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　２ 年　 ５ 月 ２８ 日現在

機関番号：１２３０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００００４

研究課題名（和文）変換型無ひずみデータ圧縮法の新潮流

研究課題名（英文）New Trends in Transform-based Lossless Compression Algorithms

研究代表者
横尾　英俊（Yokoo, Hidetoshi）

群馬大学・大学院理工学府・教授

研究者番号：７０１３４１５３

交付決定額（研究期間全体）：（直接経費）　　1,500,000円

研究成果の概要（和文）：本研究では，無ひずみデータ圧縮の新規手法の確立を通じた体系全体の見直しを目標とし，その基盤となるエントロピー符号化法とユニバーサル符号化法から，それぞれ Asymmetric Numeral Systems (ANS) と部分列数え上げデータ圧縮法を研究の中心に据えて，理論的解析と従来法との関係のより一層の明確化を行った。
ANSについては，その圧縮性能が理論的に最適となる十分条件を与えた。それを基礎として，各種バリエーションの解析及び改良の提案を行った。部分列数え上げ法については，文字列の変換法であるBW変換との関係を明らかにし，両者を共に修正した手法を提案した。

研究成果の学術的意義や社会的意義
高度情報社会にあって，データ通信の重要性は言を俟たない。データ通信の効率化と高信頼性化のための様々な技術の基盤をなすエントロピー符号化とユニバーサル符号化とを理論的に研究したものである。特に，Apple等のIT大企業が製品に組み込んでいるとされながら，その性能の理論的根拠に乏しかったエントロピー符号の例について，種々の性質を世界に先駆けて明らかにした。その結果，データの蓄積や通信のためのこれからのシステムにとって有用な知見を得た。

研究成果の概要（英文）：This study aims at the research and development of novel lossless compression algorithms, which can be classified into entropy codes and universal codes. We focus on asymmetric numeral systems (ANS) and compression by substring enumeration (CSE) out of both classes. We give a sufficient condition for ANS to be asymptotically optimal in compression performance. We establish the relation between CSE and the BW-transform, and derive a new CSE implementation from their improved variations.

研究分野：情報源符号化とデータ圧縮

キーワード：情報基礎　情報理論　データ圧縮　ユニバーサル符号　エントロピー符号

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

エントロピー符号化とユニバーサル符号化に大別される無ひずみデータ圧縮法は成熟期にありながらも，双方において新規手法の提案や開発も続いている。それらには，従来法の単なる改良ではなく，成熟期にあるデータ圧縮に新たな視点を提供する手法が含まれている。特に本研究で着目したのは，エントロピー符号化法としては Asymmetric Binary Systems/Asymmetric Numeral Systems (ABS/ANS)，ユニバーサル符号としては部分列数え上げデータ圧縮法である。両者とも 2010 年前後から開発が進められており，既にある程度の実用化や理論研究も進んではいるが，傾向としては大きな偏りがある。ANS は Google，Apple，Facebook といった新興 IT 企業が製品に組み込んでいるとされているが，提案者以外による理論研究はほぼ皆無といってよい。一方，部分列数え上げデータ圧縮法については，情報理論的な研究論文がいくつも発表されているが，他分野や実用化に関しての取組みはほとんど見られない。共に，データ圧縮という目的を共有する以上，相互に何らかの関係性を有すると考えられるが，十分な解明がなされないままに，手法それぞれに独自に展開がはかられている状況である。

ABS/ANS は，ポーランドの Jarek Duda によって 2009 年に発表されたエントロピー符号化法である。エントロピー符号化には，既に定番の手法が存在し，それらに代わる手法の出現は見込まれない中にあっての考案であり，実用上も理論的にも大きな価値の見込まれる手法である。しかし，実用上の応用にとどまり，理論研究としては ABS に関する研究代表者自身の解析が存在する程度で，より実用上の意義の大きい ANS に解析対象を拡張する必要があった。

部分列数え上げデータ圧縮法は，カナダの Danny Dube と Vincent Beaudoin によって 2010 年に提案されたユニバーサルデータ圧縮法である。Danny Dube と研究代表者はそれ以前から共同研究を実施しており，部分列数え上げデータ圧縮法についても，開発当初から共同してその解析等にあたっている。理論的解析についてはかなりの進展がみられてはいたが，実用可能性の見極め，および，理論的意義の更なる追及が必要であり，周辺諸手法との連携も求められている。また可能ならば，ANS 等のエントロピー符号化法の導入も試みる必要があった。

２．研究の目的

本研究では，ANS および部分列数え上げデータ圧縮法のより深い理解を意図した解析を目的とした。ANS については，理論的な解析がほとんどなされないままに既に実用に供されている。裏を返せば，それだけ採用実績のある性能を有しながら，その性能の根拠が明らかになっていないということである。性能の根拠を解析し，定量的に評価することで，更なる高性能化を図ることが本研究の目的である。部分列数え上げデータ圧縮法については，既にいくつもの理論研究が存在している。その多くは情報理論的な視点での解析であり，改良の提案である。一方，本研究の予備研究として，部分列数え上げデータ圧縮法が BW 変換を経由して導出できることが判明している。BW 変換は，情報理論的な解析対象でもあるが，アルゴリズム分野や生物情報学等でより活用が進んでいる。そのように視点を拡張することで，部分列数え上げデータ圧縮法についての理解も一層深まると期待される。本研究では，BW 変換に代えて類似の変換を導入した場合の部分列数え上げデータ圧縮法の振る舞いを明らかにし，得られた示唆に基づいて，高効率実装法や改良法を開発することを目的とした。

３．研究の方法

ANS については情報理論的な解析を主体に，部分列数え上げデータ圧縮法については文字列処理の観点から研究を遂行した。さらに，理論的な予測の検証と実際の性能の検証のために，コンピュータプログラムによる実装によっても研究を遂行した。

ANS は既に多くの IT 関連企業が製品に組み込んでいるとされている。その場合の対象となるデータは現実のデータであるが，本研究では，理論的に扱いやすい無記憶情報源を対象とする解析を行った。コンピュータシミュレーションの場合，最初に２元無記憶情報源に対する確率近似問題を扱い，次にそれを一般の有限情報源に拡張するというアプローチをとった。また，ANS には複数の変種があるため，最初にもっとも基本的と考えられる rANS を解析の対象とし，次に rANS を規定する制約を緩和するという手法をとった。

部分列数え上げデータ圧縮法に対してとった文字列処理の視点は，主として，アルゴリズムの分野で研究されている。そのため，計算量（主として時間計算量，次に領域計算量）が第一義的な興味となることが多い。本研究でも高速な実装法を探るという点では計算量にも注意を払ったが，むしろ，圧縮力の実性能を評価するための高速化であり，アルゴリズム的視点よりは，やはり情報理論的視点に立ち位置を置いた研究方法を採用した。

４．研究成果

ANS にはいくつもの変種が存在する。そのような変種を決定づける中心的な要素は，ANS の記号分布である。記号分布の与え方を決めることによって ANS の具体的な実現例が決まり，同時に，その長所と短所が決定づけられる。本研究では変種の一つである rANS (range variant) に対しては，その圧縮性能を最大化する確率近似法を提案し，tANS (tabled variant) に対しては漸近的最良性を示し，さらに，圧縮性能を最適化するための記号分布決定法を提案した。

ANS の記号分布を与えると，整数上の有限区間に値を取る状態空間と状態遷移が決まる。すると，状態遷移と入力データの確率構造によって状態空間上の定常分布が決まる。定常分布は ANS

の圧縮性能を左右する。本研究では，圧縮性能が理論的に最適となるための定常分布の十分条件を明らかにした。そのような十分条件を与える定常分布と実際の定常分布が異なる場合には，入力データの確率構造を変換することで分布の食い違いを矯正することができる。rANS に対して提案した確率近似法はこの考えを基礎としている。一方，十分条件を与える定常分布は，整数区間上単調減少であることが分かっている。記号分布を与えることで決まる定常分布が単調減少でない場合には，記号分布の非減少の要素を逐次的に更新することで定常分布を徐々に単調減少にすることができる。圧縮性能を最適化するための記号分布決定法は，このような逐次更新を基礎としている。さらに，状態遷移を表によって明示的に与える tANS の一例として，ANS の提案者が厳密な初期化として定式化した手法を理論的に解析し，状態空間の大きさを限りなく大きくすることで，その定常分布が圧縮性能を最適化する十分条件を漸近的に達成することを示した。

以上の各結果を理論的に導出すると同時にコンピュータシミュレーションによる検証を加え，６件の学会発表として発表した。

部分列数え上げデータ圧縮法と BW 変換の関係については，両者を関係づけるために提案した Aカウンタと呼ぶ数え上げ法を解析した。部分列数え上げデータ圧縮法がデータの巡回性を本質的に要求するのに対し，Aカウンタは巡回データを前提としないという特徴がある。BW 変換後のデータにAカウンタを作用させることで部分列数え上げ法と等価になることを示した上で，その際の BW 変換をほかの変換に置き換えることで部分列数え上げデータ圧縮法の亜種が生成できる。そのような亜種の中でも，BW 変換のかわりに全単射 BW 変換を採用した方法は，部分列数え上げデータ圧縮法に必要な巡回記号列としての同値類から対象データを一意に特定するための情報を必要としない方法となっている。さらに，Aカウンタの効率的な実装法を提案し，提案した実装法を応用して全単射部分列数え上げ法も効率的に実現できることを示した。提案した Aカウンタの実装法は，部分列数え上げを BW 変換によって実現する手法に基づいている。BW 変換後のデータにAカウンタを作用させることが部分列数え上げであり，Aカウンタの実装に BW 変換を用いるというのは，模式的には恒等変換ではあるが，実装という観点では，BW 変換を二重に利用したことになる。実際に提案法をプログラムし，Aカウンタの高速実現が可能であることを確認した。ただし，全単射 BW 変換を採用した改良では，一部の特徴的なデータを除いては圧縮性能の有意な改善は見られなかった。これらの成果は１件の解説論文のほか，3 件の学会発表として発表した。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 横尾英俊	４．巻 12
２．論文標題 部分列数え上げデータ圧縮法とその周辺	５．発行年 2018年
３．雑誌名 電子情報通信学会基礎・境界ソサイエティ Fundamentals Review	６．最初と最後の頁 21-29
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1587/essfr.12.1_21	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9件（うち招待講演　1件／うち国際学会　2件）

１．発表者名 Danny Dube and Hidetoshi Yokoo
２．発表標題 Fast construction of almost optimal symbol distributions for asymmetric numeral systems
３．学会等名 2019 IEEE International Symposium on Information Theory, ISIT 2019（国際学会）
４．発表年 2019年

１．発表者名 Hidetoshi Yokoo and Danny Dube
２．発表標題 Asymptotic optimality of asymmetric numeral systems
３．学会等名 42nd Symposium on Information Theory and its Applications, SITA2019
４．発表年 2019年

１．発表者名 入江明宏, 横尾英俊
２．発表標題 部分列数え上げデータ圧縮法拡張のための実装法
３．学会等名 電子情報通信学会情報理論研究会
４．発表年 2020年

１．発表者名
Hidetoshi Yokoo and Toshiki Shimizu

２．発表標題
Probability approximation in asymmetric numeral systems

３．学会等名
International Symposium on Information Theory and Its Applications (ISITA2018) (国際学会)

４．発表年
2018年

１．発表者名
横尾英俊

２．発表標題
部分列数え上げデータ圧縮法のある拡張について

３．学会等名
第41回情報理論とその応用シンポジウム

４．発表年
2018年

１．発表者名
井上レオナルド, 横尾英俊

２．発表標題
エントロピー符号化法ANSのための確率近似法の多元情報源での評価

３．学会等名
電子情報通信学会技術研究報告, vol. 118, no. 477, IT2018-93

４．発表年
2019年

１．発表者名
Xin Qi and Hidetoshi Yokoo

２．発表標題
A new variation of asymmetric numeral systems

３．学会等名
電子情報通信学会技術研究報告, vol. 118, no. 477, IT2018-94

４．発表年
2019年

１．発表者名 横尾英俊
２．発表標題 部分列数え上げデータ圧縮法とその関連法について
３．学会等名 電子情報通信学会 情報理論研究会（招待講演）
４．発表年 2017年

１．発表者名 清水寿樹，横尾英俊
２．発表標題 エントロピー符号化法ABS/ANSの冗長度評価
３．学会等名 第40回情報理論とその応用シンポジウム
４．発表年 2017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