研　究　報　告　書

「透過的データ圧縮による高速かつ省メモリなビッグデータ活用技術の創出」

研究タイプ：通常型
研究期間：　平成25年10月～平成29年3月
研　究　者：　田部井　靖生

1．　研究のねらい

近年の様々な情報処理分野において，データは高度に大規模化・複雑化している.データは主に文字列，木，グラフで表現される．大規模化に伴いこのようなデータを従来の表現で取り扱うことは困難になりつつあり，データ中に潜む有意義な情報を抽出する手法の研究開発が，現代社会における緊急の課題となっている.

このような状況から，データから統計的な情報をもとにデータの背後に潜む規則を自動的に抽出するデータマイニング及び機械学習の研究が近年盛んに行われている．従来の研究に共通することは，　入力の規模をあまり考慮しないで研究が行われているために，近年の大規模化・複雑化するデータに必ずしも適応可能であるとは限らない.

透過的データ圧縮とは，データを圧縮した状態のまま定数時間でランダムアクセスを可能にする圧縮方式のことで，近年，文字列，木，グラフを含む様々なデータ表現のための透過的データ圧縮法が盛んに研究されている．代表的な応用例として ローマ字として入力されたひらがな単語を漢字に変換する Input Method Editor (IME)があり，IME では各ひらがな単語を漢字に対応付けるために大規模なトライとして表現される辞書が必要となる．　通常トライのポインターを用いた標準的な実装では，トライの各ノード当たり log(n)ビット(n:全ノード数)のメモリが必要となり，全ノードで nlog(n)ビットのメモリが必要で，数億からなる日本語の単語を登録するには適していない．代表的な IME である GoogleIME では，木の代表的な透過的データ圧縮法 LOUDS を用いることにより，トライの各ノードを 2 ビットで表現している．全単語を登録するためのトライのサイズはわずか 2n ビットである．さらにトライをコンパクトに表現することは，キャッシュアクセスによる高速化の役割も担っている．　このように透過的データ圧縮はデータ処理における省メモリ化と高速化の両方において有効であるにもかかわらず，これまでの研究では代表的なデータ構造の操作をいかに圧縮したまま実現するかに焦点があてられており，より高度な処理であるデータ中に潜む有意義な情報を抽出する処理に関しては手法が開発されていないのが現状である.

本研究のねらいは，複雑な構造を持つ膨大な量のデータから，その背後に潜む有益な規則を自動的に見つけ出す高速かつ省メモリなデータ処理手法を透過的データ圧縮法に基づき開発することである．本研究プロジェクトでは，バイオインフォマティクス，ケモインフォマティクスなどの分野に存在するビッグデータを対象として，普遍的なデータ処理課題を抽出し，その課題を数理的な立場から問題を定式化する．そして，ビッグデータ処理に適応可能な高速かつ省メモリなデータ処理手法を開発する.

2. 研究成果

(1)概要

本研究課題における成果は主に 3 つの領域に分けられる. (i) 大規模反復テキスト処理のための文法圧縮法, (ii) 大規模化合物データベースの類似度検索法, (iii) データ圧縮技術によるスケーラブルな機械学習法に分けられる. いずれの成果も最新のデータ圧縮技術に関するものである.

(2)詳細

研究テーマ A「大規模反復テキスト処理のための文法圧縮法」

文法圧縮とは, 与えられたテキストを一意に表現する文脈自由文法を構築することで圧縮する技術である. 文法圧縮は, 繰り返しの多いテキストに対して高い圧縮率を達成するが可能である. そのようなテキストの実例としては, DNA 配列やバージョン管理されたテキストなどがある.特に DNA 配列の個体間での違いは 0.1%ほどと言われており, 文法圧縮は DNA 配列に対して高い圧縮率を達成することが期待できるので, DNA 配列を処理するための有効な手段として期待されている. 本研究プロジェクトでは, Edit Sensitive Parsing (ESP)と呼ばれる文法圧縮アルゴリズムをテキストビッグデータ利活用のために, (i)スケーラブルなオンライン文法圧縮技術, (ii) 文法圧縮されたテキストの類似度検索, (iii) 文法圧縮されたテキスト上での rank, select, access 操作などの様々なテキスト処理アルゴリズムを開発した.

研究テーマ B「大規模化合物データベースの類似度検索技術」

大規模化合物データベースの類似度検索は, 創薬における重要なタスクである. データベース中の各化合物は, 0 または 1 を要素とするバイナリーベクトルで表現される. 化合物のバイナリーベクトル表現はフィンガープリントと呼ばれ, フィンガープリント表現の類似度検索技術も数多く提案されてきた. 近年, ディスクリプターと呼ばれる化合物の整数ベクトル表現が提案され注目を集めている. 代表的なディスクリプターに ringo や KCF-S があるが, ディスクリプター表現された化合物の類似度検索技術はまだないのが現状である. そこで, 本研究課題では, ディスクリプターとして表現された化合物データベースの類似度検索技術を wavelet 木などの透過的データ圧縮技術を応用することで開発した. 本手法により, 米国国立生物工学情報センターの化合物データベース PubChem 中の 4 千万化合物に対して, 高速に類似度検索を行うことを可能にした.

研究テーマ C「データ圧縮技術によるスケーラブルな機械学習法」

解釈可能な統計モデルを学習させることは, 大規模データから有益な情報を抽出するための有効な手段の一つである. 機械学習では, 学習データは特徴ベクトルを要素として持つデータ行列として表現され, データ行列を入力として統計モデルの学習が行われる. しかし, 学習データが大規模になる時, 学習には大量のメモリが必要になり統計モデルの学習が困難になる傾向がある. このような問題に対して, 我々はデータ圧縮されたデータ行列上でのスケーラブルな統計モデル学習法を開発した. 提案手法により約 100GB のデータ行列を 4GB にまで圧縮した状態で統計モデル学習が可能となった. 提案手法の応用として, 創薬におけるバーチャルスクリーニングの応用を行った. バーチャルスクリーニングは, 化合物とタンパク質の

相互作用予測として定式化され，大規模機械学習問題となる傾向にある．提案手法により大規模な化合物とタンパク質のデータからでも効率よくモデル学習が可能となることを示した．

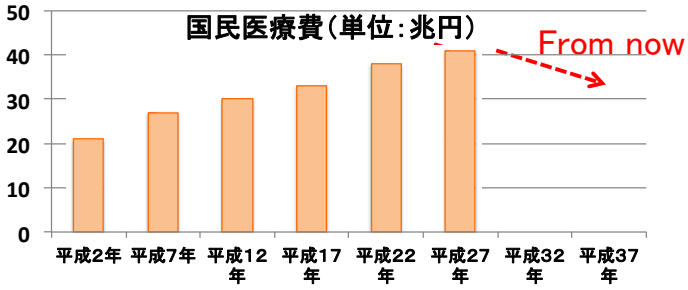
3．　今後の展開

開発したデータ圧縮技術はビッグデータ処理の様々な場面で有効であると考えられる．そこで，考えられる今後の展開は開発した技術を様々なビッグデータ処理に適応していくことである．研究テーマ B の類似度検索に関しては，要求されている制約条件も複雑になってきおり本研究課題で開発した手法だけでは不十分である．そこで，今後はより複雑な制約条件にも適応可能な類似度検索技術を開発する予定である．研究テーマ C の「機械学習応用」に関しては，開発したデータ圧縮技術を様々な機械学習法に適応していくことが考えられる．また，本研究課題では，機械学習アルゴリズムのメモリ削減に焦点を当てて手法を開発したが，機械学習アルゴリズムの速度を向上する研究の方向性も考えられる．機械学習アルゴリズムの高速化はビッグデータからの統計モデル学習において有効な手段と考えられるので，今後，開発して行く予定である．

4．　評価

（1）自己評価

研究当初の目的は概ね達成した．研究の進め方に関しては，研究補助者を見付けられなく，研究を加速させることができなかったことが残念な点である．本研究プロジェクトで開発した様々な技術は新規創薬の場面で有効な技術ある．現在，新規薬の開発コストの増加に伴い薬の価格も増加している．さらに，薬の価格の増加と高齢社会により国民医療費も増加していることが現代社会における重要な問題となっている．本研究課題で開発した技術は創薬におけるバーチャルスクリーニングの効率を上げることが可能であるので，薬の低価格化，さらには，国民医療費を下げることが期待できる．



（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

ビッグデータの統合利活用のためには大きなメモリを必要とする．その結果，メモリのコストが増加したり，メモリ階層の上位のメモリのヒット率が低下することで，処理性能が劣化したりしやすい．

本研究では，このようなビッグデータの巨大さがもたらす問題の解決策として，データを圧縮した状態で基本演算を高速に実行する方式を提案し，さらに応用分野の実データを用いた実証実験も行っている．主な成果は，類似列が繰り返し現れる場合に有効性が高

い文法圧縮技術，圧縮に wavelet 木を用いた化合物データベースの類似度検索技術，圧縮したデータ行列を入力とする機械学習技術をあげることができる．これらに関する論文を，トップカンファレンスなどで多数発表しており，優れた学術成果をあげている．これらの成果の中には，バイオインフォマティクスやケモインフォマティクスなどの応用分野の実データを用いた性能評価も含まれる．

今後，ケモインフォマティクスで真に実用に耐える制約の少ない圧縮方式，多様な機械学習アルゴリズムで利用可能な圧縮方式などの研究を継続し，社会的インパクトの大きな成果をあげることを期待したい．

5. 主な研究成果リスト

（1）論文（原著論文）発表

研究テーマA「大規模反復テキスト処理のための文法圧縮法」

1. Djamal Belazzougui, Patrick Coding, Simon J. Puglisi, Yasuo Tabei, Access, Rank and Select in Grammar-compressed strings, In Proceedings of the 23rd European Symposium on Algorithms (ESA), 2015.
2. Shirou Maruyama and Yasuo Tabei,, Fully-online Grammar Compression in Constant Space, In Proceedings of Data Compression Conference, 2014.
3. Yoshimasa Takabatake, Yasuo Tabei, Hiroshi Sakamoto, Online Self-indexed Grammar Compression, In Proceedings of the 22nd edition of the International Symposium on String Processing and Information Retrieval (SPIRE), 2015
4. Djamal Belazzougui, Travis Gagie, Paweł Gawrychowski, Juha Kärkkäinen, Alberto Ordóñez, Simon J. Puglisi, Yasuo Tabei: Queries on LZ-Bounded Encodings, In Proceedings of the Data Compression Conference (DCC), 2015.
5. Yoshimasa Takabatake, Yasuo Tabei, Hiroshi Sakamoto: Improved ESP-index: a practical self-index for highly repetitive texts, 13th International Symposium on Experimental Algorithms (SEA), 2014.

研究テーマ C「データ圧縮技術によるスケーラブルな機械学習法」

1. Yasuo Tabei, Hiroto Saigo, Yoshihiro Yamanishi, Simon J. Puglisi: Scalable partial least squares regression on grammar-compressed data matrices, In Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2016.
2. Yasuo Tabei, Yoshihiro Yamanishi, Masaaki Kotera, Simultaneous Prediction of Enzyme Orthologs from Chemical Transformation Patterns for De Novo Metabolic Pathway Reconstruction, In Proceedings of the 23rd International Conference on Intelligent Systems for Molecular Biology (ISMB), 2016.
3. Yoshihiro Yamanishi, Yasuo Tabei, Masaaki Kotera: Metabolome-scale de novo pathway reconstruction using regioisomer-sensitive graph alignments, In Proceedings of ISMB/ECCB, 2015.
4. Masaaki Kotera*, Yasuo Tabei*, Yoshihiro Yamanishi*, Ai Muto, Yuki Moriya, Toshiaki Tokimatsu, Susumu Goto: Metabolome-scale prediction of intermediate compounds in

multi-step metabolic pathways with a recursive supervised approach, 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), 2014.

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

・ Succinct Data Structure for Scalable Knowledge Discoveries, チュートリアルセッション, The 20th Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2016 年 4 月 19 日(火)

・ 文法圧縮の理論と実践, 第 27 回 RAMP シンポジウム, セッション「離散構造とアルゴリズム」, 2015 年 10 月 15 日(木)

・コンパクトなデータ表現による機械学習, 第 14 回情報科学フォーラム(FIT), イベント企画, ビッグデータ解析のための機械学習技術, 2015 年 9 月 17 日(木)

・透過的データ圧縮法による高速かつ省メモリーなビッグデータ活用技術の創出, 第 77 回情報処理学会全国大会, CREST・さきがけ「ビッグデータ」2 領域 成果報告会, 2015 年 3 月 18 日(火)

・ Dictionary based compression for processing massive genome sequences, ゲノムテクノロジー164 委員, 2014 年 12 月 18 日(木)

・ 透過的データ圧縮法による高速かつ省メモリーなビッグデータ活用技術の創出, ビッグデータ時代に向けた革新的アルゴリズム基盤, 京都リサーチパーク, 2014 年 1 月 11 日(土)12 日(日)