様式 C-19

科学研究費補助金研究成果報告書

平成　21　年　3　月　31　日現在

研究種目：基盤研究(C)
研究期間：2007 - 2008
課題番号：19560370
研究課題名（和文）　反辞書確率モデルを用いた同期誤り耐性をもつ適応情報源符号化方式の開発

研究課題名（英文）　Development of Adaptive Source Coding Scheme Using Stochastic Model Based on Antidictionary for Synchronization Error Resilience

研究代表者　電気通信大学・大学院情報システム学研究科・教授　森田　啓義（Hiroyoshi Morita）（研究者番号 80166420）

研究成果の概要：
反辞書とは，情報記号列において，その記号列には出現しない極小禁止部分記号列全体の集まりをいう．従来研究では，反辞書で扱えるアルファベットは，２値に限られていたが，本研究において，一般の多値アルファベットも扱えるように拡張し，適応算術符号化の確率モデルとして利用するなど，反辞書符号化法の適用範囲を広げた．さらに，反辞書と辞書の対応関係から接尾辞木の上で反辞書法の動作を実現することによって，符号化・復号化処理時間を情報記号列の長さに比例するまでに高速化した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,800,000	540,000	2,340,000
2008 年度	1,000,000	300,000	1,300,000
年度			
年度			
年度			
総　計	2,800,000	840,000	3,640,000

研究分野：情報理論
科研費の分科・細目：電気電子工学・通信・ネットワーク
キーワード：データ圧縮，反辞書，極小禁止系列，接尾辞木生成算法

１．研究開始当初の背景

反辞書とは，ディジタル情報データ（テキスト，音声・画像，生体計測データなど）を表す有限アルファベットの情報記号列において，その記号列には出現しない極小禁止部分記号列(以下では MFW とよぶ)全体の集まりをいう．たとえば，２値記号列 $\boldsymbol{x}$=01011 に対する反辞書は，00，110，111，1010 の４つの MFW から構成される．これらの MFW はどれも元の記号列 $\boldsymbol{x}$ には含まれないが，どの MFW も両端のシンボルのいずれか一方を削除すると，$\boldsymbol{x}$ に含まれるという意味で極小である．いまの場合，例えば，1010 の右端と左端のシンボルをそれぞれ削除すると，101，010 が得られるが，いずれも，$\boldsymbol{x}$ の部分列である．この反辞書を用いた情報源符号化法（DCA 法とよぶ）が，1998 年に Crochemore らによって提案された．

反辞書を一旦求めておけば，元の記号列を先頭から順に走査しながら，次に出現するシンボルを一意に特定できる場合が頻繁に生じるので，これらのシンボルの符号化処理を省略することができる．上の例における $\boldsymbol{x}$=01011 では，特定できるシンボルを下線で

表すと，01011 となる．したがって，反辞書の情報と元の系列の長さが分かっていれば，残った 00 から，01011 を復元することができる．これが反辞書を用いた符号化の基本的なアイデアである．

DCA 法は，従来から知られる一般的な圧縮手法と同等以上の圧縮率が得られることに加えて，

(1) 復号化と固定系列に対する符号化処理が高速に行えること

(2) 同期特性を持つこと

(3) ハードウェア化に適していること

などの実用的にも優れた特徴をもつ．しかし，これらの利点があるにもかかわらず，比較的新しい手法であることや，形式言語の分野からのアプローチであること，反辞書の生成に非常に多くの計算量を必要とするなどの理由から，情報理論関係の研究者から注目されることはほとんどなく，その基礎的な性質評価や性能向上などの研究は非常に少ない．

ところで，情報理論の分野では，反辞書と逆の概念である辞書，すなわち，情報記号列に出現する部分列全体の集まりを利用した符号化法が，1976 年の Lempel と Ziv の論文に端を発して，これまで数多く研究されてきた．すぐれた圧縮性能をもつフリーウェアの gzip もその流れを組む代表的な符号化法である．また，計算機科学の分野では，辞書を表すデータ構造である接尾辞木についても盛んに研究されている．

本研究代表者は，2004 年ビーレフェルト大で行った招待講演において，接尾辞木の中間ノードと MFW の関係を明らかにする不等式を発表し，それに基づいて，接尾辞木上で MFW を効率的に探索する手法を編み出した．これによって，反辞書を情報記号列の長さに比例する時間で生成する算法が生まれた．心電図データの圧縮でも gzip と同等の圧縮性能を達成できた．しかし，一般の計算機ファイルに対しては，gzip や DCA 法と比べて同等の圧縮性能を達成するには至らずにいたが，2005 年に発表された大川らの DCA 法に算術符号化を組み合わせた符号化方式から，反辞書を算術符号化の確率モデルとして用いるアイデアを得て，今回の申請に至っている．

２．研究の目的

本研究の目的は，情報記号列から反辞書を生成する高速かつ省メモリな算法を確立することと，その算法に基づいて，情報記号列を適応算術符号化するのに必要な，優れた確率モデルを構築することである．具体的には，次の三つの課題を掲げる．

(1) 情報記号列から反辞書を生成する高速かつ省メモリな算法を確立することと，

(2) その算法に基づいて，情報記号列を適応算術符号化するのに必要な，優れた確率モデルを構築すること．

(3) さらに，提案方式の計算コストの評価，反辞書の理論的解析を行うこと．

３．研究の方法

接尾辞木上で極小禁止語がどのように表わされるかを明確にし，系列の長さに比例した時間で系列に含まれるすべての接尾辞を表す木構造を求める既知の算法を利用して，極小禁止語を表す木構造を動的に構成し，それをもとに確率モデルを構築する．

本研究代表者らは，反辞書と接尾辞木との関係について，すでに「反辞書生成時に接尾辞木上で走査しなければならない節点の数は，木上で二つ以上の枝を持つ節点と最も短い経路を持つ葉に対応し，それらの総数+1 個以下になる」という結果をえており，これを用いれば，接尾辞木の更新の過程で，これら特定の節点のみを選んで探索することによって，線形時間で反辞書を生成することが可能である．プログラムの実装，検証を行いつつ，反辞書生成算法について，生成ノード数，MFW の平均長について確率的漸近解析を行う．

４．研究成果

(1) 反辞書の線形時間生成法の確立

反辞書生成に関しては，従来から知られている，Ukkonen による接尾辞木生成算法に基づいて，反辞書木生成のための算法を与えた．この算法は，

① 多値アルファベットに対応できる．

② 情報記号列から長さに比例した時間・メモリ量で反辞書木を生成できる．

③ 情報記号列を先頭からシンボルを読みつつ、反辞書木を動的に更新する．

という特色をもつ．これらの結果は論文にまとめ公表された．

(2) 反辞書を用いた適応型算術符号の開発

(1)で提案した算法では，反辞書を更新していく過程で，探索する節点への訪問頻度を情報記号の出現確率のモデルとして利用することができる．そこで，従来法のように情報記号列全体から確率モデルを作成して再度符号化を行う２パス方式ではなく，反辞書を動的に生成しつつ，それまでに得た頻度情報を基に確率モデルを構成する１パス方式を考案し，プログラムへの実装を行った．

(3) 反辞書確率モデルの改良

(2)の反辞書木生成法においては，入力データ列長に比例した時間で処理可能であるが，符号化に反辞書木を用いるには，反辞書生成，符号器への変換という操作が必要であるため，処理を高速化するために，反辞書の情報をすべて接尾辞木上で表現した符号化法を提案した．

とくに接尾辞木生成に利用されるアクテ

ィブポイントと呼ばれる節点に着目し，MFW を調べ上げることなく次に出現するシンボルが削除できるかどうかを容易に予測できることを明らかにし，その予測法を実装した．
実装したプログラムは，２値アルファベットだけでなく直接，多値アルファベットのデータに適用することができ，さまざまなデータのバイト単位での処理に適している．
(4) 反辞書配列
反辞書木のメモリ量は入力系列長に比例するが，ポインタ演算を多用しているため，比例係数は大きくなり，系列長が長くなると，符号化への適用な難しくなる．そこで，木構造でなく配列を用いて反辞書を表現する方法を考案した．
(5) 分岐予測への反辞書応用
マイクロプロセッサの高速化のために，分岐成立の有無を高い精度で予測する問題に，反辞書木を適用した．その結果，パターンマッチングを用いる従来法に比べ，実行時間は 50 分の１から 4600 分の１程度に短縮された．
(6) 心電図波形圧縮への適用
スライド窓を設けた動的反辞書木を動的算術符号の確率モデルに利用することによって，心電図波形の圧縮を行った．心電図波形はほぼ周期的なデータであるが，周期ならびに振幅は一定ではなく時間的に少しずつ変動する非定常なデータであるため，効果的な圧縮は困難で，従来法では，元のデータ量の 40%までしか削減できなかったが，提案したスライド窓をもつ反辞書木確率モデルを用いた動的算術符号化を行うと，データ量を 30%まで削減することができた．
(7) 極小禁止語の長さの解析
定常エルゴード情報源から生成される系列に対する極小禁止語の平均長について，Wyner & Ziv の結果を適用して，極小禁止語の平均長は，系列長をエントロピーレートで割った値に確率収束することを示した．また，この結果の妥当性を計算機実験によって確認した．

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５件，すべて査読あり）

1. Takahiro Ota and Hiroyoshi Morita, ``On-line Electrocardiogram Lossless Compression Using Antidictionary-Based Methods,'' Electrical Proc. of Third International Symposium on Medical Information and Communication Technologies, Feb. 25, 11:30 - 11:45, Montreal, Canada, Feb. 24 -- Feb. 27, 2009.

2. Takahiro Ota and Hiroyoshi Morita, ``On the Sliding Window Variations of Antidictionary Data Compression Using Dynamic Suffix Trees,'' Proc. of International Symposium on Information Theory and its Applications, pp.1105--1110, Auckland, New Zealand, Dec. 7 -- Dec. 10, 2008.

3. Takahiro Ota and Hiroyoshi Morita, ``On the Electrocardiogram Lossless Data Compression Using Almost Antidictionaries,'' 長野県工科短期大学校紀要, Vol. 11, pp. 21--27, 2008.

4. Takahiro Ota and Hiroyoshi Morita, ``On the Construction of an antidictionary with Linear Complexity Using the Suffix Tree,'' IEICE Trans. on Fundamentals, Vol. E90-A, no. 11, pp. 2533--2539, Nov. 2007.

5. Takahiro Ota and Hiroyoshi Morita, ``On the On-line Arithmetic Coding Based on Antidictionaries with Linear Complexity,'' Proc. of 2007 International Symposium on Information Theory, pp.86--90, Nice, France, June 24 -- June 29, 2007.

〔学会発表〕（計８件）

1. 太田隆博，森田啓義，
``長さ制限のある極小禁止語を用いた動的な反辞書データ圧縮法，''
信学技報，IT2008-99，pp. 363--370， 函館，北海道， Mar. 9 - Mar. 10， 2009.

2. Takahiro Ota and Hiroyoshi Morita, ``On the Antidictionary-Based Data Compression and its Applications - On-line Electrocardiogram Lossless Compression -,'' 2008 長野県工科短期大学校研究成果発表会講演論文集，pp. 21--24，上田，長野，Nov. 26， 2008.

3. 森田啓義，“エントロピーレートを達成する実際的な情報源符号について -- 算術符号の確率モデルに関する最近の話題を中心に --,″信学技報，IT2008-41，pp.13 -- 22，鬼怒川，栃木，Oct.7，2008.（招待講演）

4. 太田隆博，森田啓義，“定常エルゴード情報源に対する極小禁止語の長さ,″第 31 回情報理論とその応用シンポジウム予稿集，pp.688-691，鬼怒川，栃木，Oct. 7 -- Oct. 10，2008.

5. 太田隆博，森田啓義，
“スライド窓を用いた反辞書データ圧縮

法,”信学技報, IT2007-68, pp. 127--132, 調布, 東京, Feb. 28 - Feb. 29, 2008.

6.深江裕忠, 森田啓義, ``Suffix array を利用した動的な反辞書生成,″第30回情報理論とその応用シンポジウム予稿集, pp.383-387, 賢島, 三重, Nov. 27 -- Nov. 30, 2007.

7.太田隆博, 森田啓義, ``反辞書による適応型データ圧縮法のモデリング,″第30回情報理論とその応用シンポジウム予稿集, pp.388-391, 賢島, 三重, Nov. 27 -- Nov. 30, 2007.

8.西新幹彦, 森田啓義, 太田隆博, ``反辞書木を用いた分岐予測手法,″信学技報, CAS2007-38, pp.21-24, 世田谷, 東京, Oct.18, 2007.

〔図書〕(計0件)

〔産業財産権〕
○出願状況(計0件)

○取得状況(計　件)

〔その他〕

６．研究組織
(1)研究代表者
電気通信大学・大学院情報システム学研究科・教授　森田　啓義 (Hiroyoshi Morita)
(研究者番号 80166420)

(2)研究分担者

(3)連携研究者