Vol.2010-CSEC-50 No.48
2010/7/2

情報大航海プロジェクトにおける個人情報匿名化基盤の構築と検証

廣田啓一[†] 保木野昌稔[††] 藤木由里[††]
松崎和賢[†††] 吉田圭吾[†††] 赤井健一郎[†††]
高橋克巳[†] 白井康之[††††] 山口利恵[†††††]

GIS や電子マネーなどの技術の普及にともない，事業者の手元にはユーザの属性情報や履歴など様々な情報が蓄積されるようになった．こうしたパーソナル情報を有効かつ安全に利活用するためには，個人が識別されないよう匿名化するとともに，匿名化されたデータの安全性を判断する基準となる評価指標を付与できることが望ましい．情報大航海プロジェクトでは，k-匿名性の考え方の下に，大規模データに対して実用的な処理時間で匿名化を施す，個人情報匿名化基盤を構築した．本稿では，汎用的なプラットフォームである個人情報匿名化基盤の設計思想と実現機能，および実規模データを用いて評価した結果について報告する．

Development and Evaluation of Personal Data Anonymization Platform in Information Grand Voyage Project

Keiichi Hirota[†] , Masatoshi Hokino[††] , Yuri Fujiki[††] ,
Kazutaka Matsuzaki[†††] , Keigo Yoshida[†††] ,
Kenichirou Akai[†††] , Katsumi Takahashi[†] ,
Yasuyuki Shirai[††††] and Rie Yamaguchi[†††††]

With the popularization of the useful technologies such as GIS, digital cash and so on, various types of information including user's attributes and behavior histories can be gathered and accumulated at servicer's hand as personal data. For the purpose of valid and safe utilization of such data, it is recommended to anonymize personal data avoiding the risk of individual identification, and estimate the safety of anonymized data. In Information Grand Voyage Project, we have developed the personal data anonymization platform which anonymize large-scale database within practical processing time, regarding the concept of k-anonytmity. In this paper, we describe the architecture and functions of the versatile anonymization platform and report the evaluation result of its performance using real-scale personal data.

1. はじめに

近年 GIS や電子マネーなどの技術の普及にともなって，サービス提供事業者の手元には，ユーザの氏名や性別，年齢や住所といった属性情報だけでなく，移動や購買，閲覧といった行動の履歴など様々な情報が取得・蓄積されるようになった．こうしたパーソナル情報を効果的に利用することで，ユーザにとって便利で有用なサービスが提供できることが期待され，積極的な利活用によるイノベーションを起こし，新たなビジネスやサービスを産み出そうとする気運が高まっている．

他方，こうしたパーソナル情報は，個人の日常生活に関する情報を多く含んでいるため，利活用にあたってはセキュリティを保護するだけでなく，プライバシーや個人情報の保護に十分な配慮をする必要がある．配慮を欠いた利活用は社会的非難を浴びるだけでなく，経営的にも深刻な影響を与える恐れがあり，こうした事業者の懸念が新たなサービス実現の足かせになっている現状がある．

経済産業省「情報大航海プロジェクト」[1]では，情報爆発の時代ともいえる現代において，十分なプライバシーや個人情報の保護を確保しつつ，行動履歴・閲覧履歴など個人に関する情報を有効に利活用し，パーソナライズドサービスを代表とするイノベーションを実現するための方策について検討してきた．パーソナル情報の利活用において，プライバシー・個人情報保護とサービスの利便性のバランスをとるための方法の一つが，個人情報の匿名化である．匿名化とは，収集・蓄積したパーソナル情報に対して，容易に個人の識別ができないように加工する処理のことをいう．パーソナライズドサービスにおいては,個人識別性のない形でもサービスに利用しうる情報(位置情報，視聴履歴など）があり，たとえ，情報を取得した事業者が事前に利用を想定していなかったような場合でも，個人識別性のない形で利用するのであれば，プライバシーに配慮しつつ，情報利活用を促進しうる可能性が生じてくる．

情報大航海プロジェクトでは，こうしたパーソナル情報の利活用に対するニーズを背景に，法制度や事業との連携を図りながら，安全といえる匿名化の要件を整理し，大規模データに対して一定の安全性の保証の下に個人情報を匿名化する，個人情報匿名化基盤の構築を行った．以下では，汎用的なプラットフォームである個人情報匿名

[†] NTT 情報流通プラットフォーム研究所
NTT Information Sharing Platform Laboratories
[††] （財）日本情報処理開発協会
Japan Information Processing Development Corporation
[†††]（株）三菱総合研究所
Mitsubishi Research Institute, Inc.
[††††]（株）三菱総合研究所（現在（独）科学技術振興機構）
Mitsubishi Research Institute, Inc. (Recently, Japan Science and Technology Agency)
[†††††]（独）産業技術総合研究所
Advanced Industrial Science and Technology

化基盤の設計思想と実現機能について述べるとともに，実規模データを想定した同基盤の検証結果について報告する．

2. パーソナル情報の利活用と匿名化

ユーザ個別のニーズを満たすサービスやマーケティングを可能にするためには，ユーザ自身のニーズを的確に捉える必要があることから，個人情報も含めた幅広い情報の有効利用が不可欠である．近年では技術の進展にともなって，取り扱うことができる情報の種類や量が増大し，一人一人の個性に応じたサービスを提供するパーソナライズドサービスが隆盛している．例えば，以下のような事例がある．

- 複数の方法により入手した消費者・利用者の情報（売買・検索の履歴等）を一体化・分析した新たなサービスの提供
- ポイント・電子マネーの企業間連携に伴う個人の購買情報等の交換・流通
- 血圧・脈拍などのデータを利用し，健康サービスを行う際の事業者による個人情報の一括管理

こうしたパーソナル情報の利活用は，従来あらかじめ定めた目的の範囲内で，ユーザの同意の下に行われてきた．しかしながら，こうした一次利用の枠組みは利用目的の制限や同意取得の方法など様々な制限があるため，より有効にパーソナル情報を利活用し，新たなビジネスやサービスを実現するために，パーソナル情報の収集・蓄積を行う一次事業者と，収集・蓄積されたパーソナル情報の利活用を考える二次事業者との間でパーソナル情報を流通させ，二次利用を可能とすることが望まれている．

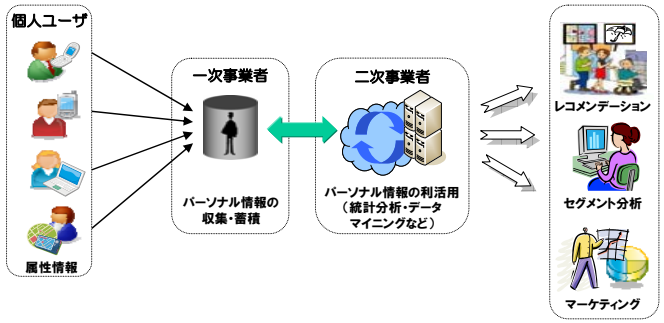

図 1　パーソナル情報の収集・蓄積と利活用

しかしながら，個人に関連付けて収集した各種の属性情報や，サービスの利用にともなって蓄積される個人に関する履歴情報は，個人の日常生活や行動，生活習慣に関する情報が多く含まれているため，個人を特定可能な状態のままで二次利用（たとえば，第三者への提供や一般公開など）することは，個人情報保護の観点から問題があるだけでなく,ユーザのプライバシーの侵害となる様々なリスクが想定される.特に，このような情報を事業やサービスで利用する場合には，情報から直接的あるいは間接的に個人が特定され，個人に関する様々な属性情報や履歴情報が第三者に知られるリスクや，特定の個人に関するプライベートな，あるいはセンシティブな情報が推定され，個人のプライバシーが侵害されるリスクについて注意する必要がある．パーソナル情報の利活用における想定リスクを表 1 に示す.

表 1　パーソナル情報の利活用における想定リスク

想定リスク	リスクの定義
直接個人識別	レコードに氏名・住所などの直接個人を識別可能な属性が含まれている場合，その個人（の氏名・住所）を知る第三者が，その個人に関する他の属性を知ることができる．
間接個人識別	ある属性の組み合わせに対応するレコードが唯一に絞り込める場合，それが誰であるかの情報を持つ第三者は，個人の識別ができ，その個人に関する他の属性を知ることができる．
属性推定	ある属性の組み合わせに対応する複数のレコードに対して，必ず共通の値を持つ他の属性情報がある場合，属性の組み合わせに該当する個人を知る第三者は，その個人に関する新たな属性情報を推定することができる．

こうしたパーソナル情報の利活用におけるリスクを低減し，安全な二次利用を可能にする手段の一つが，匿名化である．

匿名化とは「個人が識別されることのないように，パーソナル情報を加工すること」をいい，代表的な手段として，個人単位のレコードから，直接個人を識別できる氏名や ID などの情報を切り落としや仮名化，あいまい化などによって取り除く「単純匿名化」がある．しかし，氏名など個人を直接的に識別できる情報（以下，識別情報と呼ぶ）を取り除いたとしても，住所や生年月日などの情報から間接的に個人を識別できる可能性が残る場合があることが知られている[2].

そのため，安全にパーソナル情報を利活用するためには，住所や性別，年齢など，組み合わせによって個人識別につながる可能性のある情報（以下，準識別情報と呼ぶ）を加工し，どう組み合わせても個人を特定できないよう，すなわち一定数未満の人数

に対象を絞り込めないように匿名化する「集合匿名化」が望ましいと考えられる．また，匿名化したパーソナル情報の円滑な利活用を促進するためには，匿名化の安全性の基準を定量的に示せることが望ましい．

情報大航海プロジェクトでは，匿名化したパーソナル情報の安全性を保証する指標として，代表的な「k-匿名性」[2][3]を採用した．k-匿名性は，匿名化したパーソナル情報が個人識別リスクをどの程度回避できるかを示す評価指標で，同じ準識別情報の組み合わせを持つレコードがパーソナル情報中に少なくとも k 個ある時，そのパーソナル情報は「k-匿名性を満たす」といい，個人識別リスクが少なくとも k 分の 1 となる．匿名化が必要なパーソナル情報に対し，準識別情報を一般化し，同じ準識別情報の値の組み合わせを持つレコードが最低でも k 個以上存在するよう「k-匿名化」[4]することで，パーソナル情報を安全に利活用できるようになると考えられる．

3. 個人情報匿名化基盤の設計と構築

3.1 基盤設計の考え方

こうした検討を背景に，パーソナル情報の安全な利活用を促進するための基盤技術として，k-匿名性に基づく集合匿名化機能を提供する個人情報匿名化基盤（以下，匿名化基盤と記す）の構築を行った．

匿名化基盤は，収集したパーソナル情報を保有する事業者が，個人識別性のないようにパーソナル情報を加工し，安全に利活用できる匿名化データを得るための機能群を提供するプラットフォームである．そのため，事業者自身の手で実事業・サービスに即した匿名化処理を主体的に実施し，匿名化データの利活用を行えるように，単に匿名化処理を施す機能を提供するだけでなく，匿名化後のデータの安全性や有用性を評価する指標を算出する機能を実現するものとした．

また，匿名化基盤では，実事業への適用を考慮し，汎用的，かつ実用レベルの処理機能を提供することを目的とした．そのため，機能の実現にあたっては，実用レベルを踏まえた実装方式を前提に実規模データへの対応を優先し，サービス共通利用を考慮した特徴の異なる 2 つの代表的なアルゴリズムを採用した．

また，事業者によって異なる多様なサービスや利用形態に適応，共通利用できるように，様々な匿名化の手法や評価尺度，基盤機能などを追加・変更可能な柔軟な基盤アーキテクチャとする．そのため，各機能はライブラリとして提供し，新たなアルゴリズムや評価機能の追加を容易に可能とする設計とした．

3.2 実現機能の概要

個人情報匿名化基盤は，多様な利用形態に柔軟に適用できる機能構成とするため，共通実行基盤機能とプラグ＆プレイ機能群から構成するアーキテクチャとした．このアーキテクチャを採用することにより，共通実行基盤機能上に利用形態に応じて必要な機能をプラグ＆プレイで選択可能とした．

匿名化基盤の機能構成を図 2 に，各実現機能の概要を表 2 に，それぞれ示す．

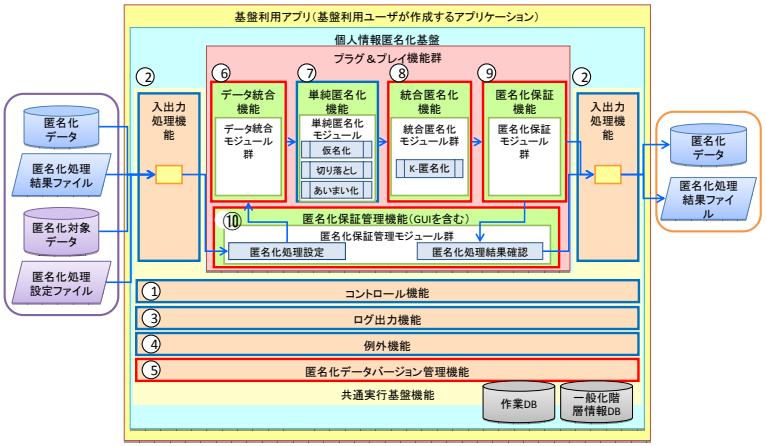

図 2　個人情報匿名化基盤の機能構成

表 2　個人情報匿名化基盤における実現機能と概要

No.	機能名称	機能内容
①	コントロール機能	・匿名化処理設定ファイルの入力，匿名化処理結果ファイルの出力を行う． ・匿名化処理設定ファイルに記述された内容を解析し，適用する匿名化処理を呼び出す．
②	入出力処理機能	・匿名化対象データの入力元指定や匿名化データの出力先指定を行う．
③	ログ出力機能	・匿名化処理機能の実行結果を監査ログとして出力する．また，各処理の実行内容及び例外内容をシステムログとして出力する．
④	例外機能	・各処理の実行時に発生した例外を処理する．
⑤	匿名化データバージョン管理機能	・匿名化処理を実行した際に使用した匿名化処理設定ファイル，匿名化処理結果ファイルを，実行単位で履歴を管理する．

⑥	データ統合機能	・複数の匿名化対象データまたは匿名化データを結合する.
⑦	単純匿名化機能	・匿名化対象データに仮名化，切り落とし，あいまい化を用いて匿名化を施す.
⑧	統合匿名化機能	・特定の匿名化アルゴリズムを用いて匿名化を施す
⑨	匿名化保証機能	・匿名性の評価や有用性の保証のための指標を算出する.
⑩	匿名化保証管理機能	・匿名化処理設定ファイルの策定を補助するための GUI と，匿名化処理結果を確認するための GUI を提供する.

匿名化基盤では，直接的な個人識別につながる識別情報を加工する単純匿名化機能に加えて，準識別情報を加工して k-匿名性を実現するための統合匿名化機能として，「階層探索型」「クラスタリング型」「対話型」の 3 通りの k-匿名化機能を実装した.

(1) 階層探索型 k-匿名化機能

階層探索型の k-匿名化は，「指定された準識別情報に対し，一般化階層情報にしたがって，k-匿名性を満たすことができる，可能な一般化の程度の組み合わせを探索して提示する」機能である.

一般化階層情報とは，属性の値をより上位の一般的な（情報量が少ない）値に置換する一般化（Generalization）のための，属性の値の階層関係や包含関係をあらかじめ定義した情報である．たとえば，14 歳，25 歳といった年齢に対し 10 代，20 代といった年代に置換する，新宿区や渋谷区といった地名に対し東京都に置換する，などの関係を与えておくことで，対象データに対し，一般化によって複数のレコードが同じ値を持つように加工し，k-匿名性を満たすように匿名化することができる．こうした加工処理は対象データ全体に対して行なわれるため，大域的再符号化（Global Recoding）と呼ばれる.

複数の属性情報を対象に匿名化する場合，k-匿名性を満たす一般化のパターン（以下，匿名化プランと記す）が複数考えられ，匿名化データの利用目的や求める有用性によってどの属性をどの程度一般化するかが異なってくる．そのため，汎用性を目的とした匿名化基盤では，k-匿名性を満たす可能な匿名化プランを網羅的に抽出するものとし，効率よく匿名化プランを列挙するアルゴリズムを文献[5]を参考に実装した．実装アルゴリズムは，与えられた一般化階層情報に基づくラティス構造を構築して，k-匿名性を満たす匿名化プランを探索するため，階層探索型と呼ばれる．また，実規模データに対応するために探索処理を並列的に行うスレッド化や最終的に出力する匿名化プラン数を制限する匿名化プラン抑制などの機能追加を行なった.

階層探索型の場合，指定した k の値を満たす匿名化プランを列挙するため，実現される k-匿名性や，有用性や安全性に関するその他の指標を参考にしながら，基盤利用者にとって望ましい匿名化データを選択することができる.

階層探索型の k-匿名化では，頻度の少ない準識別情報の組み合わせに合わせて全体を一般化するため，全体として情報の粒度が粗い匿名化データとなることが知られている．そのため，指定された数以下のレコードを削除すれば k-匿名性を満たすような匿名化プランを抽出する，レコード削除（Suppression）の機能を設けた.

(2) クラスタリング型 k-匿名化機能

クラスタリング型の k-匿名化は，「指定された準識別情報を参照して，同じ値あるいは近い値の組み合わせを持つ k 個以上のレコード同士を 1 つのグループにまとめ，共通の値を持つように一般化することで，k-匿名性を満たす匿名化データを作る」機能である．階層探索型とは違い，準識別情報の値が近い k 個以上のレコードをグループ化し，小さく分割されたグループの中で共通の値を持つように加工するため，より粒度の細かい，有用性の高い匿名化データができる．ただし，元は同じ値でもグループが異なれば異なる加工がされるため，同じ準識別情報の中でも粒度の異なる値が存在する匿名化データとなる．これを局所再符号化（Local Recoding）という．また，距離の計算によって近いレコード同士をグループ化するため，処理時間がかかり，大規模なデータの匿名化には向かないと考えられる.

クラスタリング型の k-匿名化アルゴリズムには大きく分けて分枝型と凝集型の 2 方式があるが，より小さなグループを形成し，粒度の細かい，有用性の高い匿名化データを得ることを目的として，凝集型のアルゴリズムを採用し，文献[6]を参考として実装を行なった．また，グループ化するレコード間の近さに閾値を設け，他のどのレコードに対しても閾値より遠い，孤立したレコードはグループ化せずに削除する，レコード削除の機能を設けた．クラスタリング型の k-匿名化は，階層探索型と違い，1 回の処理でただ 1 つの匿名化データを作成する.

(3) 対話型 k-匿名化

上記の 2 機能と違い，対話型の k-匿名化は，「目標とする k-匿名性と処理の対象とする属性情報を設定し，評価指標を参照しながら，各属性情報に対する匿名化処理を選択して実行する」ための機能で，匿名化基盤が提供する GUI に従って，利用者自身が対話的に k-匿名性を満たす匿名化データを作ることができる.

利用者に匿名化に関する十分な知識とノウハウがあり，あらかじめ用いる匿名化手段やその程度が分かっている場合には，対話型 k-匿名化を使って，利活用の目的や用途に合った匿名化データを効率よく作ることができると考えられる.

さらに，匿名化基盤では，匿名化データの安全性や有用性を評価するための指標として，「k-匿名性」「情報損失」「母集団一意性」の 3 つの評価指標を求める匿名化保証機能を実装した．

k-匿名性は，前述したように，匿名化したデータの中に，同じ属性情報の組み合わせを持つレコードが少なくとも k 個以上存在することを評価する，個人識別リスクの回避を示す安全性の指標である．匿名化基盤では，匿名化した後のデータ中に同じ準識別情報の値の組み合わせを持つレコードが何件あるかを求め，目標値として設定した k の値に対し，実際の処理により達成された最小の k の値を出力する．

情報損失は，匿名化前のレコードに対して匿名化後のレコードがどのくらい情報を失っているかを評価する，匿名化データの有用性を示す指標である[7]．情報損失の算出は，各レコードの匿名化前後の値の変化を差として求め，全レコードの差を平均して求める．したがって，情報損失の値が小さければ，匿名化データは元のパーソナル情報に近く，その有用性も高いと考えることができる．

レコードが持つ属性情報の値は，数値属性や文字列属性，一般化階層情報の定義されたカテゴリ属性など様々であるため，一様に損失の度合いを量ることは難しい．そのため，値の種別毎に異なる計算方法を用いるものとした[8]．数値属性の場合は，匿名化前後の数値の差分を求め，属性全体の最大値と最小値の差分である値域で割った値とした．文字列属性の場合は，匿名化前後の文字列で前方一致した文字数を匿名化前の文字数で割った値を 1 から減算した値とした．カテゴリ属性の場合は，匿名化前後の値の一般化階層上での階層の差を求め，一般化階層情報の高さで割った値とした．

母集団一意性は，個々のレコードに対し，母集団の中からどの程度一意に個人が特定できるかを，母集団の大きさの推定値に基づいて算出する指標で，母集団の中での「個人の識別のされにくさ」を表している．母集団一意性の算出には，有名な統計的開示制御ソフトである μ-ARGUS で用いられている Individual Risk の計算式を用いた[9][10]．母集団一意性の値は，母集団全体に対する匿名化データの割合を示す抽出率と，匿名化データの中で同じ準識別情報の組み合わせを持つレコードの頻度によって決まり，それぞれのレコードから個人が特定できる確率を表す値となる．

3.3 匿名化処理の流れ

個人情報匿名化基盤を使った，匿名化処理の流れを簡単に説明する．匿名化基盤における匿名化処理の流れを図 3 に示す．

基盤利用者は，まず入力とする匿名化対象データの構成，対象データに施す匿名化処理，求める評価指標を決め，匿名化処理設定ファイルに記述する．

匿名化処理設定ファイルは，匿名化データ情報と匿名化処理定義からなり，匿名化データ情報には，処理の対象とする匿名化対象データの情報項目と，各情報項目に対する識別情報，準識別情報，その他の情報の区分を指定する．識別情報は主に単純匿名化による加工が必要な情報項目，準識別情報は k-匿名化の対象とする情報項目，その他の項目は加工の対象としない情報項目を表す．

匿名化処理定義には，匿名化対象データに対しどのような処理を適用するかを定義する．まず，単純匿名化機能による処理として，識別情報，準識別情報に対して適用する手法とそのパラメータを指定する．また，統合匿名化機能による処理として，階層探索型，クラスタリング型，対話型を選択し，各 k-匿名化機能に応じて必要なパラメータや満たすべき k の値を設定する．また，匿名化保証機能において算出する，k-匿名性，情報損失，母集団一意性などの評価指標を指定することができる．

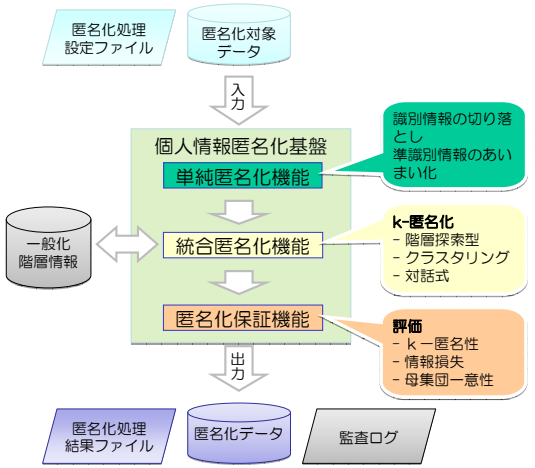

図 3　個人情報匿名化基盤における匿名化処理の流れ

匿名化基盤は，匿名化処理設定ファイルの内容に基づいて，匿名化対象データに対し，単純匿名化機能による識別情報・準識別情報の加工，統合匿名化機能による k-匿名化，匿名化保証機能による匿名化指標の算出の順に処理を実行し，一連の処理によって匿名化処理が施された匿名化データ，匿名化処理の結果を記した匿名化処理結果ファイル，および匿名化処理のログにあたる監査ログを出力する．

基盤利用者は匿名化データおよび匿名化処理結果ファイルを参照し，匿名化されたデータの内容や k-匿名性，情報損失などの評価指標を参考にして，匿名化データの安全性や有用性を判断し，必要に応じて異なるパラメータで異なる匿名化処理を実行し，利用の目的にあった匿名化データを得ることができる．

4. 匿名化基盤の検証

4.1 検証の目的とデータセット

構築した個人情報匿名化基盤の機能ならびに実用性を検証することを目的に，実用規模を想定したデータでの性能評価を行ない，階層探索型 k-匿名化およびクラスタリング型 k-匿名化のそれぞれの実装機能についての性能要件の検証を行なった．

階層探索型の k-匿名化について，その機能性を検証するとともに，レコード数や準識別情報の数，k の値といった，性能要件に影響を与えるパラメータ毎の特性を検証した．また，スレッド化や匿名化プラン抑制といった，実用性の向上に向けた追加機能の効果についても検証した．

同様に，クラスタリング型の k-匿名化についても，レコード数や準識別情報の数，k の値といった，性能要件に影響を与えるパラメータ毎の特性を検証した．

また，階層探索型とクラスタリング型の簡単な比較検証を実施して性能を比較し，生成した匿名化データの情報損失，母集団一意性を評価して，機能による評価値の違いについて考察した．

検証には，実規模データの匿名化を前提に，公開されているデータセットなどを用いた．検証で用いたデータセットを表 2 に示す．Adult Data Set および US Census Data は，機械学習に関する公開データリポジトリ[11]から取得した，個人の性別や年齢，職業，国籍といった属性を含む，それぞれ 15 属性 45,222 件，68 属性 2,458,285 件のデータセットである．Person Trip Data は，東京都市圏交通計画協議会のパーソントリップ調査データを東京大学空間情報科学研究センターにおいて加工した，15 分毎の位置座標の記録で 1 日分の人の移動を表した 722,000 件のデータセットである[12].

表 2　検証で使用したデータセット

データセット名	レコード数	利用属性	概要
Adult Data Set	45,222 件	年齢，職業，最終学歴，夫婦，職種，関係性（家族構成），人種，性別，本国	米国勢調査局の 1994 年のデータベースから抽出した，収入と各種の属性との関係を表した，14 属性からなる機械学習用データ
US Census Data	2,458,285 件	年齢，先代の民族，先々代の民族，出勤時間，産業，職業，出身地	1990 年の米国センサス公開ミクロデータ(PUMS)から 1 %抽出した 68 属性からなるデータ
Person Trip Data	722,000 件	位置（緯度経度），性別，年齢，住所，職業，移動の目的，交通手段	アンケート調査に基づいて，人の 1 日分の移動履歴を 15 分単位の位置座標で記録したデータ
擬似 Adult Data Set	任意	Adult Data Set に準拠	各属性の値を一様ランダム分布に従って決定した擬似データ

これらのデータは，一般的な統計調査やアンケート調査に基づくものであり，属性の値の分布に偏りがあることが分かっている．そのため，k-匿名性の満たしやすさによる処理速度のバイアスを避けるために，Adult Data Set の構成に準拠した 9 種類の属性の取り得る値を一様ランダム分布に従って決定した擬似データ（擬似 Adult Data Set）を生成し，属性の値の分布が一様であるようなデータに対しても各機能が有効な処理性能を有することを検証した．

性能評価にあたっては，Xeon X5160(3GHz) × 2CPU(4 コア)，16GB，RedhatEnterpriseLinuxAS4.5(x64), PostgreSQL 8.4.1 と Xeon X5550(2.66GHz) × 2CPU(8 コア)，12GB，CentOS4.7(x64)，PostgreSQL 8.4.1 の 2 種類の検証環境を用いた．前者を検証環境 1，後者を検証環境 2 とし，特に記述のない限り前者を実験環境とする．

4.2 階層探索型 k-匿名化の機能検証

(1) アルゴリズム実装の妥当性検証

実装の参考とした文献[5]の筆者により公開されているサンプルプログラム（以下，Incognito サンプルと記す）と，階層探索型の k-匿名化機能を比較し，実装の妥当性を検証した. Incognito サンプルを匿名化基盤と同じ PostgreSQL を用いた環境で動作するよう修正し，Adult Data Set 9 属性 45,222 件を対象に，k の値を 2 として処理時間の比較を行なった．なお，条件を合わせるために，階層探索型 k-匿名化の設定を，スレッド数 1，匿名化プランの抑制なしとした．

Incognito サンプルの処理時間は 13 分 24 秒で，264 個の匿名化プランを出力した．一方，階層探索型は 19 分 34 秒で，73 個の匿名化プランを出力し，匿名化データを 1 つ生成した．

出力された匿名化プランを確認したところ，階層探索型による匿名化プランは Incognito サンプルが出力した匿名化プランに全て含まれ，機能として妥当なことが確認できた．また，階層探索型の処理時間は匿名化データの生成処理も含むため，その差分を考慮すると Incognito サンプルと同等の処理性能を有していると考えられる．

(2) 実規模データにおける動作検証

実規模データに対する総合的な処理性能について，妥当な範囲内の処理時間で終了し，実利用に耐えうる性能を有していることを検証した．想定する実データの規模としては，某電子マネーの利用状況として月間利用件数が数千万件というデータが公表されており，一日当たり利用件数は数十万から数百万件と考えられる．本検証では，実データ規模として十分な 7 属性 2,458,285 件の US Census Data を用いた．

k 値を 2 に指定して，匿名化プラン抑制なし，スレッド数 1 と，匿名化プラン抑制数 1，スレッド数 4 での 2 通りの条件での匿名化を実行した．結果を図 4 に示す．

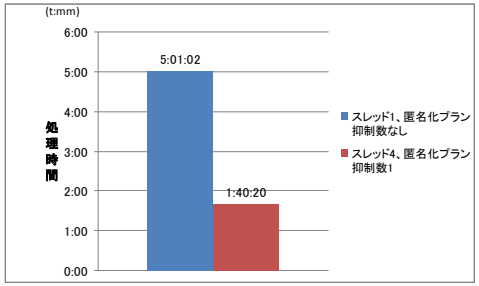

図 4　準識別情報の数と処理時間（PTD）

従来アルゴリズムを想定したスレッド数 1，匿名化プラン抑制なしでの処理では，処理時間に約 5 時間を要した．処理時間はレコード数に比例するので，1,000 万件規模のデータを処理した場合，約 1 日かかるとの想定になる．これに対し，スレッド数 4，匿名化プラン抑制数 1 に設定することで，処理時間は 1 時間 40 分に短縮された．1,000 万件規模のデータ処理が 6 時間程度で完了するとの結果となり，階層探索型 k-匿名化がバッチ処理として十分な処理性能を有することが検証できた．

4.3　階層探索型 k-匿名化の性能要件

階層探索型 k-匿名化の処理時間に影響を与えると考えられる，レコード数，許容匿名度指定値（k の値），準識別情報の数といった要因について処理時間の測定を行い，階層探索型 k-匿名化の基本的な性能要件について検証した．

検証データとしては 9 属性 722,000 件の Person Trip Data と擬似 Adult Data Set を用い，属性の値に偏りのあるデータと，一様ランダム分布に従って決定した，値が平均的に分散したデータの双方で同様の傾向を示すことを検証する．

(1)　レコード数と処理時間

まず，処理対象のデータのレコード数と階層探索型の処理性能の関係を検証した．

PTD から抽出した 100,000～700,000 件のレコードと，擬似 ADS として生成した 50,000～1,000,000 件のレコードに対し，匿名化処理を行った．いずれも k の値は 2 とし，スレッド数 4，匿名化プランの抑制なしとした．結果を図 5，図 6 に示す．

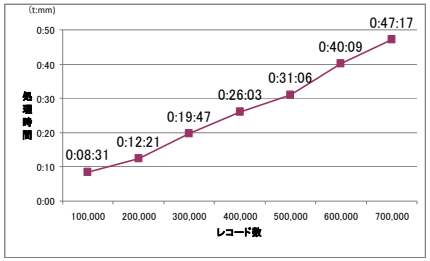

図 5　レコード数と処理時間（PTD）

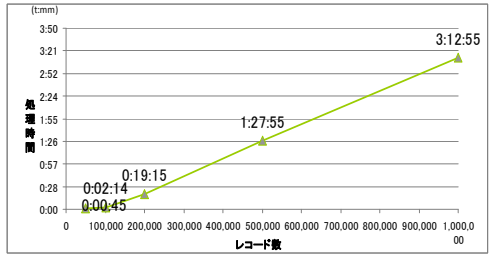

図 6　レコード数と処理時間（擬似 ADS）

結果から，階層探索型 k-匿名化の処理時間は，レコード数の増加にほぼ比例して増加することが明らかとなった．また，属性の値の分布は比例傾向に影響がないものと考えられる．ただし，データに偏りがある PTD での検証結果では 50 万件で約 30 分弱であったのに対し，擬似 ADS では 50 万件に約 88 分を要した．これは，データに偏りがある PTD では探索過程で匿名化プランの枝狩りが発生し，処理時間が減少するが，一様分布データでは枝狩りが発生しづらく処理時間がかかったものと考えられる．

(2)　k 値と処理時間

許容匿名度指定値（k 値）が，処理性能に与える影響を検証した．

レコード数 722,000 件の PTD に対し，k の値を 2，5，10，15，50，100，1,000 と変えて匿名化を行った．また，擬似 ADS についてはレコード数 50,000 件，100,000 件，200,000 件の 3 セットを用意し，k の値を 10，20，50，100 と変えて匿名化を行った．いずれもスレッド数 4，匿名化プランの抑制なしとした．結果を図 7，図 8 に示す．

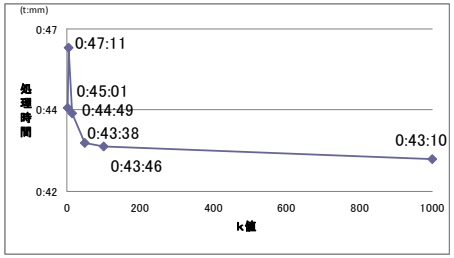

図 7　k の値と処理時間（PTD）

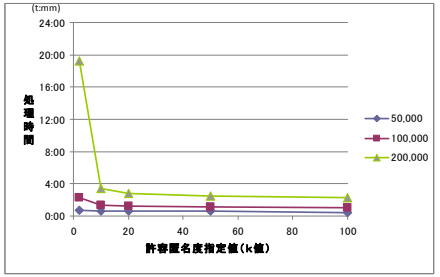

図 8　k の値と処理時間（擬似 ADS）

基本的には k の値を大きくするほど，処理が早くなる傾向が見られた．これは，匿

名化プラン探索の途中におけるラティス構造の探索過程で，k-匿名性を満たさない匿名化プラン候補を切り捨てる枝狩りが多く発生し，ラティスの探索空間が小さくなることで結果として処理時間が短くなったためと考えられる．

なお，図 7 において k=5 の処理時間が k=2 よりも突出しているのは，探索の過程で構築されるラティス構造の枝狩りの状況によって，頻度表を作成する回数が増えたのが原因と考えられる．これは，Person Trip Data は，準識別情報の取る値に偏りがあるデータであるためと考えられる．

(3) 準識別情報数と処理時間

複数の準識別情報（対象項目）の組み合わせでの探索処理時間の測定を行った．PTD 722,000 件を対象として，準識別情報の指定を 1 つずつ増やしていった場合の処理時間を測定した．結果を図 9 に示す．

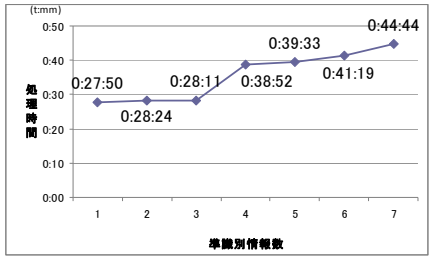

図 9　準識別情報の数と処理時間（PTD）

準識別情報数の増加にともなって，漸増的に処理時間が増加した．なお，3 属性と 4 属性で処理時間に差があるが，4 属性以降には padd（住所）が含まれている．padd は，階層の幅が広い一般化階層情報に従って k-匿名化される．階層の幅が広く，頻度表が大きくなったことから padd の匿名化プランの探索に時間がかかる．このため，3 属性と 4 属性には処理時間の差が生まれたと考えられる．

本検証結果から，準識別情報数が増えると頻度表の構築数が大幅に増えるため，匿名化プラン探索の処理時間は増加する．

(4) k 値と情報損失

一般に，安全性と有用性はトレードオフの関係にあると考えられていることから，階層探索型 k-匿名化によって生成される匿名化データの，k の値が情報損失に与える影響を評価した．PTD を対象に，k の値を 2，5，10，15，50，100，1,000 と変えて匿名化を行い，得られた匿名化データの情報損失を求めた．また，匿名化プランにおける各属性の一般化の程度（遷移した階層の数）を求めた．結果を図 10，図 11 に示す．

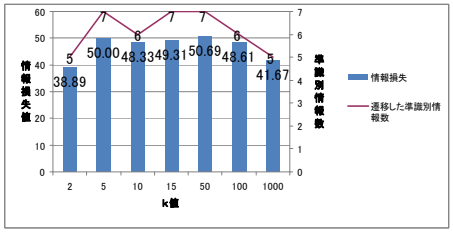

図 10　k の値と情報損失（PTD）

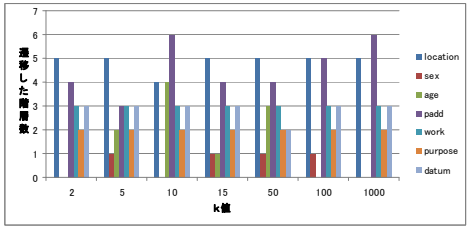

図 11　k の値と匿名化プラン（PTD）

図 10 から，指定した k の値と情報損失の値に直接の関係性は見られなかった．この結果は情報損失の算出方法に依存し，カテゴリ型の属性情報の一般化階層の定義によるものと考えられる．実装した算出方法では，階層数が 2 である属性情報が 1 階層一般化された時の情報損失値は 100%と算出されるが，階層数が 4 である属性情報が 1 階層一般化された時の情報損失値は 25%と算出される．このため，適用された匿名化プランによって情報損失の値が異なる結果となる．図 11 に示した結果から，k の値を 100 とした時と k の値を 1000 とした時の匿名化プランを比較すると，前者では属性 sex が 1 階層一般化されたのに対し，後者では属性 sex の一般化が起きていない．また，前者は 7 属性中 6 属性で一般化が起きているが，後者では 5 属性の一般化で k-匿名性が満たせている．これにより，後者の情報損失が低い結果になったと考えられる．

階層探索型においては，情報損失の値は匿名化プランの違いによる各属性の一般化の程度によって決まり，一概にトレードオフの関係にあるとは言えない結果となった．

4.4 機能追加の効果検証

階層探索型の実装に際し，実用規模のデータ量に対して処理を行うために導入したマルチスレッド対応や匿名化プラン抑制が有効に機能し，大量データに対する処理時間を短縮できることを検証した．

(1) マルチスレッド対応の効果

マルチスレッド化が，処理性能に与える効果を検証した．Person Trip Data を対象に k の値を 2 とし，スレッド数を 1～4 として匿名化処理を行った．処理時間における匿名化プラン抽出と一般化の処理時間の測定結果を図 12 に，匿名化プラン抽出における探索処理の処理時間の内訳を図 13 に示す．

Vol.2010-CSEC-50 No.48
2010/7/2

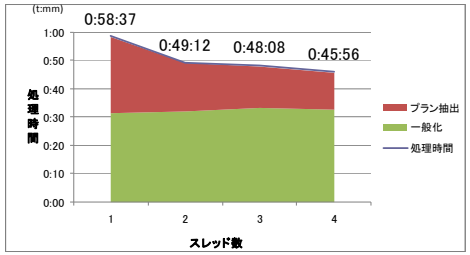

図 12　スレッド数と処理時間（PTD）

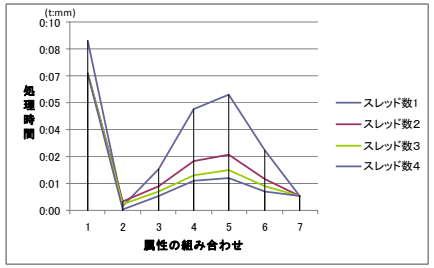

図 13　処理時間の詳細（PTD）

図 12 から，スレッド数の増加により，匿名化プラン抽出に要する時間が短縮できたことが確認できる．一方，匿名化プランにしたがってデータ全体を加工する一般化処理はマルチスレッド化していないため，ほぼ同じ処理時間を要する．

また，匿名化プラン抽出における処理時間の内訳（図 13）を見ると，探索処理の過程において，処理時間を要する組合せ属性数 3～6 属性の時に，スレッド数増加の効果が良く現れている．

全体としてスレッド数に応じた処理時間の短縮が見られ，マルチスレッド化の効果が確認された．

(2) 匿名化プラン抑制の効果

匿名化プラン抑制が処理性能に与える効果を検証した．Person Trip Data を対象に k の値を 2 とし，匿名化プラン数を 1，5，10，抑制なしとして匿名化を行った．結果を図 14 に示す．

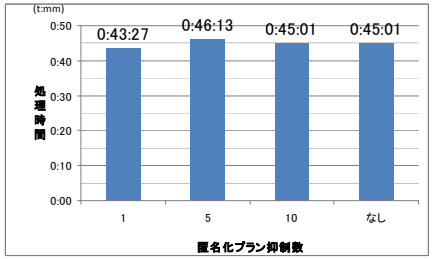

図 14　匿名化プラン抑制数と処理時間（PTD）

匿名化プランを抑制した場合と抑制なしとした場合の差分は少なく，本検証では，匿名化プラン抑制の明確な効果は見られなかった．

4.5 クラスタリング型 k-匿名化の性能要件

クラスタリング型 k-匿名化に対して，処理時間に影響を与えると考えられるレコード数，許容匿名度指定値（k の値），準識別情報の数，グループ化閾値といった要因について処理時間の測定を行い，基本的な性能要件について検証した．

検証データとしては 9 属性 45,222 件の Adult Data Set と擬似 Adult Data Set を用いた．なお，以下の評価は検証環境 2 で行なった．

(1) レコード数と処理時間

匿名化対象データのレコード数とクラスタリング型の処理時間の関係を検証した．

ADS から抽出した 10,000 件～45,222 件のレコードと，擬似 ADS として生成した 10,000 件～40,000 件のレコードに対し，匿名化処理を行った．いずれも 9 属性を対象として k の値は 2 とし，グループ化閾値を 0.5 とした．結果を図 15，図 16 に示す．

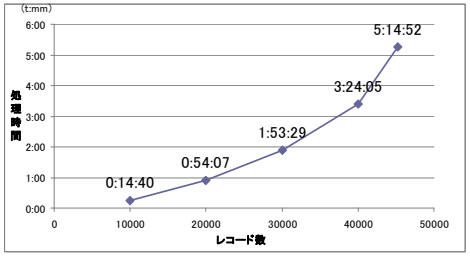

図 15　レコード数と処理時間（ADS）

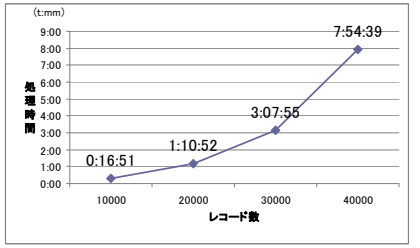

図 16　レコード数と処理時間（擬似 ADS）

図 15，図 16 から明らかに，レコード数が増加すると処理時間は増加し，その増加は指数倍数的である．凝集型のクラスタリングアルゴリズムでは，レコード数を n とし，共通の準識別情報の値の組み合わせを持つレコード集合の数を|E|とおくと，処理時間が $O(n \log n + |E|^2)$で表せる[6]ため，概ね妥当な結果といえる．

なお，9 属性 45,222 件での処理時間が 5 時間 14 分を要していることから，夜間バッチ処理の許容時間を 6 時間とした場合，9 属性 5 万件程度のデータ規模がクラスタリング型 k-匿名化の性能限界と考えられる．

(2) k の値と処理時間

次に，許容匿名度指定値（k 値）が，処理性能に与える影響について検証した．

ADS 45,222 件に対し，k の値を 2，5，10，15，50，100 と変えて匿名化を行った．

また，擬似 ADS についてはレコード数を 10,000 件とし，k の値を 10，20，50，100 と変えて匿名化を行った．いずれも 9 属性を対象に，グループ化閾値を 0.5 とした．結果を図 17，図 18 に示す．

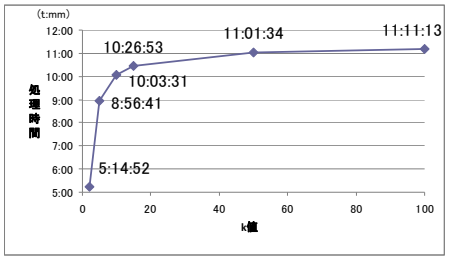

図 17　k の値と処理時間（ADS）

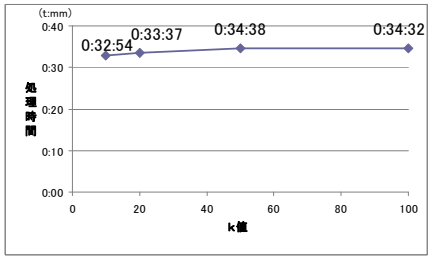

図 18　k の値と処理時間（擬似 ADS）

クラスタリング型では階層探索型と異なり，k 値を大きくするほど処理時間が増加する結果となった．なお，k=50 以降の処理時間は横這いになっている．

これは，許容匿名度指定値が大きいほど k を満たさないグループの数が増加し，これらのグループの間で再帰的にグループ化を繰り返すために，処理時間が増加する．ただし，k を満たさないグループ数が全グループ数と一致した時，処理時間の最大値になる．k=50 以降は k を満たさないグループ数と全グループ数がほとんど同じため処理時間が増加しない．

(3) 準識別情報の数と処理時間

ADS 45,222 件を対象に，準識別情報の指定を 1 つずつ増やした時の処理時間を測定した．結果を図 19 に示す．

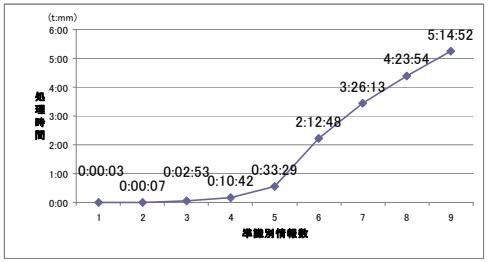

図 19　準識別情報の数と処理時間（ADS）

図 19 から，準識別情報の数が増えると処理時間が増加することは明らかである．クラスタリング型の k-匿名化は，k-匿名性を満たさないレコードのグループ化を繰り返すため，準識別情報数が増えるほど k を満たさないグループ数が増加し，処理時間が増加する．9 属性 45,222 件の ADS においては 5 属性を前後に処理時間が跳ね上がっており，この傾向が顕著であったといえる．

(4) グループ化閾値と処理時間

グループ化閾値が処理性能に与える影響について検証した．ADS 45,222 件を対象に，k の値を 2 とし，グループ化閾値を 0.1，0.3，0.5，0.7，0.9 と変えて匿名化した．結果を図 20 に示す．

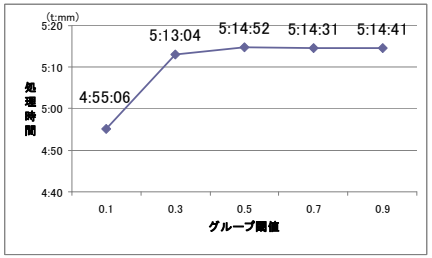

図 20　グループ化閾値と処理時間（ADS）

グループ化閾値を 0.1 にした場合，グループ化する距離が短く，近傍レコードと結合されない孤立レコードが多数削除されるため，処理時間が短くなる結果となった．0.5 以降の値ではほとんど影響は見られなかった．本検証では明確な傾向は確認できなかったが，対象データにおける値の分布に依存すると考えられる．

(5) k の値と情報損失

クラスタリング型において，k の値が情報損失に与える影響について検証した．ADS 45,222 件に対し，k の値を 2，5，10，15，50，100 と変えて匿名化を行った．また，擬似 ADS に対し，レコード数を 10,000 件，k の値を 10，20，50，100 とした．いずれも 9 属性を対象に，グループ化閾値を 0.5 とした．結果を図 21，図 22 に示す．

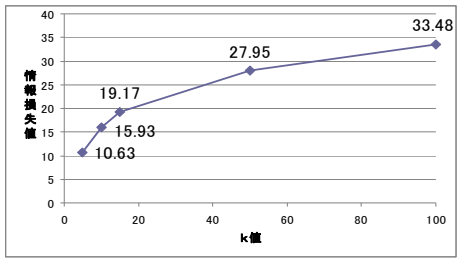

図 21　k 値と情報損失（ADS）

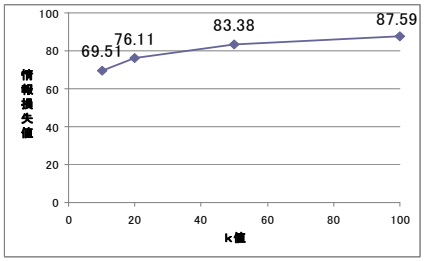

図 22　k 値と情報損失（擬似 ADS）

k 値が大きくなると，情報損失が大きくなる．検証結果より，クラスタリング型 k-匿名化機能では，安全性（k-匿名性）が大きくなるほど，有用性が下がる（情報損失が大きくなる）トレードオフの関係が確認できた．

4.6　階層探索型とクラスタリング型の性能比較

ADS 45,222 件 9 属性に対し，階層探索型でレコード削除なし，10%レコード削除，20%レコード削除の 3 通りの設定と，クラスタリング型を用いて匿名化を行なった結果を示す．なお，検証環境 2 で，k 値は 2 とし，階層探索型はスレッド数 4，匿名化プラン抑制なし，クラスタリング型はグループ化閾値 0.5 とした．結果を表 4 に示す．

表 4　階層探索型とクラスタリング型の性能比較

	階層探索型（レコード削除なし）	階層探索型（レコード 10%削除）	階層探索型（レコード 20%削除）	クラスタリング型
k 値	2	2	2	2
レコード削除件数	0	4460	8682	0
処理時間	00:05:44	01:06:29	00:20:31	05:14:52
情報損失	45.00000	29.95415	24.98607	3.76743
母集団一意性	0.00020	0.01117	0.01117	0.04390

レコード削除機能の有無による階層探索型の処理性能を比較すると，一定数以下のレコード削除を許容して k-匿名性を満たさない匿名化プランも探索する，レコード削除ありの階層探索型は処理時間がかかる結果となった．

また，レコード削除ありの場合でも，10%削除と 20%削除とで，後者の方が早い結果となった．これは，20%削除の場合は 10%削除に比べラティスの枝刈りが発生せず，頻度表の作成回数が減ったため処理時間が早くなったものと考えられる．

k-匿名化機能の処理時間の比較という面では，クラスタリング型は，5 時間強と大幅に遅い結果となり，両者のアルゴリズムの特性通りの結果が出たといえる．

5.　検証結果の考察

検証結果から，階層探索型の k-匿名化機能について，以下のことが明らかになった．

- 1,000 万件規模のデータを処理する場合に，6 時間程度で完了すると想定され，バッチ処理として十分実用的な処理性能を有していると言える．
- 匿名化の処理時間は，主にレコード数，準識別情報の数，許容匿名度指定値（k の値）によって影響され，レコード数，準識別情報の数が増えると処理時間が増加する．一方，k の値は大きくなるほど処理時間が減少する結果となった．
- k の値と情報損失の間に明確な関係性はなく，抽出された匿名化プランによる各属性情報の一般化の程度によって情報損失の値が異なる結果となった．
- 追加機能によるマルチスレッド化，匿名化プラン抑制により，処理時間が短縮された．特に，マルチスレッド化による属性の組み合わせに対する匿名化プラン探索の処理時間を削減する効果が見られた．

階層探索型の k-匿名化機能は，与えられた一般化階層情報に基づくラティス構造を構築して，k-匿名性を満たす匿名化プランを探索するアルゴリズムであり，ラティス構造が大きい程処理時間を要する．ラティス構造の大きさは，準識別情報の数と各準識別情報の一般化の深さ（階層数）によって一意に定まる．ただし，アルゴリズムはラティス構造の枝狩りによって探索の効率化を図るため，実際に行われる探索処理の範囲は匿名化対象のデータでの準識別情報の値の分布と指定された k の値によって異なり，処理時間が大きく異なる結果となるものと考えられる．

処理時間について詳細なログを分析したところ，匿名化プラン探索の処理時間は主に，頻度表の作成処理の時間と回数の積，頻度表の更新処理の時間と回数の積の和で表すことができることが分かった．レコード数，準識別情報の数は頻度表の作成処理時間に影響し，一般化階層情報の深さや幅（一つの階層に含まれる値の種類数）が頻度表の更新処理時間に影響する．前者の方が処理として重いため，主にレコード数と準識別情報の数によって，全体的な処理時間が影響される．なお，頻度表作成処理および更新処理の回数はラティス構造上での探索処理により決まるため，前述のように準識別情報の値の分布と指定された k の値によって処理時間が異なる結果となる．

一方，クラスタリング型の k-匿名化機能について，以下のことが明らかになった．

- 匿名化の処理時間は，主にレコード数，準識別情報の数，許容匿名度指定値（k の値）によって影響され，レコード数，準識別情報の数が増えると指数倍数的に処理時間が増加する．一方，階層探索型と違って，k の値が大きいほど処理時間は反比例的に増加し，一定以上の値で処理時間が飽和する結果となった．
- k の値と情報損失の間には関係性があり，k の値が大きいほど有用性が下がる（情報損失が大きくなる）トレードオフが確認できた．

クラスタリング型の k-匿名化機能は，共通の値または近い値の組み合わせを持つレコード同士の距離を計算して再帰的にグループを作る凝集型のアルゴリズムであるため，異なる値を持つレコードの数が多く，匿名化対象のデータ中で形成されるグループの数が大きくなるほど，処理時間を要する．特に，レコード数や準識別情報の数はグループ数に影響し，処理時間が増加する．一方，k の値は大きいほど再帰的なグループ化の回数が増えるため処理時間が増加するが，グループ数としては徐々に減少するため，処理時間が飽和する結果となったと考えられる．

階層探索型とクラスタリング型の比較では，アルゴリズムの特性通りの結果が見られ，クラスタリング型の方が処理時間は遅いが，有用性の高い（情報損失の低い）匿名化データが得られることが確認できた．事業者によるパーソナル情報の利用目的や用途によって求められる匿名化データの一般化の仕方や有用性は異なるため，特徴の異なる両アルゴリズムの実装は汎用的な匿名化基盤を考える上で重要である．

6. おわりに

本稿では，情報大航海プロジェクトにおいて構築した個人情報匿名化基盤について，その設計思想と実現機能，および実規模データを想定した検証の結果について，整理を行なった．構築した個人情報匿名化基盤は，以下の特徴を持つ．

- 共通実行基盤機能とプラグ＆プレイ機能群から構成するアーキテクチャにより，事業者による多様な利用形態に柔軟に適用できる機能構成とした．
- 階層探索型の k-匿名化をベースに，マルチスレッド化，匿名化プラン抑制といった機能追加により，実規模データに対応可能な性能を持つ匿名化技術を実装した．
- k-匿名性とあわせて情報損失や母集団一意性などの評価指標により，事業者による安全性と有用性のバランスの取れた匿名化処理を可能とした．

情報大航海プロジェクトの成果はオープンソースとして公開されている 1]．技術的な機能や性能のブラッシュアップ，更なる検証などについて「次世代パーソナルサービス推進コンソーシアム」[13]での検討が予定されている．

一方，パーソナル情報の利活用にともなう市場動向の変化は法制度面で制約されることが想定されるため，個人情報利活用におけるルール化の推進が必要である．そのために，技術開発に加え，パーソナル情報の利用に係るガイドライン，品質評価基準，認証方法などの検討が両輪として進められる必要がある．

本稿がこうした活動の一助となれば幸いである．

謝辞 本研究は経済産業省「情報大航海プロジェクト」における「パーソナル情報保護・解析基盤の開発・改良と検証」の一環として実施した．多大なご協力をいただいた関係者の皆様に謝意を表する．また，パーソントリップデータの利用は，東京大学空間情報科学研究センターとの共同研究「個人情報の匿名化とその 2 次利用について（情報大航海プロジェクト）」に基づく．貴重かつ有用なデータのご提供に謹んで感謝の意を表する．

参考文献

1) 情報大航海プロジェクト，http://www.meti.go.jp/policy/it_policy/daikoukai/igvp/index/index.html
2) Sweeney, L.: k-anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge Based Systems, 10(5), pp.557-570 (2002).
3) Samarati, P. and Sweeney, L.: Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression, Technical report, SRI International (1998).
4) Sweeney, L.: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 10(5), pp.571-588 (2002).
5) LeFevre, K., DeWitt, D.J. and Ramakrishnan, R.: Incognito: Efficient Full-Domain K-Anonymity, In Proc. of the ACM SIGMOD Conference on Management of Data (2005).
6) Jiuyong Li, Raymond Chi-Wing Wong, AdaWai-Chee Fu, and Jian Pei: Achieving k-Anonymity by Clustering in Attribute Hierarchical Structures, DaWaK 2006, LNCS 4081, pp.405-416 (2006).
7) J. Domingo-Ferrer, J.M. Mateo-Sanz, V. Torra: Comparing SDC Methods for Microdata on The Basis of Information Loss and Disclosure Risk, Pre-proceeding of ETK-NTTS, pp.807-826 (2001).
8) Byun, J.W., Kamra, A., Bertino, E., and Li, N.: Efficient k-Anonymization Using Clustering Techniques, In DASFAA 2007, LNCS 4443, pp.188-200 (2007).
9) Franconi, L. and Polettini, S.: Individual Risk Estimation in μ-Argus: A Review, In Proceedings of Privacy in Statistical Databases 2004 (PSD2004), LNCS 2316, pp.262-272 (2004).
10) Polettini, S.: A Note on The Individual Risk of Disclosure, Istituto Nazionale di Statistica (2003).
11) UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/
12) 東京大学空間情報科学研究センター：人の流れプロジェクト, http://pflow.csis.u-tokyo.ac.jp
13) 次世代パーソナルサービス推進コンソーシアム, http://www.coneps.org/