学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　　氏名　吉田 諭史

審査担当者　主　査　准教授　喜田 拓也
副　査　教　授　有村 博紀
副　査　教　授　湊 真一
副　査　教　授　Zeugmann Thomas

学位論文題名

Efficient Variable-to-Fixed Length Coding Algorithms for Text Compression
（テキスト圧縮に対する効率よい可変長-固定長符号化アルゴリズム）

データ圧縮技術は、情報科学・計算機工学において、古くより議論されている重要なデータ処理技術の一つである。データ圧縮の圧縮率の理論的限界は C. E. Shannon によって定義された情報源のエントロピーで示されるが、実データに対してその理論限界に近づけるための効率よい符号化アルゴリズムについては、今なお多くの研究者らによって議論されている。実応用されているデータ圧縮法としては、古典的には、Huffman 符号や算術符号といった、記号の発生確率にしたがってデータ系列に可変長の符号語を割り当てるエントロピー符号化が主流であった。現在では、J. Ziv と A. Lempel によって開発された LZ 法やその変種にエントロピー符号化を組み合わせる手法が代表的である。

これに対し、本論文の著者は、「圧縮データの取り扱いやすさ」という観点から、固定長符号を用いてデータ圧縮を行う可変長-固定長符号化 (VF 符号化) に着目し、その効率よい符号化アルゴリズムの開発に取り組んできた。「圧縮データの取り扱いやすさ」とは、定量的にはパタン照合やデータアクセス等に要する時間で示される。事実、2000 年ごろより、高速なパタン照合を目的としたデータ圧縮法がいくつか提案されているが、その観点から最も成功したものの一つである Byte Pair 符号化は、固定長符号を用いた VF 符号化の一種である。しかしながら、既存の VF 符号化は主に圧縮率において難点があった。

本論文の前半 (3 章、4 章) では、既存 VF 符号化の圧縮速度や圧縮率を改善するために、著者がこれまでに提案した手法がまとめられている。3 章では、VF 符号の圧縮率を改善するために H. Yamamoto と H. Yokoo らが提案した AIVF 符号化について、そのメモリ使用量と圧縮速度を改善する手法を提案している。AIVF 符号化では、入力テキストの使用文字集合の大きさに比例する本数の木構造 (分節木) を辞書としていたが、これを一本の木構造に統合し、仮想的に元の辞書を模倣することで、大幅なメモリ使用量削減と圧縮速度の向上に成功している。また、提案手法によるメモリの削減量に関する理論的解析も与えている。4 章では、VF 符号化で用いられる分節木を、入力テキストを繰り返し走査することで洗練させる手法を提案している。また、それにより、自然言語文章などの実データに対して、可変長符号を用いたデータ圧縮法に匹敵する圧縮率を VF 符号化で達成できることを実証している。

本論文の 5 章では、文法変換と固定長符号を組み合わせることで VF 符号化を実現する手法について論じている。著者は、N. J. Larsson と A. Moffat らが提案した文法変換アルゴリズムである

Re-Pair アルゴリズムに固定長符号を適用する際、最終的な出力サイズを正確に見積もる計算式を示した。それにより Re-Pair アルゴリズムの文法変換処理を中途で打ち切り、固定長符号化する場合において最適なサイズの文法を得る手法を提案している。また、本提案手法により、高速な圧縮・展開処理と極めて良好な圧縮率とを達成できることを実証している。

本論文の 6 章では、VF 符号化を大規模なテキストに適用する際に問題となる点について論じている。大規模なテキストは、通常、複数のブロックに分割してデータ圧縮される。ブロック分割に対する、圧縮率に優れた圧縮法としては、R. Wan と A. Moffat が提案した Re-Merge 法があるが、圧縮に極めて多くの時間を要するという問題があった。著者は、5 章で提案した VF 符号化を基に、ブロック間で辞書を共有する手法を提案し、比較的良好な圧縮率を得ながら Re-Merge 法の数十倍もの高速な圧縮処理が可能であることを実証した。また、著者は本章において、VF 符号化されたデータに対する、元のデータ位置を指定した直接的なデータアクセス手法についても提案している。さらに、本手法による部分文字列抽出が、可変長符号を用いた圧縮法上での場合に比べて 10 倍以上高速に行えることを実証している。

本論文の成果は、次のようにまとめられる。

1. 既存 VF 符号化に対し、その圧縮速度や圧縮率を改善する手法を提案し、その有効性を示した。
2. 文法変換と固定長符号を組み合わせる VF 符号化アルゴリズムを提案し、従来の分節木を用いる VF 符号化では到達が困難であった高い水準の性能を持つ VF 符号化を実現した。
3. VF 符号化を大規模テキストに適用する場合の問題点について有効な解決策を示し、VF 符号化の実応用上の有用性を実証した。

これを要するに、著者は、大規模テキストに対するデータ圧縮において、効率よい VF 符号化アルゴリズムを提案し、圧縮データへのアクセスの容易さと高い圧縮性能とを両立するデータ圧縮法に関する新知見を得たものであり、情報科学におけるデータ圧縮技術分野において貢献するところ大なるものがある。よって著者は、北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。