HOKKAIDO UNIVERSITY

Title	Repetition-Aware Lossless Compression [an abstract of dissertation and a summary of dissertation review]
Author(s)	古谷, 勇
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第14281号
Issue Date	2020-09-25
Doc URL	https://hdl.handle.net/2115/79531
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	doctoral thesis
File Information	Isamu_Furuya_review.pdf, 審査の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

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　古谷　勇

審査担当者　主　査　教　授　有村 博紀
副　査　教　授　吉岡 真治
副　査　教　授　堀山 貴史

学位論文題名

Repetition-Aware Lossless Compression
（反復構造のための可逆圧縮）

データ圧縮 (非可逆圧縮とも呼ぶ) は、データから冗長な部分を取り除いて表現することで、元データを完全に復元可能な形でデータ容量を減少させ、データの保管や通信のコストを削減する技術である。既に 20 世紀初頭のシャノンの情報源符号化の研究により、データ圧縮の可能性と限界が示されているが、現実世界の大規模データに対して圧縮効率と計算効率が高い高効率な圧縮法を開発することは、現在もデータ圧縮分野の主要な目標である。

本学位論文では、現在のデータ圧縮分野において最も注目を集めている研究課題の一つである反復構造をもつデータの圧縮の問題に焦点をあてて、研究を行っている。ここで、「反復構造をもつデータ」とは、その生成過程や内容に起因する複雑な繰り返し構造をもつデータを指し、例として、医療データ、ウェブ百科事典の更新記録、センサーのログデータなどの構造データや、テキストデータや遺伝子情報データなどの非構造データなどがあげられる。このような反復構造を持つデータに対して、2000 年前後に「文法圧縮」(grammar compression) と呼ばれる可逆圧縮の新しい枠組みが提案され、圧縮データとして入力データを一意に導出するようなサイズの小さい文法 (とくに文脈自由文法) を構築することで、高効率なデータ圧縮の実現をはかっている。比較として、従来のジフとレンペルの LZ 圧縮法や、リサネンの算術符号などの「エントロピー圧縮」手法では、文脈とよばれるデータ塊の出現構造をモデル化することで、古典的なハフマン符号に比べて比較的高い圧縮性能を達成したが、遺伝子やテキストなどの複雑な反復構造をもつデータに対しては、圧縮効率が高くなかった。これに対して文法圧縮は、階層的な文法を用いて反復構造を表現することで、これらのデータに対して高い圧縮性能を発揮できると期待されている。一方で、どのような文法族を用いれば高い圧縮率と計算効率を同時に達成できるかは、未解決の問題であり、現在もさかんに研究が行われている。

本研究では、従来の枠組みである文脈自由文法による文法圧縮を拡張して、連長文脈自由文法や高階プログラムなどの拡張された文法を用いた圧縮法を検討している。代表的な文法圧縮手法であるラーソンらによる RePair アルゴリズムでは、チョムスキー標準形の文脈自由文法を用いており、データ中に出現する隣接した文字対を繰り返し一つにまとめることでデータを圧縮する。これに対して、より表現力の高い文法族を用いることで、より高い圧縮率を達成できると期待される。そこで、本学位論文では、圧縮率と計算資源のトレードオフを考慮しつつ、計算効率を保ちながら、高い圧縮性能を達成する新しい文法圧縮手法の開発を目標として研究を行っている。

まず、第 2 章で、本研究に必要な用語や定義の準備を行ったあと、第 3 章では、極大反復部分文

字列と呼ばれる可変長かつ極大な文字列に拡張することで、文法圧縮の高圧縮率化をはかるアプローチをとっている。現在、最も高い圧縮性能で知られる文法圧縮法 RePair の理論的解析を行い、反復の表現である二つの文字の対を、極大反復部分文字列 (maximalrepeat) と呼ばれるデータ中の特徴的な構造に拡張することで、新たな文法圧縮アルゴリズム MR-RePair を提案し、実験評価によって反復構造に対して圧縮性能の向上を確認している。

第 4 章では、チョムスキー標準形の文脈自由文法を、要素の繰り返し回数を直接指定できるような連長文脈自由文法 (run-lengthCFG、RLCFG) に拡張し、第 3 章の提案アルゴリズム MR-RePair をベースに、連長文脈自由文法の枠組みとしては初となる文法圧縮アルゴリズム RL-MR-RePair を提案している。既存手法との比較実験により、反復構造を持つ実データに対する提案手法の有効性を確認している。

第 5 章では、文脈自由文法より表現力が高く、文脈依存な言語を表現可能な文法体系として、関数型プログラミングにおけるラムダ計算式を用いる高階圧縮を研究している。とくに、反復構造の最も基本的な対象である繰り返しを表す整数の効率的圧縮を動機とし、先行研究による二進数と算術演算による分解法を改良して、計算可能性理論で用いられるチャーチ数表現を用いることで、超冪と呼ばれる拡張された指数表現を効率良く圧縮可能な手法を提案している。最後に、第 6 章で本研究を総括し、今後の展望を提示している。

これを要するに、著者は、大規模データ処理分野におけるデータ圧縮の問題を考察し、データがもつ種々の反復構造に着目することで、反復構造をもつデータを効率良く表現する文法表現、および、高い圧縮率と計算効率を達成するデータ圧縮アルゴリズムの設計と実装が可能であるという新知見を得たものであり、情報科学における大規模データ処理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。