科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１０　日現在

機関番号：　１０１０１
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００００２

研究課題名（和文）高速・高度なパターン照合と高圧縮率とを実現するＶＦ符号化の研究

研究課題名（英文）On high performance VF coding allowing fast and and sophisticated pattern matching

研究代表者
喜田　拓也（Kida, Takuya）

北海道大学・情報科学研究科・准教授

研究者番号：７０３４３３１６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円、（間接経費）　　960,000円

研究成果の概要（和文）：申請者は，可変長-固定長符号（VF符号）の性能改善に取り組んだ．VF符号は，固定長の符号化であるため，圧縮データへのアクセスが容易であるという特長がある．しかし，既存のVF符号は，可変長符号化と比べて圧縮率が劣るという欠点が存在する．これに対し，申請者らは，1999年にLarssonとMoffatらによって提案されたRe-Pairアルゴリズムに固定長符号を組み合わせた新しいVF符号化を提案した．本手法をRe-Pair-VF符号化と名付け，既存のVF符号より優れた圧縮性能を持つことを実証した．

研究成果の概要（英文）：In this study, the applicant addressed the problem of improving variable-length-to-fixed-length codes (VF codes). A VF code is an encoding scheme that uses a fixed-length code, which provides easy access to compressed data. However, conventional VF codes generally have an inferior compression ratio compared with variable-length codes. The applicant et al. proposed a new VF coding method that applies a fixed-length code to a set of rules extracted using the Re-Pair algorithm, which was proposed by Larsson and Moffat in 1999. The Re-Pair algorithm is a simple offline grammar-based compression method, which has good compression-ratio performance with moderate compression speed. The proposed method is named as Re-Pair-VF coding. The experimental results showed that Re-Pair-VF coding is superior to the existing VF coding method.

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：　圧縮照合　データ圧縮　文法圧縮　VF符号

１．研究開始当初の背景

高速インターネット技術と大容量記憶装置技術の進展に伴い，近年，膨大な量の機械可読な文書が生成され，また流通するようになった．個人が持つ電子メールのアーカイブや私的な情報データベースでさえもギガバイト以上の容量が必要となってきている．したがって，効率のよい情報検索技術の開発が求められている．

データを先頭から走査して検索クエリに一致する部分を探索するパターン照合技術（文字列照合技術）は情報検索の重要な基本技術の一つであり，申請者自身，これまでに様々な検索要求に対するパターン照合アルゴリズムの高速化および高度化に取り組んできた（「半構造化データに対する文字列処理の高速化に関する研究【平成 14 年度～16 年度；若手研究 B】」，「背景知識を考慮した文字列照合に関する研究【平成 17 年度～19 年度；若手研究 B】」，「連続データストリームに対する高度なパターン照合の研究【平成 20 年度～22 年度；若手研究 B】」）．

一方で，このような大量の文書データ（テキストデータ）は，保存コストあるいはその通信コストを低減するためにデータ圧縮して保存されることが多い．データ圧縮とは，データ中に含まれる冗長性をコンパクトに表現することで，記憶のための領域を削減する技術である．そこで，このような圧縮されたテキストに対して，それを元の文書データに展開することなくパターン照合処理を行うという要求が生じた．この問題は，1990 年代初頭に A. Amir と G. Benson らにより提案され，以降盛んに研究が行われている．

申請者はこれまでに，既存の圧縮法である LZ 法に対するパターン照合アルゴリズムや Byte pair encoding (BPE)圧縮法に対するパターン照合アルゴリズムの開発に従事してきた．特に，BPE 圧縮法に対するパターン照合アルゴリズムでは，圧縮していない元の文書テキストに対してパターン照合する場合に比べて，およそ圧縮率程度の時間でパターン照合を行えることを示している．これは視点を変えると，データ圧縮がパターン照合処理を高速化するという見方ができる．

この事実は，本研究分野の研究者らに大きなパラダイムシフトをもたらした．すなわち，パターン照合アルゴリズム自体を改善するというこれまでの考え方から，パターン照合に都合の良いデータ圧縮法を設計することでパターン照合の高速化を達成するという考え方への移行である．実際，2000 年以降，パターン照合の高速化を目的としたデータ圧縮方式が，申請者が当初所属していた九州大学の竹田らのグループの他に，ヘルシンキ工科大学の J. Tarhio らのグループや，チリ大学の G. Navarro らのグループ，バル＝イラン大学の S. T. Klein らのグループなどから相次いで提案されている．

しかしながら，パターン照合を高速に行うことと，テキストを効率良く圧縮することの両立は，非常に困難な課題である．これまでに提案されてきたパターン照合用のデータ圧縮方式には，次に挙げる共通する特徴がある．

① 静的（かつコンパクト）な辞書を利用している．
② 符号語の境界が圧縮データの途中からでも識別しやすい符号化を採用している．

一方で，現在主流の圧縮法の多くは，動的（適応的）に更新される辞書を用い，ビット単位で可変な非等長の符号語を採用している．なぜならば，それらは圧縮率を第一義として設計されているためである．これら，上述した二つの特徴は，圧縮率を向上させる上での大きな制約となっている．

申請者は，数ある圧縮手法の中でも特に，VF 符号(Variable Length to Fixed Length 符号)に着目して研究をすすめてきた．VF 符号は，文書を長さが異なる文字列（ブロックとよばれる）に分解して，それぞれに固定長の符号語を割り当てる圧縮方法である．可変長の符号語を採用している圧縮法とは異なり，VF 符号はすべての符号語が同じ長さ（固定長符号化）であるため，任意の符号語の開始位置と終了位置が明白である．さらに，圧縮データ上の任意の位置からパターン照合を開始できるほか，部分的なデータの復元・再圧縮処理も比較的容易に行えるなど，実応用の観点から非常に優れた性質を備えている．VF 符号化されたテキストに対するパターン照合アルゴリズムは，前述した BPE 圧縮法に対するパターン照合アルゴリズムと同様の手法によって組織的に導出できることが分かっている．

最も古典的かつ基本的な VF 符号としては Tunstall 符号がある．Tunstall 符号は，分節木と呼ばれる静的な木構造を辞書として用いる VF 符号であり，無記憶情報源に対して理論的に最適な効率を得る VF 符号であることが証明されている．しかし，Tunstall 符号の圧縮率は，理論的に同等の性能をもち良く知られた可変長符号化である Huffman 符号と比べると，実際の圧縮率で劣ることが申請者らの調査により判明している．実際的なテキストデータの多くは，自然言語文書やゲノムデータなど，そのモデルは記憶のある情報源であると考えられており，そうしたモデルに対応するための改善策もいくつか提案されてはいるが，それらの実用性を示す実証実験等の報告は皆無である．

このような背景から，申請者は，高い圧縮率を達成するために，接尾辞木と呼ばれる木構造を利用した新しい VF 符号である STVF 符号を提案した．接尾辞木は，与えられたテキストのすべての部分文字列を格納するデータ構造であり，任意の部分文字列の頻度を接尾辞木上であらかじめ計算しておくことができる．STVF 符号では，テキスト中に出

現する各部分文字列の頻度に基づいて接尾辞木を刈り込み，それを分節木として用いる．STVF 符号を実装し，その圧縮性能の評価実験を行った結果，Huffman 符号や BPE 圧縮法よりも優れた圧縮率を達成することが確認された．

２．研究の目的

本研究の目的は，圧縮率の高い VF 符号を確立し，それを実現する効率良い符号化・復号化アルゴリズムを開発することである．それによって，大規模データに対する多様な検索要求を高速に処理するシステムの構築を目標とする．

申請者らは，先に述べた STVF 符号のさらなる性能向上を目指して改良を続けている．現状では，圧縮速度をある程度犠牲にしてよければ，よく知られた圧縮ツールである gzip よりも高い圧縮率を達成できることが判明している．本研究期間内では，圧縮率を向上もしくは維持しつつ高速に圧縮処理できる符号化アルゴリズムを開発し，その性能を理論的・実験的に解析することを第一の目標とする．第二に，複雑な検索を高速に処理するために，VF 符号化テキストに対する多様なパターン照合アルゴリズムの開発を行う．具体的には，近似文字列照合や正規表現パターンの照合などを VF 符号上で高速に行うアルゴリズムの開発に取り組む．最後に，圧縮ツールとその検索ツールを組み合わせたシステムの構築を行う．

３．研究の方法

本研究では，これまでに申請者が提案した STVF 符号を足掛かりとし，高い圧縮率を達成する実用的な VF 符号化方法の研究・開発を行う．それを通して，高速・高度なパターン照合を可能とする大規模テキスト検索システムの構築を目指している．

そのためのマイルストーンとして，第一には，STVF 符号の改善あるいは新規アイデアによる VF 符号のさらなる圧縮率の向上を図る．第二に，その圧縮率を維持しつつ高速な符号化・復号化アルゴリズムを開発する．第三に，VF 符号化テキストに対する複雑なパターン照合アルゴリズムの開発を行う．

初年度にあたる平成 23 年度は，本分野における関連研究の調査を行うと共に，STVF 符号の圧縮率改善と圧縮・復元アルゴリズムの高速化に取り組む．平成 24 年度以降は，VF 符号上での複雑なパターン照合を効率よく処理するアルゴリズムの開発に取り組む．

４．研究成果

STVF 符号は，圧縮対象となるテキストが自然言語のように，文脈がある場合（すなわち記憶のある情報源の場合）には高い圧縮率を達成できる符号である．しかしながら，gzip や bzip2 など最新の圧縮方法と比較すれば，まだ圧縮率の点で劣っていた．ここまでの改善手法では，未使用な符号語の割合を 10～20%程度にしか抑えられず，これ以上の圧縮率改善には，根本的な発想の転換が必要であった．

これに対し，初年度では，文法変換に基づくデータ圧縮法のアイデアを用い，Re-Pair アルゴリズムによって文法変換されたデータに VF 符号を適用することで，データ圧縮率と圧縮速度の向上を図った．提案したデータ圧縮法を Re-Pair-VF と名付け，これを実装し，実証実験を行った．この実験結果から，圧縮率において gzip を上回り，圧縮速度において STVF 符号の 2 倍の速度を達成することを示した．

初年度において，VF 符号の大幅な改善を達成することができたため，次年度以降では，Re-Pair-VF 符号を実応用する際の問題点解消について取り組みを開始した．

Re-pair-VF 符号は，Re-pair アルゴリズムに基づいた手法であるため，Re-pair アルゴリズムの制約を受ける．すなわち，基本的にはオフラインの処理アルゴリズムであり，また（入力データに対して線形時間での）高速な圧縮処理を実現するために，元データの 20 倍程度のメモリを消費する．このメモリ消費量の問題が，Re-Pair-VF 符号を大規模テキストへ適用する際のネックとなっていた．

これに対し，分割されたブロック毎に Re-Pair-VF 符号を行うブロック分割手法について研究を行った．テキストのブロック分割は既に多くのデータ圧縮ツールで採用されている手法だが，単純にブロック毎に圧縮を行うだけでは圧縮率の低下を起こしてしまう．よって，各ブロックで共通の辞書を用いるという手法を提案し，実現した．このことにより，大規模なデータに対しても，現実的な圧縮時間で gzip を凌ぐ非常に良好な圧縮率を得ることに成功した．

また，Re-pair-VF 符号上でのパターン照合アルゴリズムを実現し，従来法（zgrep 等）と比較実験を行った結果，2 倍程度の速度向上を達成することに成功した．

最終年度も，上述の Re-Pair-VF 符号の改善を推進した．前年度で実現した辞書共有の手法は，テキスト全体から一旦情報を集めたのちに共有辞書を構築するという静的な辞書構築法であった．その場合，効率よく圧縮を行える共有辞書を構築するためには，テキスト全体から偏りなく情報を集めてくる必要があり，圧縮速度と圧縮率とのトレードオフが存在する．そこで，共有する辞書をブロック間で動的に構築する手法について研究・開発を行い，Adaptive Dictionary Sharing 法(ADS 法)という動的な辞書構築手法を実現した．このことにより，テキスト全体から情報を集める場合よりも速度が向上した．さらに，途中で入力データの傾向が変化した場合にも適切な辞書構造を保つことが可能となり，全体の圧縮率が向上した．

また最終年度では，Re-pair アルゴリズム

による圧縮データに対して，元のデータ位置を指定した直接的なアクセスを可能にする手法についても研究・開発を行った．通常，圧縮されたデータに対し，元のデータ位置を特定するには，前方から逐次的にデータを展開もしくは解析する必要がある．これに対し，符号語の切れ目を認識するビット列を完備辞書として保持することで，既存手法よりもコンパクトなデータ量を維持しつつ，高速な直接アクセスを実現することができた．

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 2 件）

① S. Yoshida, T. Uemura, T. Kida, et al.: Improving Parse Trees for Efficient VF Codes, 査読有, Journal of Information Processing, Vol.20, No.1, pp.238-249, Jan., 2012, 10.2197/ipsjjip.20.238
② Satoshi Yoshida and Takuya Kida: A Variable-length-to-fixed-length Coding Method Using a Re-Pair Algorithm, 査読有, IPSJ Transactions on Databases, Vol. 6, No. 4, pp. 17-23, Sep., 2013.

〔学会発表〕（計 13 件）
① Satoshi Yoshida, Hirohito Sasakawa, Kei Sekine, Takuya Kida: Direct Access to Variable-to-Fixed Length Codes with a Succinct Index, In Proceedings of Data Compression Conference 2014 (DCC 2014), p.436, Snowbird, Utah, USA, 27 March, 2014.
② Kei Sekine, Hirohito Sasakawa, Satoshi Yoshida, Takuya Kida: Adaptive Dictionary Sharing Method for Re-Pair Algorithm, In Proceedings of Data Compression Conference 2014 (DCC 2014), p.425, Snowbird, Utah, USA, 27 March, 2014.
③ 関根渓, 笹川裕人, 吉田論史, 喜田拓也: 大規模テキストに対する共有辞書を用いた Re-Pair 圧縮法, 第 5 回データ工学と情報マネジメントに関するフォーラム(DEIM2013), No. C10-2, 福島県, 2013 年 3 月 5 日.
④ 関根渓, 笹川裕人, 吉田論史, 喜田拓也: 共有辞書を用いた効率の良い圧縮アルゴリズム, 第 156 回データベースシステム研究発表会, Vol.2012-DBS-156, No. 7, 京都府, 2012 年 12 月 12 日.
⑤ Satoshi Yoshida and Takuya Kida: An Efficient VF Coding via Re-Pair Algorithm, In International Workshop on Information Search, Integration and Personalization (ISIP'2012), Sapporo, Japan, 12 Oct., 2012.
⑥ Satoshi Yoshida and Takuya Kida: A Multiplexed Parse Tree for Almost Instantaneous VF Codes, In the 15th Japan-Korea Joint Workshop on Algorithms and Computation (WAAC 2012), Tokyo, Japan, 10 July, 2012.
⑦ 笹川 裕人, 喜田 拓也, 有村 博紀: 長大な拡張文字列パターンに対する GPU による高速な文字列照合, 第 5 回データ工学と情報マネジメントに関するフォーラム (DEIM2013), No. E2-4, 福島県, 2013 年 3 月 3 日.
⑧ Satoshi Yoshida and Takuya Kida: Effective Variable-Length-to-Fixed-Length Coding via a Re-Pair Algorithm, In Proceedings of Data Compression Conference 2013 (DCC 2013), p. 532, Snowbird, Utah, USA, 21 Mar., 2013.
⑨ Kei Sekine, Hirohito Sasakawa, Satoshi Yoshida, and Takuya Kida: Variable-to-Fixed-Length Encoding for Large Texts Using Re-Pair Algorithm with Shared Dictionaries, In Proceedings of Data Compression Conference 2013 (DCC 2013), p.518, Snowbird, Utah, USA, 21 Mar., 2013.
⑩ Satoshi Yoshida and Takuya Kida: Analysis of Multiplexed Parse Trees for Almost Instantaneous VF codes, In Proceedings of the 3rd IIAI International Conference on e-Services and Knowledge Management (IIAI ESKM 2012), pp. 36-41, Fukuoka, Japan, 20 Sep., 2012.
⑪ 吉田論史, 喜田拓也: Re-pair アルゴリズムを用いた効率よい VF 符号, 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM2012), D7-1, 兵庫県, 2012 年 3 月 4 日.
⑫ 吉田論史, 喜田拓也, 効率よい VF 符号化のための分節木を訓練する新手法, 第 10 回情報科学技術フォーラム, No. A-008, 函館, 2011 年 9 月 7 日.
⑬ 吉田論史, 喜田拓也, 効率よい VF 符号のための MDL 原理に基づく分節木の訓練手法, 情報処理学会 第 152 回 データベースシステム・第 103 回 情報基礎とアクセス技術 合同研究発表会, Vol. 2011-IFAT-103 No. 14, 京都府, 2011 年 8 月 3 日.

〔図書〕（計 0 件）

〔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-ikn.ist.hokudai.ac.jp/~kida/publication.html

６．研究組織

(1)研究代表者

喜田　拓也（KIDA TAKUYA）
北海道大学・大学院情報科学研究科・准教授
研究者番号：70343316

(2)研究分担者
なし

(3)連携研究者
なし