様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月２６日現在

研究種目：基盤研究（Ｂ）
研究期間：2007～2009
課題番号：１９３０００００８
研究課題名（和文）
高速圧縮パターン照合に基づく組込み機器向けＸＭＬデータベース基盤技術
研究課題名（英文）
Key Technology for XML DB in Embedded Device
Based on Efficient Compressed Pattern Matching
研究代表者
竹田　正幸（TAKEDA MASAYUKI）
九州大学・大学院システム情報科学研究院・教授
研究者番号：５０２１６９０９

研究成果の概要（和文）：
組込み機器では，メモリやストレージ等の計算資源が乏しいため，従来型の DB 技術では，ローカルな DB をもたせることが難しい．そこで本研究では，独自の高速圧縮パターン照合技術に基づき，組込み機器向けの XML-DB 基盤技術を開発した．

研究成果の概要（英文）：
Standard RDB technique does not perform well in embedded device because of less amount of memory and/or disk storage. Based on our technique of very fast compressed pattern matching, we developed key technique for XML-DB running in embedded devices such as smart phones.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	5,100,000	1,530,000	6,630,000
2008 年度	3,700,000	1,110,000	4,810,000
2009 年度	3,600,000	1,080,000	4,680,000
年度			
年度			
総　計	12,400,000	3,720,000	16,120,000

研究分野：総合領域
科研費の分科・細目：情報学・ソフトウェア
キーワード：アルゴリズム，ＸＭＬ，ＸＭＬストリーム，半構造データ，パターン照合，
データ圧縮，圧縮パターン照合，パターン発見

１．研究開始当初の背景

コンピュータが社会の「どこにでも」存在し，「誰もが」「何時でも」利用可能なユビキタスコンピューティグの時代を迎えている．その主役は，机上や膝上で使う汎用のパーソナルコンピュータではなく，機器に組み込まれ特定の機能を遂行する「組込み型」コンピュータである．このようなコンピュータには，特定の機能を実行する組込み型ソフトウェアが実装されている．高機能化した携帯電話や個人用携帯情報端末(PDA)，自動車に搭載されるカー・ナビゲーション・システムがその典型的な例である．

ユビキタス社会においては，現在のインターネットを幹線とし，そこから毛細血管のように張り巡らされたユビキタスネットワー

クによって，あらゆる組込み機器が接続され，データが頻繁にやり取りされる．したがって，データの検索や格納・管理といったデータベース(DB)技術がユビキタス社会においていっそう重要性を増すことには，疑いの余地がない．

しかし従来技術には次の問題がある．

- 従来のDB技術は大量のメモリとストレージを要するが，組込み機器はそれらの計算資源に乏しい．このため，ローカルにDBをもたせることが難しい．
- 一方，中央のDB管理サーバへオンラインDB問合せを行う場合には，無線ネットワークの通信速度が遅く，問合せの際に遅延が発生し，これがボトルネックとなる．

今後，ネットワークインフラの整備や通信技術の発達に伴い通信速度の向上が期待されるが，その向上は新たな利用を促して通信データ量を爆発的に増大させ，上記の問題はいっそう顕在化するであろう．

上記の２つの問題は，組込み機器上で動作可能な超軽量 DB 技術の開発により，根本的に解決できる．すなわち，組込み機器上の DB 技術により通常利用時はオフライン DB 問合せを行い，データの更新は夜間などにバッチ処理で行う方式が可能となる．この方式には，次の３つの利点がある．

(1) ネットワーク障害発生時においても業務を継続することが可能である．
(2) ネットワーク遅延がないために，高速な応答が得られる．
(3) 無線通信の回数を抑えることで，低コスト環境を実現できる．

反面，データの更新を頻繁に行わないため，必ずしもデータが最新でないという問題が生じる．しかし，データ在庫管理や顧客情報など実世界の多くのデータ交換業務では，データの最新性を常時保証する必要はない．

ユビキタス情報通信におけるデータ交換形式として，XML 形式が標準的になりつつある．XML 形式のデータを扱う XML-DB 技術は，(1) XML 対応の関係 DB 技術，(2)ネイティブ XML-DB 技術の２つに大別される．前者が木構造である XML データを複数の表形式データに分解し関係 DB として格納するのに対し，後者は XML データを木構造としてそのまま格納するため DB システム全体の仕組みもシンプルになる，という利点をもつ．この理由から，後者のネイティブ XML-DB 技術が近年注目を集めており，国内外で盛んに研究されている．ユビキタス機器の観点から見ると，「入力される XML データをそのまま DB 化でき，かつ，DB から検索されたデータをそのまま送信できることによって，計算資源を節約可能」というネイティブ XML-DB の利点は，必要不可欠な要素である．

２．研究の目的

そこで本研究課題では，組込み機器上で動作可能な超軽量ネイティブXML-DB技術を開発する．すなわち「省ストレージ化」「省メモリ化」「省電力化」「機密性保護」といった，組込み機器特有の要件を満たすDB基盤技術を構築する．具体的には，次の５つを実現する．(i) プログラムが使用するメモリ量の削減，(ii) プログラムコード量の削減，(iii) DB サイズの削減，(iv) CPU使用時間の削減，(v) 暗号データ上のパターン照合技術．このうち，(i)によって省メモリ化を達成でき，(ii)(iii)によって省ストレージ化を達成できる．また，(i)-(iv)はすべて省電力化に繋がる．さらに，(v)により機密性保護機能を実現できる．

通常，これらは単なる実装技術の問題として扱われがちであるが，本研究では，これをアルゴリズムの効率化の問題と捉え，理論と実際の両面から，本質的な解決に取り組む．

３．研究の方法

本研究課題は，ユビキタス機器上で動作可能な超軽量 XML-DB 技術の開発を目的とする．すなわち，「省メモリ化」「省ストレージ化」「省電力化」「機密性保護」という，ユビキタス機器独自の要件を満たす DB 技術の確立である．そのために，次の５つの具体的目標を掲げて研究に臨む．

(i)　プログラムが使用するメモリ量の削減．
(ii)　プログラムコード量の削減．
(iii)　DB サイズの削減．
(iv)　CPU 使用時間の削減．
(v)　暗号データ上のパターン照合技術．

通常，(i)-(iv)は実装技術の問題とみなされがちであるが，本研究では，アルゴリズムの効率の問題と捉え，理論と実際の両面から，これに取り組む．その際に鍵となるのが，申請者らが独自に開発した，超高速ストリーム走査技術，圧縮パターン照合技術である．

また，その研究の経験から，圧縮パターン照合技術の発展として(v)の暗号データ上のパターン照合技術を開発することにより，DB ファイルの機密性保護機能を実現できる，という独自の着想を得た．

そこで，本研究では，次の３つを研究項目として研究を遂行する．

A. 超高速ストリーム走査に基づく省メモリ･省電力型 XML データ検索技術.
B. 圧縮パターン照合による省ストレージ型 XML データ格納技術.
C. 軽量圧縮パターン照合によるセキュアな XML データ保護技術.

４．研究成果

ここでは，スペースの都合から，研究項目 B についてのみ報告する．

代表者らはこれまでに世界に先駆けて「テキスト圧縮による高速化」技術を開発している．この研究では，「圧縮パターン照合」の観点から既存の圧縮法の再評価を行い，無名だった BPE 圧縮法に注目し，パターン照合の高速化を達成した．さらなる高速化を図るためには，圧縮パターン照合の観点から有効な圧縮法を新たに開発する必要がある. そこで, 本研究では, BPE 法を拡張した圧縮法を二つ開発し，圧縮率と高速化の両方が劇的に向上することを確認した (Matsumoto 2009; Maruyama 2010).

４．１．Repair+バイトハフマン符号

BPE は, 文脈自由文法変換(Kieffer 2000)に基づく圧縮法のひとつで，Repair (Larsson DCC99)と同様，頻度優先戦略に基づく文法変換によるものである．すなわち，与えられたテキストからそれのみを生成する文脈自由文法を求める際に，頻度の高い部分文字列から非終端記号に置き換えてゆく．Repair との相違点は，テキスト中に生起する終端記号数と開始記号 S を除く非終端記号数との合計が 256 に達した時点で，再帰的な文法の構成を打ち切る点にある．これにより，終端記号と非終端記号がすべて１バイトで表現されることになり，圧縮テキストの処理にビット演算が不要になるという大きな利点がある．この点は，圧縮パターン照合の高速化にとって重要である．一方, BPE では文法の構成を途中で打ち切るため，得られる文法のサイズは比較的大きく，したがって圧縮率もあまり良くない. BPE 圧縮テキスト上のパターン照合では，走査時間の短縮率は，基本的に圧縮率とほぼ同等であるから，圧縮率の向上がさらなる高速化のための鍵といえる．

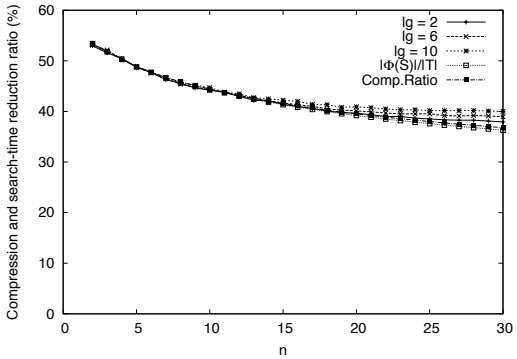
そこで，開始記号以外の非終端記号数の上限 256 を大きい値にすることとし，終端記号と非終端記号の符号化に，1 バイトを符号単位とするハフマン符号を採用する．すると，ハフマン木は 256 分木となり，内部節点数を n とするときの葉の個数，すなわち，表現できる記号数は $255n+1$ となる．ハフマン木の内部節点が根のみ(n=1)のときが BPE 圧縮法である．圧縮率をほかの圧縮法と比較した結果を下表に示す.

	通常の圧縮ツール		
	compress	gzip	bzip2
Medline	42.34	33.29	24.13
Genbank	26.80	21.98	222.71

	Goal2 志向の圧縮ツール				
	BPE	SE	n=10	n=20	n=30
Medline	56.41	66.51	44.42	39.53	36.79
Genbank	31.37	51.54	29.21	29.41	29.74

表より，本手法の圧縮率は，通常の圧縮ツールである gzip や bzip2 には劣るものの，Goal 2 志向の圧縮ツールに比べるとはるかに高いことがわかる．

パラメータ n の値を 2 から 30 まで変化させたときの圧縮率を図に示した(Medline の場合)．また，辞書 D とハフマン木Φを無視した場合の圧縮率である比 $|\Phi(S)|/|T|$ の値も示している．比 $|\Phi(S)|/|T|$ は n の増加につれ単調に増加しているが，圧縮率は必ずしも単調ではない．その理由は，圧縮データは辞書 D とハフマン符号Φを含むがそのサイズは n の増加に伴い増加するからである．



図には，非圧縮テキスト上の KMP アルゴリズムと比べた走査時間短縮率も示してある．走査時間の短縮率は，$|\Phi(S)|/|T|$ にほぼ等しい．しかし，照合に用いる有限状態機械のサイズは，n に比例して増加するため，L2

キャッシュミス率を増加させてしまう．n が大きい場合には，パターン長の増加による速度低下が認められる．

４．２．文脈依存文法変換

上述の手法では，BPE 法の文法規則数の上限を取り払うことによって圧縮率を向上させ，それによって高速化が達成できることが判明した．しかしながら，文法規則をひとつ増やすことは非終端記号をひとつ増やすことを意味しており，その結果，照合に用いる有限状態機械のサイズの増大を招く．そこで，非終端記号の数は BPE と同等に抑えたまま文法規則だけを増加させることを考える．そのために，文法のクラスを文脈自由文法から文脈依存文法(単調文法)へ広げ，生成規則は以下のいずれかの形であるものとする．

$$aA \rightarrow \gamma, \qquad A \rightarrow \gamma$$

ここに，a は終端記号，A は非終端記号であり，γ は非終端記号もしくは終端記号からなり，左辺の長さは右辺の長さを超えないものとする．この文法のクラスを Σ-依存文法と名付けた．さらに，$uA \rightarrow \gamma$ $(u \in \Sigma^n)$ という形の生成規則を許すことでこれを拡張したものを Σ^n-依存文法とよぶ．

終端記号の個数を $|\Sigma|=k$ とする．w を Σ 上の空でない文字列とし，$|w|=n$ とおく．G_s を w に対する最小サイズの Σ−依存文法とし，G_f をそれと等価な最小サイズの文脈自由文法とする．以下の定理が成り立つ．

【定理１】h を G_s の導出木の高さとするとき $|G_f|/|G_s|=O(kh)$.

【定理２】$|G_f|/|G_s|=\Omega((k\log(n/k))/(k+\log(n/k)))$.

【定理３】w を $\Sigma=\{a,b,c\}$ 上の文字列とし，G^*_s をこの w に対する最小サイズの Σ^n−依存文法とする．このとき，以下が成り立つ．

$$|G_f|/|G^*_s|=\Omega(n^{1/3}\log n)/(n^{1/3}+\log n).$$

この新しい文法変換に対して，Repair を拡張した圧縮アルゴリズムを開発した．データセットとして，DBLP, Sources, Pitches を加えて実験を行った．ここで，Sources は linux-2.6.11.6 および gcc-4.00 のソースファイルを連接したファイルであり，Pitches は Web 上から入手した MIDI ファイルから得たピッチ列である．圧縮率を下に示す．

	通常の圧縮ツール		
	gzip	bzip2	Repair
Medline	33.29	23.57	33.83
Genbank	21.98	22.17	31.32
DBLP	17.48	11.66	17.67
Source	23.29	19.79	31.07
Pitches	30.27	35.73	58.23

	Goal 2 志向の圧縮ツール		
	SE	BPE	提案手法
Medline	66.50	56.41	32.94
Genbank	51.74	31.37	28.22
DBLP	70.05	40.83	20.24
Source	71.93	80.54	55.56
Pitches	74.77	78.34	63.36

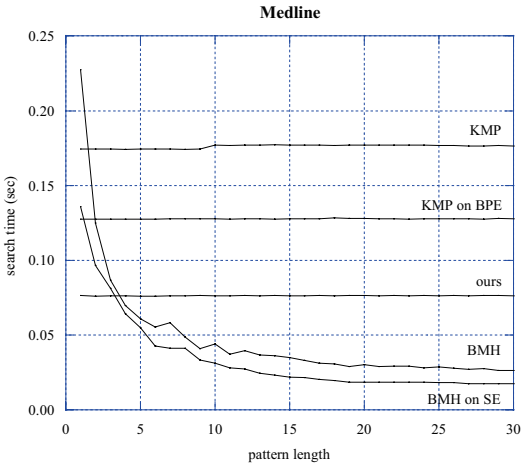
生成規則の上限は $|\Sigma|(256-|\Sigma|)$ となる．そこで，Medline, Genbank, DBLP など $|\Sigma|$ の値が 128 に近いデータに対しては gzip 並みの高圧縮率を達成できる．一方 Sources では，$|\Sigma|=227$ であり圧縮率は良くない．

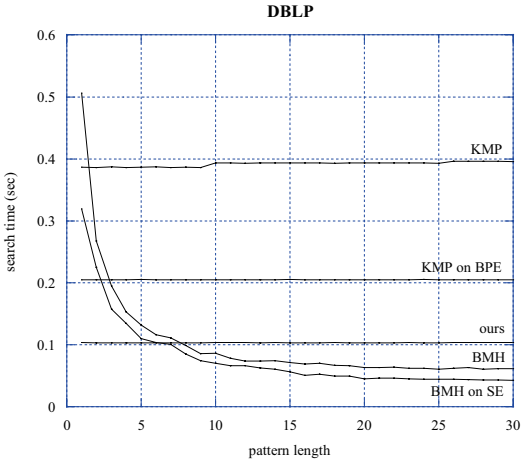
さらに，この圧縮法で圧縮されたテキスト上で動作する圧縮パターン照合アルゴリズムを開発した．速度の比較を図に示す．

以上のように，圧縮率向上とともにパターン照合処理の高速化が達成でき，かつ，有限状態機械のサイズの増大を抑えることができることが判明した．

４．３．携帯機器上での性能評価

上記のプログラムを携帯機器上に移植し，性能評価を行い，本手法の有効性を確認した．





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２９件）

① T. I, S. Inenaga, H. Bannai and M. Takeda, Verifying a Parameterized Border Array in $O(n^{1.5})$ Time, *Proc. the 21st Annual Symposium on Combinatorial Pattern Matching (CPM'10)*, 2010, to appear (査読有).

② S. Maruyama, Y. Tanaka, H. Sakamoto, and M. Takeda, Context-sensitive grammar transform: compression and pattern matching, *IEICE Trans. on Information and Systems*, 2010, to appear (査読有).

③ S. Angelov, S. Inenaga, T. Kivioja, and V. Mäkinen, Finding Missing Patterns, *J. Discrete Algorithms*, 2010, to appear (査読有).

④ T. Nakamura, S. Inenaga, D. Ikeda, K. Baba, and H. Yasuura, Password Based Anonymous Authentication with Private Information Retrieval, *J. Digital Information Management*, 2010, to appear (査読有).

⑤ W. Matsubara, S. Inenaga, and A. Shinohara, An Efficient Algorithm to Test Square-Freeness of Strings Compressed by Balanced Straight Line Programs, Chicago Journal of *Theor. Comput. Sci.*, 2010, to appear (査読有).

⑥ 多比良嘉成, 岸上直也, 田中洋平, 坂本比呂志, データ圧縮の理論に基づく効率的な索引構造, 日本データベース学会論文誌 **8**(3):7-12 (2010)(査読有).

⑦ H. Hyyrö, K. Narisawa, and S. Inenaga, Dynamic Edit Distance Table under a General Weighted Cost Function, *Proc. 36th Int. Conf. on Current Trends in Theory and Practice of Computer Science (SOFSEM2010)*, LNCS 5901, pp. 515-527, 2010 (査読有).

⑧ T. I, S. Deguchi, H. Bannai, S. Inenaga, and M. Takeda, Lightweight Parameterized Suffix Array Construction, *Proc. 20th International Workshop on* Combinatorial Algorithms (IWOCA'09), pp. 312-323, 2009 (査読有).

⑨ T. I, S. Inenaga, H. Bannai and M. Takeda, Counting Parameterized Border Arrays for a Binary Alphabet. *Proc. 3rd Int. Conf. on Language and Automata Theory and Applications (LATA2009)*, LNCS 5457, pp. 422-433, 2009 (査読有).

⑩ S. Inenaga and H. Bannai, Finding Characteristic Substrings from Compressed Texts, *Proc. The Prague Stringology Conference 2009 (PSC 2009),* pp. 40-54, 2009 (査読有).

⑪ K. Hirashima, H. Bannai, W. Matsubara, A. Ishino and A. Shinohara, Bit-parallel algorithms for computing all the runs in a string, *Proc. The Prague Stringology Conference 2009 (PSC 2009)*, pp.203-213, 2009 (査読有).

⑫ W. Matsubara, K. Kusano, H. Bannai and A. Shinohara, A Series of Run-rich Strings, *Proc. 3rd Int. Conf. on Language and Automata Theory and Applications (LATA 2009)*, LNCS 5457, pp. 578-587, 2009 (査読有).

⑬ T. Matsumoto, K. Hagio, and M. Takeda. A Run-Time Efficient Implementation of Compressed Pattern Matching Automata. *Int. J. Found. Comput. Sci.* **20**(4):717-733 (2009)(査読有).

⑭ W. Matsubara, S. Inenaga, A. Ishino, A. Shinohara, T. Nakamura, and K. Hashimoto, Efficient algorithms to compute compressed longest common substrings and compressed palindromes, *Theor. Comput. Sci.* **410**(8-10):900-913 (2009)(査読有).

⑮ W. Matsubara, S. Inenaga, and A. Shinohara, Testing Square-Freeness of Strings Compressed by Balanced Straight Line Program, *Proc. 15th Computing: The Australasian Theory Symposium (CATS2009)*, CRPIT'94, pp.19-28, 2009 (査読有).

⑯ S. Maruyama, Y. Tanaka, H. Sakamoto, and M. Takeda, Context-Sensitive Grammar Transform: Compression and Pattern Matching, *Proc. 15th International Symposium on String Processing and Information Retrieval (SPIRE2008)*, LNCS 5280, pp.27-38, 2008 (査読有).

⑰ S. Deguchi, F. Higashijima, H. Bannai, S. Inenaga, and M. Takeda, Parameterized Suffix Arrays for Binary Strings, *Proc. Prague Stringology Conference 2008*, pp. 84-94, 2008 (査読有).

⑱ W. Matsubara, K. Kusano, H. Bannai, A. Ishino, and A. Shinohara, New Lower Bounds for the Maximum Number of Runs in a String, *The Prague Stringology Conference 2008 (PSC'08)*, pp.140-145, 2008 (査読有).
⑲ T. Matsumoto, K. Hagio, and M. Takeda, A Run-Time Efficient Implementation of Compressed Pattern Matching Automata, *Proc. 13th International Conference on Implementation and Application of Automata (CIAA 2008)*, LNCS 5148, pp.201-211, 2008 (査読有).
⑳ H. Sakamoto, S. Maruyama, T. Kida, and S. Shimozono, A Space-Saving Approximation Algorithm for Grammar-Based Compression, *IEICE Trans. on Information and Systems* **E92-D**(2):158-165 (2008) (査読有).
21 Y. Higa H. Bannai, S. Inenaga, and M. Takeda, Reachability on Suffix Tree Graphs, *Int. J. Found. Comput. Sci.* **19**(1):147-162 (2008) (査読有).
22 W. Matsubara, S. Inenaga, A. Ishino, A. Shinohara, T. Nakamura, and K. Hashimoto, Computing longest common substring and all palindromes from compressed strings, *Int. Conf. on Current Trends in Theory and Practice of Computer Science (SOFSEM'08)*, LNCS 4910, pp.364-375, 2008 (査読有).
23 K. Narisawa, S. Inenaga, H. Bannai, and M. Takeda, Efficient Computation of Substring Equivalence Classes with Suffix Arrays, *Proc. the 18th Annual Symposium on Combinatorial Pattern Matching (CPM'07)*, LNCS 4580, pp.340-351, 2007 (査読有).
24 Y. Nakamura, T. Maita, and H. Sakamoto, Efficient Reachability Test on Directed Graphs and Its Application to Large XML Data, *Proc. 3rd IEEE Int. Workshop on Databases for Next-Generation Researchers (SWOD2007)*, 2007 (査読有).
25 中村 有作, 舞田 哲哉, 坂本 比呂志, 高速な到達可能性判定のための規模耐性の高い索引付け, *DBSJ Letters* **6**(1):77-80 (2007).
26 中村 有作, 舞田 哲哉, 坂本 比呂志, 参照構造を持つ XML 上の高速な到達可能性判定, 人工知能学会論文誌 22(2):191-199 (2007) (査読有).

〔学会発表〕（計５件）

① 松原　渉，圧縮文字列における最長共通部分文字列および回文を求める多項式時間アルゴリズム，コンピュテーション研究会，2008 年 3 月，IBM 東京基礎研究所
② 中村　徹，認証システムのプライバシ保護評価のためのフレームワークの提案，2008 年 暗号と情報セキュリティシンポジウム，2008 年１月，フェニックス・シーガイア・リゾート
③ 石野　明，セキュアな全文検索手法の提案，暗号と情報セキュリティシンポジウム (SCIS2008)，2008 年１月，フェニックス・シーガイア・リゾート

６．研究組織

(1)研究代表者

竹田　正幸（TAKEDA MASAYUKI）
九州大学・大学院システム情報科学研究院・教授
研究者番号：５０２１６９０９

(2)研究分担者

坂本　比呂志（SAKAMOTO HIROSHI）
九州工業大学・情報工学部・准教授
研究者番号：５０３１５１２５

坂内　英夫（BANNAI HIDEO）
九州大学・大学院システム情報科学研究院・准教授
研究者番号：２０３２３６４２

馬場　謙介（BABA KENSUKE）
九州大学・大学院システム情報科学研究院・助教
研究者番号：７０３８０６８７

稲永　俊介（INENAGA SHUNSUKE）
九州大学・大学院システム情報科学研究院・特任准教授
研究者番号：６０４４８４０６

篠原　歩（SHINOHARA AYUMI）
東北大学・大学院情報科学研究科・教授
研究者番号：００２２６１５０

石野　明（ISHINO AKIRA）
東北大学・大学院情報科学研究科・助教
研究者番号：１０３１５１２９
（H19 のみ->H20 連携研究者）