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

平成　２９　年　　６　月　　５　日現在

機関番号：　１２６１２
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６８７０２０１

研究課題名（和文）ユビキタスコンピューティング環境におけるユーザ情報匿名化手法の研究

研究課題名（英文）Anonymization of Personal Data in Ubiquitous Computing

研究代表者
清　雄一（Sei, Yuichi）

電気通信大学・大学院情報理工学研究科・助教

研究者番号：２０７００１５７

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：年齢，性別，病歴，位置情報等，個人に関連する情報を活用するビジネスが拡大している．データ連携によって新たな知識創出が期待されるが，プライバシの問題が懸念される．情報の有用性をできるだけ損なわずに匿名化を行う研究が，病院等の閉じた環境で管理されている少数の正確な項目から成るデータベースに対しては，国内外で盛んに行われている．
今後はユビキタスコンピューティングの進展により，1．多数の個人属性を含むデータベースの収集と共有，2．誤差を含む個人属性値の収集と共有，3．個人所有の端末が観測した個人属性値の収集と共有，が行われるようになると考えられ，この収集と共有を可能にする技術を提案する．

研究成果の概要（英文）：In recent years, numerous organizations have begun to provide services that collect large amounts of personal information. This personal information can be shared with other organizations so that they can subsequently create new services.
Owing to the development of ubiquitous computing and sensing technologies, numerous research methods for crowdsensing have been proposed to collect and analyze sensed environmental information from mobile phone users.
Our contributions are as follows: (1) we propose a simple but effective general anonymization algorithm for large databases, (2) we propose a novel privacy metric and a utility metric that can treat the location error and propose an efficient anonymization algorithm for the proposed metrics, and (3) we propose S2M and S2Mb, both of which can make a better trade-off between privacy and utility in crowdsensing.

研究分野：　プライバシ保護

キーワード：　プライバシ　データマイニング　IoT

１．研究開始当初の背景

年齢，性別，病歴，位置情報，周辺の音声等，個人に関連する情報（以下，「個人属性」と記す）を活用するビジネスが拡大している．現在は，自組織で収集した個人属性を自組織のサービス提供に利用するに留まっているが，自組織で保有する個人属性のみではデータ分析に限界がある．今後は，より詳細な分析を行うために，複数組織が持つ多種多様な個人属性を統合して分析を行い，その分析結果から新たなサービスを創出することが期待される．

このようにデータ連携によって新たな知識創出が期待されるが，プライバシの問題が懸念される．顧客から収集したデータは，個人を直接特定できる情報を削除し，匿名化した上で他事業者に提供されることになるが，たとえ個人を直接特定できる情報を削除したとしても，匿名化が不十分であった場合，データ内で個人が特定され，センシティブな個人属性が漏洩する危険がある．日本では，改正個人情報保護法が 2017 年 5 月 30 日に全面施行され，匿名化を行えば本人の同意無く第三者提供が可能となるが，不十分な匿名化によって個人に関する情報が判明してしまうリスクが存在する．実際，匿名化を行ってデータを公開したものの匿名化が不十分であったために，個人属性が高精度に判明してしまい，公開を取りやめた事例は多く存在する．情報の有用性をできるだけ損なわずに匿名化を行う研究が，病院等の閉じた環境で管理されている少数の正確な項目から成るデータベースに対しては，国内外で盛んに行われている．

今後はユビキタスコンピューティングの進展により，1．多数の個人属性を含むデータベースの収集と共有，2．誤差を含む個人属性値の収集と共有，3．個人所有の端末が観測した個人属性値の収集と共有，が行われるようになると考えられ，この収集と共有を可能にする技術が必要となる．

２．研究の目的

本研究で取り組む主要な課題及び目的を以下に挙げる．

(1)個人属性数の増加による匿名化後のデータの有用性低下を低減する匿名化アルゴリズムの提案

これまでは,属性数が数個から十数個程度のデータベースを匿名化する研究が主に行われているのみである．今後は，何らかのサービスを受けるために必要である等，各個人の同意が得られた上で，ユビキタスコンピューティング環境により多数の個人属性が収集されるようになる．収集された個人属性を適切に匿名化することができれば，個人の同意を得ることなく，複数組織間で自由に共有して分析を行うことで，新しいサービス創出につなげることができる．したがって，ユビキタスコンピューティング環境により取得され得る，数十以上の属性を含むデータベースを匿名化する技術が必要である．

(2)誤差を含む属性を起因とする不十分な匿名化リスクを定量的に取り扱う匿名性指標及び匿名化アルゴリズムの提案

ユビキタスコンピューティング環境で得られる個人属性値は，誤差を含み得る．これまでの研究では，誤差の無い正確な個人属性を対象として提案されており，誤差を考慮した匿名化を行う技術が必要である．

(3)個人属性を匿名化しながら収集する際のデータの有用性低下を低減する匿名化アルゴリズムの提案

ユビキタスコンピューティングの進展により，個人所有のスマートフォン等で周囲の環境データをセンシングし，その値を収集する，クラウドセンシングの活用が広まっている（本稿では，個人周辺の環境情報も個人属性とみなす）．しかし，クラウドセンシングを行うためには，各個人のプライバシに配慮する必要がある．より多くの個人に参加してもらうために，個人属性を収集後ではなく，個人所有の端末において匿名化し，匿名化後の個人属性を収集する技術が必要である．

３．研究の方法

(1) 既存研究においては,多数の個人属性を含むデータベースを匿名化すると，有用性の大幅な低下を招いてしまうという問題がある．また，個人属性数の増加により，匿名化を行うための計算量が爆発的に増加してしまう場合もある．個人属性数の増加に対して，有用性の低下を低減する，高速な匿名化アルゴリズムを提案する．

(2) 既存研究は誤差がない前提で匿名化を行っており，誤差がある場合は誤った匿名化を行うリスクがある．このリスクを明らかにするとともに，誤差を定量的に扱う匿名性指標を提案し，この指標に基づく新たな匿名化アルゴリズムを提案する．

(3) 個人属性を匿名化しながら収集することは，1 レコードから構成される多数のデータベースをそれぞれ独立に匿名化して共有することと等しい．多数のレコードからなるデータベースを，その統計的性質をできるだけ失わないように匿名化する場合と比べて，有用性の大幅な低下を招いてしまう．各

個人属性を独立に匿名化を行う場合においても，有用性の低下を低減することのできる，匿名化アルゴリズムを提案する．

また，研究の前提を以下のように整理する．

匿名化を必要以上に行うと，匿名化後のデータから分析できる情報が減少してしまう．一方，匿名化が不十分であると，個人属性が高精度に判明してしまうリスクが高まる．この関係はトレードオフにある．

プライバシ保護データマイニングの研究分野では，データベースに対し，どの程度匿名性が保証されているかを計測する指標が多数提案されている．その中でも，k-匿名性、l-多様性、ε-差分プライバシの 3 つが主要な指標である．

k-匿名性：他データベースと組み合わせて個人を特定し得る個人属性を「準識別子」と呼び，他データベースと組み合わせても個人特定につながらず，かつ，保護すべき個人属性を「センシティブ属性」と呼ぶ．各レコードにおいて，全ての準識別子の値が同じであるレコードが他に k-1 個以上存在している場合，k-匿名性が満たされていると定義される．

l-多様性：全ての準識別子の値が同じであるレコードの各集合において，センシティブ属性値が 1 種類以上存在している場合，l-多様性が満たされていると定義される．

ε-差分プライバシ：オリジナルのデータベースを D とし，D と 1 レコードだけ異なるデータベースを D’とおく．またオリジナルのデータベースとして取り得る空間を E とおく．ランダムメカニズム R が，全ての D,D’∈E について，また，全ての S⊆Range(R)について，

$$\Pr(R(D) \in S) \geq e^{\epsilon}\Pr(R(D') \in S)$$

を満たすとき，R(D)はε-差分プライバシが満たされていると定義される．

一方，有用性を計測する指標は，匿名化前であるオリジナルのデータベースと，匿名化後のデータベースの距離を直接計測するものや，各データベースから得られる解析結果同士の距離を計測するものがある．実用を考慮した場合は後者のものが望ましく，代表的な解析としてクロス集計表を対象とし，有用性を計測する指標として，オリジナルのデータベースから得られるクロス集計表と匿名化後のデータベースから得られるクロス集計表間の，L1 距離，平均二乗誤差，JS-divergence 等が用いられる．

４．研究成果

以下の研究成果が得られた．

(1)提案手法では，個人属性数を a とおき，各個人属性に対して ki-多様性を実現したい場合，概念的には，データベース中の各レコードに対して$\prod_{i=1}^{a}(k_i - 1)$個のダミーレコードを追加する．この手法が各個人属性に対して ki-多様性を実現していることの証明を行ったとともに，a や ki の値の増加に対して計算量の増加を線形に抑えるための匿名化アルゴリズムを提案した．また，多数のダミーレコードが追加されているため，各レコードを独立に見るとほとんど有用な情報を得ることができないため，匿名化されたデータベース全体から，クロス集計表を高精度に構築するアルゴリズムも併せて提案した．

図 1 に，約 240 万レコード，68 属性を含むデータベースに対して匿名化及びクロス集計表構築を行い，その誤差を計測した結果を示す．

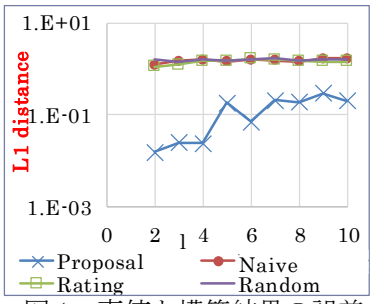

図 1：真値と構築結果の誤差

図 1 より，提案手法の誤差が既存手法（Rating）やベースライン手法（Naïve 及び Random）よりも小さいことが分かる．

(2)誤差を含む個人属性に対してその誤差を確率的に取り扱うことのできる，(k,w)-匿名性という匿名性指標を提案した．これは，準識別子の値が同一である個人が k 人以上存在する確率が w 以上であることを保証する指標である．また，誤差を考慮することのできる有用性指標も新たに提案した．研究では特に位置情報に焦点を当て，各個人が存在し得る空間を確率的に考慮した上で(k,w)-匿名性を満たすことのできる匿名化を行う手法を提案した．厳密解を求めることは NP 困難であることを示し，高速に計算可能なアルゴリズムを提案した．

(3) 提案手法では，あるパラメータ si 及び pi を用意し，各個人の a 個の個人属性に対して$\prod_{i=1}^{a}(s_i - 1)$個のダミーの個人属性群を追加する．ここで，ダミーの個人属性値全てについて，確率 pi でオリジナルの個人属性値をそのまま維持し，確率(1-pi)でオリジナルの個人属性値を一切含めない．si 及び pi は，ε-差分プライバシを満たし，かつ，匿名化後のデータの有用性を理論上最大化

するよう算出される．これを実現するため，匿名化後のデータの有用性の期待値を算出するアルゴリズムも併せて提案している．提案手法は，同時に，$\prod_{i=1}^{a}(s_i-1)$-匿名性，また，$\prod_{i=1}^{a}(s_i-1)$-多様性を満たしている．

図 2 に，10 万人を想定したデータに対して匿名化及びクロス集計表の構築を行い，その誤差を計測した結果を示す．

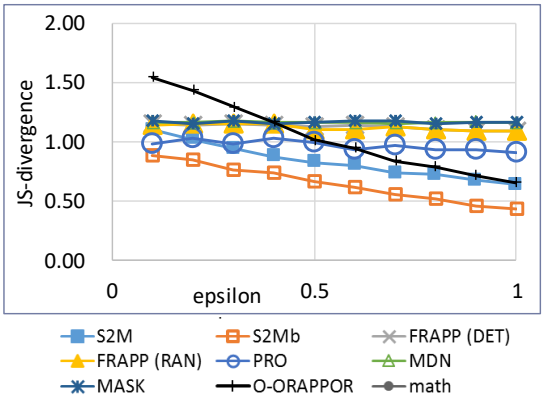

図 2: 真値と構築結果の誤差

図 2 より，提案手法（S2Mb）の誤差が，各種既存手法の誤差よりも小さいことが確認できる．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 20 件）
①Yuichi Sei and Akihiko Ohsuga: Location Anonymization with Considering Errors and Existence Probability, IEEE Transactions on System, Man, and Cybernetics: Systems (in press), 査読有, doi: 10.1109/TSMC.2016.2564928
② Yuichi Sei and Akihiko Ohsuga: Differential Private Data Collection and Analysis Based on Randomized Multiple Dummies for Untrusted Mobile Crowdsensing, IEEE Transactions on Information Forensics and Security, Vol.12, No.4, pp.926-939, 2017, 査読有, doi: 10.1109/TIFS.2016.2632069
③清雄一，竹之内隆夫，大須賀昭彦：クラウド上の安全で高速なキーワード検索アルゴリズムの提案，情報処理学会論文誌，Vol.56, No.10, pp.1977-1987, 2015, 査読有［情報処理学会論文誌ジャーナル特選論文，2016 年度情報処理学会論文賞］, http://id.nii.ac.jp/1001/00145517/
④清雄一，稲葉緑，大須賀昭彦：安心できるプライバシ指標の調査，情報処理学会論文誌，Vol.56, No.12, pp.2230-2243, 2015, 査読有, http://id.nii.ac.jp/1001/00146607/
⑤Yuichi Sei and Akihiko Ohsuga: Malicious Node Detection in Mobile Wireless Sensor Networks, Journal of Information Processing (JIP), Vol.23, No.4, pp.476-487, 2015, 査読有, doi: 10.2197/ipsjjip.23.476

〔学会発表〕（計 21 件）
①Yuichi Sei and Akihiko Ohsuga: Privacy Preservation for Participatory Sensing Applications, 30th IEEE International Conference on Advanced Information Networking and Applications (AINA), pp.653-660, 2016.3.24, Crans-Montana (Switzerland), 査読有, doi: 10.1109/AINA.2016.19
②Yuichi Sei, Takao Takenouchi, Akihiko Ohsuga: (l1, ..., lq)-diversity for Anonymizing Sensitive Quasi-Identifiers, 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp.596-603, 2015.8.20, Helsinki (Finland), 査読有, doi: 10.1109/Trustcom.2015.424
③Yuichi Sei and Akihiko Ohsuga: Locating Malicious Agents in Mobile Wireless Sensor Networks, 17th Principles and Practice of Multi-Agent Systems (PRIMA), pp.206-221, 2014.12.3, Queensland (Australia), 査読有, doi: 10.1007/978-3-319-13191-7_17
④Yuichi Sei, Akihiko Ohsuga: Randomized Addition of Sensitive Attributes for l-diversity, 11th International Conference on Security and Cryptography (SECRYPT), 2014.8.29, Vienna (Austria), 査読有, pp.350-360, 2014, doi: 10.5220/0005058203500360

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 1 件）

名称：プライバシ保護データ提供システム及びプライバシ保護データ提供方法
発明者：清雄一，奥村拓史，大須賀昭彦
権利者：同上
種類：特許
番号：特願 2016-239460
出願年月日：2016 年 12 月 9 日
国内外の別：国内

６．研究組織
(1)研究代表者
清　雄一（SEI, Yuichi）
電気通信大学・大学院情報理工学研究科・助教
研究者番号：20700157