

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7686923号
(P7686923)

(45)発行日 令和7年6月3日(2025.6.3)

(24)登録日 令和7年5月26日(2025.5.26)

(51)国際特許分類

F I

G 0 6 F 16/2455(2019.01)

G 0 6 F 16/2455

請求項の数 9 (全35頁)

(21)出願番号	特願2024-502726(P2024-502726)	(73)特許権者	000004237
(86)(22)出願日	令和4年2月28日(2022.2.28)		日本電気株式会社
(86)国際出願番号	PCT/JP2022/008227		東京都港区芝五丁目7番1号
(87)国際公開番号	WO2023/162206	(74)代理人	100103894
(87)国際公開日	令和5年8月31日(2023.8.31)		弁理士 家入 健
審査請求日	令和6年7月30日(2024.7.30)	(72)発明者	林 勝悟
			東京都港区芝五丁目7番1号 日本電気株式会社内
		(72)発明者	小山田 昌史
			東京都港区芝五丁目7番1号 日本電気株式会社内
		(72)発明者	草野 元紀
			東京都港区芝五丁目7番1号 日本電気株式会社内
		審査官	早川 学

最終頁に続く

(54)【発明の名称】 情報処理装置、情報処理方法及び情報処理プログラム

(57)【特許請求の範囲】

【請求項1】

レコード対を取得する取得手段と、
前記レコード対を変換することによって変換済レコード対を生成する変換手段と、
前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出する類似度算出手段と、
前記類似度算出手段が算出した類似度を出力する出力手段と
を備え、
前記変換手段は、前記レコード対の特徴に応じて複数のモデル候補からモデルを選択し、
選択したモデルの入力に対応する形式に当該レコード対を変換することによって前記変換済レコード対を生成する、
情報処理装置。

【請求項2】

レコード対を取得する取得手段と、
前記レコード対を変換することによって変換済レコード対を生成する変換手段と、
前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出する類似度算出手段と、
前記類似度算出手段が算出した類似度を出力する出力手段と、
を備え、
前記モデルには、画像分類モデルが含まれており、

前記変換手段による変換処理には、前記レコード対を画像に変換することによって前記変換済レコード対を生成する処理が含まれる
情報処理装置。

【請求項 3】

レコード対を取得する取得手段と、
前記レコード対を変換することによって変換済レコード対を生成する変換手段と、
前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する
類似度を算出する類似度算出手段と、
前記類似度算出手段が算出した類似度を出力する出力手段と、
を備え、

10

前記モデルには、音声分類モデルが含まれており、

前記変換手段による変換処理には、前記レコード対を音声に変換することによって前記変換済レコード対を生成する処理が含まれる
情報処理装置。

【請求項 4】

レコード対を取得する取得手段と、
前記レコード対を変換することによって変換済レコード対を生成する変換手段と、
前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する
類似度を算出する類似度算出手段と、
前記類似度算出手段が算出した類似度を出力する出力手段と、
を備え、

20

前記モデルには、グラフ分類モデルが含まれており、

前記変換手段による変換処理には、前記レコード対をグラフに変換することによって前記変換済レコード対を生成する処理が含まれる、
情報処理装置。

【請求項 5】

レコード対を取得する取得手段と、
前記レコード対を変換することによって変換済レコード対を生成する変換手段と、
前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する
類似度を算出する類似度算出手段と、
前記類似度算出手段が算出した類似度を出力する出力手段と、
を備え、

30

前記取得手段は、補助レコードを更に取得し、

前記変換手段は、前記補助レコードを変換することによって変換済補助レコードを生成し、

前記類似度算出手段は、前記変換済レコード対及び前記変換済補助レコードを前記モデルに入力することによって、前記変換済レコード対に関する類似度を算出する
情報処理装置。

【請求項 6】

前記モデルには、質問文と応答文とを入力とする質問応答モデルが含まれており、

40

前記変換手段は、前記レコード対に含まれる一方のレコードを質問文に変換し、前記レコード対に含まれる他方のレコード及び前記補助レコードの各々を応答文に変換することによって前記変換済レコード対を生成する
請求項 5 に記載の情報処理装置。

【請求項 7】

レコード対を取得する取得手段と、
前記レコード対を変換することによって変換済レコード対を生成する変換手段と、
前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する
類似度を算出する類似度算出手段と、
前記類似度算出手段が算出した類似度を出力する出力手段と、

50

を備え、

前記類似度算出手段は、前記レコード対に関して複数の類似度を算出し、

前記出力手段は、前記複数の類似度を統合して得られる統合後の類似度を出力し、

前記モデルは、当該モデルに入力される２つの要素の互いの入れ替えに対して非対称性を有するモデルであり、

前記類似度算出手段は、

前記モデルに対して、前記レコード対に含まれる２つのレコードを互いに入れ替えずに入力することによって第１の類似度を算出し、

前記モデルに対して、前記レコード対に含まれる２つのレコードを互いに入れ替えてから入力することによって第２の類似度を算出する

情報処理装置。

10

【請求項 8】

少なくとも１つのプロセッサが、

レコード対を取得することと、

前記レコード対を変換することによって変換済レコード対を生成することと、

前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出することと、

前記算出した類似度を出力することと

を含み、

前記変換済レコード対を生成する工程において、前記少なくとも１つのプロセッサが、
前記レコード対の特徴に応じて複数のモデル候補からモデルを選択し、選択したモデルの入力に対応する形式に当該レコード対を変換することによって前記変換済レコード対を生成する、

20

情報処理方法。

【請求項 9】

コンピュータに、

レコード対を取得する取得処理と、

前記レコード対を変換することによって変換済レコード対を生成する変換処理と、

前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出する類似度算出処理と、

30

前記類似度算出処理において算出した類似度を出力する出力処理と
を実行させ、

前記変換処理は、前記レコード対の特徴に応じて複数のモデル候補からモデルを選択し、
選択したモデルの入力に対応する形式に当該レコード対を変換することによって前記変換済レコード対を生成する、

情報処理プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、情報処理装置、情報処理方法及び情報処理プログラムに関する。

40

【背景技術】

【0002】

異なるデータセットに格納されたレコードから同一の又は類似するレコードの組み合わせを特定して対応付ける処理が行われている。このような処理は名寄せ処理とも呼ばれる。名寄せ処理によりテーブルの一元管理及びデータの拡張等が可能となる。名寄せ処理を行う技術として、機械学習によるマッチングを行う技術がある。例えば、特許文献１には、レコード対の類似度を計算する類似度関数を複数用いてレコード対の類似度を計算し、訓練データを用いた教師あり機械学習により類似度の重みを学習する装置が記載されている。ここで、訓練データは、レコードの組み合わせとそれらが同一であるかどうかを示すラベルが付与されたデータセットである。また、非特許文献１には、教師あり機械学習に

50

より名寄せを行うD I T T Oと呼ばれる技術が記載されている。また、非特許文献2には、訓練データを用いない教師なし機械学習によりレコードを照合するZ e r o E Rと呼ばれる技術が記載されている。

【0003】

また、近年、機械学習により生成されるモデルとして言語モデル（例えば、非特許文献3～5）及び画像分類モデル（例えば、非特許文献6）等が提案されている。

【先行技術文献】

【特許文献】

【0004】

【文献】日本国特開2019-185244号公報

10

【非特許文献】

【0005】

【文献】Yuliang Li, et. al., Deep Entity Matching with Pre-Trained Language Models, PVLDB 2021

【文献】Renzhi Wu, et. al., ZeroER: Entity Resolution using Zero Labeled Examples, SIGMOD, 2020

【文献】Yinhan Liu, et. al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv 2019

【文献】Sinong Wang, et. al., Entailment as Few-Shot Learner, arXiv 2021

【文献】Siddhant Garg, et. al., TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection, AAAI 2020

20

【文献】Kaiming He, et. al., Deep Residual Learning for Image Recognition, CVPR 2016

【発明の概要】

【発明が解決しようとする課題】

【0006】

しかしながら、教師あり機械学習は多量の訓練データを必要とするため、特許文献1及び非特許文献1に記載の技術においては、訓練データの収集のためのコスト及び時間がかさんでしまい、また異質データに対応できないという問題があった。ここで、異質データとは、レコードの組み合わせであり、データの形式が同一でないものを指す。また、非特許文献2に記載の教師なし機械学習であるZ e r o E Rでは、訓練データは不要であるものの、属性がアライメントされている必要があるため、属性が異なる異質データに対し適用できない、という問題があった。

30

【0007】

本発明の一態様は、上記の問題に鑑みてなされたものであり、その目的の一例は、レコード対の類似度を算出する技術として、レコード対に関する訓練データを必要とせず、かつ、異質データにも対応可能な技術を提供することである。

【課題を解決するための手段】

【0008】

本発明の一側面に係る情報処理装置は、レコード対を取得する取得手段と、前記レコード対を変換することによって変換済レコード対を生成する変換手段と、前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出する類似度算出手段と、前記類似度算出手段が算出した類似度を出力する出力手段とを備える。

40

【0009】

本発明の一側面に係る情報処理方法は、少なくとも1つのプロセッサが、レコード対を取得することと、前記レコード対を変換することによって変換済レコード対を生成することと、前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出することと、前記算出した類似度を出力することとを含む。

【0010】

本発明の一側面に係る情報処理プログラムは、コンピュータに、レコード対を取得する

50

取得処理と、前記レコード対を変換することによって変換済レコード対を生成する変換処理と、前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出する類似度算出処理と、前記類似度算出処理において算出した類似度を出力する出力処理とを実行させる。

【発明の効果】

【 0 0 1 1 】

本発明の一態様によれば、レコード対の類似度を算出する技術として、レコード対に関する訓練データを必要とせず、かつ、異質データにも対応可能な技術を提供することができる。

【図面の簡単な説明】

10

【 0 0 1 2 】

【図 1】例示的实施形態 1 に係る情報処理装置の構成を示すブロック図である。

【図 2】例示的实施形態 1 に係る情報処理方法の流れを示すフロー図である。

【図 3】例示的实施形態 2 に係る情報処理装置の構成を示すブロック図である。

【図 4】例示的实施形態 2 に係るレコードが含まれるデータの具体例を示す図である。

【図 5】例示的实施形態 2 に係る情報処理装置が行う処理の流れの概要を示す図である。

【図 6】例示的实施形態 2 に係る同一性の判定結果の具体例を示す図である。

【図 7】例示的实施形態 2 に係る情報処理方法の流れを示すフロー図である。

【図 8】例示的实施形態 2 に係る文書の含意関係を概略的に示す図である。

【図 9】例示的实施形態 2 に係る変換部により変換された画像の一例を示す図である。

20

【図 10】例示的实施形態 3 に係る情報処理装置の構成を示すブロック図である。

【図 11】例示的实施形態 3 に係る情報処理方法の流れを示すフロー図である。

【図 12】例示的实施形態 3 に係る質問応答モデルを用いた類似度の算出処理の概念図である。

【図 13】例示的实施形態 4 に係る情報処理装置の構成を示すブロック図である。

【図 14】例示的实施形態 4 に係る類似度算出部が算出する類似度の具体例を示す図である。

【図 15】例示的实施形態 4 に係る類似度算出部が算出する類似度の具体例を示す図である。

【図 16】例示的实施形態 5 に係る情報処理装置の構成を示すブロック図である。

30

【図 17】例示的实施形態 5 に係る画面表示の具体例を示す図である。

【図 18】グラフデータの一例を示す図である。

【図 19】グラフデータベースの一例を示す図である。

【図 20】学習済み変換器をモデルの出力の前後に設けた構成を概略的に示す図である。

【図 21】各例示的实施形態に係る情報処理装置として機能するコンピュータの構成を示すブロック図である。

【発明を実施するための形態】

【 0 0 1 3 】

〔例示的实施形態 1〕

本発明の第 1 の例示的实施形態について、図面を参照して詳細に説明する。本例示的实施形態は、後述する例示的实施形態の基本となる形態である。

40

【 0 0 1 4 】

< 情報処理装置 1 の構成 >

本例示的实施形態に係る情報処理装置 1 の構成について、図 1 を参照して説明する。図 1 は、情報処理装置 1 の構成を示すブロック図である。情報処理装置 1 は、レコード同士の類似度を算出する装置である。ここで、レコードは、類似度の算出対象であるデータの単位である。レコードが含まれるデータとしては、例えば、テーブルデータ等の構造データ、J S O N (JavaScript Object Notation : 登録商標) 又は X M L (Extensible Markup Language) 等のデータ記述言語で記述された半構造データ、及び自然言語で記された文書を表す非構造データが挙げられる。レコードは一例として、テーブルの行であり、

50

テーブルの列に対応する 1 又は複数の属性名及び属性値のセットを含む。また、レコードはグラフデータであってもよい。

【0015】

情報処理装置 1 は、取得部 11、変換部 12、類似度算出部 13 及び出力部 14 を備える。

【0016】

(取得部 11)

取得部 11 は、レコード対を取得する。レコード対はレコードのセットであり、一例として、第 1 のテーブルに含まれるレコードと、第 2 のテーブルに含まれるレコードとのセットである。第 1 のテーブル及び第 2 のテーブルは、一例として、事業者の顧客情報を保存したテーブル、又は、商品情報を保存したテーブルである。ただし、第 1 のテーブル及び第 2 のテーブルは上述した例に限られず、他のテーブルであってもよい。また、第 1 のテーブルと第 2 のテーブルとは同じであってもよく、また、異なってもよい。

10

【0017】

レコード対に含まれる複数のレコードは、データ形式がそれぞれ異なってもよい。より具体的には例えば、レコードがテーブルの行である場合、レコードに含まれる属性名のうちの一部が異なってもよく、また、レコードに含まれる全ての属性名が異なってもよい。

【0018】

取得部 11 は、記憶装置からレコード対を読み出すことによりレコード対を取得してもよく、また、通信インタフェースを介して接続された他の装置からレコード対を受信することによりレコード対を取得してもよい。また、取得部 11 は、入出力インタフェースを介して入力装置から入力されたレコード対を取得してもよい。

20

【0019】

(変換部 12)

変換部 12 は、上記レコード対を変換することによって変換済レコード対を生成する。変換部 12 は、一例として、レコード対に含まれるレコードを文書、画像、音声又はグラフを表すデータに変換する。より具体的には、変換部 12 は、一例として、レコードを肯定文又は質問文に変換する。ただし、変換部 12 がレコード対を変換する手法は上述した例に限られず、変換部 12 は他の手法によりレコード対を変換してもよい。

30

【0020】

(類似度算出部 13)

類似度算出部 13 は、上記変換済レコード対をモデルに入力することによって、上記変換済レコード対に関する類似度を算出する。ここで、モデルは、類似度を算出するためのモデルであり、一例として、一般に公開されて任意のユーザが利用可能なモデルである。モデルは、機械学習により生成されたモデルであってもよく、また、人間が作成したルールベースのモデルであってもよい。具体的には、モデルは一例として、文書分類モデル、画像分類モデル、音声分類モデル、又はグラフ分類モデルである。文書分類モデルは、文書データを分類するモデルである。画像分類モデルは、画像データを分類するモデルである。音声分類モデルは、音声データを分類するモデルである。グラフ分類モデルは、グラフデータを分類するモデルである。

40

【0021】

文書分類モデルとしては、例えば、文書埋め込みモデル、含意認識モデル、言い換え予測モデル、質問応答モデル、及びマスク言語モデルが挙げられる。文書埋め込みモデルは、文書又は単語をベクトル空間に埋め込むモデルである。含意認識モデルは、複数の文書の含意関係を予測するモデルである。言い換え予測モデルは、2 つの文書が言い換え表現かどうかを予測するモデルである。質問応答モデルは、質問に対し与えられた文書の中から回答を抽出し出力するモデルである。マスク言語モデルは、文書内のマスクに当てはまる単語を予測するためのモデルである。

【0022】

50

画像分類モデルとしては、例えば、画像埋め込みモデルが挙げられる。画像埋め込みモデルは、画像データをベクトル空間に埋め込むモデルである。音声分類モデルとしては、例えば、音声埋め込みモデルが挙げられる。音声埋め込みモデルは、音声データをベクトル空間に埋め込むモデルである。

【0023】

上記モデルの入力は、一例として、テキストデータ、画像データ、音声データ、グラフ、及びベクトルの少なくともいずれかを含む。上記モデルの出力は、一例として、ベクトル、又は確信度を示すスコアを含む。スコアは、一例として、文書の包含関係に関する確信度を示すスコア、又は、言い換え表現であるかの確信度を示すスコアである。ただし、モデルの入力及び出力は上述した例に限られず、他の情報を含んでもよい。

10

【0024】

上記モデルが機械学習により生成される場合、上記モデルとして、例えば、非特許文献3～5に記載の言語モデル、非特許文献6に記載の画像分類モデル、又は音声データを分類する音声分類モデル等が挙げられるが、これらに限られない。また、上記モデルは、情報処理装置1のメモリに記憶されていてもよいし、情報処理装置1と通信可能な他の装置に記憶されていてもよい。

【0025】

類似度は、レコード対に含まれるレコード同士の類似の度合いに関する情報であり、一例として、ベクトル対のコサイン類似度である。また、類似度は、モデルが出力するスコアから算出される値であってもよい。

20

【0026】

(出力部14)

出力部14は、類似度算出部13が算出した類似度を出力する。出力部14は、一例として、類似度を記憶装置に書き込むことにより出力してもよく、また、通信インタフェースを介して他の装置に類似度を送信することにより類似度を出力してもよい。また、出力部14は、入出力インタフェースを介して接続された出力装置(図示略)に類似度を出力してもよい。出力装置は、一例として、ディスプレイ、プリンタ、プロジェクタ又はスピーカである。

【0027】

出力部14が出力する類似度は、例えばテーブルの統合処理、又は情報検索処理に用いられる。テーブルの統合処理の場合、類似度算出部13が算出した類似度に基づき同一であると予測されたレコードを連携することで、複数のテーブルを統合しデータの一元管理を行うことができる。また、情報検索において、検索キーとするレコード(例えば、ユーザにより指定されたレコード)と、所定のテーブルに登録された他の任意のレコードとのレコード対について類似度算出部13が類似度の算出を行ってもよい。この場合、類似度算出部13が算出した類似度に基づき同一であると予測されたレコード対に含まれるレコードを、検索結果として情報処理装置1が出力してもよい。これにより、検索キーであるレコードと連携されていないテーブルにおいても、検索キーによる検索処理が可能となる。

30

【0028】

<情報処理装置1の効果>

40

以上のように、本例示的实施形態に係る情報処理装置1においては、レコード対を取得する取得部11と、上記レコード対を変換することによって変換済レコード対を生成する変換部12と、上記変換済レコード対をモデルに入力することによって、上記変換済レコード対に関する類似度を算出する類似度算出部13と、類似度算出部13が算出した類似度を出力する出力部14とを備える構成が採用されている。このため、本例示的实施形態に係る情報処理装置1によれば、レコード対の類似度を算出する技術として、レコード対に関する訓練データを必要とせず、かつ、異質データにも対応可能な技術を提供できるといふ効果が得られる。

【0029】

<情報処理プログラム>

50

上述の情報処理装置１の機能は、プログラムによって実現することもできる。本例示的实施形態に係る情報処理プログラムは、コンピュータに、レコード対を取得する取得処理と、上記レコード対を変換することによって変換済レコード対を生成する変換処理と、上記変換済レコード対をモデルに入力することによって、上記変換済レコード対に関する類似度を算出する類似度算出処理と、上記類似度算出処理において算出した類似度を出力する出力処理とを実行させる。

【００３０】

< 情報処理方法Ｓ１の流れ >

本例示的实施形態に係る情報処理方法Ｓ１の流れについて、図２を参照して説明する。図２は、情報処理方法Ｓ１の流れを示すフロー図である。情報処理方法Ｓ１における各ステップの実行主体は、情報処理装置１が備えるプロセッサであってもよいし、他の装置が備えるプロセッサであってもよく、各ステップの実行主体がそれぞれ異なる装置に設けられたプロセッサであってもよい。

10

【００３１】

ステップＳ１１では、少なくとも１つのプロセッサが、レコード対を取得する。ステップＳ１２では、少なくとも１つのプロセッサが、上記レコード対を変換することによって変換済レコード対を生成する。ステップＳ１３では、少なくとも１つのプロセッサが、上記変換済レコード対をモデルに入力することによって、上記変換済レコード対に関する類似度を算出する。ステップＳ１４では、少なくとも１つのプロセッサが、上記算出した類似度を出力する。

20

【００３２】

< 情報処理方法Ｓ１の効果 >

以上のように、本例示的实施形態に係る情報処理方法Ｓ１においては、少なくとも１つのプロセッサが、レコード対を取得することと、上記レコード対を変換することによって変換済レコード対を生成することと、上記変換済レコード対をモデルに入力することによって、上記変換済レコード対に関する類似度を算出することと、上記算出した類似度を出力することとを含む構成が採用されている。このため、本例示的实施形態に係る情報処理方法Ｓ１によれば、レコード対の類似度を算出する技術として、レコード対に関する訓練データを必要とせず、かつ、異質データにも対応可能な技術を提供できるという効果が得られる。

30

【００３３】

〔例示的实施形態２〕

本発明の第２の例示的实施形態について、図面を参照して詳細に説明する。なお、例示的实施形態１にて説明した構成要素と同じ機能を有する構成要素については、同じ符号を付し、その説明を繰り返さない。

【００３４】

< 情報処理装置１Ａの概要 >

図３は、本例示的实施形態に係る情報処理装置１Ａの構成を示すブロック図である。情報処理装置１Ａは、レコード同士の同一性を判定する機能を有する。レコードが含まれるデータは、一例として、テーブルデータ等の構造データ、ＪＳＯＮ又はＸＭＬ等のデータ記述言語で記述された半構造データ、又は、自然言語で記された文書を表す非構造データである。

40

【００３５】

図４は、レコードが含まれるデータの具体例を示す図である。図４において、データＤ１はテーブルである。この場合、レコードはテーブルの各行である。また、図４において、データＤ２は、マークアップ言語等のデータ記述言語で記述された半構造データである。この場合、レコードは一例としてウェブページである。データＤ３は、自然言語で記された文書を表す非構造データである。この場合、レコードは一例として、所定のファイル形式で生成されたファイルである。

【００３６】

50

ここで、本例示的实施形態に係る情報処理装置 1 A が行う処理の概要について、図 5 を参照しつつ説明する。図 5 は、情報処理装置 1 A が行う処理の流れの概要を示す図である。情報処理装置 1 A は、大別して (i) レコード対の生成処理、(i i) 類似度の算出処理、及び (i i i) 同一性の判定処理、を行う。

【 0 0 3 7 】

(i) レコード対の生成処理において、情報処理装置 1 A は、複数のレコード e を含む第 1 データ x と、複数のレコード e ' を含む第 2 データ x ' とからレコード対を生成する。情報処理装置 1 A は一例として、第 1 データ x に含まれるレコード e と第 2 データ x ' に含まれるレコード e ' との全ての組み合わせを生成する。また、情報処理装置 1 A は、レコード対の生成において、ブロッキングと呼ばれる技術により第 1 データ x のレコード e についての第 2 データ x ' の同一性判定の候補をしぼってもよい。

10

【 0 0 3 8 】

(i i) 類似度の算出処理において、情報処理装置 1 A は、レコード対に含まれるレコード同士の類似度を算出する。本例示的实施形態において、情報処理装置 1 A は、レコードを変換した変換済レコード対をモデルに入力することにより類似度を算出する。類似度の算出処理の詳細については後述する。

【 0 0 3 9 】

(i i i) 同一性の判定処理において、情報処理装置 1 A は、算出した類似度に基づき、レコード対に含まれるレコード同士の同一性を判定する。一例として、情報処理装置 1 A は、類似度が閾値以上である場合にレコード同士が同一であると判定する。ただし、同一性の判定の手法は上述した手法に限定されず、情報処理装置 1 A は他の手法によりレコード同士の同一性を判定してもよい。

20

【 0 0 4 0 】

図 6 は、同一性の判定結果の具体例を示す図である。図 6 において、テーブル T B L 1 は第 1 データ x の一例であり、複数の行及び複数の列を含む。また、テーブル T B L 2 は第 2 データ x ' の一例であり、複数の行及び複数の列を含む。テーブル T B L 1 及びテーブル T B L 2 において、レコードはテーブルの行である。テーブル T B L 1 はレコード l 1、l 2、l 3 及び l 4 を含み、テーブル T B L 2 はレコード r 1、r 2、r 3 を含む。

【 0 0 4 1 】

図 6 の例では、情報処理装置 1 A は、上記 (i) ~ (i i i) の処理により、レコード l 1 とレコード r 2 とが同一であると判定し、レコード l 2 とレコード r 3 とが同一であると判定し、レコード l 3 とレコード r 1 とが同一であると判定する。

30

【 0 0 4 2 】

< 情報処理装置 1 A の構成 >

情報処理装置 1 A は、図 3 に示すように、制御部 1 0 A、記憶部 2 0 A、通信部 3 0 A 及び入出力部 4 0 A を備える。

【 0 0 4 3 】

(通信部 3 0 A)

通信部 3 0 A は、情報処理装置 1 A の外部の装置と通信回線を介して通信する。通信回線の具体的構成は本例示的实施形態を限定するものではないが、通信回線は一例として、無線 L A N (Local Area Network)、有線 L A N、W A N (Wide Area Network)、公衆回線網、モバイルデータ通信網、又は、これらの組み合わせである。通信部 3 0 A は、制御部 1 0 A から供給されたデータを他の装置に送信したり、他の装置から受信したデータを制御部 1 0 A に供給したりする。

40

【 0 0 4 4 】

(入出力部 4 0 A)

入出力部 4 0 A には、キーボード、マウス、ディスプレイ、プリンタ、タッチパネル等の入出力機器が接続される。入出力部 4 0 A は、接続された入力機器から情報処理装置 1 A に対する各種の情報の入力を受け付ける。また、入出力部 4 0 A は、制御部 1 0 A の制御の下、接続された出力機器に各種の情報を出力する。入出力部 4 0 A としては、例えば

50

ＵＳＢ（Universal Serial Bus）などのインタフェースが挙げられる。

【００４５】

（制御部１０Ａ）

制御部１０Ａは、図３に示すように、取得部１１、変換部１２、類似度算出部１３、出力部１４、同一性判定部１５Ａ及び統合部１６Ａを備える。

【００４６】

（取得部１１）

本例示の実施形態において、取得部１１は、レコード e を含む第１データ x とレコード e' を含む第２データ x' とから、レコード e とレコード e' を含むレコード対を生成する。ただし、取得部１１がレコード対を生成する処理を行わなくてもよい。取得部１１は、一例として、記憶部２０Ａ又は他の外部記憶装置からレコード対を読み出すことにより取得してもよく、また、通信部３０Ａを介して他の装置から受信されるレコード対を取得してもよい。また、取得部１１は、入出力部４０Ａに接続された入力装置から入力されるレコード対を取得してもよい。

10

【００４７】

（変換部１２）

変換部１２は、レコード対を変換することによって変換済レコード対を生成する。変換部１２は一例として、レコード対に含まれるレコードを文書データ、画像データ、音声データ又はグラフに変換する。変換部１２が実行する変換処理については後述する。

【００４８】

（類似度算出部１３）

類似度算出部１３は、上記変換済レコード対をモデルＭＡに入力することによって、上記変換済レコード対に関する類似度 s を算出する。類似度 s は、レコード対に含まれるレコード同士の類似の度合いに関する情報であり、一例として、ベクトル対のコサイン類似度、又は、モデルＭＡが出力するスコアに基づき計算される値である。類似度算出部１３が類似度 s を算出する処理の詳細については後述する。

20

【００４９】

（出力部１４）

出力部１４は、類似度算出部１３が算出した類似度 s を出力する。出力部１４は、一例として、類似度を記憶部２０Ａに書き込むことにより出力する。ただし、出力部１４が類似度を出力する手法は上述した例に限定されず、他の手法により類似度 s を出力してもよい。一例として、出力部１４は、通信部３０Ａを介して接続された他の装置に類似度を送信してもよく、また、入出力部４０Ａを介して接続された出力装置に類似度を出力してもよい。

30

【００５０】

（同一性判定部１５Ａ）

同一性判定部１５Ａは、類似度 s に基づきレコード対に含まれるレコード同士の同一性を判定する。一例として、同一性判定部１５Ａは、類似度 s が閾値以上である場合にレコード同士が同一であると判定する。また、同一性判定部１５Ａは、類似度の上位 x 個のレコード対を同一と判定する、といったように、レコード対を類似度の高い順位ソートした場合の順位に基づき同一性を判定してもよい。また、同一性判定部１５Ａは、一例として、安定結婚問題アルゴリズム等のマッチングアルゴリズムを適用して同一性を判定してもよい。

40

【００５１】

また、同一性判定部１５Ａが同一性を判定する手法は上述した例に限られず、同一性判定部１５Ａは他の手法により同一性の判定を行ってもよい。一例として、同一性判定部１５Ａは、機械学習により生成された予測モデルにレコード対と類似度とを入力することによりレコード対の同一性予測を行ってもよい。この場合、予測モデルの入力は、一例として、レコード対と類似度とを含む。また、予測モデルの出力は、一例として、同一性の予測結果を含む。この場合、予測モデルの機械学習の手法は限定されず、一例として、決定

50

木ベース、線形回帰、又はニューラルネットワークの手法が用いられてもよく、また、これらのうちの2以上の手法が用いられてもよい。

【0052】

(統合部16A)

統合部16Aは、同一性判定部15Aの判定結果に基づき、第1データ x と第2データ x' とを統合する。統合部16Aは、一例として、レコード数を増やす、及び/又はデータの属性数を増やすことにより第1データ x と第2データ x' とを統合する。統合部16Aが実行するデータ統合としては、例えば、(i)エンティティ統合、(ii)データクレンジング、(iii)スキーママッチング、が挙げられる。(i)エンティティ統合は、同一のレコード集合が与えられたときに異なる属性とその値の表記を統一することをいう。(ii)データクレンジングは、社名、住所、局番等の記載形式の違い(「(株)」と「株式会社」、等)を統一することをいう。(iii)スキーママッチングは、表記が異なる複数の属性のアライメント(マッチング)をとることをいう。

10

【0053】

(記憶部20A)

記憶部20Aには、第1データ x 及び第2データ x' が記憶されるとともに、類似度算出部13が算出する類似度 s が記憶される。また、記憶部20Aには、モデルMAが記憶される。なお、モデルMAが記憶部20Aに記憶されているとは、モデルMAを規定するパラメータが記憶部20Aに記憶されていることをいう。

20

【0054】

(モデルMA)

モデルMAは、類似度を算出するためのモデルであり、一例として、一般に公開されて任意のユーザが利用可能なモデルである。モデルMAは、機械学習により生成されたモデルであってもよく、また、人間が作成したルールベースのモデルであってもよい。

【0055】

モデルMAは一例として、文書分類モデル、画像分類モデル、音声分類モデル及びグラフ分類モデルの少なくともいずれかを含む。文書分類モデルとしては例えば、文書埋め込みモデル、含意認識モデル、言い換え予測モデル、及びマスク言語モデルが挙げられる。画像分類モデルとしては例えば、画像データをベクトル空間に埋め込む画像埋め込みモデルが挙げられる。音声分類モデルとしては例えば、音声データをベクトル空間に埋め込む音声埋め込みモデルが挙げられる。

30

【0056】

モデルMAの入力は、一例として、文書データ、画像データ、音声データ、グラフ、及びベクトルの少なくともいずれかを含む。モデルMAの出力は、一例として、ベクトル及びスコアの少なくともいずれかを含む。

【0057】

(文書分類モデルの例1：文書埋め込みモデル)

文書埋め込みモデルは、文書又は単語をベクトル空間に埋め込むモデルである。文書埋め込みモデルは、一例として非特許文献3に記載されたRoBERTaにより生成される。文書埋め込みモデルの入力は、一例として文書又は単語(例えば、「老人男性が公園で歩いている。」という文)である。出力は、一例としてベクトルである。

40

【0058】

(文書分類モデルの例2：含意認識モデル)

含意認識モデルは「文書1であれば文書2である」という含意関係かどうかを予測するモデルである。含意認識モデルは、一例として非特許文献4に記載された技術により生成される。含意認識モデルの入力は、一例として2つの文書である。2つの文書は、例えば、「老人男性が公園で歩いている」という文書1、及び「男性が公園にいる」という文書2である。この場合、文書2は文書1を含意する。図8は、文書の含意関係を概略的に示す図である。図8の例では、文書2が文書1を含意する。

【0059】

50

また、含意認識モデルの出力は、一例として含意性スコアである。含意性スコアは、含意関係の確信度を示す数値であり、例えば0～1の実数である。含意性スコアは、一例として、値が大きいほど、含意関係の確信度が高いことを示す。

【0060】

（文書分類モデルの例3：言い換え予測モデル）

言い換え予測モデルは、2つの文書が言い換え表現かどうかを予測するモデルである。言い換え予測モデルは、一例として非特許文献3に記載されたR o B E R T aにより生成される。言い換え予測モデルの入力は、一例として2つの文書である。2つの文書は例えば「NECはITの会社である。」という文書1、及び「日本電気株式会社はIT事業を行っている。」という文書2である。モデルの出力は、一例として、言い換えスコアを含む。言い換えスコアは、2つの文書が言い換え表現であることの確信度を示すスコアであり、一例として、0～1の実数である。言い換えスコアは、一例として、値が大きいほど2つの文書が言い換え表現であることの確信度が高いことを示す。

10

【0061】

（文書分類モデルの例4：マスク言語モデル）

マスク言語モデルは、文書内のマスクに当てはまる単語を予測するモデルである。マスク言語モデルは、一例として例えば非特許文献3に記載されたR o B E R T aにより生成されたモデルである。文書分類モデルの入力は、一例として、文書（例えば、「このピザはとても美味しい。私はこのピザが[mask]。」という文）である。モデルの出力は、単語（例えば、「好き」）とスコアとを含む。スコアは、その単語がマスクに当てはまる確信度を示す値であり、一例として、0～1の実数である。

20

【0062】

（画像分類モデル）

画像分類モデルは、画像を分類するモデルである。画像分類モデルは、一例として非特許文献6に記載された技術により生成されたモデルである。画像分類モデルの入力は、一例として、画像（例えば、犬の画像）である。また、モデルの中間出力は、一例として、画像のベクトル表現であり、モデルの出力は、一例として、ラベル（例えば、「犬」を示すラベル）とスコアである。スコアはラベルの確信度を示す値であり、例えば0～1の実数である。本例示的实施形態では、類似度算出部13は、画像のベクトル表現を利用して類似度sを算出する。

30

【0063】

（音声分類モデル）

音声分類モデルは、音声を分類するモデルである。音声分類モデルの入力は、一例として音声データ（例えば犬の鳴き声）である。モデルの中間出力は、一例として音声のベクトル表現であり、モデルの出力は、一例としてラベル（例えば、音声が犬の鳴き声であることを示すラベル）とスコアである。スコアはラベルの確信度を示す値であり、例えば0～1の実数である。本例示的实施形態では、類似度算出部13は、中間出力である音声のベクトル表現を利用して類似度sを算出する。

【0064】

（グラフ分類モデル）

グラフ分類モデルは、グラフを分類するモデルである。グラフ分類モデルの入力は、一例としてグラフデータ（例えば顔の特徴を表すグラフ）である。モデルの中間出力は、一例としてグラフのベクトル表現であり、モデルの出力は、一例としてラベル（例えば人物を示すラベル）とスコアである。スコアはラベルの確信度を示す値であり、例えば0～1の実数である。本例示的实施形態では、類似度算出部13は、中間出力であるグラフのベクトル表現を利用して類似度sを算出する。

40

【0065】

<情報処理方法S100Aの流れ>

図7は、情報処理装置1Aが実行する情報処理方法S100Aの流れを示すフロー図である。なお、情報処理方法S100Aに含まれるステップのうち、一部のステップは並行

50

して又は順序を換えて実行されてもよい。また、既に説明した内容についてはその説明を繰り返さない。

【0066】

(ステップS101)

ステップS101において、取得部11は、モデルMAを読み込む。本例示的实施形態において、モデルMAは、複数のモデル候補から選択されたものである。複数のモデル候補は、一例として、文書埋め込みモデル、含意認識モデル等の文書分類モデル、画像分類モデル、及び音声分類モデルの少なくともいずれかを含む。モデルMAの選択は、一例としてユーザ操作に基づき行われてもよく、また、所定のアルゴリズムに従って行われてもよい。モデルMAは、ひとつのモデルであってもよく、また、複数のモデルの集合であつてもよい。

10

【0067】

(ステップS102)

ステップS102において、取得部11は、レコード対を読み込む。レコード対に含まれるレコードeとレコードe'とは、一例として、

$$e = ((a_j, v_j))_{j=1, \dots, d},$$

$$e' = ((a'_j, v'_j))_{j=1, \dots, d'}$$

と表される。ここで、属性名a_j A_jであり、A_jは、一例として文字列空間である。属性値v_j V_jであり、V_jは、一例として文字列空間又は実数空間である。この例で、図6のレコードl1は、

20

l1 = ((title, sims 2 glamour life stuff pack), (manufacturer, aspyr media), (price, 24.99))であり、レコードr2は、

r2 = ((title, aspyr media inc sims 2 glamour life stuff pack), (price, NaN))である。

【0068】

本例示的实施形態において、取得部11は、第1データx及び第2データx'からレコード対を生成する。取得部11は、一例として、第1データxに含まれるレコードeと第2データx'に含まれるレコードe'の全ての組み合わせを生成する。また、取得部11は、レコード対の生成において、ブロッキングと呼ばれる技術により第1データxのレコードeについての第2データx'の同一性判定の候補をしぼってもよい。

30

【0069】

(ステップS103)

ステップS103において、変換部12は、レコード対(e, e')をモデルMAの入力に対応する形式に変換することによって変換済レコード対を生成する。一例として、モデルMAには、文書分類モデルが含まれており、変換部12による変換処理には、レコード対(e, e')を文書に変換することによって変換済レコード対を生成する処理が含まれる。また、一例として、モデルMAには、画像分類モデルが含まれており、変換部12による変換処理には、レコード対(e, e')を画像に変換することによって変換済レコード対を生成する処理が含まれる。また、一例として、モデルMAには、音声分類モデルが含まれており、変換部12による変換処理には、レコード対を音声に変換することによって変換済レコード対を生成する処理が含まれる。また、一例として、モデルMAには、グラフ分類モデルが含まれており、変換部12による変換処理には、レコード対をグラフに変換することによって変換済レコード対を生成する処理が含まれる。

40

【0070】

(ステップS104)

ステップS104において、類似度算出部13は、変換済レコード対をモデルMAに入力することによって、変換済レコード対に関する類似度sを算出する。

【0071】

(ステップS101～S104の処理例1～5)

ここで、ステップS101～S104の処理例として、処理例1～5を説明する。処理

50

例 1 は、文書埋め込みモデルを用いる場合の処理例である。処理例 2 は、画像分類モデルを用いる場合の処理例である。処理例 3 は、音声分類モデルを用いる場合の処理例である。処理例 4 は、含意認識モデルを用いる場合の処理例である。処理例 5 は、言い換え予測モデルを用いる場合の処理例である。

【 0 0 7 2 】

(処理例 1 : 文書埋め込みモデル)

この例では、ステップ S 1 0 1 において、取得部 1 1 は、文書埋め込みモデルを読み込む。また、ステップ S 1 0 3 において、変換部 1 2 は、レコード対 (e , e ´) を文書対 (t , t ´) に変換する。変換部 1 2 は、一例として、

e = ((title, sims 2 glamour life stuff pack) , (manufacturer, aspyr media) , (price, 24.99))

10

e ´ = ((title, aspyr media inc sims 2 glamour life stuff pack) , (price, NaN))

であるレコード e とレコード e ´ とを含むレコード対 (e , e ´) を、

t = " Title is sims 2 glamour life stuff pack. Manufacturer is aspyr media. Price is 24.99. "

t ´ = " Title is aspyr media inc sims 2 glamour life stuff pack. "

である文書 t 及び文書 t ´ を含む文書対 (t , t ´) に変換する。

【 0 0 7 3 】

また、ステップ S 1 0 4 において、類似度算出部 1 3 は、文書埋め込みモデルにより文書対 (t , t ´) をベクトル対 (v , v ´) に変換する。ここで、 $v = M(t)$ 、 $v ´ = M(t ´)$ である。また、類似度算出部 1 3 は、ベクトル対 (v , v ´) から類似度 s を

20

算出する。類似度 s は、一例として、 $s = \exp(-||v - v ´|| / c)$ であり、ここで

$c > 0$ である。また、類似度 s は、コサイン類似度 $s = v^T v ´ / (||v|| ||v ´||)$ であってもよい。ここで、 T は転置を表す記号である。

【 0 0 7 4 】

(処理例 2 : 画像埋め込みモデル)

この例では、ステップ S 1 0 1 において、取得部 1 1 は、画像埋め込みモデルを読み込む。ステップ S 1 0 3 において、変換部 1 2 は、レコード対 (e , e ´) を画像対 (i , i ´) に変換する。

【 0 0 7 5 】

30

図 9 は、変換部 1 2 により変換された画像の一例を示す図である。変換部 1 2 は、一例として、

e = ((title, sims 2 glamour life stuff pack) , (manufacturer, aspyr media) , (price, 24.99))

e ´ = ((title, aspyr media inc sims 2 glamour life stuff pack) , (price, NaN))

であるレコード e とレコード e ´ とを含むレコード対 (e , e ´) を、図 9 に示す画像 i 、 i ´ に変換する。

【 0 0 7 6 】

また、ステップ S 1 0 4 において、類似度算出部 1 3 は、画像埋め込みモデルにより画像対 (i , i ´) をベクトル対 (v , v ´) に変換する。ここで、 $v = M(i)$ 、 $v ´ = M(i ´)$ である。更に、類似度算出部 1 3 は、ベクトル対 (v , v ´) から類似度 s を

40

算出する。類似度 s は、一例として、 $s = \exp(-||v - v ´|| / c)$ であり、ここで、 $c > 0$ である。

【 0 0 7 7 】

画像埋め込みモデルを用いる場合において、変換部 1 2 は、1つのレコードを1つの画像に変換してもよく、また、レコードに含まれる要素 (例えば、単語) ごとに画像変換を行ってもよい。変換部 1 2 が要素ごとに画像変換する場合、類似度算出部 1 3 は、要素ごとの画像の集合を用いて類似度 s を算出する。また、要素ごとに画像変換を行う場合において、変換部 1 2 は、レコードの欠損値については画像変換しないようにしてもよい。

【 0 0 7 8 】

50

(処理例3：音声埋め込みモデル)

この例では、ステップS101において、取得部11は、音声埋め込みモデルを読み込む。また、ステップS103において、変換部12は、レコード対 (e, e') を音声対 (i, i') に変換する。変換部12は、一例として、

$e = ((\text{title, sims 2 glamour life stuff pack}), (\text{manufacturer, aspyr media}), (\text{price, 24.99}))$

$e' = ((\text{title, aspyr media inc sims 2 glamour life stuff pack}), (\text{price, NaN}))$

であるレコード e とレコード e' とを含むレコード対 (e, e') を、レコード e を読み上げた音声を表す音声データ i と、レコード e' を読み上げた音声を表す音声データ i' とに変換する。

10

【0079】

また、ステップS104において、類似度算出部13は、音声埋め込みモデルにより音声データ対 (i, i') をベクトル対 (v, v') に変換する。ここで、 $v = M(i)$ 、 $v' = M(i')$ である。更に、類似度算出部13は、ベクトル対 (v, v') から類似度 s を算出する。類似度 s は、一例として、 $s = \exp(-||v - v'||/c)$ であり、ここで、 $c > 0$ である。

【0080】

(処理例4：含意認識モデル)

この例では、ステップS101において、取得部11は、含意認識モデルを読み込む。また、ステップS103において、変換部12は、レコード対 (e, e') を文書対 (t, t') に変換する。

20

【0081】

変換部12は、一例として、

レコード $e = ((a_j, v_j))_{j=1, \dots, d}$ を、

文書 $t = \text{「} a_1 \text{ is } v_1 \dots a_d \text{ is } v_d \text{」}$ に変換する。ただし、変換部12は欠損値を文書 t に含めない。

【0082】

変換部12は、一例として、

$e = ((\text{title, sims 2 glamour life stuff pack}), (\text{manufacturer, aspyr media}), (\text{price, 24.99}))$

30

$e' = ((\text{title, aspyr media inc sims 2 glamour life stuff pack}), (\text{price, NaN}))$

であるレコード e とレコード e' とを含むレコード対 (e, e') を、

$t = \text{「Title is sims 2 glamour life stuff pack. Manufacturer is aspyr media. Price is 24.99.」}$

$t' = \text{「Title is aspyr media inc sims 2 glamour life stuff pack.」}$

である文書 t と文書 t' とを含む文書対 (t, t') に変換する。

【0083】

また、ステップS104において、類似度算出部13は、含意認識モデルを用いて文書対 (t, t') の含意性スコアを計算する。更に、類似度算出部13は、含意性スコアを用いて類似度 s を算出する。類似度 s は、一例として、 $s = M(t, t') \times M(t', t)$ である。換言すると、類似度 s は、「文書 t であれば文書 t' である」という含意関係の含意性スコア $M(t, t')$ と、「文書 t' であれば文書 t である」という含意関係の含意性スコア $M(t', t)$ との乗算値である。ただし、類似度 s は上述した例に限られず、他の値であってもよい。類似度 s は、例えば、含意性スコア $M(t, t')$ と含意性スコア $M(t', t)$ のうちの最大値、又は、含意性スコア $M(t, t')$ と含意性スコア $M(t', t)$ との和であってもよい。

40

【0084】

レコード対 (e, e') に含まれるレコード e とレコード e' とが同一であれば、 $e = e'$ かつ $e' = e$ である。この処理例では、類似度算出部13はこの関係を利用して類似性の算出を行う。

50

【0085】

(処理例5：言い換え予測モデル)

この例では、ステップS101において、取得部11は言い換え予測モデルを読み込む。また、ステップS103において、変換部12は、レコード対(e, e')を文書対(t, t')に変換する。

【0086】

変換部12は、一例として、

レコード $e = ((a_j, v_j))_{j=1, \dots, d}$ を、

文書 $t = \text{「} v_1 \dots v_d \text{」}$ に変換する。ただし、変換部12は欠損値を文書に含めない。

【0087】

変換部12は、一例として、

$e = ((\text{title, sims 2 glamour life stuff pack}), (\text{manufacturer, aspyr media}), (\text{price, 24.99}))$

$e' = ((\text{title, aspyr media inc sims 2 glamour life stuff pack}), (\text{price, NaN}))$

であるレコードeとレコードe'を含むレコード対(e, e')を、

$t = \text{「sims 2 glamour life stuff pack aspyr media 24.99」}$

$t' = \text{「aspyr media inc sims 2 glamour life stuff pack」}$

である文書tと文書t'を含む文書対(t, t')に変換する。

【0088】

また、ステップS104において、類似度算出部13は、言い換え予測モデルを用いて文書対(t, t')の言い換えスコアを計算し、計算した言い換えスコアをレコード対の類似度sとする。すなわち、この処理例では、類似度算出部13は、レコード対を言い換え表現かを問う形式に落とし込むことにより、類似度sを算出する。

【0089】

(ステップS105・S106)

ステップS105において、出力部14は、類似度算出部13が算出した類似度sを出力する。出力部14は、一例として、類似度sを記憶部20Aに書き込むことにより出力する。ステップS106において、同一性判定部15Aは、類似度sに基づきレコード対に含まれるレコード同士の同一性を判定する。

【0090】

(ステップS107)

ステップS106において、統合部16Aは、同一性判定部15Aの判定結果を参照して、第1データxと第2データx'とから統合済データを生成する。統合済データは、一例として、同一性判定部15Aが同一であると判定したレコード対に含まれるレコード同士を統合したレコードを含む。

【0091】

<情報処理装置1Aの効果>

ところで、近年、自然言語処理及び画像処理分野の様々なタスク(テキスト分類、質問応答、画像分類など)に対して、巨大データセットで訓練された高精度なモデル(推論モデル)が公開され、応用されている。例えば、教師あり機械学習の名寄せモデルであるDITTO(非特許文献1を参照。)は、BERT(Bidirectional Encoder Representations from Transformers)の事前学習済み言語モデルを部分的に応用している。しかしながら、タスクによりデータの形式は様々であるため、既存の推論モデルだけで名寄せを行うことはできなかった。特に、想定されていない属性を含むレコードについて既存の推論モデルで名寄せを行うことはできなかった。

【0092】

それに対し本例示の実施形態によれば、情報処理装置1Aはレコード対をモデルMAの入力に対応する形式に変換してモデルMAに入力することによってレコード対に関する類似度sを算出する。レコードを変換して名寄せタスクを自然言語処理分野又は画像処理分

10

20

30

40

50

野等のタスクに落とし込むことにより、大量のデータで学習された推論モデルの活用を可能にする。すなわち、本例示的实施形態によれば、モデルM Aの訓練データを必要とすることなく、かつ、様々な属性を有するレコードについて同一性を判定することができる。

【0093】

また、本例示的实施形態に係る情報処理装置1 Aにおいては、モデルM Aは、複数のモデル候補から選択されたものであり、変換部1 2は、レコード対(e, e')をモデルM Aの入力に対応する形式に変換することによって変換済レコード対を生成する構成が採用されている。変換部1 2がレコード対を変換することにより、レコード対はモデルM Aに入力可能な形式のデータとなる。すなわち、類似度の算出対象であるレコードがどのような属性を含む場合であっても、類似度算出部1 3はモデルM Aに変換済レコード対を入力することにより類似度sを算出することができる。これにより、本例示的实施形態に係る情報処理装置1 Aによれば、モデルM Aの訓練データを必要とすることなく、モデルM Aを用いて様々な属性を有するレコードについて類似度を算出することができるという効果が得られる。

10

【0094】

また、本例示的实施形態に係る情報処理装置1 Aにおいては、モデルM Aには、文書分類モデルが含まれており、変換部1 2による変換処理には、レコード対を文書に変換することによって変換済レコード対を生成する処理が含まれるという構成が採用されている。このため、本例示的实施形態に係る情報処理装置1 Aによれば、文書分類モデルであるモデルM Aを訓練することなく、モデルM Aを用いて様々な属性を有するレコードの類似度を算出できるという効果が得られる。

20

【0095】

また、本例示的实施形態に係る情報処理装置1 Aにおいては、モデルM Aには、画像分類モデルが含まれており、変換部1 2による変換処理には、レコード対を画像に変換することによって変換済レコード対を生成する処理が含まれるという構成が採用されている。情報処理装置1 Aがレコード対を画像に変換することにより、文字としては異なるが表記が似ているレコード同士の類似度をより好適に算出できる。例えば、「glamour」という単語を含むレコードと「glqmour」という単語を含むレコードの場合、レコードに含まれる文字列は異なっているものの、文字の形状が類似しているため、高い類似度が算出される。このように、本例示的实施形態に係る情報処理装置1 Aによれば、例示的实施形態1に係る情報処理装置1の奏する効果に加えて、文字の形状の類似の度合いを反映した類似度を算出できるという効果が得られる。

30

【0096】

また、本例示的实施形態に係る情報処理装置1 Aにおいては、モデルM Aには、音声分類モデルが含まれており、変換部1 2による変換処理には、レコード対を音声に変換することによって変換済レコード対を生成する処理が含まれるという構成が採用されている。情報処理装置1 Aがレコード対を音声に変換することにより、文字としては異なるが音素が似ているレコード同士の類似度をより好適に算出できる。例えば、「glamour」という単語を含むレコードと「glamar」という単語を含むレコードの場合、レコードに含まれる文字列は異なっているものの、単語の発音が類似しているため、高い類似度が算出される。このように、本例示的实施形態に係る情報処理装置1 Aによれば、例示的实施形態1に係る情報処理装置1の奏する効果に加えて、音の類似度の度合いを反映した類似度を算出できるという効果が得られる。

40

【0097】

また、本例示的实施形態に係る情報処理装置1 Aにおいては、モデルM Aには、グラフ分類モデルが含まれており、変換部1 2による変換処理には、レコード対をグラフに変換することによって変換済レコード対を生成する処理が含まれるという構成が採用されている。情報処理装置1 Aがレコード対をグラフに変換することにより、グラフ分類モデルであるモデルM Aを訓練することなく、モデルM Aを用いて様々な属性を有するレコードの類似度を算出できるという効果が得られる。

50

【 0 0 9 8 】

〔 例示的实施形態 3 〕

本発明の第 3 の例示的实施形態について、図面を参照して詳細に説明する。なお、例示的实施形態 1 ~ 2 にて説明した構成要素と同じ機能を有する構成要素については、同じ符号を付記し、その説明を繰り返さない。

【 0 0 9 9 】

< 情報処理装置 1 B の構成 >

図 1 0 は、本例示的实施形態に係る情報処理装置 1 B の構成を示すブロック図である。情報処理装置 1 B の制御部 1 0 B は、取得部 1 1 B、変換部 1 2 B、類似度算出部 1 3 B、出力部 1 4、同一性判定部 1 5 A、及び統合部 1 6 A を備える。また、記憶部 2 0 B は、第 1 データ x 、第 2 データ x' 及び類似度 s に加えてモデル $M B$ を記憶する。

10

【 0 1 0 0 】

取得部 1 1 B は、レコード対 (e, e') に加えて、補助レコードを更に取得する。補助レコードは、レコード対 (e, e') の類似度を計算するために用いられる補助的なレコードである。補助レコードは、一例として、第 1 データ x に含まれるレコードであってレコード対 (e, e') に含まれるレコード e 以外のレコードである。また、補助レコードは、一例として、第 2 データ x' に含まれるレコードであってレコード対 (e, e') に含まれるレコード e' 以外のレコードである。

【 0 1 0 1 】

変換部 1 2 B は、取得部 1 1 B が取得したレコード対を変換して変換済レコード対を生成する。また、変換部 1 2 B は、補助レコードを変換することによって変換済補助レコードを生成する。変換部 1 2 B は、一例として、補助レコードを文書、画像、音声、又はグラフを表すデータに変換する。

20

【 0 1 0 2 】

より具体的には、変換部 1 2 B は、一例として、レコード対 (e, e') に含まれる一方のレコード e を質問文に変換し、レコード対 (e, e') に含まれる他方のレコード e' 及び補助レコードの各々を応答文に変換することによって上記変換済レコード対を生成する。

【 0 1 0 3 】

類似度算出部 1 3 B は、変換済レコード対及び変換済補助レコードをモデル $M B$ に入力することによって、変換済レコード対に関する類似度を算出する。モデル $M B$ は、一例として、質問文と応答文とを入力とする質問応答モデルを含む。

30

【 0 1 0 4 】

(質問応答モデル)

質問応答モデルは、質問文に対し与えられた文書の中から回答文を抽出して出力するモデルである。質問応答モデルは、一例として非特許文献 5 に記載された T A N D A と呼ばれる技術により生成されたモデルである。質問応答モデルの入力は、一例として、質問文と、文書とを含む。質問文は、一例として「NEC の本社はどこですか?」という文である。文書は、一例として「日本電気株式会社 (にっぽんでんき、英: NEC Corporation) は、東京都港区芝五丁目に本社を置く住友グループの電機メーカー。日経平均株価の構成銘柄の一つ。」という文書である。

40

【 0 1 0 5 】

モデルの出力は、一例として、回答文及びスコアを含む。回答文は、一例として「東京都港区芝五丁目」という文である。スコアは一例として、0 ~ 1 の実数である。ここで、モデルの出力を求める際に、各単語のスコアが計算されてもよい。例えば、質問応答モデルにより、日本電気株式会社」のスコア「0.1」、「住友グループ」のスコア「0.02」、及び「日経平均株価」のスコア「0.08」が計算される。

【 0 1 0 6 】

< 情報処理方法 S 1 0 0 B の流れ >

図 1 1 は、情報処理装置 1 B が実行する情報処理方法 S 1 0 0 B の流れを示すフロー図

50

である。なお、一部のステップは並行して、又は順序を換えて実行されてもよい。また、既に説明した内容についてはその説明を繰り返さない。

【0107】

情報処理方法S100Bは、ステップS101B、ステップS102、ステップS102B、ステップS103B、ステップS104B、ステップS105、ステップS106、及びステップS107を含む。

【0108】

ステップS101Bにおいて、取得部11BはモデルMBを読み込む。ステップS102Bにおいて、取得部11Bは補助レコードを読み込む。

【0109】

ステップS103Bにおいて、変換部12Bは、レコード対を変換することによって変換済レコード対を生成するとともに、補助レコードを変換することによって変換済補助レコードを生成する。ステップS104Bにおいて、類似度算出部13Bは、変換済レコード対及び変換済補助レコードをモデルMBに入力することによって、変換済レコード対に関する類似度を算出する。

【0110】

(ステップS101～S104Bの処理例6：質問応答モデル)

ここで、ステップS101からS104Bの処理例として、質問応答モデルを用いる場合の処理例を説明する。この例では、ステップS101において、類似度算出部13Bは、質問応答モデルを読み込む。また、ステップS102及びS102Bにおいて、取得部11Bは、レコード対 (e, e') 及び補助レコード $R = \{e_{_1}, \dots, e_{_k}\}$ (k は自然数)を読み込む。補助レコード R は、一例として、第2データ x' に含まれる全てのレコードの集合である。ただし、補助レコード R は上述した例に限られず、他のレコードの集合であってもよい。例えば、補助レコード R は、第2データ x' から乱択アルゴリズムにより選択されたレコードの集合であってもよい。また、補助レコード R は、第2データ x' からレコード e と共通する単語を含むレコードを抽出したレコード集合等、ブロッキングされたレコード集合であってもよい。

【0111】

この処理例において、補助レコード R は、レコード対 (e, e') に含まれるレコード e' を含む。補助レコードは、一例として、図6のテーブルTBL2のレコード $r_1 \sim r_3$ の集合、すなわち、

$R = \{((\text{title}, \text{adobe photoshop elements 4.0 photo-editing software for mac}), (\text{price}, 85.95)), ((\text{title}, \text{aspyr media inc sims 2 glamour life stuff pack}), (\text{price}, \text{NaN})), ((\text{title}, \text{final-draft final draft av 2.5 screenwriting software mac/win screen writing software}), (\text{price}, 199.95))\}$

である。

【0112】

この処理例では、ステップS103Bにおいて、変換部12Bは、レコード e と補助レコード $R(e', R)$ とを変換する。より具体的には、変換部12Bは、レコード e を質問文 $q = T_1(e)$ に変換する。ここで、質問文 q は、いわゆる5W1Hのオープンクエスチョン型が好ましい。また、変換部12Bは、補助レコード R を複数の回答文を含む文書に変換する。

【0113】

変換部12Bは、一例として、レコード $e = ((a_1, v_1), \dots, (a_d, v_d))$ を、 $T_1(e) = \text{"What is characterized as } v_1 \text{ of } a_1, \dots \text{ and } v_d \text{ of } a_d?"$ に変換する。

また、変換部12Bは、補助レコード $R = \{e_{_1}, \dots, e_{_k}\}$ を、文書 $T_2(R) = \text{"T3}(e_{_1}). \text{T3}(e_{_2}). \dots \text{T3}(e_{_k})."$ に変換する。

ここで、文書 $T_2(R)$ に含まれる回答文 $T_3(e_{_j})(1 \leq j \leq k)$ は、 $T_3(e_{_j}) = \text{"\{ID of } e_{_j}\} \text{ is characterized as } v_1 \text{ of } a_1, \dots, \text{ and } v_d \text{ of } a_d"$ である。

ここで、 $\{\text{ID of } e_{_j}\}$ は、レコード $e_{_j}$ に割り当てられる固有のIDであ

10

20

30

40

50

る。ただし、変換部 1 2 B は、変換の際に欠損値を文書に含めない。

【 0 1 1 4 】

変換部 1 2 B は、一例として、

$e = ((\text{title, sims 2 glamour life stuff pack}) , (\text{manufacturer, aspyr media}) , (\text{price, 24.99}))$

を、

$q = \text{"What is characterized as title of sims 2 glamour life stuff pack, manufacturer of aspyr media, and price of 24.99?"}$

に変換する。また、補助レコード $R = \{ ((\text{title, adobe photoshop elements 4.0 photo-editing software for mac}), (\text{price, 85.95})), ((\text{title, aspyr media inc sims 2 glamour life stuff pack}), (\text{price, NaN})), ((\text{title, final-draft final draft av 2.5 screenwriting software mac/win screen writing software}), (\text{price, 199.95})) \}$

を、文書 $c = \text{"r1 is characterized as title of adobe photoshop elements 4.0 photo-editing software for mac and price of 85.95. r2 is characterized as title of aspyr media inc sims 2 glamour life stuff pack. r3 is characterized as title of final-draft final draft av 2.5 screenwriting software mac/win screen writing software and price of 199.95."}$

に変換する。

【 0 1 1 5 】

また、ステップ S 1 0 4 B において、類似度算出部 1 3 B は、質問応答モデルに質問文 q と文書 c とを入力する。質問応答モデルは、入力された質問文 q の回答が文書 c から抽出した回答文 $T 3 (e_j) (1 \leq j \leq k)$ である確信度を示すスコアを出力する。類似度算出部 1 3 B は、質問応答モデルが出力したスコアに基づき類似度 s を算出する。類似度 s は、一例として、 $MB(q, c, \{ID \text{ of } e'\})$ 、すなわちレコード対 (e, e') に含まれるレコード e' が回答文である確信度である。ただし、類似度 s はこの例に限られず、類似度算出部 1 3 B は他の手法により類似度 s を算出してもよい。類似度算出部 1 3 B は、一例として、レコード e を質問文にした際のスコアと、レコード e' を質問文にした際のスコアとの和を類似度としてもよい。

【 0 1 1 6 】

図 1 2 は、質問応答モデルを用いた類似度の算出処理の概念図である。図 1 2 の例では、変換部 1 2 B がレコード e と補助レコード R とを質問文と文書に変換し、類似度算出部 1 3 B が質問文と文書とを質問応答モデルであるモデル MB に入力することにより、類似度 s を算出する。このように、本処理例では、情報処理装置 1 B は、レコードを質問応答の形式に落とし込むことによって、レコードの類似度を算出する。

【 0 1 1 7 】

< 情報処理装置 1 B の効果 >

以上のように、本例示の実施形態に係る情報処理装置 1 B においては、取得部 1 1 B は補助レコードを更に取得し、変換部 1 2 B は補助レコードを変換することによって変換済補助レコードを生成し、類似度算出部 1 3 B は上記変換済レコード対及び上記変換済補助レコードをモデル MB に入力することによって、上記変換済レコード対に関する類似度を算出する構成が採用されている。このため、本例示の実施形態に係る情報処理装置 1 B によれば、モデル MB を訓練することなく、モデル MB を用いて様々な属性を有するレコードについて類似度を算出できるという効果が得られる。

【 0 1 1 8 】

また、本例示の実施形態に係る情報処理装置 1 B においては、モデル MB には、質問文と応答文とを入力とする質問応答モデルが含まれており、変換部 1 2 B は、上記レコード対に含まれる一方のレコードを質問文に変換し、上記レコード対に含まれる他方のレコード及び上記補助レコードの各々を応答文に変換することによって上記変換済レコード対を生成するという構成が採用されている。このため、本例示の実施形態に係る情報処理装置 1 B によれば、質問応答モデルを訓練することなく、様々な属性を有するレコードについ

10

20

30

40

50

て質疑応答モデルを用いて類似度を算出できるという効果が得られる。

【0119】

〔例示的实施形態4〕

本発明の第4の例示的实施形態について、図面を参照して詳細に説明する。なお、例示的实施形態1～3にて説明した構成要素と同じ機能を有する構成要素については、同じ符号を付記し、その説明を繰り返さない。

【0120】

＜情報処理装置1Cの構成＞

図13は、本例示的实施形態に係る情報処理装置1Cの構成を示すブロック図である。情報処理装置1Cの制御部10Cは、取得部11、変換部12、類似度算出部13C、類似度統合部17C、出力部14C、同一性判定部15A及び統合部16Aを備える。また、記憶部20Cは、第1データx、第2データx'及び類似度sに加えてモデルMCを記憶する。

10

【0121】

（類似度算出部13C）

類似度算出部13Cは、ひとつのレコード対(e, e')に関して複数の類似度s_iを算出する。類似度算出部13Cは、一例として、モデルMCに対して、レコード対(e, e')に含まれる2つのレコードを互いに入れ替えずに入力することによって第1の類似度s₁を算出する。また、類似度算出部13Cは、モデルMCに対して、レコード対(e, e')に含まれる2つのレコードを互いに入れ替えてから入力することによって第2の類似度s₂を算出する。

20

【0122】

上述の例示的实施形態2で説明した処理例4～5、及び例示的实施形態3で説明した処理例6では、レコード対(e, e')の類似度は、レコード対(e', e)の類似度と異なる。そこで、本例示的实施形態では、類似度算出部13Cは、レコード対(e, e')の類似度とレコード対(e', e)の類似度とをそれぞれ算出し、それらの類似度を参照して同一性を判定する。

【0123】

ただし、類似度算出部13Cが複数の類似度s_iを算出する手法は上述した例に限られず、類似度算出部13Cは他の手法により複数の類似度s_iを算出してもよい。例えば、類似度算出部13Cは複数のモデルを用いて複数の類似度s_iを算出してもよい。この場合、例えば、変換部12が1つのレコード対に対して複数の変換を施し、類似度算出部13Cが変換後のレコード対をそれぞれのモデル(文書分類モデル、画像分類モデル、・・・)に入力することにより、複数の類似度s_iを算出してもよい。

30

【0124】

また、類似度算出部13Cは、ひとつのレコード対を複数の変換方法で変換することにより複数の変換済レコード対を生成し、複数の変換済レコード対をひとつのモデルに入力することにより、複数の類似度s_iを算出してもよい。

【0125】

（類似度統合部17C）

類似度統合部17Cは、複数の類似度s_iを統合して統合後の類似度sとする。類似度統合部17Cは、一例として、複数の類似度s_iの平均又は加重平均をとることによって統合後の類似度sを算出する。ただし、類似度統合部17Cが複数の類似度s_iを統合する手法は上述した例に限られず、類似度統合部17Cは他の手法により統合後の類似度sを算出してもよい。例えば、類似度統合部17Cは、複数の類似度s_iの総和又は積算値を統合後の類似度sとしてもよい。

40

【0126】

本明細書において、類似度統合部17Cは、対象レコード対に関する複数の類似度s_iに基づき、当該対象レコード対に関する同一性を判定する構成であるともいえる。

【0127】

50

(出力部 14C)

出力部 14C は、上記複数の類似度 s_i を統合して得られる統合後の類似度 s を出力する。出力部 14C は、一例として、記憶部 20C に類似度 s を書き込むことにより出力する。

【0128】

(モデル MC)

モデル MC は、類似度を算出するためのモデルである。モデル MC は、一例として、当該モデルに入力される 2 つの要素の互いの入れ替えに対して非対称性を有するモデルである。モデル MC は、一例として、含意認識モデル、言い換え予測モデル、及び質問応答モデルの少なくともいずれかひとつを含む。

10

【0129】

図 14 は、類似度算出部 13C が算出する類似度 s_i の具体例を示す図である。図 14 において、レコード対 (L1, R1) について類似度算出部 13C が算出した第 1 の類似度 s_1 は「9」であり、2 つのレコードを互いに入れ替えたレコード対 (R1, L1) について類似度算出部 13C が算出した第 2 の類似度 s_2 は「10」である。このように、類似度算出部 13C はひとつのレコード対について第 1 の類似度 s_1 と第 2 の類似度 s_2 とを算出し、同一性判定部 15A は第 1 の類似度 s_1 と第 2 の類似度 s_2 とが共に他のレコード対と比較して最も高ければレコードが同一であると判定する。図 14 の例では、同一性判定部 15A は、レコード L1 とレコード R1 が同一であると判定するとともに、レコード L2 とレコード R3 が同一であると判定する。

20

【0130】

図 15 は、類似度算出部 13C が算出する類似度 s_i の他の例を示す図である。図 15 において、類似度統合部 17C は、双方向の類似度を集約する。類似度統合部 17C は、一例として、レコード対 (L1, R1) の類似度 s_1 と、レコード対 (R1, L1) の類似度 s_2 との和を類似度 s とする。図 15 の例では、レコード対 (L1, R1) の類似度 s は「10」と「9」の和、すなわち「19」であり、レコード対 (L1, R2) の類似度 s は「9」と「7」の和、すなわち「16」である。また、レコード対 (L2, R2) の類似度 s は「9」と「4」の和、すなわち「13」であり、レコード対 (L2, R3) の類似度 s は「8」と「8」の和、すなわち「16」である。

【0131】

30

図 15 の例では、同一性判定部 15A は、図 14 の例と同様に、レコード L1 とレコード R1 が同一であると判定するとともに、レコード L2 とレコード R3 が同一であると判定する。この例では更に、同一性判定部 15A は、同一であると判定したレコード対の中で類似度 s が所定の閾値以上であるレコード対についても、同一であると判定する。ここで、閾値は例えば、同一であると判定されたレコード対の類似度 s のうちの最小値 (図 15 の例では、「13」) である。閾値は、同一と非同一の割合が既知であるなら、その割合に基づいて決められてもよい。図 15 の例において閾値が「13」である場合、同一性判定部 15A は、類似度 s が「13」以上であるレコード対、すなわち類似度 s が「16」であるレコード対 (L1, R2) についてもレコード同士が同一であると判定する。

【0132】

40

< 情報処理装置 1C の効果 >

以上のように、本例示的实施形態に係る情報処理装置 1C においては、類似度算出部 13C は、上記レコード対に関して複数の類似度 s_i を算出し、出力部 14C は、複数の類似度 s_i を統合して得られる統合後の類似度 s を出力する構成が採用されている。このため、本例示的实施形態に係る情報処理装置 1C によれば、レコード対の類似度 s をより精度よく算出できるという効果が得られる。

【0133】

また、本例示的实施形態に係る情報処理装置 1C においては、モデル MC は、当該モデルに入力される 2 つの要素の互いの入れ替えに対して非対称性を有するモデルであり、類似度算出部 13C は、モデル MC に対して、レコード対 (e, e') に含まれる 2 つのレ

50

コードを互いに入れ替えずに入力することによって第1の類似度 s_1 を算出し、モデルM Cに対して、レコード対 (e, e') に含まれる2つのレコードを互いに入れ替えてから入力することによって第2の類似度 s_2 を算出する構成が採用されている。このため、本例示的实施形態に係る情報処理装置1 Cによれば、第1の類似度 s_1 と第2の類似度 s_2 とを統合することにより、レコードの類似度 s をより精度よく算出できるという効果が得られる。

【0134】

〔例示的实施形態5〕

本発明の第5の例示的实施形態について、図面を参照して詳細に説明する。なお、例示的实施形態1～3にて説明した構成要素と同じ機能を有する構成要素については、同じ符号を付記し、その説明を繰り返さない。

10

【0135】

<情報処理装置1 Dの構成>

図16は、本例示的实施形態に係る情報処理装置1 Dの構成を示すブロック図である。情報処理装置1 Dの制御部10 Dは、取得部11、変換部12、類似度算出部13、出力部14、同一性判定部15 A及び検索結果出力部18 Dを備える。

【0136】

本例示的实施形態に係る取得部11は、レコード対 (e, e') に含まれる第1のレコード e として、ユーザからの入力データを取得する。ユーザからの入力データは、一例として、入出力部40 Aに接続された入力装置（例えば、キーボード、マウス、等）により入力される。

20

【0137】

また、取得部11は、レコード対 (e, e') に含まれる第2のレコード e' として、対象データに含まれる複数のレコードの1つを取得する。対象データは、検索対象のデータであり、一例として、1又は複数のテーブルを含む。

【0138】

同一性判定部15 Aは、第1のレコード e と、対象データに含まれる複数のレコードの各々とのレコード対に対して同一性予測を行う。

【0139】

検索結果出力部18 Dは、類似度算出部13が算出した類似度 s に基づき、入力データに基づく検索結果であって、対象データを検索対象とする検索結果を出力する。検索結果出力部18 Dは、一例として、同一性判定部15 Aの判定結果を参照して、入力データに基づく検索結果であって、対象データを検索対象とする検索結果を出力する。検索結果出力部18 Dは、一例として、入出力部40 Aに接続された出力装置（ディスプレイ、プリンタ、等）に検索結果を出力する。また、検索結果出力部18 Dは、通信部30 Aを介して接続された他の装置に検索結果を送信することにより、検索結果を出力してもよい。また、検索結果出力部18 Dは、検索結果を記憶部20 A又は外部記憶装置に記憶することにより検索結果を出力してもよい。

30

【0140】

図17は、検索結果出力部18 Dが出力する画面表示の具体例を示す図である。図17の例で、入力データは、ユーザがテキストボックス51に入力する文字列であり、対象データは複数のレコードを有するテーブルT1及びテーブルT2である。同一性判定部15 Aは、ユーザの入力データである第1のレコード e と、テーブルT1に含まれるレコード及びテーブルT2に含まれるレコード e' の各々とのレコード対に対して同一性を判定する。

40

【0141】

図17の例において、検索結果出力部18 Dは、同一性判定部15 Aの判定結果を参照して、入力データに基づく検索結果53、及び検索結果54を出力する。検索結果53は、「ポテチ」の文字列を入力データとして、テーブルT1から検索された検索結果である。検索結果54は、「ポテチ」の文字列を入力データとして、テーブルT2から検索され

50

た検索結果である。

【0142】

< 情報処理装置 1 D の効果 >

以上のように、本例示的实施形態に係る情報処理装置 1 D においては、同一性判定部 1 5 A の判定結果を参照して、入力データに基づく検索結果であって、対象データを検索対象とする検索結果を出力する構成が採用されている。このため、本例示的实施形態に係る情報処理装置 1 D によれば、例示的实施形態 1 に係る情報処理装置 1 の奏する効果に加えて、入力データに基づく対象データからの検索をより好適に行うことができるという効果が得られる。

【0143】

情報処理装置 1 D は、以下のようにも記載され得る。

ユーザからの入力データと、対象データに含まれる複数のレコードの 1 つとをレコード対として取得する取得手段と、

前記レコード対を変換することによって変換済レコード対を生成する変換手段と、

前記変換済レコード対をモデルに入力することによって、変換済レコード対に関する類似度を算出する類似度算出手段と、

前記類似度算出手段が算出した類似度を参照して、前記入力データに基づく検索結果であって、前記対象データを検索対象とする検索結果を出力する出力手段と、
を備えている情報処理装置。

【0144】

〔変形例〕

< 変形例 1 >

上述の各例示的实施形態では、情報処理装置 1、1 A、1 B、1 C、1 D（以下「情報処理装置 1 等」という）は、第 1 データ x に含まれるレコード e と第 2 データ x' に含まれるレコード e' との同一性を判定した。情報処理装置 1 等が判定の対象とする複数のレコードはそれぞれ異なるデータに含まれるレコードであってもよく、また、共通のデータに含まれるレコードであってもよい。換言すると、情報処理装置 1 等は、1 つのデータベースから同一のレコードを探し出す処理を実行してもよい。また、上述の例示的实施形態では、第 1 データ x と第 2 データ x' とを統合する場合について説明したが、情報処理装置 1 等は 3 つ以上のデータを統合してもよい。

【0145】

< 変形例 2 >

上述の各例示的实施形態において、情報処理装置 1 等が複数のモデル候補からモデル M A、M B、M C（以下「モデル M」という）を選択してもよく、また、ユーザが複数のモデル候補からモデル M を選択してもよい。情報処理装置 1 等がモデル M を選択するアルゴリズムは限定されないが、一例として、情報処理装置 1 等はルールベースでモデル M を選択してもよい。例えば、情報処理装置 1 等はレコード対の特徴に応じてモデル M を選択してもよい。ここで、レコード対の特徴は、一例として、レコード対に含まれるレコードの属性、レコードのデータサイズ、レコードが属するデータベースの種別、データベースの属性を含む。

【0146】

< 変形例 3 >

上述の各例示的实施形態において、レコード e、e' が含まれるデータは、JSON 又は XML 等の半構造データであってもよい。半構造データに上記例示的实施形態に係る情報処理装置を適用することにより、書類データ又はウェブページの同一性判定を行うことができる。例えば、住宅情報を提供する住宅情報サイトにおいては、同一の物件についてのウェブページが複数作成されている場合がある。この場合、ウェブページについて同一性判定を行うことにより、ウェブページを物件ごとにまとめることができる。

【0147】

この例において、レコードは一例として、対象であるサイトに含まれるウェブページで

10

20

30

40

50

ある。例えば、レコード $e = \{id1: value1, id2: \{id2-1: value2-1, id2-2: value2-1\}, id3: value3\}$

である場合、変換後の文書は、一例として、

“ id1 is value1. id2-1 of id2 is value2-1. id2-2 of id2 is value2-1. id3 is value3. ”
である。

【 0 1 4 8 】

< 変形例 4 >

また、本明細書に係るレコードは、例えば図 1 8 に示すようなグラフデータであってもよい。図 1 8 は、グラフデータの一例を示す図である。グラフデータに本明細書に係る情報処理装置 1 等を適用することにより、例えば顔照合を行うことができる。例えば、レコードが図 1 8 に示すグラフである場合、変換後の文書は、一例として、

“ 1 and 2 are linked. 1 and 4 are linked. 2 and 3 are linked. 2 and 4 are linked. ”
である。

【 0 1 4 9 】

< 変形例 5 >

また、レコードが含まれるデータは、例えば図 1 9 に示すようなグラフデータベースであってもよい。グラフデータベースに本明細書に係る情報処理装置 1 等を適用することにより、例えば異なる SNS (Social Networking Service) のコミュニティの同一性を判定することができ、例えば犯罪組織の調査に応用できる。この例で、グラフデータベースが図 1 9 に示すものである場合、変換後の文書は一例として、

“ Taro of age 23 follows Sakura of age 26. Taro of age 23 follows Emi of age 25. Sakura of age 26 follows Emi of age 25. Sakura of age 26 wrote via smart phone tweet of text “ I ’ m sleepy. ” date 20XX/YY/ZZ. Emi of age 25 follows Sakura of age 26. Emi of age 25 follows Taro of age 23. ”

である。

【 0 1 5 0 】

< 変形例 6 >

上述の各例示の実施形態において、情報処理装置 1 等がモデル M を学習する学習フェーズを実行する構成であってもよい。モデル M の機械学習の手法は限定されないが、一例として、決定木ベース、線形回帰、又はニューラルネットワークの手法が用いられてもよく、また、これらのうちの 2 以上の手法が用いられてもよい。

【 0 1 5 1 】

< 変形例 7 >

上述の各例示の実施形態において、モデル M の出力の前後に、学習可能なパラメータを備えた変換器を加える構成としてもよい。図 2 0 は、学習可能なパラメータを備えた学習済み変換器 1 2 1、1 2 2 をモデル M の出力の前後に設けた構成を概略的に示す図である。学習済み変換器 1 2 1、1 2 2 は学習可能なパラメータを備え、学習部 (図示略) が訓練データを用いてレコード変換の仕方 (文章の作り方又は補助レコードの数、等)、及び / 又は、変換のパラメータを最適化するモデルである。学習済み変換器 1 2 1、1 2 2 を設けることにより、レコードの類似度をより精度よく算出することができる。

【 0 1 5 2 】

学習済み変換器 1 2 1、1 2 2 の機械学習の手法は限定されないが、一例として、決定木ベース、線形回帰、又はニューラルネットワークの手法が用いられてもよく、また、これらのうちの 2 以上の手法が用いられてもよい。また、学習済み変換器 1 2 1、1 2 2 は能動学習により生成されたモデルであってもよい。

【 0 1 5 3 】

〔 ソフトウェアによる実現例 〕

情報処理装置 1、1 A、1 B、1 C、1 D の一部又は全部の機能は、集積回路 (IC チップ) 等のハードウェアによって実現してもよいし、ソフトウェアによって実現してもよい。

10

20

30

40

50

【 0 1 5 4 】

後者の場合、情報処理装置 1、1 A、1 B、1 C、1 D は、例えば、各機能を実現するソフトウェアであるプログラムの命令を実行するコンピュータによって実現される。このようなコンピュータの一例（以下、コンピュータ C と記載する）を図 1 8 に示す。コンピュータ C は、少なくとも 1 つのプロセッサ C 1 と、少なくとも 1 つのメモリ C 2 と、を備えている。メモリ C 2 には、コンピュータ C を情報処理装置 1、1 A、1 B、1 C、1 D として動作させるためのプログラム P が記録されている。コンピュータ C において、プロセッサ C 1 は、プログラム P をメモリ C 2 から読み取って実行することにより、情報処理装置 1、1 A、1 B、1 C、1 D の各機能が実現される。

【 0 1 5 5 】

プロセッサ C 1 としては、例えば、C P U (Central Processing Unit)、G P U (Graphic Processing Unit)、D S P (Digital Signal Processor)、M P U (Micro Processing Unit)、F P U (Floating point number Processing Unit)、P P U (Physics Processing Unit)、マイクロコントローラ、又は、これらの組み合わせなどを用いることができる。メモリ C 2 としては、例えば、フラッシュメモリ、H D D (Hard Disk Drive)、S S D (Solid State Drive)、又は、これらの組み合わせなどを用いることができる。

【 0 1 5 6 】

なお、コンピュータ C は、プログラム P を実行時に展開したり、各種データを一時的に記憶したりするための R A M (Random Access Memory) を更に備えていてもよい。また、コンピュータ C は、他の装置との間でデータを送受信するための通信インタフェースを更に備えていてもよい。また、コンピュータ C は、キーボードやマウス、ディスプレイやプリンタなどの入出力機器を接続するための入出力インタフェースを更に備えていてもよい。

【 0 1 5 7 】

また、プログラム P は、コンピュータ C が読み取り可能な、一時的でない有形の記録媒体 M に記録することができる。このような記録媒体 M としては、例えば、テープ、ディスク、カード、半導体メモリ、又はプログラマブルな論理回路などを用いることができる。コンピュータ C は、このような記録媒体 M を介してプログラム P を取得することができる。また、プログラム P は、伝送媒体を介して伝送することができる。このような伝送媒体としては、例えば、通信ネットワーク、又は放送波などを用いることができる。コンピュータ C は、このような伝送媒体を介してプログラム P を取得することもできる。

【 0 1 5 8 】

〔 付記事項 1 〕

本発明は、上述した実施形態に限定されるものでなく、請求項に示した範囲で種々の変更が可能である。例えば、上述した実施形態に開示された技術的手段を適宜組み合わせて得られる実施形態についても、本発明の技術的範囲に含まれる。

【 0 1 5 9 】

〔 付記事項 2 〕

上述した実施形態の一部又は全部は、以下のようにも記載され得る。ただし、本発明は、以下の記載する態様に限定されるものではない。

(付記 1)

レコード対を取得する取得手段と、
前記レコード対を変換することによって変換済レコード対を生成する変換手段と、
前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出する類似度算出手段と、
前記類似度算出手段が算出した類似度を出力する出力手段と、
を備えている情報処理装置。

【 0 1 6 0 】

(付記 2)

前記モデルは、複数のモデル候補から選択されたものであり、

前記変換手段は、前記レコード対を前記モデルの入力に対応する形式に変換することによって前記変換済レコード対を生成する、
付記 1 に記載の情報処理装置。

【0161】

(付記 3)

前記モデルには、文書分類モデルが含まれており、

前記変換手段による変換処理には、前記レコード対を文書に変換することによって前記変換済レコード対を生成する処理が含まれる、
付記 1 又は 2 に記載の情報処理装置。

10

【0162】

(付記 4)

前記モデルには、画像分類モデルが含まれており、

前記変換手段による変換処理には、前記レコード対を画像に変換することによって前記変換済レコード対を生成する処理が含まれる、
付記 1 から 3 の何れか 1 つに記載の情報処理装置。

【0163】

(付記 5)

前記モデルには、音声分類モデルが含まれており、

前記変換手段による変換処理には、前記レコード対を音声に変換することによって前記変換済レコード対を生成する処理が含まれる、
付記 1 から 4 の何れか 1 つに記載の情報処理装置。

20

【0164】

(付記 6)

前記モデルには、グラフ分類モデルが含まれており、

前記変換手段による変換処理には、前記レコード対をグラフに変換することによって前記変換済レコード対を生成する処理が含まれる、
付記 1 から 5 の何れか 1 つに記載の情報処理装置。

【0165】

(付記 7)

前記取得手段は、補助レコードを更に取得し、

前記変換手段は、前記補助レコードを変換することによって変換済補助レコードを生成し、

30

前記類似度算出手段は、前記変換済レコード対及び前記変換済補助レコードを前記モデルに入力することによって、前記変換済レコード対に関する類似度を算出する、
付記 1 から 6 の何れか 1 つに記載の情報処理装置。

【0166】

(付記 8)

前記モデルには、質問文と応答文とを入力とする質問応答モデルが含まれており、

前記変換手段は、前記レコード対に含まれる一方のレコードを質問文に変換し、前記レコード対に含まれる他方のレコード及び前記補助レコードの各々を応答文に変換することによって前記変換済レコード対を生成する、
付記 7 に記載の情報処理装置。

40

【0167】

(付記 9)

前記類似度算出手段は、前記レコード対に関して複数の類似度を算出し、

前記出力手段は、前記複数の類似度を統合して得られる統合後の類似度を出力する、
付記 1 から 8 の何れか 1 つに記載の情報処理装置。

【0168】

(付記 10)

50

前記モデルは、当該モデルに入力される２つの要素の互いの入れ替えに対して非対称性を有するモデルであり、

前記類似度算出手段は、

前記モデルに対して、前記レコード対に含まれる２つのレコードを互いに入れ替えずに入力することによって第１の類似度を算出し、

前記モデルに対して、前記レコード対に含まれる２つのレコードを互いに入れ替えてから入力することによって第２の類似度を算出する、

付記９に記載の情報処理装置。

【０１６９】

（付記１１）

少なくとも１つのプロセッサが、

レコード対を取得することと、

前記レコード対を変換することによって変換済レコード対を生成することと、

前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出することと、

前記算出した類似度を出力することと、
を含む情報処理方法。

【０１７０】

（付記１２）

コンピュータに、

レコード対を取得する取得処理と、

前記レコード対を変換することによって変換済レコード対を生成する変換処理と、

前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出する類似度算出処理と、

前記類似度算出処理において算出した類似度を出力する出力処理と、
を実行させる情報処理プログラム。

【０１７１】

（付記１３）

ユーザからの入力データと、対象データに含まれる複数のレコードの１つとをレコード対として取得する取得手段と、

前記レコード対を変換することによって変換済レコード対を生成する変換手段と、

前記変換済レコード対をモデルに入力することによって、変換済レコード対に関する類似度を算出する類似度算出手段と、

前記類似度算出手段が算出した類似度を参照して、前記入力データに基づく検索結果であって、前記対象データを検索対象とする検索結果を出力する出力手段と、
を備えている情報処理装置。

【０１７２】

〔付記事項３〕

上述した実施形態の一部又は全部は、更に、以下のように表現することもできる。

【０１７３】

少なくとも１つのプロセッサを備え、前記プロセッサは、レコード対を取得する取得処理と、前記レコード対を変換することによって変換済レコード対を生成する変換処理と、前記変換済レコード対をモデルに入力することによって、前記変換済レコード対に関する類似度を算出する類似度算出処理と、前記類似度算出処理において算出した類似度を出力する出力処理とを実行する情報処理装置。

【０１７４】

なお、この情報処理装置は、更にメモリを備えていてもよく、このメモリには、前記取得処理と、前記変換処理と、前記類似度算出処理と、前記出力処理とを前記プロセッサに実行させるためのプログラムが記憶されていてもよい。また、このプログラムは、コンピュータ読み取り可能な一時的でない有形の記録媒体に記録されていてもよい。

10

20

30

40

50

【符号の説明】

【 0 1 7 5 】

- 1、1 A、1 B、1 C、1 D 情報処理装置
- 1 1、1 1 B 取得部
- 1 2、1 2 B 変換部
- 1 3、1 3 B、1 3 C 類似度算出部
- 1 4、1 4 C 出力部
- 1 6 A 統合部

10

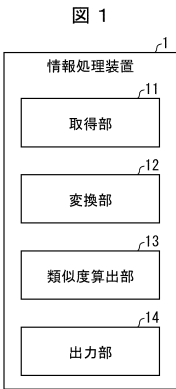
20

30

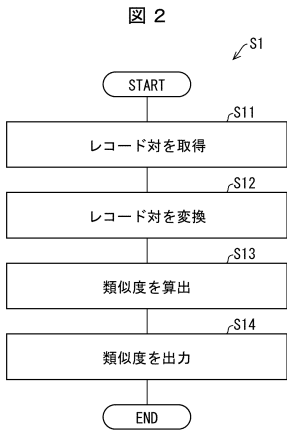
40

50

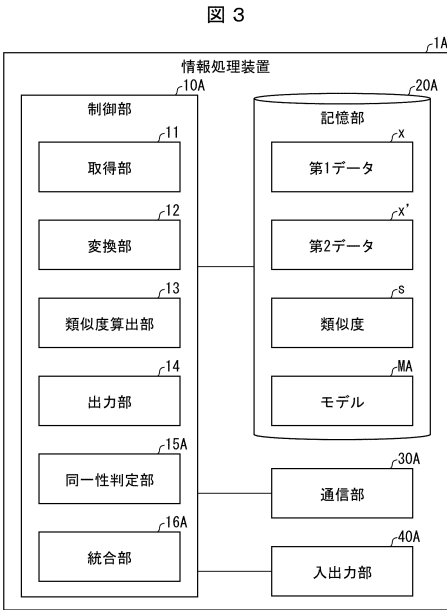
【図面】
【図 1】



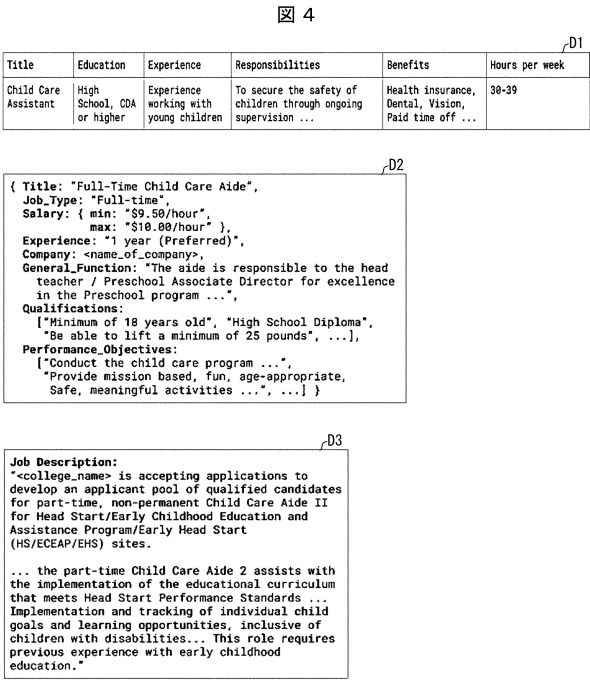
【図 2】



【図 3】



【図 4】



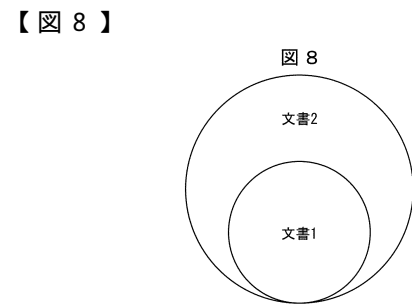
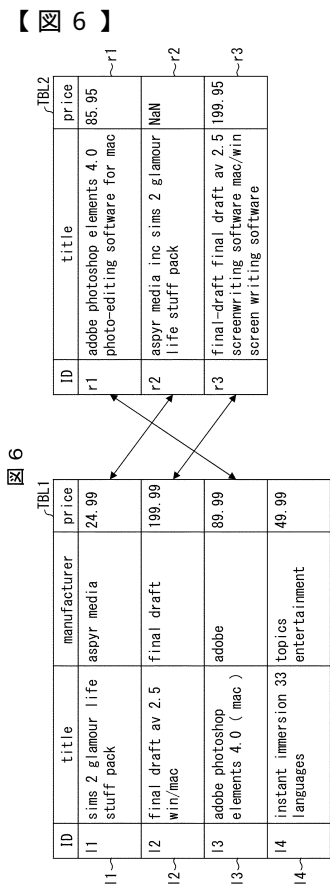
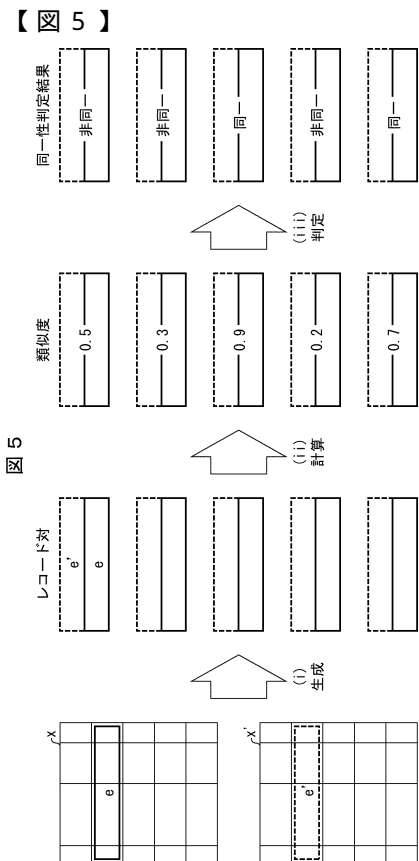
10

20

30

40

50



10

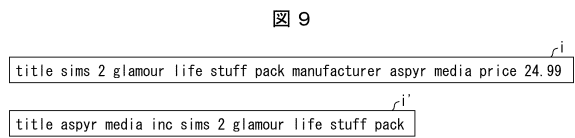
20

30

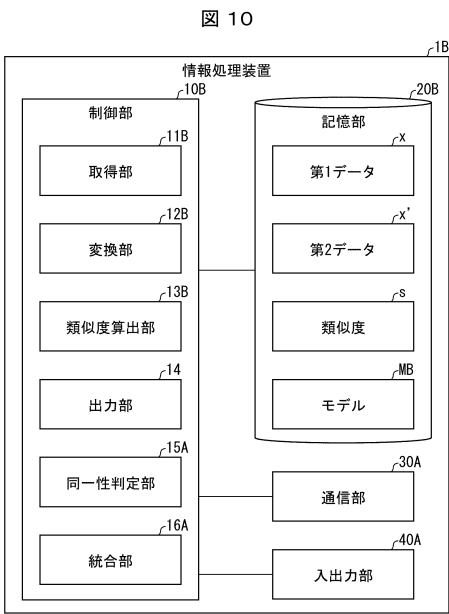
40

50

【図 9】

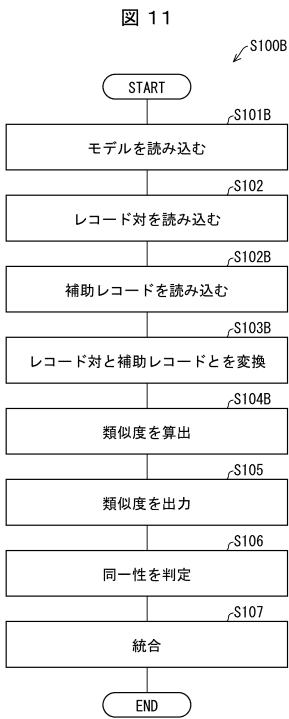


【図 10】

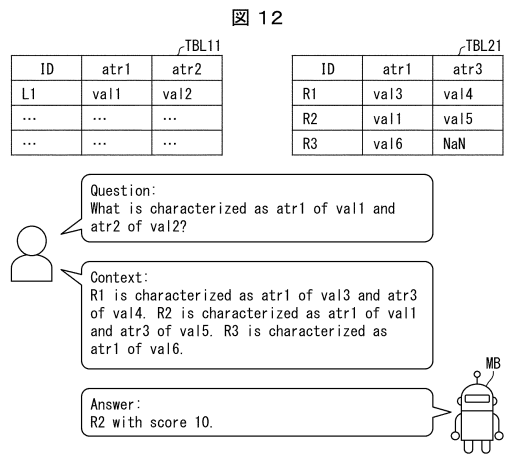


10

【図 11】



【図 12】



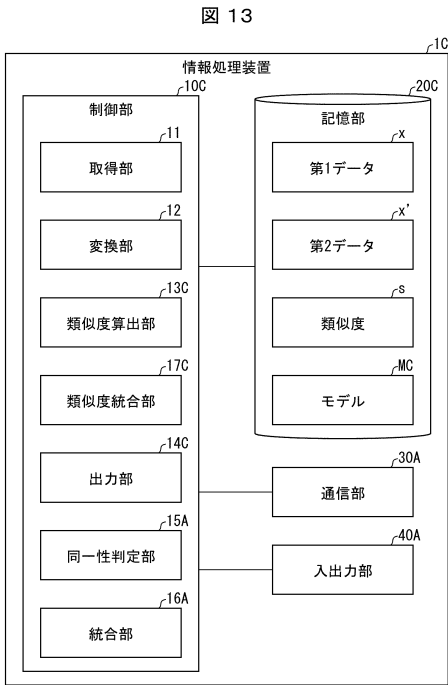
20

30

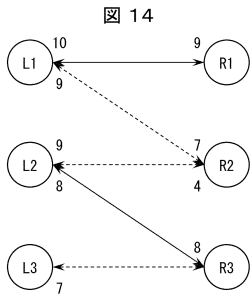
40

50

【図 1 3】



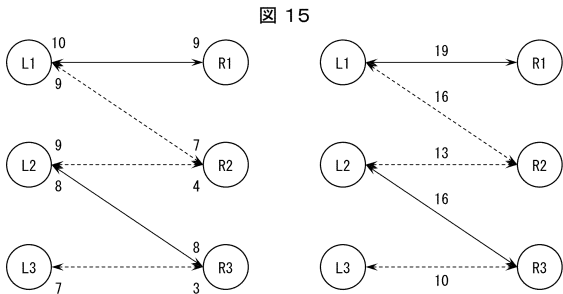
【図 1 4】



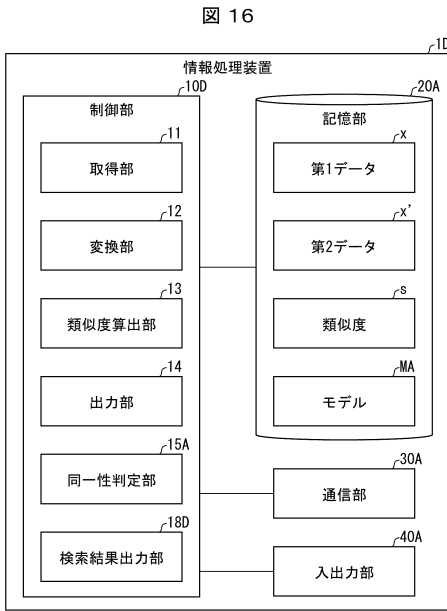
10

20

【図 1 5】



【図 1 6】

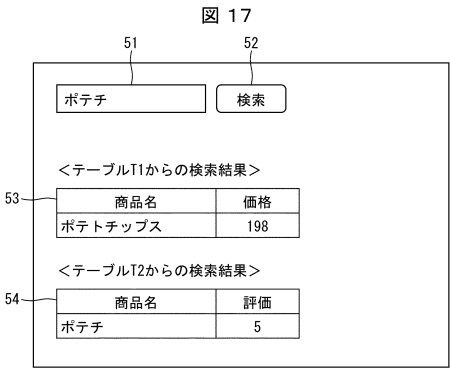


30

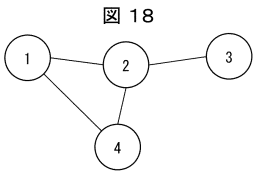
40

50

【図 17】

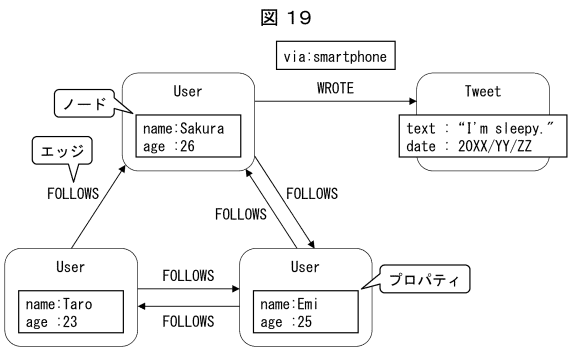


【図 18】

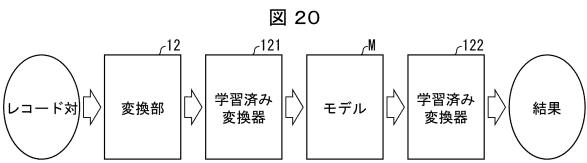


10

【図 19】

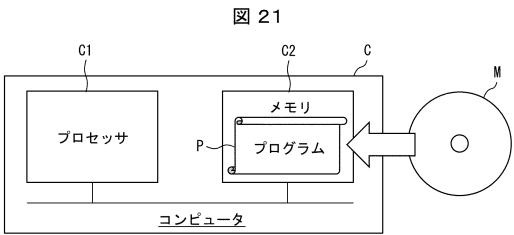


【図 20】



20

【図 21】



30

40

50

フロントページの続き

- (56)参考文献 特開 2 0 1 9 - 1 8 5 2 4 4 (J P , A)
 米国特許出願公開第 2 0 1 7 / 0 0 9 1 2 7 4 (U S , A 1)
 米国特許出願公開第 2 0 2 1 / 0 3 7 4 1 8 6 (U S , A 1)
 米国特許出願公開第 2 0 2 1 / 0 3 7 4 1 6 4 (U S , A 1)
 特開 2 0 2 1 - 1 7 4 3 0 0 (J P , A)
 特表 2 0 2 2 - 5 1 0 8 1 8 (J P , A)
 LI, Yuliang et al. , Deep Entity Matching with Pre-Trained Language Models , Cornell University , 2020年09月02日 , pp.1-15 , [検索日 2022.05.02], Internet: URL: <https://arxiv.org/pdf/2004.00584.pdf>
- (58)調査した分野 (Int.Cl. , D B 名)
 G 0 6 F 1 6 / 0 0 - 1 6 / 9 5 8