九州大学学術情報リポジトリ
Kyushu University Institutional Repository

Compact Data Structures for Faster String Processing

鶴田, 和弥

https://hdl.handle.net/2324/4475148

出版情報：九州大学, 2020, 博士（情報科学）, 課程博士
バージョン：
権利関係：

（別紙様式２）

氏　　名　：　鶴田 和弥

論 文 名　：　Compact Data Structures for Faster String Processing
（高速文字列処理のための省領域データ構造）

区　　分　：　甲

論　文　内　容　の　要　旨

インターネット上を流通するデータ量が指数関数的に増大し，科学技術分野ではセンシング技術等の発達を背景に種々の観測・実験データが巨大化している．今後の5Gの普及等により，流通するデータはさらに爆発的に増大することが予想される．このような大規模データから価値ある知識を抽出し利活用したいという要求が，学界のみならず産業界でも高まっている．これらのデータの多くは定まった形式を持たない非定型データ，すなわち，文字列データと捉えることができる．

本研究では，(1) 部分文字列探索，(2) 前方一致探索，(3) 最短ユニーク部分文字列探索，の３つの問題に取り組み，各々について高速かつ省領域なデータ構造を開発した．

(1)の部分文字列探索問題とは，テキスト T とパターン P が与えられて，T 中の P の出現位置を全て求める問題である．T が事前に与えられる場合には，あらかじめ T から索引と呼ばれるデータ構造を構築しておくことにより，高速なクエリ応答が可能となる．そのような索引として，接尾辞木やDAWG，接尾辞配列などが古くから知られているが，いずれも T の数倍から十数倍の領域を要する．このため，圧縮索引の研究が盛んになった．圧縮索引の研究には２つの流れがある．１つは索引自体を情報理論的に圧縮する流れであり，もう１つは，圧縮テキストに補助的情報を付加して索引とする流れである．本研究では，後者のうち文法圧縮に基づく索引に着目する．文法圧縮とは，T からそれを導出する直線的文法を構築し符号化する手法である．ここで，直線的文法とは，単一の文字列を導出する文脈自由文法をいう．文法圧縮は，反復の多いテキストに対して有効な圧縮法として知られている．

文法圧縮に基づく索引に関する先行研究として，ClaudeとNavarroは，任意の直線的文法 G から圧縮索引を構築する手法を提案している．G のサイズを g とするとき，索引サイズは $O(g)$，部分文字列クエリ応答時間は $O(m^2 \log \log_g n + (m + occ) \log g)$ である．ここで，n は T の長さ，m は P の長さ，occ は T における P の出現回数を表す．この手法は，任意の直線的文法に適用可能であるが，m が大きいときには項 $m^2 \log \log_g n$ が大きくなり実用的ではない．

本研究では，Lyndon文法圧縮という新しい圧縮法を提案し，それに基づく文法圧縮索引を開発した．T に対するLyndon文法のサイズを g_L とするとき，索引のサイズは $O(g_L)$，クエリ応答時間は $O(m + \log m \log n + occ \log g_L)$となる．すなわち，$m^2$ の項を除去することに成功している．索引の構築には，$O(n \log n)$期待時間・$O(n)$領域を要する．さらに，いくつかの主要な文法圧縮法に匹敵する圧縮率を示すことを実験により示した．

(2)の前方一致探索問題とは，テキストの有限集合 S とパターン P が与えられて，P を接頭辞にもつ S の全要素を求める問題である．この問題を高速に解くための索引構造として，パトリシア木などの索引構造が古くから知られている．本研究では，$O(N)$ 領域の動的索引を開発した．w ビットのワードRAMモデルの仮定の下，長さ m のパターン P に対する前方一致探索を $O(m/\alpha+\log\alpha+o)$期待時間で，長さ m のテキストの追加・削除を $O(m/\alpha+\log\alpha)$ 期待時間で，それぞれ行うことができる．ここで，N は S のテキスト長の総和，$\alpha=w/\log\sigma$，o は 出力となる要素数，σ はアルファベットサイズである．この結果は $m\geq\alpha$ であるとき既存の動的索引より高速である．また，計算機実験によりその優位性を確認した．

(3) 最短ユニーク部分文字列探索とは，長さ n のテキスト T と整数 p $(1\leq p\leq n)$ が与えられて，p を含む最短ユニーク文字列の集合 SUS(p) を求める問題である．ここで，T の部分文字列 w がユニークであるとは， w が T に1回しか出現しないときをいう．Peiらは SUS(p) の要素の1つを求める問題に取り組み，定数時間で応答可能な $O(n)$ 領域の索引構造を提案した．しかし，索引の構築には $O(n^2)$ 時間を要する．本研究は，SUS(p) の全要素を求める問題に取り組み，SUS(p) の全要素を要素数に比例した時間で出力する索引構造を提案するとともに，この索引を $O(n)$ 時間・領域で構築するアルゴリズムを示した．すなわち，より一般的な問題に対して最適解を示すことに成功している．