科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２９　日現在

機関番号：　１７１０２
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７０００２２

研究課題名（和文）データ圧縮に基づく高速パラメタ化文字列照合技法の開発

研究課題名（英文）Fast parameterized pattern matching algorithms based on data compression

研究代表者
稲永　俊介（Inenaga, Shunsuke）

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：６０４４８４０４

交付決定額（研究期間全体）：（直接経費）　　3,200,000円、（間接経費）　　960,000円

研究成果の概要（和文）：機械可読なデータの多くは，文字列とみなすことができる．大規模データの効率的な処理・活用を実現するため，本研究では，高速かつ省領域に動作する文字列処理アルゴリズムを開発した．特に，SLP という圧縮形式で表現された大規模文字列に対して，高度情報処理を高速かつ省領域に行うアルゴリズムを複数開発した．圧縮データを陽に展開することなく処理することで，高速化と省領域化を同時に達成した点が，本研究の最大の特徴である．本研究の成果を，国内・国際会議，国際論文誌において発表し，国内外に発信した．

研究成果の概要（英文）：A sequence of characters or symbols is called a string. Many digital data can be regarded as a string. In order to process and utilize large-scale data, we proposed efficient string processing algorithms that are both fast and memory efficient. In particular, we developed efficient algorithms that work on strings given as straight-line programs (SLPs). We achieved fast and memory efficient solutions by an approach called compressed string processing, where given compressed data is not explicitly decompressed. The results of our work have been published in internal journals/proceedings, and presented in international/domestic conferences.

研究分野：　総合領域

科研費の分科・細目：　情報学・情報学基礎

キーワード：　アルゴリズム　文字列処理　データ圧縮

１．研究開始当初の背景

文字列とは，記号の連鎖のことである．コンピュータ上で取り扱うデータの多くは，文字列とみなすことができる．そのため，文字列データを高速かつ省領域で処理する基盤技術の開発は，情報爆発時代における喫緊の課題となっている．

２．研究の目的

本研究では，テキスト T とパターン P の２つの文字列が与えられたとき，T における P の出現位置を求める文字列照合問題を取り扱う．特に，パラメタ化文字列照合問題という，文字の置換を許した照合問題に着目する．文字列 X 中の文字を置き換えることで，文字列 Y と合致するとき，文字列 X と Y はパラメタ化合致するという．パラメタ化文字列照合問題とは，テキスト文字列 T とパターン文字列 P が与えられたとき，P が T 中でパラメタ化合致する位置の集合を求める問題である．パラメタ化文字列照合は，ソフトウェアメンテナンスや盗作検出，RNA 配列の２次構造照合など，計算機科学やバイオ情報学の重要課題の基盤となるものである．

３．研究の方法

近年，爆発的に増加し続けるデータの省領域な格納方法，および有効な活用方法の開発に注目が集まっている．本研究では，データに内在する冗長性を削除し，データの記述長を短縮するデータ圧縮技術を活用する．特に，直線的プログラム（Straight Line Program, SLP）と呼ばれる圧縮形式で文字列データを表現する方法に着目した．SLP のサイズを n としたとき，展開文字列長 N が n に対して指数的に大きくなる場合が存在する．したがって，圧縮サイズ n の多項式時間で SLP を処理することで，最悪時の処理時間を短縮することが可能となる．

４．研究成果

本研究では，主に以下の研究成果を達成した．

(1) パターン文字列と同じ回文構造を持つテキスト中の部分文字列を探す回文照合問題を提案し，文字種類数が３以下のとき，回文照合問題とパラメタ化文字列照合問題が等価であることを示した．さらに，文字種類数が４以上のときに，回文照合問題を最適時間で解く世界初のアルゴリズムを与えた．

(2) サイズ n の SLP 圧縮文字列が反復部分文字列を含むかどうかの判定を，n の多項式時間で行うアルゴリズムを与えた．これは，イスラエルおよびフィンランドの研究者との国際共同研究成果である．

(3) (2) の研究成果を発展させ，SLP 圧縮文字列に出現する反復部分文字列を高速に求めるアルゴリズムを開発した．また，この技術を応用し，SLP 圧縮文字列中に出現するギャップ付き回文を高速に計算するアルゴリズムを与えた．この成果は，圧縮された大規模データの規則性を，陽に展開することなく発見するための基盤技術となりうる．

(4) 任意の SLP 圧縮文字列が与えられたとき，これを陽に展開することなく，LZ78 圧縮形式に変換する効率的アルゴリズムを与えた．本成果は，内容に編集操作が行われる動的データの圧縮管理に応用可能である．

(5) 文字列の LZ77 分解は，最も基本的な文字列処理の一種であり，gzip 等の圧縮プログラムに利用されている．入力文字列 T が与えられたとき，T の LZ77 分解を省領域かつ高速にオンライン計算するアルゴリズムを与えた．文字列の組み合わせ的性質と高度データ構造の融合により，本研究成果を達成した．

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２７件）（すべて査読有り）

[1] Jun'ichi Yamamoto, Tomohiro I, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, *Faster Compact On-Line Lempel-Ziv Factorization*, Proc. the 31st Symposium on Theoretical Aspects of Computer Science (STACS 2014), pp. 675-686, March 2014. http://dx.doi.org/10.4230/LIPIcs.STACS.2014.675

[2] Kazuya Tsuruta, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Shortest Unique Substrings Queries in Optimal Time*, Proc. the 40th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2014), Lecture Notes in Computer Science 8327, pp. 503-513, January 2014. http://dx.doi.org/10.1007/978-3-319-04298-5_44

[3] Tomohiro I, Hideo Bannai, Shunsuke Inenaga and Masayuki Takeda, *Inferring Strings from Suffix Trees and Links on a Binary Alphabet*, Discrete Applied Mathematics, 163(3): 316 325, January 2014.

http://dx.doi.org/10.1016/j.dam.2013.02.033

[4] Tomohiro I, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, *Faster Lyndon Factorization Algorithms for SLP and LZ78 Compressed Text*, In Proc. the 20th Symposium on String Processing and Information Retrieval (SPIRE 2013), Lecture Notes in Computer Science 8214, pp. 174-185, October 2013. http://dx.doi.org/10.1007/978-3-319-02432-5_21

[5] Shiho Sugimoto, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Computing Reversed Lempel-Ziv Factorization Online*, In Proc. the Prague Stringology Conference 2013 (PSC 2013), pp. 107-118, August 2013. http://www.stringology.org/event/2013/p10.html

[6] Tomohiro I, Wataru Matsubara, Kouji Shimohira, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, Kazuyuki Narisawa, Ayumi Shinohara, *Detecting Regularities on Grammar-compressed Strings*, In Proc. the 38th International Symposium on Mathematical Foundations of Computer Science (MFCS 2013), Lecture Notes in Computer Science 8087, pp. 571-582, August 2013. http://dx.doi.org/10.1007/978-3-642-40313-2_51

[7] Tomohiro I, Takaaki Nishimoto, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, *Compressed Automata for Dictionary Matching*, In Proc. the 18th International Conference on Implementation and Application of Automata (CIAA 2013), Lecture Notes in Computer Science 7982, pp. 319-330, July 2013. http://dx.doi.org/10.1007/978-3-642-39274-0_28

[8] Hideo Bannai, Pawel Gawrychowski, Shunsuke Inenaga, and Masayuki Takeda, *Converting SLP to LZ78 in almost linear time*, In Proc. 24th Annual Symposium on Combinatorial Pattern Matching (CPM 2013), Lecture Notes in Computer Science 7922, pp. 38-49, June 2013. http://dx.doi.org/10.1007/978-3-642-38905-4_6

[9] Tomohiro I, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, *Efficient Lyndon factorization of grammar compressed text*, In Proc. 24th Annual Symposium on Combinatorial Pattern Matching (CPM 2013), Lecture Notes in Computer Science 7922, pp. 153-164, June 2013. http://dx.doi.org/10.1007/978-3-642-38905-4_16

[10] Tomohiro I, Shunsuke Inenaga and Masayuki Takeda, *Palindrome Pattern Matching*, Theoretical Computer Science, 483: 162-170, April 2013. http://dx.doi.org/10.1016/j.tcs.2012.01.047

[11] Toshiya Tanaka, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Computing convolution on grammar-compressed text*, In Proc. Data Compression Conference 2013 (DCC 2013), IEEE Computer Society Press, pp. 451-460, March 2013. http://dx.doi.org/10.1109/DCC.2013.53

[12] Yuya Tamakoshi, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *From Run Length Encoding to LZ78 and Back Again*, In Proc. Data Compression Conference 2013 (DCC 2013), pp. 143-152, IEEE Computer Society Press, March 2013. http://dx.doi.org/10.1109/DCC.2013.22

[13] Takashi Katsura, Kazuyuki Narisawa, Ayumi Shinohara, Hideo Bannai, and Shunsuke Inenaga, *Permuted Pattern Matching on Multi-Track Strings*, In Proc. the 39th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2013), Lecture Notes in Computer Science 7741, pp. 280-291, Springer-Verlag, January 2013. http://dx.doi.org/10.1007/978-3-642-35843-2_25

[14] Keisuke Goto, Hideo Bannai, Shunsuke Inenaga and Masayuki Takeda, *Fast q-gram mining on SLP compressed strings*, Journal of Discrete Algorithms, 18: 89-99, January 2013. http://dx.doi.org/10.1016/j.jda.2012.07.006

[15] Hideo Bannai, Travis Gagie, Tomohiro I, Shunsuke Inenaga, Gad M. Landau, and Moshe Lewenstein, *An efficient algorithm to test square-freeness of*

strings compressed by straight-line programs, Information Processing Letters, 122(9):711-714, October 2012. http://dx.doi.org/10.1016/j.ipl.2012.06.017

[16] Yuto Nakashima, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *The Position Heap of a Trie*, In Proc. the 19th Symposium on String Processing and Information Retrieval (SPIRE 2012), Lecture Notes in Computer Science 7608, pp. 360-371, Springer-Verlag, October 2012. http://dx.doi.org/10.1007/978-3-642-34109-0_38

[17] Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, *Efficient LZ78 Factorization of Grammar Compressed Text*, In Proc. the 19th Symposium on String Processing and Information Retrieval (SPIRE 2012), Lecture Notes in Computer Science 7608, pp. 86-98, Springer-Verlag, October 2012. http://dx.doi.org/10.1007/978-3-642-34109-0_10

[18] Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, *Speeding-up q-gram mining on grammar-based compressed texts*, In Proc. the 23rd Annual Symposium on Combinatorial Pattern Matching (CPM 2012), Lecture Notes in Computer Science 7354, pp. 220-231, Springer-Verlag, July 2012. http://dx.doi.org/10.1007/978-3-642-31265-6_18

[19] Shunsuke Inenaga and Hideo Bannai, *Finding Characteristic Substrings from Compressed Texts*, International Journal of Foundations of Computer Science, 23(2):261-280, February 2012. http://dx.doi.org/10.1142/S0129054112400126

[20] Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, *Computing q-gram Non-overlapping Frequencies on SLP Compressed Texts*, In Proc. the 38th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2012), Lecture Notes in Computer Science 7147, pp. 301-312, Springer-Verlag, January 2012. http://dx.doi.org/10.1007/978-3-642-27660-6_25

[21] Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Verifying and Enumerating Parameterized Border Arrays*, Theoretical Computer Science, 412(50):6959-6981, November 2011. http://dx.doi.org/10.1016/j.tcs.2011.09.008

[22] Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, *Fast q-gram Mining on SLP Compressed Strings*, In Proc. the 18th Symposium on String Processing and Information Retrieval (SPIRE 2011), Lecture Notes in Computer Science 7024, pp. 278-289, Springer-Verlag, October 2011. http://dx.doi.org/10.1007/978-3-642-24583-1_27

[23] Kouji Shimohira, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Computing Longest Common Substring/Subsequence of Non-linear Texts*, In Proc. The Prague Stringology Conference 2011 (PSC 2011), pp. 197-208, Czech Technical University, August 2011. http://www.stringology.org/event/2011/p17.html

[24] Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Inferring Strings from Suffix Trees and Links on a Binary Alphabet*, In Proc. The Prague Stringology Conference 2011 (PSC 2011), pp. 121-131, Czech Technical University, August 2011. http://www.stringology.org/event/2011/p11.html

[25] Takanori Yamamoto, Hideo Bannai, Shunsuke Inenaga and Masayuki Takeda, *Faster Subsequence and Don't-Care Pattern Matching on Compressed Texts*, In Proc. the 22nd Annual Symposium on Combinatorial Pattern Matching (CPM 2011), Lecture Notes in Computer Science 6661, pp. 309-322, Springer-Verlag, June 2011. http://dx.doi.org/10.1007/978-3-642-21458-5_27

[26] Tomohiro I, Shunsuke Inenaga and Masayuki Takeda, *Palindrome Pattern Matching*, In Proc. the 22nd Annual Symposium on Combinatorial Pattern Matching (CPM 2011), Lecture Notes in Computer Science 6661, pp. 232-245, Springer-Verlag, June 2011. http://dx.doi.org/10.1007/978-3-642-21458-5_21

[27] Stanislav Angelov, Shunsuke Inenaga, Teemu Kivioja, and Veli Mäkinen, *Finding Missing Patterns*, Journal of Discrete Algorithms, 9(2):153-165, June 2011. http://dx.doi.org/10.1016/j.jda.2010.08.005

〔学会発表〕（計２０件）

[1] Jun'ichi Yamamoto, Tomohiro I, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, *Faster Compact On-Line Lempel-Ziv Factorization*, 31st Symposium on Theoretical Aspects of Computer Science (STACS 2014), 5-8 March 2014, Lyon, France.

[2] Kazuya Tsuruta, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Shortest Unique Substrings Queries in Optimal Time*, 40th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2014), 25-30 January 2014, High Tatras, Slovakia.

[3] Tomohiro I, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, *Faster Lyndon Factorization Algorithms for SLP and LZ78 Compressed Text*, 20th Symposium on String Processing and Information Retrieval (SPIRE 2013), 7-9 October 2013, Jerusalem, Israel.

[4] Shiho Sugimoto, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Computing Reversed Lempel-Ziv Factorization Online*, Prague Stringology Conference 2013 (PSC 2013), 2-3 September 2013, Prague, Czech Republic.

[5] Tomohiro I, Wataru Matsubara, Kouji Shimohira, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, Kazuyuki Narisawa, Ayumi Shinohara, Detecting *Regularities on Grammar-compressed Strings*, 38th International Symposium on Mathematical Foundations of Computer Science (MFCS 2013), 26-30 August 2013, Klosterneuburg, Austria.

[6] Tomohiro I, Takaaki Nishimoto, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, *Compressed Automata for Dictionary Matching*, 18th International Conference on Implementation and Application of Automata (CIAA 2013), 16-19 July 2013, Halifax, Canada.

[7] Hideo Bannai, Pawel Gawrychowski, Shunsuke Inenaga, and Masayuki Takeda, *Converting SLP to LZ78 in almost linear time*, 24th Annual Symposium on Combinatorial Pattern Matching (CPM 2013), 17-19 June 2013, Bad Herrenalb, Germany.

[8] Tomohiro I, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai and Masayuki Takeda, Efficient Lyndon factorization of grammar compressed text, 24th Annual Symposium on Combinatorial Pattern Matching (CPM 2013), 17-19 June 2013, Bad Herrenalb, Germany.

[9] Toshiya Tanaka, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Computing convolution on grammar-compressed text*, Data Compression Conference 2013 (DCC 2013), 20-22 March 2013, Snowbird, USA.

[10] Yuya Tamakoshi, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *From Run Length Encoding to LZ78 and Back Again*, Data Compression Conference 2013 (DCC 2013), 20-22 March 2013, Snowbird, USA.

[11] Takashi Katsura, Kazuyuki Narisawa, Ayumi Shinohara, Hideo Bannai, and Shunsuke Inenaga, *Permuted Pattern Matching on Multi-Track Strings*, 39th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2013), 26-31 January 2013, Spindleruv Mlyn, Czech Republic.

[12] Yuto Nakashima, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, The Position Heap of a Trie, 19th Symposium on String Processing and Information Retrieval (SPIRE 2012), 21-25 October 2012, Cartagena, Colombia.

[13] Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, *Efficient LZ78 Factorization of Grammar Compressed Text*, 19th Symposium on String Processing and Information Retrieval (SPIRE 2012), 21-25 October 2012, Cartagena, Colombia.

[14] Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, *Speeding-up q-gram mining on grammar-based compressed texts*, 23rd

Annual Symposium on Combinatorial Pattern Matching (CPM 2012), 3-5 July 2012, Helsinki, Finland.

[15] Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, *Computing q-gram Non-overlapping Frequencies on SLP Compressed Texts*, 38th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2012), 21-27 January 2012, Spindleruv Mlyn, Czech Republic.

[16] Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, *Fast q-gram Mining on SLP Compressed Strings*, 18th Symposium on String Processing and Information Retrieval (SPIRE 2011), 17-21 October 2011, Pisa, Italy.

[17] Kouji Shimohira, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Computing Longest Common Substring/Subsequence of Non-linear Texts*, Prague Stringology Conference 2011 (PSC 2011), 29-31 August 2011, Prague, Czech Republic.

[18] Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, *Inferring Strings from Suffix Trees and Links on a Binary Alphabet*, Prague Stringology Conference 2011 (PSC 2011), 29-31 August 2011, Czech Republic.

[19] Takanori Yamamoto, Hideo Bannai, Shunsuke Inenaga and Masayuki Takeda, *Faster Subsequence and Don't-Care Pattern Matching on Compressed Texts*, 22nd Annual Symposium on Combinatorial Pattern Matching (CPM 2011), 27-29 June 2011, Palermo, Italy.

[20] Tomohiro I, Shunsuke Inenaga and Masayuki Takeda, Palindrome Pattern Matching, 22nd Annual Symposium on Combinatorial Pattern Matching (CPM 2011), 27-19 June 2011, Palermo, Italy.

６．研究組織

(1)研究代表者

稲永 俊介 （Shunsuke Inenaga）
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