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　５ 年　 ５ 月 ２９ 日現在

機関番号：１２１０２
研究種目：基盤研究(B)（一般）
研究期間：2019～2022
課題番号：１９Ｈ０４１１４

研究課題名（和文）高水準仮想化機能をもつAugmentedリアルビッグデータ利活用基盤の構築

研究課題名（英文）Research on Augmented Real Big Data Processing Frameworks with High-level Virtualization Facilities

研究代表者

北川　博之（Kitagawa, Hiroyuki）

筑波大学・国際統合睡眠医科学研究機構・教授

研究者番号：００２０４８７６

交付決定額（研究期間全体）：（直接経費）　　13,200,000円

研究成果の概要（和文）：本研究は，ファクトデータとしてデータベースに蓄積されたリアルデータに，AI・ML等によって得られるAugmentedデータをシームレスに統合可能なAugmentedリアルビッグデータ基盤の実現に向けた技術の構築を目的に研究を実施した．成果として，シークエンスデータに対する時系列パターン処理，データベースにおける複合分析処理，ストリーム処理における複合分析処理，境界点検出，外部情報源と知識ベースの統合，ストリーム処理における集約値計算，空間ストリーム処理基盤の各テーマに関して，Augmentedデータが関わるデータ記述，整合性管理，処理効率化等の視点から新たな概念や手法の提案，検証を行った．

研究成果の学術的意義や社会的意義
本研究の多くの研究成果は，査読付きの国内外ジャーナル論文，国際会議論文等で発表済である．特に，「リレーショナルデータベースにおけるAI・ML等による Augmented データ生成を含む複合分析処理」の成果はVLDB Journal，「ストリーム処理における集約値 Augmented データ計算」の成果はIEEE TKDEという，当該分野を代表するトップレベル国際ジャーナルに採択され，学術的に高い評価を得ている．本研究で対象としたAI・ML等によるAugmentedデータ生成・利用は，今後急速に社会へ普及することが予想され，本研究の成果は社会的意義も大きいものと確信する．

研究成果の概要（英文）：This research was conducted with the aim of building a technology for realizing an augmented real big data infrastructure that can seamlessly integrate augmented data obtained by AI, machine learning, etc. with real data accumulated in a database as fact data. As a result, we devised new concepts and methods from the viewpoint of data description, consistency management, and processing efficiency related to augmented data on such topics as time-series pattern processing for sequence data, complex data analysis in databases, complex stream analysis, boundary point detection, integration of external information sources with knowledge bases, aggregate calculation in stream processing, and spatial stream processing infrastructure.

研究分野：データ工学

キーワード：ビッグデータ　Augmentedデータ　仮想化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ビッグデータ処理では，蓄積データやストリーム等の大規模異種データを対象に，結合・集約処理等に加え，AI・機械学習（ML）等を用いた判定・推定処理等の多様な処理を有機的に適用することが求められている．これまで，データベース，ストリーム処理，ビッグデータ分析，ML といったビッグデータ処理基盤技術が研究開発され，実用に供されてきた．しかし，これらの要素技術が立脚するデータ処理パラダイムはそれぞれで大きく異なっており，対象データを選択し，その詳細を理解し，様々な要素技術を自ら組合せることがユーザに求められている．

複雑なデータに対する複合的データ処理を支援する上で，データ構造や処理の詳細を隠ぺいする仮想化技術が極めて重要である．データベース分野では，その初期からこのような仮想化の仕組みが研究されてきた．データベースにおけるビューは，実在のデータベースから導出可能な仮想的データをユーザに提供する．また，実データとルールに基づく演繹推論によって導出されるデータを一体として扱える演繹データベースの研究もある．しかし，急速に進展しつつある AI・ML 等を用いたより高度なデータ拡充・補填技術により得られるデータ（本研究では，Augmented データと呼ぶ）を統合的に扱える技術はまだ未成熟である．

一方，データベース以外のビッグデータ処理一般においては，仮想化の機構は一般に脆弱である．例えば，各種センサストリーム等を対象としたストリーム処理では，センサデバイス等を用いて実世界から直接観測・取得されるデータのうち処理対象とするストリームを選択し，その内容を理解し，どのような処理を行うかをユーザ自身が詳細に指定してしなければ，目的とするデータ処理を実現することができない．

ビッグデータ処理の進展のためには，ビッグデータ処理基盤技術の統合とより高度な Augmented データ処理の仕組みが必要である．

２．研究の目的

本研究は，センサ等で直接観測され，ファクトデータとしてデータベースに蓄積されたリアルデータに，AI・ML 等の高度な拡充・補填処理によって得られるデータを Augmented データとしてシームレスに統合可能な，Augmented リアルビッグデータ基盤の実現に向けた要素技術やシステム技術の構築を目的とする．本研究が目指すデータ基盤では，Augmented データがその源となっている元データと相互関係管理も含めて一体として管理され，Augmented リアルデータ全体の生成，維持，検索，更新等がデータ基盤内で統合的に行なわれる．これにより，以下のようなことが可能となる．(1) 高水準ビッグデータ処理記述：Augmented データを含むデータ処理の高水準な記述が可能となり，ユーザがデータ処理の全ての詳細を記述する負荷がなくなる．(2) 曖昧性や整合性管理：元データと Augmented データの一体管理により Augmented データ生成過程において生じる曖昧性，整合性の管理等が可能となる．(3) 全体処理効率化：Augmented ビッグデータ処理全体が記述されることで，ストリーム処理，データベース処理等の個別の要素技術レベルではなく，Augmented ビッグデータ処理全体を捉えた最適化等の処理効率化が実現できる．

具体的なアプローチとしては，データベース分野をルーツとする SQL が，ビッグデータ分析基盤やストリーム処理基盤においても急速に広がりつつあることに着目し，リレーショナルモデルにオブジェクト指向モデルを融合したオブジェクトリレーショナルモデル等を基盤とした枠組みに基づき，種々のビッグデータ処理を対象した，AI・ML 等に関わるデータ Augmentation を対象に新たな技術を開発する．

３．研究の方法

データベース，ストリーム処理，ビッグデータ分析，ML 等のビッグデータ処理基盤におけるデータ Augmentation に関して，以下の視点から研究を推進する．

○研究項目 A：Augmented リアルビッグデータ基盤アーキテクチャの研究

【研究項目 A-1：データモデル・処理記述系】元データと Augmented データをシームレスに統合するデータモデル，両データを横断的に扱うデータ処理や曖昧性に対応した処理記述．

【研究項目 A-2：Augmented データマッピング方式】Augmented データの定義，導出方法，元データと Augmented データの相互関係管理，メタデータ管理，自動モデル更新等．

【研究項目 A-3：基盤アーキテクチャの設計・実装】上記を踏まえた Augmented リアルビッグデータ基盤システムの設計とプロトタイプ実装．

○研究項目 B：曖昧性や整合性管理機構の研究

Augmented データに関わる曖昧性等をモデル化するための仕組みや，Augmented データの相互依存性，整合性，一貫性を管理するための仕組み．

○研究項目 C：全体処理効率化

【研究項目 C-1：処理最適化】Augmented リアルデータ処理全体を捉えた処理やリソース割当の最適化方式.
【研究項目 C-2：処理高速化】Augmented リアルデータ計算や処理の高速化.
具体的には，以下のような研究トピックを対象に，上記の視点を踏まえた研究を行った.
（1）シークエンスデータに対する時系列パターンオカレンス Augmented データ処理（主要研究項目：A，C）
（2）リレーショナルデータベースにおける AI・ML 等による Augmented データ生成を含む複合分析処理（主要研究項目：A，B，C）
（3）ストリーム処理における AI・ML 等による Augmented データ生成を含む複合分析処理（主要研究項目：A，B，C）
（4）静的データ，ストリームデータに対する境界点検出（主要研究項目：A，C）
（5）外部情報源を Augmented データとして活用可能な知識ベース仮想化（主要研究項目：A，B）
（6）ストリーム処理における集約値 Augmented データ計算（主要研究項目：A，C）
（7）空間ストリーム処理基盤（主要研究項目：A，C）

４．研究成果

（1）シークエンスデータに対する時系列パターンオカレンス Augmented データ処理

(A)フィルタリングによるパターンマッチング処理の効率化

IoT の進展により，時系列データ，ログデータ等の大量のシークエンスデータが生成されている．シークエンスデータ分析における基本操作として，特定のパターンにマッチするサブシークエンス（パターンオカレンス）を発見するパターンマッチングがある．シークエンスデータ分析では，パターンオカレンスは代表的な Augmented データであり，しばしば主たる分析対象となる．リレーショナルデータベースでは，行からなるシークエンスに対して行パターンマッチングを行うための MATCH_RECOGNIZE 句（SQL/RPR）が 2016 年に SQL に導入され，その後，ストリーム処理のための Flink SQL にも導入される等，パターンオカレンスを元データと統合的に扱うための問合せ記述のフレームワークが用意された．しかしながら，その処理の効率化についてはこれまで検討が十分行われていない．最も単純には，パターンマッチングはファイルのスキャンを必要とする処理であり，大規模データではその処理コストが極めて高く，その削減は重要な課題である．

本研究においては，SQL/RPR において，一般に行パターン指定に関わる様々な条件が指定されることに着目し，これらの条件の活用によりマッチングの処理コストを削減するための 2 つの方法を提案した．いずれも，行パターンマッチングを行う前にフィルタリング処理を実行し，パターンマッチング結果に貢献しない無駄なデータを削減することで，行パターンマッチングのコストを削減するものである．1 つ目の方法は，MATCH_RECOGNIZE 句の条件に基づき，パターンオカレンスを 1 つも生成しないシークエンスを削減する手法で，シークエンスフィルタリングと呼ぶ．もう 1 つの方法は，同様に MATCH_RECOGNIZE 句の条件に基づき，シークエンス中で結果に貢献しない行を削除する手法で，行フィルタリングと呼ぶ．本研究では，リレーショナル DBMS の PostgreSQL とビッグデータ処理基盤の Spark SQL を用いて提案するフィルタリング手法を組み込んだ処理システムを実装し，実験によりその有効性を確認した．また，処理コストモデルにより，最適なフィルタリング手法の選択を可能とする手法を考案した.

(B)シークエンス OLAP のためのパターンマッチング

シーケンス OLAP(Online Analytical Proceedings Processing)は，シークエンスデータに対する OLAP 手法の一つである．シーケンス OLAP はシーケンスデータから抽出したパターンオカレンスを対象に，通常の OLAP と同様の OLAP 操作（drill-down，roll-up 等）やパターン OLAP 操作（pattern-drill-down，patten-roll-up 等）を行う．パターン OLAP 操作はシーケンス OLAP 固有のもので，複数パターンの階層構造をたどる操作となる．シークエンスデータがリレーショナルデータベースに行シークエンスとして格納される場合，対象パターンに対するパターンオカレンスとなる行サブシークエンスを発見する行パターンマッチングが必要となる．また，パターン OLAP 操作を可能とするためには，階層関係にある複数のパターンに対するパターンオカレンスを発見すると共に，それらのパターンオカレンス間の親子関係の抽出も必要となる．通常，リレーショナルデータベース上での行パターンマッチングには，リレーション全体のスキャンという高コストの処理が必要である．複数のパターンに対して個別にこの処理を実行するのは非効率的であり，階層関係にある複数のパターンに対する行パターンマッチングを同時に効率よく実行することが望まれる.

上記の通り，SQL では MATCH_RECOGNIZE 句が導入されたが，これは 1 つのパターンのみを対象とする．本研究では，MATCH_RECOGNIZE 句を拡張して階層関係にある複数のパターンを記述可能な MULTI_MATCH_RECOGNIZE 句を導入し，その効率的な実現方法を提案した．具体的には，シークエンス OLAP をサポートするためのパターン階層を定式化し，SP-NFA (Shared Prefix Nondeterministic Finite Automaton)を用いた複数パターンマッチングの同時処理とパターンオカレンス間の親子関係の抽出を実現するアルゴリズムを提案した.

（2）リレーショナルデータベースにおける AI・ML 等による Augmented データ生成を含む複合分析処理

近年のデータ分析では，データの選択・集約・結合等の基本演算に加えて，データ固有の AI・ML モデルやデータコンテンツ処理を活用した複合的データ分析が一般化しつつある．本研究では，このような AI・ML モデル等によって生成される Augmented データを扱うためのモデルを定式化すると共に，Augmented データの曖昧性や整合性に関わる問題としてトレーサビリティに注目し，複合的データ分析のトレーサビリティを向上するための拡張来歴（Augmented Lineage）を提案し，その導出手法について研究を行った．データベース処理のトレーサビリティとしては，分析結果の元になった入力データを提示するデータ来歴がある．しかし，複合的データ分析においては，従来のデータ来歴により元になった入力データを提示するだけでは，どのような根拠で分析結果が導出されたかは理解できない．例えば，機械学習による分類が分析処理含まれる場合には，どのような根拠でその分類結果が得られたかも併せて提示する必要がある．本研究提案の拡張来歴は，そのような要請に対応するものである．

複合的データ解析を記述するためのモデルとして，リレーショナルモデルの基本演算子に UDF（User Defined Function）の適用を記述するための Function 演算子を追加したモデルを定式化した．AI・ML モデルやデータコンテンツ処理は，UDF として記述可能である．本モデルでデータ分析タスクが記述された時，分析結果に含まれる各タプル t に対して拡張来歴を導出することができる．タプル t に対する拡張来歴には，t を導出する元となったデータに加えて，UDF として記述された AI・ML 処理等における結果の導出根拠が併せて含まれる．本研究では，指定されたタプル t に対する拡張来歴を求めるためのアルゴリズムを新たに定式化した．本アルゴリズムは，データベース処理の来歴を求めるために Cui ら（Cui 他 TODS2000）が提案した Tracing Query を用いる手法を大幅に改良，拡張したもので，分析処理の実行後に指定された分析結果の拡張来歴をピンポイントに導出することが可能である．

さらに効率的な拡張来歴の導出方法についても検討を行なった．提案手法で使用する Tracing Query は，分析結果と入力データで共通する属性の値をキーとして元になった入力データを追跡する．この際，分析結果と入力データで共通する属性が存在しない場合は，分析処理の中間結果を経由して来歴を求める手順となる．そのため拡張来歴を導出するためには分析処理の中間結果が必要である．中間結果を作成する際，単純には（i）Tracing Query 実行時に必要な中間結果を分析処理の再実行によって作成する（Rerun），（ii）分析処理の実行時にあらかじめ全ての中間結果をストアしておく Full Materialization (Full)の 2 つの方法が存在する．しかし，前者には分析処理を再実行するため拡張来歴導出に時間がかかるという問題点があり，後者には全ての中間結果をストアするためのストレージコストが必要となるという問題点がある．そこで「AI・ML 等の Augmented データ生成の実行コストは他のリレーショナル演算子の実行コストより多くの場合遥かに大きいため，それらを UDF として実行する Function 演算子の再実行を避けられれば十分に処理時間の短縮が見込める」というアイデアに基づき，分析処理実行時には Function 演算子の中間結果のみあらかじめストアする Function Materialization (FM)を提案した．FM では，その他の中間結果は後から再計算によって作成することで導出時間とストレージコストのトレードオフを図る．

PostgreSQL 上に実装した拡張来歴導出システムを用いた実験により，複合的データ分析に含まれる Augmented データ生成の処理コストの差が拡張来歴導出処理時間に与える影響を検証した．実験から AI・ML 等の Augmented データ導出処理コストが重い時ほど，FM が中間結果ストレージスペースの削減をしつつ拡張来歴導出時間の短縮に大きく寄与することを確認した．

（3）ストリーム処理における AI・ML 等による Augmented データ生成を含む複合分析処理

近年センサや IoT 機器等から多様なリアルタイムデータが生成されており，それらに対する AI・ML 処理等を含めた複合的ストリーム処理も一般化しつつある．そのため，リアルタイムな複合的ストリーム分析において Augmented データが生成される場合の拡張来歴導出も重要である．リアルタイム分析を行うストリーム処理は，分析対象が無限長のデータシーケンスであることや，各オペレータの内部状態が刻々と変化する特徴があるため，上記のデータベースに対する枠組みをストリーム処理に直接適用することはできない．そこで，ストリーム処理において拡張来歴を導出する枠組みを提案した．

本研究のベースとしては，代表的ストリーム処理基盤である Flink において来歴を導出する GeneaLog システム（Palyvos-Giannas 他 Parallel Computing2019）を用いた．GeneaLog は大きく 3 つのアイデアからなる．(i) 分析処理中の生成される全てのタプルに対して，元になったタプルを指すためのポインタ領域を追加し，(ii) 分析処理の各オペレータで出力タプルのポインタ領域に入力タプルをセットする．(iii) 最後に分析結果の来歴は分析結果からポインタを再帰的に辿ることで導出する．

これを踏まえ 3 点拡張を行った．(i) 全てのタプルに判断根拠データへのポインタをさらに

追加.（ii）オペレータで AI・ML 等の判断根拠が必要な処理を行った場合，分析結果の判断根拠ポインタに判断根拠データをセット.（iii）分析結果からポインタを辿る場合は入力タプルへのポインタだけでなく判断根拠ポインタも併せて辿る．以上 3 つの拡張を行うことでストリーム処理に対して拡張来歴を導出する枠組みを提案し，そのプロトタイプシステムを実装した.

プロトタイプシステムを用いた実験では，Amazon レビューデータをベースに作成した分析処理を用いて，拡張来歴導出のとき発生するオーバヘッドをレイテンシとスループットの 2 つの指標で評価した．その結果，分析処理中の AI・ML 等による Augmented データ生成コストが大きい場合は本提案方式による拡張来歴導出のオーバヘッドの割合が小さくなる傾向が認められた．実用的な AI・ML 処理は多くの場合高コストであることが想定される．したがって，本手法により AI･ML 処理を伴う複合的ストリーム分析処理に対して拡張来歴を導出した場合でも，通常処理に対するオーバヘッドの割合は小さいことが示唆される.

（4）静的データ，ストリームデータに対する境界点検出

データセットにおける境界点（Boundary point）とは，密なクラスタの境界領域に位置する点のことである．クラスタ中の中心に近い位置にある点は一般性をもった典型的データであり，コアポイント（Core point）と呼ばれる．一方，クラスタから遠く離れた点は外れ値（Outlier）を表す．境界点は，外れ値とは言えないもののコアポイントとは少し外れた境界的な性質のデータとなる．データ応用では，境界点の検出が有用なケースが存在する．例えば，健康状態を表すデータでは，境界データは疾病を発症するには至っていないものの，そのリスクが高まっている被験者を表す可能性がある．ビッグデータ分析においては，それぞれのデータがコア，境界点，外れ値のいずれかであるかを判別し，その区分を Augmented データとして付与することで，その後のデータ分析の見通しを良くすることが可能である.

本研究では，境界点の検出のための新たな手法を提案した．これまでも境界点検出手法の提案はあるものの既存の手法には種々の問題点がある．例えば，一部の手法では境界点と外れ値を十分区別することができない．したがって，境界点検出の精度は，データセット内の外れ値の存在によって影響を受ける．また，一部の手法は，データセット内の外れ値に関する知識がないと決定が難しいパラメタを含む.

本研究では，これらの問題を解決する新たな境界点検出手法である Boundary Point Factor 手法（BPF）を提案した．BPF では，Gravity と Local Outlier Factor (LOF) を組み合わせて BPF スコアを各点に対して計算し，そのスコアの高い点を境界点として検出する．Gravity は，対象点からその k 最近傍点方向の単位ベクトルの平均ベクトル長として計算される 0 から 1 までの間の値を取る値であり，k 最近傍点が特定の方向に偏って存在する場合に 1 に近い値を，全方向に満遍なく存在する場合に 0 に近い値を取る．LOF（Breunig 他 SIGMOD2000）は良く知られた密度ベースの外れ値検出手法であり，外れ値は大きな LOF スコアを取る一方，それ以外の点は 1 に比較的近い値を取るという性質を有する．点 p の Gravity スコア G(p)，LOF スコア LOF(p) を用いて，BPF スコアは，BPF(p)=G(p)/LOF(p)により求められる．境界点 p，コア点 q，外れ値 r とすると，G(p)>G(q)ならびに LOF(p)，LOF(q)<< LOF(r)が成り立つ可能性が高いため，BPF(p)>BPF(q)，BPF(r)となり，境界点 p の BPF スコアが高くなることが期待される.

StaticBPF は，静的データセット内の全データに対して BPF スコアを計算し，上位 m 個の境界点を検出するアルゴリズムである．StaticBPF の検出精度を，人工データおよび実データを用いた実験により既存手法と比較した結果，StaticBPF が既存方法と比較してより効果的に境界点を検出できることを確認した.

本研究では，さらにストリームに対して境界点検出を行う StreamBPF を開発した．これは，ストリーム処理における直近のウィンドウ内の境界点を連続的に検出する手法であり，グリッド索引と差分計算を組み合わせることで，効率的な計算を実現している.

（5）上記以外のトピックに関する研究成果

上記の他,「外部情報源を Augmented データとして活用可能な知識ベース仮想化」に関しては，RDF 知識ベース中に外部情報源を Augmented データとして統合する手法を提案した．これは，ユーザ定義述語として外部情報源へのアクセスを抽象化し，外部情報源から得たオブジェクトを知識ベース中のエンティティに自動的にリンクするものである．プロトタイプシステムを構築し，エンティティリンキングの精度を評価して，有効性を確認した.「ストリーム処理における集約値 Augmented データ計算」に関しては，遅延して到着するタプルがある場合のストリーム処理集約計算を効率的に実行するためのアルゴリズムを提案し，実験によりその有効性を確認した.「空間ストリーム処理基盤」に関しては，位置情報等の空間情報を有するリアルタイムストリームデータに対して，空間的な距離等の条件に基づく選択・結合処理等を，空間的なグリッド索引を用いて効率的に実行可能な分散ストリーム処理基盤システムの構築を行った．クラスタマシンを用いた実験により，その有効性を確認した.

５．主な発表論文等

〔雑誌論文〕　計20件（うち査読付論文　20件／うち国際共著　1件／うちオープンアクセス　7件）

１．著者名 Savong Bou, Hiroyuki Kitagawa, Toshiyuki Amagasa	４．巻 34
２．論文標題 CPiX: Real-Time Analytics Over Out-of-Order Data Streams by Incremental Sliding-Window Aggregation	５．発行年 2022年
３．雑誌名 IEEE Transactions on Knowledge and Data Engineering	６．最初と最後の頁 5239-5250
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TKDE.2021.3054898	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 Salman Ahmed Shaikh, Hiroyuki Kitagawa, Akiyoshi Matono, Komal Mariam, and Kyoung-Sook Kim	４．巻 10
２．論文標題 GeoFlink: An Efficient and Scalable Spatial Data Stream Management System	５．発行年 2022年
３．雑誌名 IEEE Access	６．最初と最後の頁 24909-24935
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2022.3154063	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 Masaya Yamada, Hiroyuki Kitagawa, Toshiyuki Amagasa, Akiyoshi Maton	４．巻 -
２．論文標題 Augmented Lineage: Traceability of Data Analysis Including Complex UDF Processing	５．発行年 2022年
３．雑誌名 The VLDB Journal	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00778-022-00769-7	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 Vijdan Khalique, Hiroyuki Kitagawa, Toshiyuki Amagas	４．巻 65
２．論文標題 BPF: A Novel Cluster Boundary Points Detection Method for Static and Streaming Data	５．発行年 2023年
３．雑誌名 Knowledge and Information System	６．最初と最後の頁 2991-3022
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10115-023-01854-1	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 Vijdan Khalique and Hiroyuki Kitagawa	4．巻 1
2．論文標題 BPF: An Effective Cluster Boundary Points Detection Technique	5．発行年 2022年
3．雑誌名 Proc. 33rd International Conference on Database and Expert Systems Applications (DEXA 2022)	6．最初と最後の頁 404-416
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-12423-5_31	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 Savong Bou, Toshiyuki Amagasa, Hiroyuki Kitagawa	4．巻 2
2．論文標題 InTrans: Fast Incremental Transformer for Time Series Data Prediction	5．発行年 2022年
3．雑誌名 Proc. 33rd International Conference on Database and Expert Systems Applications (DEXA 2022	6．最初と最後の頁 47-61
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-12426-6_4	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 Salman Ahmed Shaikh, Hiroyuki Kitagawa, Akiyoshi Matono, Kyoung-Sook Kim	4．巻 -
2．論文標題 TStream: A Framework for Real-time and Scalable Trajectory Stream Processing and Analysis	5．発行年 2022年
3．雑誌名 Proc. 30th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 2022 (ACM SIGSPATIAL 2022)	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） 10.1145/3557915.3560964	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 Masaya Yamada, Hiroyuki Kitagawa, Salman Ahmed Shaikh, Toshiyuki Amagasa, Akiyoshi Matono	4．巻 -
2．論文標題 Streaming Augmented Lineage: Traceability of Complex Stream Data Analysis	5．発行年 2022年
3．雑誌名 Proc. 24th International Conference on Information Integration and Web Intelligence (iiWAS2022)	6．最初と最後の頁 224-236
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-21047-1_20	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 Savong Bou, Toshiyuki Amagasa, Hiroyuki Kitagawa, Salman Ahmed Shaikh, Akiyoshi Matono	4．巻 -
2．論文標題 PR-MVI: Efficient Missing Value Imputation over Data Streams by Distance Likelihood	5．発行年 2022年
3．雑誌名 Proc. 24th International Conference on Information Integration and Web Intelligence (iiWAS2022)	6．最初と最後の頁 338-351
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-21047-1_28	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 Kento Miura, Ryohei Kobayashi, Toshiyuki Amagasa, Hiroyuki Kitagawa, Norihisa Fujita, Taisuke Boku	4．巻 -
2．論文標題 An FPGA-based Accelerator for Regular Path Queries over Edge-labeled Graphs	5．発行年 2022年
3．雑誌名 Proceedings of 2022 IEEE International Conference on Big Data (IEEE BigData2022)	6．最初と最後の頁 415-422
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/BigData55660.2022.10020406	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 Takdir, Hiroyuki Kitagawa, Toshiyuki Amagasa	4．巻 -
2．論文標題 Region-based Sub-Snapshot (RegSnap): Enhanced Fault Tolerance in Distributed Stream Processing with Partial Snapshot	5．発行年 2022年
3．雑誌名 Proceedings of 2022 IEEE International Conference on Big Data (IEEE BigData2022)	6．最初と最後の頁 3374-3382
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/BigData55660.2022.10020607	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 Khalique Vijdan、Kitagawa Hiroyuki	4．巻 -
2．論文標題 VOA*: Fast Angle-Based Outlier Detection over High-Dimensional Data Streams	5．発行年 2021年
3．雑誌名 Proc. 25th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2021	6．最初と最後の頁 40-52
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-75762-5_4	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 Yamada Masaya、Kitagawa Hiroyuki、Amagasa Toshiyuki、Matono Akiyoshi	4．巻 -
2．論文標題 Augmented Lineage: Traceability of Data Analysis Including Complex UDFs	5．発行年 2021年
3．雑誌名 Proc. 32nd International Conference on Database and Expert Systems Applications (DEXA2021)	6．最初と最後の頁 65-77
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-86472-9_6	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 Shaikh Salman Ahmed、Kitagawa Hiroyuki	4．巻 8
2．論文標題 StreamingCube: Seamless Integration of Stream Processing and OLAP Analysis	5．発行年 2020年
3．雑誌名 IEEE Access	6．最初と最後の頁 104632-104649
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2020.2999572	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 Dong Yuyang、Xiao Chuan、Chen Hanxiong、Yu Jeffrey Xu、Takeoka Kunihiro、Oyamada Masafumi、Kitagawa Hiroyuki	4．巻 30
2．論文標題 Continuous Top-k Spatial-Keyword Search on Dynamic Objects	5．発行年 2020年
3．雑誌名 The VLDB Journal	6．最初と最後の頁 141-161
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00778-020-00627-4	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 該当する

1．著者名 Shaikh Salman Ahmed、Mariam Komal、Kitagawa Hiroyuki、Kim Kyoung-Sook	4．巻 -
2．論文標題 GeoFlink: A Distributed and Scalable Framework for the Real-time Processing of Spatial Streams	5．発行年 2020年
3．雑誌名 Proc. 29th ACM International Conference on Information and Knowledge Management (CIKM2020)	6．最初と最後の頁 3149-3156
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3340531.3412761	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 中挾晃介，北川博之	4．巻 62
2．論文標題 シーケンスデータに対する行パターンマッチングの効率化	5．発行年 2021年
3．雑誌名 情報処理学会論文誌	6．最初と最後の頁 302-320
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 Nasu Yuya、Kitagawa Hiroyuki、Nakabasami Kosuke	4．巻 11708
2．論文標題 Efficient Row Pattern Matching Using Pattern Hierarchies for Sequence OLAP	5．発行年 2019年
3．雑誌名 Proc. 21st International Conference on Big Data Analytics and Knowledge Discovery (DaWak2019)	6．最初と最後の頁 89-104
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-27520-4_7	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 Nakabasami Kosuke、Kitagawa Hiroyuki、Nasu Yuya	4．巻 11706
2．論文標題 Optimization of Row Pattern Matching over Sequence Data in Spark SQL	5．発行年 2019年
3．雑誌名 Proc. 30th International Conference on Database and Expert Systems Applications (DEXA2019)	6．最初と最後の頁 3-17
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-27615-7_1	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 Bou Savong、Kitagawa Hiroyuki、Amagasa Toshiyuki	4．巻 62
2．論文標題 L-BiX: incremental sliding-window aggregation over data streams using linear bidirectional aggregating indexes	5．発行年 2020年
3．雑誌名 Knowledge and Information Systems	6．最初と最後の頁 3107-3131
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10115-020-01444-5	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14件（うち招待講演　4件／うち国際学会　2件）

１．発表者名
Hiroyuki Kitagawa

２．発表標題
Big Sequence Data Analysis: From Stream Processing Technology to Applications in Sleep Medicine

３．学会等名
IEEE IRI2022（招待講演）（国際学会）

４．発表年
2022年

１．発表者名
大森雄基，北川博之，天笠俊之

２．発表標題
エンティティリンキング機能を有する知識ベースと外部情報源の統合利用手法

３．学会等名
第15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 2023)

４．発表年
2023年

１．発表者名
山田真也，北川博之，Salman Ahmed Shaikh，天笠俊之，的野晃

２．発表標題
複合的ストリーム処理に対するトレーサビリティの研究

３．学会等名
第15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 2023)

４．発表年
2023年

１．発表者名
国生泰資，山田空，堀江和正，阿部高志，北川博之

２．発表標題
リアルタイム性を考慮した自動睡眠ステージ判定システムの設計

３．学会等名
第15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 2023)

４．発表年
2023年

１．発表者名
北川博之

２．発表標題
ストリーム処理の基礎：Velocityへのたゆまざる挑戦

３．学会等名
最強データベース講義シリーズ#9，日本データベース学会（招待講演）

４．発表年
2021年

１．発表者名
大森雄基，北川博之，天笠俊之

２．発表標題
ユーザ定義関数を利用した知識ベースと外部情報源の統合利用手法

３．学会等名
第14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 2022）

４．発表年
2022年

１．発表者名
山田真也，北川博之，天笠俊之，的野晃整

２．発表標題
複合的データ分析処理に対する拡張来歴導出手法と性能評価

３．学会等名
第14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 2022）

４．発表年
2022年

１．発表者名
大森雄基，北川博之，天笠俊之

２．発表標題
知識ベースと外部情報源の統合利用環境

３．学会等名
情報処理学会第84回全国大会（IPSJ全国大会 2022）

４．発表年
2022年

１．発表者名
北川博之

２．発表標題
Computing as a Scienceを担うデータベース研究

３．学会等名
情報処理学会コンピュータサイエンス領域功績賞受賞記念講演,情報処理学会第171回データベースシステム研究会・情報処理学会第140回情報基礎とアクセス技術研究会・電子情報通信学会データ工学研究会合同研究会（招待講演）

４．発表年
2020年

１．発表者名
山田真也，北川博之，天笠俊之

２．発表標題
複合的データ解析を伴う分析処理に対するトレーサビリティの研究

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 (DEIM 2021)

４．発表年
2021年

１．発表者名
大森雄基，北川博之，天笠俊之

２．発表標題
バンディットアルゴリズムとメンション関係を利用した特定トピックに関する特定の地域のツイートの収集

３．学会等名
情報処理学会第83回全国大会 (IPSJ全国大会 2021)

４．発表年
2021年

１．発表者名
Hiroyuki Kitagawa

２．発表標題
Big Data Analytics and Management: Perspectives from Big Sequence Data Analysis and Research Projects in Japan

３．学会等名
The 36th CCF National Database Conference (NDBC2019)（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Carina Miwa Yoshimura, Hiroyuki Kitagawa

２．発表標題
Topic-aware Scheme for Collecting Local Tweets

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 2020)

４．発表年
2020年

１．発表者名
山田真也，天笠俊之，北川博之

２．発表標題
コンテンツ解析を含む大規模データ分析処理に対するトレーサビリティ

３．学会等名
情報処理学会第82回全国大会（IPSJ全国大会 2020)

４．発表年
2020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
研究分担者	天笠　俊之 (Amagasa Toshiyuki) (70314531)	筑波大学・計算科学研究センター・教授 (12102)	
研究分担者	塩川　浩昭 (Shiokawa Hiroaki) (90775248)	筑波大学・計算科学研究センター・准教授 (12102)	
研究分担者	早瀬　康裕 (Hayase Yasuhiro) (40423090)	筑波大学・システム情報系・助教 (12102)	

６．研究組織（つづき）

	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
研究分担者	堀江　和正 (Horie Kazumasa) (60817112)	筑波大学・計算科学研究センター・助教 (12102)	

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