様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２７ 日現在

機関番号：２５４０３
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１２１０３

研究課題名（和文）大規模グラフ構造圧縮データに対する並列グラフマイニングシステムの開発

研究課題名（英文）Development of parallel graph mining systems for compressed large-scale graph structured data

研究代表者
内田　智之（Uchida, Tomoyuki）

広島市立大学・情報科学研究科・准教授

研究者番号：７０２６４９３４

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：可逆圧縮された大規模グラフ構造データから、陽には解凍せずに構造的特徴を表すグラフパターンを獲得する、GPUを活用したグラフマイニングシステムの開発について研究を行った。成果として、多重圧縮順序木集合に対する頻出項木パターンマッチングアルゴリズムおよび並列枚挙アルゴリズム、さらに文字列、順序木、無順序木の二値分類集合をそれぞれ学習させた超高精度グラフ畳み込みネットワーク(GCN)をオラクルとし、1つの正例と線形回数の所属性質問を用いて正則パターン、順序項木パターン、木パターンをそれぞれ同定する質問学習アルゴリズムを提案・解析を行い、GCNを活用した並列グラフマイニングシステムの基盤を構築した。

研究成果の学術的意義や社会的意義
グラフ構造を有しかつ大規模化しているソーシャル・ネットワークやタンパク質相互作用ネットワークなどの大規模グラフ構造データを可逆圧縮してデータサイズを小さくしたグラフ構造圧縮データから、陽に解凍することなく知識を獲得する高速グラフマイニング手法の開発を行った。何度もデータ走査を繰り返しながら特徴的なグラフ構造を抽出する計算時間の削減に貢献した。また、計算論的学習理論に基づき、GPUを用いたグラフ構造データに対する深層学習手法であるグラフ畳み込みネットワークを活用した特徴的な木パターンの抽出手法を開発することで、GPUをパターン抽出に活用する基盤を構築した。

研究成果の概要（英文）：We studied the development of GPU-based graph mining systems that acquire feature graph patterns from losslessly compressed large-scale graph structured data without decompressing the data. As results, we proposed frequent term tree pattern matching algorithms and parallel enumeration algorithms for multiply compressed ordered tree sets. Furthermore, we developed and analyzed query learning algorithms that identify regular patterns, ordered term tree patterns, or tree patterns, using a single positive example and a linear number of membership queries for oracles, which is ultra-high precision graph convolutional networks (GCNs) learned by giving binary classification sets of strings, ordered trees, or unordered trees as training data, respectively. By doing so, we developed a foundation for parallel graph mining systems utilizing GCNs.

研究分野：知能情報学

キーワード：グラフマイニング　グラフ構造圧縮　機械学習　グラフ畳み込みネットワーク

１．研究開始当初の背景

ICT 技術の発展に伴い、Web 文書の構文、ソーシャル・ネットワーク、タンパク質相互作用ネットワークといったグラフ構造を有するグラフ構造データが大規模化していた。部分グラフ同型問題が NP 完全であることが知られている一方で、大規模化したグラフ構造データから知識を効率的に獲得し、得られた知識を効果的に活用したいという要求が高まっていた。その要求に応えるべく高速・省メモリグラフマイニング手法の開発に取り組んだ。特に、画像処理を行う GPU を使って汎用計算を行う GPGPU (General-purpose computing on graphics processing units) を用いることで、大規模グラフ構造データに頻出する部分グラフを高速に枚挙する並列グラフマイニング手法がいくつか提案されていた。また、省メモリ化のため DFUDS (Depth-First Unary Degree Sequence) 表現やウェブレット木 (Wavelet Tree) を簡潔ビットベクトル上で実現した簡潔データ構造についての研究も進んでいた。これらの研究背景のもと、報告者らは、木構造データを可逆圧縮することでデータサイズを小さくし、圧縮された大規模な木構造データから、陽に解凍することなく頻出パスおよび頻出部分木を枚挙するグラフマイニングアルゴリズムを提案していた。

２．研究の目的

Web 文書の構文、ソーシャル・ネットワークやタンパク質相互作用ネットワークといったグラフ構造を有するグラフ構造データの解析には、1 回のデータ走査では難しく、何度もデータ走査を繰り返す必要がある。このため、ICT 技術の発展に伴い大規模化したグラフ構造データの解析に要する計算コストは大きくなる。本研究課題では、走査時間を削減しかつグラフマイニングに要する計算時間を短縮するため、画像処理を高速に実行する GPU を活用し、グラフ構造データを可逆圧縮して得られるグラフ構造圧縮データから、陽に解凍することなくより高度な特徴を表すグラフパターンを高速かつ省メモリで獲得する並列グラフマイニングシステムを開発することを目的とする。

３．研究の方法

大規模なグラフあるいはグラフの集合を可逆圧縮して得られるグラフ (圧縮グラフ)あるいは圧縮グラフの集合に現れる特徴的な部分グラフ間の接続関係を表現するグラフパターン (項グラフパターン) あるいはその集合を知識としてそれぞれ獲得するグラフマイニング手法を、本研究課題の核心をなす学術的問いである (i) 圧縮グラフをウェブレット木に基づく簡潔データ構造を用いることで省メモリ化できるか、(ii) GPGPU を用いて並列化することで高速化できるか、(iii) GPU を活用した並列グラフマイニングシステムを構築できるかの問いに、理論面と実用面の両方で答えを見出すべく、以下の方法で研究に取り組んだ。① グラフ構造圧縮データに対する並列グラフパターンマッチングアルゴリズムの理論展開、② GPGPU に対応したグラフ構造圧縮データの簡潔データ構造の開発、③ グラフ構造圧縮データに対する並列グラフマイニング手法の開発、④ GPGPU を用いたグラフ構造圧縮データに対する並列グラフパターンマッチング手法の開発と実装、および、⑤ GPGPU を用いた並列グラフマイニングシステムの構築を行う予定であった。しかし、研究を進めていく中で、グラフ構造圧縮データの簡潔データ構造上の関数演算に GPU を活用することで、圧縮木と順序木パターンとのグラフパターンマッチングアルゴリズムを高速化することは難しいと判断し次のような方針転換を行った。簡潔データ構造を用いた関数計算に GPU を活用するのではなく、GPGPU の一つであるグラフ畳み込みネットワーク（Graph Convolution Network, GCN）を活用したグラフマイニング手法の開発を行うこととした。

４．研究成果

本研究課題で得られた研究成果について述べる。

① グラフ構造圧縮データに対する並列グラフパターンマッチングアルゴリズムの理論展開

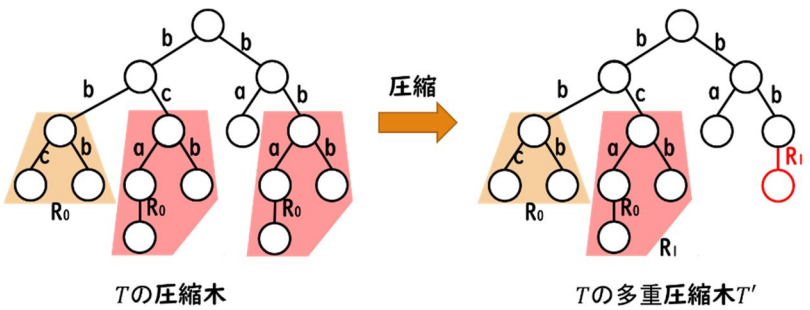

図１　順序木 T の多重圧縮木 T’

簡潔データ構造で表現された可逆圧縮された順序木（圧縮木：図１参照）から頻出パスを並列に枚挙するアルゴリズムおよび圧縮木と構造的特徴を表す項木パターンに対するマッチングアルゴリズムを提案した。さらに、変数ラベルが一つしかない１変数項木パターンの枚挙アルゴリズムの提案および形式グラフ体系（Formal Graph System, FGS）に関する多項式時間 PAC 学習可能性について示した。これにより、多重圧縮木の集合を対象としたグラフマイニング手法の基礎的理論を構築することができた。

② GPGPU に対応したグラフ構造圧縮データの簡潔データ構造の開発

図１に示した多重圧縮木のデータ表現として簡潔データ構造を採用しているため、省メモリ化と高速化を同時に実現しているが、高速化に大きく寄与している簡潔データ構造上の関数演算（Rank や Select など）のさらなる高速化のため、GPU の活用を試みた。研究を進めていくにつれ、簡潔データ構造上の関数演算の高速化に GPU を活用することが難しいという結論を得た。このため、関数演算の高速化に GPU を活用するのではなく、計算論的学習理論に基づいた質問学習アルゴリズムに GPU を活用したグラフマイニング手法の開発に舵を切った。そのため①の基礎理論で FGS に関する多項式時間 PAC 学習可能性について考察を行った。

③ グラフ構造圧縮データに対する並列グラフマイニング手法の開発

研究テーマ①の研究成果に基づき、多重圧縮木集合に対する頻出項木パターン並列枚挙アルゴリズム ParaEnuFPforMultiCompTree を開発した。図２には、ParaEnuFPforMultiCompTree における並列化の手続き PMFORMCT の流れを示している。評価実験を行い、提案アルゴリズム ParaEnuFPforMultiCompTree が十分効率的であることを示した。加えて、進化的学習手法についても検討を行った。

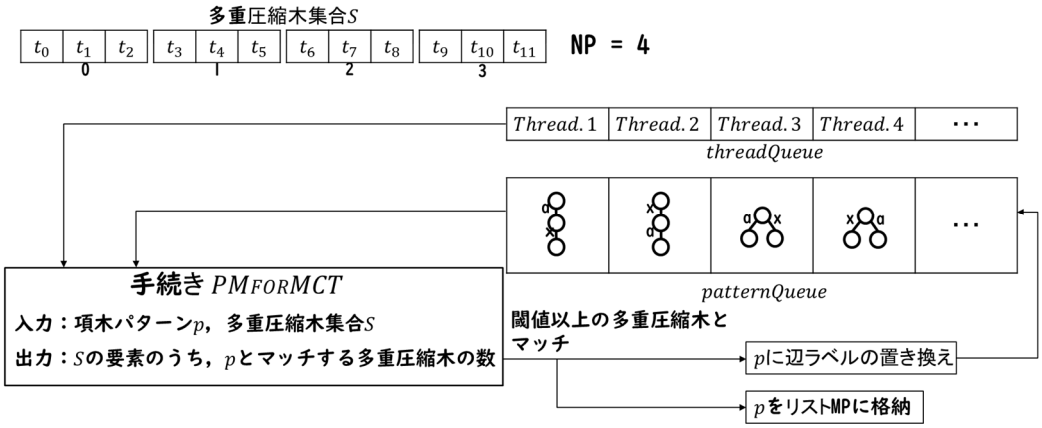

図２ 多重圧縮木集合からの頻出項木パターン並列枚挙アルゴリズムの概略

④ GPGPU を用いたグラフ構造圧縮データに対する並列グラフパターンマッチング手法の開発と実装および⑤ GPGPU を用いた並列グラフマイニングシステムの構築

研究テーマ②の考察に基づき、簡潔データ構造上の関数計算に GPU を活用するのではなく、計算論的学習理論に基づいた質問学習アルゴリズムに GPU を活用するマイニングシステムの開発に方針転換を行った。Angluin により提唱された質問学習モデルは、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。本研究では、順序木データを学習した高精度なグラフ畳み込みネットワーク(GCN)モデルをオラクルとし、順序木データに共通する概念の表現である順序項木パターンを獲得する質問学習アルゴリズム $VA_{GCN}^{LearnOTT}$（図３参照）を提案した。さらに、$VA_{GCN}^{LearnOTT}$ により獲得した順序項木パターンについて二値分類精度である F 値に関して分析することで、超高精度 GCN モデルをオラクルとした質問学習アルゴリズム $VA_{GCN}^{LearnOTT}$ を評価した。また、Wikipedia などの Web ページを学習させた超高精度 GCN モデルをオラクルとした質問学習アルゴリズム $VA_{GCN}^{LearnOTT}$ により獲得した高い F 値を有する順序項木パターンについて報告した。これにより、実データ に対する超高精度 GCN をオラクルとした質問学習アルゴリズムの有効性を示すことができた。この研究で採用した圧縮木の表現形式を FGS における項として扱う項グラフに採用することで、グラフ$\mathbf{G}$のみを生成する FGS を定義することができる。これを拡張し、圧縮木の集合$\mathbf{S}$に対し、$\mathbf{S}$だけを生成する FGS Γ も定義することができる。この$\mathbf{S}$だけを生成する FGS Γ をターゲットとして、本研究課題で開発した質問学習アルゴリズム $VA_{GCN}^{LearnOTT}$ を適用することで並列グラフマイニングシステムを構築することが今後の課題である。

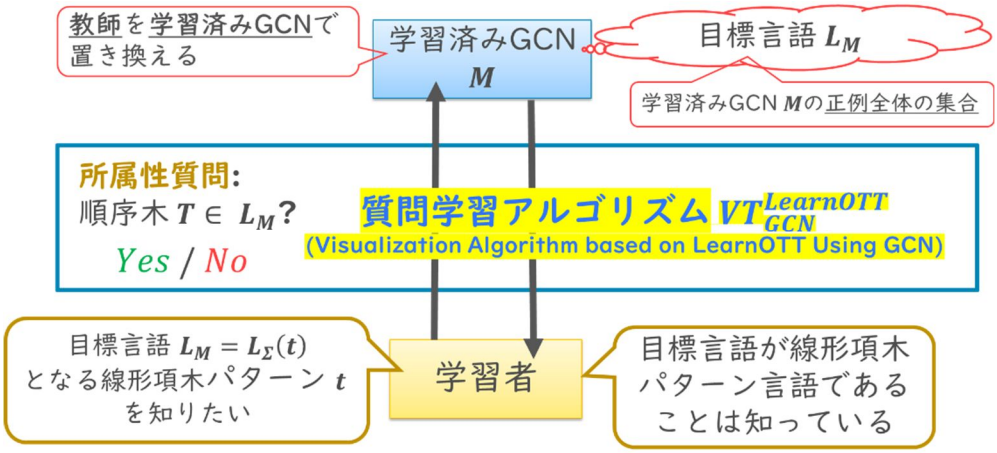

図３　高精度学習済み GCN をオラクルとする質問学習アルゴリズム $VA_{GCN}^{LearnOTT}$

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Takayoshi SHOUDAI, Satoshi MATSUMOTO, Yusuke SUZUKI, Tomoyuki UCHIDA, Tetsuhiro MIYAHARA	４．巻 E106-A(6)
２．論文標題 Parameterized Formal Graph Systems and Their Polynomial-Time PAC learnability	５．発行年 2023年
３．雑誌名 IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences	６．最初と最後の頁 to appear
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.2022EAP1052	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kawasaki Yuma、Miyahara Tetsuhiro、Kuboyama Tetsuji、Suzuki Yusuke、Uchida Tomoyuki	４．巻 -
２．論文標題 Evolutionary Acquisition of Multiple TTSP Graph Patterns with Wildcards by Clustering TTSP Graphs	５．発行年 2021年
３．雑誌名 2021 IEEE 12th International Workshop on Computational Intelligence and Applications (IWCIA)	６．最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/IWCIA52852.2021.9626029	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計11件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 石原大吾，　内田智之，糸川裕子，鈴木祐介，宮原哲浩
２．発表標題 多重圧縮順序木集合に対する頻出項木パターン並列枚挙アルゴリズム
３．学会等名 2021年度（第72回）電気・情報関連学会中国支部連合大会
４．発表年 2021年

１．発表者名 田中知希、鈴木祐介，内田智之，宮原哲浩
２．発表標題 頻出1変数項木パターンの枚挙アルゴリズム
３．学会等名 第120回人工知能基本問題研究会(SIG-FPAI)
４．発表年 2022年

１．発表者名
松本哲志、鈴木祐介、内田智之、正代隆義、宮原哲浩

２．発表標題
1つの正例と線形回数の所属性質問による変数次数が定数である線形順序項木パターンの言語族に対する質問学習アルゴリズム

３．学会等名
2022年電子情報通信学会総合大会

４．発表年
2022年

１．発表者名
川崎有馬、宮原哲浩、久保山哲二、鈴木祐介、内田智之

２．発表標題
TTSPグラフのクラスタリングによる複合的なワイルドカード付きTTSPグラフパターンの進化的獲得

３．学会等名
2021年度 人工知能学会全国大会（第35回）

４．発表年
2021年

１．発表者名
横山駿介、宮原哲浩、鈴木祐介、内田智之、久保山哲二

２．発表標題
ラベル情報を利用した進化的学習によるワイルドカードを持つ頂点ラベル付きタグ木パターンの獲得

３．学会等名
2021年度 人工知能学会全国大会（第35回）

４．発表年
2021年

１．発表者名
門田大輝、鈴木祐介、内田智之、宮原哲浩

２．発表標題
物語文からの人物間の関係と動作を表す人物相関グラフ抽出手法の開発

３．学会等名
2021年度（第72回）電気・情報関連学会中国支部連合大会

４．発表年
2021年

１．発表者名
聶翰、内田智之、宮原哲浩、鈴木祐介

２．発表標題
学習済み深層学習モデルに対する木パターンによる説明可能表現とそ の抽出手法の提案

３．学会等名
2020 IEEE SMC Hiroshima Chapter 若手研究会

４．発表年
2020年

１．発表者名
徳原史也、沖永志帆、宮原哲浩、鈴木祐介、久保山哲二、内田智之

２．発表標題
ラベル情報を利用した進化的学習による複合的なワイルドカード付きブロック保存型外平面的グラフパターンの獲得

３．学会等名
2020年度 人工知能学会全国大会（第34回）

４．発表年
2020年

１．発表者名
Fumiya Tokuhara, Shiho Okinaga, Tetsuhiro Miyahara, Yusuke Suzuki, Tetsuji Kuboyama, Tomoyuki Uchida

２．発表標題
Using Label Information in a Genetic Programming Based Method for Acquiring Block Preserving Outerplanar Graph Patterns with Wildcards

３．学会等名
2019 IEEE 11th International Workshop on Computational Intelligence and Applications (IEEE IWCIA2019)（国際学会）

４．発表年
2019年

１．発表者名
石原大吾，内田智之，畝田知典，糸川裕子，鈴木祐介，宮原哲浩

２．発表標題
多重圧縮順序木に対する頻出パス枚挙アルゴリズムの並列化

３．学会等名
令和元年度（第70回）電気・情報関連学会中国支部連合大会

４．発表年
2019年

１．発表者名
舛井 里帆、池森 千尋、鈴木 祐介、内田 智之、宮原 哲浩

２．発表標題
1変数項木パターンに対する多項式時間マッチングアルゴリズム

３．学会等名
火の国情報シンポジウム2020

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	正代　隆義 (Shoudai Takayoshi) (50226304)	福岡工業大学・情報工学部・教授 (37112)	
研究分担者	宮原　哲浩 (Miyahara Tetsuhiro) (90209932)	広島市立大学・情報科学研究科・准教授 (25403)	
研究分担者	鈴木　祐介 (Suzuki Yusuke) (10398464)	広島市立大学・情報科学研究科・助教 (25403)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関