【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】令和6年9月13日(2024.9.13)

【国際公開番号】WO2023/132029
【出願番号】特願2023-572292(P2023-572292)
【国際特許分類】
G06F 16/215(2019.01)
G06F 16/24(2019.01)
G06N 20/00(2019.01)
【FI】
G06F 16/215
G06F 16/24
G06N 20/00 160

【手続補正書】
【提出日】令和6年7月1日(2024.7.1)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【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、
前記レコード対について、複数の類似度関数を用いて複数の類似度を算出する類似度算出手段と、
前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行う予測手段と、
前記予測手段による予測結果を出力する出力手段と、
を備えている情報処理装置。
【請求項２】
前記取得手段は、補助データを更に取得し、
前記予測手段は、前記レコード対と、前記複数の類似度と、前記補助データとを参照して、前記レコード対と前記補助データとに応じて定まる重要度を用いて前記レコード対の同一性予測を行う
請求項１に記載の情報処理装置。
【請求項３】
前記予測手段は、前記レコード対を参照して前記重要度を算出する重要度算出手段を備えている
請求項１又は２に記載の情報処理装置。
【請求項４】
前記重要度算出手段は、前記複数の類似度の各々に関する重要度を算出し、
前記予測手段は、前記複数の類似度に関する線形和であって、前記各重要度を重み係数とする線形和を用いて、前記同一性予測を行う
請求項３に記載の情報処理装置。
【請求項５】
レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得する取得手段と、
予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が

有する１又は複数のパラメータ、及び
前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータ
の少なくとも何れかのパラメータを、前記訓練データを参照して生成するパラメータ生成手段と
を備えている情報処理装置。
【請求項６】
レコード対を取得することと、
前記レコード対について、複数の類似度関数を用いて複数の類似度を算出することと、
前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行うことと、
前記レコード対の同一性予測による予測結果を出力することと、
を含む情報処理方法。
【請求項７】
レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得することと、
予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び
前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータ
の少なくとも何れかのパラメータを、前記訓練データを参照して生成することと、
を含む情報処理方法。
【請求項８】
レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得することと、
予測対象のレコード対について複数の類似度を算出するための複数の類似度算出モデル、及び
前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデル
の少なくとも何れかのモデルを、前記訓練データを参照して生成することと、
を含む学習済モデルの製造方法。
【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、
前記レコード対について、複数の類似度関数を用いて複数の類似度を算出する類似度算出処理と、
前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行う予測処理と、
前記予測処理による予測結果を出力する出力処理と、
を実行させるプログラム。
【請求項１０】
コンピュータに、
レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得する取得処理と、
予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が

有する１又は複数のパラメータ、及び
前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータ
の少なくとも何れかのパラメータを、前記訓練データを参照して生成するパラメータ生成処理と、
を実行させるプログラム。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、レコード対の同一性予測を行う技術に関する。
【背景技術】
【０００２】
異なるテーブル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が可能となる。名寄せ処理を行う技術として、機械学習又はルールベースによるマッチングを行う技術が存在する。例えば、特許文献１及び非特許文献１には、機械学習により名寄せ処理を行う技術が記載されている。特に、特許文献１に記載の名寄せ処理装置は、情報処理装置と、記憶部と操作端末とから構成されている。この名寄せ処理装置は、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機械学習により類似度の重みを学習する。
【先行技術文献】
【特許文献】
【０００３】
【特許文献１】日本国特開２０１９－１８５２４４号公報
【非特許文献】
【０００４】
【非特許文献１】Pradap Konda, et. al., Magellan: Toward Building Entity Matching Management Systems, Proceedings of the VLDB Endowment, 2016
【発明の概要】
【発明が解決しようとする課題】
【０００５】
ところで、レコード対の同一性を判定する手法として様々な手法が存在する。例えば、「アイス」と「あいす」のレコード対はカタカナをひらがなに表記変更することで同一性の判定精度を高くすることができる。また、「ポテトチップス」と「ポテチ」のレコード対は部分文字列を抽出することで同一性の判定精度を高くすることができる。このように、レコード対の同一性の判定に適した手法はレコード対のそれぞれで異なる場合がある。特許文献１及び非特許文献１に記載の技術では、レコード対によっては同一性の判定を適切に行えないという問題があった。
【０００６】
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、レコード対の同一性をより好適に予測できる技術を提供することである。
【課題を解決するための手段】

【０００７】
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、前記レコード対について、複数の類似度関数を用いて複数の類似度を算出する類似度算出手段と、前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行う予測手段と、前記予測手段による予測結果を出力する出力手段と、を備えている。
【０００８】
また、本発明の一側面に係る情報処理装置は、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得する取得手段と、予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータの少なくとも何れかのパラメータを、前記訓練データを参照して生成するパラメータ生成手段と、を備えている。
【０００９】
また、本発明の一側面に係る情報処理方法は、レコード対を取得することと、前記レコード対について、複数の類似度関数を用いて複数の類似度を算出することと、前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行うことと、前記レコード対の同一性予測による予測結果を出力することと、を含む。
【００１０】
また、本発明の一側面に係る情報処理方法は、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得することと、予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータの少なくとも何れかのパラメータを、前記訓練データを参照して生成することと、を含む。
【００１１】
また、本発明の一側面に係る製造方法は、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得することと、予測対象のレコード対について複数の類似度を算出するための複数の類似度算出モデル、及び前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルの少なくとも何れかのモデルを、前記訓練データを参照して生成することと、を含む。
【００１２】
また、本発明の一側面に係るプログラムは、コンピュータに、レコード対を取得する取得処理と、前記レコード対について、複数の類似度関数を用いて複数の類似度を算出する類似度算出処理と、前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行う予測処理と、前記予測処理による予測結果を出力する出力処理と、を実行させる。
【００１３】
また、本発明の一側面に係るプログラムは、コンピュータに、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得する取得処理と、予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコー

ド対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータの少なくとも何れかのパラメータを、前記訓練データを参照して生成するパラメータ生成処理と、を実行させる。

【発明の効果】

【００１４】

本発明の一態様によれば、レコード対の同一性をより好適に予測できる。

【図面の簡単な説明】

【００１５】

【図１】例示的実施形態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】例示的実施形態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】例示的実施形態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４】例示的実施形態１に係る情報処理方法の流れを示すフロー図である。

【図５】例示的実施形態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６】例示的実施形態２に係る第１のデータと第２のデータの具体例を示す図である。

【図７】例示的実施形態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】例示的実施形態２に係る統合済データの具体例を示す図である。

【図９】例示的実施形態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１０】例示的実施形態３に係る情報処理方法の流れを示すフロー図である。

【図１１】例示的実施形態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１２】例示的実施形態４に係る画面表示例を示す図である。

【図１３】各例示的実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】

【００１６】

〔例示的実施形態１〕

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

【００１７】

＜情報処理装置１の構成＞

本例示的実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、レコード対の同一性予測を行う装置である。情報処理装置１は、取得部１１、類似度算出部１２、予測部１３及び出力部１４を備える。

【００１８】

（取得部１１）

取得部１１は、レコード対を取得する。

【００１９】

（レコード対・レコード）

レコード対は複数のレコードのセットである。レコードは、一例として、テーブルの行であり、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。レコード対に含まれるレコードの数は２であってもよく、また、３以上であってもよい。レコード対は、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、異なっていてもよい。

【００２０】

（類似度算出部１２）

類似度算出部１２は、取得部１１が取得したレコード対について、複数の類似度関数を

用いて複数の類似度を算出する。換言すると、類似度算出部１２は、ｋ個（ｋは２以上の整数）の類似度関数φ$_i$（１≦ｉ≦ｋ）を用いて、１つのレコード対についてｋ個の類似度を算出する。

【００２１】

（類似度関数）

類似度関数φ$_i$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ための関数である。以下では、類似度関数φ$_i$を「類似度算出モデル」とも呼ぶ。類似度関数φ$_i$の入力はレコード対であり、類似度関数φ$_i$の出力はレコード対に含まれるレコード同士の類似度である。複数の類似度関数φ$_i$は、後述する情報処理装置２による学習の対象であり得る。類似度関数φ$_i$が機械学習により生成される場合、類似度関数φ$_i$の機械学習の手法は限定されず、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。決定木ベースとしては、例えば、ＬｉｇｈｔＧＢＭ（Light Gradient Boosting Machine）、ランダムフォレスト、及びＸＧＢｏｏｓｔが挙げられる。線形回帰としては、例えば、ベイズ回帰、サポートベクター回帰、Ｒｉｄｇｅ回帰、Ｌａｓｓｏ回帰、及びＥｌａｓｔｉｃＮｅｔが挙げられる。ニューラルネットワークとしては、例えばディープラーニングが挙げられる。

【００２２】

類似度関数φ$_i$は、一例として、０～１の数値を類似度として出力する。類似度関数φ$_i$としては、例えば、Jaccard係数を用いることができる。Jaccard係数は、集合Ａ＝｛ａ１，ａ２，…｝と集合Ｂ＝｛ｂ１，ｂ２，…｝に対し、｜Ａ∩Ｂ｜／｜Ａ∪Ｂ｜を計算するものである。また、類似度関数φ$_i$としては例えば、非特許文献１に記載された手法が用いられてもよい。また、他の例として、類似度関数φ$_i$として、例えば文献「Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, Wang-Chiew Tan, Deep Entity Matching with Pre-Trained Language Models, Proceedings of the VLDB Endowment, 2016」（以下「非特許文献２」という）に記載された手法が用いられてもよい。ただし、類似度関数φ$_i$は上述した例に限られず、他の手法によりレコード対の類似度を算出するものであってもよい。

【００２３】

（予測部１３）

予測部１３は、レコード対と、複数の類似度とを参照して、レコード対に応じて定まる重要度を用いてレコード対の同一性予測を行う。

【００２４】

（重要度）

重要度は、レコード対に応じて定まる情報である。重要度は、一例として、レコード対を参照して算出される。より具体的には、一例として、重要度を算出するための重要度算出モデルを用いて予測部１３が重要度を算出する。この場合、重要度算出モデルの入力はレコード対である。また、重要度算出モデルの出力は重要度である。重要度算出モデルは、後述する情報処理装置２による学習の対象であり得る。重要度算出モデルが機械学習により生成される場合、重要度算出モデルの機械学習の手法は限定されず、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。

【００２５】

重要度算出モデルは、一例として、ＢＥＲＴ（Bidirectional encoder representations from Transformers）、fastText、word2vec、tf-idf、BM25、等の言語モデルを用いて生成される。また、重要度算出モデルは言語モデルを含んでもよい。言語モデルを用いる場合の重要度の算出処理の具体例について説明する。予測部１３は、一例として、言語モデルを用いてレコード対をベクトルに変換し、このベクトルを更に別の特徴量空間上のベクトルに変換する。更に、予測部１３は、このベクトルをｋクラス分類器（ソフトマックス関数、等）に入力することで、ｋ個の重要度を算出する。算出されるｋ

個の重要度のそれぞれは、ｋ個の類似度関数φ ｉのそれぞれに対応する。
【００２６】
ただし、重要度を算出する手法は上述した例に限られず、予測部１３は他の手法により重要度を算出してもよい。予測部１３は、一例として、ルールベースの処理により重要度を算出してもよい。例えば、予測部１３は、重要度とレコード対に関する情報とを対応付けたテーブルを参照することにより重要度を算出してもよい。ここで、レコード対に関する情報は、一例として、レコード対に含まれるレコードの特徴量、レコードの分類結果、又はレコードの名称、等を含んでもよい。
【００２７】
予測部１３は、一例として、類似度算出部１２が算出した複数の類似度に関する線形和であって、各重要度を重み係数とする線形和を用いて、レコード対の同一性予測を行う。ただし、予測部１３が同一性予測を行う手法は線形和を用いる手法に限られず、予測部１３は他の手法によりレコード対の同一性予測を行ってもよい。
【００２８】
予測部１３は、一例として、機械学習により生成される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、ｋ個の類似度のセットとレコード対とを含む。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。また、予測部１３は、予測モデルが有するパラメータを重要度として算出する。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。
【００２９】
（出力部１４）
出力部１４は、予測部１３による予測結果を出力する。予測結果は、一例として、レコード対に含まれるレコードが同一であるかを示す情報、又は、レコード対に含まれるレコードの類似度を示す情報を含む。
【００３０】
予測部１３による予測結果は、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。テーブルを統合する場合、予測部１３により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予測部１３が同一性予測を行ってもよい。この場合、予測部１３により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ける検索処理が可能となる。
【００３１】
＜情報処理装置１の効果＞
以上のように、本例示的実施形態に係る情報処理装置１においては、レコード対について複数の類似度関数を用いて複数の類似度を算出し、レコード対と、複数の類似度とを参照して、レコード対に応じて定まる重要度を用いてレコード対の同一性予測を行う構成が採用されている。ここで、重要度はレコード対に応じて定まるため、複数の類似度に基づく同一性予測の結果は、画一的な手法によるものではなく、レコード対毎の重要度が反映されたものとなる。このため、本例示的実施形態に係る情報処理装置１によれば、レコード対の同一性をより好適に予測できるという効果が得られる。
【００３２】
＜情報処理方法Ｓ１の流れ＞
本例示的実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。ステップＳ１１において、取得部１１はレコード対を取得する。ステップＳ１２において、類似度算出部１２は、レコー

ド対について、複数の類似度関数を用いて複数の類似度を算出する。ステップＳ１３において、予測部１３は、レコード対と複数の類似度とを参照して、レコード対に応じて定まる重要度を用いてレコード対の同一性予測を行う。ステップＳ１４において、出力部１４は予測部１３による予測結果を出力する。

【００３３】

＜情報処理方法Ｓ１の効果＞

以上のように、本例示的実施形態に係る情報処理方法Ｓ１においては、レコード対について複数の類似度関数を用いて複数の類似度を算出し、レコード対と、複数の類似度とを参照して、レコード対に応じて定まる重要度を用いてレコード対の同一性予測を行う構成が採用されている。このため、本例示的実施形態に係る情報処理方法Ｓ１によれば、レコード対の同一性をより好適に予測できるという効果が得られる。

【００３４】

＜情報処理装置２の構成＞

次いで、本例示的実施形態に係る情報処理装置２の構成について、図３を参照して説明する。図３は、情報処理装置２の構成を示すブロック図である。情報処理装置２は、レコード対の同一性を予測するために用いるパラメータを生成する装置である。情報処理装置２は、取得部２１及びパラメータ生成部２２を備える。

【００３５】

取得部２１は、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得する。同一性に関するラベルは、一例として、レコード対に含まれるレコードが同一であるか否かを示す。

【００３６】

パラメータ生成部２２は、（ｉ）予測対象のレコード対について複数の類似度を算出するための複数の類似度関数ϕ_iの各々が有する１又は複数のパラメータ、及び（ｉｉ）予測対象のレコード対と、複数の類似度とを参照して、予測対象のレコード対に応じて定まる重要度を用いて予測対象のレコード対の同一性予測を行う予測部１３が、重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータ、の少なくとも何れかのパラメータを、訓練データを参照して生成する。

【００３７】

＜情報処理装置２の効果＞

以上のように、本例示的実施形態に係る情報処理装置２においては、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得し、予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び予測対象のレコード対と、複数の類似度とを参照して、予測対象のレコード対に応じて定まる重要度を用いて予測対象のレコード対の同一性予測を行う予測手段が、重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータの少なくとも何れかのパラメータを、訓練データを参照して生成する構成が採用されている。このため、本例示的実施形態に係る情報処理装置２によれば、レコード対の同一性をより好適に予測可能なパラメータを生成できるという効果が得られる。

【００３８】

＜情報処理方法Ｓ２の流れ＞

本例示的実施形態に係る情報処理方法Ｓ２の流れについて、図４を参照して説明する。図４は、情報処理方法Ｓ２の流れを示すフロー図である。ステップＳ２１において、取得部２１は、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得する。ステップＳ２２において、パラメータ生成部２２は、（ｉ）予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び（ｉｉ）予測対象のレコード対と、複数の類似度とを参照して、予測対象のレコード対に応じて定まる重要度を用いて予測対象のレコード対の同一性予測を行う予測手段が、重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータ、の少なくとも何れかのパラメータを、訓練データを参照して生成する

。

【００３９】

＜情報処理方法Ｓ２の効果＞

以上のように、本例示的実施形態に係る情報処理方法Ｓ２においては、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得し、予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び予測対象のレコード対と、複数の類似度とを参照して、予測対象のレコード対に応じて定まる重要度を用いて予測対象のレコード対の同一性予測を行う予測手段が、重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータの少なくとも何れかのパラメータを、訓練データを参照して生成する構成が採用されている。このため、本例示的実施形態に係る情報処理方法Ｓ２によれば、レコード対の同一性をより好適に予測可能なパラメータを生成できるという効果が得られる。

【００４０】

＜製造方法＞

情報処理装置２は、学習済モデルの製造方法を実行する装置として特定することもできる。ここで、学習済モデルの製造方法は、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得することと、複数の類似度算出モデル及び重要度算出モデルの少なくとも何れかのモデルを、訓練データを参照して生成することと、を含む。

【００４１】

〔例示的実施形態２〕

本発明の第２の例示的実施形態について、図面を参照して詳細に説明する。なお、例示的実施形態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。

【００４２】

＜情報処理装置１Ａの構成＞

図５は、本例示的実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、制御部１０Ａ、記憶部２０Ａ、通信部３０Ａ及び入出力部４０Ａを備える。

【００４３】

（通信部３０Ａ）

通信部３０Ａは、情報処理装置１Ａ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実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ＬＡＮ（Local Area Network）、有線ＬＡＮ、ＷＡＮ（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３０Ａは、制御部１０Ａ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１０Ａに供給したりする。

【００４４】

（入出力部４０Ａ）

入出力部４０Ａ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４０Ａは、接続された入力機器から情報処理装置１Ａに対する各種の情報の入力を受け付ける。また、入出力部４０Ａは、制御部１０Ａ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４０Ａとしては、例えば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５に示すように、取得部１１、類似度算出部１２、予測部１３、出力部１４、及び統合部１５Ａを備える。

【００４６】

（取得部１１）

取得部１１は、レコード対に含まれる第１のレコードｅを含む第１のデータｘと、レコード対に含まれる第２のレコードｅ´を含む第２のデータｘ´とを取得する。第１のデータｘ及び第２のデータｘ´は、例えば複数のレコードを含むテーブルである。第１のレコードｅ∈ｘと、第２のレコードｅ´∈ｘ´とは、一例として、以下のように表現される。

$e=(a_1:v_1, a_2:v_2, \cdots, a_d:v_d)$

$e'=(a_1:v_1, a_2:v_2, \cdots, a_{d'}:v_{d'})$

【００４７】

ここで、$a_l \in A_l$（ｌ＝１、２、…ｄ）、及び$a'_m \in A'_m$（ｍ＝１、２、…ｄ´）、は属性名であり、A_l及びA'_mは、例えば文字列空間である。$v_l \in V_l$及び$v'_m \in V'_m$は属性値であり、V_l及びV'_mは、例えば文字列空間又は実数空間である。ｄはレコードｅが有する属性の数であり、ｄ´はレコードｅ´が有する属性の数である。換言すると、第１のレコードｅ及び第２のレコードｅ´はそれぞれ、属性名と属性値とのセットを複数含む。

【００４８】

図６は、第１のデータｘ及び第２のデータｘ´の具体例であるテーブルＴ１及びテーブルＴ２を示す図である。テーブルＴ１及びテーブルＴ２は行と列からなり、行はレコード、列は属性に対応する。換言すると、テーブルＴ１は、複数の第１のレコードe_1、e_2、…を含む。また、テーブルＴ２は、複数の第２のレコードe'_1、e'_2、…を含む。

【００４９】

図６の第１のレコードe_2は、e_2＝（商品名：ポテトチップス，価格：１９８）と表される。第１のレコードe_2において、属性名が「商品名」である属性の属性値は「ポテトチップス」であり、また、属性名が「価格」である属性の属性値は「１９８」である。

【００５０】

テーブルＴ１の属性名及び属性値と、テーブルＴ２の属性名及び属性値とは、同じであってもよく、また、異なっていてもよい。図６の例で、取得部１１が取得するレコード対（ｅ，ｅ´）は、テーブルＴ１に含まれる第１のレコードe_1、e_2、…のいずれかと、テーブルＴ２に含まれる第２のレコードe'_1、e'_2、…のいずれかとの対である。

【００５１】

（類似度算出部１２）

類似度算出部１２は、ｋ個（ｋは２以上の整数）の類似度関数ϕ_i（$1 \leqq i \leqq k$）を用いて、１つのレコード対（ｅ，ｅ´）についてｋ個の類似度s_iを算出する。類似度算出部１２がｋ個の類似度s_iを算出する処理の詳細については後述する。

【００５２】

（予測部１３）

予測部１３は、レコード対（ｅ，ｅ´）と、複数の類似度s_iとを参照して、レコード対（ｅ，ｅ´）に応じて定まる重要度を用いてレコード対の同一性予測を行う。本例示的実施形態において、予測部１３は、レコード対（ｅ，ｅ´）を参照して重要度を算出する重要度算出部１３１Ａを備えている。予測部１３が行う同一性の予測処理、及び重要度算出部１３１Ａが行う重要度の算出処理の詳細については後述する。

【００５３】

（出力部１４）

出力部１４は、予測部１３による予測結果を出力する。予測結果は、一例として、レコード対に含まれるレコード同士が同一であるか否かを示す情報を含む。また、予測結果は、レコード対に含まれるレコード同士の類似の程度を示す情報を含んでもよい。出力部１４は、予測結果を記憶部２０Ａ又は外部記憶装置に書き込むことにより出力してもよく、また、入出力部４０Ａに接続された出力装置（表示装置、印刷装置、等）に出力してもよい。また、出力部１４は、通信部３０Ａを介して予測結果を他の装置に送信することにより予測結果を出力してもよい。

【００５４】

（統合部１５Ａ）

統合部１５Ａは、出力部１４が出力する予測結果を参照して、第１のデータと第２のデータとから、統合済データを生成する。統合部１５Ａが行う統合済データの生成処理の詳細については後述する。

【００５５】

（記憶部２０Ａ）

記憶部２０Ａには、取得部１１が取得する第１のデータｘ及び第２のデータｘ´が記憶されるとともに、予測部１３の予測結果ＰＲが記憶される。また、記憶部２０Ａには、複数の類似度関数φ$_i$、重要度算出モデルｇ、及びパラメータＰが記憶される。

【００５６】

類似度関数｛φ$_1$、…、φ$_k$｝は、上述の例示的実施形態１で示したように、例えばJaccard係数、又は非特許文献１若しくは非特許文献２に記載された手法により類似度を算出する関数である。類似度関数φ$_i$は、一例として、情報処理装置１Ａのユーザ等により入力される。類似度関数φ$_i$は、一例として、レコード対（ｅ，ｅ´）に対して０から１の数値を類似度として出力する。この場合、例えば、出力値が１に近いほど類似性が高く、０に近いほど類似性が低い。類似度関数φ$_i$は、一例として、学習可能なパラメータを備えた関数である。

【００５７】

重要度算出モデルｇは、重要度算出部１３１Ａが重要度を算出するために用いるモデルである。重要度算出モデルｇは、上述の例示的実施形態１で示したように、例えばＢＥＲＴ、fastText、word2vec、tf-idf、BM25、等の言語モデルを用いて生成される。また、重要度算出モデルｇは言語モデルを含んでもよい。

【００５８】

記憶部２０Ａに記憶されたパラメータＰは、ｋ個の類似度関数φ$_i$の各々が有する１又は複数のパラメータθ$_i$、及び、重要度算出モデルｇが有する１又は複数のパラメータｗ、の少なくとも何れかのパラメータを含む。

【００５９】

＜情報処理方法Ｓ１Ａの流れ＞

図７は、情報処理装置１Ａが実行する情報処理方法の一例である情報処理方法Ｓ１Ａの流れを示すフロー図である。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。

【００６０】

（ステップＳ１０１）

ステップＳ１０１において、取得部１１は、第１のデータ及び第２のデータを取得する。取得部１１は、一例として、情報処理装置１Ａのユーザ等が入出力部４０Ａに接続された入力装置を用いて入力した第１のデータ及び第２のデータを取得する。また、取得部１１は、通信部３０Ａを介して他の装置から第１のデータ及び第２のデータを受信することにより、第１のデータ及び第２のデータを取得してもよい。また、取得部１１は、外部接続された記憶装置から第１のデータ及び第２のデータを読み出すことにより、第１のデータ及び第２のデータを取得してもよい。取得部１１は、取得した第１のデータ及び第２のデータを記憶部２０Ａに記憶する。

【００６１】

（ステップＳ１０２）

ステップＳ１０２において、取得部１１は、記憶部２０Ａに記憶されたパラメータＰを取得する。

【００６２】

（ステップＳ１０３）

ステップＳ１０３において、取得部１１は、予測対象であるレコード対（ｅ，ｅ´）を取得する。

【００６３】

（ステップＳ１０４）

ステップＳ１０４において、類似度算出部１２は、ｋ個の類似度関数φ$_i$を用いて、レコード対（ｅ，ｅ´）についてｋ個の類似度ｓ$_i$を算出する。ｋ個の類似度関数φ$_i$がそれぞれ異なるため、算出されるｋ個の類似度ｓ$_i$もそれぞれ異なった値となり得る。例えば、「アイス」と「あいす」のレコード対の場合、表記変更を行って算出される類似度ｓ$_i$は、類似性が高いことを示す値となる一方、部分文字列を抽出して算出される類似度ｓ$_i$は、類似性が低いことを示す値となる。また、「ポテトチップス」と「ポテチ」のレコード対の場合、表記変更を行って算出される類似度ｓ$_i$は、類似性が低いことを示す値となる一方、部分文字列を抽出して算出される類似度ｓ$_i$は、類似性が高いことを示す値となる。

【００６４】

（ステップＳ１０５）

ステップＳ１０５において、重要度算出部１３１Ａは、レコード対（ｅ，ｅ´）を参照して、複数の類似度ｓ$_i$の各々に関する重要度ｇ$_i$を算出する。重要度算出部１３１Ａは、一例として、重要度算出モデルｇを用いて重要度ｇ$_i$を算出する。

【００６５】

重要度算出モデルｇは、複数の類似度ｓ$_i$のそれぞれについて重要度ｇ$_i$を算出するためのモデルである。重要度算出モデルｇは、一例として、

【数１】

$$g: \mathcal{X} \times \mathcal{X}' \to [0,1]^k, \ \sum_{i=1}^{k} \{g(e, e')\}_i = 1$$

と表される。換言すると、重要度算出モデルｇにより算出されるｋ個の重要度｛ｇ（ｅ，ｅ´）｝$_i$の総和は１である。

【００６６】

重要度算出部１３１Ａが行う重要度ｇ$_i$の算出処理の具体例について説明する。まず、重要度算出部１３１Ａは、言語モデルで第１のレコードｅ及び第２のレコードｅ´の各属性値の文字列をベクトルに変換する。具体的には、例えば、重要度算出部１３１Ａは、レコード対（ｅ＝（商品名：ポテトチップス，価格：１９８）、ｅ´＝（商品名：ポテチ，評価：５））を、レコード対（ｅ，ｅ´）を文字列に変換する関数serialize（ｅ，ｅ´）により、「[CLS][COL]商品名[VAL]ポテトチップス[COL]価格[VAL]１９８[SEP][COL]商品名[VAL]ポテチ[COL]評価[VAL]５[SEP]」という文字列に変換する。ここで、[CLS]と[COL]と[VAL]と[SEP]は、それぞれ文章の始まりと、属性名と、属性値と、レコードの区切りを示す記号である。

【００６７】

更に、重要度算出部１３１Ａは、生成した文字列を言語モデル（例えば、ＢＥＲＴ）によりベクトルに変換する。続いて、重要度算出部１３１Ａは、言語モデルにより得られたベクトルに対し、連結、和、深層学習等を適用することで、新たなＬ次元ベクトルｚに変換する。

【００６８】

更に、重要度算出部１３１Ａは、変換したＬ次元ベクトルｚを、ｋクラス分類器に入力することで、ｋ個の重要度｛ｇ（ｅ，ｅ´）｝$_i$を算出する。ｋクラス分類器としては、例えば線形分類器、深層学習等の技術が用いられる。ｋクラス分類器として、例えば文献「Robert A. Jacobs, Michael Jordan, Geoffrey Hinton: Adaptive Mixtures of Local Experts, Neural Computation 3, 79-87 (1991)」の文献に記載された技術、又は、「Noam Shazeer, Quoc Le, Geoffrey Hinton: Jeffrey Dean: OUTRAGEOUSLY LARGE NEURAL NETWORKS: THE SPARSELY-GATED MIXTURE-OF-EXPERTS LAYER, ICLR 2017」の文献に記載された技術が用いられてもよい。

【００６９】

例えば、ｉ＝１，…，ｋにおいて、Ｌ次元ベクトルｗ$_{i}$に対し、線形ソフトマックス関数のｉ次元の出力である重要度｛ｇ（ｅ，ｅ´）｝$_{i}$は、
ｅｘｐ（ｗ$_{i}$＾Ｔ・ｚ）／（ｅｘｐ（ｗ$_{1}$＾Ｔ・ｚ）＋ｅｘｐ（ｗ$_{2}$＾Ｔ・ｚ）＋…＋ｅｘｐ（ｗ$_{k}$＾Ｔ・ｚ））
により算出される。ここで、Ｌ次元ベクトルｗ$_{i}$は、重要度算出モデルｇの学習可能なパラメータｗの一例である。また、「ｗ$_{i}$＾Ｔ・ｚ」はＬ次元ベクトルｗ$_{i}$とＬ次元ベクトルｚの内積である。

【００７０】

（ステップＳ１０６）

ステップＳ１０６において、予測部１３は、類似度算出部１２が計算した類似度ｓ$_{i}$とレコード対（ｅ、ｅ´）とを用いて、レコード対（ｅ、ｅ´）の同一性を予測する。予測部１３は、一例として、ｋ個の類似度ｓ$_{i}$を用いてレコード対（ｅ，ｅ´）に含まれるレコード同士の類似度を算出し、算出した類似度が閾値ｑ（例えば、ｑ＝０．５）より大きい場合に、レコードｅとレコードｅ´とが同一であると予測し、算出した確率が閾値ｑ以下である場合に同一でないと予測する。

【００７１】

予測部１３が算出する確率は、レコード対（ｅ，ｅ´）についてｋ個の類似度ｓ$_{i}$を統合し予測した結果を示すものであり、一例として、０～１の数値である。本例示的実施形態において、予測部１３は、レコード対（ｅ，ｅ´）と類似度ｓ$_{i}$とを入力とする確率関数ｈにより、確率を算出する。確率関数ｈは、一例として、ｋ個の類似度ｓ$_{i}$＝φ$_{i}$（ｅ，ｅ´）を用いて以下の（数式１）で表される。

【数２】

$$h\left(e,e' \mid \{\phi_i\}_{i=1}^{k}\right) = \sum_{i=1}^{k} \{g(e,e')\}_i \phi_i(e,e') \quad \cdots (\text{数式}1)$$

【００７２】

上述の（数式１）において、重要度｛ｇ（ｅ，ｅ´）｝$_{i}$は重要度算出部１３１Ａが算出する重要度であり、類似度ｓ$_{i}$＝φ$_{i}$（ｅ，ｅ´）は、類似度関数φ$_{i}$によりレコード対（ｅ，ｅ´）について算出された類似度である。（数式１）を用いる場合、換言すると、予測部１３は、複数の類似度ｓ$_{i}$に関する線形和であって、各重要度｛ｇ（ｅ，ｅ´）｝$_{i}$を重み係数とする線形和を用いて、同一性予測を行う。

【００７３】

本例示的実施形態では、異なる複数のレコード対（ｅ，ｅ´）のそれぞれについて算出されたｋ個の類似度ｓ$_{i}$が同じであっても、重要度｛ｇ（ｅ，ｅ´）｝$_{i}$はレコード対のそれぞれで異なり得る。換言すると、予測部１３による予測結果には、類似度ｓ$_{i}$だけでなく、レコード対により定まる重要度ｇ$_{i}$が反映される。このように、予測部１３が同一性を予測する手法はレコード対によって異なり得る。

【００７４】

（ステップＳ１０７）

ステップＳ１０７において、出力部１４は、予測部１３の予測結果を出力する。一例として、出力部１４は、予測結果を記憶部２０Ａに記憶する。

【００７５】

（ステップＳ１０８）

ステップＳ１０８において、予測部１３は、予測対象である全てのレコード対（ｅ，ｅ´）について同一性の予測を行ったかを判定する。予測対象である全てのレコード対（ｅ，ｅ´）について予測処理が完了した場合（ステップＳ１０８；ＹＥＳ）、予測部１３はステップＳ１０９の処理に進む。一方、また予測対象であるレコード対（ｅ，ｅ´）が残っている場合（ステップＳ１０８；ＮＯ）、予測部１３はステップＳ１０３の処理に戻り

、次のレコード対（ｅ，ｅ´）について同一性の予測を行う。すなわち、情報処理装置１Ａは、予測対象である全てのレコード対（ｅ，ｅ´）について、ステップＳ１０３～Ｓ１０７処理を実行する。

【００７６】

（ステップＳ１０９）

ステップＳ１０９において、統合部１５Ａは、出力部１４が出力する予測結果を参照して、第１のデータと第２のデータとから、統合済データを生成する。統合済データは、一例として、統合部１５Ａは、予測部１３が同一であると予測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。

【００７７】

図８は、統合済データの一例であるテーブルＴ３を示す図である。テーブルＴ３は、複数のレコードf_1、f_2、…を含む。レコードf_1は、図６の第１のレコードe_1と第２のレコードe'_2を統合したレコードである。レコードf_2は、図６の第１のレコードe_2と第２のレコードe'_3を統合したレコードである。レコードf_3は、図６の第１のレコードe_3と第２のレコードe'_1を統合したレコードである。

【００７８】

＜実施例＞

次に、本例示的実施形態の具体的な実施例を説明する。この例で、類似度関数$\{\phi_i\}$として、類似度関数ϕ_1～ϕ_3を用いる。類似度関数ϕ_1は、レコード対の商品名のJaccard係数を算出する関数である。類似度関数ϕ_2は、レコード対の商品名がひらがなであればカタカナに変換してからJaccard係数を算出する関数である。類似度関数ϕ_3は、上述の非特許文献２に記載された手法により類似度を算出する関数である。ここで、類似度関数ϕ_3は学習可能なパラメータθ_3を持つ。

【００７９】

（ステップＳ１０１）

図７のステップＳ１０１において、取得部１１は、同一性が未知のレコード対の集合であるテストデータD_{test}＝｛（（商品名：しょうゆせんべい，価格：268），（商品名：ショウユセンベイ，評価：４））、…、（（商品名：ヨモギ団子，価格：190）、（商品名：みたらしだんご，評価：３））｝を取得する。

【００８０】

（ステップＳ１０２～Ｓ１０４）

類似度算出部１２は、類似度Ｓ＝（s_1，s_2，s_3）を算出する。ここで、類似度算出部１２は、パラメータθ_3を記憶部２０Ａから読み取り、読み取ったパラメータθ_3を用いて類似度s_3を算出する。具体的には、類似度算出部１２は、テストデータD_{test}のレコード対｛ｅ＝（商品名：しょうゆせんべい，価格：２６８），ｅ´＝（商品名：ショウユセンベイ，評価：４）｝の類似度Ｓ＝（ϕ_1（ｅ，ｅ´），ϕ_2（ｅ，ｅ´），ϕ_3（ｅ，ｅ´））＾Ｔ＝（０，１，０．７）＾Ｔを計算する。

【００８１】

（ステップＳ１０５）

予測部１３は、レコード対（ｅ，ｅ´）の属性名と属性値を連結する関数serialize（ｅ，ｅ´）を用いて、レコード対（ｅ，ｅ´）から文字列「[CLS][COL]商品名[VAL]しょうゆせんべい[COL]価格[VAL]268[SEP][COL]商品名[VAL]ショウユセンベイ[COL]評価[VAL]４[SEP]」を作成する。また、予測部１３は、事前学習済み言語モデルであるＢＥＲＴによりこの文字列のベクトル表現であるＬ次元ベクトルｖを得る。更に、予測部１３は、線形ソフトマックス関数を用いてｉ＝１，２，３に対し類似度関数ϕ_iの重みである重要度g_iを、

$g_i = e^{(w_i^T \cdot v)} / (e^{(w_1^T \cdot v)} + e^{(w_2^T \cdot v)} + e^{(w_3^T \cdot v)})$

と計算し、（g_1，g_2，g_3）＝（０．１，０．６，０．３）を得る。ここで、w_1、w_2及びw_3は実数ベクトルであり、重要度算出モデルｇの学習可能なパラメータｗの一

例である。

【００８２】

（ステップＳ１０６）

ステップＳ１０６において、予測部１３は、類似度算出部１２が計算した類似度Ｓに重要度$_i$をかけた和を確率として算出する。類似度Ｓ＝（０，１，０．７）^{T}であり、重要度（g_1，g_2，g_3）＝（０．１，０．６，０．３）であるから、

ｈ（ｅ，ｅ´）＝０．１×０＋０．６×１＋０．３×０．７＝０．８１

となる。算出された値「０．８１」が予め定められた閾値ｑ＝０．５よりも大きいため、予測部１３は、レコード対（ｅ，ｅ´）に含まれるレコードｅとレコードｅ´とが同一であると予測する。

【００８３】

（ステップＳ１０７）

ステップＳ１０７において、出力部１４がレコード対（ｅ、ｅ´）の同一性予測結果を出力する。以上の同一性予測と出力をテストデータD_{test}の全てのレコード対に対して適用する。

【００８４】

＜情報処理装置１Ａの効果＞

以上のように、本例示的実施形態に係る情報処理装置１Ａにおいては、レコード対（ｅ，ｅ´）を参照して重要度ｇ$_i$を算出し、算出した重要度ｇ$_i$を用いてレコード対の同一性予測を行う構成が採用されている。このため、本例示的実施形態に係る情報処理装置１Ａによれば、例示的実施形態１に係る情報処理装置１の奏する効果に加えて、レコード対（ｅ，ｅ´）を用いて算出される重要度ｇ$_i$を加味した同一性予測を行うことができ、レコード対（ｅ，ｅ´）の同一性をより適切に予測できるという効果が得られる。

【００８５】

＜変形例＞

上述の例示的実施形態において、取得部１１は補助データｕを更に取得し、予測部１３は、レコード対（ｅ，ｅ´）と、複数の類似度ｓ$_i$と、補助データｕとを参照して、レコード対（ｅ，ｅ´）と補助データｕとに応じて定まる重要度ｇ$_i$を用いてレコード対（ｅ，ｅ´）の同一性予測を行ってもよい。

【００８６】

補助データｕは、一例として、レコードの名前、レコードの特徴量、及び／又はレコードの分類結果（お菓子、人名、等）、を示す情報を含む。ここで、補助データｕは、一例として、Ｗｉｋｉｐｅｄｉａ（登録商標）等の外部データから得られるレコードに関する情報を含んでもよい。また、補助データｕは、一例として、類似度関数φ$_i$のパラメータθ及び／又は重要度算出モデルｇのパラメータｗの学習で用いられた訓練データの数を含んでもよい。ただし、補助データｕは上述した例に限られず、他の情報を含んでもよい。補助データｕは、一例として、離散的な情報を表すワンホットベクトルである。

【００８７】

この場合、重要度算出モデルｇには、レコード対（ｅ，ｅ´）に加えて補助データｕが入力される。一例として、ベクトルである補助データｕは、上述のＬ次元ベクトルｚに連結され、連結されたベクトルとパラメータｗを用いて重要度ｇ$_i$が算出される。

【００８８】

本変形例では、予測部１３は、レコード対（ｅ，ｅ´）と、複数の類似度ｓ$_i$と、補助データｕとを参照して、レコード対（ｅ，ｅ´）と補助データｕとに応じて定まる重要度ｇ$_i$を用いてレコード対（ｅ，ｅ´）の同一性予測を行う。これにより、予測部１３はレコード対（ｅ，ｅ´）の同一性の予測精度をより高くすることができる。

【００８９】

〔例示的実施形態３〕

本発明の第３の例示的実施形態について、図面を参照して詳細に説明する。なお、例示的実施形態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。

【００９０】

＜情報処理装置１Ｂの構成＞

図９は、本例示的実施形態に係る情報処理装置１Ｂの構成を示すブロック図である。情報処理装置１Ｂの制御部１０Ａは、取得部１１、類似度算出部１２、予測部１３、出力部１４、統合部１５Ａに加えて、学習部１６Ｂを備える。

【００９１】

本例示的実施形態に係る取得部１１は、レコード対（ｅｊ，ｅ´ｊ）と、当該レコード対（ｅｊ，ｅ´ｊ）の同一性に関するラベルｙｊとの組を複数含む訓練データＤｔｒを更に取得する。訓練データＤｔｒは、上述のパラメータＰを学習するために用いられる。訓練データＤｔｒは、一例として、

【数３】

$$\mathcal{D}_{tr} = \{(e_j, e'_j,\ y_j)\}_{j=1}^{n}$$

と表現される。ここで、ｎは、レコード対（ｅｊ，ｅ´ｊ）の総数である。ラベルｙｊは、一例として、「０」又は「１」である。「１」は、第１のレコードｅｊと第２のレコードｅ´ｊとが同一である旨を示し、「０」は、第１のレコードｅｊと第２のレコードｅ´ｊとが同一でない旨を示す。

【００９２】

学習部１６Ｂは、（ｉ）類似度算出部１２が類似度ｓｉを算出するために用いる複数の類似度関数φｉの各々が有する１又は複数のパラメータθｉ、及び（ｉｉ）重要度算出部１３１Ａが重要度を算出するために用いる重要度算出モデルｇが有する１又は複数のパラメータｗ、の少なくとも何れかのパラメータＰを、前記訓練データを参照して生成する。学習部１６Ｂは、本明細書に係る「パラメータ生成手段」の一例である。

【００９３】

＜情報処理方法Ｓ２Ｂの流れ＞

図１０は、情報処理装置１Ｂが実行する情報処理方法の一例である情報処理方法Ｓ２Ｂの流れを示すフロー図である。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。

【００９４】

（ステップＳ２０１・Ｓ２０２）

ステップＳ２０１において、取得部１１は、訓練データＤｔｒを取得する。訓練データＤｔｒは、一例として、情報処理装置１Ｂのユーザにより入力される。また、ステップＳ２０２において、取得部１１は、複数の類似度関数φｉを取得する。類似度関数φｉは、一例として、情報処理装置１Ｂのユーザにより入力される。

【００９５】

（ステップＳ２０３）

ステップＳ２０３において、学習部１６Ｂは、訓練データＤｔｒを用いて、パラメータθｉ及びパラメータｗの少なくとも何れかを学習する。ここで、パラメータθｉは、類似度関数φｉが有するパラメータの集合である。また、パラメータｗは、重要度算出モデルｇが有するパラメータの集合である。

【００９６】

学習部１６Ｂは、一例として、目的関数Ｌによりパラメータθｉとパラメータｗとを最適化する。この最適化は、一例として、

【数４】

$$\min_{w,\ \{\theta_i\}_{i=1}^{k}} \frac{1}{n}\sum_{j=1}^{n} l\left(h_w\left(e_j,\ e_j' \mid \{\phi_{i,\theta_i}\}_{i=1}^{k}\right),\ y_j\right) + \alpha\Omega\left(w, \{\theta_i\}_{i=1}^{k}\right)$$

と表される。ここで、評価指標ｌは、

【数５】

$$l: [0,1] \times \{0,1\} \to [0,\infty]$$

である。すなわち、評価指標ｌは、

訓練データＤｔｒのレコード対（ｅｊ，ｅ´ｊ）に含まれるレコード同士が同一である確率（確率関数ｈｗの出力）と、

「０」又は「１」のラベルｙｊと、

を入力とし、０以上の値を出力する損失関数である。評価指標ｌとしては、例えばクロスエントロピー誤差を用いることができる。

【００９７】

また、目的関数Ｌにおいて、αは非負値のハイパーパラメータである。ハイパーパラメータαは、情報処理装置１Ｂのユーザ等が定めてもよいし、訓練データＤｔｒとは別の同一性が既知のレコード対の集合を用いて自動的に決定された値であってもよい。Ωはパラメータに対する正則化項であり、Ｌ２ノルムを用いてもよい。上の式においてパラメータθｉを固定してパラメータｗのみを最適化してもよい。

【００９８】

学習部１６Ｂは、生成したパラメータｗ及びパラメータθｉを記憶部２０Ａに保存する。学習部１６Ｂが生成したパラメータｗ及びパラメータθｉは、類似度算出部１２による類似度ｓｉの算出処理、及び／又は予測部１３による同一性の予測処理において用いられる。

【００９９】

＜実施例＞

次に、本例示的実施形態の具体的な実施例について説明する。例えば、テーブルＴ１の第１のレコードｅ１＝（商品名：ポテトチップス，価格：１９８）、第１のレコードｅ２＝（商品名：アイス，価格：１４８）と、テーブルＴ２の第２のレコードｅ´１＝（商品名：ポテチ，評価：５）、第２のレコードｅ´２＝（商品名：あいす，評価：４）について、訓練データＤｔｒを、

Ｄｔｒ＝｛（ｅ１，ｅ´１，１），（ｅ２，ｅ´２，１），（ｅ１，ｅ´２，０），（ｅ２，ｅ´１，０）｝

とする。

【０１００】

また、類似度関数｛φｉ｝として、類似度関数φ１～φ３を用いる。類似度関数φ１～φ３は、上述の例示的実施形態１の実施例で示した類似度関数φ１～φ３と同様である。類似度関数φ３は学習可能なパラメータθ３を有する。

【０１０１】

ステップＳ２０１において、取得部１１は訓練データＤｔｒを取得する。また、ステップＳ２０３において、学習部１６Ｂは、予測部１３による訓練データＤｔｒのレコード対（ｅｊ，ｅ´ｊ）の同一性予測がよく正解するように、クロスエントロピー誤差に基づいて、重要度算出モデルｇのパラメータｗと類似度関数φｉのパラメータθｉを、確率的勾配降下法を用いて最適化する。最適化されたパラメータｗとパラメータθｉとは、記憶部２０Ａに保存される。

【０１０２】

＜情報処理装置１Ｂの効果＞

以上のように、本例示的実施形態に係る情報処理装置１Ｂにおいては、重要度算出モデルｇが有するパラメータｗ及び類似度関数φ$_i$が有するパラメータθ$_i$の少なくとも何れかのパラメータを、訓練データＤ$_{tr}$を参照して生成する構成が採用されている。このため、本例示的実施形態に係る情報処理装置１Ｂによれば、例示的実施形態１に係る情報処理装置１の奏する効果に加えて、レコード対の同一性をより好適に予測可能なパラメータを生成できるという効果が得られる。

【０１０３】

＜変形例＞

上述の例示的実施形態において、訓練データＤ$_{tr}$は、補助データｕを含んでいてもよい。この場合、訓練データＤ$_{tr}$は、一例として、

【数６】

$$\mathcal{D}_{tr} = \{(e_j, e'_j, u_j, y_j)\}_{j=1}^{n}$$

と表される。学習部１６Ｂは、補助データｕを含む訓練データＤ$_{tr}$を用いてパラメータｗとパラメータθ$_i$とを最適化する。

【０１０４】

〔例示的実施形態４〕

本発明の第４の例示的実施形態について、図面を参照して詳細に説明する。なお、例示的実施形態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。

【０１０５】

＜情報処理装置１Ｃの構成＞

図１１は、本例示的実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１０Ａは、取得部１１、類似度算出部１２、予測部１３、出力部１４、学習部１６Ｂに加えて、検索結果出力部１７Ｃを備える。

【０１０６】

本例示的実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４０Ａ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。

【０１０７】

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。

【０１０８】

予測部１３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。検索結果出力部１７Ｃは、出力部１４が出力する各々の予測結果ＰＲ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１７Ｃは、一例として、入出力部４０Ａ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１７Ｃは、通信部３０Ａ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１７Ｃは、検索結果を記憶部２０Ａ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。

【０１０９】

図１２は、検索結果出力部１７Ｃが出力する画面表示の具体例を示す図である。図１２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、上述の例示的実施形態１において図６に示したテーブルＴ１及びテーブルＴ２で

ある。予測部１３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予測を行う。予測部１３が行う同一性の予測処理は、上述の例示的実施形態２で説明したため、その説明を繰り返さない。

【０１１０】

図１２の例において、検索結果出力部１７Ｃは、予測部１３の予測結果ＰＲ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から検索された検索結果である。

【０１１１】

＜情報処理装置１Ｃの効果＞

以上のように、本例示的実施形態に係る情報処理装置１Ｃにおいては、出力部１４が出力する各々の予測結果を参照して、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、本例示的実施形態に係る情報処理装置１Ｃによれば、例示的実施形態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。

【０１１２】

情報処理装置１Ｃ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、

前記レコード対について、複数の類似度関数を用いて複数の類似度を算出する類似度算出手段と、

前記入力データと、前記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、前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行う予測手段と、

前記予測手段による予測結果を参照して、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。

【０１１３】

〔ソフトウェアによる実現例〕

情報処理装置１、１Ａ、１Ｂ、１Ｃ、２（以下「情報処理装置１等」という）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。

【０１１４】

後者の場合、情報処理装置１等は、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１３に示す。コンピュータＣは、少なくとも１つのプロセッサＣ１と、少なくとも１つのメモリＣ２と、を備えている。メモリＣ２には、コンピュータＣを情報処理装置１等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Ｃ１は、プログラムＰをメモリＣ２から読み取って実行することにより、情報処理装置１等の各機能が実現される。

【０１１５】

プロセッサＣ１としては、例えば、ＣＰＵ（Central Processing Unit）、ＧＰＵ（Graphic Processing Unit）、ＤＳＰ（Digital Signal Processor）、ＭＰＵ（Micro Processing Unit）、ＦＰＵ（Floating point number Processing Unit）、ＰＰＵ（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Ｃ２としては、例えば、フラッシュメモリ、ＨＤＤ（Hard Disk Drive）、ＳＳＤ（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。
【０１１６】
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ＲＡＭ（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。
【０１１７】
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。
【０１１８】
〔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。
【０１１９】
〔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。
【０１２０】
（付記１）
レコード対を取得する取得手段と、
前記レコード対について、複数の類似度関数を用いて複数の類似度を算出する類似度算出手段と、
前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行う予測手段と、
前記予測手段による予測結果を出力する出力手段と、
を備えている情報処理装置。
【０１２１】
上記の構成によれば、レコード対の同一性をより好適に予測できる。
【０１２２】
（付記２）
前記取得手段は、補助データを更に取得し、
前記予測手段は、前記レコード対と、前記複数の類似度と、前記補助データとを参照して、前記レコード対と前記補助データとに応じて定まる重要度を用いて前記レコード対の同一性予測を行う、
付記１に記載の情報処理装置。
【０１２３】
上記の構成によれば、重要度はレコード対だけでなく補助データの内容を反映した情報となる。このような重要度を用いてレコード対の同一性を予測することにより、レコード対の同一性の予測精度をより高くすることができる。
【０１２４】
（付記３）
前記予測手段は、前記レコード対を参照して前記重要度を算出する重要度算出手段を備

えている、
付記１又は２に記載の情報処理装置。
【０１２５】
上記の構成によれば、レコード対を参照して算出される重要度を用いてレコード対の同一性予測を行うことにより、レコード対の同一性の予測精度をより高くすることができる。
【０１２６】
（付記４）
前記重要度算出手段は、前記複数の類似度の各々に関する重要度を算出し、
前記予測手段は、前記複数の類似度に関する線形和であって、前記各重要度を重み係数とする線形和を用いて、前記同一性予測を行う、
付記３に記載の情報処理装置。
【０１２７】
上記の構成によれば、重要度を重み係数とする類似度の線形和を用いて同一性予測を行うことにより、レコード対の同一性の予測精度を高くすることができる。
【０１２８】
（付記５）
前記取得手段は、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを更に取得し、
当該情報処理装置は、
前記類似度算出手段が前記類似度を算出するために用いる前記複数の類似度関数の各々が有する１又は複数のパラメータ、及び、
前記重要度算出手段が前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータ、
の少なくとも何れかのパラメータを、前記訓練データを参照して生成するパラメータ生成手段を更に備えている、
付記３又は４に記載の情報処理装置。
【０１２９】
上記の構成によれば、訓練データを参照して生成したパラメータを用いることで、レコード対の同一性をより好適に予測することができる。
【０１３０】
（付記６）
前記取得手段は、前記レコード対に含まれる第１のレコードを含む第１のデータと、前記レコード対に含まれる第２のレコードを含む第２のデータとを取得し、
当該情報処理装置は、前記出力手段が出力する前記予測結果を参照して、前記第１のデータと前記第２のデータとから、統合済データを生成する統合手段を備えている、
付記１から５の何れか１つに記載の情報処理装置。
【０１３１】
上記の構成によれば、第１のデータと第２のデータとをより好適に統合することができる。
【０１３２】
（付記７）
前記取得手段は、
前記レコード対に含まれる第１のレコードとして、ユーザからの入力データを取得し、
前記レコード対に含まれる第２のレコードとして、対象データに含まれる複数のレコードの１つを取得し、
前記予測手段は、前記第１のレコードと、前記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前記同一性予測を行い、
当該情報処理装置は、前記出力手段が出力する各々の前記予測結果を参照して、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検索結果出力手段を備えている、
付記１から５の何れか１つに記載の情報処理装置。
【０１３３】
上記の構成によれば、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。
【０１３４】
（付記８）
レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得する取得手段と、
予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び、
前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータ、
の少なくとも何れかのパラメータを、前記訓練データを参照して生成するパラメータ生成手段と、
を備えている情報処理装置。
【０１３５】
上記の構成によれば、レコード対の同一性をより好適に予測可能なパラメータを生成できる。
【０１３６】
（付記９）
レコード対を取得することと、
前記レコード対について、複数の類似度関数を用いて複数の類似度を算出することと、
前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行うことと、
前記レコード対の同一性予測による予測結果を出力することと、
を含む情報処理方法。
【０１３７】
上記の情報処理方法によれば、上述した情報処理装置と同様の効果を奏する。
【０１３８】
（付記１０）
レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得することと、
予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び、
前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータ
の少なくとも何れかのパラメータを、前記訓練データを参照して生成することと、
を含む情報処理方法。
【０１３９】
上記の情報処理方法によれば、上述した情報処理装置と同様の効果を奏する。
【０１４０】
（付記１１）
レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得することと、

予測対象のレコード対について複数の類似度を算出するための複数の類似度算出モデル、及び、
前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデル、
の少なくとも何れかのモデルを、前記訓練データを参照して生成することと、
を含む学習済モデルの製造方法。
【０１４１】
上記の構成によれば、レコード対の同一性をより好適に予測可能なモデルを製造することができる。
【０１４２】
（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、
前記レコード対について、複数の類似度関数を用いて複数の類似度を算出する類似度算出処理と、
前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行う予測処理と、
前記予測処理による予測結果を出力する出力処理と、
を実行させるプログラム。
【０１４３】
上記の構成によれば、上述した情報処理装置と同様の効果を奏する。
【０１４４】
（付記１３）
コンピュータに、
レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得する取得処理と、
予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び、
前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータ
の少なくとも何れかのパラメータを、前記訓練データを参照して生成するパラメータ生成処理と、
を実行させるプログラム。
【０１４５】
上記の構成によれば、上述した情報処理装置と同様の効果を奏する。
【０１４６】
〔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。
【０１４７】
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、前記レコード対について、複数の類似度関数を用いて複数の類似度を算出する類似度算出処理と、前記レコード対と、前記複数の類似度とを参照して、前記レコード対に応じて定まる重要度を用いて前記レコード対の同一性予測を行う予測処理と、前記予測処理による予測結果を出力する出力処理とを実行する情報処理装置。
【０１４８】
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、前記類似度算出処理と、前記予測処理と、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。

【０１４９】

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。

少なくとも１つのプロセッサを備え、前記プロセッサは、レコード対と、当該レコード対の同一性に関するラベルとの組を複数含む訓練データを取得する取得処理と、予測対象のレコード対について複数の類似度を算出するための複数の類似度関数の各々が有する１又は複数のパラメータ、及び前記予測対象のレコード対と、前記複数の類似度とを参照して、前記予測対象のレコード対に応じて定まる重要度を用いて前記予測対象のレコード対の同一性予測を行う予測手段が、前記重要度を算出するために用いる重要度算出モデルが有する１又は複数のパラメータの少なくとも何れかのパラメータを、前記訓練データを参照して生成するパラメータ生成処理とを実行する情報処理装置。

【０１５０】

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、前記パラメータ生成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。

【符号の説明】

【０１５１】

１、１Ａ、１Ｂ、１Ｃ、２　情報処理装置
１０Ａ　制御部
１１、２１　取得部
１２　類似度算出部
１３　予測部
１４　出力部
１５Ａ　統合部
１６Ｂ　学習部
１７Ｃ　検索結果出力部
２０Ａ　記憶部
２２　パラメータ生成部
３０Ａ　通信部
４０Ａ　入出力部
１３１Ａ　重要度算出部
Ｓ１、Ｓ１Ａ、Ｓ２、Ｓ２Ｂ　情報処理方法