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　 ２ 日現在

機関番号：１３９０１
研究種目：研究活動スタート支援
研究期間：2018～2020
課題番号：１８Ｈ０６４６１・１９Ｋ２１５３０

研究課題名（和文）確率的イベントストリームにおけるパターンマイニングのための索引構築に関する研究

研究課題名（英文）Index Construction for Pattern Mining over Probabilistic Event Streams

研究代表者
杉浦　健人（Sugiura, Kento）

名古屋大学・情報学研究科・特任助教

研究者番号：１０８２１６６３

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本課題では，近年注目されている機械学習技術によって得られる確率的イベントストリームを対象に，パターンマイニングなどの分析処理を補助可能な索引構造の開発を目指した．正規表現により柔軟に記述されるパターンを主に扱い，その適切な生起確率を効率的に計算する手法を提案した．また，最新のロックフリー索引について調査及び再現実装を行い，その性能特性や未解決の課題について明らかにし，確率的イベントストリームのための索引構造を開発するための基礎を築いた．

研究成果の学術的意義や社会的意義
機械学習技術は近年大きな注目を集めた一方で，それによって得られる不確実なデータの処理方法は未だ発展途上である．本課題の目指すところは入力データの不確実性を考慮した最終的な分析結果の取得及びその不確実性の算出であり，不確実なデータからの妥当かつ実用的な結果の取得を補助するという意義がある．また，最新の索引構造の再現実装をとおして元論文では述べられていない知見も得ており，新たな未解決課題の提示を行ったという点でも意義がある．

研究成果の概要（英文）：In this research, we aimed to develop an index structure that can assist analytical processing such as pattern mining for probabilistic event streams. We proposed a method for efficiently calculating the appropriate probability of occurrence of patterns described by regular expressions. We also surveyed the state-of-the-art lock-free indexes and re-implemented some of them, and clarified their performance characteristics and unsolved problems. These results provide a basis for developing an index structure for probabilistic event streams.

研究分野：データベース・データ工学

キーワード：確率的イベントストリーム　ストリーム処理　索引構造

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

確率的イベントストリームは各時刻におけるイベントの生起を離散確率分布で表すデータストリームであり，機械学習による解析結果など，本質的に不確実性を含むデータの有効な表現方法である．例えば，スマートフォンのセンシングデータを用いた人の行動モニタリング，監視カメラやドローンの映像を用いた異常イベント検知，サーバログを用いたシステムの状態分析など，確率的イベントストリームを生成する解析処理は数多く行われており，さらなる大規模化が予想される．

一方，確率的イベントストリームに対する分析手法の開発，特に有益な知識を自動で発見するパターンマイニング手法の開発は不十分である．ビッグデータの分析において，具体的に何を発見したいのか分析者自身にもわからないときマイニング処理が用いられる．例えば，行動モニタリング結果の分析であれば，病気や健康につながる習慣的行動が無意識に行われているときパターンマイニングが有効である．しかし，確率的データベースにおいて頻出パターンマイニングを行った既存研究では，各 11 分（時間ステップ数 690）程度の 6 つの移動軌跡から頻出パターンを抽出するのに約 2 時間 30 分も必要としている．つまり，実世界に存在する大規模な確率的イベントストリームにパターンマイニングを適用するには，マイニング手法の大幅な効率化が必要不可欠である．

２．研究の目的

本研究では，効率的なパターンマイニング処理の基盤として，確率的イベントストリームに対する索引構築手法の開発を目的とする．具体的には，「(1) 確率的イベントストリームからのモデル（有限オートマトン）抽出」，「(2) 有限オートマトンからの正規表現索引の構築」，「(3) システムとしての実装及び評価」をサブテーマとして設定し，それぞれの達成を目指す．

３．研究の方法

本課題において実際に取った方法について述べる．

(1) 確率的イベントストリームからの正規表現によるパターン抽出手法の開発：確率的イベントストリームから有限オートマトンをモデルとして抽出するために，正規表現に基づくパターン抽出に取り組んだ．加えて，提案手法を既存の並列分散ストリーム処理システム上で実装するための調査と，不確実性の考慮や高性能な DB の活用によるストリーム処理システムの拡張について取り組んだ．

(2) 最新の索引構造の調査及び再現実装による評価：本課題を取り組むうちに，広く利用されている一般的な索引構造では並列処理による提案手法の高速化を考える上で機能が不十分であることが判明した．そのため，並列（マルチスレッド）処理向けに提案された最新の索引構造について調査し，一部再現実装をとおしてその基礎性能の調査と提案手法への適用に関する検討を行った．

４．研究成果

(1) 確率的イベントストリームからの正規表現によるパターン抽出手法の開発：

確率的イベントストリームからモデルとして有限オートマトンを抽出するために，ある特定の滑り窓（スライディングウィンドウ）内で生起した正規表現パターンの生起確率を効率的に計算する手法を開発した．正規表現などで記述された時系列イベントの生起確率を計算する手法は今までも存在していたが，生起確率の計算対象がある特定のイベント系列のみであり，クリーネ閉包などで記述される曖昧なイベントの生起確率を効率的に計算できないという課題があった．つまり，本質的には同じ時系列イベントを指すと考えられるが，イベント生起の不確実性により系列としては別のものとして認識されるイベント系列を適切に扱えなかった．そのため，与えられた正規表現パターンをそうした曖昧なイベント生起を網羅できる形に変形することで，与えられたパターンの適切な生起確率を効率的に計算する手法を提案した．また，否定表現への対応やスライシング手法の適用による効率化なども行い，柔軟なパターン記述への対応及びさらなる高速化を行った．

また，派生的な研究として，分散並列ストリーム処理システムの調査及びその拡張について検討した．既存のストリーム処理システムでは耐障害性を厳密に保証する仕組みが取られているが，入力の不確実性などを考慮することで，耐障害性を近似的な保証に緩和し処理性能を向上する枠組みを提案した．また，ストリーム処理システム内のステート管理において，最新の DB 技術を使用することでサーバ内の全てのスレッドでステートを共有する枠組みについても提案した．

(2) 最新の索引構造の調査及び再現実装による評価：

まず，提案当初の想定を超えた課題として，膨大なパターンに対する効率的な索引構造の構築が挙げられる．正規表現は実世界のイベントを柔軟に記述できる一方で，パターンの記述には曖昧性があり同じイベントを指すパターンとして複数のパターンが考えられる．加えて，確率的イベントストリームでは情報の不確実性（イベントの生起確率の考慮）を考慮するため扱うパターン数は更に増加する．これらのパターンを網羅的に考慮するには並列処理による高速化が必要となるが，広く用いられている索引構造では索引への書き込み時にロックが取られ，並列処理による恩恵が限定的である．

そのため，近年注目されているロックフリーに基づく同時実行制御を取り入れた索引構造について調査し，一部再現実装を行いその性能及び提案手法で使用する上での有効性を確認した．特に，ロックフリー索引の一つである Bz 木及び内部で使用されている multi-words compare-and-swap アルゴリズムについては再現実装から行い，その性能を調査した．

結果として，ロックフリーに基づく同時実行制御は索引への並列書き込みの性能を向上する十分な潜在性を秘めている一方で，既存のロックフリー索引はその性能を引き出しきれていないことがわかった．これは，個々の書き込みが木構造の葉ノードの変更のみであるためロックフリー化が比較的容易であるのに対し，中間ノードの変更，つまり索引の木構造自体の変更をロックフリー化するのが困難なためである．既存のロックフリー索引は同時に行われている木の構造変更を発見した際にそれを後追いする（全く同じ変更処理を行い，先に処理が終了した方が変更を反映する）ことで明示的なロックの取得を回避している．しかし，後追い処理が極めて少ない命令で済むロックフリーキューなどの場合とは異なり，木の構造変更に必要な処理時間は長く，木構造の変更における後追い処理は本質的には粒度の粗いスピンロックとなってしまっている．つまり，挿入するキーに偏りがなく木構造の全域に渡って変更が発生する場合は並列処理による高速化が見込めるが，変更が伝搬し木構造の根付近の変更が必要となった際や，キーに偏りがありある特定の部分木に集中して変更が発生した場合などに既存のロックフリー索引の性能は大きく減少してしまう．

今後の展望として，今回の調査及び再現実装で得られた知見を基に，より実用なロックフリー索引の構築及びそれに基づく確率的なパターン索引の構築が挙げられる．再現実装をとおして最新のロックフリー索引の構造をおおむね把握できたが，同時に上述した木の構造変更に関する疑問及び改善案も得られた．今後は，得られた知見を基により並列処理に適した索引構造を開発し，それをベースとしたパターン索引の構築について取り組む予定である．

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 高尾 大樹、杉浦 健人、石川 佳治	４．巻 J104-D
２．論文標題 エッジコンピューティングにおける低遅延かつ高信頼度なデータストリームの近似的集約処理	５．発行年 2021年
３．雑誌名 電子情報通信学会論文誌D 情報・システム	６．最初と最後の頁 463～475
掲載論文のDOI（デジタルオブジェクト識別子） 10.14923/transinfj.2020DEP0004	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 Kento Sugiura, Yoshiharu Ishikawa	４．巻 E103.D
２．論文標題 Multiple Regular Expression Pattern Monitoring over Probabilistic Event Streams	５．発行年 2020年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 982～991
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2019DAP0009	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 －

〔学会発表〕　計26件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 倪 天嘉、石川 佳治、杉浦 健人
２．発表標題 機械学習を用いた近似的問合せ処理
３．学会等名 第19回情報科学技術フォーラム
４．発表年 2020年

１．発表者名 笠井 雄太、杉浦 健人、石川 佳治
２．発表標題 3次元TINデータ上での空間的スカイライン問合せ
３．学会等名 第19回情報科学技術フォーラム
４．発表年 2020年

１．発表者名
倪 天嘉、杉浦 健人、石川 佳治

２．発表標題
誤差を保証する近似的問合せについて

３．学会等名
第13回データ工学と情報マネジメントに関するフォーラム

４．発表年
2021年

１．発表者名
笠井 雄太、杉浦 健人、石川 佳治

２．発表標題
TIN上での空間的スカイライン問合せ

３．学会等名
第13回データ工学と情報マネジメントに関するフォーラム

４．発表年
2021年

１．発表者名
高尾 大樹、杉浦 健人、石川 佳治

２．発表標題
エッジコンピューティング環境における低遅延かつ高可用な耐障害性保証

３．学会等名
第13回データ工学と情報マネジメントに関するフォーラム

４．発表年
2021年

１．発表者名
野原 健汰、杉浦 健人、石川 佳治

２．発表標題
マルチバージョン索引構造P-Treeの性能評価

３．学会等名
情報処理学会第83回全国大会

４．発表年
2021年

１．発表者名
西村 学、杉浦 健人、石川 佳治

２．発表標題
不揮発性メモリのための索引手法の分析

３．学会等名
情報処理学会第83回全国大会

４．発表年
2021年

１．発表者名
鈴木 駿也、杉浦 健人、石川 佳治

２．発表標題
機械学習による空間索引の性能評価

３．学会等名
情報処理学会第83回全国大会

４．発表年
2021年

１．発表者名
山本 孝生、石川 佳治、杉浦 健人、朴 秀日、加藤 博和

２．発表標題
都市のサステナビリティ及びレジリエンス分析のためのインタフェースの開発

３．学会等名
情報処理学会第83回全国大会

４．発表年
2021年

１．発表者名
牧田 直樹，杉浦 健人，石川 佳治

２．発表標題
メニーコアシステムにおける分散ストリーム処理システムの性能評価 - 遅延に関する評価 -

３．学会等名
情報処理学会第82回全国大会

４．発表年
2020年

１．発表者名
徳増 直紀，杉浦 健人，石川 佳治

２．発表標題
メニーコアシステムにおける分散ストリーム処理システムの性能評価 - スループットに関する評価 -

３．学会等名
情報処理学会第82回全国大会

４．発表年
2020年

１．発表者名
田中 玲史，杉浦 健人，石川 佳治

２．発表標題
RDBMSによる3D TINデータベース実装手法

３．学会等名
情報処理学会第82回全国大会

４．発表年
2020年

１．発表者名
野田 昌太郎，杉浦 健人，石川 佳治

２．発表標題
多次元データの探索分析のための多様性を考慮した可視化システム

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2020）

４．発表年
2020年

１．発表者名
杉浦 健人，石川 佳治

２．発表標題
並列ストリーム処理システムにおけるDBを用いた内部状態の共有手法

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2020）

４．発表年
2020年

１．発表者名
志村 薫，杉浦 健人，石川 佳治

２．発表標題
データベースのスキーマ情報を活用した機械学習

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2020）

４．発表年
2020年

１．発表者名
高尾 大樹，杉浦 健人，石川 佳治

２．発表標題
チェックポインティングを考慮した近似的耐障害性保証

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2020）

４．発表年
2020年

１．発表者名
Daiki Takao, Kento Sugiura, Yoshiharu Ishikawa

２．発表標題
Approximate Fault Tolerance for Sensor Stream Processing

３．学会等名
The 31st Australasian Database Conference (ADC 2020)（国際学会）

４．発表年
2020年

１．発表者名
高尾 大樹，石川 佳治，杉浦 健人

２．発表標題
センサストリーム処理のための近似的耐障害性保証

３．学会等名
第169回データベースシステム・第136回情報基礎とアクセス技術合同研究発表会

４．発表年
2019年

１．発表者名
野田 昌太郎，杉浦 健人，石川 佳治

２．発表標題
多次元データ分析のための可視化推薦システム

３．学会等名
第18回情報科学技術フォーラム （FIT 2019）

４．発表年
2019年

１．発表者名
志村 薫，杉浦 健人，石川 佳治

２．発表標題
データベースのスキーマ情報を活用した機械学習

３．学会等名
第18回情報科学技術フォーラム （FIT 2019）

４．発表年
2019年

１．発表者名
笠井 雄太，石川 佳治，杉浦 健人

２．発表標題
大規模点群データ分析のためのデータベースの検討

３．学会等名
第81回情報処理学会全国大会

４．発表年
2019年

１．発表者名
杉浦 健人，椎名 健，石川 佳治

２．発表標題
データベース管理システムにおける3D TIN 管理の検討

３．学会等名
第81回情報処理学会全国大会

４．発表年
2019年

１．発表者名
杉浦 健人，石川 佳治

２．発表標題
データストリーム管理システムに関する再考

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム (DEIM 2019)

４．発表年
2019年

１．発表者名
高尾 大樹，石川 佳治，杉浦 健人

２．発表標題
確率モデルに基づく近似的な耐障害性の保証

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム (DEIM 2019)

４．発表年
2019年

１．発表者名
Kento Sugiura, Yoshiharu Ishikawa

２．発表標題
Regular Expression Pattern Matching with Sliding Windows over Probabilistic Event Streams

３．学会等名
The 6th IEEE International Conference on Big Data and Smart Computing (IEEE BigComp 2019)（国際学会）

４．発表年
2019年

１．発表者名
高尾 大樹，石川 佳治，杉浦 健人

２．発表標題
データストリームの集約処理における近似的耐障害性に関する一考察

３．学会等名
第17回情報科学技術フォーラム （FIT 2018）

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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