様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　　９　日現在

機関番号：　１７１０４
研究種目：　若手研究(A)
研究期間：　2011～2014
課題番号：　２３６８００１６

研究課題名（和文）圧縮マイニング：超大規模テキストに埋もれている知識の顕在化

研究課題名（英文）Compressed Mining: exposing hidden knowledge in large-scale 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）　　15,100,000円

研究成果の概要（和文）：本研究は，データ圧縮によって巨大テキストの俯瞰を可能にし，気づかれずに埋もれている知識を顕在化する圧縮マイニングを実現する．具体的には，これまでに申請者が開発した，テキスト中のパターンの関係を保存しながら圧縮する技術をマイニングに応用することで，巨大テキスト同士の直接比較を可能にする．今年度（最終年度）は，前年度までに開発した手法を実装し，様々な分野の研究者にプログラムを配布し，基盤研究（B)として共同研究をスタートした．

研究成果の概要（英文）：This study focuses on exposing hidden knowledge in large-scale data by the data compression technique. Concretely, we propose a novel method for directly comparing of large texts using our algorithm which allow us to compress data keeping the relationship among frequent patterns. By this, we start a new research project as the Scientific Research (B).

研究分野：　知能情報学

キーワード：　データ圧縮　簡潔データ構造　ストリームデータ　文法圧縮

１．研究開始当初の背景

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，データ圧縮によって巨大テキストの俯瞰を可能にし，気づかれずに埋もれている知識を顕在化する圧縮マイニングを実現する．具体的には，これまでに申請者が開発した，テキスト中のパターンの関係を保存しながら圧縮する技術をマイニングに応用することで，GB 超～TB クラスの巨大テキスト同士の直接比較を可能にする．そして，これまでは歯が立たなかった超大規模テキストから知識を掘り起こし，まとまりごとに再構成することで知識を顕在化する．最終的には開発した手法の実世界応用をめざし，プログラムの公開を含めて成果を社会に向けて発信する．

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に対して曖昧検索と部分構造抽出を可能とする理論拡張を行い，圧縮マイニングとして定式化し，実世界への応用を目指す．この研究は，基礎理論の構築，アルゴリズムの実装，実世界への応用からなっている．

【基礎理論の構築】

・木分解による構造索引の構築：グラフ上のさまざまな探索問題を木構造における単純な演算(頂点の比較やランク計算)に置き換えることで圧縮マイニングの基礎理論を構築する．

・簡潔データ構造のデータ圧縮への応用: 簡潔データ構造のテクニックを利用してポインタなどの冗長なデータ構造を使わない圧縮索引を実現する．

【アルゴリズムの実装】

木構造の類似性判定アルゴリズムのグラフ構造への応用: 木構造の類似性を高速に計算する近似アルゴリズムを応用することで，グラフ構造の類似性を判定する近似アルゴリズムを設計する．従来のグラフマイニングでは，グラフの部分構造を取り出すことで類似性を計算していたが，本研究では，ノードのラベルやノード間の先祖子孫関係などの単純な比較のみで類似性を近似計算する

【実世界への応用と情報発信】

剽窃検出: テキストデータがネットワークによって広く公開されることで剽窃や著作権侵害が問題となっている．これを検知する市販のソフトウエアなどがあるが，アルゴリズムとデータ構造の最先端の技術により，本研究は規模や精度でそれらを大幅に上回ることを目指す．

４．研究成果

【基礎理論の構築】

申請者らが本研究を開始前に開発した大規模データ圧縮アルゴリズムをストリームデータに対するオンラインアルゴリズムに拡張し，同時に，このアルゴリズムの性能を飛躍的に高めるための着想を得た．この研究を進めた結果，入力データ全体をメモリに読み込むことなく大規模データの効率的な圧縮が可能となり，この理論を足がかりに，テキスト中のパターンの関係を保存しながら圧縮する技術によってパターンマイニングに応用することが可能となった．この手法を引き続き洗練し，構築したデータ構造とアルゴリズムに高速・軽量の照合技術を組み合わせることで圧縮マイニングの基本的枠組みを完成させた．さらに，ネットワーククラスタリングへの応用として，申請者がデータ圧縮と平行して進めているネットワークマイニングの成果を応用して大規模グラフデータからのパターン獲得に応用できることを予備的な実験によって示した．

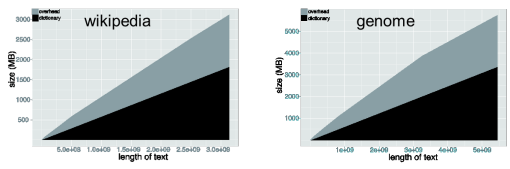
【アルゴリズムの実装】

基礎理論で構築したアルゴリズムとデータ構造を元に，大規模データへ適用可能な実装を行った．下記の表およびグラフは開発したアルゴリズムの性能であり，入力データに対して消費メモリが非常に少ないことを示している．さらに，このアルゴリズムを応用し，大規模データから曖昧検索の実現を可能にした．これまでの基本的な枠組みでは，圧縮データからパターン検索には曖昧な検索ができない．現在は，このアルゴリズムは，部分的な一致を検出することで曖昧な検索が可能となっている．今回の成果によって，これまでは困難であった，GB 超～TB クラスの巨大テキスト同士の直接比較が可能になる．そして，これまでは歯が立たなかった超大規模テキストから知識を掘り起こし，まとまりごとに再構成することが可能となった．

Description of data

source	size (MB)	length	#alphabet
wikipedia (en)	5,533	5,442,222,932	209
genome	3,199	3,137,162,547	38

Memory usage of VLD and overhead of hashtable



【実世界への応用と情報発信】
さらに本研究では，開発したアルゴリズムを実際の巨大データに応用可能であることを示した．特に，大規模ストリームデータ処理を限られたメモリ上で実現するための新しい手法を開発し，理論及び実験の両方でその有効性を確認した．この成果は複数の国際会議において発表し，高い評価を得た．ゲノムデータや twitter など幅広い実世界データに対して，本手法の有効性を確認した．この成果は，ビッグデータ専門の国際会議で採択され，評価を受けた．また，本研究課題は，4年間の計画であったが，最終年度の前年度に基盤研究（B)に採択されたため，引き続き発展的課題について取り組むこととなり，社会的にインパクトのある応用を目指して現在研究中である．

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）

① "Scalable Detection of Frequent Substrings by Grammar-Based Compression",
M.Nakahara, S.Maruyama, T.Kuboyama, H.Sakamoto,
IEICE Trans. on Information and Systems, E96-D(3):457-464 (2013). **査読有り**

②"ESP-Index: A Compressed Index Based on Edit-Sensitive Parsing",
S.Maruyama, M.Nakahara, N.Kishiue, H.Sakamoto,
Journal of Discrete Algorithms 18:100-112 (2013), Elsevier. **査読有り**

③"An Online Algorithm for Lightweight Grammar-Based Compression",
S. Maruyama, H. Sakamoto, M. Takeda,
Algorithms 5(2):214-235 (2012-4). **査読有り**

④"Extracting research communities from bibliographic data",
Y.Nakamura, T.Horiike, T.Kuboyama, H.Sakamoto,
KES Journal 16(1): 25-34 (2012-1), IOS Press. **査読有り**

〔学会発表〕（計 10 件）

① "Scalable Pattern Discovery on Knot Theory",
Y. Takabatake, T. Kuboyama, A. Yasuhara, H. Sakamoto,
Workshop on Data Discretization and Segmentation for Knowledge Discovery (DDS2013), Keio University, October 27th – 28th, 2013.

② "An Implementation of Truss Decomposition of Bipartite Graph",
Y. Li, T. Kuboyama, H. Sakamoto,
Workshop on Data Discretization and Segmentation for Knowledge Discovery (DDS2013), Keio University, October 27th – 28th, 2013.

③"A Reconfigurable Stream Compression Hardware based on Static Symbol-Lookup Table",
S. Yamagiwa, H. Sakamoto,
The First Workshop on Benchmarks, Performance Optimization, and Emerging hardware of Big Data Systems and Applications (BPOE 2013), 86-93, October 8, 2013, Silicon Valley, CA, USA.

④"Fully-Online Grammar Compression",
S. Maruyama, Y. Tabei, H. Sakamoto, K. Sadakane,
20th International Symposium on String Processing and Information Retrieval (SPIRE2013), 218-229, October 7-9, Jerusalem, Israel.

⑤"A Succinct Grammar Compression",
Y. Tabei, Y. Takabatake, H. Sakamoto,
24th Annual Symposium on Combinatorial Pattern Matching (CPM2013), 235-246, Bad Herrenalb, Germany, June 17–19, 2013.

⑥ "Variable-Length Codes for Space-Efficient Grammar-Based Compression",
Y. Takabatake, Y. Tabei, H. Sakamoto,
19th International Symposium on String Processing and Information Retrieval (SPIRE2012), 398-410, Cartagena de Indias, Colombia, October 21-25, 2012.

⑦ "Grammar-Based Compression for Frequent Pattern Mining",
M. Nakahara, S. Maruyama, T. Kuboyama, H. Sakamoto,
Second Workshop on Algorithms for Large-Scale Information Processing in Knowledge Discovery (ALSIP 2011), Takamatsu, Kagawa, Japan.
December 1-2, 2011.

⑧ "Scalable Detection of Frequent Substrings by Grammar-Based Compression",
M.Nakahara, S.Maruyama, T.Kuboyama, H.Sakamoto,
The 14th International Conference on

Discovery Science (DS 2011), 236-246, Espoo, Finland, 5-7 October, 2011.

⑨"ESP-Index: A Compressed Index Based on Edit-Sensitive Parsing",
S.Maruyama, M.Nakahara, N.Kishiue, H.Sakamoto,
18th International Symposium on String Processing and Information Retrieval (SPIRE2011), 398-409, Pisa, Italy, October 17-21, 2011.

⑩"An Online Algorithm for Lightweight Grammar-Based Compression",
S.Maruyama, T.Takeda, M.Nakahara, H.Sakamoto,
1st International Conference on Data Compression, Communication, and Processing (CCP2011), 19-28, June 21-24, 2011 King's Residence Hotel Palinuro (Cilento Coast), Italy.

〔図書〕（計　1　件）
A chapter in "Multimedia Services in Intelligent Environments", H. Sakamoto, T.Kuboyama, Springer, 2013.
ISBN: 978-3-319-00371-9

〔産業財産権〕
○出願状況（計　0　件）

○取得状況（計　0　件）

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

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
なし

(3)連携研究者
久保山哲二（KUBOYAMA, Tetsuji）
学習院大学・計算機センター・教授
研究者番号：80302660