

Static and Dynamic Compressed Data Structures for String Processing

西本, 崇晃

<https://doi.org/10.15017/1807062>

出版情報 : 九州大学, 2016, 博士 (理学), 課程博士
バージョン :
権利関係 : 全文ファイル公表済



氏 名 : 西本 崇晃

論 文 名 : Static and Dynamic Compressed Data Structures for String
Processing
(文字列処理のための静的・動的圧縮データ構造)

区 分 : 甲

論 文 内 容 の 要 旨

近年のコンピュータ関連の技術は他の分野とは比べようもないほどに目まぐるしく発達している。特に記憶装置の容量の増加により大規模なデータの蓄積が可能になり、大量のデータを処理して新たな知識を発見するデータマイニングの研究が盛んとなった。大規模データからのデータマイニングを現実的な計算時間と計算資源で行うためには、高速かつ省領域なデータアクセス技術が必要である。しかし、日々増加の一途をたどるデータの大部分は、定まった形式をもたない非定型データであり、表形式の定型データ(関係データ)を対象にした従来のデータベース技術は有効ではない。非定型データは単なる記号の連鎖(すなわち、文字列)とみなすことができる。このため、近年では文字列処理アルゴリズムの研究が盛んに行われている。本論文では以下に述べる基本的な文字列処理を効率よく行うことができる省領域な静的・動的データ構造をいくつか提案する。

(1) 動的最長共通延長(Longest Common Extension; LCE)問題とは、与えられたテキスト T から、LCEクエリと T のテキスト更新操作をサポートするデータ構造を構築する問題である。ここでのLCEクエリとは T の2つの位置が与えられたとき、その位置から始まる2つの接尾辞の最長共通接頭辞長を返すクエリである。また、 T の更新操作としては、任意の位置に長さ y の文字列を挿入すること、もしくは、任意の位置から始まる長さ y の部分文字列を削除することを考える。 T の更新操作を考えない場合は静的なLCE問題と呼ぶ。

LCEクエリには多くの応用が存在する。代表的な応用例として、近似文字列照合、文字列中に含まれる回文構造の検出などが挙げられる。動的なLCE問題に対していくつかの動的データ構造が既に提案されているが、いずれも必要領域が更新回数に依存する問題をもつ。

本論文では T を表現するシグネチャエンコーディング(Signature Encoding)と呼ばれるサイズ $O(z \log |T| \log^* M)$ ワード領域の圧縮表現形式を用いて、LCEクエリが $O(\log |T| + \log L \log^* M)$ 時間、 T の更新操作が $O((y + \log |T| \log^* M) q \log |T| \log^* M)$ 時間で行える新しいアルゴリズムを提案する。 L はLCEクエリの共通接頭辞長、 q はシグネチャエンコーディング内で用いられる後続クエリ応答時間、 $z(\leq |T|)$ は T をLZ77分解したときのファクタ数、 M は T の取りうるテキスト長である。 T に文字列の繰り返しが多く含まれているとき z の値は非常に小さくなるので、そのような T の場合にこのデータ構造は省領域である。提案手法は動的LCE問題における初の省領域な手法であり、必要領域は更新回数に依存しない。さらに、本論文ではこの結果を用いて既存のいくつかの応用問題の計算量を改善できることを示すとともに、この結果を用いた新しい応用を提案する。

(2) 動的索引問題とは、与えられたテキスト T から検索クエリと(1)で述べた T の更新操作をサポートするデータ構造を構築する問題である。ここでの検索クエリとはパターン P を与えられたとき T 中の P の出現位置をすべて返すクエリである。検索クエリは最も基本的な文字列処理の一つであるので幅広い応用が存在する。

本論文ではシグネチャエンコーディングを用いた新しい圧縮動的索引を提案する。具体的には、提案する索引が検索クエリを $O(q|P| + \log |T| \log w \log |P| (\log^* M)^2 + occ \log |T|)$ 時間、 T の更新操作を

$O((y + \log |T| \log^* M) \log w \log |T| \log^* M)$ 均し時間で答えられることを示す．ここでの occ は T 中出现する P の数を表す．また，必要な領域はオーダーに関して(1)と同じである．省領域な動的索引において，このテキスト更新時間と検索クエリの計算時間は既存手法よりも優れている．さらに，本論文ではこのアイデアを応用してテキストのLZ77分解が省領域で行えることを示す．LZ77分解は主にデータ圧縮に用いられ，他にも様々な応用が知られている．

(3) 圧縮辞書照合問題とは，サイズ n の文脈自由文法 S の形式で表現されたパターン集合 Π から辞書照合クエリを答えるデータ構造を構築する問題である．ここでの辞書照合クエリとは与えられたテキスト T に対して T 中出现している Π のパターンの種類と出現位置をすべて返すクエリである． S ではなく Π が直接与えられた場合は辞書照合問題と呼ばれる．

辞書照合クエリには多くの応用が存在する．代表的な応用例として記憶装置に含まれているコンピュータウイルスの検出やSDI (Selective Dissemination of Information)サービスなどが挙げられる．辞書照合問題に対しては多くの手法が既に提案されているが，圧縮辞書照合問題に対する既存研究は著者の知る限り存在しない．

本論文は $O(|T|(h + |\Pi|))$ 時間で辞書照合クエリを答えることができる $O(n^2 \log N)$ ワード領域の圧縮データ構造を提案し，さらに $O(n^3 \log n \log N)$ 時間で構築できることを示す． h は S の導出木の高さ， N はパターン長の総和である． n が小さい時，提案手法は省領域である．多くの圧縮アルゴリズムの出力は文脈自由文法で表現できるのでこの結果は有用である．