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　２９　年　　６　月　　７　日現在

機関番号：　１２３０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０００５

研究課題名（和文）データ圧縮法CSEを核とする圧縮情報処理基盤技術の再組織化

研究課題名（英文）Systematic unification of fundamental data compression methods related to CSE

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）　　2,400,000円

研究成果の概要（和文）：部分列数え上げデータ圧縮法（Compression by Substring Enumeration. 以下，CSE法）と呼ばれる無ひずみデータ圧縮法の高効率実現法を確立し，関連する諸手法との関係を明らかにすることで，体系化をはかった。また，CSE法を真に実用法とするためには，実際にエントロピー符号化を組み込む必要がある。そのようなエントロピー符号化法として，ANSと呼ばれる手法が有効であることを指摘し，ANSとその関連手法の理論解析を行って強力な性能を証明した。

研究成果の概要（英文）：We have developed and analyzed an efficient implementation for the target method known as compression by substring enumeration (CSE). We have established a system of data compression techniques related to the CSE. First, we introduced the Burrows-Wheeler (BW) transform to realize a faster implementation of CSE. Second, we developed a new enumeration method which can be applied to a BW-transformed string. We have shown that the combination of the new method with the BW transform is completely equivalent to the CSE. Finally, we analyzed two entropy coders: asymmetric binary systems and asymmetric numeral systems as a possible final step encoder for CSE. We have shown that both entropy coders asymptotically attain the entropy bound for any memoryless source.

研究分野：　情報数理工学

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　エントロピー符号化

１．研究開始当初の背景

(1) ビッグデータ等のキーワードに象徴されるデータ爆発の時代にあって，データの圧縮処理技術の重要性は著しく高まっている。様々な圧縮基盤技術が実用に供されていると同時に，新規手法の提案や開発も活発に続いている。それらは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の展開を進めているものが少なくない。また，新規の手法ほど，関連する諸手法と結び付けてシステムとして利用する必要がある。

(2) 発案されて間もないデータ圧縮法である部分列数え上げデータ圧縮法 (Compression by Substring Enumeration。以下，CSE 法) は，圧縮情報処理のコア技術の一つとして可能性が期待されるものの，十分な解析や実用化が進められてはいなかった。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。特に，具体的なエントロピー符号化法導入はまったく検討されていない状況であった。算術符号と呼ばれる従来型のエントロピー符号化法の存在を暗黙の前提として CSE 法が開発されてきたため，その真の有用性の検証が必要であった。

２．研究の目的

本研究では，CSE 法を中心にすえて，関連する諸手法の体系化およびシステムとして利用可能とするための要素技術の開発を目的とした。特に，他の手法との関連の解明にも力点をおいた。ユニバーサル・データ圧縮のための従来法であるブロック－ソート法との関連の解明や CSE 法に組み合わせることの可能なエントロピー符号化法の検討である。

３．研究の方法

(1) CSE 法の実用可能性の検証のために，効率的な実装法の開発に取り組んだ。同時に，ブロックソート法の基礎にある BW 変換との関係を追及し，BW 変換が CSE 法の効率的実現にとっても有用なものであることを見出した。この結果を受け，BW 変換による CSE 法実現のための要素技術の開発およびプログラムによる実現に取り組んだ。実現結果の種々の最適化後，実データおよびシミュレーション用データを用いた圧縮実験を行い，実性能を評価した。

一方，CSE 法の効率的実現法が BW 変換に基づくブロックソート法の実現にもなっていることを踏まえ，BW 変換後のデータの符号化法を CSE 法の観点から見直すことにした。主として文字列処理アルゴリズムの視点での理論的考察である。

(2) CSE 法の実用化には，実際の符号化を可能とする具体的なエントロピー符号化が必要であるため，その候補の調査と候補例についての理論解析を行った。調査の結果，Asymmetric Binary Systems (ABS) および Asymmetric Numeral Systems (ANS) と呼ばれる手法が候補となり得ることが判明したので，その妥当性を検証するため，両手法の理論解析を進めることにした。

４．研究成果

(1) BW 変換を利用した CSE 法の高効率実現法を実装し，大規模実データを含む種々のデータでの圧縮性能評価を可能とした。

CSE 法では，長さ n ビットの巡回的な 2 元系列が符号化対象であり，以下では，これを x で表す。符号化対象 x に含まれる部分列 w の個数を C(w)で表すと，巡回列を考えていることから

$$C(w) = C(w0) + C(w1)$$
$$= C(0w) + C(1w)$$

が成り立つ。これを整合性条件という。整合性条件と各個数の非負性を組み合わせることで，次を導くことができる。

$$\max\{0, C(0w)-C(w1)\} \leqq C(0w0) \leqq \min\{C(0w), C(w0)\}.$$

CSE 法では，より短い部分列に対する出現回数を利用して C(0w0)の値を符号化することで **x** 自身を符号化する。そのため，C(0w0)等の値を高速に数え上げることが必要である。そのための新手法として，BW 変換に基づく次のような手法を実現した。

BW 変換とは，x の巡回シフトを辞書順に整列し，それらの最後尾の記号だけを連接する記号列の変換である。一例として，x = 00001101 を考えたとする。入力系列 **x** のすべての巡回シフトを辞書順に整列し，行列状に配置する。下の図では，これを BWT 行列と呼んでいる。図の w[i]と LCP[i]は，それぞれ，BWT 行列のその行とすぐ上の行との最長共通接頭辞とその長さを表している。配列 S[i]と E[i]は，w[i]を接頭辞として有する行が BWT 行列の S[i]行から E[i]行にあることを示す。R[i]は，BWT 行列の右端の 1 列における第 i 行までの 0 の個数である。すると，次を示すことができる。

$$C(w[i]) = E[i] - S[i]+1,$$
$$C(w[i]0) = i - S[i],$$

i	w	LCP	S	E
0	$-$	-1	$-$	$-$
1	000	3	0	1
2	00	2	0	2
3	0	1	0	4
4	01	2	3	4
5	λ	0	0	7
6	10	2	5	6
7	1	1	5	7

BWT 行列							BWT(x)	R
0	0	0	0	1	1	0	**1**	0
0	0	0	1	1	0	1	**0**	1
0	0	1	1	0	1	0	**0**	2
0	1	0	0	0	0	1	**1**	2
0	1	1	0	1	0	0	**0**	3
1	0	0	0	0	1	1	**0**	4
1	0	1	0	0	0	0	**1**	4
1	1	0	1	0	0	0	**0**	5

図．BWT 行列とその他の配列

C(0w[i]) = R[E[i]] – R[S[i]-1],
C(0w[i]0) = R[i-1] – R[S[i]-1].

以上の関係式を利用することにより，従来の CSE 法の実現法に比べてはるかに高速かつ省メモリの実現法が可能になった。符号化法の全体の流れは次のとおりである。

1. **x** の BW 変換，および LCP の計算
2. 配列 S, E, R の計算
3. LCP 値の安定整列。この結果を
   0=LCP[i1]<LCP[i2]≦LCP[i3]≦…
   とする。
4. j=1, 2, 3, …の順に必要な C(0w[ij]0)の値を符号化する。

以上の符号化法に加え，これまで明確にされなかった復号法についても明確なアルゴリズムを与えた。

提案した新手法を実際にプログラムとして実装し，種々の圧縮実験を行い，これまで困難だった大規模データに対しても，CSE 法の実性能の確認が可能になった。理論的に予測されていたエントロピーへの収束やその際の冗長度について，理論の妥当性を検証することができた。

(2) 以上の結果とは独立に，CSE 法と BW 変換のより直接的な関係を導くことができた。

BW 変換とは記号列の変換法であり，変換後の記号列にさらに別の変換や符号化法を組み合わせることで最終的な圧縮結果を得ることができる。変換後のデータにさらに加える手法がこれまでいくつも提案され，理論解析や実際的な性能の評価の対象となっている。そのような手法の一つとして今回新たに提案した手法を用いることによって，BW 変換と新手法を組み合わせたシステムが CSE 法と等価であることを示した。このことは，BW 変換を利用したブロックソート法の変種の一つとして，CSE 法に成立する種々の性質がそのまま成立する方法の存在を示す重要な結果である。

(3) CSE 法では，C(0w0)の値の列を符号化することで入力系列を符号化している。C(0w0)の値の符号化では，可能な値の集合上に特定の確率分布が仮定される。仮定した確率分布にしたがって C(0w0)を符号化するためには，この目的を達成することのできるエントロピー符号化法が必要である。

従来のエントロピー符号化法の代表例である算術符号は基数変換の一般化とみなすことができる。ABS/ANS も同様の考えを共有しているが，算術符号が MSB から出力するのに対し，ABS/ANS では LSB から出力するという違いがある。さらに，算術符号の符号化過程では，符号語を数値の区間によって表している。一方，ABS/ANS では，LSB から出力を行うため，区間という概念を必要としない。そのため，符号化の状態は 1 個の整数変数で記述することができる。具体的な符号化に応じて整数変数の取り得る範囲が固定されるため，この範囲内の各整数値の生起する定常分布が明らかになれば，これらの符号化法の精密な解析が可能になる。しかし，定常分布を一般的に記述することは困難であり，そのため，ABS/ANS のエントロピー符号化法としての理論的確立は不十分である。この点を克服するため，定常分布の近似を試み，近似分布と真の分布の関係の理論解析を行った。

まず，解析の対象とする ABS，ANS 双方の符号化・復号アルゴリズムを明確にした。ABS は 2 元データを対象とするため，情報源のパラメータは例えば 0 の出現確率の p である。ABS では p をある条件を満たす q で近似する。すると，ABS の出力の平均符号語長は，情報源のエントロピー，および p と q の近似の度合いを示すダイバージェンス，そして，状態の定常分布で決まる量という 3 量の和で与えられることが判明した。ただし，定常分布を解析的に表現するのは困難なため，それに対する近似分布を提案した。近似分布が単調で滑らかな分布であるのに対し，真の分布はパラメータに応じて滑らかな分布から遠ざかる。しかし，状態の取り得る範囲を大きくしたとき，両分布が漸近的に一致することが示せた。同時に，状態の取り得る範囲を無限に大きくすることで，q を任意の精度で p に近づけることができる。この結果，ABS の平均符号語長はエントロピーに漸近することが理論的に証明できたことになる。証明に利用した技法は ANS にも応用可能なものであり，ANS においても同様の漸近的最良性を証明することができた。

以上の結果は，ANS が CSE 法に限らず，種々のモデルに対して強力なエントロピー符号化法となることを示唆するものである。しかし，実際に CSE 法に組み合わせる点については今後の課題である。

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 1 件）

① Sho Kanai, Hidetoshi Yokoo, Kosumo Yamazaki, Hideaki Kaneyasu, Efficient implementation and empirical evaluation of compression by substring enumeration, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 査読有, Vol. E99-A, No. 2, pp. 601-611, 2016.
DOI: 10.1587/transfun.E99.A.601

〔学会発表〕（計 5 件）

① 佐藤 源，横尾英俊，BW 変換による部分列数え上げデータ圧縮法の導出とその関連法，第 39 回情報理論とその応用シンポジウム，査読無，ポスター，高山市，2016 年 12 月.

② Hidetoshi Yokoo, On the stationary distribution of asymmetric numeral systems, International Symposium on Information Theory and Its Applications, ISITA 2016, **査読有,** pp. 662-666, Monterey, USA, 2016 **年** 11 **月**.
③ Hidetoshi Yokoo, On the stationary distribution of asymmetric binary systems, 2016 IEEE International Symposium on Information Theory, ISIT 2016, **査読有,** pp. 11-15, Barcelona, Spain, 2016 **年** 7 **月**.
DOI: 10.1109/ISIT.2016.7541051
④ **横尾英俊, エントロピー符号化法** ABS **についての一考察, 第** 38 **回情報理論とその応用シンポジウム, 査読無,** pp. 439-444, **倉敷市,** 2015 **年** 11 **月**.
⑤ **金井　翔, 横尾英俊,** Compression by substring enumeration **符号化法の** BWT **行列による実現, 情報処理学会研究報告, 査読無,** Vol. 2014-AL-148, No. 7, pp. 39-46, **松山市,** 2014 **年** 6 **月**.

６．研究組織
(1)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