応専

k-匿名化技術と実用化に向けた取り組み

竹之内隆夫（日本電気（株）クラウドシステム研究所）

パーソナルデータの二次利用における k-匿名化への期待

医療機関や通信事業者などさまざまな機関では，サービス提供のために個人に関する情報（パーソナルデータ）を収集している（本稿では，個人情報保護法が定める「個人情報」に限らず，個人に関する情報を「パーソナルデータ」と呼ぶ）．通常，これらのパーソナルデータは，収集した機関内のみで利用（一次利用）されることが多いが，今後は，より良いサービス提供や社会生活のために，収集した機関以外のほかの機関に提供し利用（二次利用）されることが期待されている．たとえば，医療機関が診察した患者の診療情報を医学研究機関で二次利用することで，薬の副作用分析や医療費分析を行い，医療の質向上や効率化を行うことが期待されている[1)]．また，通信事業者が収集した個人の位置情報を二次利用することで，災害時の避難対策などに活用することが期待されている．

しかし，パーソナルデータをほかの機関に提供することは，個人のプライバシーを侵害してしまう恐れがある．たとえば，米国のビデオストリーミングサービス会社のNetflix社では，レコメンドのアルゴリズム開発のコンテスト「Netflix Prize」を開催し，約50万人の顧客の視聴履歴と視聴した映画の評価情報を個人特定が困難になるように加工して公開した．しかし，個人特定ができないはずであった視聴履歴は，ほかのサイトで公開されている映画批評のコメント内容と比較することで，個人特定ができてしまうことが指摘された．この問題は，訴訟にまで発展し，コンテストの続編は中止となった．

そこでパーソナルデータをほかの機関に提供する際のプライバシーを保護するために，パーソナルデータに含まれる個人に紐付く情報を加工し，個人を1/k以下に特定されることを防ぐというk-匿名化技術が注目されている．k-匿名化されたデータは，個人を特定した分析には利用できないが，個人特定が不要な統計的な分析には利用できる．しかし，データは加工されるため，分析の精度は低下する．つまり，データの有用性は低下する．匿名化技術は，いかにデータの加工を抑え，データの有用性を保ちつつも，個人特定ができないような安全なデータに加工するかが重要となる．そして，プライバシーの保護とデータの有用性の維持を両立させることを目指している．

本稿では，パーソナルデータを収集した機関以外へ提供する際の個人特定の問題について説明し，k-匿名化技術の概要を説明する．そして，k-匿名化技術の実用化に向けた取り組みの例として，医療情報や位置情報の匿名化技術の研究開発の例を紹介する．

個人特定の問題とプライバシー保護の方法

k-匿名化では，パーソナルデータは以下のような属性で構成されると整理されている．

- 識別子：単独で個人を識別できる属性（例：氏名，電話番号，メールアドレス）
- 準識別子：組み合わせて個人を識別できる属性（例：年齢，性別，生年月日）
- センシティブ属性：他人に知られたくない属性（例：病名，滞在場所）
- その他の属性：上記以外の属性

表-1（a）に，パーソナルデータをテーブル形式で表現した例を示す．この例では，各レコードが個人のパーソナルデータに対応し，各カラムが属性に

(a) 識別子を削除したテーブル

No.	ZIPコード	年齢	職業	病状
1	13068	28	ダンサー	心臓病
2	13068	29	技術者	心臓病
3	13053	21	法律家	感染症
4	13053	23	技術者	感染症
5	14853	31	技術者	風邪
6	14853	37	作家	風邪
7	14850	36	法律家	がん
8	14850	35	技術者	がん

準識別子　センシティブ情報

(b) k-匿名化したテーブル (k=2)

No.	ZIPコード	年齢	職業	病状
1	13068	28-29	*	心臓病
2	13068	28-29	*	心臓病
3	13053	21-23	*	感染症
4	13053	21-23	*	感染症
5	14853	31-37	*	風邪
6	14853	31-37	*	風邪
7	14850	35-36	*	がん
8	14850	35-36	*	がん

(c) ℓ-多様化したテーブル (ℓ=2)

No.	ZIPコード	年齢	職業	病状
1	130**	21-29	*	心臓病
2	130**	21-29	*	心臓病
3	130**	21-29	*	感染症
4	130**	21-29	*	感染症
5	148**	31-37	*	風邪
6	148**	31-37	*	風邪
7	148**	31-37	*	がん
8	148**	31-37	*	がん

表 -1　匿名化の例（k- 匿名化，ℓ - 多様化）

対応する．また，「ZIP コード」「年齢」「職業」が準識別子，「病状」がセンシティブ属性としている．このテーブルでは，氏名のような識別子が削除されているので，どのレコードが誰のパーソナルデータであるかを特定できないように見える．しかし，このテーブルがある病院の全患者の診療情報であり，このテーブルを受け取った分析者（攻撃者）が「A さんの ZIP コードは 14850 であり，年齢 35 歳，職業が技術者であり，この病院に通院している」ことを前提知識として知っていたとする．すると，このテーブルを受け取った分析者は表 -1(a) の No.8 のレコードが A さんのレコードであることを特定できる．その結果，A さんの病状が「がん」であることを特定できてしまう．この例のように，たとえ識別子を削除したとしても，準識別子によって個人を特定できてしまう可能性があり，その結果センシティブ属性が，知られてしまう恐れがある．たとえば，文献 2）では ZIP コード，性別，生年月日の 3 つの属性の値の組合せから約 87% の米国居住者を 1 名に識別できるとされている．

k- 匿名化では，個人の特定を防ぐために，準識別子を加工する．つまり，「誰の」パーソナルデータであるかを隠すことにより，個人のプライバシーを守るという発想である．

k- 匿名化では，個人のプライバシーを侵害しようとしている攻撃者から，どのようにプライバシーを守るかを以下のように整理している．

- 攻撃モデル：攻撃者がどのようなプライバシー侵害の攻撃を仕掛けてくるか？
- プライバシーモデル：どのような攻撃に対して，どのような情報が漏洩しないことを保証するか？
- 匿名化処理：プライバシーモデルを実現するためにデータをどのように加工するか？

以降で，これらについて，代表的なものをいくつか紹介する．

攻撃モデル	プライバシーモデル
レコード特定 (Record Linkage)	k- 匿名性 (k-anonymity)
属性特定 (Attribute Linkage)	ℓ - 多様性 (ℓ -diversity) t- 近似性 (t-closeness)

表 -2　攻撃モデルとプライバシーモデル

攻撃モデルとプライバシーモデル

代表的な攻撃モデルとプライバシーモデルを**表 -2** にまとめた．レコード特定とは，準識別子を用いてテーブルの中からターゲット（被害者）のレコードを特定するという攻撃である．この攻撃によって，攻撃者にターゲットのセンシティブ属性や準識別子を知られる恐れがある．レコード特定を防ぐためのプライバシーモデルが，k- 匿名性である．k- 匿名性とは，テーブル内の準識別子で識別できるレコードが少なくとも k 個以上あるという性質である ($k > 1$)．k- 匿名化とは k- 匿名性を満たすようにテーブルを加工することである．**表 -1 (b)** は，2- 匿名化した例である．

加工方法の名前	加工内容
切落し（Suppression）	一部の属性またはレコードを削除する
汎化（Generalization）	属性の値をより一般化した値に置き換える
分離（Anatomization）	準識別子とセンシティブ属性とでテーブル分割する
置換（Permutation）	レコード間で属性の値を置き換える
摂動（Perturbation）	属性の値に揺らぎを与える

表 -3　データの加工方法

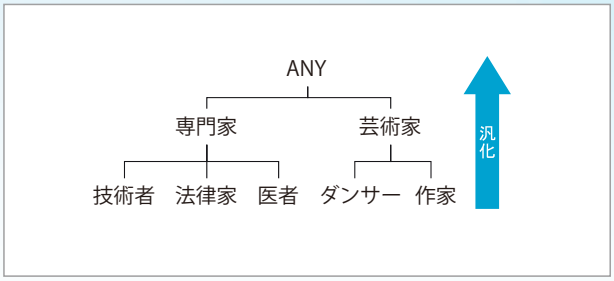

図 -1　汎化ツリーの例

しかし，2- 匿名化した表 -1(b) のテーブルでは，No.7, 8 のレコードは両方とも「がん」である．つまり，k- 匿名化することでレコード特定は防げたとしても，センシティブ属性を特定することができてしまう．このような攻撃を属性特定と呼ぶ．そこで，属性特定を防ぐためのプライバシーモデルとして ℓ-多様性が提案されている．ℓ-多様性とは，k- 匿名性を満たすテーブルにおいて，準識別子で識別できるレコードのセンシティブ属性の値が少なくとも ℓ 種類以上あるという性質である（$k \geqq \ell > 1$）．**表 -1（c）**は，2- 多様化した例である．

しかし，ℓ- 多様化を行ったとしても，準識別子で識別されるレコードにおけるセンシティブ属性の分布が，テーブル全体における分布と大きく異なっていると，テーブル全体における分布から推測できる以上に，センシティブ属性を推測できてしまうため，プライバシーを侵害してしまう恐れがある．たとえば，あるテーブルのテーブル全体における分布が，「がん」のレコード数が全体の 5%，「かぜ」が 95% であったとする．ここで，もし攻撃者がこの分布を知っていた場合，この攻撃者は，このテーブルに含まれる患者は 5% の確率で「がん」であると推測できる．しかし，もし，このテーブルを 2- 多様化した結果，あるターゲットの準識別子で識別されるレコードにおける分布が，「がん」が 50%，「かぜ」が 50% であった場合，この攻撃者は，そのターゲットは 50% の確率で「がん」であると推測できてしまう．

そこで，このような属性の推測にも耐えられるプライバシーモデルとして提案されているのが，t- 近似性である．t- 近似性とは，準識別子で識別されるレコードにおけるセンシティブ属性の分布とテーブル全体におけるセンシティブ属性の分布の差が t 以内であるという性質である．ほかにも，δ- 存在性や m- 不変性などさまざまなプライバシーモデルが提案されている [3]．

どのプライバシーモデルを適用するかや，どの属性を準識別子やセンシティブ属性とするかは，アプリケーションによって異なる．攻撃者やデータの特性に応じて，適切に決定する必要がある．

匿名化処理

匿名化処理は，プライバシーモデルを充足させつつも，可能な限りデータの有用性を向上させることを目的としている．ここでは，匿名性を満たすために，どのようにデータを加工するかについて説明する．代表的なデータの加工方法を**表 -3** にまとめる．

最も簡単な匿名化処理は，切落としである．この処理では，単にレコードや属性を切り落とすだけであるので，たとえば準識別子で識別できるレコード数が k 以下となるレコードを削除すれば，k- 匿名性を満たすテーブルを生成することができる．しかし，削除するレコード数が多くなると，統計的な性質を保たなくなり，匿名化したテーブルを用いて統計的な分析を行うことができなくなってしまう．

そこで，データの加工方法としてよく使われるのが，汎化である．汎化では，**図 -1** に示したような汎化ツリー（一般化の階層）に従って，属性の値を一般化する．汎化方法には，いくつか種類が存在する．**表 -4** に代表的な汎化方法を示す．全領域汎化は，テーブル内の全レコードで汎化レベルを統一する

(a) 元のデータ

No.	…	職業
1	…	法律家
2	…	法律家
3	…	法律家
4	…	技術者
5	…	医者
6	…	作家
7	…	作家

(b) 全領域汎化 (Full-domain generalization)

No.	…	職業
1	…	専門家
2	…	専門家
3	…	専門家
4	…	専門家
5	…	専門家
6	…	芸術家
7	…	芸術家

(c) 部分ツリー汎化 (Subtree generalization)

No.	…	職業
1	…	専門家
2	…	専門家
3	…	専門家
4	…	専門家
5	…	専門家
6	…	作家
7	…	作家

(d) セル汎化 (Cell generalization)

No.	…	職業
1	…	法律家
2	…	法律家
3	…	法律家
4	…	専門家
5	…	専門家
6	…	作家
7	…	作家

表 -4　汎化の例

という汎化方法である．表 -4 (a) に示した元データを全領域汎化したのが表 -4(b) である．この例では，全レコードの値が汎化ツリーにおける専門家や芸術家という汎化レベルに統一されている．これを，より柔軟にした汎化方法が部分ツリー汎化である．この汎化方法では，汎化ツリーのカテゴリごとに汎化レベルを変えることを許容する（表 -4 (c)）．さらにセル汎化では，レコードごとに汎化レベルを変えることを許す（表 -4 (d)）．

汎化方法によっては，データの加工を最小限に抑えた最適な k- 匿名化を実現するには，計算量が膨大になってしまう．たとえば，セル汎化を用いた最適な k- 匿名化は NP 困難であることが証明されている．

そこで，数多くの匿名化のアルゴリズムが研究されている．たとえば汎化を用いた k- 匿名化のアルゴリズムとしては，徐々に汎化レベルを上げていくボトムアップと呼ばれるアプローチや，徐々に汎化レベルを下げていくトップダウンと呼ばれるアプローチのアルゴリズムが提案されている．詳細は，文献 3）などを参照してほしい．

実用化に向けた取り組み

匿名化技術を実用化するためにいくつかの研究開発が進んでいる．カナダの，Privacy Analytics 社では，Privacy Analytics Risk Assessment Tool（PARAT）という匿名化ツールを商用化している．PARAT はボトムアップアプローチの匿名化アルゴリズムを実装しており，主に医療情報を対象としている．PARAT は，匿名化を行うだけでなく，個人特定のリスク評価も行えるツールとなっている．

筆者らの研究グループでは，レセプト（診療報酬明細書）データを匿名化するための研究を行っている．レセプトデータとは，医療機関が医療費の一部を保険者（市町村や健康保険組合等）に請求する際の明細書に記載されている情報のことである．このデータは，患者の疾病や投薬に関する情報が含まれる．患者は複数の病気にかかったり複数の医薬品が処方されたりするため，1 人の患者に対して複数の疾病や医薬品の情報が関連付く．筆者らは，攻撃者が患者の一部の疾病や医薬品の情報を知っている場合を想定し，ある患者について複数の疾病や医薬品が含まれるようなデータを匿名化するためのシステムを構築した．そして，実際のレセプトデータを用いて有用性の評価を行った [4]．評価の結果，特定の医薬品の処方パターンの推移を調べるような分析において，匿名化後のデータを用いた分析結果は元データを用いた分析結果とほぼ一致し，十分な精度を持った分析が可能であることが分かった（**図 -2, 3**, 文献 4）より引用）．また，匿名化されたレセプトデータを病院の医師 8 名に提示して匿名化技術の医学研究への適用可能性についてアンケートを実施した．アンケート結果では，一部の属性が過度に汎化されてしまう場合に元データの持つ統計的な性質（分布など）に大きな影響があるという懸念が指摘された．

位置情報の匿名化技術の研究もいくつか行われている．たとえば情報大航海プロジェクトでは，個人の頻繁に滞留する場所（以降，滞留点と呼ぶ）に対する匿名化の研究とその実証実験が行われた [5]．個

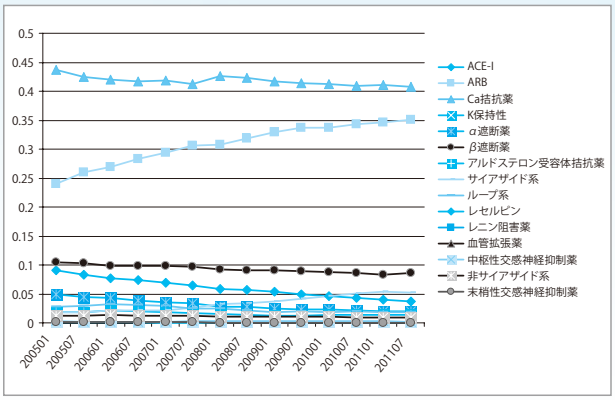

図 -2　元データでの集計結果（著者の許諾を得て，文献 4）から引用）

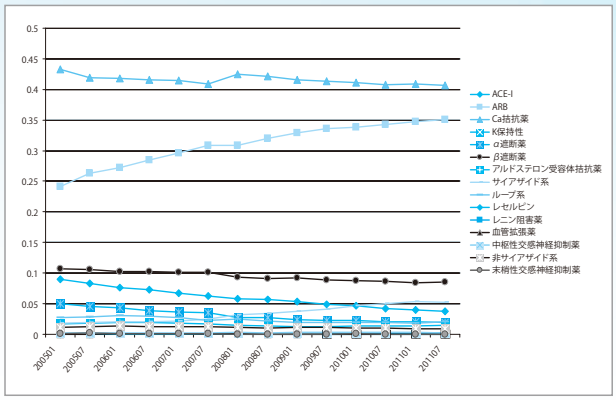

図 -3　匿名化後のデータでの集計結果（著者の許諾を得て，文献 4）から引用）

人の位置情報を継続的に取得すると，自宅や会社やよく行く店や病院等の位置を滞留点として推測することができる．もし攻撃者がある個人の滞留点の一部を知っていたとすると，その個人のほかの滞留点を知ることができてしまう恐れがある．そこで，この研究では滞留点のピンポイントの位置情報をエリア情報に拡大するなどして匿名化している．実証実験では，首都圏ユーザ約 3,000 人の実際の滞留点を匿名化し，サービスに活用できることを実証した．

また，クラウド上で匿名化機能を提供するための国家プロジェクトも行われている [6]．このプロジェクトでは，Hadoop を用いた分散処理で匿名化を実現するための研究などが行われている．

今後の期待

匿名化技術は実用化段階に入っており，実用化に向けた研究が活発化している．今後は，さらなる実案件への適用とパーソナルデータ活用の促進が期待される．

参考文献

1) 内閣府，「日本再生加速プログラム」について（平成 24 年 11 月 30 日閣議決定）．
2) Sweeney, L. : k-anonymity : A Model for Protecting Privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), pp. 555-570 (2002).
3) Fung, B. C. M., Wang, K., Fu, A. W. C. and Yu., P. S. : Privacy-Preserving Data Publishing : Concepts and Techniques CRC Press (2010).
4) 側高，高橋，豊田，竹之内，森，興梠：レセプト匿名化システムの実証と評価，第 32 回医療情報学連合大会（2012）．
5) 宮川，森，岡田，佐治：プライバシ情報の安全な流通と利活用を実現するシステムのアーキテクチャと評価，FIT2011.
6) 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（2013）．

（2013 年 6 月 10 日受付）

竹之内隆夫（正会員）｜ takenouchi@bu.jp.nec.com
2005 年 NEC 入社．博士（工学）．現在 NEC クラウドシステム研究所にて，プライバシー保護技術に関する研究開発に従事．