【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5年6月6日(2023.6.6)

【国際公開番号】WO2020/243499
【公表番号】特表2022-535792(P2022-535792A)
【公表日】令和4年8月10日(2022.8.10)
【年通号数】公開公報(特許)2022-146
【出願番号】特願2021-571432(P2021-571432)
【国際特許分類】
Ｇ０６Ｆ 16/26(2019.01)
【ＦＩ】
Ｇ０６Ｆ 16/26

【手続補正書】
【提出日】令和5年5月29日(2023.5.29)
【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】
１つ又は複数のデータセットに含まれるフィールドのデータ値の意味論的意味を発見するための、データ処理システムによって実行される方法であって、前記方法は、
前記１つ又は複数のデータセットに含まれるフィールドを識別することであって、前記フィールドは、識別子に関連付けられる、識別することと、
前記フィールドの１つ又は複数のデータ値を識別することであって、少なくとも前記データ値のうちの第1のデータ値は、前記データ値のうちの第2のデータ値と異なる、識別することと、
前記フィールドの前記１つ又は複数のデータ値に基づいて、前記フィールドの前記１つ又は複数のデータ値の１つ又は複数の属性を特定することと、
複数のテストにアクセスすることであって、テストは、１つ又は複数の所定の属性に対して提案された意味論的意味を指定する、アクセスすることと、
意味論的意味は、所定のフィールドにどのような種類のデータ値が含まれるかを示し、
前記フィールドに対して前記特定された１つ又は複数の属性に少なくとも前記複数のテストを適用することに基づいて、前記フィールドの前記１つ又は複数のデータ値の１つ又は複数の提案された意味論的意味を指定することと、
前記フィールドに対して前記１つ又は複数の提案された意味論的意味の間の類似性を特定することと、
前記１つ又は複数の提案された意味論的意味の間の前記類似性に基づいて、前記提案された意味論的意味の1つを、前記フィールドの前記１つ又は複数のデータ値の前記意味論的意味を識別するものとして識別することと、
データストア内において、前記フィールドの前記識別子を、前記フィールドの前記１つ又は複数のデータ値の前記識別された意味論的意味と共に保存することと
を含む、方法。
【請求項２】
前記フィールドのデータ値のフォーマットを特定することをさらに含み、前記１つ又は複数の属性は前記フォーマットを示す、請求項１に記載の方法。
【請求項３】

前記フィールドに含まれる前記データ値を表す統計値を特定することをさらに含み、前記１つ又は複数の属性は統計値を示す、請求項１に記載の方法。
【請求項４】
前記統計値は、前記フィールドの前記１つ又は複数のデータ値の最小長さ、前記フィールドの前記１つ又は複数のデータ値の最大長さ、前記フィールドの最も一般的なデータ値、前記フィールドの最も一般的でないデータ値、前記フィールドの最大データ値及び前記フィールドの最小データ値の少なくとも１つを含む、請求項３に記載の方法。
【請求項５】
前記複数のテストを適用することは、
前記フィールドが、前記１つ又は複数のデータセットのうちのデータセットのためのプライマリキーを含むことを特定することと、
前記プライマリキーに関連する前記複数のテストのうち1つのテストを選択することと
を含む、請求項１に記載の方法。
【請求項６】
前記複数のテストを適用することは、前記フィールドのデータ値の、用語集内の用語とのメタデータ比較を行うことを含む、請求項１に記載の方法。
【請求項７】
前記複数のテストを適用することは、
前記フィールドの前記データ値の前記１つ又は複数の属性から、前記フィールドの前記１つ又は複数のデータ値によって表されるパターンを特定することと、
前記パターンにマッピングされる特定の意味論的意味を特定することと、
前記フィールドを前記特定の意味論的意味でラベル付けすることと
を含む、請求項１に記載の方法。
【請求項８】
前記複数のテストを適用することは、
データ集合を表す値のリストを検索することと、
前記フィールドの前記１つ又は複数のデータ値を前記値のリストと比較することと、
前記比較することに応答して、前記データ値の閾値数が値の前記リストの前記値と一致することを特定することと、
前記特定することに応答して、前記フィールドを、前記データ集合を指定する特定の意味論的意味でラベル付けすることと
を含む、請求項１に記載の方法。
【請求項９】
前記複数のテストを適用することは、
前記フィールドのための少なくとも２つの意味論的意味を生成することと、
前記少なくとも２つの意味論的意味が相互に排他的であるか又は包含的であるかを特定することと
を含む、請求項１に記載の方法。
【請求項１０】
前記複数のテストを適用することに応答して、前記フィールドと、前記１つ又は複数のデータセットの別のフィールドとの間の関係を特定することをさらに含む、請求項１に記載の方法。
【請求項１１】
前記関係は、前記フィールドの第一のデータ値が、他のフィールドに保存された第二のデータ値を決定するという表示、前記第一のデータ値が前記第二のデータ値と相関するという表示又は前記第一のデータ値が前記第二のデータ値と同一であるという表示の１つを含む、請求項１０に記載の方法。
【請求項１２】
前記複数のテストは、少なくとも１つの重み値にそれぞれ関連付けられ、前記方法は、
少なくとも１つのテストに関連付けられた重み値を更新することと、

前記更新された重み値を使用して、前記フィールドの前記データ値の前記１つ又は複数の属性に基づいて前記テストを再び適用することと
をさらに含む、請求項１に記載の方法。
【請求項１３】
機械学習プロセスを使用して前記複数のテストを訓練することをさらに含む、請求項１に記載の方法。
【請求項１４】
データ品質ルール環境から、前記提案された意味論的意味に割り当てられる１つ又は複数のデータ品質ルールを検索することと、
前記１つ又は複数のデータ品質ルールのうちのデータ品質ルールを前記フィールドに割り当てることと
をさらに含む、請求項１に記載の方法。
【請求項１５】
スコア値を各提案された意味論的意味に適用することと、
前記提案された意味論的意味の各意味論的意味について、前記意味論的意味に関連付けられた前記スコア値を結合することと、
各提案された意味論的意味に関連付けられた前記スコア値に従って前記提案された意味論的意味をランク付けすることと
をさらに含む、請求項１に記載の方法。
【請求項１６】
前記複数のテストから前記提案された意味論的意味のバリデーションを受け取ることと、
前記バリデーションを受け取ることに応答して、前記複数のテストを重み付けすることと
をさらに含む、請求項１に記載の方法。
【請求項１７】
前記データストアは、データ辞書を含む、請求項１に記載の方法。
【請求項１８】
前記提案された意味論的意味をデータ品質ルール環境に出力することをさらに含む、請求項１に記載の方法。
【請求項１９】
前記提案された意味論的意味の前記識別された１つに基づいて、前記データ品質ルール環境からのデータ品質ルールを使用して前記フィールドのデータを処理することに関するエラーの数を、前記提案された意味論的意味の前記識別された１つを使用することなく前記フィールドの前記データを処理することに関するエラーの別の数に対して減少させることをさらに含む、請求項１８に記載の方法。
【請求項２０】
特定の識別子を含む前記１つ又は複数のデータセットを処理するためのリクエストを受信することと、
前記リクエストに応答して、前記データストアから、前記特定の識別子と、前記特定の識別子に関連付けられた特定の意味論的意味とにアクセスすることと、
前記特定の意味論的意味に基づいて、前記１つ又は複数のデータセットの少なくとも一部を処理するための１つ又は複数のデータ処理ルールを特定することと、
少なくとも前記１つ又は複数のデータセットの前記一部を、前記選択された１つ又は複数のデータ処理ルールに従って処理することと
をさらに含む、請求項1に記載の方法。
【請求項２１】
１つ又は複数のデータセットに含まれるフィールドの意味論的意味を発見するためのデータ処理システムであって、前記データ処理システムは、
命令を保存するデータストレージと、

前記データストレージによって保存された前記命令を実行して、
１つ又は複数のデータセットに含まれるフィールドを識別することであって、前記フィールドは、識別子に関連付けられている、識別することと、
前記フィールドの１つ又は複数のデータ値を識別することであって、少なくとも前記データ値のうちの第1のデータ値は、前記データ値のうちの第2のデータ値と異なる、識別することと、
前記フィールドの前記１つ又は複数のデータ値に基づいて、前記フィールドの前記データ値の１つ又は複数の属性を特定することと、
複数のテストにアクセスすることであって、テストは、１つ又は複数の所定の属性に対して提案された意味論的意味を指定する、アクセスすることと、
意味論的意味は、所定のフィールドにどのような種類のデータ値が含まれるかを示し、
前記フィールドに対して前記特定された１つ又は複数の属性に少なくとも前記複数のテストを適用することに基づいて、前記フィールドの前記１つ又は複数のデータ値の１つ又は複数の提案された意味論的意味を指定することと、
前記フィールドに対して前記１つ又は複数の提案された意味論的意味の間の類似性を特定することと、
前記１つ又は複数の提案された意味論的意味の間の前記類似性に基づいて、前記提案された意味論的意味の1つを、前記フィールドの前記１つ又は複数のデータ値の前記意味論的意味を識別するものとして識別することと、
データストア内において、前記フィールドの前記識別子を、前記フィールドの前記１つ又は複数のデータ値の前記識別された意味論的意味と共に保存することと
を含む動作を行うように構成された少なくとも１つのプロセッサと
を含むデータ処理システム。
【請求項２２】
前記動作は、
特定の識別子を含む前記１つ又は複数のデータセットを処理するためのリクエストを受信することと、
前記リクエストに応答して、前記データストアから、前記特定の識別子と、前記特定の識別子に関連付けられた特定の意味論的意味とにアクセスすることと、
前記特定の意味論的意味に基づいて、前記１つ又は複数のデータセットの少なくとも一部を処理するための１つ又は複数のデータ処理ルールを特定することと、
少なくとも前記１つ又は複数のデータセットの前記一部を、前記選択された１つ又は複数のデータ処理ルールに従って処理することと
をさらに含む請求項21に記載のデータ処理システム。
【請求項２３】
前記動作は、前記フィールドのデータ値のフォーマットを特定することをさらに含み、前記１つ又は複数の属性は前記フォーマットを示す、請求項21に記載のデータ処理システム。
【請求項２４】
前記動作は、前記フィールドに含まれる前記データ値を表す統計値を特定することをさらに含み、前記１つ又は複数の属性は統計値を示す、請求項21に記載のデータ処理システム。
【請求項２５】
前記統計値は、前記フィールドの前記１つ又は複数のデータ値の最小長さ、前記フィールドの前記１つ又は複数のデータ値の最大長さ、前記フィールドの最も一般的なデータ値、前記フィールドの最も一般的でないデータ値、前記フィールドの最大データ値及び前記フィールドの最小データ値の少なくとも１つを含む、請求項24に記載のデータ処理システム。
【請求項２６】
前記複数のテストを適用することは、

前記フィールドが、前記１つ又は複数のデータセットのうちのデータセットのためのプライマリキーを含むことを特定することと、
前記プライマリキーに関連する前記複数のテストのうち1つのテストを選択することとを含む、請求項２１に記載のデータ処理システム。
【請求項２７】
前記複数のテストを適用することは、前記フィールドのデータ値の、用語集内の用語とのメタデータ比較を行うことを含む、請求項２１に記載のデータ処理システム。
【請求項２８】
前記複数のテストを適用することは、
前記フィールドの前記データ値の前記１つ又は複数の属性から、前記フィールドの前記１つ又は複数のデータ値によって表されるパターンを特定することと、
前記パターンにマッピングされる特定の意味論的意味を特定することと、
前記フィールドを前記特定の意味論的意味でラベル付けすることと
を含む、請求項２１に記載のデータ処理システム。
【請求項２９】
前記複数のテストを適用することは、
データ集合を表す値のリストを検索することと、
前記フィールドの前記１つ又は複数のデータ値を前記値のリストと比較することと、
前記比較することに応答して、前記データ値の閾値数が値の前記リストの前記値と一致することを特定することと、
前記特定することに応答して、前記フィールドを、前記データ集合を指定する特定の意味論的意味でラベル付けすることと
を含む、請求項２１に記載のデータ処理システム。
【請求項３０】
１つ又は複数のデータセットに含まれるフィールドの意味論的意味を発見するための命令を記憶する１つ又は複数の非一時的コンピュータ可読媒体であって、前記命令は、
１つ又は複数のデータセットに含まれるフィールドを識別することであって、前記フィールドは、識別子に関連付けられている、識別することと、
前記フィールドの１つ又は複数のデータ値を識別することであって、少なくとも前記データ値のうちの第1のデータ値は、前記データ値のうちの第2のデータ値と異なる、識別することと、
前記フィールドの前記１つ又は複数のデータ値に基づいて、前記フィールドの前記データ値の１つ又は複数の属性を特定することと、
複数のテストにアクセスすることであって、テストは、１つ又は複数の所定の属性に対して提案された意味論的意味を指定する、アクセスすることと、
意味論的意味は、所定のフィールドにどのような種類のデータ値が含まれるかを示し、
前記フィールドに対して前記特定された１つ又は複数の属性に少なくとも前記複数のテストを適用することに基づいて、前記フィールドの前記１つ又は複数のデータ値の１つ又は複数の提案された意味論的意味を指定することと、
前記フィールドに対して前記１つ又は複数の提案された意味論的意味の間の類似性を特定することと、
前記１つ又は複数の提案された意味論的意味の間の前記類似性に基づいて、前記提案された意味論的意味の1つを、前記フィールドの前記１つ又は複数のデータ値の前記意味論的意味を識別するものとして識別することと、
データストア内において、前記フィールドの前記識別子を、前記フィールドの前記１つ又は複数のデータ値の前記識別された意味論的意味と共に保存することと
を含む動作を行うように構成された１つ又は複数のプロセッサによって実行可能である、１つ又は複数の非一時的コンピュータ可読媒体。
【請求項３１】
前記動作は、

特定の識別子を含む前記１つ又は複数のデータセットを処理するためのリクエストを受信することと、
前記リクエストに応答して、前記データストアから、前記特定の識別子と、前記特定の識別子に関連付けられた特定の意味論的意味とにアクセスすることと、
前記特定の意味論的意味に基づいて、前記１つ又は複数のデータセットの少なくとも一部を処理するための１つ又は複数のデータ処理ルールを特定することと、
少なくとも前記１つ又は複数のデータセットの前記一部を、前記選択された１つ又は複数のデータ処理ルールに従って処理することと
をさらに含む、請求項30に記載の１つ又は複数の非一時的コンピュータ可読媒体。