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　２ 年　 ６ 月　 ８ 日現在

機関番号：２５４０３
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００１８８

研究課題名（和文）ストリームデータに対する匿名化処理の情報損失低減と高速化に関する研究

研究課題名（英文）A Study on Reduction of Information Loss and Improvement of Performance of Anonymization Processing for Streaming Data

研究代表者

若林　真一（Wakabayashi, Shin'ichi）

広島市立大学・情報科学研究科・教授

研究者番号：５０２１０８６０

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：ストリームデータに対する匿名化に対し、FDH(Flexible Distance-based Hashing)を改良したδ-FDHに基づく近似最近傍データ探索に基づくk-匿名化手法を提案し、計算機実験により有効性を示した。δ-FDHのハードウェア化について研究し、近似最近傍探索ハードウェアエンジンを開発した。提案エンジンはソフトウェアプログラムと比較して、26倍から56倍の高速化を実現した。FDHの拡張として、複数のアンカー集合を持ち、アンカー集合ごとに近似最近傍探索を行うMA-FDHを提案し、計算機実験により有効性を示した。

研究成果の学術的意義や社会的意義
一般に、ストリームデータの処理は高速性が要求される。また、匿名化は計算負荷の大きい処理である。本研究で得られた研究成果により、ストリームデータに対する、効率がよく、情報損失が少なく、プライバシー保護に対する安全性が強化されたデータ匿名化が可能となる。急速に普及するインターネットショッピングやSNSなどで生成されるインターネット上を伝送される膨大なストリームデータから有用な知識を獲得することで、新しいビジネスの創出が期待できる。

研究成果の概要（英文）：For k-anonymization of stream data, we proposed a k-anonymization method based on δ-FDH, which was an extension of FDH (Flexible Distance-based Hashing) for approximate nearest neighbor search, and showed the effectiveness by computer experiment. We studied hardware implementation of δ-FDH, and developed a hardware engine for approximate nearest neighbor search. The proposed engine realized speedup from 26 to 56 times in comparison with a software program. As an expansion of FDH, we proposed MA-FDH, in which a number of anchor sets were used for approximate nearest neighbor search, and showed the effectiveness by computer experiment.

研究分野：情報工学

キーワード：プライバシー保護　k-匿名化　l-多様性　ストリームデータ　FDH　専用ハードウェア

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1)　近年、ビッグデータの活用に注目が集まっている。特に、インターネットの普及に伴い、急速に普及したインターネットショッピングや SNS などの情報がサーバやデータベースに膨大に蓄積されつつあり、そうしたデータを活用し、データから有用な知識を獲得することで、新しいビジネスの創出が期待されている。一方、こうしたビッグデータの活用においては個人情報の保護は不可欠である。個人情報を適切に保護しつつビッグデータから有用な知識を獲得する技術は一般に PPDM（Privacy Preserving Data Mining）と呼ばれる。特に、個人情報が保護された状態でデータを公開する技術は PPDP（Privacy Preserving Data Publishing）と呼ばれ、近年、急速に研究が進んでいる[2]。

(2)　PPDP の基本となる手法はデータの匿名化である。匿名化には様々な方法が提案されているが、基本となる概念は k-匿名化とℓ-多様性である。PPDP が対象とするデータモデルは、タプル（各属性に対する値の系列）を個別データとするデータテーブルである。k-匿名化とは、タプルから個別識別子（データを一意に特定可能な属性であり、個人の氏名、クレジットカードのカード番号など）を除去し、注目属性（データ解析の対象となる属性）以外のタプルの属性（非注目属性）の値を変換することで、変換されたデータテーブル中に、同じ値を持つ非注目属性の組み合わせを持つタプルが少なくとも k 個以上となるようにデータ変換する操作である。k-匿名化することで、変換されたデータテーブルから、各タプルが帰属する個人（個別識別子）を推定することが困難になる。一方、ℓ-多様性とは、k-匿名化された結果、同じ非注目属性値を持つタプルの集合において、注目属性値が少なくともℓ通り以上あることを保証することである。

(3)　PPDP に対する従来研究の多くは、処理の対象となるデータテーブルがすべて事前に与えられることを仮定している。一方、近年のインターネットの普及に伴い、個人情報を含む膨大なストリームデータが生成されており、こうしたデータを利活用することで、ビッグデータの新しい活用が期待できる[1]。ストリームデータに対する従来研究としては、周期的にバッファに入力データストリームを保管し、バッファ内のデータに対してデータテーブルに対する従来匿名化手法を適用する手法等が知られているが、バッファサイズと情報損失にはトレードオフが存在し、匿名化に伴う情報損失を小さくしようとすると匿名化処理に伴う遅延が増大する、という問題点がある。

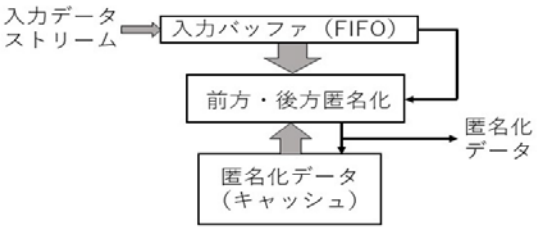
２．研究の目的

(1)　本研究では、ストリームデータに対する匿名化処理における情報損失の低減と高速化に関して研究し、新しい匿名化処理手法を提案する。提案手法においては、匿名化処理におけるデータ管理に LSH（Locality Sensitive Hashing）を応用することで、情報損失の低減と匿名化処理の高速化を同時に実現する匿名化手法を開発する。

(2)　ストリームデータに対する匿名化においては、高速処理が求められることが一般的であることを考慮し、提案手法をハードウェア化することで、匿名化処理の高速化を実現する。

３．研究の方法

(1) 提案匿名化処理手法の概要を次図に示す。入力データストリームは FIFO で構成される入力



バッファに格納され、FIFO 長に相当する遅延を経て、匿名化処理部に入力される。FIFO から出力されたタプルは、２つの手法で匿名化される。ひとつは入力バッファ中のデータを用いた匿名化（前方匿名化という）であり、もう一つはすでに匿名化され、システム中にキャッシュされている過去の匿名化データを用いた匿名化（後方匿名化という）である。匿名化条件は k-匿名化とℓ-多様性である（k とℓの値はユーザー指定）。

(2) 提案手法の中心となる最近傍探索について、ハッシュ探索に基づく近似最近傍探索手法を提案し、実験的に評価すると共に、(1)で示したストリームデータに対する匿名化手法に組み込む。

(3) (2)で提案したストリームデータに対する匿名化手法をワークステーション上にシングルスレッドのプログラムとして実装し、実験的に評価する。実験においては、入力 FIFO 長と情報損失、計算時間の関係等について実験的に評価する。

(4) (2)の近似最近傍探索手法のハードウェア実装を提案し、FPGA 上に実装することで、提案手法の高速化を実現する。

４．研究成果

(1) 本研究では、ストリームデータに対する匿名化処理に対し、データ管理にLSH（Locality Sensitive Hashing）を応用することで、情報損失の低減と匿名化処理の高速化を同時に実現する匿名化手法を開発した。提案手法においては、入力されたデータを一度バッファに格納し、匿名化処理を行った上で、入力順に匿名化データを出力する。バッファ内のデータに対する匿名化処理を前方匿名化、すでに出力された匿名化データに基づくデータの匿名化を後方匿名化と呼ぶ。提案手法ではこれら２つの匿名化をLSHに基づく手法で高速に実行する。本研究においては、LSHの一種であるFDH（Flexible Distance-based Hashing）[3]に基づく近傍データ探索をk-匿名化手法に組み込むことで計算時間の短縮を実現する手法を提案した。提案手法をC言語を用いてプログラムとして実装し、計算機実験によってその有効性を示した。

(2) FDHに最遠δ近傍を導入した新しい近似最近傍データ探索手法δ-FDHを提案した。提案手法はFDHと比較して、効率よく精度の高い近似最近傍データを探索できる。提案手法をプログラムとして実現し、計算機実験により有効性を示した。

(3) δ-FDHに基づく近似最近傍データ探索手法のハードウェア化について研究を行い、パイプライン処理と並列処理に基づく近傍データ探索ハードウェアを開発した。開発したハードウェアはFPGA上の専用回路として実現し、回路規模と計算時間の高速化についてシミュレーション実験により評価を行った。評価実験の結果、シングルCPUで実現された近似最近傍データ探索プログラムと比較して、提案ハードウェアは26倍から56倍の高速化を実現した。

(4) FDHの改良として、マルチアンカーFDH（Multi-Anchor FDH, MA-FDH）を提案した。通常の

FDHが１個のアンカー集合に基づいて多次元空間を複数の領域に分割し、与えられたクエリに近い領域を探索領域として最近傍探索を行うのに対し、MA-FDHでは複数のアンカー集合を用意し、アンカー集合ごとに多次元空間を領域に分割し、与えられたクエリに対して、アンカー集合ごとのそれぞれの近傍領域において最近傍探索を行うことで、１つのアンカー集合を用いる場合よりもよりよい近似解を得ようとするものである。計算機実験の結果、提案手法の有効性を示した。

(5) FDHを拡張したFDHQ (Flexible Distance-based Hashing with Quadrisection) を提案した。FDHが探索領域を定義する各アンカーに対してデータ空間を2分割するのに対し、FDHQはデータ空間を4分割することで探索対象データの絞り込みと部分空間ごとのデータ数の均等化を実現した。また、FDHQのアンカー集合を複数にすることで計算時間と正答率のトレードオフの柔軟な制御を可能にする手法を提案し、計算機実験により有効性を示した。

(6) FDHQに基づくストリームデータに対するk-匿名化手法を提案した。提案手法においては、ℓ-多様性を考慮したk-匿名化も実現できるようにした。提案手法をプログラムとして実装し、計算機実験により提案手法の有効性を示した。

(7) 今後の課題としては、本研究で提案したストリームデータに対するk-匿名化手法に対する並列処理の導入による処理時間の短縮、数値データと非数値データを共に属性として持つストリームデータに対するk-匿名化への提案手法の拡張等が上げられる。

＜引用文献＞

[1] Esmaeil Mohammadian, Morteza Noferesti, Rasool Jalili, “Fast Anonymization of Big Data Streams”, Proc. BigDataScience’14, 2014.
[2] 佐久間淳, データ解析におけるプライバシー保護, 講談社, 2016.
[3] Man Lung Yiu, Ira Assent, Christian S. Jensen, and Panos Kalnis, “Outsourced similarity search on metric data assets”, IEEE Trans. Knowledge and Data Engineering, Vol.24, No.2, pp.338-352, 2012.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Yuri Itotani, Shin’ichi Wakabayashi, Shinobu Nagayama, and Masato Inagi
２．発表標題 An Approximate Nearest Neighbor Search Algorithm Using Distance-Based Hashing
３．学会等名 29th International Conference on Database and Expert Systems Applications (DEXA 2018)（国際学会）
４．発表年 2018年

１．発表者名 Toshitaka Ito, Yuri Itotani, Shin’ichi Wakabayashi, Shinobu Nagayama, Masato Inagi
２．発表標題 A Nearest Neighbor Search Engine UsingDistance-based Hashing
３．学会等名 2018 International Conference on Field-Programmable Technology (FPT)（国際学会）
４．発表年 2018年

１．発表者名 糸谷友里，森啓輔，若林真一，永山忍，稲木雅人，上土井陽子
２．発表標題 Flexible Distance-based Hashingに基づく大規模多次元データ集合に対する近似最近傍探索手法の改良
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
４．発表年 2019年

１．発表者名 糸谷友里、若林真一、永山忍、稲木雅人
２．発表標題 距離に基づくハッシングを用いた高次元ストリームデータに対する効率の良いk-匿名化手法
３．学会等名 IEEE広島支部学生シンポジウム
４．発表年 2017年

１．発表者名
坂田奈々子、上土井陽子、村上頼太、若林真一

２．発表標題
動的データセットにおけるプライバシ保護の厳密な安全性評価と妥当な安全性評価について

３．学会等名
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
2018年

１．発表者名
Toshitaka Ito, Yuri Itotani, Shin’ichi Wakabayashi, Shinobu Nagayama, Masato Inagi

２．発表標題
An FPGA-based Nearest Neighbor Search Engine Using Distance-based Hashing for High-Dimensional Data

３．学会等名
Workshop on Synthesis And System Integration of Mixed Information technologies（国際学会）

４．発表年
2018年

１．発表者名
山吉勇輝，若林真一，上土井陽子

２．発表標題
高次元データに対する近似最近傍探索手法とストリームデータに対するk-匿名化への応用

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

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	上土井　陽子 (Kamidoi Yoko) (80264935)	広島市立大学・情報科学研究科・講師 (25403)	