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

平成　２９　年　　６　月　１４　日現在

機関番号：　１７１０２
研究種目：　基盤研究(B)（一般）
研究期間：　2013～2016
課題番号：　２５２８００８６

研究課題名（和文）文字列圧縮に基づく知識発見とデータ分類の研究

研究課題名（英文）Knowledge discovery and data classification based on string compression

研究代表者
坂内　英夫（Bannai, Hideo）

九州大学・システム情報科学研究科（研究院・准教授

研究者番号：２０３２３６４４

交付決定額（研究期間全体）：（直接経費）　　11,100,000円

研究成果の概要（和文）：文字列の特徴抽出に関する様々な問題について，主に直線的プログラム（SLP）という文字列の圧縮表現を対象に，圧縮文字列処理アルゴリズムの開発に取り組み，極大繰り返し構造（連）の発見問題，Lyndon 分解，LZ78 分解(圧縮)/LZ77 分解(圧縮)などの問題について，省スペースで効率的なアルゴリズムを考案した．また，15年来の未解決問題であった連予想を肯定的に解決することに成功した．

研究成果の概要（英文）：We worked mainly on developing compressed string processing algorithms for various problems for extracting characteristics of strings from straight line program (SLP) representation of strings. We developed time and space efficient algorithms for computing runs, Lyndon factorizations, LZ78 factorization, LZ77 factorizations, etc. Furthermore, we affirmatively solved the “runs conjecture”, which was open for 15 years.

研究分野：　情報科学

キーワード：　圧縮文字列処理　SLP　繰り返し構造　LZ78　LZ77　連

１．研究開始当初の背景

近年，インターネットや計算機器の発達により，ウェブ文書，ゲノム配列など，多岐にわたる分野で膨大な量の文字列データが生み出されており，利用可能となっている．このような巨大な文字列データを現実的な時間で処理し得る，より高速かつ省メモリな知識発見技術が求められている．巨大文字列データは通信・記憶容量の節約の観点から，通常はそのまま保存されることはなく，データをより短い表現に変換する操作，すなわち圧縮が施された後に蓄積される．しかし，データを利活用する際に圧縮表現を展開してしまうと結局は元の巨大なデータと対峙することになり，その処理に莫大な計算資源が必要となることに変わりはない．この問題に対し，「圧縮文字列処理」のアプローチは圧縮表現を陽に展開せずに直接処理するものである．圧縮は，データに内在する規則性に基づき短い記述を得る操作であるため，圧縮表現が十分にデータに含まれる規則性を抽出できているならば，それを上手に利用し，効率的な処理が可能となる．特に文字列照合問題に関しては多くの研究がされており，非圧縮の文字列での照合よりも高速で実用的なアルゴリズムが開発されている．しかし，文字列照合問題以外においてはまださほど利用されておらず，多くの課題が残っている．

２．研究の目的

本研究の目的は，このような巨大な文字列データから知識発見や文字列データ分類を可能とするために，文字列照合以外の様々な問題に対して圧縮文字列処理のアプローチによる効率的な手法を開発することである．

３．研究の方法

文字列の特徴抽出に関する様々な問題について，主に単一の文字列を導出するチョムスキー標準形の文脈自由文法である直線的プログラム(Straight Line Program - SLP)という文字列の圧縮表現を対象に，圧縮表現を陽に展開することなく，効率よく処理するアルゴリズムの開発に取り組んだ．

４．研究成果

本研究の主な成果は以下の通りである．

(1) サイズ n，構文木の高さが h の SLP で表現された文字列に含まれる極大繰り返し構造(連)を発見する問題について，$O(n^3h)$ 時間・$O(n^2)$ 領域のアルゴリズムを提案した(成果⑪，23)．

(2) サイズ n，構文木の高さが h の SLP で表現された文字列の Lyndon 分解を求める問題について $O(nh(n+\log M \log n))$ 時間 $O(n^2)$ 領域のアルゴリズムを考案した．更に，入力の SLP が LZ78 分解に対応している時，$O(n \log n)$ 時間・領域で動作するアルゴリズムを示した．これらのアルゴリズムを得る過程で，二つの文字列それぞれの Lyndon 分解が分かっている時に，その連接の Lyndon 分解を求める方法について，既存研究の致命的な誤りを発見し，これを正した．また，文字列の Lyndon 分解のサイズはその文字列を表現する任意の SLP のサイズの下界となっているという理論的に興味深い知見を得た(成果⑳，25)．

(3) サイズ n の SLP で表現された文字列の LZ78 分解（圧縮）を行う $O(n+m \log m)$ 時間のアルゴリズムを提案した(成果 21)．ここで，m はその文字列の LZ78 分解のサイズである．これは従来の $O(n|T|^{1/2}+m \log |T|)$ 時間アルゴリズムを改善するものである．また，整数アルファベット上で LZ78 分解を線形時間で計算する初のアルゴリズムを提案した(成果②)．

(4) アルファベット $\Sigma = \{1, \cdots, \sigma\}$ 上の長さ N の文字列に対して LZ77 分解(圧縮)を行う $O(N)$ 時間，$N \log N + O(\sigma \log N)$ ビット領域のアルゴリズムを提案した(成果⑯)．これは，線形時間 LZ77 分解アルゴリズムの中で最も省領域なアルゴリズムである．また，新しい文字が次々に追加されるというオンラインな問題設定に対して，$O(N \log \sigma)$ ビット領域のみを使用したアルゴリズムの中で初めて $O(N \log N)$ 時間で動作するアルゴリズムを提案した(成果⑱)．

(5) 長さ N の文字列に含まれる極大な繰り返し構造(連)を発見する問題について，Lyndon 語および Lyndon 木に基づく連の新しい特徴づけを発見し，連の最大数 $\rho(N)$ が N 未満であるという 15 年来未解決であった「連予想」を肯定的に解決した．これまで $\rho(N)$ の上界を解析した既存研究はいずれ非常に難解であったが，考案した証明は極めて簡潔であり，ブレークスルーをもたらした(成果⑫)．また，この特徴づけから連を計算する新しい線形時間アルゴリズムを提案した．従来のアルゴリズムは文字列の LZ77 分解を求めることが必要であったが、提案アルゴリズムは LZ77 分解を必要としない初めてのアルゴリズムである．

(6) 長さ N の文字列の任意の位置の組が与えられた時，それらの位置で始まる文字列の最長共通接頭辞の長さを求める LCE クエリ問題について，任意のパラメータ $1 \leqq \tau \leqq N$ に対して $O(N\tau)$時間で構築できる $O(N/\tau)$ 領域のデータ構造を用い，任意の LCE クエリに $O(\tau \min\{\log \tau, \log (N/\tau)\})$ 時間で答える手法を考案した．既存の LCE 劣線形領域データ構造は構築に $\Omega(N^2)$ 時間必要であったが，提案手法は構築時間を大幅に改善したものである(成果①)．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 25 件）

① Yuka Tanimura, Tomohiro I, Hideo

Bannai, Shunsuke Inenaga, Simon Puglisi, Masayuki Takeda, "Deterministic sub-linear space LCE data structures with efficient construction", Proc. 27th Annual Symposium on Combinatorial Pattern Matching (CPM 2016), LIPIcs 54, 1:1-1:10, 2016.
DOI: 10.4230/LIPIcs.CPM.2016.1

② Yuto Nakashima, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, "Constructing LZ78 tries and position heaps in linear time for large alphabets", Information Processing Letters, 115(9): 655-659, 2015.
DOI: 10.1016/j.ipl.2015.04.002

③ Tomohiro I, Takaaki Nishimoto, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, "Compressed automata for dictionary matching", Theoretical Computer Science, 578:30-41, 2015.
DOI: 10.1016/j.tcs.2015.01.019

④ Makoto Nishida, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, "Inferring strings from full Abelian period", Proc. 26th International Symposium on Algorithms and Computation (ISAAC 2015), LNCS 9472:758-779, 2015.
DOI: 10.1007/978-3-662-48971-0_64

⑤ Hideo Bannai, Shunsuke Inenaga, Tomasz Kociumaka, Arnaud Lefebvre, Jakub Radoszewski, Wojciech Rytter, Shiho Sugimoto, Tomasz Walen, "Efficient algorithms for longest closed factor array", Proc. 22nd International Symposium on String Processing and Information Retrieval (SPIRE 2015), LNCS 9309:95-102, 2015.
DOI: 10.1007/978-3-319-23826-5_10

⑥ Yuka Tanimura, Yuta Fujishige, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, "A faster algorithm for computing maximal α-gapped repeats in a string", Proc. 22nd International Symposium on String Processing and Information Retrieval (SPIRE 2015), LNCS 9309:124-136,2 2015.
DOI: 10.1007/978-3-319-23826-5_13

⑦ Takaaki Nishimoto, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, "Computing left-right maximal generic words", Proc. Prague Stringology Conference 2015 (PSC 2015), 5-16, 2015.
http://www.stringology.org/event/2015/p02.html

⑧ Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, Masayuki Takeda, "LZD factorization: simple and practical online grammar compression with variable-to-fixed encoding", Proc. 26th Annual Symposium on Combinatorial Pattern matching (CPM 2015), LNCS 9133:219-230, 2015.
DOI: 10.1007/978-3-319-19929-0_19

⑨ Yoshiaki Matsuoka, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, "Semi-dynamic compact index for short patterns and succinct van Emde Boas tree", Proc. 26th Annual Symposium on Combinatorial Pattern matching (CPM 2015), LNCS 9133:355-366, 2015.
DOI: 10.1007/978-3-319-19929-0_30

⑩ Yuya Tamakoshi, Keisuke Goto, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, "An opportunistic text indexing structure based on run length encoding", Proc. 9th International Conference on Algorithms and Complexity (CIAC 2015), LNCS 9079:390-402, 2015.
DOI: 10.1007/978-3-319-18173-8_29

⑪ Tomohiro I, Wataru Matsubara, Kouji Shimohira, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, Kazuyuki Narisawa, Ayumi Shinohara, "Detecting regularities on grammar-compressed strings", Information and Computation, 240:74-89, 2015.
DOI: 10.1016/j.ic.2014.09.009

⑫ Hideo Bannai, Tomohiro I, Shunsuke Inenaga, Yuto Nakashima, Masayuki Takeda, Kazuya Tsuruta, "A new characterization of maximal repetitions by Lyndon trees", Proc. 26th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2015), 562-571, 2015.
DOI: 10.1137/1.9781611973730.38

⑬ Shohei Matsuda, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, "Computing Abelian covers and Abelian runs", Proc. Prague Stringology Conference 2014 (PSC 2014), 43-51, 2014.
http://www.stringology.org/event/2014/p05.html

⑭ Golnaz Badkobeh, Hideo Bannai, Keisuke Goto, Tomohiro I, Costas S. Iliopoulos, Shunsuke Inenaga, Simon J. Puglisi, Shiho Sugimoto, "Closed factorization", Proc. Prague Stringology Conference 2014 (PSC 2014), 162-168, 2014.

http://www.stringology.org/event/2014/p15.html
⑮ Yuto Nakashima, Takashi Okabe, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, “Inferring strings from Lyndon factorization”, Proc. 39th International Symposium on Mathematical Foundations of Computer Science (MFCS 2014), LNCS 8635:565-576, 2014.
DOI: 10.1007/978-3-662-44465-8_48
⑯ Keisuke Goto, Hideo Bannai, “Space efficient linear time Lempel-Ziv factorization for small alphabets”, Proc. Data Compression Conference (DCC 2014), 163-172, 2014.
DOI: 10.1109/DCC.2014.62
⑰ Tomohiro I, Shiho Sugimoto, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, “Computing palindromic factorizations and palindromic covers on-line”, Proc. 25th Annual Symposium on Combinatorial Pattern Matching (CPM 2014), LNCS 8486:150-161, 2014.
DOI: 10.1007/978-3-319-07566-2_16
⑱ Jun-ichi Yamamoto, Tomohiro I, Hideo Bannai, Shunsuke Inenaga, Masayuki Takeda, “Faster compact on-line Lempel-Ziv factorization”, Proc. 31st Symposium on Theoretical Aspects of Computer Science (STACS 2014), LIPIcs 25:675-686, 2014.
DOI: 10.4230/LIPIcs.STACS.2014.675
⑲ Kazuya Tsuruta, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, “Shortest unique substrings queries in optimal time”, Proc. 40th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2014), LNCS 8327:503-513, 2014.
DOI: 10.1007/978-3-319-04298-5_44
⑳ Tomohiro I, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, “Efficient Lyndon factorization of grammar compressed text”, Proc. 24th Annual Symposium on Combinatorial Pattern Matching (CPM 2013), LNCS 7922:153-164, 2013.
DOI: 10.1007/978-3-642-38905-4_16
21 Hideo Bannai, Pawel Gawrychowski, Shunsuke Inenaga, Masayuki Takeda, “Converting SLP to LZ78 in almost linear time”, Proc. 24th Annual Symposium on Combinatorial Pattern matching (CPM 2013), LNCS 7922:38-49, 2013.
DOI: 10.1007/978-3-642-38905-4_6
22 Tomohiro I, Takaaki Nishimoto, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, “Compressed automata for dictionary matching”, Proc. 18th International Conference on Implementation and Appilcations of Automata (CIAA 2013), LNCS 7982:319-330, 2013.
DOI: 10.1007/978-3-642-39274-0_28
23 Tomohiro I, Wataru Matsubara, Kouji Shimohira, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, Kazuyuki Narisawa, Ayumi Shinohara, “Detecting regularities on grammar-compressed strings”, Proc. 38th International Symposium on Mathematical Foundations of Computer Science (MFCS 2013), LNCS 8087:571-582, 2013.
DOI: 10.1007/978-3-642-40313-2_51
24 Shiho Sugimoto, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, “Computing reversed Lempel-Ziv factorization online”, Proc. Prague Stringology Conference 2013 (PSC 2013), 107-118, 2013.
http://www.stringology.org/event/2013/p10.html
25 Tomohiro I, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, “Faster Lyndon factorization algorithms for SLP and LZ78 compressed text”, Proc. 20th International Symposium on String Processing and Information Retrieval (SPIRE 2013), LNCS 8214:174-185, 2013.
DOI: 10.1007/978-3-319-02432-5_21

〔学会発表〕（計 28 件）

〔その他〕
ホームページ等

６．研究組織

(1)研究代表者
坂内 英夫（BANNAI, Hideo）
九州大学・大学院システム情報科学研究院・准教授
研究者番号：20323644

(2)研究分担者
稲永 俊介（INENAGA, Shunsuke）
九州大学・大学院システム情報科学研究院・准教授
研究者番号： 60448404

(4)研究協力者
後藤 啓介（GOTO, Keisuke）
中島 祐人（NAKASHIMA, Yuto）
西本 崇晃（NISHIMOTO, Takaaki）