クラウドコンピューティング・ビッグデータ利活用を支える先進セキュリティ技術

Advanced Security Technologies for Cloud Computing and Utilization of Big Data

藤井 康広 Fujii Yasuhiro
佐藤 尚宜 Sato Hisayoshi
吉野 雅之 Yoshino Masayuki
原田 邦彦 Harada Kunihiko

クラウドコンピューティングの普及に伴い，クラウド上に集積された膨大な量の情報を分析して新たな価値創出につなげる，ビッグデータの利活用への期待が高まっている。一方で，個人情報などといった機密情報の漏洩（えい）防止も大きな課題であり，収集した膨大なデータを強固に保護しつつ，安全な利活用を可能にするような新たなセキュリティ技術が求められている。
日立グループは，データの強固な保護と利活用を両立する先進セキュリティ技術の研究開発を推進し，安全性の高いプラットフォームを提供していく。

1. はじめに

クラウドコンピューティングの普及により，データベースサーバへの情報の預託が活発になってきている。また，クラウドに集まった膨大な量の情報を分析することで新たな顧客価値を創出する，いわゆるビッグデータ利活用への期待も高まっている。

その一方で，個人情報などといった機密情報の流出も大きな社会問題となってきている。例えば，マルウェアなどで巧妙な攻撃を仕掛けて，企業・官公庁などから機密情報を盗み出すといった不正が日々メディアをにぎわせている。また，不正の目的はなくとも，データの二次利用のためにモバイル機器などから通信履歴などの個人情報を許可なく収集し，そのことが発覚したためにブランドが毀損され，存亡の危機に瀕（ひん）した企業も少なくない。

クラウドコンピューティングやビッグデータといった新たな技術潮流の浸透により，データを安全に保管するだけではなく，収集した膨大なデータを安全に利活用できるようにする新たなセキュリティ技術が求められるようになってきた。このようなセキュリティ技術があれば，情報漏洩に関する事業者への不安を解消してデータの預託を促進し，かつプライバシーを適切に保護することでデータの二次利用への不安を解消して，新たな顧客価値を創出することが期待できる（図1参照）。

ここでは，データの強固な保護と利活用を両立する最新

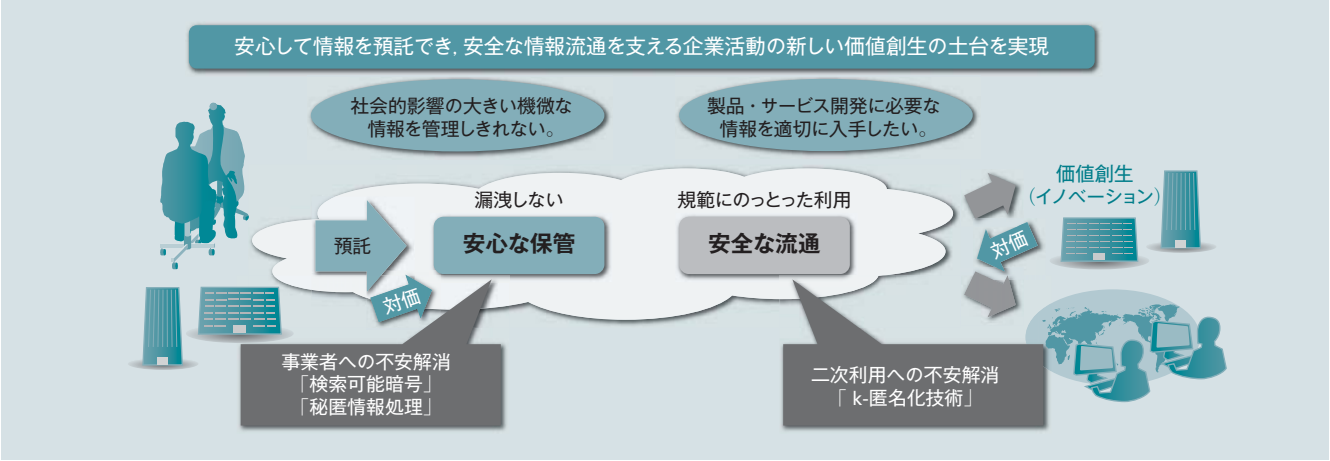


図1｜ビッグデータ利活用を支える先進セキュリティ技術
社会的影響の大きい機微なデータの安全な保管を実現する「検索可能暗号」，「秘匿情報処理」により，事業者への不安を解消し，データの預託を促進する。プライバシー保護を実現する「k-匿名化技術」によってデータの二次利用への懸念を払拭し，ビッグデータ利活用ならではの顧客価値を創生する。

のセキュリティ技術の中から，データの預託先にもデータを秘匿したままで検索や数値演算を可能にする「検索可能暗号」と「秘匿演算技術」，個人情報を曖昧（あいまい）化してデータの安全な二次利用を可能にする「k-匿名化技術」の特長や効果について述べる。

2. 検索可能暗号

まず，通信路上の第三者やデータセンター管理者に対して一切の情報を漏らさずに検索が可能な「検索可能暗号」について述べる。

2.1 従来方式とその問題点

データを安全にやり取りするため，これまではVPN（Virtual Private Network）やSSL（Secure Socket Layer）などで通信路を暗号化し，かつ格納するデータベースも暗号化しておく方式が多く用いられてきた。この方式により，第三者による盗聴を防止することはできる。

しかし，このような方式では暗号鍵がデータセンター内に保管されるため，鍵にアクセス可能なデータセンター管理者は，簡単にデータを復号できてしまう。また，海外のデータセンターにおいて，データセンターへ強制的な情報開示命令が出されたときにも，データセンター内にある鍵を用いて暗号が復号されて当局にデータが筒抜けになってしまうおそれがある〔**図2**（a）参照〕。

こうした課題を解決するため，最近ではデータセンター管理者にも一切の情報を漏らさずに所望の処理ができるような新しいセキュリティ技術が検討されるようになってきている。特に検索や照合は最も基本的な処理であるため，暗号化を解かずに検索できる「検索可能暗号」の研究が急速に脚光を浴びている。

検索可能暗号の最も単純な実現手法は，まずクライアントでデータを暗号化してからデータセンターに登録し，検索時には同じ暗号鍵で検索クエリを暗号化してデータセンターに問い合わせるというものである[1)]。このような暗号化は一般に「トークン化」と呼ばれる。同一データから同一の暗号文が生成されるため，暗号を解かなくても，暗号文の一致を照合することで完全一致検索ができる。

しかし，このトークン化の方式にはセキュリティ上の問題がある。例えば「男」が暗号化されて「88a」に，「女」が「7b6」に変換されたとする。当然「88a」が「男」を意味しているとは分からない。しかし，データセンター管理者が預託されたデータを観察し，もしもあるカラムが「88a」と「7b6」の2種類のデータしかないと気づいた場合，元データが性別を表していると容易に特定できてしまうだろう。このような解読法は一般に頻度分析と呼ばれる。従来の検索可能暗号は頻度分析によって元データが推測できてしまうという問題がある〔**図2**（b）参照〕。

2.2 日立の提案方式

従来の検索可能暗号は頻度分析に弱く，依然として情報漏洩のおそれがある。そこで日立製作所は，頻度分析に耐性を持たせることを第一目標として研究を進めてきた。また，実用上の観点からデータの検索速度が通常の検索と同レベルに達するよう，処理の高速化を第二の目標とした。

第一目標である頻度分析への耐性を実現するために，暗号化と乱数を組み合わせた新方式を提案した[2), 3)]。乱数と

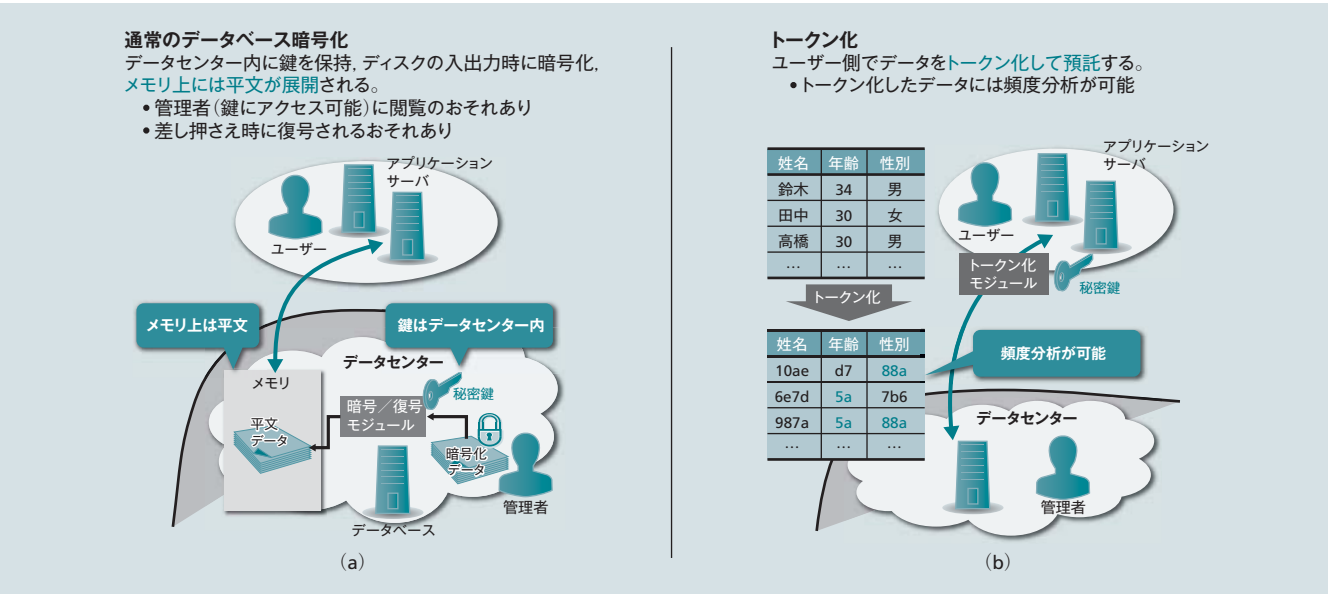
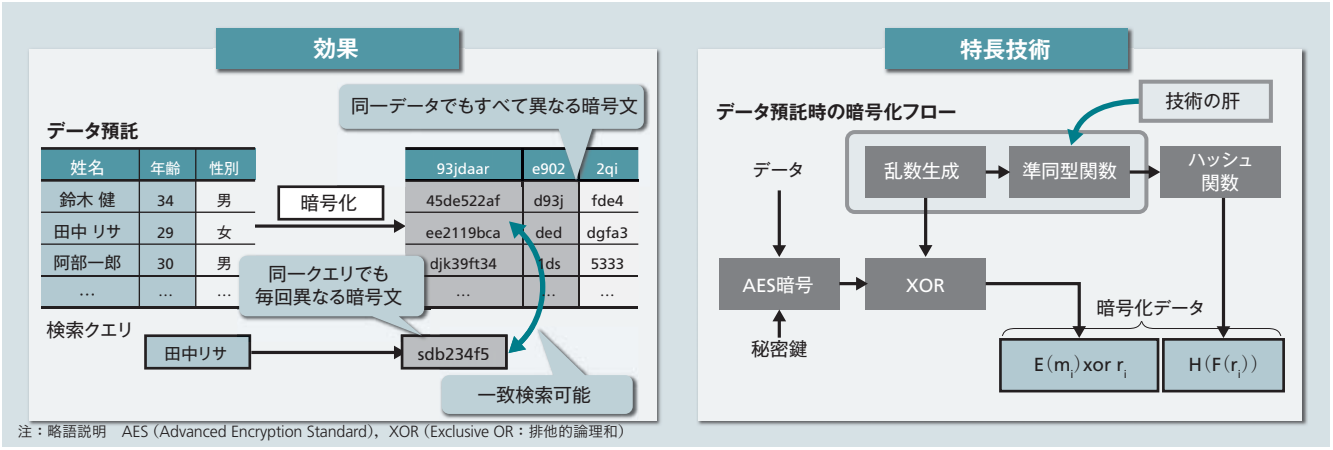


図2｜従来の検索可能暗号

通常のデータベース暗号化では，管理者に解読される可能性がある。従来の検索可能暗号では，クライアントでデータをまず暗号化（トークン化）してからデータセンターに登録し，検索時に同じ暗号鍵で検索クエリをトークン化してデータセンターに問い合わせるが，この方式には頻度分析に弱いという課題がある。



注：略語説明　AES（Advanced Encryption Standard），XOR（Exclusive OR：排他的論理和）

図3｜日立の検索可能暗号

乱数生成と準同型関数（行列の拡張）を導入し，同一キーワードでも毎回，異なる暗号文を生成して安全性を向上している。

排他的論理和をとってデータや検索クエリをランダム化することで，同一データを暗号化しても毎回異なる暗号文が生成されるようにした。

検索処理を可能にするためには，ランダム化されて互いに異なる暗号文であっても元データの同一性を判定しなければならない。そこで，「準同型関数」と呼ばれる数学的概念を導入し，元データのランダム化に用いた乱数と検索クエリの乱数とが暗号を解かずに相殺するような新しい関数を定式化した。これにより，乱数生成との連携と準同型関数の定式化により，第一目標である頻度分析に対する耐性を持たせることに成功した[2), 3)]（**図3**参照）。

さらに，暗号化には高速な共通鍵暗号方式を用い，準同型関数による演算を最小化することで，第二目標である処理の高速化を達成した。データ1万件当たり10ミリ秒という通常の検索と同レベルの処理時間で検索を実行することが可能である。

2.3　今後の展開

日立製作所の検索可能暗号方式は頻度分析に耐性を持つ高安全な方式であり，データ1万件当たり10ミリ秒と高速に検索を実行することができる。現状は完全一致検索しか実行できないが，今後対応できる検索の種類を増やしていく。また，ビッグデータの適用に備え，検索処理をさらに高速化する，検索用のインデックスをセキュアに生成する方式についても研究を進めている。

この方式によって情報漏洩に関する事業者への不安を解消して，クラウドへのデータの預託を促進していく。

3.　秘匿演算技術

次に，通信路上の第三者やデータセンター管理者に対して一切の情報を漏らさず足し算やかけ算などの数値演算が可能な「秘匿演算技術」について述べる。

3.1　従来方式とその問題点

通常の暗号では，暗号データに対して演算はできない。暗号データに対して任意の演算を許可すると，ノイズが非常に大きくなり，元のデータに復号できなくなるからである。そのため，実用的な秘匿演算技術を実現するのは非常に困難であると考えられていた。

ところが近年，2種類の暗号アルゴリズムを導入し，数値演算によって発生したノイズを互いに打ち消しあう巧妙な手法（完全準同型暗号）が提案された[4)]。この方式により，理論的には任意の数値演算を何回でも暗号を解かずに行うことが可能であることが示された。

しかし，この理想的な機能を安全に達成するために導入した種々のアイデアは，効率性に大きく影響する。実際にこの従来方式では，安全性を確保しつつ1ビット送信するのに，少なくとも10,000ビット近くダミーデータを導入する必要があるなど，データ転送量や計算量が爆発的に増大してしまう。

3.2　日立のアプローチ

任意の演算について秘匿演算技術を実現することは理想ではあるが，上述のように計算量の観点で本質的な問題があり，従来方式のアプローチでは実用には適さないと考えられる。そこで，秘匿化して行わなければならない演算を絞り込むというアプローチをとって研究を進めている。例えば，ある業務において平均値だけが分析に用いられていることが明らかになった場合，任意回数の足し算と1回の割り算のみ秘匿化し，ほかは平文のまま演算できるようにアルゴリズムを設計する。

すでに加法の秘匿化については確立しており，クラウドサービスへの適用を進めている。並行してアプリケーション探索を行って必須演算を絞り込み，サービス立ち上げを加速していく（**図4**参照）。

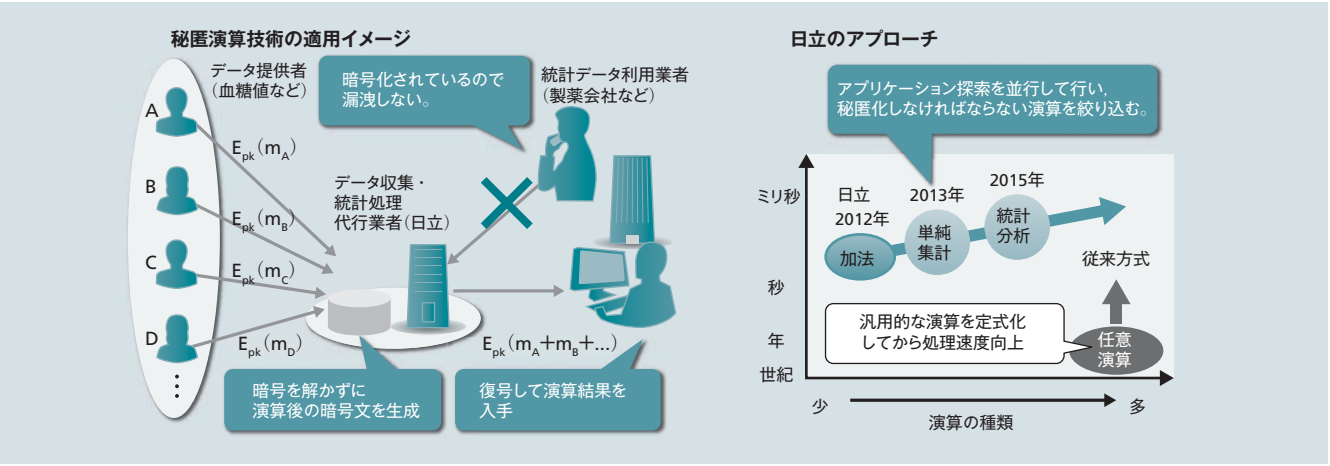


図4｜日立の秘匿演算技術の研究アプローチ

任意の演算について秘匿演算技術を実現することは理想ではあるが，計算量の観点で本質的な問題がある。日立製作所はアプリケーション探索を並行して行い，秘匿化しなければならない演算を絞り込んで，サービス立ち上げを加速する。

4. k-匿名化技術

上述の検索可能暗号および秘匿演算技術はデータを強固に保護するものであるが，検索や数値演算など所定の処理しか許可しないため，データの利活用の面で十分ではない場合もある。データ利活用に重点を置き，個人情報など特にプライバシー面での配慮が要求される機微な情報のみを曖昧化してデータの安全かつ自由な二次利用を可能にするものとして「k-匿名化技術」を研究している。次に「k-匿名化技術」の特長や効果を述べる。

4.1 従来方式とその問題点

データの二次利用のための従来の匿名化は，氏名や電話番号など明らかに本人を特定できるような属性を削除したり，住所などを途中までしか開示しないよう曖昧にしたりするものであった。しかし，まれなレコードが存在すると唯一に特定可能な情報が残ってしまい，データが安全でないということが起こりうる。例えば，氏名と住所と年齢からなる表データに対して氏名を削除し，住所を県名だけにしたとしても，仮に県名が愛知県で年齢が100歳のレコードが存在した場合，そのレコードが誰を指すか容易に特定できてしまう（**図5**参照）。

k-匿名化技術はこのようなことを防ぐ技術で，同一レコードが少なくともk件以上存在するように各セルを曖昧にするというものである。同図ではk = 2の例を示してい

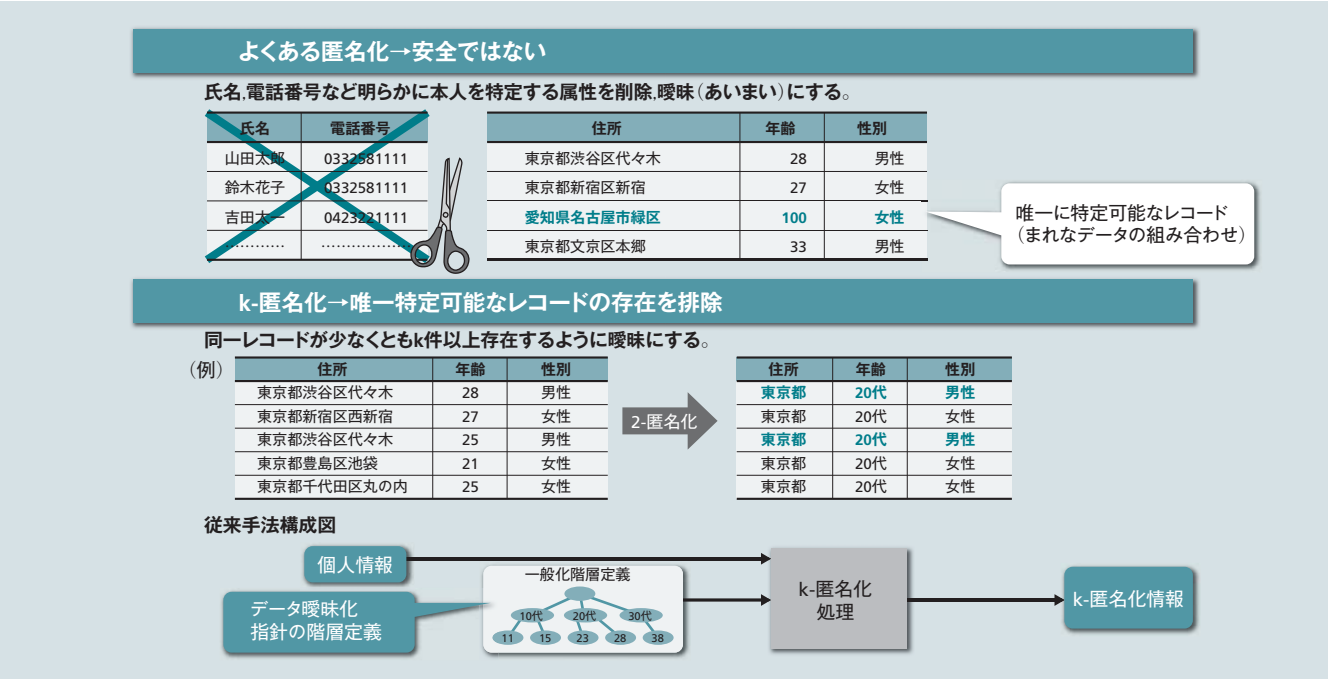


図5｜従来の匿名化技術

単純に氏名などを削除する方式では，まれなレコードが存在すると個人を特定できてしまう。k-匿名化技術により，同一レコードが少なくともk個以上存在するように曖昧にする。ユーザーが辞書情報（一般化階層）を準備する必要がある。

る。各セルを曖昧化することで同一のレコードが2件以上存在することが見て取れる。

従来のk-匿名化技術では，匿名化を実現するために，あらかじめユーザーがデータをどのように曖昧にするかという指針を与える辞書情報（一般化階層）を準備する必要がある[5]。個人情報が大量の場合には準備に時間がかかり，また準備した一般化階層によっては，k-匿名化に伴い必要以上の情報を落としてしまう可能性がある。

まとめると，従来のk-匿名化技術には一般化階層をユーザーが準備しなければならないという課題と，匿名化に伴う情報損失が大きいという課題があった。

4.2　日立の提案方式

従来のk-匿名化技術の課題を解決するために，日立製作所は，情報損失を低減するような一般化階層を自動生成することを目標として研究を進めてきた。

一般化階層の自動生成を実現するために，文字の出現頻度の活用によってデータ圧縮を達成するハフマン符号木などを用いた新方式を提案した[6]（**図6**参照）。この方式は，まれな情報を優先的に曖昧化するような階層を自動的に生成する。

さらに，情報損失の低減を達成するため，k-匿名化処理の内部でどれだけの情報量が失われたかを随時定量的に評価するための指標として，情報理論で用いられる「エントロピー」の概念を導入した。これらの手段により，情報損失を大きく抑えることに成功した[6]。

4.3　今後の展開

k-匿名化技術は，従来，ユーザーが生成する必要があった辞書情報（一般化階層）を，ハフマン符号木とエントロピーを導入することで，情報損失をできるだけ抑えるように自動生成する技術である。この方式により，プライバシー侵害に対する事業者の不安を解消して，データの二次利用を活発化させ，ビッグデータ利活用を加速させていく。

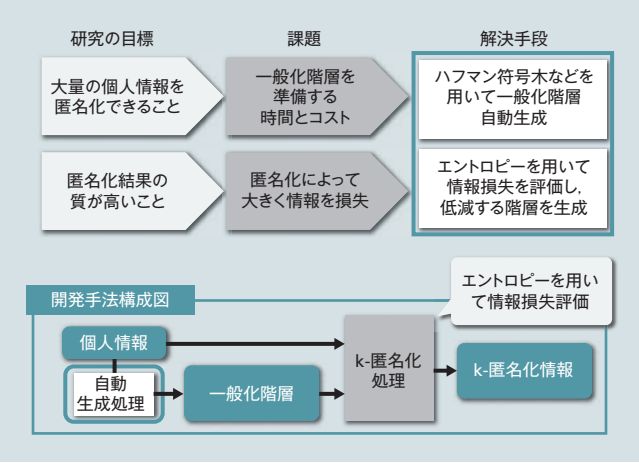


図6｜日立のk-匿名化技術
ハフマン符号木などを用いて一般化階層を自動生成し，エントロピーを用いて情報損失を評価し，低減する一般化階層を生成する。

5.　おわりに

ここでは，データの強固な保護と利活用を両立する最新のセキュリティ技術の中から，データの預託先にもデータを秘匿したままで検索や数値演算を可能にする「検索可能暗号」と「秘匿演算技術」，個人情報を曖昧化してデータの安全な二次利用を可能にする「k-匿名化技術」の特長や効果について述べた。

今後も，日立グループは，データの強固な保護と利活用を両立する先進セキュリティ技術の研究開発を推進し，安全性の高いプラットフォームを提供していく。

参考文献など

1) クラウドデータ保護ソリューション，株式会社サンブリッジ，http://www.sunbridge.com/cloud/cloud-security/ciphercloud/
2) M. Yoshino, et al. : Symmetric Searchable Encryption for Database Applications, Network-Based Information Systems - NBIS, pp. 657-662 (2011)
3) 日立ニュースリリース，クラウド上での情報漏えい防止に貢献する検索可能暗号技術を開発（2012.3）http://www.hitachi.co.jp/New/cnews/month/2012/03/0312.html
4) C. Gentry : Fully homomorphic encryption using ideal lattices, In STOC 2009, pp. 169-178 (2009)
5) L. Sweeney : Achieving *k*-anonymity privacy protection using generalization and suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, pp. 571-588 (2002.10)
6) 原田，外：一般化階層木の自動生成と情報エントロピーによる歪度評価を伴うk-匿名化手法，電子情報通信学会技術研究報告，110，115，p. 289-295 (2010.7)

執筆者紹介

藤井 康広
2001年日立製作所入社，横浜研究所 情報サービス研究センタ サービスイノベーション研究部 所属
現在，クラウドコンピューティング・ビッグデータ利活用を支えるセキュリティ技術の研究開発に従事
博士（理学）

佐藤 尚宣
1999年日立製作所入社，横浜研究所 情報サービス研究センタ サービスイノベーション研究部 所属
現在，クラウドコンピューティング・ビッグデータ利活用を支えるセキュリティ技術，特に検索可能暗号と秘匿演算技術の研究開発に従事
博士（数理学）
電子情報通信学会会員

吉野 雅之
2002年日立製作所入社，横浜研究所 情報サービス研究センタ サービスイノベーション研究部 所属
現在，クラウドコンピューティング・ビッグデータ利活用を支えるセキュリティ技術，特に検索可能暗号と秘匿演算技術の研究開発に従事
電子情報通信学会会員，情報処理学会会員

原田 邦彦
2009年日立製作所入社，横浜研究所 情報サービス研究センタ サービスイノベーション研究部 所属
現在，クラウドコンピューティング・ビッグデータ利活用を支えるセキュリティ技術，特にk-匿名化技術の研究開発に従事
博士（情報理工学）