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　３０　年　　６　月　　７　日現在

機関番号：　１７１０４
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2017
課題番号：　２６２８００８８

研究課題名（和文）圧縮情報処理によるストリームデータからの知識発見

研究課題名（英文）Knowledge discovery from compressed stream data

研究代表者
坂本　比呂志（Sakamoto, Hiroshi）

九州工業大学・大学院情報工学研究院・教授

研究者番号：５０３１５１２３

交付決定額（研究期間全体）：（直接経費）　　10,300,000円

研究成果の概要（和文）：本研究の目的は，圧縮情報処理をハードウェア上で実現し，ストリームデータをリアルタイム処理しながら知識発見に繋がる新しい仕組みを生み出すことである．これまでの手法では圧縮率とメモリサイズのトレードオフの影響により，データサイズによっては実行できないほどメモリを消費するか，断片的な入力に対する圧縮しかできなかった．本研究はこのトレードオフをこれまでで最小に抑えることに成功し，その結果，圧縮･転送・復号によるネットワークのスループット拡大に成功した．これらの成果は，関連技術の出願，論文発表，開発したハードウエアの展示会でのデモンストレーション等で社会へ発信している．

研究成果の概要（英文）：The aim of this research is to produce a novel technology for stream data compression realizing the real time processing of stream data to obtain deep knowledge hidden big data stream. In the previous studies, because of the tradeoff between compression ratio and memory consumption, either algorithm requires so huge memory to load whole input or input data must be decomposed into a sequence of small segments so that the algorithm can load it into the restricted memory. In this proposal, we can minimize the nonnegligible tradeoff, and we can expand the throughput of the network constructed by the compression-transmission-decompression algorithm implemented by FPGA. We propose and demonstrate these outcomes to the real world via patent applications, research articles, and industrial demonstrations.

研究分野：　データ圧縮

キーワード：　ストリームデータ　圧縮伝送　FPGA　可逆圧縮

１．研究開始当初の背景
データ圧縮によって知的情報処理が新展開を迎えている．ネットワーク上のストリームデータは増加する一方であるが，ハードウェアの性能向上は限界に近づいている．ネットワークを流れる多様で大量のデータは今後ますます増加し，それらのデータから重要な情報を素早く発見することが求められる．しかし，そのようなデータサイズの増加に対し，ハードウェアの性能の増加はほとんど止まっているに等しく，このような問題を解決するアルゴリズムが必要である．

２．研究の目的
あまりにも巨大なテキストは，読むことができないデータとほぼ同じであり，このようなデータの洪水に立ち向かうための次世代基盤技術の確立が急務である．本研究は，新しい可逆圧縮の理論を武器にこの問題に取り組み，圧縮情報処理による大量データからのリアルタイム知識発見を実現する．具体的には，極めて小さい遅延時間で圧縮，伝送，復号が可能なストリーム圧縮アルゴリズムを開発し，ハードウェア上に実装することで，ネットワークの物理的限界を超えたスループットを達成する．そして，このアルゴリズムを高密度画像圧縮へ拡張し，文字列と動画像ストリームからの高精度イベント抽出の実現によって圧縮情報処理技術の普及と標準化を目指す．

３．研究の方法
本課題では，申請者がこれまでに開発した圧縮アルゴリズムを基盤として，様々なストリーム圧縮へ適用可能な応用を目指し，当初の予定では，研究期間中の主な研究課題は，ストリーム圧縮理論の構築，ストリーム圧縮器の実装，ソーシャルメディアへの拡張，データ収集と実証実験からなっていた．しかし，研究途中で当初予期しなかった新しい知見が得られたため，ソーシャルメディアへの拡張と実証実験については中断し，新しい知見基づいて新たな研究計画を策定し，最終年度前年度申請を行った．その結果，新しい申請仮題が採択されたため，この成果報告では，最終年度前年度申請前までの成果についてまとめる．本研究で推進した研究課題は以下の２項目である．
【ストリーム圧縮理論の構築】
データを読み込みながら処理する通常のオンライン圧縮と本研究の課題であるストリーム圧縮の違いはレイテンシ(遅延時間)にある．オンラインの枠組みでは，出力の遅延時間が O(1) で抑えられればよいが，本研究のストリーム圧縮では，遅延時間が１クロック(1CPU 時間) 程度に収まることを目指す．これを実現するためには，シンプルで高速に動作するアルゴリズムを FPGA 上に実装しなければならないが，本研究では文法圧縮と呼ばれる圧縮モデル上でのストリーム圧縮を行う．ここで，構文木によって文字列を表現し，同じ部分木の繰り返しを除去する圧縮法を文法圧縮と呼ぶ．この構文木は文字を固定長で表現するため，符号の最適化ためにはさらに文字を可変長符号化しなければならない．FPGA の試作機ではこの可変長符号化は未達成であり，これを可能にすることで情報理論的に最適な圧縮を達成できる．
【ストリーム圧縮器の実装】
ゲート数やメモリの制限により，既存アルゴリズムの複雑なデータ構造はハードウェア上では実現困難である．本研究では CAM(Content Addressable Memory) のみによる単純な論理回路によってソフトウェアと同等の圧縮率を達成する．CAM はハードウェア上のメモリであり，大規模に搭載することが難しい．したがって，圧縮処理に必要な途中の記憶をすべて保存することができないため，頻度の高い情報を残し，そうでないものを捨てるなど情報を適切に更新しなければならない．このようにして高頻度のパターンを優先的に処理することで圧縮率が高まることが知られている．高頻度パターンは変換テーブルによって記憶するが，頻度は入力に応じて変化する．本研究では，ストリーム上で変換テーブルを動的に更新する(試作機では静的テーブル)．この機能の実現よって，変換テーブル自身は伝送不要となり，高速伝送に大きく貢献する．
【メディア情報からのリアルタイムパターン抽出】(この研究課題は最終年度前年度申請のため新規課題に延期)
観測されたデータからイベントを抽出するとき，バックエンドで特徴抽出や比較が行われる．このとき，できるだけ高密度の情報を圧縮伝送することで，現象の理解をより確実に行うことができる．さらに，前記の技術を拡張し，SNS などのテキストや位置情報と統合することで，画像とテキストの連携による知識発見を可能にする．

４．研究成果
これまでに【ストリーム圧縮理論の構築】および【ストリーム圧縮器の実装】のそれぞれの項目に当てはまる以下のような成果を上げた．当該研究の目的は，圧縮情報処理をハードウェア上で実現し，ストリームデータをリアルタイム処理しながら知識発見に繋がる新しい仕組みを生み出すことである．これまでの手法では圧縮率とメモリサイズのトレードオフの影響により，データサイズによっては実行できないほどメモリを消費するか，細切れ入力を圧縮することを余儀なくされる．本研究はこのトレードオフをこれまでで最小に抑えることに成功し，その結果，圧縮･転送・復号によるネットワークのスループット拡大に成功した．これまでに関連技術の出願（特開 2014，特願 2015）や論文発表，展示会でのデモンストレーションを行っている．これらの成果は，IEEE 関連の国際

会議における最優秀論文賞(BPOE2015)や国内の組込み総合技術展特別賞（ET/IoT Tech. 2015）などで高く評価されている．また，研究分担者（筑波大）が大学発ベンチャー（ストリームテクノロジ(株)）を起業し，技術の普及を目指しており，研究代表者は技術顧問として新製品の開発に貢献している．また，ソフトウェアとしても，世界初のオンライン型圧縮索引を実装して人工知能学会２０１３年度研究会優秀賞を受賞している．なお，出願した特許はそれぞれ 2017 年 7 月 7 日および同年 12 月 15 日に登録された．

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　4　件）

K. Marumo, S. Yamagiwa, R. Morita, H. Sakamoto. Lazy Management for Frequency Table on Hardware-based Stream Lossless. Information 7(4)63(16 pages), 2016

Y. Takabatake1, K. Nakashima, T. Kuboyama, Y. Tabei, H. Sakamoto. siEDM: an efficient string index and search algorithm for edit distance with moves. Algorithms 9(2)26(18 pages), 2016

Yoshimasa Takabatake, Yasuo Tabei, Hiroshi Sakamoto. Online Pattern Matching for String Edit Distance with Moves. Lecture Notes in Computer Science 8799:203-214, 2014.

Yoshimasa Takabatake, Yasuo Tabei, Hiroshi Sakamoto. Improved ESP-index: A Practical Self-index for Highly Repetitive Texts. Lecture Notes in Computer Science 8504: 338-350, 2014.

〔学会発表〕（計　20　件）

S. Fukunaga, Y. Takabatake, T. I, H. Sakamoto. Online Grammar Compression for Frequent Pattern Discovery. ICGI 2016, Delft, The Netherlands, from October 5 through October 7, 2016.

S. Iwasaki, Y. Takabatake, T. Kuboyama, H. Sakamoto. Finding Frequent Patterns from Stream Data with Small Space. SISA2016, 2016 年 09 月 14 日～2016 年 09 月 17 日, タイ王国, アユタヤ市

福永祥平, 坂本比呂志. 文法圧縮における逆引き辞書の省スペース化. 第 100 回 SIG-FPAI 研究会, 2016 年 03 月 27 日～2016 年 03 月 28 日, 熊本市

徳永啓太郎, 坂本比呂志. 定数領域の頻度計算を用いたオンライン文法圧縮アルゴリズム. 第 100 回 SIG-FPAI 研究会, 2016 年 03 月 27 日～2016 年 03 月 28 日, 熊本市

大西孝典, 坂本比呂志. 文法圧縮を用いた類似度計算の大規模データへの適用. 第 100 回 SIG-FPAI 研究会, 2016 年 03 月 27 日～2016 年 03 月 28 日, 熊本市

水野仁人,高畠嘉将,坂本比呂志. 文法圧縮のハッシュ領域の削減. 第99回SIG-FPAI研究会, 2016 年 01 月 21 日～2016 年 01 月 22 日, 仙台市

青山友紀, 高畠嘉将,坂本比呂志. ストリームデータからの頻出パターンの近似発見. 第 99 回 SIG-FPAI 研究会, 2016 年 01 月 21 日～2016 年 01 月 22 日, 仙台市

Shinichi Yamagiwa, Yoshinobu Kawahara, Noriyuki Tabuchi, Yoshinobu Watanabe and Takeshi Naruo. Skill Grouping Method: Mining and Clustering Skill Differences from Body Movement BigData. BigData 2015, 2015 年 10 月 29 日～2015 年 11 月 01 日, Santa Clara, CA, USA

Yoshimasa Takabatake, Yasuo Tabei, Hiroshi Sakamoto. Online Self-Indexed Grammar Compression. SPIRE 2015. 2015 年 09 月 01 日～2015 年 09 月 04 日, London, UK

Shinichi Yamagiwa, Koichi Marumo, Hiroshi Sakamoto. Stream-Based Lossless Data Compression Hardware Using Adaptive Frequency Table Management. BPOE 2015, 2015 年 08 月 31 日～
2015 年 09 月 04 日, Kohala, HI, USA

高畠嘉将, 坂本比呂志. 文法圧縮のための逆引き辞書の省スペース化. 第 98 回 SIG-FPAI 研究会, 2015 年 08 月 07 日～2015 年 08 月 08 日, 和歌山市

高畠嘉将, 中島健太, 田部井靖生, 坂本比呂志. siEDM: 移動付き編集距離の為の効率的な索引. 第 98 回 SIG-FPAI 研究会, 2015 年 08 月 07 日～2015 年 08 月 08 日, 和歌山市

岩崎暁, 坂本比呂志. ストリーム中の頻出アイテムの発見のための Simple Algorithm の高速化. 第 98 回 SIG-FPAI 研究会, 2015 年 08 月 07 日～2015 年 08 月 08 日, 和歌山市

中島 健太, 前田 幸司, 高畠 嘉将, 坂本 比呂志. 移動付き編集距離に基づく曖昧検索が可能な圧縮索引. 第 97 回 SIG-FPAI 研究会, 2015 年 03 月 22 日～2015 年 03 月 23 日, 別

府国際コンベンションセンター

高畠 嘉将, 田部井 靖生, 坂本 比呂志. 移動付き編集距離のオンラインパターンマッチング. 第 97 回 SIG-FPAI 研究会, 2015 年 03 月 22 日～2015 年 03 月 23 日, 別府国際コンベンションセンター

高畠 嘉将, 田部井 靖生, 坂本 比呂志. 文法圧縮に基づく自己索引のオンライン構築について. 第 96 回 SIG-FPAI 研究会, 2015 年 01 月 13 日～2015 年 01 月 14 日, 名古屋工業大学

前田幸司, 高畠嘉将, 田部井靖男, 坂本比呂志. 文法圧縮を応用したハミング距離計算の高速化. 大 95 回 SIG-FPAI 研究会. 2014 年 10 月 10 日～2014 年 10 月 10 日, 大阪大学産業科学研究所

Hiroshi Sakamoto. Grammar Compression: Grammatical Inference by Compression and Its Application to Real Data. 12th International Conference on Grammatical Inference（招待講演）, 2014 年 09 月 17 日～2014 年 09 月 19 日, Kyoto University

宮木 亮洋, 坂本 比呂志. 乱択アルゴリズムを用いた頻出文字列の近似数え上げに基づく省スペース文法圧縮. 第 94 回 SIG-FPAI 研究会, 2014 年 07 月 24 日～2014 年 07 月 24 日, 根室市総合文化会館

前田幸司, 高畠嘉将, 田部井靖生, 坂本比呂志. 文法圧縮を応用したハミング距離の短い文字列列挙アルゴリズム. 第 94 回 SIG-FPAI 研究会, 2014 年 07 月 24 日～2014 年 07 月 24 日, 根室市総合文化会館

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計　2　件）
名称：データ圧縮･解凍システム、データ圧縮方法及びデータ解凍方法、並びにデータ圧縮器及びデータ解凍器
発明者：坂本比呂志、山際伸一
権利者：同上
種類：特許
番号：PCT/JP2016/59372
出願年月日：2016 年 03 月 24 日
国内外の別：国外

名称：データ圧縮･解凍システム、データ圧縮方法及びデータ解凍方法、並びにデータ圧縮器及びデータ解凍器
発明者：坂本比呂志、山際伸一
権利者：同上
種類：特許
番号：特願 2015-063449
出願年月日：2015 年 03 月 25 日
国内外の別：国内

○取得状況（計　2　件）
名称：データ圧縮･解凍システム、データ圧縮方法及びデータ解凍方法、並びにデータ圧縮器及びデータ解凍器
発明者：坂本比呂志、山際伸一
権利者：同上
種類：特許
番号：特許第 6256883
取得年月日：2017 年 12 月 15 日
国内外の別：国内

名称：データ圧縮器及びデータ解凍器
発明者：坂本比呂志、山際伸一
権利者：同上
種類：特許
番号：特許第 6168595
取得年月日：2017 年 7 月 7 日
国内外の別：国内

〔その他〕
ホームページ等
http://www.donald.ai.kyutech.ac.jp/~hiroshi/index.html

６．研究組織
(1)研究代表者
坂本 比呂志（SAKAMOTO, Hiroshi）
九州工業大学・大学院情報工学研究院・教授
研究者番号：50315123

(2)研究分担者
山際 伸一（YAMAGIWA, Shinichi）
筑波大学・システム情報系・准教授
研究者番号：10574725

(3)研究分担者
榎田 修一（Enokida, Shuichi）
九州工業大学・大学院情報工学研究院・教授
研究者番号：40346862

(4)連携研究者
田部 井靖生（TABEI, Yasuo）
理化学研究所・革新知能統合研究センター・ユニットリーダー
研究者番号：20589824

(5)研究協力者
なし