様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月　 ６ 日現在

機関番号：１７１０４
研究種目：若手研究
研究期間：2019～2023
課題番号：１９Ｋ２０２１３

研究課題名（和文）BW変換技術の深化による大規模データ処理基盤技術の開発

研究課題名（英文）Deepening BWT for massive data processing

研究代表者
井　智弘（I, Tomohiro）

九州工業大学・大学院情報工学研究院・准教授

研究者番号：２０７７３３６０

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：BW変換は文字列中の各文字をその後ろに続く文字列（文脈）によってソートして得られる可逆変換であり，文字列圧縮や圧縮文字列処理に利用されている．本研究の主な成果は以下の通りである．(1) BW変換文字列を連長圧縮した領域（RLBWT領域）で索引を実現するデータ構造（r-index）の実用性と逐次的な構築の速度を向上させた．(2) RLBWTからLZ77圧縮形式に変換する実用的なアルゴリズムを開発した．(3) 回文照合問題に対するBW変換に基づいた索引を開発した．(4) パラメタ化文字列照合に対するBW変換に基づいた索引を効率的に構築する手法を開発した．

研究成果の学術的意義や社会的意義
データ処理において，データをどのように表現するかは処理の効率に大きく関わる最重要かつ根源的な問題である．圧縮のためのデータ変換手法として提案されたBurrows-Wheeler変換（BW変換）は，後の研究によりデータ処理において様々な利点を有していることが明らかになっている．本研究は，BW変換文字列を連長圧縮した領域で動作するアルゴリズムや一般化文字列照合におけるBW変換の応用技術の発展に寄与した．

研究成果の概要（英文）：The Burrows-Wheeler Transform (BWT) of a string is obtained by sorting each character in the string with its subsequent suffix, which has been used for data compression and compressed data processing. In this project we obtained the following results: (1) We simplified the index based on Run-length BWT (RLBWT) and improved its throughput for direct construction. (2) We proposed a practical algorithm for converting RLBWT to LZ77. (3) We proposed a BWT-based index for palindrome pattern matching. (4) We proposed an efficient algorithm to construct BWT-based indexes for parameterized pattern matching.

研究分野：文字列処理

キーワード：文字列処理　BW変換　圧縮文字列処理

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

データ処理において，データをどのように表現するかは処理の効率に大きく関わる最重要かつ根源的な問題である．データ圧縮がデータの冗長性を捉えより簡潔な表現に変換するように，データ処理における“計算の冗長性”を捉え排除した表現を用いることで計算の効率が向上する．データと計算の冗長性は密接に関係しているため，データ処理の本質を捉えた良い表現は圧縮と高速化の両方を達成することができる．近年，計算機に蓄積されるデータ量が爆発的に増加していることから，データ処理の本質を捉えた表現を追求することの重要性が増している．本研究では，Burrows-Wheeler 変換（BW 変換）に着目する．BW 変換は，圧縮のためのデータ変換手法として提案されたが，後にデータ処理において様々な利点を有していることが判明した．そこで，BW 変換をデータ処理に適した表現への変換技術と位置付け，その可能性を徹底的に追求することで，大規模データ解析の基盤技術を開発する．

２．研究の目的

本研究は BW 変換をデータ処理に適した表現への変換技術と位置付け，BW 変換の技術をさらに深化させることを目的とする．

３．研究の方法

代表者が持つ文字列処理に関する知識を用いて，BW 変換に関する諸問題に取り組む．

４．研究成果

本研究の主な成果として以下の 4 つをあげる．

(1) R-index は，BW 変換文字列を連長圧縮した領域（RLBWT 領域）で索引を実現するデータ構造である．本研究では r-index の実用性と逐次的な構築の速度を向上させる手法を提案し実装を行った．

(2) RLBWT から LZ77 圧縮形式に変換する実用的なアルゴリズムを開発した．本成果をまとめた論文は Data Compression Conference 2022 に採択され発表を行った．LZ77 は高い圧縮率を誇る辞書式圧縮手法でありその変種は広く一般の圧縮アルゴリズムとして利用されている．本研究の成果を使えば，RLBWT から圧縮したまま LZ77 に変換できるので，それぞれの圧縮手法の利点を最大限に活かすことが可能となる．

(3) FM-index は BW 変換に基づいた索引である．FM-index はテキスト中でパタン文字列と完全に一致する部分文字列の位置を検索できるが，完全一致の照合問題を一般化した照合に対してはそのまま使うことはできない．本研究では，回文構造の一致に基づいた一般化文字列照合に対する FM-index を提案した．成果をまとめた論文は Annual Symposium on Combinatorial Pattern Matching 2023 に投稿し採択された．この結果は，FM-index 的な索引構造を設計するために必要な条件を明らかにするための一助になると考えられる．

(4) パラメタ化文字列照合に対する FM-index を効率的に省スペースで構築する問題に取り組ん

だ．パラメタ化文字列照合に対する索引は FM-index ベースのもの以外にも様々提案されており，それらを効率的に構築する手法も広く研究されているが，最も省スペースな索引である FM-index ベースの索引を省スペースで構築する手法はこれまで提案されていなかった．本研究では，この問題に内在する技術的な課題を解決し，パラメタ化文字列照合に対する FM-index ベースの索引を省スペースで構築する初の手法を提案した．

５．主な発表論文等

〔雑誌論文〕　計15件（うち査読付論文　15件／うち国際共著　6件／うちオープンアクセス　5件）

１．著者名 Shinya Nagashita, Tomohiro I	４．巻 －
２．論文標題 PalFM-index: FM-index for Palindrome Pattern Matching	５．発行年 2023年
３．雑誌名 Proc. 34th Annual Symposium on Combinatorial Pattern Matching (CPM) 2023	６．最初と最後の頁 23:1-23:15
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Shigekuni Masaki, Tomohiro I	４．巻 -
２．論文標題 Converting RLBWT to LZ77 in smaller space	５．発行年 2022年
３．雑誌名 Proc. Data Compression Conference (DCC) 2022	６．最初と最後の頁 242-251
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/DCC52660.2022.00032	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Christina Boucher, Travis Gagie, Tomohiro I, Dominik Koeppl, Ben Langmead, Giovanni Manzini, Gonzalo Navarro, Alejandro Pacheco, Massimiliano Rossi	４．巻 -
２．論文標題 PHONI: Streamed Matching Statistics with Multi-genome References	５．発行年 2021年
３．雑誌名 Proc. Data Compression Conference (DCC) 2021	６．最初と最後の頁 193～202
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Kento Iseri, Tomohiro I, Diptarama Hendrian, Dominik Koeppl, Ryo Yoshinaka, Ayumi Shinohara	４．巻 -
２．論文標題 Breaking a Barrier in Constructing Compact Indexes for Parameterized Pattern Matching	５．発行年 2024年
３．雑誌名 Proc. 51st International Colloquium on Automata, Languages, and Programming (ICALP) 2024	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Bannai Hideo、Gagie Travis、I Tomohiro	4．巻 812
2．論文標題 Refining the r-index	5．発行年 2020年
3．雑誌名 Theoretical Computer Science	6．最初と最後の頁 96～108
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.tcs.2019.08.005	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　9件）

1．発表者名 Shinya Nagashita, Tomohiro I
2．発表標題 PalFM-index: FM-index for Palindrome Pattern Matching
3．学会等名 34th Annual Symposium on Combinatorial Pattern Matching (CPM) 2023（国際学会）
4．発表年 2023年

1．発表者名 Masaki Shigekuni, Tomohiro I
2．発表標題 Converting RLBWT to LZ77 in smaller space
3．学会等名 Data Compression Conference 2022（国際学会）
4．発表年 2022年

1．発表者名 Christina Boucher, Travis Gagie, Tomohiro I, Dominik Koeppl, Ben Langmead, Giovanni Manzini, Gonzalo Navarro, Alejandro Pacheco, Massimiliano Rossi
2．発表標題 PHONI: Streamed Matching Statistics with Multi-genome References
3．学会等名 Data Compression Conference (DCC) 2021（国際学会）
4．発表年 2021年

１．発表者名 Kento Iseri, Tomohiro I, Diptarama Hendrian, Dominik Koeppl, Ryo Yoshinaka, Ayumi Shinohara
２．発表標題 Breaking a Barrier in Constructing Compact Indexes for Parameterized Pattern Matching
３．学会等名 51st International Colloquium on Automata, Languages, and Programming (ICALP) 2024（国際学会）
４．発表年 2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関