Combinatorial Approaches for Compact String Indexing and Efficient Pattern Discovery

藤重, 雄大

https://doi.org/10.15017/4060184

出版情報：九州大学, 2019, 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　：　藤重　雄大

論 文 名　：　Combinatorial Approaches for Compact String Indexing and Efficient Pattern Discovery
（文字列索引の省領域化とパターン発見の効率化のための組み合わせ論的アプローチ）

区　　分　：　甲

論　文　内　容　の　要　旨

近年のネットワークの普及に加え，センサ技術の発達やモバイル端末の普及により，大規模かつ多様なデータが産出され続けている．このような大規模なデータには，社会や経済などの問題を解決し得る情報が潜んでおり，その利活用が世界中で注目を集めている．ところが，これらのデータの多くは定まった形式をもたないため，定型データを対象に発展してきた従来のデータ解析技術の適用が困難であり，非定型データのための新しいデータ解析技術の確立が急務である．非定型データは，陽に構造を持たない記号の列，すなわち文字列と捉えることができるため，本研究では，文字列を扱うデータ処理技術の効率化に挑む．この効率化には，通常のアルゴリズムとデータ構造に関する知識に加え，文字列データが潜在的にもつ組み合わせ的性質の活用が必須である．例えば，著名なKMPパタン照合アルゴリズムは，パタンの周期性に関する性質に基づく．また，重要な部分文字列索引として知られる接尾辞木とDAWGは，入力文字列の部分文字列上に導入された同値関係 $\equiv_L$，$\equiv_R$ に基づく．

本研究では，文字列の組み合わせ的性質の活用により，(A) 索引構造の高速な構築・省領域化と(B) パタン発見アルゴリズムの効率化の研究を行った．

(A) では，(A-1) DAWG構築アルゴリズムの高速化，(A-2) DAWGの省領域化，(A-3) 頻出部分文字列パタン列挙のための索引構造の省領域化に取り組んだ．(A-1) については，整数アルファベット上の入力文字列に対する世界初の線形時間DAWG構築アルゴリズムの開発に成功した．この成果は，接尾辞木とDAWGを基礎付ける同値関係$\equiv_L$，$\equiv_R$のあいだに成り立つ性質の究明によるものである．双方向部分文字列索引として知られる affix tree についても同じ性質を用いて線形時間構築アルゴリズムを示した．

(A-2) については，部分文字列探索におけるクエリ文字列は十分短いという実用的観点から，長さ k 以下のクエリ文字列に対して正しく動作する部分文字列索引として，k-truncated DAWG を定義し，そのオンライン構築アルゴリズムを開発した．DAWGのサイズは入力文字列 n に関して線形であるのに対し，k-truncated DAWG のサイズは $O(|\mathrm{Sub}_k| + k)$ となる．ここで，$|\mathrm{Sub}_k|$ は入力文字列中の長さ k の部分文字列の異なり数を表す．また，本研究において $|\mathrm{Sub}_k| \leq k\gamma$ を示した．ここで，γ は最小の文字列アトラクタのサイズであり，あらゆる辞書式圧縮手法の圧縮サイズの下界となることが知られている．したがって，入力文字列が十分高い圧縮率で圧縮可能なとき，k-truncated DAWGは DAWG に対して小さくなる．

(A-3) については，頻出部分文字列パタン列挙問題の変種のための索引構造の省領域化に成功した．

頻出部分文字列パタン列挙問題とは，文書(文字列)の有限集合 D と正整数 d が与えられて， d 個以上の文書に生起する文字列をすべて列挙する問題である．本研究では，その変種として，クエリ文字列 p を部分文字列として含み，かつ，文字列の包含関係に関して極大な文字列のみを列挙する問題に取り組み，$O(n\log|D|)$ 領域・$O(|p| + o\cdot\log\log|D|)$ クエリ応答時間の索引構造を開発した．ここで，n は D 中の文字列長の総和，o は解のサイズである．この結果は，Nishimoto らの先行研究を，領域・クエリ応答時間ともに大幅に改善している．

(B) では，入力文字列に含まれる組み合わせ的オブジェクト(combinatorial objects)を列挙する問題に取り組んだ. 組み合わせ的オブジェクトとして, (B-1) 長さ制約付きギャップ付き回文と，(B-2) 極大反復の２つを取り上げた．長さ制約付きギャップ付き回文とは，回文の中央にギャップを許したものであり, 形式的には, vuv^{R} と表せる文字列で, $g_{\min} \le |u| \le g_{\max}$, $u[1] \neq u[|v|]$, $|v| \ge a_{\min}$ という制約を満たすものを指す．また，文字列 w の極大反復とは，$w[i..j]$ が反復文字列となるような区間 $I=[i, j]$ $(1 \le i \le j \le |w|)$ のうち区間の包含関係に関して極大なものをいい，文字列 x が反復文字列であるとは $x = y^e$ となる文字列 y と有理数 $e \ge 2$ が存在するときをいう．

(B-1) の列挙に関しては, 先行研究として Kolpakov と Kucherov の $O(n\log\sigma + occ)$ 時間・$O(n)$ 領域で動作するオフラインアルゴリズムが知られている．本研究では，$O(n\,((g_{\max} - g_{\min}) / a_{\min} + \log\sigma) + occ)$ 時間で動作するオンラインアルゴリズムを開発した． ここで，σ はアルファベットサイズ，occ は出力の個数である．$(g_{\max} - g_{\min}) / a_{\min} \in O(\log\sigma)$ であるとき，計算時間は先行研究のオフラインアルゴリズムと同等である．

一方，(B-2) の列挙に関しては，先行研究として，順序付きアルファベット上の長さ n の文字列に対して $O(n\,\mathcal{A}(n))$ 時間・$O(n)$ 領域で動作する Crochemore らのアルゴリズムが知られている．ここで，$\mathcal{A}$ は逆アッカーマン関数である．長さ n の任意の文字列に含まれる極大反復の個数は n 未満であることが知られているが, 本研究では, まず, 入力文字列の連長圧縮サイズ m に着目し，新たな上界 $2m$ を示した．ここで，連長圧縮とは文字列中に連続して出現する文字をその文字と連続長の対で置き換える圧縮法をいう．次に，連長圧縮形式で与えられた入力文字列に対して極大反復を $O(m\,\mathcal{A}(m))$ 時間・$O(m)$ 領域で列挙するアルゴリズムを開発した．連長圧縮は $O(n)$ 時間・$O(m)$ 追加領域で行うことができ，かつ， $m \le n$ が常に成り立つことから，連長圧縮後に提案アルゴリズムを適用する手法は，先行研究と同等または高速・省領域である．