科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２１　日現在

機関番号：　１２３０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００００６

研究課題名（和文）新規な変換型無ひずみデータ圧縮法の展開と応用

研究課題名（英文）Development of novel transform-based lossless compression schemes

研究代表者
横尾　英俊（Yokoo, Hidetoshi）

群馬大学・理工学研究院・教授

研究者番号：７０１３４１５３

交付決定額（研究期間全体）：（直接経費）　　3,000,000円、（間接経費）　　900,000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ための二つの記号列変換法のGRPとCSEについて，解析と拡張を加えた。GRP変換は，ブロックソートデータ圧縮法のためのBWT変換のパラメトリックな拡張である。これを任意のパラメータに対応できるように更に拡張し，記号列長の線形時間で変換可能なアルゴリズムを導いた。次に，CSE法について解析を行い，定常エルゴード情報源およびマルコフ情報源に対して漸近最良な符号化モデルを導出した。さらに，CSEの高効率な実現法を明らかにし，プログラムとして実装した。最後に，これらの記号列の変換の応用として，秘密分散への応用を提案した。

研究成果の概要（英文）：We have developed and analyzed two transformations for lossless compression: generalized radix permutation (GRP) and compression by substring enumeration (CSE). The GRP transform was proposed as a parametric generalization of the BWT of the block-sorting data compression algorithm. Our proposed extension can be applied to arbitrary parameter values and can transform a string in time linear in the string length. We then analyze the CSE algorithm, and propose encoding models that achieve asymptotic optimality for stationary ergodic sources and Markov sources of any order. We also establish a concrete way for efficient implementation of CSE. CSE often produces more codewords than necessary. In order to reduce such redundancy, we propose a method for computing the maximum length of the substrings that should be enumerated in CSE to uniquely identify the input string. As for an application of these transformations, we propose a practical secret sharing scheme for string data.

研究分野：　総合領域

科研費の分科・細目：　情報学・情報学基礎

キーワード：　情報数理　情報理論　データ圧縮　符号化　ユニバーサル符号

１．研究開始当初の背景

(1) データの量的側面での問題を圧縮という技術によって解決しようとする圧縮技術の進展に伴い，そのような技術の中には単に圧縮にとどまらない意義を有する手法が登場している。代表例として，Burrows-Wheeler 変換（以下，BWT と略す）に基づくデータ圧縮法が著名である。BWT は圧縮対象をいったん別のデータに変換する手法であり，データ圧縮のための変換法としてだけでなく，簡潔データ構造，ゲノム情報処理等の著しい発展に中心的な貢献をしている。一方において，BWT そのものの高性能化や一般化はほとんど試みられておらず，また BWT 以外の同様の目的の変換法もほとんど知られていない。

(2) 筆者らが開発した Generalized Radix Permute (GRP) 変換は BWT 変換の拡張であり，数少ない BWT 変換の一般化のうちの一つである。さらに，GRP 変換とは独立に考案された Compression by Substring Enumeration (CSE) 符号化法の復号に GRP 変換の逆変換が利用できることが判明していた。無ひずみデータ圧縮法の諸性能の向上が進み，目だった進展が報告されにくくなっている状況において，これらの諸手法の相互関係を深く理解することは，このような変換法の高性能化や新規応用の開拓に新しい道筋をつけることにつながると考えられる。

２．研究の目的

(1) CSE 符号化法は無ひずみデータ圧縮のための最新の手法であり，経験的な性能がわずかに示されてはいるものの，効率的な符号化法も復号法も未開発である。そのため，その経験的性能の検証も，さらに多くのデータを対象とした網羅的な評価も困難な状況にある。また，理論的な性能の評価にいたっては全く着手されていない状況にある。しかしそれでも，BWT をはじめとする既存の有力な手法との関連が部分的に明らかにされており，CSE 法の解析と評価は理論と実践の両面から進める必要がある。本研究では，CSE 法の漸近的最良性の理論的検証と実性能を評価するための符号化および復号アルゴリズムの開発を第一の目的とした。

(2) BWT 変換も CSE 法も対象データを別のデータに変換するという点において類似している。連続データに対するフーリエ変換がそうであるように，これらの変換はデータのある種の側面を抽出する機能を有している。また，抽出した側面すべてがそろって初めて元データが再現できるという点においても共通性がある。変換後のデータに対する符号化モデルの検討を通じて，これらの変換が抽出しようとしているデータの側面を明らかにする。同時に，このような共通性を利用した新規応用を開発する。

３．研究の方法

(1) CSE 法の理論的評価のために CSE 法自身のより明確な記述と解決すべき問題群の列挙を行った。CSE 法はその名が示すように，部分列の列挙が基本になっている。まず，どのような部分列が列挙の対象となるのかをできるだけ明示的に述べることを試みた。同時に，列挙に伴って条件を満たす記号列の集合が狭まっていく過程をモデル化した。さらに，漸近的最良性を示すために必要な補助定理等を整備し，漸近最良が達成できる符号化モデルを導入した。

一方，CSE の実装用として Compacted Substring Tree (CST) と呼ばれるデータ構造が提案されている。この構造と BWT 変換との関係を調べることで，CSE 法の線形性の証明の切り口とした。

以上の考察を統合することで CSE 法の理論的特徴づけと漸近的最良性の証明を行った。さらに漸近的最良性達成のために利用する符号化モデルに対して情報理論的解釈を加えることで，モデルの意義の把握を試みた。

(2) CSE 法の実際的な性能の評価と改善を目標として，本来の CSE 法に含まれるむだの解明を行った。さらに，そのようなむだの除去のために，CSE 法と GRP 変換との関係を利用した復号法を開発した。

次に，CSE 法のむだを CST や関連するデータ構造との関係で理解することで，より効率的な実現法を設計することを試みた。設計した実現法はコンピュータ・プログラムとして実装し，性能の実際的な評価を行った。さらに，入力系列から直接 CST を生成するアルゴリズムや BWT に現れる行列を利用した CSE 符号化の開発にも着手した。

(3) 以上述べてきた BWT, GRP, CSE のいずれもが，入力データを別のデータに変換するという共通性を有している。本側面に着目した新たな応用例として，秘密分散に相当するデータ分散管理手法を設計した。たとえば，同じデータを複数箇所に分散することでデータの保全性は高まるが，情報漏洩の可能性が高まることで安全性は低下する。このような安全性低下を防ぐには，分散した個別のデータからだけでは元のデータが再現できない仕組みを導入すればよい。このことを本研究の枠組みの中で具体化することにした。

４．研究成果

(1) CSE 法の解析のために，まず，CSE 法自身の符号化法の明示的な記述を与えた。

CSE 法の符号化対象を D とすると，D は長さ n ビットの 2 元系列である。D に含まれる部分列 w の個数を C(w)で表し，D の末尾と先頭がつながった巡回列を考えると

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

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。

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

このときの C(0w0)のとりうる可能性は

$$\min\{C(0w), C(1w), C(w0), C(w1)\}+1$$

とおりである。したがって，

$$U(D) = \{w \mid C(w0)>0 \text{ and } C(w1)>0\},$$
$$V(D) = \{w \mid C(0w)>0 \text{ and } C(1w)>0\},$$
$$I(D) = U(D) \cap V(D)$$

を定義すると，n=|D|であるような 2 元系列 D の符号化手続きは次のようになる。

```
Encode n; Encode C(0);
For i=0 to n-2 do
  For every w∈D of length i do
    If w∈I(D) then Encode C(0w0).
```

上に示したように，C(0w0)はその下限と上限が符号化直前に判明するような区間に値を取る整数値であるため，「Encode C(0w0)」は当該区間におけるモデルに応じた符号化を意味する。そのようなモデルとして，次の 4 種を検討した。

1) 一様分布モデル
2) 可能な D の個数から決まるモデル
3) 超幾何分布モデル
4) w の長さに応じて切り替えるモデル

これらのモデルのうち，4)の切り替えモデルの前半に 1)の一様分布モデルを使い，切り替え後に 2)もしくは 3)のモデルを利用するモデルが次の意味での漸近的最良性を有することを証明した。

定理：エントロピーH(X)の k 次マルコフ情報源 X からの出力に CSE 符号化法を適用すると，入力長 1 ビットあたりの平均符号語長が n→∞の極限で H(X)に漸近する。

定理：エントロピーレート H の定常エルゴード情報源からの出力に CSE 符号化法を適用すると，入力長 1 ビットあたりの符号語長が確率 1 で H に漸近する。

なお本成果を契機として，CSE 法の冗長度の評価，本手法を応用した反辞書データ圧縮法の漸近的最良性の証明，CSE 法の多元アルファベットへの一般化，上述のモデルの変形等，他の研究者による研究が活発に行われるようになっていることを付記する。

(2) CSE 法は上述の漸近的最良性を持つという意味で理論的には優れたデータ圧縮法であるが，実際に利用するには，実用的なアルゴリズムを開発する必要がある。

まず，符号化アルゴリズム用に開発された Compacted Substring Tree (CST) というデータ構造の解析，ならびに BWT との関係解明を行った。その結果，次の事実が判明した。

定理：D を表す CST において，記号列 0w に対応する頂点が存在するための必要十分条件は，D において 0w と 1w の両者が生起していることである。同様に，記号列 1w に対応する頂点が存在する必要十分条件は，D において 0w と 1w の両者が生起していることである。

定理：長さ n の非反復的な D に対応する CST から後退辺を除去すると，0 に対応する頂点を根とする部分木と 1 に対応する頂点を根とする部分木は同型となり，頂点数はともに n-1 である。

系：長さ n の非反復的な D に対応する CST の頂点数は 2n-1 である。

CST は CSE を提案した原論文で提案されたデータ構造である。原論文では，上記の「系」が「予想」の段階に留まっていた。この成立が厳密に示せたことにより，CSE 実行の線形性が確立できたことになる。とは言え，これは理論的な見積りに過ぎず，現実的にも実行可能なものにするには，より実際的な実現法を開発する必要がある。そこで本研究では，CSE のむだに着目し，その除去を実現法開発の糸口とした。

上述の CSE の符号化手続きでは，すべての w∈I(D)について C(0w0)の値を符号化している。本研究では，その必要がないことを明らかにした上で，符号化に必要な w の長さが満たすべき十分条件とその長さを具体的に計算するアルゴリズムを開発した。これを利用することで，符号化で利用される CST をはじめとする木構造の高さを前もって制限することができ，効率的な符号化が可能となった。同時に，原論文において CST の生成後に符号化していた点を見直し，符号化では CST 経由の必要がないことを明らかにした。むしろ CST が活用できるのは復号においてであり，これまで明確に議論されることの少なかった復号法に対しても，効率的なアルゴリズムを与えた。

上記の符号化および復号のアルゴリズムを計算機プログラムとして実装して CSE 法の実際的な性能を評価した。その結果，マルコフ情報源からの出力に対しては，理論が予想するとおりの性能を示すことが検証できた。1 例として，図 1 にはエントロピー 0.8819 [bit]，1 次，2 次，3 次の条件付エントロピーが，それぞれ，0.8447，0.7121，0.6507 [bit] の 3 重マルコフ情報源からの出力を対象として実際に圧縮率を評価した結果を示す。使用した符号化モデルは，w の長さに応じて一様分布から超幾何分布に切り

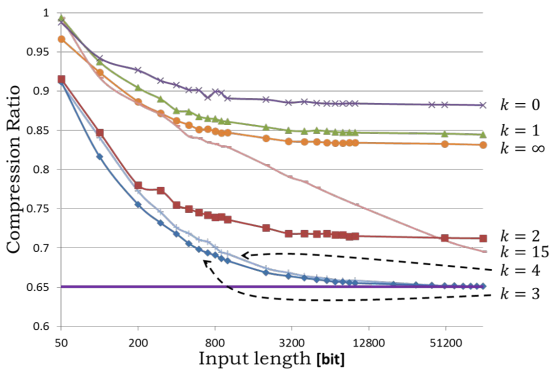

図1. マルコフ情報源での圧縮率の測定結果

替えるモデルである。図中，k の値が切り替えたときの w の長さを表す。

(3) 現実のデータの多くは，本質的には記号列である。BWT, GRP, CSE は記号列に対する可逆変換であるため，もとのデータを完全に復元することができる一方で，変換後のデータは原データの記号列としての特徴を隠してしまうという特徴がある。この性質を秘密分散法に組み合わせることで，鍵やテーブルを必要としない，簡素な非完全秘密分散法を実現した。

提案法では，データをいくつかの分散情報に分ける。このときの個数を n とすると，$1 < k \leqq n$ であるような与えられた k に対し，k 個以上の分散情報がそろったときに原データが復元できる方式である。そのために，最初に原データを BWT（もしくは，GRP）変換し，それを n 個ずつのブロックに分割する。分割後のブロックを 2 元データとみて，これに排他的論理和を利用する秘密分散法を適用する。本研究では，この提案法を実装し，画像データに応用することで所望の目的が実現できていることを確認した。

５．主な発表論文等

〔学会発表〕（計 5 件）

① 山崎世界，金安英明，横尾英俊，Compression by substring enumeration データ圧縮法の効率的実現，電子情報通信学会技術研究報告，査読無，Vol.113, No.411, IT2013-51, pp. 35-40, 大阪市，2014 年 1 月.

② 津久井香奈，横尾英俊，記号列に対する変換に基づく秘密分散法，電子情報通信学会技術研究報告，査読無，Vol.112, No.420, IT2012-100, pp. 53-58, 仙台市，2013 年 1 月.

③ 横尾英俊，CSE 無ひずみデータ圧縮法の情報理論的解釈，電子情報通信学会技術研究報告，査読無，Vol.111, No.390, IT2011-45, pp. 37-42, つくば市，2012 年 1 月.

④ Danny Dúbe and Hidetoshi Yokoo, The universality and linearity of compression by substring enumeration, 2011 IEEE International Symposium on Information Theory, ISIT 2011, 査読有，pp. 1619-1623, Saint-Petersburg, Russia, 2011 年 8 月.

⑤ Hidetoshi Yokoo, Asymptotic optimal compression via the CSE technique, 2011 First International Conference on Data Compression, Communication and Processing, CCP 2011, 査読有，pp. 1-8, Plinuro, Italy, 2011 年 6 月.

６．研究組織

(1)研究代表者

横尾　英俊（YOKOO HIDETOSHI）
群馬大学・理工学研究院・教授
研究者番号：70134153