並列計算量理論に基づく木構造パターンの並列照合アルゴリズムと並列機械学習
ー順序木構造パターンに対する学習アルゴリズムとその応用ー

並列計算量理論に基づく木構造パターンの並列照合アルゴリズムと並列機械学習

－順序木構造パターンに対する学習アルゴリズムとその応用－

正代　隆義（情報工学部情報工学科）

An Efficient Parallel Matching Algorithm for Linear Ordered Term Tree Pattern Matching Problem

－Learning Algorithms for Ordered Tree Structured Patterns and Their Applications－

SHOUDAI Takayoshi （Department of Computer Science and Engineering, Faculty of Information Engineering）

Abstract

The pattern matching problem for linear ordered term tree patterns (LOTT-patterns, for short) is the problem of deciding, given an LOTT-pattern t and an ordered tree T, whether t matches T or not. In this research project, an efficient parallel algorithm for the pattern matching problem for LOTT-patterns was proposed. Moreover, computer experiments were conducted on a GPU-equipped PC running a sequential polynomial-time algorithm that computes the pattern matching problem for LOTT-patterns to extract characteristic ordered tree structure patterns for real data.

Keywords：Ordered term tree pattern, Pattern matching, Efficient parallel algorithm, Query learning, Graph convolutional network.

1.　研究の内容

グラフデータから意味のある構造的知識を抽出するためには、(1) データ中に潜むパターンをどのようにグラフパターンとして表現するか、そして、(2) データを高精度に表現するグラフパターンをどのようにして高速に発見するかを議論する必要がある。本研究課題では、既存のグラフ構造パターンとして構造的変数を持つ順序木構造パターンに注目し、主として(2)の観点から研究を行った。

HTML/XML のようなデータは関係データベースのように明確な構造を持つわけではないが、タグ付けによる順序木構造を持つ。順序木構造データから構造的なパターンを発見するために、Suzuki ら[1]は線形順序項木パターン(*Linear Ordered Term Tree Pattern*、LOTT-パターンと略す)を提案した。LOTT-パターンは順序木構造データに対して設計された高い表現力を持つグラフパターンである。

並列ランダムアクセス機械(PRAM)とは、共有メモリをプロセッサ間の通信手段とする並列計算の理論的モデルである。効率のよい並列アルゴリズムとは、入力サイズnに対して、$O(n^k)$個のプロセッサを持つ PRAM 上で、$O((\log n)^c)$時間で計算するアルゴリズム(kとcはnに依存しない定数)のことをいう。効率のよい並列アルゴリズムを持つ計算問題のクラスを NC という。クラス P に含まれる計算問題がいつでも NC に含まれるかという問いは並列計算量理論における未解決問題である。LOTT-パターン照合問題に対する最初の効率のよい並列アルゴリズムは財津[2]によって提案された。財津のアルゴリズムは、宮野[3]による推論の並列化手法と Suzuki ら[1]によって提案された$O(nN)$時間逐次アルゴリズムの正当性に依存して設計されている。ここで、nとNはLOTT-パターン及び順序木の頂点数である。宮野[3]による推論の並列化手法は一般的な推論に関するもので、財津はLOTT-パターン照合並列アルゴリズムの計算量解析を厳密に行なっていない。本研究課題で提案した効率のよい並列アルゴリズムは、Suzuki らの逐次アルゴリズム[1]の正当性には依存しない。第 2 章で、LOTT-パターンと LOTT-パターン照合問題を定義し、本研究課題で提案した効率のよい並列アルゴリズムの概略及びプロセッサ数と並列時間計算量の理論的解析の結果を述べる。

大規模なデータを対象としてデータマイニングを行うときには、一般に、多くのパターン照合問題を計算する必要がある。質問学習とは、Angluin (1988)により提唱された計算論的学習モデルで、学習者が常に正答を返す教師(オラクル)に質問を繰り返すことで、教師の有する概念を同定する学習手法である。LOTT-パターン照合問題の応用として、本研究課題では、小田ら[4]が提案したグラフ畳み込みネットワーク(GCN)と質問学習の協調学習モデルについて、実データによる有効性を確認した。具体的には、二値分類精度である F 値を求めることで、高精度 GCN モデルをオラクルとした質問学習アルゴリズムを HTML データ上で評価した。評価を行うには、順序木データ数の多項式回数の LOTT-パターン

照合問題を計算する必要がある。本研究課題では、GPU 付き PC において LOTT-パターン照合問題を計算する逐次アルゴリズムを動作させて、実データに特徴的な順序木構造パターンの抽出を行った。その実験において得られた LOTT-パターンと実データによる実験的評価を第 3 章で述べる。

2. LOTT-パターン並列照合アルゴリズム

HTML/XML のようなデータは関係データベースのように明確な構造を持つわけではないが、タグ付けによる順序木構造を持つ。順序木構造データから構造的なパターンを発見するために、我々は過去の研究[1]で、超辺置換に基づく順序木構造パターンとして、任意の変数次数を持つ順序項木パターンを提案した。本研究課題では、変数次数 2 の順序項木パターンのみを扱う。

$T = (V_T, E_T)$を頂点集合V_Tと辺集合E_Tを持つ順序木とする。Tにおいて、頂点u_1が頂点u_0の子であれば、それらを繋ぐ辺を順序対(u_0, u_1)のように表す。Xを変数ラベルの無限集合とする。Tの次数 2 の変数とは、次の条件を満たすV_Tの頂点の順序対$h = [u_0, u_1]$である: u_1はu_0の子であり、Xに属す変数ラベルが唯一つ対応する。これ以降、次数 2 の変数を単に変数と呼ぶ。H_TをTの変数の集合とし、$V_t = V_T$、$E_t = E_T \setminus \left(\bigcup_{[u_0,u_1]\in H_t}\{(u_0, u_1) \in E_t\}\right)$、$E_t = E_T$とするとき、3 つ組$t = (V_t, E_t, H_t)$を順序項木パターンと呼ぶ。さらに、$H_t$の全ての変数が$X$に属す互いに異なる変数ラベルに対応しているとき、$t = (V_t, E_t, H_t)$を線形順序項木パターン(*Linear Ordered Term Tree Pattern*、LOTT-パターンと略す)と呼ぶ。頂点$u_0, u_1 \in V_t$に対して、$(u_0, u_1) \in E_t$または$[u_0, u_1] \in H_t$であるとき、u_1はu_0の子、u_0はu_1の親と呼ぶ。

$t = (V_t, E_t, H_t)$を LOTT-パターンとする。$h = [u_0, u_1]$を変数ラベルxを持つtの変数とする。$g = (V_g, E_g)$を頂点数 2 以上の順序木とする。hへの順序木gの代入を定義する。$\rho = [v_0, v_1]$をv_0がgの根であり、v_1がgの葉であるようなgの頂点の順序対とする。代入式$x := [g, \rho]$をxに関する束縛と呼ぶ。tに束縛$x := [g, \rho]$を適用して、新しい LOTT-パターン$t\{x := [g, \rho]\}$を次のように得る: 変数ラベルxを持つ変数$h = [u_0, u_1]$について、H_tから変数hを削除し、Tの頂点v_0, v_1をこの順序でtの頂点u_0, u_1と同一視することにより、gをtに貼り付ける。代入θとは、異なる変数ラベルに関する束縛の有限集合$\theta = \{x_1 := [g_1, \rho_1], \ldots, x_n := [g_n, \rho_n]\}$ $(x_i \in X, 1 \le i \le n)$のことである。$t$に$\theta$の束縛を全て適用して得られる LOTT-パターンを$t\theta$と書く。図 1 に LOTT-パターンtと照合する順序木Tの例を示す。図 1 において、代入$\theta = \{x := [g_1, [v_0, v_1]], y := [g_2, [w_0, w_1]]\}$を適用することにより得られる$t\theta$は順序木$T$と同型となる。

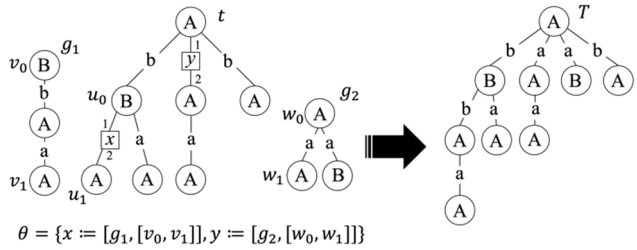


図 1 LOTT-パターンtと照合する順序木Tの例: xを変数ラベルとする変数$[u_0, u_1]$は、変数ラベルxを囲む四角と、その四角から変数に属す頂点u_0, u_1に向かう 2 つの線で示される。線上の数字は変数における頂点の順番を表す。

LOTT-パターンtと順序木Tに対して、$t\theta$がTが同型となるような代入θが存在するとき、tはTに照合する(マッチする)という。LOTT-パターン照合問題は次のように定義される決定問題である:

LOTT-パターン照合問題
入力: LOTT-パターンtと順序木T;
問題: tはTに照合するか?

本研究課題では LOTT-パターン照合問題を計算する新しい効率のよい並列アルゴリズムを提案した。アルゴリズムのアイデアは次のとおりである。

(Step 1) 入力の線形順序項木パターンtと順序木Tを、二分均衡化する(詳細は省略する)。順序木Tを二分均衡化した順序木を証明木と呼ぶ。頂点数Nが 6 以上の順序木Tに対して、高さが高々16 $log\ N$であるような証明木を構築できることが示される。順序木とその順序木を二分均衡化することによって得られる証明木の例を図 2 にあげる。

(Step 2) 二分均衡化後のtとTに対して、動的計画法を用いて葉から根へ向けて頂点の対応を定める。

Tの証明木の同じ深さの頂点に対して並列にtの証明木との対応関係を計算することができるので、計算時間はTの証

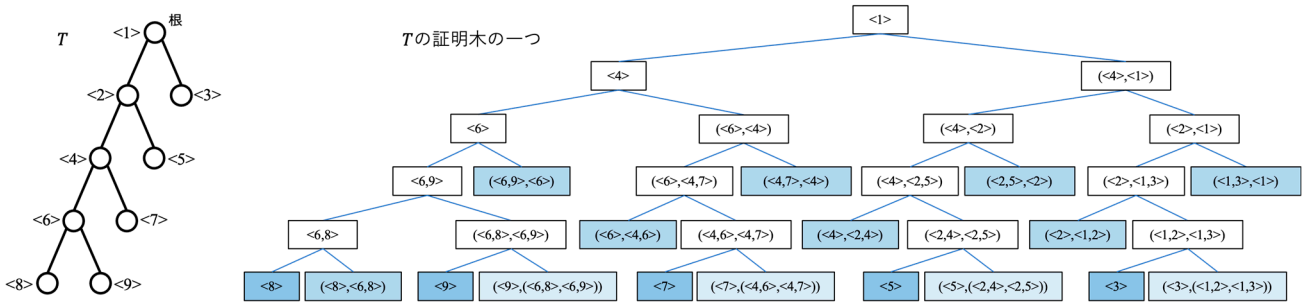


図 2 順序木Tとその証明木の一つ: 順序木の頂点には、その頂点を識別する固有の番号または記号(列)が与えられているものとする。その固有の記号を頂点識別子と呼び、頂点vに対して$\langle v \rangle$で表す。

明木の高さに依存する。このことより、次の定理を得る(証明は省略する)。

定理: LOTT-パターン照合問題は、頂点数nの LOTT-パターンと頂点数Nの順序木に対して、$O(n^2N)$個のプロセッサを持つ PRAM を用いて$O(log\ N)$時間で計算可能である。

3. HTML 木構造データの LOTT-パターン抽出

Web スクレイピング

Web スクレイピング(クローリング)とは、Web サイトから自動的に情報を収集する処理である。主として Web ブラウザを使った人間以外の手段で、データを収集する作業を指す。本研究課題での目標は次の 2 つである。(I) Python の BeautifulSoup4 ライブラリを利用して、Web サイトのトップページの HTML データを収集し、順序木構造データベースを構築する。(II) 順序木構造データのための質問学習モデルを、(I)で収集した順序木構造データベースに適用し、モデルの有効性を確認する。

本研究課題では次の 3 つの Web サイトを対象とした。

(1) Wikipedia: 福岡工業大学から始めて Wikipedia 内のページをクローリングし、1807 ページ(38.2MB)を収集した。

(2) Wikipedia: 良質記事のページ内の Wikipedia の記事リンクを 1806 ページ(25.9MB)収集した。

(3) 日本株の配当金データベース: 日本株の配当金データベースから 3649 ページ(13.1MB)の上場企業 HP を収集した。

本研究では、収集した HTML データを、HTML タグを辺ラベルとする順序木に変換し、さらにタグの役割に基づいてグループ分けした番号を辺ラベルとした順序木を順序木パターン学習実験の対象データとした。HTML 文書の順序木への変換過程を図 3 に示す。

グラフ畳み込みネットワーク(RGCNConv)

実験では、Python のグラフニューラルネットワークライブラリである Pytorch Geometric に実装されているグラフ畳み込みネットワーク(GCN と略す)のレイヤ RGCNConv による 6 層構造を用いた。簡単に RGCNConv について説明する。Λを辺ラベルの集合とする。グラフの頂点vの辺ラベル$r \in \Lambda$を持つ辺によって接続された隣接頂点の集合を$N_r(v)$とする。RGCNConv は、頂点vの特徴ベクトル$\boldsymbol{x}_v$と重み行列W_1を乗じたものと、辺ラベル$r \in \Lambda$に対する重み行列W_rを$N_r(v)$に属す頂点v'の特徴ベクトル$\boldsymbol{x}_{v'}$に乗じて平均したものを、全ての辺ラベル$r \in \Lambda$について足し合せて、vの新たな特徴ベクトル$\boldsymbol{x'}_v$とする。まとめると次の式で表される。

$$\boldsymbol{x'}_v = W_1\boldsymbol{x}_v + \sum_{r\in R}\left(\sum_{v'\in N_r(v)}\frac{1}{|N_r(v)|}W_r\boldsymbol{x}_{v'}\right).$$

小田ら[4]は、順序木データを超高精度で二値分類する GCN をオラクルとし、それを用いた順序木構造パターンの質問学習モデルを提案した。実験では、Pytorch Geometric の GCN レイヤである RGCNConv による 6 層構造に対して、HTML 木構造データの各頂点には深さなど次の 6 種類の特徴量を入力層の特徴ベクトルとして与えた: (i) vの根からの深さ、(ii) vの兄弟関係における順番、(iii) vの子の数、(iv) vの子孫の数(vから深さ 2 まで)、(v) vの子孫の数(vから深さ 3 まで)、(vi) vの子孫の数。

RGCNConv と質問学習による LOTT-パターンの学習

実験では、正例を(2)により構築した順序木(1806 ページ)、負例を(3)により構築した順序木(3649 ページ)からランダムに選択した順序木(1806 ページ)とした。そして、それらを学習データとして学習済み RGCNConv を構成した。学習済み RGCNConv を複数回構成して、各学習済み RGCNConv をオラクルとする質問学習により LOTT-パターンを求めた。高精度を達成した LOTT-パターンの一部を図 4 に示す。

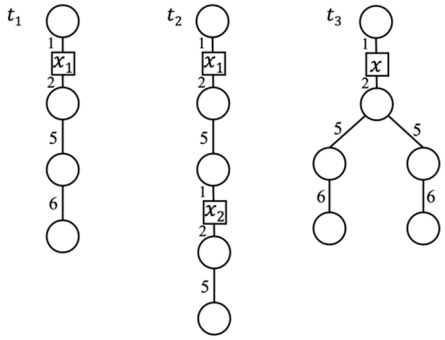
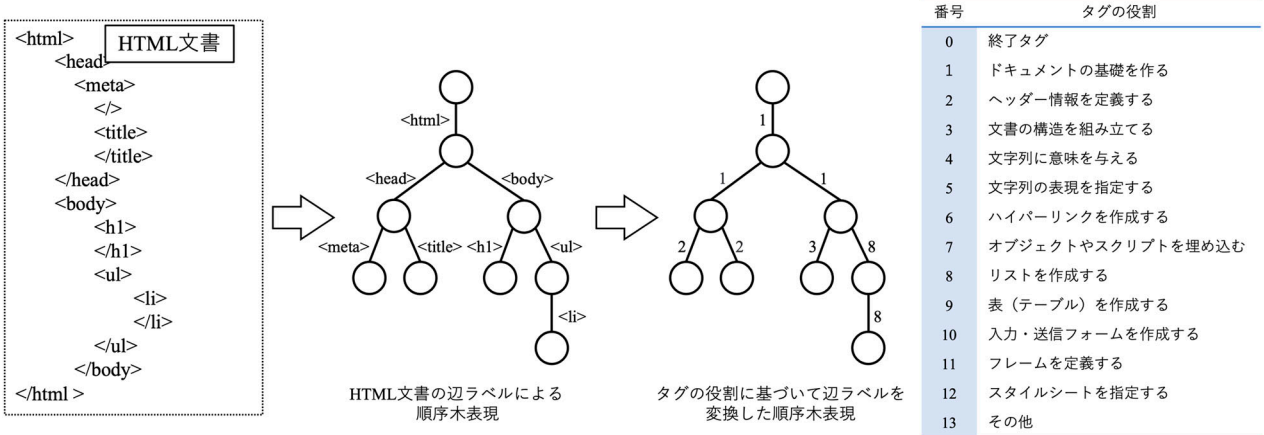


図 4 学習済み RGCNConv をオラクルとする質問学習により得られた LOTT-パターンt_1, t_2, t_3



番号	タグの役割
0	終了タグ
1	ドキュメントの基礎を作る
2	ヘッダー情報を定義する
3	文書の構造を組み立てる
4	文字列に意味を与える
5	文字列の表現を指定する
6	ハイパーリンクを作成する
7	オブジェクトやスクリプトを埋め込む
8	リストを作成する
9	表（テーブル）を作成する
10	入力・送信フォームを作成する
11	フレームを定義する
12	スタイルシートを指定する
13	その他

図 3 Web スクレイピングにより収集した HTML 文書を辺ラベル付き順序木データに変換する過程

	再現率	適合率	F値
RGCNConv M_1	1.000	0.997	0.999
LOTT-パターン t_1	0.990	0.931	0.960
TP: 1788, FN: 18, FP: 132, TN: 1674			
	再現率	**適合率**	**F値**
RGCNConv M_2	1.000	0.997	0.999
LOTT-パターン t_2	0.968	0.994	0.981
TP: 1748, FN: 58, FP: 11, TN: 1795			
	再現率	**適合率**	**F値**
RGCNConv M_3	1.000	0.999	1.000
LOTT-パターン t_3	0.993	0.966	0.979
TP: 1793, FN: 13, FP: 63, TN: 1743			

図 5 学習済み RGCNConv M_1, M_2, M_3 の分類精度と各 RGCNConv をオラクルとする質問学習により得られた LOTT-パターンt_1, t_2, t_3(図 4)の分類精度

t_1, t_2, t_3を求めるためにオラクルとして用いた学習済み RGCNConv M_1, M_2, M_3 の分類精度とt_1, t_2, t_3の分類精度を図 5 に示す。図 5 に示された分類精度から、図 4 に示された LOTT-パターンt_1, t_2, t_3は Wikipedia の良質記事に高頻度で照合し、上場企業 HP にはほぼ照合しないことがわかる。

4. 今後の展開

本研究課題では、次の 2 つの研究成果を報告した。

1. LOTT-パターン照合問題に対して、効率のよい並列アルゴリズムを提案し、プロセッサ数と並列時間計算量の理論的解析を行った。
2. LOTT-パターンの実データに対する有効性を確認するために、高精度 GCN モデルをオラクルとした質問学習アルゴリズムを HTML データ上で評価した。

結果 1 は、2023 年度(第 76 回)電気・情報関係学会九州支部連合大会において発表予定である。また、結果 2 は、東山ら[5]による研究発表の一部である。

今後は、形式グラフ体系(FGS)[6]等のグラフパターン表現に対して、並列化による学習アルゴリズムの高速化を行う。石灘ら[7]は無順序木パターンに対する GCN と質問学習の協調学習が、順序木パターンと同様に有効であることを示した。そのため無順序木パターンに対する並列化による高速化は喫緊の課題である。PRAM は通信手段の制約により現実的な並列計算モデルではない。本研究課題で述べた並列アルゴリズムをベースに、GPU などを利用した現実的な並列計算モデルを用いて、実データに即した並列化を行うことが今後の課題である。

謝辞

本研究は本学情報科学研究所の 2022 年度研究スタートアップ支援制度により実施したものである。

文　献

（1） Y. Suzuki, T. Shoudai, T. Uchida, and T. Miyahara: “Ordered term tree languages which are polynomial time inductively inferable from positive data,” Theoretical Computer Science, 350(1), (2006) pp.63-90.
（2） 財津 恭太:「順序木構造パターンの並列マッチングアルゴリズムについて」、平成 13 年度九州大学大学院システム情報科学府修士論文 (2002)
（3） 宮野 悟:「並列アルゴリズム」、近代科学社 (1993)
（4） 小田 直季、内田 智之、正代 隆義、松本 哲志、鈴木 祐介、宮原 哲浩:「順序木パターンの質問学習アルゴリズムによるグラフ畳み込みネットワークの予測根拠の可視化」、2022 年度人工知能学会全国大会(第 36 回)論文集, (2022) p.2G4GS201.
（5） 東山 的生、野口 大悟、内田 智之、正代 隆義、松本 哲志:「学習済超高精度 GCN をオラクルとする順序項木パターンの質問学習モデルの解析と実データでの評価」、情報処理学会第 85 回全国大会講演論文集, 1 (2023) pp.371-372.
（6） T. Shoudai, S. Matsumoto, Y. Suzuki, T. Uchida, and T. Miyahara: “Parameterized Formal Graph Systems and Their Polynomial-Time PAC learnability,” IEICE Trans. Fundamentals, E106-A(6), (2023) pp.896-906.
（7） 石灘 洸樹、正代 隆義、内田 智之、松本 哲志:「超高精度グラフ畳み込みネットワークをオラクルとする無順序木パターンの質問学習モデル」、情報処理学会第 85 回全国大会講演論文集, 1 (2023) pp.373-374.