様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年５月２２日現在

機関番号：１７１０２
研究種目：基盤研究（Ｂ）
研究期間：2010～2012
課題番号：２２３０００１０
研究課題名（和文）超高速圧縮データストリーム処理に基づく
軽量ＸＭＬデータベース管理システム基盤技術
研究課題名（英文）Foundational technology for light-weight XML-DBMS based on very fast compressed data stream processing
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）：
本研究では，圧縮データ処理に基づいて軽量 XML データベース管理システム（DBMS）のための基盤技術を確立することを目標とし，主として以下の成果を得た．
(1) 高速で軽量なオンライン文法圧縮アルゴリズムの開発．
(2) 圧縮データ上で動作する q-グラム頻度計算アルゴリズムの開発．
(3) 高速 XML データストリームフィルタリング技術の開発．
この他，DBMS の備えるべき知的データ処理機能として，パターンの効率的な枚挙，分類，オンライン予測等に関する研究を行い，多くの成果を得ている．

研究成果の概要（英文）：
In this project we aimed to establish a foundational technology for lightweight DBMS and successfully developed:
(1) Fast and lightweight online grammar based compression algorithms.
(2) *q*-gram mining algorithms over compressed data.
(3) Efficient filtering algorithms for XML data streams.
Additionally, we tackled some problems such as efficient pattern enumeration, data classification, online prediction, which will be intelligent data processing primitives of DBMS.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	5,400,000	1,620,000	7,020,000
２０１１年度	4,400,000	1,320,000	5,720,000
２０１２年度	4,000,000	1,200,000	5,200,000
年度			
年度			
総　計	13,800,000	4,140,000	17,940,000

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：XML データベース，圧縮データストリーム，アルゴリズム，暗号・認証等，情報工学，ディレクトリ・情報検索，ソフトウェア開発効率化

１．研究開始当初の背景

世界的な不況が継続し，多くの企業が事業見直しや投資抑制を迫られる中にあっても，データの増加はとどまるところを知らず，指

数関数的に増え続けている．経済産業省は，2025 年のインターネット情報流通量は 2006 年の 190 倍に達すると予測しているが，インターネット上に限らず，企業内部で扱うデータもまた増加し続けている．特に日本版 SOX 法が施行された 2008 年度以降は，いざというときに証跡の提出が求められることもあり，保管が必要なデータは増加の一途を辿っている．このように増加し続けるデータを，膨大なコストをかけてデータベース（DB）に格納していくことは，事実上不可能である．そこで，より低コストで，データを格納・管理・検索・分析することのできる新しい DB 基盤技術の開発が，強く期待されている．

２．研究の目的

情報爆発時代とよばれる今日，日々発生し続けるデータを従来のデータベース（DB）によって格納管理することは，もはや限界に達している．そこで本研究では，より低コストで，データを格納・管理・検索・分析することのできる新しい DB 基盤技術を開発する．すなわち「高速な質問処理」「省ストレージ化 」「省メモリ化」「高速な格納」という 4 つの要件を満たす XML-DB 基盤技術を構築する．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．申請者らの有する３つの純国産技術「高速パターン照合技術」「超高速 XML データストリーム処理技術」「圧縮によるデータ処理の高速化」を核に据えて，研究に臨む．

３．研究の方法

本研究課題は，低コストでデータを格納・管理・検索・分析することのできる，軽量 XML-DB 基盤技術の開発を目的とする．すなわち，「高速な質問処理」「省ストレージ化」「省メモリ化」「高速な格納」という４つの要件を満たす DB 基盤技術の確立である．本研究では，これらを実装技術の問題としてではなく，アルゴリズム効率化の問題と捉え，理論と実際の両面から解決に取り組む．その際，次の３つを研究項目として研究を遂行する．

(1) 超高速ストリーム走査に基づく省メモリ型 XML データ高速質問処理．
(2) 圧縮パターン照合に基づく省ストレージ型 XML データ高速質問処理．
(3) 高速オンライン圧縮と圧縮形式変換に基づく XML データ格納技術．

４．研究成果

本研究では，圧縮データ処理に基づく軽量 XML データベース管理システム基盤技術の確立を目指しており，主として以下の成果を得ることができた．

(1) 高速で軽量なオンライン文法圧縮アルゴリズムの開発．
圧縮パターン照合に適した圧縮形式である Straight-Line Program (SLP) をとりあげ，テキストからサイズの小さい SLP を得るための高速軽量な文法圧縮アルゴリズムの開発を行い，既存の文法圧縮アルゴリズムに比べて高い性能を示すことを実証した．

(2) 圧縮データ上で動作する q-グラム頻度計算アルゴリズムの開発．
前年度に引き続き，SLP 上で q-グラム統計を求める問題に取り組んだ．前年度に開発した多項式時間アルゴリズムは q の値が大きいときに速度が低下する問題を抱えていたが，SLP を q グラムの観点から見た本質である「隣接木」という構造を提案し，これを用いることで，q の値によらず高速化することに成功した．これにより，理論と実用の両面から「圧縮による高速化」という目標をも達成したということができる．

(3) 高速 XML データストリームフィルタリング技術の開発．
XML データストリーム処理アルゴリズムにはリアルタイム処理と省スペースが求められる．既存アルゴリズムは，計算時間・領域は効率的であるものの，lazy であるという欠点があった．本研究では，計算量を改善し，かつ，eager に改善することに成功した．

(4) その他．
DBMS の備えるべき知的データ処理機能として，パターンの効率的な枚挙，分類，オンライン予測等に関する研究を行い，多くの成果を得た．

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４９件）すべて査読有

(1) Y. Tabei, Y. Takabatake, H. Sakamoto, A Succinct Grammar Compression, 24th Annual Symposium on Combinatorial Pattern Matching (CPM2013), 2013, to appear.

(2) T.I, Y. Nakashima, S. Inenaga, H. Bannai, and M. Takeda, Efficient Lyndon factorization of grammar compressed text, Proceedings of the 24th Annual Symposium on Combinatorial Pattern Matching (CPM 2013), 2013, to appear.

(3) H. Bannai, P. Gawrychowski, S. Inenaga and M. Takeda, Converting SLP to LZ78 in almost linear time, Proceedings of the 24th

Annual Symposium on Combinatorial Pattern Matching (CPM 2013), 2013, to appear.

(4) T. Tanaka, T.I, S. Inenaga, H. Bannai, and M. Takeda, Computing convolution on grammar-compressed text, Proc. Data Compression Conference 2013 (DCC 2013), 2013, pp.451-460, http://arxiv.org/abs/1303.3945

(5) Y. Tamakoshi, T.I, S. Inenaga, H. Bannai, and M. Takeda, From Run Length Encoding to LZ78 and Back Again, Proc. Data Compression Conference 2013 (DCC 2013), 2013 , pp.143-152.

(6) K. Goto and H. Bannai, Simpler and Faster Lempel Ziv Factorization,In Proc. Data Compression Conference 2013 (DCC 2013), 2013 , pp.133-142, http://arxiv.org/abs/1211.3642

(7) T. Katsura, K. Narisawa, A. Shinohara, H. Bannai, S. Inenaga, Permuted Pattern Matching on Multi-track Strings. SOFSEM 2013, 2013, pp.280-291, 10.1007/978-3-642-35843-2_25.

(8) K. Goto, H. Bannai, S. Inenaga, M. Takeda, Fast q-gram mining on SLP compressed strings. J. Discrete Algorithms 18, 2013, pp.89-99 , 10.1016/j.jda.2012.07.006

(9) S. Maruyama, M. Nakahara, N. Kishiue, H. Sakamoto, ESP-index: A compressed index based on edit-sensitive parsing, J. Discrete Algorithms 18, 2013, pp.100-112, 10.1016/j.jda.2012.07.009

(10) M. Nakahara, S. Maruyama, T. Kuboyama, H. Sakamoto, Scalable Detection of Frequent Substrings by Grammar-Based Compression, IEICE Transactions 96-D(3), 2013, pp.457-464. http://search.ieice.org/bin/summary.php?id=e96-d_3_457

(11) S. Yasutake, K. Hatano, E. Takimoto, M. Takeda, Online Rank Aggregation, Journal of Machine Learning Research - Proceedings Track 25, 2012, pp.539-553, proceedings/papers

(12) S. Maruyama, H. Sakamoto, M. Takeda, An Online Algorithm for Lightweight Grammar-Based Compression. Algorithms 5(2), 2012, pp. 214-235, 10.3390/a5020214

(13) Y. Nakamura, T. Horiike, T. Kuboyama, H. Sakamoto, Extracting research communities from bibliographic data. KES Journal 16(1), 2012, pp.25-34. 10.3233/KES-2012-0230

(14) N. Kobayashi, K. Matsuda, A. Shinohara: Functional programs as compressed data, PEPM 2012, pp.121-130, 10.1145/2103746.2103770

(15) S. Inenaga, H. Bannai, Finding Characteristic Substrings from Compressed Texts, Int. J. Found. Comput. Sci. 23(2), 2012, pp.261-280, 10.1142/S0129054112400126

(16) H. Bannai, T. Gagie, T.I, S. Inenaga, G. M. Landau, M. Lewenstein, An efficient algorithm to test square-freeness of strings compressed by straight-line programs, Inf. Process. Lett. 112(19), 2012, pp.711-714, 10.1016/j.ipl.2012.06.017

(17) S. Yasutake, K. Hatano, E. Takimoto, and M. Takeda, Online Rank Aggregation, Proceedings of the 4th Asian Conference on Machine Learning (ACML 2012), JMLR W&CP 25, 2012, pp.539-553. http://jmlr.org/proceedings/papers/v25/yasutake12/yasutake12.pdf

(18) H. Bannai, S. Inenaga, M. Takeda, Efficient LZ78 Factorization of Grammar Compressed Text. SPIRE 2012, pp.86-98, 10.1007/978-3-642-34109-0_10

(19) K. Hagio, T. Ohgami, H. Bannai, M. Takeda, Eager XPath Evaluation over XML Streams, SPIRE 2012, pp.245-250, 10.1007/978-3-642-34109-0_26

(20) K, Kusano, K. Narisawa, A. Shinohara, Computing Maximum Number of Runs in Strings. SPIRE 2012, pp.318-329, 10.1007/978-3-642-34109-0_33

(21) Y. Nakashima, T.I, S. Inenaga, H. Bannai, M. Takeda, The Position Heap of a Trie. SPIRE 2012, pp.360-371, 10.1007/978-3-642-34109-0_38

(22) Y. Takabatake, Y. Tabei, H. Sakamoto, Variable-Length Codes for Space-Efficient Grammar-Based Compression. SPIRE 2012, pp.398-410, 10.1007/978-3-642-34109-0_42

(23) D. Suehiro, K. Hatano, S. Kijima, E. Takimoto, K. Nagano, Online Prediction under Submodular Constraints. ALT 2012, pp.260-274, 10.1007/978-3-642-34106-9_22

(24) Y. Anan, K. Hatano, H. Bannai, M. Takeda, K. Satoh, Polyphonic Music Classification on Symbolic Data Using Dissimilarity Functions. ISMIR 2012, pp.229-234, event/papers

(25) T.I, Y. Enokuma, H. Bannai, M. Takeda, General Algorithms for Mining Closed Flexible Patterns under Various Equivalence Relations. ECML/PKDD (2), 2012, pp.435-450, 10.1007/978-3-642-33486-3_28

(26) K. Goto, H. Bannai, S. Inenaga, M. Takeda, Speeding Up q-Gram Mining on Grammar-Based Compressed Texts. CPM 2012, pp.220-231, 10.1007/978-3-642-31265-6_18

(27) J. Saito, K. Narisawa, A. Shinohara, Prediction for Control Delay on Reinforcement Learning. ICAART (1), 2012, pp.579-586, http://www.icaart.org/Abstracts/2012/SSML_2012_Abstracts.htm

(28) K. Goto, H. Bannai, S. Inenaga, M. Takeda, Computing q-Gram Non-overlapping Frequencies on SLP Compressed Texts. SOFSEM 2012, pp.301-312, 10.1007/978-3-642-27660-6_25

(29) S. Yoshida, K. Hatano, E. Takimoto, M. Takeda, Adaptive Online Prediction Using Weighted Windows. IEICE Transactions 94-D(10), 2011, pp1917-1923, http://search.ieice.org/bin/summary.php?id=e94-d_10_1917

(30) S. Maruyama, M. Takeda, M. Nakahara, H. Sakamoto, An Online Algorithm for Lightweight Grammar-Based Compression. CCP 2011, pp.19-28, 10.1109/CCP.2011.40

(31) T.I, S. Inenaga, M. Takeda, Palindrome Pattern Matching. CPM 2011, pp.232-245, 10.1007/978-3-642-21458-5_21

(32) S. Yasutake, K. Hatano, S. Kijima, E. Takimoto, M. Takeda, Online Linear Optimization over Permutations. ISAAC 2011, pp.534-543, 10.1007/978-3-642-25591-5_55

(33) S. Angelov, S. Inenaga, T. Kivioja, V. Mäkinen, Missing pattern discovery. J. Discrete Algorithms 9(2), 2011,pp.153-165, 10.1016/j.jda.2010.08.005

(34) T. Nakamura, S. Inenaga, D. Ikeda, K. Baba, H. Yasuura, Password Based Anonymous Authentication with Private Information Retrieval. JDIM 9(2), 2011, pp.72-78, http://www.c.csce.kyushu-u.ac.jp/~toru/paper/toru09_5.pdf

(35) M. Nakahara, S. Maruyama, T. Kuboyama, H. Sakamoto, Scalable Detection of Frequent Substrings by Grammar-Based Compression. Discovery Science 2011, pp.236-246, 10.1007/978-3-642-24477-3_20

(36) T.I, S. Inenaga, H. Bannai, M. Takeda, Verifying and enumerating parameterized border arrays. Theor. Comput. Sci. 412(50), 2011, pp.6959-6981, 10.1016/j.tcs.2011.09.008

(37) T. Yamamoto, H. Bannai, S. Inenaga, M. Takeda, Faster Subsequence and Don't-Care Pattern Matching on Compressed Texts. CPM 2011, pp.309-322, 10.1007/978-3-642-21458-5_27

(38) Y. Anan, K. Hatano, H. Bannai, M. Takeda, Music Genre Classification using Similarity Functions. ISMIR 2011, pp.693-698, papers/PS6-7.pdf

(39) K. Goto, H. Bannai, S. Inenaga, M. Takeda, Fast q-gram Mining on SLP Compressed Strings. SPIRE 2011, pp.278-289, 10.1007/978-3-642-24583-1_27

(40) K. Hagio, T. Ohgami, H. Bannai, M. Takeda, Efficient Eager XPath Filtering over XML Streams. Stringology 2011, pp.30-44, event/2011

(41) T.I, S. Inenaga, H. Bannai, M. Takeda, Inferring Strings from Suffix Trees and Links on a Binary Alphabet. Stringology 2011, pp.121-130, event/2011

(42) K. Shimohira, S. Inenaga, H. Bannai, M. Takeda, Computing Longest Common Substring/Subsequence of Non-linear Texts. Stringology 2011, pp.197-208, event/2011

(43) T.I, S. Inenaga, H. Bannai, M. Takeda, Counting and Verifying Maximal Palindromes. SPIRE 2010, pp.135-146, 10.1007/978-3-642-16321-0_13

(44) K. Kashihara, K. Hatano, H. Bannai, M. Takeda, Sparse Substring Pattern Set Discovery Using Linear Programming Boosting. Discovery Science 2010, pp.132-143, 10.1007/978-3-642-16184-1_10

〔学会発表〕（計３３件）

(1) 立石大悟，バイアス付き Passive Agressive アルゴリズム，第 12 回 情報論的学習理論と機械学習研究会（IBISML），2013.3.4,名古屋工業大学

(2) 藤田隆寛， オフラインアルゴリズムを用いた離散構造のオンライン予測，，冬の LA シンポジウム，2013.1.28,京都大学

(3) Daiki Suehiro, Online Prediction over Base Polyhedra, NIPS 2012 Workshop on Discrete Optimization in Machine Learning (DISCML), 2012.12.7, Lake Tahoe,U.S.A.

(4) Daiki Suehiro, Efficient AUC Maximization by Approximate Reduction of Ranking SVMs, 第１５回情報論的学習理論ワークショップ（IBIS2012），2012.11.7，筑波大学

(5) 末廣大貴，基多面体上のオンライン予測，夏の LA シンポジウム，2012.7.17,

宮津ロイヤルホテル
(6) 松本一成, Bradley-Terry モデルのオンライン予測, 夏の LA シンポジウム, 2012.7.17, 宮津ロイヤルホテル
(7) Daiki Suehiro, Online Prediction under Submodular Constraints, 第９回情報論的学習理論と機械学習研究会(IBISML), 2012.6.19, キャンパスプラザ京都
(8) 阿南陽子, 類似度に基づくポリフォニックな楽曲の分類, 情報処理学会 第94回音楽情報科学研究会(IPSJ-SIGMUS), 2012.2.3, ホテルウェルシーズン浜名湖
(9) 寺岡和紀, モンテカルロ木探索問題に対する効率的サンプリング手法,冬の LA シンポジウム, 2012.1.30, 京都大学
(10) Shota Yasutake, Online Rank Aggregation, NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning (COST), 2011.12.16, Montebajo Basketball Court, Spain
(11) Daiki Suehiro, Approximate Reduction from AUC Maximization to 1-norm Soft Margin Optimization, NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning (COST), 2011.12.16, Montebajo Basketball Court, Spain
(12) 安武翔太, トップ k リストのオンライン予測, 第１４回情報論的学習理論ワークショップ(IBIS2011), 2011.11.9, 奈良女子大学
(13) Daiki Suehiro, Approximate Reduction from AUC Maximization to 1-norm Soft Margin Optimization, 第１４回情報論的学習理論ワークショップ（IBIS2011）, 2011.11.9, 奈良女子大学
(14) 阿南陽子, 類似性指標を用いた楽曲のジャンル分類, 人工知能学会 第 82 回基本問題研究会 (SIG-FPAI), 2011.8.4, 釧路工業高等専門学校
(15) 金城瞬, k-選択多腕バンディット問題, 夏の LA シンポジウム, 2011.7.19, ザヴィラ浜名湖
(16) 安武翔太, Online Prediction over Permutahedron, 情報処理学会 第134回アルゴリズム研究会 (SIGAL), 2011.3.4, 琉球大学
(17) 奥山洋平, 確率的評価値をもつゲーム木における最善手探索,冬の LA シンポジウム, 2011.2.1, 京都大学
(18) 樫原和昭, Sparse Substring Pattern Set Discovery using Linear Programming Boosting, 第９回情報科学技術フォーラム(FIT 2010), 2010.9.7, 九州大学
(19) 安武翔太, "Online Rang Aggregation," 第９回情報科学技術フォーラム(FIT 2010), 2010.9.7, 九州大学

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

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：００２２６１５１

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：２０３２３６４４

瀧本　英二（TAKIMOTO　EIJI）
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：５０２３６３９５

坂本　比呂志（SAKAMOTO　HIROSHI）
九州工業大学・大学院情報工学研究院・教授
研究者番号：５０３１５１２３

畑埜　晃平（HATANO　KOHEI）
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：６０４０４０２６

稲永　俊介（INENAGA　SYUNSUKE）
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：６０４４８４０４

(3)連携研究者

なし