様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　３月３１日現在

機関番号：１７１０２
研究種目：若手研究（A）
研究期間：2010～2012
課題番号：２２６８００１４
研究課題名（和文）　文字列圧縮に基づく文字列パターン発見・文字列データ分類の研究

研究課題名（英文）　Pattern Discovery and Data Classification Based on String Compression

研究代表者
坂内　英夫（BANNAI HIDEO）
九州大学・システム情報研究院・准教授
研究者番号：20323644

研究成果の概要（和文）：本研究は，文字列の圧縮表現が与えられたときに圧縮表現を展開せずに直接処理をする圧縮文字列処理のアプローチを，文字列パターン発見・文字列データ分類の分野に導入し,様々な関連問題に対して効率的なアルゴリズムを開発した.特に,文字列中の全 q-グラムの頻度問題に対しては,非圧縮の文字列から計算するよりも高速なアルゴリズムの開発に成功し，当該分野における圧縮文字列処理の有効性・実用可能性を初めて示した.

研究成果の概要（英文）：Compressed string processing is an approach that aims to process a compressed representation of the string without explicitly decompressing it. In this study, we investigated the application of this approach to the problem of string pattern discovery and string data classification, and developed various efficient algorithms. Especially for the q-gram frequencies problem, we succeeded in developing a practically efficient algorithm that can be faster than directly processing the uncompressed text, showing, the effectiveness of the approach to the string pattern discovery and string classification problems.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	2,100,000	630,000	2,730,000
２０１１年度	1,900,000	570,000	2,470,000
２０１２年度	1,800,000	540,000	2,340,000
年度			
年度			
総　計	5,800,000	1,740,000	7,540,000

研究分野：知能情報学
科研費の分科・細目：
キーワード：圧縮文字列処理、パターン発見

１．研究開始当初の背景

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．与えられた文字列データからデータの特徴を捉えた意味のあるパターンを効率的に発見する手法は，非常に応用性が高い重要な技術であり，

様々な研究が行われてきた．しかし，例えばゲノム配列解析の分野においては，次世代配列シーケンサが登場したことで，更に桁違いな量の文字列データが短時間で生み出されており，その効率的な処理・解析が大きな問題となっている．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強く求められている．

従来，データ圧縮は通信・記憶容量の節約が主な目標であった．しかし，１０年程前から圧縮された文字列を展開せずに直接処理することで，場合によってはより高速な処理が可能であることが分かり，注目されている．特に文字列パターン照合問題に関しては通常通り非圧縮の文字列に対してパターン照合する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，ほとんどの研究は文字列照合問題しか扱っておらず，圧縮文字列処理は文字列からのパターン発見問題等の知識発見手法における諸技術にほとんど適用されていない．

２．研究の目的

本研究の目的は，文字列を圧縮表現上で直接処理することで，従来では扱うことが事実上不可能な大規模な文字列データにも適用可能な，高速・省メモリな文字列パターン発見・文字列データ分類のアルゴリズムを開発することである．

３．研究の方法

本研究においては文字列の圧縮表現として主に Straight Line Program (SLP) と呼ばれる，単一の文字列を導出するチョムスキー標準形文脈自由文法を対象に，文字列のパターン発見・文字列のデータ分類問題に関連する様々な圧縮文字列処理アルゴリズムの検討を行う．ある SLP が表現するテキスト文字列 T の長さ N はその SLP の大きさ n に対して指数的に大きくなり得るため，n に関して多項式時間で動作するアルゴリズムは，SLP 全体を展開してから非圧縮文字列を処理する任意のアルゴリズムよりも最悪の場合に高速である．

４．研究成果

本研究の主な成果は以下の通りである．

(1) SLP で表現されたテキストに対する部分列および Don’t Care 文字を含んだパターン照合問題を解く高速なアルゴリズム：
テキスト文字列 T を表現する大きさ n の SLP と，長さ m のパターン文字列 P に対して，P の T における部分列としての極小な出現を数え上げる $O(nm)$ 時間アルゴリズムを開発した(論文⑪)．これは[Tiskin 2011] によって提案された $O(nm \log m)$ アルゴリズムよりも効率的である．また，このアルゴリズムを拡張することで，SLP 表現されたテキスト文字列に対して don’t care 文字を含んだパターン照合を行う初のアルゴリズムを提案した(論文⑪)．

(2) SLP で表現されたテキストに対する q-グラム頻度問題を解く高速なアルゴリズム：
テキスト文字列 T を表現する大きさ n の SLP と整数 q に対して，T に出現するすべての q-グラム(長さが q の部分文字列) とその出現頻度を求める $O(qn)$ 時間アルゴリズムを開発した(論文⑩)．これは従来アルファベットの大きさに対して指数的であった $O(|\Sigma|^q qn^2)$ 時間アルゴリズムに比べて大幅に改善されている．実データに対する計算機実験においても，q がある程度小さい場合には SLP 表現から q-グラム頻度を求める方が，非圧縮のテキスト文字列から q-グラムを求めるよりも高速であることを示した．また，このアルゴリズムを応用することで，
① 大きさ n の SLP で表現されたテキスト T の最頻出 q-グラムを求める問題を $O(qn)$ 時間で解くアルゴリズム
② 大きさがそれぞれ n_1, n_2 である SLP で表現された二つのテキストに対して q-グラムスペクトラムカーネルの計算を $O(q(n_1+n_2))$ で解くアルゴリズム
③ 総 SLP サイズ n の，SLP で表現された２つの文字列集合を区別する最適弁別 q-グラムパターン発見問題を $O(qn)$ 時間で解くアルゴリズム
を得ることができ，圧縮文字列処理のパターン発見・テキスト分類分野への有用性を初めて示すことに成功した．

この成果に加え，長さ N のテキスト文字列 T を表現する大きさ n の SLP と整数 q に対して，T に含まれる q-グラムの情報をすべて含む，大きさが $O(N-\alpha)$ のトライ木を構築することで，q-グラム頻度問題が $O(N_\alpha)$ 時間・領域で解けることを示した(論文⑦)．ここで $N_\alpha = N-\alpha$ であり，α は q と SLP 表現の圧縮性能に関連した非負整を表しており，$N_\alpha = N-\alpha \leqq qn$ が成り立つ．この成果は，SLP を対象としたアルゴリズムが理論的に非圧縮文字列を対象にしたアルゴリズムに劣らず，圧縮表現のサイズが小さいときにはより高速であることを示している．計算機実験におい

て,新しいアルゴリズムはほとんどの場合に前述のアルゴリズムよりも高速であること,また,q が大きい場合でも非圧縮のテキスト文字列から q-グラム頻度を求める線形時間アルゴリズムと比べて大幅に遅くなることはなく,場合によってはより高速になることを示した.q-グラム頻度は応用によっては重複する出現を数えないことが望ましい場合がある.この q-グラム非重複頻度問題に対しては,$O(q^2n)$ 時間 $O(qn)$ 領域アルゴリズムを開発し,$q>2$ の場合の初めての多項式時間アルゴリズムを与えた(論文⑧).

(3) SLP で表現されたテキストに対する畳み込み計算問題を解く高速なアルゴリズム:
畳み込み計算は,近似文字列の照合に応用され,高速フーリエ変換(FFT)を用いることで長さ m のパターン文字列と長さ N のテキスト文字列の間の畳み込みは $O(N \log m)$時間で計算できることが知られている.本研究では,長さ m のパターン文字列と大きさ N' の木に対する畳み込み計算が $O(N' \log m)$ 時間で計算できることを示した.この結果を研究成果の(2)で得られるトライ木に応用することにより,パターンと長さ N の文字列との畳み込みの時間計算量を従来の $O(N \log m)$ 時間から $O(N + N_\alpha \log m)$ 時間に改善することに成功した(論文③).

(4) 圧縮表現変換のための効率的なアルゴリズム:
SLP で表現された文字列を陽に展開することなく,その文字列の LZ78 分解を求めるアルゴリズムを提案した.この成果により,LZ78 圧縮を用いた正規化圧縮距離 (NCD) の計算,ひいては NCD に基づくデータの分類などを,圧縮表現のまま効率良く行うことを可能となる(論文⑤).更に,連長圧縮表現された文字列を LZ78 分解する効率的なアルゴリズム,また,逆に SLP 表現から連長圧縮表現への変換を高速に行うアルゴリズムを考案した(論文②).

(5) 圧縮文字列上の繰り返し構造発見問題を解く効率的なアルゴリズム:
SLP で表現された文字列が,連続して 2 回出現する部分文字列 (square) を含むかどうかを検証する多項式時間アルゴリズムを考案した(論文⑥).

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 12 件）

① Keisuke Goto and Hideo Bannai, Simpler and Faster Lempel Ziv Factorization, Proc. Data Compression Conference 2013 (DCC 2013), 133-142, 2013.

② Yuya Tamakoshi, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, From Run Length Encoding to LZ78 and Back Again, Proc. Data Compression Conference 2013 (DCC 2013), 143-152, 2013.

③ Toshiya Tanaka, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, Computing convolution on grammar-compressed text, Proc. Data Compression Conference 2013 (DCC 2013), 451-460, 2013.

④ Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, Fast q-gram mining on SLP compressed strings, Journal of Discrete Algorithms, 18:89-99,2013.

⑤ Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, Efficient LZ78 factorization of grammar compressed text, Proc. 19th International Symposium on String Processing and Information Retrieval (SPIRE 2012), LNCS 7608:86-98, 2012.

⑥ 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, 112(19):711-714, 2012.

⑦ Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, Masayuki Takeda, Speeding up q-gram mining on grammar-based compressed texts, Proc. 23rd Annual Symposium on Combinatorial Pattern Matching (CPM 2012), LNCS 7354:220-231, 2012.

⑧ Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, Masayuki Takeda, Computing q-gram Non-overlapping Frequencies on SLP Compressed Texts, Proc. 38th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2012), LNCS 7147:301-312, 2012.

⑨ Shunsuke Inenaga, Hideo Bannai, Finding Characteristic Substrings from Compressed Texts, International Journal of Foundations of Computer Science, 23(2):261-280, 2012.

⑩ Keisuke Goto, Hideo Bannai, Shunsuke

Inenaga, Masayuki Takeda, Fast q-gram Mining on SLP Compressed Strings, Proc. 18th International Symposium on String Processing and Information Retrieval (SPIRE 2011), LNCS 7024:278-289, 2011.

⑪ Takanori Yamamoto, Hideo Bannai, Shunsuke Inenaga, Masayuki Takeda, Faster Subsequence and Don't-Care Pattern Matching on Compressed Texts, Proc. 22nd Annual Symposium on Combinatorial Pattern Matching (CPM 2011), LNCS 6661:309-322, 2011.

⑫ Kazuaki Kashihara, Kohei Hatano, Hideo Bannai, Masayuki Takeda, Sparse Substring Pattern Set Discovery using Linear Programming Boosting, Proc. 13th International Conference on Discovery Science (DS 2010), LNAI 6332:132-143, 2010.

〔学会発表〕(計 9件)

① Keisuke Goto and Hideo Bannai, Simpler and Faster Lempel Ziv Factorization, Data Compression Conference 2013 (DCC 2013), 2013年3月, Snowbird, USA.

② 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), 2013年3月, Snowbird, USA.

③ Toshiya Tanaka, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, Computing convolution on grammar-compressed text, Data Compression Conference 2013 (DCC 2013),2013年3月, Snowbird, USA.

④ Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda, Efficient LZ78 factorization of grammar compressed text, 19th International Symposium on String Processing and Information Retrieval (SPIRE 2012), 2012年10月, Cartagena, Colombia.

⑤ Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, Masayuki Takeda, Speeding up q-gram mining on grammar-based compressed texts, 23rd Annual Symposium on Combinatorial Pattern Matching (CPM 2012), 2012年7月, Helsinki, Finland.

⑥ Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, 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), 2012年1月, Špindlerův Mlýn, Czech Republic.

⑦ Keisuke Goto, Hideo Bannai, Shunsuke Inenaga, Masayuki Takeda, Fast q-gram Mining on SLP Compressed Strings, 18th International Symposium on String Processing and Information Retrieval (SPIRE 2011), 2011年10月, Pisa, Italy.

⑧ Takanori Yamamoto, Hideo Bannai, Shunsuke Inenaga, Masayuki Takeda, Faster Subsequence and Don't-Care Pattern Matching on Compressed Texts, 22nd Annual Symposium on Combinatorial Pattern Matching (CPM 2011), 2011年6月, Palermo, Italy.

⑨ Kazuaki Kashihara, Kohei Hatano, Hideo Bannai, Masayuki Takeda, Sparse Substring Pattern Set Discovery using Linear Programming Boosting, 13th International Conference on Discovery Science (DS 2010), 2010年10月, Canberra, Australia.

〔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・システム情報研究院・准教授
研究者番号：20323644

(2)研究分担者
(　　)
研究者番号：

(3)連携研究者
(　　)
研究者番号：