様式 C-19

科学研究費補助金研究成果報告書

平成２３年　４月２５日現在

機関番号：　１０１０１
研究種目：　若手研究（B）
研究期間：　2008～2010
課題番号：　20700001
研究課題名（和文）　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

研究課題名（英文）　Studies on Advanced Pattern Matching over Continuous Data Streams

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

研究成果の概要（和文）：連続データストリームに対する高速・高度なパターン照合技術およびそのためのデータ圧縮技術について研究を行った．前者については，ビットパラレル手法に基づいた，多次元データストリームに対する複雑なクエリを許すパターン照合アルゴリズム BPS (Bit-Parallel on Streams)を提案した．これにより，文字情報のみならず，数値データや分類データなどが複雑に組み合わさったクエリをデータストリームに対して行うことができるようになった．後者については，VF 符号（Variable-to-fixed-length code）に基づいた，パターン照合に適した新規のデータ圧縮法 STVF 符号(Suffix Tree based VF coding)を開発した．この圧縮法は，既存の著名な圧縮法と同程度の圧縮率を達成しながらも，文書中のキーワード検索が高速・簡便に行えるという優れた特徴を持つ．

研究成果の概要（英文）：I have studied high speed and advanced pattern matching over continuous data streams and also about compression technique for realizing that. For the former, I have proposed a pattern matching algorithm, named BPS, which is based on bit-parallel techniques and allows complex queries for multi-dimension data streams. By the algorithm, we can search over data streams for queries that highly combined with numerical data and categorical data as well as text data. For the latter, I have developed a novel data compression method, named STVF coding, which is based on VF coding and suitable for pattern matching. The method has a good feature of allowing doing keyword search in simple and quick manners, as it gains high compression ratios as well as existent well-known compression methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,200,000	360,000	1,560,000
2009 年度	1,300,000	390,000	1,690,000
2010 年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

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

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，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ギガバイトの容量を必要としきている．したがって，効率のよい情報検索技術の開発は必須である．パターン照合技術は情報検索の重要な基本技術の一つであり，これまでにも多くの研究者らによって研究が行われている．

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圧縮して保存されることが多い．そこで，このような圧縮テキストに対して，それを元の文書データに展開することなく文字列の検索（パターン照合）を行う要求が生じた．これに対し，申請者らはこれまで，LZW 圧縮法に対するパターン照合アルゴリズム（Kida 他，1998）や Byte pair encoding（BPE)圧縮法に対するアルゴリズム（Shibata 他，2000）を開発してきた．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た．すなわち，BPE 圧縮法はパターン照合処理を高速化するといえる．

また一方では，XML ファイルや HTML ファイルのような，タグを単位とした木構造を内部表現に持つ半構造データと呼ばれる文書データが急増し，これら半構造データに対して構造を考慮したパターン照合を行う必要が出てきた．これまで，半構造データに対してパターン照合を行うには，一旦元のテキストデータから木構造を抽出しなければならなかった．しかしながら，この方法では非常に時間がかかる上に巨大なメモリを必要とするため，大量のテキストデータに対して実用的ではなかった．これに対して，申請者らは，木構造を抽出することなしに高速にパターン照合を行う手法を開発した（Takeda 他，2002）．

さらには，様々な分野のテキスト情報に関する知識体系がシソーラスや分類階層といった形でデータベース化され手に入るようになってきたことから，テキストデータを単なる文字列として扱うのではなく，それに関する背景知識を考慮にいれた複雑なパターン照合についても取り組んできた(Kida 他，2004～)．すなわち，これはテキストの意味的な構造を考慮したパターン照合の技術である．

このように，申請者は，パターン照合技術に関してその高速化および高度化に取り組んできた．

近年，自動測定技術の発展により，センサーデータや通信記録などの連続データストリームに対する大規模データ処理が重要になっている．こうしたストリーム型のデータに対してパターン照合を行う場合，単純な文字列の照合とは異なる困難さがある．第一に，入力データ系列の各要素は，検索パターンの各要素に対してある程度の誤差を許して一致していればよく，厳密に一致する性質を利用した文字列照合の技術がそのままでは適用できない．第二に，照合処理の過程においてリアルタイム性が要求され，また過去のデータに対してアクセスすることが困難であることが挙げられる．

サンプリングされるデータは実数の場合が多いが，この例では簡単のために整数の列としている．Harada[2002]や Sadri-Zaniolo [2001]らは，よく知られた文字列照合手法である KMP 法および BM 法を関係データストリーム上のパターン照合に適用する手法を提案した．それらの手法では，パターン中の述語間の依存関係を静的に解析することで実行の高速化を図っている．しかしながら，これらは最悪時の時間計算量がパターン長 m とテキスト長 n に対して O(mn)時間であり，Naïve な手法と変らない．また，元にしたアルゴリズムの制約のため，拡張性にも乏しいという欠点がある．

こうした状況から，高速かつ柔軟性が高い，実用に耐えうるアルゴリズムの開発が切望されている．

２．研究の目的

本研究では，連続データストリームに対する高速・高度なパターン照合技術の確立を目指している．具体的には，連続データストリームに適した検索パターンの形式を定式化し，理論的にすぐれた計算量を持つ照合アルゴリズムを開発する．そして実際に実装し，アルゴリズムの性能を実証する．

申請者らは既に，古典的な文字列照合手法である KMP 法や BM 法とは異なる考え方に基づいた，ビットパラレル手法と呼ばれる照合手法を連続データストリーム上のパターン照合に適用し，理論的に効率のよい照合アルゴリズムを一つ得ている(Saito 他, 2007)．今回の研究期間内には，このアルゴリズムを元に以下のような拡張を課題とする．

１）複数の本数のデータストリームが同時に流れてくる状況へ対応する．

２）複数のパターンを同時に照合可能にする．

３）パターンに対して，時系列方向に多少の伸縮を許した一致を可能にする．

こうした拡張により，幅広い実問題への応用が可能となる．また，より複雑なパターンへの拡張について検討し，その理論的解析を行う．最終的には，上述したアルゴリズムを

実際に実装し，オープンに利用可能なパターン照合ライブラリの整備を行う．

３．研究の方法

本研究では，これまで申請者らが研究を進めてきたパターン照合技術に基づき，新たに連続データストリームに対する高速・高度なパターン照合を開発することを目的としている．そのため，以下の二つの項目について研究を行う．

(1) 複数の本数のデータストリームが同時に流れてくる状況への対応．

これまでの研究では，データストリームのモデルとして最も基本的なもの，すなわち，単一の実数値データが時間ごとに入力される場合のみが扱われてきた．気温や震度などの一次元のデータストリームに対してはそれで十分であるが，例えば風向・風速データや多重音声データ，モーションデータのような多次元のデータストリームに対しては一次元用に開発したアルゴリズムを単純には適用できない．また，数値データに伴って属性値データが共にデータストリームとなって入力される場合など，複数種類のデータが同時に流れてくるといった状況が考えられる．このように，より実用的な用途に対して適用するには，多次元データストリームに対する効率よいアルゴリズムが不可欠である．

これに対する最も単純な方法としては，多次元のデータストリームを一次元ごとにパターン照合し，すべての次元においてパターンが一致する部分を検出する方法が考えられる．しかしながら，これでは次元数倍かそれ以上の時間がかかり高速化が望めない．一方で，多次元データを一次元に写像することで照合を行う方法もあるが，入力データの値域が莫大になるばかりか，検索パターンが複雑なものになってしまい，パターン照合に必要となる補助領域が爆発的に増えてしまうという欠点がある．加えて，パターンの述語間の依存関係も複雑なものとなってしまうため，照合速度も低下する．

こうした問題を解決し，多次元データストリームに対する少メモリで効率のよいパターン照合アルゴリズムの開発を目指す．現在のところ，各次元のデータストリームに対するパターン中の述語が他の次元の述語と依存関係をもたない場合には，アルゴリズム BPS を，メモリ使用量を抑えつつ多次元に拡張できることの見通しが立っている．

(2) 連続データストリームに対する複数パターンの同時照合．

一つの連続データストリームに対して，検知したいパターンは大抵の場合に複数個存在するので，複数の検索パターンを同時に照合できると実用的にも都合が良い．複数パターンを同時に照合する文字列照合アルゴリズムとしては Aho-Corasick 照合機械が良く知られているが，これは入力データとパターンとが厳密に一致する場合でしか用いることができない．

最も単純な解決策としては，入力データストリームに対して，パターンの個数分だけ照合アルゴリズムを走らせる方法が考えられる．しかし，これでは当然ながら，パターンの個数に比例して照合時間が増大してしまうという問題が生じる．パターン数が少ない間は使用に耐えうるかもしれないが，実用的には Aho-Corasick 照合機械と同様に，パターン数になるべく依存しない手法が望ましい．アルゴリズム BPS の元となるビットパラレル手法では，複数パターンへ拡張する手法が既に確立されている．ただし，同じようにアルゴリズム BPS を複数パターンへ拡張できるかどうかについてはまだ判っていない．

(3) パターンの正規表現への拡張．

文字列パターンの照合の場合と同様に，ある種の正規表現でもってパターンを記述できることが望ましい．例えば，任意の要素に一致するワイルドカードや，パターンのある部分の繰り返しを記述できるようになるとユーザーの利便性が向上する．ビットパラレル手法自体は非常に拡張性の高いアルゴリズムであるので，アルゴリズム BPS も同様に拡張できることが期待できる．ただし，ビットパラレル手法のパフォーマンスを保ったまま，文字列照合における正規表現と同程度にパターンの記述力を高めることは困難であることが判明している．そこで，実用的な照合速度とメモリ使用量に対して，どこまでパターンの記述力を高められるかについて検討する．

(4) 時系列方向へ伸縮を許したパターン照合．

センサーデータ等の連続データストリームにおいては，変動の仕方が同じ傾向であるとみなせる部分であっても，その時間間隔は一定でないことが多い．すなわち，時系列方向に対してある程度の伸縮を考慮した上でパターンと一致するかどうかを判断できることが望ましい．これを解決するための手法として Dynamic Time Warping (DTW)法がよく知られているが，動的計画法に基づいているために計算コストが高いという欠点がある．したがって，連続データストリームに対するパターン照合に適した，より高速な手法に関して研究を行う．

４．研究成果

H20 年度は，これまでに得られた多次元の数値データストリームに対するパターン照合に，文字列型データストリームや分類階層概念型データストリームに対するパターン照合を組み合わせ，より複雑なクエリに対す

る統合的な照合システムの枠組みを提案した.

一方で，巨大なストリームデータを蓄えるための検索可能なデータ圧縮技術の要求も高まっており，検索効率を保ちつつ圧縮率の高い圧縮法の開発に取り組むことになった.その結果，刈り込み接尾辞木を利用することで，圧縮後の符号語がすべて等しいというパターン照合に適した特徴を備えつつ，既存のハフマン符号などよりも高い圧縮率を得られる STVF 符号化と名付けた新しい圧縮法の開発に成功した.

H21 年度は，その STVF 符号上で効率良くパターン照合を行う手法について，新規なアルゴリズムを考案し，理論的観点から考察した.また，STVF 符号のさらなる改良を行うために，既存手法の調査および詳細な実験を行った.

H22 年度は，STVF 符号の圧縮率を高めるために，辞書木を学習によって強化する手法を提案し，実際に既存の最も良いと考えられている圧縮ツールである gzip 並みに圧縮率を高めることに成功した.また，STVF 符号上での実際の情報検索のパフォーマンスについて総合的な試験を行った.

上記に加えて，本研究成果の各種データストリーム（ブログ，音楽データ，圧縮画像データなど）への技術応用について，検討と試験的な実験を行い，それぞれにおいて実用上の有望な結果を得た.

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 5 件）

① Takashi Uemura, Daisuke Ikeda, Takuya Kida, and Hiroki Arimura, Unsupervised Spam Detection by Document Probability Estimation with Maximal Overlap Method, 査読有, 人工知能学会論文誌, Vol. 26, No. 1, 297-306, Jan. 2011.
② 喜田拓也, 分節木と共用文字列で表現される符号上での効率良い圧縮照合アルゴリズム, 査読有, 電子情報通信学会和文論文誌, Vol.J93-D, No.6, 733-741, Jun. 2010.
③ 喜田拓也, STVF 符号： 頻度刈り込み接尾辞木を用いた効率良い VF 符号化, 査読有, 日本データベース学会論文誌 DBSJ Journal, Vol.8, No.1, 125-130, June 2009.
④ H. Sakamoto, S. Maruyama, T. Kida, S. Shimozono: A space-saving approximation algorithm for grammar-based compression, 査読有, IEICE Trans. on Information and Systems E92-D(2):158-165(2009-2).
⑤ 上村卓史, 喜田拓也, 有村博紀, ウェブ閲覧における効率的なキーワード抽出とその利用, 査読有, 情報処理学会論文誌：データベース(TOD), Vol.38, pp.49-60, 2008 年 6 月.

〔学会発表〕（計 15 件）

① Satoshi Yoshida and Takuya Kida, On Performance of Compressed Pattern Matching on VF Codes, Proc. of Data Compression Conference 2011, p.486, Utah, USA, March 30, 2011.
② Takashi Uemura, Takuya Kida, Satoshi Yoshida, Tatsuya Asai and Seishi Okamoto, Training Parse Trees for Efficient VF Coding, Proc. of the 17th Symposium on String Processing and Information Retrieval (SPIRE2010), LNCS 6393, pp.179-184, Los Cabos, México, October 12, 2010.
③ Satoshi Yoshida and Takuya Kida, An Efficient Algorithm for Almost Instantaneous VF Code Using Multiplexed Parse Tree, In Proc. of Data Compression Conference 2010 (DCC 2010), 219-228, Utah, USA, March 25, 2010.
④ Takuya Kida, Suffix Tree Based VF-Coding for Compressed Pattern Matching, In Proc. Data Compression Conference 2009, IEEE press, p.449, Utah, USA, March 17, 2009.
⑤ Hideyuki Ohtani, Takuya Kida, Takeaki Uno, Hiroki Arimura, Efficient Serial Episode Mining with Minimal Occurrences, Proc. of The 3rd International Conference on Ubiquitous Information Management and Communication (ICUIMC 2009), 471-479, Suwon, Korea, January 16, 2009.
⑥ Takuya Kida, Tomoya Saito, and Hiroki Arimura, Flexible Framework for Time-Series Pattern Matching over Multi-Dimension Data Stream, Proc. the First International Workshop on Algorithms for Large-Scale Information Processing in Knowledge Discovery (ALSIP 2008), in conjunction with PAKDD 2008, 5-16, Hotel Seagull Tempozan, Osaka, May 20, 2008.

〔図書〕（計１件）

① 中野 智晴, 喜田 拓也： JPEG 画像に対する 2 次元近似パターンマッチング, 画像ラボ 日本工業出版, 2009/09/05 発売号 (9 月号), pp.6-11.

〔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