科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１８　日現在

機関番号：　１７１０４
研究種目：　挑戦的萌芽研究
研究期間：　2011～2013
課題番号：　２３６５００７４

研究課題名（和文）データ圧縮はテキストを要約できるか？

研究課題名（英文）Can data compression algorithm do abstraction?

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）　　2,800,000円、（間接経費）　　840,000円

研究成果の概要（和文）：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には、次世代情報基盤技術の確率が急務である。本研究は、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超の巨大テキスト同士の直接比較を可能にし、超大規模テキストからの知識発見を実現する。

研究成果の概要（英文）：When data is too big to read by machine, such data is useless. For such huge data flood, we need to develop the next generation information technology. In this study, adopting our technique for directly comparing texts, we propose an efficient algorithm of knowledge discovery for huge text data beyond GB.

研究分野：　複合領域

科研費の分科・細目：　知能情報学

キーワード：　データ圧縮　データマイニング　簡潔データ構造　グラフ構造

１．研究開始当初の背景

ネットワークを流れるデータの氾濫によって，近年データ圧縮が再び注目されている．例えば、インターネット上の画像共有サイトでは，差分記憶によって容量を削減することでストレージコストを抑えたり，科学技術の分野では，センサデータや遺伝子データなどのやりとりは記録媒体の郵送に頼っているため，データを圧縮して作業を効率化している．また，テキストデータに限っても，検索エンジンの精度低下や剽窃によるテキストデータの二次利用などが問題となっている．このように，データ圧縮のみならず，そこに新しい付加価値を付けるための高度情報処理の研究が盛んに行われ，ICDM，WWW, SIGMOD 等の主要国際会議で発表が行われている．一方で申請者は，データ圧縮の理論(Sakamoto et al. CPM ’03, SPIRE ’04)，省メモリな圧縮照合への応用(Sakamoto J. Discrete Algorithms ’05) および有効性の検証(Sakamoto et al. SPIRE ’08) を進めて，現在は圧縮データ長いパターンを検索するための手法(Sakamoto et al. IEICE Trans. ’09) に取り組んでいる．本研究は、これらの研究成果を発展させて、氾濫しているデータを有効利用するための技術を目指す。

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，データ圧縮を要約するための技術として発展させることで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発掘する．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し，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のまとまりを再構成する．

テキストの洪水に立ち向かう技術として，キーワード検索によって網羅的に得られた情報を再構築し，それらが大まかに意味するものを端的に提示する高度な情報処理が必要である．このとき，一次的な検索結果全体をパターンとして再検索することで，情報のフィードバックが掛かり，データ間の深い関連性を見いだせると期待できる．そこで本研究ではデータ圧縮による情報の要約のための枠組みを提案し，その有用性を実証することで，データ圧縮に新しい価値を見いだそうとしている．また，本研究の成果を一般に広く周知するため，プログラムソースの公開と可視化ツールの作成・配布を計画している．以上をまとめると，(A) 大規模テキストを圧縮によって直接比較する手法の開発および (B) システムの実装と実世界データでの実証実験および情報発信が焦点である．

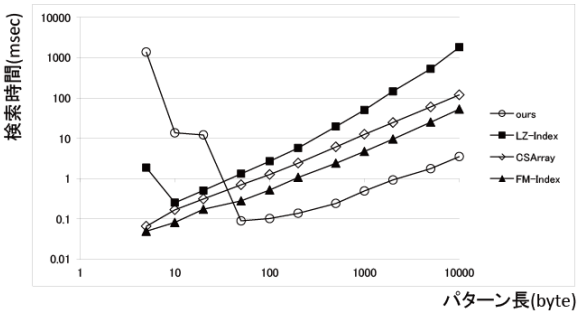
３．研究の方法

本研究は，計算機科学における基本問題であるデータ圧縮に新しい価値を付加しようとしている．これまでの研究は，(1) サイズを小さくする，(2) 圧縮データを高速照合する，(3) 圧縮データを索引化する，(4) 圧縮データから類似性を求める，などに大別される．このうち本研究と関連が深いテーマは(3)と(4) であるが，(3) は圧縮されたテキストから短いパターンの出現を検出することが目的であるため，テキストとパターンを特に区別せず圧縮データ同士を高速比較する本研究の要請をクリアできない．また(4) では，データを不可逆な数値にマッピングするため，まったく異なるデータ同士が偶然近い値になることが避けられない．さらに，圧縮データのどの部分が類似しているかを取り出して提示することも困難である．このように，データ圧縮は情報の氾濫を緩和するために有効な手段である反面，取り扱いが難しく解決すべき多くの困難性を持っている．

これらの問題は，データ圧縮がテキストの内容をシャッフルしてしまうことに起因する．申請者は，この問題を回避するために，テキスト中のパターンを崩さないようにうまく符号化する手法を考案した(Sakamoto et al. IEICE ’09 等)．本研究ではさらにそれを発展させ，時間やメモリなどの限られたリソースにおいてこれらの課題を解決し，超大規模テキストを俯瞰する技術の創出を目指している．本研究の基本アルゴリズムは，テキストデータをコンパクトなグラフ構造に変換することでよい圧縮を達成する．したがって，グラフや木構造に対して開発されてきたテクニックがそのまま利用可能である．申請者が開発した有向グラフ上の探索手法 (Sakamoto et al. SWORD ’07, AMT ’05, 信学会論文誌 ’08 等) もその一部であり，これらを圧縮データ上で活用することによって，ネットワーククラスタリングに対する新しい応用が開拓できる可能性が高い．

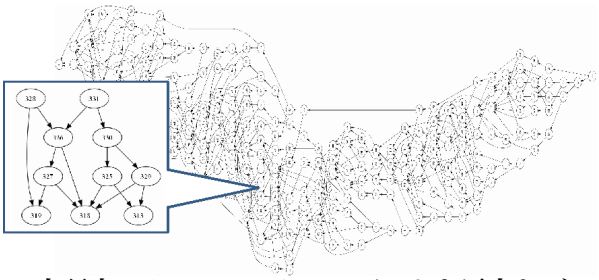
本研究は新しい原理に基づく圧縮索引を応用することで目標を達成する．図は，提案手法の基本アルゴリズムの性能を圧縮接尾辞(CSArray) 等の他の代表的な手法と比較した実験結果である．圧縮索引は約 200MB の英文テキストに対して構築され，そこからパターンを検索する時間を測定した．その結果，パターンが十分に長いときは極めて高速であることが確認できた．この特性は，圧縮手法に秘密があり，パターンの出現回数などの統計情報の復元性能を犠牲にしている代わりに，長いパターンの符号化に強いためである．この特性はテキストデータの種類によらず他言語，XML データ，ソースコード，DNA シーケンス等でも同様に観測された．これは他

の手法には見られない特性であり，この利点をさらに追求することで，現在は困難である共通構造や曖昧情報の抽出により，高度情報検索を支援する．具体的には，コピーによる著作権の侵害やオークションでの不正行為などの高精度検出，超大規模データのコンパクトな記録手法，論文・特許データからの思いがけない関連技術検索など，今後のネットワーク社会におけるキラーアプリケーションが見込まれる．



図および関連の実験結果が示すように，本研究の基本アルゴリズムはデータに関する背景知識を必要とせず，得られる結果も十分なパフォーマンスを持っている．したがって，本研究課題が提案するデータ圧縮による情報の要約技術は，ネットワーク上に日々蓄積され続けている利用困難なテキストデータのなかからこれまでは発見できなかった価値を見いだすことが期待できる．

４．研究成果



本研究によって、テキストデータを巨大なグラフ構造で表現したものをコンパクトに圧縮することで、図のような複雑なネットワークからデータの関係性を高速に取り出すことが可能となった。このグラフ構造は、実際のデータ（約１GB）を圧縮したものを可視化した結果である。本研究によって開発した技術は、このようなデータの俯瞰に応用できる。

さらに本研究では、本研究実施期間内に以下の成果を上げた。【基礎理論の構築】最終年度以前では、木構造の分解による索引構造の構築および簡潔データ構造のデータ圧縮への応用について理論の拡張を行った。この理論に基づき、最終年度では圧縮による情報の抽象化を定式化し、ネットワーク分析への応用について集中的に研究を行った。その結果、再規模グラフ構造からの知識を抽出する新しいアルゴリズムを構築した。

【アルゴリズムの実装】これまでの曖昧検索を圧縮データ上で実現した。またこのアルゴリズムを並列化し、より大規模データに適用可能とした。

【実世界応用と情報発信】これまでに開発した類似性判定や高速照合を用いて、最終年度では、文字列の類似度を高速に判定するオンラインアルゴリズムを開発し、ツイッターデータなどに適用してその規模耐性や有効性を確認した。以上の成果は、国際会議等で発表し、レビュワー等から高い評価を受けた。

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　　3 件）

① S.Yamagiwa, H.Sakamoto, A reconfigurable compression hardware based on static symbol-lookup table, 1st workshop on benchmarks, performance optimization, and emerging hardware of big data systems and applications, 86-93, 2014. 査読有り

② S. Maruyama, Y.Tabei, H.Sakamoto, K.Sadakane, Fully-online compression, 20th international symbosium on string processing and information retrieval, 218-229, 2013. 査読有り

③ Y.Tabei, Y.Takabatake,H.Sakamoto, A succinct grammar compression, 24th annual symposium on combinatorial pattern matching, 235-246, 2013. 査読有り

〔学会発表〕（計　　2 件）

① 前田幸司、髙畠嘉将、坂本比呂志、頻度情報に基づく省スペースなオンライン文法圧縮、第 92 回 SIGFPAI 研究会、2014 年 1 月 30 日、函館市民会館

② 髙畠嘉将、坂本比呂志、文法圧縮に基づく自己索引の省スペース化、第 90 回 SIGFPAI 研究会、2013 年 7 月 18 日、稚内日ロ友好会館

〔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
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