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　元 年　 ６ 月 ２６ 日現在

機関番号：２５４０３
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ００３１３

研究課題名（和文）グラフ文法圧縮データからの省メモリ高速グラフマイニング手法の開発

研究課題名（英文）Development of memory-saving high-speed graph mining method for graph grammar-compressed 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）　　3,600,000円

研究成果の概要（和文）：研究課題の目的は、グラフ構造データを可逆圧縮するグラフ文法圧縮手法を開発し、グラフ文法圧縮されたグラフ構造データに対する省メモリ高速グラフマイニングアルゴリズムを開発することである。目的を達成するために、まず頻出部分木が多重圧縮された順序木を多重圧縮木と定義し、簡潔データ構造表現を用いて多重圧縮木のコードを与えた。また、コード化された多重圧縮木上で頻出するパスや部分木を陽に展開することなく高速に枚挙する省メモリ高速アルゴリズムを提案した。さらに、計算論的学習理論に基づき、多重圧縮木データのグラフ文法モデルであるグラフ文法システムにより定義される言語のクラスが高速に同定できることを示した。

研究成果の学術的意義や社会的意義
ICT関連技術の発達に伴い、1 兆を超えるといわれるWebページを頂点に、ハイパーリンクを辺とするWebグラフや、爆発的な人気を誇っているFacebookやLINEの利用者を頂点、友人関係を辺としたソーシャル・ネットワークなど、グラフ構造を有する大規模なデータ（ビッググラフデータ)は日々拡大している。 このビッググラフデータの解析には膨大な時間と資産が必要となる。木構造を有するビックグラフデータを可逆圧縮し、陽に展開することなく構造的特徴を抽出するグラフマイニング手法を提案した本研究成果は、ビッググラフデータの解析時間の短縮および使用メモリ量の削減に寄与するものである。

研究成果の概要（英文）：The purpose of this research is to develop graph grammar compression method for lossless compression of graph structure data and to develop a memory-saving high-speed graph mining algorithm for graph structure data that have been graph grammar-compressed. In order to achieve the purpose, we defined a multi-compressed ordered tree obtained by compressing frequent subtrees as a multi-compressed tree. Then, we also proposed a memory-saving high-speed algorithm that enumerates all frequent paths and subtrees without explicitly expanding given coded multi-compressed trees that represent ordered tree structured data. Furthermore, based on computational learning theory, it was shown that classes of graph languages defined by a special type of Formal Graph System, which is a graph grammar model of multiple compressed tree data, can be identified from one positive example by using a polynomial number of membership queries.

研究分野：グラフアルゴリズム

キーワード：グラフアルゴリズム　グラフ文法圧縮　グラフマイニング　計算論的学習理論　機械学習

１．研究開始当初の背景

ICT 関連技術の発達に伴い、Web グラフ、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大規模なデータ (ビッググラフデータ) が日々電子データとして蓄積されている。これらビッググラフデータを解析した結果を様々な分野で利活用しようという研究が行われ始めている。ビッググラフデータの例として、Google の推定で 1 兆を超えるといわれる Web ページを頂点に、ハイパーリンクを辺に持つ Web グラフや、爆発的な人気を誇っている Facebook や LINE の利用者を頂点、友人関係を辺としたソーシャル・ネットワークなどがある。これらビッググラフデータの解析には膨大な時間と資産が必要となる。文字列でモデル化できるビッグデータに対しては、解析に要する時間の短縮や使用メモリ量の削減のための戦略として、文脈自由文法を用いてビッグデータを可逆圧縮し、圧縮したデータを陽に展開することなく解析しようという研究が注目を集めていた。これら文字列に対する手法を拡張し、グラフ文法を用いて圧縮されたビッググラフデータに対するグラフマイニング手法についての研究はまだ少なかった。

２．研究の目的

本研究課題の目的は、グラフ構造を有するビッググラフデータからより広くより深い知識を抽出する、省メモリ高速グラフマイニング手法の開発である。近年、ICT 技術の発展に伴い、Web グラフ、 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ビッググラフデータを効率的に解析する研究が盛んに行われている。そこで、本研究課題では、文字列上のビッグデータに対する既存の文法圧縮法を、グラフ文法を用いてビッググラフデータを可逆圧縮するグラフ文法圧縮法に拡張し（図１参照)、グラフ文法圧縮がもたらすマイニング手法の高効率化について研究し、高効率グラフマイニングアルゴリズムを開発することを目的とする。

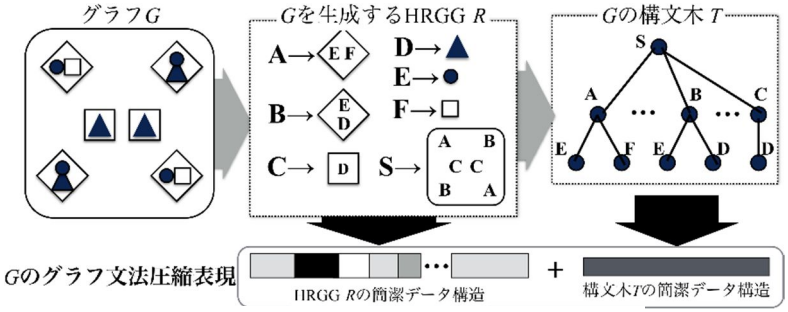


図 1: グラフ文法圧縮概念図

３．研究の方法

研究目的を達成するために、以下の方法で研究を遂行した。

(1) グラフ文法圧縮の理論展開：文字列に対する文法圧縮手法について調査し、文脈自由グラフ文法の１つである超辺書換グラフ文法(Hyperedge Replacement Graph Grammar, HRGG)(参考論文[1])および変数を持つ超グラフを項として持つ１階述語論理プログラミングシステムである形式グラフ体系（Formal Graph System, FGS）(参考論文[2])をもとにしたグラフ文法圧縮について検討する。

(2) 計算論的学習理論に基づくグラフ文法圧縮データからの知識発見手法の開発：計算論的学習理論の観点から、FGS を用いて表現されたグラフ文法圧縮データからの知識発見について考察する。計算量的困難性からこれまでに知見を有しているグラフの族である順序木について検討を加える。HRGG と同様に正則 FGS により生成されるグラフ（例えば、Two-Terminal Series Parallel(TTSP)グラフやコグラフ(cograph)など）の構造的特徴は、そのグラフの生成過程を表す構文木上の特徴と対応付けることができる。このため、順序木をグラフ文法圧縮して得られる圧縮順序木（圧縮木）を対象にした研究成果は、TTSP グラフやコグラフなどのグラフ族に対して拡張することができる。

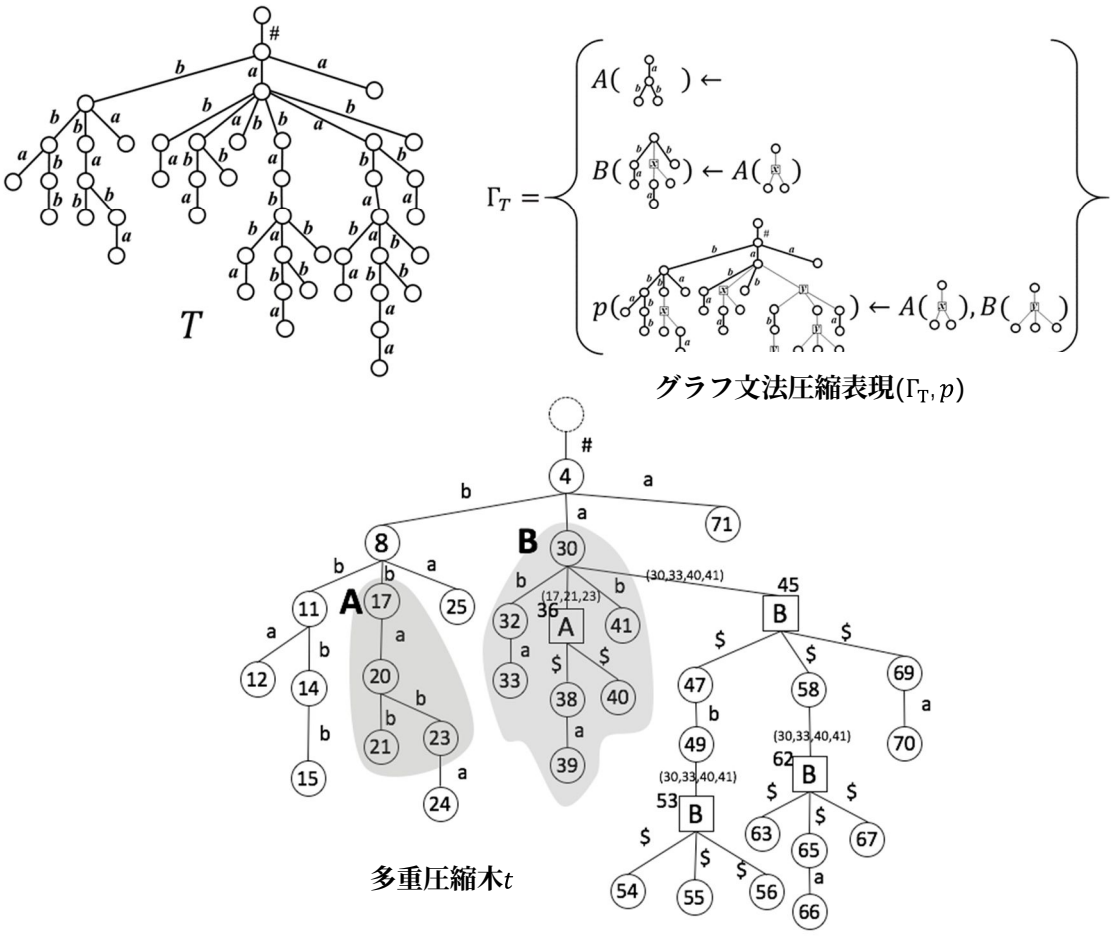
(3) 圧縮木に対する省メモリ高速グラフマイニング手法の開発：グラフ文法圧縮された圧縮木のコンパクトなデータ表現として簡潔データ表現を用いる。簡潔データ表現を用いることにより、圧縮木のサイズをおさえることができる。さらに select，rank 操作が定数で行えるため、高速なグラフマイニング手法を構築することが期待できる。本研究課題においては、圧縮木を陽に解凍することなく頻出するパスと部分木を枚挙する省メモリ高速グラフマイニングアルゴリズムを提案する。

(4) クラウド・コンピューティングへの応用可能性についての検討：圧縮木に対する省メモリグラフマイニング手法のさらなる高速化を目指し、手法の並列実行について検討を行い、クラウド・コンピューティングへの応用可能性について検討を加える。

４．研究成果

以下に本研究で得られた成果を以下に述べる。

(1) グラフ文法圧縮の理論展開というテーマにおいては、順序木を生成する正則 FGS のグラフ圧縮表現、そのコンパクト表現である多重圧縮木とそのコードを、次のように定義した。Λを辺ラベルの集合とする。内部ノードが順序付けられた子を持ち、辺がΛ内の要素でラベル付けられている根付き木を辺ラベル付き順序木という。参考論文[2]に基づき、与えられた辺ラベル付き順序木Tのみを生成する FGS Γ_{T}と述語記号pの組(Γ_{T}, p)をTのグラフ文法圧縮表現という。ヘッド部の述語記号がpであるグラフ書き換え規則$r = {}'p(t_0) \leftarrow A_1(t_1), A_2(t_2), \ldots, A_k(t_k)' \in \Gamma_{\mathrm{T}}$において、項木$t_0$に同じ変数ラベルの出現回数が多いほど高い圧縮効果が期待できる。t_0の各変数ラベルxに対して、rのボディ部で変数ラベルxを持つアトムの述語記号をA_xとする。Γ_{T}のグラフ書き換え規則の中でヘッド部に述語記号を持つアトムを$A_x(t_x)$とする。このとき、深さ優先走査で最初に出現する変数ラベルxを持つ変数をt_xで、それ以降に現れる変数ラベルxを持つ変数は、A_xをラベルとして持つ辺とその子供として\$をラベルとして持つ辺からなる順序木で置き換えてt_0から新しい順序木を得る操作をすべての変数ラベルに再帰的に行って得られる順序木を多重圧縮木という。簡潔データ表現の１つである DFUDS(Depth-First Unary Degree Sequence)(参考論文[3])を用いてこの多重圧縮木を表現することで、順序木Tのグラフ文法圧縮表現(Γ_{T}, p)のコードを定義した。図 2 に順序木Tのグラフ文法圧縮表現(Γ_{T}, p)に対する多重圧縮木tとそのコードの例を示す。



多重圧縮木tのコード

((((#(((b((ba(bb(bb((ab(baa((((aba((A(ab(((B($(b(((B$$$($(((B$(a($aa

図２：順序木Tのグラフ文法圧縮表現(Γ_{T}, p)、多重圧縮木tとそのコード

(2) 計算論的学習理論に基づくグラフ文法圧縮データからの知識発見手法の開発というテーマにおいては、順序木のグラフ文法圧縮表現の集合を定義する正則 FGS の部分クラスである primitive formal ordered tree system (pFOTS)を定義し、計算論的学習理論の観点からグラフ文法圧縮データを管理するデータベースから学習する手法について研究を行った。具体的には、１つの正例が与えられたとき、データベースへの多項式回の問い合わせでターゲットとなる pFOTS を同定する学習アルゴリズムを提案した（雑誌論文①）。さらに、2-限定的木置換文

法により定義される言語のクラスについても研究を行い学会発表を行った（学会発表③）.これらの結果は、グラフ文法圧縮データからの知識発見に関する理論的な結果である。

(3) 圧縮木に対する省メモリ高速グラフマイニング手法の開発というテーマにおいては、次に定義する「多重圧縮木に対する頻出パス枚挙問題」と「多重圧縮木集合に対する頻出部分木枚挙問題」についてそれぞれ効率的なグラフマイニングアルゴリズムを提案した。自然数$k(k \geq 1)$に対し、パスpが k頻出であるとは、多重圧縮木を解凍して得られる順序木にパスpがk回以上出現するときをいう。順序木Tの多重圧縮木tと自然数 $k(k \geq 1)$が与えられたとき、tを解凍することなくTに出現するk頻出パスをすべて枚挙する問題を多重圧縮木に対する頻出パス枚挙問題という。さらに、順序木集合$D = \{T_1, T_2, \ldots, T_m\}$の各要素に対する多重圧縮木の集合$\{t_1, t_2, \ldots, t_m\}$と自然数$k$ $(1 \leq k \leq m)$が与えられたとき、$t_1, t_2, \ldots, t_m$を解凍することなくD中のk個以上の順序木に出現する部分木をすべて枚挙する問題を多重圧縮木集合に対する頻出部分木枚挙問題という。多重圧縮木に対する頻出パス枚挙問題を解くアルゴリズムを実装し、解凍した順序木上で頻出パスをすべて枚挙するアルゴリズムとの比較実験を通して提案したアルゴリズムの効率性を示した(雑誌論文②,③,学会発表①,④)。同様に多重圧縮木集合に対する頻出部分木枚挙問題をそれぞれ解くアルゴリズムを提案し、実装した上での評価実験を通して提案アルゴリズムの効率性を示した(雑誌論文②,学会発表②)。

(4) クラウド・コンピューティングへの応用可能性についての検討というテーマにおいては、3で与えたグラフマイニングアルゴリズムを頻出パスと頻出部分木を並列に枚挙するアルゴリズムへと拡張する研究を行った。現在、提案した並列枚挙グラフマイニングアルゴリズムを実装中であり、実装終了後逐次枚挙版との比較実験を行い、並列枚挙グラフマイニングアルゴリズムの高速性を実証する予定である。本研究テーマの根幹をなす並列枚挙グラフマイニングアルゴリズムとその評価結果を速やかに公表する予定にしている。

参考論文

[1] F. Drewes, et al., “Hyperedge Replacement Graph Grammars”, Chapter 2 of Handbook of Graph Grammars and Computing by Graph Transformation (Vol.1), Ed. G. Rozenberg, pp.95-162, 1997.

[2] T. Uchida, et al., “Parallel Algorithms for Refutation Tree Problem on Formal Graph Systems”, IEICE Transactions on Information and Systems, E78-D, pp.99-112, 1995.

[3] 定兼邦彦，「超簡潔データ構造」，電子情報通信学会誌 Vol.92, No.2, pp.97-104, 2009.

５．主な発表論文等

（研究代表者、研究分担者及び研究協力者には下線）

〔雑誌論文〕（計3件）

① Tomoyuki Uchida, Satoshi Matsumoto, Takayoshi Shoudai, Yusuke Suzuki, and Tetsuhiro Miyahara, “Exact Learning of Primitive Formal Systems Defining Labeled Ordered Tree Languages via Queries”, IEICE Transactions on Information and Systems, 査読有, E102.D, 2019, 470-482, DOI:10.1587/transinf.2018FCP0011

② Tomoya Horibe, Yuko Itokawa, Tomoyuki Uchida, Yusuke Suzuki, and Tetsuhiro Miyahara, “Enumeration Algorithms for All Characteristic Paths and Subtrees from Structurally Compressed Tree-Structured Data”, IAENG International Journal of Computer Science, 査読有, 45(1), 2018, 206-218

③ Tomoya Horibe, Yuko Itokawa, Tomoyuki Uchida, Yusuke Suzuki, and Tetsuhiro Miyahara, “Algorithm for Enumerating all Frequent Paths from Structurally Compressed Tree- Structured Data”, Proceedings of The International MultiConference of Engineers and Computer Scientists 2017 (IMECS 2017), 査読有, 2017, 69-74

〔学会発表〕（計 4 件）

① 畝田知典，堀部智也，糸川裕子，内田智之，宮原哲浩，鈴木祐介，多重圧縮された順序木構造データに対する頻出パス枚挙アルゴリズム，平成 30 年度(第 69 回)電気・情報関連学会中国支部連合大会，2018 年 10 月

② 堀部智也，糸川裕子，内田智之，鈴木祐介，宮原哲浩，構造圧縮された木構造データからの頻出部分木枚挙アルゴリズム，2017 人工知能学会全国大会，2017 年 6 月

③ 李起春，鈴木祐介，内田智之，宮原哲浩，2-限定的木置換文法に対する非終端記号に関する所属性質問を用いた質問学習，平成 28 年度(第 67 回)電気・情報関連学会中国支部連合大，2016 年 10 月

④ 堀部智也，糸川裕子，内田智之，宮原哲浩，鈴木祐介，LZ 法に基づいて構造圧縮された順序木構造データに対する頻出パス枚挙アルゴリズム，2015 年度 火の国情報シンポジウム 2016，2016 年 3 月

６．研究組織

(1)研究分担者

研究分担者氏名：正代 隆義

ローマ字氏名：Takayoshi Shoudai

所属研究機関名：九州国際大学

部局名：現代ビジネス学部

職名：教授

研究者番号（8 桁）：50226304

研究分担者氏名：宮原 哲浩

ローマ字氏名：Tetsuhiro Miyahara

所属研究機関名：広島市立大学

部局名：情報科学研究科

職名：准教授

研究者番号（8 桁）：90209932

(2)研究協力者

研究協力者氏名： 鈴木 祐介

ローマ字氏名：Yusuke Suzuki

研究協力者氏名： 糸川 裕子

ローマ字氏名：Yuko Itokawa