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　３ 年　 ６ 月 １０ 日現在

機関番号：３０１０７
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１１１４９

研究課題名（和文）高階圧縮実用化に向けた基盤技術開発

研究課題名（英文）Development of fundamental technology for practical use of high-order compression

研究代表者
喜田　拓也（Kida, Takuya）

北海学園大学・工学部・教授

研究者番号：７０３４３３１６

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）：本研究では，高階圧縮と呼ばれるデータ圧縮方式について，効率よい処理アルゴリズムの開発を行った．既存の高階圧縮の圧縮処理アルゴリズムは処理速度に一番の難点がある．高速に圧縮処理を行うためには，入力データから共通する部分構造をすばやく見つけ出し，ラムダ式として抽出しなければならない．今回，入力データの繰り返し部分を表現するラムダ式を高速に抽出するアルゴリズムの開発に成功した．また，高階圧縮のサブクラスである文法圧縮についても研究を行い，理論的に優れた文法を生成することのできる手法MR-RePairアルゴリズムの開発に成功した．

研究成果の学術的意義や社会的意義
本研究の特色は，単に圧縮率もしくは処理速度に優れたデータ圧縮法を開発するのではなく，同時に，圧縮されたデータが活用しやすいものとなるようなデータ圧縮法を目指している点にある．圧縮率，処理速度，データ活用の簡便さはトレードオフの関係にあり両立することが難しい．高階圧縮やそのサブクラスである文法圧縮は，圧縮率とデータ活用の簡便さにおいて優れたものであった．今回の研究で，処理速度についても大きく向上することができた．この研究成果は，現在のインターネット社会の中で日々増大する膨大なデータの保存コストを下げると同時に，データ解析のコストも下げることができる．

研究成果の概要（英文）：In this study, we develop efficient processing algorithms for a data compression method called higher-order compression. Existing compression algorithms for higher-order compression have the greatest difficulty in processing speed. To perform compression processing at high speed, it is necessary to quickly find common substructures in the input data and extract them as lambda expressions. Finally, we have succeeded in developing an algorithm to quickly extract lambda expressions that represent repetitive parts of the input data. We also studied grammar compression, a subclass of higher-order compression, and developed an efficient algorithm named MR-RePair algorithm, which is a method that can generate theoretically superior grammars.

研究分野：情報学基礎論関連

キーワード：高階圧縮　ラムダ計算　文法圧縮　大規模データ　透過的データ圧縮法

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

データ圧縮とは，データ中に含まれる冗長な部分を簡潔に表現することで，そのデータを保持するために必要な記憶容量を削減する技術である．大量のデータは，保存コストあるいはその通信コストを低減するために，データ圧縮を施してから保存されることが多い．

1990 年代頃より，圧縮されたデータ上で直接に様々な処理を行う研究が盛んになった．例えば，圧縮データを展開することなくキーワード検索を行ったり，データマイニング処理やデータ集計処理を行ったりするためのアルゴリズムが考えられるようになった．

2000 年以降になると，検索やデータ解析処理が容易になることを目的としたデータ圧縮の手法に主眼を置く研究が起こった．しかしながら，そうした圧縮法は良く知られたものと比べて圧縮率の点で大幅に劣っていたため実用的に用いられる場面が少なかった．データ処理を簡便に行えるようにすることと，データを効率良く圧縮することの両立は，従来，非常に困難な課題であった．

こうした中，2000 年にミネソタ大の J. C. Kieffer と E. Yang らにより提案された文法圧縮が次第に注目を浴びるようになった．文法圧縮とは，入力データを形式文法の形へと変換し，抽出された文法を符号化することでデータ圧縮を行う手法である．ある種の文法圧縮は，圧縮データへの直接アクセスをサポートする索引構造的な側面を持つことが判明している．以降，文法圧縮の研究は花開き，実用的な文法圧縮方式がいくつか提案されるに至った．

文法圧縮の研究が成熟しつつある中，既存の形式文法より表現力の高い中間表現の模索が始まった．東京大学の小林直樹と東北大学の篠原歩らのグループは共同し，2012 年に，高階プログラム（ラムダ式）を中間表現として用いた高階圧縮を提案した．中間表現であるラムダ式の性質から，データの変換やクエリ操作が比較的容易であるという特長がある．彼らのグループと議論を重ねていく中で，高階圧縮の課題点である圧縮処理の高速化について取り組むこととなった．

２．研究の目的

本研究の目的は，高階圧縮と呼ばれるデータ圧縮方式の効率よい処理アルゴリズムを開発することである．ここで「効率よい」とは，次の三つの観点において優れていることである．第一には，データをどれだけコンパクトに表現できるかという圧縮率の観点である．第二には，処理時間とメモリ消費量をどれだけ抑えられるかという計算量の観点である．第三には，圧縮後のデータ自体が，後の情報検索やデータ解析の際にどのくらい利用しやすいかというデータ活用の観点である．高階圧縮は，ラムダ計算に基づく高階プログラムを中間表現とした可逆なデータ圧縮法である．既存のデータ圧縮法に比べ，データに内在するより複雑な構造を捉えられるため，次世代のデータ圧縮方式として期待されている．しかし，先に述べた三つの観点を高いレベルで満たす処理アルゴリズムの開発が課題として残っている．特に，本研究の特徴は，第三の観点である「データ活用」に着目している点にある．

３．研究の方法

既存の高階圧縮の圧縮処理アルゴリズムは処理速度に一番の難点がある．高速に圧縮処理を行うためには，入力データから共通する部分構造をすばやく見つけ出し，ラムダ式として抽出しなければならない．ラムダ式の構造は，構文木と呼ばれる木構造で表される．また，ラムダ計算は，この構文木上での木構造の変換操作としてとらえることができる．小林・篠原らが最初に示したラムダ式の抽出方法は，まず入力データを直線状の木構造に変換し，その後，頻出するすべての可能な部分木を探索して，ラムダ計算の逆計算にあたる操作によって木構造をコンパクトにまとめていくという手順を取る．この手法の問題点は，探索すべき部分木の種類が組み合わせ的に増大してしまうことである．

そこで本研究では，以下の点に注力して研究を推進した．

①　高速で省メモリなラムダ式抽出アルゴリズムの開発
②　データ活用に適した符号化方式の開発
③　圧縮後のデータを活用する基盤技術の開発

しかしながら，研究を進める中で，文法圧縮について目覚ましい進展があったため，以下の点に注力することとなった．

④　文法圧縮の新手法に関する研究

４．研究成果

研究①の高速で省メモリなラムダ式抽出アルゴリズムの開発について，データが連続している部分のラムダ式を高速にかつコンパクトに抽出する効率よいアルゴリズムの開発を行った．我々の手法は，連続するパターンに対して非負整数の超冪による分解表現を定義し，連続パターンのラン長をその分解表現に対応するラムダ式に置き換えることでデータを圧縮する．ラムダ式によるラン長の単純な表現としては，小林らの高階圧縮の論文の中でバイナリ表現が示されている．新手法は，ほとんどの場合においてこのバイナリ表現よりも小さなラムダ式を生成する．また，このラムダ式置き換えを既存手法である矢口らの圧縮処理に組み込むことで，より圧縮処理を短縮することが期待できる．本研究成果は，研究期間内において理論的解析の成果を改めてとりまとめ直し，国際雑誌 Algorithms（MDPI）の「Data Compression Algorithms and their Applications」特集号への掲載に至った．

データ圧縮は，データのモデル化と符号化の二つの処理からなる．研究②と④に関して，これまで，文法圧縮の一つである Re-Pair アルゴリズムをもとにより高性能なデータ圧縮法の研究開発を行ってきた．また，データ利活用に適した符号化方式として VF 符号化に着目し，これらを組み合わせた Repair-VF 符号を得ている．その後も Repair-VF の改良を続けており，開発した一連の圧縮法は，展開速度に優れるばかりでなく，高速検索にも適した性質を持つ．こうした研究成果の中で，文字列中に含まれる最長反復部分文字列（maximal repeat）が Re-Pair の圧縮法において重要な役割を担うことに気づいた．

我々は，本研究期間において，最長反復部分文字列に基づく文法変換アルゴリズム MR-RePair の開発に成功した．MR-RePair は，最頻の文字ペアではなく，最頻の最長反復部分文字列を優先的に置き換えることで文法変換を行う．このことは，Re-Pair アルゴリズムで生じていた無駄な規則の生成を抑制し，最終的な文法サイズを低減する．実際，遺伝子データセットや文書履歴データなど長い文字列の繰り返しが多いテキストデータに対して，Re-Pair よりもかなり小さな文法サイズを生成することを確認していた．このことを改めて理論的に解析し，MR-RePair は，Re-Pair よりも真に小さい文法を生成することを証明した．これらの成果を，データ圧縮分野のトップ国際会議（CORE Rank A*）である Data Compression Conference (DCC2019)にて発表した．さらに，理論的解析を深めて雑誌論文としてとりまとめ直し，長く推敲を重ねた結果，国際雑誌 Algorithms (MDPI)の「Lossless Data Compression」特集号に掲載されるに至った。

研究③に関して，データストリーム上のデータマイニング問題の一つである文字列計数アルゴリズムの研究開発を行った．データストリームとはオンラインに再現なく流れてくるデータの列のことである．データストリームは，そのデータ全てをメモリ上に保持することが難しく，過去に流れてきたデータの一部または全てにアクセスすることはできない．したがって，データストリームに対するアルゴリズムは，データをオンラインに処理する必要があり，実行時間が高速かつ省メモリであることが求められる．データストリームに対するマイニングの重要な問題の一つに，与えられた閾値より多く出現するアイテムを見つける頻出アイテム問題がある．この問題に対して，2002 年に Demaine ら（ESA2002）は k-reduced Bag に基づいた Frequent と呼ばれるアルゴリズムを提案している．ただし，Frequent の出力には偽陽性の解が含まれており，偽陽性の解が持つ頻度の最小値に保証がない．そこで，Frequent を改善し，偽陽性の解の最小頻度に保証を持つ KRB アルゴリズムを開発した．実験の結果，KRB は Frequent と比べて，適合率が高くなっていることを確認した．さらに，KRB と他の既存アルゴリズムとを比較した結果，時間・空間計算量で Manku and Motowani の Lossy Counting（VLDB2002）より優れており，エラー範囲で Metwally らの Space Saving（ICDT2005）より優れていることが分かった．さらに，実実験においても，それらの手法と比べて同程度以上の速度，精度であることを確認することができた．本研究成果については，情報処理学会第 173 回アルゴリズム研究会にて発表を行っている．

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　2件）

１．著者名 Isamu Furuya and Takuya Kida	４．巻 vol. 12, no. 8, 159
２．論文標題 Compaction of Church Numerals	５．発行年 2019年
３．雑誌名 Algorithms (MDPI Journal)	６．最初と最後の頁 1-16
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/a12080159	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Furuya Isamu、Takagi Takuya、Nakashima Yuto、Inenaga Shunsuke、Bannai Hideo、Kida Takuya	４．巻 vol. 13, no. 4, 103
２．論文標題 Practical Grammar Compression Based on Maximal Repeats	５．発行年 2020年
３．雑誌名 Algorithms (MDPI Journal)	６．最初と最後の頁 1-18
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/a13040103	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計7件（うち招待講演　1件／うち国際学会　1件）

１．発表者名 鳥谷部 直弥, 古谷 勇, 喜田 拓也
２．発表標題 データストリームのための頻出部分文字列発見アルゴリズム
３．学会等名 第173回アルゴリズム研究会
４．発表年 2019年

１．発表者名 Isamu Furuya, Takuya Takagi, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Takuya Kida
２．発表標題 MR-RePair: Grammar Compression Based on Maximal Repeats
３．学会等名 Data Compression Conference (DCC2019)（国際学会）
４．発表年 2019年

１．発表者名
古谷 勇，髙木 拓也，中島 祐人，稲永 俊介，坂内 英夫，喜田 拓也

２．発表標題
極大反復部分文字列に基づく文法圧縮

３．学会等名
第171回アルゴリズム研究会

４．発表年
2019年

１．発表者名
鳥谷部 直弥，喜田 拓也

２．発表標題
データストリームに対する効率良い頻出アイテム発見アルゴリズム

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム(DEIM2019)

４．発表年
2019年

１．発表者名
鳥谷部 直弥，谷 陽太，喜田 拓也

２．発表標題
データストリームに対する頻出値問題を解くアルゴリズムの実証実験

３．学会等名
第17回情報科学フォーラム（FIT2018）

４．発表年
2018年

１．発表者名
喜田拓也

２．発表標題
データ圧縮とパターン照合

３．学会等名
スマートインフォメディアシステム研究会（SIS，IPSJ-AVM，ITE-3DIT合同研究会）（招待講演）

４．発表年
2020年

１．発表者名 小畠教寛，喜田拓也
２．発表標題 最大クリーク発見問題の省メモリアルゴリズム
３．学会等名 人工知能基本問題研究会
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
連携研究者	篠原　歩 (Shinohara Ayumi) (00226151)	東北大学・情報科学研究科・教授 (11301)	博士（理学） アルゴリズム設計
連携研究者	坂本　比呂志 (Sakamoto Hiroshi) (50315123)	九州工業大学・大学院情報工学研究院・教授 (17104)	博士（理学） アルゴリズム設計

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