解凍不要:　圧縮データの直接解析

● 研究の特徴･独自性

ストレージと処理速度を同時に最適化

私の研究は、圧縮データを解凍せずに処理することを目指しています。データが生（ナマ)で保存された場合にストレージに負担をかける大規模なデータセットを扱う際に有効です。現在では小規模なビジネスでも課題になっています。データを解凍せずに処理する一般的なアプローチは圧縮索引であり、以下のことが可能になります。

- 解凍せずに圧縮データを直接検索する
- ストレージ使用量を節約し、解凍の計算コストを省く
- 生データのサイズに関係なく、大規模データの高速処理を可能にする

圧縮索引は複数で提案されており、それぞれ異なる特性を持ち、異なる機能を提供します。これらの特性を改善し、操作性を向上させることが活発な研究分野です。 実際、圧縮索引はストレージの負担を軽減しますが、特定のタスクのためにデータを分析することを一層困難にしています。

将来の研究課題は、圧縮索引の特性に基づいて、より高度な検索と分析を可能にすることを目指しています。例えば、図1はパターンマッチングをルートから下に向かってパターンを読み取ることで可能にする接尾辞木と呼ばれる索引を示しています。残念ながら、プレーン形式の接尾辞木は入力のサイズの倍数を必要とするため、大規模データには実現不可能です。しかしながら、最近の研究では、入力のわずかな部分のスペースしか必要とせず、大部分の操作を維持する圧縮表現の接尾辞木が明らかにされました。その中でも一般的な技術は、入力文字列をより圧縮しやすく、索引化しやすくするために順列を行うBWTです(図2) 。将来的には、圧縮索引の構築、圧縮性能および操作性の向上を目指しています。

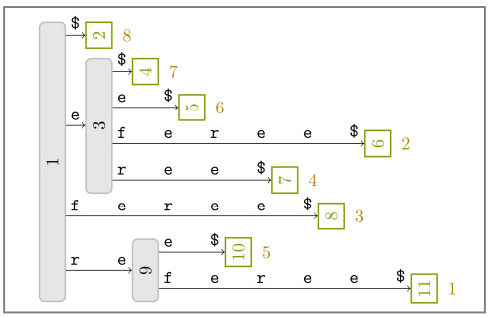

図１：文字列「referee$」 の接尾辞木

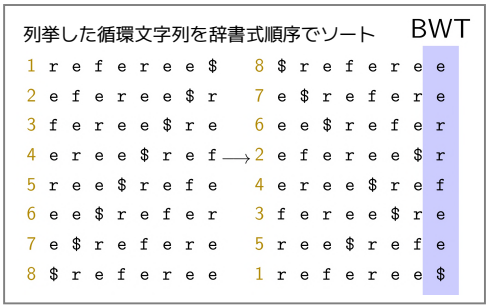

図２：文字列「referee$」 の Burrows–Wheeler変換 (BWT)

● 社会実装･応用例

●産業界へのアピール

- ストレージ削減と高速処理の両立を実現し、クラウドやエッジ環境など、リソース制限のある現場で有効です。
- バイオ・AI・通信・メディア・ゲーム業界との連携を希望しています。

●応用・活用例

- バイオインフォマティクス：遺伝子配列データの圧縮と検索により、新型ウイルスの分類や疾患関連遺伝子の探索を高速化できます。
- 機械学習：事前のデータ圧縮により、学習データの転送や読み込みを効率化し、大規模モデルの学習コストを削減できます。
- IT通信：ウェブデータやログの圧縮・検索によって、通信量と復元コストの両方を最小化できます。
- エンタメ・メディア：SNSや楽曲データにおけるパターン検出やバージョン管理に応用可能で、分析や検索の精度と効率を向上させます。

研究キーワード ：可逆圧縮、圧縮索引、データ構造的特徴、大規模データ解析

大学院 総合研究部 工学域
電気電子情報工学系（コンピュータ理工学）
特任准教授
クップル ドミニク

山梨大学 研究者総覧

論文:
1. Eric M. Osterkamp, Dominik Köppl: Extending the Burrows–Wheeler Transform for Cartesian Tree Matching and Constructing It. Proc. CPM, LIPIcs 331, pages 26:1–26:17. (2025)
2. Dominik Köppl, Florian Kurpicz, Daniel Meyer: Faster Block Tree Construction. Proc. ESA, LIPIcs 274, pages 74:1–74:20. (2023)

お問い合わせ：山梨大学　社会連携・知財戦略室　055-220-8759　renkei-as@yamanashi.ac.jp