

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7512402号
(P7512402)

(45)発行日 令和6年7月8日(2024.7.8)

(24)登録日 令和6年6月28日(2024.6.28)

(51)国際特許分類

F I

G 0 6 F 16/215(2019.01)

G 0 6 F 16/215

請求項の数 40 (全57頁)

(21)出願番号	特願2022-550704(P2022-550704)	(73)特許権者	509123208
(86)(22)出願日	令和3年2月25日(2021.2.25)		アビニシオ テクノロジー エルエルシー
(65)公表番号	特表2023-516139(P2023-516139 A)		アメリカ合衆国 0 2 4 2 1 マサチュー
			セッツ州 レキシントン スプリング ス
(43)公表日	令和5年4月18日(2023.4.18)		トリート 2 0 1
(86)国際出願番号	PCT/US2021/019572	(74)代理人	100079108
(87)国際公開番号	WO2021/173777		弁理士 稲葉 良幸
(87)国際公開日	令和3年9月2日(2021.9.2)	(74)代理人	100109346
審査請求日	令和6年2月22日(2024.2.22)		弁理士 大貫 敏史
(31)優先権主張番号	62/981,646	(74)代理人	100117189
(32)優先日	令和2年2月26日(2020.2.26)		弁理士 江口 昭彦
(33)優先権主張国・地域又は機関	米国(US)	(74)代理人	100134120
			弁理士 内藤 和彦
(31)優先権主張番号	17/006,504	(72)発明者	ジョイス, ジョン
(32)優先日	令和2年8月28日(2020.8.28)		アメリカ合衆国, マサチューセッツ州
	最終頁に続く		最終頁に続く

(54)【発明の名称】 データフィールドの意味論的ラベルからのデータフィールドのデータ処理値に対するルールの生成

(57)【特許請求の範囲】

【請求項1】

データ記録のセット内のデータ記録のフィールド内の値に対するデータ品質ルールを、前記フィールドに関連付けられたラベルに基づいて決定するための方法であって、前記ラベルは、前記フィールドの前記値の特性を示し、前記方法は、データ処理システムによって実装され、

ラベルをデータ記録内の1つ以上のフィールドのセットに関連付けるラベルインデックスを取り出すことであって、前記ラベルは、前記1つ以上のフィールドのセットのそれぞれのフィールドで予想される情報のタイプを識別する、ことと、

前記ラベルによって示される前記情報のタイプを、前記ラベルに関連付けられた前記1つ以上のフィールドの値に対する要件を表す属性値のセットに関連付けるデータ辞書にアクセスすることであって、前記要件は、前記1つ以上のフィールドの前記値の論理的又は構文的特性を含む、ことと、

特定のデータ記録のフィールドについて、

前記ラベルインデックスにアクセスすることによって、前記特定のデータ記録の前記フィールドに関連付けられた特定のラベルを識別することと、

前記データ辞書から、前記特定のラベルの属性値を取り出すことであって、前記属性値は、前記フィールドに対する特定の要件を指定する、ことと、

データ品質ルールを生成することであって、前記データ品質ルールは、実行されると、

前記フィールドの値が前記属性値によって表される前記特定の要件を満たすかどうか

を検証し、

前記特定の要件が満たされているかどうかを示す出力データを生成するように構成されている、ことと、を含む、方法。

【請求項 2】

前記データ品質ルールは、

前記要件からの前記フィールド内の前記値の許容偏差、

前記フィールド内の 1 つ以上の許容値、及び

前記フィールド内の 1 つ以上の禁止値のうちの 1 つ以上を示す、請求項 1 に記載の方法。

【請求項 3】

前記フィールド内の前記 1 つ以上の許容値又は前記 1 つ以上の禁止値は、前記フィールドのフィールド名に関連付けられている、請求項 2 に記載の方法。

【請求項 4】

前記フィールドは第 1 のフィールドであり、前記フィールド内の前記 1 つ以上の許容値又は前記 1 つ以上の禁止値は、前記特定のデータ記録内の第 2 のフィールド内の値に基づいて決定され、前記第 2 のフィールドは、前記特定のデータ記録の前記第 1 のフィールドに結び付けられている、請求項 2 に記載の方法。

【請求項 5】

前記 1 つ以上の許容値又は前記 1 つ以上の禁止値は、前記第 2 のフィールド内の前記値と前記第 1 のフィールド内の前記値との組み合わせに基づく、請求項 4 に記載の方法。

【請求項 6】

前記 1 つ以上の許容値は、数値関数を満たす値に対応し、前記 1 つ以上の禁止値は、前記数値関数を満たさない値に対応する、請求項 2 に記載の方法。

【請求項 7】

前記フィールドは第 1 のフィールドであり、前記データ記録は第 2 のフィールドを含み、前記データ品質ルールは第 1 のデータ品質ルールであり、前記方法は、前記特定のラベルの前記属性値に基づいて、前記第 1 のフィールドと前記第 2 のフィールドとの間に関係が存在すると決定することと、

前記第 2 のフィールドと前記第 1 のフィールドとの間の前記関係に基づいて、前記第 2 のフィールドに対する第 2 のデータ品質ルールを生成することと、を含む、請求項 1 に記載の方法。

【請求項 8】

前記関係は、前記第 1 のフィールドの値に対する前記第 2 のフィールドの値の依存性、又は前記第 2 のフィールドの前記値に対する前記第 1 のフィールドの前記値の依存性を示す、請求項 7 に記載の方法。

【請求項 9】

前記関係は、前記第 1 のフィールドの値と前記第 2 のフィールドの値との間の相関関係を示す、請求項 7 に記載の方法。

【請求項 10】

前記第 1 のフィールド及び前記第 2 のフィールドのそれぞれの値の前記関係を検証する検証データを取得することを更に含む、請求項 7 に記載の方法。

【請求項 11】

前記関係は、前記第 1 のフィールド及び前記第 2 のフィールドの値の閾値数について検証される、請求項 7 に記載の方法。

【請求項 12】

前記第 2 のデータ品質ルールは、前記第 1 のフィールドの値に基づいて、前記第 2 のフィールドの値に制約を強制するように構成されている、請求項 7 に記載の方法。

【請求項 13】

前記第 1 のフィールドと前記第 2 のフィールドとの間に前記関係が存在すると決定することは、前記第 2 のフィールドの値が前記第 1 のフィールドの値によって参照されるキー値を含むと決定することを含み、前記データ品質ルールは、前記第 2 のフィールドのそれ

10

20

30

40

50

それぞれの値が有効なキー値であることを要求するように構成されている、請求項 7 に記載の方法。

【請求項 14】

前記第 1 のフィールド及び前記第 2 のフィールドはそれぞれ数値を含み、前記第 1 のフィールドと前記第 2 のフィールドとの間に前記関係が存在すると決定することは、前記第 1 のフィールドの値を前記第 2 のフィールドの値に結び付ける数値関数を決定することを含む、請求項 7 に記載の方法。

【請求項 15】

前記第 1 のフィールドと前記第 2 のフィールドとの間に前記関係が存在すると決定することは、機械学習プロセスによって構成された少なくとも 1 つの分類子を使用することを含む、請求項 7 に記載の方法。

10

【請求項 16】

前記特定のラベルに関連付けられた前記属性値が、前記フィールドが前記特定のデータ記録の一次キー値を含むことを示すと決定することと、

前記一次キー値が前記フィールド内でそれぞれ固有であることを要求するようにデータ品質ルールを構成することと、を含む、請求項 1 に記載の方法。

【請求項 17】

前記特定のラベルの前記属性値は、前記フィールド内の前記値の平均、前記値の最大長、前記値の最小長、前記値のデータタイプ、及び前記値のフォーマットのうちの少なくとも 1 つを表す、請求項 1 に記載の方法。

20

【請求項 18】

前記特定のデータ記録は第 1 のデータ記録であり、前記方法は、前記第 1 のデータ記録とは異なる第 2 のデータ記録内の前記特定のラベルに関連付けられた別のフィールドに前記データ品質ルールを適用することを含む、請求項 1 に記載の方法。

【請求項 19】

前記方法は、前記データ品質ルール及び前記特定のラベルに関連付けるデータを記憶することを含む、請求項 1 に記載の方法。

【請求項 20】

前記データ品質ルールの承認を要求するためのデータを生成することと、

前記データ品質ルールの承認を示す承認データを取得することに応答して、前記データ品質ルールを承認することと、を更に含む、請求項 1 に記載の方法。

30

【請求項 21】

前記データ品質ルールを生成することは、

前記フィールド内の値の履歴傾向を決定することと、

前記フィールド内の前記値の前記履歴傾向に基づいて、要件を生成することと、を含む、請求項 1 に記載の方法。

【請求項 22】

前記フィールドに対する前記データ品質ルールを生成することは、

前記フィールド内の値の履歴平均を識別することと、

前記フィールド内の前記値の前記履歴平均に基づいて、要件を生成することと、を含む、請求項 1 に記載の方法。

40

【請求項 23】

1 つ以上のデータセット内の値に対するデータ品質ルールを決定するための方法であって、前記方法は、データ処理システムによって実装され、

前記データ処理システムによって、1 つ以上のデータセットのデータ記録に含まれるフィールドに記憶された値をプロファイルすることと、

前記データ処理システムによって、1 つ以上の分類子を前記プロファイルされた値に適用することと、

前記 1 つ以上の分類子を適用することに基づいて、前記フィールドの前記値の論理的又は構文的特性を示す 1 つ以上の属性を識別することであって、前記 1 つ以上の属性のそれ

50

それは、前記1つ以上の分類子のそれぞれの出力に基づく、対応の信頼レベルを有する、ことと、

前記信頼レベルが閾値レベルを満たす、前記識別された属性のうちの1つ以上を前記フィールドに関連付けることと、

前記フィールドに関連付けられた前記1つ以上の属性に基づいて、前記フィールドに含まれる値に対する1つ以上の制約を決定することと、

前記決定された1つ以上の制約に基づいて、前記データセットの前記フィールドに対するデータ品質ルールを決定することであって、前記データ品質ルールは、前記フィールドの値と前記1つ以上の制約との間の関係を示す、ことと、を含む、方法。

【請求項24】

前記データ品質ルールは、

前記1つ以上の制約からの前記フィールド内の前記値の許容偏差、

前記フィールド内の1つ以上の許容値、及び

前記フィールド内の1つ以上の禁止値のうちの1つ以上を示す、請求項23に記載の方法。

【請求項25】

前記フィールドは第1のフィールドであり、前記データ記録は第2のフィールドを含み、前記データ品質ルールは第1のデータ品質ルールであり、前記方法は、

前記フィールドに関連付けられた前記1つ以上の属性に基づいて、前記第1のフィールドと前記第2のフィールドとの間に関係が存在すると決定することと、

前記第2のフィールドと前記第1のフィールドとの間の前記関係に基づいて、前記第2のフィールドに対する第2のデータ品質ルールを生成することと、を含む、請求項23に記載の方法。

【請求項26】

前記フィールドに関連付けられた前記1つ以上の属性が、前記フィールドが特定のデータ記録の一次キー値を含むことを示すと決定することと、

前記一次キー値が前記フィールド内でそれぞれ固有であることを要求するようにデータ品質ルールを構成することと、を含む、請求項23に記載の方法。

【請求項27】

前記フィールドに関連付けられた前記1つ以上の属性は、前記フィールド内の前記値の平均、前記値の最大長、前記値の最小長、前記値のデータタイプ、及び前記値のフォーマットのうちの少なくとも1つを表す、請求項23に記載の方法。

【請求項28】

前記方法は、別のデータ記録内の別のフィールドに前記データ品質ルールを適用することを含む、請求項23に記載の方法。

【請求項29】

データ記録のセット内のデータ記録のフィールド内の値に対するデータ品質ルールを、前記フィールドに関連付けられたラベルに基づいて決定するためのデータ処理システムであって、前記ラベルは、前記フィールドの前記値の特性を示し、前記データ処理システムは、

少なくとも1つの処理デバイスと、

前記少なくとも1つの処理デバイスと通信する少なくとも1つのメモリであって、前記少なくとも1つのメモリは、命令を記憶し、前記命令は、前記少なくとも1つの処理デバイスによって実行されると、前記少なくとも1つの処理デバイスに、

ラベルをデータ記録内の1つ以上のフィールドのセットに関連付けるラベルインデックスを取り出すことであって、前記ラベルは、前記1つ以上のフィールドのセットのそれぞれのフィールドで予想される情報のタイプを識別する、ことと、

前記ラベルによって示される前記情報のタイプを、前記ラベルに関連付けられた前記1つ以上のフィールドの値に対する要件を表す属性値のセットに関連付けるデータ辞書にアクセスすることであって、前記要件は、前記1つ以上のフィールドの前記値の論理的又は

10

20

30

40

50

構文的特性を含む、ことと、

特定のデータ記録のフィールドについて、

前記ラベルインデックスにアクセスすることによって、前記特定のデータ記録の前記フィールドに関連付けられた特定のラベルを識別することと、

前記データ辞書から、前記特定のラベルの属性値を取り出すことであって、前記属性値は、前記フィールドに対する特定の要件を指定する、ことと、

データ品質ルールを生成することであって、前記データ品質ルールは、実行されると、

前記フィールドの値が前記属性値によって表される前記特定の要件を満たすかどうかを検証し、

前記特定の要件が満たされているかどうかを示す出力データを生成するように構成されている、ことと、含む動作を実行させる、少なくとも1つのメモリと、を含む、データ処理システム。

10

【請求項30】

前記データ品質ルールは、

前記要件からの前記フィールド内の前記値の許容偏差、

前記フィールド内の1つ以上の許容値、及び

前記フィールド内の1つ以上の禁止値のうちの1つ以上を示す、請求項29に記載のデータ処理システム。

【請求項31】

前記フィールドは第1のフィールドであり、前記データ記録は第2のフィールドを含み、前記データ品質ルールは第1のデータ品質ルールであり、前記動作は、

20

前記特定のラベルの前記属性値に基づいて、前記第1のフィールドと前記第2のフィールドとの間に関係が存在すると決定することと、

前記第2のフィールドと前記第1のフィールドとの間の前記関係に基づいて、前記第2のフィールドに対する第2のデータ品質ルールを生成することと、を含む、請求項29に記載のデータ処理システム。

【請求項32】

前記動作は、

前記特定のラベルに関連付けられた前記属性値が、前記フィールドが前記特定のデータ記録の一次キー値を含むことを示すと決定することと、

30

前記一次キー値が前記フィールド内でそれぞれ固有であることを要求するようにデータ品質ルールを構成することと、を含む、請求項29に記載のデータ処理システム。

【請求項33】

前記特定のラベルの前記属性値は、前記フィールド内の前記値の平均、前記値の最大長、前記値の最小長、前記値のデータタイプ、及び前記値のフォーマットのうちの少なくとも1つを表す、請求項29に記載のデータ処理システム。

【請求項34】

前記特定のデータ記録は第1のデータ記録であり、前記動作は、前記第1のデータ記録とは異なる第2のデータ記録内の前記特定のラベルに関連付けられた別のフィールドに前記データ品質ルールを適用することを含む、請求項29に記載のデータ処理システム。

40

【請求項35】

データ記録のセット内のデータ記録のフィールド内の値に対するデータ品質ルールを、前記フィールドに関連付けられたラベルに基づいて決定するための命令を記憶する1つ以上の非一時的コンピュータ可読媒体であって、前記ラベルは、前記フィールドの前記値の特性を示し、前記命令は、少なくとも1つの処理デバイスによって実行されると、前記少なくとも1つの処理デバイスに、

ラベルをデータ記録内の1つ以上のフィールドのセットに関連付けるラベルインデックスを取り出すことであって、前記ラベルは、前記1つ以上のフィールドのセットのそれぞれのフィールドで予想される情報のタイプを識別する、ことと、

前記ラベルによって示される前記情報のタイプを、前記ラベルに関連付けられた前記1

50

つ以上のフィールドの値に対する要件を表す属性値のセットに関連付けるデータ辞書にアクセスすることであって、前記要件は、前記1つ以上のフィールドの前記値の論理的又は構文的特性を含む、ことと、

特定のデータ記録のフィールドについて、

前記ラベルインデックスにアクセスすることによって、前記特定のデータ記録の前記フィールドに関連付けられた特定のラベルを識別することと、

前記データ辞書から、前記特定のラベルの属性値を取り出すことであって、前記属性値は、前記フィールドに対する特定の要件を指定する、ことと、

データ品質ルールを生成することであって、前記データ品質ルールは、実行されると、

前記フィールドの値が前記属性値によって表される前記特定の要件を満たすかどうかを検証し、

前記特定の要件が満たされているかどうかを示す出力データを生成するように構成されている、ことと、を含む動作を実行させるように構成されている、1つ以上の非一時的コンピュータ可読媒体。

【請求項36】

前記データ品質ルールは、

前記要件からの前記フィールド内の前記値の許容偏差、

前記フィールド内の1つ以上の許容値、及び

前記フィールド内の1つ以上の禁止値のうちの1つ以上を示す、請求項35に記載の1つ以上の非一時的コンピュータ可読媒体。

【請求項37】

前記フィールドは第1のフィールドであり、前記データ記録は第2のフィールドを含み、前記データ品質ルールは第1のデータ品質ルールであり、前記動作は、

前記特定のラベルの前記属性値に基づいて、前記第1のフィールドと前記第2のフィールドとの間に関係が存在すると決定することと、

前記第2のフィールドと前記第1のフィールドとの間の前記関係に基づいて、前記第2のフィールドに対する第2のデータ品質ルールを生成することと、を含む、請求項35に記載の1つ以上の非一時的コンピュータ可読媒体。

【請求項38】

前記動作は、

前記特定のラベルに関連付けられた前記属性値が、前記フィールドが前記特定のデータ記録の一次キー値を含むことを示すと決定することと、

前記一次キー値が前記フィールド内でそれぞれ固有であることを要求するようにデータ品質ルールを構成することと、を含む、請求項35に記載の1つ以上の非一時的コンピュータ可読媒体。

【請求項39】

前記特定のラベルの前記属性値は、前記フィールド内の前記値の平均、前記値の最大長、前記値の最小長、前記値のデータタイプ、及び前記値のフォーマットのうちの少なくとも1つを表す、請求項35に記載の1つ以上の非一時的コンピュータ可読媒体。

【請求項40】

前記特定のデータ記録は第1のデータ記録であり、前記動作は、前記第1のデータ記録とは異なる第2のデータ記録内の前記特定のラベルに関連付けられた別のフィールドに前記データ品質ルールを適用することを含む、請求項35に記載の1つ以上の非一時的コンピュータ可読媒体。

【発明の詳細な説明】

【技術分野】

【0001】

(優先権の主張)

本出願は、2020年8月28日に出願された米国特許出願第17／006,504号及び2020年2月26日に出願された米国特許出願第62／981,646号に対する

10

20

30

40

50

優先権を主張し、いずれの内容もその全体が本明細書に組み込まれる。

【0002】

(発明の分野)

本開示は、データフィールドの値の分類の結果に基づいて、データフィールドのデータを処理するためのルールを生成することに関する。より具体的には、本開示は、データフィールドのデータ及びメタデータのデータプロファイル进行分析することによってデータフィールドを分類し、データフィールドのデータを処理するための1つ以上のルールを生成することに関する。

【背景技術】

【0003】

コンピュータシステムは、データを送信、受信、及び／又は処理するために使用され得る。例えば、サーバコンピュータシステムは、リソース（例えば、ウェブページなどのウェブコンテンツ）を受信及び記憶し、コンテンツを1つ以上のクライアントコンピュータシステムで利用可能にするために使用され得る。クライアントコンピュータシステムからコンテンツの要求を受信すると、サーバコンピュータシステムは、要求されたコンテンツを取り出し、コンテンツをクライアントコンピュータシステムに送信して要求を満たし得る。

【0004】

データのセットは、データフィールドに記憶されたデータ値を含み得る。データフィールドは、データ値のうちの1つ以上を含み得る。データフィールドの多くのインスタンスが、データセット内に（例えば、異なるフィールド内に）存在し得る。

【発明の概要】

【0005】

本文書に記載のシステムは、処理されているデータの意味論的ラベルからデータを処理するための1つ以上のルールを生成するように構成されている。ラベルは、ラベル付けされているデータの意味論的意味を示す。データの意味論的意味は、データが実際の用語で表すものである。例えば、意味論的意味は、特定のフィールドが日付を表すこと、及び表されるデータがユーザプロファイルの誕生日であることを示し得る。データの意味論的意味は、データがデータ処理システムによってどのように処理されるかを示し得る。より具体的には、データ処理システムは、意味論的ラベルでラベル付けされたデータフィールドのデータ値を受信するように構成されている。データ処理システムは、データフィールドのデータ値を処理するための1つ以上のルールを、そのデータフィールドのラベルが何であるかに応じて生成するように構成されている。ルールとしては、データマスキングルール、データ品質ルール、試験データを生成するためのルール、スキーマを決定するためのルール、又は他のそのようなルールが挙げられ得る。一般に、ラベル付けされたデータに適用されるルールは、全体的なルールのクラス（例えば、データ品質ルール）が全てのデータに適用されている場合でも、データの意味論的意味に依存する。例えば、誕生日に対するデータ品質ルールは、有効期限日に対するデータ品質ルールとは異なり得るが、両方とも日付であり、同様のフォーマットを有する。データ処理システムは、データの意味論的意味を決定し、意味論的意味を示すラベルでデータをラベル付けし、ラベルに基づいてそのデータを処理するための1つ以上のルールを生成する。

【0006】

実施形態は、以下の特徴のうちの1つ以上を含むことができる。

【0007】

一態様では、プロセスは、フィールドに関連付けられたラベルに基づいてデータ記録のセット内のデータ記録のフィールド内の値に対するデータ品質ルールを決定するためのものであり、ラベルは、フィールドの値の特性を示す。プロセスは、データ処理システムによって実施される。プロセスは、ラベルを、データ記録内の1つ以上のフィールドのセットに関連付けるラベルインデックスを取り出すことを含み、ラベルは、1つ以上のフィールドのセットのそれぞれのフィールドで予想される情報のタイプを識別する。プロセスは

10

20

30

40

50

、ラベルによって示される情報のタイプを、ラベルに関連付けられた1つ以上のフィールドの値に対する要件を表す属性値のセットに関連付けるデータ辞書にアクセスすることを含み、要件は、1つ以上のフィールドの値の論理的又は構文的特性を含む。プロセスは、特定のデータ記録のフィールドについて、ラベルインデックスにアクセスすることによって、特定のデータ記録のフィールドに関連付けられた特定のラベルを識別することと、データ辞書から、特定のラベルの属性値を取り出すことであって、属性値は、フィールドに対する特定の要件を指定する、ことと、を含む。プロセスは、データ品質ルールを生成することであって、データ品質ルールは、実行されると、フィールドの値が属性値によって表される特定の要件を満たしているかどうかを検証し、特定の要件が満たされているかどうかを示す出力データを生成するように構成されている、ことを含む。

10

【0008】

いくつかの実装形態では、データ品質ルールは、要件からのフィールド内の値の許容偏差、フィールド内の1つ以上の許容値、及びフィールド内の1つ以上の禁止値のうちの1つ以上を示す。いくつかの実装形態では、フィールド内の1つ以上の許容値又は禁止値は、フィールドのフィールド名に関連付けられる。いくつかの実装形態では、フィールドは第1のフィールドであり、フィールド内の1つ以上の許容値又は禁止値は、特定のデータ記録内の第2のフィールド内の値に基づいて決定され、第2のフィールドは、特定のデータ記録の第1のフィールドに結び付けられている。いくつかの実装形態では、1つ以上の許容値又は禁止値は、第2のフィールド内の値と第1のフィールド内の値との組み合わせに基づく。いくつかの実装形態では、1つ以上の許容値は、数値関数を満たす値に対応し、1つ以上の禁止値は、数値関数を満たさない値に対応する。

20

【0009】

いくつかの実装形態では、フィールドは第1のフィールドであり、データ記録は第2のフィールドを含み、データ品質ルールは第1のデータ品質ルールであり、方法は、特定のラベルの属性値に基づいて、第1のフィールドと第2のフィールドとの間に関係が存在すると決定することと、第2のフィールドと第1のフィールドとの間に関係に基づいて第2のフィールドに対する第2のデータ品質ルールを生成することと、を含む。いくつかの実装形態では、関係は、第1のフィールドの値に対する第2のフィールドの値の依存性、又は第2のフィールドの値に対する第1のフィールドの値の依存性を示す。いくつかの実装形態では、関係は、第1のフィールドの値と第2のフィールドの値との間の相関関係を示す。いくつかの実装形態では、プロセスは、第1のフィールド及び第2のフィールドのそれぞれの値の関係を検証する検証データを取得することを含む。いくつかの実装形態では、関係は、第1のフィールド及び第2のフィールドの値の閾値数について検証される。いくつかの実装形態では、第2のデータ品質ルールは、第1のフィールドの値に基づいて、第2のフィールドの値に制約を強制するように構成されている。いくつかの実装形態では、第1のフィールドと第2のフィールドとの間に関係が存在すると決定することは、第2のフィールドの値が第1のフィールドの値によって参照されるキー値を含むと決定することを含み、データ品質ルールは、第2のフィールドのそれぞれの値が有効なキー値であることを要求するように構成されている。いくつかの実装形態では、第1のフィールド及び第2のフィールドはそれぞれ数値を含み、第1のフィールドと第2のフィールドとの間に関係が存在すると決定することは、第1のフィールドの値を第2のフィールドの値に結び付ける数値関数を決定することを含む。いくつかの実装形態では、第1のフィールドと第2のフィールドとの間に関係が存在すると決定することは、機械学習プロセスによって構成された少なくとも1つの分類子を使用することを含む。

30

40

【0010】

いくつかの実装形態では、プロセスは、特定のラベルに関連付けられた属性値が、フィールドが特定のデータ記録の一次キー値を含むことを示すと決定することを含む。いくつかの実装形態では、プロセスは、一次キー値がフィールド内でそれぞれ固有であることを要求するようにデータ品質ルールを構成することを含む。

【0011】

50

いくつかの実装形態では、特定のラベルの属性値は、フィールド内の値の平均、値の最大長、値の最小長、値のデータタイプ、及び値のフォーマットのうちの少なくとも1つを表す。いくつかの実装形態では、特定のデータ記録は第1のデータ記録であり、方法は、第1のデータ記録とは異なる第2のデータ記録内の特定のラベルに関連付けられた別のフィールドにデータ品質ルールを適用することを含む。いくつかの実装形態では、方法は、データ品質ルール及び特定のラベルに関連付けるデータを記憶することを含む。

【0012】

いくつかの実装形態では、プロセスは、データ品質ルールの承認を要求するためのデータを生成することと、データ品質ルールの承認を示す承認データを取得することに応答してデータ品質ルールを承認することと、を含む。

10

【0013】

いくつかの実装形態では、データ品質ルールを生成することは、フィールド内の値の履歴傾向を決定することと、フィールド内の値の履歴傾向に基づいて要件を生成することと、を含む。

【0014】

いくつかの実装形態では、フィールドに対するデータ品質ルールを生成することは、フィールド内の値の履歴平均を識別することと、フィールド内の値の履歴平均に基づいて要件を生成することと、を含む。

【0015】

実施形態は、以下の特徴のうちの1つ以上を含むことができる。

20

【0016】

データ処理アプリケーションの試験中に処理ルールの実行を引き起こすための試験データを選択するためのプロセスが、データ処理システムによって実行される。このプロセスは、ラベルを、データ記録内の1つ以上のフィールドのセットに関連付けるラベルインデックスを取り出すことと、ラベルは、1つ以上のフィールドのセットのそれぞれのフィールドで予想される情報のタイプを識別する。プロセスは、ラベルによって示される情報のタイプを、ラベルに関連付けられた1つ以上のフィールドの値に対する要件を表す属性値のセットに関連付けるデータ辞書にアクセスすることと、要件は、1つ以上のフィールドの値の論理的特性又は構文的特性を含む。プロセスは、特定のデータ記録のフィールドについて、ラベルインデックスにアクセスすることによって、特定のデータ記録のフィールドに関連付けられた特定のラベルを識別することと、データ辞書から、特定のラベルの属性値を取り出すことと、属性値は、フィールドに対する特定の要件を指定する、ことと、サブ設定ルールを生成することと、該サブ設定ルールは、実行されたときに、フィールドの値が属性値を含むかどうかを指定する、ことと、サブ設定ルールに従って特定のデータ記録からフィールドのサブセットを選択することと、それぞれの選択されたフィールドは、属性値を有する値を含む、ことと、データ処理アプリケーションを試験するために、選択されたサブセットをデータ処理アプリケーションに提供することと、を含む。

30

【0017】

いくつかの実装形態では、プロセスは、第1のデータ記録のフィールドの値で動作し、少なくとも1つの出力値を生成するように構成された処理ルールを含むデータ処理アプリケーションを使用して、第1のデータ記録を処理することと、フィールドは、ラベルに関連付けられている。プロセスは、第1のデータ記録の処理と関連して処理ルールが実行された回数を示す実行情報を取得することと、処理ルールが、第1のデータ記録の処理中にデータ処理アプリケーションによって実行されるかどうかは、ラベルに関連付けられたフィールドの値に直接的又は間接的に依存する。いくつかの実装形態では、サブ設定ルールは、第1のデータ記録の処理と関連して処理ルールが実行された回数を示す実行情報に基づいて決定され、サブ設定ルールは、第1のデータ記録のフィールドの識別を含む。

40

【0018】

50

いくつかの実装形態では、サブ設定ルールは、フィールドを、特定のデータ記録のデータ要素の少なくとも1つのキー値を含む特定のデータ記録のキーフィールドとして識別する。いくつかの実装形態では、特定のデータ記録のフィールドのサブセットは、所定の値と共にキー値を有するフィールドを含む。いくつかの実装形態では、サブ設定ルールは、ラベルに対応するキーフィールド名のリストを識別し、サブ設定ルールは、ラベルの属性値をキーフィールドのリストと比較することによって、フィールドをキーフィールドとして識別する。いくつかの実装形態では、特定のデータ記録のフィールドは、個人情報 (personally identifying information、PII) を含み、方法は、PII 情報を含むフィールドが存在しない特定のデータ記録のフィールドのサブセットを選択することを更に含む。いくつかの実装形態では、特定のデータ記録のフィールドは、個人情報 (PII) を含み、方法は、マスキング関数を PII に適用して、マスクされたデータを生成することと、マスクされたデータを含む、特定のデータ記録のフィールドのサブセットを選択することと、を更に含む。

10

【0019】

実施形態は、以下の特徴のうちの1つ以上を含むことができる。

【0020】

一般的な態様では、プロセスは、データ処理アプリケーションのデータをマスクするように構成されたデータ処理システムによって実行される。このプロセスは、ラベルを、データ記録内の1つ以上のフィールドのセットに関連付けるラベルインデックスを取り出すことを含み、ラベルは、1つ以上のフィールドのセットのそれぞれのフィールドで予想される情報のタイプを識別する。プロセスは、ラベルによって示される情報のタイプを、ラベルに関連付けられた1つ以上のフィールドの値に対する要件を表す属性値のセットに関連付けるデータ辞書にアクセスすることを含み、要件は、1つ以上のフィールドの値の論理的特性又は構文的特性を含む。プロセスは、特定のデータ記録のフィールドについて、ラベルインデックスにアクセスすることによって、特定のデータ記録のフィールドに関連付けられた特定のラベルを識別することと、データ辞書から、特定のラベルの属性値を取り出すことであって、属性値は、フィールドに対する特定の要件を指定する、ことと、属性値に基づいて、特定のデータ記録のフィールドが機密データを表すと決定することと、を含む。プロセスは、この決定に応答して、データマスキング関数を実行して、フィールドの機密データを含む値をマスクされた値に変換することを含む。

20

30

【0021】

いくつかの実装形態では、フィールドは第1のフィールドであり、ラベルは第1のラベルである。プロセスは、第1のラベルを有する第1のフィールドと第2のラベルを有する第2のフィールドとの間に関係が存在すると決定することを含む。プロセスは、関係が存在すると決定することに応答して、データマスキング関数を実行して、第2のフィールド内の値をマスクされた値に変換することを含む。いくつかの実装形態では、関係が存在すると決定することは、第1のラベル及び第2のラベルが共通のデータソースに関連付けられていると決定することを含む。いくつかの実装形態では、共通データソースは、ユーザプロファイルを含む。いくつかの実装形態では、関係は、第1のフィールドの値に対する第2のフィールドの値の依存性又は第2のフィールドの値に対する第1のフィールドの値の依存性を示す。いくつかの実装形態では、関係は、第1のフィールドの第1の値と第2のフィールドの第2の値との間の相関関係を示す。いくつかの実装形態では、関係は算術関数を含む。いくつかの実装形態では、プロセスは、関係のタイプに基づいて、データマスキング関数のタイプを選択することを含む。

40

【0022】

いくつかの実装形態では、データマスキング関数のタイプは、シャッフル関数、データ暗号化、文字スクランブル、及びデータ置換のうちの少なくとも1つを含む。いくつかの実装形態では、プロセスは、特定のデータ記録をスキャンして、少なくとも1つの他のフィールドの1つ以上の特定の値がマスクされた値に変換されるかどうかを決定することを含む。いくつかの実装形態では、プロセスは、ラベルの属性値に基づいて、データマスキ

50

ング関数のタイプを選択することを含む。いくつかの実装形態では、第1のデータマスキング関数は、数値を含む、特定のデータ記録の第1のフィールドに対して選択され、第2の異なるマスキング関数は、非数値を含む、特定のデータ記録の第2のフィールドに対して選択される。

【0023】

いくつかの実装形態では、プロセスは、フィールドに結び付けられた別のフィールドのラベルに基づいてデータマスキング関数のタイプを選択することを含む。いくつかの実装形態では、機密データは個人情報（PII）を含む。

【0024】

実施形態は、以下の特徴のうちの1つ以上を含むことができる。

【0025】

一般的な態様では、プロセスは、データ記録のスキーマを決定するように構成されたデータ処理システムによって実行される。このプロセスは、ラベルを、データ記録内の1つ以上のフィールドのセットに関連付けるラベルインデックスを取り出すことを含み、ラベルは、1つ以上のフィールドのセットのそれぞれのフィールドで予想される情報のタイプを識別する。プロセスは、ラベルによって示される情報のタイプを、ラベルに関連付けられた1つ以上のフィールドの値に対する要件を表す属性値のセットに関連付けるデータ辞書にアクセスすることを含み、要件は、1つ以上のフィールドの値の論理的特性又は構文的特性を含む。プロセスは、特定のデータ記録の第1のフィールドについて、ラベルインデックスにアクセスすることによって、特定のデータ記録の第1のフィールドに関連付けられた特定のラベルを識別することと、データ辞書から、特定のラベルの属性値を取り出すことであって、属性値は、第1のフィールドに対する特定の要件を指定する、ことと、属性値が、第1のフィールドに含まれる値のスキーマ特徴を表すと決定することと、スキーマ特徴に基づいて、ラベルを有する第1のフィールドが、第2のフィールドの値によって参照されるキー値を含むと決定することと、この決定に応答して、第1のフィールドが第2のフィールドの値によって参照されるキー値を含むことを参照するように、特定のデータ記録を記述するスキーマデータを更新することと、を含む。

【0026】

いくつかの実装形態では、プロセスは、スキーマデータに基づいてデータフローグラフを生成することを含む。いくつかの実装形態では、第2のフィールドは、特定のデータ記録とは異なる第2のデータ記録にある。

【0027】

いくつかの実装形態では、プロセスは、キー値に基づいて、特定のデータ記録を第2のデータ記録に結合するように構成された結合機能を生成することを含む。いくつかの実装形態では、第2のフィールドは、そのラベルとは異なる第2のラベルに関連付けられる。

【0028】

実施形態は、以下の特徴のうちの1つ以上を含むことができる。

【0029】

一般的な態様では、1つ以上のデータセット内の値に対するデータ品質ルールを決定するためのプロセスは、データ処理システムによって実装され、データ処理システムによって、1つ以上のデータセットのデータ記録に含まれるフィールドに記憶された値をプロファイルすることを含む。プロセスは、データ処理システムによって、1つ以上の分類子をプロファイルされた値に適用することを含む。プロセスは、1つ以上の分類子を適用することに基づいて、フィールドの値の論理的又は構文的特性を示す1つ以上の属性を識別することを含み、1つ以上の属性のそれぞれは、1つ以上の分類子のそれぞれの出力に基づく、対応の信頼レベルを有する。プロセスは、信頼レベルが閾値レベルを満たす識別された属性のうちの1つ以上をフィールドに関連付けることを含む。プロセスは、フィールドに関連付けられた1つ以上の属性に基づいて、フィールドに含まれる値に対する1つ以上の制約を決定することを含む。プロセスは、決定された1つ以上の制約に基づいて、データセットのフィールドに対するデータ品質ルールを決定することを含み、データ品質ル

10

20

30

40

50

ルは、フィールドの値と1つ以上の制約との間の関係を示す。

【0030】

これらのプロセスのいずれも、1つ以上の処理デバイスと、命令を記憶するメモリとを含むシステムとして実装され得、該命令は、1つ以上の処理デバイスによって実行されると、1つ以上の処理デバイスにプロセスの動作を実行させるように構成されている。いくつかの実装形態では、1つ以上の非一時的コンピュータ可読媒体は、命令を記憶するように構成され得、該命令は、1つ以上の処理デバイスによって実行されると、1つ以上の処理デバイスにプロセスの動作を実行させるように構成されている。

【0031】

態様は、1つ以上の利点を含み得る。例えば、本明細書に記載の技術は、データ処理システムが、データセットのデータフィールドのデータを処理するための1つ以上のルールを自動的に生成することを可能にする。データの意味論的意味が認識されると、データ処理システムは、アプリケーションの指定された目標を達成するためにどのデータ処理動作をデータ値に適用するかを決定する。このようにして、データ処理システムは、データ値のマスキング、データ品質ルールの強制、データセットのスキーマの識別、及び／又は別のアプリケーションを試験するための試験データの選択など、データセット全体の目標を達成するために、データセットの異なるデータフィールドを処理する方法を自動的に決定し得る。どのデータフィールドがどの意味論的意味を有するかを自動的に決定することは、ユーザがそれぞれのフィールドを手動でラベル付けすることが实际的でない又は不可能であり得る大きなデータセットにこれらのルールが適用されることを可能にする。例えば、数百、数千、又は数百万個のテーブルを有するデータセットの場合、データマスキングルールなどのルールを強制するために、それぞれのフィールドを手動でラベル付けすることは实际的でない場合がある。別の例では、いずれかのユーザが機密データを見る前に、機密データを自動的にマスクすることが有用であり得る。別の例では、データ処理システムは、機密データを含むことを意図しないデータフィールドに記憶された機密データをマスクし得る。

【0032】

データ処理システムはまた、データ処理ルールに従って、どのデータフィールドが処理される必要がないかを決定する。例えば、データ処理システムは、データフィールドがマスクされる必要がないと決定し得る。どのデータフィールドがデータ処理操作によって処理される必要がないかを決定することは、アプリケーションの指定された目標（例えば、マスキング）が満たされることを確実にするためのデータセット全体の処理時間を短縮する。

【0033】

データ処理システムは、データセットにルールを適用するための以下の課題のうちの少なくとも1つ以上を克服する。多くの場合、データセットのデータフィールドは、データフィールドの名前がデータフィールドのコンテンツの信頼性のある指標であるように、標準化された方法では命名されていない。アプリケーションの指定された目標を達成するためのデータフィールドの処理要件は、多くの場合、データフィールド内のデータ値の意味に依存するため、データ処理システムがデータセットのデータフィールドのデータ値の意味を知ることが多くの場合重要である。例えば、アプリケーションの目標が個人情報（PII）をマスクすることである場合、データ処理システムは、データセットのどのデータがPIIを含むかを決定してから、それらのデータにマスキング関数を適用する。いくつかの実装形態では、特定のデータフィールドは、（例えば、そのデータフィールドのデータ値の保障要件を満たすために）特定のマスキング関数の適用を必要とし得る。更に、データ処理システムは、データフィールドのデータ値の意味を決定するように構成されるが、これは、意味が、フィールド名、フィールドのデータ値のフォーマット、又はデータ値自体から必ずしも明らかであるとは限らないため、困難であり得る。

【0034】

1つ以上の実施形態の詳細を、添付の図面及び以下の説明に記載する。その他の特徴及

10

20

30

40

50

び利点は、明細書及び図面、並びに特許請求の範囲から明らかとなるであろう。

【図面の簡単な説明】

【0035】

【図1】データ処理システムのブロック図である。

【図2A】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図2B】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図2C】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図3A】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図3B】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図3C】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図4A】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図4B】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図5A】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図5B】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図6A】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図6B】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図6C】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図7A】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図7B】データ品質ルール処理のための図1の例示的なデータ処理モジュールをそれぞれ含むブロック図である。

【図8】データ処理システムのブロック図である。

【図9】例示的なプロセスをそれぞれ含むフロー図を示す。

【図10】例示的なプロセスをそれぞれ含むフロー図を示す。

【図11】例示的なプロセスをそれぞれ含むフロー図を示す。

【図12】例示的なプロセスをそれぞれ含むフロー図を示す。

【発明を実施するための形態】

【0036】

図1は、データ処理システム100のブロック図である。データ処理システム100は、入力データ112からデータ記録を処理するように構成されている。データ処理システムは、意味論的発見システム602及びデータ処理デバイス102を含む。意味論的発見システム602は、データ記録の1つ以上のフィールドの値の意味（例えば、意味論の意味）を決定するように構成されている。意味論的発見システム602は、データ辞書データベース614から選択された意味論的ラベル118でフィールドのそれぞれをラベル付けし得る。意味論的ラベル118（ラベルとも呼ばれる）は、データ辞書データベース614内の1つ以上の属性120に関連付けられている。意味論的ラベル118に関連付けられた属性は、ラベルの意味論の意味を定義する。

10

20

30

40

50

【0037】

データ処理デバイス102は、意味論的発見システム602からのラベルインデックス610及びデータ辞書データベース614、並びに入力データ112を受信するように構成されている。データ処理デバイス102は、ラベルインデックス614のラベル及びデータ辞書データベース614のラベル及び属性120に基づいて、入力データ112を処理するように構成されている。データベース614のラベル118及びラベルインデックス610は、意味論的発見システム602の意味論的発見プロセスによって生成される。

【0038】

意味論的発見システム602は、入力データ112を受信し、入力データのフィールドの意味論的意味を決定するように構成されている。フィールドの意味論的意味は、フィールドの値の実際的な（例えば、ビジネスの）意味の説明であり、これについては後で更に詳しく記載する。入力データ112のフィールドの意味論的意味を決定するために、意味論的発見システム602、意味論的発見システム602は、入力データ112をプロファイルし、プロファイルデータ及びフィールドの値を分析する複数の分類を実行する。プロファイルデータ及びフィールドの値の分類は、意味論的発見システム602が、どの属性がフィールドに最も関連付けられているかを決定することを可能にする。属性120は、所与のフィールドの特性を示すメタデータ（又は他のデータ）を含む。例えば、属性120は、特定のフォーマット、フィールドと別のフィールド（単数又は複数）との特定の関係、フィールドの許容値又は禁止値、関連するキー用語又はビジネス用語、関連するフィールドの値（複数可）の統計基準（個別のもの、フィールドの他の値に関するもの、又はグループとしてのフィールドの値に対するもののいずれか）などを示し得る。

【0039】

意味論的発見システム602は、入力データ112のデータ記録のフィールドをデータ辞書データベース614内のラベル118に関連付けるラベルインデックス610を生成するように構成されている。データ記録のフィールドがデータ処理デバイス102によって処理されているとき、フィールドに関連付けられた意味論的ラベル118（存在する場合）は、ラベルインデックス610を使用して取り出され得る。

【0040】

意味論的ラベル（例えば、ラベル118の）は、意味論的ラベルによってラベル付けされたフィールドのデータの意味論的意味を示す用語又は用語のセット（一般に、一連の単語又は英数字文字）を含む。意味論的ラベル118は、データ処理システムに関連付けられたデータ辞書に現れる、認識された用語又は用語のセットである。それぞれの意味論的ラベル118は、データセットの1つ以上のフィールドに関連付けられる。例えば、テーブルは、意味論的ラベルの意味をもつデータ値を有するデータセット内のそれぞれのフィールド（又はそれぞれのフィールドのそれぞれのインスタンス）に意味論的ラベル118を結び付け得る。

【0041】

いくつかの実装形態では、意味論的発見システム602は、必ずしも、フィールドの意味論的意味を識別するために実際のラベルを生成するのではなく、むしろ、意味論的意味を示す対応する属性120にフィールドを関連付ける。例えば、ラベルインデックス610は、フィールドの指定されたラベルを含めることなく、1つ以上の属性120を含むデータ辞書データベース614内のエントリにフィールドを関連付け得る。他の実装形態では、ラベルは、フィールドの意味論的意味を表すデータ辞書データベース614の属性120のうちの1つ以上のセットを示す略記として使用される。

【0042】

一般に、フィールドの意味論的意味は、フィールドの値の実際的な意味又は文脈上の意味に対応し得る。実際的な意味（例えば、ビジネスの意味）又は文脈上の意味は、データ記録内のデータ値の1つ以上の構文的特性及び／又はデータ値の文脈上の指標を表す。1つ以上の構文的特性は、一緒に解釈されるとき、様々なアプリケーションにデータ記録の値を使用しているデータ処理デバイス102の1つ以上のモジュールによってどのように

値が処理されるかを伝える。構文的特性は、データ辞書データベース 6 1 4 内のそれぞれのラベル 1 1 8 に関連付けられた属性 1 2 0 によって示される。例えば、日付に対応するデータ値について、意味論的意味は、日付値が誕生日（例えば、データ記録に関連付けられた顧客又はユーザに関する）を表すことを示し得る。別の例では、意味論的意味は、数値識別子が、特定の種類の数値識別子であり、社会保障番号特有の特定のプロパティを有する社会保障番号を表すことを示し得る。別の例では、意味論的意味は、2つの異なるフィールドの値と一緒にグループ化されたときにビジネスの意味を伝えることを示し得る。例えば、数字が郵便番号として識別され、その数字が都市名及び州名を表すものとして識別されている他のフィールドと対である場合、意味論的発見システム 6 0 2 は、3つの全フィールドをまとめた意味論的意味が住所であると決定するように構成されている。意味論的意味の更なる例については、後に詳しく記載する。

10

【0 0 4 3】

いくつかの実装形態では、フィールドの意味論的意味は、フィールド内のデータの実的な意味（例えば、ビジネスの意味又は現実世界の意味）を含む。意味論的意味は、フィールドがどの種類のデータを含むかを示すことよりも特異的であり得る。例えば、データプロファイルから日付を含むフィールドが識別され得る。データ処理システム 1 0 0 は、フィールド内の値の長さ（例えば、値の平均長さ）により、フィールドが日付を含むと決定し得る。データ処理システム 1 0 0 は、フォーマット（例えば、値が全て数字であるか、又は ## ¥ # # ¥ # # # のフォーマットに適合する、など）により、日付を含むと決定し得る。他の統計データは、フィールドが日付を含むことをデータ処理システム 1 0 0

20

【0 0 4 4】

前述のように、意味論的ラベル 1 1 8 は、データ辞書データベース 1 1 6 内の1つ以上の属性 1 2 0 に関連付けられる。属性 1 2 0 は、データフィールド内のデータ値の意味を示す。属性 1 2 0 は、データ値が特定のプロパティを有することを示し得る。例えば、属性 1 2 0 は、データ値が個人情報（P I I）を含むことを示し得る。属性 1 2 0 は、データフィールドのデータ値が他のデータフィールドと特定の関係を有することを示し得る。例えば、属性 1 2 0 は、データフィールドがデータベースのキー値を含むことを示し得る。属性 1 2 0 は、データ値のフォーマットによって、データ値自体によって、データフィールドの名前によって、及びデータフィールドのデータ値とデータセット内の他のデータフィールドの他の値との比較によって、暗示されるフィールド内のデータ値の特定の意味を示し得る。例えば、属性 1 2 0 は、データフィールドのデータ値が（何らかの他の目的のために単に日付（単数又は複数）ではなく）誕生日を含むことを示し得る。意味論的発見システム 6 0 2 の意味論的発見プロセスについては、図 8 に関して後により詳しく記載する。

30

【0 0 4 5】

いくつかの実装形態では、意味論的ラベル（又はラベルに関連付けられた属性 1 2 0）は、単一のフィールドではなく、フィールドのグループに関連付けられ得る。例えば、フィールドのグループ内のフィールドが通り名、番地、郵便番号、及び都市をそれぞれ表すと決定された場合、意味論的発見プロセスは、このフィールドのグループが全体として住所を表すと決定し、住所属性をそのフィールドのグループに関連付け得る。例えば、属性 1 2 0 は、これらのフィールドの全てが、所与のデータのセット内のグループに含まれるべきであること（例えば、住所の欠落部分についてのデータ品質チェックのため）を示し得る。いくつかの実装形態では、属性 1 2 0 は、これらのフィールドは全体として P I I を表すが、個々には P I I を表さないことを示し得る。これは、データマスキングチェックに有用であり得る。例えば、全てのフィールドがデータセットに含まれる場合、データマスキングモジュール 1 0 8 は、データマスキングが必要とされると決定し得、フィールドのうちの1つのみがデータセットに含まれる場合、データマスキングは必要とされない

40

50

。これはまた、属性が、グループ内のそれぞれのフィールドが他のフィールドと相関していることを示すスキーマ分析にも有用であり得る。他のそのような例がマルチフィールド分析に可能であり、これについては以下に詳しく記載する。

【0046】

図1に示すように、データ処理システム100が意味論的ラベル118でフィールドをラベル付けするとき、フィールドは、意味論的ラベルの属性120に関連付けられる。データ処理デバイス102のルールモジュール104、106、108、及び110は、属性120を使用して、どのようなルール（複数可）をデータセットの所与のフィールドに適用するか、又はどのようなルールをデータセットの所与のフィールドに生成するかを決定する。データ処理デバイス102は、それぞれのフィールドに異なるルールを適用することによって、それぞれのフィールドのデータを固有な方法で自動的に処理することができる。データ処理デバイス102は、データセットの所与のフィールドを処理するためのルールを生成することができる。

【0047】

データ処理デバイス102のモジュール104、106、108、及び110は、1つ以上の処理ルールに従って入力データを処理するように構成されている。一般に、データセットのデータ値を処理するためのルールは、データ値が特定の意味論的ラベルに関連付けられたフィールドにあるときにデータ値に対して実行される1つ以上のデータ処理動作を含み得る。ルールは、単一の動作又は2つ以上の動作のリストを含み得る。意味論的ラベル118は、アプリケーションの所与のデータフィールドのデータ値に（例えば、ルールのライブラリの）どのルール（複数可）が適用されるべきであるか、又は任意のルールが所与のフィールドのデータ値に適用されるべきであるかどうかを示し得る。例えば、ルールは、所与のフィールドのそれぞれのデータ値にデータ要件を課すデータ品質ルールを含み得る。データ要件は、フォーマット要件、値の範囲要件などであり得る。ルールは、下流アプリケーションがデータ値を処理し得る前に、データ値が異なるフォーマットに変換されることを必要とするデータマスキングルールを含み得る。後に記載されるように、多くの他のタイプのルールが可能である。

【0048】

データ処理デバイス102は、データ辞書614及びラベルインデックス610から、入力データ112を処理するためのデータを受信するように構成されている。データ処理デバイス102は、いくつかの異なるアプリケーションのうちの1つ以上のために意味論的ラベル118を使用するように構成されている。データ処理デバイス102は、データ品質ルールモジュール104（データ品質ルールモジュール104とも呼ばれる）を含む。データ品質ルールモジュール104は、フィールドのそれぞれに関連付けられた意味論的ラベル（複数可）に基づいて（例えば、入力データ112の）データ記録のフィールドのデータ品質ルールを生成するように構成されている。一例では、第1のデータ品質ルールのセットは、第1の意味論的ラベルを有する第1のフィールドに対して自動的に生成され得る（かつ、第1のフィールドを処理するために使用され得る）。例えば、第1のフィールドが社会保障番号としてラベル付けされている場合、データ品質ルールモジュール104は、そのフィールドの値が実際に有効な社会保障番号であるかどうかを試験するように構成された第1のデータ品質ルールのセットを生成するように構成され得る。別の例では、第2のフィールドは、出生日（data of birth）フィールドとしてラベル付けされ得る。データ品質ルールモジュール104は、その第2のフィールドの値が実際に有効な出生日（他の日付又は他の種類のデータではなく）であるかどうかを試験するように構成された、第2の、異なるデータ品質ルールのセットを生成するように構成され得る。データ品質ルールモジュール104については、図2A～図3Cに関連して更に詳しく記載する。

【0049】

データ処理デバイス102は、試験データセットモジュール106を含む。試験データセットモジュール106は、データ処理アプリケーションを試験するための試験データセットを生成するように構成されている。一般に、アプリケーションを試験するために、現

10

20

30

40

50

実的な試験データを使用して、アプリケーションの全ての特徴が意図されるように機能していることが確認される。試験される必要がある多くの処理順列、ケース構造などが存在し得るため、アプリケーションの全ての機能を試験する試験データを生成することは困難であり、時間がかかる可能性がある。更に、試験データは、参照整合性と内部論理一貫性に適合するという点で、実データと同様に機能するべきである。例えば、試験データ内のリンクは、同じく試験可能である有効なデータを指すべきである。試験データセットモジュール106は、試験データがこれらの要件を満たすことを確実にするように構成されている。例えば、試験データセットモジュール106は、試験データが参照整合性を満たしている（例えば、ポインタが有効なフィールドを指している）試験データ、適切なデータが含まれている（例えば、必要に応じて、社会保障番号フィールドが試験データに含まれている）試験データ、無関係なデータは含まれない（例えば、試験データに必要とされないフィールドは試験データから除外される）試験データなどを生成するように構成されている。試験データセットモジュール106については、図4A～図5Bに関して更に詳しく記載する。

10

【0050】

データマスキングモジュール108は、入力データ112内のデータフィールドをマスクするためのルールを生成するように構成されている。データマスキングは、個人情報（PII）などの機密データが、データ記録内の実際の人々又は本物のエントリを表さなくなるように、機密データがマスクされることを確実にするために使用され得る。例えば、データマスキングは、データ記録のフィールドの値を匿名化することを伴い得る。データマスキングは、特定のフィールドの値を暗号化することを含み得る。後に記載されるように、データをマスクするための多数の他の方法も可能である。データマスキングモジュール106は、特定のマスキング関数を使用してフィールドをマスクするように構成されている。フィールドの特定のマスキング関数は、フィールドの意味論的ラベルに基づいて選択され得る。例えば、フィールドに関連付けられた意味論的ラベルは、データがどのようにマスクされるか（例えば、どのマスキング関数又はマスキング技術がそのフィールドに使用されるべきか）を指定し得る。このようにして、意味論的ラベルは、フィールドがマスクされるべきであることのみではなく、フィールドがどのようにマスクされるべきかを指定し得る。

20

【0051】

データセットスキーマ分析モジュール110は、入力データ112のフィールド間の関係を識別し、これらの関係から入力データ112のスキーマを決定するように構成されている。スキーマは、入力データ112の構造を示す。例えば、入力データ112のスキーマは、フィールド間のキー値関係、他のフィールドに対するフィールドの依存性、フィールド間の値の相関などを示し得る。データ処理デバイス102の他のモジュール104、106、108のうちの1つ以上は、データ品質ルールの生成又は適用、データマスキングポリシーの生成又は適用、及び試験データセットの生成を含む、識別されたスキーマに基づいて入力データ112を処理する。いくつかの実装形態では、入力データ112のスキーマは、意味論的発見システム602のプロセスと並行して決定される。いくつかの実装形態では、意味論的発見システム602によって生成された意味論的ラベルは、データ

30

40

【0052】

一般に、データ処理デバイス102は、入力データストア112からデータセットを受信し、処理されたデータを出力データストア114に出力する。いくつかの実装形態では、入力データストア112及び出力データストア114は、同じデータストアに含まれる。いくつかの実装形態では、入力データストア112及び出力データストア114は、データベースである。データ処理デバイス102によって受信されたデータセットは、データ記録、データテーブル、又は1つ以上の値のグループが名前又はラベルに関連付けられている他のデータを含み得る。例えば、データセットは、それぞれが1つ以上の値を含む

50

1つ以上のフィールドを含み得る。それぞれのフィールドは、フィールド名を含み得る。フィールド名は、技術名、コード名、ビジネス用語、数値、又は任意の他の値であり得る。いくつかの実装形態では、フィールドは、フィールド名を有さない。一般に、入力データ112のフィールドのうちの1つ以上は、対応の意味論的ラベルに関連付けられる。それぞれの意味論的ラベルは、前述のように、フィールドの構文的又は意味論的意味が何であるかを示す。

【0053】

データ処理システム100は、参照データストア116を含む。参照データストア116は、モジュール104、106、108、及び110を動作させるデータ処理システム100によって使用されるデータを記憶する。例えば、データ品質ルールライブラリからのデータ品質ルールがデータ品質ルールモジュール104によって必要とされる場合、データ処理デバイス102は、以下に記載される基準に基づいて、参照データストア116からデータ品質ルールを取り出す。参照データストア116は、データマスキングプロセス中に使用するためのマスキング関数などの情報を記憶する。参照データストア116は、ラベルインデックス610及びデータ辞書データベース614のデータを記憶し得る。参照データストア116としては、インメモリデータストア又は永続的データストアが挙げられ得る。参照データストア116は、単一のデータストア又はデータストアの分散ネットワークであり得る。本質的に、参照データストア116は、データ処理システム100の様々なプロセスで使用するためのデータを保持するための場所である。

【0054】

図2A～図7Bは、図1のデータ処理デバイス102のデータ処理モジュールをそれぞれ記載したブロック図である。図2A～図3Cを参照すると、データ品質ルールモジュール104は、データセットのフィールドの意味論的意味を使用して、データセットに対するデータ品質ルールを生成するか、又はデータセット（入力データ112など）にデータ品質ルールを適用するように構成されている。フィールドの意味論的意味は、図1に関連して前述したように、及び図8に関連して後に詳しく記載されるように、意味論的発見システム602によって決定される。

【0055】

図2Aは、データ品質ルールエンジン104の属性分析モジュール204によって実行されるプロセス200aを示す。属性分析モジュール204は、データ品質ルールモジュール104によって処理される値を有するフィールドに、どのような属性が関連付けられているかを決定するように構成され、その結果、そのフィールドに対して1つ以上のデータ品質ルールがモジュール104によって生成され得る。属性分析モジュール204は、ラベルインデックス610によってデータ辞書614の属性に関連付けられたフィールドを含むデータセット201を受信する。前述のように、意味論的ラベル228は、属性及びフィールドに関連付け得るが、実際のラベルは必要とされない。簡単にするために、後続の例は、属性によって示される意味論的意味をフィールドと関連付けるための意味論的ラベルを含むが、後続の例のそれぞれにおいて、実際のラベル（例えば、記述的な単語又は用語）は必要とされない。

【0056】

属性分析モジュール204は、データ品質ルールが生成される所与のフィールド、又はデータ品質ルールが適用される所与のフィールドの属性が何であるかを決定する。属性分析モジュール204は、データセット201を受信する（250）ように構成されている。属性分析モジュールは、データセット201のフィールドを選択する（252）。図2Aの例では、選択されたフィールド218は、「SSN」と命名されているが、フィールド名は必要とされない。フィールド「SSN」の値220は、「55-555-5555」である。

【0057】

フィールドが選択されると、属性分析モジュールは、ラベルインデックス610を使用して、ラベルインデックス610内の意味論的ラベル228をルックアップする（254

）。いくつかの実装形態では、ラベルは受信されないが、データ辞書データベース 6 1 4 内のエントリ識別子（又はラベルを使用せずに、どの属性が選択されたフィールドに結び付けられているかを示す何らかの他の指標）が受信される。属性分析モジュール 2 0 4 は、ルックアップを実行して（2 5 6）、どの属性がラベル 2 2 8 a「社会保障番号」に結び付けられているかを決定する。ここで、属性 2 3 0 は、データ辞書データベース 6 1 4 内のラベル 2 2 8 a に結び付けられている。次いで、属性 2 3 0 は、データ品質ルール生成モジュール 1 0 4 のルール生成モジュール 2 0 6 又はルール適用モジュール 2 0 8 によるデータ品質ルールの生成又は適用のために（例えば、メモリ内に）記憶される。

【0 0 5 8】

図 2 B を参照すると、プロセス 2 0 0 b では、ルール生成モジュール 2 0 6 は、（例えば、属性 2 3 0 によって示される）フィールドの意味論的意味に基づいて、フィールドに対する 1 つ以上のデータ品質ルールを決定するように構成されている。例えば、特定の属性は、選択されたフィールドに適用され得る 1 つ以上のルールを示し得る。この文脈において、属性 2 3 0 は、許容される値のセット又は範囲、フィールドの値と 1 つ以上の他のフィールドの値との関係（例えば、常により大きい、常により小さい、等しい、～の関数など）、データタイプ、予想される値の長さ、又はフィールドの値の他の特性を示し得る。例えば、後に記載されるように、属性は、フィールド内でどのような値が許可されるか、及びフィールド内でどのような値が許可されないかを示し得る。例えば、フィールドのラベルに関連付けられた属性に示される所与の範囲はまた、そのフィールドに対するデータ品質ルールであり得る。しかしながら、フィールドに関連付けられた属性及びフィールドに対して生成されるルールは、必ずしも一致しない。例えば、フィールドに対するデータ品質ルールは、複製値がないか、又は空の値がないことを要求し得るが、属性 2 3 0 はこのルールを示す必要はない。逆に、属性は、選択されたフィールドが特定の要件に適合することを示し得るが、生成されるデータ品質ルールは、この要件を強制する必要はない。

【0 0 5 9】

このルール生成モジュール 2 0 6 は、フィールドに関連付けられた属性 2 3 0 を受信する（2 6 0）ように構成されている。ルール生成は、前述のように、属性に基づいて選択されたフィールドに対するデータ品質ルールを生成する（2 6 2）ように構成されている。属性 2 3 0 の任意の組み合わせを使用して、任意の数のデータ品質ルールが生成され得る。ルールは、データセット 2 0 1 の任意の他のフィールド又は値とは独立して、フィールドのそれぞれの値に対する要件を強制し得る。例えば、データ品質ルール 2 1 4 a は、ラベル「社会保障番号」に関連付けられたフィールドの値のフォーマットが、フォーマット「X X X - X X - X X X X」を有することを要求し得、ここで、X は、0 ～ 9 の任意の数字であり得る。この要件は、それぞれの個々の値に基づくものであり、強制するために更なる分析を必要としない。別の例では、ルール 2 1 4 b は、値のそれぞれの文字のタイプをチェックし、全ての X 値が他の文字ではなく、実際に 0 ～ 9 の数字であることを確実にし得る。別の例では、データ品質ルールは、選択されたフィールド全体に対するチェックを要求し得る。例えば、データ品質ルール 2 1 4 c は、選択されたフィールド内のそれぞれの値が固有であることを要求する。これには、現在処理されている値に加えて、選択されたフィールド内の値をチェックして、それぞれの値が固有であるかどうかを決定することが必要である。

【0 0 6 0】

データ品質ルールの他の例が可能である。いくつかの例では、データのセットの品質を特徴付けるために使用されるルールは、データセット内のデータ記録のプロファイルの許容属性又は禁止属性を示し得る。プロファイルの属性は、値又は値の範囲であり得る。プロファイルの許容属性を示すルールは、プロファイルが許容属性を含むときに満たされる。フィールドの許容属性の例は、そのフィールドの許容可能な最大値及び最小値であり得、フィールドの平均値が許容可能な最大値と最小値との間に入る場合、ルールは満たされる。プロファイルの禁止属性を示すルールは、プロファイルが禁止属性を含まない限り満たされる。フィールドの禁止属性の例は、そのフィールドに対して禁止された値のリスト

10

20

30

40

50

であり得、フィールドが禁止値のいずれかを含む場合、ルールは満たされない。

【0061】

データ品質ルールは、フィールドの値（単数又は複数）とフィールドのプロファイルとの間の許容偏差を示し得る。対応するルールによって示される許容偏差よりも大きい、プロファイルとフィールドの値との間の偏差は、データセットにおけるデータ品質問題の指標であり得、したがって、データセットが、データの下流セットにおける既存のデータ品質問題又は潜在的なデータ品質問題の考えられる根本原因であるという指標であり得る。いくつかの例では、許容偏差は、最大許容値及び最小許容値などの値の範囲として指定され得る。いくつかの例では、許容偏差は、平均値（例えば、過去のデータセットにおける値の平均又は中央値）であり得る単一の値から標準偏差として指定され得る。

10

【0062】

いくつかの例では、データのセットの品質を特徴付けるために使用されるルールは、フィールド内の値の有効性に基づくなど、データ記録の1つ以上のフィールドのそれぞれにおける値の許容された特性又は禁止された特性を示し得る。フィールドの許容された特性を示すルールは、フィールド内の値が許容された特性に見合う場合に満たされる。フィールドの禁止された特性を示すルールは、フィールド内の値が禁止された特性に見合わない限り満たされる。ルールを満たす値は、有効な値と称されることがあり、ルールを満たさない値は、無効な値と称されることがある。フィールド内の値の様々な特性は、ルールによって許容された特性又は禁止された特性として示され得る。例示的なルールは、許容された値の範囲又は禁止された値の範囲、許容可能な最大値、許容可能な最小値、又は許容された若しくは禁止された1つ以上の特定の値のリストなど、フィールドのコンテンツの許容された特性又は禁止された特性を示し得る。例えば、1900未満又は2016超の値を有する出生年フィールドは、無効と見なされ得る。例示的なルールは、フィールドのデータタイプの許容された特性又は禁止された特性を示し得る。例示的なルールは、ある特定のフィールドに値がないこと（又はNULLがあること）が許容されているか又は禁止されているかどうかを示し得る。例えば、文字列値（例えば、「Smith」）を含むラストネームフィールドは、有効と見なされ得るが、空白であるか又は数値を含むラストネームフィールドは、無効と見なされ得る。例示的なルールは、同じデータ記録内の2つ以上のフィールド間に許容された関係又は禁止された関係を示し得る。例えば、ルールは、州フィールドのそれぞれの可能な値に対応する郵便番号フィールドの値のリストを指定してもよく、リストによってサポートされない郵便番号フィールド及び州フィールドの値のいずれの組み合わせも無効であることを指定してもよい。

20

30

【0063】

いくつかの例では、ルールは、履歴データの分析に基づいて生成され得る。ルール生成モジュール206によって生成されたルールは、データのセット内のデータ記録のプロファイルの許容属性又は禁止属性を示し得る。例えば、プロファイルに対するルールは、特定のデータセットのフィールドのプロファイルとそのデータセットのフィールドの決定された履歴プロファイルとの間の許容偏差を示し得る。データセットの履歴プロファイルは、履歴データに基づき得、例えば、履歴プロファイルは、前日からの同じデータセットのプロファイル、数日前からの（例えば、過去の週又は月にわたる）同じデータセットの平均プロファイル、同じデータセットの寿命平均プロファイルであり得る。より一般的には、プロファイルは、様々な種類の統計分析を利用するために、多種多様な参照情報を保持し得る。例えば、プロファイルは、標準偏差又は値の分布の他の指標に関する情報を含み得る。以下の例の目的で、本出願の一般性を限定することなく、プロファイルは、従来のデータセットの数値平均、及び場合によっては標準偏差も含み得る。

40

【0064】

生成されたルールは、データ記録のフィールド内の値の決定されている許容された特性又は禁止された特性を示し得る。一例では、フィールドに対して生成されたルールは、フィールドの過去の最大値又は最小値の分析に基づいて、フィールドの許容可能な最大値又は最小値を示し得る。一例では、フィールドに対して生成されたルールは、フィールドに

50

対して以前に発生した値の分析に基づいて、フィールドの許容値のリストを示し得る。

【0065】

いくつかの例では、機械学習技術が、データ品質ルールを生成するために用いられる。例えば、データは、ユーザ設定又はアプリケーションのために、どの特性がそのアプリケーションのデータ品質に影響を与え、ゆえにその特性を強制するためのデータ品質ルールを必要とするかについて学習期間にわたって分析され得る。例えば、アプリケーションが、フィールドの値が誤ったフォーマットであるために失敗することが多い場合、ルール生成モジュールは、特定のフォーマットを強制するためのルールを生成する。数値が許容可能範囲外であるために失敗する場合、ルール生成モジュール206は、フィールドの特定の値範囲を強制するためのデータ品質ルールを生成する。他のそのような例が可能である。学習期間は、特定の期間であり得るか、又は平均値若しくは予想される値が安定した値に収束するまでの時間量であり得る。

10

【0066】

ルール生成モジュール206は、システムによって処理されるデータ記録の予想される特徴に基づいて、1つ以上のデータ品質ルールを生成するように構成されている。特定の例では、ソースデータは、米国で発生する取引に関するクレジットカード取引記録である。ソースデータは、1時間の増分で処理されるストリーミングデータである。フィールドに対して識別された属性230、及びクレジットカード取引記録を処理するときに実行される動作を示すアプリケーションからのデータに基づいて、ユーザは、取引識別子フィールド、カード識別子フィールド、状態フィールド、日付フィールド、及び総額フィールドを、プロファイルされる重要なデータ要素として識別し得る。

20

【0067】

ソースデータがクレジットカード取引記録である特定の例では、ルール生成モジュール206は、状態フィールドの50個の許容値のみが存在することを示す属性を受信する。ルール生成モジュール206は、ソースデータのセットのプロファイルが、基準に対するソースデータのセットのプロファイルの標準偏差に関係なく、状態フィールド内の50個を超える値を識別する場合に、警告フラグが使用されるようにするルールを作成し得る。ルール生成モジュール206は、処理がソースデータのセットに存在すべきである日と同日に完了した取引に関するクレジットカード取引記録のみを示す属性を受信し得る。ルール生成モジュール206は、任意のソースデータ記録が、処理の日付と一致しない日付を有する場合、警告メッセージが送信されるようにするルールを作成し得る。

30

【0068】

いくつかの例では、ルール生成モジュール206は、ユーザインターフェースを通して生成される1つ以上のルールを指定し得る。ユーザインターフェースを通じて、ユーザは、1つ以上のフィールドに対するルールのフィードバックを提供することができ、又はフィールドについて事前に入力されたデフォルトルールを承認することができる。ユーザインターフェースの更なる説明は、2012年10月17日に出願された米国出願第13/653,995号に見つけることができ、その内容は、参照によりその全体が本明細書に組み込まれている。ユーザインターフェースの他の実装形態も可能である。

【0069】

40

いくつかの例では、考えられるデータ品質問題が、参照データのセットの新規バージョン内又はソースデータのセット内などのデータセット内で検出される場合、考えられるデータ品質問題を有するデータセットの識別子が、データベースに記憶された根本原因のデータセットのリストに配置される。出力データのセットに関するデータ品質問題が後に検出されると、その出力データのセットの上流データ系統要素を識別し、存在する場合は、それらの上流データ系統要素のうちのどれが根本原因データセットのリストに含まれるかを決定するために、データベースが照会され得る。

【0070】

いくつかの例では、考えられるデータ品質問題が、参照データのセットの新規バージョン又はソースデータのセットなどのデータセット内で検出される場合、ユーザ通知が有効

50

にされ得る。いくつかの例では、警告フラグは、データ品質問題を示すために記憶され得る。例えば、考えられるデータ品質問題が、参照データのセットの新規バージョン内で検出された場合、警告フラグは、参照データの新規バージョンのプロファイルデータと併せて記憶され得る。考えられるデータ品質問題が、ソースデータのセット内で検出された場合、警告フラグは、そのソースデータのセットのプロファイルデータと併せて記憶され得る。いくつかの例では、警告メッセージは、考えられるデータ品質問題の存在を示すためにユーザに伝達され得る。警告メッセージは、例えば、メッセージ、アイコン、若しくはユーザインターフェース上のポップアップウィンドウとして、電子メール若しくはショートメッセージサービス（SMS）メッセージとして、又は別の形態であり得る。

【0071】

いくつかの例では、ルールは、警告フラグ又は警告メッセージが使用されるプロファイルからの1つ以上の閾値偏差を指定し得る。例えば、現在のデータセットのプロファイルとそのデータセットのプロファイルとの間の偏差が、1及び2の標準偏差などのように小さい場合、警告フラグは記憶され得、偏差が2を超える場合、警告メッセージは伝達され得る。閾値偏差は、ソースデータ及び参照データのそれぞれのセットに特有であり得る。

【0072】

いくつかの例では、偏差が重度、例えば、プロファイルから3を超える標準偏差である場合などは、ユーザが介入するまで、データ処理システムによる更なる処理は停止され得る。例えば、重度の偏差を有するソースデータ又は参照データによって影響を受ける任意の更なる処理は停止される。停止される変換は、影響を受けるソースデータ又は参照データの下流にあるデータ系統要素を参照するデータによって識別され得る。

【0073】

いくつかの例では、プロファイルデータは自動的に決定される。例えば、所与のデータセットのプロファイルデータは、例えば、そのデータセットの新しいプロファイルデータが決定される場合は常にプロファイルデータを再計算することによって、そのデータセットの過去のプロファイルデータの実行中の履歴平均として自動的に更新され得る。いくつかの例では、ユーザは、例えば、所望の特性を有するデータセットをプロファイルすることによって、初期プロファイルデータを供給し得る。

【0074】

引き続き図2Bを参照すると、データ品質ルール214a～cは、1回生成され、後の再利用のために特定の意味論的ラベルに関連付けられ得る。属性230によって示される制約の全てが、明示的なデータ品質ルール214として強制される必要があるわけではないが、これは必要に応じて実行され得る。むしろ、所与のアプリケーションに必要なとされる場合に、制約を強制するデータ品質ルール214のセットが生成され得る。例えば、意味論的ラベルの属性は、意味論的ラベルを有するフィールドの全てのデータが長さ=9であることを示し得る。属性に関連付けられたデータ品質ルールは、フィールドのデータ値の長さがそれぞれ正確に9であることを必要とし得る。長さ要件は、プロファイル216で指定され得る。関連付けられたデータ品質ルールは、この情報から生成され得る。例えば、データ品質ルール214は、長さ要件1を指定するデータ品質ルールを含み得る。ルール生成モジュールは、この利用可能なルールから、プロファイル216を参照した後、正確に9桁の長さを要求するデータ品質ルールを生成し得る。

【0075】

ルール生成モジュール206は、フィールドに対して少なくとも1つのデータ品質ルールを生成する。生成されるデータ品質ルールは、フィールドに固有であるか、又は2つ以上のフィールドに共通であり得る。データ品質ルールは、データセットのフィールド内の値の特徴と、フィールドのプロファイルに含まれる1つ以上の特徴との間の許容偏差、フィールド内の1つ以上の許容値、及びフィールド内の1つ以上の禁止値のうちの1つ以上を示し得る。

【0076】

前述のように、データ品質ルールモジュール104は、意味論的ラベル228に関連付

10

20

30

40

50

けられた属性（複数可）を使用してデータ品質ルールを自動的に生成するように構成されている。これを行うために、ルール生成モジュール 206 は、属性のうちの 1 つ以上を論理ルールに変換する。例えば、属性が、フィールドの全ての値が 1 ～ 100 の間であることを示す場合、全ての値が 1 ～ 100 であることを要求するデータ品質ルールが生成され得る。しかしながら、そのようなデータ品質ルールは必ずしも生成される必要はない。例えば、意味論的ラベル 228 に関連付けられた別の属性は、フィールドのそれぞれの値の最終桁が 7 で終わることを示し得る。したがって、その値が 1 ～ 100 であり、かつ 7 で終わることの両方を要求するデータ品質ルールが生成され得る。これらは、2 つの異なるルールとしても表現され得る。別の例では、属性は、値が、社会保障番号のフォーマットである ###-##-#### のようなフォーマットを有することなど、値の特定のフォーマットを示し得る。データ品質ルール生成器は、この正確なフォーマット、その一部、又は一般的な変形（例えば、ダッシュなしで 9 桁）に適合する値を要求するデータ品質ルールを生成し得る。

【0077】

意味論的ラベル 228 は、フィールドの値と別のフィールドの値との関係を示し得る。ルール生成モジュール 206 は、これらの関係に基づいてデータ品質ルールを生成し得る。例えば、フィールドが第 1 のフィールドであり、データセットが第 2 のフィールドを含む場合、ルール生成モジュールは、第 2 のフィールドと第 1 のフィールドとの間の関係に基づいて、第 1 のフィールドに対するデータ品質ルールを生成するように構成され得る。例えば、第 2 のフィールドは、フィールドにユーザ名が入っていることを示す意味論的ラベルに関連付けられ得る。データ品質ルールは、第 1 のフィールドがデータ品質試験を通過するように、第 2 のフィールド内の有効な名前が第 1 のフィールド内のそれぞれの値に関連付けられることを要求し得る。ルール 214 a ～ c の組み合わせが使用され得る。例えば、フィールド SSN について、フォーマット = ###-##-####、かつ SSN フィールド 220 のそれぞれの社会保障番号 218 の値が固有である場合、データ品質ルールの出力 = 合格 (PASS) となることを指定するルールが生成され得る。これは、制約 214 a 及び 214 c を含む、3 つの提案されたルール（制約とも呼ばれ得る）のうちの 2 つを組み合わせる。

【0078】

いくつかの実装形態では、ルールは、フィールドのデータ値がどのように経時変化するかに基づき得る。例えば、フィールドのプロファイルは、フィールド内の値の履歴傾向に基づいて選択され得る。別の例では、フィールドのプロファイルは、フィールド内の値の履歴平均を識別することに基づいて決定される。

【0079】

ルール生成モジュール 206 は、データ品質ルールの承認を要求するためのデータを生成し得る。例えば、ユーザは、ユーザインターフェース内のルールを承認又は拒否することができる。ルール生成モジュール 206 は、データ品質ルールの承認を示す承認データを取得することに応答して、アプリケーションのデータ品質ルールを承認し得る。いくつかの実装形態では、ルールを適用するために承認データは必要とされない。ルール生成モジュール 206 がルール（例えば、ルール 214 a ～ c）を生成すると、ルール 214 は、データ品質ルールデータストア 250 などのデータストアに記憶される（264）。

【0080】

図 2C を参照すると、プロセス 200 c において、ルール適用モジュール 208 は、意味論的ラベル 228 を使用してフィールドに関連付けられた属性 230 などによって、受信されたデータに関連付けられた意味論的意味に基づいて、受信された任意のデータ 112 に生成されたデータ品質ルール 214 を適用するように構成されている。

【0081】

ルール適用モジュール 208 は、データセット 201 を受信する（270）ように構成されている。ルール適用は、テーブル 1、列 1 のデータセットの場所 221 を有する「SSN」と命名されたフィールド 218 など、データセット 201 のフィールドを選択する

10

20

30

40

50

(272)。ルール適用モジュール208は、ラベルインデックス610を参照し(274)、どのルールがデータセット201内のテーブル1、列1のフィールドに関連付けられているかを見つける。ここで、意味論的ラベル228「社会保障番号」は、フィールドを表すために使用される。ルール適用モジュール208は、データ品質ルールデータストア250にアクセスし(276)、フィールドラベル「社会保障番号」に対するデータ品質ルール214を取り出す。ルール適用モジュール208は、データ品質ルールをフィールド218の値に適用する(278)。一例では、値555-55-5555は、フォーマットルール、文字制限ルール、及び固有の値ルールを通過する。ルール214a~cの全てを通過した結果は、合格(PASS)である。適用モジュール208は、データ品質ルールの適用の結果224を出力データストア114に記憶する(280)。

10

【0082】

ラベル228に関連付けられたフィールドに対して生成されたデータ品質ルールは、データセット201の第2のインスタンス内又は新しい異なるデータセット内のいずれかの同じラベルに関連付けられた任意のフィールドに適用され得る。例えば、ラベル228の社会保障番号は、ルール生成モジュール206が1つ以上のデータ品質ルールを生成したフィールドをラベル付けする。同じラベル228を有する任意のフィールドは同じ意味論的意味を有するため、生成されたデータ品質ルールは、それらのフィールドに適用され得る。これはまた、データセットにわたってラベル228でラベル付けされた異なるフィールドにも当てはまり得る。例えば、新しいデータセット内の新しいフィールドがラベル228の社会保障番号を有する場合、データ記録202のために生成されたデータ品質ルールは、新しいデータセット内のその新しいフィールドに依然として適用可能である。

20

【0083】

上記の実施例は、単一のフィールド分析から生成されたルールを示す。マルチテーブル(例えば、マルチフィールド)分析に基づいてルールを生成することも可能である。図3Aを参照すると、一態様では、データ品質ルールは、データ記録202及び203などの複数のフィールド又はデータ記録間の関係に基づいて生成され得る。フィールド間の関係は、第1のフィールドの値に対する第2のフィールドの値の依存性又は第2のフィールドの値に対する第1のフィールドの値の依存性を示し得る。一態様では、この関係は、第1のフィールドの値と第2のフィールドの値との間の相関関係を示し得る。いくつかの例では、データ品質ルールモジュール104は、関係を確認する検証データを取得することによって、2つ又は3つ以上のフィールド間の関係を検証する。検証データは、ユーザ入力であり得るか、又は関連するフィールドの意味論的ラベルの属性に基づき得る。例えば、ルール生成中に、生成されたルールを試験するために、第1のフィールド及び第2のフィールドの値の閾値数について関係が検証され得る。例えば、第1のフィールド及び第2のフィールドがそれぞれ数値を含む場合、第1のフィールドと第2のフィールドとの間に関係が存在すると決定することは、第1のフィールドの値を第2のフィールドの値に結び付ける数値関数を決定することを含み得る。いくつかの実装形態では、生成されるデータ品質ルールは、1つ以上の許容値が、数値関数を満たす値に対応することを要求し得、1つ以上の禁止値は、数値関数を満たさない値に対応する。

30

【0084】

図2Aに関して記載されたプロセス200aと同様のプロセス300aでは、属性分析モジュール204は、データ品質ルールが生成される所与のフィールド、又はデータ品質ルールが適用される所与のフィールドのいずれかに対する属性が何であるかを決定する。属性分析モジュール204は、データセット201を受信する(282)ように構成されている。属性分析モジュールは、データセット201のフィールドを選択する(284)。図3Aの例では、選択されたフィールド218は、「SSN」と命名されているが、フィールド名は必要とされない。フィールド「SSN」の値220は、「55-5555-5555」である。

40

【0085】

フィールドが選択されると、属性分析モジュールは、ラベルインデックス610を使用

50

して、ラベルインデックス610内の意味論的ラベル228をルックアップする(286)。いくつかの実装形態では、ラベルは受信されないが、データ辞書データベース614内のエントリ識別子(又はラベルを使用せずに、どの属性が選択されたフィールドに結び付けられているかを示す何らかの他の指標)が受信される。属性分析モジュール204は、ルックアップを実行して(288)、どの属性がラベル228a「社会保障番号」に結び付けられているかを決定する。ここで、属性230は、データ辞書データベース614内のラベル228aに結び付けられている。次いで、属性231は、データ品質ルール生成モジュール104のルール生成モジュール206又はルール適用モジュール208によるデータ品質ルールの生成又は適用のために(例えば、メモリ内に)記憶される。この例では、属性231は、テーブル1の「SSN」と命名されたフィールド218(例えば、記録202)が、テーブル2の「ユーザID」と命名されたフィールド(記録203内)を参照することを示す。データ品質ルール生成モジュール206は、この情報(図7A～図7Bに関連して記載されるスキーマ分析中に発見され得る)を使用して、記録202、203のフィールド間の関係に基づいてデータ品質ルールを生成し得る。

【0086】

図3Bは、図2Bのプロセス200bと同様に、モジュール206によってデータ品質ルールを生成するためのプロセス300bを示す。データ品質ルールは、第1のフィールド(例えば、記録203の)の値に基づいて、第2のフィールドの値(例えば、記録201の)に制約を強制するように、ルール生成モジュール206によって構成され得る。例えば、第1のフィールドと第2のフィールドとの間に関係が存在すると決定することは、第2のフィールドの値が第1のフィールドの値によって参照されるキー値を含むと決定することを含み得る。データ品質ルール生成モジュール206は、第2のフィールドのそれぞれの値が有効なキー値であることを要求するように構成されたデータ品質ルールを生成するように構成されている。

【0087】

ルール生成モジュール206は、システム602によって実行される意味論的発見プロセスの出力によって、又は図7A～図7Bに関連して記載されるスキーマ分析によって、第1のフィールドと第2のフィールドとの間に関係が存在すると決定するように構成され得る。一例では、記録202、203のフィールド間の関係は、機械学習プロセスによって構成された少なくとも1つの分類子を使用して決定され得る。例えば、フィールドの意味論的ラベルを決定するために使用される1つ以上の試験(後に記載される)は、フィールド間の関係も決定し得る。フィールド間の関係は、フィールドのそれぞれの意味論的ラベルに関連付けられた属性231に含まれる。例えば、属性は、フィールドの意味論的ラベルが、フィールドがデータセットの一次キー値を含むことを示すと決定することを示し得る。一例では、ルール生成モジュール206は、一次キー値がそれぞれそのフィールド内で固有であることを要求するように、データ品質ルールを構成する。

【0088】

引き続き図3Bを参照すると、マルチテーブル分析を使用してルール生成モジュール206によって生成された例示的なデータ品質ルール233が示されている。前述のように、属性231は、図2A～図2Cに関連して記載される単一フィールドの制約とは異なり、マルチテーブル制約を含む。複数の記録202及び203は、記録202、203に関連するマルチテーブル属性231を決定するために、属性分析モジュール204によって分析される。いくつかの実装形態では、データセットスキーマ分析モジュール110(図7A～図7Bに関連して後に記載される)によって実行されるスキーマ分析は、マルチテーブル属性231を決定するために使用される。属性230と同様に、マルチテーブル属性231は、図8に関連して記載される意味論的発見プロセス中に決定され得る。ルール生成モジュール206は、図2Aに関して前述した属性230を使用するときと同様の方法で、マルチテーブル属性231を使用する。

【0089】

マルチテーブル制約(この例では、属性231)は、ルール生成モジュール206によ

って見出される。マルチテーブル制約は、複数のフィールドに適用される任意の制約を含み得る。マルチテーブル属性 2 3 1 から決定されるマルチテーブル制約から、ルール生成モジュール 2 0 6 は、フィールド間の関係をチェックするデータ品質ルール 2 3 3 b を生成するように構成され得る。ここで、マルチテーブル属性 2 3 1 は、SSN フィールドがユーザ ID フィールドを参照することを示す。ルール生成モジュール 2 0 6 は、データ品質ルール 2 3 3 b を生成する。例えば、データ品質ルール 2 3 3 b は、選択されたフィールドに結び付けられたユーザ ID フィールド内に対応する値が存在するかどうかを（選択されたフィールドの値ごとに）決定するためのチェックを行う。このデータ品質ルール 2 3 3 b は、対応するユーザ ID フィールドが選択されたフィールドに結び付けられており、それぞれの値がユーザ ID フィールド内の対応する既存の値を有すると決定することを要求する。ユーザ ID フィールドに値が存在しない場合、SSN フィールドに対するデータ品質ルール 2 3 3 は失敗となる。

【0090】

この例示的なルールは、結び付けられたフィールド内の値の存在のみをチェックする。しかしながら、より複雑なマルチテーブルデータ品質ルールが可能である。例えば、ルール生成モジュール 2 0 6 は、SSN フィールド内の値ごとに、有効な対応するユーザ ID がユーザ ID フィールド内に存在するかどうかを確認するためのチェックを行い得る。これは、ID が固有である、所与の数値範囲内にある、などのように、ユーザ ID フィールドに関連付けられたデータ品質ルールをチェックすることによって行われ得る。ユーザ ID フィールドに適用されるデータ品質ルールのいずれかが失敗となった場合、この例では、SSN フィールドに対するデータ品質ルールも失敗となる。

【0091】

図 3 C では、マルチファイルのデータ品質ルール 2 3 3 b を適用するためのプロセス 3 0 0 c が示されている。5 5 5 - 5 5 - 5 5 5 5 を含む SSN フィールドの値について、ルール適用モジュール 2 0 8 によるチェックがユーザ ID フィールドに対して実行される。ルール適用モジュール 2 0 8 は、テーブル 1 を含む記録 2 0 2 を受信する (2 7 1) ように構成されている。ルール適用は、テーブル 1、列 1 のデータセット内の場所 2 2 1 を有する「SSN」と命名されたフィールド 2 1 8 などのデータセット 2 0 2 ~ 2 0 3 のフィールドを選択する (2 7 3)。ルール適用モジュール 2 0 8 は、ラベルインデックス 6 1 0 を参照し (2 7 5)、どのルールがデータセット 2 0 1 内のテーブル 1、列 1 のフィールドに関連付けられているかを見つける。ここで、意味論的ラベル 2 2 8 「社会保障番号」は、フィールドを表すために使用される。ルール適用モジュール 2 0 8 は、データ品質ルールデータストア 2 5 0 にアクセスし (2 7 7)、フィールドラベル「社会保障番号」に関するデータ品質ルール 2 1 4 を取り出す。ルール適用モジュール 2 0 8 は、データ品質ルールをフィールド 2 1 8 の値に適用する (2 7 9)。一例では、値 5 5 5 - 5 5 - 5 5 5 5 は、フォーマットルール 2 3 3 a、文字制限ルール、及び固有の値ルールを通過する。

【0092】

マルチフィールドフィールド要件を含む第 2 のデータ品質ルール 2 3 3 b を適用するとき、ルール適用モジュール 2 0 8 は、第 2 のフィールドにアクセスし、ルール 2 3 3 b を適用するためにそのフィールドをチェックする。この例では、データ記録 2 0 3 のテーブル 2 内の対応するユーザ ID フィールドの値が<空>であり、ユーザ ID は存在しないことを示しているため、ルール 2 3 3 b は失敗となる。このように、フィールド 2 0 2 は全ての単一フィールドデータ品質ルールを通過した、マルチテーブルデータ品質ルール 2 3 3 b に失敗した。これは、単にそれぞれの個々のフィールドをチェックすることよりも、システムがデータ品質チェックをどのように失敗しているかをより明確にユーザに示し得る。ルール 2 3 3 a、b、c の全てを通過した結果は、失敗 (FAIL) である。適用モジュール 2 0 8 は、データ品質ルールの適用の結果 2 2 4 を出力データストア 1 1 4 に記憶する (2 8 1)。

【0093】

ラベル 2 2 8 に関連付けられたフィールドについて生成されたデータ品質ルールは、データセット 2 0 1 の第 2 のインスタンス又は新しい異なるデータセットのいずれかで、同じラベルに関連付けられた任意のフィールドに適用され得る。例えば、ラベル 2 2 8 の社会保障番号は、ルール生成モジュール 2 0 6 が 1 つ以上のデータ品質ルールを生成したフィールドをラベル付けする。同じラベル 2 2 8 を有する任意のフィールドは同じ意味論的意味を有するため、生成されたデータ品質ルールは、それらのフィールドに適用され得る。これはまた、データセットにわたってラベル 2 2 8 でラベル付けされた異なるフィールドにも当てはまり得る。例えば、新しいデータセット内の新しいフィールドがラベル 2 2 8 の社会保障番号を有する場合、データ記録 2 0 2 のために生成されたデータ品質ルールは、新しいデータセット内のその新しいフィールドに依然として適用可能である。

10

【 0 0 9 4 】

前述のように、他のデータ品質ルールは、フィールドに関連付けられた属性 2 3 1 から生成され得る。例えば、属性は、値がフィールド内の値の平均、値の最大長、値の最小長、値のデータタイプ、及び値のフォーマットのうちの少なくとも 1 つなどの対応するフィールドの値に基づくことを示し得る。例えば、フィールド内の 1 つ以上の許容値又は禁止値は、フィールドをラベル付けするフィールド名に関連付けられている。例えば、フィールド内の 1 つ以上の許容値又は禁止値は、データセット内の第 2 のフィールド内の値に基づいて決定され、第 2 のフィールドは、データセットの第 1 のフィールドに結び付けられる。例えば、1 つ以上の許容値又は禁止値は、第 2 のフィールド内の値と第 1 のフィールド内の値との組み合わせに基づく。

20

【 0 0 9 5 】

図 4 A ～ 図 4 B は、システム 6 0 2 の意味論的発見プロセスの結果を使用して試験データを生成するためのプロセス 4 0 0 a 及び 4 0 0 b を示す。試験データは、試験データセットモジュール 1 0 6 によって生成される。試験データセットモジュール 1 0 6 は、試験処理モジュール 3 0 4 及び試験データ生成モジュール 3 0 6 を含む。

【 0 0 9 6 】

図 4 A は、データセットのフィールドの意味論的意味を使用してアプリケーションの論理を試験するためのデータセット（例えば、データセット 4 0 1）から試験データを生成するように構成された試験処理モジュール 4 0 5 を示す。意味論的意味は、前述のように、属性 4 3 0 によって、また、いくつかの態様では意味論的ラベルによって表される。試験データセットモジュール 1 0 6 は、アプリケーション用の試験データのセットを生成するように構成されている。試験データは、アプリケーションによって処理される実際のデータを表し、可能な限り多くのアプリケーションの機能（例えば、アプリケーションの全ての機能）を試験し、かつ、試験データが有効であるように（例えば、全てのキーが有効である、フィールドが有効な値を含む、など）生成される。試験データに含まれるフィールドの選択、及び試験データの検証は、データセット 4 0 1 のフィールドのそれぞれの意味論的意味に基づいて実行され得る。

30

【 0 0 9 7 】

一般に、試験データを生成するために、試験データセットモジュール 1 0 6 は、意味論的ラベルでラベル付けされた（ないしは別の方法で、データセット 4 0 1 のフィールドのそれぞれについての意味論的意味を示す属性 4 3 0 に関連付けられた）フィールドを含むデータセット 4 0 1 を受信する（4 0 2）。この例では、試験処理モジュール 4 0 5 は、「1 2 - 2 2 - 1 9 6 0」の値 4 2 0 を含む「S S N」と命名されたフィールド 4 1 8 など、データセット 4 0 1 の 1 つ以上のフィールドを選択する（4 0 4）。試験処理モジュール 4 0 5 は、インデックス 6 1 0 内の意味論的ラベル 4 2 8 のルックアップ（4 0 6）を実行する。意味論的ラベル 4 2 8 を取り出すことに応答して、ラベル 4 2 8 a は、フィールド 4 1 8 をラベル付けするものとして見つけれられる。ラベル 4 2 8 a の属性 4 3 0 は、試験処理モジュール 4 0 5 によって受信される（4 0 8）。属性 4 3 0 は、フィールド 4 1 8 が 1 0 桁及び M M / D D / Y Y Y Y フォーマットを含む数字を含むことを示す。

40

【 0 0 9 8 】

50

属性 430 が受信されると、試験処理モジュール 405 は、試験構成データストア 403 から試験構成データ 407 を取り出す (410)。試験構成データストア 403 は、試験されるアプリケーションに関連付けられ得る。試験構成 407 は、生成される試験データに対する要件の仕様を含む。例えば、試験構成 407 は、無視するフィールドのリストを含み得る。アプリケーションは、データセット 401 を受信する前にフィールド名を示すデータを有さない場合があるため、無視するフィールドは、フィールドの意味論的意味に基づき得る。例えば、試験構成 407 を生成するとき、ユーザは、社会保障フィールドが試験について無視されるべきであることを指定することができる。試験データセットモジュール 106 は、どのフィールドが社会保障番号を表すかを決定し、それらを試験データから除去し得る。これは、試験データのデータフットプリントを低減し、記憶スペースを節約し得る。別の例では、試験構成データ 407 は、試験データの社会保障番号が固有であるべきであることを示し得る。試験データセットモジュール 106 は、データセット 401 のどのフィールドが社会保障番号を表すかを見つけ、フィールド内の重複エントリを除去して、試験データサイズを低減し得る。別の例では、試験構成データ 407 は、どのフィールドがキー値であるべきか、及びそれらが他のデータセット内のどのようなフィールドを指すべきかを示し得る。試験データセットモジュール 106 は、後に記載されるように、生成された試験データにおいてその構造を強制し得る。試験構成データ 407 には任意の他の同様の要件が含まれ得、該要件としては、スキーマ要件、データ品質要件、フィールド包含要件などが挙げられ得る。

【0099】

試験処理モジュール 405 は、試験構成データ 407 のルックアップ (410) を実行した後、フィールドのそれぞれについて識別された意味論的意味に基づいて、データセット 401 のフィールドに対して試験構成データに示される要件を強制する。例えば、図 4A の要件 412 は、データセット 401 内の誕生日を無視し、固有の社会保障番号を強制するための要件を含む。試験処理モジュール 405 は、属性 430 を試験構成データ 407 の要件 412 に関連付ける。

【0100】

いくつかの実装形態では、意味論的ラベル 428 の属性 430 は、意味論的ラベルを有するフィールドが、別の特定の意味論的ラベルを有する別のフィールドに対してどのような関係にあるべきかを含み得る。例えば、「誕生日」とラベル付けされたフィールド 418 の属性 430 は、所与のデータエントリの誕生日の値が、「登録日」とラベル付けされたフィールド内の別の値よりも常に早い時間であるべきであることを示し得る。意味論的ラベル 428 の属性 430 は、フィールドのデータ値がどのようなフォーマットであるか、どのような値が許可されるか (例えば、属性は、フィールドの全ての可能な値を含むルックアップテーブルを参照し得る) を示し得る。属性 430 は、フィールドの値がどの程度一般的であるべきか、フィールドに予想される値の範囲、及びフィールドのデータ値の他のそのような特徴を示し得る。属性 430 は、試験生成プロセス 400b 中に使用される。

【0101】

図 4B は、試験データ生成モジュール 415 による試験データ 422 の生成のプロセス 400b を示す。試験データ生成モジュール 415 は、選択されたフィールドに関連付けられた属性 430 を受信する (440)。試験データ生成モジュール 415 はまた、該当する場合、処理ルールが、従来の試験でフィールドをどのように処理してきたかを決定し得る。例えば、意味論的ラベル 428 でラベル付けされたフィールド上で処理ルールが何回実行されたかを示す処理ルールデータが受信され得る。いくつかの実装形態では、このデータは、試験構成データ 407 に示されている。いくつかの実装形態では、例えば、フィールドが以前の試験で試験されていない場合、フィールドは、試験データから除去され得る。

【0102】

試験データ生成モジュール 415 は、試験構成データ 407 を受信し (442)、試験

構成データ 407 及び属性 430 の要件に基づいて試験データを生成する (446)。一例では、それぞれ固有であり、試験構成データ 407 のフォーマット要件 412 に適合する 2 つの値が、SSN フィールド 418 に含まれる。SSN フィールド内の値は、第 1 のユーザ ID 値 (値 1) に関連付けられた「555-55-5555」、及び第 2 のユーザ ID 値 (値 2) に関連付けられた「666-66-6666」である。試験データ生成モジュール 415 は、1 つ以上のアプリケーションによる使用のために、試験データ 422 を試験データデータストア 450 に記憶する (448)。

【0103】

いくつかの実装形態では、試験構成データ 407 は、試験データ 422 のサブ設定ルールを示し得る。この例では、属性 430 が識別されると、試験データ生成モジュール 415 は、属性値及び処理ルールデータを受信し、どのサブ設定ルールをそのフィールドに適用すべきかを決定する。試験データ生成モジュール 415 は、フィールドの意味論的ラベル 428 に基づいて、フィールドのサブ設定ルールを決定する。サブ設定ルールは、フィールドの全ての値又は値の一部が、アプリケーションを試験するための試験データに含まれるべきかどうかを示す。例えば、特定の意味論的ラベルを使用して、フィールドの値が所与の試験での試験に有用であるか、又は試験中にフィールドが無視されるべきかが決定され得る。例えば、特定のアプリケーションにとって対象外であるフィールドは、アプリケーションを試験するときに無視され得る。前述のように、意味論的ラベルは、データ辞書内の 1 つ以上の属性に関連付けられる。属性は、フィールドの特性がどのようなものであるべきかを示す。例えば、属性は合わせて、フィールドがデータにとってどのような現実世界の意味であるかを示し得る。試験を構成する場合、ユーザは、試験のためにより大きなデータセットから、どのような現実世界のデータが関連しているかを示すことができる。データセットを受信されると、試験データセットモジュール 106 は自動的に、関連するフィールドを選択し、非関連フィールドを無視し得る。

【0104】

試験データ生成モジュール 415 は、意味論的ラベル 428 の属性に基づいてサブ設定ルールを選択する。いくつかの実装形態では、サブ設定ルールは、意味論的ラベル 428 に直接関連付けられ得る。例えば、サブ設定ルールは、特定のフィールドが、構成データ 407 に含まれる値に基づいて試験データに含まれるべきであるか否かを示し得る。例えば、構成データ 407 が、閾値数未満のフィールド上の動作の実行カウントを示す場合、フィールドは、試験データから除去され得る。

【0105】

他のデータが、構成データ 407 に含まれ得る。例えば、データ 407 は、フィールドが 1 つ以上のデータ品質ルール試験に失敗した場合、フィールドが試験データから除去され得ることを指定してもよい。例えば、フィールド内の空の値、誤ったフォーマットでデータを有する値、又は他のそのようなデータは、フィールドから除去され得る。別の例では、属性 430 が、フィールドが PII を含むことを示す場合、フィールドは、試験データから除去され得る。別の例では、特定の範囲外にある値は、試験データから除去され得る。別の例では、フィールドが、それぞれ重複値を有するいくつかのカテゴリを含む場合、サブ設定ルールは、固有の値を有するデータ記録のセットを抽出し得る。試験データが可能な限り重複値を有さないことを確実にすることは、システムの可能な試験入力 of 完全な選択を表しながら、試験データのフットプリントを可能な限り小さくすることを確実にする。

【0106】

いくつかの実装形態では、試験データ生成モジュール 415 は、フィールドに適用するためのサブ設定ルールのうちの 1 つ以上を選択し得る。サブ設定ルールの選択は、ユーザによって (例えば、ユーザインターフェースを介して) 検証され得る。サブ設定ルールの選択は、任意の目標のためであり得るが、一般に、目標は、可能な限り小さいデータフットプリントを有しながら、試験されるアプリケーションの完全な動作カバレッジを提供する試験データセット 422 を生成することである。このように、サブ設定ルールは、重複

値、誤った値、間違ってフォーマットされた値など、データセットから無関係な値を除去するために選択される。いくつかの実装形態では、選択されたサブ設定ルールの一覧は、出力データベース 114 又は試験構成データストア 403 に送信され得る。

【0107】

一態様では、試験データ 422 は、試験データセットのフィールドの値で動作し、少なくとも 1 つの出力値を生成するように構成された処理ルールを含むデータ処理アプリケーションに送信され得、フィールドは、ラベル提案でラベル付けされる。試験データセット生成モジュールは、第 1 のデータセットの処理と関連して処理ルールが実行された回数を示す実行情報を取得するように構成されている。第 1 のデータセットの処理中にデータ処理アプリケーションによって処理ルールが実行されるかどうかは、ラベル提案を有するフィールドの値に直接的又は間接的に依存し得る。サブ設定ルールは、第 1 のデータセットの処理と関連して処理ルールが実行された回数を示す実行情報に基づいて決定され得る。この例では、サブ設定ルールは、データセットのフィールドの識別を含む。例えば、サブ設定ルールは、データセットのデータ要素の少なくとも 1 つのキー値を含む、データセットのキーフィールドとしてフィールドを識別し得る。いくつかの実装形態では、データセットのデータ記録のサブセットは、所定の値を有するキー値を有するデータ記録を含む。いくつかの実装形態では、サブ設定ルールは、ラベル提案に対応するキーフィールド名のリストを識別する。サブ設定ルールは、フィールドのラベル提案をキーフィールドのリストと比較することによって、フィールドをキーフィールドとして識別する。

【0108】

一態様では、前述のように、データセットのフィールドが個人情報 (PII) を含む場合、サブ設定ルール生成モジュールは、PII 情報を含まないデータセットのデータ記録のサブセットを選択するように構成されたサブ設定ルールを生成し得る。代替的又は追加的に、データセットのフィールドが個人情報 (PII) を含む場合、試験データセット生成モジュールは、マスキング関数を PII に適用して、マスクされたデータを生成し、マスクされたデータを含むデータセットのデータ記録のサブセットを選択するように更に構成され得る。

【0109】

図 5A ~ 図 5B は、試験データセットモジュール 106 が、試験データ 422 が所与のシステムを試験するための実際のデータを正しく表すことを確認するように構成されているプロセス 500a 及び 500b を示し、これは、試験データ 422 のスキーマの検証を含み得る。例えば、アプリケーションは、動作中に特定の種類のデータを予想し得る。したがって、試験データは、アプリケーションによって処理される実データに近似すべきである。実データに近似する試験データを生成するために、試験処理モジュール 405 は、データセット 401 を受信し (452)、フィールド 418 及びフィールドの値 420 を選択する (454)。フィールド 418 は、社会保障番号と呼ばれる意味論的ラベル 428 でラベル付けされる。モジュール 405 は、ラベルを受信し (456)、フィールド 418 を説明する属性 430 を取り出す (458)。試験処理モジュール 405 は、意味論的ラベル 428 から、SSN フィールドが関連付けられている属性 430 を決定する。属性 430 は、現実のデータでフィールドの特徴を示し得る。例えば、実際の社会保障番号は、ユーザ ID テーブルで表され得る実在の人物とリンクしている。この特定の場合において、全ての実際の社会保障番号は、実在の人物に関連付けられているが、実在の全人物が社会保障番号を有するわけではない。試験データ生成モジュール 408 は、これらのメトリックを満たす試験データテーブルを生成するように構成されている。

【0110】

処理モジュール 405 は、スキーマデータ 409 を試験構成データストア 403 から取り出す (460)。スキーマデータ 409 は、意味論的発見プロセス (図 68 に関連して記載される) の最中又はその後、データセットスキーマ分析モジュール 110 (図 7A ~ 図 7B を参照して記載される) によって決定され得る。スキーマデータ 409 は、どのフィールドがデータセット内で結合キーであるか、どのフィールドが外部キーであるか、

などを示し得る。例えば、スキーマデータ409は、キーであるフィールドの名前を含み得る。このように、試験データセットモジュール106は、予想されるように、試験データ内の他のテーブルを適切に参照するため、参照に適した試験データを生成するように構成される。より具体的には、実データにおいて、社会保障番号を有するテーブルがユーザIDテーブルへの参照を含む場合、試験データセット生成モジュールは、試験データ内の既存のユーザIDテーブルを適切に参照するキーを含むフィールドを有する社会保障番号のテーブルを生成するように構成される。試験データがアプリケーションによって使用される場合、試験データは、アプリケーションによって予想される実データに近似する。アプリケーションを試験しているユーザは、試験データの処理結果におけるエラーが、試験データにおけるエラーではなく、アプリケーション自体におけるエラーの結果であると確信することができる。

10

【0111】

図5Bは、試験データ検証モジュール417による試験データ検証のためのプロセス500bを示す。検証モジュール417は、試験データ423がスキーマルール409を満たすかどうかを検証する。試験データ423は、テーブル内の重複した又は正しくない値及び記録を除去するためにテーブル402及び403から更新されたテーブル411及び413を含む。例えば、テーブル402は、エントリ402a～cを含み、それぞれ、テーブル403内の対応するエントリ403a～cの一次キーを参照する外部キーを含む。しかしながら、図5Aに示すように、テーブル403の一次キーは、キー「BBB」のエントリを重複させている。加えて、エントリ402a／403aのSSN 444-44-4444にはユーザIDの欠落がある。試験データ生成モジュール415は、テーブル402、403からエントリ402a及び403aを除去し、エントリ403cのキー値を「BBB」ではなく「CCC」に修正し、その結果、試験構成データ407の一次キー要件（例えば、それぞれの一次キーは固有である）が満たされる。モジュール417によって検証される場合、テーブル413の一次キーは、有効であるものとして検証され、テーブル411及び413の全てのエントリは、有効なユーザID値を含むものとして検証される。検証された試験データ423は、試験データセットデータストア450に記憶される（470）。

20

【0112】

図6A～図6Cを参照すると、データマスキングモジュール108によってデータをマスクするためのプロセス600a、600b、及び600cが示されている。データマスキングモジュール108は、（例えば、入力データ112から）データセット501のどのフィールドがマスクされるべきか、及びその場合、どのマスキング関数がそれらのフィールドに適用されるかを決定するように構成されている。マスキング関数は、データセット501の属性530及び意味論的ラベル528を使用して、データマスキングモジュール108によってフィールドに関連付けられる。

30

【0113】

一般に、マスクされたデータを生成するために、データマスキングモジュール108の属性分析モジュール504は、値（例えば、値520）を有するフィールド（例えば、フィールド518）を含むデータセット501を受信する（550）。属性分析モジュール504は、フィールド518などの所与のフィールドを選択する。次いで、属性分析モジュールは、インデックス610内の関連付けられた意味論的ラベルを取り出し（554）、データ辞書データベース614からどの属性530を取り出すか（556）を決定する。

40

【0114】

意味論的ラベル528は、その属性530に利用可能なそれぞれの意味論的ラベルに関連付けるデータ辞書に記憶される。フィールドがラベル付けされている場合、意味論的ラベル528がデータ辞書から選択され、フィールドは、意味論的ラベルの属性530に関連付けられる。データ辞書の意味論的ラベル528ごとに、データ処理デバイス102によって処理されるデータセット内のその意味論的ラベルに関連付けられたフィールド（複数可）（存在する場合）を示すインデックスが記憶される。

50

【0115】

それぞれの意味論的ラベル528は、意味論的ラベルの意味論的意味を示す1つ以上の属性530に関連付けられている。意味論的ラベル528の属性530は、意味論的ラベルを有するフィールドが、別の特定の意味論的ラベルを有する別のフィールドに対してどのような関係にあるべきかを含み得る。例えば、「誕生日」とラベル付けされたフィールドの属性530は、所与のデータエントリの誕生日の値が、「登録日」とラベル付けされたフィールド内の別の値よりも常に早い時間にあるべきであることを示し得る。意味論的ラベル528の属性530は、フィールドのデータ値がどのようなフォーマットであるか、どのような値が許可されるか（例えば、属性は、フィールドの全ての可能な値を含むルックアップテーブルを参照し得る）を示し得る。属性530は、フィールドの値がどの程度一般的であるべきか、フィールドに予想される値の範囲、及びフィールドのデータ値の他のそのような特徴を示し得る。

10

【0116】

属性は、マスクされたフィールドの特徴がどのようなものであるべきかを示す。例えば、属性は合わせて、フィールドがデータにとってどのような現実世界の意味であるかを示し得る。データをマスクするとき、ユーザは、どのような現実世界のデータがマスキングに関連するかを示すことができる。データセットが受信されると、データマスキングモジュール108は自動的に、マスキングに関連するフィールドを選択し、非関連フィールドを無視し得る。

20

【0117】

データマスキングモジュール108は、フィールドの属性530に基づいて、フィールドに対するマスキング関数を決定する。マスキング関数は、フィールドの全ての値又は値の一部が所与のアプリケーションに対してマスクされるべきかどうかを示す。例えば、特定の意味論的ラベルを使用して、フィールドの値がPIIを含むかどうか、又はマスキング中にフィールドが無視されるべきかどうか決定され得る。例えば、名前を含むフィールドの値は、スクランブルされ、それらが異なるインデックスに関連付けられるように混合されるか、ないしは別の方法で後続のアプリケーションのために匿名化され得る。

【0118】

図6Bは、マスキングモジュール506が、マスキング関数をデータセット501のフィールドに適用するプロセス600bを示す。マスキングモジュール506は、属性530が受信されると（560）、属性に基づいてマスキング関数を取得する（562）ように構成されている。属性530値は、フィールド518がマスクされるべきであること、及びどのようにフィールドがマスクされるべきかを示し得る。例えば、マスキング関数のリストが属性によって参照され得るか、又は特定のマスキング関数が示され得る。いくつかの実装形態では、属性530は、選択されたフィールドのデータがPIIを含むことを単に示し得、マスキングモジュール506は、フィールドの値をマスクする方法を決定する。いくつかの実装形態では、マスキング関数が、フィールドと一緒に提示されたときにフィールドのグループのうちの1つ以上に適用されるように、マスキング関数はフィールドのグループに基づき得る。これは、フィールドのいずれかが個々に提示されるときではなく、フィールドと一緒に提示されるときにのみ、PIIが明らかになる可能性がある場合に起こり得る。

30

40

【0119】

データマスキングモジュール108は、データをマスクするために1つ以上のマスキング関数514を使用し得る。マスキング関数514a～cは、フィールド内の値のシャッフル、フィールド内の値の代替値との置換、フィールド内の数値への分散の適用、暗号化機能、マスクされた値の削除、又は他のマスキング関数を含み得る。いくつかの実装形態では、マスキング関数は、マスクされたデータが有効な試験データであり、かつ／又は依然としてフィールドに対する1つ以上のデータ品質ルールに適合するように、前述の試験データセット生成及び／又はデータ品質プロセスと組み合わせられ得る。これらのプロセスについての異なる要件は、フィールドの意味論的ラベル528を使用して結び付けられ得

50

る。

【0120】

マスキングモジュール506は、存在する場合、どのマスキング関数（複数可）がフィールドに適用されるべきであることを決定する（564）。選択された関数514a及び選択されたフィールドは、マスキング関数データストア590内に記憶される（566）。したがって、マスキングのためにフィールド518のインスタンスが受信されると、マスキング適用モジュール508は、どのマスキング関数をマスキングに使用するか、又は（適用可能な場合に、データをマスク解除するために）マスキング関数のうちのどれが使用されているかを決定し得る。

【0121】

図6Cは、プロセス600c中に、受信されたデータセット501のフィールドに関連付けられたマスキング関数に従って、受信されたデータをマスクする又はマスク解除するように構成されたマスキング適用モジュール508を示す。マスキング適用モジュール508は、データセット501を受信する（570）ように構成されている。マスキングモジュール508は、データセット510内の場所520によってフィールド（例えば、フィールド518）を選択する（572）。モジュール508は、インデックス610内の意味論的ラベルをルックアップし（574）、どの属性530がフィールド518に関連付けられているかを決定する。属性530は、マスキング関数データストア590に表されるように、どのマスキング関数が使用されるかを示す。モジュール508は、マスキング関数データ516をデータストア590から取り出し（576）、マスキング関数をフィールド518のフィールド値に適用する（578）。例えば、「555-55-5555」であるフィールド値521は、関数F(x)によってマスクされた値522に変換され、「W2YSQP4LKED」（又は使用されるマスキング関数に応じて別の値）となる。モジュール508は、マスキングのマスク結果524を出力データストア114に記憶する（580）。

【0122】

一態様では、データマスキングモジュール108は、意味論的ラベルを有するフィールドと異なる意味論的ラベルを有する第2のフィールドとの間に関係が存在すると決定するように構成されている。関係が存在すると決定することに対応して、ルール適用モジュール408は、データマスキング関数を実行して、第2のフィールド内の値をマスクされた値に変換する。いくつかの実装形態では、第1のフィールドをマスクするために使用されたものとは異なる関数が、第2のフィールドをマスクするために使用され得る。

【0123】

一態様では、関係が存在すると決定することは、第1のラベル及び第2のラベルが共通のデータソースに関連付けられていると決定することを含む。例えば、共通のデータソースは、ユーザプロファイルを含み得る。関係は、第1のフィールドの値に対する第2のフィールドの値の依存性又は第2のフィールドの値に対する第1のフィールドの値の依存性を示し得る。いくつかの実装形態では、関係は、第1のフィールドの第1の値と第2のフィールドの第2の値との間の相関関係を示す。更に別の例では、関係は算術関数を含む。マスキングモジュール508は、関係のタイプに基づいて、データマスキング関数のタイプを選択するように構成され得る。前述のように、データマスキング関数のタイプは、シャッフル関数、データ暗号化、文字スクランブル、及びデータ置換のうちの1つ以上を含み得る。

【0124】

一態様では、モジュール508は、少なくとも1つの他のフィールドの1つ以上の特定の値がマスクされた値に変換されるかどうかを決定するために、データセット501をスキャンするように構成され得る。いくつかの実装形態では、モジュール508は、数値を含むデータセットの第1のフィールドに対するマスキング関数を選択するように構成され得、非数値を含むデータセットの第2のフィールドには、第2の異なるマスキング関数が選択される。

10

20

30

40

50

【0125】

図7A～図7Bは、データセットのフィールドに関連付けられた意味論的ラベルに基づいて、データセット701のスキーマを決定するためのプロセス700a～700bを示す。データセットスキーマ分析モジュール110は、データセットのスキーマが何であるかを決定するように構成されている。例えば、データセットスキーマ分析モジュール110は、どのフィールドがキー値又はインデックスであるか、どのようにフィールドが互いに参照するか、などを決定するように構成されている。

【0126】

一般に、データセットのスキーマを決定するために、属性分析モジュール704は、1つ以上のテーブル701a～bと、意味論的ラベルでラベル付けされたフィールドを含むデータセット701を受信する。データセットスキーマ分析モジュール110は、SSNフィールド及びフィールド__インデックスなど、テーブル701a～bのフィールドの意味論的ラベルに基づいて、フィールドの値間の関係を含めて決定する。データセット701のスキーマは、どのフィールドがインデックスであり、どの値がデータセット内の他の値を参照するかを示し得る。この情報は、データセットをマージする、データウェアハウスに記憶するためにデータセットを変換する、などのために、下流のアプリケーションで使用され得る。前述のように、意味論的ラベルは、データ辞書614内の1つ以上の属性に関連付けられる。属性730は、フィールドの特性がどのようなものであるべきかを示す。例えば、属性730は合わせて、フィールドがデータにとってどのような現実世界の意味を持つかを示し得る。データセットのスキーマを決定するとき、ユーザは、どのような現実世界のデータがスキーマに関連するかを示し得る。データセットが受信されると、データセットスキーマ分析モジュール110は自動的に、関連フィールドをインデックスとして選択し、非関連フィールドを無視し得る。

【0127】

一態様では、分析モジュール705は、データセットのフィールドを含むデータ記録701を受信する(702)。それぞれのフィールドは、一般に、フィールド名及び1つ以上のデータ値を含む。特定のフィールド718について、モジュール705は、フィールドを選択し(704)、フィールドに関連付けられた意味論的ラベル728を決定する(706)。データセットの場合、異なるフィールドの値は、共通のデータ記録にあることによって結び付けられ得る。データ記録は、データセット内のテーブルの行であり得るか、又は単にフィールドのうちの1つ以上からのデータ値の集合であり得る。フィールドに対する意味論的ラベル728がどのようなものであるかを決定するためのステップについては、図8に関連して以下に記載する。

【0128】

モジュール705は、フィールド718の属性730を決定する(708)。属性730は、社会保障番号フィールドとユーザIDフィールドとの間の関係を示し得る。モジュール705は、データセット内の発見された関係をスキーマ更新モジュール508に送信する。

【0129】

図7Bは、スキーマ更新モジュール707を使用してデータセットのスキーマを決定又は更新するためのプロセス700bを示す。モジュール707は、属性730で参照されるスキーマに適合するように、データセットの値を更新し得る。いくつかの実装形態では、モジュール707は、そのフィールド間で発見された関係を明示的に含むように、特定のデータセット701を記述するスキーマを更新する。モジュール707は、データセット701の属性730を受信し(712)、属性730によって与えられた情報に基づいてフィールドをスキャンすることによって関係を決定する(714)。例えば、モジュール707は、データ701においてテーブル1のフィールド__キーがテーブル2のフィールド__インデックスを指すことを決定し得る。モジュール707は、スキーマデータ735とこの関係を含むように、データセット701を記述するスキーマを更新する(716)。例えば、スキーマデータは、「テーブル2のフィールド__インデックスは一次キー

である。テーブル 2 にとってフィールド__キーは、テーブル 1 の外部キーである」ことを示し得る。生成されたスキーマ 7 3 5 は、スキーマデータストア 7 5 0 に記憶される (7 2 1)。下流アプリケーションは、スキーマデータ 7 3 5 を使用するためにデータストア 7 5 0 にアクセスし得る。

【0 1 3 0】

一態様では、スキーマ更新モジュール 7 0 7 は、スキーマデータに基づいてデータフローグラフを生成するように構成されている。データフローグラフは、スキーマデータ 7 3 5 内の関係に基づいてデータセットで動作し得る。例えば、スキーマ更新モジュールは、キー値に基づいて第 1 のデータセットを第 2 のデータセットに結合するように構成された結合関数を生成するように構成され得る。

【0 1 3 1】

図 8 を参照すると、データフィールドのデータから生成されたデータプロファイルを分析することによって、データフィールドを発見、分類、及びラベル付けするためのネットワーク化システム 6 0 0 が示されている。ネットワーク化システム 6 0 0 は、実行システム 6 0 2、1 つ以上のデータソース 6 1 2、参照データベース 2 1 6、開発環境 6 2 0、及び本番環境 6 1 4 を含む。実行システム 6 0 2 は、ネットワーク化システム 6 0 0 内のデータソース 6 1 2 a、6 1 2 b (集合的にデータソース 6 1 2) からソースデータを受信するように構成されている。実行システム 6 0 2 は、データソース 6 1 2 から受信したソースデータをプロファイルして、データソース 6 1 2 のソースデータを表すデータプロファイルを生成するように構成されている。実行システム 6 0 2 は、プロファイルデータを分析して、ソースデータの部分を発見、分類、及びラベル付けするように構成されている。より具体的には、実行システム 6 0 2 は、プロファイルデータを使用して、ソースデータの部分を分類する。ソースデータを分類することは、確率をソースデータの一部分に関連付けることを含む。確率は、ソースデータの部分が実行システム 6 0 2 によって認識されるラベルに対応する可能性を指定する。実行システム 6 0 2 は、自動的又はユーザ入力への応答のいずれかによって、ソースデータの部分を 1 つ以上のラベルでタグ付けするように構成されている。ソースデータの発見、分類、及びラベル付けは、ソースデータの分類を改善し、ソースデータのラベル付けの精度を高めるために繰り返され得る。ソースデータのラベルは、ソースデータに関連付けられたメタデータを含む。ソースデータの発見、分類、及びラベル付けが完了すると、ラベルは、1 つ以上の下流アプリケーションによって使用するために本番環境 6 1 4 にロードされる。意味論的発見プロセスについては、参照によりその全体が本明細書に組み込まれる、2 0 2 0 年 2 月 1 9 日に出願された「Discovering a Semantic Meaning of Data Fields From Profile Data of the Data Fields」と題する米国特許出願第 1 6 / 7 9 4 , 3 6 1 号に詳しく記載されている。

【0 1 3 2】

データソース 6 1 2 のソースデータは、いくつかの異なる種類のデータを含み得る。一例では、データソース (複数可) 6 1 2 のソースデータは、データフィールドを有するテーブルを含む。実行システム 6 0 2 は、テーブルのデータフィールドを発見、分類、及びラベル付けするように構成されている。例えば、実行システム 6 0 2 は、ソースデータのそれぞれの発見されたフィールドのデータコンテンツを分析し、データフィールドのデータコンテンツが何を表しているかを決定する。実行システム 6 0 2 は、データフィールドを既知のラベルに関連付けることによって (例えば、確率値を割り当てることによって) それぞれのデータフィールドを分類する。テーブルのラベル付けされたデータは、ラベル付けされたデータで動作するために他のアプリケーション及びシステムによってアクセス可能なデータ記憶装置 6 1 4 に出力され得る。データテーブルのデータフィールドを発見、分類、及びラベル付けするプロセスについては、後に詳しく記載する。データソース 6 1 2 は、任意のタイプのコンピューティングシステムを含み得る。例えば、データソース 6 1 2 としては、メインフレーム、データベース、サードパーティから供給された非構造化データ、データレイク、パーソナルコンピュータ、ハイスケールネットワークなどが挙

げられ得る。

【0133】

本開示では、データテーブルのデータフィールドを発見、分類、及びラベル付けするためのプロセスは、実行システム602及びネットワーク化システム600の機能を説明するための例として使用される。しかしながら、データフィールドは、実行システム602が発見、分類、及び検出するように構成されているものの一例であるが、実行システム602は、他のタイプのデータで動作し得る。例えば、実行システム602は、アプリケーションからアプリケーションデータを受信する。実行システム602は、アプリケーションのために異なるデータを発見、分類、及びラベル付けするように構成されている。データは、アプリケーションを動作させるために記憶され、アクセスされる異なるファイルを含み得る。実行システム602は、アプリケーションによって記憶されたPIEを発見し、マルウェア、アプリケーションファイルに対する変更などを発見するために使用され得る。別の例では、実行システム602は、ファイルシステムに記憶されたファイルを（例えば、パーソナルコンピュータ上で）分析し得る。実行システム602は、ファイルシステムをスキャンして、ユーザによって定義された特定のファイルサブセットを識別し得る。例えば、ユーザは、ホームコンピュータを再フォーマットする前に、ホームコンピュータから作業ファイルを除去することを望む場合がある。実行システム602は、ホームコンピュータのファイルシステムをスキャンし、全ての作業ファイルをタグ付けするように構成され得る。実行システム602は、忘れる権利、データ消去、サブジェクトアクセス要求、データ修正要求、データ停止、データポータビリティ、及び同意などのデータ主体の権利に関してデータをラベル付けするように構成され得る。多数の他のアプリケーションが可能である。

10

20

【0134】

ソースデータの部分（データフィールドなど）を発見、分類、及びラベル付けするために、実行システム602は、1つ以上のファイル618を求めて参照データベース216にアクセスするように構成される。ファイル618は、ソースデータの発見、分類、及びラベル付けを実行するためのコンテキストを実行システム602に提供する。いくつかの実装形態では、参照データベース216は、データフィールドのエントリ内に見られる値間の関係を記憶するルックアップテーブルを記憶し得る。例えば、データストレージは、コードを、ソースデータの発見、分類、及びラベル付けの最中に実行システム602によって参照され得る用語集にマッチングさせるルックアップテーブルを含み得る。参照データベース216のファイル618は、分類に使用される重み値を含み得る。例えば、重み値は、実行システム602に、2つの用語（例えば、ビジネス用語）が、ラベル付けされているソースデータ612に関して互いに結び付けられている確率を示し得る。これらの値は、ソースデータ618の発見、分類、及びラベル付けの初期反復中に生成され、その後の反復中に、ユーザによるか又は自動的のいずれかによって更新され得る。

30

【0135】

ファイル618は、実行システム602によるソースデータの発見、分類、及びラベル付けの前に、発見、分類、及びラベル付けの最中に、又は反復プロセスにおいて発見、分類、及びラベル付けの後に定義され得る。ネットワーク化システム600の開発環境620は、ユーザが参照データベース216に書き込むことを可能にするか、又はデータ記憶装置のファイル618を更新することを可能にする手段を提供し得る。例えば、開発環境620は、実行システム602のユーザにフィードバックを提供するユーザインターフェースを含み得る。例えば、開発環境620のユーザインターフェースは、どのようにデータフィールドがラベル付けされているか、及び実行システム602によってそれぞれの分類がどのような確率が行われるか、など、実行システム602がどのように実行されているかを示すレポートを表示し得る。ユーザ及びユーザインターフェースに提供されるフィードバックの例については、後に詳しく記載する。

40

【0136】

一般に、実行システム602は、プロファイルデータモジュール604、分類モジュール

50

ル 605、試験モジュール 606、結果確証モジュール 608、及びロードデータモジュール 660 の論理を実行するように構成された 1 つ以上のプロセッサを含む。プロファイルデータモジュール 604、分類モジュール 605、試験モジュール 606、結果確証モジュール 608、及びロードデータモジュール 610 のそれぞれの動作は、バッチ処理又はリアルタイムのいずれかで実行され得る。加えて、実行システム 602 は、モジュール 604、605、606、608、610 のそれぞれの動作をほぼ同時に又は異なる期間中のいずれかで実行し得る。例えば、いくつかの実装形態では、プロファイルデータモジュール 604 は、最初にデータソース 612 からソースデータのプロファイルを表すプロファイルデータを生成する。後に、所与の期間のデータソース 612 からの全てのデータがプロファイルデータモジュール 604 によってプロファイルされると、分類モジュール 605、試験モジュール 606、結果確証モジュール 608、及びロードデータモジュール 610 は、図 2A～図 7B に関連して前述したように、プロファイルデータを分析して、ソースデータのデータフィールドを発見、分類、及びラベル付けし、1 つ以上の下流アプリケーションのためにデータを参照データベース 216 にロードし得る。

【0137】

プロファイルデータモジュール 604 は、ソースデータ（例えば、テーブル、ファイルなど）を受信し、ソースデータのデータプロファイルを生成するように構成されている。プロファイルデータモジュール 604 は、ソースデータのフィールド（例えば、1 つ以上のデータセット）を発見する。プロファイルデータモジュール 604 は、ソースデータ内のテーブルの行を識別するか、フィールド名を検出するか、フィールドを参照するか、又は任意の同様のプロセスを使用することによってフィールドを発見し得る。プロファイルデータモジュール 604 は、データフィールドの統計的属性（複数可）を決定し、それらの統計的属性を含むプロファイルデータを生成する。プロファイルデータは、ソースデータのパターンを識別する。より具体的には、プロファイルデータは、ソースデータのテーブルのデータフィールドの値に関する統計を含む。例えば、プロファイルデータは、データフィールドのデータ値が数値データ、文字列などを含むかどうかを指定する情報を含み得る。例えば、データ値に関する統計は、データフィールドのそれぞれに含まれる値の最大値、最小値、標準偏差、平均などを含み得る（データが数値である場合）。いくつかの実装形態では、データに関する統計は、データ値のそれぞれのエントリにいくつかの数字又は文字があるかを含み得る。例えば、データプロファイルは、データフィールドのそれぞれのデータ値が 7 つ（又は 10 個）の数字を含むことを示し得、データフィールドが電話番号を含むことを示す文脈上の手がかりを提供し得る。例えば、データプロファイルは、データフィールドのそれぞれのエントリが、試験モジュール 606 によってルックアップテーブルとの比較をトリガするために使用され得る小さい値セットからの値を含むことを示し得る。

【0138】

プロファイルデータモジュール 604 のデータ取り込みは、フィールドのフィールド名、テーブル内のフィールド（又はファイルシステム内のファイル）の場所、及びデータのスキーマの分析を分析することを含む。言い換えれば、データの取り込みは、フィールドレベル、データセットレベル、及びスキーマレベルで行われる。

【0139】

フィールドレベルについては、プロファイルデータモジュール 604 は、フィールドの値及びフィールドのエントリを分析して、プロファイルデータを生成する。プロファイルデータモジュール 604 は、フィールド又はそのエントリの値がヌルであるか、空白であるか、特定のデータタイプに有効かどうかなどを決定し得る。プロファイルデータは、ヌルのパーセンテージ、空白のパーセンテージ、及びフィールドごとの値のパーセンテージに関する統計を含み得る。プロファイルデータモジュール 604 はまた、ベースラインパーセンテージ（開発環境を通じてユーザによって指定され得るか、又は自動的に生成され得る）からのこれらのパーセンテージの変化を示すデータを生成し得る。別の例では、プロファイルデータは、エントリのデータが暗黙のデータタイプに有効であるかど

うかの指標を含み得る。例えば、データフィールドが文字列フィールドであると認識されているが、日付データが見つかった場合、そのデータはそのエントリに対して無効であると推測され得る。別の例では、プロファイルデータは、エントリのデータが指定されたフォーマット（例えば、小数点以下2桁が必要であるが、明示的なタイプは指定されない）に有効であるという指標を含み得る。いくつかの実装形態では、この機能の一部は、プロファイルデータが生成された後、試験モジュール606によって実行される。

【0140】

データセットレベルについては、プロファイルデータモジュール604は、ソースデータの一部をソースデータ全体のデータセットに結び付ける統計を提供し得る。例えば、プロファイルデータモジュール604は、自然なキーの固有性及びキーのカーディナリティの指標を含む。プロファイルデータモジュール604は、ソースデータのキー値に重複があるかどうかを示す。プロファイルデータモジュール604は、近似的な自然キーの一致に重複があるかどうかを示す。プロファイルデータモジュール604は、これらの特徴のうちの1つ以上を有する記録のカウントを示す。いくつかの実装形態では、この情報は、例えば、フィールドがキー値であると決定された後に、結果確認モジュール608によって決定される。

【0141】

スキーマレベルについては、プロファイルデータモジュール604は、経時的にソースデータに対するデータフィールドの統計情報を決定するように構成されている。例えば、プロファイルデータモジュール604は、ソースデータの新しいバージョンに追加された列又は除去された列があるかどうかを検出する。プロファイルデータモジュール604は、区切られたフィールドを検出する。いくつかの実装形態では、プロファイルデータモジュール604は、ソースデータの以前のバージョンからいずれかの記録が脱落しているかどうかを決定するために、システムカウントを実行する。他のスキーマの展開が検出され得る。いくつかの実装形態では、この機能は、データフィールドがプロファイルデータにおいて発見された後、結果確認モジュール608によって実行される。

【0142】

いくつかの実装形態では、コンテキストデータが参照データベース216で利用可能である場合、プロファイルデータが改善され得る。例えば、ユーザがソースデータの1つ以上のデータフィールドのフォーマットを指定する場合、プロファイルデータは、それらのデータフィールドに関する追加の統計情報を生成し得る。これは、反復プロセスの一部であり得る。例えば、第1の反復でデータフィールドが発見されると（但し、その反復に対する分類が失敗した場合）、ユーザは、データフィールドのデータコンテンツを見つけ、プロファイルデータモジュール604（及び他のデータモジュール）による分析のための追加情報を実行システム602に提供することができる。例えば、ユーザが、全てのデータフィールドが数値データのみを含むべきであることを指定すると、プロファイルデータモジュール604は、どのデータが無効であるかを迅速に決定し、データプロファイル内のその情報の統計的測定値を提供し得る。

【0143】

プロファイルデータモジュール604は、プロファイルデータを生成する前に、ソースデータのデータセット全体をスキャンすることによって、ソースデータからプロファイルデータを生成する。プロファイルデータモジュール604は、ソースデータからデータセット全体をコピーする必要はないが、これは、プロファイルデータを生成するために実行され得る。データセットは大きくなる可能性があるため（例えば、何ギガバイトものサイズ、又は更には何テラバイトものサイズ）、分析のためにデータをローカルシステムに大規模にコピーすることは実際的ではない場合がある。一般に、プロファイルデータモジュール604は、データソースに対する低活動期間中にソースデータをスキャンする。

【0144】

分類モジュール605は、プロファイルデータを受信し、フィールドを含むソースデータを受信するように構成されている。ソースデータのフィールドごとに、分類モジュール

10

20

30

40

50

605は、ソースデータの（例えば、参照データベース216から）発見されたフィールドの既存のラベルを含むラベルインデックスをルックアップするように構成されている。これらのラベルは、ラベル付けプロセスの以前の反復からのものであるか、又はラベルインデックス（例えば、初期ラベルインデックス）は、手動で生成、インポート、ないしは別の方法で獲得され得る。しかしながら、ラベル付けプロセスを実行する前に、ラベルインデックスが存在する必要はない。

【0145】

フィールドについて、分類モジュールは、フィールドがラベルインデックス内のラベルにすでに関連付けられているかどうかを決定する。フィールドがまだラベル付けされていない場合、又はラベルインデックスが存在しない場合、分類モジュール605は、フィールドに関するラベルは見つからなかったと決定する。必要に応じて、分類モジュール605は、新しいラベルインデックスを生成して、意味論的ラベルに追加する。分類モジュール605は、フィールドデータタイプの分類を実行する。分類は、フィールドのプロファイルデータ、フィールド名、及びフィールドの値に基づき得る。例えば、分類モジュール605は、フィールドが「日付」フィールドであると決定し得る。いくつかの実装形態では、分類モジュール605は、フィールドが数値フィールド、文字列フィールド、又は他のそのようなデータタイプであると決定し得る。分類モジュール605は、フィールドのデータタイプを決定するが、フィールド（したがって、意味論的ラベル）の意味論的意味は、後に記載されるように、試験モジュール606によって決定される。例えば、分類モジュール605は、フィールドが日付フィールドであると決定し得、試験モジュール606は、日付フィールドの日付が顧客の「誕生日」であると決定する。別の例では、試験モジュール606は、数値フィールドが「ユーザID」フィールドであると決定する。多くの他のそのような例が可能である。分類モジュール605は、意味論的意味を見つけるための分類出力として試験モジュール606に送信される分類されたデータを生成する。分類されたデータは、分類モジュール605によって決定されたデータタイプでタグ付けされる。

【0146】

ラベルが見つかった場合、分類モジュールは、試験モジュール606及び結果確認モジュール608を通過し得るラベルデータを生成する。ラベルデータは、フィールドがすでにラベル付けされていることを試験モジュール606及び結果確認モジュール608に通知する。これは、フィールドに適用されるか、又はラベルを示唆する分類子に重み付けするために使用され得る。しかしながら、フィールドは、ラベルが正確であることを確認し、データ辞書データベース614内のそのラベルのラベル属性を潜在的に更新するために、分類モジュール605によって再分類され、試験モジュール606によって再試験され得る。例えば、試験モジュール606が、既存のラベルに適合性がないことを見出した場合、新しいラベルが示唆され得る。ユーザが既存のラベル（例えば、結果確認モジュール608によって提示されるもの）を選択した場合、データ辞書データベース614内のラベルのラベル属性が、そのラベルによってラベル付けされているデータ値を表していないラベルデータは、実行システム602へのフラグとして使用され得、これらの属性は更新され得る。

【0147】

いくつかの実装形態では、分類モジュール605は、機械学習アプローチを使用して複数の反復を介して更新され得る。例えば、発見されたフィールドがすでにラベル付けされている場合、分類子は、更なる分類が回避され得ることを決定し得る。別の例では、フィールドに適用されたスコアは、データソース612から、又はユーザ入力から受信される追加のデータに基づいて更新され得る。分類モジュール605は、以前の反復と比較して、異なる試験（複数可）が試験モジュール608によって実行されるべきであると決定し得る。例えば、フィンガープリント分析が以前の反復で不確定であった場合、分類子は、この試験が後続の反復で回避されるべきである（又は別の試験と交換されるべきである）ことを決定し得る。

10

20

30

40

50

【0148】

試験モジュール606は、プロファイルデータの統計を使用し、かつ参照データベース216に提供される追加のコンテキスト情報（ルックアップテーブル442など）を使用して、データセットのソースデータを分類するように構成されている。分類モジュール605の分類出力は、それぞれのフィールドにデータタイプのコンテキストを提供するため、及びフィールドの既存のラベル（存在する場合）を提供するために使用される。試験モジュール606は、データ辞書データベース614から候補ラベル440を受信するように構成されている。候補ラベルは、既存のラベルのライブラリであり、それぞれが、データフィールドの（例えば、下流アプリケーション又はユーザにとっての）意味論的意味を識別する1つ以上の属性に関連付けられる。前述のように、データ辞書データベース614内のそれぞれのラベルに関連付けられた属性は、ラベルの意味論的意味を提供する。試験モジュール606は、それらの候補ラベルのうちのどれがデータプロファイル、フィールド名、及びフィールドのデータ値によって決定されたデータフィールドの属性に最も密接に関連付けられているかを決定する。

10

【0149】

試験モジュール606は、1つ以上のラベルを試験されるそれぞれのフィールドに関連付けるために、それぞれが異なる試験モジュールによって実行される、複数の試験（又は分類子）を含む。例えば、データセットのフィールドがプロファイルデータモジュール604によって発見された後、試験モジュールは、フィールドの属性が候補ラベルのそれぞれの属性にどの程度近接して対応するかを決定する。それぞれの試験は、異なるデータ及びアプローチを使用して、1つ以上のラベルを提案する。異なる試験は分類のための異なるアプローチを使用するため、それぞれの試験から提案されたラベルは必ずしも同じではない場合がある。提案されたラベルは、後に記載される結果確証モジュール608で確証される。フィールドの意味論的意味を示すものとして信頼される単一の要因（例えば、フィールド名、又はフィールド内の特定の値若しくは値のセットの含有など）は存在しないため、フィールドの意味論的意味を識別するために異なる試験を使用することで、任意の単一試験を使用するよりもはるかに強固な意味論的意味の決定がもたらされる。

20

【0150】

それぞれのデータフィールドにおけるデータの試験は、データセットのデータの母集団レベル（データフィールド内で値がどのくらいの頻度で発生するか）、発見されたデータタイプ（例えば、日付又は数字は、文字列として保持される）、データフィールドのドメイン、キーフィールドの発見、フィールドが単一の単語であるか又は説明フィールドであるかの決定などのうちの1つ以上の決定を含み得る。例えば、データの分類は、日付及び時間の分析を含み得る。したがって、試験モジュール606は、プロファイルデータモジュール604からプロファイルデータを受信し、一連の統計ベースの関数を実行して、既知のラベルタイプのセットに対してフィールドの詳細を識別、分類、及び試験する。試験のルールは、分類モジュール605によって分類出力で識別され得るか、又は場合によってはプロファイルデータモジュール604によって生成されたプロファイルデータに含まれ得るデータタイプによって異なる。

30

【0151】

試験モジュール606は、データフィールド内のフィールド名及びエントリに対して複数の異なる分類試験を実行して、データフィールドをラベル付けする方法を決定するように構成されている。試験モジュール606は、参照データベース216から分類出力、候補ラベル、及び任意の参照データを受信し、これらのデータを試験に提供する。試験には、パターン分析、ビジネス用語分析、フィンガープリント分析、及びキーワード検索が含まれる。前述のように、フィールド名及びフィールドデータを有するテーブルは、実例として記載されているが、試験モジュール606の試験は他のデータタイプで実行され得る。ソースデータのデータフィールド及びデータエントリに対して実行される分類試験の例としては、ファジーマッチング計画、列データ計画、ビジネス用語マッチング計画、キーワードマッチング、フィンガープリント計画（例えば、コンテキストデータルックアップ

40

50

）、パターンマッチング、及び確証が挙げられ得る。

【0152】

試験モジュール606のファジーマッチング論理は、用語の辞書からのフィールド名のファジーマッチングの論理を含む。一般に、ファジーマッチングは、正確な一致を見つけれない場合、フィールド名と辞書内の用語との間の一致を見つけるように構成される。システムは、正確ではない辞書の用語を見つける。例えば、試験モジュール606が、あいまいな一致の閾値を100未満のパーセンテージ値に設定すると、辞書データベース（例えば、データベース216）は、そのパーセンテージに対応する（例えば、より大きい）そのメモリ内の任意の一致を返すことになる。いくつかの実装形態では、確率スコアがそれぞれの一致に割り当てられる。確率スコアは、ユーザに提示され得る。

10

【0153】

試験モジュール606のビジネス用語分析は、データフィールド名を用語集で既知のビジネス用語に一致させるための論理を含む。一般に、ビジネス用語は、異なるコンテキスト又はビジネス用語グループに入れられ得る。試験モジュール606は、別の単語又は用語内の特定の単語又は用語の出現を見つけるためのチェックを実行する。これには、特定の仕様の参照が含まれ得る。例えば、試験モジュール606は、「st」、「ln」、「ave」、「pl」、「ct」など、通り名の異なる略語を示す仕様を受信する。試験モジュール606は、それらの略語のいずれかがデータフィールド内に含まれるかどうかを決定するためのチェックを実行する。仕様のデータ要素のうちの1つ以上がデータフィールドに含まれる場合、試験モジュール606は、データフィールドが通り名を含むと決定する。この情報は、データフィールドが通り名を含み、したがってそのようにラベル付けされるべきであると決定するためなど、それ自体で使用され得る。この情報はまた、データフィールドが住所などの他の情報を含むことを示し得る。試験モジュール606は、他のデータに加えて、列に通り名が存在するという決定を使用して、データフィールドが住所、通り名のみ、又はいくつかの他のデータを含むかどうかを決定する。別の例では、「誕生日」という語句は、認識され、かかるラベルに関連付けられ得る。他のマッチング戦略としては、ファジーマッチング、同義語などを使用するマッチングが挙げられ得る。一般に、試験モジュール606は、確率スコアをそれぞれの一致に関連付ける。ユーザは、開発環境を通じてなど、この論理を構成するための仕様を生成することができる。

20

【0154】

試験モジュール606のパターンマッチング分析は、（フィールド名に加えて又はその代わりに）フィールドのデータコンテンツを使用する。パターンマッチングに使用されるパターンマッチングのタイプは、分類データの結果に基づいて、試験モジュール606によって決定され得る。例えば、分類データは、データが数値であるなど、フィールドのデータタイプを識別し得る。この例では、プロファイルデータはまた、データフィールド内のそれぞれのエントリが13～18文字の長さであることを示す。これは、データフィールドがクレジットカード番号データフィールドであり得ることを試験モジュール606に示し得る。これを確認するために、1つ以上のパターン試験が、疑わしいデータフィールドのデータに対して試験モジュール606によって実行され得る。例えば、それぞれのエントリの最初の4～6桁が、発行者コードのテーブルに対してチェックされ得る。最後の数字は、Luhn試験によって定義されるチェックディジットを含み得る。データフィールドのエントリの閾値パーセンテージがこれらのパターンのそれぞれを満たす場合、試験モジュール606は、フィールドがクレジットカード番号を保持するものと結論付け、フィールド名を適切なラベル及び確率に関連付け得る。パターンマッチング論理の場合、所与のフィールドのデータ自体及びフィールド内のデータのパターン（例えば、プロファイルデータで識別される）の両方を使用して、どのパターン試験を実行し、どのラベルを所与のデータフィールドに適用するかが見分けられ得る。

40

【0155】

試験モジュール606は、ソースデータ612でパターン分析試験を実行するかどうかを決定する。この決定は、分類データの結果であり得る。パターン一致分析は、プロファ

50

イルデータを使用して、ソースデータ 6 1 2 が候補フィールドラベルを示す所定のパターンに適合するかどうかを決定する。例えば、フィールドのデータが特定の長さ及び組成を有する場合、パターン一致分析は、対応する候補ラベルを識別する。ラベルに適用されるパターンスコアは、識別されたパターンが所定のパターンにどの程度密接に一致しているか、パターンがどれだけ特徴的であるか、又は任意の数の要因の関数であり得る。重みは、パターンの特徴性の関数として調整され得る。例えば、非常に珍しいパターンは、より高い重み値に対応し得る。発見されたフィールドの値がパターンに密接に（値の全て又は一部にわたって）一致する場合、スコアは、少数の値のみがパターンと一致する場合よりも高くなり得る。

【0156】

試験モジュール 6 0 6 は、キーワード検索試験を含み得る。キーワード試験（ビジネス用語マッチング試験と同様である）は、データフィールド内の特定のキーワードの検索を含むデータに基づいた試験を含む。例えば、住所を見つけるために、試験モジュール 6 0 6 は、「通り」、「道」、「大通り」などの共通の住所単語を検索する。試験は、ユーザによって拡張され得、ユーザは、新しいキーワードファイルを参照データベース 2 1 6 の仕様に追加することができる。キーワード試験は、データフィールドを固有に識別し得る限られた一組の共通単語が存在する住所及び会社名など、語句内の単語又は単語の一部を見つけるために使用され得る。

【0157】

試験モジュール 6 0 6 は、フィールド名でキーワード分析を実行するかどうかを決定し得る。この例では、試験モジュール 6 0 6 は、いくつかのフィールドが依然としてラベル値に関連付けられていない場合にキーワードマッチング試験を実行する。フィールド名（及び場合によってはフィールド値）は、テーブルから 1 つ以上のキーワードを含むかどうかをチェックされ、これは、試験モジュール 6 0 6 が特定のラベルをフィールドに関連付ける際の助けになり得る。試験モジュール 6 0 6 は、キーワード検索を実行し、キーワード検索スコアを生成する。

【0158】

試験モジュール 6 0 6 は、ソースデータのデータ値を分析するためのフィンガープリント試験を含み得る。フィンガープリントの論理は、一般に、データフィールド全体のデータルックアップを含む。フィンガープリントの論理は、データ値に基づいた試験を含む。データフィールドのフィンガープリントの論理は、既知のリスト（例えば、参照データベース 2 1 6 から）をデータフィールドのデータと比較して、データフィールドのデータがリストのデータと相関するかどうかを決定することを含む。例えば、データフィールドからのデータは、ファーストネームのリスト、州名のリスト、都市名のリストなどと比較され得る。フィンガープリント（既知のデータリスト）は、概して、包括的というよりは代表的なものである。言い換えれば、フィンガープリントは、フィンガープリントのセットの一部である値のそれぞれ及び全ての例を含む必要はない。一般に、フィンガープリントは、データフィールドのデータエントリに現れるべきであるほぼ最も一般的な値を表す、選択された例示的な値を含み得る。例えば、フィンガープリントは、ファーストネームのテーブルについて米国内にあり得る全てのファーストネームを必要とするわけではなく、むしろ、最も一般的な名前の選択されたグループで十分であり得る。例えば、上位 1 0 0 個の名前は、一般に、データフィールドとデータフィンガープリントとの間の相関を示すのに十分なデータをもたらす。データフィンガープリントは、マスターデータから生成され得る。例えば、システムは、データフィールドが都市名に対応するかどうかをチェックするために、5 0 ～ 1 0 0 個の最も人口の多い米国都市を含み得る。ユーザは、特定のシステムに対するフィンガープリント試験の機能を高めるために、新しいドメインを仕様に追加することができる。

【0159】

試験モジュール 6 0 6 は、フィールドのフィンガープリントテーブルをチェックするかどうかを決定する。フィンガープリントは、フィールドに可能な値の長いリストが存在す

10

20

30

40

50

るが、値のうちのいくつかは他の値よりも一般的であることが予想される状況でうまく作用し得る。例えば、都市名、通り名、更にはファーストネーム及びラストネームは、フィンガープリント分析の良好な候補である。試験モジュール606は、フィンガープリント検索を実行し、フィンガープリントスコアを生成する。

【0160】

試験モジュール606は、ビジネス用語分析を実行するかどうかを決定する。ビジネス用語分析は、フィールドをラベル付けするためのビジネス用語に対応し得るソースデータ612に多くの固有な用語が存在する場合に有用であり得る。ビジネス分析が実行される場合、試験モジュール606は、フィールド名をビジネス用語と比較して一致を見つけ、ラベル及びそれらの関連する確率を生成することによって、ビジネス用語分析を実行する。ビジネス用語分析の結果は、パターン一致分析と同様に、重み及びスコアに関連付けられる。

10

【0161】

試験モジュール606は、試験全体で結果を比較して、確証論理を使用して結果を改善するように構成され得る。例えば、確証論理は、別の分類試験を使用してビジネス用語の一致を検証するために使用され得る。例えば、データフィールドが旧姓値を含むことを指定する試験は、ラストネームフィールドである分類試験も通過するべきである。

【0162】

試験モジュール606は、機械学習論理を実行するように構成されており、機械学習論理では、（例えば、特定のソースからの）以前のデータセットの分類又は同じデータセットの以前の反復の分類が辿られ、どの試験が後続の反復に選択されるか、及びそれらの後続の反復の確率値がどのように決定されるかが左右される。機械学習論理は、データセット上で訓練され、訓練データを使用して開発された重みを適用して、データセットの新しいデータを分類し得る。

20

【0163】

試験のそれぞれは、分析されるデータフィールドについて1つ以上の提案されたラベルを出力し得る。試験は合致している必要はない。それぞれの提案されたラベルは、スコア（図示せず）及び重み値（図示せず）に関連付けられ得る。それぞれのラベルのスコア及び重みは、提案されたラベルの特定のラベルを、フィールドの意味論的意味を識別するものとして示唆するため、及びラベルをカテゴリに分類する（又はラベルをバケットに分ける（bucket））ために、結果確証モジュール608によって使用され得る。カテゴリ（後に記載）は、試験の間にいくつかの合致が存在するかを示し、そうして、フィールドのデータの意味論的意味を識別するものとして提案されたラベルの信頼性を示唆する。

30

【0164】

試験を実行するために、試験モジュール606は、プロファイルデータ、分類データ、及び他の参照データを受信し、それぞれの試験が実行されるかどうかを決定する。例えば、受信されているデータのタイプが特定の試験用にフォーマットされていない場合、試験は実行され得ない。試験の任意の組み合わせが実行され得る。試験モジュールは、パターン分析を実行するかどうかを決定し、ビジネス用語分析を実行するかどうかを決定し、フィンガープリント分析を実行するかどうかを決定し、キーワード検索を実行するかどうかを決定する。それぞれの試験は、1つ以上の提案されたラベルを生成し、提案されたラベルを出力する。提案された全てのラベルを含む試験結果は結合され、試験結果として結果確証モジュール608に送信される。

40

【0165】

いくつかの実装形態では、早期の試験からまだ結果が見出されていない場合、後続の試験が実行され得る。いくつかの実装形態では、試験モジュール606は、分析されるソースデータ612に基づいて、参照データベース216からデータを取り出すことによって、参照データを準備する。例えば、試験モジュール606は、ソースデータ612の場所、ソースデータ内のテーブル名、ユーザ入力などに基づいて、参照データベース216からデータを取り出し得る。試験モジュール606は、どのルックアップテーブルがアクセ

50

スされ、ソースデータのそれぞれのフィールドに対してチェックされるかを決定し得る。ルックアップの選択は、プロファイルデータモジュール 6 0 4 から受信したプロファイルデータに基づいて実行され得る。

【0166】

いくつかの実装形態では、試験モジュール 6 0 6 は、スキーマ分析を実行するかどうかを決定する。スキーマ分析は、ソースデータ 6 1 2 を経時的に、かつ全体として分析するために使用され得る。例えば、フィールドの欠落、追加、除去などが生じている場合、情報は、他のデータフィールドをラベル付けするために使用され得る。

【0167】

いくつかの実装形態では、試験がそれぞれ完了すると、試験モジュール 6 0 6 は、関連するスコアを組み合わせる最終スコアを生成し、最終スコアは、スコアが適用される関連する提案されたラベルと共に表示される。いくつかの実装形態では、上流スコアは、下流スコアに影響を与えるため、スコアは互いに区別されないが、分析が行われるたびに更新されるスコアを表す。いくつかの実装形態では、それぞれの提案されたラベル及びそのスコアは、結果確認モジュール 6 0 8 に別々に報告され、結果確認モジュール 6 0 8 は、次いで、試験結果 3 0 6 を分類する方法を決定する。

【0168】

いくつかの実装形態では、試験の順序は、より処理集約的な試験が最後にスケジュールされるようなものであり得る。他の試験が失敗した場合、より時間集約的な試験が最後の手段となり得る。このように試験を順序付けすると、ソースデータ 6 1 2 をラベル付けするための実行システム 6 0 2 に対する処理時間が短縮され得る。

【0169】

試験モジュール 6 0 6 が、ソースデータと、ソースデータのデータフィールドのラベルについて決定された確率値との分類試験を実行すると、結果確認モジュール 6 0 8 は、分類された結果が高い信頼性を示すかどうか、又は更なる試験が実行されるべきかどうかを決定するためのチェックを実行する。

【0170】

結果確認モジュール 6 0 8 は、試験モジュール 6 0 6 によってソースデータで実行された異なる試験の試験結果を受信し、結果が互いを確認するものであるか又は互いに矛盾しているかどうかを決定する。試験モジュール 6 0 6 の試験の結果は、結果確認モジュール 6 0 8 によっていくつかの分類カテゴリに区分される。カテゴリとしては、一致カテゴリ、推奨カテゴリ、調査カテゴリ、及び無視カテゴリが挙げられる。カテゴリのそれぞれは、試験結果におけるラベル提案間の類似性を示す。例えば、全てのラベル提案が同一である場合、試験結果は高レベルの類似性を有する。それぞれのラベル提案が異なる場合、試験結果は、低レベルの類似性を有する。類似性は、それぞれの試験による単なる投票機構以上のものであり得る。それぞれのラベル提案は、重み付けされたスコア値に関連付けられる。試験のあるラベル提案が、他のラベル提案に一致していないが、他の試験の他の提案されたラベルに比べて比較的高いスコア値及び大きい重みに関連付けられている場合、4 つの試験のうちの 3 つが合致しているにもかかわらず、類似性は低いものとして識別され得る。

【0171】

識別されたカテゴリに応じて、結果確認モジュール 6 0 8 は、フィールドの意味論的意味を識別するときにラベルを自動的に検証するか、又はラベルを手動で検証するようにユーザを促すかのいずれかを行い得る。後に記載されるように、検証は、ユーザインターフェースでクライアントデバイスを介して行われ得る。

【0172】

一致カテゴリは、ラベルに対する最大の信頼レベル（例えば、一致度 1 0 0）を示す。いくつかの実装形態では、一致したラベルは、ソースデータに関連付けられたメタデータに自動的に含まれる。しかしながら、結果確認モジュール 6 0 8 は、依然として、この結果を手動検証のためにユーザに提示するように構成され得る。一般に、一致カテゴリは、

10

20

30

40

50

実行される全ての試験が同じラベルを提案したことを示す。いくつかの実装形態では、一致カテゴリは、ラベルの全ては一致していないが、一致していないいずれのラベルもスコア閾値を下回り、試験間の一般的な合致を示している場合に選択され得る。

【0173】

推奨カテゴリは一般に、少なくとも1つのラベルがデータフィールドとの高い品質関連を有することを示す。しかしながら、推奨されるラベルは、一般に、最大の信頼に設定された閾値レベルを下回り、更なる検証が好ましい。いくつかの実装形態では、推奨カテゴリは、データフィールドに関連付けられているいくつかの高品質ラベルを示す。いくつかの実装形態では、結果確認モジュール608は、推奨されるラベルをランク付けしてリストを作成し、それぞれに確率スコアを示し、ユーザがデータフィールドの最良のラベル（複数可）を選択する際の助けになり得る。

10

【0174】

調査カテゴリは、任意の特定の試験を通過しないいくつかの興味深い統計的プロパティを有する値又はデータフィールドを示す。例えば、データフィールドは、意味を有しているかのように見えるが、いずれの試験もラベルを提案していないか、又は提案しているが、提案されたラベルのスコアは所与の閾値を下回っている場合がある。例えば、データフィールドは、データフィールドがソースデータ内の重要なドメイン又はフィールドであるが、推奨される（又は閾値確率を超えて推奨される）ラベルがないことを示すプロファイル属性を含み得る。一般に、そのような結果は、追加のルールが試験に追加され（例えば、試験は、何らかの方法で拡張又は変更されるべきである）、再実行されるべきであることを示す。

20

【0175】

無視カテゴリは、データフィールドが対象外のものとして指定されるか、又はフィールドが試験をトリガせず、データフィールドが重要であることを示唆するプロファイル属性を表示していないかのいずれかであることを示す。一般に、無視カテゴリは、受信されたデータがエラーを有するか、又は認識可能なパターンを有さないときに示される。無視カテゴリは、一般に、フィールドが特定の構造を有さない種々のデータを含むことを示すと同時に、新しいラベルが定義されるべきであることを示し得る。

【0176】

一般に、試験結果をカテゴリに分類する（例えば、バケットに分ける）ために、結果確認モジュール608はプロセスを実行する。結果確認性モジュール608は、フィールドからの試験結果を比較する。フィールドごとに、結果確認モジュールは、類似性のレベルに基づいて試験結果をカテゴリに割り当てる。

30

【0177】

分類カテゴリを指定する分類ファイルを生成することに加えて、結果確認モジュール608は、他のファイルを生成するように構成される。他のファイルには、プロファイルコメントファイルが含まれる。プロファイルコメントファイルは、分類分析プロセスの結果を記述するファイルである。他のファイルは、ビジネス用語マッチングプロセス及びフィンガープリントプロセスからの組み合わせられた出力を記述するものである、発見結果の要約を含む。

40

【0178】

結果確認モジュール608は、それぞれのデータフィールドの集計及び範囲に基づいて、データ値の外れ値が存在するかどうかを決定するように構成されている。外れ値は、データフィールドの識別された関係又はフォーマットに適合しないデータ値を含む。結果確認モジュール608は、外れ値が、試験モジュール606の分類からの関係のクラスタリング及び予測に基づいて決定されたことを決定する。

【0179】

上記のプロセスは、分類の精度を高め、ユーザが分類試験を更新してより良好な結果を得るために反復され得る。前述のように、いくつかの実装形態では、機械学習論理を使用して、このプロセスを容易にするために、それぞれの反復中に分類子（複数可）が訓練さ

50

れ得る。一般に、プロファイルデータモジュール 604 によってプロファイルデータが生成されると、試験モジュール 606 及び結果確認モジュール 608 のプロセスが反復され、ソースデータ自体が更新されない限り、新しいプロファイルデータが繰り返し生成される必要はない。

【0180】

プロファイルデータモジュール 604、分類モジュール 605、試験モジュール 606、及び結果確認モジュール 608 が、データフィールドのそれぞれのラベルを生成すると、ロードデータモジュール 610 は、ラベルインデックスを含むメタデータを参照データベース 216 にロードし得る。ロードデータモジュール 610 は、ラベルインデックスを更新し、データを参照データベース 216 にロードするためのプロセスを実行する。ロードデータモジュール 610 は、フィールド名を受信し、手動又は自動のいずれかで検証された、フィールドの提案されたラベルを受信する。ロードデータモジュール 610 は、フィールド名と提案されたラベルとを結合する。ロードデータモジュールは、ラベルをデータセット内のフィールドの場所に関連付けることによって、ラベルインデックスを更新する。ロードデータモジュールは、分析されるデータセットのラベルを、実行システム 602 によって及び下流アプリケーションによって、データセットに関して参照され得るラベルインデックスに結合する。

【0181】

一般に、参照データベース 216 は、様々なアプリケーションのために 1 つ以上の下流コンピューティングシステムによってアクセスされ得る。例えば、生成されたデータセットのラベルは、データ品質の強制、個人データの匿名化、データマスキング、(PII) レポート、試験データ管理、データセットアノテーションなどに使用され得る。

【0182】

ロードデータモジュール 610 は、メタデータ及びソースデータを、1 つ以上の他のコンピューティングシステムによって使用可能なパッケージにパッケージ化するように構成されている。例えば、プロファイルデータが生成されると、分類モジュール 605、試験モジュール 606、及び結果確認モジュール 608 の動作はそれぞれ、それらのプロセスの複数のインスタンスをソースデータの異なるパーティションで並列に実行するように構成され得る。例えば、ソースデータが複数のテーブルを含む場合、ソースデータは、テーブルによってパーティション化され得る。試験モジュール 606 及び結果確認モジュール 608 は、実行システム 602 のプロセスのスループットを向上させるために、複数のテーブルに対してそれらの論理プロセスのインスタンスを同時に実行し得る。ソースデータがラベル付けされると、ロードデータモジュール 610 は、パーティションと一緒にマージし、ラベル付けされたソースデータを参照データベース 216 に記憶するように構成され得る。

【0183】

図 9～図 12 は、例示的なプロセスをそれぞれ含むフロー図を示す。図 9 は、1 つ以上のデータセットにおける値に対するデータ品質ルールを決定するためのプロセス 900 を示す。プロセス 900 は、前述のようなデータ処理システムによって実装され得る。プロセス 900 は、データ処理システムによって、1 つ以上のデータセットを受信する (902) ことを含み、1 つ以上のデータセットはフィールドを含み、フィールドは値を記憶する。プロセス 900 は、データ処理システムによって、1 つ以上のデータセットに含まれるフィールドに記憶された値をプロファイルする (904) ことを含む。プロセスは、データ処理システムによって、1 つ以上の分類子をプロファイルされた値に適用する (906) ことを含む。プロセス 900 は、フィールドの意味論的意味を識別する 1 つ以上のラベル提案を識別する (908) ことを含み、1 つ以上のラベル提案のそれぞれは、計算された信頼レベルを有する。プロセス 900 は、フィールドをラベル付けする (910) ことを含み、ラベル提案は、閾値レベルを満たす計算された信頼レベルを有する。プロセス 900 は、フィールドをラベル付けするラベル提案に基づいて、フィールドのプロファイルを決

含まれる値の1つ以上の特徴を表す。プロセス900は、決定されたプロファイルに基づいて、データセットのフィールドに対するデータ品質ルールを決定する(914)ことを含む。

【0184】

図10は、データ処理アプリケーションの試験中に処理ルールの実行を引き起こすように試験データを選択するためのプロセス1000を示す。プロセス1000は、前述のようなデータ処理システムによって実装され得る。プロセス1000は、データ処理システムによって、1つ以上のデータセットを受信する(1002)ことを含み、1つ以上のデータセットはフィールドを含み、フィールドは値を記憶する。プロセス1000は、データ処理システムによって、1つ以上のデータセットに含まれるフィールドに記憶された値をプロファイルする(1004)ことを含む。プロセスは、データ処理システムによって、1つ以上の分類子をプロファイルされた値に適用する(1006)ことを含む。プロセス1000は、フィールドの意味論的意味を識別する1つ以上のラベル提案を識別する(1008)ことを含み、1つ以上のラベル提案のそれぞれは、計算された信頼レベルを有し、閾値レベルを満たす計算された信頼レベルを有するラベル提案でフィールドをラベル付けする。プロセス1000は、フィールドをラベル付けするラベル提案に基づいてサブ設定ルールを決定する(1010)ことを含む。プロセス1000は、サブ設定ルールに従ってデータセットからデータ記録のサブセットを選択する(1012)ことを含み、サブセットの選択は、データセットのフィールドをラベル付けするラベル提案に基づく。プロセス1000は、選択されたサブセットを、データ処理アプリケーションの試験のためにデータ処理アプリケーションに提供する(1014)ことを含む。

【0185】

図11は、データ処理アプリケーションのデータをマスクするためのプロセス1100を示す。プロセス1100は、前述のようなデータ処理システムによって実装され得る。プロセス1100は、データ処理システムによって、1つ以上のデータセットを受信する(1102)ことを含み、1つ以上のデータセットはフィールドを含み、フィールドは値を記憶する。プロセス1100は、データ処理システムによって、1つ以上のデータセットに含まれるフィールドに記憶された値をプロファイルする(1104)ことを含む。プロセスは、データ処理システムによって、1つ以上の分類子をプロファイルされた値に適用する(1106)ことを含む。プロセス1100は、フィールドの意味論的意味を識別する1つ以上のラベル提案を識別する(1108)ことを含み、1つ以上のラベル提案のそれぞれは、計算された信頼レベルを有し、閾値レベルを満たす計算された信頼レベルを有するラベル提案でフィールドをラベル付けする。プロセス1100は、ラベル提案に基づいて、データセットのフィールドが機密データを表すと決定する(1110)ことを含む。プロセス1100は、この決定に回答して、フィールドの機密データを含む値をマスクされた値に変換するためにデータマスキング関数を実行する(1112)ことを含む。

【0186】

図12は、データセットのスキーマを決定するためのプロセス1200を示す。プロセス1200は、前述のようなデータ処理システムによって実装され得る。プロセス1200は、データ処理システムによって、1つ以上のデータセットを受信する(1202)ことを含み、1つ以上のデータセットはフィールドを含み、フィールドは値を記憶する。プロセス1200は、データ処理システムによって、1つ以上のデータセットに含まれるフィールドに記憶された値をプロファイルする(1204)ことを含む。プロセスは、データ処理システムによって、1つ以上の分類子をプロファイルされた値に適用する(1206)ことを含む。プロセス1200は、フィールドの意味論的意味を識別する1つ以上のラベル提案を識別する(1208)ことを含み、1つ以上のラベル提案のそれぞれは、計算された信頼レベルを有し、閾値レベルを満たす計算された信頼レベルを有するラベル提案でフィールドをラベル付けする。プロセス1200は、フィールドをラベル付けするラベル提案に基づいて、ラベル提案を有するフィールドに含まれる値の1つ以上のスキーマ特徴を決定する(1210)ことを含む。プロセス1200は、1つ以上のスキーマ特徴

に基づいて、第1のラベル提案を有する第1のフィールドが、第2のラベル提案を有する第2のフィールドの値によって参照されるキー値を含むと決定する(1212)ことを含む。プロセス1200は、この決定に応答して、第1のフィールドが第2のフィールドの値によって参照されるキー値を含むことを参照するためにデータセットを記述するスキーマデータを更新する(1214)ことを含む。

【0187】

本明細書に記載された主題及び動作のいくつかの実装は、本明細書に開示された構造及びそれらの構造的均等物を含むデジタル電子回路、又はコンピュータソフトウェア、ファームウェア、若しくはハードウェア、又はそれらの1つ以上の組み合わせに実装することができる。例えば、いくつかの実装形態では、データ処理デバイス102のモジュールは、デジタル電子回路を使用して、又はコンピュータソフトウェア、ファームウェア、若しくはハードウェア、又はそれらのうちの1つ以上の組み合わせに実装することができる。別の例では、プロセス900、1000、1100、及び1200は、デジタル電子回路を使用して、又はコンピュータソフトウェア、ファームウェア、若しくはハードウェア、又はそれらのうちの1つ以上の組み合わせに実装することができる。

【0188】

本明細書に記載のいくつかの実装形態(例えば、データ品質ルールモジュール104、試験データセットモジュール106、データマスキングモジュール108、データセットスキーマ分析モジュール110など)は、デジタル電子回路、コンピュータソフトウェア、ファームウェア、又はハードウェアの1つ以上のグループ又はモジュールとして、又はそれらの1つ以上の組み合わせに実装することができる。異なるモジュールを使用することができるが、各モジュールは別個である必要はなく、複数のモジュールを同じデジタル電子回路、コンピュータソフトウェア、ファームウェア、若しくはハードウェア、又はそれらの組み合わせに実装することができる。

【0189】

本明細書に記載のいくつかの実装形態は、1つ以上のコンピュータプログラム、すなわち、データ処理装置によって実行するために、又はデータ処理装置の動作を制御するためにコンピュータ記憶媒体上に符号化されたコンピュータプログラム命令の1つ以上のモジュールとして実装することができる。コンピュータ記憶媒体は、コンピュータ可読記憶デバイス、コンピュータ可読記憶基板、ランダムアクセス若しくはシリアルアクセスメモリアレイ若しくはデバイス、又はそれらの1つ以上の組み合わせであり得るか、又はそれらに含まれ得る。更に、コンピュータ記憶媒体は伝搬信号ではないが、コンピュータ記憶媒体は、人工的に生成された伝搬信号に符号化されたコンピュータプログラム命令のソース又は対象とすることができる。コンピュータ記憶媒体はまた、1つ以上の別個の物理的コンポーネント又は媒体(例えば、複数のCD、ディスク、又は他の記憶デバイス)であってもよく、又はそれらに含まれてもよい。

【0190】

「データ処理装置」という用語は、例として、プログラマブルプロセッサ、コンピュータ、システムオンチップ、又は上記の複数のもの、又は組み合わせを含む、データを処理するためのあらゆる種類のデバイス、デバイス、及び機械を包含する。いくつかの実装形態では、データ品質ルールモジュール104及び／又はデータ構造モジュール106は、本明細書に記載のデータ処理装置を備える。装置は、例えばFPGA(フィールドプログラマブルゲートアレイ)又はASIC(特定用途向け集積回路)などの専用ロジック回路を含むことができる。装置はまた、ハードウェアに加えて、問題のコンピュータプログラムの実行環境を作成するコード、例えば、プロセッサファームウェア、プロトコルスタック、データベース管理システム、オペレーティングシステム、クロスプラットフォームランタイム環境、仮想マシン、又はそれらの1つ以上の組み合わせを構成するコードを含むことができる。装置及び実行環境は、ウェブサービス、分散コンピューティング、及びグリッドコンピューティングインフラストラクチャなどの様々な異なるコンピューティングモデルインフラストラクチャを実現することができる。

【0191】

コンピュータプログラム（プログラム、ソフトウェア、ソフトウェアアプリケーション、スクリプト、又はコードとしても知られている）は、コンパイル型言語又はインタプリタ型言語、宣言型言語又は手続き型言語を含む、任意の形式のプログラミング言語で記述することができる。コンピュータプログラムは、ファイルシステム内のファイルに対応してもよいが、対応しなくてもよい。プログラムは、他のプログラム又はデータ（例えば、マークアップ言語文書に記憶された1つ以上のスクリプト）を保持するファイルの一部、問題のプログラム専用の単一のファイル、又は複数の協調ファイル（例えば、1つ以上のモジュール、サブプログラム、又はコードの一部を記憶するファイル）に記憶することができる。コンピュータプログラムは、1つのコンピュータ上で、又は1つのサイトに配置された、又は複数のサイトに分散され、通信ネットワークによって相互接続された複数のコンピュータ上で実行するために展開することができる。

10

【0192】

本明細書で説明されるプロセス及びロジックフローのいくつかは、1つ以上のコンピュータプログラムを実行する1つ以上のプログラマブルプロセッサによって実行され、入力データを操作して出力を生成することによって動作を実行することができる。プロセス及びロジックフローはまた、例えばFPGA（フィールドプログラマブルゲートアレイ）又はASIC（特定用途向け集積回路）などの専用ロジック回路によって実行することができる。装置は専用ロジック回路として実装することができる。

【0193】

コンピュータプログラムの実行に適したプロセッサとしては、例えば、汎用マイクロプロセッサ及び専用マイクロプロセッサの両方、並びに任意の種類のデジタルコンピュータのプロセッサが挙げられる。概して、プロセッサは、リードオンリーメモリ若しくはランダムアクセスメモリ又はその両方から命令及びデータを受信するであろう。コンピュータは、命令に従って動作を実行するためのプロセッサと、命令及びデータを記憶するための1つ以上のメモリデバイスとを含む。コンピュータはまた、データを記憶するための1つ以上の大容量記憶デバイス、例えば、磁気、光磁気ディスク、又は光ディスクを含むか、それらからデータを受信するか、それらにデータを転送するか、それらの両方を行うように動作可能に結合することができる。しかしながら、コンピュータは、そのようなデバイスを有する必要はない。コンピュータプログラム命令及びデータを記憶するのに適した装置は、例として、半導体メモリ装置（例えば、EPROM、EEPROM、フラッシュメモリデバイスなど）、磁気ディスク（例えば、内蔵ハードディスク、リムーバブルディスクなど）、光磁気ディスク、並びにCD-ROM及びDVD-ROMディスクを含む、あらゆる形態の不揮発性メモリ、媒体及びメモリ装置を含む。プロセッサ及びメモリは、専用ロジック回路によって補完され得るか、又は専用ロジック回路に組み込まれ得る。

20

30

【0194】

ユーザとの対話を提供するために、ユーザに情報を表示するためのディスプレイデバイス（例えば、モニタ、又は別の種類のディスプレイデバイス）と、ユーザがコンピュータに入力を提供することができるキーボード及びポインティングデバイス（例えば、マウス、トラックボール、タブレット、タッチセンシティブスクリーン、又は別のタイプのポインティングデバイス）とを有するコンピュータ上で動作を実施することができる。他の種類のデバイスを使用して、ユーザとの対話を提供することもできる。例えば、ユーザに提供されるフィードバックは、視覚フィードバック、聴覚フィードバック、又は触覚フィードバックなどの任意の形態の感覚フィードバックとすることができ、ユーザからの入力は、音響、音声、又は触覚入力を含む任意の形態で受け取ることができる。更に、コンピュータは、ユーザによって使用されるデバイスに文書を送信し、デバイスから文書を受信することによって、例えば、ウェブブラウザから受信した要求に回答して、ユーザのクライアントデバイス上のウェブブラウザにウェブページを送信することによって、ユーザと対話することができる。

40

【0195】

50

コンピュータシステムは、単一のコンピューティングデバイス、又は互いに近接して又は一般に遠隔で動作し、通常は通信ネットワークを介して対話する複数のコンピュータを含むことができる。通信ネットワークの例には、ローカルエリアネットワーク（「LAN」）及びワイドエリアネットワーク（「WAN」）、インターネットワーク（例えば、インターネット）、衛星リンクを含むネットワーク、及びピアツーピアネットワーク（例えば、アドホックピアツーピアネットワーク）が含まれる。クライアントとサーバとの関係は、それぞれのコンピュータ上で実行され、互いにクライアントーサーバ関係を有するコンピュータプログラムによって生じ得る。

【0196】

例示的なコンピュータシステムは、プロセッサ、メモリ、記憶デバイス、及び入出力デバイスを含む。構成要素のそれぞれは、例えば、システムバスによって相互接続され得る。プロセッサは、システム内で実行するための命令を処理することができる。いくつかの実装形態では、プロセッサは、シングルスレッドプロセッサ、マルチフィールドスレッドプロセッサ、又は別のタイプのプロセッサである。プロセッサは、メモリ又は記憶デバイスに記憶された命令を処理することができる。メモリ及び記憶デバイスは、システム内に情報を記憶し得る。

【0197】

入出力デバイスは、システムに入出力動作を提供する。いくつかの実装形態では、入出力デバイスは、ネットワークインタフェースデバイス、例えばイーサネットカード、シリアル通信デバイス、例えばRS-232ポート、及び／又は無線インタフェースデバイス、例えば802.11カード、3G無線モデム、4G無線モデム、5G無線モデムなどのうちの1つ以上を含み得る。いくつかの実装形態では、入出力デバイスは、入力データを受信し、他の入出力デバイス、例えばキーボード、プリンタ及びディスプレイデバイスに出力データを送信するように構成されたドライバデバイスを含み得る。いくつかの実装形態では、モバイルコンピューティングデバイス、モバイル通信デバイス、及び他のデバイスを使用することができる。

【0198】

本明細書は多くの詳細を含むが、これらは特許請求され得るものの範囲に対する限定として解釈されるべきではなく、特定の例に特有の特徴の説明として解釈されるべきである。別個の実施態様の文脈で本明細書に記載されている特定の特徴を組み合わせることもできる。逆に、単一の実施態様の文脈で説明されている様々な特徴は、複数の実施形態において別々に、又は任意の適切な副組み合わせで実施することもできる。

【0199】

いくつかの実施形態が説明された。それにもかかわらず、本明細書に記載のデータ処理システムの趣旨及び範囲から逸脱することなく、様々な修正を行うことができることが理解されよう。したがって、他の実施形態も、以下の特許請求の範囲の範囲内にある。

10

20

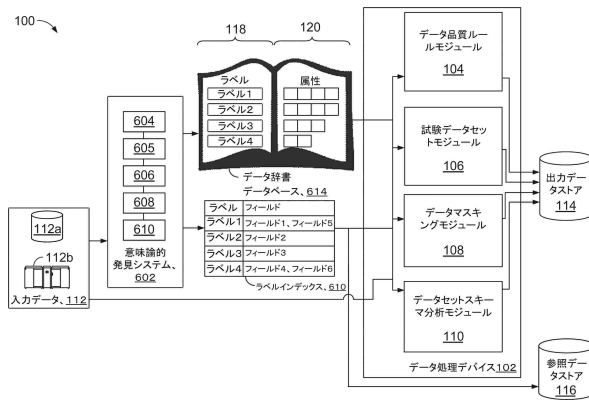
30

40

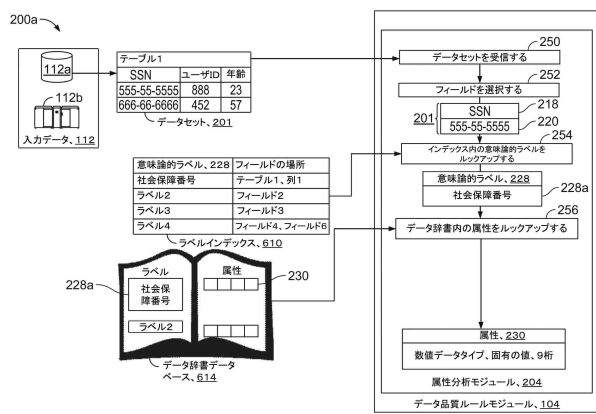
50

【図面】

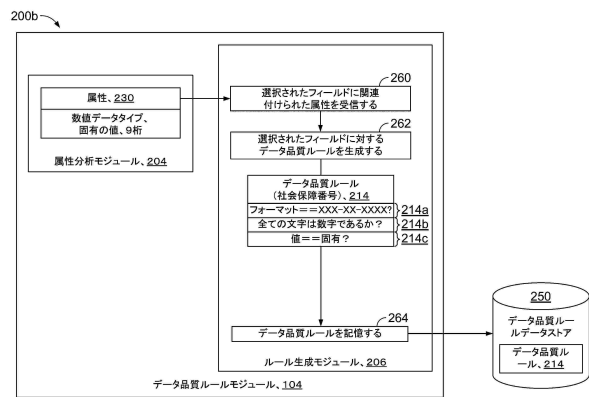
【図 1】



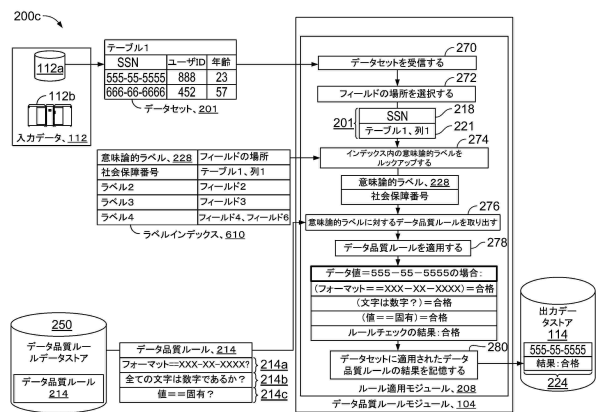
【図 2 A】



【図 2 B】



【図 2 C】



10

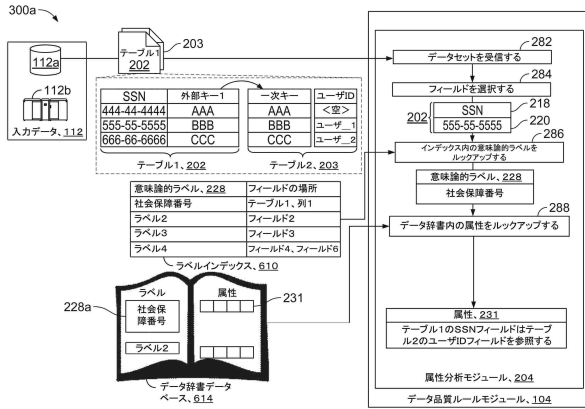
20

30

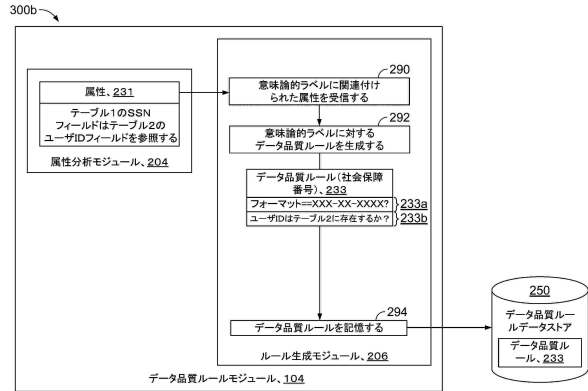
40

50

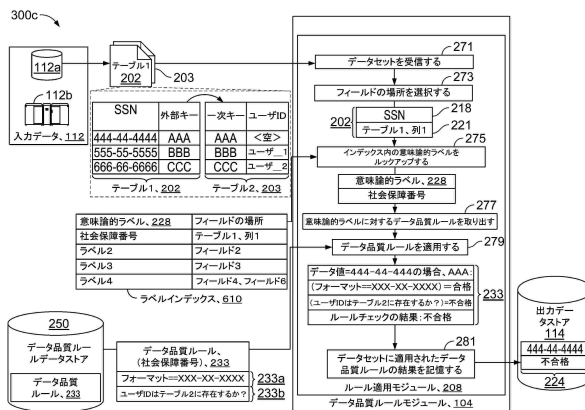
【図 3 A】



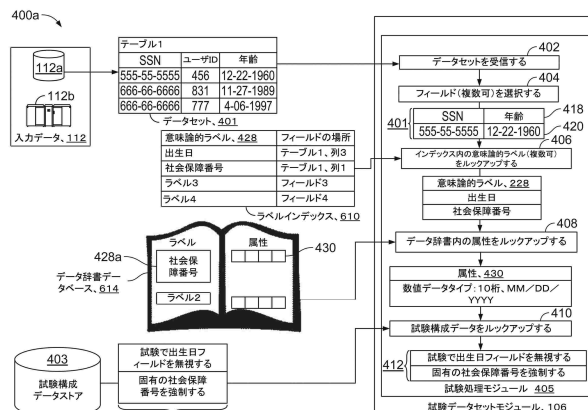
【図 3 B】



【図 3 C】



【図 4 A】



10

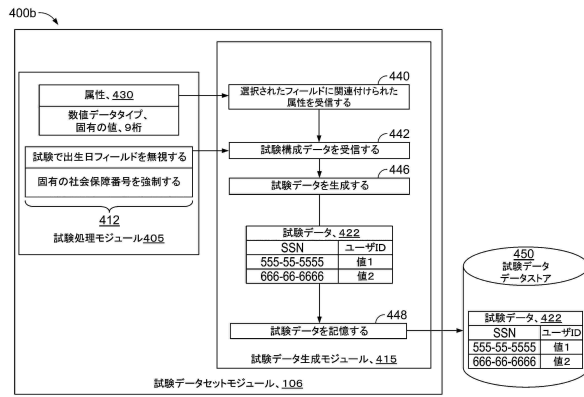
20

30

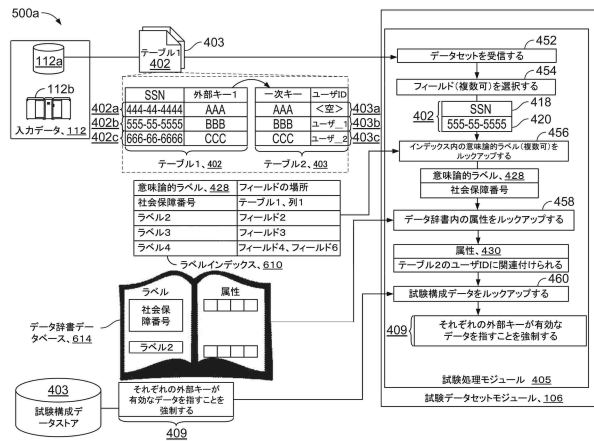
40

50

【図 4 B】

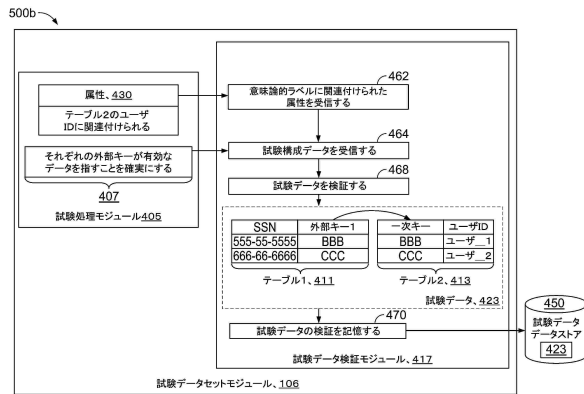


【図 5 A】

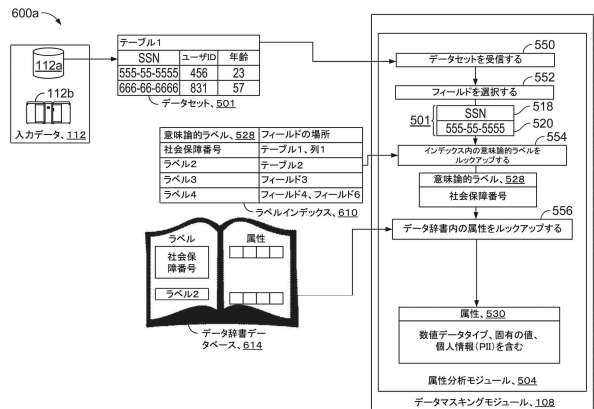


10

【図 5 B】



【図 6 A】



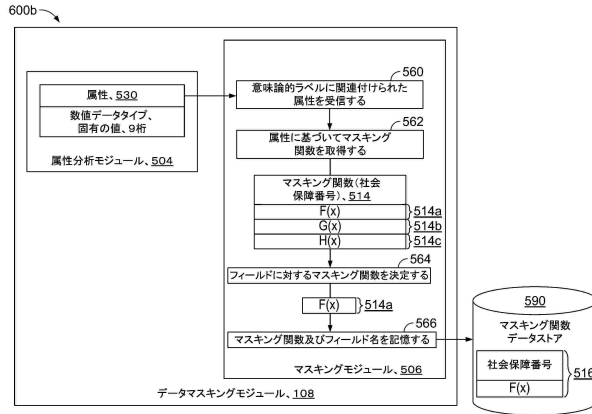
20

30

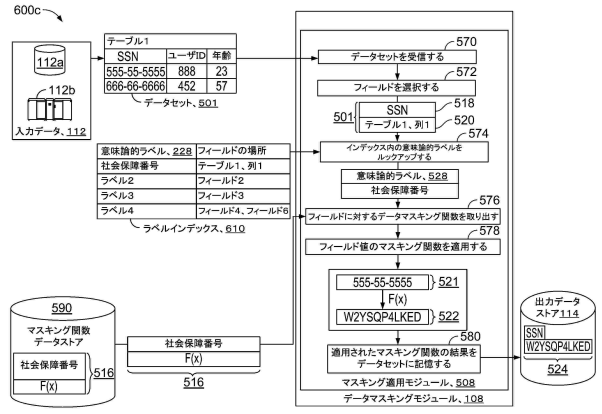
40

50

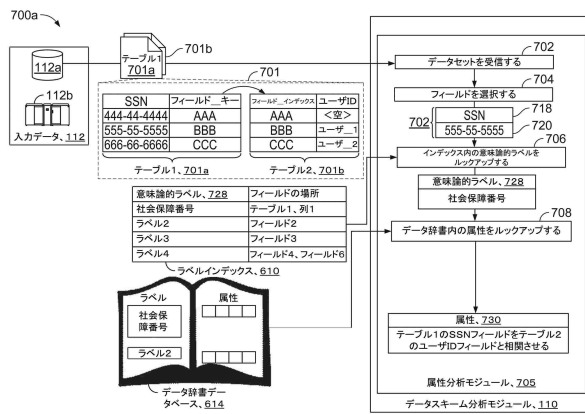
【図 6 B】



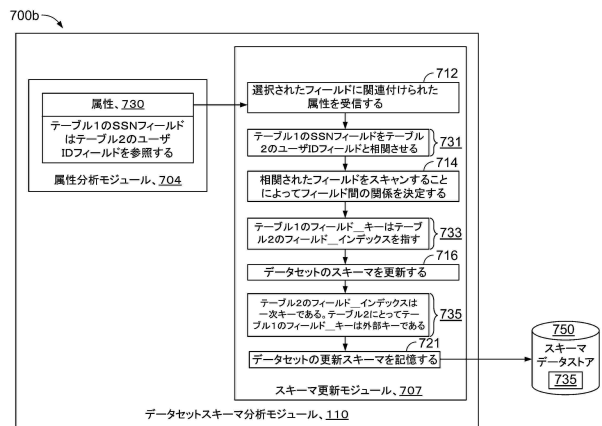
【図 6 C】



【図 7 A】



【図 7 B】



10

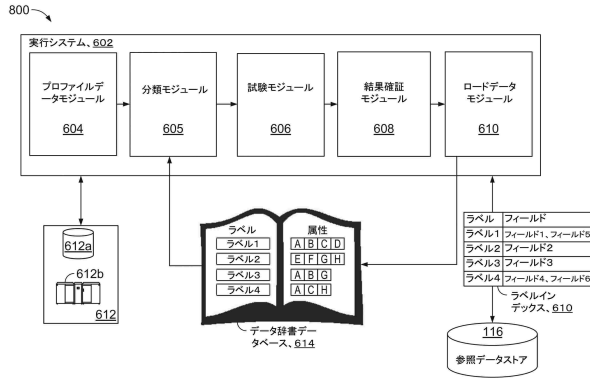
20

30

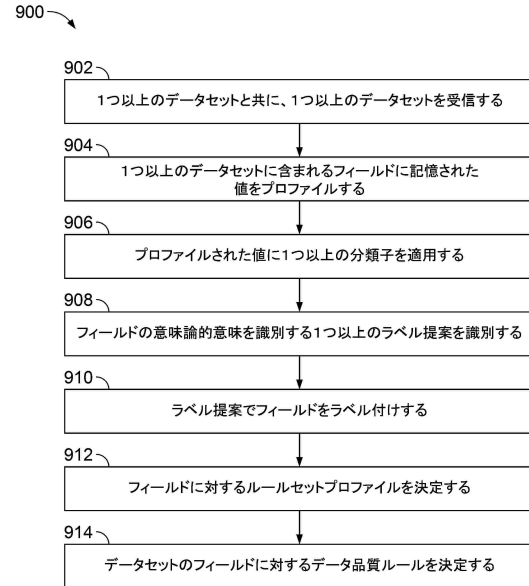
40

50

【図 8】

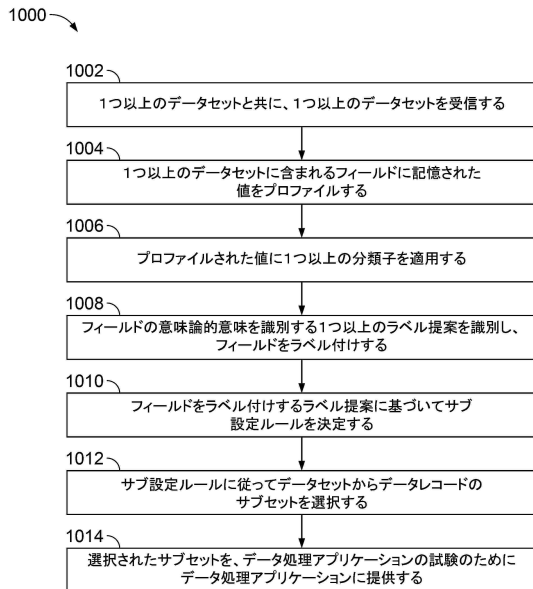


【図 9】

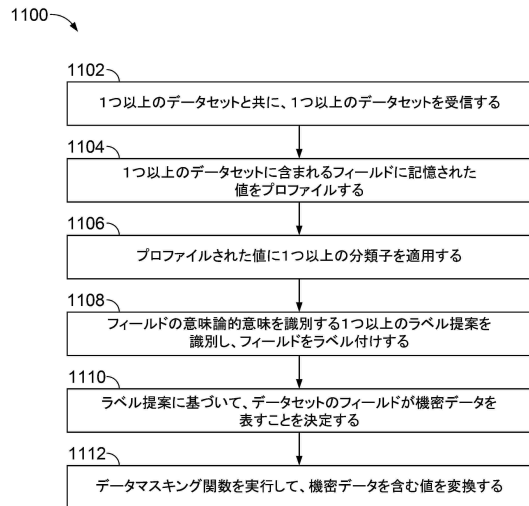


10

【図 10】



【図 11】



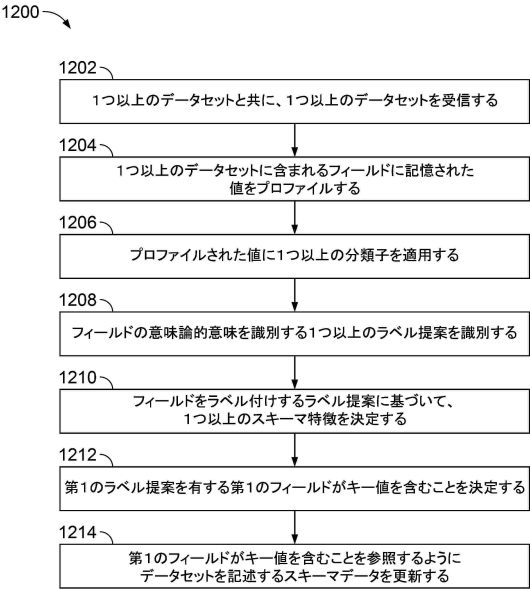
20

30

40

50

【図 12】



10

20

30

40

50

フロントページの続き

(33)優先権主張国・地域又は機関
米国(US)

早期審査対象出願

0 2 4 6 0, ニュートン, ウォルナット ストリート 5 0 3

(72)発明者 イスマン, マーシャル エー.
アメリカ合衆国, マサチューセッツ州 0 2 4 5 8, ニュートン, バレー スプリング ロード 1 1

(72)発明者 メルバウチ, サンドリック
アメリカ合衆国, メリーランド州 2 1 7 7 3 マイヤーズビル イースターデイ ロード 1 0 4 0 9

審査官 甲斐 哲雄

(56)参考文献 米国特許出願公開第 2 0 1 8 / 0 1 0 7 6 9 4 (US, A 1)
米国特許出願公開第 2 0 1 3 / 0 0 7 3 5 9 4 (US, A 1)
米国特許出願公開第 2 0 1 6 / 0 3 6 4 3 2 5 (US, A 1)

(58)調査した分野 (Int.Cl., DB名)
G 0 6 F 1 6 / 0 0 - 1 6 / 9 5 8