様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 １０ 日現在

機関番号：１３８０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ００３０１

研究課題名（和文）オンライン近似圧縮に基づく次世代ストリームデータマイニング法の開発

研究課題名（英文）Development of Integrated Approximation and Compression Techniques for Next Generation Streaming Data Mining

研究代表者
山本　泰生（Yamamoto, Yoshitaka）

静岡大学・情報学部・准教授

研究者番号：３０５５０７９３

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では，複数のエッジノードから随時到着するストリームデータを受理し，その頻出集合系列を高速かつ省メモリに抽出する手法の開発ならびにその応用課題に取り組んだ．集合系列を扱う問題は，解候補の組み合わせ爆発やリアルタイム処理など，離散構造パターンを扱う問題に共通する難しさと制約を有する基礎問題である．これに対し，オンライン近似圧縮と呼ぶ技術を導入することにより，アルゴリズムのスケーラビリティを改善させた．また提案法によって抽出された圧縮パターンを用いてイベント予測を行う応用課題に取り組んだ．

研究成果の学術的意義や社会的意義
クラウドサービスやIoTの発展に伴い，多くのストリームデータが生み出されている．ストリームデータのインパクトはリアルタイム分析にあるが，他方，大量のデータを高速・省メモリで処理する必要がある．本研究で扱う問題は，組み合わせ爆発やリアルタイム性などオンライン処理を実現するストリームデータマイニングに共通する技術的制約や難しさを含んでおり重要な基礎問題に位置付けられる．本研究を通して，適用困難だった大規模データへのデータマイニング法の可用性が高められ、安価な計算資源でビッグデータの相関分析や時系列解析を行えるようになっている．

研究成果の概要（英文）：In this research, we developed a fast and memory-efficient algorithm for frequent sequential pattern mining from streaming data (FSP-SD). Streaming data analysis is a central issue in many domains. FSP-SD is one of the most fundamental tasks in streaming data analysis dealing with discrete structures. It exhibits two important issues; (1) the real time property to process a huge volume of transactions continuously arriving at high speed and simultaneously output the frequent sequences (FSs); and (2) memory efficiency to enumerate FSs while managing an exponential number of candidates with limited memory resource. We have addressed these two issues based on a novel technique, which is achieved by integrating approximation and compression. Our proposed algorithm and implementation, called PARASOL, is published in Journal of Intelligent Information Systems, and now available freely for academic. We also applied PARASOL to the event prediction problem.

研究分野：知能情報学

キーワード：ストリームデータ　オンラインアルゴリズム　系列予測　頻出パターンマイニング

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

クラウドサービスや IoT の発展に伴い，気象，小売 (POS)，製造，インフラ，観光，医療，スポーツ等の多岐にわたる分野において，多種多様なストリームデータが生み出されている．ストリームデータとは，複数のエッジノードから高速に生成される無限長のデータ系列のことであり，さまざまな応用可能性を持つとされる．　高速に流れ続けるストリームデータを扱う際，時間経過とともに蓄積されるデータ総量は急速に増加し，それに伴い，データ全体の走査処理は低速化する．このため，ストリームデータをリアルタイム分析するようなタスクでは，ディスク上のデータ走査は極力避けながら，到着するデータを逐次的にインメモリ処理することが普通である．このような処理は一般にオンライン処理と呼ばれる．Walmart では毎年，数十億の販売履歴データが生成され，Twitter では毎日，100TB の Web ログデータが蓄積される今日，オンライン処理はストリームデータだけでなく大規模データベースの管理・運用技術としても，ますます不可欠なものになってきている．

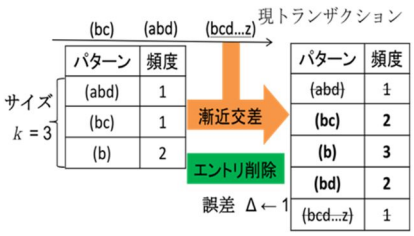
このような背景のなか，研究代表者らは，トランザクションストリーム上の頻出集合を抽出する高速・省メモリなオンラインマイニング法を開発している．トランザクションとはアイテムと呼ばれる説明変数の集合であり，そのストリームは無限に連なるトランザクション列に相当する．頻出集合とは，ストリーム上に頻繁に出現する部分集合のことであり，計算タスクは各時刻の全頻出集合を求めることである．このタスクは，解候補の組み合わせ爆発現象やユーザー要求に対するリアルタイム応答などが課題となる．オンライン処理を実現するストリームデータマイニング全体に関わる技術的制約と難しさを含んでおり，離散構造パターンを扱うデータマイニング問題の基礎問題に位置付けられている．

この問題に対し，研究代表者らはオンライン処理のなかで生じる頻度誤差（Δと書く）を用いて，頻出集合を効率的に圧縮するオンライン近似圧縮法を提案している．提案法は，メモリで管理するパターン数の上限 (k と書く) を定め，各トランザクションを O(kL) で逐次処理することができる (L は最大トランザクション長)．提案法は，Δを許容しながら，任意の頻出集合を復元抽出できる．このような理論的性質を持つ反面，提案法を大規模データで利用するには，可用性をさらに高める必要があった．

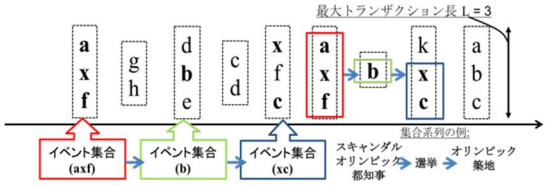
２．研究の目的

そこで本研究では，近似圧縮法の処理性能を改善するデータ構造ならびに枝刈り法を確立し，これをオンライン頻出集合系列マイニング法に拡張する取り組みを推進する．また抽出される頻出集合系列パターンの応用として，ストリームデータ上の時系列分析の課題に取り組む．

本研究のコア技術であるオンライン近似圧縮法は，漸近交差演算とエントリ削除演算の 2 つから構成される（右図）．漸近交差演算では，現トランザクションに対し，メモリ上の各パターンとの共通部分を計算する．得られた共通部分をもとにパターンと頻度を更新する．エントリ削除演算は，各パターンについて，頻度誤差Δ以下の低頻度のものを削除する処理である．漸近交差は飽和アイテム集合族を求める逐次型厳密解法として，2011 年 C.Borgelt らにより提案，再評価された演算である．研究代表者らは，エントリ削除演算を導入することで頻度誤差Δのもとで任意の頻出集合を復元可能な圧縮表現を求められることを明らかにしている．他方，2 つの演算処理を無駄なく効率的に行うための工夫については十分検討されていない．



頻出集合系列とは，ストリーム上に頻繁に出現する部分集合系列のことである．下図のように飛び飛びに出現するイベント系列も対象としており，複数のコンテキストが混在するような多重ストリームデータの知識発見や素性合成，時系列相関分析など応用可能性は高い．他方，解候補の組み合わせ爆発現象が大きな問題となり，特にオンラインアルゴリズムの枠組みでは有効な手法が提案されていないのが現状であった．



３．研究の方法

以上のような背景・目的のもと，本研究では次の 2 つの研究課題に取り組んだ．

(1) 頻出集合マイニングの効率化
木ベースのデータ構造を導入することで漸近交差演算とエントリ削除演算処理の枝刈りを行う．また近似圧縮法の空間使用効率を上げるための工夫を検討する．
(2) 頻出集合系列マイニングへの拡張
漸近交差法を集合系列間の共通極大系列を求める演算に拡張することで，近似圧縮法に基づく頻出集合系列マイニング法を開発する．また抽出された頻出集合系列をストリームデータの素性として利用する新しい系列予測法を開発し，その性能を実証的に評価する．

４．研究成果

(1) 頻出集合マイニングの効率化
Binomial Spanning Tree を拡張した「枝垂れ木」データ構造を導入した．枝垂れ木は，エントリ間の一部の包含関係を低コストで管理することができ，これを用いて漸近交差とエントリ削除の両演算処理を効率化した．作成したソフトウェア (PARASOL) は GitHub 上に公開している．またオンライン頻出集合マイニングに関する研究成果を論文に取りまとめた (Journal of Intelligent Information Systems に採録)．PARASOL は，従来の近似アプローチ「パラメータ指向近似法」と「リソース指向近似法」を融合したハイブリッド型の新しい近似アプローチに基づいており，定数時間計算量のリアルタイム処理を実現している．高い可用性を持つマイニングツールとして今後，利用が増えることを期待している．
近似圧縮法の空間使用効率を改善する課題として，Quantile サマリ計算で用いられている drop & merge 操作に基づくオンラインアルゴリズムを検討し，その実装評価を行なった．結果として，従来法では十分な圧縮効果を得られていなかった稠密データセットにおいて，25%程度までメモリ使用量を削減できることを実証的に確認した．組み合わせ爆発により生じる大量パターンの管理問題は依然大きく，これを解決する新しい取り組みとして，Random Sampling に基づく確率的アプローチを導入する検討を行った．その中で，ストリームデータの圧縮表現（サマリ）を抽出する課題に取り組んだ．

(2) 頻出集合系列マイニングへの拡張
集合系列を対象とするよう漸近交差演算を一般化し，頻出集合マイニングから頻出集合系列マイニングへの拡張を実現した．拡張において 2 つの集合系列間の極大共通部分系列を求める必要があった．この中で入力の系列より構成される 2 次元行列を用いて部分系列を列挙する効率的な列挙アルゴリズムを提案している．
提案法は，頻度が上位 k 個の圧縮集合系列パターンを常に管理している．本研究では，これら頻度上位の系列パターンをストリームデータの素性とみなし，次のイベントをオンライン予測する課題に取り組んだ．現ウィンドウのトランザクション列によって支持される系列パターンを用いて次イベントを予測する手法を実際に実装し，性能評価を行なっている．実験には，Yahoo! Research が提供する Hadoop クラスタのシステムログデータを用いた．イベント系列の予測法として Variable-order Markov Model (VMM) が知られている．本研究では VMM をリファレンスとして提案法の性能評価を行なった．結果として，良好な予測性能を示すことがわかった．VMM モデルは連続系列を対象とするのに対し，提案法は，飛び飛びに出現するイベント系列も予測対象に含まれる．利用したデータは，多数のクライアント IP からのアクセスイベントが入れ子になっており，このデータの特性から実験結果の差異を生じたものと考えられる．このように多くのコンテキストが混在するデータでのイベント予測に対して効果を発揮することがわかった．

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Yoshitaka Yamamoto, Yasuo Tabei, Koji Iwanuma	４．巻 17
２．論文標題 PARASOL: a hybrid approximation approach for scalable frequent itemset mining in streaming data	５．発行年 2019年
３．雑誌名 Journal of Intelligent Information Systems	６．最初と最後の頁 1-29
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10844-019-00590-9	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Takumi Nishina, Koji Iwanuma, Yoshitaka Yamamoto	４．巻 -
２．論文標題 A Skipping FP-Tree for Incrementally Intersecting Closed Itemsets in On-Line Stream Mining	５．発行年 2019年
３．雑誌名 Proc. of BigComp2019	６．最初と最後の頁 1-4
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yoshitaka Yamamoto, Yasuo Tabei, Koji Iwanuma	４．巻 -
２．論文標題 Approximate-Closed-Itemset Mining for Streaming Data Under Resource Constraint	５．発行年 2018年
３．雑誌名 CoRR abs/1901.01710 (2019)	６．最初と最後の頁 1-14
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Takumi Nishina, Koji Iwanuma, Yoshitaka Yamamoto	４．巻 -
２．論文標題 On-Line Approximation Mining for Frequent Closed Itemsets Greater than or Equal to Size K	５．発行年 2018年
３．雑誌名 Proc of BCD2018	６．最初と最後の頁 61-66
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 山本泰生	４．巻 44
２．論文標題 深層知識を獲得するストリームデータマイニングの研究	５．発行年 2017年
３．雑誌名 山梨科学アカデミー会報	６．最初と最後の頁 15-22
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計9件（うち招待講演　1件／うち国際学会　1件）

１．発表者名 山本泰生
２．発表標題 射影積算法による劣線形メンバーシップサマリの構築に向けて
３．学会等名 第127回情報処理学会プログラミング研究会
４．発表年 2020年

１．発表者名 山本泰生，錦戸彩
２．発表標題 確率的メンバーシップサマリの構築に向けて
３．学会等名 第117回知識ベースシステム研究会
４．発表年 2019年

１．発表者名 Koji Iwanuma, Takumi Nishina, Yoshitaka Yamamoto
２．発表標題 Accelerating an On-Line Approximation Mining for Large Closed Itemsets
３．学会等名 IEEE BigData2019（国際学会）
４．発表年 2019年

１．発表者名
山本泰生，岩沼宏治，今井友輝

２．発表標題
半順序ストリームデータのサマリ構築

３．学会等名
人工知能学会知識ベースシステム研究会

４．発表年
2018年

１．発表者名
雨宮晶良，岩沼宏治，谷島健斗，山本泰生

２．発表標題
正負の相関ルールの妥当性の再考察と正負ルールの高速抽出手法

３．学会等名
人工知能学会知識ベースシステム研究会

４．発表年
2018年

１．発表者名
山本泰生

２．発表標題
リソース指向型計算に基づくストリームデータマイニングの研究

３．学会等名
人工知能学会 合同研究会2017 知識ベースシステム研究会（招待講演）

４．発表年
2017年

１．発表者名
谷島健斗，岩沼宏治，山本 泰生

２．発表標題
負の相関ルールマイニングの効率化のための飽和アイテム集合からの極小生成子の高速抽出

３．学会等名
人工知能学会 合同研究会2017 知識ベースシステム研究会

４．発表年
2017年

１．発表者名
仁科拓巳，岩沼宏治，山本泰生

２．発表標題
逆順走査FP木とトライ木を併用した ストリーム上の飽和集合のオンライン抽出

３．学会等名
人工知能基本問題研究会（SIG-FPAI）

４．発表年
2018年

１．発表者名
谷島 健斗，岩沼宏治，山本泰生

２．発表標題
飽和集合上の極小生成子の支持度計算を行わない高速抽出ー負の相関ルール抽出の効率化にむけてー

３．学会等名
人工知能基本問題研究会（SIG-FPAI）

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
公開ソフトウェア PARASOL (ver. 1.00)
https://github.com/Yoshitaka-Yamamoto/parasol
研究代表者HP
http://www.iwlab.org/our-lab/our-staff/yy

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関