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　元 年　 ６ 月 １７ 日現在

機関番号：１７１０４
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１６００９

研究課題名（和文）高効率で多様な文字列処理を実現する圧縮変換の理論

研究課題名（英文）On the study of recompressing strings

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）　　3,100,000円

研究成果の概要（和文）：巨大文字列データを解析・利用するために，文字列を圧縮したまま様々な処理を効率的に行う「圧縮文字列処理」の研究を行った．主な成果として以下の五つをあげる．(1) 理論的に最良な領域で文法圧縮を行う手法の提案．(2) 圧縮したまま最長共通接頭辞クエリに応えるためのデータ構造の提案．(3) 連長圧縮された Burrows-Wheeler 変換（RLBWT）を圧縮領域で高速に構築する手法の提案．(4) RLBWT から Lempel-Ziv 77 の変換をオンラインで行う手法の提案．(5) 高い圧縮率を達成することで知られる文法圧縮 RePair を圧縮領域で計算する手法の提案．

研究成果の学術的意義や社会的意義
巨大データを解析・利用するために圧縮は欠かせない技術であり，様々な圧縮手法が提案されている．しかし，一般に圧縮手法には利点と欠点があるため，必要なデータ処理に応じて圧縮形式を変換する技術の発展が望まれる．本研究では，圧縮されたデータの利用価値を上げるために，圧縮形式間の効率的な変換技法を提案した．

研究成果の概要（英文）：We studied the theory of compression to process huge string data. We obtained the following results: (1) We proposed a grammar compression method that works in optimal working space. (2) We showed how to answer longest common extension queries in grammar compressed strings (3) We proposed a practical algorithm to compute run-length encoded Burrows-Wheeler transform (RLBWT) in an online manner. (4) We showed how to compute Lempel-Ziv 77 via RLBWT in an online manner. (5) We proposed the first algorithm to compute RePair (the most notable grammar compression for its high compression performance) in compressed space.

研究分野：文字列処理

キーワード：文字列処理　可逆圧縮

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

データ圧縮は本来データを効率的に蓄積するための技術だが，近年，圧縮データ上で比較，検索，特徴発見など様々な処理を効率的に行う手法が登場し，汎用的前処理としての圧縮の価値が注目されるようになってきた．しかし問題は，処理の内容によって最適な圧縮表現が異なるという点である．例えば，検索を高速に行える圧縮表現が特徴発見に有用な構造を捉えているとは限らない．従って，処理に応じて適した圧縮表現に変換する必要が生まれるが，単純にデータを展開して圧縮し直すという方法では使用領域・時間が元データのサイズに爆発してしまい圧縮の利点を台無しにしてしまう．このような背景から，必要なデータ処理に応じて圧縮形式を効率的に変換する技術の発展が望まれている．

２．研究の目的

本研究の目的は，圧縮表現の変換を圧縮したまま行う圧縮変換の理論を追求し，高効率で多様なデータ処理のための基盤技術を開発することである．

３．研究の方法

連長圧縮された Burrows-Wheeler 変換（RLBWT）や，高い圧縮性能を誇る Lempel-Ziv 77 (LZ77) 型圧縮，圧縮データ上での操作のしやすさに定評がある文法圧縮など，いくつかの主要なデータ圧縮手法に着目し，それらの変換技法を研究する．

４．研究成果

本研究では，三年間で数多くの成果を上げることができた．主な成果として以下の五つをあげる．

(1) 理論的に最良な領域で文法圧縮を行う手法の提案.

本研究では，省スペースで動作する従来手法の作業領域をさらに理論的限界まで絞ることに成功した．この成果は，アルゴリズムの分野におけるトップ会議の一つである Annual European Symposium on Algorithms (ESA) 2017 に採択されるなど，高い評価を得ている．

(2) 圧縮したまま最長共通接頭辞クエリに応えるためのデータ構造の提案.

最長共通接頭辞クエリは，様々な文字列処理に応用がある重要な処理である．本研究では，文法圧縮されたデータ上で最長共通接頭辞クエリに応えるためのデータ構造を提案した．

(3) RLBWT を圧縮領域で高速に構築する手法の提案.

本研究では，RLBWT を圧縮領域で構築する手法を提案し，計算機実験により従来手法より数十倍高速に動作することを示した．この結果は，RLBWT に基づいたデータ構造とアルゴリズムを実用化するにあたって重要な意味を持つ成果である．

(4) RLBWT から LZ77 への変換をオンラインで行う手法の提案.

RLBWT から LZ77 への変換をオンラインで行う手法を提案した．LZ77 は高い圧縮性能を誇り，計算方法が盛んに研究されている．本研究では，RLBWT のオンライン構築を経由することによって，省スペースかつ高速に LZ77 計算を実現した．

(5) RePair を圧縮領域で計算する手法の提案.

RePair は文法圧縮手法の中で高い圧縮率を達成する手法として知られているが，圧縮にかかる計算領域が大きく大規模データに適用する際の問題となっている．本研究では，省スペースで動作する文法圧縮を経由することで RePair を圧縮領域で計算する手法を提案した．

５．主な発表論文等

〔雑誌論文〕（計 15 件）すべて査読あり

① Kensuke Sakai, Tatsuya Ohno, Keisuke Goto, Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, RePair in Compressed Space and Time, In Proc. the Data Compression Conference 2019 (DCC 2019), IEEE Computer Society Press CPS Online, pp.518-527, 2019 https://doi.org/10.1109/DCC.2019.00060

② Tatsuya Ohno, Kensuke Sakai, Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, A faster implementation of online RLBWT and its application to LZ77 parsing, Journal of Discrete Algorithms, 52:18-28, 2018

https://doi.org/10.1016/j.jda.2018.11.002

③ Johannes Fischer, Tomohiro I, Dominik Köppl and Kunihiko Sadakane, Lempel-Ziv Factorization Powered by Space Efficient Suffix Trees, Algorithmica, 80(7):2048-2081, 2018
https://doi.org/10.1007/s00453-017-0333-1

④ Shouhei Fukunaga, Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, Approximate Frequent Pattern Discovery in Compressed Space, IEICE Transactions on Information and Systems, 101-D(3):593-601, 2018
http://search.ieice.org/bin/summary.php?id=e101-d_3_593

⑤ Shunta Nakagawa, Tokio Sakamoto, Yoshimasa Takabatake, Tomohiro I, Kilho Shin and Hiroshi Sakamoto, Privacy-Preserving String Edit Distance with Moves, In Proc. 11th International Conference on Similarity Search and Applications (SISAP 2018), Lecture Notes in Computer Science (LNCS 11223), pp. 226-240, 2018
https://doi.org/10.1007/978-3-030-02224-2_18

⑥ Tatsuya Ohno, Keisuke Goto, Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, LZ-ABT: A Practical Algorithm for alpha-Balanced Grammar Compression, In Proc. 29th International Workshop on Combinatorial Algorithms (IWOCA 2018), Lecture Notes in Computer Science (LNCS 10979), pp. 323-335, Springer-Verlag, 2018
http://doi.org/10.1007/978-3-319-94667-2_27

⑦ Hideo Bannai, Travis Gagie and Tomohiro I, Online LZ77 Parsing and Matching Statistics with RLBWTs, In Proc. the 29th Annual Symposium on Combinatorial Pattern Matching (CPM 2018), pp. 7:1-7:12, 2018
http://doi.org/10.4230/LIPIcs.CPM.2018.7

⑧ Isamu Furuya, Yuto Nakashima, Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Lyndon Factorization of Grammar Compressed Texts Revisited, In Proc. the 29th Annual Symposium on Combinatorial Pattern Matching (CPM 2018), pp. 24:1-24:10, 2018
http://doi.org/10.4230/LIPIcs.CPM.2018.24

⑨ Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, A Space-Optimal Grammar Compression, In Proc. 25th Annual European Symposium on Algorithms (ESA 2017), pp. 67:1-67:15, 2017
http://doi.org/10.4230/LIPIcs.ESA.2017.67

⑩ Tatsuya Ohno, Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, A Faster Implementation of Online Run-Length Burrows-Wheeler Transform, In Proc. 28th International Workshop on Combinatorial Algorithms (IWOCA 2017), Lecture Notes in Computer Science (LNCS 10765), pp. 409-419, Springer-Verlag, 2017
https://doi.org/10.1007/978-3-319-78825-8_33

⑪ Tomohiro I, Longest Common Extensions with Recompression, In Proc. the 28th Annual Symposium on Combinatorial Pattern Matching (CPM 2017), pp. 18:1-18:15, 2017
https://doi.org/10.4230/LIPIcs.CPM.2017.18

⑫ Tomohiro I, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Faster Lyndon factorization algorithms for SLP and LZ78 compressed text, Theoretical Computer Science, 656:215-224, 2016
http://doi.org/10.1016/j.tcs.2016.03.005

⑬ Shouhei Fukunaga, Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, Online Grammar Compression for Frequent Pattern Discovery, In Proc. the 13th International Conference on Grammatical Inference (ICGI 2016), pp. 93-104, 2016

http://jmlr.org/proceedings/papers/v57/fukunaga16.html

⑭ Takaaki Nishimoto, Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Dynamic index and LZ factorization in compressed space, In Proc. The Prague Stringology Conference (PSC 2016), pp. 158-170, 2016
http://www.stringology.org/event/2016/p14.html

⑮ Takaaki Nishimoto, Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Fully dynamic data structure for LCE queries in compressed space, In Proc. the 41st International Symposium on Mathematical Foundations of Computer Science (MFCS 2016), pp. 72:1-72:15, 2016
http://doi.org/10.4230/LIPIcs.MFCS.2016.72

〔学会発表〕(計 11 件)

① Kensuke Sakai, Tatsuya Ohno, Keisuke Goto, Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, RePair in Compressed Space and Time, In Proc. the Data Compression Conference 2019 (DCC 2019), IEEE Computer Society Press CPS Online, pp.518-527, 2019

② Shunta Nakagawa, Tokio Sakamoto, Yoshimasa Takabatake, Tomohiro I, Kilho Shin and Hiroshi Sakamoto, Privacy-Preserving String Edit Distance with Moves, In Proc. 11th International Conference on Similarity Search and Applications (SISAP 2018), Lecture Notes in Computer Science (LNCS 11223), pp. 226-240, 2018

③ Tatsuya Ohno, Keisuke Goto, Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, LZ-ABT: A Practical Algorithm for alpha-Balanced Grammar Compression, In Proc. 29th International Workshop on Combinatorial Algorithms (IWOCA 2018), Lecture Notes in Computer Science (LNCS 10979), pp. 323-335, Springer-Verlag, 2018

④ Hideo Bannai, Travis Gagie and Tomohiro I, Online LZ77 Parsing and Matching Statistics with RLBWTs, In Proc. the 29th Annual Symposium on Combinatorial Pattern Matching (CPM 2018), pp. 7:1-7:12, 2018

⑤ Isamu Furuya, Yuto Nakashima, Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Lyndon Factorization of Grammar Compressed Texts Revisited, In Proc. the 29th Annual Symposium on Combinatorial Pattern Matching (CPM 2018), pp. 24:1-24:10, 2018

⑥ Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, A Space-Optimal Grammar Compression, In Proc. 25th Annual European Symposium on Algorithms (ESA 2017), pp. 67:1-67:15, 2017

⑦ Tatsuya Ohno, Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, A Faster Implementation of Online Run-Length Burrows-Wheeler Transform, In Proc. 28th International Workshop on Combinatorial Algorithms (IWOCA 2017), Lecture Notes in Computer Science (LNCS 10765), pp. 409-419, Springer-Verlag, 2017

⑧ Tomohiro I, Longest Common Extensions with Recompression, In Proc. the 28th Annual Symposium on Combinatorial Pattern Matching (CPM 2017), pp. 18:1-18:15, 2017

⑨ Shouhei Fukunaga, Yoshimasa Takabatake, Tomohiro I and Hiroshi Sakamoto, Online Grammar Compression for Frequent Pattern Discovery, In Proc. the 13th International Conference on Grammatical Inference (ICGI 2016), pp. 93-104, 2016

⑩ Takaaki Nishimoto, Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Dynamic index and LZ factorization in compressed space, In Proc. The Prague Stringology Conference (PSC 2016), pp. 158-170, 2016

⑪ Takaaki Nishimoto, Tomohiro I, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Fully dynamic data structure for LCE queries in compressed space, In Proc. the 41st International Symposium on Mathematical Foundations of Computer Science (MFCS 2016), pp. 72:1-72:15, 2016