日本データベース学会和文論文誌
Vol. 18-J, Article No. 6, 2020 年 3 月

非順序型データベースエンジンを用いた大規模データの対話的非特定化手法の性能評価

西川 記史 [1]　礒田 有哉 [2]　茂木 和彦 [3]
清水 晃 [4]　早水 悠登 [5]　合田 和生 [6]
喜連川 優 [7]

近年，情報化社会の進展に伴い，様々な機関が保有するデータを有効活用した施策の立案や，新たなサービスの創出が求められている．データの利活用の推進には，データの有用性を維持しつつ再識別のリスクを低減するための対話的なデータ加工プロセスが不可欠であり，当該データ加工に要する時間の短縮が課題となっている．著者らは，データ加工に要する時間の短縮を目的に，非順序型実行原理を用いたデータ加工方法を提案している．本研究では，データ加工における非順序型実行原理の効果を実証すべく，その性能特性を明らかにする．模擬データによる実験結果から，データの特性や非特定化の要件の差異による性能特性を明らかにでき，データ加工に要する時間を短縮できる見込みを得た．これにより，データを有効活用した施策の立案や新たなサービスの創出が期待できる．

1　はじめに

近年，情報化社会の進展に伴い，地方自治体や公共機関，医療機関，民間企業などが保有するさまざまなデータを有効活用した施策の立案や，新たなサービスの創出が求められている [1] [2] [3] [4]．例えば，自動車の自動運転技術，産業における歩留り向上技術や故障予兆診断技術，医療では遺伝子解析や製薬開発など幅広い分野で活用されている．この潮流から，異分野・異業種のデータ利活用を促進するために，2017 年 5 月に匿名加工情報に関する規則を加えた改正個人情報保護法が施行され，事業者間でデータを流通させる場合は個人を識別できないように加工した匿名加工情報とすることが規定される [5] など，活用のための環境整備が進んでいる．

データを匿名加工する手法としては，同じ属性を持つデータが一定数以上存在するようにデータを変換・加工することにより個人が特定される確率を低減する k 匿名化 [6] [7] [8] [9] 等が知られている．これまで，データを非特定化するためには，データベースから対象となるデータを抽出し，データベース外部の専用ソフトウェアを用いてデータの加工や検証を行なっていた [14] [15]．しかし，非特定化により情報が失われる可能性があり，データの利用者に有用な情報量を確保するためには，データの抽出範囲や加工単位などを細かく調整しながら，データの抽出・加工・検証を繰り返さなければならない．専用ソフトウェアは，データベースから抽出したデータを主記憶に展開し加工や検証を行うため，データの規模や種類が増えると主記憶内では処理できずデータ非特定化に費やす時間が膨大となっていた．

著者らは，これまでデータの分析の高速化に用いられてきた非順序型実行原理 [10] [11] [12] をデータの非特定化処理に適用する方法を提案している [13]．本研究では，電子レセプト情報 [28] を模擬して人工的に生成したデータセットを用いて提案手法の性能を評価し，その性能特性を明らかにする．模擬データによる実験結果から，データの特性や非特定化の要件の差異による性能特性を明らかにでき，データ加工に要する時間を短縮できる見込みを得た．これにより，データを有効活用した施策の立案や新たなサービスの創出が期待できる．

2　対話的データ非特定化のためのデータの加工・検証方法

2.1　データ加工方法

匿名加工情報とは，特定の個人を識別することができないように個人情報を加工し，その個人情報を復元できないようにした情報である [26]．匿名加工情報を生成するためのデータの加工方法には様々あるが，文献 [26] では，表 1 に示すような方法が挙げられている．本論文では，これらのデータの加工を非順序型実行原理を用いて行う方法を提案する．

2.2　データ検証方法

加工されたデータの検証方法には，安全性の検証と有用性の検証がある．安全性の検証とは，加工されたデータが再特定される可能性を検証するものであり，代表的な手法として k 匿名化が知られている [6] [7] [8]．k 匿名化とは，対象となるデータ内に，同じ属性を持つデータが k 件以上存在するようにデータを加工する技術であり，これにより個人が特定される確率を $1/k$ 以下に低減させる．

有用性の検証方法としては，文献 [27] では，情報損失メトリクスとしてエントロピーと欠損という 2 つのメトリクスを用いる方法が紹介されている．本論文では，これらのデータ検証方法を非順序型実行原理を用いて行う方法を提案する．

3　非順序型実行原理のデータ加工への適用

3.1　非順序型実行原理

非順序型実行原理とは，主記憶には収まらない大規模なデータベースから，特定の条件に合致する個別データの選択を高速に実行することを目的としており，近年の多コア・多ストレージデバイスのハードウェアの特徴を生かすべく，並列演算性能と

[1] 正会員　株式会社 日立製作所
norifumi.nishikawa.mn@hitachi.com
[2] 非会員　株式会社 日立製作所
yuuya.isoda.sj@hitachi.com
[3] 正会員　株式会社 日立製作所
kazuhiko.mogi.uv@hitachi.com
[4] 非会員　株式会社 日立製作所
akira.shimizu.wv@hitachi.com
[5] 正会員　東京大学
haya@tkl.iis.u-tokyo.ac.jp
[6] 正会員　東京大学
kgoda@tkl.iis.u-tokyo.ac.jp
[7] 会長　東京大学, 国立情報学研究所
kitsure@tkl.iis.u-tokyo.ac.jp

表 1　データの加工方法の例

加工方法	説明
項目削除/レコード削除/セル削除	加工対象となる個人情報データベース等に含まれる個人情報の記述等を削除する
一般化	加工対象となる情報に含まれる記述を，上位概念に置き換える，数値を四捨五入で丸める
トップ (ボトム) コーディング	加工対象となる個人情報データベース等に含まれる数値に対して，特に大きい又は小さい数値をまとめる
ミクロアグリゲーション	加工対象となる個人情報データベース等を構成する個人情報をグループ化，した後グループの代表的な記述等に置き換える
データ交換 (スワップ)	加工対象となる個人情報データベース等を構成する個人情報相互に含まれる記述等を（確率的に）入れ替える
ノイズ (誤差) 付加	一定の分布に従った乱数的な数値を付加することにより，他の任意の数値へと置き換える
疑似データ生成	人工的な合成データを作成し，これを加工対象となる個人情報データベース等に含ませる

入出力帯域を高効率に活用することを可能とした実行原理である [10] [11] [12].

大規模データベースから特定の条件に合致する個別データの選択や結合処理には通常 B+-Tree などの索引が使用される．従来の DBMS では，問合せ処理において B+-Tree の各ページやリーフページからポイントされるデータページのストレージからの読み取りを少数のタスクで逐次的に実行している．このため多コア・多ストレージデバイスの入出力性能を十分に使いきることが困難であった．

非順序型実行原理では，前述のストレージからの読み取り要求が発生する都度新たなタスクを生成する．各タスクは，既に発行済みの読み取り要求の結果を待たずに次々に読み取り要求をストレージデバイスに発行する．この多重 I/O は，ストレージデバイス内のディスクアレイコントローラや OS（Operating System）内の高度なスケジューリング機構により，論理的な入出力発行順序とは異なる順序で処理される．これにより，特定の条件に合致する個別データの選択や結合処理の飛躍的な高速化を実現する．

3.2　非順序型実行原理を用いたデータの加工

非順序型実行原理は，Hitachi Advanced Data Binder（以下 HADB) において実装されている．HADB は関係データベースシステムであり SQL をサポートしている．このため，データの非特定化に必要なデータの加工や検証を SQL により記述できるようにすることで，非順序型実行原理をデータの加工及び検証に適用する．

3.2.1　データ加工方法の SQL での実現

表 1 に示したデータ加工方法は，セルの値の変形とセルの値の置き換えにより実現できる．セルの値の変形は，SQL で用意されている関数をセルの値に適用することにより行うことができる．3.1 節で述べたように，非順序型実行原理は特定の条件に合致する個別データの選択や結合処理に効果を発揮する．そこで我々は，セルの値の変形と置き換えを B+-Tree を用いた選択および結合演算で実現することにより非順序型実行原理を適用できるようにした．具体的には，セルの値の置き換えのために，データの置き換え前後の値を記述した置換表をデータベース内に定義しするとともに，置換対象のデータの列に B+-Tree 索引を定義することによりこれを実現した．表 2 に SQL によるデータ加工の実現方法の例を示す．

表 2　SQL によるデータ加工方法の実現例

加工方法	SQL による記述
項目削除/レコード削除/セル削除	select 句による選択/where 句条件指定/case 式による削除 (NULL 値へ置換)，または置換表との結合
一般化	置換表との結合
トップ (ボトム) コーディング	置換表との結合，case 式による置換，算術関数
ミクロアグリゲーション	丸め関数，置換表との結合
データ交換 (スワップ)	乱数生成関数と case 式の組合せ
ノイズ (誤差) 付加	乱数生成関数，置換表との結合
疑似データ生成	乱数生成関数，置換表との結合

3.2.2　SQL によるデータ検証の実現

データ検証に用いるメトリクスのは，以下のように算出する．

1. **k 値の計算**
   データ項目の k 値は、項目 i ごとの行数 n_i である．
2. **エントロピーの計算**
   全体行数 N，項目 i ごとの行数 n_i を集計演算により求めた後，以下を計算することにより求める．ここで，$p_i = n_i/N$ である．
   $$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i$$
3. **欠損の計算**
   欠損 d は，加工前の全体行数 N_b と加工後の k 値を満たす全行数 N_a とから，$d = 1 - N_a/N_b$ により求める．

4　性能評価

非順序型実行原理を適用したデータの加工方法の性能特性を明らかにするために，著者らはデータの加工・検証を SQL で記述し，データセット内のデータの分布，データの選択率，及びデータ非特定化の際に用いるパラメタである k 値を変化させ，データ

内閣府の最先端研究開発支援プログラム「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」（中心研究者：喜連川 東大教授／国立情報学研究所所長）の成果を利用.

の加工・検証処理の応答時間を計測した．

4.1 データ加工・検証処理

性能評価にあたり，著者らは対話的なデータの非特定化で必要となる 2 種類の処理を定義した．

1. **抽出データの検証**
   データの利用者が要求したデータが，安全性を満たしているかを検証するために，準識別子の組合せ毎の行数を計算した．同時に，有用性の評価に必要となる行数及びエントロピーを計算した．本処理では，利用者が要求するデータの抽出条件をビュー表として定義し，そのビュー表に対して k 値・行数・エントロピーを計算する SQL を発行しその応答時間を計測した．
2. **データの加工・再検証**
   データの加工に必要な情報を置換表としてデータベース内に定義した．その後，前述のビュー表と置換表を結合することによりデータを加工 (置換) し，その結果に対して k 値・行数・エントロピーを計算する SQL を発行しその応答時間を計測した．

性能評価では，非順序型実行原理を用いない従来方式と，非順序型実行原理を適用した提案方式について，上記処理を直列に実行し，その合計応答時間を計測した．評価では，データの選択率と k 値を変えた 6 種類の加工・検証処理 (それぞれ Q1 から Q6 とする) を用いた．Q1 から Q3 は k 値を 8 に固定し選択率をそれぞれ 0.006%, 0.06%, 0.6% に変化させ，Q5 から Q8 は選択率を 0.06% に固定し k 値をそれぞれ 2, 4, 8 に変化させた．データの選択率は、データが分析に用いられることを想定し数百万件から数億件になるように，k 値は 2 から 10 程度と想定されるためそれに近い範囲で値を選択した．

4.2 データセット

試験データセットとしては電子レセプト情報 [28] を模擬して人工的に生成したデータセット（約 1,000 億レコード）を用いた．

4.2.1 データ生成方法

疑似データを生成するに当たり，以下の分布を考慮した．

- 保険者番号当りの保険者数
- 年齢 (5 歳幅) 毎の保険者数の分布
- 性別・年齢 (5 歳幅) 毎の傷病発生確率．医科 (入院，入院外) 及び DPC 毎を区別
- 性別・年齢区分　・傷病毎の医薬品の発生確率．院内，及び調剤を区別

本評価では，上記の分布が一様分布であるデータを生成した．

4.3 システム構成

性能評価に用いたハードウェアは 44 コア/88 スレッド，512GB メモリのサーバ及び 7 TB のフラッシュドライブを 16 台搭載したストレージであった．サーバとストレージ間は 16Gbps のファイバチャネル 16 本で接続されている．本ハードウェア上に，前述のデータセットを用いてデータベースを作成した．関係データベースシステムとして HADB を用いた．

4.4 性能評価結果

性能評価の結果を図 1 に示す．

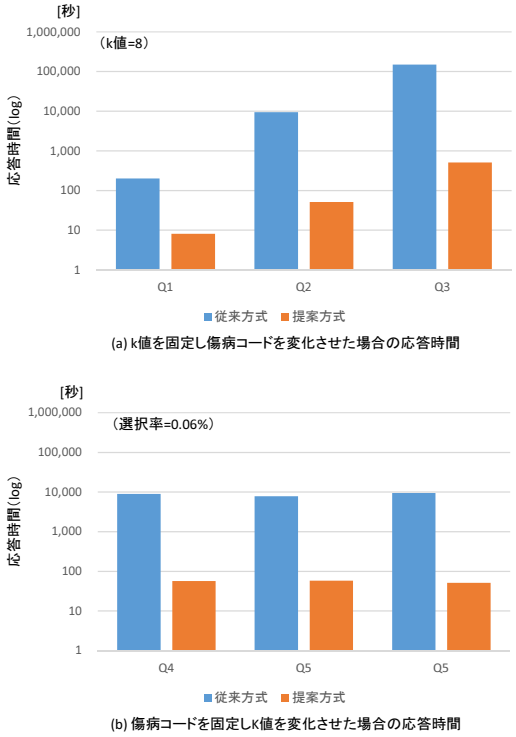

図 1　匿名加工処理の性能評価結果．

図 (a) は Q1 から Q3 についての選択率ごとの応答時間を示している．図から分かるように，どの選択率においても提案方式の方が応答時間が短く，選択率が高くなるにつれ差が大きくなっていることが分かる．また，データの選択率が高くなるに従い，両方式とも応答時間は長くなっている．

図 (b) は，Q4 から Q6 についてのデータ分布毎の応答時間を示している．図 (a) と同様，提案手法の方が応答時間は短い．しかし図 (a) と異なり異なり k 値を変化させた場合は応答時間はそれほど変化していない．

4.5 考察

前節の結果はデータ加工および k 値，エントロピー，欠損率の計算は加工・検証対象レコードの件数の影響を受けるが，k 値の大小の影響はあまり受けないことを示している．k 値の大小の影響が小さい理由は，データの加工や検証は選択された全てのデータを対象としており k 値の大小に関わらず計算量がほぼ一定になるためと考えられる．

また，図 1(a) に示す通り選択率が高くなるにつれて高速化率が向上している．これは、選択率の増加に伴い処理対象レコード数が増加し，入出力帯域を高効率に使用できるようになるためと考えられる．

図 1 に示すように，提案手法は全ての加工・検証処理において応答時間を短縮できており，このことから提案した非順序型実行原理に基づく加工処理手法は，匿名加工処理への適正があると考

4 週未満，1 歳未満，7 歳未満，15 歳未満，15 歳以上の 5 区分とした．

えられる．

5 関連研究

データベース管理システムによる k 匿名化の研究は，データベース管理システムによるデータのプライバシー保護の研究の一機能として研究されている．データベース管理システムによるデータのプライバシー保護の研究としては，プライバシーメタデータをデータベースに保持し，それを用いてデータ収集・クエリ・保持に関する様々な制御を行う DBMS [18] [19]，プライバシーポリシーやセキュリティポリシーを宣言的に与え，それを満たすようクエリを書き換える DBMS [20] [21] などがある．文献 [21] では，一般化階層木を DBMS 内部に保持し，セキュリティポリシーの宣言に基づき一般化階層木を用いたクエリに書き換えを行う．データベース管理システムによる k 匿名化の研究には，以下のものがある [22] [23]．これらは，k 値を満たすようにデータを加工する DBMS オペレータを定義し，DBMS 内部で k 値を満たすよう自動匿名化を実行する．このため，本研究が狙う対話的なデータ加工プロセスに用いること困難と考えられる．

また，関係データベースシステム上で時系列データの k 匿名化を行うためのアルゴリズムの提案 [24]，匿名化を SQL に似た言語で記述するツールの研究 [25] 等も行われている．

6 まとめ

本研究では，大規模データを対象とした対話的なデータの非特定化のための非順序型実行原理を用いたデータの加工方法の性能を模擬データを用いて計測した．実験結果から，データの特性や非特定化の要件の差異による性能特性を明らかにするとともに，提案した非順序型実行原理に基づく加工処理手法が匿名加工処理に有用であることを示した．この結果非順序型実行原理を用いることでデータ加工に要する時間を短縮できる見込みを得た．これにより，データを有効活用した施策の立案や新たなサービスの創出が期待できる．

謝辞

本研究の一部は，総合科学技術・イノベーション会議が主導する 革新的研究開発推進プログラム（ImPACT）の一環として実施したものです．

参考文献

[1] 平成 30 年版情報通信白書, 総務省, 2018.
[2] オープンデータ戦略の推進, 総務省, 2018.
[3] Klaus Schwab, The Fourth Industrial Revolution, World Economic Forum, 2016.
[4] 経済産業省, 新産業構造ビジョン, http://www.meti.go.jp/ committee/sankoushin/shin_sangyoukouzou/pdf/ 017_05_00.pdf, 2017.
[5] e-Gov, 個人情報の保護に関する法律, http://elaws.e-gov.go.jp/search/elawsSearch/elaws_search/lsg0500/detail?lawId=415AC0000000057&openerCode=1, 2018.
[6] Latanya Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No.5, pp.571–588, 2002.
[7] Latanya Sweeney, A Model for Protecting Privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol.10, No.5, pp.557-570, 2002.
[8] 原田邦彦, 佐藤嘉則, 一般化階層木の自動生成と情報エントロピーによる歪度評価を伴う k-匿名化手法, 情報処理学会研究報告, Vol.2010-CSEC-50 No.47, 2010.
[9] 竹之内隆夫, $k-$ 匿名化技術と実用化に向けた取り組み, 情報処理 Vol.54, No.11, pp.1125–1129, 2013.
[10] 喜連川優, 合田和生, アウトオブオーダ型データベースエンジン OoODE の構想と初期実験, 日本データベース学会論文誌，Vol. 8, No. 1，pp.131–136, 2009.
[11] 合田和生, 豊田正史, 喜連川優, アウトオブオーダ型データベースエンジン OoODE の試作実装と小規模実験環境におけるソフトウェア実行挙動の観測，日本データベース学会論文誌, Vol. 12，No. 1, pp.25–30, 2013.
[12] 清水晃, 茂木和彦, 合田和生, 喜連川優, 非順序型実行原理に基づく超高速データベースエンジンの詳細分析処理における性能評価, 日立評論 イノベイティブ R&D レポート, pp.83–89, 2014.
[13] 西川記史, 礒田有哉, 出射英臣, 茂木和彦, 吉野雅之, 清水晃, 早水悠登, 合田和生, 喜連川優, 非順序型データベースエンジンを用いた大規模データの対話的な非特定化手法の検討と初期評価, 電子情報通信学会データ工学研究会第一種研究会・情報処理学会データベースシステム研究会合同研究会, pp.61–64, 2018.
[14] FUJITSU ビジネスアプリケーション NESTGate, http://www.fujitsu.com/jp/solutions/business-technology/intelligent-data-services/bigdata/ba-solutions/nestgate/download/, 2017.
[15] 長谷川 聡, 正木 彰伍, 岡田莉奈, 大規模データを実用的な速度で処理可能な匿名化ライブラリの設計と実装評価, コンピュータセキュリティシンポジウム (CSS), 2017
[16] NTT テクノロス, 匿名加工情報作成ソフトウェア, https://www.ntt-tx.co.jp/products/anontool/, 2017.
[17] NEC データ匿名化ソリューション, https://www.nec-solutioninnovators.co.jp/sl/danony/, 2017.
[18] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, Yirong Xu, Hippocratic Databases, Proceedings of the 28th VLDB Conference, 2002.
[19] Kristen LeFevre, Rakesh Agrawal, Vuk Ercegovac, Raghu Ramakrishnan, Yirong Xu, David DeWitt, Limiting Disclosure in Hippocratic Databases, Proceedings of the 30th VLDB Conference, 2004.
[20] Rakesh Agrawal, Paul Bird, Tyrone Grandison, Jerry Kiernan, Scott Logan, Walid Rjaibi, Extending Relational Database Systems to Automatically Enforce Privacy Policies, 21st International Conference on Data Engineering (ICDE'05), pp.1013–1022, 2005.
[21] Yasin Laura-Silva, Walid G. Aref, Realizing Privacy-Preserving Features in Hippocratic Databases, IEEE 23rd International Conference on Data Engineering Workshop, pp.198–206, 2007.
[22] Jalaja Padma, Yasin N. Silva, Muhammad U. Arshad, Walid G. Aref, Hippocratic PostgreSQL, IEEE 25th International Conference on Data Engineering , pp.1555–1558, 2009.
[23] Mohamed Nassar, Adel Al-Rahal Orabi, Marwan Doha, Bechara AL Bouna, An SQL-like Query Tool for Data Anonymization and Outsourcing, International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), pp.1–3, 2015.
[24] Sergio Mascetti, Claudio Bettini, X. Sean Wang, k-Anonymity in Databases with Timestamped Data, Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06), pp.177–186, 2006
[25] Mohamed Nassar, Adel Al-Rahal Orabi, Marwan Doha, Bechara AL Bouna, An SQL-like Query Tool for Data Anonymization and Outsourcing, International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), pp.1–3, 2015.
[26] 個人情報の保護に関する法律についてのガイドライン（匿名加工情報編），個人情報保護委員会, http://www.ppc.go.jp/files/pdf/guidelines04.pdf, 2016.
[27] Khaled El Eman, Luk Arbucke 著, 木村映善, 魔狸 監訳, 笹井崇司訳, データ匿名化手法 ヘルスデータ事例に学ぶ個人情報保護, オライリー・ジャパン, 2015.
[28] 合田和生, 山田浩之, 喜連川優, 満武巨裕, 大規模電子レセプト情報の解析のためのデータベース基盤の性能ベンチマークの検討, 電子情報通信学会第 10 回データ工学と情報マネジメントに関するフォーラム／第 16 回日本データベース学会年次大会 (DEIM2018)，C6-1, 2018.