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

令和　４ 年　 ５ 月　 ９ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2018～2021
課題番号：１８Ｋ１８１０２

研究課題名（和文）圧縮索引と文字列圧縮の組合せによる大規模データ高速情報処理技術

研究課題名（英文）Fast Information Processing of Large-scale Data Based on a Combination of Compressed Indices and String Compression

研究代表者

伝住　周平（Denzumi, Shuhei）

東京大学・大学院情報理工学系研究科・助教

研究者番号：９０７５５７２９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では，巨大なデータを予め圧縮して小さくしてから処理することで計算時間や計算資源の劇的な削減を実現する圧縮表現上での計算技術の開発を行った．系列二分決定グラフに限らずより広い種類の決定グラフに適用可能な圧縮方法を提案し性能評価を行った．これらの成果により大規模な文字列集合を表す系列二分決定グラフを更にコンパクトなサイズに圧縮し，かつ高速に処理することが可能なデータ構造とアルゴリズムが得られた．

研究成果の学術的意義や社会的意義
人々の生活のあらゆるところに電子機器が浸透し，それらがネットワークにつながることで日々膨大な量のデータが生み出され続けている．そういったデータを解析処理しようとしても爆発的なデータの生成速度に対し通常のアルゴリズムでは処理が追いつかないという問題が広く顕在化している．そのため，文字列集合のみならず集合族なども圧縮して表現することで効率良く扱えるようにする本研究の成果は計算機科学を利用する広範な分野において共通して重要な基盤技術でありその社会的意義も大きい．また，従来の決定グラフの性能をさらに向上させ，より一層の省領域化や多機能化，理論解析を進展させたことは学術的な側面からも意義深い成果である．

研究成果の概要（英文）：In this study, I developed a computational technique for compressed representations that dramatically reduces computation time and space by compressing large-scale data in advance. We proposed compression methods that apply not only to sequence binary decision diagrams but also to various types of decision diagrams and evaluated their performance. These results provide a data structure and algorithm that can further compress a sequence binary decision diagrams representing large-scale sets of strings into a compact size and process them at high speed.

研究分野：知能情報学

キーワード：データ構造　圧縮　索引　文字列　集合族　二分決定グラフ　項分岐決定グラフ　簡潔データ構造

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年，World Wide Web (ウェブ) や大容量記憶装置，Internet of Things (IoT) に代表される高速ネットワークの急速な発展によって爆発的に増大する一次データを解析処理し役立てていくことについて社会的関心が高まっているが，こういったデータはそのままの形で扱うことが困難なほど膨大であるために効率よくかつ高速に処理を行う手法の開発が急務となっている．しかし，データが増えたからといってそこに含まれる有用な情報も比例して増えるとは限らない．例えば，24 時間稼働し続けるセンサから送られてくるデータはほとんどの時間で変化がなかったり，人間のゲノムでもサンプルが増えるほど新しいサンプルは既にあるゲノムと類似する部分が多くなったりする，といったことが考えられる．つまり，巨大なデータほど何らかの「冗長性」を強く持つようになる．冗長な部分があるならばそこを圧縮してやることでデータをよりコンパクトな形態に変化させて表現することが可能となる．データの圧縮というとサイズが小さくなればそれでよいというイメージを持たれることがあるが，データを保存するだけならまだしも利活用を考えた場合にはそれだけでは不十分だ．圧縮されたデータを展開せずにそのまま処理することが可能であれば時間や計算資源の大幅な節約につながるためである．圧縮とは元のデータにおける冗長な部分をまとめることと捉えられる．元のデータをそのまま処理しようとしていたら有用な情報がある部分も冗長な部分も区別なく扱われてしまうが，圧縮表現上で同じ処理を実現できたならば圧縮後のサイズに依存した時間で計算を終えることができるようになる．圧縮するために時間がかかったとしてもデータを指数的にコンパクトにできる場合や，行いたい処理にかかる計算時間のオーダーがサイズの二乗や三乗といった場合はこの圧縮表現上での計算の利点が非常に大きくなる．このように圧縮表現上での計算は日々生み出される大量のデータを処理するための重要な基盤技術となるものである．

２．研究の目的

系列二分決定グラフというデータ構造がある．これは有限な文字列集合を巡回のないグラフ構造として表現するもので，場合によってはグラフのサイズの指数倍もの文字列集合を表現することができる．さらに文字列集合を圧縮したまま検索，和集合や積集合などの基本的集合演算，さらにはより複雑な演算を実現することができる．しかし，系列二分決定グラフはグラフ構造に基く他のデータ構造と比べた際に，特に検索などの基本的な操作において，時間がかかることが知られている．本研究の目的は，系列二分決定グラフに対し文字列圧縮の技法を適用することで，よりコンパクトな表現を得つつ検索や集合同士の演算の高速化も実現することである．文字列圧縮とは与えられた文字列中に繰り返し出現する部分文字列に同じ変数を割り当てるなどにより元の文字列より短い記述長を実現する技法である．

文字列圧縮アルゴリズムは LZ 符号化や Re-pair，TtoG など数多く提案されており，圧縮したまま検索する手法も数十年にわたり研究されてきている．それら既存の文字列圧縮アルゴリズムの中でどのような手法が本研究の目的に適しているかの検討も重要な課題である．系列二分決定グラフのようなグラフ構造による圧縮はメモリに収めるのが考えられないほど多くの数の文字列を表現するのに適しており，一方で文字列圧縮はメモリ上に収まるくらいの文字列をできるかぎり小さくするのに適している．それぞれ得意とする範囲が異なるためか両者をうまく組合せようという研究は少ない．加えて，圧縮した後での演算，それも単なる検索だけではなく基本的な集合演算，さらにより複雑な演算をサポートするデータ構造となると存在せずそれを新たに創造することが本研究の目的である．

３．研究の方法

グラフ構造による圧縮索引と文字列圧縮を組合せることを考えた際に最も単純な方策は索引を構築する入力となる文字列に先んじて文字列圧縮をかけてからグラフ構造にする，というものである．しかし，この方法では圧縮表現上での計算がうまくいかなくなってしまう．なぜなら最終的にはグラフ構造上での計算を行いたいのにそれを考慮せずいきなり圧縮してしまうとグラフにした後で扱いにくいような圧縮されかたになってしまうからである．そのため本研究ではまず文字列集合が系列二分決定グラフとして与えられるものとし，その後に文字列圧縮の技法を用いてそれをさらにコンパクトにする方法を開発する．

文字列索引としてまずできなくてはならない操作，それは与えられた文字列パターンが索引の表す文字列集合に含まれているかどうかを判定するメンバーシップ演算である．現在の系列二分決定グラフでは最悪の場合でパターン 1 文字ごとに存在する文字の種類の数だけグラフの節点を移動しなければならないことが知られている．文字の種類数が多い場合，この移動にかかる時間が無視できないほど大きくなるのでまずはこの問題を解決したい．方法としては各文字を二進列などに変換して移動回数を文字種の対数回に抑えるなどが考えられる．

次に考えるべきは分岐することのない直線的なグラフ節点の連なりである． Compact Directed Acyclic Word Graph のような系列二分決定グラフと同じようにグラフ構造により文字列集合を

表す索引ではこのような節点列はまとめて連続してアクセスできるようにして高速化する工夫が導入されていることがあるが，これをそのまま系列二分決定グラフに適用することは難しい．そこでこのようなグラフ上で直線的に連続している文字列に対し文字列圧縮を適用することで節点の数を削減しアクセスの高速化を目指す．ただし，検索の際などに文字列圧縮した節点をすべて展開しなければならなくなっては元も子もないので数ある文字列圧縮アルゴリズムの中から適した手法を選択する．
そして系列二分決定グラフの重要な特徴であり，本研究の独自性に大きく関わるのが文字列集合同士の演算である．これは 2 つ以上の系列二分決定グラフが与えられたときにそれらが表す文字列集合同士に何らかの演算を実行した結果得られる文字列集合を系列二分決定グラフの形式で出力するというものである．肝要なのは入力の系列二分決定グラフを展開せずに処理することで計算にかかる時間を圧縮された状態であるグラフのサイズのみに依存するようにすることである．これを実現するには系列二分決定グラフとは異なりより文字列圧縮に近い考えでグラフを圧縮する top-tree compression の技法が参考になると思われる．

４．研究成果
巨大なデータを予め圧縮してから処理することで計算資源を劇的に削減するための圧縮表現上での計算技術の開発を目的とし以下のような研究を行った．特に二分決定グラフというデータ構造の一群を主要な対象とした．二分決定グラフは離散構造を圧縮して表現し，その表現対象同士の演算もサポートするデータ構造である．ゼロサプレス型二分決定グラフは組合せ集合を表現し，その変種である系列二分決定グラフは文字列集合を表す．系列二分決定グラフは指数的な数の文字列を圧縮して表現可能で，文字列集合同士の多様な演算をサポートするデータ構造である．

(1) 論理関数を表現する二分決定グラフの一般化である項分岐二分決定グラフの技法を系列二分決定グラフに適用することで，文字列集合を表す項分岐二分決定グラフという新しいデータ構造を提案した．項分岐二分決定グラフは決定性有限オートマトンと同様のデータ構造である系列二分決定グラフを一般化したデータ構造であるが，同じ文字列集合を系列二分決定グラフよりも指数的に小さく表現できる場合がある．さらに，項分岐二分決定グラフによって表される文字列集合を操作するためのアルゴリズムも提案した．これらのアルゴリズムにより項分岐二分決定グラフによって圧縮表現された文字列表現を展開することなく和集合，積集合，連接を計算することができる．また，提案アルゴリズムの空間計算量と時間計算量の解析を行った．

(2) 組合せ集合を効率良く表現するゼロサプレス型二分決定グラフ，それを更新するための演算が必要ない場合により簡潔な表現を実現する密集ゼロサプレス型二分決定グラフに対する新たなアルゴリズムを提案した．これにより組合せ集合からランダムに組合せを取り出すランダムサンプリングを従来のゼロサプレス型二分決定グラフより高速に行うことが可能になった．また，密集ゼロサプレス型二分決定グラフの有効性を実験により示した．この技法は系列二分決定グラフにも適用できる．

(3) 文字列は繰り返し可能な要素が一列に並ぶもので，組合せと比べてより複雑である．系列二分決定グラフはゼロサプレス型二分決定グラフから継承した集合演算を有してはいるが，文字列処理に際し要求される多様な操作を実現するには不十分であった．そこで５０を超える新たな系列二分決定グラフ操作アルゴリズムとともに，その時間・空間計算量の解析を行った．

(4) 二分決定グラフは離散データを指数的に圧縮できる場合があるといえど，巨大なデータに対してはそれでもメモリに収まらないことがある．しかし，情報論的には可逆圧縮では指数的に圧縮することが限界である．そこで，ゼロサプレス型二分決定グラフにあえて本来は含まれない組合せをある程度追加することでデータ構造のメモリ使用量を大幅に削減する技術を開発した．本来の組合せ集合との差異は新たに挿入された誤った組合せのみであるため，得られるデータ構造は偽陽性を許容した索引として利用可能である．特に，ある組合せが本来の要素であるか容易に確かめることができる場合に効力を発揮する．さらに，元のゼロサプレス型二分決定グラフを構築しながらこの圧縮を実行するオンライン手法も実現した．この技術は部分的に系列二分決定グラフにも適用可能である．

(5) 文字列 x の k-anticover とは x の長さ k の相異なる部分文字列からなる集合であり，x の全ての位置をいずれかの部分文字列を高々1 回使うことで被覆できるようなものである．文字列に k-anticover が存在することは冗長性の無さを示し，計算生物学に応用が可能である．この論文では与えられた文字列に k-anticover が存在するかどうか決定する問題が $k \geqq 3$ で NP 完全であることを示し，k=2 のとき多項式時間で解けると示した．また，指数時間で解を求めるアルゴリズムを提案した．
(6) 項分岐決定グラフ（SDD）は二分決定グラフを一般化したデータ構造で論理関数を簡潔に正規形で表現し，解の数え上げや論理演算などの操作を行える．この論文では変数シフト SDD（VS-

SDD）という更に簡潔な変種を提案した．VS-SDD は SDD より大きくなることはない上に指数関数的に小さくなりうることや，多数の演算が VS-SDD でも多項式時間実行可能と示した．実験においても VS-SDD が SDD よりかなり小さくなることを確認した．

(7) ゼロサプレス型二分決定グラフ（ZDD）は集合族を圧縮表現するデータ構造である．この論文は ZDD をより省空間で表現する Top ZDD を提案した．Top ZDD は同一の部分グラフをまとめることで ZDD を圧縮する．Top ZDD 上の遷移がサイズの対数時間で行えることを示し，top ZDD のサイズが ZDD より指数関数的に小さくなりうると示した．さらに，複雑な Top ZDD の構成やその構築方法をよりわかりやすく整理し疑似コードに書き起こした．Top ZDD が実データに対し ZDD より小さくなることを実験で確認した．また，従来のゼロサプレス型二分決定グラフやそれを簡潔データ構造の技法を融合させることで省領域化を実現したデータ構造である DenseZDD とも比較し，詳細な Top ZDD の性能評価を行った．

(8) データを分散して格納しつつも正確性と耐障害性を有するブロックチェーンは今後多くの分野で活用が見込まれる有用な技術であり，悪意により生じる問題に対処する方法の開発が求められている．ここでは一般的なブロックチェーンを処理能力が向上するように拡張した DAG 型ブロックチェーンを対象としその誠実なブロック，つまり特定のユーザーに有利になるように追加されたわけではない悪意のないデータ，を特定する高速なアルゴリズムの設計を行った．誠実なブロックを特定する問題は大きな k-独立集合を求める問題として定式化できるが，k-独立集合問題は NP 困難に属するため理論的に高速なアルゴリズムの設計は難しい．そこでゼロサプレス型二分決定グラフを用いて大きな k-独立集合を全列挙する実用的に高速な手法を提案した．この問題は k-独立集合を１つ求めるだけであれば整数計画法で解くための定式化も可能である．計算機実験により提案手法と整数計画法を用いたソルバーを比較し，提案手法の性能を検討した．

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Matsuda Kotaro、Denzumi Shuhei、Sadakane Kunihiko	４．巻 14(6), 172
２．論文標題 Storing Set Families More Compactly with Top ZDDs	５．発行年 2021年
３．雑誌名 Algorithms	６．最初と最後の頁 1～23
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/a14060172	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Shuhei Denzumi, Jun Kawahara, Koji Tsuda, Hiroki Arimura, Shin-ichi Minato and Kunihiko Sadakane	４．巻 11
２．論文標題 DenseZDD: A Compact and Fast Index for Families of Sets	５．発行年 2018年
３．雑誌名 Algorithms	６．最初と最後の頁 1-23
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/a11080128	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　6件）

１．発表者名 伝住 周平、川原 純
２．発表標題 ブロックDAGに対する最大k-独立集合問題の二分決定グラフを用いた解法
３．学会等名 アルゴリズム研究発表会
４．発表年 2022年

１．発表者名 Mai Alzamel, Alessio Conte, Shuhei Denzumi, Roberto Grossi, Costas S. Iliopoulos, Kazuhiro Kurita and Kunihiro Wasa
２．発表標題 Finding the Anticover of a String
３．学会等名 The 31th Annual Symposium on Combinatorial Pattern Matching (CPM 2020), Leibniz International Proceedings in Informatics, Vol. 161, No. 2, pp. 1-11, Copenhagen, Denmark, June 17-19, 2020 (国際学会)
４．発表年 2020年

１．発表者名
Kengo Nakamura, Shuhei Denzumi and Masaaki Nishino

２．発表標題
Variable Shift SDD: A More Succinct Sentential Decision Diagram

３．学会等名
The 18th Symposium on Experimental Algorithms (SEA 2020), Leibniz International Proceedings in Informatics, Vol. 160, No. 22, pp. 1-13, Catania, Italy (held online), June 16-18, 2020 (国際学会)

４．発表年
2020年

１．発表者名
Kotaro Matsuda, Shuhei Denzumi and Kunihiko Sadakane

２．発表標題
Storing Set Families More Compactly with Top ZDDs

３．学会等名
The 18th Symposium on Experimental Algorithms (SEA 2020), Leibniz International Proceedings in Informatics, Vol. 160, No. 6, pp. 1-13, Catania, Italy (held online), June 16-18, 2020 (国際学会)

４．発表年
2020年

１．発表者名
Shuhei Denzumi

２．発表標題
New Algorithms for Manipulating Sequence BDDs

３．学会等名
24th International Conference on Implementation and Application of Automata (CIAA 2019) (国際学会)

４．発表年
2019年

１．発表者名
Kotaro Matsuda, Shuhei Denzumi, Kengo Nakamura, Masaaki Nishino and Norihito Yasuda

２．発表標題
Approximated ZDD Construction Considering Inclusion Relations of Models

３．学会等名
Special Event on Analysis of Experimental Algorithms (SEA^2 2019) (国際学会)

４．発表年
2019年

１．発表者名 Shuhei Denzumi
２．発表標題 Sequence Sentential Decision Diagrams
３．学会等名 The 12th Annual International Conference on Combinatorial Optimization and Applications (COCOA'18)（国際学会）
４．発表年 2018年

１．発表者名 伝住周平，堀山貴史，栗田和宏，中畑　裕，鈴木浩史，○和佐州洋，山崎一明
２．発表標題 非同型な2端子直並列グラフの列挙とランダムサンプリング
３．学会等名 コンピュテーション研究会　（2018年9月18日開催，会場：九州工業大学（飯塚キャンパス））
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
イタリア	Universita di Pisa			
英国	King ’s College London			
英国	King Saud University			