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

平成　３０　年　　６　月　　７　日現在

機関番号：　１０１０１
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００００２

研究課題名（和文）オンライン型文法圧縮とVF符号化アルゴリズムによるストリーム型データ圧縮

研究課題名（英文）Stream Data Compression by Online Grammar Compression and VF Coding Algorithm

研究代表者
喜田　拓也（Kida, Takuya）

北海道大学・情報科学研究科・准教授

研究者番号：７０３４３３１６

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）：申請者は，VF符号と文法圧縮を組み合わせたRe-Pair-VFアルゴリズムの改善に取り組んだ．Re-Pair-VFは，既存のVF符号よりも優れた圧縮性能を持つが，オフライン型の文法変換であるRe-Pairを基にしているため，大規模なデータに適用することが困難である．そこで，申請者らは，Re-Pairを改変し，オンライン型の文法置換アルゴリズムを開発した．また，そのアルゴリズムを利用して，ブロック長を適応的に伸長させながら圧縮するALT（Adaptive LT-Repair）法の提案を行った．

研究成果の概要（英文）：In this study, the applicant addressed the problem of improving the Re-Pair-VF algorithm which is a combination of VF code and grammar compression. Re-Pair-VF has better compression performance than the existing VF code, but it is difficult to apply it to large-scale data because it is based on Re-Pair which is offline type grammar transformation. Therefore, the applicant modified Re-Pair and developed online grammar transformation algorithm. Moreover, the applicant proposed an ALT (Adaptive LT-Repair) method that compresses while adaptively expanding the block length using the algorithm.

研究分野：　情報学基礎

キーワード：　データ圧縮　VF符号化　圧縮照合　文法圧縮

１．研究開始当初の背景

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．データ圧縮とは，データ中に含まれる冗長性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．今日，世界中の人間が生み出すデータの量も膨大であるが，それよりもはるかに多くのデータがネットワーク機器やセンサーなどの機械によって自動的に生成され，インターネット上を流通している．こうしたデータは，量は膨大でも冗長な部分が多く，データ圧縮して保存するのが適当である．その一方で，データの大半は時間的あるいは位置的に順序のあるストリーム型データであり，絶え間なく高速にシステムへと到来するデータをオンラインで高速に処理する必要がある．

ZIP や zlib, gzip などの有名なデータ圧縮ソフトウェアは，1970 年代後半にイスラエル工科大学の Jacob Ziv と Abraham Lempel によって提案された LZ 圧縮法を土台としている．より直接的には，LZ 圧縮法の変種の一つで，ブランダイス大の James Storer と Thomas Szymanski らによって 1982 年に提案された LZSS 法と，1952 年に David Huffman によって提案されたハフマン符号とを組み合わせた Deflate（デフレート）と呼ばれるデータ圧縮アルゴリズムに基づいている．これらのソフトウェアは，処理速度が速く，また良好な圧縮率を達成できるため，現在，世界中でデファクト・スタンダードとして用いられている．しかしながら，圧縮後のデータを利用する際には，一旦，元の形に復元する（展開する）必要があり，データの解析や検索には余分な手間がかかる．

このような背景から，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（パターン照合）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．申請者も，1996 年頃より，圧縮データに対する直接的なパターン照合処理の高速化の研究開発を行ってきた（「半構造化データに対する文字列処理の高速化に関する研究【平成 14 年度～16 年度；若手研究 B】」，「連続データストリームに対する高度なパターン照合の研究【平成 20 年度～22 年度;若手研究 B】」）．この研究は，圧縮パターン照合（Compressed Pattern Matching）と呼ばれ，LZ 圧縮法などの既存手法によって圧縮されたデータに対して，効率よく検索を行う「パターン照合のアルゴリズム」をどのように構築するかが焦点であった．

2000 年以降になると，パターン照合アルゴリズムではなく，検索やデータ解析処理が容易になることを目的とした「データ圧縮アルゴリズム」に主眼を置いた研究が起こった．申請者が当時所属していた九州大学の竹田正幸らのグループの他に，ヘルシンキ工科大学（当時）の Jorma Tarhio らや，チリ大学の Gonzalo Navarro ら，バル＝イラン大学の Shmuel Klein らなど，いずれも本分野で著名な研究者グループから相次いで新規なデータ圧縮法が提案されている．しかしながら，これらデータ処理を見越した圧縮法は，上述した Deflate をベースとしたソフトウェア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．

この課題に対し，申請者は，特に可変長-固定長符号化 (Variable-to-Fixed length coding, VF 符号化)と呼ばれる符号化に着目して研究をすすめてきた（「高速・高度なパターン照合と高圧縮率とを実現するＶＦ符号化の研究【平成 23 年度～25 年度；若手研究 B】」）．VF 符号化は，入力データを長さが異なる文字列（ブロック）に分解したのち，各ブロックに同じ長さの符号語を割り当てる圧縮方法である．ブロック毎に異なる長さの符号語を割り当てる可変長符号化とは異なり，VF 符号はすべての符号語が固定長の符号であるため，任意の符号語の開始位置と終了位置が明白である．この性質により，圧縮データの部分的な展開や再圧縮処理も比較的容易に行えるほか，圧縮データ上でパターン照合を行う際にも符号語の切り出しが高速に行えるなど，VF 符号化は実応用の観点から多くの優れた利点を持つ．ただしその一方で，符号語が固定長であることは，圧縮率を向上させるうえでは大きな制約となる．

申請者は，全文索引のためのデータ構造である接尾辞木を利用した新しい VF 符号化（STVF 符号化）を提案した．接尾辞木は，データの任意の連続する部分を索引付けできる優れたデータ構造である．ちなみに，同時期に，Klein らのグループもほぼ同じアイデアの符号化方法を提案している．この STVF 符号化により，Deflate の実装の一つである gzip と同程度の圧縮率を達成したが，構築コストの高い接尾辞木を用いるため圧縮速度は極めて遅かった．そこで申請者らは次に，データを高度にモデル化することのできる文法圧縮に着目した．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．デンマーク工科大の Philip Bille やハイファ大学の Gad M. Landau，ランチェスター大の Rajeev Raman，東京大学の定兼邦彦らの近年の研究により，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．

2013 年に申請者らは，Jesper Larsson と Alistair Moffat らによって提案された Re-Pair アルゴリズムと呼ばれる文法圧縮法

と VF 符号化を組み合わせた Repair-VF 符号化を開発した．Repair-VF 符号化は，gzip を凌ぐ圧縮率を達成し，さらに圧縮速度を STVF 符号化の 2 倍以上に向上させた．また，展開速度は gzip と同程度に高速である．

２．研究の目的

本研究の目的は，可変長-固定長符号化（VF 符号化）による効率よいデータ圧縮法を開発することである．ここで「効率よい」とは，次の三つの観点で優れていることを指す．第一に，データ圧縮としての基本性能である圧縮率・処理速度・メモリ消費量について，高いレベルでバランスしていること．第二に，ストリーム型データに対して，逐次的（オンライン）に符号化が行えること．そして第三に，圧縮後のデータ自体が，後の情報検索やデータ解析を補助する索引能力を持つことである．これらを兼ね備えたデータ圧縮法を確立することで，増加し続けるストリーム型データをコンパクトに格納しつつ，効果的に活用できる情報基盤システムを構築する．

申請者らがこれまでに提案した Repair-VF 符号は，メモリ使用量や圧縮速度に関してまだ十分な性能には至っていなかった．その主な原因は，元にした Re-Pair アルゴリズムがオフライン型である点にある．Re-Pair アルゴリズムは，理論的には入力データ長に線形に比例した時間しか要しない効率よいアルゴリズムであるが，実際には元データの 5 倍程度のメモリ領域を消費し，その上をランダムにアクセスしなければならない．したがって，現状ではギガバイト以上の巨大なデータを一度に取り扱うことができない．これに対し，Deflate に基づく圧縮法は，ほぼ逐次処理的なオンライン型のデータ圧縮を行っている．

そこで本研究では，オンライン型の文法圧縮と VF 符号化を組み合わせることで，大規模なデータに対しても高い圧縮率で高速にデータ圧縮を行うことのできる VF 符号化の開発を目的とする．関連研究に，Preferred Infrastructure 社の研究員である丸山史郎氏や九州工業大学の坂本比呂志，科学技術振興機構さきがけ研究員の田部井靖生らが共同で開発した FOLCA がある．FOLCA はオンライン型の文法圧縮に適応的可変長符号化を組み合わせている．圧縮率では Repair-VF に軍配が上がるが，圧縮速度においては FOLCA が勝っている．彼らとの連携を通じて，効率よいデータ圧縮法の実現を目指す．

３．研究の方法

本研究では，大きく分けて三つの項目について研究を行う．

まず，第一に「オンライン文法変換の VF 符号化」について研究を行う．本研究項目では，既存のオンライン型文法圧縮に VF 符号化を組み合わせるアプローチについて研究を行う．オンライン型の文法変換としては，Sequitur や連携研究者らによる FOLCA などが知られているが，これらは，あらかじめ最適な符号語長を設定することが困難であり，単純な方法では VF 符号化することができない．そこで，オンラインで処理しつつも，あらかじめ符号長を決めておける新たな文法変換の枠組みについて開発を行う．

第二に，「ブロック分割と辞書の共有」による半オンラインなデータ圧縮の枠組みについて検討を行う．そのため，ブロック間の辞書の共有化手法に関する研究や，共有辞書の構築方法に関する研究を行う．

第三に，「大規模ストリーム型データへの実応用」について研究を行う．本研究項目では，上述した項目の研究を通して実現するデータ圧縮法を，索引データ構造として利用する手法について研究を行う．

４．研究成果

Re-Pair アルゴリズムは，入力テキスト長に対して線形時間で動作し，優れた圧縮率を達成することのできる文法圧縮アルゴリズムである．ただし，その動作はオフライン的であるため，テキスト全体を一度にメモリ上に読み込む必要がある．この問題に対し，従来，入力テキストを固定長のブロックに分割し，ブロック毎に Re-Pair を適用する手法が取られている．その手法で良好な圧縮率を達成するには，あらかじめ適切なブロック長を与える必要がある．

初年度では，Re-Pair アルゴリズムによって生成される文法にある種の条件を設けることで，Re-Pair アルゴリズムと同等のテキスト置換を与えられた辞書を用いてオンライン的に実行することのできるアルゴリズムを提案した．さらに，そのオンライン置換アルゴリズムを基に，ブロック長を適応的に伸長させながら圧縮する ALT（Adaptive LT-Repair）法の提案を行った．その実証実験において，既存の文法圧縮手法と比較し，格段に省メモリで動作しつつも圧縮率や圧縮速度の犠牲がほとんどないことを示した．

これまでに開発した一連の Re-Pair 型の圧縮法は，展開速度に優れ，また高速検索にも適した性質を持つ．初年度には，Re-Pair アルゴリズムを基にオンライン化したデータ圧縮手法を考案したが，圧縮率を第一とした既存手法と比較すると圧縮率は若干劣る．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．よりよいモデル化はより優れた圧縮率につながる．

近年，文法の代わりに高階プログラムを生成することで圧縮を行う高階圧縮と呼ばれる手法が東京大学の小林直樹教授，東北大学の篠原歩教授らの共同研究グループによって提案された．彼らは高階プログラムとしてラムダ式を用いている．ラムダ式によるラムダ計算はチューリング完全であることが知られている．すなわち，高階圧縮とは，入力テキストから，それを生成するプログラムに

変換して，そのプログラム自体を符号化することで圧縮を行う方式であるといえる．文法圧縮では行えなかった柔軟なモデル化を行う能力を有しており，より高度なデータ圧縮が期待できる．また，この圧縮の際に抽出されるラムダ式は，入力データの構造を説明するコンパクトなプログラム表現であることから，高階圧縮はデータを高度に学習する機械とも見ることができる．

高階圧縮の現状の課題点は，効率よい圧縮アルゴリズム（ラムダ式への変換アルゴリズム）の開発である．そこで第二年度では，連続パターンという限定した場合において，効率よくラムダ式を生成する手法を開発し，そのアルゴリズムの実装と計算量の解析を行った．

さらに最終年度では，文字列中に含まれる最長反復部分文字列（maximal repeat）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た．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

1. Iku Ohama, Hiromi Iida, Takuya Kida, Hiroki Arimura: The Relevance Dependent Infinite Relational Model for Discovering Co-Cluster Structure from Relationships with Structured Noise, 査読有, IEICE Transactions, 99-D(4), pp. 1139-1152, 2016.
2. Satoshi Yoshida and Takuya Kida: An efficient Variable-to-Fixed-length encoding using multiplexed parse trees, 査読有, In Journal of Discrete Algorithms, Vol. 32, pp. 75-86, May 2015.

〔学会発表〕（計 9 件）

1. Isamu Furuya and Takuya Kida: Compaction of Church Numerals for Higher-Order Compression, Data Compression Conference (DCC2018), IEEE Press, p.410, Cliff Lodge, Snowbird, UT, March 2018.
2. Iku Ohama, Issei Sato, Takuya Kida, Hiroki Arimura: On the Model Shrinkage Effect of Gamma Process Edge Partition Models, Proc. the 31st Annual Conference on Neural Information Processing Systems (NIPS2017), pp. 396-404, December 2017.
3. Iku Ohama, Takuya Kida, Hiroki Arimura: Discovering Relevance-Dependent Bicluster Structure from Relational Data, Proc. the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017), Melbourne, pp.2578-2584, August 2017.
4. Isamu Furuya and Takuya Kida: A Compact Expression of Church Numerals and Its Application to Higher-Order Compression, The 20th Korea-Japan Joint Workshop on Algorithms and Computation (WAAC2017), Hanyang University, Seoul, Korea, August 2017.
5. 古谷勇, 喜田拓也: A Compact Expression of Church Numerals and Its Application to Higher-Order Compression, 情報処理北海道シンポジウム 2017, 北海道大学, 北海道, 2017 年 10 月.
6. 古谷勇, 喜田拓也: 高階圧縮における連続パターンのコンパクトな表現法, 第 162 回アルゴリズム研究会, 2017-AL-162, 大分県由布市, 2017 年 3 月 6 日.
7. Takuya Masaki and Takuya Kida: Online Grammar Transformation based on Re-Pair Algorithm, Data Compression Conference 2016 (DCC2016), IEEE Press, pp.349-358, Snowbird, UT, March 29 - April 1, 2016.
8. 正木拓也, 喜田拓也: 制約付き Repair に基づいた適応型ブロック伸長法によるデータ圧縮アルゴリズム, 情報処理学会 第 154 回アルゴリズム研究会, 福岡県 九大西新プラザ, 2015 年 9 月.
9. 正木拓也, 喜田拓也: 制約付き Repair アルゴリズムと等価な半オンライン型置換アルゴリズム, 情報処理学会 第 153 回アルゴリズム研究会, 北海道 定山渓ビューホテル, 2015 年 6 月.

〔図書〕（計 1 件）

1. Akihiro Yamamoto, Takuya Kida, Takeaki Uno, Tetsuji Kuboyama: Discovery Science, Springer, 357 ページ, 2017.

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

ホームページ等

http://www-ikn.ist.hokudai.ac.jp/~kida/publication.html

６．研究組織

(1)研究代表者
喜田 拓也（KIDA TAKUYA）
北海道大学・大学院情報科学研究科・准教授
研究者番号：70343316

(2)研究分担者
なし

(3)連携研究者
坂本　比呂志（SAKAMOTO HIROSHI）
九州工業大学・大学院情報工学研究院・教授
研究者番号：50315123

(4)研究協力者
なし