様式 C-19

# 科学研究費補助金研究成果報告書

平成２１年　５月２１日現在

研究種目：基盤研究（C）
研究期間：2007～2008
課題番号：19500052
研究課題名（和文）　検索が高速なデータ圧縮法とその高信頼化

研究課題名（英文）　String matching and error control for compressed data

研究代表者
氏　名（ローマ字）：北神　正人　(Kitakami Masato)
所属機関・部局・職：千葉大学・大学院融合科学研究科・准教授
研　究　者　番　号：20282832

研究成果の概要：
本研究は圧縮ファイルを伸長することなく高速に検索してコンピュータウィルスの検索等を効率化する手法に関するものである。PPM 圧縮法に圧縮の過程で生成される文脈情報をヘッダとして付加し、それを検索することにより高速に圧縮データの検索を行う。また、そのヘッダ情報を用いた誤り回復手法も提案した。本手法は静的 PPM に対するものである。元データに一定間隔で特別な記号を挿入してから圧縮する。圧縮データをブロックに分割し、そのパリティを生成する。伸長時には挿入した特別な記号の出現間隔で誤りブロックを特定しパリティで誤りを回復する。

交付額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2007 年度 | 2,100,000 | 630,000 | 2,730,000 |
| 2008 年度 | 1,400,000 | 420,000 | 1,820,000 |
| 年度 | | | |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 3,500,000 | 1,050,000 | 4,550,000 |

研究分野：総合領域
科研費の分科・細目：情報学・計算機システムネットワーク
キーワード：データ圧縮、文字列検索、誤り回復

## １．研究開始当初の背景

データ圧縮技術は記憶容量の節約と通信時間の短縮を目的として現在盛んに利用されている。文書データや実行ファイルなどに対しては無歪圧縮と呼ばれる伸長後のデータが元データと完全に一致する圧縮法がとられる。現在では様々な無歪圧縮法が提案され実用に供せられている。圧縮データは誤りに対する耐性が低く圧縮データに生じたわずかの誤りでも伸長結果の全体に影響が及ぶことが多いため、圧縮データに対する誤り回復手法はこれまで多数提案されている。研究代表者も多数提案してきた。

一方、近年のディスク装置の大容量化や計算機の処理能力の増大により、蓄積されるデータの量が膨大になっている。そのため、必

要なデータ列・文字列の含まれるファイルを高速に特定するデータ検索が必要とされている。また、コンピュータウィルスの脅威が拡大していることより保持しているファイルすべてに対して¥E{定期的にウィルス検査を行う}必要性も増大している。ウィルス検査は通常特定のパターンのデータがファイルに含まれていないか検索することによって行うため高速なファイル検索が求められる。上記の理由より圧縮ファイルに対しても高速な検索を行う必要があるが、圧縮ファイルの一部を伸長するのみで、あるいは全く伸長せずに検索が行えれば検索時間および使用メモリ量の点から好ましい。このような手法をCPM (Compressed Pattern Matching)手法と呼ぶ。CPM手法は数種類の圧縮法に対して提案されており、研究代表者も最近手法を提案している。

研究代表者は上述の圧縮データに対する検索と誤り回復の手法の検討の過程で¥代表的な圧縮手法の本質（圧縮データの構造、その構成要素の間の種々の関係等）の解析をかなり行っている。その視点でCPM手法を見たときに付加情報の有効活用法に関する着想を得た。すなわち、CPM手法では通常圧縮データに対して検索を容易にするために付加情報を挿入する。この情報は伸長時には冗長であるが、伸長の過程で参照することによって伸長過程が正しいか否か判別することができ、付加情報が誤り検出に利用できると考えられる。これにより誤り回復機能を検査情報の追加なし、あるいはわずかな検査情報の追加で実現可能であると考え、本応募課題の着想を得た。

２．研究の目的

本研究の目的は高速に文字列検索が可能であり、圧縮データに誤りが含まれても正しく伸長できるデータ圧縮法を開発することである。対象はテキストデータおよび実行ファイルなどのバイナリデータである。これらのデータを圧縮したデータから、それを伸長するよりも短い時間で与えられた文字列あるいはバイナリデータのパターンが元データに含まれているか否かを判定する手法を既存の圧縮法を改良することによって開発する。さらに、改良の際に付加される情報を利用して圧縮データに生じた¥誤りを回復する手法を開発し、誤りに対する耐性が低いという圧縮データの欠点を克服する。この誤り回復と上述の高速検索技術によるウィルス検査の高速化により圧縮データの信頼性および安全性を高める。

３．研究の方法

①検索が高速な圧縮法の開発

高速な検索法が明らかになっていないデータ圧縮法は多数存在する。そこで、これらに対して高速な検索手法を開発する。具体的には、高速な検索を行うためにどのような情報が必要であるかを検討し、その情報をなるべくサイズが小さくなるように形式を工夫して付加する。付加情報の位置も検索速度に依存するので検討し、適切な位置に挿入する。

②誤り回復手法の検討

既存の高速検索可能な圧縮手法、あるいは上記①で開発した手法に対して検討を行う。付加した情報を用いて伸長過程の整合性を検査する手法を検討することにより誤り検出手法を開発する。その検出手法をさらに発展させ(場合によってはさらに情報を付加して)、誤りの位置とパターンを導出する手法を検討し、誤り回復手法を開発する。

③評価

得られた手法を様々なファイルに適用して提案手法の検索時間、誤り耐性、圧縮率、圧縮伸長時間、使用メモリ量等のデータを採取する。

４．研究成果

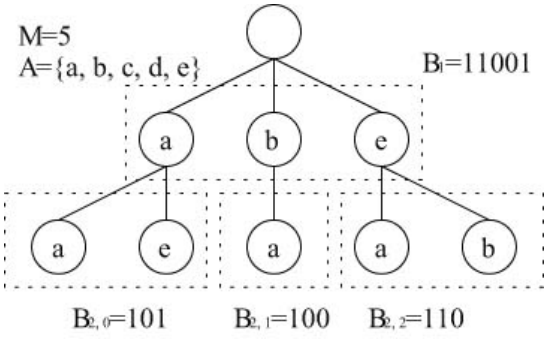
①検索が容易な圧縮法

本研究ではPPM圧縮法に対する検索手法を提案した。PPM 圧縮法はマルコフ過程に基づくモデル化を行って圧縮する手法である。具体的には圧縮対象の文字の直前の一定数の文字列を文脈とし、文脈ごとに皇族文字の生起確率の表（頻度表）を持たせる。文脈に応じた頻度法を元にエントロピー符号化することにより優れた圧縮率を達成する。頻度表および文脈情報は圧縮しながら更新され、伸長の過程でも同一のものが得られるため圧縮情報にこれらの情報を付加する必要はない。

提案手法はこの文脈情報を検索に利用する。すなわち圧縮終了後の文脈情報をヘッダ情報として圧縮データに負荷し、検索はこのヘッダの探索によって行う。文脈情報のデータ構造として木が用いられることが多いので、この木（文脈木）を符号化してヘッダ情報とする。符号化は根から順に広さ優先探索の順に葉以外のノードに対して行う。根は要素数が情報源アルファベットの大きさに等しい２元ベクトルに符号化され、根の子についたラベルの位置の要素のみ″1″となる。太ノードも同様にベクトルに符号化するが、ベクトルの次元は根の子の個数にして、データ量を削減している。これらのベクトルを連接したものがヘッダ情報となる。図に文脈木とその符号化の例を示す。ここでは情報源アルファベットがa,b,c,d,eの５文字としている。

表１に本手法の圧縮率を示す。比較のため、従来のPPMにより圧縮率、検索機能を持つ圧縮法 FM-index（ブロックソート法ベース）

LZ-index（Ziv-Lempel 符号ベース）の圧縮率もあわせて示した。パラメータ $l$ は本手法における文脈木の高さを表している。表より本手法による PPM 圧縮に対する圧縮率の増加は 3～5 ポイント程度であることが分かる。また、既存の検索可能な圧縮法 FM-index, LZ-index よりも圧縮率が優れていることが分かる。表 2 に提案手法の符号化時間を示した。表よりヘッダの生成時間は PPM 圧縮の時間に比べて無視できるほど小さいことがわかる。




図：文脈木とその符号化の例

②誤り回復手法

次に、付加したヘッダ情報を用いて誤りを回復する手法について検討を行った。本手法は文脈木を１パス目で生成し２パス目でそれに変更を加えずに圧縮する静的 PPM に対して行った。

元データに一定間隔で特別な記号を挿入してから圧縮する。圧縮データをブロックに分割し、そのパリティを生成する。伸長時には挿入した記号の出現間隔で誤りブロックを特定しパリティで誤りを回復する。本手法は現在評価中である。評価がまとまり次第論文として発表する。

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）

①Masato Kitakami and Toshihiro Okura, "Dependability Improvement for PPM Compressed Data by Using Compression Pattern Matching" *IEICE Trans. Inf. Syst.*, vol.E91-D, No.10, pp.2435-2439, October 2008.. 査読有

② Masato Kitakami, Bochuan Cai, and Hideo Ito, "A Checkpointing Method with Small Checkpoint Latency" *IEICE Trans. Inf. Syst.*, vol.E91-D, No.3, pp.857-861, March 2008. 査読有

〔学会発表〕（計８件）

① Masato Kitakami and Kensuke Tai, "Lossless Image Compression by PPM-Based Prediction Coding," *Proc. 2009 Data Compression Conference*, pp. 452, March 16-18, 200p, Snowbird UT, USA.

②今野宏, 北神正人, 難波一輝, 伊藤秀男: "インターネット利用システムにおける Integrity の定量的評価," *信学技報*, DC2008-63, December 2008.ディペンダブルコンピューティング研究会（サンライフ萩）

③ Masato Kitakami and Shou-Man Yu, "Router Architecture for Wormhole Switching with Backtracking Capability," *IEICE Technical Report*, FIIS-2008, no.237, June 2008.機能集積情報システム研究会（筑波大学）

④田井健介, 北神正人: "予測符号化にＰＰＭを利用した可逆画像圧縮法," *信学技報*, FIIS-2008, no.231, March 2008. 機能集積情報システム研究会（千葉大学）

⑤Masato Kitakami and Yuta Noguchi, "Error Recovery Method for Multiple-Dictionary Compression Method," *Proc. 2007 Pacific Rim International Symposium on Dependable Computing*, pp. 11-18, Deceomber 17-19, 2007, Melbourne, Victoria, Australia.

⑥片多昭裕, 北神正人, 難波一輝, 伊藤秀男:"インターネット利用遠隔システムの信頼性評価法," *信学技報*, DC2007-60, December 2007. ディペンダブルコンピューティング研究会（下関勤労福祉会館）

⑦大西徹治, 北神正人: "分散ハッシュテーブルにおける Token-Based 相互排他制御," *信学技報*, DE2007-121, DC2007-18, October 2007. ディペンダブルコンピューティング研究会（機械振興会館）

⑧野口雄太, 青木朋之, 北神 正人: "多次元コンテキストを用いた PPM による可逆画像圧縮アルゴリズム," *信学技報*, FIIS-2007, no.211, June 2007. 機能集積情報システム研究会（大分大学）

６．研究組織

(1)研究代表者

北神　正人　（Kitakami Masato）
千葉大学・大学院融合科学研究科・准教授
研究者番号：20282832

(2)研究分担者

(3)連携研究者

表 1: 提案手法の圧縮率 (%).

| ファイル名 | サイズ (バイト) | PPM | | | proposed method | | | FM-index | LZ-index |
|---|---|---|---|---|---|---|---|---|---|
| | | $l=3$ | $l=4$ | $l=5$ | $l=3$ | $l=4$ | $l=5$ | | |
| alice29.txt | 152,089 | 29.24 | 29.42 | 30.46 | 29.24 | 44.70 | 51.98 | 46.75 | 171.64 |
| asyoulik.txt | 125,179 | 31.96 | 32.14 | 33.07 | 31.96 | 49.35 | 57.64 | 50.43 | 181.00 |
| bible.txt | 4,047,392 | 24.68 | 23.27 | 24.25 | 24.68 | 23.74 | 24.91 | 34.81 | 125.18 |
| cp.html | 24,603 | 29.96 | 30.02 | 30.63 | 29.96 | 157.15 | 194.51 | 31.76 | 194.85 |
| E.coli | 4,638,690 | 24.41 | 24.35 | 24.38 | 24.41 | 24.35 | 24.39 | 40.55 | 106.80 |
| fields.c | 11,150 | 28.11 | 27.97 | 28.58 | 28.11 | 194.23 | 362.06 | 29.00 | 211.61 |
| grammar.lsp | 3,721 | 31.49 | 31.41 | 32.05 | 31.49 | 267.41 | 472.30 | 37.87 | 263.80 |
| lcet10.txt | 426,754 | 27.70 | 27.10 | 28.68 | 27.70 | 33.40 | 37.54 | 43.54 | 166.59 |
| plrabn12.txt | 481,861 | 30.73 | 30.73 | 31.95 | 30.73 | 35.84 | 39.35 | 48.70 | 168.73 |
| world192.txt | 2,473,400 | 27.14 | 28.45 | 31.03 | 27.14 | 29.73 | 32.76 | 34.56 | 135.18 |
| xargs.1 | 4,227 | 38.36 | 38.46 | 38.74 | 38.36 | 317.08 | 580.06 | 46.39 | 276.72 |
| english.50MB | 52,428,800 | 30.71 | 30.34 | 31.72 | 30.71 | 30.44 | 31.86 | * | 132.16 |
| english.100MB | 104,857,600 | 30.95 | 30.49 | 31.86 | 30.95 | 30.55 | 31.95 | * | 129.99 |
| english.200MB | 209,715,200 | 30.81 | 30.23 | 31.62 | 30.81 | 30.26 | 31.67 | * | 126.74 |

表中 “*” のケースはプログラムが異常終了して計測できなかった.

表 2: 提案手法の符号化時間 (秒).

| ファイル名 | PPM 圧縮 | | | ヘッダ生成 | | |
|---|---|---|---|---|---|---|
| | $l=3$ | $l=4$ | $l=5$ | $l=3$ | $l=4$ | $l=5$ |
| alice29.txt | 0.768 | 0.992 | 1.502 | 0.017 | 0.020 | 0.024 |
| asyoulik.txt | 0.625 | 0.862 | 1.291 | 0.016 | 0.017 | 0.021 |
| bible.txt | 13.099 | 15.919 | 24.285 | 0.459 | 0.194 | 0.203 |
| cp.html | 0.139 | 0.170 | 0.207 | 0.008 | 0.012 | 0.017 |
| E.coli | 12.435 | 12.500 | 12.708 | 0.234 | 0.227 | 0.188 |
| fields.c | 0.072 | 0.082 | 0.089 | 0.006 | 0.009 | 0.014 |
| grammar.lsp | 0.032 | 0.035 | 0.038 | 0.004 | 0.006 | 0.008 |
| lcet10.txt | 1.482 | 2.091 | 3.199 | 0.031 | 0.033 | 0.039 |
| plrabn12.txt | 1.630 | 2.262 | 3.568 | 0.036 | 0.038 | 0.045 |
| world192.txt | 8.556 | 14.840 | 23.039 | 0.384 | 0.149 | 0.111 |
| xargs.1 | 0.055 | 0.053 | 0.052 | 0.005 | 0.007 | 0.009 |
| english.50MB | 184.059 | 269.597 | 417.765 | 4.220 | 1.827 | 1.872 |
| english.100MB | 373.906 | 543.597 | 842.373 | 8.313 | 3.723 | 8.579 |
| english.200MB | 747.488 | 1,075.834 | 1,677.713 | 17.000 | 7.214 | 15.160 |