公開

研 究 終 了 報 告 書

「辞書式順序に基づいた文字列データ処理法の構築」

研究期間： 2020 年 11 月～2023 年 3 月
研 究 者： 中島 祐人

1. 研究のねらい

本研究では，辞書式順序に基づいた文字列構造に注目する．辞書式順序に基づいた古典的な文字列構造には，文字列照合問題に対する索引構造の一つとして知られる接尾辞木や接尾辞配列，有名な圧縮ツールの一つである bzip2 の核である Burrows-Wheeler 変換（BW 変換）などがある．文字列処理におけるこれまでの辞書式順序の利用は，暗に固定された順序の上で文字列を整列することに過ぎなかった．しかし，辞書式順序に基づいて定義される文字列構造は，その順序が異なれば一般に構造も異なる．この特徴に注目し，本研究では，固定された辞書式順序にとらわれない文字列データ処理法を構築する．ある文字列処理 X を行う際に，その前処理として処理 X に適した辞書式順序を与えることで効率的な文字列データ処理を実現する．例えば，BW 変換の圧縮表現である連長 BW 変換に基づいた全文検索のための索引構造や文字列処理アルゴリズムが盛んに研究されており，連長 BW 変換のサイズに依存した計算量での解析が行われている．つまり，連長 BW 変換のサイズが小さくなるような辞書式順序を与えることが，この手法における計算資源の削減に直結する．つまり、文字列の辞書式順序による文字列アルゴリズムやデータ構造の最適化問題を中心に取り組むことを主目的としている。

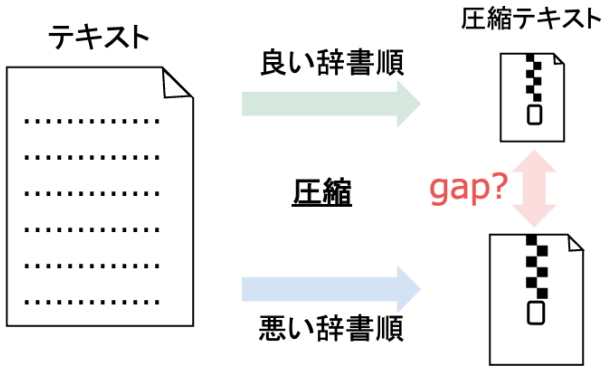
2. 研究成果

（1）概要

文字列の辞書式順序による文字列アルゴリズムやデータ構造の最適化問題に対して，次の二つのテーマに分けて研究を実施した．

A） 辞書式順序による最適化問題の理解と厳密解法の開発
B） 厳密解法の開発を支える文字列構造の数理的性質の解明

主目的であるテーマ A に対しては，大きく二つの成果を得ている．本研究期間で得られた成果は，いずれも文字列圧縮を対象としている．つまり，辞書式順序の変更により，圧縮構造や圧縮サイズがどのように変化するかに注目している．その一つは，lex-parse と呼ばれる辞書式順序依存の文字列分解の一種に対して，辞書



式順序に変更を加える前後での lex-parse のサイズ比の解析である（成果リスト(3)-1）. このサイズ比に対して，タイトな上下界を示した. この成果により，辞書式順序によるサイズ比のタイトな上下界をはじめて明らかにすることに成功している. テーマ B に対しても，大きく二つの成果を得ている. 一つ目は，文字列の部分列として出現する Lyndon 文字列の数について解析を行い，最大延べ数，期待延べ数，期待異なり数を示した. これらの解析により，Lyndon 文字列と呼ばれる辞書式順序依存の文字列クラスの数理的性質に関する知見を得た. 二つ目は，LZ-End 分解と呼ばれる文字列分解について，分解のサイズが最小となる分解を計算する問題の困難性を示した. また，この最小サイズに対して貪欲に得られる分解のサイズの比を考え，非自明な下界を与えた（成果リスト(3)-2）. LZ-End 分解は，本研究で着目している辞書式順序依存の構造ではなく，辞書式順序非依存の構造であるが，間接的に関係があると考えられる Lyndon 分解などの辞書式順序依存の構造の性質を解明するために取り組んだトピックである. また，テーマ B に対しては，研究期間内に様々な辞書式順序依存の文字列構造に対する問題に取り組み，小さいながらも知見を蓄えてきた.

（2）詳細

概要で挙げた各成果について詳細を述べる.

A） 辞書式順序による最適化問題の理解と厳密解法の開発

研究成果 A-1「辞書式順序による lex-parse サイズ比」

lex-parse は，辞書式順序依存の文字列分解の一種である. 一方で，文字列の圧縮スキームの一つである macro scheme は，部分文字列を別の位置での出現へのポインタで表現することにより，文字列をコンパクトに表現することを可能にするが，最小サイズの macro scheme を求めることは NP 困難であることが知られている. lex-parse は文字列長に対して線形時間で計算可能であり，lex-parse サイズの macro scheme を簡単に構成できる. つまり，辞書式順序を変更することでより小さいサイズの lex-parse を選ぶことができれば，より最小に近い macro scheme を構成できることとなり，圧縮率の改善を期待できる. 本研究では，良い辞書式順序を選ぶアルゴリズム開発への第一歩として，辞書式順序に変更を加える前後での lex-parse のサイズ比の最大値について考え，タイトな上下界を与えた. 後に記述する問題などに関連するが，辞書式順序の変化によるサイズ比の非自明なタイトな上下界が明らかになっている文字列構造は現在 lex-parse のみであり，非常に興味深い成果であると言える.

B） 厳密解法の開発を支える文字列構造の数理的性質の解明

研究成果 B-1「Lyndon 部分列の数え上げ」

Lyndon 文字列とは，自身を巡回して得られる文字列集合の中で，自身が辞書式順序最小である文字列のことである. つまり，Lyndon 文字列もまた辞書式順序依存の構造を持つ. 本研究では，文字列の部分列として出現する Lyndon 文字列の数について解析を行い，最大延べ数，期待延べ数，期待異なり数を示した. これらの解析により，Lyndon 文字列の数理的性質

に関する知見を得た.

研究成果 B-2「最適 LZ-End 分解」
LZ 分解とは，部分文字列を前方の出現へのポインタによって文字列を表現する分解である（上記の macro scheme の特殊形である）. この LZ 分解と，本研究課題で対象としている辞書式順序依存の Lyndon 分解との間には非自明な関係が知られている. LZ-End 分解は，LZ 分解の特殊形である. 本研究では，辞書式順序非依存の文字列構造（LZ，LZ-End 分解など）の性質を利用することで，辞書式順序依存の文字列構造（Lyndon 分解など）の性質を間接的に明らかにしようとするアプローチを考え，第一に LZ-End 分解の性質の解明に取り組んだ. 具体的には，LZ-End 分解のサイズが最小となる分解を計算する問題の困難性を示した. また，この最小サイズに対して貪欲に得られる分解のサイズの比を考え，非自明な下界を与えた. 本研究は，領域三期生の栗田和宏氏らとの共同研究である.

研究目標の達成状況等
本研究課題における当初の目的は，辞書式順序による文字列構造の最適化問題を考え，それに対する効率的なアルゴリズムを開発することにあった. ここまでに説明したように，本研究では，アルゴリズムの開発には至らなかったため，当初の目的を十分に達成できたとは言い難い. 文字列の辞書式順序を変更することで，文字列処理の効率化を目指す枠組みは，分野においても歴史が浅い（本研究課題の前後でいくつかの成果が発表されたのみである）こともあり，本質的な進展を導くだけの知見が不足していると考える. したがって，プロジェクトの後半ではこの問題を打破するために，アルゴリズム開発に主眼を置くのではなく，サイズ比解析などの問題に取り組むことで，対象とする問題群の難しさや性質を理解することを目指した. その結果，いくつかの成果を得ることができ，少しずつではあるが進展していると感じている.

3. 今後の展開

直近の展開としては，プロジェクト終盤で得られた成果を論文等にまとめ公開することである. 前項で述べたように，辞書式順序の変更により文字列処理の効率化を目指す枠組みにおける問題に対しては，解決するだけの知見や理解が足りていないと考える. 本質的な解決を目指すため，このプロジェクトの後半で実施した，問題群の難しさや性質を理解するという方針で，引き続き取り組むことが重要であると考えている. 今後数年（3～5 年程度）をかけて，このような知見を蓄えつつ，最適化問題としての難しさの解明やアルゴリズム開発へと発展させていく. さらに将来的には，得られた基盤技術を応用することで，実データや実問題を対象とした問題設定などを考慮したアルゴリズム開発や実装を目指す.

4. 自己評価

研究目的の達成状況は，上記で述べた通りであり，当初目標に対しては多くの課題を残してしまった. コロナ禍での研究開始となり，思うように研究を進めることができない時期もあったが，約2年半という研究期間に対して，適切な目標設定ではなかったのかもしれない

と感じている. しかし, この経験は今後の研究活動に活かしたいと思う.

研究開始時から, 他研究者との交流の場である領域会議などは(最後を除き)すべてがオンラインでの開催となり, 直接的にネットワークを広げるのは困難であったが, これまでに聞いたことのない多種多様な分野の研究を目にすることで, 自身の研究や活動に対する考え方などに新たな視点を取り入れることができたと実感している.

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数:1件

1. Ryo Hirakawa, Yuto Nakashima, Shunsuke Inenaga, Masayuki Takeda. Counting Lyndon Subsequence. Proc. of the Prague Stringology Conference 2021, pp. 53–60 (2021).
文字列中に部分列として現れる Lyndon 文字列(Lyndon 部分列)の数について解析を行った. 最大延べ数, 期待延べ数, 期待異なり数のそれぞれについて解析を与えることで, Lyndon 文字列の数理的性質を得た.

(2)特許出願

該当なし.

(3)その他の成果(主要な学会発表, 受賞, 著作物, プレスリリース等)

口頭発表

1. アルファベット順による lex-parse サイズ比, 第 191 回アルゴリズム研究発表会, 九州工業大学, 2023 年 1 月
2. 最適 LZ-End 分解, 2022 年度 冬の LA シンポジウム, 京都大学, 2023 年 1 月