Computation over Compressed Strings: Indexing, Sensitivity, and Beyond

赤木, 亨

（様式 2）

氏　　名　：　赤木　亨

論 文 名　：Computation over Compressed Strings:
Indexing, Sensitivity, and Beyond
（圧縮文字列処理：索引，感度，および発展）

区　　分　：　甲

論　文　内　容　の　要　旨

近年のネットワーク技術・センサ技術の発展により，世界各地で膨大な情報のやりとりが盛んに行われるようになった．これらの大規模データは社会や経済などの問題を解決し得るほどの重大な情報が潜んでいる可能性があり，それらの利活用が大きな課題となっている．しかし，これらのデータはサイズが非常に大きいため，解析する際のメモリ不足や通信帯域圧迫が課題となっている．また，Wikipedia や git など，膨大でありながらも頻繁に編集されるデータも数多く存在するため，編集操作に対して堅牢なデータ構造や圧縮形式の開発が課題となっている．

計算機上の多くのデータは記号の列，つまり文字列とみなすこと可能であるため，文字列データ処理の理論とアルゴリズムが適用可能である．そこで，本研究では文字列データ処理において最も基本的かつ重要なパターン照合問題のための圧縮索引構造の開発，および，文字列データ構造が 1 文字編集によって受ける影響 (感度) の理論解析を行った．

本研究では，圧縮手法においては特に文法圧縮と連長圧縮を題材の中心とした。また，1 文字編集に対する圧縮率の変化や，極小不在文字列 (MAW) の個数の変化などについての研究を行った．文字列 T に対する MAW とは，T に出現しない部分文字列でありながら，末尾か先頭の 1 文字を削除してしまうと T に出現するような文字列のことを指す．MAW は，圧縮に直接的に利用されるだけでなく，音楽情報検索やバイオインフォマティクスにも応用を持つ，文字列組合せ論の主題の一つである．さらには，文字列編集の特殊なケースとして，左一文字削除と右一文字追加を連続して行うスライド窓というオンラインストリーミングを模した枠組みにも焦点を当てた．本研究の３つの柱となるテーマは以下の通りである．

(A) 新たな文法圧縮索引の開発と実性能評価
(B) テキスト圧縮・文字列データ構造に 1 文字編集が与える影響の解析
(C) 連長圧縮文字列上の極小不在文字列 (MAW) の個数の上界・下界

(A) では，(A-1) GCIS-Index の開発・理論性能評価と，(A-2) GCIS-Index の実験的性能評価に取り組んだ．(A-1) については，展開に長い時間を要する GCIS 文法圧縮の課題を克服すべく，GCIS が生成する圧縮データそのものからパターンの出現位置を抽出する手法を発見した．その手法を索引構造に発展させた GCIS-Index を考案するとともに，その理論的領域サイズと理論的照合時間を評価した．GCIS-Index の実用的性能を評価するため，最も関連深い既存手法である ESP-Index などとの比較実験を行い，領域や構築時間などを比較評価した．GCIS-Index は，1000 文字を超えるような長いパターンの照合において高速な検索を実現した．また，類似手法の ESP-Index よりも理論的にはコア発見後にパターン出現を判定

する計算量が少なく，実験においても多くのパターン出現が存在するようなデータセットでは，開発した GCIS-Index の処理速度 が ESP-Index を凌駕した．

(B) では，文字列の圧縮手法において 1 文字の編集（置換・追加・削除）が圧縮のサイズに対しどの程度の影響を与えうるかを定量化した「圧縮感度」の概念を提唱し，各種圧縮法の感度の上界・下界の解明に取り組んだ．(B-1) 圧縮文法の中でも特に(A) に関連する GCIS，(B-2) LZ77 の亜種である LZ-End，及び (B-3) Bisection の圧縮感度について新たな上界または下界を証明した．さらに，(B-4) LZSS の上界を用いた，AVL grammar などの他の文法圧縮における非自明な上界の導出を行なった．(B-1) においては，GCIS-Index の技術の根幹となる GCIS が生成する文法圧縮のサイズが，圧縮前文字列の 1 文字編集に対し高々4 倍までの変化であることを証明するとともに，実際に 4 倍を実現する文字列の系列を与えた．これは，GCIS が他の圧縮手法である LZ77 などのように「1 文字の編集に対し，高々定数倍にしか領域が変化しない」という，微小な編集に対し頑健な圧縮率を持つことを示した成果である．(B-2) については，LZ-End の領域が 1 文字の置換・追加・削除によって 2 倍に変化する文字列の系列を与えた．

(C) では，文字列中における MAW を組合せ的性質に基づいてグループ分けし，それぞれのグループに属す MAW の個数の上界・下界を解析した．(C-1) スライド窓上における 1 文字スライド(先頭 1 文字の削除と末尾への 1 文字追加）によって発生する MAW の変化量に対する厳密な上界・下界を与えた．さらに，(C-2) 連長圧縮表現 RLE 上における MAW を 5 つに分類し，それぞれの個数の上界・下界を与え，さらにその連長圧縮表現のサイズに線形なメモリ領域で保持可能な MAW のコンパクト表現を考案した．また，(C-3) 既存データ構造の技術を応用することにより，(C-2) で扱った MAW のコンパクト表現を， O(m log m) 時間で構築する手法を考案した．ここで，m は入力文字列の連長圧縮サイズである．