論文の内容の要旨

論文題目	Antidictionary Data Compression Using Dynamic Suffix Trees
学位申請者	太田　隆博

データ圧縮は，ディジタル機器における記憶容量・通信コストの削減，伝送速度の高速化を実現する手法として，情報システムの基幹技術の一つである．

圧縮手法には，大きく分けて，データを読み込みながら逐次的に圧縮を行う動的手法と，データ全体を読み込んでから圧縮を行う静的手法の二つがある．とくに，動的手法には，静的手法のようにデータを二度読みせずに済む点や符号器で用いる符号テーブルを復号器に送らなくてすむなど，実用的に優れた特色を持つ．

これまでの動的データ圧縮手法の多くでは，入力系列を効率よく圧縮するために，辞書と呼ばれる，入力系列に現れる部分系列の集まりを表す接尾辞木というデータ構造が用いられてきた．また，接尾辞木を構築するためには高速な動的構築手法が知られている．一方，2000 年に，辞書と逆の概念である反辞書と呼ばれる入力 2 値系列に出現しない極小禁止語の集まりを利用したデータ圧縮手法(反辞書法)が，Crochemore らにより提案された．

従来研究の結果から，辞書および反辞書は，それぞれデータ圧縮に有効なデータ構造であり，これらを組み合わせた手法は，さらによい圧縮率を与えることが期待される．

本論文の主要な目的は，

(1) 接尾辞木による高速かつメモリ使用量を抑えた反辞書構築手法の実現，

(2) 接尾辞木を用いた高速な動的反辞書法の実現，

(3) 反辞書を利用した高速かつ逐次的な無ひずみ心電図データ圧縮法の実現，

の 3 点である．

接尾辞木を用いて反辞書を構築する方法は従来より知られていたが，入力系列長の 2 乗に比例した計算量がかかる問題点があった．そこで，本論文では，この計算量を削減するために，接尾辞木に新しいポインタ構造を導入することによって，入力系列長に比例する計算量による反辞書構築手法を実現した．また，計算機実験により，少ないメモリ使用量で高速に反辞書構築が行えることを示した．

しかしながら，この構築法を用いて，入力系列を1記号読み込むたびに反辞書

の更新と符号化を同時に行う動的反辞書法を実現するためには，入力系列長の2乗に比例した計算時間がかかる問題点があった．この計算時間を削減するために，本論文では，反辞書の構築を行わずに，接尾辞木だけを用いて反辞書による符号化が行える条件を明らかにした．さらに，この条件を用いることによって，計算時間を線形量に削減した非常に高速な動的反辞書法を実現し，さらに算術符号と組み合わせることによって圧縮効果の向上を図っている．

これまでに，2値系列に対する算術符号の静的確率モデルとして，あらかじめ入力系列から構築した反辞書を用いる静的なデータ圧縮法(OHY法)が提案されているが，反辞書構築の高速化については考えられていない．また，OHY法は，2値系列に対しては辞書を用いた高性能な圧縮手法(LZ法)に匹敵する性能を持つことが示されているが，そのまま多値系列に対して適用すると復号器に送る反辞書のコストが大きくなるため，圧縮性能が悪化する問題点がある．

一方，本論文で提案している算術符号を用いた動的反辞書法は，実用的に有用な動的手法で多値系列を高速に処理できる特長を持つ．データ圧縮手法の性能比較によく用いられるデータベース(Calgary Corpus)に対して，実験による性能評価を行った結果，本論文で提案している動的反辞書法は，従来の動的反辞書法に比べて，ほぼすべてのファイルに対して圧縮率が向上した．また，平均圧縮率による比較では，従来の動的反辞書法に比べて3%向上し，ファイルを2値系列に変換して，ファイルごとに2つのパラメータ(分割ブロックサイズ，極小禁止語の最大長)を変化させて計算機実験で得られたOHY法の最良の報告結果と比較しても1%向上した．

最後に，本論文では，心電図データの無ひずみ圧縮への動的反辞書法の適用を考察している．心電図データの圧縮に関しては，測定と符号化を同時に行う動的手法で長時間のデータを少ないメモリ使用量で処理できる性能が求められる．さきに提案した動的反辞書法を心電図データ圧縮に適用すると，メモリ使用量が測定時間に比例する問題点がある．しかしながら，心電図データの大半は概周期的な波形なので，心電図全体を用いなくても，その一部分だけを用いて反辞書を構築することによって，メモリ使用量の削減と符号化処理の高速化が図れる．そこで，心電図全体に対する反辞書とほぼ同等な圧縮性能をもつためにどれくらいの長さの部分系列から反辞書を構築すればよいかについての計算式を，クーポンコレクターズ問題の適用により導出した．

心電図データ圧縮の性能比較によく用いられるデータベース(MIT-BIH Arrhythmia Database)に対して，実験による性能評価を行ったところ，提案手法は，リアルタイム伝送が可能な処理速度で，LZ法と比較して圧縮率が15%向上した．医学的には望まれていたが，高能率な無ひずみ圧縮が困難であった心電図データに対して，高性能でかつリアルタイム処理可能な無ひずみ圧縮手法を示した．これにより，他の生体情報などの高能率圧縮手法への応用が期待できる．

別紙２

論文審査の結果の要旨

学位申請者氏名　太田　隆博

審査委員主査　森田　啓義

委員　高瀬　國克

委員　田中　健次

委員　星　守

委員　長岡　浩司

委員

委員

本論文は，２値系列に対して高圧縮性能を有することで知られていた反辞書法を，圧縮率・計算量の観点から，多値系列へ一般化する問題に取り組み，一般的なファイルと心電図データに対する高性能な動的反辞書法を新たに提案するものである．

本論文は，全７章から構成されている．

第１章は，序章であり，研究背景と目的について述べている．まず，データ圧縮に対する研究背景と従来研究について述べている．続いて，反辞書法に関する従来研究と問題点について触れ，本論文の目的と主要な結果を提示したのち，最後に論文の構成を述べている．

第２章は，データ構造の定義と反辞書構築法について述べている．最初に，本論文で用いる基本的な表記法，反辞書および接尾辞トライ・接尾辞木などのデータ構造について定義を行っている．続いて，反辞書構築の従来手法について説明している．最後に，従来，２値系列しか扱えなかった接尾辞トライからの反辞書構築法を多値系列に拡張した手法の提案を行い，反辞書をトライ表現した場合の必要主記憶量の上限についての定理を示している．

第３章は，反辞書法について述べている．最初に，Crochemoreらの反辞書法の基本アイデアと彼らの手法について説明している．続いて，大川らにより示された２値系列に対する反辞書法に算術符号を組み合わせた手法(OHY法)を説明している．最後に，２値系列しか扱えなかったOHY法を多値系列に拡張した手法を提案している．

第4章は，接尾辞木を用いた高速かつメモリ使用量を抑えた反辞書構築法について提案している．4.1節では，反辞書構築の従来手法とその問題点について述べている．4.2節では，多値系列から線形計算量で反辞書を構築する手法について提案している．まず，反辞書の要素である極小禁止語を接尾辞木上で効率的に探索するための新しいポインタ構造(MF-link)の定義を行い，一つのMF-linkを定数時間で探索するための命題を証明している．続いて，この命題の結果を用いて，従来，系列長の2乗に比例していた反辞書構築の計算量を線形量に削減する手法を提案している．併せて2値系列から多値系列への拡張を行い，さらに，Crochemoreらの方法においては，記号列としてそのまま出力されていた極小禁止語を，入力系列

のインデックスのペアによって表現することにより，定数メモリ量で効率的に出力する手法も示している．最後に，計算機実験により，アルファベットサイズ(2, 16, 64 値)上のランダム系列に対する性能評価を行っている．4.3 節では，4.2 節で得られた結果を用いて，トライ表現された反辞書を，線形計算量で構築する手法について提案している．

第 5 章では，接尾辞木を用いた高速な動的反辞書法について提案している．5.1 節では，反辞書法の従来手法とその問題点について述べている．5.2 節では，辞書を用いた静的反辞書法について提案している．まず，反辞書法の符号器(AD-automaton)と辞書を表現する接尾辞木の関係について論じ，AD-automaton による記号削除処理を接尾辞木により行う条件について示している．加えて，逆 MF-link を接尾辞木に導入したデータ構造(AD-tree)を用いることで，AD-automaton 上での遷移と同等の動作を行えることを示し，AD-tree を用いた静的反辞書法を提案している．

5.3 節では，接尾辞木を用いた高速な動的反辞書法を提案している．まず，動的反辞書法と線形計算量の動的接尾辞木構築法(Ukkonen 法)の関係について論じ，定理として，Ukkonen 法において新しい枝の挿入位置(アクティブポイント)を AD-automaton 上の遷移位置の代わりに利用できることを示している．この定理と 5.2 節で得られた記号削除の条件を組み合わせることで，線形計算量の動的反辞書法を提案している．5.4 節では，5.3 節での提案手法に算術符号を組み合わせたデータ圧縮手法を提案している．さらに，この提案手法を，データ圧縮手法の性能比較によく用いられるデータベース(Calgary Corpus) に対して，実験による性能評価を行い，従来の動的反辞書法に比べて，ほぼすべてのファイルに対して圧縮率が向上した結果を報告している．また，平均圧縮率による比較では，従来の動的反辞書法に比べて 3%，OHY 法と比較して 1%向上する結果を報告している．

第 6 章では，反辞書を用いた心電図データの高速な 1 パス無ひずみ圧縮法を提案している．6.1 節では，心電図データの特徴，臨床的な要求事項と従来手法の問題点が述べられている．6.2 節では，心電図データの部分系列を反辞書構築用の学習系列として用いた無ひずみ圧縮法(ECG-DCA 法)と ECG-DCA 法に算術符号を組み合わせた(ECG-ACDCA 法)を提案している．6.3 節では，心電図全体から得られる反辞書とほぼ同等の反辞書を得るための学習系列長の導出式を与えている．6.4 節では，心電図データ圧縮の性能比較によく用いられるデータベース(MIT-BIH Arrhythmia Database)の 24 個のファイル(それぞれが 30 分間の ECG ファイル)に対して，実験による性能評価を行い，リアルタイム処理が可能な計算時間で，LZ 法に比べて，ECG-DCA 法が 10%，ECG-ACDCA 法が 15%，圧縮率が向上する結果を報告している．

第 7 章では，本論文の主要結果について述べている．

以上をまとめると，これまで，静的反辞書法には，多値系列に対する圧縮率が悪い問題点があり，この問題を回避できる動的反辞書法にも，系列長が長くなると膨大な計算時間がかかる問題点があった．本論文で得られた成果により，これら二つの問題は同時に解決された．すなわち，動的反辞書法の計算時間が実用レベルまで低減され，一般的なファイルに対する反辞書法の利用を可能にした．加えて，接尾辞木の動的生成法を利用することにより，反辞書と辞書の同時利用が容易に行える手法が確立され，さらなる圧縮率改良が期待できるようになった．これらの成果は，データ圧縮の新しい展開に貢献するだけでなく，実用性を高めるものとして評価できる．よって，学位論文として相応しいと認める．