応専

プライバシーのルールを扱う技術
—制御・検証から説明・理解の支援へ—

石川冬樹（国立情報学研究所）

ルールに基づくプライバシーへのアプローチ

情報を扱う技術が飛躍的に発展し，社会がその影響を受け大きく変化を続ける中，プライバシーに関する議論がますます活発となっている．研究対象となる技術としては，個人情報が特定できないようにする仕組みがよく取り上げられる．典型的には，データの通信や検索，統計処理などに際し，仮想的な識別子を用いたり，ノイズを交えたりして，匿名性などを実現する仕組みである．

一方で，データへのアクセスに関し，

> どのデータを，どういう条件下で，どのように扱うことができるのか（できないのか）という権利や，その際に発生する義務

をルールとして定義し，それに従ってデータを活用するアプローチもある．

ルールを誰がいつ定義するかはさまざまである．多くの場合，情報を活用するサービスの提供者などが，「ポリシー」などとして定義する．普遍性，重要性が高い領域では，国などにより「法令」として与えられることもある．一方で，個人情報の対象となる利用者などの個人が，自分自身で都度「設定」として与えることができることも多い．

いずれにしても，個人情報保護に関する社会や法の要請が高まる中，こういったルールの扱いがますます重要になると考えられる．

なお，ルールに基づくアプローチは，現状では実際に受け入れられやすいように思える．個人のプライバシーを強化する技術（PETs：Privacy Enhancing Technologies）にはさまざまなものがあるが，それらの導入を促す要因を Rubinstein が議論している [1]．この議論も参考にし，ルールに基づくアプローチの特徴を挙げる．

- 技術の観点からは，現状普及している技術を置き換えて導入するのではなく，サービス提供者側で追加の制御を加えるという形式で実現しやすい．また，アクセス制御ルールやビジネスルールなど，情報システムの開発，運用においてすでに馴染みのある考え方に近い．
- 活用の観点からは，個人情報をそもそも渡さない・保持しないという方針に限らず，サービス提供側の要求やビジネスモデルも踏まえ，データを十分に活用する余地を残すことができる．一方，活用範囲の制限，選択肢の提供，通知や同意の徹底などの形で，プライバシーへの配慮を行うこともできる．

これらの特徴は，特別なツールを各個人が用い，IPアドレスやアクセス元の国までもほぼ追跡できなくしてしまうような極端な場合と対比してみると，明確になるであろう．

本稿ではルールに基づくアプローチに関する研究として，ルールを形式言語で与え，計算機に処理させるものを紹介する．開発者が仕様書およびその実装プログラムに対してルールを反映するようなやり方に比べ，より系統的，効率的にルールの実現や管理などを行える．

そもそも「適切」なルールをどう「決める」かについては，本特集にも要求分析に関する解説があるため，そちらをご参照いただきたい．

アクセス制御ルールとの関連

プライバシーの中心となる考え方は，「知られたくない個人情報などが，他者に知られないようにする（そのことを留意しつつ他者が活用する）」というこ

とである．これは，情報の機密性と可用性を扱うアクセス制御の考え方に強く関連する．ただしプライバシーの場合，OECDによるプライバシー保護に関するガイドラインや社会の要請を踏まえ，目的との合致，同意や通知の義務なども扱う必要がある．

古典的なアクセス制御では，アクセスが発生したタイミングで，アクセスをする主体と対象データの属性を基に，その可否の判断をする．このため上記のように，さまざまな条件判定や，通知の義務などを扱うことは想定していない．また状況変化も踏まえて，アクセス発生時に限らない継続的な制御を行うことも想定していない．

これらの限界は，DRM（Digital Rights Management）にも関連して指摘されていた．これに対し，古典的なアクセス制御を拡張したUsage Control（UCON）という制御モデルが提案されている[2]．UCONにおいては，古典的なアクセス制御に加えて，通知などの義務や，時間などの環境条件を扱うことにより，多様な要件を表現できる．また状況変化に対応する継続的な制御も扱っている．

UCONでは，プライバシーの扱いにも言及している．個人データの場合，データを保持するサービスなどの提供者と利用者だけでなく，対象となる個人も権利義務を持つことを想定している．このため，他者が管理している自己の情報について訂正・削除を求める権利（「忘れられる権利」や「積極的プライバシー」と呼ばれる）も表現し得る．

以上のように，標準的，汎用的な基礎モデルや言語の表現能力という観点では，アクセス制御の発展形が十分強力になっている．これに対し，プライバシーポリシー記述言語などの提案は，目的や同意に関する語彙を与えるなど，義務や環境条件の記述方法を特化させ詳細化したものであることも多い．たとえばアクセス制御の標準言語XACMLにおいては，義務や環境条件を扱うことができる（eXtensible Access Control Markup Language，現在バージョン3.0）．XACMLにてプライバシーを扱うための拡張Privacy Policy Profile 1.0では，目的とその合致判断方法に関する記述欄が追加されるだけのものにとどまっている．

ただし，アクセス制御のモデルや言語における表現能力が十分だとしても，具体的にどのようなルールをどう実現するかについて，プライバシー固有の議論が必要である．上記では「忘れられる権利」に簡単に触れたが，これはEUで予定されるデータ保護指令の改定案にも含まれている．一方，現状のシステム開発・運用では，そのような個人情報の削除権利を保証するような想定はしていないであろう．当然ながら，あるべき姿や現実的な実現方法の模索など，社会的な議論，取り組みが引き続き必要である．

ルールの形式表現の活用

以降では，前章で述べたような，条件付きの権利義務などを記したルールを，形式表現し，計算機に処理させる研究について紹介していく．

まず典型的な活用としては，XACMLなどで記述されたルールを実行時の制御判断に用いる．アクセス要求が発生した際に，その扱いが対応エンジンに問い合わされ，可否判断や，義務達成のための動作起動などが行われる．

一方，具体的なルールを直接制御判断に用いるのではなく，システム実装に対する仕様と見なすこともある．この場合，さまざまな挙動やユーザ設定の可能性を踏まえ，システムの設計によりルールが正しく実現されるか検証する必要がある．

たとえば，S&P，TrustBus，PETS☆1といった関連国際会議における最近の研究発表を見てみると，以下のような研究発表がある．

- UCONにおける，属性の変化を踏まえた制御を行うエンジン
- 複数のSNS（Social Networking Service）における，友人登録や組織情報を踏まえたつながりの強さなどを環境条件として考慮した制御の設定と実現

☆1
- IEEE Symposium on Security and Privacy
- International Conference on Trust, Privacy & Security in Digital Business
- Privacy Enhancing Technologies Symposium

- 目的の合致性判断に関する，マルコフ決定プロセスに基づいた形式化

以上のような制御や検証は，形式表現が計算機で処理可能である点に基づく典型的な活用である．次章からは異なる方向性として，法令など外部から与えられるルールの理解支援，ルールが実現されることの根拠の説明という 2 つを紹介する．

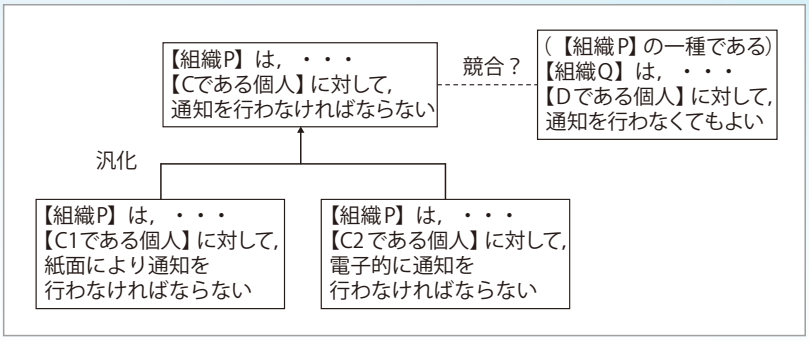

図 -1　法令における条例間の関係整理

ルールの抽出や理解の支援

プライバシーに関するルールを計算機により扱う際には，自然言語で与えられた記述を，形式言語による表現に対応づけることとなる．

自然言語によるルール記述として代表的なものは，特定の領域を対象とした具体的な法令である．たとえば，米国において医療情報の扱いを定めた HIPAA（Health Insurance Portability and Accountability Act）がある．HIPAA では，医療提供者や健康保険会社による，個人の健康情報に対する扱いを定めている．この中には，心理療法記録など患者が閲覧できる権利が発生しない例外，通知を電子的に行う場合と紙面により行う場合の指定など，具体的な内容も含まれる．

Breaux らは，HIPAA における記述を，制約自然言語による表現に（人手で）対応づける分析手法を示している [3]．法律の記述には典型的な語句と構造がある．各条文では，「せねばならない」，「してはならない」，「することができる」といった語句で，権利や義務，およびそれらの委譲を示している．それに対し，「に該当するときは」「の場合を除き」といった語句は，適用の条件となる制約を示している．Breaux らの方法では，これらの言い回しに着目して，統一された制約自然言語表現への対応づけを行う．この際，概念間の包含関係の整理，複数の行為に言及する一文の分解，例外の説明が適用される範囲の判定なども行う．

このように形式言語での表現を行う過程で，あるいは行った結果，法の定めるルールに対する理解支援がなされる．簡単には，さまざまな状況をクエリとして入力し，どのような判断がされるのかシミュレーションすることが考えられる．また，概念（主語や目的語）間の包含関係も踏まえ，**図 -1** のように条文間の汎化関係や，競合の可能性を自動で示すこともできる．実際の条文は，多少の階層化はあるものの箇条書きとして列挙されるため，こういった構造整理は重要であろう．Breaux らはそのほかにも，権利と義務のバランスや，条文記述の曖昧さに関する議論も行っている．

なお，国が定めた法律により与えられるルールについては，十分に検証されていると期待し，「遵守する」というスタンスで臨むことが多いだろう．これに対し，個人情報保護法などの抽象的な法律を踏まえて，組織やシステムごとに独自の具体的なルール（規則，ポリシー，仕様など）を決める場合もある．こういった場合，そのルール自体に対する検証も重要である．Breaux らの取り組みでも競合や曖昧さに関する言及があったが，形式言語を用いたアプローチは，ルール自体の検証にも活用することができる．特に法令を対象としたものについては，「法令（を扱う）工学」という概念も提起されている [4]．

実現根拠に関する表現と推論

実現したいゴールや要求があったとき，それがどうして実現されると言えるのかを論理的に示すことは重要である．このための仕組みとしては，ゴールをその実現に必要となる具体的なサブゴールに分

【ユーザ】は，その同僚【対象者】の位置情報を知ることができる
Warranted by
- 【ユーザ】が【対象者】のアイコンをタップすると，彼らが同僚であるかチェックされる
- 【ユーザ】と【対象者】が同僚であると，【対象者】の位置情報が問い合わせされる
- 位置情報が問い合わされると，最後の GPS 位置が返される
- GPS 情報が返されると，【ユーザ】に提示される
- 【ユーザ】に【対象者】の GPS 情報が提示されると，その位置を知ることになる

図 -2　Argument における根拠の記述

解し整理するゴール指向要求分析手法が挙げられる．また，ある主張が成り立つことを，より具体的な根拠を揃えて示す Argument（議論）として構成，表現することもある．Argument のモデルとしては，想定条件（仮定）や例外条件，考えられる反論と再反論などを含めることも多い．

Tun らは，Argument モデルに基づき，ルールの実現根拠となる動作設計や仮定を表現する言語を提案している [5]．**図 -2** に Argument の根拠（Warrant）部分に関する例を示す．このような根拠の整理は，典型的には開発時の検証に用いられる．Tun らの取り組みにおいては，このような Argument をあくまで骨組みと見なし，実行時の状況，特にユーザの設定によって具体化，上書きされるものとしている．

具体的には，図 -2 における【ユーザ】や【対象者】の部分が個人に置き換わり，個別の設定が反映される．図 -2 は根拠のみ示しているが，Argument には適用条件あるいは例外も含むことが多い．Tun らの取り組みにおいても，平日の勤務時間帯のみ同僚が位置情報を知ることができる，特定の同僚には知らせない，といった個別の設定を適用条件や例外として反映するようになっている．

こういった実現根拠を含む Argument に対しても，形式表現を行えば，厳密な検証や，さまざまな推論も扱うことができるようになる．Tun らの取り組みにおいては，特に実行時の活用として，設定変更を受けて情報取得可否を判定したり，その可否の理由を説明したりすることを想定している．

実行時活用への期待

最後に，実行時の活用に対する個人的な期待を述べたい．Tun らの取り組みは，要求工学に関する国際会議 RE における，“RE@runtime”というセッションにて発表された．この言葉は，情報システムが達成すべきゴール間の依存関係や代替関係，前提条件などを含む要求モデルの形式表現を，実行時にも活用するアプローチを指す．システム自身が，要求モデルを実行状況や環境情報と対応させつつ推論を行うことにより，高度な診断や対応を系統的に行いやすくなる．逆に言うと，要求やその実現根拠が，開発者の頭の中に暗黙的にとどまると，その後「何がどうしてどううまくいくのか」を把握し，再検証や説明，要求や環境の変化への対応などを行うのが難しい．

本稿で扱ったようなルールに対しては，状況や意味を把握しないまま同意を機械的にしてしまう，複雑な制御に対し意図に合う設定方法が分からない，といった問題も取り上げられている．一方，プライバシーに関し，説明責任や透明性の実現に対する要求も高まっている．RE@runtime のビジョンのように，プログラムの作り込みではない，実行時の系統的なモデル活用による対応は，これらの問題に対しても有用ではないかと期待している．

参考文献

1) Rubinstein, I. : Regulating Privacy by Design, Berkeley Technology Law Journal, Vol.26, p.1409 (2012).
2) Park, J. and Sandhu, R. : The UCON ABC Usage Control Model, ACM Transactions on Information and System Security, Vol.7, Issue 1, pp.128-274 (2004).
3) Breaux, T. D., Vail, M.W. and Anton, A. I. : Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations, The 14th IEEE International Requirements Engineering Conference, pp.49-58 (2006).
4) 法令工学：安心な社会システム設計のための総合ソフトウェア科学，情報処理，Vol. 51, No.5, pp.487-490 (2010).
5) Tun, T. T., Bandara, A. K., Price, B. A., Yu, Y., Haley, C., Omoronyia, I. and Nuseibeh, B. : Privacy Arguments : Analysing Selective Disclosure Requirements for Mobile Applications, The 20th IEEE International Requirements Engineering Conference, pp.131-140 (2012).

（2013 年 5 月 17 日受付）

石川冬樹（正会員）｜f-ishikawa@nii.ac.jp

国立情報学研究所 コンテンツ科学研究系 准教授．2007 年東京大学大学院 情報理工学系研究科 コンピュータ科学専攻 博士課程修了．博士（情報理工学）．サービスコンピューティングおよびソフトウェア工学の研究に従事．

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

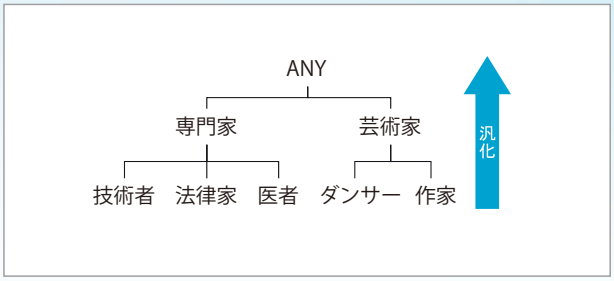

図 -1　汎化ツリーの例

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

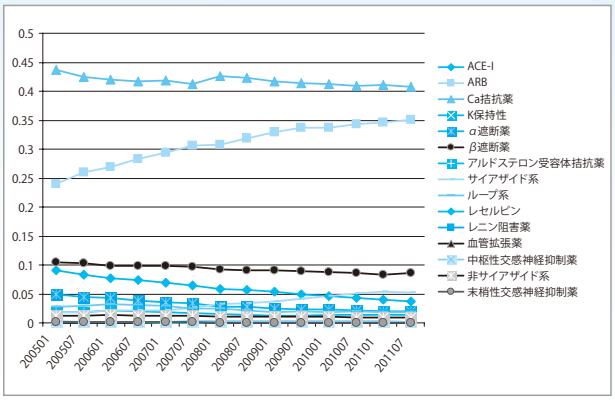

図 -2　元データでの集計結果（著者の許諾を得て，文献 4）から引用）

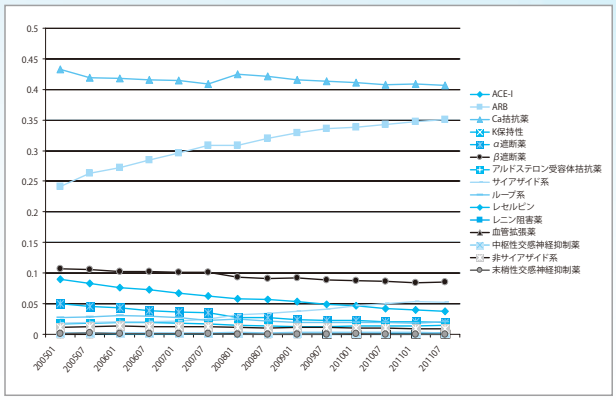

図 -3　匿名化後のデータでの集計結果（著者の許諾を得て，文献 4）から引用）

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．