(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6712529号
(P6712529)
(45)発行日　令和2年6月24日(2020.6.24)　(24)登録日　令和2年6月3日(2020.6.3)

(51)Int.Cl.　F I
G06F 21/62 (2013.01)　G06F 21/62 354

請求項の数 10 (全 17 頁)

(21)出願番号　特願2016-205505(P2016-205505)
(22)出願日　平成28年10月19日(2016.10.19)
(65)公開番号　特開2017-90905(P2017-90905A)
(43)公開日　平成29年5月25日(2017.5.25)
審査請求日　令和1年10月15日(2019.10.15)
(31)優先権主張番号　14/931,774
(32)優先日　平成27年11月3日(2015.11.3)
(33)優先権主張国・地域又は機関
米国(US)

早期審査対象出願

(73)特許権者　504407000
パロ　アルト　リサーチ　センター　インコーポレイテッド
アメリカ合衆国　カリフォルニア州　９４３０４　パロ　アルト　カイオーテ　ヒル　ロード　３３３３
(74)代理人　100086771
弁理士　西島　孝喜
(74)代理人　100088694
弁理士　弟子丸　健
(74)代理人　100094569
弁理士　田中　伸一郎
(74)代理人　100067013
弁理士　大塚　文昭

最終頁に続く

(54)【発明の名称】暗号化済データを匿名化するためのコンピュータで実施されるシステムおよび方法

(57)【特許請求の範囲】
【請求項１】
暗号化済データを匿名化するためのコンピュータで実施されるシステムであって、
匿名化するためのデータセット内の少なくとも１つの属性を特定する特定モジュールであって、各属性が複数のデータ値に関連する、特定モジュールと、
特定される属性ごとに各データ値を暗号化し、その一方で、前記暗号化済データ値の順序を維持する暗号化モジュールと、
前記暗号化される値を順序付けする順序付けモジュールと、
前記暗号化済データ値の順序に基づいて、前記順序付けされる暗号化済データ値を２つ以上のクラスに分割する分割モジュールと、
分割される各クラス内の前記暗号化済データ値の範囲を決定する決定モジュールと、
前記クラスのうちの１つクラスの前記範囲を匿名化データとして、そのクラス内の各暗号化済データ値に割り当てる割り当てモジュールと、を含み、
モジュールがプロセッサを介して実行される、システム。
【請求項２】
前記匿名化済データ値を信頼のおけない第三者に供給する供給モジュールをさらに含む、請求項１に記載のシステム。
【請求項３】
匿名化するための前記データセット内の別の属性を特定し、前記別の属性に関連するマスキングされた値を前記データ項目にそれぞれランダムに割り当てるマスキングモジュー

ルをさらに含む、請求項１に記載のシステム。
【請求項４】
前記少なくとも１つの属性と一緒に匿名化するための、前記データセット内の別の属性を特定し、前記別のデータ値を暗号化し、前記分割された各クラスに関する前記別の暗号化済データ値をさらに分割されたクラスに分割し、前記少なくとも１つの属性および前記別の属性に関するクラスのグループを生成する連結分割モジュールであって、前記少なくとも１つの属性に関する各データ値が、前記別の属性に関する別のデータ値に対応する、連結分割モジュールをさらに含む、請求項１に記載のシステム。
【請求項５】
ｎ個の分割されたクラスを特定するクラス識別子をさらに含む、請求項１に記載のシステム。
【請求項６】
暗号化済データを匿名化するためのコンピュータで実施される方法であって、
中央処理ユニット、メモリ、インプットポート、及び出力ポートを備えた信頼のおけるサーバによって、匿名化するためのデータセット内の少なくとも１つの属性を特定するステップであって、各属性が複数のデータ値に関連する、ステップと、
前記信頼のおけるサーバによって、特定される属性ごとに各データ値を暗号化し、その一方で、前記暗号化済データ値の順序を維持するステップと、
匿名化装置を介して、前記暗号化される値を順序付けするステップと、
前記匿名化装置によって、前記暗号化済データ値の順序に基づいて、前記順序付けされる暗号化済データ値を２つ以上のクラスに分割するステップと、
前記匿名化装置によって、分割される各クラス内の前記暗号化済データ値の範囲を特定するステップと、
前記匿名化装置によって、前記クラスのうちの１つの前記範囲を匿名化データとしてそのクラス内の各暗号化済データ値に割り当てるステップと、を含む方法。
【請求項７】
前記匿名化済データ値を信頼のおけない第三者に供給するステップをさらに含む、請求項６に記載の方法。
【請求項８】
匿名化するための前記データセット内の別の属性を特定するステップと、
前記別の属性に関連するマスキングされた値を前記データ項目にそれぞれランダムに割り当てるステップと、をさらに含む、請求項６に記載の方法。
【請求項９】
前記少なくとも１つの属性と一緒に匿名化するための、前記データセット内の別の属性を特定するステップであって、前記少なくとも１つの属性に関する各データ値が、前記別の属性に関する別のデータ値に対応する、ステップと、
前記別のデータ値を暗号化するステップと、
前記分割された各クラスに関する前記別の暗号化済データ値をさらに分割されたクラスに分割するステップと、
前記少なくとも１つの属性および前記別の属性に関するクラスのグループを生成するステップと、をさらに含む、請求項６に記載の方法。
【請求項１０】
ｎ個の分割されたクラスを特定するステップをさらに含む、請求項６に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本出願は、一般に機密性の高いデータを保護することに関し、より詳細には、暗号化済データを匿名化するための、コンピュータで実施されるシステムおよび方法に関する。
【背景技術】
【０００２】

通常業務において、企業は大量のデータを蓄積している。近年、所有データを広告者、研究者、または共同パートナーなどの第三者と共有することにより、これらのデータを貨幣化する企業が現れ始めてきている。第三者は、適切な料金を支払い、データの所有者から関連するデータを受け取る。次いで、第三者は、これらのデータを用いて、広告する対象をしぼる、あるいは研究を行うことができる。しかし、大抵の場合、第三者が要求するデータには、１人以上の個人のプライベートに関する情報が含まれており、その情報からデータが集められてしまう。

【０００３】

例えば、病院は患者のカルテを管理しており、これらのカルテには、患者の身元情報、年齢、住所、社会保障番号、および医療診断が記載されている。糖尿病に関する研究を行う第三者機関では、米国において、ＩＩ型糖尿病と診断された４０才未満の患者が最も多い地域と最も少ない地域を特定する必要がある。データ所有者は、要求データを送信する前に、信頼のおけない第三者が供給データを用いて、個人情報をアクセスしたり、個々の身元情報を特定したりできないようにする処理を施さなければならない。

【０００４】

データの匿名化とは、データを変換して機密情報を保護する一方で、そのデータ要求した第三者が使用可能な特徴を維持することである。このデータ変換には、そのデータの精度を低減させること、あるいは、そのデータの一部を削除することが含まれ得る。一般に、データの所有者は、匿名化に関して十分な知識がなく、データを第三者に供給する前に、そのデータを匿名化するよう第三者に依頼している。あるアプローチでは、匿名化サービス行うプロバイダを利用し、このプロバイダにより匿名化されたデータが個々のスタッフに提供される。匿名化されたデータを割り当られたスタッフは、信頼のおけない第三者であっても、これらのデータへのアクセス権を有する。現在、多くの企業は、データ匿名化の前後で、匿名化サービスに対して、基本合意書や守秘義務合意などの秘密保持契約の締結を要求してデータを保護している。

【０００５】

データの匿名化を行う方法は従来から存在するが、これらの方法では、信頼のおけない第三者が匿名化された情報を用いる可能性があるという問題には対処しきれていない。Ｓｗｅｅｎｅｙの米国特許第７，２６９，５７８号明細書では、特定の分野および記録、受信者のプロフィール、および最小匿名レベルなどのユーザの要求に基づいて表の項目が変更されている。ｋの値を算出し、準識別子（ある属性のグループに渡って割り当てられる同じ値を有するｋ個のタプルから成る）を特定して発行する。各属性の機密性を決定し、機密性の高い属性（等価クラス代入などの）ごとに、一方向ハッシングまたは一般化置換などの置換方法を決定する。一般化置換とは、最大数の個別の値を有する属性を特定し、その値に対して供給される情報量を削減することにより、その属性の各値を一般化することである。例えば、月、日、および年を有する日付情報は、月と年、年のみ、あるいは年の期間に一般化可能である。しかし、Ｓｗｅｅｎｅｙでは、信頼のおけない当事者により匿名化が行われる可能性があり、匿名化される予定のデータが保護されていないことが考慮されていない。さらに、Ｓｗｅｅｎｅｙでは、匿名化されるデータセットが分割される複数のクラスを特定し、そのデータ値が属するクラスに基づいて、各データ値を匿名化することが記載されていない。

【０００６】

さらに、ＬｅＦｅｖｒｅ ｅｔ ａｌの「Ｍｏｎｄｒｉａｎ Ｍｕｌｔｉｄｉｍｅｎｓｉｏｎａｌ Ｋ－Ａｎｏｎｙｍｉｔｙ」と題する論文には、各領域が、ｋ個以上のポイントを含むよう、一次元分割または多次元分割を用いて、データセットを分割することが記載されている。ある例では、この分割は、中間分割を用いて行われ得る。しかし、このＬｅＦｅｖｒｅの論文には、信頼のおけない第三者がデータを匿名化する場合、匿名化を行う前に、そのデータを保護するステップが記載されていない。さらに、ＬｅＦｅｖｒｅには、データの機密保持の手段を提供して、匿名化する属性を自動的に特定することに関する記載がなく、さらにマスキングも考慮されていない。

【発明の概要】
【発明が解決しようとする課題】
【０００７】
したがって、データを集めることができる個人のプライバシーを危険にさらすことなく、第三者が匿名化を行うことができる機密データを作成するアプローチが必要である。
【課題を解決するための手段】
【０００８】
企業が所有データを商業化できるようにするために、データの所有者は、データセット内の機密情報が確実に保護される手段を施さなければならない。こういったデータを保護するために、データセット内の各属性を分析して、その属性の機密性を判定する。これらの属性の機密性に基づいて、一般化やマスキングなどにより、データの匿名化を行って、これらのデータを難読化することができる。匿名化を選択する際、属性に関するデータ値を暗号化し、信頼のおけないデータ匿名化サイトに送信する。この匿名化サイトが、属性に関する暗号化済データ値をクラスに分割することにより、データを匿名化し、各クラス内の暗号化された値の範囲を特定する。最後に、これら各クラスの範囲を、そのクラスが属する暗号化された各データ値に割り当てる。
【０００９】
本発明の実施形態により、暗号化済データを匿名化するためのコンピュータで実施されるシステムおよび方法が提供される。複数のデータ値に関連する、匿名化するためのデータセット内の少なくとも１つの属性を特定する。特定された属性ごとに各データ値を暗号化し、一方で、暗号化済データ値の順序を維持する。暗号化済値を順序付けし、暗号化済データ値の順序に基づいて、順序付けされ暗号化済データ値を２つ以上のクラスに分割する。各分割クラス内の暗号化済データ値の範囲を特定し、匿名化されたデータとして、それらのクラスのうちの１つの範囲をそのクラス内の暗号化された各データ値に割り当てる。
【００１０】
以下の詳細な説明により、本発明のさらに別の実施形態は、当業者にとって明らかとなり、本発明の実施形態は、本発明を実施するために考えられた最良のモードを示すことにより説明される。気付かれている通り、本発明は、その他の実施形態および異なる実施形態も包含可能であり、いくつかの詳細は、全て本発明の趣旨と範囲を逸脱することなく種々の明白な点で変更可能である。したがって、これらの図面や詳細な説明は、本来、説明をその目的とし、限定することを意図していない。
【図面の簡単な説明】
【００１１】
【図１】図１は、一実施形態に従った、暗号化済データを匿名化するためのコンピュータで実施されるシステムを示すブロック図である。
【図２】図２は、一実施形態に従った、暗号化済データを匿名化するためのコンピュータで実施される方法を示す流れ図である。
【図３】図３は、順序維持暗号化の処理を例として示すブロック図である。
【図４】図４は、一般化を用いて、データを匿名化する処理を例として示す流れ図である。
【図５】図５は、一般化を用いる匿名化に関するデータ値の表を例として示すブロック図である。
【図６】図６は、暗号化済データ値を分割する処理を例として示す流れ図である。
【図７】図７は、一般化を介して、単一属性に関するデータ値を匿名化する処理を例として示す流れ図である。
【図８】図８は、単一属性の暗号化済データ値に関する木を例として示すブロック図である。
【図９】図９は、一般化を介して、２つ以上の属性のデータを共に匿名化する処理を例として示す流れ図である。

【図１０】図１０は、第２の属性に関するデータ値の分割化を例として示すブロック図である。
【図１１】図１１は、属性のうちの２つに関する一般化された匿名化の値を有する図５のデータセットを例として示すブロック図である。
【図１２】図１２は、マスキングを介してデータ値を匿名化する処理を例として示す流れ図である。
【図１３】図１３は、マスキングされた匿名化済データ値を有する図１１のデータセットを例として示すブロック図である。
【発明を実施するための形態】
【００１２】
所有しているデータの商業化に関心を示す企業が増加している中、データを確実に保護することが極めて重要となっている。通常、企業は、データの機密性に関して十分な知識を持っておらず、データを第三者の要求に応じて供給する前にデータを匿名化する業者を抱えている。しかし、大抵の場合、外部の匿名化サービスは信頼がおけず、前もってデータを保護する目的で、守秘義務合意などの秘密保持契約を締結している。秘密保持契約の代わりデータの保護を確実にするために、好適なレベル匿名化を選択し、必要な場合、見えなくするようにその匿名化をやみくもに行うことができる。見えなくする匿名化を行うとき、第三者の匿名化サービスにデータを送信する前に、データの暗号化する。次いで、暗号化済データを複数のクラスに分割し、各クラスにおいて暗号化された範囲を特定する。これらの範囲を、匿名化値として、そのクラス内の各暗号化済データ項目に割り当てる。
【００１３】
見えなくする匿名化により、企業はデータを確実に保護でき、データを集めることができる個人のプライバシーを危険にさらすことなく、外部業者に自分たちのデータを提供することができるようになる。図１は、一実施形態に従った、暗号化済データを匿名化するためのコンピュータで実施されるシステムを示すブロック図である。データ所有者は、長年の間蓄積された大量のデータ２２を保持している。データ２２は、業者内のデータ所有者用ロケーションや遠隔地に位置する、信頼のおけるサーバ１７に相互接続するデータベース２１に格納され得る。あるいは、データは、複数のサーバ上のプールを含むクラウドに格納され得る。データ所有者は、デスクトップ・コンピュータ１１Ａ、ラップトップ・コンピュータ１１Ｂ、またはモバイルコンピュータ装置（図示せず）などの、その他の種類のコンピュータ装置を介してデータ２２にアクセス可能である。格納されているデータ２２には、１つ以上のデータセット２２が含まれ得、これらのデータセットがそれぞれ１人以上の個人の複数の属性に関連している。さらに、各属性は各個人に関するデータ値に関連している。
【００１４】
この信頼のおけるサーバ１７は、機密性特定モジュール１８、暗号化モジュール１９、および送信モジュール２０を含む。この機密性特定モジュール１８は、要求されたデータセット内で、機密である可能性の高い属性を特定し、匿名化を要求することができる。機密である可能性のある属性が特定されると、暗号化モジュール１９は、それらの属性に関する各データ値を暗号化することができる。その後、送信モジュール２０は、暗号化された全てのデータ値を信頼のおけない第三者の匿名化サイト１２に送信し、そこで匿名化が行われる。匿名化サイト１２は、機密性割り当てモジュール１３、順序付けモジュール１４、クラス生成モジュール１５、およびラベル割り当てモジュール１６を含む。この機密性割り当てモジュール１３は、信頼のおけるサーバ１７から受け取る暗号化済データ値に関する各属性を分析し、機密性の値を各属性に割り当てる。これらの機密性の値により、そのデータが、社会保障番号、クレジットカード番号、または住所などの個人情報を含んでいるかどうかといったデータ機密性を測定することができる。その他の種類の個人情報が含まれている可能性がある。第三者が個人データにアクセスできなくするために、属性に割り当てる機密性の値によって、その属性に関するデータ値の匿名化が必要となり得る

。匿名化には、データの一般化やマスキングが含まれ得る。機密性の値に基づく属性に対する好適な種類の匿名化に関しては、図２を参照して、以下にさらに詳しく説明する。

【００１５】

一般化する匿名化では、順序付けモジュール１４は、属性に関する暗号化済データ値の順序付けを昇順あるいは降順で行う。次いで、クラス生成モジュール１５が、順序付けされたデータを２つ以上のクラスに分割し、ラベル割り当てモジュール１６が、暗号化済データ値の範囲をクラスごとに決定し、匿名化される値として、そのクラス内で暗号化された各データ値の範囲に割り当てる。次いで、匿名化されたデータは、信頼のおけるサーバ１７に送信され、データベース２１に格納されるかまたは、要求元の第三者に提供される。あるいは、各データ値をマスキングすることで、属性のデータ値を匿名化することも可能である。マスキングモジュール２４は、例えば、ハッシングを用いて、機密性の高い属性を擬似乱数値に置き換える。

【００１６】

コンピュータ装置およびサーバは、中央処理装置、ランダム・アクセス・メモリ（ＲＡＭ）、ハードドライブまたはＣＤ－ＲＯＭドライブなどの不揮発性の補助記憶装置、ネットワークインターフェース、およびキーボードおよびディスプレイなどのユーザインターフェーシング手段を含む周辺装置をそれぞれ含むことができる。プログラムコード（ソフトウェアプログラムを含む）とデータがＲＡＭに読み込まれて、ＣＰＵにより実行および処理され、その結果が生成されて、表示、出力、送信、または格納される。本明細書で開示されている実施形態を実行するための１つ以上のモジュールについて説明する。

【００１７】

これらのモジュールは、コンピュータプログラム、または従来のプログラム言語のソースコードを書き込まれたプロージャとして実装され得、オブジェクトコードまたはバイトコードとして中央処理装置により実行されるために提示される。あるいは、集積回路またはリード・オンリ・メモリのコンポーネントに焼き付けられる回路として、モジュールをハードウェア内で実装することも可能であり、各コンピュータ装置やサーバが専用コンピュータとして機能する。例えば、モジュールがハードウェアとして実装される場合、その特定のハードウェアがメッセージの優先順位付けの実行に特化し、その他のコンピュータは使用することができない。それに加えて、モジュールがリード・オンリ・メモリのコンポーネントに焼き付けられる場合、そのリード・オンリ・メモリを格納するコンピュータ装置またはサーバが、メッセージの優先順位付けの実行に特化し、その他のコンピュータは使用することができない。その他の種類の専用コンピュータも使用可能である。さらに、管理システムを特定のクライアントおよび特定のハードウェアに限定し、これらの上で管理システムを実行することができる、さらに、管理システムをサブスクリプションサービスによりサブスクライビング・クラアントだけに限定することもできる。ソースコード、オブジェクトコード、およびバイトコードの種々の実装形態は、フロッピーディスク、ハードドライブ、デジタル・ビデオ・ディスク（ＤＶＤ）、ランダム・アクセス・メモリ（ＲＡＭ）、リード・オンリ・メモリ（ＲＯＭ）、および同様の記憶媒体などのコンピュータ可読記憶媒体に保持可能である。その他の種類のモジュールおよびモジュール機能、ならびにその他の物理的ハードウェアコンポーネントも使用可能である。

【００１８】

先の研究で匿名化されていないマスキングデータベースの可能性が示されている通り、マスキングデータを供給することにより、一般化よりも低いレベルの機密保護がもたらされる。したがって、マスキングを単独で使用するのではなく、一般化と組み合わせて使用するべきである。データ一般化により、若干のデータ保護がもたらされると同時に、研究や広告のために一般化データを利用する第三者にとってもそのデータの有用性が保たれる。データの一般化を行う際、匿名化を行う前に、属性に関するデータ値を暗号化して、データ保護の付加的な層を加える。図２は、一実施形態に従った、暗号化済データを匿名化するためのコンピュータで実施される方法を示す流れ図である。データ所有者は、特定の種類のデータおよびデータセットに対する要求を受け取り、要求されたデータを特定する

（ブロック３１）。データセットには、個人のリスト、および各個人に関連する属性のデータ値が含まれ得る。データの所有者は、そのデータセットを分析して、１つ以上の属性が機密データまたは個人情報を含んでいる可能性があるかどうか判定し、そうならば、そのデータの機密性がどの程度なのかを判定する。機密データには、特定の個人を特定することができる情報が含まれ、この情報が開示された場合、セキュリティの損失になりかねない。その後、機密性の高い属性に関連するデータ値を暗号化する（ブロック３２）。さらに別の実施形態では、データセット内の全てのデータ値を暗号化して、匿名化サイトに供給し、この匿名化サイトが、どの属性の機密性が高いかを単独で判定することができる。

【００１９】

一実施形態では、順序維持暗号化を用いることができ、この順序維持暗号化が行われている間、暗号化された値を各データ値に割り当てる擬似乱数生成モジュールを通してデータ値が置き換えられるが、非暗号化済データ値の順序は維持される。先進的な暗号規格などの大部分の暗号化アルゴリズムとは異なり、順序維持暗号化では暗号化された形式での平文データの順序は維持される。図３は、順序維持暗号化を行うための処理を例として示すブロック図である。データ値４１の平文バージョンには数字０～５が含まれる。順序維持暗号化の後も平文のデータ値４１の順序を維持して、順序付けされた暗号文のデータ値４２を生成する。具体的には、擬似乱数生成モジュールが、平文値に対して擬似乱数マッピングを行って順序を維持した暗号化値を暗号文値として生成する。順序維持暗号化の唯一の基準は、平文値の順序と暗号文値の順序が交差することができないことである。例えば、数字０は４５７に、数字１は４７３に、数字２は５１０に、数字３は６２５に、数字４は６３５に、数字５は１００１にそれぞれマッピングされる。ゼロは最も小さい平文データ値であり、ゼロの暗号化された値である４５７はデータセットの最も小さい暗号文値である。さらに、平文値５が最も大きく、５を暗号化された値である１００１も最も大きい。

【００２０】

データ値が、色、病状またはその他の種類の数字で表せない値などのテキストを表す場合、暗号化を行う前に、各データ値に対して数値が割り当てられる。例えば、症状や病的状態の厳しさに基づいて数値を病状に割り当て、その後、暗号化することができる。暗号化された値の順序を維持することにより、データは暗号化されていても、依然として有用であることを確認することができる。その他の暗号化アルゴリズムの使用可能であるが、少なくとも、暗号化済データ値の順序は平文値と一貫性があるよう保持されなければならない。

【００２１】

図２に戻ると、暗号化済データセットを信頼のおけない匿名化サイトに送信し（ブロック３３）、そこでデータの処理が行われる（ブロック３４）。具体的には、匿名化サイトが、機密性の値に基づいて匿名化が必要かどうかを判定し、必要な場合、機密性の高い属性のあるデータを保護するために必要な変更の量を決定する。具体的には、ある属性からユーザを特定可能である、あるいは、準識別子として知られている場合その属性は機密として認定される。属性の機密性を判定するために、匿名化サイトは属性を分析し、機密性の値を割り当てるが、これに関しては、２０１５年１１月３日出願の「Ｃｏｍｐｕｔｅｒ－Ｉｍｐｌｅｍｅｎｔｅｄ Ｓｙｓｔｅｍ ａｎｄ Ｍｅｔｈｏｄ ｆｏｒ Ａｕｔｏｍａｔｉｃａｌｌｙ Ｉｄｅｎｔｉｆｙｉｎｇ Ａｔｔｒｉｂｕｔｅｓ ｆｏｒ Ａｎｏｎｙｍｉｚａｔｉｏｎ」と題する米国特許出願第１４／９３１，８０２号明細書（代理人整理番号第０２２．１４５４．ＵＳ．ＵＴＬ）に詳しく記載されている。一実施形態では、属性を重みに関連付けすることができ、この重みが機密性の値を決定するために使用可能となる。どの種類データの匿名化を行うかについては、各属性の機密性の値に基づいてデータの一般化やデータのマスキングなどから決定される。例えば、各機密性の値は０から１の範囲でよく、この場合０は非機密データを表し、１は極めて機密性の高いデータを表す。属性に関する機密性のレベルが非常に高い場合、データ値をマスキングすることが唯

一の選択肢であり得る。しかし、機密性の値が低いかまたは中間の場合、属性を準識別子と認定することができ、その属性だけでは関連する個人を特定することはできないが、１つ以上の準識別子などの他の情報と組み合わせることにより、個人が特定可能となる。例えば、通常、年齢だけでは人物を特定することはできないが、年齢、性別、住所を組み合わせることにより関連する人物が特定される可能性が出てくる。準識別子に関して、データの保護および開示の防止のために、一般化などの匿名化技術が好適であり得る。

【００２２】

データ処理では、一般化の値またはマスクキング値を機密属性に関連するデータ値ごとに決定するが、これに関しては、図４および図１２を参照して後程詳しく説明する。その後、匿名化されたデータ値をデータ所有者に送信することができ（ブロック３５）、データ所有者は、そのデータ値を随意的に第三者に供給することができる。データ受け取り後、データ所有者は、データセット内の平文のデータ値を匿名化サイトからの匿名化された値またはマスキングされた値に置き換えることができる。さらに別の実施形態では、一般化された値に関して、データ所有者は、暗号化済データ値の一般化された範囲を復号化し、各データ値に割り当てられた一般化された範囲を暗号文ではなく平文で供給することができる。例えば、匿名化サイトが、データ所有者に暗号化済データ値を一般化した範囲を送信する。次いで、データ所有者は、匿名化された範囲を復号化して平文値の範囲を取得し、これを要求元の第三者に供給する。この実施形態では、この範囲によりデータ値の一般化が提供され、平文により第三者は一般化された範囲が暗号文で供給される場合よりも数多くデータにアクセス可能となる。

【００２３】

一般化では、データ値のグループ分けを行い、匿名化される値として、グループごとの値の範囲をそのグループ内のデータ値に割り当てる。図４は、一般化によりデータを匿名化するための処理を例として示す流れ図である。匿名化サイトが、１つ以上の機密属性に関する暗号化済データ値を受け取り、単一属性に関する暗号化済データ値を順序付けする（ブロック５１）。次いで、匿名化サイトが、順序付けされたデータを２つ以上のクラスに分割する（ブロック５２）。この分割は、暗号化済データ値の中間でランダムに行われ得る、あるいは、所定のｎの値に基づいて行われ得るが、これに関しては、後程図６を参照して詳細に説明する。あるいは、データ所有者が暗号文の範囲を匿名化サイトに供給し、その範囲に基づいて、匿名化サイトが、暗号化済データ値をグループに分けることもできる。暗号化済データ値がクラスに分割された後、クラスごとの暗号化済データ値の範囲を決定し（ブロック５３）、決定された範囲をその範囲に対応するクラスに属する暗号化された各データ値に、匿名化された値として、割り当てる（ブロック５４）。

【００２４】

例えば、広告調査会社は百貨店と接触して、顧客の年齢、郵便番号、買い物の金額を含む、顧客の消費に関するデータを求める。百貨店は、要求されているデータが維持されているデータセットを特定しアクセスする。図５は、匿名化を行うためのデータセット６０を例として示すブロック図である。データセット６０は、Ｘ軸に沿って表示されている属性６１～６６と、表示されている各属性の下の縦列に並べられているデータ値と、を有するチャート図を含む。これらの属性には、名前６１、苗字６２、年齢６３、口座番号６４、郵便番号６５、および買い物の金額６６が含まれる。データ所有者はデータセットを見て、どの属性の機密性が高いか事前に判定することができる。この例では、データ所有者は、年齢の属性の機密性が若干高く、匿名化が必要ではないかと判定している。年齢の属性に関するデータ値を順序維持暗号化によりそれぞれ暗号化する。さらに別の例では、データ所有者はデータセット内の全ての属性に関するデータ値を暗号化し、どの属性の関連データ値の匿名化が必要かということを判定するよう匿名化サイトに要求する。

【００２５】

データを暗号した後、百貨店は、暗号化済データを匿名化サイトに送信することができる、あるいは、百貨店は暗号化済データ値に直接アクセスするための仮想プライベートネットワーク接続を有する匿名化サイトを設けることができる。匿名化サイトは、年齢の属

性に関する暗号化済データ値を受け取ると、順序付けされた暗号化済データ値を複数のクラスに分割し、クラスのラベルにそのクラス内の暗号化された各データ値を割り当てることにより暗号化済データ値を匿名化する。

【００２６】

暗号化済データ値を分割するために別の方法を行うことも可能性である。図６は、暗号化済データ値を分割するための処理７０を例として示す流れ図である。暗号化済データ値に適応する分割方法を決定し（ブロック７１）、分割停止ポイントを選択する（ブロック７２）。選択された分割方法を用いて、暗号化済データ値を複数のクラスに分割し（ブロック７３）、これを選択された停止ポイントが満たされるまで（ブロック７４）続ける。

【００２７】

別の分割方法には、ランダム分割、ｎ分割、および中間分割が含まれ得るが、この他の分割方法も使用可能である。ランダム分割では、分割停止ポイントに達するまで暗号化済データ値をランダムに分割する。これらの分割とクラスは、均一でも非均一でもよく、匿名化サイトによって決定されるか、またはデータ所有者により依頼される。あるいは、暗号化済データ値は、所定のｎ値および停止ポイントを含む１つ以上の分割パラメータを介して分割することができる。ｎ分割値の場合、ｎ＝２のとき、暗号化済データ値は各分割すなわちデータ値の分割で半分に分割される。さらに、ｎ＝３のとき、暗号化済データ値は、各分割で３つのグループに分割される。ｎ個の分割は、停止ポイントが満たされるまで続けられる。データ所有者、データ所有者に関連する個人または匿名化サイトは、ｎの値を決定し、停止ポイントを選択することができる。最終的に、中間分割では、暗号化済データ値の中間で暗号化済データ値を２つのクラスに分割する（ｎ＝２のときと同様に）。暗号化済データ値の分割は、データ値をすることにより中間で停止ポイントに達するまで続けられる。

【００２８】

一実施形態では、停止ポイントとは、全てのクラスがｋ個の要素を有するとき、ｋがプライバシーの所望のレベルである。別の実施形態では、停止ポイントは、分割を終了させるために満たされなければならない所定の数のクラスである。さらに別の実施形態では、所定の数の分割が完了した時点で分割が終了する。データ所有者または匿名化サイトは、所定の停止ポイントを決定することができる。匿名化サイトは、先のデータセット、属性、および難読化のレベルに基づくなどの特定分野の専門知識（ｄｏｍａｉｎ　ｅｘｐｅｒｔｉｓｅ）を用いて、停止ポイントを算出する。形成されるクラスの数は、データの匿名化の強度に直接関係する。例えば、より少ないクラスが生成されると、各クラス内のデータ値の範囲と数が大きくなるため、匿名性、すなわちクラスのそれぞれの匿名化の値は高くなる。これとは対照的に、生成されるクラスの数が小さいと、匿名化の値の匿名性は低くなる。さらに、分割停止ポイントは、第三者が必要とするデータの特殊性に依存し得る。より特殊なデータが必要な場合、特殊性の低いデータを必要としている場合に比べて、クラスの数をより多く設定することができる。

【００２９】

分割は順序付けされた暗号化済データ値のリスト上で行われ得るか、あるいは順序付けされた暗号化済データ値の木を介して行われ得る。図５のデータセットを参照する上記の例に戻ると、匿名化に対して年齢の属性が選択されている。図７は、単一属性に関するデータ値を匿名化するための処理を例として示す流れ図である。年齢の属性８１では、百貨店の顧客の個々の年齢がデータ値として表示されている。その後、データ値を、順序維持暗号化により、個々に暗号化して、暗号化済データ値８２を生成する。暗号化済データ値８２を、例えば、昇順あるいは降順で順序付けする（８３）。次に、順序付けされた暗号化済データ値を８４ａ～８４ｂに分割する。

【００３０】

一実施形態では、順序付けされた暗号化済データ値でリストを形成することができ、このリストで分割を行う。分割パラメータ（停止ポイントとしても知られている）を供給して分割を行う。この分割パラメータには、例えば、ｎ分割値（作成された複数のクラスを

表す）、暗号化済データ値が分割されるクラスごとのｋ個の要素またはデータ値が含まれる。ｎ分割値はｎ＝３と設定され、データ値の数はｋ＝２と設定される。ｋ値は、最小の数のデータ値であり、ｎ分割すなわちクラスはいずれも、最小のｋの数以上のデータ値を含み得る。

【００３１】

順序付けされたリストに関して、第１の分割処理で暗号化済データ値を２つのクラスに分割して、暗号化済データ値０８５７、１０５６、２７６４、および４７９８が一方のクラスに入り、暗号化済データ値６３２１、７７４４、８７９１、および９９２１がもう一方のクラスに入るようにする。ｎ＝３のため、３つのクラスを作成するためにさらなる分割処理が必要となる。したがって、最も低く順序付けされた暗号化値を有するクラスをさらに分割して、０８５７～１０５６で第１のクラスを形成し、２７６４～４７９８で第２のクラスを形成し、６３２１～９９２１で第３のクラスを形成するようにする。さらに別の実施形態では、最も低く順序付けされた値ではなく、最も高く順序付けされた暗号化値を分割することもできる。分割の順序は、所定のものでも、ユーザ入力によるものでも、自動的に決定されるものでもよい。

【００３２】

さらに別の実施形態では、ｎの値により分割処理の数を表すことができる、すなわち分割処理は、最終的な分割またはクラスの数ではなくデータ値に適用される。例えば、ｎにより分割処理の数が３と表される場合、第１の分割処理を行って、０８５７、１０５６、２７６４、４７９８が一方のクラスで、暗号化済データ値６３２１、７７４４、８７９１、９９２１がもう一方のクラスとなるよう、２つのクラスを作成する。次いで、最も低く順序された値に第２の分割処理を行って、０８５７と１０５６が一方のクラスとなり、２７６４と４７９８がもう一方のクラスとなるようにする。最終的に、第３の分割処置を行って、６３２１と７７４４が一方のクラスとなり、８７９１と９９２１がもう一方のクラスとなるようにする。したがって、３つの分割処理が行われると、４つのクラスが生成される。

【００３３】

さらに別の実施形態では、木を用いて属性に関する暗号化済データ値を分割することができる。図８は、暗号化された単一の属性のデータ値に関する木９０を例として示すブロック図である。この木９０の頂点のノード９６により順序付けされた、暗号化済データ値が示されている。第１の分割処理９１により、暗号化済データ値０８５７、１０５６、２７６４、４７９８が一方のグループ９７になり、暗号化済データ値６３２１、７７４４、８７９１、９９２１がもう一方のグループ９５になるよう、順序付けされた暗号化済データを２つのグループに分割する。所定のクラスの数が３であるため、分割処理９２がさらに必要となる。したがって、低い値のグループをさらに２つのグループに分割して、０８５７と１０５６が一方のグループ９３に、２７６４と４７９８が異なるグループ９４になるようにし、合計で３つの暗号化済データ値のクラス９３～９５を形成する。

【００３４】

図７に戻って議論すると、分割処理終了後、各クラスに関する範囲８５を決定する。各クラスで暗号化済データ値を昇順に並べて、割り当てる範囲を形成することができるか、あるいは、別の範囲と重ならなければ、これらの範囲を拡げて追加のデータ値を含ませることができる。例えば、クラスＩに関する範囲は、データ値に基づいて０８５７～１０５６でよいが、この範囲は、１０５６以下または２７００以下でもよい。その他の範囲も可能である。クラスＩＩの暗号化済データ値の範囲は２７６４～４７９８であり、クラスＩＩＩの暗号化済データ値の範囲は６３２１～９９２１である。しかし、これらの範囲はこれより広くてもよい、例えば、クラスＩＩでは２７０５～６３２０で、クラスＩＩＩでは６３２１以上でもよい。その他の範囲も可能である。範囲決定の方法は、所定のものでも、自動的に決定されるものでも、ユーザが選択するものでもよい。

【００３５】

範囲を決定した後、各クラスの範囲を匿名化された値として、そのクラスに所属する暗

号化された各データ値に割り当てる。したがって、年齢６２は、暗号化済データ値８７９１を有し、クラスＩＩＩに属する。範囲６３２１～９９２１は、匿名化される値として年齢６２に割り当てられる。さらに、年齢２７は、暗号化された値１０５６を有しクラスＩに属する、したがって、範囲０８５７～１０５６は、匿名化されたデータ値８６として年齢２７に割り当てられる。

【００３６】

１つ以上の属性に関するデータ値が匿名化された後、データセットを第三者に供給して、処理または分析を行うことができる。順序維持暗号化により、データセットの暗号化済データ値の範囲に基づいて、データセット内の若者、お年寄り、および中年の人口を特定するなど、第三者はリサーチを行うためにこのデータを用いることができる。例えば、低く順序付けされた範囲は低い年齢に関連し、高く順序付けされた範囲は高い年齢に関連する。

【００３７】

さらに別の実施形態では、木に基づいて、これらのデータ値を下から上にクラス分けすることができ、ｎ個のクラスおよび各クラスのｋ個のデータ値が確実に満たされるには、木の終端ノードをどれだけ多く組み合わせるべきかを決定する。

【００３８】

さらに別の実施形態では、２つ以上の属性に関するデータ値を一緒に匿名化して、属性のデータ値間の関係が確実に維持されるようにする。図９は、２つ以上の属性のデータを一緒に匿名化するための処理１００を例として示す流れ図である。匿名化するための２つの属性を選択する。機密性の値、データ値の内容または重要性、およびその他の要因に基づいて、匿名化する属性を選択することができる。例えば、順序維持暗号化を用いて、２つの属性のデータ値を個々に暗号化する（ブロック１０１）。次に、匿名化処理を行うための属性のうちの１つを選択する（ブロック１０２）。可能性のある選択の方法は、いくつか存在する。例えば、機密性の値に基づく場合、機密性が最も高い属性を第１の属性として選択することができ、またその逆も同様である。属性を選択する別の方法として、その属性に関する一意の値の数に基づいて、ランダムに選択することも可能である。しかし、その他の方法で選択することも可能である。次いで、選択された属性のデータ値は、複数のクラスに分割され（１０３）、この分割処理はｎ個の最終クラスの値が満たされるなどの停止ポイントの条件が満たされるまで繰り返される。属性データの分割処理に基づいて、第１の選択属性に関する暗号化済データ値からクラスを形成し（ブロック１０４）、各クラスに関する範囲を決定する。停止ポイントが満たされ、それ以上データ値を分割しなくてもよいとき、各クラスに関する範囲を決定し、匿名化済データ値として、そのクラス内に含まれる暗号化済データ値に割り当てる（ブロック１０５）。次に、それ以外に処理する属性が残っているかどうかを判定し（ブロック１０６）、残っている場合、その属性を選択し（ブロック１０２）、分割して（ブロック１０３）、その分割処理から生成されたクラスにラベルを割り当てる（ブロック１０４）。しかし、残っていない場合、２つ以上の属性に関して生成されたクラスを、組み合わされたクラスにグループ分けし（ブロック１０７）、組み合わされた属性値の匿名化済の値に対しラベルを生成するが、これに関しては、後程図１０を参照してさらに詳しく説明する。匿名化の値を決定した後、要求元の第三者にデータを供給する前に、データセット内の組み合わされた属性のデータ値ごとに平文のデータ値を決定された匿名化の値に置き換えることができる。

【００３９】

２つ以上の属性を組み合わせるクラスのグループ分けは、ランダムに行うことも、所定のベースで行うことも、あるいはユーザが命令して行うこともできる。図１０は、２つの異なる属性に関するクラスのチャート図１２０を示すブロック図である。属性１などが属する１つの属性１２１に関するクラスが、チャート図の上段の横列に沿って示されており、属性２などのその他の属性１２２のクラスがチャート図１２０の縦列に沿って示されている。異なる属性に関するクラスのグループ分けは、所定のｎ個の分割値に基づく。この例では、ｎ＝３であり、したがって、２つの属性に関するクラスは、組み合わされて３つ

のグループにならなければならない。この分割は、ランダムに行うことも、所定のベースで行うことも、あるいはユーザが命令して行うこともできる。図１０で示される通り、この例では、全部で３つのグループが形成される。これはｎ色を有する表のセルを書くことと同等である。なおｎはグループまたは所望の分割の数である。その後、最終的なグループに基づいて、匿名化の値を各グループに割り当てる。上記に説明した実施形態と同様に、順序維持暗号化をグループの指数またはグループの指数の数学的関数に適用することにより、ユーザは匿名化に関する値を生成することができる。

【００４０】

連結匿名化は、第三者が複数の属性を満たす個人を特定しようとするときに便利である。例えば、ある調査員が病院からデータを取得して、年齢６２～８０才のカリフォルニア州に在住のアルツハイマー患者の数を特定しようとする場合。この例では、病状に関するデータ値を平文のまま残し、年齢と郵便番号の値を匿名化することができる。一実施形態では、最初に年齢を匿名化し、各年齢の値に関連する郵便番号を匿名化して、それぞれ特定された郵便番号の地域に住む年齢６２～８１才のアルツハイマー患者を特定する。次いで、カリフォルニア州の各郵便番号の地域の患者の数を合計することにより、カリフォルニア州在住の年齢が６２～８１才のアルツハイマー患者の総数を特定する。さらに別の実施形態では、最初に年齢に関するデータ値を匿名化し、次いで、各郵便番号の地域に関連する州を匿名化して、カリフォルニア州在住の年齢が６２～８１才のアルツハイマー患者と、カリフォルニア州在住でない年齢が６２～８１才のアルツハイマー患者との対比を特定する。

【００４１】

さらに、上記の百貨店の例に戻ると、第三者のデータ要求者が、平均的な顧客の消費を判定しようとする、あるいは、最もお金を使う顧客の年齢層、最も頻繁に訪れる年齢層、その百貨店が発行するクレジットカードを所持する顧客の年齢層を特定しようとする。図５に示されるように、この例では、第三者が、匿名化された年齢のデータと買い物の金額の平文のデータを組み合わせて、年齢の高い個人と年齢の若い個人のどちらがより多くお金を使うかを特定することができる。さらに別の例では、第三者は、最もお金を使う個人が住む地域も特定しようとする。要求されたデータを第三者に供給するには、匿名化する間、データ値同士の関係は維持しなければならない。上記の通り、匿名化の値を第１の属性に関する暗号化済データ値のグループに割り当て、第２の属性を分離し、第１の属性のクラスと第２の属性のクラスにグループ分けすることにより、この関係を維持することができる。

【００４２】

データ値を決定し割り当てた後、データセット内の機密性の高い属性の平文データ値を匿名化の値と置き換える。図１１は、匿名化される値を有する図５のデータセットを例として示すブロック図である。図５に示される通り、このデータセットには、名前、苗字、年齢、口座番号、郵便番号、および買い物の金額に関する属性６１が含まれる。各属性は、データセットにより示される個人に関する平文のデータ値６２に関連する。データセット内のデータ値により表される特定の個人を信頼のおけない第三者が特定しないよう、年齢の属性と郵便番号の属性に関する平文のデータ値は、匿名化されたデータ値６３に置き換えられている。例えば、図７および図８を参照して上記に説明した通り、年齢の属性に関して匿名化されたデータ値を決定する。

【００４３】

このデータセット内では、各個人に関して名前と苗字が供給されているため、依然として、信頼のおけない第三者が個人を特定する可能性がある。個人の身元を保護するために、名前と苗字に関するデータ値をマスキングすることができる。データの機密性が高い場合、あるいは、州や連邦の規制によりマスキングが義務付けられている場合には、特定の属性に対してマスキングが必要となり得る。その他の選択肢としては、単純にそれらの属性を隠すことが挙げられる。図１２は、データ値をマスキングするための処理１４０を例として示す流れ図である。機密性が高いか、あるいは州や連邦の規制によりマスキングが

必要とされているかをデータセット内の少なくとも１つの属性に対して決定する（ブロック１４１）。一例を挙げてみると、データ所有者は、医療保険の相互運用性と説明責任に関する法令（Ｈｅａｌｔｈ　Ｉｎｓｕｒａｎｃｅ　Ｐｏｒｔａｂｉｌｉｔｙ　ａｎｄ　Ａｃｃｏｕｎｔａｂｉｌｉｔｙ　Ａｃｔ（ＨＩＰＡＡ））に該当するデータを含むデータセットを匿名化するよう匿名化サイトに命令する。このＨＩＰＡＡでは、特定の属性に関するデータ値を完全に難読化するよう求められている。匿名化サイトはＨＩＰＡＡの規制を用いてプログラムされ、ＨＩＰＡＡの要求事項に基づいてマスキングが必要な属性を特定する（ブロック１４１）。

【００４４】

機密性の高い属性に関するデータ値を選択し（ブロック１４２）、選択されたデータ値にマスキング値を割り当てることにより、ランダムにマスキングを行う（ブロック１４３）。マスキングでは、置き換えられるデータ値と類似するその他の値とデータ値を置き換えることができ（ブロック１４４）、データ値の置き換えでは、暗号学ハッシュ関数または暗号化などの擬似乱数関数が用いられる。その他の種類のマスキングも使用可能である。その属性に関するマスキングされていないデータ値が残っている場合（ブロック１４５）、マスキングをするために別のデータ値を選択する（ブロック１４２）。

【００４５】

データセット内で匿名化されたデータとして、マスキングを施されたデータ値を機密性の高い属性に関する平文のデータ値と置き換える。図１３は、２つの属性に関するマスキングを施されたデータ値を有する図１１のデータセット１５０を例として示すブロック図である。このデータセット１５０では、名前と苗字に関する属性１５１が、機密性の高い属性として特定され、データ値は、プライバシーの機密性に従って処理されている。具体的には、名前と苗字の属性に関するデータ値は、ハッシングによりマスキングを施される。さらに、準識別子と見なされる年齢の属性と郵便番号の属性に関するデータ値が、一般化により匿名化されている。データセット内に残る平文のデータ値は、平文のまま残される、あるいは、そのデータセットを第三者に供給する前に暗号化することもできる。

【００４６】

一般化と同様に、データ所有者は、機密性の高い属性を暗号化し、匿名化サイトにマスキングを施すよう依頼することができる。この場合、匿名化サイトは、暗号化済データ上で処理を行う。

【図１】

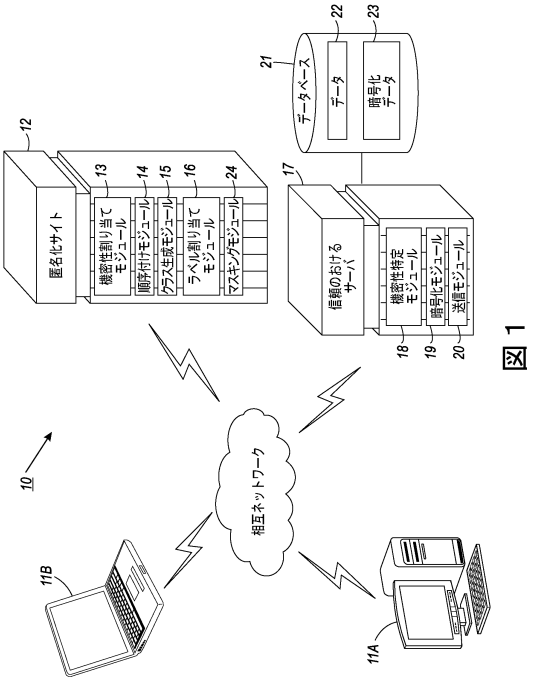

図１

【図２】

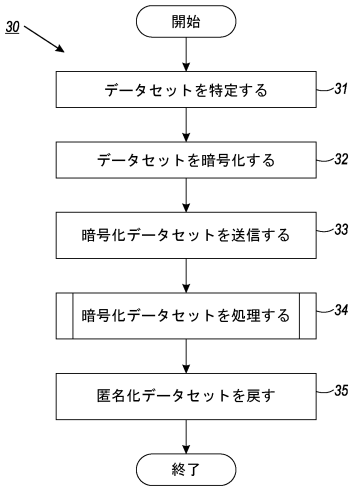

図２

【図３】

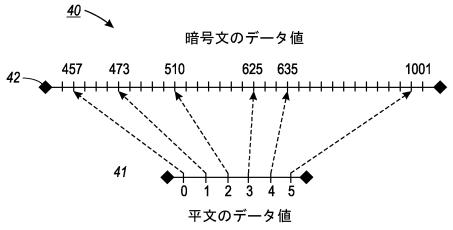

図３

【図４】

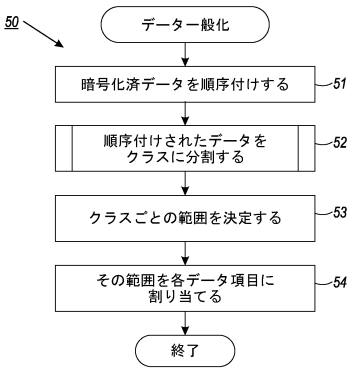

図４

【図５】

60

61 名前	62 苗字	63 年齢	64 口座番号	65 郵便番号	66 買い物の金額
Betty	Wu	62	14578	21768	$ 1,057.21
John	Ireland	68	16743	01764	$ 4,109.15
Maurice	Thomas	27	09673	94602	$ 11,012.17
Russell	Smith	36	37810	98125	$ 19,429.08
Susan	Yamamoto	45	41137	14557	$ 7,120.54
Marshawn	Sherman	43	28759	98109	$ 3,235.07
Joe	Lee	52	84221	57214	$ 2,250.14
Braylon	Wilson	22	02356	78265	$ 9,283.16

図５

【図６】

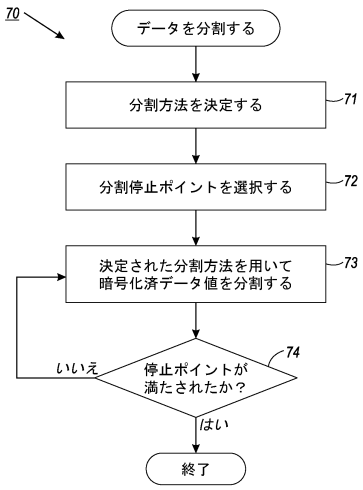

図６

【図７】

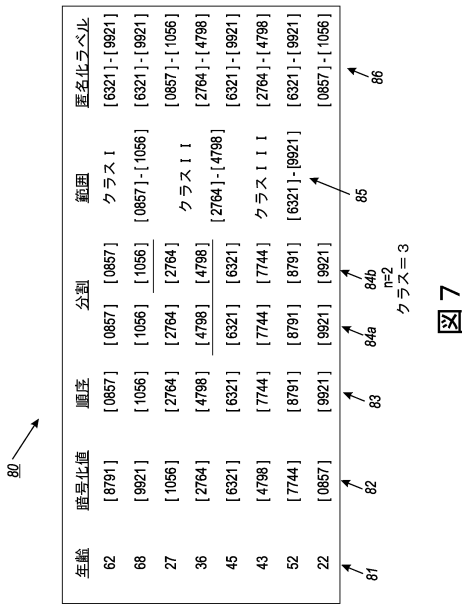

図７

【図８】

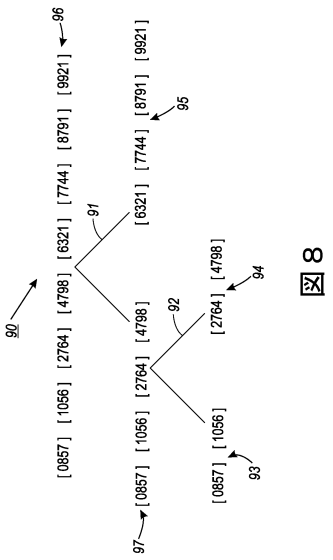

図８

【図９】

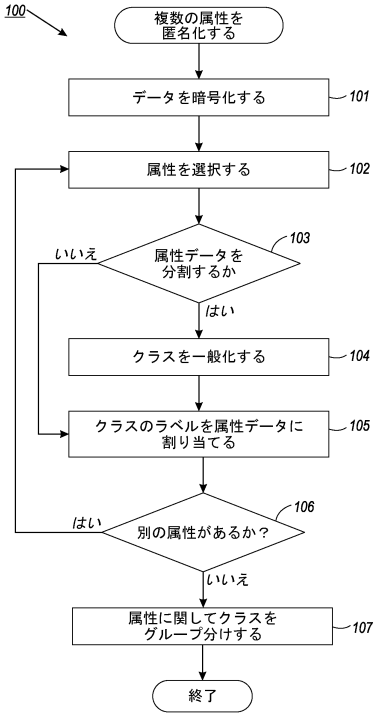

図９

【図１０】

図１０

【図１１】

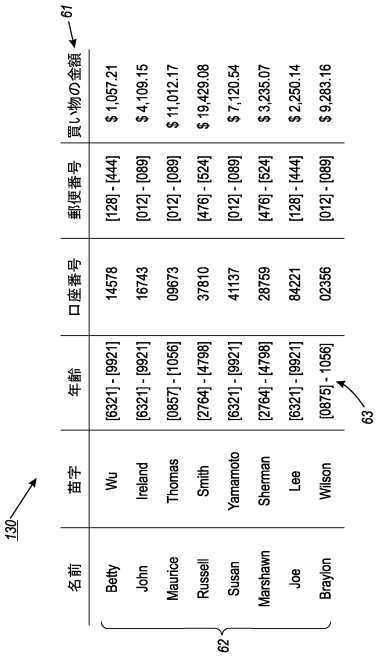

図１１

【図１２】

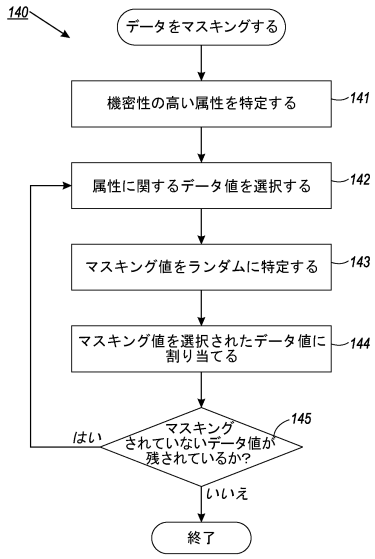

図１２

【図１３】

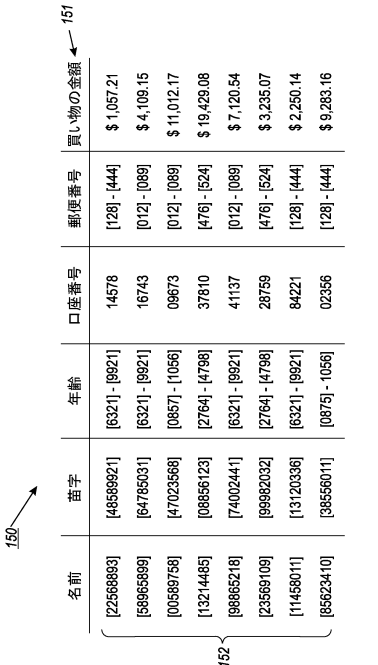

図１３

フロントページの続き

(74)代理人 100109070
弁理士 須田 洋之
(74)代理人 100109335
弁理士 上杉 浩
(74)代理人 100120525
弁理士 近藤 直樹
(74)代理人 100158551
弁理士 山崎 貴明
(72)発明者 ジュリアン・フロイトガー
アメリカ合衆国 カリフォルニア州 ９４０４３ マウンテン・ビュー ロック・ストリート ２３０９
(72)発明者 アレハンドロ・イー・ブリトー
アメリカ合衆国 カリフォルニア州 ９４０４０ マウンテン・ビュー オルテガ・アベニュー １６３
(72)発明者 シャンタヌ・レイン
アメリカ合衆国 カリフォルニア州 ９４０２５ メンロー・パーク シャロン・パーク・ドライブ ６７５ アパートメント ２０１
(72)発明者 アーシン・ウズン
アメリカ合衆国 カリフォルニア州 ９５００８ キャンベル カプリ・ドライブ １１８６

審査官 中里 裕正

(56)参考文献 特開２００６－０４３０１（ＪＰ，Ａ）
米国特許出願公開第２００７／２３９７０５（ＵＳ，Ａ１）
伊沢亮一，森井昌克，順序性を保持した匿名化方式の提案，電子情報通信学会技術研究報告，２０１１年 3月１８日，第110巻 第475号，pp.5-10

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ ２１／６２
ＪＳＴＰｌｕｓ／ＪＭＥＤＰｌｕｓ／ＪＳＴ７５８０（ＪＤｒｅａｍＩＩＩ）
ＩＥＥＥ Ｘｐｌｏｒｅ