様式 C-7-2

自己評価報告書

平成２２年　５月２１日現在

研究種目：若手研究(A)
研究期間：2007～2010
課題番号：１９６８０００１
研究課題名（和文）　実際的簡潔データ構造の開発と大量データ活用

研究課題名（英文）　Development of Practical Succinct Data Structures with Applications to Large-Scale Data Processing

研究代表者
定兼　邦彦（SADAKANE KUNIHIKO）
国立情報学研究所・情報学プリンシプル研究系・准教授
研究者番号：２０３２３０９０

研究分野：総合領域
科研費の分科・細目：情報学・情報学基礎
キーワード：アルゴリズム理論，データ構造，情報検索

１．研究計画の概要

Web ページ, 関係データベース, 新聞記事, ゲノム配列など，現在大量のデータが存在し，それらを有効活用するための技術の開発が重要となっている．古典的な手法では，データをディスクに格納し，それを逐次的に読み込みながら処理を行うが，これでは速度が非常に遅く，また，限られた処理しか行えない．そこで現在は「主記憶データベース」と呼ばれる，全データを計算機のメモリ(主記憶)に格納して高速処理を行う手法が取られている．しかしこの手法でもデータの検索手法は古典的なアルゴリズムとデータ構造を用いているため，以下のような問題が存在する：

(1)　検索速度と，検索を行うために必要なメモリ量にはトレードオフが存在する．つまり，高速な検索を行うには大量のメモリが必要となる．

(2)　主記憶量の制限により，高速な検索を行えるデータ量は非常に小さくなる．

(3)　データを圧縮することで必要メモリを減らすことができるが，データのランダムアクセスができなくなるため，検索が遅くなる．

これらの問題を解決するために提案されたものが簡潔データ構造である．

簡潔データ構造とは，データおよびそれを高速に処理するためのデータ構造（索引）のサイズを極限まで小さくし，なおかつ従来のデータ構造と同じ処理が同じ計算量で行えるものであり，申請者や海外の研究者らによりここ数年盛んに研究されている新しい概念である．簡潔データ構造を用いれば，大量のデータと索引をメモリに格納でき，高速処理が実現できる．

しかし，簡潔データ構造はまだ理論的な研究が始まったばかりであり，それを実際に活用する段階には至っていない．現在の理論的な結果をそのままプログラムとして実装すると，実行速度，索引サイズの点で満足のいくものにはなっていない．本研究では，理論的にも実際的にも優れた簡潔データ構造を開発し，それを大量データ処理に活用する．

２．研究の進捗状況

これまでに次のような簡潔データ構造を開発している．

(1) DNA 配列検索のための圧縮索引

既存の文字列検索のための圧縮索引は一般の文字列を対象にしているため，速度が遅い．しかし対象を DNA 配列に限定することで検索速度を大幅に向上できる．索引サイズを固定した場合に既存手法よりも検索が 10 倍高速，検索速度を固定した場合に索引サイズが半分になるような索引を開発した．

(2) ビット列に対する簡潔データ構造

ビット列において，列中の 1 の数を数えたり，i 番目の 1 の位置を求める演算は全ての簡潔データ構造で用いられる基本的な演算であるため，実用的なデータ構造の開発は重要である．そこで圧縮されたビット列に対する簡潔データ構造を開発した．これは既存手法よりも圧縮率，演算速度共に勝っている．

(3) 順序木に対する簡潔データ構造

木構造も情報検索において基本的なデータ構造であり，様々なデータ構造が提案されているが，それらは理論的な結果のみであり，

実用的なものは存在しなかった．そこで，理論的にも実際的にも優れている順序木の簡潔データ構造を開発した．理論的には，知られている全ての演算を定数時間で実現し，データ構造のサイズも既存手法よりもはるかに小さい．実際的には，既存手法はデータ構造が複雑であるためサイズも大きくなり，限定された演算しか実装されていなかったが，本研究で開発したデータ構造は単純であるため実装が簡単であり，全ての演算を実装することができた．また，全ての演算をいくつかの基本演算で実現できることを示したため，索引のサイズも非常に小さくできた．

３．現在までの達成度
①当初の計画以上に進展している．
データ構造のサイズ，検索速度ともに当初の計画よりも優れたものが開発できている．

４．今後の研究の推進方策
いくつかの基本的なデータ構造が開発できたため，それらを用いて大量データ処理を高速に行う．具体的には，日本の特許データベース過去５年分(約 90 ギガバイト)，DNA 配列データ，Web グラフデータなどに対する圧縮索引を作成し，その上での検索や知識発見を行う．その際に，検索処理では完全一致だけではなくあいまい一致を見つけることが重要であるため，圧縮索引上であいまい検索を高速に行うアルゴリズムの開発が重要となる．

５． 代表的な研究成果
（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 13 件）

① K. Sadakane, G. Navarro: Fully-Functional Succinct Trees, SODA 2010, 134-149, 査読有
②D. Arroyuelo, R. Canovas, G. Navarro, K. Sadakane: Succinct Trees in Practice, ALENEX, 84-97, 2010, 査読有
③ Daisuke Okanohara, Kunihiko Sadakane: A Linear-Time Burrows-Wheeler Transform Using Induced Sorting. SPIRE 2009: 90-101, 査読有
④ Wing-Kai Hon, Kunihiko Sadakane, Wing-Kin Sung: Breaking a Time-and-Space Barrier in Constructing Full-Text Indices. SIAM J. Comput. 38(6): 2162-2178 (2009), 査読有
⑤Daisuke Okanohara, Kunihiko Sadakane: An Online Algorithm for Finding the Longest Previous Factors. ESA 2008: 696-707, 査読有
⑥Daisuke Okanohara, Kunihiko Sadakane: Practical Entropy-Compressed Rank/Select Dictionary. ALENEX 2007, 査読有
⑦ Jesper Jansson, Kunihiko Sadakane, Wing-Kin Sung: Compressed Dynamic Tries with Applications to LZ-Compression in Sublinear Time and Space. FSTTCS 2007: 424-435, 査読有
⑧ Jesper Jansson, Kunihiko Sadakane, Wing-Kin Sung: Ultra-succinct representation of ordered trees. SODA 2007: 575-584, 査読有
⑨Ho-Leung Chan, Wing-Kai Hon, Tak Wah Lam, Kunihiko Sadakane: Compressed indexes for dynamic text collections. ACM Transactions on Algorithms 3(2): (2007), 査読有
⑩ Wing-Kai Hon, Tak Wah Lam, Kunihiko Sadakane, Wing-Kin Sung, Siu-Ming Yiu: A Space and Time Efficient Algorithm for Constructing Compressed Suffix Arrays. Algorithmica 48(1): 23-36 (2007), 査読有
⑪ Kunihiko Sadakane: Succinct data structures for flexible text retrieval systems. J. Discrete Algorithms 5(1): 12-22 (2007), 査読有
⑫N. Jesper Larsson, Kunihiko Sadakane: Faster suffix sorting. Theoretical Computer Science 387(3): 258-272 (2007), 査読有
⑬ Kunihiko Sadakane: Compressed Suffix Trees with Full Functionality. Theory of Computing Systems 41(4): 589-607 (2007), 査読有

〔学会発表〕（計 4 件）

①定兼邦彦: 順序木の簡単・簡潔な表現法, 日本オペレーションズ・リサーチ学会研究部会 画期における最適化, 2009 年 11 月 26 日, 京都大学
②定兼邦彦: 動的簡潔順序木, 電子情報通信学会コンピュテーション研究会, 2009 年 4 月 17 日, 京都大学
③定兼邦彦: 括弧列の簡単・簡潔な表現法, 電子情報通信学会コンピュテーション研究会, 2008 年 10 月 10 日, 東北大学
④定兼邦彦: DNA配列に適した圧縮全文索引, 電子情報通信学会コンピュテーション研究会, 2008 年 3 月 10 日, 日本IBM東京基礎研究所

〔その他〕
ホームページ http://researchmap.jp/sada