情報処理学会論文誌 Vol.57 No.9 2034–2044 (Sep. 2016)

累乗近似式を用いたk-匿名化処理の効率化

小栗 秀暢[1,3,a)] 曽根原 登[2] 松井 くにお[3] モハマド ラスール サラフィ アグダム[1]

受付日 2015年12月4日, 採録日 2016年6月2日

概要：ビッグデータ分析においてパーソナルデータの個人識別性を減少させ，安全性を高める手段として，k-匿名化処理が利用される．属性の出現数を統計的に利用するk-匿名化処理を行う場合，処理負荷が高く，処理結果のk-匿名性の予測ができないという難点がある．そこで，大規模なデータから一定規模の情報を抽出する場合，出力された結果の多くが正規分布に準じるという性質から累乗近似型の予測式を提案した．提案した予測式は重相関係数0.9以上の値で実測値と近似したが，予測誤差が発生するため正確な数値が必要な場合に利用が難しい．そこで，匿名化可能な情報区分を予測し，予測地点から匿名化処理を開始するアルゴリズムを提案した．実験によって，一般的な匿名化アルゴリズムと比較したところ，$k \geqq 50$のとき3.5%～12.5%の処理量で匿名化処理が達成できることを確認した．

キーワード：プライバシー保護，k-匿名性，匿名化処理

An Efficient k-anonymization Algorithm by Predictive Model of the Power Approximation

Hidenobu Oguri[1,3,a)] Noboru Sonehara[2] Kunio Matsui[3]
Mohammad Rasool Sarrafi Aghdam[1]

Received: December 4, 2015, Accepted: June 2, 2016

Abstract: Privacy is an important issue in big data analysis. An enormous amount of calculation is required to decrease the privacy risks associated with various data. The k-anonymity model is widely used to protect privacy, but it is difficult to predict the k-anonymity of datasets that are processed statistically. We proposed a k-anonymity predictive model that uses a power approximation based on the property that most data have a normal distribution when extracted at a constant scale from large-scale data. This predictive model took out multiple correlation coefficient scores of more than 0.9. However, if there are prediction errors, the model cannot be used when correct numerical values are required. We therefore proposed an anonymizing algorithm that starts processing from a prediction spot and compared it experimentally with a general anonymizing algorithm. We found that processing could be achieved with a throughput of 3.5%–12.5% at $k \geqq 50$.

Keywords: Privacy preserving, k-anonymisation, Anonymising method

1. はじめに

近年の個人情報保護意識の高まりによって，個人情報を保持する事業者は，情報の有効利用を促進する施策と，情報の漏洩や不正利用を防止する施策を両立させることが求められるようになってきた．

情報内からセンシティブな要素を排除することでコンプライアンスリスクを軽減させ，データを他社に提供し，分析やマーケティングなどに利用する手段として，個人情報

[1] 総合研究大学院大学複合科学研究科情報学専攻
The Graduate University for Advanced Studies, School of Multidisciplinary, Informatics Department, SOKENDAI, Chiyoda, Tokyo 101–8430, Japan
[2] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[3] ニフティ株式会社
NIFTY Corporation, Shinjuku, Tokyo 169–8333 Japan
[a)] oguri.hidenobu@nifty.co.jp

の匿名化技術が有望視されている．特に k-匿名化処理 [1] をはじめとする個人の特定性・識別性を低減させる手法は，他のデータベースとの名寄せによる結合や公開情報どうしの再結合と再識別化を防ぐ手段として効果的である．

高いレベルで k-匿名化が施されたデータ群は，再識別化による攻撃や，悪用の可能性が低くなるため，個人情報よりも簡単な手続きで利用可能となり，第三者へのデータ提供によって，ノウハウの共有やマーケティング分析，協調フィルタリングによるレコメンドエンジン [2] などへの活用が期待できる．そのため，政府や企業が保持する情報について，利用目的に合致した形で匿名化処理を行い，安全性基準を満たして提供する手法が必要とされている．

だが一方で，匿名化処理は計算コストが高く，最適な k-匿名化の実現は NP 困難な問題として知られている [3]．

k-匿名化処理は，属性情報どうしを組み合わせた際の最小出現数（k 値）が，求める基準以上に存在するかを調査する．

もし組み合わせた属性情報が k-匿名性を満たさない場合，属性値をより粗い粒度のデータに抽象化し，属性値の書き換えを行うことで，安全性を高める処理を行う．

通常，ある匿名化データの提供要求があった場合，その条件に沿って処理されたデータが，k-匿名化状態を満たすか予測できない．そのため，匿名化処理した結果が，利用者のデータ利用目的に合致しない場合は，属性値を変更するなど，処理条件を変更して複数回の処理が行われる．このようなデータ提供者の負担を軽減するための，軽量な匿名化処理アルゴリズムが求められている．

そこで本稿では k-匿名化状態を満たす情報を効率的に導くアルゴリズムを提案し，実データを用いて評価する．具体的には，k-匿名性を予測する近似式を用いて，匿名化処理の適切な開始地点を導き，他の匿名化処理アルゴリズムと組み合わせることで不必要な匿名化処理を排除する方式である．

本稿の構成は次のとおりである．2 章で k-匿名化における情報加工の特徴と問題点について述べる．3 章で匿名化処理の従来研究．4 章で k 値と属性情報を組み合わせた際の区分数との関係性について分析し，k 値の予測モデルを提案．5 章でその予測モデルを実データで検証．6 章でその予測式を用いた処理削減アルゴリズムを提案し，7 章で結果をまとめる．

2. k-匿名化処理の特徴と問題点

まず，匿名化とは，パーソナルデータを加工して，個人識別性を減少させる処理のことである．

パーソナルデータとは「属性」と「属性値」として表現される個人に関するデータであり，ある個人のパーソナルデータをテーブルのレコードとして表現する．

そして，単一の属性では個人を特定できないが，複数

表 1 一般化階層の例
Table 1 Sample of generalization taxonomy.

性別	2区分	男				女			
年齢1	2区分	10代				20代			
年齢2	4区分	10–14才		15–19才		20–24才		25–29才	
年齢3	20区分	10才	‥	15才	‥	20才	‥	25才	‥

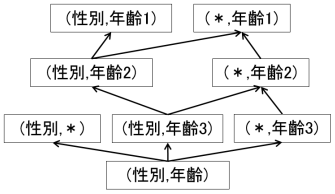

図 1 (性別, 年齢 1, 2, 3) を用いた Lattice Structure の例
Fig. 1 Sample of lattice structure (Gender, Age1, 2, 3).

組み合わせると特定できる可能性のある属性を準識別子（quasi-identifier：QID）と呼ぶ．また，個人を特定された状態で開示されることが望ましくない属性をセンシティブ属性（sensitive attribute：SA）と呼ぶ．このとき，もし攻撃者がある個人の QID の属性値を知っていたとすると，そのレコードを特定できてしまい，SA の属性値を知られてしまう．

これを防ぐために，QID の属性値を一般化して，より抽象的な値にする．そして，QID の属性値によって識別されるレコードが少なくとも k 個（$k > 1$）以上ある場合，そのテーブルは k-匿名性を満たすという [1]．

匿名化処理によって QID を書き換える際には，匿名化条件を満たさない属性値を抽象度の高い候補に書き換える．

あるパーソナルデータに性別と年齢属性が含まれており，**表 1** の一般化階層を用いて k-匿名化処理を行う場合，**図 1** のような Lattice Structure [4], [5]（格子構造）を作成し，属性どうしの全組合せを作成し，それぞれの属性組合せにおける k 値を調査する．

だが，データの抽象化処理によって，分析対象の削除や過度な変更が発生し，データ利用目的が損なわれる場合がある．たとえば，目的が自動車免許に関係する場合，18 歳以上という属性値の区切りが必要となる．そのため，図 1 で作成した Lattice Structure から，(性別, 年齢 2) の結果データを出力されても，分析目的が達成できない．

データ利用者は，年齢 3 を含むデータの出力を求め，それが匿名化条件を満たさない場合，新たに年齢 4 = (18 歳未満, 18 歳以上) などの新しい一般化階層を提案して，再度処理を要求する場合がある．

そこで，独立行政法人統計センターなどの機関では，学術利用に限定して，個別利用者が必要とするデータ区分をリクエストしてデータを加工する「オーダーメード集計」[6], [7] を提供することで，再識別可能性とデータ漏えいリスクを制御しつつ，個別のデータ利用ニーズに対応している．

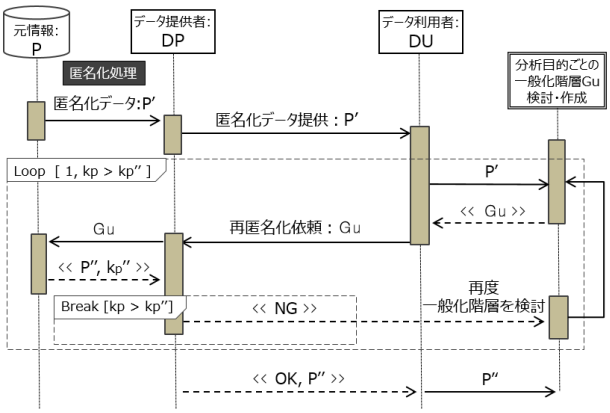

図 2 オーダーメード型匿名化処理のシーケンス図
Fig. 2 Sequence of tailor-made anonymization process.

オーダーメード集計では，データ利用者が，パーソナルデータの書き換え要求をまとめた統計作成仕様書を作成し，提供者と利用者間で折衝を繰り返してデータを提供する方式を採用している．統計作成仕様書には，加工する統計調査名，および集計対象となる属性項目とその区分種類，属性区分数を記載し，データの再集計を依頼する．

図 2 はこの方式を参考に，オーダーメード型匿名化のシーケンス図を検討した例である．

図 2 の処理は，以下の手順で行う．データ提供者 DP はパーソナルデータ P に対して，k-匿名性における k 値 kp（kp > 1 の整数）を満たす匿名化データ P' を作成し公開した．

だが，匿名化データ P' はデータ利用者 DU が求める分析目的が達成できず，DU は新しい一般化階層 Gu を作成し，再匿名化処理を依頼した．DU は P についての知識はなく，また目的外利用を禁止するなどの利用条件があるものとする．

DP は P に対して Gu を適用した結果が k-匿名性を満たすかが不明であることから匿名化処理を複数回試行する．その際に個人情報を含むデータベースに対して処理量の大きい匿名化処理を要求するため，作業負荷が大きい．

一方 DU は P のユーザ分布を知りえないため，新たに目的を達成するための，妥当な Gu を検討する指標がなく，双方の不一致問題が繰り返し発生する可能性がある．

これらの DP，DU 双方の作業負荷が匿名化データの流通を妨げている原因の 1 つと考える．

他の機関からの多様な要求に対応して匿名化データを作成するため，データ提供者 DP が求める k-匿名性を満たし，かつデータ利用者 DU のデータニーズに合致した匿名化データを，軽量に作成するアルゴリズムが必要とされている．

3. 情報の匿名化に関する従来研究

匿名化データは，多くの研究や分析に活用できるが，データの特徴や分析目的に対応するため，多様な匿名化処理アルゴリズムが考案されている．

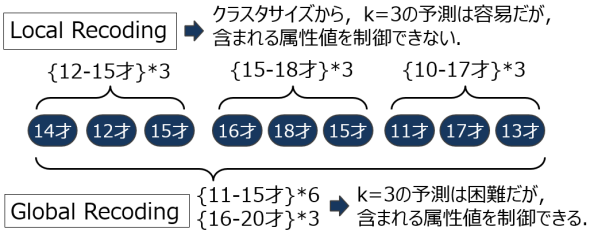

図 3 Recoding 手法の特徴
Fig. 3 Characteristics of recoding method.

表 2 匿名化処理における Recoding 手法の比較
Table 2 Comparison of recoding methods.

	データ処理量	含まれる属性値の制御	k-匿名性の予測
Local Recoding	少ない	困難	容易
Global Recoding	多い	容易	困難

データの k-匿名化を実現するための手法として，Datafly 方式 [1] や μ-Argus 方式 [5], [8], [9] などのアルゴリズムによって情報を書き換える（Recoding）処理が主に使われており，公共データや医療データの配信システムとして利用されている．

情報の Recoding の方法は，大きく分けて，局所的な処理である Local Recoding と，属性値全体の統計情報から処理を行う Global Recoding の 2 種類が存在する．

Local Recoding はデータベースの部分集合に対して匿名化処理を行うため，単位あたりの処理量が少なく，分散処理に向く．また，各クラスタサイズを自由に設定できるため，k-匿名性の予測は容易だが，クラスタに含まれる属性値の制御が難しい．

Global Recoding はデータベース全体を用いて統計情報を作成するため，処理量は多いが，利用者が求める属性値について，統計情報を参照しながら調整できるという利点がある．だが，求める属性値による処理の結果が，求める k-匿名性を実現するかについて，予測は難しい（**図 3**，**表 2**）．

一般的に，位置情報や購買ログなどのトランザクション型データは，各属性値の出現数が頻繁に変化するため，局所的な特徴で情報をクラスタ化する方が効率的である．そこで有用性と匿名性を維持しつつクラスタ化処理する [10], [11], [12] などの Local Recoding アルゴリズムを採用し，行動分析や機械学習に利用することが多い．

Global Recoding は，住民台帳やサービス登録情報など，ある程度固定化されているマスタ型（ミクロデータ型）の情報に対して採用し，各属性値の出現数を統計化したうえで匿名化処理する．結果データは統計処理化や，回帰分析などに活用される．

本稿では，公共情報やサービス会員情報など，大量の情報から，分析目的に沿った匿名化データを抽出する方式を

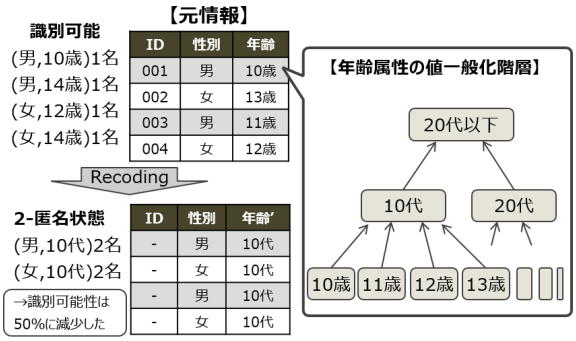

図 4 2-匿名化処理の書き換え例
Fig. 4 Sample of 2-anonymous process.

想定していることから，Global Recoding について詳細を記述する．

Datafly 方式をはじめとする Global Recoding では，主に各属性値の出現数を調査しながら，匿名化条件を満たさない属性値を抽象度の高い候補に書き換えるという，一般化階層型の集合匿名化処理を行う．

たとえば，**図 4** において元情報の (男, 10 歳) という準識別子を組み合わせた属性値の出現数は 1 である．そのため ID を消去したとしても，属性の組合せによる再識別が可能であり，k-匿名状態の条件（$k > 1$）を満たしていない．

そこで，匿名化処理として値一般化階層 VGH [1] や，属性一般化階層 DGH [13] などを利用し，出現数の少ない属性値を抽象度の高い属性値に書き換える．

VGH と DGH の差は，書き換えの際にすべての属性値を同じ階層で処理するか否かにある．たとえば，VGH は値単位での抽象化を行うため，図 4 の例の場合，(10 代, 20 歳, 21 歳) のように異なる階層が混在した結果が出力される．逆に DGH を用いた場合，1 属性値でも匿名化が達成できない場合，階層全体を変更するため，(10 代, 20 代) のような，同一の階層による出力結果が得られる．本稿では基本的には VGH の考え方で処理を行うものとする．

Global Recoding では値の書き換え前後に各属性値の出現数の検証を行い，匿名化処理の条件やデータの利用用途の条件を満たすまで，処理を繰り返す．

出現数の検証処理は作成される属性どうしの組合せの数だけ必要となる．Meyerson らの研究によると，最適な k-匿名化の実現は NP 困難 [3] であり，属性数の多い個人情報では容易に計算困難となる．

そこで，このような匿名化処理にともなう組合せ爆発状態を回避するためのアルゴリズムが多く提案されている．

Incognito [4] は，トップダウン型で属性の抽象化候補を探索する中で，匿名化処理ができない属性が判明した場合，その属性を含む組合せをその後の計算から排除し，不必要な処理量を減少させている．**図 5** にて Incognito の概要を示す．

また OLA [14] (Optimal Lattice Anonymization) では，

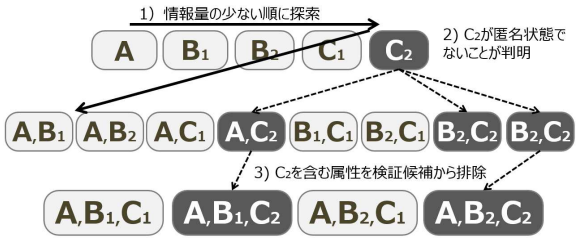

図 5 Incognito 方式による検証量削減方式
Fig. 5 Sample of Incognito method.

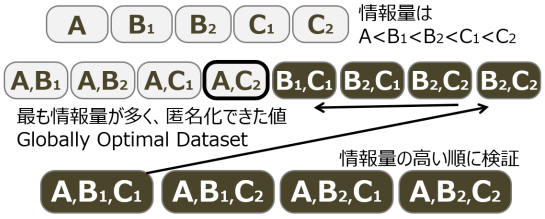

図 6 OLA 方式における最適値の検証順番
Fig. 6 Sample of OLA method.

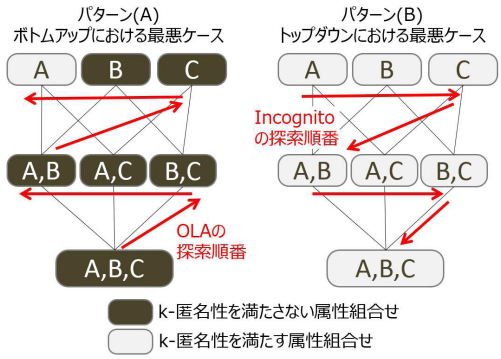

図 7 各アルゴリズムにおける最悪ケース
Fig. 7 Worst case on bottom-up and top-down.

情報量に着目して最適な匿名化可能な属性の組合せを，ボトムアップ型で導き出す方式を提案している．出現した属性値の組合せを情報量の多い順番にソートし，最も情報量が多く，匿名化条件を満たした群を “Globally Optimal Dataset” として利用する．**図 6** にて OLA の概要を示す．

これらの方式は，ボトムアップやトップダウンなどの順に匿名化処理可能な組合せの探索を行い，条件を満たした場合にその後の処理を省略することで処理量を削減する．そのため処理削減効果は属性値の出現数の特徴に依存し，作業量を事前に予測することが難しい．

たとえば，**図 7** におけるパターン (A) はボトムアップにおける最悪ケースの例である．最も抽象的な情報でしか k-匿名化処理ができないデータを匿名化処理するとき，OLA 方式では実質的に全組合せの探索処理が必要となる．同じくパターン (B) は，トップダウン方式における最悪ケースである．すべての組合せにおいて匿名性を満たせる場合，最も詳細な組合せまで全探索を行う．

このような現象は，データの特徴や達成するべき k 値な

どによって変化するため，最適なアルゴリズムを選択することは困難である．

そこで我々は，データの特徴に依存せず，安定して処理削減効果を高めることができる匿名化処理アルゴリズムを検討した．

4. k-匿名性の予測近似式の提案

まず，k-匿名性の性質について検討する．

本研究は，公共データなどの大きなデータ群から，特定の条件に合致した群を抽出し，その情報に対して一般化階層を適用して匿名化処理を行う方式を想定する．

通常，複数の抽出条件によってデータの出現量を予測する場合は，多次元正規分布の同時密度関数によって出現率を予測する．だが，その際には，全組合せによる分散と相関係数を導く必要があるため，各一般化階層における属性値の出現数調査を，Lattice Structure における組合せ回数だけ行い，同時密度を計算する必要がある．これは匿名化処理以上のコストがかかるため，実用的でない．

逆に，出力された結果から考えると，抽出条件によって，分析に耐えうる十分なデータ量が出力される場合，出力データを基準化すると中心極限定理によって標準正規分布に近似する．その分布の特徴から k 値を予測する方が効率的である．

出力されたデータが正規分布である場合，k 値は，データの全体数 P に対して，属性区分数 x 個のクラスタ化を行った場合のクラスタサイズの最小値として定義でき，その場合の k 値は平均値から最も遠い位置にある．

まず，各属性の区分によって生成されたクラスタの出現確率が，標準正規分布であった場合の k 値を検討する．

全体数 P に対して，属性区分数 x でクラスタ化した際の結果分布がつねに標準正規分布であると設定する．**図 8** にてその分布の例を示す．そのときのクラスタサイズの最小値 $k_{(x)}$ は，標準正規分布の最端値として存在し，平均値から最端値までの距離を $a\sigma_{(x)}$ と設定する．

この $k_{(x)}$ と $k_{(x+n)}$ の関係性を調査するため，区分数が均一に増加する正規分布によるクラスタサイズの最小値の推移について実験を行った．

図 9 は分散が 1 である正規分布で 10 万サンプルを生成し，最端を 4σ とした場合のクラスタサイズの最小値（k 値）と設定し，属性区分数を増加させ，各区分数で 50 回試行した際の平均 k 値の推移である．x 軸が属性区分数，y 軸が k 値である．図 9 によって，平均 k 値は，$48770x^{-3.21}$ で $R^2 = 0.9612$ で近似することが判明した．

実際には，それぞれのデータの傾向によって異なる分散の特徴を持つため，実データにおける k 値は $\boldsymbol{P} * \boldsymbol{x}^{-\boldsymbol{n}}$ のような，べき乗型で漸減する傾向があると仮説を設定する．

そこで，少量のサンプルを用いた匿名化処理の結果を用いて，属性の区分数からその後の k 値を導くため，累乗近似型の予測式を提案する．

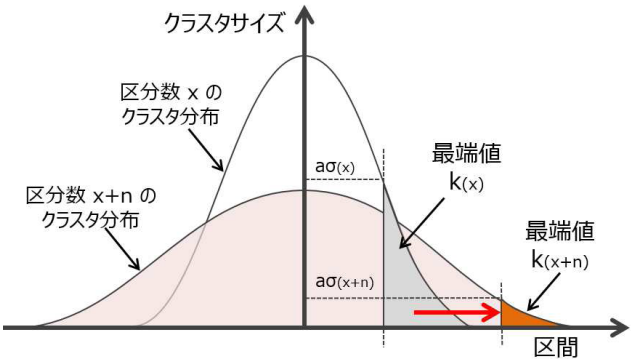

図 8 区分数の増加と最端値の変化
Fig. 8 Distribution on division and the most distant value.

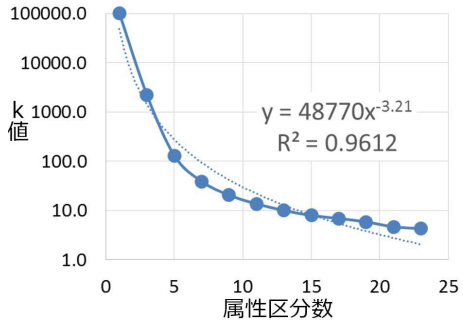

図 9 正規分布による k 値推移の実験結果
Fig. 9 k-values based on test results.

一般的なデータの匿名化処理を想定した場合，匿名化処理結果が 1 つも存在しないことは稀であり，通常は少量の匿名化の結果を保持していると考えられる．

それらの属性区分数を既知の x，その区分数における最小クラスタのサイズ（k 値）を既知の y と設定し，累乗近似式によって，その後の区分数における k 値を予測する．累乗近似式は，既知の x，y によって最小二乗法における切片 α と傾き β を求め，式 (1) に代入する形で求める．

$$\begin{aligned} \alpha &= \bar{y} - \beta\bar{x} \\ \beta &= \ln\left\{\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}\right\} \\ k &= \alpha x^{\beta} + 1 \end{aligned} \tag{1}$$

式 (1) では累乗近似式の結果に 1 を加えている．これは負の β 値を持つ累乗近似値は最終的に 0 に収束するのに対し，k-匿名性は $k = 1$ に収束するためである．これによって精度が高くなり，誤差計算も容易となる．

また，予測式の利用法としては，定められた k 値を満たす区分数 x を予測する場合もある．その場合は式 (1) に求める k 値を代入する形となり，式 (2) で表す．

$$\begin{aligned} x^{\beta} &= (k-1)/\alpha \\ x &= \frac{(k-1)^{1/\beta}}{\alpha^{1/\beta}} \end{aligned} \tag{2}$$

本予測式が成立する条件をまとめる．

1) 対象データが正規分布である．
2) 一般化階層の区分数と分散の増加量に規則性がある．
3) 既知の x，y として設定可能な，属性区分数と k 値の実測値が存在する．

我々は，国勢調査と N 社が運営するサービス群に登録している会員データに対し，区分数の異なる複数の一般化階層を適用し，k 値の推移を計測．その結果と予測式が出力する結果とを比較する実験を行った．

5. 実験：k-匿名性累乗近似式の比較と検証

実験は国勢調査，および N 社サービス会員データから，性別，年齢，地域の 3 属性を抽出して実施した．

N 社の会員データ群は，1635 サービスに対して 2013 年 10 月に 1 度以上課金決済を行った会員データの中から，対象となる 3 属性が含まれる 434 万人を抽出し作成した．

企業の持つサービス会員データは，多様な分布パターンがあるが，通常は，そのサービスがターゲットとしている顧客層を中心とした正規分布の性質を持つ．本対象データには，特定の地域にしか提供しないもの，男性の利用が多いものなど，同一の基準を用いた場合に k 値が低くなる可能性が高いものも多く含まれている．

これらのサービスごとの会員データに対し，登録人数によって階級を作成して**表 3** に分類する．本会員データの元情報は個人情報を含むため開示できないが，階級ごとに集計した統計値などは，文献 [15] などで発表しているため，必要な場合は筆者に問い合わせていただきたい．

国勢調査は，分布が会員データに比して均一に近いが，過去における「団塊世代」などは出生数が多く，現代に近づくと減少する傾向を持つ．結果として 10 年単位で出現数をクラスタリングすると，分散の小さい正規分布が成立する．属性区分数を大きくしても高い k-匿名性を維持することから，比較対象として採用した．

国勢調査（2010 年）のデータは，N 社のサービス登録人数とスケールを合わせるため 1/1000 に変更した．以下，国勢調査群はすべてこの 1/1000 の群を指すものとする．

それに対して適用する一般化階層は**表 4** の基準で作成した．この一般化階層は企業の中でマーケティング分析などに利用するものであるため，サービス会員データに対してある程度適合した情報区分であるといえる．

性別（2 区分），年代（3，5，9 区分），地域（2，9，47 区分）の 3 属性を組み合わせて 2 区分から 846 区分まで設定し，各区分における k-匿名性を計測した結果を実測値として使用する．

その実測値に対して，4 章で検討した累乗近似式が近似することを，実データを用いて検証する．

まず，国勢調査を用いて，仮説設定した累乗近似式を作成し，値の一致率を調査する．

属性区分数を既知の x，一般化階層を適用して 5 区分まで組み合わせた場合の k 値を既知の y と設定して，累乗近似式を作成し，実測値と比較したものが**図 10** である．重相関係数が 0.998 と非常に高い数値だが，標準誤差が 1790.88 と大きな値になる．

図 10 の結果を受け，他のデータ群においても累乗近似式を適用した．**表 5** は，対象データ群に対して，一律に一般化階層を適用し，9 区分までを既知の x，y として利用して作成した累乗近似式の回帰分析結果である．

結果は，サービス人数規模ごとにばらつきはあるが，す

表 3 対象ユーザの階級ごとの状況
Table 3 Property of the target classes.

階級	登録人数	サービス数	人数	平均ユーザ数
1	100001人以上	10	1,669,482	166,948
2	50001～100000人	16	1,147,872	71,742
3	10001～50000人	36	870,965	24,193
4	5001～10000人	34	241,927	7,116
5	1001～5000人	124	266,613	2,150
6	1000人以下	1415	148,063	105
	合計	1635	4,344,922	2,657

表 4 適用する一般化階層
Table 4 Values of generalization taxonomy.

属性		k 値（全体）	標準偏差（全体）	区分数	分類1	分類2	分類3	分類4	分類5	分類6	分類7	分類8	分類9
性別	性別1	998882	1645359	2区分	男性	女性							
年代	年代1	30515	1604320	3区分	未成年		成人				老人		
	年代2	175971	451613	5区分	20代以下			30代	40代	50代	60代以上		
	年代3	13166	480745	9区分	0代	10代	20代	30代	40代	50代	60代	70代	80代以上
地域	地域1	984954	1665056	2区分	東日本				西日本				
	地域2	11429	771477	9区分	北海道	東北	関東	中部	近畿	中国	四国	九州	沖縄
	地域3	6489	176167	47区分	北海道,青森…沖縄までの47都道府県								

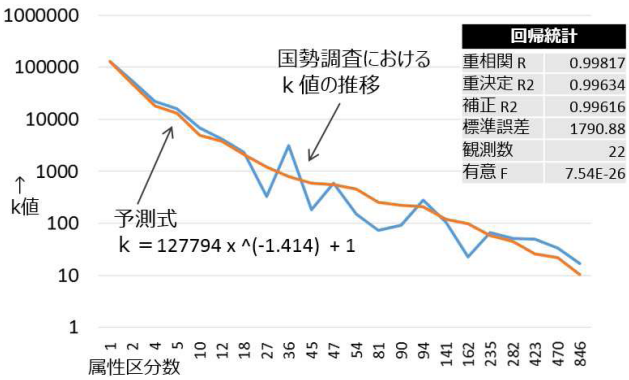

図 10 国勢調査の k-匿名性の実測値と予測値の比較
Fig. 10 Number of divisions of attributes (Census).

表 5 サービス人数ごとの回帰分析結果（人数規模比較）
Table 5 Comparison of regression investigation.

サービス人数規模	国勢調査	10万人～	5～10万人	1～5万人	5千～1万
対象サービス数	1	10	16	36	34
重相関R 平均	0.99705	0.99104	0.99143	0.99000	0.99047
重相関R 標準偏差	-	0.00344	0.00777	0.00870	0.00927
重決定R2 平均	0.99410	0.98217	0.98299	0.98018	0.98111
自由度修正決定係数 平均	0.99384	0.98139	0.98225	0.97932	0.98029
標準誤差 平均	2238.56	4267.16	1903.85	651.01	145.41
有意F 平均	3.8.E-27	4.1.E-20	1.7.E-17	2.5.E-16	7.3.E-17
平均誤差比率 平均	1.86	14.06	8.88	5.69	4.22

表 6 属性組合せ数による結果比較（全体平均）
Table 6 Comparison of attribute combination.

属性組合せ	1属性	2属性	3属性
重相関R 平均	0.98995	0.98807	0.99058
重決定R2 平均	0.98008	0.97637	0.98132
標準誤差 平均	1300.61	1852.58	1069.61
有意F 平均	1.1E−15	3.0E−16	1.2E−16
平均誤差比率 平均	6.83	8.03	6.52

べての群において，重相関係数が 0.99 を超え，自由度修正後の決定係数においても 0.97 以上の関係を持つ．また有意 F は 1.2E-16 以下と低い値となっており，回帰式としてあてはまりが良いことを示している．

また，累乗近似式を 3 属性以下の組合せで作成した場合について検証したものが**表 6** である．3 属性を組み合わせた k 値から作成された累乗近似式の重相関係数が最も高い．だが，1 または 2 属性によって得られた k 値から作成された近似式についても重相関係数が 0.98 を超えるため，簡易的なものとしては用いることが可能である．

これらの回帰分析結果の結果によると，重相関係数は高く出るが，標準誤差平均について，国勢調査で 2238.56 と非常に誤差が大きいことが判明した．だが標準誤差値は k 値が大きい場合における誤差も含めた平均値であるため，小さな k 値における誤差を表現するには不適当である．

そこで，k-匿名性が 1 に向けて収束していく性質から，各属性区分数の誤差を相対化して評価するため平均絶対比率 (3) を利用した．これは実測値 a と予測値 a' の比の大きい方を取得し，試行数 N で割り平均化したものである．

$$\frac{1}{N} * \Sigma\left(\max\left[\frac{a}{a'}, \frac{a'}{a}\right]\right) \tag{3}$$

これによって累乗近似型の予測式は，全体平均で 6.52 倍以内の値で予測が可能であることが判明した．このような誤差範囲の大きい予測式では，k-匿名性を満たす属性区分を正確に出力することは難しい．

だが，重相関係数が高いことから，匿名化可能な限界までの大まかな傾向をつかむことが可能である．

そこで，累乗近似型の予測式を用いて，匿名化処理が実現できる属性区分数を予測し，最も予測値に近い属性組合せから匿名化処理を開始する．その際に最適な匿名化アルゴリズムを選択することで，処理回数を削減するアルゴリズムを提案する．

6. 累乗近似式を用いた匿名化処理選択方式

k 値が累乗近似型に減少するという性質を持ち，その予測値と実測値が高い重相関係数を持つことが期待できる場合，最も予測値との差が小さい属性組合せを選択し，その地点から匿名化処理を開始する方式を提案する．また，その開始地点における属性組合せが k-匿名性を満たすかを確認することによって，ボトムアップ型，またはトップダウン型の匿名化処理を選択し，既存アルゴリズムの課題を解決する．

表 7 一般化階層の例
Table 7 Sample of generalization taxonomy.

A		2区分	男			女		
B	B1	2区分	10代			20代		
	B2	4区分	10–14才		16–19才	20–24才		25–29才
C	C1	2区分	東日本			西日本		
	C2	6区分	北海道・東北	関東	中部	関西	中国・四国	九州・沖縄

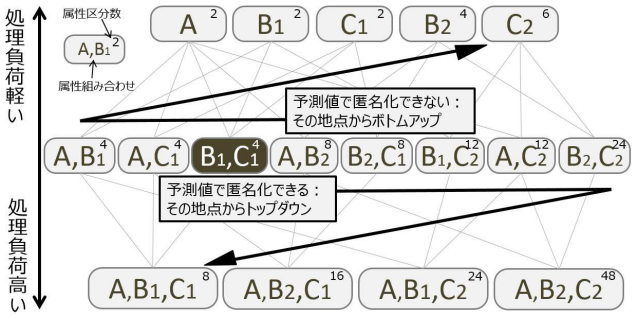

図 11 提案アルゴリズム概要
Fig. 11 Outline of proposed algorithm.

まず，Lattice Structure によって作成された属性の組合せについて，それぞれの属性区分数を調査する．

k-匿名化処理を行う候補として，属性 A，B，C に対して，一般化階層 A，B_1，B_2，C_1，C_2 を適用（**表 7**）し，最も詳細で k-匿名状態を満たす群を Lattice Structure の候補から探索する場合を想定する．また，アルゴリズム上で探索する際の優先度として C > B > A と設定する．これは，(A, B_1) と (A, C_1) は両方とも 4 区分であるため，探索の際の検証順を決定するためである．

対象データに一般化階層を適用し，属性の組合せとその区分数を記録する．たとえば，$(A, B_1) = 4$ 区分 {(男, 10 代), (女, 10 代), (男, 20 代), (女, 20 代)} を構成し，Lattice Structure を作成すると**図 11** の形で全候補が作成される．

図 11 の属性の組合せについて，区分数が少ないものから順番に，既知の y として使用する k 値を一定数取得し，記録する．

既知の x，y によって作成された累乗近似式から，k-匿名性を満たす属性区分数を予測し，Lattice Structure 内で最も差分の絶対値が少ない属性組合せを選択する．図 11 においては (B_1, C_1) が最も予測値に近い属性組合せであった場合を記載している．

最も予測値に近い属性組合せで，k-匿名化処理が達成できるならば，そこで匿名化処理終了である．そのうえで，より詳細な組合せの存在を疑うならば，その地点を開始点としてトップダウン型の匿名化処理アルゴリズム（Incognito など）を用いて匿名化可能な組合せを探索することも可能である．

逆に，指定した組合せでは匿名化処理が達成できない場合，その地点を開始点としてボトムアップ型の匿名化処理

アルゴリズム（OLA など）を用いることで検証回数を減少させることができる．

匿名化処理は組合せ数が増加するごとに処理数が増加するため，処理を行う範囲を限定することで，大きく処理回数を削減できる．

サンプルコード：累乗近似予測値による匿名化処理選択方式(PAK)

Input: 匿名化すべき複数属性(D_1,D_2...D_N)のテーブル *T*.
Power Approximation k-anonymity Method(PAK)::={
1. 処理開始：属性番号と属性値の種類を定義する．
Array[1]=D_1{ $d1_1$,$d1_2$...$d1_n$}… Array[N]= D_N{dn_1,dn_2...dn_m}
2. D_1〜D_n の一般化階層に沿った組み合わせと属性区分数を記録．
3.累乗近似式の作成に必要なサンプル数 *z* 回の k-匿名性を計測する．
for (i_1 =1, i_1 <= *z* , i_1 ++) {
query1=Select count(*) as *Count*, Array[i_1] *as Att* from *T* group by Array[i_n] ; // *query1 の結果を table R に記録.*
for (i_2 = 1, i_2 <= *z*, i_2 ++) {
query2=Select min(*Count*) as *k_val*, count(*) as *k_cnt* from *R* where *Att* = Array[i_2] ; // *query2 の結果を table C に記録.*
} end for
} end for
4. 累乗近似値 PowApp_k を求める
query3 = update *C* set *k_calc* = LN(*k_val* – AVG*(k_val)*;
query4 = update *C* set *c_calc* = LN*(k_cnt*– AVG*(k_cnt))*;
query5=select EXP{sum(k_calc*c_calc)/sum(k_calc ^ 2)} as A,
AVG(c_calc) – (***k*** * AVG(*k_calc*)) as B from *table C* ;
PowApp_k = (k-1)^(1/ B) / A^(1/ B) ; // k 値の予測値
5. 全組み合わせ中から最も予測値に近い組み合わせを取得．
for (i_3 = 0, i_3 <=*N*, i_3 ++) {
If (abs(Count(Array[i_3])***Count**(Array[i_{3+m}])−**PowApp_k)) <** *min 値* **)**
{ Array[*x*]と Array[*y*]の組み合わせが最も予測値に近いと判明 }
} **end for**
6. 最も予測値に近い組み合わせにおける k-匿名性(min(*Count*))を計測．
7. 求める k-匿名性を満たさない場合，ボトムアップでより粗い情報を探索する．満たす場合は，処理を終了しても良いが，トップダウンのアルゴリズムを用いて，より詳細な情報を検証することもできる．
If (min(*Count*) < *k*){
Use Bottom Up Algorithm } else {
GODval1 = *x*, **GODval2 =** *y*, **GODk=** min(*Count*)
Use Top Down Algorithm } }
Output: 最も情報量が多く k-匿名性を満たす Globally Optimal Dataset を含む Table *C*[表 8]

1. 属性とその属性に適用する一般化階層について定義を行う．表 7 を用いた場合，A → D_1{男, 女}，B_1 → D_2{10 代, 20 代..}，B_2 → D_3{10-14 歳, 15-19 歳...} と定義し，一般化階層を数値で取得できるよう変換する．

2. 属性の全組合せを作成し，属性区分数を求め，table C に記録する（例：[D_1：2 区分] [D_1, D_2：4 区分] [D_1, D_3：8 区分]...）．また，表 7 の場合 B ∈ D_2, D_3，C ∈ D_4, D_5 であるため，[D_2, D_3] [D_4, D_5] を含む候補は除外する．

3. 精度の高い累乗近似式を取得するために必要な k-匿名性と属性区分数を既知の x，y と設定する．

4-5. 累乗近似式により予測値を作成し，予測した属性区分数に最も近い属性組合せを求める．サンプルコードでは絶対値で差分の少ない値を求めているが，予測値よりも大きい場合に排除する手法もある．

6-7. 予測による組合せによる k-匿名性を取得し，求める k-匿名性を実現した場合は，Top Down Algorithm によって，より詳細な値を探索し，実現しない場合は Bottom Up Algorithm によって，より抽象化された値を探索する．

ここで出力された table C（**表 8**）は，既知の x，y に加え，本アルゴリズムの処理過程で得られた新たな属性区分数と k 値が記録されていき，値がより正確となる．そのため，もし再度匿名化処理を行う場合は 1〜3 までの手順を省略し，直接累乗近似式を作成することができる．

表 8 table C のサンプル
Table 8 Sample of table C.

No.	Att	k_val	k_cnt	k_calc	c_calc
	属性組合せ	k 値	属性区分数	k 平均値	属性数平均値
1	(D_1, D_2)	$k\_val(d1_n, d2_m)$	$Count(Array[D_1 * D_2])$	*k_calc*	*C_calc*
2	(D_1, D_3)	$k\_val(d1_n, d3_m)$	$Count(Array[D_1 * D_2])$	*k_calc*	*C_calc*
:	:	:	:	:	:
GOD	(D_x, D_y)	$k\_val(dx_n, dy_m)$	$Count(Array[D_x * D_y])$	-	-

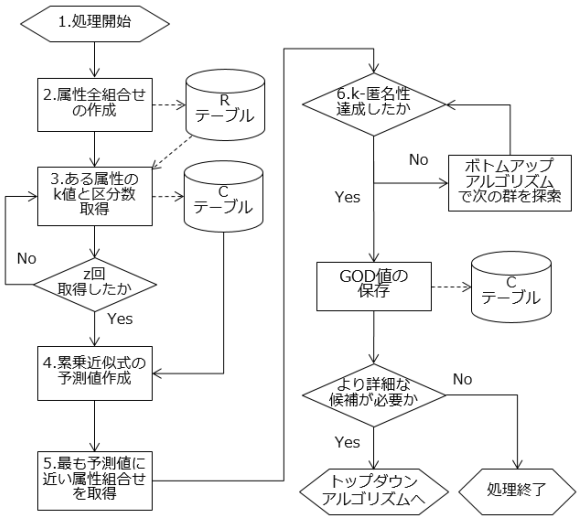

図 12 提案アルゴリズム（PAK）のフローチャート
Fig. 12 Flowchart of proposed algorithm (PAK).

本方式（Power Approximation k-anonymity method：PAK）は，直接的な匿名化処理ではなく，データの特徴に合わせた匿名化処理を選択することで処理の削減を図る方式のため，従来アルゴリズムの単独処理と比較した処理削減率で効果を計測する．

処理削減率の比較対象としてはトップダウン型の Incognito アルゴリズムと，ボトムアップ型の OLA アルゴリズムを用いる．また，PAK 方式に関しては，予測誤差発生後の修正としてトップダウン型の匿名化処理検証を行った計算回数を加えている．

処理削減率を検証する実験は，5 章で用いた国勢調査，および N 社サービス群において人数の多い上位 6 サービスに対して実施した．それぞれのデータ量，および予測式と実測値との相関係数を**表 9** に示す．総処理回数は，匿名状態を検証するべき属性数の合計（総数：3185）である．

図 13 は，k ≧ 50 における匿名化処理にかかった処理量の比較である．予測値に基づいた PAK 方式は OLA 方式と比較して平均で 3.5%，Incognito 方式と比較して平均で 12.5%の処理量で匿名化処理結果を出力した．

図 14 は，k ≧ 2 における処理量比較である．k 値を小

表 9 対象サービスの人数と予測式
Table 9 Property of service groups.

	国勢調査	A	B	C	D	E	F
総人数	127794	350527	215122	208675	190105	140826	134721
予測式 α 値	114659	50316	40048	19664	16326	13618	6641
予測式 β 値	−1.414	−1.776	−1.852	−1.735	−1.695	−1.722	−1.642
予測式と実測値の相関係数	0.9974	0.9901	0.9911	0.9887	0.9853	0.9880	0.9805

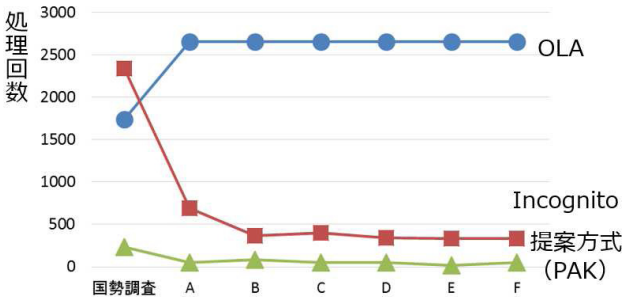

図 13 k ≧ 50 における匿名化処理量比較
Fig. 13 Throughput comparison at k ≧ 50.

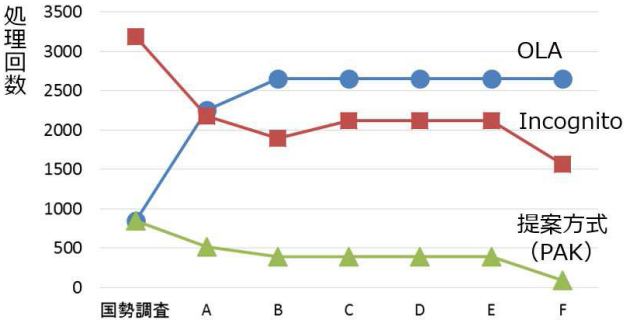

図 14 k ≧ 2 における匿名化処理量比較
Fig. 14 Throughput comparison at k ≧ 2.

さく設定した場合，OLA 方式と比較して平均で 26.6%，Incognito 方式と比較して平均で 19.1%の処理量となった．特に，国勢調査において，OLA 方式と比較すると処理量が同じとなっている．これは最も詳細な群においても k ≧ 2 が達成できたため，ボトムアップ形式における最良ケースで探索が終了したためである．PAK 方式も，最も詳細な組合せでの匿名化が可能であるとの予測ができたため，結果として処理回数が同じとなった状態を示している．**図 15** にてその概要を示す．

図 16 は，図 14 で OLA 方式と同一の値であった国勢調査を除き，人数の多い順から 50 サービスにおける PAK 方式の処理削減量と相関係数の関係性を調査した結果である．予測値と実測値の相関係数が低い群の方が，OLA 形式と比較した相対的な処理量が少ないことが分かる．

この現象は，一般に予測値と実測値の相関係数が低い群は，属性値の分散が大きく，k 値が低くなる傾向が強いために発生している．

ボトムアップ型の OLA 形式は，k 値が低い場合には，全探索を行うため，図 7 および**図 17** における最悪ケース

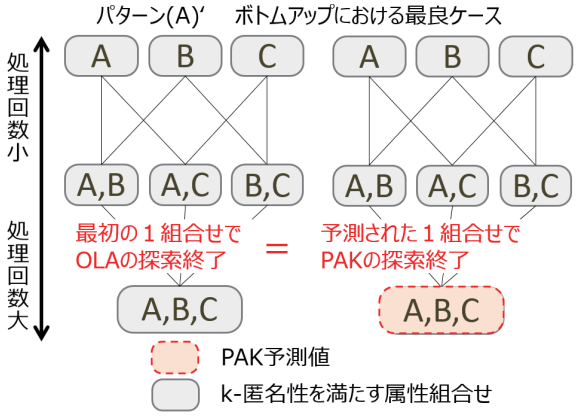

図 15 ボトムアップの最良ケースと PAK の比較
Fig. 15 Best case on Bottom-up and PAK.

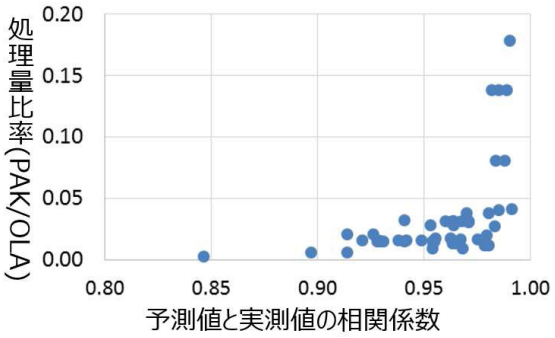

図 16 相関係数と処理量比率の関係性グラフ
Fig. 16 Processing ratio and the correlation coefficient.

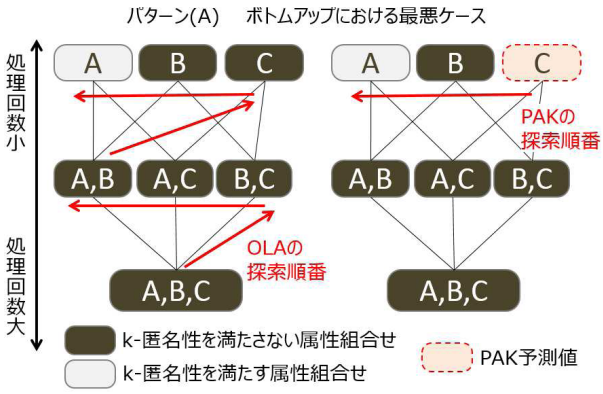

図 17 ボトムアップの最悪ケースと PAK の比較
Fig. 17 Worst case on Bottom-up and PAK.

であるパターン（A）に近い結果となる．それと比較して PAK は相関係数が低く，予測値による匿名化処理が実現できない場合でも，探索回数が少なくて済み，相対的に OLA と比べて探索効率が向上した．

同様の現象は，図 14 で，OLA が最も効率的に処理できる国勢調査の際に Incognito の効率が最も下がる結果となっていることから，トップダウン型アルゴリズムに関しても同様の現象によって処理効率の変化が発生している．

だが，PAK 方式は予測した地点から処理を行っているため，他のアルゴリズムより増減の幅は小さい．

これらの結果により，予測式を用いるPAK方式は，既存アルゴリズムと比して，多様な分散を持つ実データを用いた匿名化処理において，安定して処理回数を減少させたといえる．

7. まとめ

本稿の内容をまとめる．

1) 累乗近似型のk-匿名性予測式について，国勢調査や実データを用いて検証したところ，5,000人以上の群において重相関係数が0.99以上と，高い値が得られた．だが，全体平均で絶対誤差比率6.52であり，直接的なk値予測に利用するのは難しい．
2) k値を得られる属性組合せの情報を，匿名化処理の開始地点として利用し，最適な匿名化アルゴリズムを選択する，PAK方式を提案した．
3) PAK方式は$k \geqq 50$のとき，平均でOLA方式と比較して3.5%，Incognito方式と比較して12.5%の処理量で匿名化処理結果を出力した．
4) PAK方式は，対象データの分布が正規分布で，かつ各属性どうしが独立している場合の予測に用いることを想定していたが，実データでの検証によって，多様な分散を持つ情報に対しても，他のアルゴリズムと比べ，安定して処理回数を削減できることが判明した．

本アルゴリズムによって，個人情報を含むデータベースに対して，少ない負荷で匿名化処理を行うことが可能となる．また，匿名化処理の結果データが，属性区分数とk値として蓄積されていくことで，オーダーメード型匿名化処理を行う際に，データ利用者の求める属性値によって，k-匿名化処理が可能であるかを予測する精度も向上する．

これによりデータ提供者とデータ利用者の双方の処理が効率化し，匿名化データの流通を促進することが期待できる．

参考文献

[1] Sweeney, L.: k-anonymity: A model for protecting privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol.10, No.5, pp.557–570 (2002).
[2] 本多克宏：個人情報のクラスタリングによる匿名化と安心・安全な推薦システム（特集安全社会における情報科学の役割），ケミカルエンジニヤリング，Vol.58, No.3, pp.188–192 (2013).
[3] Meyerson, A. and Williams, R.: On the complexity of optimal k-anonymity, *Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp.223–228 (2004).
[4] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity, *Proc. 2005 ACM SIGMOD International Conference on Management of Data*, pp.49–60 (2005).
[5] 松崎和賢：データ匿名化の現状に関する一考察—医療・統計分野を中心とした国内外の動向，ERATO湊離散構造処理系プロジェクトセミナー (2011).
[6] 亀本信康，齋藤 敦：統計データの二次的利用に関する統計センターの取組状況，2013年度統計関連学会連合大会 (2013).
[7] 日本情報経済社会推進協会（JIPDEC）：パーソナル情報の利用のための調査研究報告書 (2011).
[8] Hundepool, A., Willenborg, L. and Statistics Netherlands: m-and t-ARGUS: Software for Statistical Disclosure Control, *Record Linkage Techniques—1997 Proc. an International Workshop and Exposition*, pp.142–149 (Mar. 1997).
[9] 経済産業省，（株）日立コンサルティング：「行動情報活用型クラウドサービス振興のためのデータ匿名化プラットフォーム技術開発事業」事業報告書 (2012).
[10] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R.: Mondrian multidimensional k-anonymity, *Proc. 22nd International Conference on Data Engineering, ICDE'06*, p.25 (2006).
[11] Aghdam, M.R.S. and Sonehara, N.: EFFICIENT LOCAL RECODING ANONYMIZATION FOR DATASETS WITHOUT ATTRIBUTE HIERARCHICAL STRUCTURE, *The 2nd International Conference on Cyber Security, Cyber Peacefare and Digital Forensic* (*CyberSec2013*), pp.130–140 (2013).
[12] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B. and Fu, A.W.: Utility-based anonymization using local recoding, *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785–790 (2006).
[13] 村本俊祐，上土井陽子，若林真一：k-匿名性を利用したデータ一般化によるプライバシー保護，データ工学ワークショップDEWS (2007).
[14] El Emam, K., Dankar, F.K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J., Walker, M., Chowdhury, S., Vaillancourt, R., et al.: A globally optimal k-anonymity method for the deidentification of health data, *Journal of the American Medical Informatics Association*, Vol.16, No.5, pp.670–682 (2009).
[15] 小栗秀暢，曽根原登：実サービスのデータを用いたk-匿名状態の推移調査と，合理的な匿名状態評価指標の検討，研究報告コンピュータセキュリティ（CSEC），Vol.2014, No.4, pp.1–8 (2014).

小栗 秀暢 （学生会員）

1997年早稲田大学第二文学部卒業．同年タイトー株式会社入社．2007年よりニフティ株式会社にてデータ分析，プライバシ保護技術に関する研究開発業務に従事．現在，総合研究大学院大学複合科学研究科情報学専攻に在学中．

曽根原 登

1978 年信州大学大学院修士課程修了．同年日本電信電話公社（現，NTT）入社．以来，ファクシミリ，画像処理，神経回路網システム，コンテンツ ID，コンテンツ流通システム等の研究開発に従事．1999 年東京工業大学客員教授．2004 年国立情報学研究所（NII）教授，総合研究大学院大学教授兼務．2006 年より NII 情報社会相関研究系研究主幹．博士（工学）．

松井くにお （正会員）

1980 年静岡大学工学部情報工学科卒業．同年(株)富士通研究所入社．2003 年東京工業大学大学院情報理工学研究科後期課程修了．自然言語処理，情報検索，ナレッジマネジメントの研究開発に従事．1999 年富士通（中国）研究開発中心を兼務．2007 年 Fujitsu Laboratories of America．2009 年よりニフティ株式会社．博士（工学）．

モハマド ラスール サラフィ アグダム

2005 年マルチメディア大学（マレーシア）電子工学科卒業．現在，総合研究大学院大学複合科学研究科情報学専攻博士課程に在学中．プライバシー保護データマイニング，匿名化処理技術およびアルゴリズム，ワイヤレスセンサーネットワークの研究に従事．