(19) Deutsches Patent- und Markenamt

(10) **DE 10 2014 204 827 A1** 2014.09.18

(12) # Offenlegungsschrift

(21) Aktenzeichen: **10 2014 204 827.3**
(22) Anmeldetag: **14.03.2014**
(43) Offenlegungstag: **18.09.2014**

(51) Int Cl.: ***G06F 17/30*** (2006.01)
G06N 7/00 (2006.01)

(30) Unionspriorität:
13/827,491 **14.03.2013** **US**

(71) Anmelder:
Palantir Technologies, Inc., Palo Alto, Calif., US

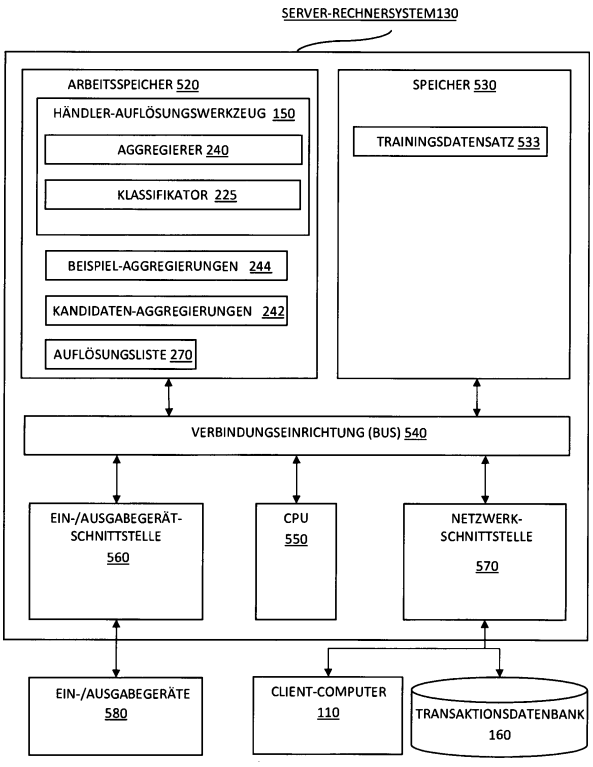
(74) Vertreter:
Dendorfer & Herrmann Patentanwälte Partnerschaft mbB, 80335 München, DE

(72) Erfinder:
Erenrich, Daniel, Palo Alto, Calif., US

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

(54) Bezeichnung: **Auflösen ähnlicher Entitäten aus einer Transaktionsdatenbank**

(57) Zusammenfassung: Ein Verfahren zum Identifizieren von in Beziehung stehenden Transaktionsdatensätzen aus einer Datenbank, die Transaktionsdatensätze für mehrere Entitäten speichert, beinhaltet: Gruppieren von Transaktionsdatensätzen mit einem gemeinsamen Attributwert zu Sätzen von Transaktionsdatensätzen, Entgegennehmen einer Auswahl eines Satzes von Beispiel-Datensätzen, und Bestimmen der Wahrscheinlichkeit, dass der Satz von Transaktionsdatensätzen Transaktionsdatensätze speichert, die mit einer ersten Entität assoziiert sind. Weitere Operationen beinhalten ein Auflösen, dass der Satz von Transaktionsdatensätzen Transaktionsdatensätze speichert, die mit der ersten Entität assoziiert sind. Dies verbessert den Prozess eines Identifizierens von in Beziehung stehenden Transaktionsdatensätzen, da in Beziehung stehende Transaktionsdatensätze, die bei Vergleichen von Zeichenketten von Transaktionsdatensatz-Attributen verpasst wurden, erfasst werden.



Beschreibung

[0001] Ausführungsformen der Erfindung betreffen allgemein eine Datenanalyse, und insbesondere ein Auflösen ähnlicher Entitäten von einer Transaktionsdatenbank.

[0002] Eine Beschaffung relevanter Information aus großen Datenbanken kann in einigen Situationen relativ unkompliziert sein. Speziell wenn die Datensätze in einer Datenbank gut strukturiert sind und angestrebt wird, Information in Datensätzen zu beschaffen, die einen speziellen Wert oder eine spezielle Zeichenkette in einem speziellen Feld aufweisen, können diese Datensätze unter Verwendung von Filterfunktionen von Datenbank-Schnittstellensoftware isoliert werden. Unter Verwendung von Kombinationen von Filterfunktionen kann die Art und Weise, in der Datensätze zur Isolierung identifiziert werden, technisch verfeinert werden. Die isolierten Datensätze können dann aggregiert werden, um einen Bericht zu liefern, der alle Datensätze enthält, die gemeinsam die gewünschte Information bilden.

[0003] Jedoch stützen sich derartige Filterfunktionen, um Gemeinsamkeiten aufweisende Datenbankdatensätze zu bezeichnen, auf Attribute, die über diese Datenbankdatensätze hinweg identisch sind. In der realen Welt haben Datenbankdatensätze möglicherweise keine Attribute, die über diese Datensätze hinweg identisch sind, obschon diese Datensätze in Beziehung zueinander stehen, oder sie haben möglicherweise identische Attribute in einer relativ kleinen Anzahl von Feldern (oder Teile von Feldern), derart, dass Filterfunktionen nicht in der Lage sind, eine Isolierung der gewünschten Datenbankdatensätze von anderen Datenbankdatensätzen zu liefern. Beispielsweise können solche Probleme auftreten, wenn eine Datenbank Datenbankdatensätze aufweist, die aus mehreren unterschiedlichen Quellen stammen. Das Isolieren von in Beziehung stehenden Datenbankdatensätzen von anderen Datenbankdatensätzen ist ein technisches Problem, das mit zunehmender Größe der Datenbank, hinsichtlich der Anzahl von vorhandenen Datensätzen (z. B. eine Datenbank, die Milliarden von Datenbankdatensätzen aufweist) schwerwiegender wird. Da die Größen von Datenbanken in der realen Welt im Verlauf der Zeit zunimmt, wird erwartet, dass dieses Problem mit der Zeit schwerwiegender wird.

[0004] Die hier beschriebenen Methoden können beispielsweise auf dem Gebiet von Finanzinstitutionen Anwendung finden, die Transaktionsdaten zur Analyse speichern. Eine Finanzinstitution erzeugt Transaktionsdaten aus mit Kredit- und Debitkarten durchgeführten Kaufvorgängen bei Firmen, die ein Händlerkonto bei der Finanzinstitution haben. Das Händlerkonto kann verwendet werden, um einzelne mit Kredit- oder Debitkarte durchgeführte Kaufvorgänge zu verarbeiten. Wiederum wird jeder derartige Kaufvorgang als Transaktionsdatensatz in einer Transaktionsdatenbank gespeichert. Ein Transaktionsdatensatz, der zu einem speziellen Händlerkonto gehört, beinhaltet häufig ein Händler-ID-Attribut, das den Transaktionsdatensatz mit dem Händlerkonto verknüpft. Eine Händler-ID kann ein beliebiger Datentyp sein, einschließlich einer Zahl, einer Zeichenkette, oder irgendeiner Kombination aus diesen. Die Finanzinstitution kann dann die Transaktionsdatensätze von einem oder mehreren Händlerkonten analysieren. Beispielsweise kann eine Analyse ein Aggregieren der Transaktionsdatensätze eines Händlerkontos oder spezieller Händlerkonten beinhalten. Die Analyse kann dann die Performanz des Händlerkontos mit derjenigen von Konten konkurrierender Händler in demselben geographischen Gebiet vergleichen.

[0005] Obschon die Finanzinstitution die Transaktionsdatensätze in einer Transaktionsdatenbank speichert, kann eine gewisse Analyse erfordern, dass die Daten auf Arten und Weisen organisiert sind, die nicht Teil der Transaktionsdatensätze in der Datenbank sind. Diese Datenbanken enthalten Transaktionsdatensatz-Sätze, die eine Analyse gemeinsam gruppieren sollte, obschon es keinen einzigen Attributwert gibt, der die Transaktionsdatensätze in Beziehung setzt. Falls beispielsweise eine Finanzinstitution eine Transaktionsdatenbank mit einem Händler-ID-Attribut konfiguriert, das jeden Transaktionsdatensatz mit einem Händlerkonto verknüpft, dann könnte eine Analyse ohne Weiteres Transaktionsdatensätze mit der gleichen Händler-ID aggregieren. Jedoch kann eine einzelne Firma mehrere Händlerkonten bei einer Finanzinstitution haben. Falls die Finanzinstitution unterschiedliche Händler-IDs für jedes einzelne Händlerkonto bietet, sogar wenn mehrere Händlerkonten zu einer einzigen Firma gehören, dann ist es schwierig, Transaktionsdatensätze von den mehreren Händlerkonten dieser Firma zu aggregieren. Beispielsweise kann eine Franchise-Firma unterschiedliche Händlerkonten mit unterschiedlichen Händler-IDs für jeden Franchisenehmer-Standort haben. In einem solchen Fall wäre eine Analyse nicht in der Lage, die Transaktionsdatensätze der Franchise-Firma allein auf Basis identischer Händler-IDs zu aggregieren. Stattdessen kann eine Analyse Ähnlichkeiten zwischen den Händler-ID-Attributwerten nutzen, um die Transaktionsdatensätze der Franchise-Firma zu aggregieren.

[0006] Bestehende Verfahren stützen sich auf einfache Überprüfungen, beispielsweise Vergleiche von Zeichenketten zwischen einem Attribut in einer Datenbank von Transaktionsdatensätzen, um Ähnlichkeiten zwischen Gruppen von Transaktionsdatensätzen zu erfassen. Transaktionsdatensätze, die Attribut-Zeichenketten beinhalten, welche einem Maß einer Ähnlichkeit genügen, werden dann zur Analy-

se aggregiert. Diese Verfahren funktionieren möglicherweise, sofern das Attribut identische oder ähnliche Zeichenketten enthält, für Gruppen von Transaktionsdatensätzen, die aggregiert werden sollten, und sofern es unterschiedliche Zeichenketten enthält, für Gruppen von Transaktionsdatensätzen, die nicht aggregiert werden sollten.

[0007] Jedoch sind derartige Identifikatoren nicht immer (oder sogar nicht einmal für gewöhnlich) verfügbar. Beispielsweise können unterschiedliche Händler-IDs für die Händlerkonten einer einzelnen Firma verhindern, dass ein Analysesystem die Transaktionsdatensätze der Firma aggregieren kann. Außerdem können Transaktionsdatensätze ähnliche Identifikatoren enthalten, auf deren Basis ein Analysesystem ein Aggregieren durchführen kann, sogar wenn die Transaktionsdatensätze nicht aggregiert werden sollten. Beispielsweise können zwei unterschiedliche Firmen Händlerkonten mit ähnlichen Händler-IDs haben, die ein Analysesystem fälschlicherweise einer einzigen Firma zuschreiben könnte. Das Analysesystem kann dann fälschlicherweise die Transaktionsdatensätze der zwei Firmen aggregieren.

[0008] Wie das zuvor Beschriebene illustriert, besteht weiter ein Bedarf nach effektiveren Methoden zum Evaluieren von Finanztransaktionsdatensätzen.

[0009] Ausführungsformen der Erfindung befassen sich mit dem Problem, in Beziehung stehende Datenbanksätze zu identifizieren, die möglicherweise keine nutzbaren identischen Attribute haben, und dabei nicht in Beziehung stehende Datenbankdatensätze auszuschließen, und lösen insbesondere das Problem, Datenbankdatensätze zu identifizieren, die in Beziehung zu einer gemeinsamen Entität stehen, jedoch möglicherweise keine identischen Attribute aufweisen.

[0010] Die vorliegende Erfindung ist in den unabhängigen Ansprüchen dargelegt. Die abhängigen Ansprüche betreffen optionale Merkmale einiger Ausführungsformen der Erfindung.

[0011] Eine Ausführungsform der Erfindung legt ein Verfahren zum Identifizieren von in Beziehung stehenden Transaktionsdatensätzen aus einer Datenbank dar, die Transaktionsdatensätze für mehrere Entitäten speichert, welches beinhaltet: Gruppieren von Transaktionsdatensätze mit gemeinsamem Attributwert zu Transaktionsdatensatz-Sätzen, Entgegennehmen einer Auswahl eines Beispiel-Datensatz-Satzes, und Bestimmen der Wahrscheinlichkeit, dass der Transaktionsdatensatz-Satz Transaktionsdatensätze speichert, die mit einer ersten Entität assoziiert sind. Das Verfahren beinhaltet auch ein Auflösen, dass der Transaktionsdatensatz-Satz Transaktionsdatensätze speichert, die mit der ersten Entität assoziiert sind.

[0012] Weitere Ausführungsformen der Erfindung beinhalten, ohne Einschränkung, ein computerlesbares Speichermedium, das Anweisungen beinhaltet, die, wenn sie durch eine Verarbeitungseinheit ausgeführt werden, die Verarbeitungseinheit veranlassen, Aspekte des hier beschriebenen Lösungsansatzes zu implementieren, sowie ein System, das unterschiedliche Elemente beinhaltet, die konfiguriert sind, Aspekte des hier beschriebenen Lösungsansatzes zu implementieren.

[0013] Ein Vorteil des offenbarten Verfahrens besteht darin, dass zwei Datensatz-Sätze in einer Datenbank von Transaktionsdatensätzen, die keine identischen Attribute haben, jedoch zu derselben gemeinsamen Entität gehören, mit der gemeinsamen Entität verknüpft werden können. Daher werden Auflösungen, die allein mit Vergleichen von Zeichenketten verpasst würden, bewerkstelligt, und inkorrekte Auflösungen, die lediglich auf ähnlichen Zeichenketten basieren, werden vermieden, was die Präzision einer Auflösung verbessert. Ein weiterer Vorteil des offenbarten Verfahrens besteht darin, dass es die Anzahl von fälschlichen Aggregierungen verringert, die von Transaktionsdatensätzen herrührt, welche ähnliche Identifikatoren aufweisen, obwohl sie unterschiedlichen Entitäten zugehörig sind. Durch Verringern der Anzahl von fälschlichen Aggregierungen wird von einer auf diese Weise gelieferten aggregierten Meldung von Datensätzen weniger Speicher belegt als von einer entsprechenden aggregierten Meldung, die mittels Filterfunktionen erzeugt wird.

[0014] Damit die zuvor beschriebenen Merkmale der Erfindung im Detail verstanden werden können, wird eine genauere Beschreibung der zuvor kurz zusammengefassten Erfindung mit Bezug auf die Ausführungsformen gegeben, von denen einige in den anliegenden Zeichnungen dargestellt sind. Es versteht sich jedoch, dass die anliegenden Zeichnungen lediglich typische Ausführungsformen der Erfindung darstellen, und daher nicht als den Schutzumfang einschränkend zu verstehen sind, da die Erfindung weitere gleichermaßen effektive Ausführungsformen zulässt.

[0015] **Fig.** 1 ist ein Blockdiagramm, das ein Computersystem darstellt, das konfiguriert ist, einen oder mehrere Aspekte der Erfindung zu implementieren.

[0016] **Fig.** 2 ist ein Blockdiagramm des Datenflusses im Anwendungsserver.

[0017] **Fig.** 3 zeigt ein Verfahren zum Trainieren des Klassifikators, gemäß einer Ausführungsform.

[0018] **Fig.** 4 zeigt ein Verfahren zum Auflösen einer Händler-ID auf einen Händler, gemäß einer Ausführungsform.

[0019] **Fig.** 5 zeigt ein Beispiel einer Rechnerumgebung, gemäß einer Ausführungsform.

[0020] Ausführungsformen der Erfindung können verwendet werden, um gewisse Finanztransaktionsdatensätze zu aggregieren, die zu einer gemeinsamen Entität aufgelöst werden, jedoch möglicherweise sonst nicht miteinander gruppiert werden könnten. Nimmt man an, dass eine Transaktionsdatenbank einer Finanzinstitution Transaktionsdatensätze von jedem unterschiedlichen Händlerkonto einer Firma anhand eines unterschiedlichen Händler-ID-Attributs identifiziert, dann ist es möglich, dass die unterschiedlichen Händler-ID-Attribute nicht in korrekter Weise zusammenpassen, um die Transaktionsdatensätze der Konten mit der Firma zu verknüpfen. Als weiteres Beispiel haben unterschiedliche Franchisenehmer eines gemeinsamen Franchisegebers eine Flut von Händlerkonten, was es schwierig macht, die Transaktionsdatensätze, die allen Franchisenehmern des Franchisegebers zugehörig sind, allein aus den Transaktionsdatensätzen zu aggregieren. Bei einer Ausführungsform kombiniert ein Finanzanalysesystem Transaktionsdatensätze zu Händler-ID-Sätzen basierend auf identischen Händler-IDs, so dass jeder Händler-ID-Satz alle Transaktionsdatensätze mit einer speziellen Händler-ID enthält. Wie dieses Beispiel darstellt, kann eine einzelne Firma durch mehrere Händler-IDs repräsentiert sein. Um den vollständigen Satz von Transaktionsdatensätzen für eine einzelne Entität (Firma) zu evaluieren, muss jede Sammlung von Finanztransaktionsdatensätzen (die Händler-ID-Sätze), die der einzelnen Entität zugehörig sind, zusammengeführt werden.

[0021] Bei einer Ausführungsform aggregiert das Analysesystem Transaktionsdatensätze aus einer großen Sammlung von Händler-ID-Sätzen. Dieses Aggregieren kann beinhalten, dass die durchschnittliche Transaktionsgröße, die Standardabweichung der Transaktionsgröße oder der durchschnittliche Betrag berechnet wird, den eine Einzelperson ausgegeben hat. Das Analysesystem verwendet die Aggregierungen, um einen Klassifikator zu trainieren. Sobald das Analysesystem trainiert wurde, erzeugt es einen Konfidenzwert, ob zwei Händler-ID-Sätze zu einer Firma gehören, basierend auf den Aggregierungen aus dem Paar von Händler-ID-Sätzen. Um die Händler-ID-Sätze mit der Firma zu assoziieren, empfängt das Analysesystem eine Auswahl eines Beispiel-Händler-ID-Satzes, der mit der Firma assoziiert werden sollte und die Kennzeichen der Firma bestmöglich repräsentiert. Das Analysesystem vergleicht den Beispiel-Händler-ID-Satz mit anderen Händler-ID-Sätzen, um einen Konfidenzwert zu bestimmen. Der Konfidenzwert stellt die Wahrscheinlichkeit dar, dass der Beispiel-Händler-ID-Satz und der andere Händler-ID-Satz mit der Firma assoziiert sind. Das Analysesystem assoziiert jeden einzelnen Händler-ID-Satz, der, bei Vergleich mit dem Beispiel, Konfidenzwerte oberhalb eines Schwellenwertes hat, mit der Firma. Dieses Vorgehen resultiert in einer Sammlung von Finanztransaktionsdatensätzen, von denen anzunehmen ist, dass sie alle zu einer einzigen Firma gehören, trotz der Tatsache, dass viele derartige Datensätze unterschiedliche Händler-IDs enthalten können.

[0022] In der folgenden Beschreibung sind zahlreiche spezifische Details dargelegt, um für ein grundlegenderes Verständnis der Erfindung zu sorgen. Jedoch versteht es sich für Fachleute, dass die Erfindung ohne eines oder mehrere dieser spezifischen Details ausgeführt werden kann.

[0023] **Fig.** 1 ist ein Blockdiagramm, das ein beispielhaftes Datenanalysesystem **100** darstellt, gemäß einer Ausführungsform der Erfindung. Wie dargestellt, beinhaltet das Datenanalysesystem **100** einen Anwendungsserver **140**, der auf einem Server-Rechnersystem **130** läuft, einen Client **120**, der auf einem Client-Computersystem **110** läuft, und mindestens eine Transaktionsdatenbank **160**. Weiter können der Client **120**, der Anwendungsserver **140** und die Transaktionsdatenbank **160** über ein Netzwerk **180** kommunizieren.

[0024] Der Client **120** repräsentiert eine oder mehrere Softwareanwendungen, die konfiguriert sind, um Daten zu präsentieren und Benutzereingaben übersetzt in Anfragen nach Datenanalysen durch den Anwendungsserver **140**. Bei dieser Ausführungsform hat der Client **120** eine Verbindung zum Anwendungsserver **140**. Jedoch ist es auch möglich, dass mehrere Clients **120** auf dem Client-Computer **110** ausgeführt werden, oder mehrere Clients **120** auf mehreren Client-Computern **110** können mit dem Anwendungsserver **140** interagieren. Bei einer Ausführungsform kann der Client **120** ein Browser sein, der auf einen Web-Dienst zugreift.

[0025] Alternativ kann der Client **120** auf demselben Server-Rechnersystem **130** wie der Anwendungsserver **140** laufen. In jedem Fall würde ein Benutzer über den Client **120** mit dem Datenanalysesystem **100** interagieren.

[0026] Der Anwendungsserver **140** ist so konfiguriert, dass er ein Händler-Auflösungswerkzeug **150** und eine Analyse-Engine **155** beinhaltet. Das Händler-Auflösungswerkzeug **150** verknüpft zusammenpassende Händler-IDs mit einer Firma. Das Händler-Auflösungswerkzeug **150** liest Daten aus der Transaktionsdatenbank **160**. Das Händler-Auflösungswerkzeug **150** kann Auflösungsdaten auf dem Server-Computer **130** oder in der Transaktionsdatenbank **160** speichern.

[0027] Die Analyse-Engine **155** verwendet die Auflösungsdaten vom Händler-Auflösungswerkzeug **150**,

um aus der Transaktionsdatenbank **160** abgerufene Daten zu analysieren. Die Analyse-Engine **155** führt ein Aggregieren und Vergleichen der Transaktionsdatensätze aus der Transaktionsdatenbank **160** durch, um Erkenntnisse betreffend eine spezielle Firma zu liefern. Beispielsweise kann eine Finanzinstitution eine Datenanalyse entwerfen, um die saisonalen Ausgabentrends für eine Franchise-Firma zu evaluieren. Jedoch kann jeder Franchisenehmer der Franchise-Firma ein unterschiedliches Händlerkonto bei der Finanzinstitution haben. Die Finanzinstitution speichert die Transaktionsdatensätze von den Händlerkonten mit unterschiedlichen Händler-IDs, die einen Transaktionsdatensatz mit einem Händlerkonto assoziieren. Um den vollständigen Satz von Transaktionsdatensätzen für die Franchise-Firma zu evaluieren, ist es erforderlich, dass die Analyse-Engine **155** jede Sammlung von Finanztransaktionsdatensätzen von jedem Franchisenehmer zusammenführt. Dazu nutzt die Analyse-Engine **155** die Auflösungsdaten von dem Händler-Auflösungswerkzeug **150**, um die Finanztransaktionsdatensätze von jedem Franchisenehmer zu einem vollständigen Satz von Transaktionsdatensätzen für die Franchise-Firma zusammenzuführen, um die saisonalen Ausgabentrends für die Franchise-Firma zu evaluieren.

[0028] Bei dieser Ausführungsform speichert die Transaktionsdatenbank **160** Datensätze von Finanztransaktionen, die einer Finanzinstitution zugehörig sind. Beispielsweise kann die Transaktionsdatenbank Datensätze für eine große Anzahl von Händlerkonten beinhalten, die Kredit- und Debitkarten-Transaktionen verarbeiten. In einem derartigen Fall enthielte jeder Datensatz Datenattribute für den ausgegebenen Betrag, das Datum und die Zeit der Transaktion, die Adresse des Händlers, und eine Händler-ID, um den Datensatz mit einem speziellen Händlerkonto zu assoziieren.

[0029] Die Transaktionsdatenbank **160** kann ein relationales Datenbankmanagementsystem (RDBMS) sein, das die Transaktionsdaten als Zeilen in relationalen Tabellen speichert. Alternativ kann die Transaktionsdatenbank **160** auf demselben Server-Rechnersystem **130** wie der Anwendungsserver **140** gespeichert sein. Die Datensätze einer Finanzinstitution [hier fehlt Text im englischen Original]

[0030] **Fig.** 2 zeigt einen Datenfluss von der Transaktionsdatenbank **160** durch das Händler-Auflösungswerkzeug **150**, gemäß einer Ausführungsform der Erfindung. Wie dargestellt, beinhaltet die Transaktionsdatenbank **160** Händler-ID-Sätze **210**. Jeder Händler-ID-Satz **210** beinhaltet Transaktionsdatensätze **215** mit der gleichen Händler-ID, beispielsweise Kredit- und Debitkartentransaktionen, die für ein einzelnes Händlerkonto bei einer Finanzinstitution verarbeitet werden. Das Händler-Auflösungswerkzeug **150** beinhaltet einen Aggregierer **240**, Kandidaten-Aggregierungen **242**, Beispiel-Aggregierungen **244**, einen Trainingsdatensatz **260** und einen Identitätsauflöser **250**. Der Identitätsauflöser **250** selbst beinhaltet einen Klassifikator **255** und eine Auflösungsliste **270**.

[0031] Bei einer Ausführungsform ist der Klassifikator **255** ein Random-Forest-Klassifikator. Ein Random-Forest-Klassifikator ist ein Maschinenlernalgorithmus, von dem allgemein bekannt ist, dass er äußerst genau bei großen Datenbanken arbeitet, die diskrete, zusammenhängende und fehlende Daten beinhalten, wie dies für Finanztransaktionsdatensätze **215** in der Transaktionsdatenbank **160** der Fall sein kann. Random-Forest-Klassifikatoren beinhalten mehrere Entscheidungsbäume. Die Entscheidungsbäume evaluieren Merkmale von Eingangsdaten. Beim vorliegenden Kontext von Finanztransaktionsdatensätzen, die mit Händlerkonten mittels einer Händler-ID assoziiert sind, können die evaluierten Merkmale beinhalten:

– Wortüberschneidungszählwert und Häufigkeit von Händler-ID-Attributen
– Wort-basierte Kosinus-Ähnlichkeit, gewichtet anhand von Werten der auf Begriffshäufigkeit bezogenen Inversen Dokumenthäufigkeit von Händler-ID-Attributen
– Zeichenbasierte Kosinus-Ähnlichkeit von Händler-ID-Attributen
– Platzierung einer Wortüberschneidung von Händler-ID-Attributen
– Identifizieren der Zeichenkette „.com”
– Ob die Händler-ID-Attribute einen Ladencode beinhalten
– Überschneidung von Präfix- oder Suffix-Ziffern in den Händler-ID-Attributen
– Ob der gelieferte Stadt-Code numerisch ist
– Zusammenpassende eindeutige Handelsklassencode
– Fraktionale Differenz der durchschnittlichen Transaktionsgröße
– Standardabweichungen von der durchschnittlichen Transaktionsgröße
– Fraktionale Differenz der Größe der Transaktionsgrößenvarianzen

[0032] Man beachte, dass der Klassifikator **255** eine Vielfalt weiterer Merkmale evaluieren kann, abhängig von den Erfordernissen eines speziellen Falls sowie von Daten, die aus den zugrundeliegenden Transaktionsdatensätzen verfügbar sind. Weiter erkennt ein Fachmann, dass ein Random-Forest-Klassifikator als Referenzbeispiel eines Klassifikators verwendet wird, und dass eine Vielfalt weiterer Maschinenlern-Klassifikatoren verwendet werden könnte.

[0033] Um die Vielfalt von Merkmalen zu evaluieren, wird im Klassifikator **255** ein Wachsen von Entscheidungsbäumen durchgeführt, basierend auf der Wahrscheinlichkeit, dass ein ausgewähltes Merkmal zu ei-

ner gewissen Klassifikation führen sollte. Im vorliegenden Kontext wird im Klassifikator **255** ein Wachsen von mehreren Entscheidungsbäumen basierend auf unterschiedlichen Kombinationen der Merkmale durchgeführt, so dass jeder Entscheidungsbaum klassifiziert, dass ein Paar von Händler-ID-Sätzen **210** zu derselben Firma passt, oder nicht. Der Ausgabewert des Klassifikators **255** ist der Prozentsatz von Entscheidungsbäumen, die klassifizieren, dass ein Paar von Händler-ID-Sätzen **210** zu derselben Firma passt.

[0034] Um vorzubereiten, Händler-IDs mit einer Firma zu verknüpfen, wird im Klassifikator **255** ein Wachsen der Entscheidungsbäume durchgeführt, mittels Trainieren am Trainingsdatendatz **260**. Der Trainingsdatensatz **260** beinhaltet Paare von Händler-ID-Sätzen **210**, die zu derselben Firma passen, sowie Paare von Händler-ID-Sätzen **210**, die nicht zu derselben Firma passen. Die Paare von Händler-ID-Sätzen **210**, die zu derselben Firma passen, werden im Trainingsdatensatz **260** als positive Beispiele klassifiziert. Die Paare von Händler-ID-Sätzen **210**, die nicht zu derselben Firma passen, werden im Trainingsdatensatz **260** als negative Beispiele klassifiziert.

[0035] Während der Klassifikator **255** die Merkmale eines jeden Paares von Händler-ID-Sätzen **210** als positives oder negatives Beispiel verarbeitet, wird die Genauigkeit des Klassifikators **255** vergrößert, dadurch dass die in den Entscheidungsbäumen verwendeten Wahrscheinlichkeiten verfeinert werden.

[0036] Der Trainingsdatensatz **260** kann auch schwierige Grenzfälle beinhalten, beispielsweise Paare von Händler-ID-Sätzen **210**, die nicht zusammenpassen, jedoch ähnliche Händler-ID-Zeichenketten aufweisen. Ein Paar von Händler-ID-Sätzen **210** mit ähnlichen Händler-ID-Zeichenketten, die nicht mit derselben Firma verknüpft werden sollten, ist ein Grenzfall, da häufig ähnliche Händler-ID-Zeichenketten von Händler-ID-Sätzen **210** kommen, die mit derselben Firma verknüpft werden sollten. Ein Hinzufügen derartiger Grenzfälle zum Trainingsdatensatz **260** bewirkt, dass der Klassifikator **255** die Wahrscheinlichkeiten in den Entscheidungsbäumen des Klassifikators **255** anpasst, um ein besseres Klassifizieren von Paaren von Händler-ID-Sätzen **210** mit ähnlichen Händler-ID-Attributen durchzuführen.

[0037] Um einen großen Trainingsdatensatz **260** zu erzeugen, kann das Händler-Auflösungswerkzeug **150** Paare von zufällig gewählten Händler-ID-Sätzen **210** erzeugen, die typischerweise negative Trainingsbeispiele liefern.

[0038] Der Trainingsdatensatz **260** kann Transaktionsdatensätze **215** beinhalten, die aus der Transaktionsdatenbank **160** ausgelesen wurden, kann künstliche Transaktionsdatensätze **215** beinhalten oder kann irgendeine Kombination von diesen beinhalten. Zwar hat sich ein Trainingsdatensatz **260** von 4000 Paaren von Händler-ID-Sätzen **210** als effektiv erwiesen, jedoch kann die tatsächliche Größe des Trainingsdatensatzes **260** je nach Präferenz festgelegt werden.

[0039] Sobald der Klassifikator **255** trainiert wurde, kann das Händler-Auflösungswerkzeug **150** verwendet werden, um Händler-IDs von einem unterschiedlichen Händlerkonto mit einer Firma zu assoziieren, so dass die Analyse-Engine **155** Datenanalysen mit vollständigen Sätzen von Transaktionsdatensätzen **215** von allen Händlerkonten der Firma durchführen kann.

[0040] Die Transaktionsdatenbank **160** ist so konfiguriert, dass sie einen Mechanismus beinhaltet, um Transaktionsdatensätze **215** mit einem gemeinsamen Händler-ID-Attribut als Händler-ID-Sätze **210** liefert. Beispielsweise kann die Transaktionsdatenbank **160** Transaktionsdatensätze **215** mit gleichen Händler-ID-Attributen gemeinsam in Händler-ID-Sätzen **210** speichern, oder die Transaktionsdatenbank **160** kann Transaktionsdatensätze **215** sequentiell anhand des Wertes eines Transaktionsdatenattributs speichern. Ungeachtet der Anordnung der Transaktionsdatensätze **215** kann das Händler-Auflösungswerkzeug **150** Händler-ID-Sätze **210** aus der Transaktionsdatenbank **160** abrufen.

[0041] Nachdem ein Benutzer einen Händler-ID-Satz **210** als Beispiel-Händler-ID-Satz **210(0)** ausgewählt hat, können weitere Händler-ID-Sätze **210** als Kandidaten-Händler-ID-Sätze **210(1)** bis **210(M – 1)** in Betracht gezogen werden. Der Benutzer wählt den Beispiel-Händler-ID-Satz **210(0)** als repräsentativ für die Kennzeichen der Firma aus, die aufgelöst werden soll. Der Beispiel-Händler-ID-Satz kann eine große Anzahl von Transaktionsdatensätzen **215** beinhalten. Eine große Anzahl von Transaktionsdatensätzen **215** kann Aggregierungen liefern, beispielsweise die durchschnittliche Transaktionsgröße, die genauer als Händler-ID-Sätze **210** mit einer geringeren Anzahl von Transaktionsdatensätzen **215** sind. Weitere Faktoren, beispielsweise geographische Standorte, die Händler-ID-Zeichenkette, oder weitere Geschäfts-Heuristik können ebenfalls die Auswahl des Beispiel-Händler-ID-Satzes **210(0)** aus den verfügbaren Händler-ID-Sätzen **210** leiten.

[0042] Wenn Händler-IDs mit einer Firma verknüpft werden, ruft das Händler-Auflösungswerkzeug **150** die Transaktionsdatensätze **215** des Beispiel-Händler-ID-Satzes **210(0)** und die Transaktionsdatensätze **215** eines Kandidaten-Händler-ID-Satzes **210(1)** ab. Der Aggregierer **240** aggregiert die Attribute der Transaktionsdatensätze **215** des Beispiel-Händler-ID-Satzes **210(0)**, um Beispiel-Aggregierungen **244** zu erzeugen. Beispielsweise berechnet der Ag-

gregierer **240** die durchschnittliche Transaktionsgröße, die Standardabweichung der Transaktionsgröße oder den durchschnittlichen Betrag, den eine Einzelperson ausgegeben hat. Das Händler-ID-Attribut des Beispiel-Händler-ID-Satzes **210(0)** ist ebenfalls in den Beispiel-Aggregierungen **244** enthalten. Der Aggregierer **240** berechnet auch die Kandidaten-Aggregierungen **242** aus dem Kandidaten-Händler-ID-Satz **210** und schließt auch das Händler-ID-Attribut des Kandidaten-Händler-ID-Satzes **210(1)** mit den Kandidaten-Aggregierungen **242** ein. Es sei angemerkt, dass der Aggregierer **240** zusätzliche Aggregierungswerte berechnen kann, gemäß zahlreichen unterschiedlichen Gestaltungen, die der Werkzeugentwickler wählen kann.

[0043] Nachdem der Aggregierer **240** die Aggregierungswerte bestimmt hat, leitet das Händler-Auflösungswerkzeug **150** die Beispiel-Aggregierung **244** und die Kandidaten-Aggregierung **242** an einen Identitätsauflöser **250** weiter. Der Klassifikator **255** bestimmt die Werte, die als Merkmale in den Entscheidungsbäumen verwendet werden, aus den Daten, die in den Beispiel-Aggregierungen **244** und den Kandidaten-Aggregierungen **242** enthalten sind. Der Klassifikator **255** verarbeitet die Beispiel-Aggregierung **244** und die Kandidaten-Aggregierung **242**, um einen Konfidenzwert zwischen Null und Eins zu erzeugen, der dem entspricht, wie wahrscheinlich der Beispiel-Händler-ID-Satz **210(0)** zum Kandidaten-Händler-ID-Satz **201(1)** passt, und daher mit derselben Firma verknüpft werden sollte. Falls der Beispiel-Händler-ID-Satz **210(0)** und der Kandidaten-Händler-ID-Satz **210(1)** einen Wert oberhalb eines gewissen Schwellenwertes erhält, beispielsweise 0,70, dann speichert der Identitätsauflöser **250** die Händler-ID des Kandidaten-Händler-ID-Satzes **201(1)** in einer Auflösungsliste **270**.

[0044] Das Händler-Auflösungswerkzeug **150** vergleicht Kandidaten-Händler-ID-Sätze **210(2)** bis **210(M – 1)** mit einem Beispiel-Händler-ID-Satz **210(0)**. Der Identitätsauflöser **250** fügt die Händler-ID eines jeden Kandidaten-Händler-ID-Satzes **210(1)** bis **210(M – 1)**, die einen hohen Konfidenzwert erzeugt, zur Auflösungsliste **270** hinzu.

[0045] Somit repräsentieren die Händler-IDs auf der Auflösungsliste **270** die Händler-ID-Sätze **210**, die zu derselben Firma wie der Beispiel-Händler-ID-Satz **210(0)** gehören.

[0046] Das Händler-Auflösungswerkzeug **150** speichert die Auflösungsliste **270** zur Verwendung durch die Analyse-Engine **155**. Wiederum kann die Analyse-Engine **155** die vollständige Sammlung von Transaktionsdatensätzen **215** der Firma unabhängig von den verschiedenen Händler-IDs analysieren, die in den Transaktionsdatensätzen **215** der Firma enthalten sind. Beispielsweise sollte, falls die verschiedenen Händler-IDs in einer Auflösungsliste **270** Transaktionsdatensätze **215** mit mehreren Händlerkonten von mehreren Franchisenehmern einer Franchise-Firma assoziieren, dann die Analyse-Engine **155** die Transaktionsdatensätze **215** mit den Händler-IDs in der Auflösungsliste **270** zusammenführen, um die vollständige Sammlung von Transaktionsdatensätzen **215** der Franchise-Firma zu analysieren.

[0047] **Fig.** 3 ist ein Ablaufdiagramm von Verfahrensschritten zum Trainieren des Klassifikators **255**, gemäß einer Ausführungsform der Erfindung. Zwar sind die Verfahrensschritte in Verbindung mit den Systemen der **Fig.** 1–**Fig.** 2 und **Fig.** 5 beschrieben, jedoch ist es für Fachleute klar, dass eine beliebige Systemkonfiguration, welche die Verfahrensschritte in beliebiger Reihenfolge durchführt, innerhalb des Schutzumfangs der Erfindung liegt.

[0048] Wie dargestellt, beginnt Verfahren **300** bei Schritt **305**, bei dem ein Händler-Auflösungswerkzeug **150** einen Trainingsdatensatz **260** positiver Beispiele von Paaren von Händler-ID-Sätzen **210** erzeugt, die mit derselben Firma verknüpft sind. Das Händler-Auflösungswerkzeug **150** fügt Grenzfälle zu dem Trainingsdatensatz **210** hinzu. Die Grenzfälle beinhalten Paare von Händler-ID-Sätzen **210**, die nicht zusammenpassen, jedoch ähnliche Händler-ID-Zeichenketten aufweisen. Die Grenzfälle können auch Paare von Händler-ID-Sätzen **210** beinhalten, die ähnliche Aggregierungswerte aufweisen, jedoch von unterschiedlichen Firmen sind, so dass sie tatsächlich negative Trainingsbeispiele sind.

[0049] Bei Schritt **310** fügt das Händler-Auflösungswerkzeug **150** zufällig ausgewählte Paare von Händler-ID-Sätzen **210** zum Trainingsdatensatz **260** hinzu. Die zufällig ausgewählten Paare von Händler-ID-Sätzen **210** sollten mehrheitlich negative Trainingsbeispiele enthalten.

[0050] Bei Schritt **315** reicht das Händler-Auflösungswerkzeug **150** jeweilige Händler-ID-Sätze **210** im Trainingsdatensatz **260** an den Aggregierer **240** weiter, um Kandidaten-Aggregierungen **242** zu erzeugen. Beim Trainieren des Klassifikators **255** gibt es keinen Beispiel-Händler-ID-Satz **210(0)**, und daher werden alle Händler-ID-Sätze **210** im Trainingsdatensatz **260** als Kandidaten-Händler-ID-Sätze **210(1)** bis **210(M – 1)** betrachtet. Ein Benutzer kann diese Kandidaten-Aggregierungen **242** überprüfen.

[0051] Bei Schritt **320** wählt ein Benutzer Paare von Händler-ID-Sätzen **210**, die mit derselben Firma verknüpft werden sollten, als positive Trainingsbeispiele aus.

[0052] Bei Schritt **325** wählt ein Benutzer Paare von Händler-ID-Sätzen **210**, die mit unterschiedlichen Fir-

men verknüpft sind, als negative Trainingsbeispiele aus. Diese negativen Trainingsbeispiele beinhalten mehrere schwierige Grenzfälle. Außerdem beinhaltet der Trainingsdatensatz **210** zum überwiegenden Teil zufällige Auswahlen, so dass der überwiegende Teil der Paare von Händler-ID-Sätzen **210** im Trainingsdatensatz **260** negative Trainingsbeispiele sind.

[0053] Bei Schritt **330** trainiert das Händler-Auflösungswerkzeug **150** den Klassifikator **255** mit dem Trainingsdatensatz **260**. Wie beschrieben, ist der Klassifikator **255** ein Random-Forest-Lernalgorithmus.

[0054] Nach einem Trainieren des Klassifikators **255** mit dem Trainingsdatensatz **260** kann der Klassifikator **255** ein Paar von Händler-ID-Sätzen **210** evaluieren, um einen Konfidenzwert, z. B. einen Wert zwischen Null und Eins, zu erzeugen. Der Konfidenzwert entspricht dem Prozentsatz von Entscheidungsbäumen in dem vom Klassifikator **255** verwendeten Random-Forest-Algorithmus, die bestimmen, dass beide Händler-ID-Sätze **210** in dem Paar mit derselben Firma verknüpft werden sollten. Daher ist der Klassifikator **255** in der Lage, einen Konfidenzwert zu erzeugen, der darstellt, ob ein Paar von Händler-ID-Sätzen **210**, die einen Beispiel-Händler-ID-Satz **210(0)** und einen Kandidaten-Händler-ID-Satz **210(1)** beinhalten, mit derselben Firma verknüpft sein sollte.

[0055] **Fig.** 4 ist ein Ablaufdiagramm von Verfahrensschritten, um Händler-IDs mit einer Firma zu verknüpfen, gemäß einer Ausführungsform der Erfindung. Zwar sind die Verfahrensschritte in Verbindung mit den Systemen der **Fig.** 1–**Fig.** 2 und **Fig.** 5 beschrieben, jedoch ist es für Fachleute klar, dass eine beliebige Systemkonfiguration, welche die Verfahrensschritte in beliebiger Reihenfolge durchführt, innerhalb des Schutzumfangs der Erfindung liegt.

[0056] Wie dargestellt, beginnt das Verfahren **400** bei Schritt **410**, wobei das Händler-Auflösungswerkzeug **150** eine Beispiel-Händler-ID als Händler-ID-Attribut für einen Beispiel-Händler-ID-Satz **210(0)** empfängt. Wie beschrieben, wählt ein Benutzer den Beispiel-Händler-ID-Satz **210(0)** als repräsentativ für die Kennzeichen der Finanztransaktionsdatensätze **215** aus, die mit einer Firma assoziiert sind, z. B. den Franchisenehmer, der eine gegebene Franchise-Firma am besten repräsentiert. Alternativ kann das System automatisch einen Beispiel-Händler-ID-Satz **210(0)** basierend auf benutzerspezifischen Kriterien auswählen.

[0057] Bei einer Ausführungsform präsentiert das Händler-Auflösungswerkzeug **150** dem Benutzer ein Beispiel-Auswahlwerkzeug. Das Beispiel-Auswahlwerkzeug bietet Unterstützung beim Auswählen einer Beispiel-Händler-ID, die repräsentativ für eine Firma ist, die aufgelöst werden soll. Das Beispiel-Auswahlwerkzeug kann vom Benutzer eine Such-Zeichenkette erhalten, um Händler-IDs zu identifizieren, die möglicherweise mit der Firma verknüpft werden sollten. Das Beispiel-Auswahlwerkzeug kann auch irgendeine Teilmenge des Firmennamens als Suchzeichenkette verwenden. Außerdem kann das Beispiel-Auswahlwerkzeug die Händler-ID-Sätze **210**, die mit den identifizierten Händler-IDs assoziiert sind, an den Aggregierer **240** weiterreichen. Der Aggregierer **240** berechnet dann Aggregierungen **242**, die den Benutzer beim Auswählen der Beispiel-Händler-ID unterstützen.

[0058] Bei Schritt **420** erzeugt das Händler-Auflösungswerkzeug **150** Beispiel-Aggregierungen **244** für den ausgewählten Beispiel-Händler-ID-Satz **210(0)**. Nachdem das Händler-Auflösungswerkzeug **150** den Beispiel-Händler-ID-Satz **210(0)** aus der Transaktionsdatenbank **160** abgerufen hat, berechnet der Aggregierer **240** die durchschnittliche Transaktionsgröße, die Standardabweichung der Transaktionsgröße und den durchschnittlichen Betrag, den eine Einzelperson ausgegeben hat.

[0059] Bei Schritt **430** erzeugt das Händler-Auflösungswerkzeug **150** Kandidaten-Aggregierungen **242** für einen Kandidaten-Händler-ID-Satz **210(1)**. Das Händler-Auflösungswerkzeug **150** identifiziert einen Händler-ID-Satz **210(1)** bis **210(M – 1)**, der nicht mit dem Beispiel-Händler-ID-Satz **210(0)** verglichen wurde, als Kandidaten-Händler-ID-Satz **210(1)**. Sobald das Identifizieren erfolgt ist, ruft das Händler-Auflösungswerkzeug **150** den Kandidaten-Händler-ID-Satz **210(1)** aus der Transaktionsdatenbank **160** ab, und reicht den Kandidaten-Händler-ID-Satz **210(1)** an den Aggregierer **240** weiter. Der Aggregierer **240** erzeugt die Kandidaten-Aggregierungen **242**.

[0060] Das Aggregieren und Vergleichen jedes möglichen Händler-ID-Datensatzes **210(1)** bis **210(M – 1)** kann sehr zeitaufwändig sein, und daher ist ein Verringern der Anzahl von Vergleichen wünschenswert. Bei einer Ausführungsform wird vom Händler-Auflösungswerkzeug **242** nicht jeder Händler-ID-Datensatz **210** verglichen. Das Händler-Auflösungswerkzeug **242** überspringt Händler-ID-Datensätze **210**, die gewissen Voraussetzungen nicht genügen. Nimmt man an, dass eine Franchise-Firma lediglich Franchisenehmer-Standorte im Staat Kalifornien hat und die Transaktionsdatensätze **215** ein Attribut für die Adresse beinhalten, bei der die Transaktion erfolgt ist, dann würde das Händler-Auflösungswerkzeug **242** solche Händler-ID-Datensätze **210** überspringen, die keine Transaktionsdatensätze **215** aus Kalifornien enthalten. In diesem Fall verringert das Händler-Auflösungswerkzeug **242** die Anzahl von Vergleichen, und zwar dadurch, dass solche Händler-ID-Sätze **210**, die nicht aus Kalifornien sind, übersprungen werden.

[0061] Bei Schritt **440** bestimmt das Händler-Auflösungswerkzeug **150**, ob der Beispiel-Händler-ID-Satz **210(0)** und der Kandidaten-Händler-ID-Satz **210(1)** zueinander passen und daher mit derselben Firma verknüpft werden sollten. Der Identitätsauflöser **250** reicht die Beispiel-Aggregierungen **244** und die Kandidaten-Aggregierungen **242** an den Klassifikator **255** weiter. Wie beschrieben, erzeugt der Klassifikator **255** einen Konfidenzwert zwischen Null und Eins, der dem Prozentsatz von Entscheidungsbäumen in dem vom Klassifikator **255** verwendeten Random-Forest-Algorithmus entspricht, welche bestimmen, dass beide Händler-ID-Sätze **210** in dem Paar mit derselben Firma verknüpft werden sollten. Falls der Klassifikator **255** einen Konfidenzwert unterhalb eines Schwellenwertes erzeugt, dann fährt das Verfahren **400** mit Schritt **460** fort. Falls jedoch der Konfidenzwert oberhalb des Schwellenwertes liegt, dann fährt das Verfahren **400** mit Schritt **450** fort. Zwar hat sich ein Schwellenwert von 0,70 für den Konfidenzwert als effektiv erwiesen, jedoch kann der tatsächliche Schwellenwert je nach Präferenz festgelegt werden.

[0062] Bei Schritt **450** speichert der Identitätsauflöser **250** das Händler-ID-Attribut des Kandidaten-Händler-ID-Satzes **201(1)** in einer Auflösungsliste **270**.

[0063] Bei einer Ausführungsform führt das Händler-Auflösungswerkzeug **242** den Beispiel-Händler-ID-Satz **240(0)** und den Kandidaten-Händler-ID-Satz **240(1)** zu einem kombinierten Händler-ID-Satz zusammen, der zu einem neuen größeren Beispiel-Händler-ID-Satz **240(0)** wird. Dann re-generiert das Händler-Auflösungswerkzeug **242** die Beispiel-Aggregierungen **244** für die verbleibenden Vergleiche. Durch dieses Vorgehen kann der neue Beispiel-Händler-ID-Satz **240(0)** die Firma besser repräsentieren und das Auflösen der verbleibenden Kandidaten-Händler-ID-Sätze **240(2)** bis **240(M – 1)** verbessern.

[0064] Bei Schritt **460** bestimmt das Händler-Auflösungswerkzeug **150**, ob es noch Händler-ID-Sätze **210** in der Transaktionsdatenbank **160** gibt, die noch nicht verglichen wurden. Falls das Händler-Auflösungswerkzeug **150** bestimmt, dass es einen weiteren Kandidaten-Händler-ID-Satz **240(2)** zu vergleichen gibt, dann geht das Verfahren **400** zurück auf Schritt **430**. Sobald keine Kandidaten-Händler-ID-Sätze **210** mehr verbleiben, die zu vergleichen sind, verknüpft das Händler-Auflösungswerkzeug **150** Händler-ID-Sätze **210**, die in der Auflösungsliste **270** für die Firma gelistet sind.

[0065] Bei Schritt **470** verknüpft das Händler-Auflösungswerkzeug **150** den Beispiel-Händler-ID-Satz **210(0)** mit den Kandidaten-Händler-ID-Sätzen **210(1)** bis **210(M – 1)**, die in der Auflösungsliste **270** gelistet sind. Wie beschrieben, kann das Auflösen der Händler-ID-Sätze beinhalten, dass eine Liste von Händler-ID-Attributen gespeichert wird, welche die Analyse-Engine **155** verwenden kann, um die Transaktionsdatensätze **215** der Firma zu identifizieren. Alternativ kann das Händler-Auflösungswerkzeug **150** die Transaktionsdatensätze **215** der Händler-ID-Sätze **210** auf der Auflösungsliste **270** mit der Firma verknüpfen, dadurch dass es in ein Attribut der Transaktionsdatensätze **215** den Firmennamen einschreibt, so dass die Analyse-Engine **155** eine Abfrage der Transaktionsdatenbank **160** nach den zu der Firma gehörenden Transaktionsdatensätzen **215** durchführen kann.

[0066] **Fig.** 5 zeigt ein beispielhaftes Server-Rechnersystem **130**, auf dem ein Händler-Auflösungswerkzeug **150** läuft, gemäß einer Ausführungsform. Wie dargestellt, beinhaltet das Server-Rechnersystem **130** eine Zentralrecheneinheit (CPU) **550**, eine Netzwerk-Schnittstelle **570**, einen Arbeitsspeicher **520** und eine Speichereinrichtung **530**, die jeweils mit einer Kopplungseinrichtung (Bus) **540** verbunden sind. Das Server-Rechnersystem **130** kann auch eine Ein-/Ausgabegerät-Schnittstelle **560** beinhalten, die Ein-/Ausgabegeräte **580** (z. B. Tastatur, Anzeigegerät und Maus) mit dem Rechnersystem **130** verbindet. Weiter können, im Kontext dieser Offenbarung, die Rechnerelemente, die im Server-Rechnersystem **130** gezeigt sind, einem physischen Rechnersystem entsprechen (z. B. einem System in einem Rechenzentrum) oder kann eine virtuelle Rechnerinstanz sein, die in einer Rechner-Cloud ausgeführt wird.

[0067] Die CPU **550** führt ein Auslesen und Speichern von im Arbeitsspeicher **520** gespeicherten Programmanweisungen durch, sowie auch ein Speichern und Auslesen von im Speicher **520** liegenden Anwendungsdaten. Der Bus **540** wird verwendet, um Programmanweisungen und Anwendungsdaten zwischen der CPU **550**, der Ein-/Ausgabegerät-Schnittstelle **560**, der Speichereinrichtung **530**, der Netzwerk-Schnittstelle **570** und dem Arbeitsspeicher **520** zu übertragen. Es sei angemerkt, dass hier inbegriffen ist, dass die CPU **550** repräsentativ ist für eine einzelne CPU, mehrere CPUs, eine einzelne CPU, die mehrere Prozessorkerne hat, eine CPU mit zugehöriger Speicherverwaltungseinheit, und dergleichen. Generell inbegriffen ist hier der Arbeitsspeicher **520** repräsentativ für einen Direktzugriffsspeicher (RAM). Die Speichereinrichtung **530** kann eine Plattenlaufwerk-Speichervorrichtung sein. Obschon als einzelne Einheit dargestellt, kann die Speichereinrichtung **530** eine Kombination aus fest eingebauten und/oder entnehmbaren Speichervorrichtungen sein, beispielsweise fest eingebaute Plattenlaufwerke, entnehmbare Speicherkarten oder eine optische Speichereinrichtung, ein netzgebundener Speicher (NAS) oder ein Speichernetzwerk (SAN).

[0068] Die Kommunikationen zwischen dem Client **120** und dem Händler-Auflösungswerkzeug **150** werden über das Netzwerk **180** via Netzwerkschnittstelle **570** übertragen.

[0069] Als erläuternde Darstellung beinhaltet der Arbeitsspeicher **520** ein Händler-Auflösungswerkzeug **150**, Beispiel-Aggregierungen **244**, Kandidaten-Aggregierungen **242** und eine Auflösungsliste **270**. Das Händler-Auflösungswerkzeug **150** selbst beinhaltet einen Aggregierer **240** und einen Klassifikator **225**. Die Speichereinrichtung **530** beinhaltet einen Trainingsdatensatz **533**, den das Händler-Auflösungswerkzeug **150** verwendet, um den Klassifikator **225** zu trainieren.

[0070] Der Aggregierer **240** erzeugt die Beispiel-Aggregierungen **244** und die Kandidaten-Aggregierungen **242** aus den Transaktionsdatensätzen **215**, die aus der Transaktionsdatenbank **160** abgerufen wurden. Das Händler-Auflösungswerkzeug **150** gibt Datenbankabfragen über das Netzwerk **180** an die Transaktionsdatenbank **160** via Netzwerkschnittstelle **570** aus. Sobald der Aggregierer **240** die Beispiel-Aggregierungen **244** und Kandidaten-Aggregierungen **242** erzeugt hat, verwendet das Händler-Auflösungswerkzeug **150** den Klassifikator **225**, um zu bestimmen, ob die Händler-ID-Sätze **240** mit einer Firma verknüpft werden sollten.

[0071] Obschon hier im Arbeitsspeicher **520** dargestellt, ist es auch möglich, dass das Händler-Auflösungswerkzeug **150**, die Beispiel-Aggregierungen **244**, die Kandidaten-Aggregierungen **242** und die Auflösungsliste **270** gespeichert sind im Arbeitsspeicher **520**, in der Speichereinrichtung **530**, oder aufgeteilt zwischen Arbeitsspeicher **520** und Speichereinrichtung **530**. In ähnlicher Weise kann der Trainingsdatensatz **533** im Arbeitsspeicher **520**, der Speichereinrichtung **530**, oder aufgeteilt zwischen Arbeitsspeicher **520** und Speichereinrichtung **530**, gespeichert sein.

[0072] Bei einigen Ausführungsformen kann sich das Datenbank-Repository **160** in der Speichereinrichtung **530** befinden. In einem derartigen Fall werden die Datenbankabfragen und anschließenden Antworten über den Bus **540** übertragen. Wie beschrieben, kann sich der Client **120** auch auf dem Server-Rechnersystem **130** befinden, in welchem Fall der Client **120** auch im Arbeitsspeicher **520** gespeichert würde und der Benutzer die Ein-/Ausgabegeräte **580** verwenden würde, um mit dem Client **120** über die Ein-/Ausgabegerät-Schnittstelle **560** zu interagieren.

[0073] Obschon das zuvor Beschriebene auf Ausführungsformen der Erfindung abzielt, können andere und weitere Ausführungsformen der Erfindung erdacht werden, ohne von deren grundlegendem Schutzumfang abzuweichen. Beispielsweise können Aspekte der Erfindung in Hardware oder in Software implementiert sein, oder in einer Kombination aus Hardware und Software. Eine Ausführungsform der Erfindung kann als Programmprodukt zur Verwendung mit einem Computersystem implementiert sein. Das/die Programm(e) des Programmproduktes definieren Funktionen der Ausführungsformen (einschließlich der hier beschriebenen Verfahren) und kann sich auf einer Vielfalt von computerlesbaren Speichermedien befinden. Beispiele von computerlesbaren Speichermedien beinhalten (i) nicht-beschreibbare Speichermedien (z. B. Nur-Lese-Speichervorrichtungen in einem Computer, CD-ROMs, die durch ein CD-ROM-Laufwerk gelesen werden können, einen Flash-Speicher, ROM-Speicherbausteine oder einen beliebigen Typ von nicht-flüchtigem Festkörper-Halbleiterspeicher); und (ii) beschreibbare Speichermedien (z. B. Floppy-Disks in einem Diskettenlaufwerk oder ein Festplatten-Laufwerk oder einen beliebigen Typ von Direktzugriff-Festkörper-Halbleiterspeicher), auf denen veränderbare Information gespeichert ist.

[0074] Die Erfindung wurde im Vorhergehenden mit Bezug auf spezielle Ausführungsformen beschrieben. Für Fachleute ist es jedoch klar, dass verschiedene Modifikationen und Änderungen an diesen vorgenommen werden können, ohne von dem weiter gefassten Geist und Schutzumfang der Erfindung abzuweichen, der in den anliegenden Ansprüchen dargelegt ist. Die vorhergehende Beschreibung und die Zeichnungen sind demgemäß vielmehr erläuternd und nicht-einschränkend zu verstehen.

[0075] Daher ist der Schutzumfang der Erfindung durch die folgenden Ansprüche bestimmt.

Patentansprüche

1. Verfahren zum Identifizieren von in Beziehung stehenden Transaktionsdatensätzen aus einer Datenbank, die Transaktionsdatensätze für mehrere Entitäten speichert, wobei das Verfahren umfasst:
Abrufen einer Mehrzahl von Sätzen von Transaktionsdatensätzen, wobei jeder Satz von Transaktionsdatensätzen einen oder mehrere der Transaktionsdatensätze beinhaltet, die einen gemeinsamen Attributwert teilen;
Entgegennehmen einer Auswahl oder Auswählen eines Satzes von Beispiel-Datensätzen, wobei der Satz von Beispiel-Datensätzen eine Mehrzahl der Transaktionsdatensätze aufweist, die mit einer ersten Entität assoziiert sind;
für jeden der Mehrzahl von Sätzen von Transaktionsdatensätzen:
Bestimmen einer Wahrscheinlichkeit, dass der Satz von Transaktionsdatensätzen Transaktionsdatensätze speichert, die mit der ersten Entität assoziiert sind, und

bei Bestimmen, dass die Wahrscheinlichkeit einen Schwellenwert übersteigt, Auflösen des Satzes von Transaktionsdatensätzen als ein Satz, der Transaktionsdatensätze speichert, die mit der ersten Entität assoziiert sind.

2. Verfahren nach Anspruch 1, wobei das Bestimmen einer Wahrscheinlichkeit umfasst, dass ein Satz von Transaktionsdatensätzen und der Satz von Beispiel-Datensätzen an einen Klassifikator weitergegeben wird, wobei der Klassifikator konfiguriert ist, um die Wahrscheinlichkeit zu bestimmen, dass der Satz von Transaktionsdatensätzen Transaktionsdatensätze speichert, die mit der ersten Entität assoziiert sind.

3. Verfahren nach Anspruch 2, wobei der Klassifikator ein Random-Forest-Klassifikator ist.

4. Verfahren nach Anspruch 2 oder Anspruch 3, das weiter umfasst, dass der Klassifikator unter Verwendung einer Mehrzahl von Trainingsbeispielen trainiert wird, wobei die Trainingsbeispiele aufweisen:
eines oder mehrere erste Paare von Sätzen von Transaktionsdatensätzen, wobei jedes erste Paar eine gemeinsame Entität repräsentiert; und
eines oder mehrere zweite Paare von Sätzen von Transaktionsdatensätzen, wobei jedes zweite Paar nicht in Beziehung stehende Entitäten repräsentiert.

5. Verfahren nach einem der Ansprüche 2 bis 4, wobei der Klassifikator Merkmale eines jeden Transaktionsdatensatzes evaluiert, einschließlich mindestens eines der Folgenden, und zwar eines Wortüberschneidungszählwerts, einer Worthäufigkeit, einer wortbasierten oder zeichenbasierten Kosinus-Ähnlichkeit, eines Händlerklassencodes, und numerischer Stadtcodes, die mit jedem Transaktionsdatensatz assoziiert sind.

6. Verfahren nach einem der Ansprüche 2 bis 5, wobei der Klassifikator Merkmale eines jeden Transaktionsdatensatzes evaluiert, einschließlich mindestens eines der Folgenden, und zwar einer fraktionalen Differenz einer Größe eines durchschnittlichen Transaktionsbetrags im Transaktionsdatensatz, einer Standardabweichung zwischen den durchschnittlichen Transaktionsbeträgen in den Transaktionsdatensätzen, und einer fraktionalen Differenz einer Größe von Transaktionsbetrag-Varianzen.

7. Verfahren nach einem der Ansprüche 1 bis 6, wobei das Auflösen des Transaktionsdatensatz-Satzes auf den Satz von Beispiel-Datensätzen umfasst:
Zusammenführen der Transaktionsdatensätze des Satzes von Transaktionsdatensätzen in den Satz von Beispiel-Datensätzen.

8. Verfahren nach einem der Ansprüche 1 bis 7, das weiter umfasst, dass eine Analyse bei einem Satz der Transaktionsdatensätze durchgeführt wird, wobei der Satz die Transaktionsdatensätze des Satzes von Beispiel-Datensätzen und die Transaktionsdatensätze beinhaltet, die als assoziiert mit der ersten Entität aufgelöst wurden.

9. Verfahren nach einem der Ansprüche 1 bis 8, wobei die Transaktionsdatensätze Finanztransaktionsdatensätze sind, insbesondere Kredit- oder Debittransaktionen, die durch eine Finanzinstitution für einen Händler verarbeitet werden.

10. Verfahren nach Anspruch 9, wobei Attribute der Finanztransaktionsdatensätze eines oder mehrere von den Folgenden beinhalten:
eine Identifikation des Händlers, von dem die Finanztransaktion stammt;
eine Identifikation des Eigentümers des Kredit- oder Debitkontos;
einen Betrag der Finanztransaktion;
ein Datum der Finanztransaktion;
einen Zeitpunkt der Finanztransaktion; und
einen Ort, von dem aus die Finanztransaktion durchgeführt wurde.

11. Verfahren nach Anspruch 10, das weiter umfasst, dass für jeden Satz von Transaktionsdatensätzen Aggregierungswerte für die Attribute des Satzes von Transaktionsdatensätzen bestimmt werden; und Bestimmen von Aggregierungswerten für Attribute des Satzes von Beispiel-Datensätzen.

12. Computerlesbares Medium, das Anweisungen speichert, die, wenn sie durch einen Prozessor ausgeführt werden, den Prozessor veranlassen, Operationen durchzuführen, welche die Operationen nach einem der Ansprüche 1 bis 11 beinhalten.

13. Computersystem, aufweisend:
einen Speicher; und
einen Prozessor, der eines oder mehrere Programme speichert, die konfiguriert sind, um Operationen durchzuführen, welche die Operationen nach einem der Ansprüche 1 bis 11 beinhalten.

Es folgen 5 Seiten Zeichnungen

Anhängende Zeichnungen

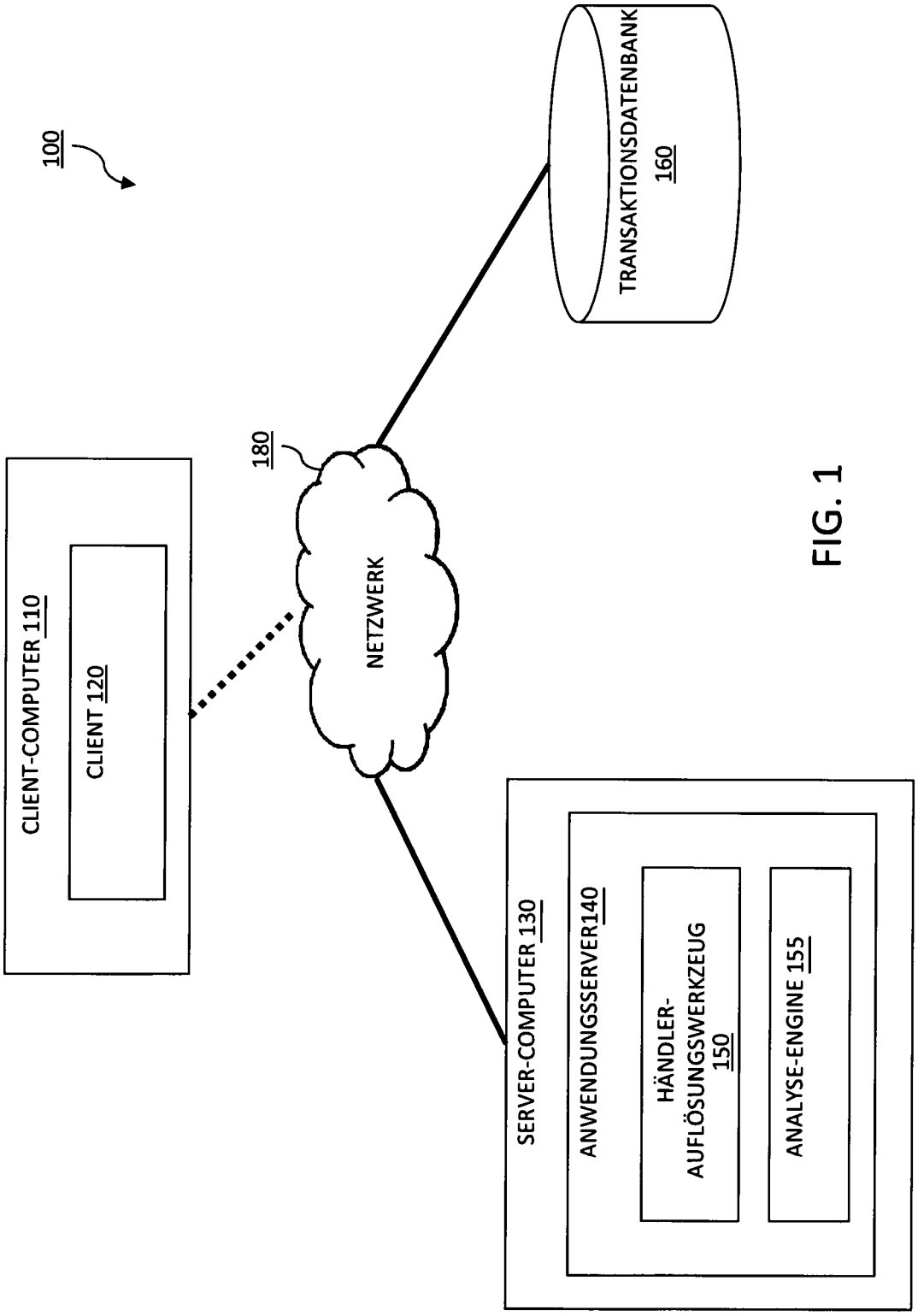


FIG. 1

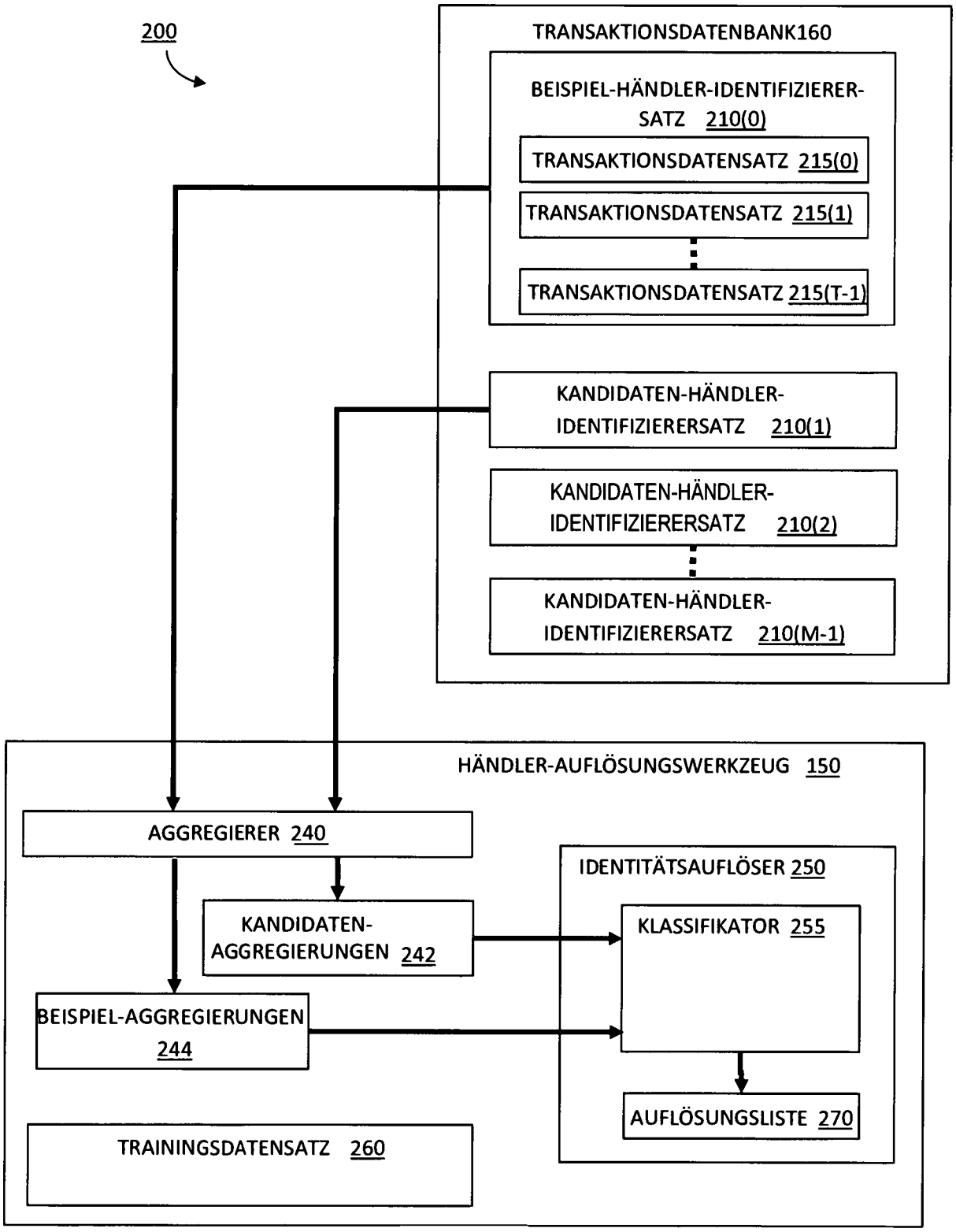

200
TRANSAKTIONSDATENBANK160
BEISPIEL-HÄNDLER-IDENTIFIZIERER-SATZ 210(0)
TRANSAKTIONSDATENSATZ 215(0)
TRANSAKTIONSDATENSATZ 215(1)
TRANSAKTIONSDATENSATZ 215(T-1)
KANDIDATEN-HÄNDLER-IDENTIFIZIERERSATZ 210(1)
KANDIDATEN-HÄNDLER-IDENTIFIZIERERSATZ 210(2)
KANDIDATEN-HÄNDLER-IDENTIFIZIERERSATZ 210(M-1)
HÄNDLER-AUFLÖSUNGSWERKZEUG 150
AGGREGIERER 240
KANDIDATEN-AGGREGIERUNGEN 242
BEISPIEL-AGGREGIERUNGEN 244
IDENTITÄTSAUFLÖSER 250
KLASSIFIKATOR 255
AUFLÖSUNGSLISTE 270
TRAININGSDATENSATZ 260

FIG. 2

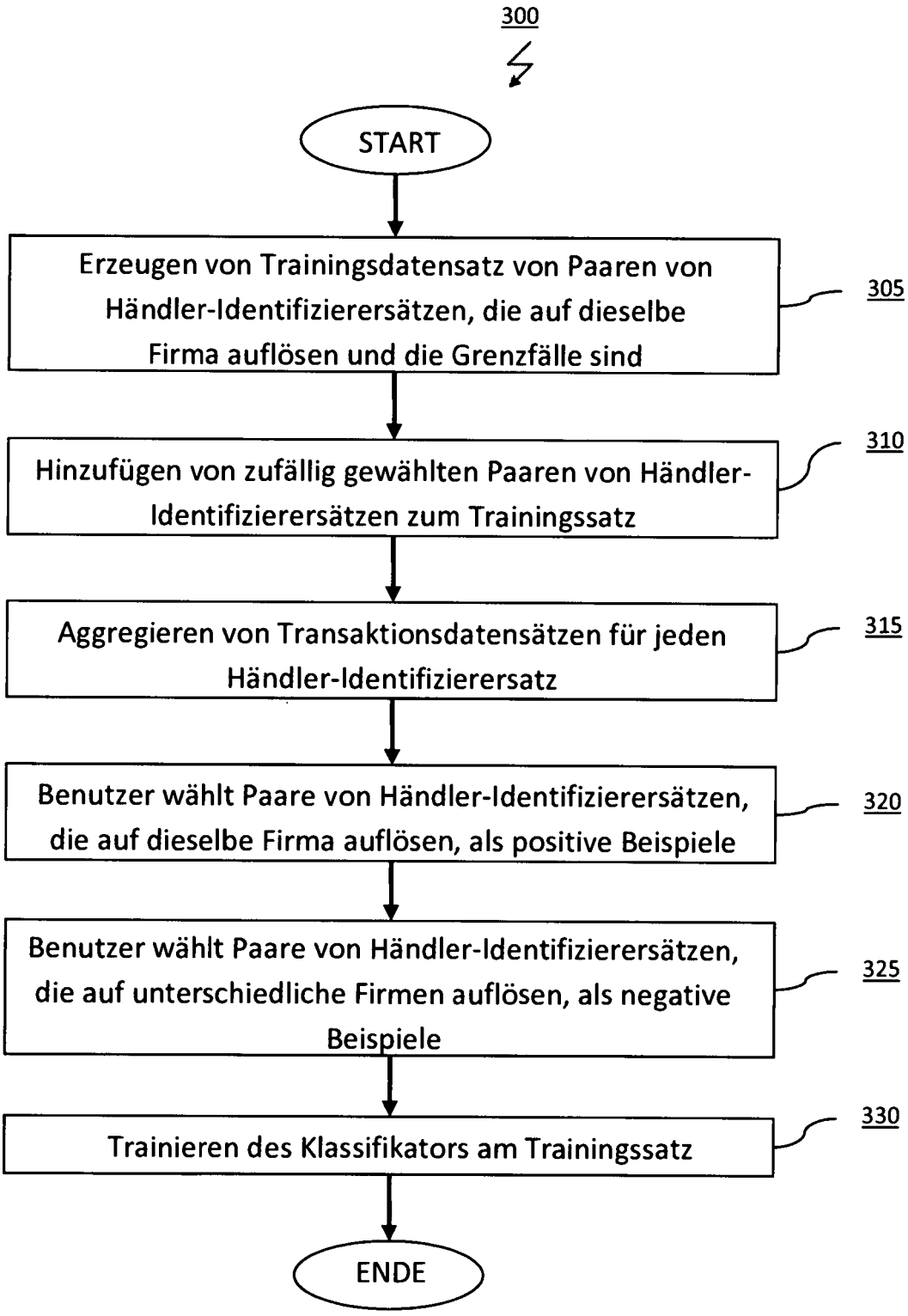

300
START
Erzeugen von Trainingsdatensatz von Paaren von Händler-Identifizierersätzen, die auf dieselbe Firma auflösen und die Grenzfälle sind
305
Hinzufügen von zufällig gewählten Paaren von Händler-Identifizierersätzen zum Trainingssatz
310
Aggregieren von Transaktionsdatensätzen für jeden Händler-Identifizierersatz
315
Benutzer wählt Paare von Händler-Identifizierersätzen, die auf dieselbe Firma auflösen, als positive Beispiele
320
Benutzer wählt Paare von Händler-Identifizierersätzen, die auf unterschiedliche Firmen auflösen, als negative Beispiele
325
Trainieren des Klassifikators am Trainingssatz
330
ENDE

FIG. 3

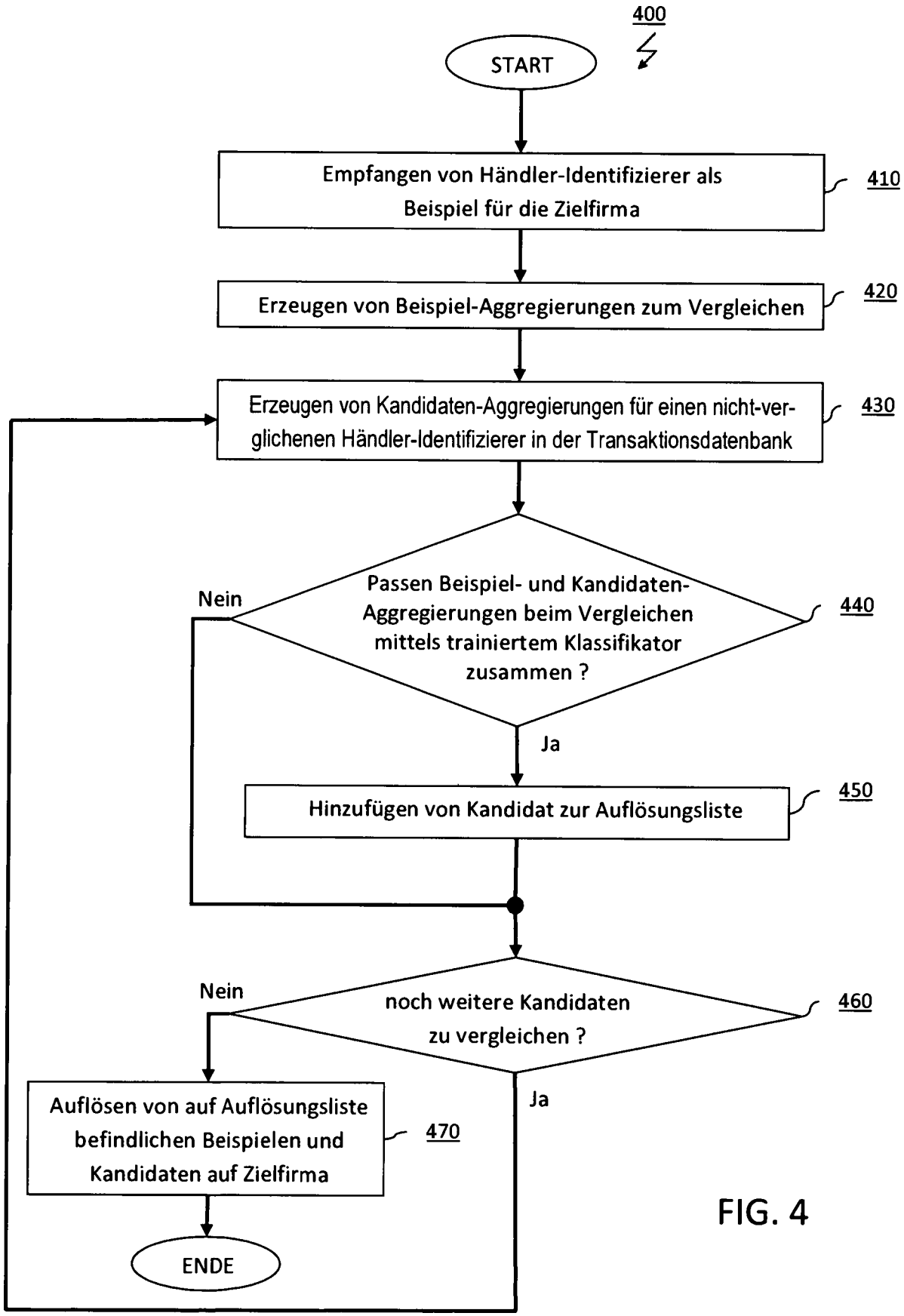

400
START
Empfangen von Händler-Identifizierer als Beispiel für die Zielfirma
410
Erzeugen von Beispiel-Aggregierungen zum Vergleichen
420
Erzeugen von Kandidaten-Aggregierungen für einen nicht-verglichenen Händler-Identifizierer in der Transaktionsdatenbank
430
Passen Beispiel- und Kandidaten-Aggregierungen beim Vergleichen mittels trainiertem Klassifikator zusammen ?
440
Nein
Ja
Hinzufügen von Kandidat zur Auflösungsliste
450
noch weitere Kandidaten zu vergleichen ?
460
Nein
Ja
Auflösen von auf Auflösungsliste befindlichen Beispielen und Kandidaten auf Zielfirma
470
ENDE

FIG. 4

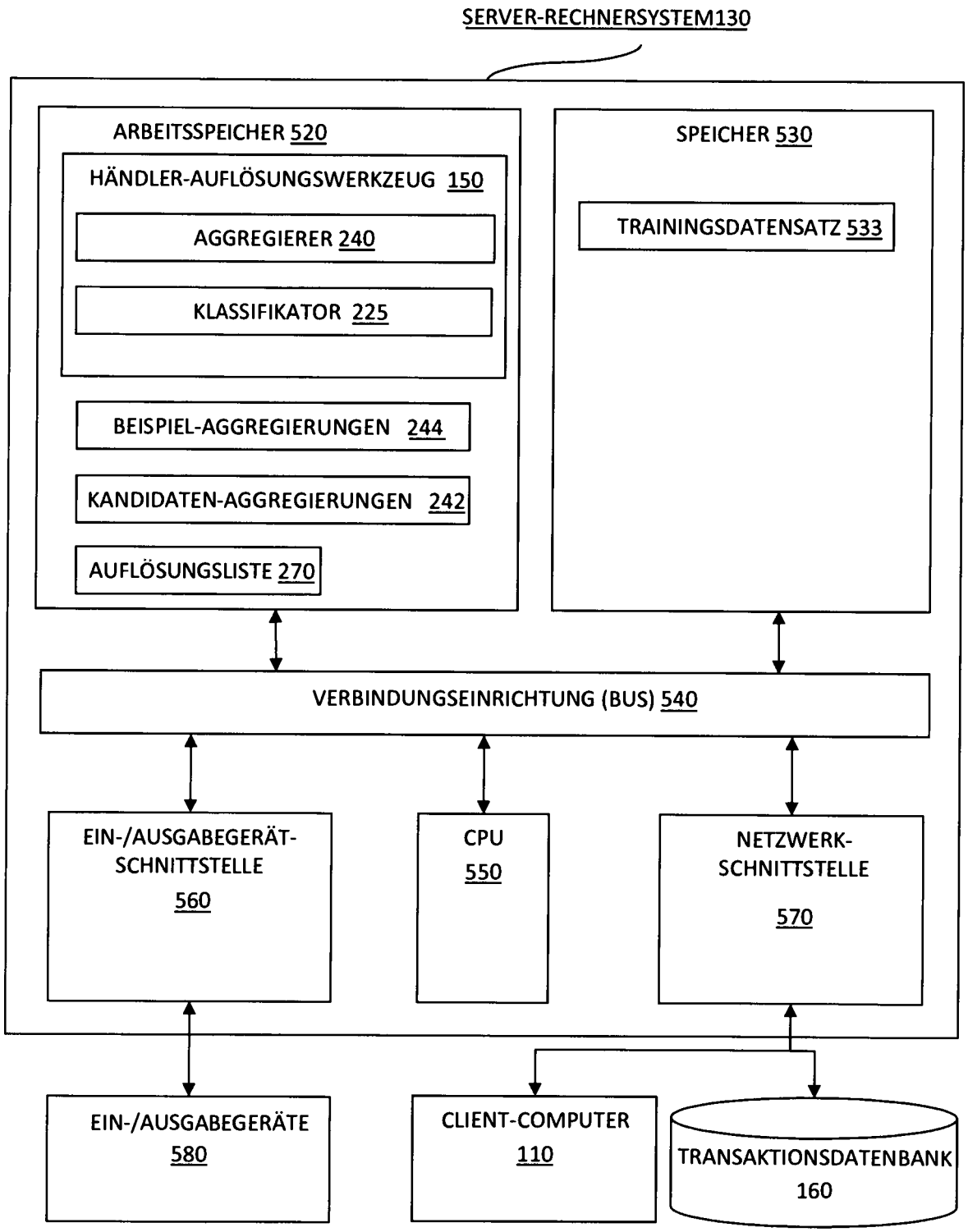

SERVER-RECHNERSYSTEM130
ARBEITSSPEICHER 520
HÄNDLER-AUFLÖSUNGSWERKZEUG 150
AGGREGIERER 240
KLASSIFIKATOR 225
BEISPIEL-AGGREGIERUNGEN 244
KANDIDATEN-AGGREGIERUNGEN 242
AUFLÖSUNGSLISTE 270
SPEICHER 530
TRAININGSDATENSATZ 533
VERBINDUNGSEINRICHTUNG (BUS) 540
EIN-/AUSGABEGERÄT-SCHNITTSTELLE 560
CPU 550
NETZWERK-SCHNITTSTELLE 570
EIN-/AUSGABEGERÄTE 580
CLIENT-COMPUTER 110
TRANSAKTIONSDATENBANK 160

FIG.5