(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2017-536601
(P2017-536601A)
(43)公表日　平成29年12月7日(2017.12.7)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	220Z
		G06F	17/30	210D
		G06F	17/30	380Z

審査請求 未請求　予備審査請求 未請求　（全 73 頁）

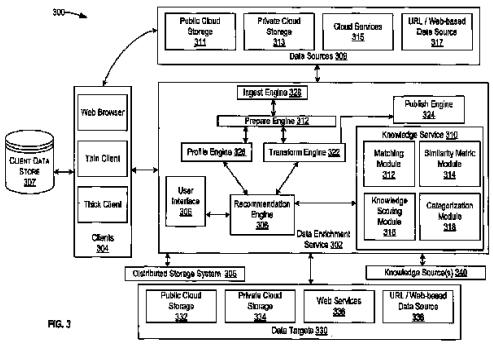
(21)出願番号	特願2017-516310 (P2017-516310)
(86)(22)出願日	平成27年9月25日 (2015.9.25)
(85)翻訳文提出日	平成29年4月27日 (2017.4.27)
(86)国際出願番号	PCT/US2015/052190
(87)国際公開番号	WO2016/049437
(87)国際公開日	平成28年3月31日 (2016.3.31)
(31)優先権主張番号	62/056,468
(32)優先日	平成26年9月26日 (2014.9.26)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	62/163,296
(32)優先日	平成27年5月18日 (2015.5.18)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	62/203,806
(32)優先日	平成27年8月11日 (2015.8.11)
(33)優先権主張国	米国 (US)

(71)出願人　502303739
オラクル・インターナショナル・コーポレイション
アメリカ合衆国カリフォルニア州９４０６５レッドウッド・シティー，オラクル・パークウェイ５００
(74)代理人　110001195
特許業務法人深見特許事務所
(72)発明者　ストジャノビク，アレクサンダー・サシャ
アメリカ合衆国、９５０３０　カリフォルニア州、ロス・ガトス、ウエスト・セントラル・アベニュ、１４

最終頁に続く

(54)【発明の名称】知識ソースを用いた類似性分析およびデータ強化の技術

(57)【要約】
本開示は、知識ソースを用いて類似性メトリック分析とデータ強化を実施することに関する。データ強化サービスは、入力データセットを知識ソースに格納されている参照データセットと比較することによって類似関連データを識別できる。類似性メトリックは、２つ以上のデータセットの意味類似性に対応するように計算してもよい。類似性メトリックを用いることにより、データセットを、これらデータセットのメタデータ属性およびデータ値に基づいて識別することができ、これは、インデックス付けとデータ値の高性能検索を簡単にすることができる。入力データセットに、入力データセットとのベストマッチを示すデータセットに基づくカテゴリでラベル付けできる。入力データセットと知識ソースから提供されたデータセットとの類似性を用いて知識ソースにクエリすることにより、データセットに関する追加情報を取得できる。追加情報を用いてユーザに推薦を提供してもよい。



【特許請求の範囲】
【請求項１】
方法であって、
入力データセットを１つ以上の入力データソースから受けるステップと、
データ強化サービスのコンピューティングシステムによって、前記入力データセットを、参照ソースから取得した１つ以上の参照データセットと比較するステップと、
前記コンピューティングシステムによって、前記１つ以上の参照データセット各々について類似性メトリックを計算するステップとを含み、前記類似性メトリックは、前記入力データセットとの比較における前記１つ以上の参照データセット各々の類似性の程度を示し、
前記コンピューティングシステムによって、前記類似性メトリックに基づいて前記入力データセットと前記１つ以上の参照データセットとの間の一致を識別するステップと、
前記コンピューティングシステムによって、前記１つ以上の参照データセット各々について計算した前記類似性メトリックを示しかつ前記入力データセットと前記１つ以上の参照データセットとの間の前記識別した一致を示すグラフィカルインターフェイスを生成するステップと、
前記グラフィカルインターフェイスを用いて、前記１つ以上の参照データセット各々について計算した前記類似性メトリックを示しかつ前記入力データセットと前記１つ以上の参照データセットとの間の前記識別した一致を示すグラフィカルなビジュアライゼーションをレンダリングするステップとを含む、方法。
【請求項２】
前記１つ以上の参照データセットは、ドメインに対応付けられた用語を含み、前記類似性メトリックは、前記１つ以上の参照データセット各々について計算されたマッチングスコアであり、前記マッチングスコアは、前記参照データセットに関するメトリックを示す第１の値と前記入力データセットと前記参照データセットとの比較に基づくメトリックを示す第２の値とを含む１つ以上の値を用いて計算される、請求項１に記載の方法。
【請求項３】
前記グラフィカルなビジュアライゼーションはレンダリングされることによって前記マッチングスコアの計算に用いられる１つ以上の値を示す、請求項２または３に記載の方法。
【請求項４】
前記１つ以上の値は、前記入力データセットと前記データセットとの間で一致する用語の度数値と、前記データセットの母集団値と、前記入力データセットと前記データセットとの間で一致する異なる用語の数を示す固有マッチング値と、前記データセット内の用語の数を示すドメイン値と、前記データセットのキュレーションの程度を示すキュレーションレベルとを含む、請求項１から４のいずれか一項に記載の方法。
【請求項５】
前記コンピューティングシステムによって、アグリゲーションサービスから取得した増補データに基づいて増補リストを生成するステップと、
前記増補リストに基づいて前記入力データセットを増補するステップとをさらに含み、
前記１つ以上の参照データセットと比較される前記入力データは、前記増補リストに基づいて増補される、請求項１に記載の方法。
【請求項６】
前記方法はさらに、
前記コンピューティングシステムによって、前記１つ以上の参照データセットに基づいてインデックス付トライグラム表を生成するステップを含み、
増補後の前記入力データセットにおけるワードごとに、
前記ワードのトライグラムを作成するステップと、
前記トライグラム各々を前記インデックス付トライグラム表と比較するステップと、
前記トライグラムのうちの第１のトライグラムと一致する、トライグラムに対応付け

られた前記インデックス付トライグラム表におけるワードを識別するステップと、
　前記ワードをトライグラム増補データセットに格納するステップとを含み、
　前記トライグラム増補データセットを前記１つ以上の参照データセットと比較するステップと、
　前記比較に基づいて前記トライグラム増補データセットと前記１つ以上の参照データセットとの間の一致を判断するステップとを含み、
　前記入力データセットと前記１つ以上の参照データセットとの間の一致を識別するステップは、前記比較に基づく前記トライグラム増補データセットと前記１つ以上の参照データセットとの間の一致を用いて実行される、請求項５に記載の方法。
【請求項７】
　前記１つ以上の参照データセットの少なくとも一部を表わすデータ構造を生成するステップをさらに含み、前記データ構造における各ノードは、前記１つ以上の参照データセットから抽出された１つ以上のストリングの中の異なる文字を表わし、
　前記入力データセットは、前記データ構造をトラバースすることによって前記１つ以上の参照データセットと比較される、請求項１から６のいずれか一項に記載の方法。
【請求項８】
　前記類似性メトリックは、前記入力データセットとの比較における前記１つ以上の参照データセットの共通部分のカーディナリティに基づく値として計算され、
　前記値は前記カーディナリティによって正規化され、
　前記値は、前記１つ以上の参照データセットのサイズに基づく第１のファクタだけ減じられ、前記値は、前記１つ以上の参照データセットのタイプに基づく第２のファクタだけ減じられる、請求項７に記載の方法。
【請求項９】
　前記類似性メトリックは、前記１つ以上の参照データセットのうちの各参照データセットについて、前記入力データセットと前記参照データセットとの間のコサイン類似度を求めることによって計算される、請求項１から８のいずれか一項に記載の方法。
【請求項１０】
　前記一致を識別するステップは、前記１つ以上の参照データセットのうち、前記１つ以上の参照データセット各々について計算した前記類似性メトリックに基づく類似性の程度が最大である参照データを求めるステップを含む、請求項１から９のいずれか一項に記載の方法。
【請求項１１】
　前記入力データセットは１つ以上のデータ列にフォーマットされる、請求項１から１０のいずれか一項に記載の方法。
【請求項１２】
　データ強化システムであって、
　複数の入力データソースと、
　クラウドコンピューティングインフラストラクチャシステムとを備え、前記クラウドコンピューティングインフラストラクチャシステムは、
　少なくとも１つの通信ネットワークを通して前記複数の入力データソースに通信可能に結合されかつ複数のデータターゲットに通信可能に結合された１つ以上のプロセッサと、
　前記１つ以上のプロセッサに結合されたメモリとを含み、前記メモリは、データ強化サービスを提供することを指示する命令を格納し、前記命令は、前記１つ以上のプロセッサによって実行されたときに、前記１つ以上のプロセッサに、
　入力データセットを前記複数の入力データソースのうちの１つ以上の入力データソースから受けることと、
　前記入力データセットを、参照ソースから取得した１つ以上の参照データセットと比較することと、
　前記１つ以上の参照データセット各々について類似性メトリックを計算することとを

実行させ、前記類似性メトリックは、前記入力データセットとの比較における前記１つ以上の参照データセット各々の類似性の程度を示し、
　前記類似性メトリックに基づいて前記入力データセットと前記１つ以上の参照データセットとの間の一致を識別することと、
　前記１つ以上の参照データセット各々について計算した前記類似性メトリックを示しかつ前記入力データセットと前記１つ以上の参照データセットとの間の前記識別した一致を示すグラフィカルインターフェイスを生成することと、
　前記１つ以上の参照データセット各々について計算した前記類似性メトリックを示しかつ前記入力データセットと前記１つ以上の参照データセットとの間の前記識別した一致を示すグラフィカルなビジュアライゼーションをレンダリングすることとを実行させる、データ強化システム。
【請求項１３】
　前記１つ以上の参照データセットは、ドメインに対応付けられた用語を含み、前記類似性メトリックは、前記１つ以上の参照データセット各々について計算されたマッチングスコアであり、前記マッチングスコアは、前記参照データセットに関するメトリックを示す第１の値と前記入力データセットと前記参照データセットとの比較に基づくメトリックを示す第２の値とを含む１つ以上の値を用いて計算され、前記グラフィカルなビジュアライゼーションはレンダリングされることによって前記マッチングスコアの計算に用いられる１つ以上の値を示す、請求項１２に記載のデータ強化システム。
【請求項１４】
　前記１つ以上の値は、前記入力データセットと前記データセットとの間で一致する用語の度数値と、前記データセットの母集団値と、前記入力データセットと前記データセットとの間で一致する異なる用語の数を示す固有マッチング値と、前記データセット内の用語の数を示すドメイン値と、前記データセットのキュレーションの程度を示すキュレーションレベルとを含む、請求項１３に記載のデータ強化システム。
【請求項１５】
　前記命令はさらに、前記１つ以上のプロセッサによって実行されたときに、前記１つ以上のプロセッサに、
　アグリゲーションサービスから取得した増補データに基づいて増補リストを生成することと、
　前記増補リストに基づいて前記入力データセットを増補することと、
　前記１つ以上の参照データセットに基づいてインデックス付トライグラム表を生成することとを実行させ、
　増補後の前記入力データセットにおけるワードごとに、
　　前記ワードのトライグラムを作成することと、
　　前記トライグラム各々を前記インデックス付トライグラム表と比較することと、
　　前記トライグラムのうちの第１のトライグラムと一致する、トライグラムに対応付けられた前記インデックス付トライグラム表におけるワードを識別することと、
　　前記ワードをトライグラム増補データセットに格納することとを実行させ、
　前記トライグラム増補データセットを前記１つ以上の参照データセットと比較することと、
　前記比較に基づいて前記トライグラム増補データセットと前記１つ以上の参照データセットとの間の一致を判断することとを実行させ、
　前記１つ以上の参照データセットと比較される前記入力データは、前記増補リストに基づいて増補され、
　前記入力データセットと前記１つ以上の参照データセットとの間の一致を識別することは、前記比較に基づく前記トライグラム増補データセットと前記１つ以上の参照データセットとの間の一致を用いて実行される、請求項１２から１４のいずれか一項に記載のデータ強化システム。
【請求項１６】

非一時的なコンピュータ可読記憶媒体であって、前記非一時的なコンピュータ可読記憶媒体に格納された命令を含み、前記命令は、１つ以上のプロセッサによって実行されたときに、前記１つ以上のプロセッサに、
入力データセットを１つ以上の入力データソースから受けることと、
データ強化サービスのコンピューティングシステムによって、前記入力データセットを、参照ソースから取得した１つ以上の参照データセットと比較することと、
前記コンピューティングシステムによって、前記１つ以上の参照データセット各々について類似性メトリックを計算することとを実行させ、前記類似性メトリックは、前記入力データセットとの比較における前記１つ以上の参照データセット各々の類似性の程度を示し、
前記コンピューティングシステムによって、前記類似性メトリックに基づいて前記入力データセットと前記１つ以上の参照データセットとの間の一致を識別することと、
前記コンピューティングシステムによって、前記１つ以上の参照データセット各々について計算した前記類似性メトリックを示しかつ前記入力データセットと前記１つ以上の参照データセットとの間の前記識別した一致を示すグラフィカルインターフェイスを生成することと、
前記グラフィカルインターフェイスを用いて、前記１つ以上の参照データセット各々について計算した前記類似性メトリックを示しかつ前記入力データセットと前記１つ以上の参照データセットとの間の前記識別した一致を示すグラフィカルなビジュアライゼーションをレンダリングすることとを実行させる、非一時的なコンピュータ可読記憶媒体。

【請求項１７】
方法であって、
入力データセットを１つ以上の入力データソースから受けるステップと、
データ強化サービスのコンピューティングシステムによって、前記入力データセットを、参照ソースから取得した１つ以上の参照データセットと比較するステップと、
前記コンピューティングシステムによって、前記１つ以上の参照データセット各々について類似性メトリックを計算するステップとを含み、前記類似性メトリックは、前記入力データセットとの比較における前記１つ以上の参照データセット各々の類似性の程度を示し、
前記コンピューティングシステムによって、前記類似性メトリックに基づいて前記入力データセットと前記１つ以上の参照データセットとの間の一致を識別するステップと、
前記入力データセットをマッチング情報とともに格納するステップとを含み、前記マッチング情報は、前記１つ以上の参照データセット各々について計算した類似性メトリックを示しかつ前記入力データセットと前記１つ以上の参照データセットとの間の前記識別した一致を示す、方法。

【請求項１８】
前記入力データセットと前記１つ以上の参照データセットとの間の一致の識別に基づいて、前記入力データセットのカテゴリラベルを識別するステップと、
前記カテゴリラベルに対応付けて前記入力データセットを格納するステップとをさらに含む、請求項１７に記載の方法。

【請求項１９】
前記類似性メトリックは、Jaccard係数、Tversky係数、またはDice-Sorensen係数のうちの１つ以上を用いて計算される、請求項１７または１８に記載の方法。

【請求項２０】
前記入力データセットは、グラフマッチングまたは意味類似性マッチングのうちの１つ以上を用いて、前記１つ以上の参照データセットと比較される、請求項１７から１９のいずれか一項に記載の方法。

【請求項２１】
データ強化システムであって、
複数の入力データソースと、

クラウドコンピューティングインフラストラクチャシステムとを備え、前記クラウドコンピューティングインフラストラクチャシステムは、

少なくとも１つの通信ネットワークを通して前記複数の入力データソースに通信可能に結合されかつ複数のデータターゲットに通信可能に結合された１つ以上のプロセッサと、

前記１つ以上のプロセッサに結合されたメモリとを含み、前記メモリは、データ強化サービスを提供することを指示する命令を格納し、前記命令は、前記１つ以上のプロセッサによって実行されたときに、前記１つ以上のプロセッサに、

入力データセットを１つ以上の入力データソースから受けることと、

前記入力データセットを、参照ソースから取得した１つ以上の参照データセットと比較することと、

前記１つ以上の参照データセット各々について類似性メトリックを計算することとを実行させ、前記類似性メトリックは、前記入力データセットとの比較における前記１つ以上の参照データセット各々の類似性の程度を示し、

前記類似性メトリックに基づいて前記入力データセットと前記１つ以上の参照データセットとの間の一致を識別することと、

前記入力データセットをマッチング情報とともに格納することとを実行させ、前記マッチング情報は、前記１つ以上の参照データセット各々について計算した類似性メトリックを示しかつ前記入力データセットと前記１つ以上の参照データセットとの間の前記識別した一致を示す、データ強化システム。

【請求項２２】

前記命令はさらに、前記１つ以上のプロセッサによって実行されたときに、前記１つ以上のプロセッサに、

前記入力データセットと前記１つ以上の参照データセットとの間の一致の識別に基づいて、前記入力データセットのカテゴリラベルを識別することと、

前記カテゴリラベルに対応付けて前記入力データセットを格納することとを実行させる、請求項２１に記載のデータ強化システム。

【請求項２３】

前記類似性メトリックは、Jaccard係数、Tversky係数、またはDice-Sorensen係数のうちの１つ以上を用いて計算される、請求項２１または２２に記載のデータ強化システム。

【請求項２４】

前記入力データセットは、グラフマッチングまたは意味類似性マッチングのうちの１つ以上を用いて、前記１つ以上の参照データセットと比較される、請求項２１または２３に記載のデータ強化システム。

【発明の詳細な説明】

【技術分野】

【０００１】

関連出願の相互参照

本願は、２０１５年９月２４日に出願され「TECHNIQUES FOR SIMILARITY ANALYSIS AND DATA ENRICHMENT USING KNOWLEDGE SOURCES」と題され以下の出願に基づく利益および優先権を主張する米国非仮特許出願第１４／８６４，４８５号に基づく利益および優先権を主張する。

【０００２】

１）２０１４年９月２６日に出願され「METHOD FOR SEMANTIC ENTITY EXTRACTION BASED ON GRAPH MATCHING WITH AN EXTERNAL KNOWLEDGEBASE AND SIMILARITY RANKING OF DATASET METADATA FOR SEMANTIC INDEXING, SEARCH, AND RETRIEVAL」と題された米国仮出願第６２／０５６，４６８号

２）２０１５年５月１８日に出願され「CATEGORY LABELING」と題された米国仮出願第６２／１６３，２９６号

３）２０１５年８月１１日に出願され「SIMILARITY METRIC ANALYSIS AND KNOWLEDGE S

CORING SYSTEM」と題された米国仮出願第６２／２０３，８０６号
本願は以下の出願に関連する。
【０００３】
１）２０１４年９月２６日に出願され「DECLARATIVE LANGUAGE AND VISUALIZATION SYSTEM FOR RECOMMENDED DATA TRANSFORMATIONS AND REPAIRS」と題された米国仮出願第６２／０５６，４７１号
２）２０１４年９月２６日に出願され「DYNAMIC VISUAL PROFILING AND VISUALIZATION OF HIGH VOLUME DATASETS AND REAL-TIME SMART SAMPLING AND STATISTICAL PROFILING OF EXTREMELY LARGE DATASETS」と題された米国仮出願第６２／０５６，４７４号
３）２０１４年９月２６日に出願され「AUTOMATED ENTITY CORRELATION AND CLASSIFICATION ACROSS HETEROGENEOUS DATASETS」と題された米国仮出願第６２／０５６，４７５号
４）２０１４年９月２６日に出願され「DECLARATIVE EXTERNAL DATA SOURCE IMPORTATION, EXPORTATION, AND METADATA REFLECTION UTILIZING HTTP AND HDFS PROTOCOLS」と題された米国仮出願第６２／０５６，４７６号
上記特許出願の内容全体を、すべての目的のために本明細書に引用により援用する。
【０００４】
背景
本開示は概してデータの準備および分析に関する。より具体的には、知識ソースを用いて類似性メトリック分析およびデータ強化を実行する技術が開示される。
【背景技術】
【０００５】
「ビッグデータ」システムがデータを分析して有用な結果を提供できるようになる以前は、データをビッグデータシステムに追加し、分析できるようにフォーマットする必要があった。このデータオンボーディング（data onboarding）は、現在のクラウドおよび「ビッグデータ」システムに対し、ある課題を示している。典型的に、ビッグデータシステムに追加されるデータはノイズが多い（たとえばデータが不正確にフォーマットされている、間違っている、失効している、重複を含む等）。データを分析するとき（たとえば報告、予測モデリング等のために）、データの信号対雑音比が不十分であるということは、結果が有用でないことを意味する。その結果、現在のソリューションは、データおよび／または分析結果をクリーンにしキュレートするために実質的にマニュアルのプロセスを必要とする。しかしながら、これらのマニュアルプロセスはスケーリングできない。追加し分析するデータの量が増すと、マニュアルプロセスは実装が不可能になる。
【０００６】
ビッグデータシステムを実装してデータを分析することにより、その他の類似関連データを識別することがある。データの処理量が課題となる。さらに、分析対象のデータの構造次第でまたはこの構造の欠落次第で、分析対象のデータは、データの内容と関係の判断における一層大きな課題を課すかもしれない。
【０００７】
機械学習を実装してデータを分析することがある。たとえば、教師なしの機械学習をデータ分析ツール（たとえばWord2Vec）を用いて実装してデータ間の類似性を判断することがある。しかしながら、教師なしの機械学習は、関連性が高いデータに対応するグループまたはカテゴリを示す情報を提供できない場合がある。このため、教師なし学習は、関連性が高い一組の種（species）（たとえば用語）の属（genus）またはカテゴリを判断できない場合がある。一方、キュレートされた知識ソース（たとえばMax Planck Institute for InformaticsのＹＡＧＯ）に基づく教師ありの機械学習は、データのグループまたはカテゴリの判断においてより優れた結果を提供し得る。教師あり学習は、矛盾するおよび／または不完全な結果をもたらす場合がある。キュレートされた知識ソースが提供するデータは疎データである場合があり品質はキュレータによって異なり得る。教師あり学習の使用に基づいて識別されたカテゴリは、類似関連データの正しいカテゴリ分類を提供しない

場合がある。複数の知識ソースが異なるカテゴリ分類を実装しておりそのために複数の知識ソースを一体化するのが困難になる場合がある。データを分析して類似性と関係とを判断することは、分析対象のデータにおける用語のスペルミスが原因で、負担になる場合がある。データにスペルミスが含まれていると、類似するデータを簡単に識別できない場合がある。

【発明の概要】

【課題を解決するための手段】

【０００８】

本発明の特定の実施形態は、上記およびその他の課題に取り組んでいる。

簡単な概要

本開示は、概してデータの準備および分析に関する。より具体的には、知識ソースを用いて類似性メトリック分析およびデータ強化（data enrichment）を実行する技術が開示される。

【０００９】

本開示は、概してデータ強化サービスに関し、このサービスは、データセットを抽出、修復、および強化することにより、後のインデックス作成およびクラスタ化のための、より精密なエンティティのレゾリューションおよび相関を得る。データ強化サービスは、異種のデータセットの大規模なデータの準備、修復、および強化を実行するための視覚推薦エンジンおよび言語を含み得る。これにより、ユーザは、推薦された強化（たとえば変換および修復）がどのようにユーザのデータに影響しどのように調整を必要に応じて実行するかを、選択し確認することができる。データ強化サービスは、ユーザインターフェイスを通してユーザからのフィードバックを受けることができ、かつ、ユーザからのフィードバックに基づいて推薦をフィルタリングすることができる。いくつかの実施形態において、データ強化サービスは、データセットを分析することにより、データにおけるパターンを識別することができる。

【００１０】

いくつかの実施形態において、データ強化サービスは、入力データセットを、知識ソースに格納されている参照データセットと比較することにより、類似関連データを識別することができる。教師ありトレーニング（たとえば機械学習）なしで入力データと参照データセットとのマッチングを実行することができ、エンドユーザからの適応フィードバックを介して徐々に抽出精度を改善することができる。いくつかの実施形態において、２つ以上のデータセットの意味類似性に対応する類似性メトリックを計算することができる。類似性メトリックを用いることにより、データセットを、そのメタデータ属性とデータ値に基づいて、識別することができる。これは、インデックス作成とデータ値の高性能検索をより簡単にすることができる。

【００１１】

上記のように、特に分析対象のデータの構造次第でまたはこの構造の欠落次第で、データの処理量は課題になる。参照データのキュレーションにおけるスペルミスおよび相違が、カテゴリ分類の誤りにつながり、それが原因で、類似するまたは関連するデータの識別が困難になる。本明細書に記載の技術は、より洗練された類似性メトリックを提供し、これは、入力データセットに対して意味類似性を有する関連性が高いデータセットの自動識別を改善することができる。より類似関連するデータセットを識別することにより、入力データセットを、関連するデータセットからのデータを用いて強化してもよい。入力データセットの強化によって、ユーザは、そうでなければ管理が難しい大量のデータを理解し管理することができる。たとえば、ユーザは、あるデータセットが特定のトピックに関連するか否か判断してもよく、関連していれば、このトピックの関連データがあるか否か判断してもよい。いくつかの実施形態において、参照データセットを更新することにより、類似性メトリックに基づいて、入力データとの関係を反映してもよい。このように、参照データセットを、後に他の入力データセットとの類似性の判断において使用するために強化することができる。

【００１２】
いくつかの実施形態において、データ強化サービスは、入力データセットと比較される複数の参照データセット各々に関する類似性メトリックを表示するグラフィカルインターフェイスをレンダリングすることができる。グラフィカルインターフェイスにより、ユーザは、類似性メトリックを示す対象である参照データセットのうちの１つに基づく変換を選択することができる。このように、類似性メトリックにより、ユーザが参照データを選択的に選んでデータソースからのデータセットを強化することが可能になる。
【００１３】
いくつかの実施形態において、本明細書に開示される技術は、データソースから受けたデータの分類をユーザに提示する方法を提供する。この技術は、関連性が高い一組の種（たとえば用語）の属またはカテゴリを判断できない場合がある教師なし機械学習に勝る利点を提供する。この技術はさらに、教師なし機械学習技術を、教師あり機械学習のための複数のソースを併合することと組合わせることによって、より安定した完璧なデータ分類を提供する。このような技術は、用語のキュレーションおよびスペルミスまたはカテゴリ分類誤りの、レベルの違いを考慮することができる。
【００１４】
いくつかの実施形態において、データ強化サービスは、入力データセットにおける用語を、知識ソースからのデータセットにおける用語と比較することにより、類似性メトリックを求めることができる。類似性メトリックは、本明細書に開示されるさまざまな技術を用いて計算し得る。類似性メトリックはスコアとして表わしてもよい。入力データセットを、各々がカテゴリ（たとえばドメイン）に対応付けられていてもよい複数のデータセットと比較してもよい。類似性メトリックは、各データセットを入力データセットと比較するために計算してもよい。よって、類似性メトリックは、より高い一致度を類似性メトリックの値に基づいて識別できるように、一致度を示してもよい（たとえば最高の類似度を最大の値によって示す）。本明細書に開示される技術のうちの１つ以上を用いて求められた類似性メトリックは、入力データセットと知識ソースから提供されたデータセットとのマッチングについて、より高い確度を提供し得る。
【００１５】
データ強化サービスは、いくつかの異なる技術を実装することにより、入力データセットと１つ以上のデータセットとの類似性を判断してもよい。入力データセットを、この入力データセットとの一致が最大（たとえば類似度が最大）であるデータセットに対応するカテゴリ（たとえばドメイン）に対応付けるまたはこのカテゴリでラベル付けすることができる。よって、入力データセットをカテゴリ名を用いて修正または強化することができる。カテゴリ名により、ユーザは入力データセットをより上手く識別できる。少なくとも１つの実施形態において、データ強化サービスは、教師なし学習技術を教師あり学習技術と組合わせることにより、一層精密に入力データのカテゴリをラベル付けすることができる。知識ソースから与えられたデータセットに対する入力データの類似度を用いることにより、知識ソースに問合せて当該データセットに関する追加情報を取得することができる。追加情報を用いて推薦をユーザに提供することができる。
【００１６】
いくつかの実施形態において、コンピューティングシステムを、知識ソースから与えられたデータセットとの比較におけるデータの類似性メトリック分析を実行するために実装してもよい。コンピューティングシステムは、データ強化サービスを実装し得る。コンピューティングシステムは、本明細書に記載の方法およびオペレーションを実装するように構成されてもよい。このシステムは、複数の入力データソースと複数のデーターゲットとを含み得る。このシステムは、少なくとも１つの通信ネットワークを通して複数の入力データソースに通信可能に結合されかつ複数のデーターゲットに通信可能に結合された１つ以上のプロセッサを備えるクラウドコンピューティングインフラストラクチャシステムを含み得る。クラウドコンピューティングインフラストラクチャシステムは、上記１つ以上のプロセッサに結合されたメモリを含み得る。メモリは、データ強化サービスを提供

することを指示する命令を含み、この命令が上記１つ以上のプロセッサによって実行されると、本明細書に記載の１つ以上の方法またはオペレーションが上記１つ以上のプロセッサによって実行される。さらに他の実施形態は、システムと、機械読取可能な有形の記憶媒体とに関し、これは、本明細書に記載の方法およびオペレーションのための命令を用いるまたは格納する。

【００１７】

少なくとも１つの実施形態において、方法は、入力データセットを１つ以上の入力データソースから受けるステップを含む。入力データセットを、１つ以上のデータ列にフォーマットしてもよい。この方法は、入力データセットを、参照ソースから取得した１つ以上の参照データセットと比較するステップを含み得る。参照ソースは、知識サービスから提供される知識ソースであってもよい。入力データセットは、グラフマッチングまたは意味類似性マッチングのうちの１つ以上を用いて、１つ以上の参照データセットと比較されてもよい。この方法は、上記１つ以上の参照データセット各々について類似性メトリックを計算するステップを含み得る。類似性メトリックは、入力データセットとの比較における１つ以上の参照データセット各々の類似性の程度を示す。この方法は、類似性メトリックに基づいて入力データセットと１つ以上の参照データセットとの間の一致を識別するステップを含み得る。いくつかの実施形態において、この方法は、上記１つ以上の参照データセット各々について計算した類似性メトリックを示しかつ上記入力データセットと上記１つ以上の参照データセットとの間の識別した一致を示すグラフィカルインターフェイスを生成するステップと、上記１つ以上の参照データセット各々について計算した類似性メトリックを示しかつ上記入力データセットと上記１つ以上の参照データセットとの間の識別した一致を示す、グラフィカルなビジュアライゼーションを、グラフィカルインターフェイスを用いてレンダリングすることとを含み得る。いくつかの実施形態において、この方法は、上記１つ以上の参照データセット各々について計算した類似性メトリックを示しかつ上記入力データセットと上記１つ以上の参照データセットとの間の識別した一致を示すマッチング情報とともに、入力データセットを格納することと、上記入力データセットと上記１つ以上の参照データセットとの間の一致の識別に基づいて、上記入力データセットのカテゴリラベルを識別することと、上記カテゴリラベルに対応付けて上記入力データセットを格納することとを含み得る。

【００１８】

いくつかの実施形態において、上記１つ以上の参照データセットは、ドメインに対応付けられた用語を含む。類似性メトリックは、上記１つ以上の参照データセット各々について計算されたマッチングスコアであってもよい。マッチングスコアは、参照データセットに関するメトリックを示す第１の値と入力データセットと参照データセットとの比較に基づくメトリックを示す第２の値とを含む１つ以上の値を用いて計算されてもよい。グラフィカルなビジュアライゼーションはレンダリングされることによってマッチングスコアの計算に用いられる上記１つ以上の値を示してもよい。上記１つ以上の値は、入力データセットとデータセットとの間で一致する用語の度数値と、データセットの母集団値と、入力データセットとデータセットとの間で一致する異なる用語の数を示す固有マッチング値と、データセット内の用語の数を示すドメイン値と、データセットのキュレーションの程度を示すキュレーションレベルとを含み得る。

【００１９】

いくつかの実施形態において、この方法はさらに、アグリゲーションサービスから取得した増補（augmentation）データに基づいて増補リストを生成するステップと、増補リストに基づいて入力データセットを増補するステップとをさらに含み得る。１つ以上の参照データセットと比較される入力データは、増補リストに基づいて増補されてもよい。この方法はさらに、上記１つ以上の参照データセットに基づいてインデックス付トライグラム表を生成するステップを含み得る。この方法は、増補後の入力データセットにおけるワードごとに、このワードのトライグラムを作成するステップと、トライグラム各々をインデックス付トライグラム表と比較するステップと、上記トライグラムのうちの第１のトライ

グラムと一致する、トライグラムに対応付けられたインデックス付トライグラム表におけるワードを識別するステップと、このワードをトライグラム増補データセットに格納するステップとを含み得る。この方法は、上記トライグラム増補データセットを１つ以上の参照データセットと比較するステップと、この比較に基づいてトライグラム増補データセットと１つ以上の参照データセットとの間の一致を判断するステップとを含み得る。上記入力データセットと１つ以上の参照データセットとの間の一致を識別するステップは、上記比較に基づくトライグラム増補データセットと１つ以上の参照データセットとの間の一致を用いて実行されてもよい。

【００２０】

いくつかの実施形態において、この方法は、上記１つ以上の参照データセットの少なくとも一部を表わすデータ構造を生成するステップを含み得る。このデータ構造における各ノードは、上記１つ以上の参照データセットから抽出された１つ以上のストリングの中の異なる文字を表わす。上記入力データセットは、上記データ構造をトラバースすることによって上記１つ以上の参照データセットと比較されてもよい。類似性メトリックは、入力データセットとの比較における上記１つ以上の参照データセットの共通部分のカーディナリティ（cardinality）に基づく値として計算されてもよい。この値はカーディナリティによって正規化されてもよい。この値は、上記１つ以上の参照データセットのサイズに基づく第１のファクタだけ減じられてもよく、この値は、上記１つ以上の参照データセットのタイプに基づく第２のファクタだけ減じられてもよい。

【００２１】

いくつかの実施形態において、上記１つ以上の参照データセットのうちの各参照データセットの類似性メトリックは、上記入力データセットと参照データセットとの間のコサイン類似度を求めることによって計算されてもよい。類似性メトリックは、Jaccard係数、Tversky係数、またはDice-Sorensen係数のうちの１つ以上を用いて計算されてもよい。上記一致を識別するステップは、上記１つ以上の参照データセットのうち、上記１つ以上の参照データセット各々について計算した類似性メトリックに基づく類似性の程度が最大である参照データを求めるステップを含み得る。

【００２２】

これまでに述べたことは、他の特徴および実施形態とともに、以下の明細書、請求項、および添付の図面を参照すれば、より明らかになるであろう。

【図面の簡単な説明】

【００２３】

【図１】本発明の実施形態に従うデータ強化システムの簡略化されたハイレベル図を示す。

【図２】本発明の実施形態に従うテクノロジースタックの簡略化されたブロック図を示す。

【図３】本発明の実施形態に従うデータ強化システムの簡略化されたブロック図を示す。

【図４Ａ】本発明の実施形態に従う対話型データ強化を提供するユーザインターフェイスの一例を示す。

【図４Ｂ】本発明の実施形態に従う対話型データ強化を提供するユーザインターフェイスの一例を示す。

【図４Ｃ】本発明の実施形態に従う対話型データ強化を提供するユーザインターフェイスの一例を示す。

【図４Ｄ】本発明の実施形態に従う対話型データ強化を提供するユーザインターフェイスの一例を示す。

【図５Ａ】本発明の実施形態に従うデータセットのビジュアライゼーションを提供するさまざまなユーザインターフェイスの一例を示す。

【図５Ｂ】本発明の実施形態に従うデータセットのビジュアライゼーションを提供するさまざまなユーザインターフェイスの一例を示す。

【図５Ｃ】本発明の実施形態に従うデータセットのビジュアライゼーションを提供するさ

まざまなユーザインターフェイスの一例を示す。
【図５Ｄ】本発明の実施形態に従うデータセットのビジュアライゼーションを提供するさまざまなユーザインターフェイスの一例を示す。
【図６】本発明の実施形態に従う代表的なグラフを示す。
【図７】本発明の実施形態に従う代表的な状態表を示す。
【図８】本発明の実施形態に従う大文字と小文字を区別しないグラフの例を示す。
【図９】本発明の実施形態に従うデータセットの類似性を示す図を示す。
【図１０】本発明の実施形態に従う異なる知識ドメインの知識スコアリングを表示するグラフィカルインターフェイスの例を示す。
【図１１】本発明の実施形態に従う自動化されたデータ分析の例を示す。
【図１２】本発明の実施形態に従うトライグラムモデリングの一例を示す。
【図１３】本発明の実施形態に従うカテゴリラベル付けの例を示す。
【図１４】本発明の実施形態に従うランク付けされたカテゴリを求めるための類似性分析を示す。
【図１５】本発明の実施形態に従うランク付けされたカテゴリを求めるための類似性分析を示す。
【図１６】本発明の実施形態に従うランク付けされたカテゴリを求めるための類似性分析を示す。
【図１７】本発明の実施形態に従う類似性分析のプロセスのフローチャートを示す。
【図１８】本発明の実施形態に従う類似性分析のプロセスのフローチャートを示す。
【図１９】実施形態を実現するための分散型システムの簡略図を示す。
【図２０】本開示の実施形態に従うクラウドサービスとしてサービスを提供し得るシステム環境の１つ以上のコンポーネントの簡略化されたブロック図である。
【図２１】本発明の実施形態を実現するのに使用し得る典型的なコンピュータシステムを示す。
【発明を実施するための形態】
【００２４】
詳細な説明
以下の記載において、説明のために、具体的な詳細事項を述べることによって本発明の実施形態が十分に理解されるようにする。しかしながら、これらの具体的な詳細事項がなくてもさまざまな実施形態を実施し得ることが明らかであろう。図面および説明は限定を意図したものではない。
【００２５】
本開示は概してデータ強化サービスに関し、このサービスは、データセットを抽出、修復、および強化することにより、後のインデックス作成およびクラスタ化のための、より精密なエンティティのレゾリューションおよび相関を得る。いくつかの実施形態において、データ強化サービスは、データの採集からデータの分析までの多数の段階でデータを処理することによってデータをデータターゲットに対して公開する拡張可能なセマンティックパイプラインを含む。
【００２６】
本発明のある実施形態において、データをデータウェアハウス（またはその他のデータターゲット）にロードする前に、さまざまな処理段を含むパイプライン（本明細書ではセマンティックパイプラインとも呼ぶ）を通して処理する。いくつかの実施形態において、パイプラインは、採集段と、準備段と、プロファイル段と、変換段と、公開段とを含み得る。処理中に、データを分析し、準備し、強化することができる。次に、結果として得られたデータを１つ以上のデータターゲット（たとえばローカルストレージシステム、クラウドベースのストレージサービス、ウェブサービス、データウェアハウス等）に公開する（たとえば下流のプロセスに与える）ことができる。このターゲットにおいて、データに対しさまざまなデータ分析を実行することができる。このデータには修復と強化が行なわれているので、それを分析することによって有用な結果が得られる。したがって、データ

オンボーディングプロセスは自動化されているので、スケーリングすることにより、その量のためにマニュアル処理できない非常に大きなデータセットを処理することができる。

【００２７】

いくつかの実施形態において、データを分析してこのデータからエンティティを抽出することができ、抽出したエンティティに基づいてデータを修復することができる。たとえば、スペルミス、アドレスの誤り、およびその他の一般的な間違いは、ビッグデータシステムに対して複雑な問題を示す。データ量が少ない場合はこのような誤りをマニュアルで識別して修正できる。しかしながら、非常に大きなデータセット（たとえば何十億ものノードまたは記録）の場合、このようなマニュアル処理は不可能である。本発明のある実施形態において、データ強化サービスは、知識サービスを用いてデータを分析することができる。知識サービスのコンテンツに基づいて、データ内のエンティティを識別することができる。たとえば、エンティティは、住所、事業所名、場所、個人名、ＩＤ番号等であってもよい。

【００２８】

図１は、本発明の実施形態に従うデータ強化サービスの簡略化されたハイレベル図１００を示す。図１に示されるように、クラウドベースのデータ強化サービス１０２は、さまざまなデータソース１０４からデータを受信することができる。いくつかの実施形態において、クライアントはデータ強化要求をデータ強化サービス１０２に対して出すことができ、データ強化サービス１０２はデータソース１０４のうちの１つ以上（またはその一部、たとえば特定の表、データセット等）を識別する。次に、データ強化サービス１０２は、識別したデータソース１０４からのデータの処理を要求してもよい。いくつかの実施形態において、データソースはサンプリングされてもよく、サンプリングされたデータは強化のために分析され、それによって大きなデータセットはより扱い易くなる。識別されたデータを受け、データ強化サービスからアクセス可能な分散記憶システム（Hadoop分散記憶（ＨＤＦＳ）システム等）に追加することができる。データは、多数の処理段（本明細書においてパイプラインまたはセマンティックパイプラインとして説明）によって意味論的に処理してもよい。これらの処理段は、準備段１０８と、強化段１１０と、公開段１１２とを含み得る。いくつかの実施形態において、データを、データ強化サービスによって１つ以上のバッチで処理することができる。いくつかの実施形態において、データを受信しながら処理するストリーミングパイプラインを提供することができる。

【００２９】

いくつかの実施形態において、準備段１０８はさまざまな処理サブ段を含み得る。これは、自動的にデータソースフォーマットを検出しコンテンツの抽出および／または修復を実行することを含み得る。データソースフォーマットが検出されると、自動的に、データソースをデータ強化サービスが処理できるフォーマットに正規化することができる。いくつかの実施形態において、データソースが準備されたら、このデータソースは強化段１１０によって処理することができる。いくつかの実施形態において、インバウンドデータソースは、データ強化サービスからアクセス可能な分散記憶システム１０５（データ強化サービスに通信可能に結合されたＨＤＦＳシステム等）にロードすることができる。分散記憶システム１０５は、採集されたデータファイルのための一時的な記憶空間を提供し、これはまた、中間処理ファイルの、および、公開前の結果の一時記憶域としての記憶域を提供することができる。いくつかの実施形態において、増大されたまたは強化された結果も分散記憶システムに格納することができる。いくつかの実施形態において、採集されたデータソースに関連する強化中に取込まれたメタデータは、分散記憶システム１０５に格納することができる。システムレベルのメタデータ（たとえばデータソースの位置、結果、処理履歴、ユーザセッション、実行履歴、および構成等を示す）は、分散記憶システムに、または、データ強化サービスからアクセス可能な独立したリポジトリに格納することができる。

【００３０】

特定の実施形態において、強化プロセス１１０は、セマンティックバス（本明細書では

パイプラインまたはセマンティックパイプラインとも呼ぶ）およびこのバスに接続する１つ以上の自然言語（ＮＬ（natural language））プロセッサを用いてデータを分析することができる。ＮＬプロセッサは、自動的にデータソース列を識別し、特定列のデータのタイプを判断し、入力にスキーマがなければこの列に命名し、および／または列および／またはデータソースを説明するメタデータを提供することができる。いくつかの実施形態において、ＮＬプロセッサは、列のテキストからエンティティ（たとえば人物、場所、物等）を識別して抽出することができる。ＮＬプロセッサは、データソース内のおよびデータソース間の関係を識別しおよび／または構築することもできる。以下でさらに説明するように、抽出したエンティティに基づいて、データを修復（たとえばタイプミスもしくはフォーマットエラーを修正）および／または強化する（たとえば抽出したエンティティに追加の関連情報を含める）ことができる。

【００３１】

いくつかの実施形態において、公開段１１２は、強化中に取込まれたデータソースのメタデータと、データソースのいかなる強化または修復も、分析のために１つ以上のビジュアライゼーションシステムに与えることができる（たとえば推奨されるデータ変換、強化、および／またはその他の修正をユーザに対して表示することができる）。公開サブシステムは、処理済のデータを１つ以上のデータターゲットに送ることができる。データターゲットは、処理済みのデータを送ることができる場所に相当し得る。この場所は、たとえば、メモリ内の場所、コンピューティングシステム、データベース、または、サービスを提供するシステムであってもよい。たとえば、データターゲットは、オラクルストレージクラウドサービス（Oracle Storage Cloud Service）（ＯＳＣＳ）、ＵＲＬ、第三者ストレージサービス、ウェブサービス、ならびに、オラクルビジネスインテリジェンス（Business Intelligence）（ＢＩ）、サービスとしてのデータベース（Database as a Service）およびサービスとしてのデータベーススキーマ（Database Schema as a Service）等のその他のクラウドサービスを、含み得る。いくつかの実施形態において、シンジケーションエンジンは、ブラウズ、選択、および結果に対するサブスクライブの対象である一組のＡＰＩを顧客に提供する。サブスクライブされ、新たな結果が生じると、結果データは、外部ウェブサービスのエンドポイントへの直接フィードとして、またはバルクファイルダウンロードとして、提供することができる。

【００３２】

図２は、本発明の実施形態に従うテクノロジースタックの簡略化されたブロック図２００を示す。いくつかの実施形態において、データ強化サービスは、図２に示される論理テクノロジースタックを用いて実現できる。このテクノロジースタックは、１つ以上のクライアントデバイスを通して（たとえばシンクライアント、シッククライアント、ウェブブラウザ、またはクライアントデバイス上で実行されるその他のアプリケーションを用いて）データ強化サービスへのアクセスを提供するユーザインターフェイス／エクスペリエンス（ＵＸ）レイヤ２０２を含み得る。スケジューラサービス２０４は、ＵＸレイヤを通して受けた結果／レスポンスを管理することができ、かつ、基礎をなすインフラストラクチャを管理することができ、データ強化サービスはこのインフラストラクチャ上で実行される。

【００３３】

いくつかの実施形態において、図１を参照して先に説明した処理段は、多数の処理エンジンを含み得る。たとえば、準備処理段１０８は、採集／準備エンジンと、プロファイリングエンジンと、推薦エンジンとを含み得る。準備処理中にデータが採集されると、このデータ（またはそのサンプル）は、分散データストレージシステム２１０（「ビッグデータ」クラスタ等）に格納することができる。強化処理段１１０は、意味／統計エンジンと、エンティティ抽出エンジンと、修復／変換エンジンとを含み得る。以下でさらに説明するように、強化処理段１１０は、強化プロセス中に知識サービス２０６から取得した情報を利用できる。強化アクション（たとえばデータの追加および／または変換）を、分散ストレージシステム２１０に格納されているデータに対して実行できる。データの変換は、

欠けているデータまたはデータを追加することによりデータを強化するための修正を含み得る。データの変換は、データ中のエラーを修正することまたはデータを修復することを含み得る。公開処理段１１２は、公開エンジンと、シンジケーションエンジンと、メタデータ結果マネージャとを含み得る。いくつかの実施形態において、さまざまなオープンソース技術を用いることにより、さまざまな処理段および／または処理エンジン内のいくつかの機能を実装できる。たとえば、ファイルフォーマット検出は、Apache Tikaを使用してもよい。

【００３４】

いくつかの実施形態において、管理サービス２０８は、強化処理１１０中にデータに対してなされる変更をモニタリングすることができる。変更のモニタリングは、どのユーザがデータにアクセスしたか、どのデータ変換が実行されたか、および、その他のデータをトラッキングすることを含み得る。これにより、データ強化サービスは強化アクションをロールバックすることができる。

【００３５】

テクノロジースタック２００は、ビッグデータオペレーションのためのクラスタ２１０（「ビッグデータクラスタ」）等の環境において実装できる。クラスタ２１０は、ＨＤＦＳ等の分散ファイルシステム（distributed file system）（ＤＦＳ）と互換性がある分散コンピューティングフレームワークを実装するための一組のライブラリを提供するApache Sparkを用いて実装できる。Apache Sparkは、マップ、低減、フィルタ、ソート、またはサンプルクラスタ処理ジョブ要求を、ＹＡＲＮのような有効なリソースマネージャに送ることができる。いくつかの実施形態において、クラスタ２１０は、たとえばCloudera（登録商標）が提供する分散ファイルシステム製品を用いて実装できる。たとえばCloudera（登録商標）が提供するＤＦＳは、ＨＤＦＳおよびＹＡＲＮを含み得る。

【００３６】

図３は、本発明の実施形態に従う対話型ビジュアライゼーションシステムの簡略化されたブロック図を示す。図３に示されるように、データ強化サービス３０２は１つ以上のクライアント３０４からデータ強化要求を受けることができる。データ強化システム３００はデータ強化サービス３０２を実装し得る。データ強化サービス３０２は１以上のクライアント３０４からデータ強化要求を受けることができる。データ強化サービス３０２は１つ以上のコンピュータおよび／またはサーバを含み得る。データ強化サービス３０２は、いくつかのサブシステムおよび／またはモジュールで構成されたモジュールであってもよく、その中に含まれるいくつかは図示されていない可能性もある。データ強化サービス３０２のサブシステムおよび／またはモジュールの数は、図示されているものの数より多くても少なくてもよく、２つ以上のサブシステムおよび／またはモジュールを組合わせてもよく、または、異なる構成または配置のサブシステムおよび／またはモジュールであってもよい。いくつかの実施形態において、データ強化サービス３０２は、ユーザインターフェイス３０６と、採集エンジン３２８と、推薦エンジン３０８と、知識サービス３１０と、プロファイルエンジン３２６と、変換エンジン３２２と、準備エンジン３１２と、公開エンジン３２４とを含み得る。データ強化サービス３０２を実装する要素は、上記のようなセマンティック処理パイプラインを実装するように機能し得る。

【００３７】

データ強化システム３００は、本発明の実施形態に従うセマンティック処理パイプラインを含み得る。セマンティック処理パイプラインのうちのすべてまたは一部を、データ強化サービス１０２によって実装してもよい。データソースを追加するとき、このデータソースおよび／またはそこに格納されるデータは、データソースをロードする前にパイプラインを通して処理することができる。パイプラインは、１つ以上のデータターゲットに対して処理済のデータを公開する前にデータおよび／またはデータソースを処理するように構成された１つ以上の処理エンジンを含み得る。処理エンジンは、新たなデータソースから生データを抽出しこの生データを準備エンジンに提供する採集エンジンを含み得る。準備エンジンは、この生データに対応付けられたフォーマットを識別することができ、この

生データを、データ強化サービス３０２が処理できるフォーマットに変換する（たとえばこの生データを正規化する）ことができる。プロファイルエンジンは、正規化されたデータに対応付けられたメタデータを抽出および／または生成することができ、変換エンジンは、メタデータに基づいて正規化されたデータを変換する（たとえば修復および／または強化する）ことができる。結果として得られた強化データは、公開エンジンに与えられて１つ以上のデータターゲットに送られてもよい。各処理エンジンについては以下でさらに説明する。

【００３８】

いくつかの実施形態において、データ強化サービス３０２は、コンピューティングインフラストラクチャシステム（たとえばクラウドコンピューティングインフラストラクチャシステム）によって与えられてもよい。コンピューティングインフラストラクチャシステムは、１つ以上のコンピューティングシステムを有するクラウドコンピューティング環境において実装し得る。コンピューティングインフラストラクチャシステムは、１つ以上の通信ネットワークを通して、本明細書に記載されているもののような１つ以上のデータソースにまたは１つ以上のデータターゲットに通信可能に結合されてもよい。

【００３９】

クライアント３０４はさまざまなクライアントデバイス（デスクトップコンピュータ、ラップトップコンピュータ、タブレットコンピュータ、モバイルデバイス等）を含み得る。各クライアントデバイスは１つ以上のクライアントアプリケーション３０４を含み得る。このアプリケーションを通してデータ強化サービス３０２にアクセスできる。たとえば、ブラウザアプリケーション、シンクライアント（たとえばモバイルアプリケーション）、および／またはシッククライアントは、クライアントデバイス上で実行することができ、ユーザがデータ強化サービス３０２と対話できるようにする。図３に示される実施形態は、単なる一例であって、本発明のクレームされている実施形態を不当に限定することは意図していない。当業者は数多くの変形、代替例、および修正を認識するであろう。たとえば、クライアントデバイスの数は図示されているデバイスの数よりも多くても少なくてもよい。

【００４０】

クライアントデバイス３０４の種類は多種多様であり得る。これは、パーソナルコンピュータ、デスクトップ、ラップトップ、携帯電話、タブレット等のモバイルまたはハンドヘルドデバイス、および、その他の種類のデバイスを含むが、これらに限定されない。通信ネットワークは、クライアントデバイス３０４とデータ強化サービス３０２との間の通信を容易にする。通信ネットワークの種類はさまざまな種類であり得る。この通信ネットワークは１つ以上の通信ネットワークを含み得る。通信ネットワーク１０６の例は、インターネット、ワイドエリアネットワーク（ＷＡＮ）、ローカルエリアネットワーク（ＬＡＮ）、イーサネット（登録商標）ネットワーク、パブリックまたはプライベートネットワーク、有線ネットワーク、無線ネットワーク等と、その組合わせを含むが、これらに限定されない。ＩＥＥＥ　８０２．ＸＸプロトコルスーツ、ＴＣＰ／ＩＰ、ＩＰＸ、ＳＡＮ、AppleTalk、Bluetooth、およびその他のプロトコル等の、有線プロトコルも無線プロトコルも含む異なる通信プロトコルを用いて通信を容易にしてもよい。一般的に、通信ネットワークはクライアントとデータ強化サービス３０２との通信を容易にするいかなる種類の通信ネットワークまたはインフラストラクチャも含み得る。

【００４１】

ユーザは、ユーザインターフェイス３０６を通してデータ強化サービス３０２と対話することができる。クライアント３０４は、グラフィカルユーザインターフェイスをレンダリングすることにより、ユーザのデータやユーザのデータを変換するための推薦を表示し、命令（「変換命令」）をユーザインターフェイス３０６を通してデータ強化サービス３０２に送信および／または受信することができる。本明細書に開示されている、図４Ａ～図４Ｄ、図５Ａ～図５Ｄ、および図１０に示されるようなユーザインターフェイスは、データ強化サービス３０２によってまたはクライアント３０４を介してレンダリングしても

よい。たとえば、ユーザインターフェイスは、ユーザインターフェイス３０６によって生成されてもよく、クライアント３０４のうちのいずれか１つでデータ強化サービス３０２によってレンダリングされてもよい。ユーザインターフェイスは、ネットワークを介してデータ強化システム３０２から、サービス（たとえばクラウドサービス）またはネットワークアクセス可能なアプリケーションの一部として提供されてもよい。少なくとも１つの例において、データ強化サービス３０２のオペレータは、クライアント３０４のうちの１つを操作することにより、本明細書に開示されるユーザインターフェイスのうちのいずれかにアクセスしこれと対話してもよい。ユーザは、命令をユーザインターフェイス３０６に送信することによりデータソースを追加してもよい（たとえばデータソースアクセスおよび／または位置情報等を提供してもよい）。

【００４２】

データ強化サービス３０２は、採集エンジン３２８を用いてデータを採集してもよい。採集エンジン３２８は、データソースが追加されたときに初期処理エンジンとして機能することができる。採集エンジン３２８は、１つ以上のデータソース３０９からデータ強化サービス３０２に、ユーザデータを、安全に、確実に、かつ信頼性高くアップロードすることを容易にすることができる。いくつかの実施形態において、採集エンジン３２８は、１つ以上のデータソース３０９からデータを抽出しデータ強化サービス３０２内の分散ストレージシステム３０５に格納することができる。１つ以上のデータソース３０９および／または１つ以上のクライアント３０４から採集したデータは、図１および図２を参照しながら先に述べたように処理して分散ストレージシステム３０５に格納することができる。データ強化サービス３０２は、クライアントデータストア３０７からおよび／または１つ以上のデータソース３０９からデータを受信できる。分散ストレージシステム３０５は、１つ以上のデータターゲット３３０に対するデータ公開の前の、パイプラインの残りの処理段の間、アップロードされたデータの一時ストレージの機能を果たすことができる。アップロードが完了すると、準備エンジン３１２を呼出し、アップロードされたデータセットを正規化することができる。

【００４３】

受信データは、構造化データ、非構造化データ、またはこれらの組合わせを含み得る。構造データは、限定されないが、アレイ、レコード、リレーショナルデータベース表、ハッシュ表、連結リスト、またはそれ以外の種類のデータ構造を含む、データ構造に基づき得る。上記のように、データソースは、パブリッククラウドストレージサービス３１１、プライベートクラウドストレージサービス３１３、さまざまな他のクラウドサービス３１５、ＵＲＬまたはウェブベースのデータソース３１７、または、その他任意のアクセス可能なデータソースを含み得る。クライアント３０４からのデータ強化要求は、データソースおよび／または特定のデータ（データソース３０９またはクライアントデータストア３０７を通して入手可能な、表、列、ファイル、またはその他任意の構造化または非構造化データ）を特定することができる。そうすると、データ強化要求サービス３０２は、特定されたデータソースにアクセスして上記データ強化要求において特定された特定のデータを取得してもよい。データソースは、アドレス（たとえばＵＲＬ）によって、ストレージプロバイダ名によって、またはその他の識別子によって特定できる。いくつかの実施形態において、データソースへのアクセスを、アクセス管理サービスによって制御してもよい。クライアント３０４は、ユーザに対し、身分証明（たとえばユーザ名とパスワード）入力要求および／またはデータ強化サービス３０２に対してデータソースにアクセスする権限を与えるための要求を示してもよい。

【００４４】

いくつかの実施形態において、１つ以上のデータソース３０９からアップロードされたデータは、多種多様なフォーマットに変更できる。準備エンジン３１２は、アップロードされたデータを、データ強化サービス３０２による処理のために、一般的な正規化されたフォーマットに変換できる。正規化は、Apache（登録商標）が供給しているApache Tikaのような命令またはコードを用いて実装されるルーチンおよび／または技術によって実行

してもよい。正規化されたフォーマットにより、データソースから取得したデータが正規化されたものを見ることができる。いくつかの実施形態において、準備エンジン３１２は、多数の異なるファイルタイプを読み出すことができる。準備エンジン３１２は、データを正規化して文字で区切られた形式（character separated form）（たとえばタブで区切られた値（tab separated values）、カンマで区切られた値（comma separated values）等）、または、階層データ用のJavaScript(登録商標)オブジェクト表記法（JavaScript Object Notation）（ＪＳＯＮ）文書にすることができる。いくつかの実施形態において、さまざまなファイルフォーマットを認識し正規化することができる。たとえば、Microsoft Excel（登録商標）フォーマット（たとえばＸＬＳまたはＸＬＳＸ）、Microsoft Word（登録商標）フォーマット（たとえばＤＯＣまたはＤＯＸ）、ポータブルドキュメントフォーマット（ＰＤＦ）、ＪＳＯＮのような階層フォーマット、および拡張マークアップ言語（ＸＭＬ）等の、標準ファイルフォーマットをサポートすることができる。いくつかの実施形態において、さまざまなバイナリ符号化ファイルフォーマットおよびシリアル化されたオブジェクトデータを読み出して復号することもできる。いくつかの実施形態において、データは、Unicodeフォーマット（ＵＴＦ－８）符号化においてパイプラインに与えることができる。準備エンジン３１２は、コンテキスト抽出と、データ強化サービス３０２が予測するファイルタイプへの変換を実行することができるとともに、データソースから文書レベルメタデータを抽出することができる。

【００４５】

データセットの正規化は、データセット内の生データを、データ強化サービス３０２、特にプロファイルエンジン３２６が処理できるフォーマットに変換することを含み得る。一例において、データセットを正規化して正規化データセットを作成することは、あるフォーマットを有するデータセットを、正規化されたデータセットとして調整されたフォーマットに修正することを含み、調整されたフォーマットは上記フォーマットと異なるフォーマットである。データセットは、このデータセット内のデータの１つ以上の列を識別し、この列に対応するデータのフォーマットを同じフォーマットに修正することによって正規化してもよい。たとえば、あるデータセット内の、フォーマットが異なる日付を有するデータを、この日付のフォーマットをプロファイルエンジン３２６が処理できる共通フォーマットに変更することによって正規化してもよい。データは、表形式でないフォーマットから１つ以上のデータ列を有する表形式のフォーマットに修正または変換することによって正規化されることもある。

【００４６】

データの正規化後、正規化されたデータはプロファイルエンジン３２６に送ることができる。プロファイルエンジン３２６は、正規化されたデータを列ごとに分析することにより、これらの列に格納されているデータのタイプを識別し、データがこれらの列にどのようにして格納されているかに関する情報を識別することができる。本開示では、プロファイルエンジン３２６を多くの場合データに対してオペレーションを実行するものとして説明しているが、プロファイルエンジン３２６によって処理されるデータは準備エンジン３１２によって既に正規化されている。いくつかの実施形態において、プロファイルエンジン３２６によって処理されるデータは、プロファイルエンジン３２６が処理できるフォーマット（たとえば正規化されたフォーマット）であるので正規化されていないデータを含み得る。プロファイルエンジン３２６の出力または結果は、ソースからのデータに関するプロファイル情報を示すメタデータ（たとえばソースプロファイル）であってもよい。メタデータは、データに関する１つ以上のパターンおよび／またはデータの分類を示し得る。以下でさらに説明するように、メタデータは、データの分析に基づく統計情報を含み得る。たとえば、プロファイルエンジン３２６は、識別された各列に関する多数のメトリックとパターン情報を出力することができ、かつ、列の名称およびタイプの形態のスキーマ情報を識別してデータとマッチングすることができる。

【００４７】

プロファイルエンジン３２６が生成したメタデータを、データ強化サービスのその他の

要素、たとえば推薦エンジン３０８および変換エンジン３２２が使用してデータ強化サービス３０２に関して本明細書で説明するオペレーションを実行してもよい。いくつかの実施形態において、プロファイルエンジン３２６はメタデータを推薦エンジン３０８に与えることができる。

【００４８】

推薦エンジン３０８は、プロファイルエンジン３２６によって処理されたデータに関する、修復、変換、およびデータ強化推薦を識別することができる。プロファイルエンジン３２６によって生成されたメタデータを用いて、このメタデータが示す統計分析および／または分類に基づいてデータに関する推薦を判断することができる。いくつかの実施形態において、推薦は、ユーザインターフェイスまたはその他のウェブサービスを通してユーザに提供できる。推薦は、どのようなデータ修復または強化を利用できるか、これらの推薦を如何にして過去のユーザアクティビティと比較するか、および／または未知のアイテムを既存の知識またはパターンに基づいて如何にして分類するかを推薦がハイレベルで記述するように、ビジネスユーザに合わせて調整することができる。知識サービス３１０は、１つ以上の知識グラフまたはその他の知識ソース３４０にアクセスできる。この知識ソースは、ウェブサイト、ウェブサービス、キュレートされた知識ストア、およびそれ以外のソースによって公開されている公的に入手できる情報を含み得る。推薦エンジン３０８は、知識サービス３１０に対し、ソースから取得したデータについてユーザに推薦できるデータを要求する（たとえば問合せる）ことができる。

【００４９】

いくつかの実施形態において、変換エンジン３２２は、ユーザインターフェイス３０６を通して、入力されたデータセットの、列ごとにサンプリングされたデータまたはサンプル行をユーザに対して示すことができる。データ強化サービス３０２は、ユーザインターフェイス３０６を通して、推薦される変換をユーザに示してもよい。この変換は、変換命令に対応付けられていてもよい。変換命令は、変換アクションを実行するためのコードおよび／または関数呼出しを含み得る。変換命令は、ユーザによって、ユーザインターフェイス３０６での選択に基づいて呼び出されてもよく、たとえば、変換に関する推薦を選択することにより、または、オペレーションを示す入力（たとえばオペレータコマンド）を受信することにより、呼び出されてもよい。一例において、変換命令は、エンティティ情報に基づいてデータの少なくとも１つの列をリネームするための命令を含み得る。データの少なくとも１つの列をデフォルト名にリネームするための他の変換命令を受けることもある。デフォルト名は、予め定められた名称を含み得る。デフォルト名は、データの列の名称を判断できないまたはこの列の名称が定義されていない場合の、規定のいかなる名称であってもよい。変換命令は、エンティティ情報に基づいて少なくとも１つの列を再フォーマットするための変換命令、および、エンティティ情報に基づいてデータの少なくとも１つの列を難読化するための命令を含み得る。いくつかの実施形態において、変換命令は、エンティティ情報に基づいて知識サービスから取得したデータの１つ以上の列を追加するための強化命令を含み得る。

【００５０】

ユーザはユーザインターフェイス３０６を通して変換アクションを実行することができ、変換エンジン３２２はデータソースから取得したデータをこれらのアクションに適用し結果を表示することができる。これは、即時フィードバックをユーザに与え、このフィードバックを用いて変換エンジン３２２の構成の効果を可視化して検証することができる。いくつかの実施形態において、変換エンジン３２２は、プロファイルエンジン３２６と、推薦する変換アクションを提供する推薦エンジン３０８とから、パターンおよび／またはメタデータ情報（たとえば列の名称とタイプ）を受けることができる。いくつかの実施形態において、変換エンジン３２２は、データに対する変更を調整しトラッキングすることにより、取り消し、やり直し、削除、および編集イベントを容易にする、ユーザイベントモデルを提供することができる。このモデルは、アクション間の従属性を捕えることにより、現在の構成が矛盾のない状態に保たれるようにすることができる。たとえば、ある列

が削除される場合は、この列に関して推薦エンジン３０８が提供する推薦変換アクションも削除すればよい。同様に、ある変換アクションの結果新たな列が挿入されこのアクションが削除される場合は、この新たな列に対して実行されるいかなるアクションも削除される。

【００５１】

上記のように、処理中に、受信データを分析することができ、推薦エンジン３０８は、このデータに対して実施する、強化、修復、およびそれ以外の変換を含む１つ以上の推薦される変換を示すことができる。データ強化のために推薦される変換は、一組の変換で構成されてもよく、各変換は、データに対して実施する、１つの変換アクションまたはアトミック変換である。変換は、上記組における別の変換によって過去に変換されたデータに対して実施されてもよい。一組の変換は、一組の変換実行後に得られるデータが強化されるように、並列に実行されても特定の順序で実行されてもよい。一組の変換は、変換仕様に従って実施されてもよい。変換仕様は、プロファイルエンジン３２６によって生成されたデータに対する一組の変換各々をどのようにいつ実施するかを示す変換命令と、推薦エンジン３０８が判断したデータを強化するための推薦とを含み得る。アトミック変換の例は、限定されないが、ヘッダへの変換、転換、削除、分割、結合、および修復を含み得る。一組の変換に従って変換されたデータに対して一連の変更がなされてもよい。これらの変更は各々、中間データが強化されるという結果をもたらす。一組の変換に対して中間ステップで生成されるデータは、耐障害性分散データセット（Resilient Distributed Dataset）（ＲＤＤ）、テキスト、データ記録フォーマット、ファイルフォーマット、その他いずれかのフォーマット、またはその組合わせ等のフォーマットで格納されてもよい。

【００５２】

いくつかの実施形態において、データ強化サービス３０２のいずれかの要素によって実行されたオペレーションの結果として生成されたデータは、限定されないがＲＤＤ、テキスト、ドキュメントフォーマット、その他任意の種類のフォーマット、またはこれらを組合わせたものを含む、中間データフォーマットで格納されてもよい。中間フォーマットで格納されたデータを用いて、データ強化サービス３０２のためのオペレーションをさらに実行してもよい。

【００５３】

以下の表は変換の例を示す。表１は変換アクションの種類の概要を示す。

【００５４】

【表１】

変換の種類	関数パラメータ	説明	例
更新	ストリング => ストリング	列の値を更新	難読化、日付フォーマット
分割	ストリング => アレイ[ストリング]	列の値を新たな列に分割	正規表現分割、デリミター分割
フィルタリング	ストリング => ブーリアン	１つの列の値に基づいて行をフィルタリング	ホワイトリストフィルタリング、日付範囲フィルタリング
多重列フィルタ	アレイ[ストリング] => ブーリアン	複数の列の値に基づいて行をフィルタリング	ＮＥＲフォールスポジティブフィルタリング
列編集	アレイ[ストリング] => アレイ[ストリング]	既存の列を編集	リオーダー、削除、列交換
抽出	(ストリング, ストリング) => アレイ[アレイ[ストリング]]	列から新たなＲＤＤに値を抽出	ＮＥＲ結果新たな表に抽出
挿入	ストリング => アレイ[ストリング]	新たな列を挿入	タイムスタンプ挿入
挿入　１：Ｍ	ストリング => アレイ[アレイ[ストリング]]	一対多の方法で新たな列を挿入	ＮＥＲ結果挿入

表１

【００５５】
表２は表１に示されるカテゴリの種類に属さない変換アクションを示す。
【００５６】
【表２】

変換アクション	説明
列をリネーム	列をリネーム
サンプル	現在のＲＤＤをそのサンプルに置換
結合	２つのＲＤＤ間で左－外結合を実施
エクスポート	たとえばＨＤＦＳに列状データとしてＲＤＤをエクスポート

表２

【００５７】

以下の表３は、変換例の種類の例を示す。具体的には、表３は、変換アクションの例を示し、これらのアクションに対応する変換の種類を説明している。たとえば、変換アクションは、データ内のホワイトリストからのワードの存在の検出に基づいてデータをフィルタリングすることを含み得る。ユーザが「Android」または「iPhone（登録商標）」を含む通信（たとえばツイート）の追跡を希望する場合、変換アクションに、与えられたホワイトリストを含む上記２つのワードを追加すればよい。これは、ユーザのためにデータを強化し得る方法の一例に過ぎない。

【００５８】

【表３】

変換アクション	説明	入力	出力	R1
難読化	たとえばクレジットカード番号、ＩＤまたは誕生日等の機密情報を難読化	123-45-6789	###-##-####	Y
日付再フォーマット	日付を含む列を再フォーマット	1330978536 2012-03-12 14:13:49	March 05, 2012 03/12/12 02:13:49 PM	Y
列をリネーム	列をリネーム	tagged_0001 text_label_0005	user_agent call_letters	Y
NER	固有表現抽出を実行し値を挿入する（次のセクション参照）	PopBooth turns your iPhone or iPad into a photo booth, prints and all	Type: Product Text: PopBooth, iPhone, iPad	Y
検索／置換	列の値に対して検索と置換を実行	Search: Mozilla Replace: Godzilla Value: Mozilla 5.0	Value: Godzilla 5.0	Y
ケース変更	ケースを小文字、大文字、または適切な文字に変更	Case: Proper Value: eden prairie	Value: Eden Prairie	Y
ホワイトリストフィルタリング	テキスト一値列のホワイトリストからのワードの存在に基づいて行をフィルタリング	List: Android, iPhone Value: I heart my iPhone	Keep all rows whose values contain “Android” or “iPhone”	Y

表３

【００５９】

推薦エンジン３０８は、知識サービス３１０および知識ソース３４０からの情報を用いることにより、変換エンジン３２２に対する推薦を生成することができ、かつ、変換エンジン３２２に対しデータを変換する変換スクリプトを生成するよう命令することができる

。変換スクリプトは、プログラム、コード、または命令を含み得る。この変換スクリプトは１つ以上の処理ユニットによって実行可能であり、そうすることによって受信データを変換できる。このように、推薦エンジン３０８は、ユーザインターフェイス３０６と知識サービス３１０との間を媒介する機能を果たすことができる。

【００６０】

上記のように、プロファイルエンジン３２６は、データソースからのデータを分析することにより、何らかのパターンがあるか否か判断することができ、何らかのパターンがある場合、そのパターンを分類できるか否か判断することができる。データソースから取得したデータが正規化されると、このデータを構文解析することにより、データの構造内の１つ以上の属性またはフィールドを識別してもよい。パターンは、各々がラベル（「タグ」）を有しカテゴリによって定義される正規表現の集合体を用いて識別し得る。データをさまざまなタイプのパターンと比較することにより、そのパターンを識別してもよい。識別可能なパターンの種類の例は、限定されないが、整数、小数、日付または日付／時間ストリング、ＵＲＬ、ドメインアドレス、ＩＰアドレス、電子メールアドレス、バージョン番号、ロケール識別子、ＵＵＩＤおよびその他の十六進法の識別子、社会保障番号、米国の私書箱番号、典型的な米国のストリートアドレスパターン、郵便番号、米国の電話番号、部屋番号、クレジットカード番号、固有名詞、個人情報、ならびにクレジットカード発行会社を含み得る。

【００６１】

いくつかの実施形態において、プロファイルエンジン３２６は、データ内のパターンを、意味制約または統語制約によって定義された一組の正規表現に基づいて識別し得る。正規表現を用いることにより、データの形状および／または構造を判断できる。プロファイルエンジン３２６は、オペレーションまたはルーチンを実装する（たとえば正規表現に対する処理を実行するルーチンのＡＰＩを呼び出す）ことにより、１つ以上の正規表現に基づいてデータ内のパターンを判別してもよい。たとえば、統語制約に基づいてあるパターンに関する正規表現をデータに適用することにより、データ内のこのパターンを識別可能か否か判断してもよい。

【００６２】

プロファイルエンジン３２６は、１つ以上の正規表現を用いて構文解析作業を実行することにより、プロファイルエンジン３２６によって処理されるデータにおけるパターンを識別することができる。正規表現は、階層に従って並べられてもよい。パターンは、正規表現の複雑度の順に基づいて識別されてもよい。複数のパターンが、分析対象のデータと一致する場合があり、複雑度がより高いパターンが選択される。以下でさらに説明するように、プロファイルエンジン３２６は、統計的分析を実行することにより、パターンとパターンを、これらのパターンの判断のために用いられる正規表現の適用に基づいて区別してもよい。

【００６３】

いくつかの実施形態において、構造化されていないデータを処理することにより、このデータ内のメタデータ記述属性を分析してもよい。メタデータ自身はデータに関する情報を示し得る。このメタデータを比較することにより、類似性を識別するおよび／または情報の種類を判断することができる。データに基づいて識別した情報を比較することにより、データのタイプ（たとえばビジネス情報、個人識別情報、または住所情報）を認識し、パターンに対応するデータを識別することができる。

【００６４】

実施形態に従い、プロファイルエンジン３２６は、統計的分析を実行することにより、データ内のパターンおよび／またはテキストを区別してもよい。プロファイルエンジン３２６は、統計的分析に基づく統計情報を含むメタデータを生成してもよい。パターンが識別されると、プロファイルエンジン３２６は、異なるパターン各々に関する統計情報（たとえばパターンメトリック）を求めることにより、複数のパターンを区別してもよい。統計情報は、認識対象の異なるパターンに関する標準偏差を含み得る。統計情報を含むメタ

データは、推薦エンジン３０８等の、データ強化サービス３０２の他のコンポーネントに提供してもよい。たとえば、メタデータを推薦エンジン３０８に提供することにより、推薦エンジン３０８が、識別されたパターンに基づいてデータの強化のための推薦を決定できるようにしてもよい。推薦エンジン３０８は、パターンを用いて知識サービス３１０に問合せを行なうことにより、パターンに関する追加情報を取得することができる。知識サービス３１０は、１つ以上の知識ソース３４０を含み得る、または、１つ以上の知識ソース３４０にアクセスできる。知識ソースは、ウェブサイト、ウェブサービス、キュレートされた知識ストア、およびその他のソースが公開する、公的に入手可能な情報を含み得る。

【００６５】

プロファイルエンジン３２６は、統計的分析を実行することにより、データ内の識別されたパターンを区別してもよい。たとえば、プロファイルエンジン３２６が分析したデータを評価することにより、データ内の識別された異なるパターン各々についてパターンメトリック（たとえばデータ内の異なるパターンの統計度数）を計算してもよい。各パターンメトリックの組は、識別されたパターンの中で異なるパターンについて計算される。プロファイルエンジン３２６は、異なるパターンについて計算されたパターンメトリック間の相違を判断してもよい。この相違に基づいて、識別されたパターンの中から１つのパターンが選択されてもよい。たとえば、データ内のパターンの度数に基づいて、あるパターンを別のパターンから区別してもよい。別の例において、複数の異なるフォーマットを有する日付でデータが構成されておりこれらのフォーマットがそれぞれ異なるパターンに対応する場合、プロファイルエンジン３２６は、日付を、正規化に加えて標準フォーマットに変換してもよく、その次に、異なるパターンから各フォーマットの標準偏差を求めてもよい。この例において、プロファイルエンジン３２６は、標準偏差が最低のフォーマットがある場合に、複数のフォーマットを統計的に区別し得る。標準偏差が最低のデータのフォーマットに対応するパターンを、データのベストパターンとして選択してもよい。

【００６６】

プロファイルエンジン３２６は、識別するパターンの分類を判断してもよい。プロファイルエンジン３２６は、知識サービス３１０と通信することにより、識別したパターンを知識ドメイン内で分類できるか否か判断してもよい。知識サービス３１０は、マッチング技術および類似性分析等の本明細書で説明する技術に基づいて、データに対応付けられた可能な１つ以上のドメインを判断してもよい。知識サービス３１０は、プロファイルエンジン３２６に、パターンで識別されたデータと類似する可能性がある１つ以上のドメインの分類を提供してもよい。知識サービス３１０は、知識サービス３１０が識別したドメイン各々について、ドメインに対する類似度を示す類似性メトリックを提供してもよい。類似性メトリック分析およびスコアリングについて本明細書に開示する技術を、推薦エンジン３０８によって適用することにより、プロファイルエンジン３２６が処理するデータの分類を判断してもよい。プロファイルエンジン３２６が生成するメタデータは、適用できるものがあれば知識ドメインに関する情報と、プロファイルエンジン３２６が分析したデータに対する類似度を示すメトリックとを含み得る。

【００６７】

プロファイルエンジン３２６は、統計的分析を実行することにより、データ内のパターンが識別されるか否かにかかわらず、データ内の識別されたテキストを区別してもよい。テキストはパターンの一部であってもよく、テキストの分析を用いることにより、識別可能なものがあればさらにパターンを識別してもよい。プロファイルエンジン３２６は、テキストに対するドメイン分析の実行を知識サービス３１０に要求することにより、テキストを１つ以上のドメインに分類できるか否か判断してもよい。知識サービス３１０は、分析しているテキストに適用できる１つ以上のドメインに関する情報を提供するように機能し得る。知識サービス３１０がドメインを判断するために実行する分析は、データのドメインを判断するために使用される類似性分析等の本明細書で説明する技術を用いて実行されてもよい。

【００６８】
いくつかの実施形態において、プロファイルエンジン３２６は、データセット内のテキストデータを識別してもよい。テキストデータは、一組のエンティティのうちの識別された各エンティティに対応し得る。識別されたエンティティごとに分類を判断してもよい。プロファイルエンジン３２６は、知識サービスに対し、エンティティの分類を識別するよう要求してもよい。一組のエンティティ（たとえば１つの列内のエンティティ）について一組の分類を判断すると、プロファイルエンジン３２６は、一組のメトリック（「分類メトリック」を計算することにより、一組の分類を区別してもよい。一組のメトリック各々は、一組の分類のうちのそれぞれの分類について計算されてもよい。プロファイルエンジン３２６は、一組のメトリックを、互いに比較することにより区別して、この一組のエンティティの分類として最も近い分類を決定してもよい。一組のエンティティの分類は、この一組のエンティティを表わす分類に基づいて選択されてもよい。

【００６９】
知識サービス３１０は、知識ソース３４０を用いて、プロファイルエンジン３２６によって識別されたパターンのコンテキストのマッチングを行なうことができる。知識サービス３１０は、データ内の識別されたパターンを、またはテキスト内にあるのであればデータを、知識ソースに格納されている各種エンティティのエンティティ情報と比較してもよい。エンティティ情報は、知識サービス３１０を用いて、１つ以上の知識ソース３４０から取得してもよい。周知のエンティティの例は、社会保障番号、電話番号、住所、固有名詞、またはその他の個人情報を含み得る。データを各種エンティティのエンティティ情報と比較することにより、識別されたパターンに基づいて１つ以上のエンティティと一致するか否か判断してもよい。たとえば、知識サービス３１０は、「ＸＸＸ－ＸＸ－ＸＸＸＸ」というパターンを、米国社会保障番号のフォーマットとマッチングすることができる。さらに、知識サービス３１０は、社会保障番号は保護されておりまたは機密情報でありその開示はさまざまな処罰につながると判断することができる。

【００７０】
いくつかの実施形態において、プロファイルエンジン３２６は、統計分析を実行することにより、プロファイルエンジン３２６が処理したデータについて識別された複数の分類を区別することができる。たとえば、テキストが複数のドメインで分類されている場合、プロファイルエンジン３２６は、データを処理することにより、知識サービス３１０が判断した適切な分類を統計的に求めることができる。分類の統計的分析は、プロファイルエンジン３２６が生成したメタデータに含まれていてもよい。

【００７１】
パターンの識別に加えて、プロファイルエンジン３２６は、データを統計的に分析することができる。プロファイルエンジン３２６は、大量のデータの内容を特徴付けることができ、かつ、このデータに関する全体統計とこのデータの内容の、たとえばその値、パターン、タイプ、構文、意味およびその統計的特性の、列ごとの分析を提供することができる。たとえば、数値データを統計的に分析することができ、これはたとえば、Ｎ、平均、最大値、最小値、標準偏差、歪度、尖度、および／または２０ビンのヒストグラム（Ｎが１００よりも大きく固有値がＫよりも大きい場合）を含む。次の分析のために内容を分類してもよい。

【００７２】
一例において、全体統計は、限定されないが、行の数、列の数、記入されていない列と記入されている列の数およびこれらがどのように変化するか、異なる行と重複する行、ヘッダ情報、タイプまたはサブタイプによって分類される列の数、ならびに、機密保護またはその他の警告付の列の数を含み得る。列固有の統計は、記入されている行（たとえばＫ最大度数、Ｋ最低度数固有値、固有パターン、および（適用可能であれば）タイプ）、度数分布、テキストメトリック（たとえば、テキスト長、トークンカウント、句読点、パターンベースのトークン、および導出されたさまざまな有用テキスト特性の、最小値、最大値、平均値）、トークンメトリック、データタイプおよびサブタイプ、数値列の統計的分

析、大部分が構造化されていないデータの列内で見出される、Ｌ最大／最小確率単純もしくは複合用語またはｎグラム、ならびに、この固有語彙によってマッチングされる参照知識カテゴリ、日付／時間パターンの発見およびフォーマッティング、参照データ一致、ならびに、原因となる列見出しラベルを、含み得る。

【００７３】

結果として得られたプロファイルを用いて、次の分析のために内容を分類することにより、直接または間接的に、データの変換を示唆して、データソース間の関係を識別するとともに、前に取得したデータのプロファイルに基づいて設計された一組の変換を適用する前に新たに取得したデータの妥当性確認を実行することができる。

【００７４】

プロファイルエンジン３２６によって作成されたメタデータを、推薦エンジン３０８に与えることにより、１つ以上の変換推薦を生成することができる。データの識別されたパターンと一致するエンティティを用いてデータを強化することができる。このデータは、知識サービス３１０を用いて判断された分類によって識別されたエンティティを用いて強化される。いくつかの実施形態において、識別されたパターン（たとえば都市および州）に関連するデータを、知識サービス３１０に与えることにより、知識ソース３４０から、識別されたパターンと一致するエンティティを取得してもよい。たとえば、知識サービス３１０を呼出し、識別されたパターンに対応するルーチン（たとえばgetCities()およびgetStates()）をコールすることにより、エンティティ情報を受けてもよい。この知識サービス３１０から受けた情報は、エンティティに関する適切なスペリングの情報（たとえば適切なスペリングの都市および州）を有する、エンティティのリスト（たとえばカノニカル（canonical）リスト）を含み得る。知識サービス３１０から取得した一致するエンティティに対応するエンティティ情報を用いて、データを強化する、たとえばデータを正規化する、データを修復する、および／またはデータを増補することができる。

【００７５】

いくつかの実施形態において、推薦エンジン３０８は、知識サービス３１０から受けた一致したパターンに基づいて、変換推薦を生成することができる。たとえば、社会保障番号を含むデータの場合、推薦エンジンは、エントリを難読化する変換を推薦することができる（たとえば、エントリのうちのすべてまたは一部の切り捨て、ランダム化、または削除）。変換のその他の例は、データの再フォーマット（たとえばデータ内の日付の再フォーマット）、データのリネーム、データの強化（たとえば値を挿入するまたはカテゴリにデータを対応付ける）、データの検索と置換（たとえばデータのスペルを修正）、文字のケースの変更（たとえばケースを大文字から小文字に変更）、および、ブラックリストまたはホワイトリスト用語に基づくフィルタリングを、含み得る。いくつかの実施形態において、特定のユーザに合わせて推薦を調整してどのデータ修復または強化を利用できるかをこの推薦がハイレベルで説明するようにしてもよい。たとえば、難読化の推薦は、エントリの最初の５桁を削除することを示し得る。いくつかの実施形態において、推薦は、過去のユーザの活動に基づいて生成してもよい（たとえば以前に機密データを識別したときに使用した推薦変換を提供）。

【００７６】

変換エンジン３２２は、推薦エンジン３０８から提供された推薦に基づいて変換スクリプト（たとえば社会保障番号を難読化するためのスクリプト）を生成することができる。変換スクリプトは、オペレーションを実行することによってデータを変換し得る。いくつかの実施形態において、変換スクリプトは、データの線形変換を実現し得る。線形変換は、ＡＰＩ（たとえばSpark API）を通して実現されてもよい。変換アクションは、ＡＰＩを用いて呼び出されたオペレーションによって実施されてもよい。変換スクリプトは、ＡＰＩを用いて定義された変換オペレーションに基づいて構成されてもよい。オペレーションは推薦に基づいて実行されてもよい。

【００７７】

いくつかの実施形態において、変換エンジン３２２は、変換スクリプトを自動的に生成

してデータソースでデータを修復することができる。修復は、自動的に列をリネームすること、列内のストリングまたはパターンを置換すること、テキストのケースを修正すること、データを再フォーマットすること等を含み得る。たとえば、変換エンジン３２２は、変換スクリプトを生成することにより、日付の列を、推薦エンジン３０８からの、列内の日付のフォーマットの修正または変換の推薦に基づいて、変換することができる。推薦を複数の推薦の中から選択して、プロファイルエンジン３２６によって処理されたデータソースからのデータを強化または修正してもよい。推薦エンジン３０８は、プロファイルエンジン３２６から提供されたメタデータまたはプロファイルに基づいて推薦を決定してもよい。メタデータは異なるフォーマットについて識別された日付の列を示し得る（たとえばＭＭ／ＤＤ／ＹＹＹＹ、ＤＤ－ＭＭ－ＹＹ等）。変換エンジン３２２によって生成された変換スクリプトは、たとえば、推薦エンジン３０８からの提案に基づいて列を分割および／または結合することができる。変換エンジン３２２はまた、プロファイルエンジン３２６から受けたデータソースプロファイルに基づいて列を削除してもよい（たとえば空の列、またはユーザが望まない情報を含む列を削除する）。

【００７８】

変換スクリプトは、１つ以上のアルゴリズム（たとえばSparkオペレータツリー）に対するオペレーションを記述する構文を用いて定義し得る。よって、構文はオペレーターツリーの変換／簡約化を記述し得る。変換スクリプトは、グラフィカルユーザインターフェイスを介した対話を通してユーザが選択した推薦に基づいてまたはユーザによって要求されて生成されてもよい。推薦される変換の例は、図４Ａ、図４Ｂ、図４Ｃ、および図４Ｄを参照しながら説明する。グラフィカルユーザインターフェイスを通してユーザが指定した変換オペレーションに基づいて、変換エンジン３２２はこのオペレーションに従って変換オペレーションを実行する。変換オペレーションをユーザに対して推薦することによりデータセットを強化してもよい。

【００７９】

以下でさらに説明するように、クライアント３０４は、推薦された各変換を記述するかそうでなければ示す推薦を表示することができる。ユーザが変換スクリプトの実行を選択した場合、選択された変換スクリプトは、推薦される変換を決定するために分析されたデータに加えてデータソースからのデータすべてまたはそれ以上に対して実行することができる。その結果変換されたデータは、次に公開エンジン３２４によって１つ以上のデータターゲット３３０に対して公開することができる。いくつかの実施形態において、データターゲットは、データソースとは異なるデータストアである。いくつかの実施形態において、データターゲットはデータソースと同一のデータストアであってもよい。データターゲット３３０は、パブリッククラウドストレージサービス３３２、プライベートクラウドストレージサービス３３４、その他さまざまなクラウドサービス３３６、ＵＲＬまたはウェブベースのデータターゲット３３８、またはその他任意のアクセス可能なデータターゲットを含み得る。

【００８０】

いくつかの実施形態において、推薦エンジン３０８は、識別されたプラットフォームに関連するその他のデータについて知識サービス３１０に問合せることができる。たとえば、データが都市名の列を含む場合、関連データ（たとえば場所、州、人口、国等）を識別することができ、関連データでデータセットを強化するという推薦を表示することができる。ユーザインターフェイスを通した推薦の表示およびデータ変換の例は、以下において図４～図４Ｄを参照しながら示す。

【００８１】

知識サービス３１０は、マッチングモジュール３１２と、類似性メトリックモジュール３１４と、知識スコアリングモジュール３１６と、カテゴリ分類モジュール３１８とを含み得る。以下でさらに説明するように、マッチングモジュール３１２は、マッチング方法を実装することにより、データを、知識サービス３１０を通して入手できる参照データと比較することができる。知識サービス３１０は、１つ以上の知識ソース３４０を含み得る

または１つ以上の知識ソース３４０にアクセスできる。知識ソースは、ウェブサイト、ウェブサービス、キュレートされた知識ストア、およびそれ以外のソースによって公開されている公的に入手できる情報を含み得る。マッチングモジュール３１２は、本開示に記載されているような１つ以上のマッチング方法を実装し得る。マッチングモジュール３１２は、適用されるマッチング方法に関連する状態を格納するためのデータ構造を実装し得る。

【００８２】

類似性メトリックモジュール３１４は、２つ以上のデータセット間の意味類似性を判断するための方法を実装することができる。これは、知識サービス３３０を通して入手できる参照データに対してユーザのデータをマッチングする場合も使用できる。類似性メトリックモジュール３１４は、図６～図１５を参照する説明を含む本開示に記載されている類似性メトリック分析を実行し得る。

【００８３】

カテゴリ分類モジュール３１８は、自動データ分析を実装するためのオペレーションを実行することができる。いくつかの実施形態において、カテゴリ分類モジュール３１８は、Word2Vec等の教師なし機械学習ツールを用いて入力データセットを分析することができる。Word2Vecは、テキスト入力（たとえば大きなデータソースからのテキストコーパス）を受けて各入力ワードのベクトル表現を生成することができる。次に、その結果得たモデルを用いて任意入力された一組のワードの関連性がどれほど高いかを識別してもよい。たとえば、大きなテキストコーパス（たとえばニュースアグリゲータまたはその他のデータソース）を用いて構築されたWord2Vecモデルを利用して、対応する数値ベクトルを入力ワードごとに求めることができる。これらのベクトルが分析される際に、ベクトルはベクトル空間内で「近い」（ユークリッドの意味で）と判断されることがある。これは入力ワードが関連していると識別することができるが（たとえばベクトル空間内で互いに近接してクラスタリングされている入力ワードを識別する）、Word2Vecは、ワードを説明するラベル（たとえば「メーカー」）を識別するには有用でない場合がある。カテゴリ分類モジュール３１８は、キュレートされた知識ソース３４０（たとえばMax Planck Institute for InformaticsのＹＡＧＯ）を用いて関連ワードをカテゴリ分類するためのオペレーションを実装してもよい。カテゴリ分類モジュール３１８は、知識ソース３４０からの情報を用いて、入力データセットに対してその他の関連データを追加することができる。

【００８４】

いくつかの実施形態において、カテゴリ分類モジュール３１８は、トライグラムモデリングを実行することによって関連する用語のカテゴリ分類をさらに精密にするためのオペレーションを実装し得る。トライグラムモデリングを用いてワードの組をカテゴリ識別のために比較することができる。入力データセットは関連する用語で増補することができる。

【００８５】

マッチングモジュール３１２は、追加データを含み得る入力データセットを用いて、マッチング方法（たとえばグラフマッチング法）を実装することにより、増補データセットからのワードを、知識ソース３４０からのデータのカテゴリと比較することができる。類似性メトリックモジュール３１４は、増補データセットと知識ソース３４０内の各カテゴリとの意味類似性を判断してそのカテゴリの名称を識別するための方法を実装することができる。カテゴリの名称は、最大類似性メトリックに基づいて選択してもよい。類似性メトリックは、カテゴリ名と一致するデータセット内の用語の数に基づいて計算されてもよい。カテゴリは、類似性メトリックに基づいて一致する最大数の用語に基づいて選択されてもよい。類似性分析およびカテゴリ分類のために実行される技術およびオペレーションを、図６～図１５を参照する説明を含む本開示においてさらに説明する。

【００８６】

いくつかの実施形態において、カテゴリ分類モジュール３１８は、入力データセットを増補することができ、知識ソース３４０からの情報を用いて入力データセットにその他の

関連データを追加することができる。たとえば、Word2Vec等のデータ分析ツールを用いて、ニュース収集サービスからのテキストコーパスのような知識ソースからの入力データセットに含まれているワードに意味的に類似するワードを識別することができる。いくつかの実施形態において、カテゴリ分類モジュール３１８は、トライグラムモデリングを実装することにより、知識ソース３４０（ＹＡＧＯ等）から取得したデータを処理して、カテゴリによってインデックスが作成されたワードの表を生成することができる。カテゴリ分類モジュール３１８は次に、増補されたデータセット内のワードごとにトライグラムを作成しそのワードをインデックス付の知識ソース３４０からのワードとマッチングすることができる。

【００８７】

カテゴリ分類モジュール３１８は、増補データセット（またはトライグラム一致増補データセット）を用いて、マッチングモジュール３１２に対し、増補データセットからのワードを、知識ソース３４０からのデータのカテゴリと比較するよう要求することができる。たとえば、知識ソース３４０内のデータの各カテゴリはツリー構造で表現することができる。ツリー構造のルート（root）ノードはカテゴリを表わし各リーフ（leaf）ノードはそのカテゴリに属するそれぞれのワードを表わす。類似性メトリックモジュール３１４は、増補データセットと知識ソース５１０内の各カテゴリとの意味類似性を判断するための方法（たとえばJaccard係数またはその他の類似性メトリック）を実装することができる。次に、増補データセットと一致する（たとえば類似性メトリックが最大である）カテゴリの名称をラベルとして入力データセットに適用することができる。

【００８８】

いくつかの実施形態において、類似性メトリックモジュール３１４は、２つのデータセットＡおよびＢの類似性を、データセットＡおよびＢの共通集合の大きさの、これらのデータセットの合併集合の大きさに対する比率を求めることによって判断できる。たとえば、類似性メトリックを、１）データセット（たとえば増補データセット）とカテゴリとの共通部分の大きさと、２）これらを合併したものの大きさとの比率に基づいて計算してもよい。類似性メトリックは、上記のように、データセットとカテゴリとの比較のために計算してもよい。よって、類似性メトリックの比較に基づいて「ベストマッチ」を判断してもよい。この比較に使用されるデータセットを、類似性メトリックを用いてベストマッチを判断したカテゴリに対応するラベルで増補することによって強化してもよい。

【００８９】

上記のように、その他の類似性メトリックを、Jaccard係数に加えてまたはその代わりに使用してもよい。上記技術に対していかなる類似性メトリックも使用し得ることを当業者は理解するであろう。代替の類似性メトリックのいくつかの例は、Dice-Sorensen係数、Tversky係数、Tanimotoメトリック、およびコサイン類似度メトリックを含むが、これらに限定される訳ではない。

【００９０】

いくつかの実施形態において、カテゴリ分類モジュール３１８は、Word2Vec等のデータ分析ツールを利用することにより、知識ソース３４０からのデータと知識ソースからのデータで増補し得る入力データとの間の一致度を示す精密なメトリック（たとえばスコア）を計算してもよい。スコア（「知識スコア」）は、入力データセットと比較対象のカテゴリとの類似度に関してより多くの知識を提供し得る。知識スコアによって、入力データを最も良く表わしているカテゴリ名をデータ強化サービス３０２が選択できるようにしてもよい。

【００９１】

上記技術において、カテゴリ分類モジュール３１８は、知識ソース３４０における候補カテゴリ（たとえば属）の名称に対する、入力データセットにおける用語の一致の数をカウントしてもよい。この比較の結果から、完全な整数（whole integer）を表わす値を得ることができる。よって、この値は、用語と用語の一致度を示すが、入力データセットと知識ソース内の各種用語との間の一致度は示さない場合がある。

【００９２】
カテゴリ分類モジュール３１８は、Word2Vecを用いることにより、知識ソース内の各用語（たとえば種を表わす用語）と入力データの用語（たとえば種）との比較類似度を判断してもよい。カテゴリ分類モジュール３１８は、Word2Vecを用いて入力データセットと知識ソースから取得した１つ以上の用語との類似性メトリック（たとえばコサイン類似度または距離）を計算することができる。コサイン類似度は、知識ソースから取得した用語のデータセット（たとえばドメインまたは属）と用語の入力データセットとの間のコサイン角度として計算してもよい。コサイン類似度メトリックは、Tanimotoメトリックと同様に計算してもよい。以下の式はコサイン類似度メトリックの一例を示す。

【００９３】
【数１】

$$\frac{X \cdot Y}{|X||Y|}$$

【００９４】
コサイン類似度に基づいて類似度メトリックを計算することにより、入力データセット内の各用語を、その用語と候補カテゴリとの間の類似性のパーセンテージを示す値のような、完全値整数（whole-value integer）分の１とみなしてもよい。たとえば、タイヤメーカーと名字との間の類似度メトリックを計算した結果、類似度メトリックは０．３かもしれない。一方、タイヤメーカーと企業名との間の類似度メトリックを計算した結果、類似度メトリックは０．５かもしれない。類似度メトリックを表わす非完全整数値を細かく比較することにより、一致度が高いカテゴリ名をより正確にすることができる。一致度が高いカテゴリ名を、値１に最も近い類似度メトリックに基づいて最も適切なカテゴリ名として選択してもよい。上記の例において、類似度メトリックに基づくと、企業名は正しいカテゴリである可能性が高い。よって、カテゴリ分類モジュール３１８は、「名字」ではなく「企業」を、タイヤメーカーを含む、ユーザから提供されたデータ列に対応付ければよい。

【００９５】
知識スコアリングモジュール３１６は、知識グループ（たとえばドメインまたはカテゴリ）に関する情報を判断することができる。知識グループに関する情報は、図１０に示される例のようなグラフィカルユーザインターフェイスに表示することができる。知識ドメインに関する情報は、知識ドメインと用語の入力データセットとの間の類似度を示すメトリック（たとえば知識スコア）を含み得る。入力データを知識ソース３４０からのデータと比較してもよい。入力データセットは、ユーザによって指定されたデータセットからのデータの列に対応する場合がある。知識スコアは、入力データセットと、知識ソースから提供される１つ以上の用語との間の類似度を示し得る。各用語は知識ドメインに対応する。データの列は、場合によっては知識ドメインに属する用語を含み得る。

【００９６】
少なくとも１つの実施形態において、知識スコアリングモジュール３１６は、より正確なマッチングスコアを求めることができる。このスコアは、スコアリングの式を用いて計算された値に対応していてもよい。スコアリングの式により、２つのデータセット、たとえば、入力データセットと知識ソースから取得したドメイン（たとえば候補カテゴリ）の用語との間の意味類似性を求めてもよい。そのマッチングスコアがベストマッチ（たとえば最大マッチングスコア）を示すドメインを、入力データセットとの類似性が最大であるドメインとして選択してもよい。よって、入力データセット内の用語は、カテゴリとしてのドメイン名に対応付けられてもよい。

【００９７】
スコアリングの式を、入力データセットとドメイン（たとえば知識ソースから取得した用語のカテゴリ）に適用することにより、この入力データとドメインとの間の一致度を示

すスコアを求めてもよい。ドメインは、集まってドメインを定義する１つ以上の用語を有し得る。スコアを用いることにより、入力データセットが最も類似するドメインを求めてもよい。入力データセットを、この入力データセットが最も類似するドメインを記述する用語に対応付けてもよい。

【００９８】

スコアリングの式は、入力データセットと比較されるドメインに関連する１つ以上の因子に基づいて定めてもよい。スコアリングの式の因子は、限定されないが、度数値（たとえば入力データセットとドメイン内の用語とが一致する用語度数）、母集団値（たとえば入力データセット内の用語の数）、固有マッチング値（たとえば入力データセットとドメインとが一致する各種用語の数）、ドメイン値（たとえばドメイン内の用語の数）、および、ドメインがどの程度キュレートされたかを示す値の範囲（たとえば０．０～１００．０）の中の一定の値を示すキュレーションレベルを含み得る。少なくとも１つの実施形態において、スコアリングの式は、関数スコア（ｆ，ｐ，ｕ，ｎ，ｃ）として定めてもよい。この場合、スコアリングの式は等式（１＋ｃ／１００）＊（ｆ／ｐ）＊（ｌｏｇ（ｕ＋１）／ｌｏｇ（ｎ＋１））によって計算され、「ｆ」は度数値を表わし、「ｃ」はキュレーションレベルを表わし、「ｐ」は母集団値を表わし、「ｕ」は固有マッチング値を表わし、「ｎ」はドメイン値を表わす。

【００９９】

スコアリングの式の計算は、図１０を参照しながらさらに説明することができる。縮小例において、入力データセット（たとえば表のデータの列）を、１００の短いテキストの値を有するものとして定めてもよく、知識ソースを、各々が都市に対応する１０００個の用語を有する都市ドメイン（たとえば「city」）と、各々が姓に対応する８００個の用語を有する姓ドメイン（たとえば「last_name」）とを含むドメインによって定めてもよい。入力データセットは、都市ドメイン内の６０個の用語（たとえば都市）と一致する８０の行（各行は１つの用語に対応）を有していてもよく、入力データセットは、姓ドメイン内の５５個の用語（たとえば姓）と一致する６５の行を有していてもよい。都市ドメインはキュレーションレベル１０で定めてもよく、姓ドメインはキュレーションレベル０（たとえばキュレートされていない）で定めてもよい。この例の値に基づいてスコアリングの式を適用し、都市ドメインの知識スコアをスコア（８０，１００，６０，１０００，１０）に従って計算すると、０．５２３６２０９８７５７７０２３１（たとえば、１００のうちの５２スコアまたは５２％）となる。姓ドメインの知識スコアをスコア（６５，１００，５５，８００，０）に従って計算すると、０．３９１３４５０５１８４７８２９７５（たとえば、１００のうちの３９スコアまたは３９％）となる。この縮小例において、知識スコアリングに基づくと、入力データセットは、一致度がより高く、姓ドメインよりも都市ドメインに対する類似度が高い。この例に従い、図１０は、入力データセットと比較される用語を有する各種ドメインと、スコアリングの式を用いて計算したスコア（「マッチング」）の一例を示す。

【０１００】

いくつかの実施形態において、スコアリングの式は、上記よりも多いまたは少ない因子に基づいて定めてもよい。この式を調整または修正することにより、一致をより適切に表わすスコアを生成してもよい。

【０１０１】

プロファイルエンジン３２６は、パターン識別とマッチングに加えて、データを統計的に分析することができる。プロファイルエンジン３２６は、大量のデータの内容を特徴付けることができ、かつ、データに関する全体統計と、データの内容の、たとえばその値、パターン、タイプ、構文、意味、およびその統計的特徴の、列ごとの分析とを提供することができる。たとえば、数値データを統計的に分析することができ、これはたとえば、Ｎ、平均、最大値、最小値、標準偏差、歪度、尖度、および／または２０ビンのヒストグラム（Ｎが１００よりも大きく固有値がＫよりも大きい場合）を含む。次の分析のために内容を分類してもよい。いくつかの実施形態において、プロファイルエンジン３２６は、１

つ以上のＮＬプロセッサによってデータを分析することができる。これは、データソースの列を自動的に識別し、特定列のデータのタイプを判断し、入力にスキーマがなければ列に命名し、および／または列および／またはデータソースを記述するメタデータを提供することができる。いくつかの実施形態において、ＮＬプロセッサは、列のテキストのエンティティ（たとえば人、場所、物等）を識別して抽出することができる。ＮＬプロセッサは、データソース内の関係およびデータソース間の関係を識別および／または構築することもできる。

【０１０２】

一例において、全体統計は、限定されないが、行の数、列の数、記入されていない列と記入されている列の数およびこれらがどのように変化するか、異なる行と重複する行、ヘッダ情報、タイプまたはサブタイプによって分類される列の数、ならびに、機密保護またはその他の警告付の列の数を含み得る。列固有の統計は、記入されている行（たとえばＫ最大度数、Ｋ最低度数固有値、固有パターン、および（適用可能であれば）タイプ）、度数分布、テキストメトリック（たとえば、テキスト長、トークンカウント、句読点、パターンベースのトークン、および導出されたさまざまな有用テキスト特性の、最小値、最大値、平均値）、トークンメトリック、データタイプおよびサブタイプ、数値列の統計的分析、大部分が構造化されていないデータの列内で見出される、Ｌ最大／最小確率単純もしくは複合用語またはｎグラム、およびこの固有語彙によってマッチングされる参照知識カテゴリ、日付／時間パターンの発見およびフォーマッティング、参照データ一致、ならびに、原因となる列見出しラベルを、含み得る。

【０１０３】

結果として得られたプロファイルを用いて、次の分析のために内容を分類することにより、直接または間接的に、データの変換を示唆して、データソース間の関係を識別するとともに、前に取得したデータのプロファイルに基づいて設計された一組の変換を適用する前に新たに取得したデータの妥当性確認を実行することができる。

【０１０４】

いくつかの実施形態において、ユーザインターフェイス３０６は、プロファイルエンジン３２６から提供されたメタデータに基づいて、グラフィカルなビジュアライゼーションを１つ以上生成することができる。上記のように、プロファイルエンジン３２６から提供されるデータは、プロファイルエンジン３２６によって処理されたデータに関するメトリックを示す統計情報を含み得る。プロファイリングされたデータのメトリックのグラフィカルなビジュアライゼーションの例は、図５Ａ～図５Ｄに示される。グラフィカルなビジュアライゼーションは、グラフィカルダッシュボード（たとえばビジュアライゼーションダッシュボード）を含み得る。グラフィカルダッシュボードは複数のメトリックを示し得る。これら複数のメトリックは各々、データがプロファイリングされた時間に対する、データのリアルタイムメトリックを示す。グラフィカルなビジュアライゼーションはユーザインターフェイスに表示されてもよい。たとえば、グラフィカルなビジュアライゼーションをクライアントデバイスに送ることにより、クライアントデバイスが、クライアントデバイスのユーザインターフェイスに、グラフィカルなビジュアライゼーションを表示できるようにする。いくつかの実施形態において、グラフィカルなビジュアライゼーションはプロファイリング結果を提供し得る。

【０１０５】

加えて、プロファイルエンジン３２６による構造分析により、推薦エンジンは、そのクエリをより適切に知識サービスに向けることができ、その結果、処理速度が改善されシステムリソースに対する負荷が低減される。たとえば、この情報を用いて、クエリ対象の知識の範囲を制限することにより、知識サービス３１０が数値データの列を場所名に対してマッチングするようなことが生じないようにすることができる。

【０１０６】

図４Ａ～図４Ｄは、本発明の実施形態に従う対話型データ強化を提供するユーザインターフェイスの例を示す。図４Ａに示されるように、代表的な対話型ユーザインターフェイ

ス４００は、変換スクリプト４０２、推薦される変換４０４、および分析／変換の対象であるデータ４０６の少なくとも一部を表示することができる。パネルに一覧表示されている変換スクリプト４０２は、既にデータに適用されパネルで見ることができる変換４０６を含み得る。各変換スクリプト４０２は、ビジネスユーザにとってわかり易い単純な宣言型言語で記述することができる。パネルに一覧表示されている変換スクリプト４０２を、自動的にデータに適用し、対話型ユーザインターフェイス４００に表示されているデータ４０６の一部に反映させてもよい。たとえば、パネルに一覧表示されている変換スクリプト４０２は、その内容を記述すべきリネーム列を含む。対話型ユーザインターフェイス４００に示される列４０８は、変換スクリプト４０２に従って既にリネームされている（たとえば、列０００３はdate_time_02にリネームされ列０００７は「url」にリネームされている等）。しかしながら、推薦される変換４０４はユーザのデータに自動的に適用されていない。

【０１０７】

図４Ｂに示されるように、ユーザは推薦パネルの推薦４０４を見ることができこの推薦に基づいて変更すべきデータを識別することができる。たとえば、推薦４１０は、「Col_0008 to city」にリネームすることを推薦している。推薦は、（たとえばコードまたは疑似コードではなく）ビジネスユーザが理解できるように記述されているので、ユーザは対応するデータ４１２を簡単に識別できる。図４Ｂに示されるように、データ４１２はストリングの列（ユーザインターフェイス４００では行として表わされる）を含む。プロファイルエンジン３２６はデータを分析することによりこれが２つ以下のワード（またはトークン）のストリングを含むと判断することができる。このパターンを、知識サービス３１０に対してクエリすることができる推薦エンジン３０８に与えることができる。この場合、知識サービス３１０は、このデータパターンを都市名に対してマッチングし、推薦４０８はそれに応じて列をリネームするために生成された。

【０１０８】

いくつかの実施形態において、パネルに一覧表示されている推薦４０４は、ユーザに向けて（たとえば変換を適用せよという命令に応じて）適用されていてもよく、または、自動的に適用されてもよい。たとえば、いくつかの実施形態において、知識サービス３１０は、所与のパターン一致に対して信頼性スコアを与えることができる。しきい値を推薦エンジン３０８に設定し、このしきい値よりも高い信頼性スコアを有する一致が自動的に適用されるようにすることができる。

【０１０９】

ユーザは、推薦を受容れる場合、この推薦に対応付けられた受容アイコン４１４（この例では上向きの矢印のアイコン）を選択すればよい。図４Ｃに示されるように、そうすると、受容された推薦４１４は、変換スクリプト４０２のパネルに移動し、自動的に変換を対応するデータ４１６に適用する。たとえば、図４Ｃに示される実施形態において、Col_0008は、選択された変換に従って「city」にリネームされている。

【０１１０】

いくつかの実施形態において、データ強化サービス３０２は、さらに他のデータ列をデータソースに加えることを提案できる。図４Ｄに示されるように、「city」の例を続けると、変換４１８は、都市の人口と経度および緯度を含む都市の位置の詳細とを含む新たな列でデータを強化することが受容されている。選択されると、ユーザのデータセットは、この追加情報４２０を含むように強化される。そうすると、このデータセットは、以前は総合的にかつ自動的にユーザが利用できなかった情報を含むことになる。この時点で、ユーザのデータセットを用いて、データセット内の他のデータに対応付けられた位置ゾーンおよび人口ゾーンからなる全国地図を作成することができる（たとえばこれを企業のウェブサイトトランザクションに対応付けてもよい）。

【０１１１】

図５Ａ～図５Ｄは、本発明の実施形態に従うデータセットのビジュアライゼーションを提供するさまざまなユーザインターフェイスの例を示す。

【０１１２】
図５Ａは、本発明の実施形態に従うデータセットのビジュアライゼーションを提供するユーザインターフェイスの一例を示す。図５Ａに示されるように、代表的な対話型ユーザインターフェイス５００は、プロファイル概要５０２（「プロファイル結果」）と、変換スクリプト５０４と、推薦される変換５０６と、分析／変換対象のデータの少なくとも一部５０８とを表示することができる。パネルに一覧表示されている変換５０４は、既にデータに適用されパネルにおいて見ることができる変換５０８を含み得る。
【０１１３】
プロファイル概要５０２は、全体統計（たとえば行総数および列総数）と、列特有の統計とを含み得る。列特有の統計は、データ強化サービス３０２によって処理されたデータの分析によって生成することができる。いくつかの実施形態において、列特有の統計は、データ強化サービス３０２によって処理されたデータの分析によって求められた列情報に基づいて生成することができる。
【０１１４】
プロファイル概要５０２は、米国の地図（たとえば「ヒートマップ」）を含み得る。この地図では、分析対象のデータ５０８から識別された統計に基づいて、米国の異なる地域が色を変えて示される。この統計は、これらの場所が、データに対応付けられていると識別された頻度を示していてもよい。説明のための一例において、データはオンライン小売店における購入トランザクションを表わしていてもよく、この場合の各トランザクションは、たとえば配送先／請求先住所に基づいてまたは記録されているＩＰアドレスに基づいて）場所に対応付けることができる。プロファイル概要５０２は、購入トランザクションを表わすデータの処理に基づいてトランザクションの場所を示してもよい。いくつかの実施形態において、ビジュアライゼーションをユーザ入力に基づいて修正することにより、ユーザがデータを検索して有益な相関関係を見出すのを支援することができる。これらの特徴を以下でさらに説明する。
【０１１５】
図５Ｂ、図５Ｃおよび図５Ｄは、データセットの対話型データ強化の結果の例を示す。図５Ｂはプロファイルメトリックパネル５４２を含み得るユーザインターフェイス５４０を示す。パネル５４２は、選択されたデータソースに対応付けられたメトリックの要約を示すことができる。図５Ｃに示されるように、いくつかの実施形態において、プロファイルメトリックパネル５６０は、データセット全体ではなく特定列のメトリック５６２を含み得る。たとえば、ユーザは、ユーザのクライアントデバイス上で特定の列を選択すればよく、そうすると対応する列のプロファイル５６４を表示することができる。この例において、プロファイラは、column_0008と、知識ソースの既知の都市との間の一致が９２％であることを示す。いくつかの実施形態において、確率が高いことにより、変換エンジンが自動的にcol_0008のラベルを「city」にするようにできる。
【０１１６】
図５Ｄは、全体的なメトリック５８２（たとえばデータセット全体に関連するメトリック）と、列ごとのビジュアライゼーション５８４とを含み得るプロファイルメトリックパネル５８０を示す。列ごとのビジュアライゼーション５８４は、ユーザによって選択されるおよび／または使用されることにより、（たとえばクリック、ドラッグ、スワイプ等によって）データをナビゲートすることができる。上記の例は、小さなデータセットへの簡単な変換を示す。同様のまたはより複雑な処理を、何十億もの記録を含む大きなデータセットに自動的に適用することもできる。
【０１１７】
図６は、本発明の実施形態に従う代表的なグラフを示す。いくつかの実施形態において、テキストデータの文字ストリングを識別することが有用な場合がある。文字ストリングは、（正規表現のような）埋込み構文がないので「文字通りの意味で」扱うことができるストリングであってもよい。データセット内の文字ストリングを検索するとき、完全ストリングマッチングを実行する。文字ストリングマッチングは、１つ以上の文字ストリング

を１つのデータ構造で表わすことによって実行できる。このデータ構造をグラフマッチング法と組合わせて用いてもよい。グラフマッチング法では、入力ストリングに対して一回のパスを実行することにより、入力ストリングの、一致するすべての文字ストリングを同時に見つけ出す。これは、すべての文字ストリングを発見するのにテキストに対して一回だけパスを実行すればよいので、マッチング効率を改善する。

【０１１８】

いくつかの実施形態において、グラフマッチング法をエイホーコラシック（Aho-Corasik）アルゴリズムの変形として実装してもよい。グラフマッチング法は、文字ストリングをツリー状のデータ構造に格納し、入力テキストにおいてそれまでにわかっている文字との可能な一致すべてを探してツリーを繰返しトラバースすることによって、機能する。データ構造は、そのノードが文字ストリングの文字であるツリーであってもよい。すべての文字ストリングの最初の文字はルートノードの子であってもよい。すべての文字ストリングの２番目の文字は、最初の文字に対応するノードの子であってもよい。図６は、以下のワード：can(1)、car(2)、cart(3)、cat(4)、catch(5)、cup(6)、cut(7)、およびten(8)のツリーを示す。各ワードの最後のノードにはそのワードに対応する数字が示されている。

【０１１９】

いくつかの実施形態において、グラフマッチング法は、特定の一致のリストを追跡することができる。部分一致は、ツリー内のノードへのポインタと、入力ストリング内の、部分一致が導入されている場所に対応する文字オフセットとを含み得る。グラフマッチング法は、オフセット１の文字のルートノードに対応する１つの部分一致で初期化することができる。最初の文字を読み出したときに、所与の文字を有する子がルートノードにあるか否かを調べる。このような子が存在する場合は、部分一致のノードを子ノードに進める。次の文字を読出す前に、ルートノードに対応する新たな部分一致およびオフセット２の文字を部分一致のリストに追加する。これをすべての文字に対して繰返す。文字ごとに、リスト内のすべての部分一致を現在の文字で評価する。部分一致のノードに、現在の文字に対応する子がある場合、部分一致は維持される。そうでなければ削除しなければならない。部分一致が維持される場合は、ノードから、現在の文字に対応する子ノードに進む。いくつかの実施形態において、部分一致のノードがワードの最後（図６において数字で示される）に対応する場合は完全一致が生成され、これを戻された値のリストに加えればよい。次の文字に進む前に、ルートノードと次の文字の文字オフセットとに対応する新たな部分一致を作成すればよい。

【０１２０】

図７は、本発明の実施形態に従う代表的な状態表を示す。説明のために、表７００は、入力ストリング「cacatch」を調べるときのグラフマッチング法の状態を示す。以下の表は、部分一致評価後の、入力ストリングの各文字においてこの方法が格納する内部状態を示す。図７に示されるように、グラフマッチング法は、文字２から４の「cat」(ワード４）と、文字２から６のワード「catch」（ワード５）を識別することができる。部分一致は、文字オフセットと、文字によって与えられる、ツリー内のノードとを含む対として表わされる。いくつかの実施形態において、文字オフセットは部分一致の識別子の役割を果たすことができ（たとえば、ノードが更新されても文字オフセットは一定のままであり得る）、各文字オフセットで導入されるの部分一致は１つだけである。

【０１２１】

いくつかの実施形態において、文字ごとに新たな部分一致を導入することができる。ツリー６００のルートノードは２つの子（「ｃ」および「ｔ」）を有するので、新たに生成された部分一致は、行１、３、５、および６に示すように文字「ｃ」および「ｔ」が見出されると進められる。行３で、部分一致「（１，ａ）」は進められない。なぜなら、「ａ」には文字「ｃ」を有する子がないからである。一方、部分一致「（３，ａ）」は行５の「（３，ｔ）」に進められる。ノード「ａ」には文字「ｔ」を有する子があるからである。この時点において、部分一致「（３，ｔ）」は、「ｔ」ノードの数字４で示される「ｃ

ａｔ」というワードの完全一致ではない。したがって、「ｃａｔ」というワードについて、３から５の一致が見出される。

【０１２２】

図８は、本発明の実施形態に従う大文字と小文字を区別しない（ケースインセンシティブ）グラフの例を示す。いくつかの実施形態において、ケースインセンシティブマッチングを、第２のケースインセンシティブツリー上で実行することができる。図７に示されるツリーのようなツリーを、各文字のアッパーケース（大文字）とロアーケース（小文字）双方を含む格子に変換し、１つの文字のいずれの「ケース」も次の文字の双方の「ケース」を指すようにすることができる。ツリー８００は、「ｃａｎ」というワードのケースインセンシティブツリー／格子を示す。図８に示されるように、「ｃａｎ」というワードについて、ケースのすべての組合わせ（たとえば、「Ｃａｎ」、「ＣＡｎ」、「ＣａＮ」、「ｃａＮ」等）の経路がツリーに存在する。

【０１２３】

しかしながら、ケースセンシティブ（大文字と小文字を区別する）エントリとケースインセンシティブエントリは、同一のツリーには存在できない。なぜなら、ケースインセンシティブエントリはケースセンシティブエントリの大文字と小文字の区別に対して有害な影響を与えるからである。加えて、ケースインセンシティブマッチングを実施するときに、すべての文字がロワーケース（小文字）とアッパーケース（大文字）を有するとは限らない。したがって、ツリーは必ずしも対応する文字の対を含まない。一例として、ツリー８０２は、ケースインセンシティブなやり方でツリーに「ｂ２ｂ」というワードを加えた場合の一致構造を示す。

【０１２４】

いくつかの実施形態において、ケースインセンシティブマッチングは、ケースインセンシティブマッチングのために追加された文字ストリングを含む第２のケースインセンシティブツリーを加えることによってサポートできる。次に、グラフマッチングを、文字ごとに２つの部分一致を、２つのツリー各々のルートノードに対応する部分マッチのリストに加えることを除いて、上記のように実行すればよい。

【０１２５】

図９は、本発明の実施形態に従うデータセットの類似性を示す図を示す。本発明の実施形態は、データセットを意味的に分析してこれらのデータセット間の意味類似性を判断することができる。データセット間の意味類似性は、意味メトリックとして表現できる。たとえば、アイテムの顧客リストＣと参照リストＲが与えられた場合、ＣとＲの間の「意味類似性」は、Jaccard係数、Sorensen-Dice係数（Diceの係数のSorensen係数とも呼ばれる）、およびTversky係数等の周知の多数の関数を用いて計算できる。しかしながら、既存の方法は、近いデータセットのマッチングを適切に実施しない。たとえば、図９に示されるように、すべての州都は都市であるが、すべての都市が州都である訳ではない。したがって、５０都市のリストを含むデータセットＣが与えられた場合、そのうちの４９が州都であり１つは州都でない都市であり、Ｃは「州都のリスト」ではなく「都市のリスト」とマッチングする必要がある。Jaccard係数およびDice係数等の従来技術の方法は、データセットを対称に扱う、すなわち、これらの方法は顧客データと参照データを区別をしない。そうすると、結果として、顧客データの中で参照データに対するマッチングが行なわれていないデータがあるという状況が生じ得る。

【０１２６】

いくつかの実施形態において、類似性メトリックを求める方法は、自然対数を用いることにより、参照データセットのサイズの変動を考慮する。結果として、顧客リストが１００個のアイテムを有する場合、顧客リストのアイテムすべてとマッチングする１０００個のアイテムの参照リストは、すべての顧客アイテムと一致する１０，０００個のアイテムの参照リストの２倍（かつ１０倍未満）のアイテムである。類似性メトリックを求める方法を記述する式を以下に示す。

【０１２７】

【数２】

$$\frac{|R\cap C|}{|C|}-\alpha\frac{\ln(1+|R-C|)}{1+\ln(1+|R-C|)}-\beta\begin{cases}0, & \text{Rはキュレートされている}\\ 1, & \text{Rはキュレートされていない}\end{cases}$$

【０１２８】

式中、Ｒは参照データセット、Ｃは顧客データセット、αおよびβは調整可能な係数である。いくつかの実施形態において、デフォルトはα＝０．１およびβ＝０．１である。

【０１２９】

この方法は、特定の望ましくないデータセットの特徴に負の重み付けをすることによって類似性マッチングを改善する。たとえば、参照セットのサイズが増すと、α項が大きくなり、類似性メトリックが減じられる。加えて、キュレートされた参照データセットの場合（一般的に高い値のデータセットであると想定される）、β項は０である。しかしながら、キュレートされていないデータセットの場合、β項は１であり、類似性メトリックが大幅に減じられる。

【０１３０】

いくつかの実施形態において、この方法に頂点ランク（vertex rank）を取入れることができる。そうすると、この方法が結果として正規化された類似性メトリックになることはない。よって、類似性メトリックは以下のように負の重みで乗算される。

【０１３１】

【数３】

$$\left(\sum_{x\in R\cap C}x_{vertex\_rank}\right)\left(1-\alpha\frac{\ln(1+|R-C|)}{1+\ln(1+|R-C|)}\right)\left(1-\beta\begin{cases}0, & \text{Rはキュレートされている}\\ 1, & \text{Rはキュレートされていない}\end{cases}\right)$$

【０１３２】

図１０は、本発明の実施形態に従う異なる知識ドメインの知識スコアリングを表示するグラフィカルインターフェイス１０００の例を示す。先に述べたように、グラフィカルインターフェイス１０００は、データ強化サービス３０２によって表示されたドメインのマッチングに関するデータをグラフィカルなビジュアライゼーションを表示できる。グラフィカルインターフェイス１０００は、異なる知識ドメインについてスコアリングの式に基づく統計をユーザに提供するデータを示す。知識ドメインは、列１００２（「ドメイン」）において識別されるもののような特定のカテゴリ（たとえばドメイン）に関連する複数の用語を含み得る。ドメイン１００２は各々、複数の用語を含み得るものであり、知識ソースによって規定されてもよい。知識ソースをキュレートすることにより、ドメイン１００２各々に対応付けられた用語を維持してもよい。グラフィカルインターフェイス１０００は、ドメイン１００２を規定するさまざまな値と、ドメイン各々に対してスコアリングの式を用いて求めたマッチングスコア１０１６（「スコア」）とを示す。ドメイン１００２は各々、度数値１００４（たとえば入力データセットとドメイン内の用語との間で一致する用語の度数）、母集団値１００６（たとえば入力データセット内の用語の数）、一マッチング値１００８（たとえば度数値１００４を母集団で割ることに基づいて計算されたドメインに一致する用語のパーセンテージを示す）、固有マッチング値１０１０（たとえば入力データセットとドメインとの間で一致する異なる用語の数）、サイズ１０１２（たとえばドメイン内の用語の数を示すドメインカウント）および、選択値（たとえばドメインに対して選択された用語のパーセンテージを示す）等の値を有し得る。図示されていないが、グラフィカルインターフェイス１０００は、ドメインがキュレートされた程度を示す値の範囲（たとえば０．０～１００．００）の中の一定の値を示すキュレーションレベルを示してもよい。ドメインに関する１つ以上の値に基づいて、スコア１０１６を、上記スコア（ｆ，ｐ，ｕ，ｎ，ｃ）等の類似性を求めるための関数を用いて計算してもよい。ドメインに関する値に加えて、スコアは、精密な測定値を提供し得る。これによってユー

ザは入力データセットに最も一致するドメインに関してより適切な評価を行なうことができる。最も近いマッチングドメインを用いて、ドメインに関するデータに名称を付けてもよくまたはドメインに関するデータを入力データセットに対応付けてもよい。

【０１３３】

図１１～図１８を参照して説明する実施形態のようないくつかの実施形態は、フローチャート、フロー図、データフロー図、構造図、またはブロック図で示されるプロセスとして説明し得る。フローチャートはオペレーションを逐次プロセスとして説明する場合があるが、これらのオペレーションのうちの多くは並列してまたは同時に実行し得る。加えて、オペレーションの順序は構成し直してもよい。プロセスは、そのオペレーションが完了したときに終了するが、図面には含まれていないさらに他のステップを有することがある。プロセスは、方法、関数、手順、サブルーチン、サブプログラム等に対応し得る。プロセスが関数に対応する場合、その終わりは、その関数の、呼出し関数またはメイン関数へのリターンに対応し得る。

【０１３４】

図１１～図１８を参照しながら説明するプロセスのような本明細書に示すプロセスは、１つ以上の処理ユニット（たとえばプロセッサコア）によって実行されるソフトウェア（たとえばコード、命令、プログラム）、ハードウェア、または、これらを組合わせたもので、実装し得る。ソフトウェアはメモリ（たとえばメモリデバイス、非一時的なコンピュータ読取可能な記憶媒体）に格納されていてもよい。いくつかの実施形態において、本明細書のフローチャートに示されるプロセスは、データ強化サービス、たとえばデータ強化サービス３０２のコンピューティングシステムによって実装できる。本開示における特定の一連の処理ステップは限定を意図しているのではない。代替の実施形態に従って他の順序のステップも実施し得る。たとえば、本発明の代替の実施形態は、先に概要を述べたステップを他の順序で実行し得る。加えて、図面に示される個々のステップは、個々のステップに適したさまざまな順序で実行し得る複数のサブステップを含み得る。さらに、特定の用途に応じてその他のステップを追加してもよく削除してもよい。当業者は数多くの変形、修正および代替例を認識するであろう。

【０１３５】

いくつかの実施形態のある側面において、図１１～図１８の各プロセスは、１つ以上の処理ユニットによって実行できる。１つの処理ユニットは、シングルコアもしくはマルチコアプロセッサ、プロセッサの１つ以上のコア、またはその組合わせを含む、１つ以上のプロセッサを含み得る。いくつかの実施形態において、１つの処理ユニットは、グラフィックプロセッサ、デジタル信号プロセッサ（ＤＳＰ）等の専用コプロセッサを１つ以上含み得る。いくつかの実施形態において、処理ユニットのうちの一部またはすべてを、特定用途向け集積回路（ＡＳＩＣ）またはフィールドプログラマブルゲートアレイ（ＦＰＧＡ）等のカスタマイズされた回路を用いて実装することができる。

【０１３６】

図１１は、自動化されたデータ分析の例を示す。１１００で示されるように、教師なし機械学習技術、たとえばWord2Vecを用いて入力データセットを分析することができる。Word2Vecは、テキスト入力（たとえば大きなデータソースからのテキストコーパス）を受けて各入力ワードのベクトル表現を生成できる。次に、得られたモデルを用いて、任意の入力ワードセットのがワードがどれほど近い関連性があるかを識別できる。たとえば、Ｋ平均クラスタリング（またはその他のベクトル分析）を用いて一組の入力ワードに対応するベクトルを分析し、入力ワードがどれほど類似するかを、ベクトル空間内の対応するベクトルがどれほど「近い」かに基づいて、判断することができる。

【０１３７】

１１００に示されるように、入力された一組のワードは、「Bridgestone」、「Firestone」、および「Michelin」を含み得る。大きなテキストコーパス（たとえばニュースアグリゲータまたはその他のデータソース）を用いて構築されたWord2Vecモデルを利用して、対応する数値ベクトルを入力ワードごとに識別することができる。これらのベクトルを分

析する際に、ベクトルはベクトル空間内で「近い」（ユークリッドの意味で）と判断してもよい。図１１に示されるように、ベクトル空間内で３つの入力ワードが互いに近接してクラスタリングされている。これによって入力ワードが関連していることを識別することができるが、Word2Vecを用いてワードを説明するラベル（たとえば「タイヤメーカー」）を識別することはできない。

【０１３８】

１１０２に、キュレートされたデータソースを用いてカテゴリ分類する方法が示される。キュレートされたデータソース（Max Planck Institute for InformaticsのＹＡＧＯ等）は、オントロジ（ontology）（たとえば特定のドメインに対して存在するエンティティのタイプ、特性、および相互関係の正式名称および定義）を提供することができる。キュレートされたデータソースを用いて、入力データセット内のエンティティをグラフマッチングを通して識別してもよい。これにより、多様な種（たとえばワード）を含む入力データセットに対して属ラベル（たとえばカテゴリ）を特定することができる。しかしながら、１１０２に示されるように、属ラベルは不完全または不正確である場合がある（たとえば異なるキュレータは異なるやり方で種をカテゴリ分類する場合がある）。３０２に示される例において、このような不正確さがあると、結果として、同一組の入力ワードが異なる属に対してマッチングされることが起こり得る（たとえばBridgestoneとMichelinがタイヤメーカーという属に対してマッチングされ、Firestoneがタイヤ産業の人々という属に対してマッチングされる）。これは、種データの入力に対して属を特定する方法を提供するが、特定される属の精度は、キュレートされたデータソースの精度と完全性によって限定される。

【０１３９】

図１２は、トライグラムモデリング１２００の一例を示す。従来、トライグラムは自動スペル訂正を実行するために使用されてきた。図１２に示されるように、各入力ワードはトライグラムに分解することができ、トライグラムでインデックスが作成されそのトライグラムを含むワードを含む表を作成することができる（たとえばトライグラム「ＡＮＴ」は「ａｎｔｉｑｕｅ」、「ｇｉａｎｔ」等に対応付けられる）。自動スペル訂正に使用されるときは、辞書がデータソースとして使用され、トライグラムを用いて、入力されたスペルミスに最も類似するワードを特定することができる。たとえば、類似性メトリックは、スペルミスがある入力ワードとデータソース（たとえば辞書）からのワードとで共有されるトライグラムの数で類似性を示すことができる。

【０１４０】

ある実施形態に従うと、カテゴリを識別するためにトライグラムモデリングを用いてワードの組を比較することができる。以下で図１３を参照しながらさらに説明するように、トライグラムを用いることにより、データソース（キュレートされたＹＡＧＯデータソース等）のインデックス付の表（データベースインデックスと同様）を作成することができる。インデックスグ付の表において、各トライグラムは、そのトライグラムを含む複数のワードに対応付けられた主キー（primary key）であってもよい。この表のそれぞれの列は、トライグラムを含むワードに関連付けられたそれぞれのカテゴリに対応していてもよい。入力データセットを受けると、データセット内の各ワードをトライグラムに分割してインデックス付の表と比較することにより、一致するワードを識別することができる。次に、一致するワードをデータソースと比較することにより、カテゴリに対するベストマッチを識別することができる。トライグラムモデリングにおける統計的な一致の判断は、「Scalable string matching as a component for unsupervised learning in semantic meta-model development」と題された米国特許出願第１３／５９６，８４４号（Philip Ogren他）に記載されている技術を用いて行なってもよい。

【０１４１】

図１３は、本発明の実施形態に従うカテゴリラベル付け１３００の一例を示す。図１３に示されるように入力データセット１３０２を受けることができる。入力データセット１３０２は、たとえばテキストストリングの列を含み得る。この例において、入力データセ

ット１３０２は、たとえばテキストストリングの列を含み得る。この例において、入力データセット１３０２は、「Bridgestone」、「Firestone」、および「Michelin」というストリングを含む。１３０４で、データ分析ツール（たとえばWord2Vec）を用いて、入力データセットと類似するデータを識別することができる。いくつかの実施形態において、データ分析ツールは、データソース、たとえばニュースアグリゲーションサービスから取得したデータを予め処理することにより、ワード増補リストを作成することができる。次に、入力データセット１３０２をワード増補リストと比較することにより、類似するワードを識別してもよい。たとえば、Word2Vecを用いて、ベクトルを、入力データセット１３０２に含まれるストリングごとに識別することができる。ベクトル分析法（たとえばK平均クラスタリング）を用いて、入力データセットのワードに「近い」、ワード増補リストのその他のワードを識別できる。ワード増補リストからの類似するワードを含む増補データセット１３０６を作成できる。図１３に示されるように、これは、「Goodyear」というワードを入力データセット１３０２に追加することによって増補データセット１３０６を作成することを含む。

【０１４２】

いくつかの実施形態において、次に増補データセット１３０６を知識ソース１３０８と比較することにより、この増補データセットと一致するカテゴリを識別することができる。図１３に示されているように、知識ソース１３０８は、カテゴリによって組織されたデータを含み得る。いくつかの実施形態において、各カテゴリをルートノードで表わすことができ、各ルートノードはそのカテゴリに属するデータを表わす１つ以上のリーフノードを有し得る。たとえば、知識ソース１３０８は少なくとも２つのカテゴリとして「タイヤメーカー」と「タイヤ産業の人々」を含む。各カテゴリはそのカテゴリに属するデータを含む（MichelinとBridgestoneはタイヤメーカーに属し、Harvey Samuel Firestoneはタイヤ産業の人々に属する）。

【０１４３】

ある実施形態に従うと、増補データセットは、Jaccard係数等の類似性メトリックを用いて知識ソースカテゴリと比較することができる。類似性メトリックは、１つのリストを別のリストと比較して、２つのデータセットの類似性を示す値を割当てることができる。

【０１４４】

【数４】

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}.$$

【０１４５】

知識ソース１３０８が不完全または不正確な場合でも、類似性メトリックによって「最も適合する」カテゴリを識別することができる。

【０１４６】

いくつかの実施形態において、類似性メトリックは、Tanimotoメトリックに基づいて計算してもよい。これは、ブールベクトルについてJaccard係数の数値ベクトルの一般化を計算する。以下の式はTanimotoメトリックを表わす。

【０１４７】

【数５】

$$\frac{X\cdot Y}{|X|^2+|Y|^2-X\cdot Y}$$

【０１４８】

ドットはベクトルドット積を表わす。

上に示されるように、Jaccard係数は、２つのデータセットＡおよびＢの類似性を、こ

れらのデータセットＡおよびＢの共通集合の大きさの、これらのデータセットの合併集合の大きさに対する比率を求めることによって判断できる。１３１４に示されるように、増補データセット１３０６と「タイヤメーカー」というカテゴリとの共通集合は２（MichelinおよびBridgestone）であり、合併集合のサイズは４なので、類似性メトリックは０．５である。増補データセット１３０６と、タイヤ産業の人々というカテゴリとの共通集合は１（Firestone）であり、合併集合のサイズは４なので、類似性メトリックは０．２５である。よって、「ベストマッチ」は、「タイヤメーカー」であり、データ強化サービスは、「タイヤメーカー」の列にラベル付けすることによって、入力データセットを強化できる。

【０１４９】

上記のように、その他の類似性メトリックを、Jaccard係数に加えてまたはその代わりに用いてもよい。当業者は、上記の技術に対してどの類似性メトリックも使用し得ることを認識するであろう。代替の類似性メトリックのいくつかの例は、限定されないが、Dice-Sorensen係数、Tversky係数、Tanimotoメトリック、およびコサイン類似度メトリックを含む。

【０１５０】

本発明の実施形態は、「ビッグデータ」システムおよびサービスを参照しながら一般的に説明されている。これは、説明を明確にすることが目的であって限定ではない。当業者は、本発明の実施形態が「ビッグデータ」という文脈の範囲外のその他のシステムおよびサービスに対しても実装し得ることを理解するであろう。

【０１５１】

いくつかの実施形態において、トライグラム統計分析１３１２を増補データセット１３０６に適用してもよい。トライグラムモデリングモジュールは、知識ソース１３０８を、前処理１３１０することにより、主キーがトライグラムであり各列が知識ソースの少なくとも１つのワードを含むインデックス付の表にする。トライグラムが同一であるそれぞれの列は、知識ソース１３０８内の異なるカテゴリに対応し得る。増補データセット１３０６内のワードごとにトライグラムを作成してインデックス付のトライグラム表と比較すればよい。この比較の結果、知識ソース１３０８の互いに関連するワードのリストが得られ、関連するワードのうち最も一致度が高いものをトライグラム一致データセットに加えればよい。次に、このトライグラム一致データセットを上述のように知識ソース１３０８のカテゴリと比較することにより、最も一致度が高いカテゴリを識別する。次に、入力データセット１３０２が下流の１つ以上の下流データターゲットに対して公開されるときに、一致カテゴリを用いて入力データセット１３０２にラベル付けすることができる。

【０１５２】

いくつかの実施形態において、知識ソース１３０８から生成されたインデックス付トライグラム表は、トライグラムを有する主インデックス列（アルファベット順にソートされる）と、そのリーフノードに同じトライグラムがある各カテゴリおよびサブカテゴリのリストを有する第２の列とを含み得る。

【０１５３】

図１４～図１６は、本発明の実施形態に従うランク付けされたカテゴリを求めるための類似性分析を示す。図１４は、自動データ分析のためのシステム１４００を示す。システム１４００により、用語の入力データセット１４０２について、カテゴリとこれらのカテゴリに関連付けられたランキングとを見出すことができる。自動データ分析を実装することにより、知識ソースから得たキュレートされたデータを用いてデータ１４０２をカテゴリ分類してもよい。

【０１５４】

図示のように、入力データセット（たとえばデータ１４０２）は、ユーザが提供する入力ソースから取得できる。データ１４０２は、ソースに応じて２つ以上の列にフォーマットしてもよい。データ強化サービス１４０８は、Ｊａｖａ（登録商標）仮想マシン（Java virtual machine）（ＪＶＭ）等の仮想コンピューティング環境を用いて実装し得る。デ

ータ強化サービス１４０８は入力としてデータ１４０２を受付てもよい。データ強化サービス１４０２は、キュレートされたデータ１４０６（たとえばキュレートされたリスト）をキュレートされたデータソース（Max Planck Institute for InformaticsのＹＡＧＯ等）から取得してもよい。キュレートされたデータ１４０６はオントロジ（たとえば特定のドメインに対して存在するエンティティのタイプ、特性、および相互関係の正式名称および定義）を提供することができる。キュレートされたデータの例は、地理的な位置が異なる郵便番号を示す地理的名称（geoname）を含み得る。

【０１５５】

データ強化サービス１４０８は、キュレートされたデータソースを用いて、データ１４０２を意味分析することにより、キュレートされたデータ１４０６に対する類似性または近接性を判断することができる。データセット間の意味類似性は、類似性メトリック（たとえば値）で表わすことができる。たとえば、入力データセットとキュレートされたリストが与えられたとすると、この入力データセットとキュレートされたリストとの間の類似性は、Tverskyメトリック等の多数の比較関数を用いて計算できる。比較によって求めた類似性メトリックに基づいて、データ１４０２とキュレートされたデータ１４０６との比較から、近接性のランクを求めることができる。比較に基づいて、キュレートされたデータ１４０６内のカテゴリ１４０４を、類似性メトリックによって判断しランク付けすることができる。ランク付けされたカテゴリ１４０４を評価することにより、データ１４０２に対応付けることができる最高ランクのカテゴリを識別することができる。

【０１５６】

別の実施形態において、図１５は、自動データ分析のためのシステムのもう１つの例１５００を示す。システム１５００により、用語の入力データセット１５０２について、カテゴリとこれらのカテゴリに関連付けられたランキングとを見出すことができる。システム１５００は、図１４を参照しながら説明した類似性分析によって識別されたカテゴリが近接しておらずこれらが良好なマッチではないと思われると判断したときに実装されてもよい。

【０１５７】

図１５を参照して、ユーザから受けたデータ１５０２は、ソースに応じて２つ以上の列にフォーマットしてもよい。データ強化サービス１５２０は、Ｊａｖａ（登録商標）仮想マシン（ＪＶＭ）等の仮想コンピューティング環境を用いて実装し得る。自動データ分析を実装することにより、知識ソースから取得したキュレートされたデータを用いてデータ１５０２をカテゴリ分類してもよい。システム１５００は、類似性分析を実行する前にアグリゲーションサービスを利用してデータ１５０２を増補することにより、実装してもよい。データ１５０２は、入力データの供給元であるデータソースとは異なるソース（たとえばアグリゲーションサービス）から取得した増補データで増補してもよい。たとえば、入力データセットは、参照データセットのソースとは異なるニュースアグリゲーションサービス（たとえばGoogle（登録商標）ニュースコーパス）からのテキストコーパス等の知識ソースから得たデータで増補してもよい。たとえば、Word2Vec等のデータ分析ツールを用いて、知識ソースからのデータセットに含まれるものと意味的に類似するワード（たとえば同義語）を識別することができる。知識ソースから取得したデータを予め処理することによってワード増補リストを生成することができる。次に、入力データセットをワード増補リストと比較することによって類似するワードを識別すればよい。たとえば、Word2Vecを用いて、データ１５０２に含まれるストリングごとにベクトルを識別できる。ベクトル分析法（たとえばＫ平均クラスタリング）を用いると、ワード増補リストのその他のワードであって入力データセットのワードに「近接する」ワードを識別できる。ワード増補リストの類似するワードを含む増補データセットを生成することができる。入力データセットは増補データセットを用いて増補することができる。増補データを有する入力データセットは、図１５に示されるプロセスの残りの部分に対して用いてもよい。

【０１５８】

データ１５０２の増補後に、データ強化サービス１５０２は、データ１５０２を意味的

に分析することによって、知識ソース１５１６から取得したデータとの類似性または近接性を判断することができる。いくつかの実施形態において、知識ソース１５１６は、キュレートされたデータソースからのデータを提供してもよい。キュレートされたデータは、１つ以上のファイル内のキュレートされたカテゴリおよびタイプを含み得る。タイプは、データ１５０２についてカテゴリをより適切に識別するための用語の分類基準（taxonomy）を含み得る。いくつかの実施形態において、中間システム１５１４を実装することにより、キュレートされたリストを知識ソース１５１６から生成してもよい。システム１５は、オフラインモードのみにおいて、開発中の一回限りのオペレーションによってキュレートされたリストを生成してもよく、または、毎回システムが初期化されると表を最初から（知識ソース１５１６から取得したデータから）考案してもよい。

【０１５９】

キュレートされたデータは、分散記憶システム（たとえばＨＤＦＳ）に格納されてもよい。いくつかの実施形態において、キュレートされたデータはインデックス付きＲＤＤ１５１２に格納されてもよい。

【０１６０】

キュレートされたデータを、既に増補されているデータ１５０２と比較することにより、データ間の意味類似性を判断してもよい。データセット間の意味類似性は、類似性メトリック（たとえば値）で表わすことができる。たとえば、たとえば、入力データセットとキュレートされたリストが与えられたとすると、この入力データセットとキュレートされたリストとの間の類似性は、Tverskyメトリック等の多数の比較関数を用いて計算できる。比較によって求めた類似性メトリックに基づいて、データ１５０２とキュレートされたデータとの比較から、近接性のランクを求めることができる。データ１５０２について近接性のランクを用いることによりカテゴリを識別できる。カテゴリのランク付け１５１０により、データ１５０２に対応付けることができる最高ランクのカテゴリを識別してもよい。

【０１６１】

図１６は、入力データセットを、ＹＡＧＯ等の知識ソースから取得したキュレートされたデータの一組の分類と比較するプロセス１６００を示す。この一組の分類は、代替スペリングおよび代替分類１６３０を含み得る。この一組の分類は、サブクラスが上位カテゴリ１６２８（たとえば生物（living thing））に含まれる階層状に並べてもよい。たとえば、生物カテゴリ１６２８はサブクラスとして有機体（organism）１６２６を有してもよく、有機体のサブクラスは人（person）１６２２、１６２４である。人のサブクラスは人の種類（たとえば知識人（intellectual））であり、人の種類のサブクラスは訓練／専門職１６１８（たとえば学者（scholar））であり、学者のサブクラスは哲学者（philosopher）１６１６である。類似性分析のためにこの一組の分類を用いて入力データセットと比較してもよい。たとえば図１４および図１５を参照して述べたような入力データセットとのマッチング中に代替スペリングを選択的に重み付けしてもよい。入力データセットとの比較のために、より広いカテゴリまたは分類を用いて重みを判断してもよい。

【０１６２】

図１６に示される例において、主要な関係１６３２は、分類１６３０のうちの１つ１６１４を用いて識別してもよい。分類１６１４は、Aristotle１６０６等の好ましいスペリングの用語１６０４と、Aristotel１６０８等のスペルミス１６０２とを含み得る。スペルミス、たとえばAristotel１６０８は、用語、たとえばAristotle１６０６の正しいスペル１６１０（たとえばAristotle）に対応するラベル１６１２にマッピングしてもよい。

【０１６３】

図１７および図１８は、本発明のいくつかの実施形態に従う類似性分析のプロセスのフローチャートを示す。いくつかの実施形態において、本明細書の図１７および図１８等のフローチャートに示されるプロセスは、データ強化サービス３０２のコンピューティングシステムによって実装することができる。フローチャート１７００は類似性分析のプロセスを示し、このプロセスでは、入力データを１つ以上の参照データセットと比較すること

により、それらの類似性を判断する。類似性は、データ強化サービスのユーザが関連するデータセットを識別して入力データセットを強化できるようにする類似の程度として示されてもよい。

【０１６４】

フローチャート１７００は、１つ以上の入力データソース（たとえば図３のデータソース３０９）から入力データセットを受けるステップ１７０２から始まる。いくつかの実施形態において、入力データセットは、１つ以上のデータ列にフォーマットされる。

【０１６５】

ステップ１７０４で、入力データセットを、参照ソースから取得した１つ以上の参照データセットと比較してもよい。たとえば、リソースソースは知識ソース３４０のような知識ソースである。入力データセットを参照データセットと比較することは、２つのデータセット間の比較において、各用語を個別にまたはまとめて比較することを含み得る。１つの入力データセットは１つ以上の用語を含み得る。１つの参照データは１つ以上の用語を含み得る。たとえば、参照データセットは、カテゴリ（たとえばドメインまたは属）に対応付けられた用語を含む。参照データセットを知識サービスによってキュレートしてもよい。

【０１６６】

いくつかの実施形態において、入力データを、入力データの供給元であるデータソースと異なるソースから取得した増補データで増補してもよい。たとえば、入力データセットを、参照データセットのソースと異なる知識ソースからのデータで増補してもよい。たとえば、Word2Vec等のデータ分析ツールを用いて、ニュースアグリゲーションサービスからのテキストコーパスのような知識ソースからの入力データセットに含まれるものと意味的に類似するワードを識別することができる。知識ソースから取得したデータを前処理してワード増補リストを生成することができる。次に、入力データセットをワード増補リストと比較することにより、類似するワードを識別してもよい。たとえば、Word2Vecを用いて、入力データセット６０２に含まれるストリングごとにベクトルを識別することができる。ベクトル分析法（たとえばＫ平均クラスタリング）を用いて、入力データセット内のワードに「近い」、ワード増補リストの他のワードを識別できる。ワード増補リストから、類似するワードを含む増補データセットを生成することができる。入力データセットは増補データセットを用いて増補することができる。増補データを有する入力データセットは、図１７に示されるプロセスの残りの部分に対して使用してもよい。

【０１６７】

いくつかの実施形態において、入力データセットを１つ以上の参照データセットと比較するのに用いるデータ構造を生成してもよい。このプロセスは、比較の対象である１つ以上の参照データセットのうちの少なくとも一部を表わすデータ構造を生成することを含み得る。データ構造内の各ノードは、１つ以上の参照データセットから抽出した１つ以上のストリングにおける異なる文字を表わしていてもよい。入力データセットは、データ構造を生成した１つ以上の参照データセットと比較してもよい。

【０１６８】

ステップ１７０６で、類似性メトリックを、１つ以上の参照データセット各々について計算してもよい。類似性メトリックは、入力データセットとの比較における１つ以上の参照データセット各々の類似の程度を示してもよい。

【０１６９】

いくつかの実施形態において、類似性メトリックは、１つ以上の参照データセット各々について計算されたマッチングスコアである。たとえば、１つの参照データセットについてのマッチングスコアは、第１の値がこの参照データセットに関するメトリックを示し第２の値が入力データセットと参照データセットとの比較に基づくメトリックを示す、１つ以上の値を用いて計算してもよい。上記１つ以上の値は、入力データセットとデータセットとの間で一致する用語の度数値と、データセットの母集団値と、データセットの固有マッチング値と、入力データセットとデータセットとの間で一致する異なる用語の数を示す

固有マッチング値と、データセット内の用語の数を示すドメイン値と、データセットのキュレーションの程度を示すキュレーションレベルとを含み得る。マッチングスコアは、１つ以上の値を用いてスコアリング関数（１＋ｃ／１００）＊（ｆ／ｐ）＊（ｌｏｇ（ｕ＋１）／ｌｏｇ（ｎ＋１））を実装することによって計算してもよい。スコアリング関数の変数は、度数値を表わす「ｆ」と、キュレーションレベルを表わす「ｃ」と、母集団値を表わす「ｐ」と、固有マッチング値を表わす「ｕ」と、ドメイン値を表わす「ｎ」とを含み得る。

【０１７０】

いくつかの実施形態において、類似性メトリックは、入力データセットとの比較における１つ以上の参照データセットの共通部分のカーディナリティに基づく値として計算されてもよい。この値はカーディナリティによって正規化されてもよい。この値は、上記１つ以上の参照データセットのサイズに基づく第１のファクタだけ減じられてもよく、この値は、上記１つ以上の参照データセットのタイプに基づく第２のファクタだけ減じられてもよい。

【０１７１】

いくつかの実施形態において、上記１つ以上の参照データセットのうちの各参照データセットの類似性メトリックを、上記入力データセットとこの参照データセットとのコサイン類似度を求めることによって計算してもよい。上記のように、入力データセットと、参照データセットの１つ以上の用語との間のコサインメトリック（たとえばコサイン類似度またはコサイン距離）は、知識ソースから取得した参照データセット（たとえばドメインまたは属）と用語の入力データセットとの間のコサイン角度として計算してもよい。コサイン類似度に基づいて類似性メトリックを計算することにより、入力データセット内の各用語を、その用語と候補カテゴリとの間の類似性のパーセンテージを示す値のような、完全値整数分の１とみなし得る。

【０１７２】

ステップ１７０８において、入力データセットと１つ以上の参照データセットとの間の一致を、類似性メトリックに基づいて識別する。いくつかの実施形態において、この一致を識別することは、上記１つ以上の参照データセット各々について計算した類似性メトリックに基づいて、類似性の程度が最大である、上記１つ以上の参照データセットのうちの参照データを決定することを含む。上記１つ以上の参照データセット各々について計算した類似性メトリックを相互に比較することにより、類似性メトリックが最も近い一致を示す参照データセットを識別してもよい。最も近い一致は、最大値を有する類似性メトリックに対応するものとして識別してもよい。入力データセットは、類似の程度が最大である参照データセットに含まれるデータを含むように修正してもよい。

【０１７３】

入力データセットを、この入力データセットを説明またはラベル付けする用語（たとえばドメインまたはカテゴリ）のような他のデータに対応付けてもよい。他のデータは、キュレートされていてもよい参照データセットに基づいて決定してもよい。他のデータは、参照データセットの提供元であるソースから取得してもよい。

【０１７４】

ステップ１７１０で、１つ以上の参照データセット各々について計算され、入力データセットとこの１つ以上の参照データセットとの間の識別された一致を表わす類似性メトリックを示す、グラフィカルインターフェイスを生成してもよい。類似性メトリックがマッチングスコアであるいくつかの実施形態において、グラフィカルインターフェイスはマッチングスコアの計算に使用された値を示す。

【０１７５】

ステップ１７１２で、グラフィカルインターフェイスを用いてグラフィカルなビジュアライゼーションをレンダリングしてもよい。たとえば、グラフィカルなビジュアライゼーションのレンダリングを生じさせるグラフィカルインターフェイスを表示してもよい。グラフィカルインターフェイスは、如何にしてグラフィカルなビジュアライゼーションをレ

ンダリングするかを判断するのに使用されるデータを含み得る。いくつかの実施形態において、グラフィカルインターフェイスを、レンダリングのために別のデバイス（たとえばクライアントデバイス）に送ってもよい。グラフィカルなビジュアライゼーションは、１つ以上の参照データセット各々について計算した類似性メトリックを示してもよく、入力データセットと１つ以上の参照データセットとの間の識別された一致を示してもよい。グラフィカルなビジュアライゼーションの例は、図５および図１０を参照しながら説明されている。

【０１７６】

いくつかの実施形態において、入力データセットを、１つ以上の参照データセット各々について計算され入力データセットとこの１つ以上の参照データセットとの間の識別された一致を表わす類似性メトリックを示すマッチング情報とともに格納してもよい。

【０１７７】

さらに、いくつかの実施形態において、フローチャート１７００に示されるプロセスは、入力データの増補後において他の１つ以上のステップを含み得る。増補された入力データセットを用いて参照データセットとの一致を識別してもよい。このような実施形態において、このプロセスは、１つ以上の参照データセットに基づいてインデックス付トライグラム表を生成することを含み得る。増補された入力データセット内のワードごとに、そのワードのトライグラムを作成し、各トライグラムをインデックス付トライグラム表と比較し、インデックス付トライグラム表の中の、トライグラムのうちの第１のトライグラムと一致するトライグラムに対応付けられたワードを識別し、そのワードをトライグラム増補データセットに格納する。トライグラム増補データセットを１つ以上の参照データセットと比較してもよい。この比較に基づいて、トライグラム増補データセットと１つ以上の参照データセットとの間の一致を判断してもよい。ステップ１７０８において入力データセットと１つ以上の参照データセットとの間の一致を識別することは、比較に基づいてトライグラム増補データセットと１つ以上の参照データセットとの間の一致を用いることを含み得る。

【０１７８】

フローチャートは、１つ以上のデータソースから入力データセットを受けるステップ１８０２から始まってもよい。ステップ１８０４で、入力データセットを、知識ソースによって格納されている１つ以上のデータセットと比較してもよい。入力データセットは１つ以上の用語を含み得る。上記１つ以上のデータセットは各々１つ以上の用語を含み得る。

【０１７９】

ステップ１８０６で、入力データセットと比較された１つ以上のデータセット各々について類似性メトリックを計算してもよい。いくつかの実施形態において、類似性メトリックを、１つ以上のデータセットのうちの各データセットについて、入力データセットとこのデータセットとの間のコサイン類似度を求めることによって計算する。コサイン類似度は、入力データセットと、この入力データセットと比較されているデータセットとの間のコサイン角度として計算してもよい。

【０１８０】

ステップ１８０８で、１つ以上のデータセットと入力データセットとの間の一致を判断してもよい。この一致は、１つ以上のデータセット各々について計算した類似性メトリックに基づいて判断してもよい。一致を判断することは、一組の類似性メトリックの中で最大値を有する類似性メトリックを識別することを含み得る。この一組の類似性メトリックは、上記１つ以上のデータセット各々について計算した類似性メトリックを含み得る。

【０１８１】

ステップ１８１０で、グラフィカルユーザインターフェイスを生成してもよい。グラフィカルユーザインターフェイスは、１つ以上のデータセット各々について計算した類似性メトリックを示してもよい。グラフィカルユーザインターフェイスは、１つ以上のデータセットと入力データセットとの間の一致を示してもよい。この一致は、１つ以上のデータセット各々について計算した類似性メトリックに基づいて判断される。ステップ１８１２

で、グラフィカルユーザインターフェイスをレンダリングすることにより、１つ以上のデータセット各々について計算した類似性メトリックを表示してもよい。グラフィカルユーザインターフェイスは、一組の類似性メトリックのうち最大値を有する類似性メトリックを示してもよい。この一組の類似性メトリックは、１つ以上のデータセット各々について計算した類似性メトリックを含み得る。

【０１８２】

図１９は、実施形態を実装するための分散型システム１９００の簡略図を示す。示されている実施形態において、分散型システム１９００は、１つ以上のクライアントコンピューティングデバイス１９０２，１９０４，１９０６，および１９０８を含み、これらは、１つ以上のネットワーク１９１０を通じて、ウェブブラウザ、専用クライアント（たとえば、Oracle Forms）等のクライアントアプリケーションを実行し操作するように構成される。サーバ１９１２は、ネットワーク１９１０を介してリモートクライアントコンピューティングデバイス１９０２，１９０４，１９０６，および１９０８と通信可能に結合されてもよい。

【０１８３】

さまざまな実施形態において、サーバ１９１２は、文書（たとえばウェブページ）の分析および修正に関連する処理を提供するサービスおよびアプリケーション等の１つ以上のサービスまたはソフトウェアアプリケーションを実行するように適合させてもよい。特定の実施形態において、サーバ１９１２はその他のサービスまたはソフトアプリケーションも提供し得る。これは非仮想および仮想環境を含み得る。いくつかの実施形態において、これらのサービスは、ウェブベースもしくはクラウドサービスとして、または、サービスとしてのソフトウェア（ＳａａＳ）モデルの下で、クライアントコンピューティングデバイス１９０２，１９０４，１９０６，および／または１９０８のユーザに提供し得る。クライアントコンピューティングデバイス１９０２、１９０４および１９０６，および／または１９０８を操作するユーザは、１つ以上のクライアントアプリケーションを利用してサーバ１９１２と対話することにより、これらのコンポーネントによって提供されるサービスを利用し得る。

【０１８４】

図１９に示される構成において、システム１９００のソフトウエアコンポーネント１９１８、１９２０および１９２２は、サーバ１９１２上で実装されるものとして示されている。他の実施形態において、システム１９００のコンポーネントのうちの１つ以上および／またはこれらのコンポーネントによって提供されるサービスも、クライアントコンピューティングデバイス１９０２，１９０４，１９０６，および／または１９０８のうちの１つ以上によって実装されてもよい。そうすると、クライアントコンピューティングデバイスを操作するユーザは、１つ以上のクライアントアプリケーションを利用して、これらのコンポーネントによって提供されるサービスを使用し得る。これらのコンポーネントは、ハードウェア、ファームウェア、ソフトウェア、またはこれらの組合わせにおいて実装し得る。分散型システム１９００とは異なり得るさまざまな異なるシステムコンフィギュレーションが可能であることが認識されるはずである。図１９に示される実施形態はしたがって、実施形態のシステムを実装するための分散型システムの一例であり、限定を意図していない。

【０１８５】

クライアントコンピューティングデバイス１９０２，１９０４，１９０６，および／または１９０８は、さまざまな種類のコンピューティングシステムを含み得る。たとえば、クライアントデバイスは、Microsoft Windows Mobile（登録商標）等のソフトウェアおよび／またはiOS、Windows Phone、Android、BlackBerry 10、Palm OS等のようなさまざまなモバイルオペレーティングシステムを実行する、ポータブルハンドヘルドデバイス（たとえばiPhone（登録商標）、携帯電話、iPad（登録商標）、コンピューティングタブレット、携帯情報端末（ＰＤＡ））またはウェアラブルデバイス（たとえばGoogle Glass（登録商標）ヘッドマウントディスプレイ）を含み得る。これらのデバイスは、さまざまなイ

ンターネット関連アプリケーション、電子メール、ショートメッセージサービス（ＳＭＳ）アプリケーションをサポートし得るものであり、その他さまざまな通信プロトコルを使用し得る。クライアントコンピューティングデバイスはまた、一例として、さまざまなバージョンのMicrosoft Windows（登録商標）、Apple Macintosh（登録商標）、および／またはLinux（登録商標）オペレーティングシステムを実行するパーソナルコンピュータおよび／またはラップトップコンピュータ含む汎用パーソナルコンピュータを含み得る。クライアントコンピューティングデバイスは、限定されないがたとえばGoogle Chrome OS等のさまざまなGNU/Linux（登録商標）オペレーティングシステムを含む市場で入手可能な多様なUNIX（登録商標）またはUNIX（登録商標）系オペレーティングシステムのうちのいずれかを実行するワークステーションコンピュータであってもよい。クライアントコンピューティングデバイスはまた、ネットワーク１９１０を通して通信可能な、シンクライアントコンピュータ、インターネット接続可能なゲームシステム（たとえばKinect（登録商標）ジェスチャー入力デバイスを有するまたは有しないMicrosoft Xboxゲームコンソール）、および／またはパーソナルメッセージングデバイス等の電子デバイスを含み得る。

【０１８６】

図１９では４つのクライアントコンピューティングデバイスを有する分散型システム１９００が示されているが、任意の数のクライアントコンピューティングデバイスをサポートし得る。センサを有するデバイス等の他のデバイスがサーバ１９１２と対話してもよい。

【０１８７】

分散型システム１９００のネットワーク１９１０は、限定されないがＴＣＰ／ＩＰ（Transmission control protocol/Internet protocol）、ＳＮＡ（systems network architecture）、ＩＰＸ（Internet packet exchange）、AppleTalk（登録商標）等を含む利用できるさまざまなプロトコルのうちのいずれかを使用してデータ通信をサポートできる、当業者によく知られた任意のタイプのネットワークであってもよい。一例に過ぎないが、ネットワーク１９１０は、ローカルエリアネットワーク（ＬＡＮ）、イーサネット（登録商標）、トークンリングに基づくネットワーク、広域ネットワーク、インターネット、仮想ネットワーク、仮想プライベートネットワーク（ＶＰＮ）、イントラネット、エクストラネット、公衆交換電話網（ＰＳＴＮ）、赤外線ネットワーク、無線ネットワーク（たとえばInstitute of Electrical and Electronics（ＩＥＥＥ）８０２．１１プロトコルスイート、Bluetooth（登録商標）、および／または任意の他の無線プロトコルのうちのいずれかの下で動作するネットワーク）、および／または上記および／またはその他のネットワークの任意の組合わせであってもよい。

【０１８８】

サーバ１９１２は、１つ以上の汎用コンピュータ、専用サーバコンピュータ（一例としてＰＣ（パーソナルコンピュータ）サーバ、UNIX（登録商標）サーバ、ミッドレンジサーバ、メインフレームコンピュータ、ラックマウントサーバ等を含む）、サーバファーム、サーバクラスタ、または任意の他の適切な構成および／または組合わせによって構成されていてもよい。サーバ１９１２は、仮想オペレーティングシステムを実行する１つ以上の仮想マシン、または、仮想化を伴うその他のコンピューティングアーキテクチャを含み得る。論理記憶デバイスの１つ以上のフレキシブルなプールを仮想化することによってサーバのための仮想記憶デバイスを維持してもよい。仮想ネットワークは、サーバ１９１２が、ソフトウェアで規定されるネットワーキングを用いて制御することができる。さまざまな実施形態において、サーバ１９１２は、これまでの開示において記載されている１つ以上のサービスまたはソフトウェアアプリケーションを実行するように適合させてもよい。たとえば、サーバ１９１２は、本開示の実施形態に係る上記処理を実行するためのサーバに対応し得る。

【０１８９】

サーバ１９１２は、上記オペレーティングシステムのうちのいずれかおよび市場で入手可能なサーバオペレーティングシステムを含むオペレーティングシステムを実行し得る。

また、サーバ１９１２は、ＨＴＴＰ（hypertext transport protocol）サーバ、ＦＴＰ（file transfer protocol）サーバ、ＣＧＩ（common gateway interface）サーバ、ＪＡＶＡ（登録商標）サーバ、データベースサーバ等を含む、さまざまな付加的なサーバアプリケーションおよび／またはミッドティアアプリケーションのうちのいずれかを実行し得る。典型的なデータベースサーバは、Oracle、Microsoft、Sybase、ＩＢＭ（International Business Machines）等から市販されているものを含むが、これらに限定されない。

【０１９０】

いくつかの実装例において、サーバ１９１２は、クライアントコンピューティングデバイス１９０２，１９０４，１９０６，および１９０８のユーザから受信したデータフィードおよび／またはイベントアップデートを分析し統合するための１つ以上のアプリケーションを含み得る。一例として、データフィードおよび／またはイベントアップデートは、限定されないが、１つ以上の第三者情報源および連続データストリームから受信したTwitter（登録商標）フィード、Facebook（登録商標）更新、またはリアルタイム更新を含み得る。これらはセンサデータアプリケーション、株式相場ディスプレイデバイス、ネットワーク性能測定ツール（たとえばネットワーク監視およびトラフィック管理アプリケーション）、クリックストリーム分析ツール、自動車トラフィック監視等に関連するリアルタイムイベントを含み得る。また、サーバ１９１２は、クライアントコンピューティングデバイス１９０２，１９０４，１９０６，および１９０８の１つ以上のディスプレイデバイスを介してデータフィードおよび／またはリアルタイムイベントを表示するための１つ以上のアプリケーションを含み得る。

【０１９１】

分散型システム１９００は、１つ以上のデータベース１９１４および１９１６も含み得る。これらのデータベースは、ユーザ対話情報、使用パターン情報、適応則情報、および本発明の実施形態で使用されるその他の情報等の情報を格納するためのメカニズムを提供し得る。データベース１９１４および１９１６はさまざまな場所に存在し得る。一例として、データベース１９１４および１９１６のうちの１つ以上は、サーバ１９１２に対してローカルな場所にある（および／またはサーバ内にある）非一時的な記憶媒体上にあってもよい。代替的に、データベース１９１４および１９１６は、サーバ１９１２から遠隔の場所に位置してネットワークベースのまたは専用接続を介してサーバ１９１２と通信してもよい。一組の実施形態において、データベース１９１４および１９１６は、ストレージエリアネットワーク（ＳＡＮ）内にあってもよい。同様に、サーバ１９１２に帰する機能を実行するために必要な任意のファイルを、適宜、サーバ１９１２に対してローカルな場所におよび／またはサーバ１９１２から遠隔の場所に格納してもよい。一組の実施形態において、データベース１９１４および１９１６は、ＳＱＬフォーマットの命令に応答してデータを記憶、更新、および検索するように適合している、Oracleによって提供されるデータベース等のリレーショナルデータベースを含み得る。

【０１９２】

いくつかの実施形態において、上記文書分析および修正サービスは、クラウド環境を介したサービスとして提供されてもよい。図２０は、本開示の実施形態に従う、サービスをクラウドサービスとして提供し得るシステム環境２０００の１つ以上のコンポーネントの簡略化されたブロック図である。図２０に示されている実施形態において、システム環境２０００は、使用パターンに応じて文書（たとえばウェブページ）を動的に修正するためのサービスを含むクラウドサービスを提供するクラウドインフラストラクチャシステム２００２と対話するためにユーザが使用し得る１つ以上のクライアントコンピューティングデバイス２００４，２００６，および２００８を含む。クラウドインフラストラクチャシステム２００２は、サーバ２０１２に関して先に述べたものを含み得る１つ以上のコンピュータおよび／またはサーバを含み得る。

【０１９３】

図２０に示されているクラウドインフラストラクチャシステム２００２は示されているもの以外のコンポーネントを有し得ることが認識されるはずである。さらに、図２０に示

される実施形態は、本発明の実施形態を組込むことができるクラウドインフラストラクチャシステムの一例に過ぎない。他のいくつかの実施形態において、クラウドインフラストラクチャシステム２００２は、図示されているよりも多いまたは少ないコンポーネントを有していてもよく、２つ以上のコンポーネントを組合わせてもよく、または異なる構成または配置のコンポーネントを有していてもよい。

【０１９４】

クライアントコンピューティングデバイス２００４，２００６，および２００８は、１９０２，１９０４，１９０６，および１９０８について先に述べたものと同様のデバイスであってもよい。クライアントコンピューティングデバイス２００４，２００６，および２００８は、以下のようなクライアントアプリケーションを操作するように構成されていてもよく、このクライアントアプリケーションは、たとえば、クライアントコンピューティングデバイスのユーザがクラウドインフラストラクチャシステム２００２と対話してクラウドインフラストラクチャシステム２００２が提供するサービスを使用するために使用し得る、ウェブブラウザ、専用クライアント（たとえばOracle Forms）、またはその他何らかのアプリケーション等である。典型的なシステム環境２０００は３つのクライアントコンピューティングデバイスとともに示されているが、任意の数のクライアントコンピューティングデバイスをサポートし得る。センサ等を有するデバイスのようなその他のデバイスがクラウドインフラストラクチャシステム２００２と対話してもよい。

【０１９５】

ネットワーク２０１０は、クライアント２００４，２００６、２００８とクラウドインフラストラクチャシステム２００２との間のデータの通信およびやり取りを容易にし得る。各ネットワークは、ネットワーク２０１０について先に述べたものを含むさまざまな市場で入手可能なプロトコルのいずれかを使用してデータ通信をサポートすることができる、当業者によく知られた任意のタイプのネットワークであってもよい。

【０１９６】

特定の実施形態において、クラウドインフラストラクチャシステム２００２によって提供されるサービスは、クラウドインフラストラクチャシステムのユーザがオンデマンドで利用できるようにされる多数のサービスを含み得る。使用パターンに応じて動的に文書を修正することに関連するサービスに加えて、その他さまざまなサービスも提供し得る。これらのサービスは、限定されないが、オンラインデータストレージおよびバックアップソリューション、ウェブベースの電子メールサービス、ホストされたオフィスパッケージおよびドキュメントコラボレーションサービス、データベース処理、管理された技術サポートサービス等である。クラウドインフラストラクチャシステムによって提供されるサービスは、そのユーザのニーズに合わせて動的にスケーリングできる。

【０１９７】

特定の実施形態において、クラウドインフラストラクチャシステム２００２によって提供されるサービスの具体的なインスタンス化は、本明細書において「サービスインスタンス」と呼ばれることがある。一般的に、クラウドサービスプロバイダのシステムからインターネット等の通信ネットワークを介してユーザが利用できるようにされる任意のサービスは、「クラウドサービス」と呼ばれる。典型的に、パブリッククラウド環境において、クラウドサービスプロバイダのシステムを構成するサーバおよびシステムは、顧客自身のオンプレミスサーバおよびシステムとは異なる。たとえば、クラウドサービスプロバイダのシステムは、アプリケーションをホストしてもよく、ユーザは、インターネット等の通信ネットワークを介してオンデマンドでアプリケーションをオーダーして使用すればよい。

【０１９８】

いくつかのの例において、コンピュータネットワーククラウドインフラストラクチャにおけるサービスは、クラウドベンダーによってまたは当該技術において周知の他のやり方でユーザに提供される、記憶装置、ホストされたデータベース、ホストされたウェブサーバ、ソフトウェアアプリケーション、または他のサービスに対する保護されたコンピュー

タネットワークアクセスを含み得る。たとえば、サービスは、インターネットを通じたクラウド上の記憶装置に対するパスワードで保護されたアクセスを含むことができる。別の例として、サービスは、ネットワーク化されたデベロッパーによる私的使用のためのウェブサービスベースのホストされたリレーショナルデータベースおよびスクリプト言語ミドルウェアエンジンを含むことができる。別の例として、サービスは、クラウドベンダーのウェブサイト上でホストされた電子メールソフトウェアアプリケーションに対するアクセスを含むことができる。

【０１９９】

特定の実施形態において、クラウドインフラストラクチャシステム２００２は、セルフサービスの、申込みに基づく、弾力的にスケーラブルで、確実で、非常に有効で、かつ安全なやり方で、顧客に与えられる、アプリケーション、ミドルウェア、およびデータベースサービス提供物一式を含み得る。そのようなクラウドインフラストラクチャシステムの一例は、本願の譲受人によって提供されるOracle Public Cloudである。

【０２００】

クラウドインフラストラクチャシステム２００２は、「ビッグデータ」に関連する計算および分析サービスも提供し得る。「ビッグデータ」という用語は一般的に、大量のデータを可視化する、傾向を発見する、および／またはそうでなければデータと対話するために、アナリストおよびリサーチャーが保存し操作することができる極めて大きなデータセットに言及するときに用いられる。このビッグデータおよび関連するアプリケーションは、多数のレベルおよびさまざまな規模でインフラストラクチャシステムがホストおよび／または操作することができる。並列にリンクされた何十、何百、または何千ものプロセッサが、このようなデータに対して機能することにより、それを示すまたはデータに対する外部からの力をもしくはそれが表わしているものをシミュレートすることができる。これらのデータセットは、データベース内でそうでなければ構造化モデルに従って組織されたデータのような構造化データ、および／または非構造化データ（たとえば電子メール、画像、データＢＬＯＢ（（binary large object）バイナリラージオブジェクト）、ウェブページ、複雑なイベント処理）を含み得る。より多くの（またはより少ない）計算リソースを比較的素早く目標物に向ける実施形態の能力を高めることにより、企業、政府機関、リサーチ組織、私人、同じ目的を有する個人もしくは組織、またはその他のエンティティからの要求に基づいて大きなデータセットに対するタスクを実行するにあたり、クラウドインフラストラクチャシステムをより有効にすることができる。

【０２０１】

さまざまな実施形態において、クラウドインフラストラクチャシステム２００２は、クラウドインフラストラクチャシステム２００２から提供されるサービスに対する顧客の申込みを自動的にプロビジョニングし、管理し、かつ追跡するように適合させることができる。クラウドインフラストラクチャシステム２００２は、異なるデプロイメントモデルを介してクラウドサービスを提供し得る。たとえば、サービスは、（たとえばOracle社所有の）クラウドサービスを販売する組織によってクラウドインフラストラクチャシステム２００２が所有されるパブリッククラウドモデルの下で提供されてもよく、サービスは、一般大衆または異なる産業企業にとって利用可能とされる。別の例として、サービスは、クラウドインフラストラクチャシステム２００２が単一の組織のためにのみ運営され、組織内の１つ以上のエンティティのためのサービスを提供し得る個人のクラウドモデルの下で提供し得る。また、クラウドサービスは、クラウドインフラストラクチャシステム２００２およびクラウドインフラストラクチャシステム２００２によって提供されるサービスが、関連するコミュニティー内の一部の組織によって共有されるコミュニティークラウドモデルの下で提供し得る。また、クラウドサービスは、２つ以上の異なるモデルの組み合わせであるハイブリッドクラウドモデルの下で提供し得る。

【０２０２】

いくつかの実施形態において、クラウドインフラストラクチャシステム２００２によって提供されるサービスは、サービスとしてのソフトウェア（ＳａａＳ）カテゴリ、サービ

スとしてのプラットフォーム（ＰａａＳ）カテゴリ、サービスとしてのインフラストラクチャ（ＩａａＳ）カテゴリ、またはハイブリッドサービスを含む他のサービスカテゴリの下で提供される１つ以上のサービスを含み得る。顧客は、クラウドインフラストラクチャシステム２００２によって提供される１つ以上のサービスを申込みオーダーによってオーダーし得る。そうすると、クラウドインフラストラクチャシステム２００２は、顧客の申込みオーダーにおけるサービスを提供するための処理を行なう。

【０２０３】

いくつかの実施形態において、クラウドインフラストラクチャシステム２００２によって提供されるサービスは、限定されないが、アプリケーションサービス、プラットフォームサービスおよびインフラストラクチャサービスを含み得る。いくつかの例において、アプリケーションサービスは、クラウドインフラストラクチャシステムによってＳａａＳプラットフォームを介して提供し得る。ＳａａＳプラットフォームは、ＳａａＳカテゴリに入るクラウドサービスを提供するように構成し得る。たとえば、ＳａａＳプラットフォームは、統合された開発およびデプロイメントプラットフォーム上のオンデマンドのアプリケーション一式を構築し、伝える能力を提供し得る。ＳａａＳプラットフォームは、ＳａａＳサービスを提供するための基礎的なソフトウェアおよびインフラストラクチャを管理し、制御し得る。ＳａａＳプラットフォームによって提供されるサービスを利用することによって、顧客は、クラウドインフラストラクチャシステム上で実行するアプリケーションを利用することができる。顧客は、顧客が別個のライセンスおよびサポートを購入する必要なしに、アプリケーションサービスを得ることができる。さまざまな異なるＳａａＳサービスが提供し得る。例は、限定されないが、大きな組織のための販売実績管理、企業統合およびビジネス上のフレキシビリティのためのソリューションを提供するサービスを含む。

【０２０４】

いくつかの実施形態において、プラットフォームサービスは、クラウドインフラストラクチャシステム２００２によってＰａａＳプラットフォームを介して提供し得る。ＰａａＳプラットフォームは、ＰａａＳカテゴリに入るクラウドサービスを提供するように構成し得る。プラットフォームサービスの例は、限定されないが、共有されている共通アーキテクチャ上の既存のアプリケーションを組織（Oracle等）が統合することを可能にするサービスと、プラットフォームによって提供される共有サービスを活用する新しいアプリケーションを構築する能力とを含み得る。ＰａａＳプラットフォームは、ＰａａＳサービスを提供するための基礎的なソフトウェアおよびインフラストラクチャを管理および制御し得る。顧客は、顧客が別個のライセンスおよびサポートを購入する必要なしに、クラウドインフラストラクチャシステム２００２によって提供されるＰａａＳサービスを得ることができる。プラットフォームサービスの例は、限定されないが、Oracle Java（登録商標） Cloud Service（ＪＣＳ）、Oracle Database Cloud Service（ＤＢＣＳ）他を含む。

【０２０５】

ＰａａＳプラットフォームによって提供されるサービスを利用することによって、顧客は、クラウドインフラストラクチャシステムによってサポートされたプログラミング言語およびツールを採用し、また、デプロイされたサービスを制御することもできる。いくつかの実施形態において、クラウドインフラストラクチャシステムによって提供されるプラットフォームサービスは、データベースクラウドサービス、ミドルウェアクラウドサービス（たとえばOracle Fusion Middleware services）、およびＪａｖａ（登録商標）クラウドサービスを含み得る。一実施形態において、データベースクラウドサービスは、組織がデータベースリソースをプールし、かつデータベースクラウドの形態でのサービスとして顧客にデータベースを提示することを可能にする共有サービスデプロイメントモデルをサポートし得る。ミドルウェアクラウドサービスは、顧客がさまざまなビジネスアプリケーションを展開しデプロイするためのプラットフォームを提供してもよく、Ｊａｖａ（登録商標）クラウドサービスは、クラウドインフラストラクチャシステムにおいて顧客がＪａｖａ（登録商標）アプリケーションをデプロイするためのプラットフォームを提供して

もよい。

【０２０６】

さまざまな異なるインフラストラクチャサービスは、クラウドインフラストラクチャシステムにおいてＩａａＳプラットフォームによって提供し得る。インフラストラクチャサービスは、ＳａａＳプラットフォームおよびＰａａＳプラットフォームによって提供されるサービスを利用する顧客のための記憶装置、ネットワーク、および他の基本のコンピューティングリソースといった基礎的なコンピューティングリソースの管理および制御を容易にする。

【０２０７】

特定の実施形態において、クラウドインフラストラクチャシステム２００２はまた、クラウドインフラストラクチャシステムの顧客にさまざまなサービスを提供するために用いられるリソースを提供するためのインフラストラクチャリソース２０３０を含み得る。一実施形態において、インフラストラクチャリソース２０３０は、ＰａａＳプラットフォームおよびＳａａＳプラットフォームによって提供されるサービスを実行するためのサーバ、記憶装置、およびネットワークのリソースといったハードウェアの予め統合され最適化された組合わせ、ならびにその他のリソースを含み得る。

【０２０８】

いくつかの実施形態において、クラウドインフラストラクチャシステム２００２におけるリソースは、複数のユーザによって共有され要求ごとに動的に再割当てされてもよい。加えて、リソースは、異なるタイムゾーンのユーザに割当てられてもよい。たとえば、クラウドインフラストラクチャシステム２００２は、第１のタイムゾーンの第１の組のユーザが特定数の時間クラウドインフラストラクチャシステムのリソースを利用できるようにし、次いで異なるタイムゾーンに位置する別の組のユーザに対して同じリソースを再度割当てることによって、リソースの利用を最大化してもよい。

【０２０９】

特定の実施形態において、クラウドインフラストラクチャシステム２００２の異なるコンポーネントまたはモジュールによって共有される多数の内部共有サービス２０３２を提供し得る。これらの内部共有サービスは、限定されないが、セキュリティおよびアイデンティティサービス、インテグレーションサービス、企業リポジトリサービス、企業マネージャーサービス、ウィルススキャンおよびホワイトリストサービス、高アベイラビリティ、保存後修復サービス、クラウドサポートを可能にするためのサービス、電子メールサービス、通知サービス、ファイル転送サービス等を含み得る。

【０２１０】

特定の実施形態において、クラウドインフラストラクチャシステム２００２は、クラウドインフラストラクチャシステムにおいてクラウドサービス（たとえばＳａａＳ、ＰａａＳおよびＩａａＳサービス）の包括的な管理を提供し得る。一実施形態において、クラウド管理機能は、クラウドインフラストラクチャシステム２００２によって受信された顧客の申込みをプロビジョニング、管理、および追跡する機能を含み得る。

【０２１１】

一実施形態において、図２０に示されるように、クラウド管理機能は、オーダー管理モジュール２０２０、オーダーオーケストレーションモジュール２０２２、オーダープロビジョニングモジュール２０２４、オーダー管理および監視モジュール２０２６、およびアイデンティティ管理モジュール２０２８といった、１つ以上のモジュールによって提供し得る。これらのモジュールは、汎用コンピュータ、専用サーバコンピュータ、サーバファーム、サーバクラスタ、または任意の他の適切な構成および／または組合わせであってもよい１つ以上のコンピュータおよび／またはサーバを含み得る、または、それらを用いて提供し得る。

【０２１２】

典型的な動作では、２０３４において、クライアントデバイス２００４、２００６、または２００８等のクライアントデバイスを用いる顧客は、クラウドインフラストラクチャ

システム２００２によって提供される１つ以上のサービスを要求し、クラウドインフラストラクチャシステム２００２によって提示される１つ以上のサービスの申込みのためのオーダーを行なうことによって、クラウドインフラストラクチャシステム２００２と対話してもよい。特定の実施形態において、顧客は、クラウドＵＩ２０１２、クラウドＵＩ２０１４、および／またはクラウドＵＩ２０１６等のクラウドユーザインターフェイス（ＵＩ）にアクセスし、これらのＵＩを介して申込みオーダーを行なってもよい。顧客がオーダーを行ったことに応じてクラウドインフラストラクチャシステム２００２によって受信されるオーダー情報は、顧客を特定する情報と、顧客が申込む予定のクラウドインフラストラクチャシステム２００２によって提供される１つ以上のサービスとを含み得る。

【０２１３】

２０３６において、顧客から受けたオーダー情報は、オーダーデータベース２０１８に格納されてもよい。これが新しいオーダーであれば、このオーダーについて新たな記録が作成されてもよい。一実施形態において、オーダーデータベース２０１８は、クラウドインフラストラクチャシステム２０１８によって操作され、他のシステムエレメントとともに操作されるいくつかのデータベースのうちの１つであってもよい。

【０２１４】

２０３８において、オーダー情報は、オーダー管理モジュール２０２０に転送されてもよい。オーダー管理モジュール２０２０は、オーダーを確認し、確認後にオーダーを記入する等の、オーダーに関連する課金および会計機能を行なうように構成し得る。

【０２１５】

２０４０において、オーダーに関する情報は、オーダーオーケストレーションモジュール２０２２に伝えられてもよい。オーダーオーケストレーションモジュール２０２２は、顧客によって出されたオーダーのためのサービスおよびリソースのプロビジョニングを調整するように構成されている。いくつかのインスタンスにおいて、オーダーオーケストレーションモジュール２０２２は、プロビジョニングのためにオーダープロビジョニングモジュール２０２４のサービスを用いてもよい。特定の実施形態において、オーダーオーケストレーションモジュール２０２２は、各オーダーに関連付けられたビジネスプロセスの管理を可能にし、オーダーがプロビジョニングに進むべきか否かを判断するためにビジネスロジックを適用する。

【０２１６】

図２０に示される実施形態において示されるように、２０４２において、新規申込みのためのオーダーを受信すると、オーダーオーケストレーションモジュール２０２２は、オーダープロビジョニングモジュール２０２４に要求を送信して、リソースを割当て、申込みオーダーを遂行するために必要とされるリソースを構成する。オーダープロビジョニングモジュール２０２４は、顧客によってオーダーされたサービスのためのリソースの割当てを可能にする。オーダープロビジョニングモジュール２０２４は、クラウドインフラストラクチャシステム２０００によって提供されるクラウドサービスと、リソースサービスを提供するためにリソースをプロビジョニングするために用いられる物理実装層との間に抽象化レベルを提供する。これにより、オーダーオーケストレーションモジュール２０２２を、サービスおよびリソースが実際にオンザフライでプロビジョニングされまたは予めプロビジョニングされ要求後に割当て／アサインされるか否かといった実施の詳細から分離することができる。

【０２１７】

２０４４において、ひとたびサービスおよびリソースがプロビジョニングされると、要求されたサービスが現在利用できる状態にあることを示す通知を、申込んだ顧客に送ってもよい。いくつかのインスタンスにおいて、顧客が要求したサービスの利用を開始できるようにする情報（たとえばリンク）を顧客に送ってもよい。

【０２１８】

２０４６において、顧客の申込みオーダーは、オーダー管理および監視モジュール２０２６によって管理および追跡されてもよい。いくつかのインスタンスにおいて、オーダー

管理および監視モジュール２０２６は、申込まれたサービスの顧客利用に関する使用統計を収集するように構成されてもよい。たとえば、記憶装置の使用量、データ転送量、ユーザ数、ならびにシステムアップタイムおよびシステムダウンタイムの量等について、統計が収集されてもよい。

【０２１９】

特定の実施形態において、クラウドインフラストラクチャシステム２０００は、アイデンティティ管理モジュール２０２８を含み得る。アイデンティティ管理モジュール２０２８は、クラウドインフラストラクチャシステム２０００におけるアクセス管理および認可サービスといったアイデンティティサービスを提供するように構成される。いくつかの実施形態において、アイデンティティ管理モジュール２０２８は、クラウドインフラストラクチャシステム２００２によって提供されるサービスを利用したい顧客に関する情報を制御し得る。そのような情報は、そのような顧客のアイデンティティを認証する情報と、さまざまなシステムリソース（たとえばファイル、ディレクトリ、アプリケーション、通信ポート、メモリセグメント等）に対しそれらの顧客が実行を認可されるアクションを記述する情報とを含むことができる。アイデンティティ管理モジュール２０２８は、各顧客に関し、かつ、どのように誰によってその記述情報のアクセスおよび修正ができるかに関する記述情報の管理も含み得る。

【０２２０】

図２１は、本発明の実施形態を実装するために使用し得る典型的なコンピュータシステム２１００を示す。いくつかの実施形態において、コンピュータシステム２１００を用いて上記さまざまなサーバおよびコンピュータシステムのうちのいずれかを実装し得る。図２１に示されるように、コンピュータシステム２１００は、バスサブシステム２１０２を介して多数の周辺サブシステムと通信する処理部２１０４を含むさまざまなサブシステムを含む。これらの周辺サブシステムは、処理加速部２１０６と、Ｉ／Ｏサブシステム２１０８と、記憶サブシステム２１１８と、通信サブシステム２１２４とを含み得る。記憶サブシステム２１１８は、有形のコンピュータ読取可能記憶媒体２１２２とシステムメモリ２１１０とを含み得る。

【０２２１】

バスサブシステム２１０２は、コンピュータシステム２１００のさまざまなコンポーネントおよびサブシステムを目的に合わせて互いに通信させるためのメカニズムを提供する。バスサブシステム２１０２は、単母線として概略的に示されるが、バスサブシステムの代替的な実施形態は複数のバスを利用し得る。バスサブシステム２１０２は、多様なバスアーキテクチャのうちのいずれかを用いる、メモリバスまたはメモリコントローラ、周辺バス、およびローカルバスを含むいくつかのタイプのバス構造のうちのいずれかであってもよい。たとえば、そのようなアーキテクチャは、ＩＥＥＥ　Ｐ１３８６．１規格等に従って製造されたMezzanineバスとして実装できる、Industry Standard Architecture（ＩＳＡ）バス、Micro Channel Architecture（ＭＣＡ）バス、Enhanced ISA（ＥＩＳＡ）バス、Video Electronics Standards Association（ＶＥＳＡ）ローカルバス、およびPeripheral Component Interconnect（ＰＣＩ）バスを含み得る。

【０２２２】

処理サブシステム２１０４は、コンピュータシステム２１００の動作を制御し、１つ以上の処理部２１３２，２１３４等を含み得る。処理部は、シングルコアもしくはマルチコアプロセッサ、プロセッサの１つ以上のコア、またはこれらの組合わせを含む、１つ以上のプロセッサを含み得る。いくつかの実施形態において、処理サブシステム２１０４は、グラフィックスプロセッサ、デジタル信号プロセッサ（ＤＳＰ）等といった１つ以上の専用コプロセッサを含み得る。いくつかの実施形態において、処理サブシステム２１０４の処理部の一部またはすべてを、特定用途向け集積回路（ＡＳＩＣ）またはフィールドプログラマブルゲートアレイ（ＦＰＧＡ）等のカスタマイズされた回路を用いて実装してもよい。

【０２２３】

いくつかの実施形態において、処理サブシステム２１０４の処理部は、システムメモリ２１１０またはコンピュータ可読記憶媒体２１２２に格納されている命令を実行できる。さまざまな実施形態において、処理部は、さまざまなプログラムまたはコード命令を実行することができ、かつ、同時に実行する複数のプログラムまたはプロセスを維持することができる。どの時点でも、実行すべきプログラムコードのうちの一部またはすべては、場合によっては１つ以上の記憶装置を含む、システムメモリ２１１０および／またはコンピュータ可読記憶媒体２１２２上に存在し得る。適切なプログラミングにより、処理サブシステム２１０４は、使用パターンに応じて文書（たとえばウェブページ）を動的に修正するための上記さまざまな機能を提供することができる。

【０２２４】

特定の実施形態において、処理加速部２１０６は、カスタマイズされた処理を実行して、または処理サブシステム２１０４が実行する処理の一部をオフロードして、コンピュータシステム２１００が実行する処理全体を加速するために、提供し得る。

【０２２５】

Ｉ／Ｏサブシステム２１０８は、情報をコンピュータシステム２１００に入力するためおよび／または情報をコンピュータシステム２１００からもしくはコンピュータシステム２１００を介して出力するためのデバイスおよびメカニズムを含み得る。一般的に、「入力デバイス」という用語を使用する場合は、コンピュータシステム２１００に情報を入力するための可能なすべての種類のデバイスおよびメカニズムを含むことを意図している。ユーザインターフェイス入力デバイスは、たとえば、キーボード、マウスまたはトラックボール等のポインティングデバイス、ディスプレイに組込まれたタッチパッドまたはタッチスクリーン、スクロールホイール、クリックホイール、ダイヤル、ボタン、スイッチ、キーパッド、音声コマンド認識システムを備えた音声入力デバイス、マイク、およびその他の種類の入力デバイスを含み得る。ユーザインターフェイス入力デバイスはまた、ユーザが入力デバイスを制御しこれと対話することを可能にするMicrosoft Kinect（登録商標）モーションセンサ、Microsoft Xbox（登録商標）３６０ゲームコントローラ、ジェスチャーおよび音声コマンドを用いた入力を受けるためのインターフェイスを提供するデバイス等の、動き検知および／またはジェスチャー認識デバイスを含み得る。ユーザインターフェイス入力デバイスはまた、Google Glass（登録商標）まばたき検出器等のアイジェスチャー認識デバイスを含み得る。これは、ユーザの目の活動（たとえば撮影中および／またはメニュー選択中の「まばたき」）を検出し、入力デバイス（たとえばGoogle Glass（登録商標））に対する入力としてのアイジェスチャーを変換する。加えて、ユーザインターフェイス入力デバイスは、ユーザが音声コマンドによって音声認識システム（たとえばSiri（登録商標）ナビゲーター）と対話することを可能にする音声認識検知装置を含み得る。

【０２２６】

ユーザインターフェイス入力デバイスのその他の例は、限定されないが、三次元（３Ｄ）マウス、ジョイスティックまたはポインティングスティック、ゲームパッドおよびグラフィックタブレット、およびスピーカ等のオーディオ／ビジュアルデバイス、デジタルカメラ、デジタルビデオカメラ、ポータブルメディアプレイヤー、ウェブカメラ、イメージスキャナ、指紋スキャナ、バーコードリーダー３Ｄスキャナ、３Ｄプリンタ、レーザ測距装置、および視線追跡デバイスを含む。加えて、ユーザインターフェイス入力デバイスは、たとえば、コンピュータ断層撮影装置、磁気共鳴撮像装置、ポジトロン断層撮影装置、医療用超音波検査装置等の医療用撮像入力デバイスを含み得る。ユーザインターフェイス入力デバイスはまた、たとえば、ＭＩＤＩキーボード、デジタル楽器等といった音声入力装置を含み得る。

【０２２７】

ユーザインターフェイス出力デバイスは、ディスプレイサブシステム、表示灯、または音声出力装置等の非視覚的ディスプレイを含み得る。ディスプレイサブシステムは、陰極線管（ＣＲＴ）、液晶ディスプレイ（ＬＣＤ）またはプラズマディスプレイを用いるもの

等のフラットパネルデバイス、投影デバイス、タッチスクリーン等であってもよい。一般的に、「出力デバイス」という用語を使用する場合は、コンピュータシステム２１００からユーザまたは他のコンピュータに情報を出力するための可能なすべての種類のデバイスおよびメカニズムを含むことを意図している。たとえば、ユーザインターフェイス出力デバイスは、限定されないが、モニタ、プリンタ、スピーカ、ヘッドホン、カーナビゲーションシステム、プロッタ、音声出力デバイス、およびモデム等の、テキスト、図形、およびオーディオ／ビデオ情報を視覚的に伝えるさまざまなディスプレイデバイスを含み得る。

【０２２８】

記憶サブシステム２１１８は、コンピュータシステム２１００によって使用される情報を格納するためのリポジトリまたはデータストアを提供する。記憶サブシステム２１１８は、いくつかの実施形態の機能を提供する基本的なプログラミングおよびデータ構造を格納するための有形の非一時的なコンピュータ可読記憶媒体を提供する。処理サブシステム２１０４によって実行されたときに上記機能を提供するソフトウェア（プログラム、コードモジュール、命令）は、記憶サブシステム２１１８に格納し得る。このソフトウェアは、処理サブシステム２１０４の１つ以上の処理部によって実行し得る。記憶サブシステム２１１８はまた、本発明に従い使用されるデータを格納するためのリポジトリを提供し得る。

【０２２９】

記憶サブシステム２１１８は、揮発性および不揮発性メモリデバイスを含む１つ以上の非一時的なメモリデバイスを含み得る。図２１に示されるように、記憶サブシステム２１１８は、システムメモリ２１１０とコンピュータ可読記憶媒体２１２２とを含む。システムメモリ２１１０は、プログラム実行中の命令およびデータの格納のための揮発性メインランダムアクセスメモリ（ＲＡＭ）、および、固定命令が格納される不揮発性読出専用メモリ（ＲＯＭ）またはフラッシュメモリを含む、多数のメモリを含み得る。いくつかの実装例において、起動中等のコンピュータシステム２１００内の要素間の情報の転送を支援する基本ルーチンを含む基本入出力システム（ＢＩＯＳ）が、典型的にはＲＯＭに格納されているであろう。ＲＡＭは典型的に、処理サブシステム２１０４が現在処理し実行しているデータおよび／またはプログラムモジュールを含む。いくつかの実装例において、システムメモリ２１１０は、スタティックランダムアクセスメモリ（ＳＲＡＭ）またはダイナミックランダムアクセスメモリ（ＤＲＡＭ）等の複数の異なる種類のメモリを含み得る。

【０２３０】

限定ではなく一例として、図２１に示されるように、システムメモリ２１１０は、クライアントアプリケーション、ウェブブラウザ、ミッドティアアプリケーション、リレーショナルデータベース管理システム（ＲＤＢＭＳ）等を含み得るアプリケーションプログラム２１１２と、プログラムデータ２１１４と、オペレーティングシステム２１１６とを含み得る。一例として、オペレーティングシステム２１１６は、さまざまなバージョンのMicrosoft Windows（登録商標）、Apple Macintosh（登録商標）、および／またはLinux（登録商標）オペレーティングシステム、市場で入手可能な多様なUNIX（登録商標）またはUNIX（登録商標）系オペレーティングシステム（限定されないが多様なGNU/Linux（登録商標）オペレーティングシステム、Google Chrome（登録商標）OS等を含む）、および／またはiOS、Windows（登録商標）Phone、Android（登録商標） OS、BlackBerry（登録商標） 10 OS、Palm（登録商標） OSオペレーティングシステム等のモバイルオペレーティングシステムを含み得る。

【０２３１】

コンピュータ可読記憶媒体２１２２は、いくつかの実施形態の機能を提供するプログラミングおよびデータ構造を格納し得る。処理サブシステム２１０４のプロセッサによって実行されたときに上記機能を提供するソフトウェア（プログラム、コード、モジュール、命令）は、記憶サブシステム２１１８に格納し得る。一例として、コンピュータ可読記憶

媒体２１２２は、ハードディスクドライブ、磁気ディスクドライブ等の不揮発性メモリ、ＣＤ　ＲＯＭ、ＤＶＤ、Blu-Ray（登録商標）ディスク、またはその他の光学媒体等の光ディスクドライブを含み得る。コンピュータ可読記憶媒体２１２２は、限定されないが、Ｚｉｐ（登録商標）ドライブ、フラッシュメモリカード、ユニバーサルシリアルバス（ＵＳＢ）フラッシュドライブ、セキュアデジタル（ＳＤ）カード、ＤＶＤディスク、デジタルビデオテープ等を含み得る。コンピュータ読取可能記憶媒体２１２２はまた、フラッシュメモリベースのＳＳＤ、企業フラッシュドライブ、ソリッドステートＲＯＭ等といった不揮発性メモリベースのソリッドステートドライブ（ＳＳＤ）、ソリッドステートＲＡＭ、ダイナミックＲＡＭ、スタティックＲＡＭ、ＤＲＡＭベースのＳＳＤ、磁気抵抗ＲＡＭ（ＭＲＡＭ）ＳＳＤ等の揮発性メモリベースのＳＳＤ、ならびにＤＲＡＭおよびフラッシュメモリベースのＳＳＤの組合わせを用いるハイブリッドＳＳＤを含み得る。コンピュータ可読媒体２１２２は、コンピュータ可読命令、データ構造、プログラムモジュール、およびコンピュータシステム２１００のためのその他のデータのための記憶部を提供し得る。

【０２３２】

特定の実施形態において、記憶サブシステム２１００はまた、コンピュータ可読記憶媒体２１２２にさらに接続できるコンピュータ可読記憶媒体読取装置２１２０を含み得る。システムメモリ２１１０とともに、また、任意でシステムメモリ２１１０と組合わされて、コンピュータ可読記憶媒体２１２２は、包括的に、コンピュータ可読情報を格納するための、遠隔、ローカル、固定、および／またはリムーバブル記憶装置プラス記憶媒体を含み得る。

【０２３３】

特定の実施形態において、コンピュータシステム２１００は、１つ以上の仮想マシンを実行するためのサポートを提供し得る。コンピュータシステム２１００は、仮想マシンの構成および管理を容易にするためのハイパーバイザのようなプログラムを実行し得る。各仮想マシンに、メモリ、計算（たとえばプロセッサ、コア）、入出力、およびネットワーキングリソースを割当ててもよい。典型的に、各仮想マシンは自身のオペレーティングシステムを実行し、これは、コンピュータシステム２１００が実行する他の仮想マシンが実行するオペレーティングシステムと同一でも異なっていてもよい。したがって、複数のオペレーティングシステムがコンピュータシステム２１００によって同時に実行される可能性がある。各仮想マシンは一般的にその他の仮想マシンから独立して実行される。

【０２３４】

通信サブシステム２１２４は、他のコンピュータシステムおよびネットワークへのインターフェイスを提供する。通信サブシステム２１２４は、コンピュータシステム２１００以外のシステムからデータを受信しコンピュータシステム２１００以外のシステムにデータを送信するためのインターフェイスとして機能する。たとえば、通信サブシステム２１２４は、クライアントデバイスとの間で情報を送受信するためのインターネットを介した１つ以上のクライアントデバイスへの通信チャネルを確立することができるようにする。

【０２３５】

通信サブシステム２１２４は、有線および／または無線通信プロトコル双方をサポートしてもよい。たとえば、特定の実施形態において、通信サブシステム２１２４は、（たとえば携帯電話技術、３Ｇ、４ＧまたはＥＤＧＥ（enhanced data rates for global evolution）等の高度データネットワーク技術、ＷｉＦｉ（ＩＥＥＥ　８０２．１１系列基準、または他の移動通信技術、またはそれらの任意の組合わせを用いて）無線音声および／またはデータネットワークにアクセスするための無線周波数（ＲＦ）トランシーバーコンポーネント、全地球測位システム（ＧＰＳ）レシーバーコンポーネント、および／または他のコンポーネントを含み得る。いくつかの実施形態において、通信サブシステム２１２４は、無線インターフェイスに加えて、またはその代わりに、有線ネットワークコネクティビティ（たとえばイーサネット（登録商標））を提供することができる。

【０２３６】

通信サブシステム２１２４は、さまざまな形態のデータを受信し送信することができる。たとえば、いくつかの実施形態において、通信サブシステム２１２４は、構造化および／または非構造化データフィード２１２６、イベントストリーム２１２８、イベントアップデート２１３０等の形態の入力通信を受信し得る。たとえば、通信サブシステム２１２４は、Twitter（登録商標）フィード、Facebook（登録商標）更新、Rich Site Summary（ＲＳＳ）フィード等のウェブフィード、および／または１つ以上の第三者情報源からのリアルタイム更新等の、ソーシャルメディアネットワークおよび／または他の通信サービスのユーザからのリアルタイムのデータフィード２１２６を、受信（または送信）するように構成してもよい。

【０２３７】

特定の実施形態において、通信サブシステム２１２４は、明示的な終わりのない本質的に連続的または無限であってもよいリアルタイムイベントのイベントストリーム２１２８および／またはイベントアップデート２１３０を含み得る、連続データストリームの形態のデータを受信するように構成し得る。連続データを生成するアプリケーションの例は、たとえば、センサデータアプリケーション、株式相場表示装置、ネットワーク性能測定ツール（たとえば、ネットワーク監視およびトラフィック管理アプリケーション）、クリックストリーム分析ツール、自動車トラフィック監視等を含み得る。

【０２３８】

通信サブシステム２１２４はまた、構造化データおよび／または非構造化データフィード２１２６、イベントストリーム２１２８、イベントアップデート２１３０等を、コンピュータシステム２１００に結合された１つ以上のストリーミングデータソースコンピュータと通信し得る１つ以上のデータベースに出力するように構成し得る。

【０２３９】

コンピュータシステム２１００は、ハンドヘルドポータブルデバイス（たとえばiPhone（登録商標）携帯電話、iPad（登録商標）コンピューティングタブレット、ＰＤＡ）、ウェアラブルデバイス（たとえばGoogle Glass（登録商標）ヘッドマウントディスプレイ）、パーソナルコンピュータ、ワークステーション、メインフレーム、キオスク、サーバラック、またはその他任意のデータ処理システムを含むさまざまな種類のうちの１つであってもよい。

【０２４０】

コンピュータおよびネットワークの性質は常に変化しているので、図２１に示されるコンピュータシステム２１００の説明は、具体的な一例を意図しているに過ぎない。図２１に示されるシステムよりも多いかまたは少ないコンポーネントを有する他の多くの構成が可能である。本明細書において提供される開示および教示に基づいて、当業者はさまざまな実施形態を実装するための他のやり方および／または方法を認識するであろう。

【０２４１】

本発明のある実施形態において、データ強化システムが提供される。データ強化システムは、コンピューティングシステムを含むクラウドコンピューティング環境において実行可能であり、データ強化システムは、少なくとも１つの通信ネットワークを通して複数の入力データソース（たとえば図１に示されるデータソース１０４）に通信可能に結合される。

【０２４２】

データ強化システムはマッチング部と、類似性メトリック部と、カテゴリ分類部とを含む。このマッチング部、類似性メトリック部、およびカテゴリ分類部はそれぞれ、たとえば図３に示される、マッチングモジュール３１２、類似性メトリックモジュール３１４、およびカテゴリ分類部３１８であってもよい。

【０２４３】

マッチング部は、複数の入力データソースから受けた入力データセットを、参照ソース（たとえば図３に示される知識ソース３４０）から取得した１つ以上の参照データセットと比較するように構成される。類似性メトリック部は、１つ以上の参照データセット各々

について類似性メトリックを計算するように構成され、類似性メトリックは、入力データセットとの比較における１つ以上の参照データセット各々の類似の程度を示し、類似性メトリック部は、類似性メトリックに基づいて入力データセットと１つ以上の参照データセットとの間の一致を識別するように構成される。カテゴリ分類部は、１つ以上の参照データセット各々について計算した類似性メトリックを示し、かつ、入力データセットと１つ以上の参照データセットとの間の識別された一致を示すグラフィカルユーザインターフェイスを生成するように構成される。加えて、１つ以上の参照データセット各々について計算した類似性メトリックを示し、かつ、入力データセットと１つ以上の参照データセットとの間の識別された一致を示す、グラフィカルなビジュアライゼーションが、ユーザインターフェイスを用いてレンダリングされる。

【０２４４】

本発明のある実施形態において、データ強化システムはさらに、たとえば図３に示される知識スコアリングモジュール３１６に対応し得る知識スコアリング部を含む。

【０２４５】

本発明ある実施形態において、上記１つ以上の参照データセットは、ドメインに対応付けられた用語を含み、類似性メトリックは、１つ以上の参照データセット各々について計算されたマッチングスコアであり、マッチングスコアは、知識スコアリング部によって、参照データセットに関するメトリックを示す第１の値と入力データセットと参照データセットとの比較に基づくメトリックを示す第２の値とを含む１つ以上の値を用いて計算され、グラフィカルなビジュアライゼーションをレンダリングすることにより、マッチングスコアの計算に用いられる１つ以上の値が表示される。

【０２４６】

本発明のある実施形態において、上記１つ以上の値は、入力データセットとデータセットとの間で一致する用語の度数値と、データセットの母集団値と、入力データセットとデータセットとの間で一致する異なる用語の数を示す固有マッチング値と、データセット内の用語の数を示すドメイン値と、データセットのキュレーションの程度を示すキュレーションレベルとを含む。

【０２４７】

本発明のある実施形態において、カテゴリ分類部はさらに、アグリゲーションサービスから取得した増補データに基づいて増補リストを生成し、増補リストに基づいて入力データセットを増補し、１つ以上の参照データセットに基づいてインデックス付トライグラム表を生成し、増補後の入力データセット内のワードごとに、そのワードのトライグラムを作成し、各トライグラムをインデックス付トライグラム表と比較し、トライグラムのうちの第１のトライグラムと一致するトライグラムに対応付けられたインデックス付トライグラム表の中のワードを識別し、このワードをトライグラム増補データセットに格納し、トライグラム増補データセットを１つ以上の参照データセットと比較し、この比較に基づいて、トライグラム増補データセットと１つ以上の参照データセットとの間の一致を判断するように、構成される。上記１つ以上の参照データセットと比較される入力データは、増補リストに基づいて増補され、入力データセットと１つ以上の参照データセットとの間の一致を識別することは、比較に基づくトライグラム増補データセットと１つ以上の参照データセットとの間の一致を用いて実行される。

【０２４８】

本発明のある実施形態において、別のデータ強化システムが提供される。データ強化システムは、コンピューティングシステムを含むクラウドコンピューティング環境において実行可能であり、データ強化システムは、少なくとも１つの通信ネットワークを通して複数の入力データソース（たとえば図１に示されるデータソース１０４）に通信可能に結合される。

【０２４９】

データ強化システムはマッチング部と、類似性メトリック部とを含む。このマッチング部および類似性メトリック部はそれぞれ、たとえば図３に示されるマッチングモジュール

３１２および類似性メトリックモジュール３１４であってもよい。
【０２５０】
マッチング部は、複数の入力データソースから受けた入力データセットを、参照ソース（たとえば図３に示される知識ソース３４０）から取得した１つ以上の参照データセットと比較するように構成される。類似性メトリック部は、１つ以上の参照データセット各々について類似性メトリックを計算するように構成され、類似性メトリックは、入力データセットとの比較における１つ以上の参照データセット各々の類似の程度を示し、類似性メトリック部は、類似性メトリックに基づいて入力データセットと１つ以上の参照データセットとの間の一致を識別するように構成される。１つ以上の参照データセット各々について計算した類似性メトリックを示し、かつ、入力データセットと１つ以上の参照データセットとの間の識別された一致を示すマッチング情報とともに、入力データセットは格納される。
【０２５１】
本発明のある実施形態において、データ強化システムはカテゴリ分類部をさらに含み、カテゴリ分類部はたとえば図３に示されるカテゴリ分類モジュール３１８に対応し得る。カテゴリ分類部は、入力データセットと１つ以上の参照データセットとの間の一致の識別に基づいて入力データセットのカテゴリラベルを識別するように構成され、入力データセットはこのカテゴリラベルに対応付けて格納される。
【０２５２】
本発明のある実施形態において、類似性メトリックは、Jaccard係数、Tversky係数、またはDice-Sorensen係数のうちの１つ以上を用いて計算される。
【０２５３】
本発明のある実施形態において、入力データセットは、グラフマッチングまたは意味類似性マッチングのうちの１つ以上を用いて、１つ以上の参照データセットと比較される。
【０２５４】
上記ユニット／モジュール（たとえばエンジン）の特定のオペレーションプロセスの代わりに、同一概念を共有する関連の方法／システムの実施形態の対応するステップ／コンポーネントを参照してもよく、この参照は、関連するユニット／モジュールの開示とみなされることが、当業者には明らかであろう。したがって、説明を適宜簡潔にするために、特定のオペレーションプロセスの中には、繰返しまたは詳細に説明しないものもある。
【０２５５】
また、上記ユニット／モジュールは、電子デバイスにおいて、ソフトウェア、ハードウェア、および／またはソフトウェアとハードウェアの組合わせとして実装できることも、当業者には明らかであろう。別々のコンポーネントとして説明されているコンポーネントは、物理的に分離されていてもいなくてもよい。特に、本発明の各実施形態に従うコンポーネントは、１つの物理的コンポーネントに一体化されていてもよく、さまざまな別々の物理的コンポーネントに存在していてもよい。電子デバイスにおけるユニットのさまざまな実装はすべて、本発明の保護範囲に含まれる。
【０２５６】
ユニット、装置、およびデバイスは、周知のまたは今後開発されるソフトウェア、ハードウェア、および／またはこのようなソフトウェアとハードウェアの組合わせの形態で実装し得ることが、理解されるはずである。
【０２５７】
図３に示されるオペレーションを、特定のアプリケーション環境に応じて、ソフトウェア、ハードウェア、および／またはこのようなソフトウェアとハードウェアの組合わせの形態で実装し得ることは、当業者には明らかである。ステップのうちの少なくともいくつかを、メモリに命令が格納されている汎用プロセッサで命令を実行することによって実装できることは、当業者には明らかである。ステップのうちの少なくともいくつかを、ＤＳＰ、ＦＰＧＡ、ＡＳＩＣを含むがこれらに限定されないさまざまなハードウェアによって実装できることも、当業者には明らかである。たとえば、いくつかの実施形態における「

オペレーション」は、「オペレーション」の機能を実装するＣＰＵ、または、ＤＳＰ、ＦＰＧＡ、ＡＳＩＣ等の専用プロセッサにおいて命令が実行されることによって実装されてもよい。

【０２５８】

本発明の特定の実施形態を説明してきたが、さまざまな修正、変更、代替構成、および均等物も本発明の範囲に含まれる。本発明の実施形態は、特定の具体的なデータ処理環境におけるオペレーションに限定されるのではなく、複数のデータ処理環境において自由に機能する。加えて、特定の一連のトランザクションおよびステップを用いて本発明の実施形態を説明してきたが、本発明の範囲が上述の一連のトランザクションおよびステップに限定されないことは当業者にとって明らかであろう。上記実施形態のさまざまな特徴および側面は、個別にまたは共同で使用してもよい。

【０２５９】

さらに、ハードウェアおよびソフトウェアの特定の組合わせを用いて本発明の実施形態を説明してきたが、ハードウェアおよびソフトウェアの他の組合わせも本発明の範囲に含まれることが認識されねばならない。本発明の実施形態は、ハードウェアのみで、ソフトウェアのみで、またはそれらの組合わせを用いて実現し得る。本明細書に記載のさまざまなプロセスは、同一のプロセッサ上で実装できる、または、任意の組合わせの異なるプロセッサ上で実装できる。したがって、コンポーネントまたはモジュールが特定のオペレーションを実行するように構成されていると説明されている場合、このような構成は、たとえば、電子回路をそのオペレーションを実行するように設計することにより、または、プログラム可能な電子回路（マイクロプロセッサ等）をそのオペレーションを実行するようにプログラムすることにより、または、これらを任意に組合わせることにより、実現することができる。プロセスは、限定されないが従来のプロセス間通信技術を含むさまざまな技術を用いてやりとりすることができ、異なるプロセス対が異なる技術を用いてもよく、または、同一のプロセス対がその時々で異なる技術を用いてもよい。

【０２６０】

したがって、明細書および図面は、限定的な意味ではなく例示的な意味で考慮されねばならない。しかしながら、特許請求の範囲に記載されている広い精神および範囲から逸脱することなく、追加、削減、削除、ならびにその他の修正および変更を行ない得ることは、明らかであろう。よって、本発明の特定の実施形態を説明したが、これらの実施形態は限定することを意図していない。さまざまな修正および均等物は以下の特許請求の範囲に含まれる。修正は、開示されている特徴の関連する任意の組合わせを含む。

【０２６１】

本明細書に記載のデータ強化サービスは、ＩＭＩ、ＯＤＥＣＳ、および／またはBig Data Prep（ビッグデータ準備）と呼ばれることもある。

【図１】

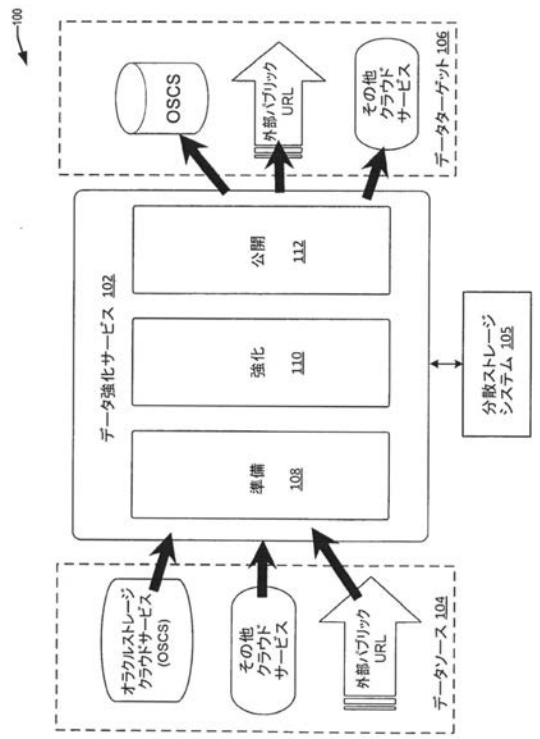


FIG. 1

【図２】

FIG. 2

【図３】

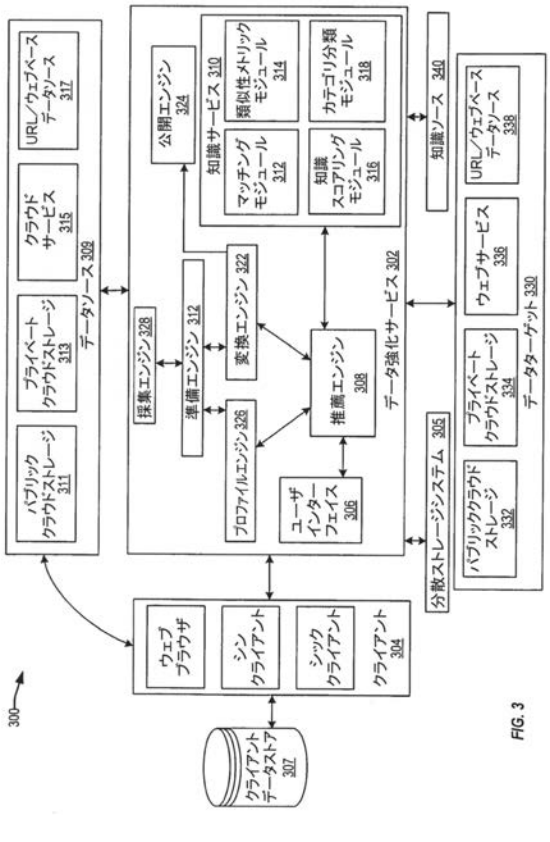


FIG. 3

【図４Ａ】

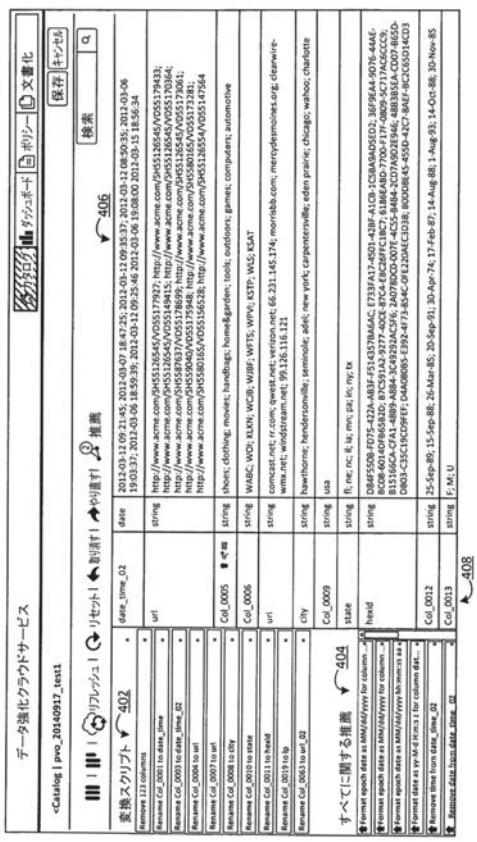
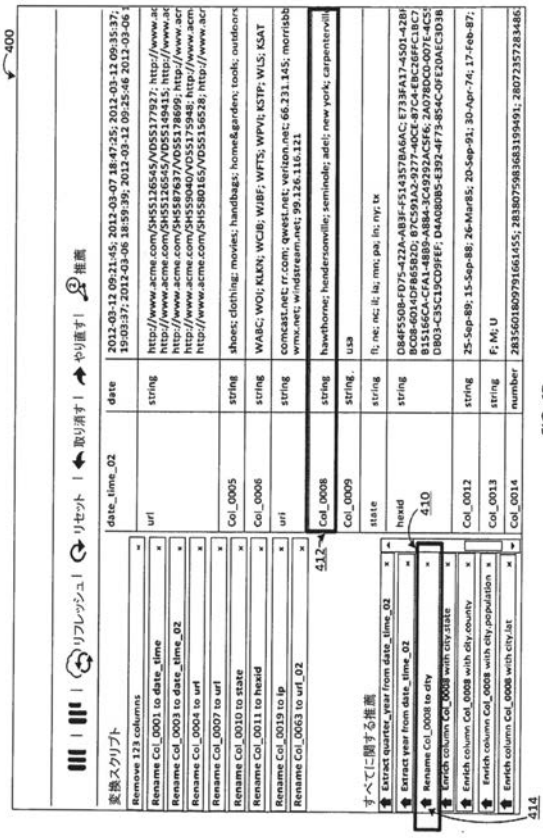
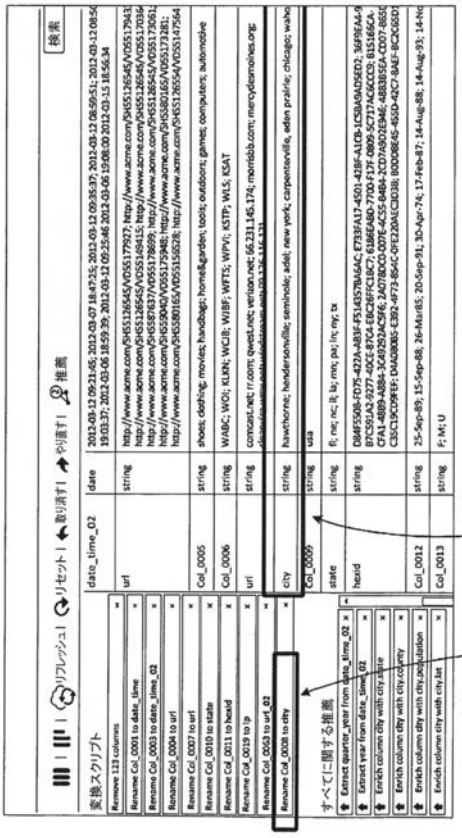


FIG. 4A

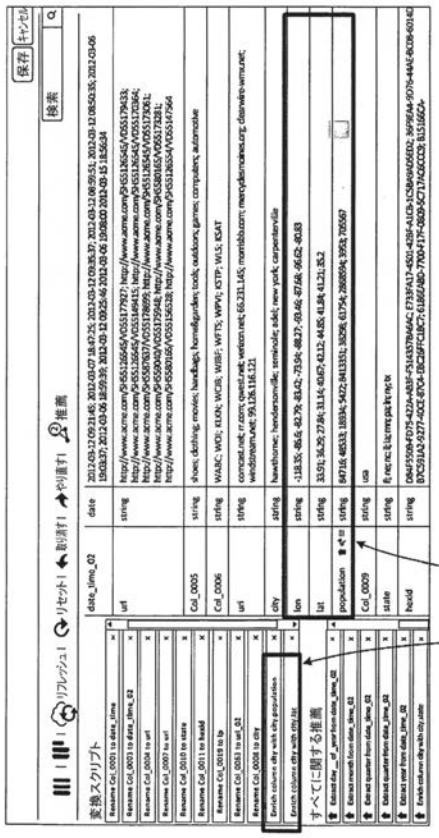
【図４Ｂ】



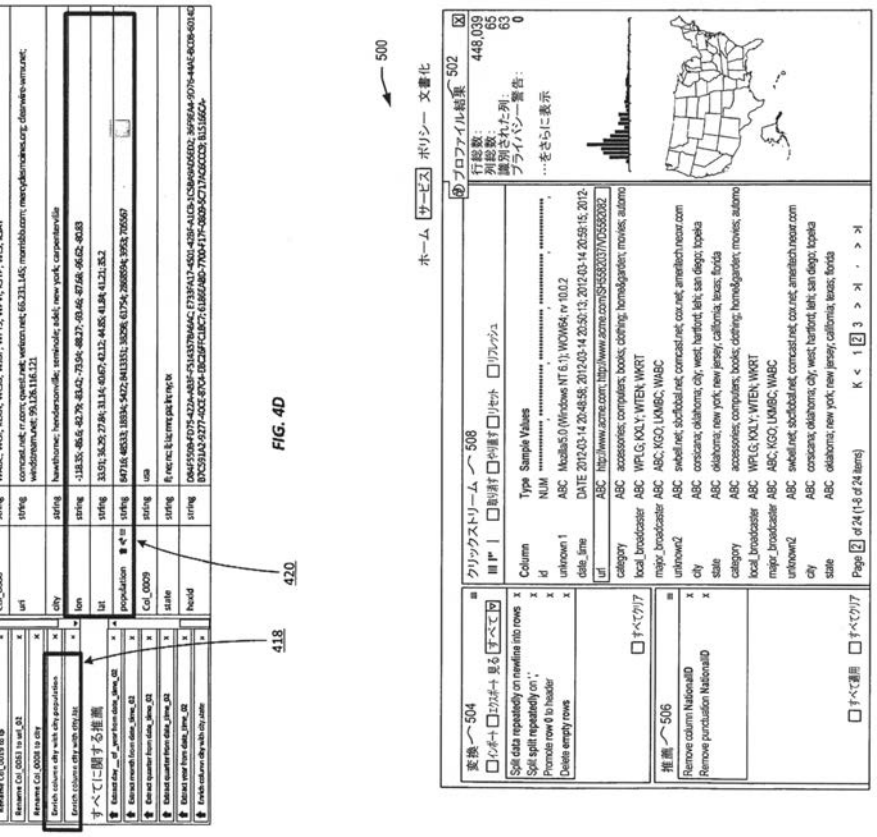
【図４Ｃ】



【図４Ｄ】



【図５Ａ】



【図５Ｂ】

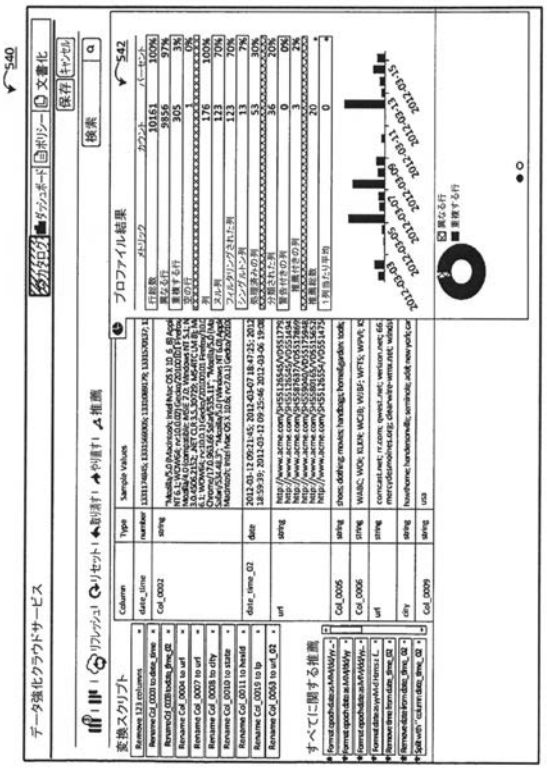


FIG. 5B

【図５Ｃ】

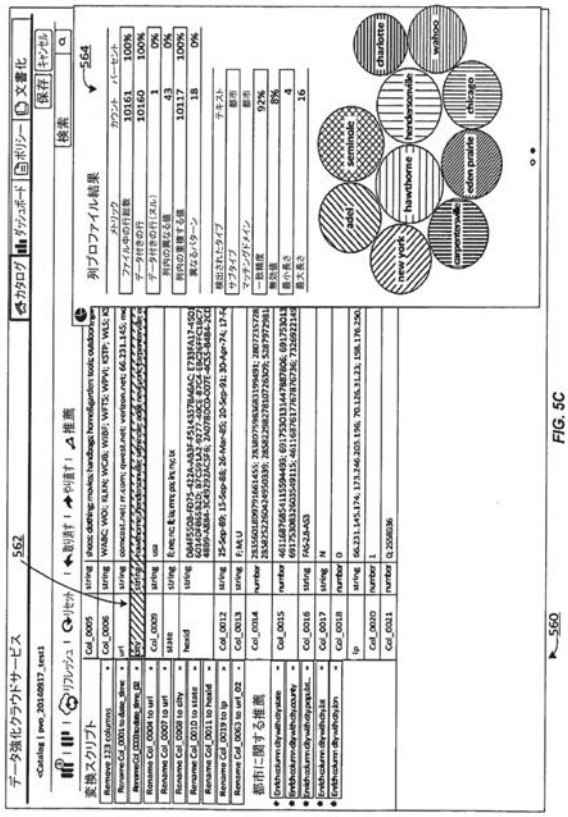


FIG. 5C

【図５Ｄ】

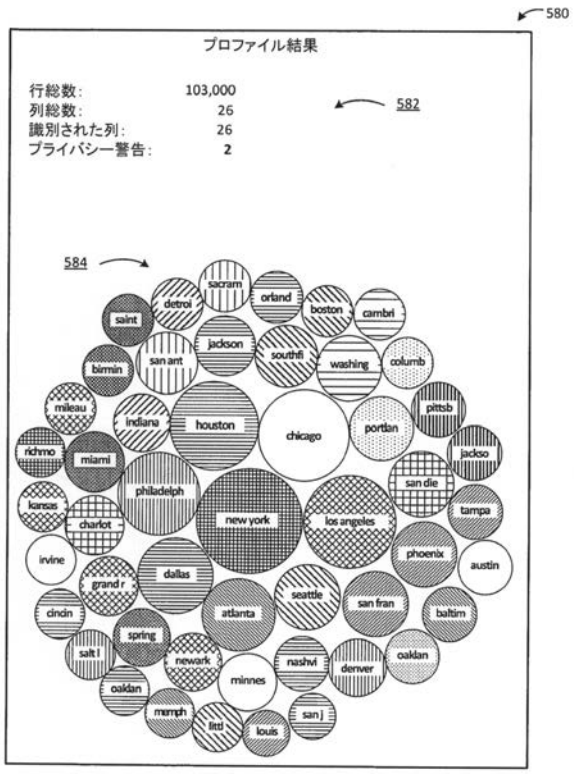


FIG. 5D

【図６】

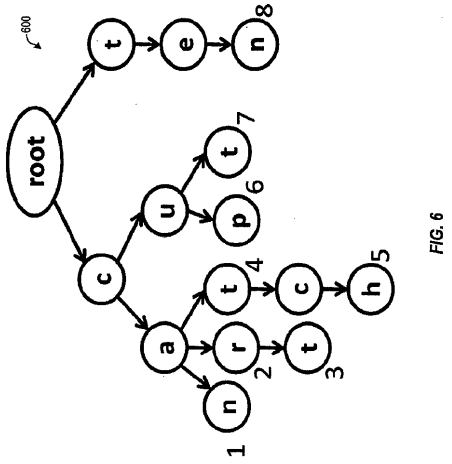


FIG. 6

【図７】

700

文字オフセット	文字	部分一致
初期状態	n/a	(1, ROOT)
1	c	(1, c) (2, ROOT)
2	a	(1, a) (3, ROOT)
3	c	(3, c) (4, ROOT)
4	a	(3, a) (5, ROOT)
5	t	(3, t) (5, t) (6, ROOT)
6	c	(3, c) (6, c) (7, ROOT)
7	h	(3, h) (8, ROOT)

FIG. 7

【図８】

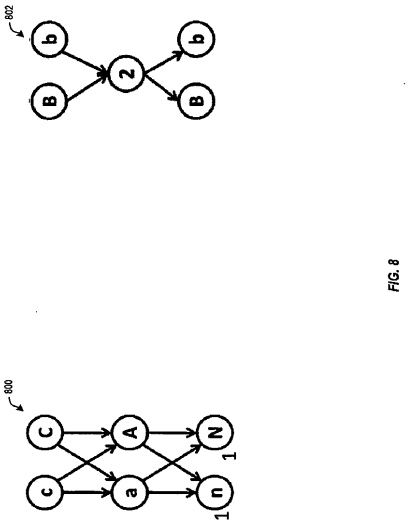


FIG. 8

【図９】

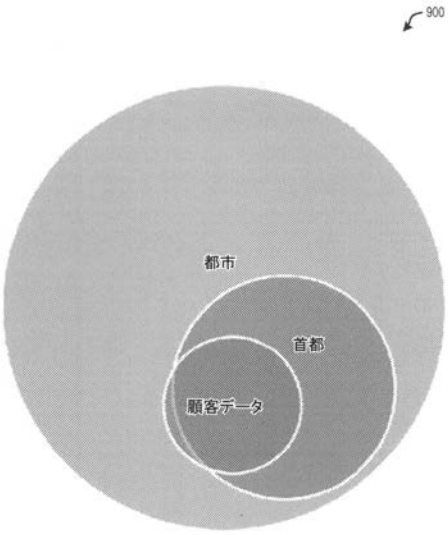


FIG. 9

【図１０】

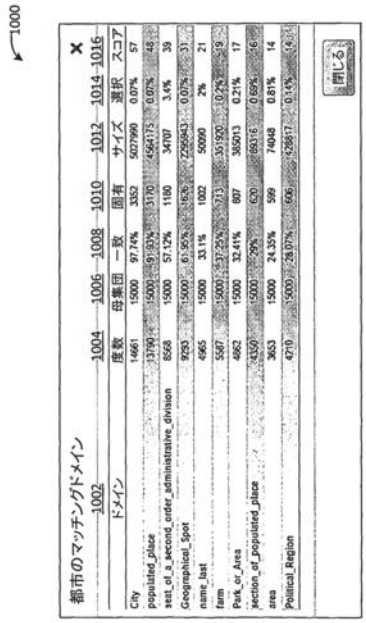


FIG. 10

【図１１】

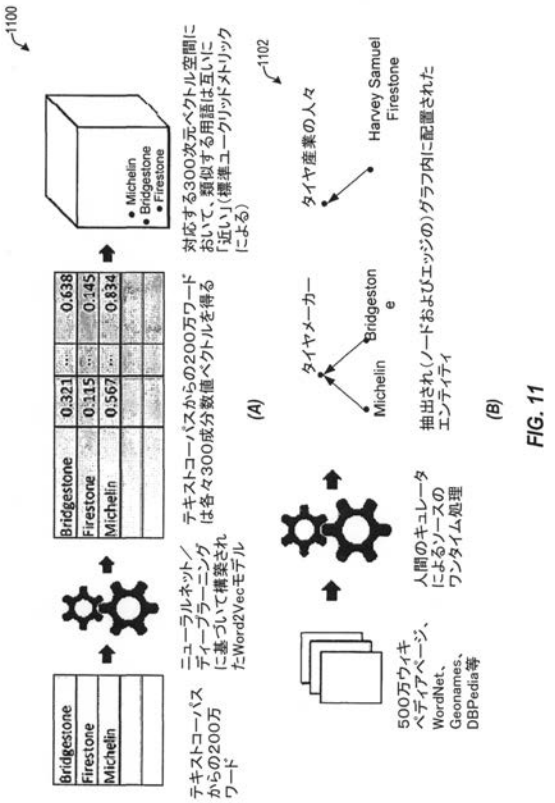


FIG. 11

【図１２】

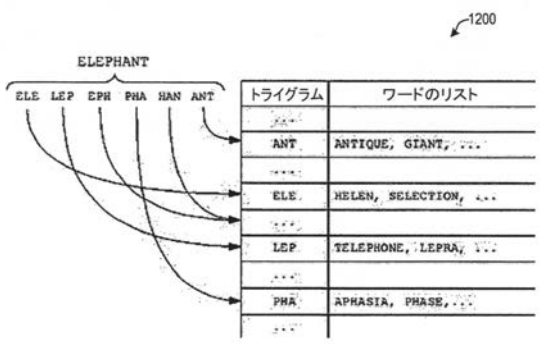


FIG.12

【図１３】

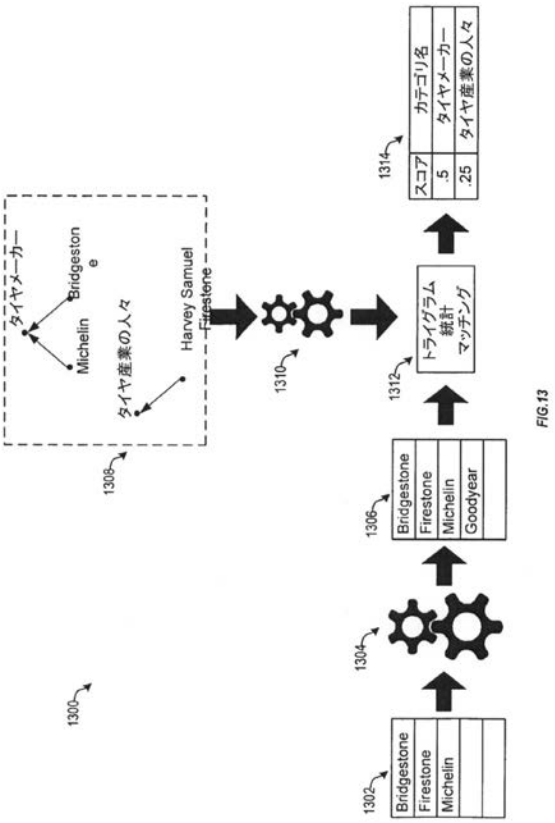


FIG.13

【図１４】

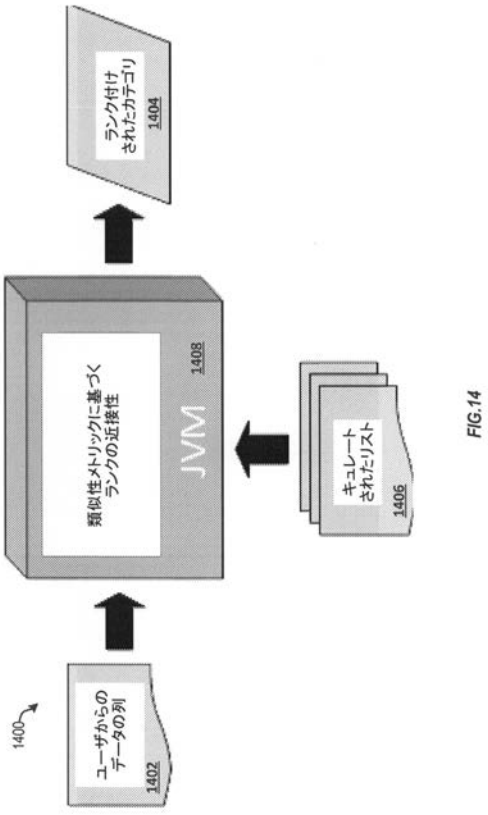


FIG.14

【図１５】

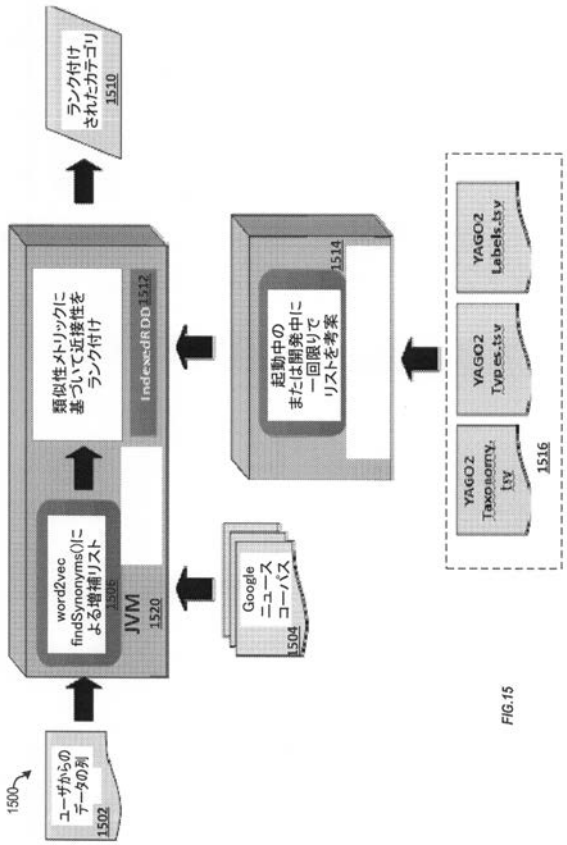


FIG.15

【図１６】

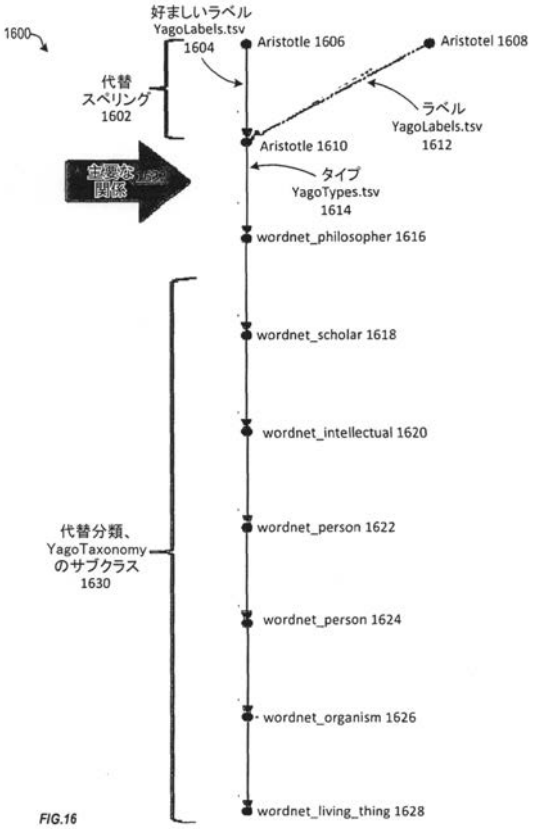


FIG.16

【図１７】

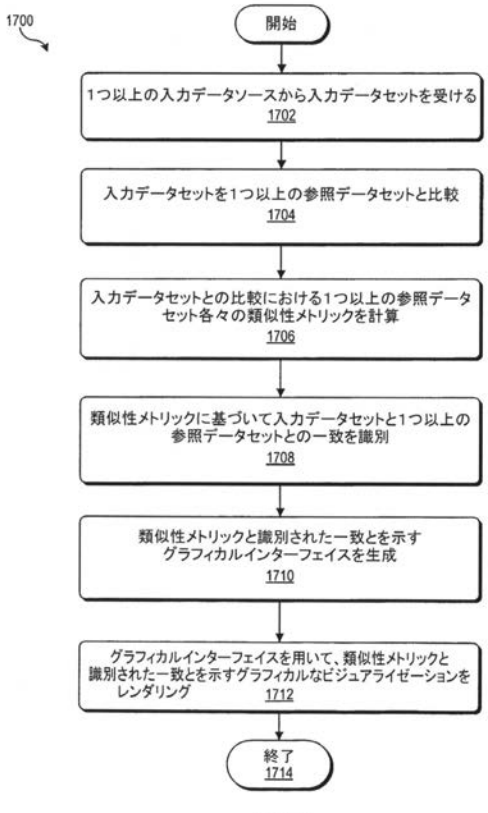


FIG. 17

【図１８】

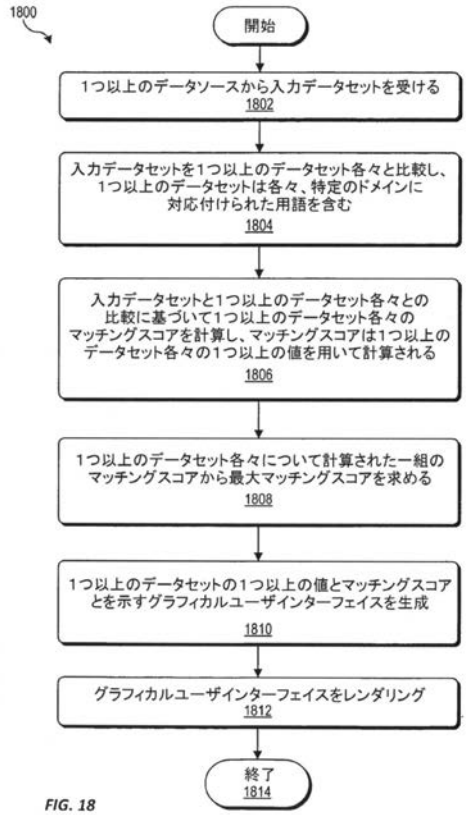


FIG. 18

【図１９】

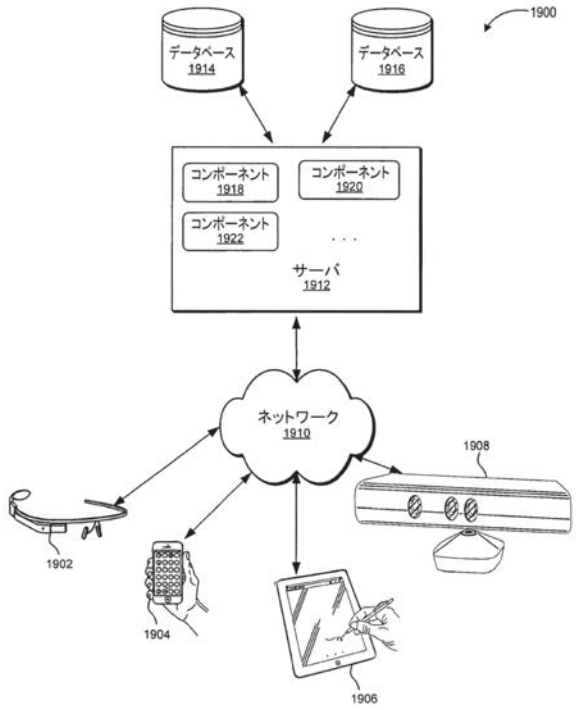


FIG. 19

【図２０】

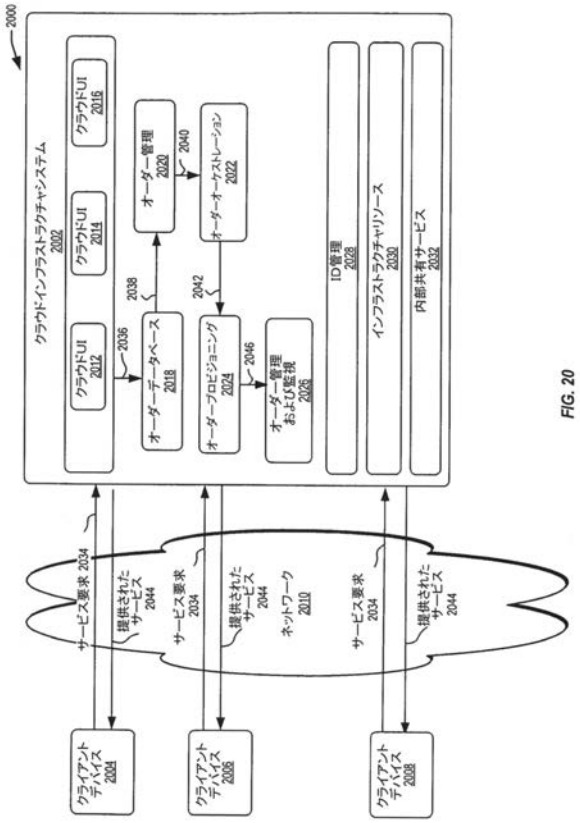


FIG. 20

【図２１】

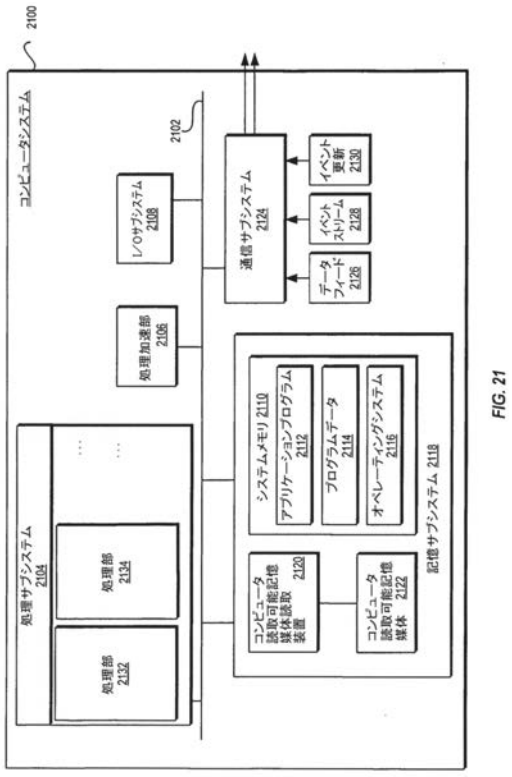


FIG. 21

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/052190

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/30 G06Q30/02
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F G06Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2013/110792 A1 (HUDIS EFIM [US] ET AL) 2 May 2013 (2013-05-02) paragraphs [0004], [0080], [0082], [0073] -----	1-24
X	US 2012/101975 A1 (KHOSRAVY MOE [US]) 26 April 2012 (2012-04-26) paragraphs [0006], [0007], [0008], [0009], [0010] -----	1-24
A	US 2014/074829 A1 (SCHMIDT MICHAEL [US]) 13 March 2014 (2014-03-13) paragraph [0006] - paragraph [0023] -----	1,12,16
	-/--	

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
2 December 2015	09/12/2015

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Stan, Johann

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/052190

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	BENJAMIN MARKINES ET AL: "Evaluating similarity measures for emergent semantics of social tagging", INTERNATIONAL WORLD WIDE WEB CONFERENCE 18TH; 20090420 - 20090424, 24 April 2009 (2009-04-24), pages 641-650, XP058025635, DOI: 10.1145/1526709.1526796 ISBN: 978-1-60558-487-4 Section 3.3 (Similarity Measures); page 645 - page 646	8,9,19, 20,23,24
A	AMINUL ISLAM ET AL: "Text Similarity Using Google Tri-grams", 28 May 2012 (2012-05-28), ADVANCES IN ARTIFICIAL INTELLIGENCE, SPRINGER BERLIN HEIDELBERG, BERLIN, HEIDELBERG, PAGE(S) 312 - 317, XP047004690, ISBN: 978-3-642-30352-4 Section 3 Proposed Method; page 313 - page 316	6,15
A	Davide Buscaldi ET AL: "LIPN-CORE: Semantic Text Similarity using n-grams, WordNet, Syntactic Analysis, ESA and Information Retrieval based Features", Second Joint Conference on Lexical and Computational Semantics Proceedings of the Main Conference and the Shared Task, 13 June 2013 (2013-06-13), pages 162-168, XP055228445, Retrieved from the Internet: URL:https://aclweb.org/anthology/S/S13/S13-1023.pdf [retrieved on 2015-11-12] the whole document	6,15
A	US 2011/106791 A1 (MAIM ENRICO [FR]) 5 May 2011 (2011-05-05) paragraph [0005] - paragraph [0060]	1-24
A	JULIAN SEDDING ET AL: "WordNet-based text document clustering", PROCEEDINGS OF THE 3RD WORKSHOP ON ROBUST METHODS IN ANALYSIS OF NATURAL LANGUAGE DATA, ROMAND '04, 1 January 2004 (2004-01-01), pages 104-113, XP055232744, Morristown, NJ, USA DOI: 10.3115/1621445.1621458 page 2, left-hand column, paragraph 2 page 3, left-hand column, paragraph 4	6,15

1

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2015/052190

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2013110792 A1	02-05-2013	CN 102999561 A	27-03-2013
		EP 2771810 A1	03-09-2014
		TW 201322024 A	01-06-2013
		US 2013110792 A1	02-05-2013
		WO 2013062877 A1	02-05-2013
US 2012101975 A1	26-04-2012	AU 2011318496 A1	02-05-2013
		CN 102419744 A	18-04-2012
		EP 2630592 A2	28-08-2013
		US 2012101975 A1	26-04-2012
		US 2015286730 A1	08-10-2015
		US 2015286731 A1	08-10-2015
		WO 2012054179 A2	26-04-2012
US 2014074829 A1	13-03-2014	US 2014074829 A1	13-03-2014
		US 2014172773 A1	19-06-2014
US 2011106791 A1	05-05-2011	EP 2181402 A1	05-05-2010
		US 2011106791 A1	05-05-2011
		US 2012117500 A1	10-05-2012
		WO 2008107338 A1	12-09-2008

フロントページの続き

(31)優先権主張番号 14/864,485
(32)優先日 平成27年9月24日(2015.9.24)
(33)優先権主張国 米国(US)

(81)指定国 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JP,KE,KG,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US

（特許庁注：以下のものは登録商標）
１．ＷＩＮＤＯＷＳ　ＰＨＯＮＥ

(72)発明者 クレイダー，マーク
アメリカ合衆国、８０００５　コロラド州、アーバダ、ウエスト・エイティーフォース・ドライブ、１２４１９
(72)発明者 マラク，マイケル
アメリカ合衆国、８０２０９　コロラド州、デンバー、サウス・ワシントン・ストリート、５７３
(72)発明者 マリー，グレン・アレン
アメリカ合衆国、８０３０５　コロラド州、ボールダー、グリネル・アベニュ、４３６５