氏　　名	小栗 秀暢
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1881 号
学位授与の日付	平成28年9月28日
学位授与の要件	複合科学研究科 情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	パーソナルデータ流通における保護と利活用を均衡させる匿名化処理技術の研究
論 文 審 査 委 員	主　査　教授　曽根原 登 教授　越前 功 教授　神門 典子 教授　計 宇生 特任教授　山田 茂樹 国立情報学研究所 教授　小舘 亮之 津田塾大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨

Summary of thesis contents

パーソナルデータ流通における保護と利活用を均衡させる匿名化処理技術の研究

個人の属性や行動に関連する情報である「パーソナルデータ」を利用したサービスが増加している．一方で，コンピューティング能力の進歩により，パーソナルデータから個人が特定，または識別されるリスクが高まっている．それらのリスクを低減した匿名化データの流通によって，有用性の高いデータの安全な利活用が期待されている．

匿名化データの安全な利活用に向け，データ利用者の分析目的に即したデータの要求を可能とする，オーダーメイド型の匿名化処理システムが必要とされている，しかし，匿名化処理は処理コストが高く，また，匿名化データの安全性と有用性の両立が難しいという課題がある．

本研究は，パーソナルデータに対して，プライバシーの保護と利活用を両立する匿名化処理システムを提案するものである．特に，安全性基準を満たす属性値の組み合わせを効率的に探索するアルゴリズムを用いて，データ保持者とデータ利用者間のデータ授受におけるシステム全体の効率化を図り，パーソナルデータの利活用を可能とするプラットフォームの実現を目的とする．

パーソナルデータを匿名化処理してデータ利用者に提供する際に発生するリスクの安全性の指標として，属性値を抽象化，統合化，削除等を行うことで，個人が再識別されるリスクを 1/k(k>1)以下にする「k-匿名性」が知られている．k-匿名化処理によって，あるユーザを一意に識別できる場合に,個人の属性情報が知られてしまうレコード結合(Record linkage)を防止することができる．

しかし，情報量の損失が最小の匿名化データを生成する処理は NP 困難であり，処理コストが大きいという課題がある．加えて，対象とするパーソナルデータの属性値の分布の傾向や，達成するべき k-匿名性によって処理量が変化するため，効率的なアルゴリズムを選択することが困難である．

本研究では，これらの課題解決に向けたアプローチとして，実データの分布を参考に作成された擬似データを用いて，k-匿名性の推移に関する検証を行い，クラスタ化による k-匿名性の減少傾向を予測する近似式を提案した．また，その予測式を用いて，安全性の基準を満たす属性組み合わせを導く匿名化処理アルゴリズムを提案し，匿名化データの授受に関わるプロセス全体の効率化を実現する手法を検討した．具体的には，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標準正規分布に近似する性質から，累乗近似型の予測式を提案した．

予測値を実際の k -匿名性と比較し，予測誤差を検証した結果，重相関係数 0.9 以上の値で実測値と近似したが，予測誤差が大きいことから，予測値を直接用いた処理の効率化は困難であることが確認できた．そこで，提案した予測式を用いて，匿名化処理が達成可能な属性の組み合わせを導き,匿名化状態の検証回数を削減するアルゴリズムを提案した．

従来の匿名化処理アルゴリズムは詳細な属性の組み合わせから匿名化処理を試行するボトムアップ方式と，抽象的な群から匿名化処理を試行するトップダウン方式に分類される．提案するアルゴリズムは，予測地点における匿名性を調査した結果が，求める k-匿名性を満たす場合は，その地点からトップダウンの匿名化処理を選択し，満たさない場合は

(別紙様式 2)
(Separate Form 2)

ボトムアップ型の匿名化処理を選択することで，匿名化状態の検証回数を削減する．

本方式を実データの分布を参考に生成した擬似データに対して適用し，従来のアルゴリズムと比較した処理削減効果を検証した．比較対象とするアルゴリズムは，属性値の削除を行わず，値単位での集合化を行わない中で最も効率の高い OLA 方式と Incognito 方式を選択した．実験の結果，k≧50 の時，ボトムアップ型の OLA 方式と比較して 3.5%，またトップダウン型の Incognito 方式と比較して 12.5%の処理量で匿名化処理が行えることを検証した．

本予測式を用いた匿名化処理アルゴリズムを活用することで，データ作成者とデータ利用者による，安全性と有用性の折衝作業を含む，全体のプロセスを効率化することが可能となる．そこで，匿名化データの利用者に対して，パーソナルデータと属性区分数に関する予測式を提供することで，利用者が求める属性区分によって匿名化処理が達成できるかを，事前に検証出来るプラットフォームを実現可能とした．これによって，パーソナルデータの属性値の分布に関する情報公開を抑制し，かつ k-匿名性を検証する回数を削減することができることを検証した．

本研究の成果は，実サービスの分布を反映した疑似パーソナルデータを生成し，単純な属性値によるクラスタリングを行った場合の k-匿名性の推移を調査し，相関係数の高い予測式を提案したことにある．加えて，その予測式を用いて得られた属性の組み合わせ地点から匿名化状態の検証処理を開始し，その結果に応じてボトムアップ，またはトップダウンアルゴリズムを選択する処理方式を提案し，実データの分布に即した擬似データ群に適用した結果，匿名化処理が効率化できることを検証したことにある．また，本予測式とアルゴリズムを用いることで，元情報に含まれるデータの特徴を公開せずに，データ利用者の求める一般化階層を事前に評価するプラットフォームの社会実装を可能とした．

(別紙様式 3)
(Separate Form 3)

博士論文の審査結果の要旨

Summary of the results of the doctoral thesis screening

パーソナルデータ流通における保護と利活用を均衡させる匿名化処理技術の研究

本博士論文は，パーソナルデータに対して，プライバシーの保護と利活用を両立する匿名化処理システムを提案するものである．特に，安全性基準を満たす属性値の組み合わせを効率的に探索するアルゴリズムを用いて，データ保持者とデータ利用者間のデータ授受におけるシステム全体の効率化を図り，パーソナルデータの利活用を可能とするプラットフォームの実現を目的とする．

本論文は 7 章から構成され，第 1 章で研究の背景について述べ，第 2 章にて従来の匿名化処理方法を分析している．具体的には，パーソナルデータに対するリスクの安全性の指標として，属性値を抽象化，統合化，削除等を行うことで，個人が再識別されるリスクを 1/k(k>1)以下にする「k-匿名性」の安全性と有用性に関わる性質を分析している．

第 3 章及び第 4 章では，実データの分布を参考に作成された擬似データを用いて，k-匿名性の推移に関する分析を行っている．その結果，属性値の組み合わせ数から k-匿名性を予測する近似式を提案し,重相関係数 0.9 以上の値で k-匿名性が予測できることを示した．

第 5 章において，提案した k-匿名性の予測式を用いて，匿名化処理が達成可能な属性の組み合わせを予測し，匿名化状態の検証を効率化するアルゴリズムを提案した．従来の匿名化処理アルゴリズムは詳細な属性の組み合わせから匿名化処理を試行するボトムアップ方式と，抽象的な群から匿名化処理を試行するトップダウン方式に分類される．提案するアルゴリズムは，予測地点における匿名性を調査した結果が，求める k-匿名性を満たす場合は，その地点からトップダウンの匿名化処理を選択し，満たさない場合はボトムアップ型の匿名化処理を選択することで，匿名化状態の検証回数を削減することができる．本方式を実データの分布を参考に生成した擬似データセットに適用し，従来のアルゴリズムと比較した処理削減効果を検証している．比較対象とするアルゴリズムは，属性値の削除を行わず，値単位での集合化を行わない方法の中で，最も効率の高い OLA 方式と Incognito 方式と比較した．実験の結果，k≧50 の時，ボトムアップ型の OLA 方式と比較して 3.5%，またトップダウン型の Incognito 方式と比較して 12.5%の処理量で匿名化処理が行えることを検証した．

第 6 章において，匿名化データの利用者に，パーソナルデータと属性区分数に関する予測式を提供することで，利用者が求める属性区分によって匿名化処理が達成できるかを事前に検証出来るプラットフォームを提案している．予測式を用いた匿名化処理アルゴリズムを活用することで，データ作成者とデータ利用者による，安全性と有用性の折衝作業を含む，全体のプロセスを効率化することが可能となることを検証した．第 7 章にて本論文の結論をまとめ，今後の展望と研究課題を提示した．

本研究の成果は，実サービスの分布を反映した疑似パーソナルデータを生成し，単純な属性値によるクラスタリングを行った場合の k-匿名性の推移を調査し，相関係数の高い予測式を提案したことにある．加えて，その予測式を用いて得られた属性の組み合わせ地点

(別紙様式 3)
(Separate Form 3)

から匿名化状態の検証処理を開始し，その結果に応じてボトムアップ，またはトップダウンアルゴリズムを選択する処理方式を提案し，実データの分布に即した擬似データ群に適用した結果，匿名化処理が効率化できることを検証したことにある．

また，本予測式とアルゴリズムを用いることで，元情報に含まれるデータの特徴を公開せずに，データ利用者の求める一般化階層を事前に評価するプラットフォームの社会実装を可能とし，新規性と高い有用性が期待される．

なお，研究成果として，出願者は主著で査読付きジャーナル論文 1 篇，査読付き国際・国内会議論文 3 篇を発表したほか，学位論文に関連する特許登録が 2 件されている．したがって，学位論文に充分なレベルであることを判断した．