数ラベルを (i,j)**-高さ制約変数ラベル** ((i,j)-height constrained vriable label) と呼ぶ．変数 $[u,v]$ が (i,j)**-高さ制約変数** ((i,j)-height constrained vriable) であるのは，変数が (i,j)-高さ制約変数ラベルを持つときであり，$[u,v]^{(i,j)}$ と記述する.

順序項木 $f = (V_f, E_f, H_f)$ と $g = (V_g, E_g, H_g)$ に対して，f と g が同型であるとは，次の条件を満たす全単射 $\varphi : V_f \to V_g$ が存在するときをいう．このとき $f \equiv g$ と書く．(1) f の根が写像 φ によって g の根に写される．(2) $\{u,v\} \in E_f$ のとき，そのときに限り $\{\varphi(u), \varphi(v)\} \in E_g$ である．さらに任意の $\{u,v\} \in E_f$ に対して，$\{u,v\}$ の辺ラベルと，$\{\varphi(u), \varphi(v)\}$ の辺ラベルが等しい．(3) $1 \leq i \leq j$ に対して，$[u,v]^{(i,j)} \in H_f$ のとき，そのときに限り $[\varphi(u), \varphi(v)]^{(i,j)} \in H_g$ である．(4) f のいかなる 1 つ以上の子を持つ頂点 u の 2 つの子 u', u'' に対して，$u' <_u^f u''$ のとき，そのときに限り $\varphi(u') <_{\varphi(u)}^g \varphi(u'')$ である.

定義 3 (順序項木の代入) 2 つの順序項木に対する代入について定義する．f と g を 2 つ以上の頂点を持つ順序項木とする．x を $X^{\mathcal{H},(i,j)}$ の要素の変数ラベルとする $(1 \leq i \leq j)$．$\sigma = [u,u']$ を g の 2つの頂点のリストとする．ここで u は g の根であり，u' は g の葉である．u から u' へのパスの長さが少なくとも i であり，g の高さが高々j であるとき，$x := [g,\sigma]$ を x に対する**束縛** (binding) という.

束縛 $x := [g,\sigma]$ を f に次のように適用して新しい順序項木 $f\{x := [g,\sigma]\}$ を得る．$h = [v,v']^{(i,j)}$ を変数ラベル x を持つ f の (i,j)-高さ制約変数とする．変数 $h = [v,v']^{i,j}$ に対して，f の変数の集合 H_f から h を削除し，頂点 v, v' と g の頂点 u, u' を同一視して g を f に追加する．束縛の有限集合 $\theta = \{x_1 := [g_1,\sigma_1], \cdots, x_n := [g_n,\sigma_n]\}$ を**代入** (substitution) と呼ぶ．ここで $x_1, \ldots, x_n$ は $X^{\mathcal{H}}$ に含まれる互いに異なる変数ラベルであり，各 g_i は順序木である．代入 θ に含まれる束縛を f に全て適用した順序項木を $f\theta$ と書き，f の代入 θ による**代入例** (instance) と呼ぶ．代入後の順序項木 $f\theta$ の頂点 v の子の順序 $<_v^{f\theta}$ を次のように定める．v は 1 つ以上の子を持ち，u', u'' は $f\theta$ の v の子と仮定する．v は f の変数 $[v,v_1]^{(i_1,j_1)}, \ldots, [v,v_k]^{(i_k,j_k)}$ $(v_1 <_v^f \cdots <_v^f v_k)$ の親ポートであるとする．このとき，以下の 4 つの場合が考えられる．ただし g_ℓ は $[v,v_\ell]$ $(\ell = 1, \ldots, k)$ に代入される順序項木とする．(1) $u', u'' \in V_f$ かつ $u' <_v^f u''$ ならば，$u' <_v^{f\theta} u''$ である．(2) ある ℓ に対して $u', u'' \in V_{g_\ell}$ かつ $u' <_v^{g_\ell} u''$ な

らば，$u' <_v^{f\theta} u''$ である．(3) $u' \in V_{g_\ell}$, $u'' \in V_f$ かつ $v_\ell <_v^f u''$ ならば，$u' <_v^{f\theta} u''$ である．同様に $u' \in V_{g_\ell}$, $u'' \in V_f$ かつ $u'' <_v^f v_\ell$ ならば，$u'' <_v^{f\theta} u'$ である．(4) $\ell \neq \ell'$ である ℓ, ℓ' に対して，$u' \in V_{g_\ell}$，$u'' \in V_{g_{\ell'}}$ かつ $v_\ell <_v^f v_{\ell'}$ ならば，$u' <_v^{f\theta} u''$ である．もし v が変数の親ポートでないなら，$u', u'' \in V_f$ であり，$u' <_v^f u''$ ならば $u' <_v^{f\theta} u''$ である．代入後の順序項木 $f\theta$ の根は f の根と同じである．

例えば，t_1 を図 1 中の順序項木とし，$\theta = \{x := [g_1, [u_1, v_1]], y := [g_2, [u_2, v_2]], z := [g_3, [u_3, u_3]]\}$ を代入とする．ここで g_1, g_2, g_3 は図 *1* 中の順序木とする．このとき t_1 の代入例 $t_1\theta$ は T_3 と同型である．

Λ を辺ラベルの集合とする．いかなる種類の変数も持たない順序項木を**基礎順序項木** (ground term tree) という．$\mathcal{OT}_\Lambda$ は Λ を辺ラベル集合として持つ基礎順序項木全体の集合を表す．$\mathcal{OTT}_\Lambda$ は Λ を辺ラベル集合として持ち，正則な順序項木全体の集合を現す．順序項木 $t \in \mathcal{OTT}_\Lambda$ に対して，t の**項木言語** (term tree language) $L_\Lambda(t)$ は $\{s \in \mathcal{OT}_\Lambda \mid$ ある代入θに対して $s \equiv t\theta\}$ と定義される．

$\mathcal{OTT}_\Lambda^{\mathcal{H}}$ は全ての変数が $X^{\mathcal{H}}$ に属するラベルを持ち，Λ を辺ラベル集合として持つ正則な順序項木の集合を表す．g を $\mathcal{OTT}_\Lambda^{\mathcal{H}}$ に属する順序項木とする．g の変数の列 $[u_0, u_1]^{(i_1,j_1)}, [u_1, u_2]^{(i_2,j_2)}, \ldots, [u_{k-1}, u_k]^{(i_k,j_k)}$ $(2 \leq k)$ が以下の条件を満たすとき，チェーン変数とよぶ．(i) u_0 が g の根である，または 2 つ以上の子を持つ，またはその親と辺で結ばれている．(ii) $1 \leq i \leq k-1$ に対し，各 u_i がただ 1 つの子を持つ．(iii) u_k が葉である，または 2 つ以上の子を持つ，またはその子と辺で結ばれている．

本論文では，$|\Lambda| \geq 2$ であるとき $\mathcal{OTT}_\Lambda^h$ のクラスが正データから多項式時間帰納推論可能であることを示す．

$$\mathcal{OTT}_\Lambda^h = \{g \in \mathcal{OTT}_\Lambda^{\mathcal{H}} \mid g \text{ はチェーン変数を持たない }\}$$

$\mathcal{OTT}_\Lambda^{\mathcal{H}}$ のクラスが正データから多項式時間帰納推論可能であるかどうかは未解決である．

2.1 正データからの多項式時間帰納推論

Λ を辺ラベルの集合，$\mathcal{TT}_\Lambda$ を Λ を辺ラベルとして持つ順序項木の集合とする．順序項木 t が順序木の集合 $S \subseteq \mathcal{OT}_\Lambda$ に対し，$S \subseteq L_\Lambda(t)$ であるとき，t は S を説明するという．また t が S を説明し，$L_\Lambda(t') \subsetneq L_\Lambda(t)$ を満たすような，$t' \in \mathcal{TT}_\Lambda$ が存在しないとき t は S を説明する**極小一般化項木** (minimally generalized term tree) であるという．Angluin [5] は**有限の厚み** (finite thickness) という帰納推論可能性に関する十分条件を与えた．空でない集合 $S \subseteq \mathcal{OT}_\Lambda$ に対して，$\{t \in \mathcal{TT}_\Lambda \mid S \subseteq L_\Lambda(t)\}$ を満たす順序項木の数が有限であるとき，$\mathcal{TT}_\Lambda$ が有限の厚みを持つ．Moriyama & Sato[8] は有限の厚みを一般化し，**M-有限の厚み** (M-finite thickness) を導入した．いかなる空でない集合 $S \subseteq \mathcal{OT}_\Lambda$ に対して，以下の条件を満たすとき，$\mathcal{TT}_\Lambda$ は M-有限の厚みを持つという．(1) S を説明する $t \in \mathcal{TT}_\Lambda$ に属する極小一般化項木の数が有限である．(2) いかなる $t \in \mathcal{TT}_\Lambda$ に対しても，$S \subseteq L_\Lambda(t)$ ならば，$S \subseteq L_\Lambda(t') \subseteq L_\Lambda(t)$ であるような極小一般化項木 t' が存在する．

$t \in \mathcal{TT}_\Lambda$ に対して，t の**有限証拠集合** (finite tell-tale) $S \subseteq \mathcal{OT}_\Lambda$ とは，いかなる $t' \in \mathcal{TT}_\Lambda$ に対しても，$S \subseteq L_\Lambda(t')$ ならば，$L_\Lambda(t') \not\subseteq L_\Lambda(t)$ となる有限集合である．ここで以下の 2 つの問題について考える．

$\mathcal{TT}_\Lambda$ に関する所属性問題.
入力: 項木 $t \in \mathcal{TT}_\Lambda$，順序木 $T \in \mathcal{OT}_\Lambda$.
問題: $T \equiv t\theta$ となるような代入 θ が存在するかどうかを判定せよ．

$\mathcal{TT}_\Lambda$ に関する極小言語問題.
入力: 空でない順序木の集合 $S \subseteq \mathcal{OT}_\Lambda$.
問題: S を説明する極小一般化項木 $t \in \mathcal{TT}_\Lambda$ を見つけよ．

Angluin[5] と Shinohara[9] は，$\mathcal{TT}_\Lambda$ が有限の厚みを持ち，$\mathcal{TT}_\Lambda$ に関する所属性問題と極小言語問題が多項式時間計算可能であるならば，$\mathcal{TT}_\Lambda$ は正データから多項式時間推論可能であることを示した．Moriyama & Sato は [8] において帰納推論のための十分条件を示した．順序項木に関しては以下のとうりである．

定理 1 (Moriyama と Sato [8]) $\mathcal{TT}_\Lambda$ が M-有限の厚みを持つとき，$\mathcal{TT}_\Lambda$ が正データから多項式時間帰納推論可能であるのは全ての順序項木 $t \in \mathcal{TT}_\Lambda$ に対して，$L_\Lambda(t)$ の有限証拠集合が存在し，$\mathcal{TT}_\Lambda$ に関する所属性問題と極小言語問題が多項式時間計算可能であるときであり，またそのときに限る

2.2 M-有限の厚み，有限証拠集合，所属性問題について

補題 1 $\mathcal{OTT}^h_\Lambda$ は M-有限の厚みを持つ．

いかなる $S \subseteq \mathcal{OT}_\Lambda$ と S を説明する極小一般化項木 t の (i,j)-高さ制約変数に対しても，i と j は S の最大の高さを超えないことから，上の補題を簡単に示すことができる．

以下の補題に関して，辺ラベルは少なくとも 2 種類以上存在する，つまり，$|\Lambda| \geq 2$ と仮定する．

補題 2 いかなる $t = (V_t, E_t, H_t) \in \mathcal{OTT}^h_\Lambda$ に対しても，t の有限証拠集合が存在する．

証明 (概略) まず，有限証拠集合 S の構成を考える．$u_0, u_1, ..., u_k$ を頂点とし，a 及び b を Λ の異なる辺ラベルとする．λ でラベル付けされた辺を $\{u, u'\}^\lambda$ と記述する．

$g^{\lambda_1\lambda_2\cdots\lambda_k}$ を $(\{u_0, u_1, ..., u_k\}, \{\{u_0, u_1\}^{\lambda_1}, \{u_1, u_2\}^{\lambda_2}, ..., \{u_{k-1}, u_k\}^{\lambda_k}\})$ であるような順序木とする．ここで $\lambda_\ell \in \{a, b\}$ $(1 \leq \ell \leq k)$ とする．t の変数 $e = [v, v']^{(i,j)}$ に対して，$x(e)$ を e の変数ラベルとし，全ての $i \leq k \leq j$ に対して，束縛 $x(e) := [g^{\lambda_1\lambda_2\cdots\lambda_k}, [u_0, u_k]]$ と定義する．$B(e)$ を変数 $e = [v, v']^{(i,j)}$ に対する全ての束縛の集合とする．このとき $|B(e)| = \sum_{k=i}^{j} 2^k$ である．有限証拠集合 $S \subseteq \mathcal{OT}_\Lambda$ を $S = \{t\theta \mid \theta = \bigcup_{e \in H_t}\{b(e)\}$ ただし $e \in H_t$に対して $b(e) \in B(e)\}$ と定義する．

次に，いかなる $t' = (V_{t'}, E_{t'}, H_{t'}) \in \mathcal{OTT}_\Lambda$ に対しても，$S \subseteq L_\Lambda(t') \subseteq L_\Lambda(t)$ ならば $t' \equiv t$ であることを示す．$V'_t := \{v \in V_t \mid v$ は次のいずれかである: t の根である，t の葉である，2 つ以上の子を持つ根以外の頂点 $\}$ と定義する．$V'_{t'} \subseteq V_{t'}$ を V'_t と同様に定める．$S \subseteq L_\Lambda(t') \subseteq L_\Lambda(t)$ なので，$V'_{t'}$ から V'_t への全単射 ξ が存在し，次の条件を満たす．$v, v' \in V'_{t'}$ に対して，v' が v の祖先なのは $\xi(v')$ が $\xi(v)$ の祖先のときであり，またそのときに限る．いかなる $v \in V'_{t'}$ に対しても，$na(v)$ を v の最も近い祖先とし，$t'[na(v), v]$ を t' の $na(v)$ から v への辺または変数から成るチェーンとする．同様に，$t[na(\xi(v)), \xi(v)]$ を t の $na(\xi(v))$ から $\xi(v)$ への辺または変数から成るチェーンとする．このとき，$S \subseteq L_\Lambda(t')$ と $L_\Lambda(t') \subseteq L_\Lambda(t)$ から $t'[na(v), v] \equiv t[na(\xi(v)), \xi(v)]$ であることを示すことができる．よって $t' \equiv t$ である．□

我々は [11] のアルゴリズムを拡張し，$\mathcal{OTT}^{\mathcal{H}}_\Lambda$ に関する所属性問題を解く多項式時間アルゴリズムを与えた [3]．$\mathcal{OTT}^h_\Lambda$ は $\mathcal{OTT}^{\mathcal{H}}_\Lambda$ のサブクラスなので，このアルゴリズムは $\mathcal{OTT}^h_\Lambda$ に関する所属性問題を多項式時間で解くことができる．

補題 3 $\mathcal{OTT}^h_\Lambda$ に関する所属性問題は多項式時間計算可能である．

第 3 節において，$|\Lambda| \geq 2$ のとき $\mathcal{OTT}^h_\Lambda$ に関する極小言語問題を解く多項式時間アルゴリズムを与える．よって，次の定理を得る．

定理 2 $|\Lambda| \geq 2$ のとき，クラス $\mathcal{OTT}^h_\Lambda$ は正データから多項式時間帰納推論可能である．

3 高さ制約変数を持つ順序項木の極小言語問題に対する多項式時間 MINL アルゴリズム

$|\Lambda| \geq 2$ のとき，$\mathcal{OTT}^h_\Lambda$ に関する極小言語問題を解く MINL アルゴリズム (図 5) を与える．

$t = (V_t, E_t, H_t)$ を $\mathcal{OTT}^h_\Lambda$ に属する順序項木とし，u は t の頂点，v は u の子であるとする．λ を辺ラベル集合 Λ の要素とする．$e = [u, v]^{(i,j)}$ を t の (i,j)-高さ制約変数 $(1 \leq i \leq j)$ とし，$x(e)$ を e の変数ラベルとする．このとき，t の変数ラベル $x(e)$ を持つ変数に対し，次の 10 個の代入を定義する．これらの代入を**精密化演算子** (refinement operator) とよぶ．

$\mathcal{R}_\ell(e) = \{x(e) := [t_\ell, [u, v]]\}$ $(\ell = 1, 2, 3, 4, 5)$ (図 2)
$\mathcal{R}_6(e)_\lambda = \{x(e) := [t_6, [u, v]]\}$ if $i = 1$ (図 3)
$\mathcal{R}_\ell(e)_\lambda = \{x(e) := [t_\ell, [u, v]]\}$ $(\ell = 7, 8, 9)$ (図 3)
$\mathcal{R}_{10}(e) = \{x(e) := [t_{10}, [u, v]]\}$ (図 4)

h_S を S 中の順序木の最大の高さとする．MINL^h はただ 1 つの $(1, h_S)$-高さ制約変数から始め，Variable-Extension，Edge-Replacing 及び Height-Constraint の 3 つのステップを行う．

順序項木 t に対し，t のすべての (i,j)-高さ制約変数を長さ i の辺で置き換えたものを $s(t)$ と記述する．順序項木 t と t' に対し，t と t' の辺ラベルを無視したときに $s(t) \equiv s(t')$ が成り立つならば $t \approx t'$ と記述する．このとき次の補題を証明することができる．

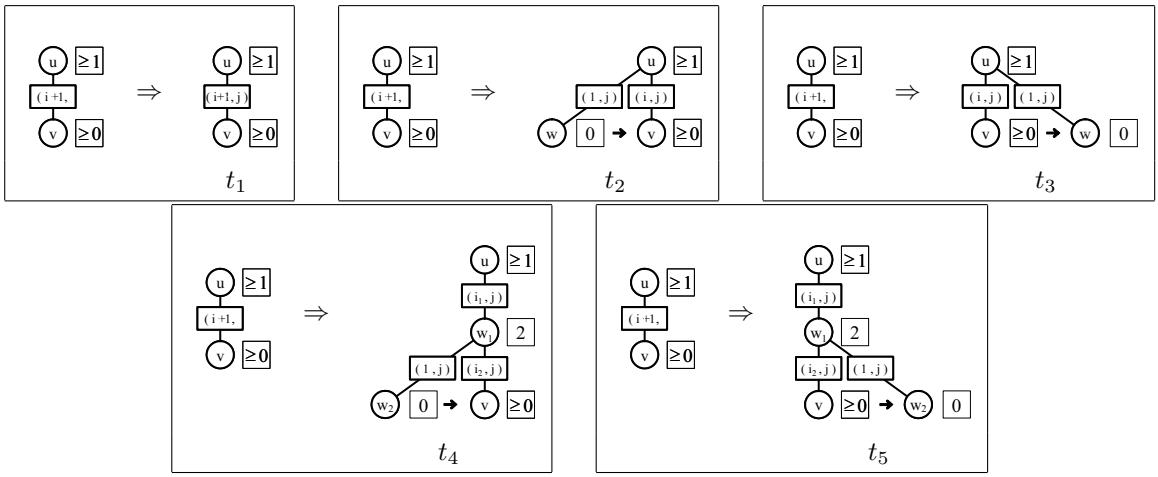

図 2: **精密化演算子** $\mathcal{R}_\ell(e)$ $(\ell = 1, 2, 3, 4, 5)$. 順序項木 t_4 及び t_5 において，$i = i_1 + i_2$ であるとする．頂点 u のそばに書かれた $\boxed{k}$ や $\boxed{\geq k}$ は，u の子の数が k，または k 以上であることを表す．右矢印は，矢印の右の頂点が，矢印の左の頂点の直後の頂点であることを表す.

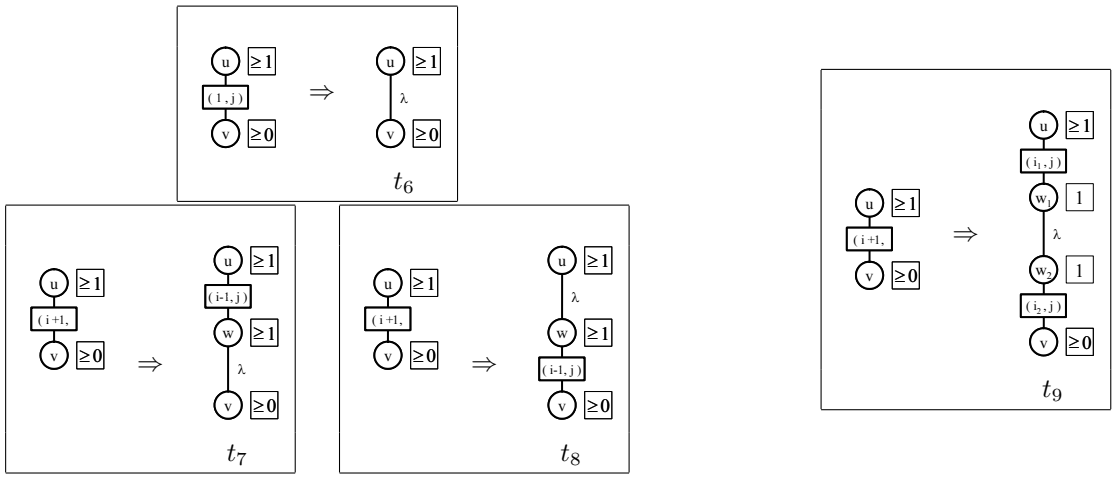

図 3: **精密化演算子** $\mathcal{R}_\ell(e)_\lambda$ $(\ell = 6, 7, 8, 9)$. 順序項木 t_9 において，$i = i_1 + i_2 + 1$ であるとする.

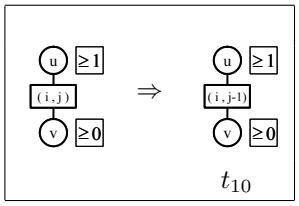

図 4: **精密化演算子** $\mathcal{R}_{10}(e)$.

Algorithm MINL$^h(S)$;
input: 順序木の集合 $S \subseteq \mathcal{OT}_\Lambda$;
begin
Λ_S は S 中の順序木に現れる全ての辺ラベルの集合とする;
h_S は S 中の順序木の最大の高さとする;
$t := (\{u, v\}, \emptyset, \{[u, v]^{(1,h_s)}\})$;
// Variable-Extension
while $S \subseteq L_\Lambda(t\mathcal{R}_\ell(e))$ であるような $e \in H_t$ 及び $1 \leq \ell \leq 5$ が存在する **do** $t := t\mathcal{R}_\ell(e)$;
// Edge-Replacing
while $S \subseteq L_\Lambda(t\mathcal{R}_\ell(e)_\lambda)$ であるような $e \in H_t$, $\lambda \in \Lambda_S$ 及び $6 \leq \ell \leq 9$ が存在する **do** $t := t\mathcal{R}_\ell(e)_\lambda$;
// Height-Constraint
while $S \subseteq L_\Lambda(t\mathcal{R}_{10}(e)_\lambda)$ であるような $e \in H_t$ が存在する **do** $t := t\mathcal{R}_{10}(e)$;
output t
end.

図 5: アルゴリズム MINL^h.

補題 4 Λ を $|\Lambda| \geq 2$ である辺ラベル集合とする．アルゴリズム MINL^h は，与えられた入力順序木の集合 $S \in \mathcal{OT}_\Lambda$ を説明する $\mathcal{OTT}^h_\Lambda$ に属する極小一般化項木を多項式時間で発見する．よって，$\mathcal{OTT}^h_\Lambda$ に関する極小言語問題は多項式時間計算可能である．

証明 (概略) アルゴリズム MINL^h の正当性を示すために以下の 3 つの主張を示す.

主張 *1.* S を入力とした Variable-Extension の出力を t とする．S を説明する極小一般化項木を t' とすると，もし $S \subseteq L_\Lambda(t') \subseteq L_\Lambda(t)$ ならば $t' \approx t$ である.

主張 *2.* Edge-Replacing の出力を t とする．t' を $S \subseteq L_\Lambda(t') \subseteq L_\Lambda(t)$ を満たす順序項木とする．このとき，$V_{t'}$ から V_t への全単射 ξ が存在し，以下の条件を満たす．t' の $\{v, v'\}$ が辺であるのは t の $\{\xi(u), \xi(v')\}$ が辺であるときであり，またそのときに限る．さらにその辺ラベルは等しい.

主張 *3.* Height-Constraint の出力を t とする．t' を $S \subseteq L_\Lambda(t') \subseteq L_\Lambda(t)$ を満たす順序項木とする．主張 2 より，$V_{t'}$ から V_t への全単射 ξ が存在する．このとき，t' の $[v, v']$ が (i, j)-高さ制約変数であるのは，t の $[\xi(v), \xi(v')]$ が (i, j)-高さ制約変数であるときであり，またそのときに限る． □

4 結論と今後の課題

本論文では，$\mathcal{OTT}^h_\Lambda$ のクラスが正データから多項式時間帰納推論可能であることを示した．現在，本手法を応用したデータマイニングシステムの開発を行っている．

参考文献

[1] S. Abiteboul, P. Buneman, and D. Suciu, Data on the web: From relations to semistructured data and XML, Morgan Kaufmann, 2000.

[2] K. Aikou, Y. Suzuki, T. Shoudai, and T. Miyahara, Metasearch with ordered tree structured patterns, Hinokuni Information Symposium, IPSJ Kyushu Chapter, 2004.

[3] K. Aikou, Y. Suzuki, and T. Shoudai, A Polynomial Time Matching Algorithm of Structured Ordered Tree Patterns with Height-Constrained Variables, *to be submitted*, 2004.

[4] T. R. Amoth, P. Cull, and P. Tadepalli, On exact learning of unordered tree patterns, *Machine Learning*, **44**, pp. 211–243, 2001.

[5] D. Angluin, Inductive inference of formal languages from positive data, *Information and Control*, **45**, pp. 117–135, 1980.

[6] H. Arimura, H. Sakamoto, and S. Arikawa, Efficient learning of semi-structured data from queries, *Proc. ALT-2001, Springer-Verlag, LNAI 2225*, pp. 315–331, 2001.

[7] T. Miyahara, Y. Suzuki, T. Shoudai, T. Uchida, S. Hirokawa, K. Takahashi, and H. Ueda, Extraction of tag tree patterns with contractible variables from irregular semistructured data, *Proc. PAKDD 2003, Springer-Verlag, LNAI 2637*, pp. 430–436, 2003.

[8] T. Moriyama and M. Sato, Properties of language classes with finite elasticity, *IEICE Transactions on Information and Systems*, **E-78-D(5)**, pp. 532–538, 1995.

[9] T. Shinohara, Polynomial time inference of extended regular pattern languages, *Proc. RIMS Symp. on Software Science and Engineering, Springer-Verlag, LNCS 147 (1982)* pp. 115–127.

[10] Y. Suzuki, R. Akanuma, T. Shoudai, T. Miyahara, and T. Uchida, Polynomial time inductive inference of ordered tree patterns with internal structured variables from positive data, *Proc. COLT-2002, Springer-Verlag, LNAI 2375*, pp. 169–184, 2002.

[11] Y. Suzuki, K. Inomae, T. Shoudai, T. Miyahara, and T. Uchida, A Polynomial Time Matching Algorithm of Structured Ordered Tree Patterns for Data Mining from Semistructured Data, *Proc. ILP-2002, Springer-Verlag, LNAI 2583*, pp. 270–284, 2003.

[12] Y. Suzuki, T. Shoudai, T. Uchida, and T. Miyahara, Ordered term tree languages which are polynomial time inductively inferable from positive data, *Proc. ALT-2002, Springer-Verlag, LNAI 2533*, pp. 188–202, 2002.

[13] Y. Suzuki, T. Shoudai, S. Matsumoto and T. Uchida, Efficient Learning of Unlabeled Term Trees with Contractible Variables from Positive Data, *Proc. ILP-2003, Springer-Verlag, LNAI 2835*, pp.347–364, 2003.