(19) Deutsches Patent- und Markenamt

(10) **DE 11 2018 005 725 T5** 2020.08.20

(12) # Veröffentlichung

der internationalen Anmeldung mit der
(87) Veröffentlichungs-Nr.: **WO 2019/116137**
in der deutschen Übersetzung (Art. III § 8 Abs. 2 IntPatÜG)
(21) Deutsches Aktenzeichen: **11 2018 005 725.9**
(86) PCT-Aktenzeichen: **PCT/IB2018/059453**
(86) PCT-Anmeldetag: **29.11.2018**
(87) PCT-Veröffentlichungstag: **20.06.2019**
(43) Veröffentlichungstag der PCT Anmeldung in deutscher Übersetzung: **20.08.2020**

(51) Int Cl.: ***G06F 21/60*** (2013.01)
G06K 9/00 (2006.01)

(30) Unionspriorität:
15/843,049 **15.12.2017** **US**

(71) Anmelder:
International Business Machines Corporation, Armonk, N.Y., US

(74) Vertreter:
Richardt Patentanwälte PartG mbB, 65185 Wiesbaden, DE

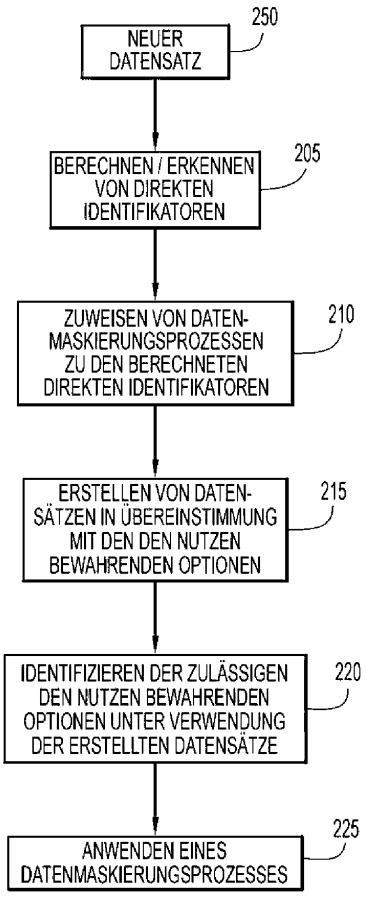
(72) Erfinder:
Gkoulalas-Divanis, Aris, Cambridge, MA, US

Prüfungsantrag gemäß § 44 PatG ist gestellt.

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen.

(54) Bezeichnung: **DATEN-DEIDENTIFIKATION AUF DER GRUNDLAGE EINES ERKENNENS VON ZULÄSSIGEN KONFIGURATIONEN FÜR DATEN-DEIDENTIFIKATIONSPROZESSE**

(57) Zusammenfassung: Ein System zum Deidentifizieren von Daten ermittelt einen oder mehrere Identifikatoren, die eine Entität eines Datensatzes identifizieren. Ein oder mehrere Daten-Deidentifikationsprozesse werden identifiziert und dem einen oder den mehreren ermittelten Identifikatoren zugewiesen. Jedem Daten-Deidentifikationsprozess werden ein oder mehrere Sätze von Konfigurationsoptionen zugewiesen, die Informationen angeben, die in dem Datensatz zu bewahren sind. Die identifizierten Daten-Deidentifikationsprozesse werden an dem Datensatz gemäß den zugehörigen Sätzen von Konfigurationsoptionen ausgeführt, um Datensätze mit unterschiedlichen bewahrten Informationen zu erstellen. Die erstellten Datensätze werden auf Datenschutz-Schwachstellen hin ausgewertet, und auf der Grundlage der Auswertung werden ein Daten-Deidentifikationsprozess und ein zugehöriger Satz von Konfigurationsoptionen ausgewählt. Der ausgewählte Daten-Deidentifikationsprozess wird an dem Datensatz gemäß dem zugehörigen Satz von Konfigurationsoptionen ausgeführt, um einen sich ergebenden deidentifizierten Datensatz zu erzeugen. Zu Ausführungsformen gehören ein Verfahren und ein Computerprogrammprodukt zum Deidentifizieren von Daten in im Wesentlichen derselben Weise wie oben beschrieben.



Beschreibung

TECHNISCHES GEBIET

[0001] Ausführungsformen der vorliegenden Erfindung betreffen ein Zugreifen auf Daten, und insbesondere ein Deidentifizieren von Daten auf der Grundlage eines Erkennens von zulässigen Konfigurationen für Daten-Deidentifikationsprozesse, die deidentifizierte Datensätze unter Beibehaltung von Datenschutz und Datennutzen erzeugen.

HINTERGRUND

[0002] Ein Prozess zum Veröffentlichen von Daten unter Wahrung des Datenschutzes besteht aus mehreren Schritten, darunter: Entdecken von direkten Identifikatoren; Maskieren von direkten Identifikatoren; Entdecken von Quasi-Identifikatoren (QIDs, quasiidentifiers); Schützen von Quasi-Identifikatoren durch Datenanonymisierungstechniken; und Datenfreigabe und Berichtserstellung. Bei direkten Identifikatoren handelt es sich um Attribute, die allein zum direkten und eindeutigen Identifizieren einer Entität verwendet werden können, während es sich bei Quasi-Identifikatoren um Gruppen von Attributen handelt, die gemeinsam zum eindeutigen Identifizieren einer Entität verwendet werden können. Die Koordination der verschiedenen Schritte in dem obigen Prozess steuert, ob ein ausreichend gut anonymisierter Datensatz wiedergegeben wird.

[0003] Der Schutz von direkten Identifikatoren in einem Datensatz wird durch Datenmaskierungsvorgänge durchgeführt. Diese Vorgänge wandeln die ursprünglichen Datenwerte in neue, fiktionalisierte Datenwerte um, die nicht mehr zum Identifizieren der entsprechenden Entitäten verwendet werden können, während sie auch speziell so gestaltet sein können, dass bestimmte Informationen der ursprünglichen Datenwerte bewahrt bleiben, wodurch ein Beibehalten eines gewissen Datennutzens in dem Datensatz ermöglicht wird. Zum Beispiel: Ein Personenname kann maskiert oder durch einen fiktiven Namen ersetzt werden, der eine Übereinstimmung mit den Geschlechtsangaben für die Person beibehält; eine eMail-Adresse kann maskiert oder durch eine andere eMail-Adresse ersetzt werden, welche die Domänen-Namensangaben der ursprünglichen eMail-Adresse beibehält; eine Kreditkartennummer kann maskiert oder durch eine andere Kreditkartennummer ersetzt werden, die Kreditkarten-Herausgeber-Angaben der ursprünglichen Kreditkartennummer widerspiegelt; eine Telefon- und/oder Faxnummer kann maskiert oder durch eine andere Telefon- und/oder Faxnummer ersetzt werden, welche die Landesvorwahl und/oder die Ortsvorwahl der ursprünglichen Telefon- und/oder Faxnummer enthält; Postleitzahlen, Städte, Landkreise, Länder und Kontinente können in einer Weise maskiert werden, welche die räumliche Nähe zu dem ursprünglichen Standort beibehält (d.h. eine geografische Korrelation zu ursprünglichen Werten); und ein Datum, das sich auf eine Person bezieht, kann maskiert oder durch ein anderes Datum innerhalb der Kalenderwoche und des Jahres, des Monats und des Jahres, des Quartals und des Jahres oder des Jahres des ursprünglichen Datums ersetzt werden, wodurch wichtige Informationen beibehalten bleiben, die für bestimmte Arten von nachfolgenden Datenanalysen wie zum Beispiel in mehreren medizinischen Fallstudien sehr nützlich sein könnten.

[0004] Das Schützen von Quasi-Identifikatoren in dem Datensatz wird üblicherweise durch Datenverallgemeinerungs- oder Datenunterdrückungsvorgänge durchgeführt. Üblicherweise werden beim Veröffentlichen von Daten unter Wahrung des Datenschutzes das Schützen von direkten Identifikatoren und das Schützen von Quasi-Identifikatoren getrennt voneinander durchgeführt. Das Schützen von direkten Identifikatoren wird mit minimaler oder gar keiner Wahrung des Nutzens durchgeführt (z.B. Ersetzen durch fiktive Werte, die keine Informationen über die ursprünglichen Datenwerte beibehalten) und beruht vollständig auf Entscheidungen von Datenexperten/Dateneignern. In derartigen Fällen muss ein Datenexperte/Dateneigner entscheiden, wie die direkten Identifikatoren in dem Datensatz in einer Weise maskiert werden können, dass der sich ergebende Datensatz ausreichend gegen Angriffe auf den Datenschutz geschützt ist, wie zum Beispiel eine erneute Identifikation des Betroffenen, Offenlegungen von vertraulichen Daten, Offenlegungen von Mitgliedschaften, Offenlegungen von Rückschlüssen usw. Ein Problem betrifft mögliche Konflikte zwischen Optionen zum Wahren des Nutzens, die für das Maskieren von direkten Identifikatoren ausgewählt werden, und Optionen, die zum Schützen von Quasi-Identifikatoren durch Datenverallgemeinerungstechniken ausgewählt werden.

[0005] Der Nutzen (oder die Informationen), der bzw. die beim Umwandeln (oder Maskieren) bestimmter direkter Identifikatoren erhalten bleibt bzw. bleiben, kann/können immer noch eine Datenschutzverletzung ermöglichen, wenn neue Werte der direkten Identifikatoren zusammen mit den verallgemeinerten (neuen) Werten der Quasi-Identifikatoren betrachtet werden.

KURZDARSTELLUNG

[0006] Gemäß einer Ausführungsform der vorliegenden Erfindung deidentifiziert ein System Daten und weist mindestens einen Prozessor auf. Das System ermittelt einen oder mehrere Identifikatoren, die eine Entität eines Datensatzes identifizieren. Ein oder mehrere Daten-Deidentifikationsprozesse werden identifiziert und dem einen oder den mehreren

ermittelten Identifikatoren zugewiesen. Jedem Daten-Deidentifikationsprozess werden ein oder mehrere Sätzen von Konfigurationsoptionen zugewiesen, die Informationen angeben, die in dem Datensatz zu bewahren sind. Die identifizierten Daten-Deidentifikationsprozesse werden an dem Datensatz gemäß den zugehörigen Sätzen von Konfigurationsoptionen ausgeführt, um Datensätze mit unterschiedlichen bewahrten Informationen zu erstellen. Die erstellten Datensätze werden auf Datenschutz-Schwachstellen hin ausgewertet, und auf der Grundlage der Auswertung werden ein Daten-Deidentifikationsprozess und ein zugehöriger Satz von Konfigurationsoptionen ausgewählt. Der ausgewählte Daten-Deidentifikationsprozess wird an dem Datensatz gemäß dem zugehörigen Satz von Konfigurationsoptionen ausgeführt, um einen sich ergebenden deidentifizierten Datensatz zu erzeugen. Zu Ausführungsformen der vorliegenden Erfindung gehören darüber hinaus ein Verfahren und ein Computerprogrammprodukt zum Deidentifizieren von Daten in im Wesentlichen derselben Weise wie oben beschrieben.

[0007] Ausführungsformen der vorliegenden Erfindung verkürzen die Verarbeitungszeit, indem sie durchführbare und/oder optimale Konfigurationen für Daten-Deidentifikationsprozesse identifizieren, anstatt Ansätze des systematischen Ausprobierens einzusetzen, um Daten-Deidentifikationsprozesse zum Deidentifizieren von Daten auszuwählen. Diese Auswahlen des systematischen Ausprobierens beruhen im Allgemeinen auf dem Wissen eines Benutzers und können zu einer suboptimalen Daten-Deidentifikation und zahlreichen Daten-Deidentifikationsversuchen führen, wodurch Verarbeitungs- und andere Ressourcen verschwendet werden.

[0008] Eine Ausführungsform der vorliegenden Erfindung kann darüber hinaus Datensätze zur Auswertung in Form einer Tabelle erstellen und zwei oder mehr Spalten eines erstellten Datensatzes zusammenfassen, um eine Spalte mit Informationen zu erzeugen, die spezifischer sind als die zwei oder mehr Spalten. Auf diese Weise kann ein Datensatz mit spezifischeren Informationen ausgewertet werden, um eine Abwesenheit einer Datenschutz-Schwachstelle sicherzustellen. Wenn der erstellte Datensatz mit spezifischeren Informationen keine Datenschutz-Schwachstelle aufweist, weisen auch andere Datensätze, die aus dem entsprechenden Daten-Deidentifikationsprozess und den Konfigurationsoptionen mit verallgemeinerteren Informationen erstellt wurden (z.B. Datensätze mit einer oder mehreren der ursprünglichen nicht zusammengefassten Spalten) keine Datenschutz-Schwachstelle auf. Dies verkürzt auch die Verarbeitungszeit, indem für Datensätze mit den spezifischeren und verallgemeinerten Informationen eine einzelne Auswertung anstelle von mehreren Auswertungen genutzt wird.

[0009] Eine Ausführungsform der vorliegenden Erfindung kann einen erstellten Datensatz auf Datenschutz-Schwachstellen hin auswerten, indem ein Vorhandensein einer Verknüpfung zwischen Daten für eine Entität in einem erstellten Datensatz und Daten für eine bekannte Entität in einem öffentlich verfügbaren Datensatz ermittelt wird, um auf eine Datenschutz-Schwachstelle für den erstellten Datensatz hinzuweisen. Bei dieser Auswertung werden die deidentifizierten Daten aus einem erstellten Datensatz genutzt und mit bekannten Entitäten in einem öffentlich verfügbaren Datensatz verglichen, um zu ermitteln, ob Identifikatoren von Entitäten in den deidentifizierten Daten durch Triangulationsangriffe ermittelt werden können, wodurch eine erhebliche Zuversicht bereitgestellt wird, dass ein empfohlener Daten-Deidentifikationsprozess mit zugehörigen Konfigurationsoptionen den Datenschutz beibehält.

[0010] Eine Ausführungsform der vorliegenden Erfindung kann einen erstellten Datensatz auf Datenschutz-Schwachstellen hin auswerten, indem ein Vorhandensein eines Satzes von Quasi-Identifikatoren in einem erstellten Datensatz ermittelt wird, der durch einen entsprechenden Daten-Deidentifikationsprozess und einen zugehörigen Satz von Konfigurationsoptionen eingebracht wurde, um auf eine Datenschutz-Schwachstelle für den erstellten Datensatz hinzuweisen. Diese Auswertung, die auf Eindeutigkeitskriterien beruht, und stellt sicher, dass keine Quasi-Identifikatoren durch einen Daten-Deidentifikationsprozess und zugehörige Konfigurationsoptionen eingebracht werden, wodurch eine erhebliche Zuversicht bereitgestellt wird, dass ein empfohlener Daten-Deidentifikationsprozess mit zugehörigen Konfigurationsoptionen den Datenschutz beibehält. Wenn der erstellte Datensatz keine Eindeutigkeiten oder Ausreißer enthält, kann er durch Triangulationsangriffe nicht mit irgendwelchen anderen (internen oder externen) Datensätzen verknüpft werden, wodurch der Datenschutz beibehalten wird.

[0011] Eine Ausführungsform der vorliegenden Erfindung kann die Verarbeitungszeit für die Deidentifikation des ursprünglichen Datensatzes verkürzen, indem ein erstellter Datensatz identifiziert wird, der keine Datenschutz-Schwachstellen aufweist, und eine Verarbeitung in Bezug auf andere zugehörige Sätze von Konfigurationsoptionen für einen entsprechenden Daten-Deidentifikationsprozess beendet wird, der einen Datensatz mit verallgemeinerteren Informationen als der identifizierte erstellte Datensatz erstellt. Dies verbessert die Leistung der Prozessoren erheblich und stellt eine optimale Daten-Deidentifikation in einer verkürzten Verarbeitungszeit bereit.

Figurenliste

[0012] Im Allgemeinen werden in den verschiedenen Figuren gleichartige Bezugszahlen verwendet, um gleichartige Komponenten zu kennzeichnen.

Fig. 1 ist eine grafische Darstellung einer beispielhaften Datenverarbeitungsumgebung einer Ausführungsform der vorliegenden Erfindung.

Fig. 2 ist ein verfahrensmäßiger Ablaufplan, der eine Art und Weise des Erkennens von zulässigen Konfigurationsoptionen für Daten-Deidentifikationsprozesse zum Erzeugen von Datensätzen unter Wahrung von Datenschutz gemäß einer Ausführungsform der vorliegenden Erfindung veranschaulicht.

Fig. 3 ist ein verfahrensmäßiger Ablaufplan einer Art und Weise des Erstellens von Datensätzen gemäß Konfigurationsoptionen von Daten-Deidentifikationsprozessen gemäß einer Ausführungsform der vorliegenden Erfindung.

Fig. 4 ist ein verfahrensmäßiger Ablaufplan einer Art und Weise des Auswertens von Konfigurationsoptionen für Daten-Deidentifikationsprozesse auf der Grundlage von öffentlich verfügbaren Daten gemäß einer Ausführungsform der vorliegenden Erfindung.

Fig. 5 ist ein verfahrensmäßiger Ablaufplan einer Art und Weise des Auswertens von Konfigurationsoptionen für Daten-Deidentifikationsprozesse auf der Grundlage eines Einbringens von Quasi-Identifikatoren in deidentifizierten Daten gemäß einer Ausführungsform der vorliegenden Erfindung.

Fig. 6 ist ein beispielhafter Datensatz, der durch einen Daten-Deidentifikationsprozess erstellt wurde, der zum Deidentifizieren eines Namensattributs unter Wahrung von Geschlechtsangaben konfiguriert wurde.

Fig. 7 ist ein beispielhafter Datensatz, der durch einen Daten-Deidentifikationsprozess erstellt wurde, der zum Deidentifizieren eines Adressattributs unter Wahrung von räumlicher Nähe konfiguriert wurde.

Fig. 8 ist eine schematische Veranschaulichung einer beispielhaften Baumstruktur, die zum Steuern des Verarbeitens von Daten-Deidentifikationsprozessen zum Verkürzen einer Verarbeitungszeit gemäß einer Ausführungsform der vorliegenden Erfindung genutzt wird.

AUSFÜHRLICHE BESCHREIBUNG

[0013] Ausführungsformen der vorliegenden Erfindung werten das Datenschutzrisiko jedes verfügbaren Satzes von Konfigurationsoptionen eines Daten-Deidentifikationsprozesses oder einer Daten-Deidentifikationstechnik aus und ermöglichen, dass ausschließlich diejenigen Konfigurationsoptionen (oder Einstellungen) verwendet werden, die Datenschutz-Schwachstellen in den Daten verhindern. Ausführungsformen der vorliegenden Erfindung analysieren einen Datensatz, um zulässige Konfigurationsoptionen (oder Einstellungen) für Daten-Deidentifikationsprozesse oder -techniken zum Durchführen von Datenanonymisierung zu entdecken und zu melden. Die Konfigurationsoptionen oder Einstellungen geben üblicherweise zu deidentifizierende Daten und entsprechende Informationen in den Daten an, die durch die deidentifizierten Werte zu bewahren sind. Als Beispiel kann eine Ausführungsform der vorliegenden Erfindung Konfigurationsoptionen für Datenmaskierungsprozesse oder -techniken für direkte Identifikatoren eines Datensatzes erkennen. Allerdings können alle Daten-Deidentifikations- oder Anonymisierungsprozesse oder -techniken für alle Typen von Identifikatoren im Wesentlichen auf die gleiche Weise wie unten beschrieben ausgewertet werden.

[0014] Ansätze des systematischen Ausprobierens werden üblicherweise von bestehenden Ansätzen eingesetzt, um Daten-Deidentifikationsprozesse zum Deidentifizieren von Daten auszuwählen. Diese Auswahlen beruhen im Allgemeinen auf dem Wissen eines Benutzers und können zu einer suboptimalen Daten-Deidentifikation und zahlreichen Daten-Deidentifikationsversuchen führen, wodurch Verarbeitungs- und andere Ressourcen verschwendet werden. Ausführungsformen der vorliegenden Erfindung verkürzen die Verarbeitungszeit, indem sie zulässige und/oder optimale Konfigurationen für Daten-Deidentifikationsprozesse zur schnellen Deidentifikation von Daten in einer Art identifizieren, die den Nutzen maximal bewahrt.

[0015] Gemäß einer Ausführungsform der vorliegenden Erfindung werden ein oder mehrere Identifikatoren (Attribute) ermittelt, die eine Entität eines Datensatzes identifizieren. Ein oder mehrere Daten-Deidentifikationsprozesse werden identifiziert und dem einen oder den mehreren ermittelten Identifikatoren zugewiesen. Jedem Daten-Deidentifikationsprozess werden ein oder mehrere Sätze von (den Nutzen bewahrenden) Konfigurationsoptionen zugewiesen, die zu bewahrende Informationen angeben. Für jeden Identifikator in dem Datensatz wird ein Daten-Deidentifikationsprozess mit einer den Nutzen bewahrenden Konfiguration ausgewählt. Der spezielle Fall eines vollständigen Unterdrückens des Identifikators wird unter den Daten-Deidentifikationsprozessen für den Identifikator berücksichtigt. Die ausgewählten Daten-Deidentifikationsprozesse werden an dem Datensatz gemäß den zugehörigen Sätzen von Konfigurationsoptionen ausgeführt, um einen Datensatz mit unterschiedlichem bewahrtem Datennutzen zu erstellen. Anschließend wird für mindestens einen Identifikator ein anderer Daten-Deidentifikationsprozess mit ei-

ner den Nutzen bewahrenden Konfiguration ausgewählt, und die neu ausgewählten Daten-Deidentifikationsprozesse werden an dem Datensatz gemäß dem zugehörigen Satz von Konfigurationsoptionen ausgeführt, um einen neuen Datensatz mit unterschiedlichem bewahrtem Datennutzen zu erstellen. Derselbe Vorgang wird so lange wiederholt, bis alle möglichen Kombinationen der verschiedenen Daten-Deidentifikationsprozesse und deren zugehöriger Konfigurationsoptionen für den einen oder die mehreren bestimmten Identifikatoren, die eine Entität des Datensatzes identifizieren, an dem Datensatz ausgeführt wurden, um Datensätze mit unterschiedlichem bewahrtem Datennutzen zu erstellen. Jeder erstellte Datensatz wird auf Datenschutz-Schwachstellen hin ausgewertet, und auf der Grundlage der Auswertung werden ein oder mehrere Daten-Deidentifikationsprozesse und zugehörige Sätze von Konfigurationsoptionen ausgewählt. Von den ausgewählten Daten-Deidentifikationsprozessen wird derjenige, der das geringste Risiko einer erneuten Identifikation und den höchsten Datennutzen erzielt, an dem Datensatz gemäß dem zugehörigen Satz von Konfigurationsoptionen ausgeführt, um einen sich ergebenden deidentifizierten Datensatz zu erzeugen.

[0016] Eine Ausführungsform der vorliegenden Erfindung kann darüber hinaus Datensätze zur Auswertung in Form einer Tabelle erstellen und zwei oder mehr Spalten eines erstellten Datensatzes zusammenfassen, um eine Spalte mit Informationen zu erzeugen, die spezifischer sind als die zwei oder mehr Spalten. Auf diese Weise kann ein Datensatz mit spezifischeren Informationen ausgewertet werden, um eine Abwesenheit einer Datenschutz-Schwachstelle sicherzustellen. Wenn der erstellte Datensatz mit spezifischeren Informationen keine Datenschutz-Schwachstelle aufweist, weisen auch andere Datensätze, die aus dem entsprechenden Daten-Deidentifikationsprozess und den Konfigurationsoptionen mit verallgemeinerteren Informationen erstellt wurden (z.B. Datensätze mit einer oder mehreren der ursprünglichen nicht zusammengefassten Spalten) keine Datenschutz-Schwachstelle auf. Dies verkürzt auch die Verarbeitungszeit, indem für Datensätze mit den spezifischeren und verallgemeinerten Informationen eine einzelne Auswertung anstelle von mehreren Auswertungen genutzt wird.

[0017] Außerdem kann eine Ausführungsform der vorliegenden Erfindung einen erstellten Datensatz auf Datenschutz-Schwachstellen hin auswerten, indem ein Vorhandensein einer Verknüpfung zwischen Daten für eine Entität in einem erstellten Datensatz und Daten für eine bekannte Entität in einem öffentlich verfügbaren Datensatz ermittelt wird, um auf eine Datenschutz-Schwachstelle für den erstellten Datensatz hinzuweisen. Bei dieser Auswertung werden die deidentifizierten Daten aus einem erstellten Datensatz genutzt und mit bekannten Entitäten in einem öffentlich verfügbaren Datensatz verglichen, um zu ermitteln, ob Identifikatoren von Entitäten in den deidentifizierten Daten ermittelt werden können, wodurch eine erhebliche Zuversicht bereitgestellt wird, dass ein empfohlener Daten-Deidentifikationsprozess mit zugehörigen Konfigurationsoptionen den Datenschutz beibehält.

[0018] Eine Ausführungsform der vorliegenden Erfindung kann darüber hinaus einen erstellten Datensatz auf Datenschutz-Schwachstellen hin auswerten, indem ein Vorhandensein eines Satzes von Quasi-Identifikatoren in einem erstellten Datensatz ermittelt wird, der durch einen entsprechenden Daten-Deidentifikationsprozess und einen zugehörigen Satz von Konfigurationsoptionen eingebracht wurde, um auf eine Datenschutz-Schwachstelle für den erstellten Datensatz hinzuweisen. Diese Auswertung stellt sicher, dass Quasi-Identifikatoren nicht durch einen Daten-Deidentifikationsprozess und zugehörige Konfigurationsoptionen eingebracht werden, wodurch eine erhebliche Zuversicht bereitgestellt wird, dass ein empfohlener Daten-Deidentifikationsprozess mit zugehörigen Konfigurationsoptionen den Datenschutz beibehält.

[0019] Außerdem kann eine Ausführungsform der vorliegenden Erfindung die Verarbeitungszeit für die Deidentifikation verkürzen, indem ein erstellter Datensatz identifiziert wird, der keine Datenschutz-Schwachstellen aufweist, und eine Verarbeitung in Bezug auf andere zugehörige Sätze von Konfigurationsoptionen für einen entsprechenden Daten-Deidentifikationsprozess beendet wird, der einen Datensatz mit verallgemeinerteren Informationen als der identifizierte erstellte Datensatz erstellt. Dies verbessert die Leistung der Prozessoren erheblich und stellt eine optimale Daten-Deidentifikation in einer verkürzten Verarbeitungszeit bereit.

[0020] Eine beispielhafte Umgebung zur Verwendung mit Ausführungsformen der vorliegenden Erfindung ist in **Fig. 1** veranschaulicht. Konkret umfasst die Umgebung ein oder mehrere Serversysteme **110** und ein oder mehrere Client- oder Endbenutzer-Systeme **114**. Die Serversysteme **110** und die Client-Systeme **114** können voneinander entfernt angeordnet sein und über ein Netzwerk **112** Daten austauschen. Das Netzwerk kann durch eine beliebige Anzahl von beliebigen geeigneten Datenübertragungsmedien umgesetzt sein (z.B. ein Weitverkehrsnetz (WAN), ein lokales Netzwerk (LAN), das Internet, ein Intranet usw.). Alternativ können sich die Serversysteme **110** und die Client-Systeme **114** nahe beieinander befinden und über ein beliebiges geeignetes lokales Datenübertragungsmedium Daten austauschen (z.B. über ein lokales Netzwerk (LAN), festverdrahtet, eine drahtlose Verbindung, ein Intranet usw.).

[0021] Die Client-Systeme **114** ermöglichen es Benutzern, mit den Serversystemen **110** zu interagieren, um gewünschte Aktionen wie zum Beispiel eine Daten-Deidentifikation durchzuführen. Die Serversysteme umfassen ein Auswertungsmodul **116**, um zulässige Konfigurationen oder Einstellungen für verschiedene Daten-Deidentifikationsprozesse oder - techniken zu erkennen, um sich ergebende Datensätze zu erzeugen, die den Datenschutz beibehalten. Ein Datenbanksystem **118** kann verschiedene Informationen für die Analyse speichern (z.B. ursprüngliche und vorläufige Datensätze, Konfigurationen oder Einstellungen, Optionen für Daten-Deidentifikationsprozesse usw.). Das Datenbanksystem kann durch jede beliebige herkömmliche oder andere Datenbank oder Speichereinheit umgesetzt sein, sich lokal bei oder entfernt angeordnet von den Serversystemen **110** und den Client-Systemen **114** befinden und über ein beliebiges geeignetes Datenübertragungsmedium Daten austauschen (z.B. über ein lokales Netzwerk (LAN), ein Weitverkehrsnetz (WAN), das Internet, festverdrahtet, eine drahtlose Verbindung, ein Intranet usw.). Die Client-Systeme können eine grafische Benutzerschnittstelle (z.B. GUI usw.) oder eine andere Schnittstelle (z.B. Befehlszeilen-Eingabeaufforderungen, Menübildschirme usw.) zum Abrufen von Informationen von die Analyse betreffenden Benutzern bereitstellen, und können verschiedene Berichte bereitstellen, die Analyseergebnisse umfassen (z.B. empfohlene Daten-Deidentifikationsprozesse, deidentifizierte Datensätze, zum Deidentifizieren von Datensätzen verwendete Optionen usw.).

[0022] Die Serversysteme **110** und die Client-Systeme **114** können durch beliebige herkömmliche oder andere Computersysteme umgesetzt sein, die vorzugsweise mit einer Anzeige oder einem Bildschirm, einer Basis, optionalen Eingabeeinheiten (z.B. einer Tastatur, einer Maus oder einer anderen Eingabeeinheit) und jeder beliebigen im Handel erhältlichen und angepassten Software (z.B. einer Server/Datenübertragungs-Software, einem Auswertungsmodul, einer Browser/Schnittstellen-Software, Daten-Deidentifikationsprozessen usw.) ausgestattet sind. Die Basis umfasst vorzugsweise mindestens einen Hardware-Prozessor **115** (z.B. einen Mikroprozessor, eine Steuereinheit, eine zentrale Verarbeitungseinheit (CPU) usw.), einen oder mehrere Hauptspeicher **135** und/oder interne oder externe Netzwerkschnittstellen oder Datenübertragungseinheiten **125** (z.B. ein Modem, Netzwerkkarten usw.).

[0023] Alternativ können ein oder mehrere Client-Systeme **114** zulässige Konfigurationen oder Einstellungen für verschiedene Daten-Deidentifikationsprozesse oder - techniken erkennen, wenn sie als eigenständige Einheit betrieben werden. In einer eigenständigen Betriebsart speichert das Client-System die Daten (z.B. Datensätze, Konfigurationen oder Einstellungen, Daten-Deidentifikationsprozesse usw.) oder hat Zugriff darauf und umfasst das Auswertungsmodul **116** zum Durchführen des Erkennens. Die grafische Benutzerschnittstelle (z.B. GUI usw.) oder die andere Schnittstelle (z.B. Befehlszeilen-Eingabeaufforderungen, Menübildschirme usw.) ruft Informationen von einem entsprechenden, die Analyse betreffenden Benutzer ab, und kann Berichte bereitstellen, die Analyseergebnisse umfassen.

[0024] Das Auswertungsmodul **116** kann ein(e) oder mehrere Module oder Einheiten zum Durchführen der verschiedenen Funktionen der unten beschriebenen Ausführungsformen der vorliegenden Erfindung umfassen. Die verschiedenen Module (z.B. das Auswertungsmodul, usw.) können durch jede beliebige Kombination aus einer beliebigen Menge von Software- und/oder Hardware-Modulen oder Einheiten umgesetzt sein und sich innerhalb des Hauptspeichers **135** der Server- und/oder Client-Systeme zur Ausführung durch den Prozessor **115** befinden.

[0025] Eine Art und Weise des Erkennens von zulässigen Konfigurationsoptionen oder Einstellungen für Daten-Deidentifikationsprozesse oder -techniken (z.B. mittels des Auswertungsmoduls **116** und eines Serversystems **110** und/oder eines Client-Systems **114**) zum Erzeugen von Datensätzen unter Wahrung von Datenschutz gemäß einer Ausführungsform der vorliegenden Erfindung ist in **Fig. 2** veranschaulicht. Zu Beginn wird jeder Daten-Deidentifikationsprozess bestimmten Datentypen zugewiesen und läuft gemäß verschiedenen Konfigurationsoptionen oder Einstellungen. Die Konfigurationsoptionen oder Einstellungen geben üblicherweise zu deidentifizierende Daten und entsprechende Informationen in den Daten an, die durch die deidentifizierten Werte zu bewahren sind. Zum Beispiel erzeugt eine Konfigurationsoption, die einen zu deidentifizierenden Namen angibt, während Geschlechtsangaben bewahrt bleiben, einen Datensatz, bei dem die Namen durch fiktionalisierte Namen ersetzt sind, die mit den ursprünglichen Geschlechtsangaben übereinstimmen (z.B. wird ein Frauenname durch einen fiktionalisierten Frauennamen ersetzt, um eine Übereinstimmung mit den Geschlechtsangaben für den ursprünglichen Namen zu bewahren oder beizubehalten, usw.). Zusätzlich können die Konfigurationsoptionen angeben, dass bestimmte Daten aus einem sich ergebenden Datensatz zu löschen sind.

[0026] Für jeden Satz von Konfigurationsoptionen, die einem Daten-Deidentifikationsprozess zugehörig sind, wird eine Vorlage erstellt. Jeder Satz von Konfigurationsoptionen kann eine oder mehrere Konfigurationsoptionen für den Daten-Deidentifikationsprozess enthalten. Somit ist jeder Daten-Deidentifikationsprozess (verfügbar auf den Server- und/oder Client-Systemen) einer Reihe von Vorlagen zugehörig, wobei jede Vorlage einem der möglichen Sätze von Konfigurationsoptionen für diesen Daten-Deidentifi-

kationsprozess entspricht (z.B. zu löschende Daten, Daten, die unter Wahrung anderer Daten zu deidentifizieren sind, Daten, die unter Wahrung bestimmter Eigenschaften wie zum Beispiel räumlicher Nähe usw. zu deidentifizieren sind, usw.). Grundsätzlich deckt die Reihe von Vorlagen für jeden Daten-Deidentifikationsprozess alle möglichen Konfigurationen für diesen Daten-Deidentifikationsprozess in Bezug auf zugehörige Attribute oder Identifikatoren ab. Die Vorlage erfasst Informationen, die in dem Datensatz erhalten bleiben, wenn ein ursprüngliches Attribut durch einen zugehörigen Daten-Deidentifikationsprozess gemäß den entsprechenden Konfigurationsoptionen verarbeitet wird.

[0027] Zum Beispiel können die Vorlagen für einen Daten-Deidentifikationsprozess in Bezug auf Namens-, Telefonnummern- und Adressattribute umfassen: Namensvorlagen (z.B. Vorlage (Name, löschen), wobei das Namensattribut zu löschen ist; Vorlage (Name, Geschlecht), wobei das Namensattribut durch Werte ersetzt wird, die eine Übereinstimmung mit Geschlechtsangaben bewahren oder beibehalten); Telefonvorlagen (z.B. Vorlage (Telefon, löschen), wobei das Telefonnummernattribut zu löschen ist; Vorlage (Telefon, Land), Vorlage (Telefon, Land und Ort), wobei das Telefonnummernattribut durch Werte ersetzt wird, welche eine Übereinstimmung mit dem Land bzw. der Landes- und Ortsvorwahl bewahren oder beibehalten); Adressvorlagen (z.B. Vorlage (Adresse, löschen), wobei das Adressattribut zu löschen ist; Vorlage (Adresse, Land), Vorlage (Adresse, Land und Stadt), Vorlage (Adresse, minimal umgebendes Rechteck (MBR minimum bounding rectangle)), wobei das Adressattribut durch Werte ersetzt wird, die eine Übereinstimmung mit dem Land, der Stadt bzw. dem Ortsgebiet innerhalb einer vorgeschriebenen Entfernung bewahren oder beibehalten). Allerdings können sich die Vorlagen auf alle gewünschten Optionen zum Löschen oder Bewahren jeglicher Attribute beziehen (z.B. Adresse, Telefonnummer, Fahrzeug-Identifizierungsnummer (VIN, vehicle identification number), Sozialversicherungsnummer (SSN, social security number), Land, URL (Uniform Resource Locator), Name, IP-Adresse, eMail-Adresse, Kreditkartennummer, internationale Bankkontonummer (IBAN, international bank account number), Datum, Stadt, medizinischer ICD-Code, Beruf, Krankenhaus, Breitengrad/Längengrad, Postleitzahl usw.). Im Hinblick auf die Wahrung von Datenschutz und Datennutzen erfasst eine Vorlage die wahrheitsgemäßen Informationen, die nach der Deidentifikation in dem Datensatz beibehalten werden. Für eine Vorlage (Attribut A, Optionen B) stellt dies das Ersetzen von Attribut A in dem Datensatz durch die (den Nutzen bewahrenden) Informationen dar, die in den Optionen B bereitgestellt werden. Zum Beispiel kann die Vorlage (Name, Geschlecht) so übersetzt werden, dass das Namensattribut in dem Datensatz durch ein Geschlechtsattribut ersetzt wird, das genaue Geschlechtsangaben über die Personen in den Daten erfasst. In ähnlicher Weise kann die Vorlage (Telefon, Land und Ort) so übersetzt werden, dass das Telefonattribut in dem Datensatz durch ein Attribut ersetzt wird, das genaue Länderinformationen beibehält, und durch ein Attribut, das genaue Ortsinformationen für die in dem Datensatz dargestellten Personen beibehält. Die Verwendung von Vorlagen stellt Informationen darüber bereit, was in den Daten erhalten wurde, die anschließend zum Berechnen eines Datenschutzrisikos und eines Datennutzens in dem sich ergebenden Datensatz verwendet werden können.

[0028] Zusätzlich kann eine Vorlage ein oder mehrere zu löschende oder zu deidentifizierende Attribute und/oder ein oder mehrere zu erhaltende Attribute angeben. Zum Beispiel kann eine Reihe von Vorlagen anfängliche Vorlagen enthalten, von denen jede ein Attribut angibt, das gemäß Konfigurationsoptionen zu löschen oder zu deidentifizieren ist. Zusätzliche Vorlagen können Konfigurationsoptionen der ursprünglichen Vorlagen oder Attribute angeben und weitere Konfigurationsoptionen in Bezug auf ein zweites Attribut enthalten (z.B. ein Bereitstellen einer Deidentifikation von zwei Attributen). Somit können die Vorlagen für einen Daten-Deidentifikationsprozess alle oder einen beliebigen Teil der verschiedenen Kombinationen von Deidentifikation abdecken, die durch den Daten-Deidentifikationsprozess für entsprechende Attribute eines Datensatzes bereitgestellt werden.

[0029] Als Beispiel wird eine Ausführungsform der vorliegenden Erfindung in Bezug auf das Erkennen von Konfigurationsoptionen für Daten-Deidentifikationsprozesse in Form von Datenmaskierungsprozessen oder -techniken für direkte Identifikatoren eines Datensatzes beschrieben. Allerdings können alle Daten-Deidentifikations- oder Anonymisierungsprozesse oder -techniken für alle Typen von Identifikatoren im Wesentlichen auf die gleiche Weise wie unten beschrieben ausgewertet werden.

[0030] Konkret wird ein Datensatz **250** empfangen und analysiert, um in Schritt **205** direkte Identifikatoren für eine Datenmaskierung zu erkennen. Bei direkten Identifikatoren handelt es sich um Attribute, die zum direkten Identifizieren einer Entität verwendet werden können (z.B. Name, Sozialversicherungsnummer, Adresse, Telefonnummer usw.). Der Datensatz liegt vorzugsweise in Form einer Tabelle vor, in der jede Zeile eine Entität darstellt und jede Spalte ein Attribut dieser Entität darstellt (z.B. Name, Adresse, Geschlecht usw.). Der Datensatz kann jedoch in jedem beliebigen gewünschten Format vorliegen. Die direkten Identifikatoren können unter Verwendung jeder beliebigen herkömmlichen oder anderen Technik erkannt werden. Zum Beispiel kann die Eindeutigkeit von Attributen in Bezug auf eine Entität ver-

wendet werden, um direkte Identifikatoren in dem Datensatz **250** zu erkennen. Alternativ können reguläre Ausdrücke oder Muster verwendet werden, um bestimmte Typen von Daten in dem Datensatz zu identifizieren, die als direkte Identifikatoren bekannt sind (z.B. Sozialversicherungsnummer, Adresse, Datumsangaben usw.). Alternativ können Nachschlagetabellen verwendet werden, um bestimmte Typen von direkten Identifikatoren wie zum Beispiel Namen zu identifizieren (z.B. durch Wählerregistrierungslisten). Zusätzlich können direkte Identifikatoren für einen Datensatz manuell durch einen Benutzer vorbestimmt werden.

[0031] Datenmaskierungsprozesse, die den erkannten direkten Identifikatoren entsprechen, werden in Schritt **210** identifiziert. Die Datenmaskierungsprozesse sind üblicherweise mit bestimmten Typen von Daten oder Attributen kompatibel, und jeder erkannte direkte Identifikator wird zur Auswertung jedem der entsprechenden kompatiblen Datenmaskierungsprozesse zugewiesen.

[0032] Die Datenmaskierungsprozesse werden auf entsprechende direkte Identifikatoren gemäß den (oben beschriebenen) Vorlagen angewendet, welche die verschiedenen Sätze von Konfigurationsoptionen für die Datenmaskierungsprozesse in Schritt **215** angeben. Hierdurch wird ein Datensatz für jeden Satz von Konfigurationsoptionen für jeden Datenmaskierungsprozess erstellt, der den direkten Identifikatoren zugehörig ist. Die erstellten Datensätze liegen vorzugsweise in Form einer Tabelle mit Zeilen und Spalten (oder Attributen) vor, können aber in jedem beliebigen gewünschten Format vorliegen. Zum Beispiel veranschaulicht **Fig. 6** einen Ausgangsdatensatz **600** in Form einer Tabelle, in der jede Zeile eine Person darstellt, und mit Spalten oder Attributen für jede Person mit Name, Adresse, Geburtsdatum, Postleitzahl und Familienstand. Ein Datenmaskierungsprozess kann es ermöglichen, das Namensattribut mit einem fiktionalisierten Namen zu maskieren, der eine Übereinstimmung mit dem Geschlechtsattribut bewahrt oder beibehält. In diesem Fall kann eine Vorlage für den Datenmaskierungsprozess den entsprechenden Satz von Konfigurationsoptionen angeben (z.B. Vorlage (Name, Geschlecht)).

[0033] Wenn der Datenmaskierungsprozess gemäß diesem Satz von Konfigurationsoptionen angewendet wird, wird ein Datensatz **620** erstellt, wobei die Namen der Personen mit fiktionalisierten Namen maskiert sind, die eine Übereinstimmung mit dem Geschlechtsattribut bewahren oder beibehalten. Im Endeffekt führt dies zu einem Datensatz **620**, in dem ein neues Geschlechtsattribut erscheint, das genaue Geschlechtsangaben enthält, die aus dem ursprünglichen Datensatz **600** berechnet wurden. In diesem Fall wurden männliche Namen in dem Datensatz **600** durch andere männliche Namen in dem Datensatz **620** ersetzt, um die Geschlechtsangaben beizubehalten. Auf ähnliche Weise wurden weibliche Namen in dem Datensatz **600** durch andere weibliche Namen in dem Datensatz **620** ersetzt, um die Geschlechtsangaben zu bewahren. Dadurch wird das Attribut oder die Spalte „Name“ effektiv durch die Spalte „Geschlecht“ ersetzt, wenn es um ein Bewerten von Datenschutz-Schwachstellen geht (da die fiktionalisierten Namen lediglich das Geschlecht der Person kennzeichnen und nicht für irgendwelche anderen Zwecke verwendet werden können, die das Datenschutzrisiko in den Daten erhöhen würden).

[0034] Als weiteres Beispiel veranschaulicht **Fig. 7** einen Ausgangsdatensatz **700** in Form einer Tabelle, in der jede Zeile eine Person darstellt, und mit Spalten oder Attributen für jede Person mit Name, Adresse, Geburtsdatum, Postleitzahl und Familienstand. Ein Datenmaskierungsprozess kann es ermöglichen, das Adressattribut mit einer anderen Adresse innerhalb eines zwei Meilen großen minimal umgebenden Rechtecks (MBR) zu maskieren. In diesem Fall kann eine Vorlage für den Datenmaskierungsprozess den entsprechenden Satz von Konfigurationsoptionen angeben (z.B. Vorlage (Adresse, minimal umgebendes Rechteck (MBR)).

[0035] Wenn der Datenmaskierungsprozess gemäß diesem Satz von Konfigurationsoptionen angewendet wird, wird ein Datensatz **720** erstellt, wobei die Adressen der Personen auf andere Adressen geändert oder maskiert werden, die sich innerhalb des zwei Meilen großen minimal umgebenden Rechtecks (MBR) befinden. Allerdings können die neuen Adressen in Kombination mit der Postleitzahl einen Quasi-Identifikator bilden und eine Datenschutz-Schwachstelle für diesen Satz von Konfigurationsoptionen schaffen. In dem erstellten Datensatz **720** muss man also die Adress- und Postleitzahlattribute kombinieren, um eine möglichst hohe Genauigkeit bezüglich des Aufenthaltsortes der Person zu erhalten (z.B. eine Privatadresse). Dann verwendet man diese Informationen, um das Datenschutzrisiko aufgrund eines Freigebens der Daten zu bewerten.

[0036] Zurückgehend zu **Fig. 2** werden die erstellten Datensätze aus den Vorlagen ausgewertet, um in Schritt **220** zulässige Datenmaskierungsprozesse und entsprechende Sätze von Konfigurationsoptionen zu identifizieren, um sich ergebende Datensätze zu erzeugen, bei denen der Datenschutz beibehalten wird. Die Auswertung analysiert einen erstellten Datensatz in Bezug auf Verknüpfungen mit öffentlich verfügbaren oder externen Datensätzen (z.B. Wählerregistrierungslisten, Gelbe Seiten, Volkszählungsdaten usw.). Wenn es eine Verknüpfung gibt (z.B. wenn ein Triangulationsangriff mit dem externen Datensatz erfolgreich ist), deutet dies darauf hin, dass eine Identität einer Person des erstellten (oder maskierten) Datensatzes ermittelt werden kann, wo-

durch eine Datenschutz-Schwachstelle im Hinblick auf den Datenmaskierungsprozess und den entsprechenden Satz von Konfigurationsoptionen identifiziert wird, die zum Erstellen des Datensatzes verwendet wurden. Zusätzlich kann der erstellte Datensatz analysiert werden, um das Vorhandensein von Quasi-Identifikatoren zu ermitteln, die auf der Grundlage des Datenmaskierungsprozesses und des entsprechenden Satzes von Konfigurationsoptionen zu dem erstellten Datensatz eingebracht wurden. Das Vorhandensein eines Quasi-Identifikators deutet auf eine Datenschutz-Schwachstelle im Hinblick auf den Datenmaskierungsprozess und den entsprechenden Satz von Konfigurationsoptionen hin, die zum Erstellen des Datensatzes verwendet wurden.

[0037] Ein sich ergebender Datenmaskierungsprozess und ein entsprechender Satz von Konfigurationsoptionen können aus den identifizierten zulässigen Datenmaskierungsprozessen (und den entsprechenden Sätzen von Konfigurationsoptionen) ausgewählt werden. Der sich ergebende Datenmaskierungsprozess kann durch einen Benutzer manuell ausgewählt werden. In diesem Fall können die zulässigen Datenmaskierungsprozesse und die zugehörigen Sätze von Konfigurationsoptionen einem Benutzer auf einem Client-System **114** zur Auswahl angeboten werden. Es können auch Empfehlungen bezüglich der zulässigen Datenmaskierungsprozesse bereitgestellt werden. Die Empfehlungen können auf verschiedenen Metriken beruhen (z.B. Datenschutzniveaus, Verarbeitungszeiten, Datenerhalt usw.).

[0038] Alternativ kann der sich ergebende Datenmaskierungsprozess automatisch ermittelt werden. Es können verschiedene Metriken genutzt werden, um den sich ergebenden Datenmaskierungsprozess zu ermitteln. Zum Beispiel kann der Datenmaskierungsprozess ausgewählt werden, der den größten Datenschutz bereitstellt, und zwar auf der Grundlage von Verknüpfungen mit öffentlich verfügbaren Datensätzen und/oder eines Einbringens der geringsten Menge von Quasi-Identifikatoren. Alternativ kann der Datenmaskierungsprozess auf der Grundlage der geringsten Menge an Ressourcennutzung und/oder Verarbeitungszeit ausgewählt werden, um die Verarbeitungszeit für die Deidentifikation des Datensatzes zu verkürzen.

[0039] Zusätzlich kann der sich ergebende Datenmaskierungsprozess auf der Grundlage von maschinellem Lernen empfohlen oder automatisch ausgewählt werden. In diesem Fall können Datenmaskierungsprozesse und zugehörige Sätze von Konfigurationsoptionen, die von einem Benutzer ausgewählt wurden, gespeichert werden, und/oder Metriken können nachverfolgt werden. Diese Informationen können verarbeitet werden, um Benutzerpräferenzen für Auswahlen und/oder Empfehlungen zu erlernen. Es können verschiedene Modelle zum Durchführen des Lernens eingesetzt werden (z.B. neuronale Netze, mathematische/statistische Modelle, Klassifikationsmerkmale usw.). Zum Beispiel kann zunächst ein Maskierungsprozess empfohlen und/oder ausgewählt werden. Allerdings bevorzugte ein Benutzer aus bestimmten Gründen wiederholt einen anderen zulässigen Datenmaskierungsprozess. Diese Aspekte und Präferenzen für Benutzer können erlernt (z.B. bevorzugt ein Benutzer möglicherweise schnellere Verarbeitungszeiten anstelle eines höheren Datenschutzniveaus usw.) und zum Auswählen und/oder Empfehlen von Datenmaskierungsprozessen verwendet werden.

[0040] Der sich ergebende Datenmaskierungsprozess wird in Schritt **225** auf den Datensatz **250** gemäß dem entsprechenden Satz von Konfigurationsoptionen angewendet (oder an ihm ausgeführt), um den Datensatz unter Wahrung des Datenschutzes zu deidentifizieren.

[0041] Eine Art und Weise des Anwendens der Vorlagen für die Datenmaskierungsprozesse zum Erstellen von Datensätzen zur Auswertung (z.B. entsprechend Schritt **215** aus **Fig. 2**) gemäß einer Ausführungsform der vorliegenden Erfindung ist in **Fig. 3** veranschaulicht. Zunächst werden eine Reihe von Datenmaskierungsprozessen und entsprechende Sätze von Konfigurationsoptionen verwendet, um Datensätze zu erstellen, die auf ein Einbringen von möglichen Datenschutzrisiken geprüft werden. Insbesondere werden in Schritt **305** verschiedene Sätze von Konfigurationsoptionen für jeden Datenmaskierungsprozess ermittelt, der den erkannten direkten Identifikatoren zugehörig ist. In Schritt **310** wird für jeden ermittelten Satz von Konfigurationsoptionen für jeden Datenmaskierungsprozess ein Datensatz erstellt. Dies kann erreicht werden, indem eine Vorlage, die einen Satz von Konfigurationsoptionen angibt, auf einen Datenmaskierungsprozess angewendet wird, um einen Datensatz zu erstellen. Mit anderen Worten, der Datenmaskierungsprozess wird gemäß dem Satz von Konfigurationsoptionen der Vorlage ausgeführt, um einen zugehörigen direkten Identifikator zu löschen oder zu maskieren. Der erstellte Datensatz liegt vorzugsweise in Form einer Tabelle mit Zeilen und Spalten (oder Attributen) vor, kann aber in jedem beliebigen gewünschten Format vorliegen.

[0042] Attribute oder Spalten in einem erstellten Datensatz, die denselben oder einen kompatiblen Typ aufweisen, können in Schritt **315** zusammengefasst werden, um eine Spalte in dem erstellten Datensatz mit präziseren oder spezifischeren Informationen bereitzustellen. Zum Beispiel kann es sich bei einer zusammengefassten Spalte um eine Schnittmenge von Bereichen oder Orten in den Ausgangsspalten handeln, die zusammengefasst werden. Als Beispiel können Spalten, die jeweils Postleitzahlen und ein minimal umgebendes Rechteck (MBR) von Adressen

enthalten, durch eine einzelne Spalte ersetzt werden, welche die präziseren Informationen in Bezug auf den Ort enthält. In diesem Fall kann, wenn das MBR ein größeres Gebiet als die Postleitzahlen abdeckt, die Postleitzahlenspalte in dem erstellten Datensatz als spezifischere Informationen in Bezug auf einen Ort bereitstellend verbleiben (z.B. decken die Postleitzahlen ein kleineres Gebiet als das MBR ab). Hierdurch wird ein erstellter Datensatz mit spezifischeren Informationen (oder ein Szenario, das anfälliger für Datenschutz-Schwachstellen ist) bereitgestellt, der auf Datenschutz-Schwachstellen hin zu prüfen ist. Wenn die spezifischeren Informationen keine Bedenken hinsichtlich des Datenschutzes aufwerfen, werfen alle verallgemeinerten oder weiter gefassten Informationen ebenfalls keine Bedenken hinsichtlich des Datenschutzes auf.

[0043] Die erstellten Datensätze für jeden der Datenmaskierungsprozesse und die zugehörigen Sätze von Konfigurationsoptionen werden auf Datenschutz-Schwachstellen hin ausgewertet.

[0044] Eine Art und Weise des Erkennens von Datenschutz-Schwachstellen für die erstellten Datensätze (z.B. entsprechend Schritt **220** aus **Fig. 2**) auf der Grundlage von öffentlich verfügbaren Daten ist in **Fig. 4** veranschaulicht. Zunächst wird jeder Datensatz, der aus einem Datenmaskierungsprozess und einer entsprechenden Vorlage, die einen Satz von Konfigurationsoptionen angibt, erstellt wurde, auf Datenschutz-Schwachstellen hin ausgewertet. Dies wird durch ein Verknüpfen von Daten in dem erstellten Datensatz mit externen oder öffentlich verfügbaren Daten erreicht. Insbesondere werden Daten innerhalb jedes erstellten Datensatzes in Schritt **405** auf eine mögliche Verknüpfung mit externen oder öffentlich verfügbaren Daten (z.B. Wählerregistrierungslisten, Gelbe Seiten, Volkszählungsdaten usw.) hin geprüft. Mit anderen Worten, Daten für eine Entität in einem erstellten Datensatz werden verwendet, um eine Verknüpfung zu Daten einer entsprechenden bekannten Entität in den öffentlich verfügbaren Daten zu ermitteln. Zum Beispiel können ein oder mehrere Attributwerte für eine Entität in dem erstellten Datensatz verwendet werden, um entsprechende Attributwerte in den öffentlich verfügbaren Daten zu finden.

[0045] Wenn eine Verknüpfung besteht (z.B. stimmen eine ausreichende Menge oder ein ausreichendes Muster von Attributen überein), weist dies darauf hin, dass die Entitätsdaten des erstellten Datensatzes der bekannten Entität in den öffentlich verfügbaren Daten entsprechen, wodurch eine Identifizierung der Entität aus dem erstellten Datensatz ermöglicht wird. Eine Menge von Verknüpfungen zwischen Entitäten eines erstellten Datensatzes und den öffentlich verfügbaren Daten kann beibehalten und mit einem Schwellenwert verglichen werden, um das Vorhandensein einer Datenschutz-Schwachstelle für den erstellten Datensatz (und den Datenmaskierungsprozess und den Satz von Konfigurationsoptionen, die zum Erzeugen des erstellten Datensatzes verwendet wurden) bei Ablauf **410** festzustellen. Der Schwellenwert kann auf jeden beliebigen gewünschten Wert gesetzt werden, wobei die Menge von Verknüpfungen in jeder beliebigen gewünschten Weise mit dem Schwellenwert verglichen werden kann, um auf eine Datenschutz-Schwachstelle hinzuweisen (z.B. größer, kleiner, größer oder gleich, kleiner oder gleich usw.). Als Beispiel kann der Schwellenwert auf null gesetzt werden, und ein erstellter Datensatz kann als Reaktion auf das Vorhandensein einer oder mehrerer Verknüpfungen zwischen Entitäten des erstellten Datensatzes und bekannten Entitäten der öffentlich verfügbaren Daten als eine Datenschutz-Schwachstelle aufweisend betrachtet werden. Die Datenmaskierungsprozesse und zugehörigen Sätze von Konfigurationsoptionen, die zum Erzeugen von erstellten Datensätzen mit Datenschutz-Schwachstellen verwendet werden, werden zum Ermitteln von Empfehlungen und/oder Auswahlen gekennzeichnet.

[0046] Sobald jeder der erstellten Datensätze im Hinblick auf die externen oder öffentlich verfügbaren Daten geprüft wurde, werden die Datenmaskierungsprozesse und zugehörigen Sätze von Konfigurationsoptionen, die zum Erzeugen von erstellten Datensätzen mit Datenschutz-Schwachstellen verwendet werden, gekennzeichnet und von der weiteren Betrachtung ausgenommen. Die verbleibenden Datenmaskierungsprozesse und zugehörigen Sätze von Konfigurationsoptionen werden in Schritt **415** analysiert, um einen empfohlenen Satz von Datenmaskierungsprozessen und zugehörigen Satz von Konfigurationsoptionen zu ermitteln, um einen nicht mit Schwachstellen behafteten Datensatz bereitzustellen. Der empfohlene Satz kann verringert werden, indem Datenmaskierungsprozesse mit zugehörigen Sätzen von Konfigurationsoptionen mit geringerer Wahrung entfernt werden. Zusätzlich können, wenn keine Datenmaskierungsprozesse und zugehörige Sätze von Konfigurationsoptionen einen Datensatz ohne Datenschutz-Schwachstelle bereitstellen, der Datenmaskierungsprozess und zugehörige Satz von Konfigurationsoptionen mit den wenigsten Datenschutz-Schwachstellen (z.B., oder der geringsten Menge von Verknüpfungen) empfohlen werden. Die empfohlenen Datenmaskierungsprozesse können einem Benutzer zur Auswahl vorgelegt werden, oder es kann, wie oben beschrieben, automatisch ein Datenmaskierungsprozess ausgewählt werden.

[0047] Zusätzlich können Datenschutz-Schwachstellen für erstellte Datensätze auf der Grundlage einer Analyse der erstellten Datensätze (z.B. entsprechend Schritt **220** aus **Fig. 2**), wie in **Fig. 5** veranschaulicht, ermittelt werden. Zunächst wird in Schritt **505** jeder erstellte Datensatz auf ein Einbringen von seltenen oder einzigartigen Werten hin untersucht.

Jeder erstellte Datensatz wird in Schritt **510** weiter ausgewertet (mined), um alle Quasi-Identifikatoren zu erfassen, die möglicherweise aufgrund des Datenmaskierungsprozesses und des entsprechenden Satzes von Konfigurationsoptionen entstanden sind. Die Quasi-Identifikatoren können in einem erstellten Datensatz auf der Grundlage von beliebigen herkömmlichen oder anderen Techniken identifiziert werden. Zum Beispiel kann eine Einzigartigkeit von Entitäten, die durch Gruppen von Attributen innerhalb des erstellten Datensatzes identifiziert werden, zum Ermitteln von Quasi-Identifikatoren verwendet werden, reguläre Ausdrücke oder Muster können zum Identifizieren von bekannten Quasi-Identifikatoren verwendet werden, usw. Zusätzlich kann ein Benutzer Quasi-Identifikatoren aus ursprünglichen Datenspalten und/oder zusammengefassten Spalten angeben (z.B. Spalten, die auf der Grundlage eines Zusammenfassens von Spalten desselben Typs hergestellt wurden (oder die gemäß kompatiblen Vorlagen erstellt wurden)).

[0048] Jede Spalte eines erstellten Datensatzes, die als eine Komponente eines Quasi-Identifikators identifiziert wird, wird als eine Datenschutz-Schwachstelle aufweisend gekennzeichnet, um Empfehlungen und/oder Auswahlen zu ermitteln. Mit anderen Worten, der Datenmaskierungsprozess und der entsprechende Satz von Konfigurationsoptionen, die zum Erzeugen des erstellten Datensatzes verwendet werden, haben in den erstellten Datensatz einen Quasi-Identifikator eingebracht. Die identifizierten Quasi-Identifikatoren und Datenschutz-Schwachstellen werden zur Präsentation auf dem Client-System **114** in Schritt **515** bereitgestellt.

[0049] Die Auswertung der erstellten Datensätze für Datenverknüpfungen und Quasi-Identifikatoren kann in einer beliebigen Reihenfolge durchgeführt werden, und kann darüber hinaus parallel durchgeführt werden, um die Verarbeitungsleistung zu verbessern. Zusätzlich können Ergebnisse dieser Auswertungen in einer beliebigen Art und Weise kombiniert werden, um das Vorhandensein einer Datenschutz-Schwachstelle innerhalb eines erstellten Datensatzes festzustellen. Zum Beispiel kann für einen erstellten Datensatz infolge einer bestimmten Menge von Verknüpfungen und einer bestimmten Menge von Quasi-Identifikatoren eine Datenschutz-Schwachstelle vorliegen. Alternativ kann entweder infolge einer bestimmten Anzahl von Datenverknüpfungen oder einer bestimmten Anzahl von Quasi-Identifikatoren festgestellt werden, dass eine Datenschutz-Schwachstelle vorliegt. In diesem Fall wird, wenn eine dieser Bedingungen eintritt, der erstellte Datensatz als Datenschutz-Schwachstellen aufweisend betrachtet, und eine weitere Verarbeitung oder Auswertung für andere Bedingungen kann beendet werden, wodurch die Verarbeitungszeit verkürzt wird.

[0050] Ein Erstellen und Auswerten von Datensätzen für Daten-Deidentifikationsprozesse oder -techniken mit zahlreichen Sätzen von zugehörigen Konfigurationsoptionen kann eine erhebliche Verarbeitungszeit erfordern. Um die Verarbeitungsleistung zu verbessern und die Verarbeitungszeit für ein Deidentifizieren von Daten zu verkürzen, können in Ausführungsformen der vorliegenden Erfindung mehrere Techniken eingesetzt werden. Zum Beispiel können verschiedene Daten-Deidentifikationsprozesse und zugehörige Sätze von Konfigurationsoptionen durch einen Benutzer bereitgestellt und ausgewertet werden. Wenn eine oder mehrere dieser Konfigurationen von Daten-Deidentifikationsprozessen einen Datensatz ohne Datenschutz-Schwachstellen erzeugen, kann die Erstellung und Auswertung von Datensätzen beendet werden, die durch verbleibende Daten-Deidentifikationsprozesse und zugehörige Konfigurationen erzeugt werden, wodurch die Verarbeitungszeit verkürzt wird und Datenverarbeitungsressourcen geschont werden. Darüber hinaus können Grenzwerte bereitgestellt werden, die eine Menge von Konfigurationen für auszuwertende Daten-Deidentifikationsprozesse angeben.

[0051] Zusätzlich kann eine Baum- oder andere Datenstruktur erstellt werden, um die Erstellung und Auswertung von Datensätzen zu steuern, die durch Daten-Deidentifikationsprozesse und zugehörige Sätze von Konfigurationsoptionen erzeugt werden, wodurch die Datenverarbeitungsleistung erhöht und die Verarbeitungszeit verkürzt wird. Eine beispielhafte Datenstruktur in Form einer Baumstruktur ist in **Fig. 8** veranschaulicht. Als Beispiel stellt die Baumstruktur **800** die Sätze von Konfigurationsoptionen für einen Deidentifikationsprozess mit zwei Konfigurationsoptionen (z.B. einer Löschoption und einer Option zur Deidentifikation mit Datenerhaltung) für jedes von zwei Attributen (z.B. Name und Adresse) dar. Allerdings kann die Baumstruktur eine beliebige Menge von Konfigurationsoptionen für einen beliebigen Deidentifikationsprozess in Bezug auf eine beliebige Menge von beliebigen Attributen darstellen.

[0052] Die Baumstruktur **800** umfasst einen Wurzelknoten **805** und Unterbäume **810** und **830**. Jeder Knoten stellt einen entsprechenden Satz von Konfigurationsoptionen für den Daten-Deidentifikationsprozess dar und ist einer entsprechenden Vorlage zugehörig. Zum Beispiel kann ein Knoten **812** des Unterbaums **810** einen ersten Satz von Konfigurationsoptionen für ein erstes Attribut (z.B. Name löschen) darstellen, während ein Knoten **816** einen zweiten Satz von Konfigurationsoptionen für das erste Attribut darstellen kann (z.B. Name unter Wahrung der Geschlechtsangaben deidentifizieren). Untergeordnete Knoten **814**, **815** des Knotens **812** können jeweils den Satz von Konfigurationsoptionen des Knotens **812** sowie entsprechende Sätze von Konfigurationsoptionen für ein zweites Attribut darstellen (z.B. Name

und Adresse löschen (Knoten **814**), Name löschen und Adresse unter Wahrung von räumlicher Nähe deidentifizieren (Knoten **815**)). Untergeordnete Knoten **817**, **818** des Knotens **816** können jeweils den Satz von Konfigurationsoptionen des Knotens **816** sowie jeweilige Sätze von Konfigurationsoptionen für ein zweites Attribut darstellen (z.B. Name unter Wahrung von Geschlechtsangaben deidentifizieren und Adresse löschen (Knoten **817**), Name unter Wahrung von Geschlechtsangaben deidentifizieren und Adresse unter Wahrung von räumlicher Nähe deidentifizieren (Knoten **817**)).

[0053] In ähnlicher Weise kann ein Knoten **832** des Unterbaums **830** einen ersten Satz von Konfigurationsoptionen für das zweite Attribut darstellen (z.B. Adresse löschen), während ein Knoten **836** einen zweiten Satz von Konfigurationsoptionen für das zweite Attribut darstellen kann (z.B. Adresse unter Wahrung von räumlicher Nähe deidentifizieren). Untergeordnete Knoten **834** und **835** des Knotens **832** können jeweils den Satz von Konfigurationsoptionen des Knotens **832** sowie jeweilige Sätze von Konfigurationsoptionen für das erste Attribut darstellen (z.B. Adresse und Name löschen (Knoten **834**), Adresse löschen und Name unter Wahrung von Geschlechtsangaben deidentifizieren (Knoten **835**)). Untergeordnete Knoten **837**, **838** des Knotens **836** können jeweils den Satz von Konfigurationsoptionen des Knotens **836** sowie jeweilige Sätze von Konfigurationsoptionen für das erste Attribut darstellen (z.B. Adresse unter Wahrung von räumlicher Nähe deidentifizieren und Name löschen (Knoten **837**), Adresse unter Wahrung von räumlicher Nähe deidentifizieren und Name unter Wahrung des Geschlechts deidentifizieren (Knoten **838**)). Knoten mit überlappenden (oder denselben) Konfigurationsoptionen können zusammengefasst oder gekürzt werden, um einen Baum zu erzeugen, bei dem jeder Knoten einen anderen Satz von Konfigurationsoptionen aufweist.

[0054] Die untergeordneten Knoten jedes übergeordneten Knotens in dem Baum **800** stellen Konfigurationsoptionen dar, die Datensätze mit verallgemeinerteren Informationen im Vergleich zu ihren übergeordneten Knoten erzeugen. Zum Beispiel kann der Knoten **812** ein Namensattribut löschen, während ein untergeordneter Knoten **814** sowohl das Namensals auch das Adressattribut löschen kann, wodurch ein Datensatz mit weniger spezifischen (oder weiter deidentifizierten) Informationen erzeugt wird. Während der Verarbeitung wird der Baum **800** von dem Wurzelknoten **805** aus durchlaufen, und eine entsprechende Vorlage eines Zielknotens wird auf den Daten-Deidentifikationsprozess angewendet, um einen Datensatz zu erstellen. Wenn der erstellte Datensatz ausgewertet und als keine Datenschutz-Schwachstelle aufweisend ermittelt wird, werden die dem Zielknoten untergeordneten Knoten ebenfalls als keine Datenschutz-Schwachstelle aufweisend betrachtet, da die untergeordneten Knoten Konfigurationsoptionen zugehörig sind, die verallgemeinertere Datensätze erzeugen. Dementsprechend werden die untergeordneten Knoten als zulässige Konfigurationen für den Daten-Deidentifikationsprozess angegeben, ohne die Auswertung durchzuführen, wodurch die Verarbeitungszeit verkürzt wird.

[0055] Zum Beispiel kann eine Vorlage, die dem Knoten **812** entspricht, auf den Daten-Deidentifikationsprozess angewendet werden, um einen Datensatz zu erstellen, bei dem das Namensattribut gelöscht wurde. Wenn dieser Datensatz ausgewertet und als keine Datenschutz-Schwachstelle aufweisend ermittelt wird, weisen auch alle untergeordneten Knoten, die über die Namenslöschung hinaus eine zusätzliche Deidentifikation bereitstellen, keine Datenschutz-Schwachstelle auf (z.B. die Knoten **814**, **815**), da diese Knoten verallgemeinertere Daten erzeugen (z.B. Löschen des Namens und Löschen der Adresse (Knoten **814**), Löschen des Namens und Deidentifikation der Adresse (Knoten **815**)). Dementsprechend ist keine zusätzliche Verarbeitung erforderlich, um den durch die untergeordneten Knoten erzeugten Datensatz auszuwerten, wodurch die Verarbeitungszeit verkürzt wird.

[0056] Der Baum **800** kann verwendet werden, um eine Verarbeitung des Erstellens und/oder Auswertens von Datensätzen zu beenden. Die Verarbeitung kann für untergeordnete Knoten beendet werden, wenn ein übergeordneter Knoten einem zulässigen Satz von Konfigurationsoptionen zugehörig ist, die einen Datensatz mit minimalen oder keinen Datenschutz-Schwachstellen erstellen, wie oben beschrieben. Zum Beispiel können Datensätze für einen oder mehrere Daten-Deidentifikationsprozesse erstellt werden, und der Baum **800** kann genutzt werden, um die Menge von erstellten Datensätzen, die verarbeitet werden, für eine schnellere Auswertung der erstellten Datensätze zu minimieren. In diesem Fall werden, wenn ein übergeordneter Knoten einem zulässigen Satz von Konfigurationsoptionen zugehörig ist, die einen Datensatz mit minimalen oder keinen Datenschutz-Schwachstellen erstellen, die untergeordneten Knoten ohne weitere Auswertung als zulässig betrachtet.

[0057] Alternativ kann der Baum **800** genutzt werden, um jeweils einen Datensatz eines oder einiger weniger Knoten für einen Deidentifikationsprozess zu erstellen und auszuwerten. Dadurch wird die Menge von Malen minimiert, die der Deidentifikationsprozess ausgeführt wird, um den Datensatz zu erstellen, und die Menge von Auswertungen wird weiter verringert. In diesem Fall werden, wenn ein übergeordneter Knoten einem zulässigen Satz von Konfigurationsoptionen zugehörig ist, die einen Datensatz mit minimalen oder keinen Datenschutz-Schwachstellen erstellen, die untergeordneten Knoten ohne Erstellung des

Datensatzes und ohne weitere Auswertung als zulässig betrachtet.

[0058] Zusätzlich kann der Baum **800** Unterbäume mit Knoten höherer Ebene für alle oder einen Teil der Attribute enthalten. Alternativ kann jeder Unterbaum einen separaten Baum zur Auswertung eines Deidentifikationsprozesses bilden.

[0059] Man wird verstehen, dass die oben beschriebenen und in den Zeichnungen veranschaulichten Ausführungsformen lediglich einige wenige der vielen Wege zum Umsetzen von Ausführungsformen für eine Daten-Deidentifikation auf der Grundlage eines Erkennens von zulässigen Konfigurationen für Daten-Deidentifikationsprozesse darstellen.

[0060] Die Umgebung der Ausführungsformen der vorliegenden Erfindung kann jede beliebige Anzahl von Computer- oder anderen Verarbeitungssystemen (z.B. Client- oder Endbenutzer-Systeme, Serversysteme usw.) und Datenbanken oder andere auf eine beliebige gewünschte Weise angeordnete Repositorys beinhalten, wobei die Ausführungsformen der vorliegenden Erfindung auf jeden beliebigen gewünschten Typ von Datenverarbeitungsumgebung (z.B. Cloud-Computing, Client-Server, Netzwerkdatenverarbeitung, Mainframes, eigenständige Systeme usw.) angewendet werden können. Die von den Ausführungsformen der vorliegenden Erfindung eingesetzten Computer- oder anderen Verarbeitungssysteme können durch eine beliebige Anzahl beliebiger Personal- oder anderer Typen von Computern oder Verarbeitungssystemen (z.B. Desktops, Laptops, PDAs, mobile Einheiten usw.) umgesetzt werden, und zu ihnen können jedes beliebige im Handel erhältliche Betriebssystem und jede beliebige Kombination aus im Handel erhältlicher und angepasster Software (z.B. Browser-Software, Datenübertragungssoftware, Server-Software, ein Auswertungsmodul, Daten-Deidentifikationsprozesse usw.) gehören. Diese Systeme können alle beliebigen Typen von Monitoren und Eingabeeinheiten (z.B. Tastatur, Maus, Spracherkennung usw.) zum Eingeben und/oder Betrachten von Informationen beinhalten.

[0061] Es wird darauf hingewiesen, dass die Software (z.B. das Auswertungsmodul usw.) der Ausführungsformen der vorliegenden Erfindung in jeder beliebigen gewünschten Computersprache umgesetzt werden kann und von jedem Fachmann auf dem Gebiet von Computern beruhend auf den Funktionsbeschreibungen entwickelt werden könnte, die in der Beschreibung und den in den Zeichnungen veranschaulichten Ablaufplänen enthalten sind. Des Weiteren beziehen sich sämtliche Hinweise auf verschiedene Funktionen durchführende Software hierin allgemein auf Computersysteme oder Prozessoren, die diese Funktionen unter der Steuerung von Software durchführen. Die Computersysteme der Ausführungsformen der vorliegenden Erfindung können alternativ durch einen beliebigen Typ von Hardware und/oder anderem Verarbeitungsschaltkreis umgesetzt werden.

[0062] Die verschiedenen Funktionen der Computer- oder anderen Verarbeitungssysteme können auf eine beliebige Weise auf eine beliebige Anzahl von Software- und/oder Hardware-Modulen oder Einheiten, Verarbeitungs- oder Computersystemen und/oder Schaltkreisen verteilt sein, wobei sich die Computer- oder Verarbeitungssysteme lokal oder entfernt voneinander befinden und über ein beliebiges Datenübertragungsmedium (z.B. LAN, WAN, Intranet, Internet, festverdrahtet, Modem-Verbindung, drahtlos usw.) Daten austauschen können. Zum Beispiel können die Funktionen der Ausführungsformen der vorliegenden Erfindung auf eine beliebige Weise zwischen den verschiedenen Endbenutzer-/Client- und Serversystemen und/oder beliebigen anderen Zwischenverarbeitungseinheiten verteilt sein. Die Software und/oder Algorithmen, die oben beschrieben und in den Ablaufplänen veranschaulicht sind, können auf jede beliebige Weise abgeändert werden, welche die hierin beschriebenen Funktionen verwirklicht. Außerdem können die Funktionen in den Ablaufplänen oder der Beschreibung in einer beliebigen Reihenfolge durchgeführt werden, durch die ein gewünschter Arbeitsschritt verwirklicht wird.

[0063] Die Software der Ausführungsformen der vorliegenden Erfindung (z.B. das Auswertungsmodul usw.) kann auf einem nicht flüchtigen, durch einen Computer verwendbaren Medium (z.B. magnetische oder optische Medien, magnetooptische Medien, Floppy-Disketten, CD-ROM, DVD, Speichereinheiten usw.) einer stationären oder tragbaren Programmproduktvorrichtung oder -einheit zur Verwendung mit eigenständigen Systemen oder Systemen, die durch ein Netzwerk oder ein anderes Datenübertragungsmedium verbunden sind, verfügbar sein.

[0064] Das Datenübertragungsnetzwerk kann durch eine beliebige Anzahl eines beliebigen Typs von Datenübertragungsnetzwerken (z.B. LAN, WAN, Internet, Intranet, VPN usw.) umgesetzt sein. Die Computer- oder anderen Verarbeitungssysteme der Ausführungsformen der vorliegenden Erfindung können beliebige herkömmliche oder andere Datenübertragungseinheiten zum Austauschen von Daten über das Netzwerk mittels jedes beliebigen herkömmlichen oder anderen Protokolls beinhalten. Die Computer- oder anderen Verarbeitungssysteme können jeden beliebigen Verbindungstyp (z.B. verdrahtet, drahtlos usw.) für ein Zugreifen auf das Netzwerk nutzen. Lokale Datenübertragungsmedien können durch ein beliebiges geeignetes Datenübertragungsmedium (z.B. ein lokales Netzwerk (LAN), festverdrahtet, eine drahtlose Verbindung, ein Intranet usw.) umgesetzt sein.

[0065] Das System kann eine beliebige Anzahl von beliebigen herkömmlichen oder anderen Datenbanken, Datenspeichern oder Speicherstrukturen (z.B. Dateien, Datenbanken, Datenstrukturen, Daten- oder andere Repositorys usw.) zum Speichern von Informationen (z.B. ursprüngliche und vorläufige Datensätze, Konfigurationen oder Einstellungen, Optionen für Daten-Deidentifikationsprozesse usw.) einsetzen. Das Datenbanksystem kann durch eine beliebige Anzahl von beliebigen herkömmlichen oder anderen Datenbanken, Datenspeichern oder Speicherstrukturen (z.B. Dateien, Datenbanken, Datenstrukturen, Daten- oder andere Repositorys usw.) zum Speichern von Informationen umgesetzt werden. Das Datenbanksystem kann innerhalb der Server- und/oder Client-Systemen beinhaltet oder mit diesen verbunden sein. Die Datenbanksysteme und/oder Speicherstrukturen können entfernt angeordnet von oder lokal bei den Computer- oder anderen Verarbeitungssystemen sein und alle beliebigen gewünschten Daten speichern.

[0066] Die Ausführungsformen der vorliegenden Erfindung können eine beliebige Anzahl eines beliebigen Typs von Benutzerschnittstelle (z.B. eine grafische Benutzerschnittstelle (GUI), eine Befehlszeile, eine Aufforderung usw.) zum Erfassen oder Bereitstellen von Informationen (z.B. Benutzerpräferenzen, empfohlene Daten-Deidentifikationsprozesse, deidentifizierte Datensätze usw.) einsetzen, wobei die Schnittstelle beliebige in einer beliebigen Weise angeordnete Informationen beinhalten kann. Die Schnittstelle kann eine beliebige Anzahl von beliebigen Typen von Eingabe- oder Betätigungsmechanismen (z.B. Schaltflächen, Symbole, Felder, Boxen, Verknüpfungen usw.) zum Eingeben/Anzeigen von Informationen und zum Initiieren von gewünschten Aktionen mittels beliebiger geeigneter Eingabeeinheiten (z.B. Maus, Tastatur usw.) beinhalten, die sich an beliebigen Stellen befinden können. Die Schnittstellenbildschirme können beliebige geeignete Aktuatoren (z.B. Verknüpfungen, Registerkarten usw.) zum Navigieren zwischen den Bildschirmen auf eine beliebige Weise beinhalten.

[0067] Der Bericht kann beliebige auf eine beliebige Weise angeordnete Informationen umfassen und beruhend auf Regeln oder anderen Kriterien konfigurierbar sein, um einem Benutzer gewünschte Informationen bereitzustellen (z.B. Empfehlungen, Datenschutzprobleme usw.).

[0068] Die Ausführungsformen der vorliegenden Erfindung sind nicht auf die konkreten oben beschriebenen Aufgaben oder Algorithmen beschränkt, sondern können zum Auswerten von beliebigen Daten-Deidentifikations- oder Anonymisierungsprozessen oder -techniken für alle beliebigen Typen von Identifikatoren verwendet werden. Die Daten-Deidentifikationsprozesse können allen beliebigen Typen von Konfigurationsoptionen zum Löschen oder Deidentifizieren von beliebigen Attributen zugehörig sein. Die Sätze von Konfigurationsoptionen und Vorlagen können eine beliebige Menge von beliebigen Konfigurationsoptionen für einen Daten-Deidentifikationsprozess angeben.

[0069] Die erstellten Datensätze können in einer beliebigen Art und Weise ausgewertet werden, um eine beliebige Menge von beliebigen Typen von Datenschutz-Schwachstellen zu identifizieren. Die Daten der erstellten Datensätze können anhand von beliebigen Typen von bekannten oder anderen Datensätzen geprüft werden (z.B. von Benutzern bereitgestellte Datensätze, öffentlich verfügbare Datensätze, organisationsinterne Datensätze usw.). Ein erstellter Datensatz kann infolge eines Identifizierens einer beliebigen Menge von beliebigen Typen von Datenschutz-Schwachstellen als mit Schwachstellen behaftet angesehen werden (z.B. einer beliebigen Menge von identifizierten Entitäten, einer beliebigen Menge von eingebrachten Quasi-Identifikatoren usw.). Der Schwellenwert zum Erkennen einer Schwachstelle kann auf alle beliebigen Werte gesetzt werden (z.B. eine Menge von Verknüpfungen, eine Menge von Quasi-Identifikatoren, eine Menge von Datenschutz-Schwachstellen usw.). Die Mengen können in jeder beliebigen gewünschten Weise mit dem Schwellenwert verglichen werden, um auf eine Datenschutz-Schwachstelle hinzuweisen (z.B. größer, kleiner, größer oder gleich, kleiner oder gleich usw.).

[0070] Es kann jede beliebige Datenstruktur genutzt werden, um Beziehungen zwischen Sätzen von Konfigurationsoptionen zu identifizieren (z.B. Baum, hierarchische Struktur usw.). Eine Verarbeitung für eine beliebige Menge von verwandten Konfigurationsoptionen kann als Reaktion darauf, dass eine Ausgangskonfiguration einen Datensatz mit minimalen oder keinen Datenschutz-Schwachstellen erstellt, beendet werden. Die Datenstruktur kann in einer beliebigen Weise durchlaufen werden, um die Konfigurationsoptionen für einen Daten-Deidentifikationsprozess auszuwerten. Für einen Satz von Attributen kann eine beliebige Menge von Daten-Deidentifikationsprozessen und zugehörigen Sätzen von Konfigurationsoptionen empfohlen oder ausgewählt werden. Zum Beispiel können die gleichen oder unterschiedliche Daten-Deidentifikationsprozesse (und entsprechende Konfigurationen) auf verschiedene Attribute in einem Datensatz angewendet werden.

[0071] Die hierin verwendete Terminologie dient lediglich dem Zweck des Beschreibens bestimmter Ausführungsformen und soll die Erfindung nicht einschränken. Die Verwendung der Singularform „ein“, „eine“ bzw. „der“, „die“, „das“ hierin soll ebenfalls die Pluralformen einschließen, es sei denn, etwas anderes ergibt sich deutlich aus dem Zusammenhang. Es wird ferner darauf hingewiesen, dass die

Begriffe „aufweisen“, „aufweisend“, „beinhaltet“, „beinhaltend“, „hat“, „haben“, „habend“, „mit“ und dergleichen, wenn sie in dieser Beschreibung verwendet werden, das Vorhandensein von aufgeführten Eigenschaften, ganzen Zahlen, Schritten, Arbeitsschritten, Elementen und/oder Komponenten angeben, jedoch nicht das Vorhandensein oder das Hinzufügen einer oder mehrerer anderer Eigenschaften, ganzer Zahlen, Schritte, Arbeitsschritte, Elemente, Komponenten und/oder Gruppen hiervon ausschließen.

[0072] Die in den nachfolgenden Ansprüchen vorhandenen, entsprechenden Strukturen, Materialien, Schritte und Entsprechungen aller Mittel oder Step-plus-function-Elemente verstehen sich dahingehend, dass sie jede beliebige Struktur, jedes beliebige Material bzw. jeden beliebigen Schritt zur Durchführung der Funktion in Kombination mit anderen beanspruchten Elementen nach Maßgabe der konkreten Beanspruchung aufweisen. Die Beschreibung der vorliegenden Erfindung wurde zum Zwecke der Veranschaulichung und Beschreibung aufgeführt, soll jedoch nicht gesamthaft stehen für bzw. begrenzt sein auf die Erfindung in der beschriebenen Form. Für Fachleute werden viele Abänderungen und Abwandlungen ersichtlich sein, ohne von dem Umfang und dem Sinngehalt der Erfindung abzuweichen. Die Ausführungsform wurde gewählt und beschrieben, um die Grundgedanken der Erfindung und die praktische Anwendung bestmöglich zu erläutern und um es anderen Fachleuten zu ermöglichen, die Erfindung für verschiedene Ausführungsformen mit verschiedenen Abänderungen, die für eine bestimmte in Betracht gezogene Verwendung geeignet sind, zu verstehen.

[0073] Die Beschreibungen der verschiedenen Ausführungsformen der vorliegenden Erfindung wurden zum Zwecke der Veranschaulichung aufgeführt, sollen jedoch nicht gesamthaft stehen für bzw. begrenzt sein auf die offenbarten Ausführungsformen. Für Fachleute werden viele Abänderungen und Abwandlungen ersichtlich sein, ohne von dem Umfang und dem Sinngehalt der beschriebenen Ausführungsformen abzuweichen. Die hierin verwendete Terminologie wurde gewählt, um die Grundgedanken der Ausführungsformen, die praktische Anwendung oder technische Verbesserung gegenüber auf dem Markt vorgefundenen Technologien bestmöglich zu erläutern oder um es anderen Fachleuten zu ermöglichen, die hierin offenbarten Ausführungsformen zu verstehen.

[0074] Bei der vorliegenden Erfindung kann es sich um ein System, ein Verfahren und/oder ein Computerprogrammprodukt auf jedem möglichen technischen Detaillierungsgrad von Integration handeln. Das Computerprogrammprodukt kann (ein) durch einen Computer lesbare(s) Speichermedium (oder -medien) umfassen, auf dem/denen durch einen Computer lesbare Programmanweisungen gespeichert ist/sind, um einen Prozessor dazu zu veranlassen, Aspekte der vorliegenden Erfindung auszuführen.

[0075] Bei dem durch einen Computer lesbaren Speichermedium kann es sich um eine physische Einheit handeln, die Anweisungen zur Verwendung durch ein System zur Ausführung von Anweisungen behalten und speichern kann. Bei dem durch einen Computer lesbaren Speichermedium kann es sich zum Beispiel um eine elektronische Speichereinheit, eine magnetische Speichereinheit, eine optische Speichereinheit, eine elektromagnetische Speichereinheit, eine Halbleiterspeichereinheit oder jede geeignete Kombination daraus handeln, ohne auf diese beschränkt zu sein. Zu einer nicht erschöpfenden Liste spezifischerer Beispiele des durch einen Computer lesbaren Speichermediums gehören die Folgenden: eine tragbare Computerdiskette, eine Festplatte, ein Direktzugriffsspeicher (RAM), ein Nur-Lese-Speicher (ROM), ein löschbarer programmierbarer Nur-Lese-Speicher (EPROM bzw. Flash-Speicher), ein statischer Direktzugriffsspeicher (SRAM), ein tragbarer Kompaktspeicherplatte-Nur-Lese-Speicher (CD-ROM), eine DVD (digital versatile disc), ein Speicher-Stick, eine Diskette, eine mechanisch kodierte Einheit wie zum Beispiel Lochkarten oder erhabene Strukturen in einer Rille, auf denen Anweisungen gespeichert sind, und jede geeignete Kombination daraus. Ein durch einen Computer lesbares Speichermedium soll in der Verwendung hierin nicht als flüchtige Signale an sich aufgefasst werden, wie zum Beispiel Funkwellen oder andere sich frei ausbreitende elektromagnetische Wellen, elektromagnetische Wellen, die sich durch einen Wellenleiter oder ein anderes Übertragungsmedium ausbreiten (z.B. durch ein Glasfaserkabel geleitete Lichtimpulse) oder durch einen Draht übertragene elektrische Signale.

[0076] Hierin beschriebene, durch einen Computer lesbare Programmanweisungen können von einem durch einen Computer lesbaren Speichermedium auf jeweilige Datenverarbeitungs-/Verarbeitungseinheiten oder über ein Netzwerk wie zum Beispiel das Internet, ein lokales Netzwerk, ein Weitverkehrsnetz und/oder ein drahtloses Netzwerk auf einen externen Computer oder eine externe Speichereinheit heruntergeladen werden. Das Netzwerk kann Kupferübertragungskabel, Lichtwellenübertragungsleiter, drahtlose Übertragung, Leitwegrechner, Firewalls, Vermittlungseinheiten, Gateway-Computer und/oder Edge-Server aufweisen. Eine Netzwerkadapterkarte oder Netzwerkschnittstelle in jeder Datenverarbeitungs-/Verarbeitungseinheit empfängt durch einen Computer lesbare Programmanweisungen aus dem Netzwerk und leitet die durch einen Computer lesbaren Programmanweisungen zur Speicherung in einem durch einen Computer lesbaren Speichermedium innerhalb der entsprechenden Datenverarbeitungs-/Verarbeitungseinheit weiter.

[0077] Bei durch einen Computer lesbaren Programmanweisungen zum Ausführen von Arbeitsschritten der vorliegenden Erfindung kann es sich um Assembler-Anweisungen, ISA-Anweisungen (Instruction-Set-Architecture), Maschinenanweisungen, maschinenabhängige Anweisungen, Mikrocode, Firmware-Anweisungen, zustandssetzende Daten, Konfigurationsdaten für integrierte Schaltungen oder entweder Quellcode oder Objektcode handeln, die in einer beliebigen Kombination aus einer oder mehreren Programmiersprachen geschrieben werden, darunter objektorientierte Programmiersprachen wie Smalltalk, C++ o.ä. sowie herkömmliche prozedurale Programmiersprachen wie die Programmiersprache „C“ oder ähnliche Programmiersprachen. Die durch einen Computer lesbaren Programmanweisungen können vollständig auf dem Computer des Benutzers, teilweise auf dem Computer des Benutzers, als eigenständiges Software-Paket, teilweise auf dem Computer des Benutzers und teilweise auf einem fernen Computer oder vollständig auf dem fernen Computer oder Server ausgeführt werden. In letzterem Fall kann der entfernt angeordnete Computer mit dem Computer des Benutzers durch eine beliebige Art Netzwerk verbunden sein, darunter ein lokales Netzwerk (LAN) oder ein Weitverkehrsnetz (WAN), oder die Verbindung kann mit einem externen Computer hergestellt werden (zum Beispiel über das Internet unter Verwendung eines Internet-Dienstanbieters). In einigen Ausführungsformen können elektronische Schaltungen, darunter zum Beispiel programmierbare Logikschaltungen, im Feld programmierbare Gatter-Anordnungen (FPGA, field programmable gate arrays) oder programmierbare Logikanordnungen (PLA, programmable logic arrays) die durch einen Computer lesbaren Programmanweisungen ausführen, indem sie Zustandsinformationen der durch einen Computer lesbaren Programmanweisungen nutzen, um die elektronischen Schaltungen zu personalisieren, um Aspekte der vorliegenden Erfindung durchzuführen.

[0078] Aspekte der vorliegenden Erfindung sind hierin unter Bezugnahme auf Ablaufpläne und/oder Blockschaltbilder bzw. Schaubilder von Verfahren, Vorrichtungen (Systemen) und Computerprogrammprodukten gemäß Ausführungsformen der Erfindung beschrieben. Es wird darauf hingewiesen, dass jeder Block der Ablaufpläne und/oder der Blockschaltbilder bzw. Schaubilder sowie Kombinationen von Blöcken in den Ablaufplänen und/oder den Blockschaltbildern bzw. Schaubildern mittels durch einen Computer lesbare Programmanweisungen ausgeführt werden können.

[0079] Diese durch einen Computer lesbaren Programmanweisungen können einem Prozessor eines Universalcomputers, eines Spezialcomputers oder einer anderen programmierbaren Datenverarbeitungsvorrichtung bereitgestellt werden, um eine Maschine zu erzeugen, so dass die über den Prozessor des Computers bzw. der anderen programmierbaren Datenverarbeitungsvorrichtung ausgeführten Anweisungen ein Mittel zur Umsetzung der in dem Block bzw. den Blöcken der Ablaufpläne und/oder der Blockschaltbilder bzw. Schaubilder festgelegten Funktionen/Schritte erzeugen. Diese durch einen Computer lesbaren Programmanweisungen können auch auf einem durch einen Computer lesbaren Speichermedium gespeichert sein, das einen Computer, eine programmierbare Datenverarbeitungsvorrichtung und/oder andere Einheiten so steuern kann, dass sie auf eine bestimmte Art funktionieren, so dass das durch einen Computer lesbare Speichermedium, auf dem Anweisungen gespeichert sind, ein Herstellungsprodukt aufweist, darunter Anweisungen, welche Aspekte der/des in dem Block bzw. den Blöcken des Ablaufplans und/oder der Blockschaltbilder bzw. Schaubilder angegebenen Funktion/Schritts umsetzen.

[0080] Die durch einen Computer lesbaren Programmanweisungen können auch auf einen Computer, eine andere programmierbare Datenverarbeitungsvorrichtung oder eine andere Einheit geladen werden, um das Ausführen einer Reihe von Prozessschritten auf dem Computer bzw. der anderen programmierbaren Vorrichtung oder anderen Einheit zu verursachen, um einen auf einem Computer ausgeführten Prozess zu erzeugen, so dass die auf dem Computer, einer anderen programmierbaren Vorrichtung oder einer anderen Einheit ausgeführten Anweisungen die in dem Block bzw. den Blöcken der Ablaufpläne und/oder der Blockschaltbilder bzw. Schaubilder festgelegten Funktionen/Schritte umsetzen.

[0081] Die Ablaufpläne und die Blockschaltbilder bzw. Schaubilder in den Figuren veranschaulichen die Architektur, die Funktionalität und den Betrieb möglicher Ausführungen von Systemen, Verfahren und Computerprogrammprodukten gemäß verschiedenen Ausführungsformen der vorliegenden Erfindung. In diesem Zusammenhang kann jeder Block in den Ablaufplänen oder Blockschaltbildern bzw. Schaubildern ein Modul, ein Segment oder einen Teil von Anweisungen darstellen, die eine oder mehrere ausführbare Anweisungen zur Ausführung der bestimmten logischen Funktion(en) aufweisen. In einigen alternativen Ausführungen können die in dem Block angegebenen Funktionen in einer anderen Reihenfolge als in den Figuren gezeigt stattfinden. Zwei nacheinander gezeigte Blöcke können zum Beispiel in Wirklichkeit im Wesentlichen gleichzeitig ausgeführt werden, oder die Blöcke können manchmal je nach entsprechender Funktionalität in umgekehrter Reihenfolge ausgeführt werden. Es ist ferner anzumerken, dass jeder Block der Blockschaltbilder bzw. Schaubilder und/oder der Ablaufpläne sowie Kombinationen aus Blöcken in den Blockschaltbildern bzw. Schaubildern und/oder den Ablaufplänen durch spe-

zielle auf Hardware beruhende Systeme umgesetzt werden können, welche die festgelegten Funktionen oder Schritte durchführen, oder Kombinationen aus Spezial-Hardware und Computeranweisungen ausführen.

Patentansprüche

1. Verfahren zum Deidentifizieren von Daten, aufweisend:
Ermitteln einer oder mehrerer Identifikatoren, die eine Entität eines Datensatzes identifizieren;
Identifizieren eines oder mehrerer Daten-Deidentifikationsprozesse, die dem ermittelten einen oder den ermittelten mehreren Identifikatoren zugehörig sind, wobei jeder Daten-Deidentifikationsprozess einem oder mehreren Sätzen von Konfigurationsoptionen zugehörig ist, die Informationen angeben, die in dem Datensatz zu bewahren sind;
Ausführen, mittels eines Prozessors, der identifizierten Daten-Deidentifikationsprozesse an dem Datensatz gemäß den zugehörigen Sätzen von Konfigurationsoptionen, um Datensätze mit unterschiedlichen bewahrten Informationen zu erstellen;
Auswerten, mittels eines Prozessors, der erstellten Datensätze auf Datenschutz-Schwachstellen hin und Auswählen eines Daten-Deidentifikationsprozesses und eines zugehörigen Satzes von Konfigurationsoptionen auf der Grundlage der Auswertung; und
Ausführen, mittels eines Prozessors, des ausgewählten Daten-Deidentifikationsprozesses an dem Datensatz gemäß dem zugehörigen Satz von Konfigurationsoptionen, um einen sich ergebenden deidentifizierten Datensatz zu erzeugen.

2. Verfahren nach Anspruch 1, wobei das Ermitteln des einen oder der mehreren Identifikatoren darüber hinaus Folgendes aufweist:
Ermitteln eines oder mehrerer direkter Identifikatoren, wobei die zugehörigen Daten-Deidentifikationsprozesse Datenmaskierungsprozesse umfassen.

3. Verfahren nach Anspruch 1, wobei das Ermitteln des einen oder der mehreren Identifikatoren darüber hinaus Folgendes aufweist:
Ermitteln einer Mehrzahl von Quasi-Identifikatoren, wobei die zugehörigen Daten-Deidentifikationsprozesse eine Datenverallgemeinerung oder Datenunterdrückung umfassen.

4. Verfahren nach Anspruch 1, wobei die erstellten Datensätze in Form einer Tabelle vorliegen und das Ausführen der identifizierten Daten-Deidentifikationsprozesse darüber hinaus Folgendes aufweist:
Zusammenfassen von zwei oder mehr Spalten eines erstellten Datensatzes, um eine Spalte mit Informationen zu erzeugen, die spezifischer sind als die zwei oder mehr Spalten.

5. Verfahren nach Anspruch 1, wobei das Auswerten der erstellten Datensätze auf Datenschutz-Schwachstellen hin darüber hinaus Folgendes aufweist:
Ermitteln eines Vorhandenseins einer Verknüpfung zwischen Daten für eine Entität in einem erstellten Datensatz und Daten für eine bekannte Entität in einem öffentlich verfügbaren Datensatz, um auf eine Datenschutz-Schwachstelle für den erstellten Datensatz hinzuweisen.

6. Verfahren nach Anspruch 1, wobei das Auswerten der erstellten Datensätze auf Datenschutz-Schwachstellen hin darüber hinaus Folgendes aufweist:
Ermitteln eines Vorhandenseins eines Satzes von Quasi-Identifikatoren in einem erstellten Datensatz, der durch einen entsprechenden Daten-Deidentifikationsprozess und einen zugehörigen Satz von Konfigurationsoptionen eingebracht wurde, um auf eine Datenschutz-Schwachstelle für den erstellten Datensatz hinzuweisen.

7. Verfahren nach Anspruch 1, darüber hinaus aufweisend:
Erzeugen einer Reihe von Vorlagen für jeden Daten-Deidentifikationsprozess, wobei jede Vorlage einen zugehörigen Satz von Konfigurationsoptionen für diesen Daten-Deidentifikationsprozess angibt.

8. Verfahren nach Anspruch 1, darüber hinaus aufweisend:
Verkürzen einer Verarbeitungszeit für die Deidentifikation durch Identifizieren eines erstellten Datensatzes, der keine Datenschutz-Schwachstellen aufweist, und Beenden einer Verarbeitung in Bezug auf andere zugehörige Sätze von Konfigurationsoptionen für einen entsprechenden Daten-Deidentifikationsprozess, der Datensätze mit verallgemeinerteren Informationen als der identifizierte erstellte Datensatz erstellt.

9. System zum Deidentifizieren von Daten, aufweisend:
mindestens einen Prozessor, der konfiguriert ist zum: Ermitteln eines oder mehrerer Identifikatoren, die eine Entität eines Datensatzes identifizieren; Identifizieren eines oder mehrerer Daten-Deidentifikationsprozesse, die dem ermittelten einen oder den ermittelten mehreren Identifikatoren zugehörig sind, wobei jeder Daten-Deidentifikationsprozess einem oder mehreren Sätzen von Konfigurationsoptionen zugehörig ist, die Informationen angeben, die in dem Datensatz zu bewahren sind;
Ausführen der identifizierten Daten-Deidentifikationsprozesse an dem Datensatz gemäß den zugehörigen Sätzen von Konfigurationsoptionen, um Datensätze mit unterschiedlichen bewahrten Informationen zu erstellen;

Auswerten der erstellten Datensätze auf Datenschutz-Schwachstellen hin und Auswählen eines Daten-Deidentifikationsprozesses und eines zugehörigen Satzes von Konfigurationsoptionen auf der Grundlage der Auswertung; und
Ausführen des ausgewählten Daten-Deidentifikationsprozesses an dem Datensatz gemäß dem zugehörigen Satz von Konfigurationsoptionen, um einen sich ergebenden deidentifizierten Datensatz zu erzeugen.

10. System nach Anspruch 9, wobei das Ermitteln des einen oder der mehreren Identifikatoren darüber hinaus Folgendes aufweist:
Ermitteln eines oder mehrerer direkter Identifikatoren, wobei die zugehörigen Daten-Deidentifikationsprozesse Datenmaskierungsprozesse umfassen.

11. System nach Anspruch 9, wobei das Ermitteln des einen oder der mehreren Identifikatoren darüber hinaus Folgendes aufweist:
Ermitteln einer Mehrzahl von Quasi-Identifikatoren, wobei die zugehörigen Daten-Deidentifikationsprozesse eine Datenverallgemeinerung oder Datenunterdrückung umfassen.

12. System nach Anspruch 9, wobei die erstellten Datensätze in Form einer Tabelle vorliegen und das Ausführen der identifizierten Daten-Deidentifikationsprozesse darüber hinaus Folgendes aufweist:
Zusammenfassen von zwei oder mehr Spalten eines erstellten Datensatzes, um eine Spalte mit Informationen zu erzeugen, die spezifischer sind als die zwei oder mehr Spalten.

13. System nach Anspruch 9, wobei das Auswerten der erstellten Datensätze auf Datenschutz-Schwachstellen hin darüber hinaus Folgendes aufweist:
Ermitteln eines Vorhandenseins einer Verknüpfung zwischen Daten für eine Entität in einem erstellten Datensatz und Daten für eine bekannte Entität in einem öffentlich verfügbaren Datensatz, um auf eine Datenschutz-Schwachstelle für den erstellten Datensatz hinzuweisen.

14. System nach Anspruch 9, wobei das Auswerten der erstellten Datensätze auf Datenschutz-Schwachstellen hin darüber hinaus Folgendes aufweist:
Ermitteln eines Vorhandenseins eines Satzes von Quasi-Identifikatoren in einem erstellten Datensatz, der durch einen entsprechenden Daten-Deidentifikationsprozess und einen zugehörigen Satz von Konfigurationsoptionen eingebracht wurde, um auf eine Datenschutz-Schwachstelle für den erstellten Datensatz hinzuweisen.

15. System nach Anspruch 9, bei dem der mindestens eine Prozessor darüber hinaus konfiguriert ist zum:
Erzeugen einer Reihe von Vorlagen für jeden Daten-Deidentifikationsprozess, wobei jede Vorlage einen zugehörigen Satz von Konfigurationsoptionen für diesen Daten-Deidentifikationsprozess festlegt.

16. System nach Anspruch 9, bei dem der mindestens eine Prozessor darüber hinaus konfiguriert ist zum:
Verkürzen einer Verarbeitungszeit für die Deidentifikation durch Identifizieren eines erstellten Datensatzes, der keine Datenschutz-Schwachstellen aufweist, und Beenden einer Verarbeitung in Bezug auf andere zugehörige Sätze von Konfigurationsoptionen für einen entsprechenden Daten-Deidentifikationsprozess, der Datensätze mit verallgemeinerteren Informationen als der identifizierte erstellte Datensatz erstellt.

17. Computerprogrammprodukt zum Deidentifizieren von Daten, wobei das Computerprogrammprodukt ein durch einen Computer lesbares Speichermedium aufweist, auf dem durch einen Computer lesbarer Programmcode enthalten ist, wobei der durch einen Computer lesbare Programmcode durch mindestens einen Prozessor ausführbar ist, um den mindestens einen Prozessor veranlassen zum:
Ermitteln eines oder mehrerer Identifikatoren, die eine Entität eines Datensatzes identifizieren;
Identifizieren eines oder mehrerer Daten-Deidentifikationsprozesse, die dem ermittelten einen oder den ermittelten mehreren Identifikatoren zugehörig sind, wobei jeder Daten-Deidentifikationsprozess einem oder mehreren Sätzen von Konfigurationsoptionen zugehörig ist, die Informationen angeben, die in dem Datensatz zu bewahren sind;
Ausführen der identifizierten Daten-Deidentifikationsprozesse an dem Datensatz gemäß den zugehörigen Sätzen von Konfigurationsoptionen, um Datensätze mit unterschiedlichen bewahrten Informationen zu erstellen;
Auswerten der erstellten Datensätze auf Datenschutz-Schwachstellen hin und Auswählen eines Daten-Deidentifikationsprozesses und eines zugehörigen Satzes von Konfigurationsoptionen auf der Grundlage der Auswertung; und
Ausführen des ausgewählten Daten-Deidentifikationsprozesses an dem Datensatz gemäß dem zugehörigen Satz von Konfigurationsoptionen, um einen sich ergebenden deidentifizierten Datensatz zu erzeugen.

18. Computerprogrammprodukt nach Anspruch 17, wobei das Ermitteln des einen oder der mehreren Identifikatoren darüber hinaus Folgendes aufweist:
Ermitteln eines oder mehrerer direkter Identifikatoren, und die zugehörigen Daten-Deidentifikationsprozesse umfassen Datenmaskierungsprozesse.

19. Computerprogrammprodukt nach Anspruch 17, wobei das Ermitteln des einen oder der mehreren Identifikatoren darüber hinaus Folgendes aufweist:
Ermitteln einer Mehrzahl von Quasi-Identifikatoren, wobei die zugehörigen Daten-Deidentifikationsprozesse eine Datenverallgemeinerung oder Datenunterdrückung umfassen.

20. Computerprogrammprodukt nach Anspruch 17, wobei die erstellten Datensätze in Form einer Tabelle vorliegen und das Ausführen der identifizierten Daten-Deidentifikationsprozesse darüber hinaus Folgendes aufweist:
Zusammenfassen von zwei oder mehr Spalten eines erstellten Datensatzes, um eine Spalte mit Informationen zu erzeugen, die spezifischer sind als die zwei oder mehr Spalten.

21. Computerprogrammprodukt nach Anspruch 17, wobei das Auswerten der erstellten Datensätze auf Datenschutz-Schwachstellen hin darüber hinaus Folgendes aufweist:
Ermitteln eines Vorhandenseins einer Verknüpfung zwischen Daten für eine Entität in einem erstellten Datensatz und Daten für eine bekannte Entität in einem öffentlich verfügbaren Datensatz, um auf eine Datenschutz-Schwachstelle für den erstellten Datensatz hinzuweisen.

22. Computerprogrammprodukt nach Anspruch 17, wobei das Auswerten der erstellten Datensätze auf Datenschutz-Schwachstellen hin darüber hinaus Folgendes aufweist:
Ermitteln eines Vorhandenseins eines Satzes von Quasi-Identifikatoren in einem erstellten Datensatz, der durch einen entsprechenden Daten-Deidentifikationsprozess und einen zugehörigen Satz von Konfigurationsoptionen eingebracht wurde, um auf eine Datenschutz-Schwachstelle für den erstellten Datensatz hinzuweisen.

23. Computerprogrammprodukt nach Anspruch 17, wobei der durch einen Computer lesbare Programmcode darüber hinaus den mindestens einen Prozessor veranlasst zum:
Erzeugen einer Reihe von Vorlagen für jeden Daten-Deidentifikationsprozess, wobei jede Vorlage einen zugehörigen Satz von Konfigurationsoptionen für diesen Daten-Deidentifikationsprozess festlegt.

24. Computerprogrammprodukt nach Anspruch 17, wobei der durch einen Computer lesbare Programmcode darüber hinaus den mindestens einen Prozessor veranlasst zum:
Verkürzen einer Verarbeitungszeit für die Deidentifikation durch Identifizieren eines erstellten Datensatzes, der keine Datenschutz-Schwachstellen aufweist, und Beenden einer Verarbeitung in Bezug auf andere zugehörige Sätze von Konfigurationsoptionen für einen entsprechenden Daten-Deidentifikationsprozess, der Datensätze mit verallgemeinerteren Informationen als der identifizierte erstellte Datensatz erstellt.

Es folgen 8 Seiten Zeichnungen

Anhängende Zeichnungen

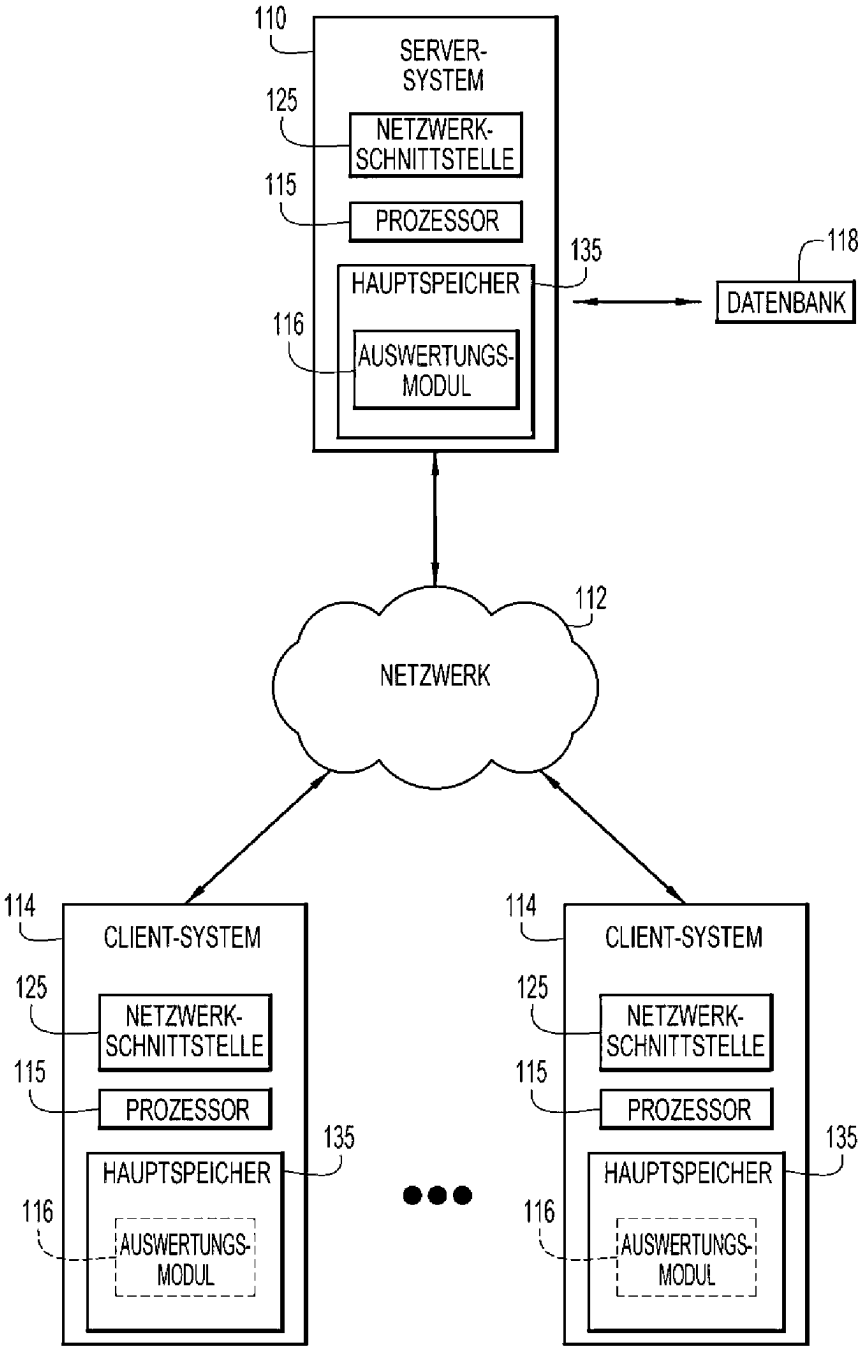


FIG.1

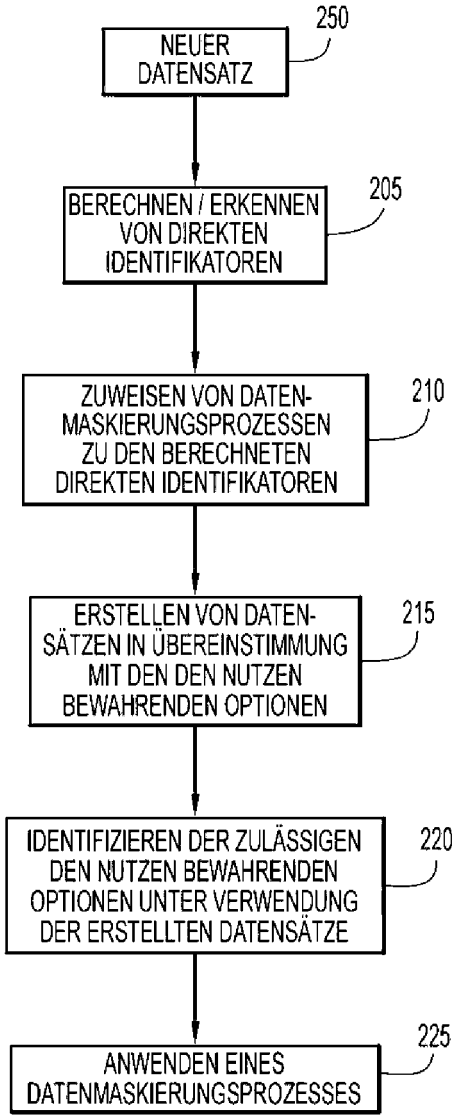

250
NEUER DATENSATZ
205
BERECHNEN / ERKENNEN VON DIREKTEN IDENTIFIKATOREN
210
ZUWEISEN VON DATEN-MASKIERUNGSPROZESSEN ZU DEN BERECHNETEN DIREKTEN IDENTIFIKATOREN
215
ERSTELLEN VON DATEN-SÄTZEN IN ÜBEREINSTIMMUNG MIT DEN DEN NUTZEN BEWAHRENDEN OPTIONEN
220
IDENTIFIZIEREN DER ZULÄSSIGEN DEN NUTZEN BEWAHRENDEN OPTIONEN UNTER VERWENDUNG DER ERSTELLTEN DATENSÄTZE
225
ANWENDEN EINES DATENMASKIERUNGSPROZESSES

FIG.2

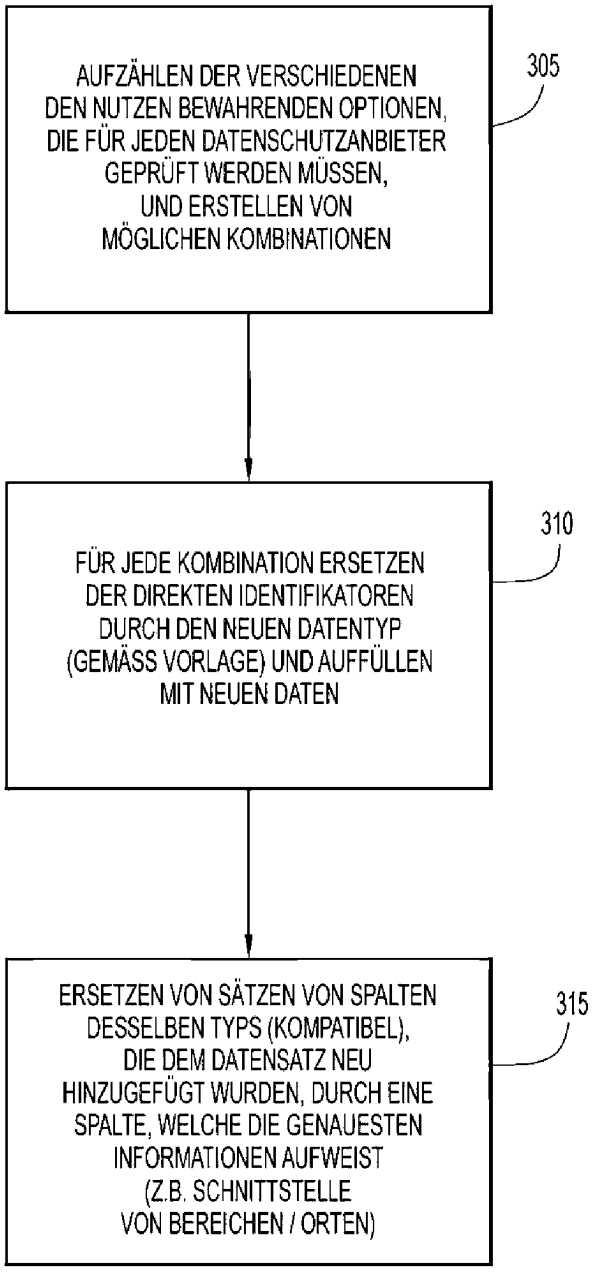

AUFZÄHLEN DER VERSCHIEDENEN DEN NUTZEN BEWAHRENDEN OPTIONEN, DIE FÜR JEDEN DATENSCHUTZANBIETER GEPRÜFT WERDEN MÜSSEN, UND ERSTELLEN VON MÖGLICHEN KOMBINATIONEN
305
FÜR JEDE KOMBINATION ERSETZEN DER DIREKTEN IDENTIFIKATOREN DURCH DEN NEUEN DATENTYP (GEMÄSS VORLAGE) UND AUFFÜLLEN MIT NEUEN DATEN
310
ERSETZEN VON SÄTZEN VON SPALTEN DESSELBEN TYPS (KOMPATIBEL), DIE DEM DATENSATZ NEU HINZUGEFÜGT WURDEN, DURCH EINE SPALTE, WELCHE DIE GENAUESTEN INFORMATIONEN AUFWEIST (Z.B. SCHNITTSTELLE VON BEREICHEN / ORTEN)
315

FIG.3

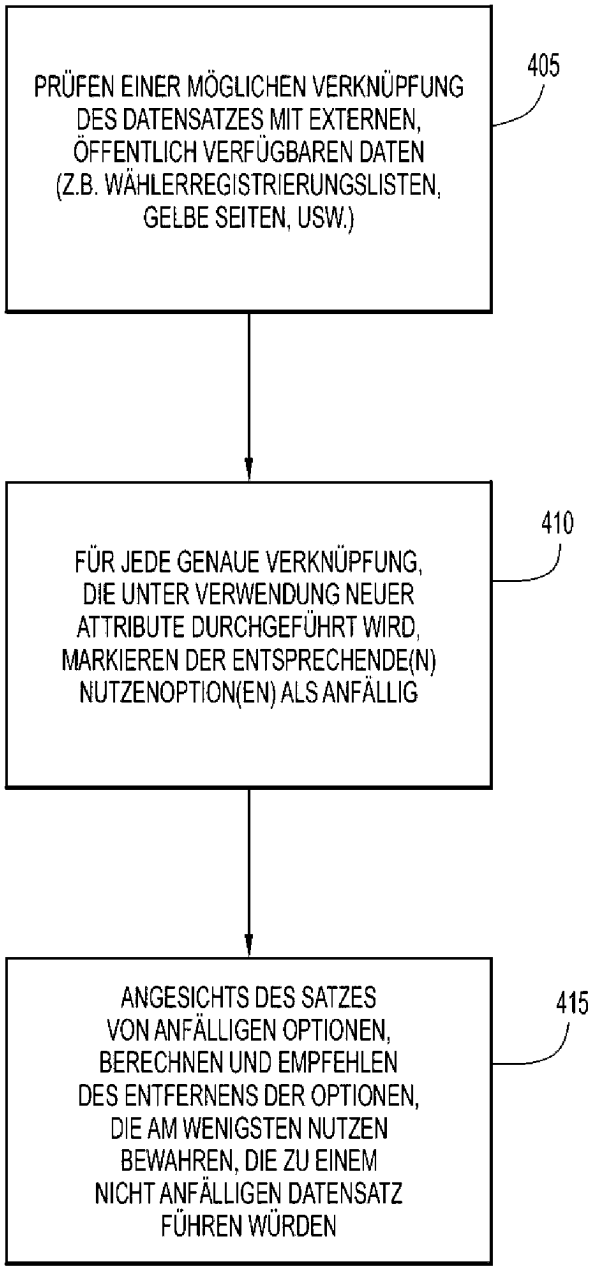

PRÜFEN EINER MÖGLICHEN VERKNÜPFUNG DES DATENSATZES MIT EXTERNEN, ÖFFENTLICH VERFÜGBAREN DATEN (Z.B. WÄHLERREGISTRIERUNGSLISTEN, GELBE SEITEN, USW.)
405
FÜR JEDE GENAUE VERKNÜPFUNG, DIE UNTER VERWENDUNG NEUER ATTRIBUTE DURCHGEFÜHRT WIRD, MARKIEREN DER ENTSPRECHENDE(N) NUTZENOPTION(EN) ALS ANFÄLLIG
410
ANGESICHTS DES SATZES VON ANFÄLLIGEN OPTIONEN, BERECHNEN UND EMPFEHLEN DES ENTFERNENS DER OPTIONEN, DIE AM WENIGSTEN NUTZEN BEWAHREN, DIE ZU EINEM NICHT ANFÄLLIGEN DATENSATZ FÜHREN WÜRDEN
415

FIG.4

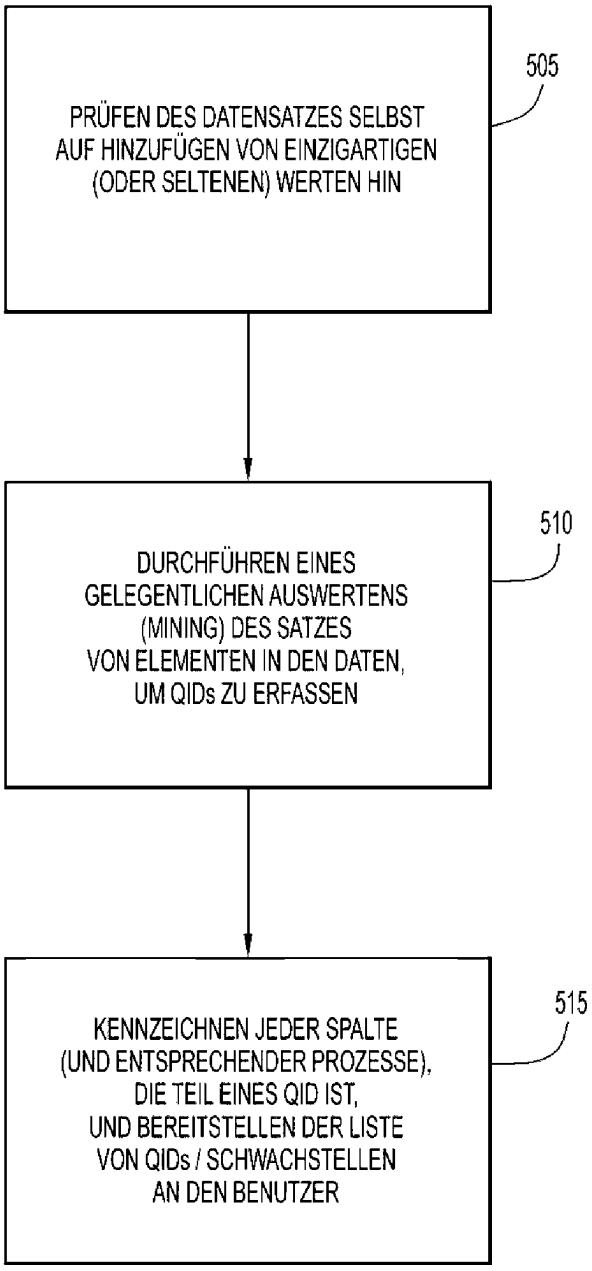
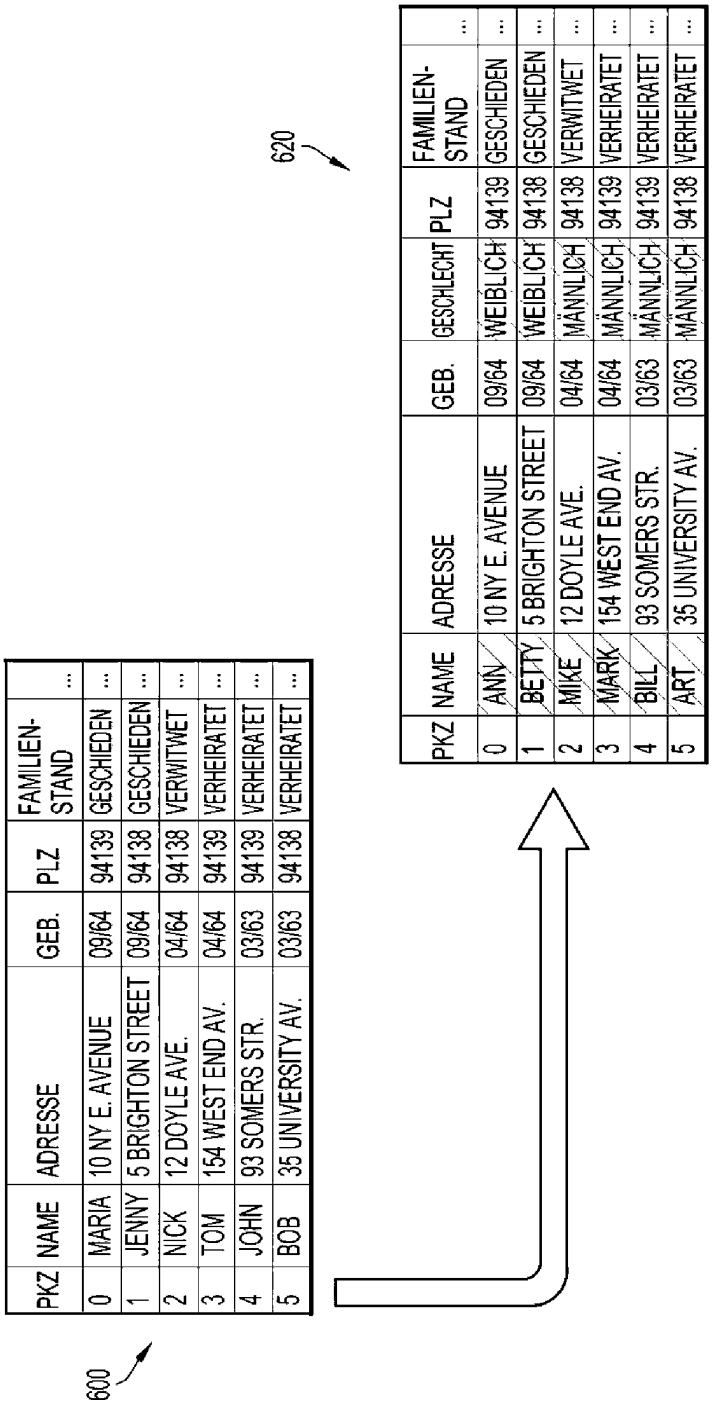

PRÜFEN DES DATENSATZES SELBST AUF HINZUFÜGEN VON EINZIGARTIGEN (ODER SELTENEN) WERTEN HIN
505
DURCHFÜHREN EINES GELEGENTLICHEN AUSWERTENS (MINING) DES SATZES VON ELEMENTEN IN DEN DATEN, UM QIDs ZU ERFASSEN
510
KENNZEICHNEN JEDER SPALTE (UND ENTSPRECHENDER PROZESSE), DIE TEIL EINES QID IST, UND BEREITSTELLEN DER LISTE VON QIDs / SCHWACHSTELLEN AN DEN BENUTZER
515

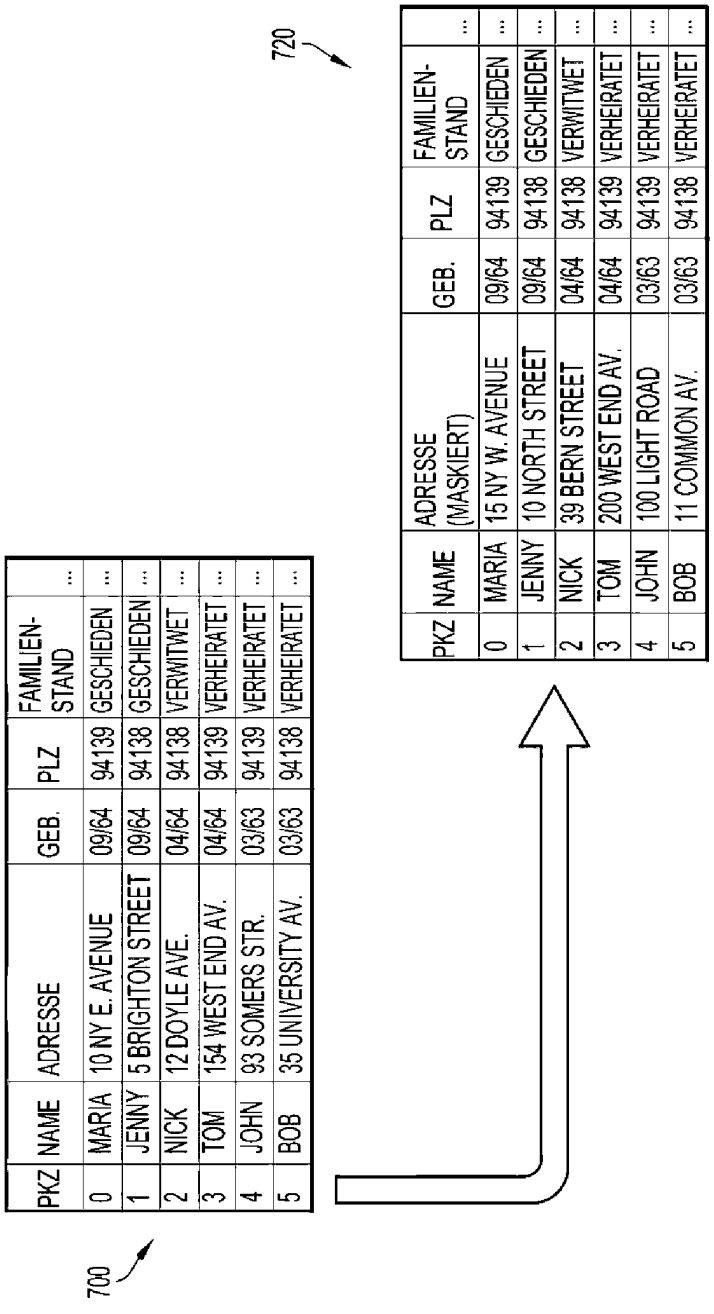
FIG.5



PKZ	NAME	ADRESSE	GEB.	PLZ	FAMILIEN-STAND	...
0	MARIA	10 NY E. AVENUE	09/64	94139	GESCHIEDEN	...
1	JENNY	5 BRIGHTON STREET	09/64	94138	GESCHIEDEN	...
2	NICK	12 DOYLE AVE.	04/64	94138	VERWITWET	...
3	TOM	154 WEST END AV.	04/64	94139	VERHEIRATET	...
4	JOHN	93 SOMERS STR.	03/63	94139	VERHEIRATET	...
5	BOB	35 UNIVERSITY AV.	03/63	94138	VERHEIRATET	...

PKZ	NAME	ADRESSE	GEB.	GESCHLECHT	PLZ	FAMILIEN-STAND	...
0	ANN	10 NY E. AVENUE	09/64	WEIBLICH	94139	GESCHIEDEN	...
1	BETTY	5 BRIGHTON STREET	09/64	WEIBLICH	94138	GESCHIEDEN	...
2	MIKE	12 DOYLE AVE.	04/64	MÄNNLICH	94138	VERWITWET	...
3	MARK	154 WEST END AV.	04/64	MÄNNLICH	94139	VERHEIRATET	...
4	BILL	93 SOMERS STR.	03/63	MÄNNLICH	94139	VERHEIRATET	...
5	ART	35 UNIVERSITY AV.	03/63	MÄNNLICH	94138	VERHEIRATET	...

FIG.6



PKZ	NAME	ADRESSE	GEB.	PLZ	FAMILIEN-STAND	...
0	MARIA	10 NY E. AVENUE	09/64	94139	GESCHIEDEN	...
1	JENNY	5 BRIGHTON STREET	09/64	94138	GESCHIEDEN	...
2	NICK	12 DOYLE AVE.	04/64	94138	VERWITWET	...
3	TOM	154 WEST END AV.	04/64	94139	VERHEIRATET	...
4	JOHN	93 SOMERS STR.	03/63	94139	VERHEIRATET	...
5	BOB	35 UNIVERSITY AV.	03/63	94138	VERHEIRATET	...

PKZ	NAME	ADRESSE (MASKIERT)	GEB.	PLZ	FAMILIEN-STAND	...
0	MARIA	15 NY W. AVENUE	09/64	94139	GESCHIEDEN	...
1	JENNY	10 NORTH STREET	09/64	94138	GESCHIEDEN	...
2	NICK	39 BERN STREET	04/64	94138	VERWITWET	...
3	TOM	200 WEST END AV.	04/64	94139	VERHEIRATET	...
4	JOHN	100 LIGHT ROAD	03/63	94139	VERHEIRATET	...
5	BOB	11 COMMON AV.	03/63	94138	VERHEIRATET	...

FIG.7

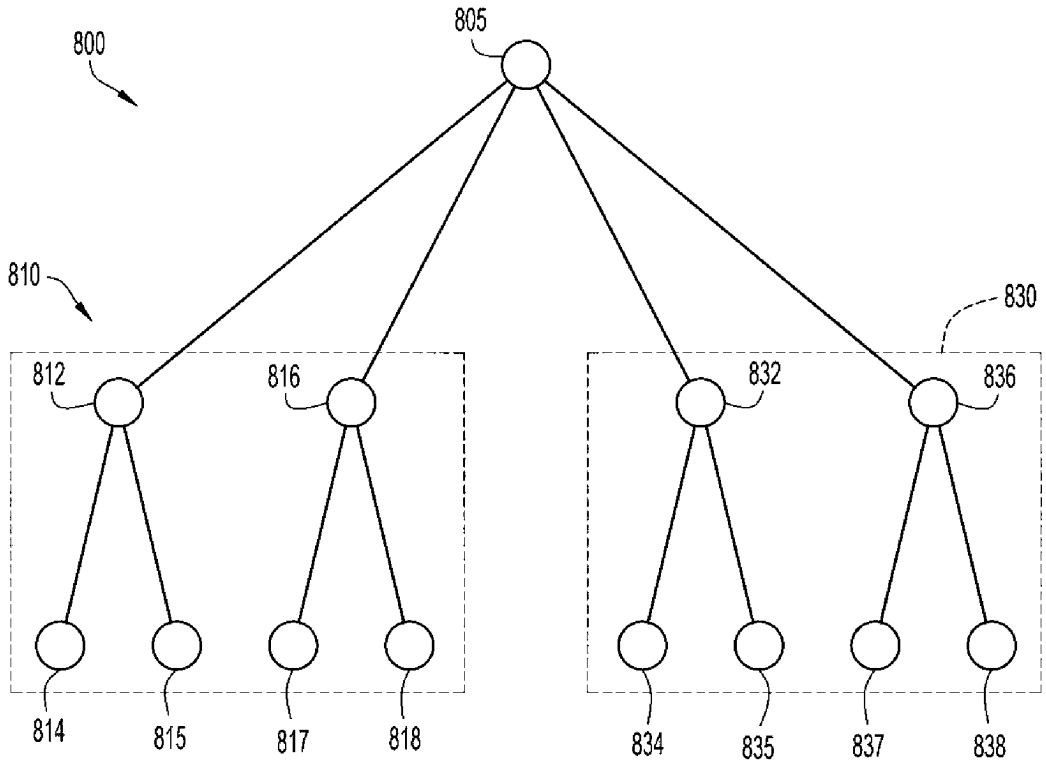


FIG.8

