(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2016 年 10 月 20 日(20.10.2016)

WIPO | PCT

(10) 国際公開番号
WO 2016/166878 A1

(51) 国際特許分類:
G06Q 50/10 (2012.01) *G06F 19/00* (2011.01)
G06F 17/30 (2006.01)

(21) 国際出願番号: PCT/JP2015/061778

(22) 国際出願日: 2015 年 4 月 17 日(17.04.2015)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人: 株式会社日立製作所 **(HITACHI, LTD.)** [JP/JP]; 〒1008280 東京都千代田区丸の内一丁目６番６号 Tokyo (JP).

(72) 発明者: 平山　淳一 **(HIRAYAMA Junichi)**; 〒1008280 東京都千代田区丸の内一丁目６番６号 株式会社日立製作所内 Tokyo (JP). 嶺　竜治 **(MINE Ryuji)**; 〒1008280 東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内 Tokyo (JP).

(74) 代理人: 井上　学, 外 **(INOUE Manabu** et al.); 〒1008220 東京都千代田区丸の内一丁目６番１号 株式会社日立製作所内 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): 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, UZ, VC, VN, ZA, ZM, ZW.

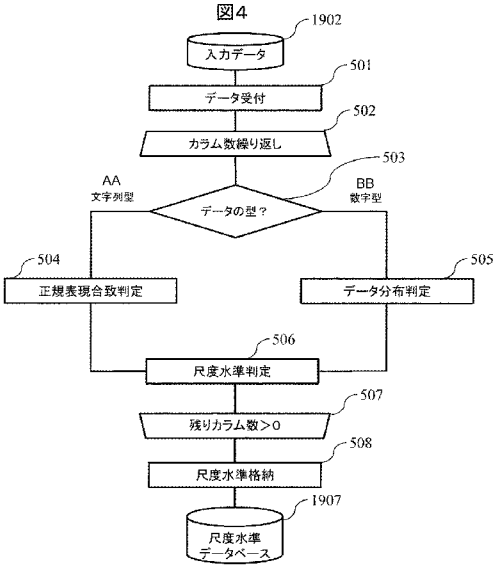
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

[続葉有]

(54) **Title**: AUTOMATIC DATA PROCESSING SYSTEM, AUTOMATIC DATA PROCESSING METHOD, AND AUTOMATIC DATA ANALYSIS SYSTEM

(54) 発明の名称 : データ自動加工システム、データ自動加工方法、およびデータ自動解析システム

図4



501 Data reception
502 Repeat column number
503 Data type?
504 Regular expression conformance determination
505 Data distribution determination
506 Scaling level determination
507 Remaining number of columns > 0
508 Scaling level storage
1902 Input data
1907 Scaling level database
AA Character string
BB Numerical

(57) **Abstract**: An automatic data processing system is configured to comprise: a reception unit for receiving data pertaining to numbers, characters, and symbols; a data type determination unit for determining the type of the data; a scaling level determination unit for determining the scaling level of the data on the basis of the distribution of the data when the data is numerical; and a data processing unit for processing data on the basis of the scaling level. Such a configuration makes it possible to provide a system for automatically determining the scaling level, which is an index for setting the properties of the data, and performing automatic processing of the data.

(57) 要約: データ自動加工システムを、数字、文字、および符号に関するデータを受け付ける受付部と、データの型を判定するデータ型判定部と、データが数字型である場合に、データの分布に基づいてデータの尺度水準を判定する尺度水準判定部と、尺度水準に基づいてデータを加工するデータ加工部からなる構成とする。係る構成により、データの性質を定める指標である尺度水準を自動で判定し、データの自動加工を行うシステムを提供することができる。

WO 2016/166878 A1

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

明　細　書

発明の名称 ：

データ自動加工システム、データ自動加工方法、およびデータ自動解析システム

技術分野

[0001] 本発明は、データ自動加工システム、データ自動加工方法、およびデータ自動解析システムに関する。

背景技術

[0002] 近年、ビッグデータと呼ばれる大量のデータを分析し、今まで人が勘と経験で行ってきた意思決定を支援するシステムの開発が急速に発展してきている。例えば、ある目的変数を変動させる他の説明変数が何であるかと見つけ出すための相関分析や、説明変数群の値から目的変数の値を予測する回帰分析や、似た傾向を持つ変数同士をグルーピングするクラスタリングといった機械学習・統計分析を主なデータ分析の手法としているシステムが開発されている。

[0003] データ分析を行う際、蓄積した生のデータそのままでは、分析に適さないことが多く、何らかの加工演算を施したデータを新たに分析用のデータとすることが多い。データ加工演算の例として、量子化や代表値などがある。量子化とは、例えば0.0～30.0で分布しているデータを、0.0～10.0⇒Low、10.0～20.0⇒Middle、20.0～30.0⇒Highのようにある区間に分け、その区間内に属する値に対して、新たにラベル化する作業である。代表値とは、あるカラム内のデータに対する平均や各値の頻度などにより、データをそのカラムを代表する１つの値にまとめた数値のことである。データ加工の例を、図１を用いて説明する。図１はＲＤＢ（リレーショナルデータベース）形式の入力テーブル１００に蓄積されたデータを、出力テーブル１１０に圧縮している例である。入力テーブル１００が”作業ID”（１０４）をキーとしているのに対し、圧縮後の出力テーブル１１０では、”作業者ID”（１１１）がキー

となっている。このとき、”作業者ID”（１０１）が同じレコードをグループとし、そのグループごとに代表値を求めている。この加工により、各カラムの値を、それぞれの作業者「700A」「700B」「700C」を代表する値に直すことができる。前述のデータ加工に関連する文献として、特許文献１がある。この公報では、テーブルに蓄積された変数を元に、事前に定めたルール・集計方法に従い、新たな変数を作成し、それを新たに説明変数として追加している。ルール・集計方法の例として、時系列を表す変数があれば１時間ごとに纏めて平均をとる集約演算などがある。このように説明変数を追加した後に、目的変数と説明変数の寄与度を計算することで、目的変数に寄与する説明変数を特定している。

[0004] また、データの性質を定める指標である尺度水準というものが知られている。例えば特許文献２には、データの尺度水準によって散布度の計算式を変えて、計算した散布度によって自社の製品・サービスの独自性を判定し、ポジショニング・マップを作成している。更に、非特許文献１には、尺度水準についての記載が存在する。

先行技術文献

特許文献

[0005] 特許文献1：特開２０１２－２７８８０号公報

特許文献2：特開２０１１－２４３０５０号公報

非特許文献

[0006] 非特許文献1：S. S. Stevens, “On the Theory of Scales of Measurement,” Science, vol.103, no.2684, pp.677-680, Jun. 1946

発明の概要

発明が解決しようとする課題

[0007] しかしながら、データの加工演算に関して見かけ上は同様に数値に見えるデータであっても、そのデータの持つ性質は異なり、適用出来る加工演算も異なる。例えば、作業ごとの所要時間：[180[s], 240[s], …]のような数量

を示すデータに対して平均を求める代表値化には意味があるが、作業者ID：[23513, 24512, …]のような符号や名前を示すデータに対して平均を求めてもその数値は意味を成さない。このように、意味を成さない演算を施してしまった場合には、適切な分析結果にはならず、誤った分析結果を招いてしまう恐れや、真に抽出したい分析結果が意味のない分析結果に埋もれてしまう恐れがある。

[0008] 前述の例のような、適用可能なデータ加工演算は、前処理としてデータ分析の知識を持つ専門家が、手作業で全てのカラムに対し適切に設定する必要があり、分析作業のコスト増大の原因となっていた。また、データ分析の知識を持たない非専門家が、これらの設定を行うことは困難であった。

[0009] さらに、特許文献２では、データの尺度水準によって散布度の計算方法を変えているが、データの尺度水準をユーザが事前に指定する必要があり、尺度水準の判定を自動で行うことはできなかった。

[0010] そこで、本発明はデータの性質を定める指標である尺度水準を自動で判定し、各データに適した方法によりデータ加工を行うシステムおよび方法、並びにデータの尺度水準を自動で判定する機能を有するデータ解析システムを提供することを目的とする。

課題を解決するための手段

[0011] 前記課題を解決するための手段のうち代表的なものを例示すれば、数字、文字、および符号に関するデータを受け付ける受付部と、データについてデータの型を判定するデータ型判定部と、データが数字型である場合にデータの分布に基づいてデータの尺度水準を判定する尺度水準判定部と、尺度水準に基づいてデータを加工するデータ加工部と、を有するデータ自動加工システムが挙げられる。

[0012] また、数字、文字、および符号に関するデータを入力とするデータ自動加工方法であって、データを受け付ける受付ステップと、データについてデータの型を判定するデータ型判定ステップと、データが数字型である場合にデータの分布に基づいてデータの尺度水準を判定する尺度水準判定ステップと

、尺度水準に基づいてデータを加工するデータ加工ステップと、を有することを特徴とするデータ自動加工方法が挙げられる。

[0013] さらに、数字、文字、および符号に関するデータを受け付ける受付部と、データについてデータの型を判定するデータ型判定部と、データが数字型である場合にデータの分布に基づいてデータの尺度水準を判定する尺度水準判定部と、尺度水準に基づいてデータを加工するデータ加工部と、加工部によって加工されたデータを解析するデータ解析部と、解析部によって解析されたデータを出力する出力部と、を有するデータ解析システムが挙げられる。

発明の効果

[0014] 本発明によれば、データの性質を定める指標である尺度水準を自動で判定し、データの自動加工を行うシステムおよび方法、並びにデータの尺度水準を自動で判定する機能を有するデータ解析システムを提供することができる。

図面の簡単な説明

[0015] [図1]データ加工の一列を示す図。

[図2]各尺度水準の説明図。

[図3]入出力テーブルの一例を示す図。

[図4]データ自動加工システムの処理フローを示す図。

[図5]データ分布判定ステップの処理フローを示す図。

[図6]正規表現判定ステップの処理フローを示す図。

[図7]加工演算判定部の処理フローを示す図。

[図8]尺度水準ごとの適用可能加工演算を示すテーブルの一例を示す図。

[図9]加工演算選択部の処理フローを示す図。

[図10]演算ロバスト性判定ステップの処理フローを示す図。

[図11]データ自動加工システムのハードウェア構成図。

[図12]各尺度水準を持つデータの分布を示す図。

[図13]等分散性と非等分散性を持つデータの分布を示す図。

[図14]単調変化の特性を持つデータの分布を示す図。

[図15]データの等分散性の判定の流れの一例を示す図。

[図16]データ加工操作を行うためのＧＵＩの一例を示す図。

[図17]データ自動解析システムの構成図。

[図18]データ自動加工システムの構成図。

[図19]加工演算データベースおよび加工演算判定部を有するデータ自動加工システムの構成図。

[図20]加工演算選択部を有するデータ自動加工システムの構成図。

発明を実施するための形態

[0016] 以下の実施の形態においては、便宜上その必要があるときは、複数のセクションまたは実施の形態に分割して説明するが、特に明示した場合を除き、それらは互いに無関係なものではなく、一方は他方の一部または全部の変形例、詳細、補足説明等の関係にある。また、以下の実施の形態において、要素の数等（個数、数値、量、範囲等を含む）に言及する場合、特に明示した場合および原理的に明らかに特定の数に限定される場合等を除き、その特定の数に限定されるものではなく、特定の数以上でも以下でもよい。

[0017] さらに、以下の実施の形態において、その構成要素（要素ステップ等も含む）は、特に明示した場合および原理的に明らかに必須であると考えられる場合等を除き、必ずしも必須のものではないことは言うまでもない。同様に、以下の実施の形態において、構成要素等の形状、位置関係等に言及するときは、特に明示した場合および原理的に明らかにそうでないと考えられる場合等を除き、実質的にその形状等に近似または類似するもの等を含むものとする。このことは、上記数値および範囲についても同様である。

実施例 1

[0018] 本実施例では、データの尺度水準を自動で決定するデータ自動加工システムの例を説明する。

[0019] 図１８は、本実施例のデータ自動加工システムの構成図の例である。データ自動加工システム１９０１は、入力データ１９０２を受け付けて、データの尺度水準を判定しデータを加工して、出力データベース１９０３に加工デ

ータを出力する。データ自動加工システム１９０１は、データ受付部１９０４、データ型判定部１９０５、尺度水準判定部１９０６、尺度水準データベース１９０７、及びデータ加工部１９０８を備えている。

[0020] データ受付部１９０４は、入力データ１９０２を受け付ける。その際、受け付けたデータをデータ自動加工システム１９０１中で取り扱うデータ形式に変換してもよい。入力データ１９０２は、数字、文字、または符号に関するデータである。入力データ１９０２の例として、例えば表形式のデータがある。表形式のデータを示した図３の入力テーブル４００は、ＲＤＢ（リレーショナルデータベース）の形式をしており、キー部４０４およびバリュー部４０５からなる。この他に、キー部４０４が省略されておりバリュー部４０５のみの形式であってもよい。ここでは、便宜上、表の形で表現しているが、ＣＳＶ（Comma Separated Value；カンマ区切りデータ）でも、スペース区切りデータでも、タブ区切りデータでも本質的には同じである。データ受付部１９０４は、受け付けたデータをデータ型判定部１９０５に送信する。

[0021] データ型判定部１９０５は、受付部１９０４から受信したデータについて各カラムに格納されたデータの型が、浮動小数点型、整数型、文字列型のいずれであるかを判定する。判定方法としては、例えば、代表的なデータベース言語であるＳＱＬにて判定された結果を用いてsmallint、integer、bigint等ならば整数型、decimal、numeric、real等ならば浮動小数点型、それ以外ならば文字列型とする方法がある。

[0022] データ型判定部１９０５は、データ自動加工システム１９０１に入力されたデータ及びデータの各カラムについて判定したデータの型についての情報を尺度水準判定部１９０６に送信する。

[0023] 図４は、入力データ１９０２を受け付けて尺度水準データベース１９０７に各カラムの尺度水準を格納する処理のフローの一例を示した図である。

[0024] ステップ５０１にて、データ受付部１９０４により入力データ１９０２を受け付ける。次に、ステップ５０３、５０４、５０５、及び５０６を入力データ１９０２のカラムの数だけ繰り返す(ステップ５０２および５０７)。

[0025] ステップ５０３では、データ型判定部１９０５により、各カラム内のデータの型を判定する。例えば、上述した代表的なデータベース言語であるＳＱＬにて判定された結果を用いて各カラム内のデータの型を判定する。ステップ５０３の判定結果において、当該カラムのデータが浮動小数点型もしくは整数型(これらを以下では数字型とする)ならばステップ５０５へ遷移し、文字列型ならばステップ５０４へ遷移する。

[0026] ステップ５０４では、所定の正規表現との合致有無の判定を行う。所定の正規表現とは、例えば、日付表現、時刻表現、時間表現、またはリスト表現などが挙げられる。

[0027] ステップ５０５では、カラム内のデータの分布の判定を行う。データの分布とは、データの統計値を基に計算されるデータの統計的な性質である。例えば、連続性、中心性、単調減少性、平滑性、または等分散性等が挙げられる。

[0028] ステップ５０６にて、ステップ５０４により判定された所定の正規表現との合致有無またはステップ５０５により判定されたデータの分布を基に各カラム内の尺度水準を判定する。

[0029] ステップ５０４～５０６は、尺度水準判定部１９０６により処理が行われる。これらの処理により、各カラムについて名義尺度、順序尺度、間隔尺度、比例尺度のうち、どれにあてはまるかを判定する。

[0030] ステップ５０７にて、尺度水準判定部１９０６により判定した尺度水準を各カラムと紐づけて尺度水準データベース１９０７に格納する。

＜尺度水準の説明＞

次に、尺度水準の例を、図２を用いて説明する

尺度水準とは、カラムに保存されているデータを、それらが表現する情報の性質に基づき数学・統計学的に分類する基準である。Stanley Stevensが提案した分類（非特許文献１）がよく用いられている。尺度には低い方から順に、図２に示す４つの水準があり、高い水準はより低い水準の性質を含む形になっている。

[0031] 名義尺度（Nominal scale）

数字・文字を単なる名前として個々のデータに割り振る。２つのデータに同じ名前がついていればそれらは同じカテゴリに属する。データ間の比較は等しいか異なるかでのみ可能である。順序はなく、加減などの算術演算もできない。代表値は最頻値で表される。データの例としては、ID、氏名、フラグなどがある。例えば作業ID＝（00001,00002,00004,00007,…）は、作業ID=00001と作業ID=00002のデータは作業が違うことのみを表し、どちらが大きいかといった比較はできない。

[0032] 順序尺度（Ordinal scale）

データに割り振られた数字・文字は順序を表す。データ間の比較は等しいか異なるかに加え、その前後・大小関係にも意味がある。一方、順序の間隔は等しくないため、加減などの算術演算には意味がない。データの例としては、作業効率Gr.やオーダー順などがある。例えば、作業効率Gr. = (5, 4, 3, …)に対して、5よりも4の方が良いといった比較はできる。一方、5→4の間隔と、4→3の間隔は均一ではなく、単純に差をとった1という値は意味を成さない。

[0033] 間隔尺度（Interval scale）

データに割り振られた数字は順序尺度の性質を全て満たし、さらに差が等しいということは間隔が等しいということを意味する。２つデータ間の差を比較しても意味がある。加減算にも意味があるが、尺度上のゼロ点は任意で負の値も使える。代表値は最頻値、中央値、算術平均などで表される。データの例としては、時刻や日付などがある。例えば、日付 =(11/4, 11/6, 11/8….)に対して、11/4→11/6の差をとった2[日間]には定量的な意味があり、同様に11/6→11/8の2[日間]との大小の比較が可能である。

[0034] 比例尺度（Ratio scale）

データに割り振られた数字は間隔尺度の性質を全て満たし、さらに２つのデータの比にも、乗除算にも意味がある。尺度上のゼロ点は絶対的である。代表値は最頻値、中央値、算術平均、幾何平均などで表される。データの例と

しては、時間や数量などがある。例えば、作業数量 = (2,5,10,…)に対して、2[個]と5[個]の比をとって、2.5倍多いといった意味づけが可能である。

[0035] 尺度水準判定部１９０６は、データが格納されている各カラムが上記４つの尺度水準のどれにあてはまるかを判定する。尺度水準判定部１９０６は、データ型判定部１９０５が、当該カラム内のデータを数字型と判定した場合にはデータ分布の判定を行い、文字列型ならば正規表現合致有無の判定を行う。

[0036] データ分布の判定では、尺度水準判定部１９０６は各カラムに格納されているデータの分布の判定を行い、そのデータの分布に基づいてカラムの尺度水準を判定する。データの分布は、データの値と当該データの値の出現頻度から計算してもよい。さらに、横軸にデータの値を、縦軸にデータの値の出現頻度を設定することでヒストグラムを作成し、その形状からデータの分布を求めてもよい。データの値とデータの値の出現頻度の組合せ以外にもデータの分布を判断できるものであれば組合せの種類は問わない。

[0037] 図５は、数字型のデータについて、図４のデータ分布判定５０５および尺度水準判定５０６の処理フローの一例を示した図である。

[0038] ステップ６０１では、カラム内のデータが十分に連続性を持つか判定する。連続性とは、カラム内のデータが飛び飛びになっておらず十分に密になっているかを表す指標である。カラム内のデータの数字が等間隔で、数量的な意味があれば、すなわち、間隔尺度や比例尺度のデータであれば、データが不規則に飛び飛びにはなりにくいといった特性を判断するための指標である。図１２は、様々なデータの分布を示す図であり、横軸がデータの値、縦軸がそのデータの出現頻度である。図１２の例の場合、ヒストグラム１３０１、１３０２が連続性を持たない例、それ以外が連続性を持つ例である。連続性を持つかどうかの判定方法の例として、例えば、以下のような方法がある。

（１）カラム内のデータを昇順もしくは降順にソートし、データの値が重複するものを排除し、１つにする。

（２）（１）のデータ列の全ての値に対し、隣り合う２つの値の差分値を求める。

（３）求めた全ての差分値の標準偏差を求める。

（４）求めた標準偏差がある閾値以下になれば、連続性を持つと判定する。より好ましくは、（２）にて差分値を求めた後に、最小の差分値で割ることで正規化することが望ましい。別の例として、データの標準偏差とレンジ（最大値－最小値）の比率がある閾値以下になるかどうかを判定する方法が考えられる。レンジに代わって、７５％点－２５％点や９０％点－１０％点を用いることもできる。このほか、データの値が連続的であるかどうかを計算する方法であれば、これらには依らない。ステップ６０１にて連続性を持つと判定された場合、ステップ６０２に遷移し、連続性を持つと判定されなかった場合、名義尺度であると判定される(ステップ６０５)。

[0039] このステップ６０１によって、データの値の間にギャップが存在するカラムを判定することができ、その結果、当該カラムが名義尺度であると判定することができる。これにより、数字型のデータが名義尺度であるかその他の尺度水準であるかを判定することができる。

[0040] ステップ６０２では、カラム内のデータが中心性もしくは単調減少性を持つか判定する。中心性とは、データの中央や平均付近のデータが多く存在し、ヒストグラムが山なりに分布するかを表す指標である。図１２の例の場合、ヒストグラム１３０１、１３０４が中心性を持つ例、それ以外が中心性を持たない例である。単調減少性とは、ヒストグラムを描いた際に、横軸の値の増加に対し、縦軸の値が徐々に減少していくかを表す指標である。これらの指標は、数量データ、特に比例尺度のデータのヒストグラムによく見られる、正規分布形状、対数正規分布形状、及び指数分布形状を判断するためのものである。図１２の例の場合、ヒストグラム１３０５、１３０６が単調減少性を持つ例、それ以外が単調減少性を持たない例である。中心性および単調減少性を持つかどうかの判定方法の例として、尖度および歪度がある閾値以上になるかどうかを判定する方法がある。尖度は式（１）で計算される値

であり、歪度は式（２）で計算される値である。

[0041] [数1]

【数1】

$$\sum_{i}^{N}(x_i-\mu)^4/N\sigma^4 \qquad \cdots 式(1)$$

[0042] [数2]

【数2】

$$\sum_{i}^{N}(x_i-\mu)^3/N\sigma^3 \qquad \cdots 式(2)$$

[0043] 式（１）（２）において、xi(i=1 to N)は各データの値、μは平均、σは標準偏差を示す。ここで、尖度とはデータの中心性を示すものであり、上記式（１）の値が大きいときに尖度が大きく、すなわち中心性があることを意味する。例えば、式（１）の値が3以上であるときに中心性を持つデータであると判断してもよい。また、歪度とはデータの単調減少性を示すものであり、上記式（２）の値が大きいときに歪度が大きく、すなわち、単調減少性があることを意味する。例えば、式（２）の値が0.5以上であるときに歪度を持つデータであると判断してもよい。このほか、ヒストグラムが大局的な山状もしくは単調減少するかを判定する方法であれば、これには依らない。ステップ６０２で中心性もしくは単調減少性を持つと判定された場合、ステップ６０４に遷移し、中心性および単調減少性を持たないと判定された場合、ス

テップ６０３に遷移する。

[0044] このステップ６０２により、カラム内のデータが不規則に存在しているかどうか判定することができる。

[0045] ステップ６０３では、カラム内のデータが平滑性を持つかどうか判定する。平滑性とは、ヒストグラムを描いた際に、横軸の値の増加に対し、縦軸の値の変化が緩やかであるかどうかを表す値である。カラム内のデータの数字に数量的な意味があれば、すなわち名義尺度でなければ、数字が隣り合うデータの頻度が近くなりやすいといった特性を判断するための指標である。図１２の例の場合、ヒストグラム１３０４、１３０７、１３０８が平滑性を持つ例、それ以外が平滑性を持たない例である。平滑性を持つかどうかの判定方法の例として、例えば以下のような方法がある。

（１）カラム内のデータの最大値から最小値までを、いくつかの均等幅の区間に分ける。

（２）分けた区間ごとに、当該区間に属するデータ数を計算する。

（３）当該区間に属するデータ数と、隣り合う区間に属するデータ数の差分を、全ての区間に対して計算する。

（４）計算した全ての区間の差分値の平均を計算する。

（５）計算した平均がある閾値以下になれば、平滑性を持つと判定する。

[0046] このほか、ヒストグラムの形状が平滑的になるかどうかを判定する方法であれば、これには依らない。ステップ６０３にて平滑性を持つと判定された場合、間隔尺度と判定され、平滑性を持たないと判定された場合、名義尺度と判定される(ステップ６０５及び６０６)。

[0047] ステップ６０２および６０３により、カラム内のデータの値の間にギャップは存在しないが、不規則にデータが存在していて、値の隣り合うデータ同士に頻度の差が大きく表れるようなカラムを判定することができ、その結果、カラムが名義尺度であると判定することができる。さらに、カラム内のデータの値の間にギャップは存在しないが、不規則にデータが存在していて、値の隣り合うデータ同士に似たような頻度をもつ傾向があるカラムを判定す

ることができ、その結果、カラムが間隔尺度であると判定することができる。これにより、連続性があり、かつ、中心性がないもしくは単調減少性がない数字型のデータについて、名義尺度か間隔尺度か判定をすることが出来る。

[0048] ステップ６０４では、カラム内のデータが等分散性を持つかどうか判定する。等分散性とは、データの平均値の変化に対して、分散値が変化しないかどうかを表す指標である。図１３は、等分散性を持つデータ分布の例と等分散性を持たないデータ分布の例を示した図である。図１３の例の場合、上段のヒストグラム１４１０が等分散性を持つ例、下段のヒストグラム１４２０が等分散性を持たない例である。ヒストグラム１４１０では、分布１４１１、１４１２、１４１３、１４１４と分布の平均値が大きくなっても、分布の分散が不変である。一方、ヒストグラム１４２０では、分布１４２１、１４２２、１４２３、１４２４と分布の平均値が大きくなるにつれて、分布の分散が大きくなる。

[0049] 等分散性を持つかどうかの判定方法の例を、図１５を用いて説明する。

（１）入力テーブル（１６１０）の着目しているカラム（例.”処理数”と”開始時刻[s]”）の値に対し、加工キー（例.”作業者ID”）が同じ行（点線内）ごとに平均と分散を求める。この加工キーは、ユーザが入力してもよいし、データ自動加工システムがランダムに選択してもよい。

（２）求めた平均と分散に着目し、平均が増加しても分散が大きく変化しないかどうかを判定する。図１５の場合、”処理数”は平均が増加すると分散が増加しており、”開始時刻[s]”は平均が増加しても分散が大きく変化していない。例えば、各加工キーの分散／平均を計算して、その差を閾値と比較して上記判断を行ってもよい。等分散性を持つデータの場合、分散／平均の値は、各加工キー間で変動する。

（３）分散が大きく変化しないと判定されれば、等分散性を持つ、それ以外ならば、等分散性を持たないと判定する。

[0050] ステップ６０４にて、等分散性を持つと判定された場合、間隔尺度と判定

され、等分散性を持たないと判定された場合、比例尺度と判定される(ステップ６０６および６０７)。このように比例尺度と間隔尺度の判定をすることが出来ることを、発明者は新たに発見した。ステップ６０４によって、連続性があり、かつ、中心性を持つもしくは単調減少性を持つ数字型のデータについて、間隔尺度か比例尺度かを判定することが出来る。

[0051] 尺度水準判定部１９０６は、上述のように各カラムの尺度水準を判定した後、カラムに格納されているデータと尺度水準とを紐づけた情報を尺度水準データベース１９０７に格納する。例えば、図３の入力テーブル４００が入力データ１９０２であるとすると、尺度水準データベース内の尺度水準テーブル４１０のバリュー部４１５に各カラムの尺度水準を格納する。また、尺度水準判定部１９０６は、データ加工部１９０８にデータの加工を行うトリガを送信する。

[0052] データ加工部１９０８は、尺度水準判定部１９０６からトリガを受け付けた後に、各カラムの尺度水準を基にデータに適用可能な演算処理を施して各カラム内のデータを加工する。図８は、データ加工部１９０８が各データの加工をする際に用いる、尺度水準９０２及び演算タイプ９０３毎に適用可能加工演算９０４を格納したテーブル９０１の図である。データ加工部１９０８内にテーブル９０１が構築されていてもよいし、データ加工部１９０８の外からデータ加工の際に適用可能演算９０４を読みだしても構わない。データ加工部１９０８は、尺度水準データベース１９０７の尺度水準を読みだして、それぞれの尺度水準に適用可能演算９０４の処理を各カラムに行い、データを加工した結果を出力データベース１９０３に格納する。例えば、図３の加工データテーブル４３０のバリュー部４３５に演算を施したデータをカラムごとに格納する。加工データテーブル４３０は、出力データベース１９０３内に構築されている。

[0053] ここまで、各カラムについてそれぞれ処理を行い、出力もテーブルにすると記述したが、必ずしもカラム形式やテーブル形式である必要はなく、一定のデータの集合を定義できるものであれば形式は問わない。例えば、カラム

形式ではなく、リスト形式のデータ、またはデータの配列に対して処理を行っても構わない。

[0054] 図１１は、実施例１におけるデータ自動加工システムを実現するハードウェア構成の一例を示す図である。

[0055] 実施例１におけるハードウェア構成は、コンピュータシステム(計算機)を用いて実現され、少なくとも1組の、ＣＰＵ１２０１、ＲＯＭ１２０２、ＲＡＭ１２０３、キーボード１２０４、表示装置１２０５、ＨＤＤ１２０６、プリンタ１２０７、マウス１２０８、バス１２０９、ＤＢ１２１０、およびネットワーク１２１１から構成される。

[0056] ＲＯＭ１２０２は、データ自動加工システムのＯＳ(オペレーティングシステム)などを記憶する。ＲＡＭ１２０３は、データ自動加工に関するコンピュータソフトウェアを格納する。キーボード１２０４は、ＣＰＵ１２０１を操作する。ＨＤＤ１２０６は、入力データや加工データを格納する。表示装置１２０５は、入力データ、加工データ、またはデータ加工の処理の過程などをユーザに示す。マウス１２０８は、ＣＰＵ１２０１を操作する。バス１２０９は、各々のデータを通信するためのものである。ＤＢ１２１０は、各データを格納しておく。ネットワーク１２１１は、バス１２０９とＤＢ１２１０を繋ぐ。

[0057] データ自動加工システム１９０１において、ＣＰＵ１２０１で、ＲＡＭに格納されたデータ自動加工に関するコンピュータソフトウェアを実行することで、図１８に示した各機能を実現することができる。

[0058] このように、実施例１に表すデータ自動加工システム１９０１は、数字、文字、および符号に関するデータを受け付ける受付部１９０４と、データについて、データの型を判定するデータ型判定部１９０５と、データが数字型である場合に、データの分布に基づいてデータの尺度水準を判定する尺度水準判定部１９０６と、尺度水準に基づいてデータを加工するデータ加工部１９０８と、を備えている。

[0059] 係る構成により、本実施例に係るデータ自動加工システム１９０１は、デ

ータの性質を定める指標である尺度水準を自動で判定し、各データに適した方法によりデータ加工を行うことが可能となる。

[0060] ＜データが文字列型の場合の例＞

尺度水準判定部１９０６は、データ型判定部１９０５が、当該カラムのデータを文字列型と判定した場合、正規表現合致有無の判定を行う。

[0061] 正規表現合致有無の判定では、尺度水準判定部１９０６は各カラムに格納されているデータと予め設定されている正規表現との合致の判定を行い、その合致の有無に基づいて尺度水準を判定する。

[0062] 図６は、図４の正規表現合致判定５０４および尺度水準判定５０６の処理のフローの一例を示した図である。

[0063] ステップ７０１では、カラム内のデータが日付表現もしくは時刻表現であるかどうかを判定する。日付表現の例として、「2014/12/20」「2014-12-20」「14/12/20」「14-12-12」「Dec.20 2014」（2014年12月20日）などが挙げられる。時刻表現の例として、「15:47」「03:47 AM」（15時47分）、「16:01:42」「04:01:42」（16時01分42秒）などが挙げられる。日付表現もしくは時刻表現を持つかどうかの判定方法として、前述の表現例を正規表現で記述し、カラムに格納されているデータ内のすべての文字列が当該正規表現に合致するかを判定する方法がある。なお、時刻表現に関しては、後述の時間表現との差異を明確にするため、取りうる時刻の範囲に注意して、正規表現を記述する必要がある。また、時刻表現および時間表現のどちらにも該当するデータの場合、前述の等分散性の判定を用いて、等分散性を持つ場合に時刻表現であると判定し、等分散性を持たない場合に時間表現であると判定する方法もある。文字列型のデータの場合には、時刻表現若しくは時間表現のデータを数字型のデータに変換して等分散性の判定を行う。例えば、「12:30:00」の場合には「750」分のように変換する。ここでは、分単位の変換としているが、秒単位でも時間単位でも構わない。その後、そのデータの値と、その出現頻度から前述のデータの等分散性に関する分布を計算することにより、時刻表現と時間表現の判定を行う。ステップ７０１にて、日付表現もしく

は時刻表現と判定された場合には、当該カラムは間隔尺度であると判定され(ステップ７０７)、日付表現および時刻表現と判定されなかった場合には、ステップ７０２に遷移する。ステップ７０１によって、文字列型のデータを格納するカラムが間隔尺度であるかその他の尺度水準であるかを判定することが出来る。

[0064] ステップ７０２では、カラム内のデータが時間表現であるかどうかを判定する。時間表現を持つ文字列の例として、「9’’58」(9秒58)、「3’26’’00」「03:26」(3分26秒00)、「2:02’57」「02:02:57」(2時間02分57秒)などが挙げられる。時間表現であるかどうかの判定方法の例として、前述の表現例を正規表現で記述し、カラムに格納されているデータ内のすべての文字列が当該正規表現に合致するかを判定する方法がある。ステップ７０２にて、時間表現と判定された場合には、当該カラムは比例尺度であると判定され(ステップ７０６)、時間表現と判定されなかった場合には、ステップ７０３に遷移する。ステップ７０２によって、文字列型のデータを格納するカラムが比例尺度であるかその他の尺度水準を持つかの判定をすることができる。

[0065] ステップ７０３では、カラム内のデータがリスト表現であり、かつ単調変化するかを判定する。リスト表現を持つ文字列の例として、「1.***, 2.***, …」「1:***, 2:***, …」「A.***, B.***, …」「I.***, II.***, …」などが挙げられる。リスト表現であるかどうかの判定方法の例として、前述の表現例を正規表現で記述し、カラムに格納されているデータ内のすべての文字列が当該正規表現に合致するかを判定する方法がある。

[0066] 図１４は、単調変化をしているデータの分布を示す図である。

[0067] ここでは、横軸を各リストの数値(文字の場合は数値に変換)、縦軸をその値の出現頻度としてヒストグラムを作成している。

[0068] 単調変化であるとは、ヒストグラム１５１０のように横軸の値の増加に対して、縦軸の値が規則的に徐々に減少する単調減少であるか、もしくはヒストグラム１５２０のように横軸の値の増加に対して、縦軸の値が規則的に徐

々に増加する単調増加であるか、もしくはヒストグラム１５３０のように横軸の値の増加に対して、ただ一つのピークを持ち、ピークの前では単調増加、ピークの後では単調減少するか、の３つのいずれかに該当するかを言う。ステップ７０３にて、データがリスト表現を持ち、かつ単調変化であると判定された場合には、当該カラムは順序尺度であると判定され、判定されなかった場合には、当該カラムは名義尺度であると判定される(ステップ７０４及び７０５)。ステップ７０３によって、文字列型のデータが順序尺度であるか名義尺度であるかの判定を行うことができる。

[0069] 上記では、ステップ７０１～７０３を順番に適用して尺度水準を判定したが、このステップの順番は変わってもよい。またその際は、ステップ７０１～７０３すべてで当てはまらないと判定されたカラムが名義尺度であると判定する。

[0070] このように、データ自動加工システム１９０１は、データが文字列型である場合に、データの所定の正規表現との合致有無に基づいてデータの尺度水準を判定する尺度水準判定部を備えている。係る構成により、データが文字列型の場合にもデータの性質を定める指標である尺度水準を自動で判定し、各データに適した方法によりデータ加工を行うことが可能となる。

[0071] ＜加工演算の提示に関する変形例＞

ここでは、判定された尺度水準に合わせたデータの加工演算の提示に関する内容を説明する。基本的なシステム構成は図１８と同じであるが、以下の点が相違する。

[0072] 図１９は、加工演算の提示を行うデータ自動加工システムを表した図である。

[0073] データ自動加工システム１９０１は、入力データ１９０２を受け付け、データの尺度水準及びデータに適用可能な加工演算を判定し、表示装置１２０５に適用可能な加工演算を表示し、加工したデータを出力データベース１９０３に出力する。また、加工したデータを表示装置に表示してもよい。

[0074] データ自動加工システム１９０１は、図１８の構成に加えて、加工演算判

定部２００１、加工演算データベース２００２、及び表示部２００３を備えている。

[0075] 尺度水準判定部１９０６は、各カラムの尺度水準を判定して尺度水準データベースに格納した後に、加工演算判定部２００１に加工演算を行うトリガを送信する。

[0076] 加工演算判定部２００１は、尺度水準判定部１９０６からトリガを受け付けた後に尺度水準データベース１９０７から各カラムの尺度水準、および、加工演算データベース２００２からそれぞれの尺度水準に適用可能な演算を受け付け、各カラムの尺度水準から該カラム内のデータに適用可能な加工演算を選択し、表示部２００３に送信する。また、各カラムに適用可能な加工演算をデータ加工部１９０８に送信する。

[0077] 図７は、加工演算判定部２００１による処理のフローを示した図である。

[0078] 尺度水準受付ステップ８０１にて、尺度水準データベース１９０７から各カラムに入力されたデータとカラムに紐づけられた尺度水準を受け付ける。受付の形式は、例えば、図３の尺度水準テーブルのようにバリュー部４１５に各カラムの尺度水準が格納されている情報を受け付ける。

[0079] 次の加工演算抽出ステップ８０３は、ステップ８０１にて受け付けたテーブルのカラムの数だけ繰り返し処理される(ステップ８０２及びステップ８０４)。加工演算抽出ステップ８０３では、演算タイプ指定ステップ８１０にてユーザが指定した演算タイプと、尺度水準受付ステップ８０１にて受け付けた尺度水準を基に、加工演算データベース２００２から適用可能な演算を抽出する。

[0080] 演算タイプ指定ステップ８１０では、ユーザが任意の加工演算のタイプを指定する。指定は、キーボード１２０４やマウス１２０８から行うことが出来る。指定された加工演算のタイプはデータ自動加工システム１９０１内のタイプ受付部が受け付ける(図示せず)。加工演算のタイプは、例えば、正規化、量子化、代表値、または散布度などが挙げられる。図１６の演算タイプ（選択）１７０２の表示がタイプの指定の際のユーザインターフェースの例

である。

[0081] 加工演算データベース２００２は、尺度水準と演算タイプごとにカラムに適用可能な加工演算が格納されている。図８は、各尺度水準と演算タイプについて、尺度水準９０２及び演算タイプ９０３毎に適用可能加工演算９０４を格納したテーブル９０１の図である。加工演算データベース２００２は例えば、図８のようなテーブル９０１を有していいてもよい。

[0082] ステップ８０３は、ステップ８０１で受け付けた各カラムに入力されているデータ並びに各カラムの尺度水準、ステップ８１０にて指定された演算タイプ、および加工演算データベース２００２に格納されている適用可能加工演算を基に、各カラムのデータに適用可能な加工演算を抽出する。例えば、カラムの尺度水準が名義尺度であり、ユーザが指定した演算タイプが代表値であった場合には、最頻値の加工演算を抽出する。

[0083] 加工演算データベース２００２に格納されている演算タイプ、適用可能加工演算は、図８に示したものに依らず、適宜、演算タイプおよび適用可能加工演算を追加・削除してもよい。また、尺度水準と各演算タイプに適用可能な加工演算が紐づけられていれば、９０１のようなテーブル形式に限らない。

[0084] 加工演算送信ステップ８０５では、加工演算判定部２００１が抽出した適用可能な演算を表示部２００３及びデータ加工部１９０８に送信する。送信の形式として、例えば図３の適用可能加工演算テーブル４２０が挙げられる。

[0085] 表示部２００３は、加工演算判定部２００１から受信した各カラムに適用可能な加工演算を表示装置１２０５に送信する。表示装置１２０５は、表示部２００３から受信した適用可能な加工演算をユーザに表示する。例えば、図１６の適用可能加工演算１７０８のように表示する。このように各カラムに適用が可能な加工演算をバリュー部１７０９に表示する。

[0086] データ加工部１９０８は、加工演算判定部２００１から各カラムに適用可能な加工演算を受け付けて、各カラムに適用可能な加工演算を適用する。そ

の際、ユーザが指定した演算タイプに該当する適用可能加工演算９０４を適用してもよい。また、データ加工部１９０８が、表示装置１２０５に加工演算適用後のデータを送信し、表示装置１２０５がそのデータをユーザに表示してもよい。その際の表示例として、図１６のデータ加工結果１７１０が挙げられる。

[0087] このように、データ自動加工システム１９０１は、数字、文字、及び符号に関する各データに対して、各データの尺度水準と紐づけて各データに適用可能な加工演算を判定する加工演算判定部２００１と、適用可能な加工演算を画面に表示する表示部２００３とを備えている。

[0088] 係る構成により、データを機械学習や統計分析可能な形式へ適切に変換することができる加工演算を提示することができる。これにより、データマイニングや統計学の知識のない非専門家でもデータの加工演算を行うことができ、また、専門家の場合でも、入力データテーブルのカラム数が数百～となる場合には、１つ１つのカラムに対し適用可能な演算を考慮し、手動で設定するのは大きなコストとなっていたが、このコストを削減することが可能となる。さらに、意味のないデータ加工による分析の無駄、および分析結果の誤解を無くすことが可能となる。

[0089] ＜最適な加工演算の選択に関する変形例＞

あるカラム内のデータに対してあるデータ加工演算が適用できる場合でも、そのデータ加工演算の結果が不安定な場合があり、加工後の値が適切なものであるかどうかを分析者が手作業と直感により毎回判断する必要があった。

[0090] ここでは、適用可能な加工演算の中からの最適な加工演算の選択に関する内容を説明する。基本的なシステム構成は図１９と同様であるが、以下の点が相違する。

[0091] 図２０は、最適な加工演算を選択するデータ自動加工システム１９０１を表した図である。

[0092] データ自動加工システム１９０１は、入力データ１９０２を受け付け、カラムの尺度水準を判定し、各カラムに最適な加工演算を選択し、最適な加工

演算により加工したデータを出力データベース１９０３に出力する。

[0093] データ自動加工システム１９０１は、図１９の構成に加えて加工演算選択部２１０１を備えている。

[0094] 加工演算選択部２１０１は、加工演算判定部２００１が抽出した適用可能な加工演算の中から、各カラムに最も適用するのが適している加工演算を選択し、選択した加工演算をデータ加工部１９０８に送信する。

[0095] 図９は、加工演算選択部２１０１の処理のフローを示した図である。

[0096] 加工演算受付ステップ１００１にて、加工演算選択部２１０１は、加工演算判定部２００１から適用可能加工演算テーブル４２０を受け付ける。

[0097] 次の演算ロバスト性判定ステップ１００３と最適加工演算選択ステップ１００４は、適用可能加工演算テーブル４２０のカラムの数だけ繰り返し処理される(ステップ１００２及び１００５)。

[0098] 演算ロバスト性判定ステップ１００３は、適用可能加工演算テーブル４２０のバリュー部４２５に格納された適用可能加工演算について、演算のロバスト性を判定する。

[0099] 最適加工演算選択ステップ１００４は、演算ロバスト性判定ステップ１００３にて判定されたロバスト性に関する値を基に各カラムに最適な加工演算を選択する。

[0100] 最後に、最適加工演算送信ステップ１００６は、加工演算選択部２１０１により最適加工演算選択ステップ１００４が選択した各カラムに最適な加工演算をデータ加工部１９０８に送信する。

[0101] データ加工部１９０８は、受信した各カラムに最適な加工演算をカラム内の各データに施してデータを加工する。

[0102] 次に、図１０を用いて、演算ロバスト性判定ステップ１００３および最適加工演算選択ステップ１００４の処理のフローを説明する。

[0103] Ｎ分割ステップ１１０２、演算適用ステップ１１０４、及び分散計算ステップ１１０６は、は適用可能加工演算テーブルの各バリュー部に格納された適用可能加工演算の数だけ繰り返し処理される。

[0104] まず、Ｎ分割ステップ１１０２にてデータをランダムにＮ個の集合に分割する。Ｎは、ユーザが指定してもよいし、任意の数でもよい。例えば、５～１０個に分割することが考えられる。

[0105] 次の演算適用ステップ１１０４は、データの分割数Ｎだけ繰り返し処理される。

[0106] 演算適用ステップ１１０４は、分割後のデータに対して加工演算受付ステップ１００１で受け付けた適用可能な加工演算を施し、加工後のデータの値を計算する。

[0107] 分散計算ステップ１１０６では、Ｎ個の加工後のデータの値について分散を計算する。分散の計算方法は、既存の方法で構わない。

[0108] 最後に、分割値最小演算選択ステップ１１０８にて、分散計算ステップ１１０６で計算した分散値が最小となる加工演算を、最もロバスト性の高い演算と判定し、最適な加工演算として選択する。ここで、演算ロバスト性とは、演算適用後の各データの値のばらつきの小ささを示す性質のことをいう。

[0109] 上記では、分散を基に演算ロバスト性の判定に利用しているが、これは標準偏差でも同様に判定をすることが出来る。

[0110] ここまで、各カラムについてそれぞれ処理を行い、出力もテーブルにすると記述したが、必ずしもカラム形式やテーブル形式である必要はなく、一定のデータの集合を定義できるものであれば形式は問わない。例えば、カラム形式ではなく、リスト形式のデータ、またはデータの配列に対して処理を行っても構わない。

[0111] このように、データ自動加工システム１９０１は、数字、文字、及び符号に関する各データに対して、各データの尺度水準と紐づけて各データに適用可能な加工演算を判定する加工演算判定部２００１と、適用可能な加工演算のうち、演算適用後の各データの値のばらつきが最も小さい加工演算を選択する加工演算選択部２１０１と、各データの値のばらつきが最も小さい加工演算を適用することによりデータを加工するデータ加工部１９０８とを備えている。

[0112] 係る構成により、適用可能なデータ加工演算が複数ある場合に、最も演算適用後の値が安定している演算によりデータの加工をすることができる。これにより、データ分析の精度を高めることができる。さらに、データ加工の試行錯誤を行うことなく、精度の高いデータ分析を行うことができる。

実施例 2

[0113] 本発明のデータ自動加工システムの別の例を示す。

[0114] 実施例２は、データ自動加工システムについてのＧＵＩ(グラフィカルユーザインタフェース)に関する内容である。基本的な構成は図１９及び図２０と同様である。

[0115] 図１６に示すように、表示装置１２０４上に、データ加工操作をユーザが行うためのＧＵＩを表示し、ユーザからの入力を元に、ユーザからの入力があるごとに、データ加工結果を変えて表示装置１２０４に表示する。ユーザからの入力は、図１１のキーボード１２０３やマウス１２０６を介して行われる。

[0116] まず、ユーザが図３の入力テーブル４００をデータ自動加工システムに入力すると、入力テーブル表示部１７０１に表示される。

[0117] 演算タイプ選択部１７０２では、図８の適用可能加工演算格納テーブル９０１に事前定義された演算タイプのうち、いずれか１つをユーザが選択することができる。ユーザが選択した演算タイプは図７の演算タイプ指定ステップ８１０に入力される。

[0118] 尺度水準判定部１９０６が、入力テーブル４００の各カラムの尺度水準を判定すると、図３の尺度水準テーブル４１０を、尺度水準判定結果表示部１７０６に表示する。尺度水準選択部１７０７では、尺度水準判定部１９０６により自動判定された尺度水準が初期状態として設定されるが、ユーザが必要に応じて書き換えて尺度水準を再設定することもできる。

[0119] 演算タイプ選択部１７０２と、尺度水準選択部１７０７により、演算タイプと尺度水準が選択されると、加工演算判定部２００１により、図３の適用可能加工演算テーブル４２０が作成され、適用可能演算表示部１７０８に表

示される。各カラムの適用可能演算は、加工演算選択部２１０１にて、選択されたロバスト性の最も高い加工演算のみを表示しても良いし、ロバスト性の高い順に加工演算を表示してもよい。

[0120] 適用演算選択部１７０９では、適用可能な演算が複数ある場合に、いずれか１つの演算をユーザが選択することができる。

[0121] 適用演算選択部１７０９にて、演算が選択されると、データ加工結果表示部１７１０に、図３の加工テーブル４３０が表示される。

[0122] 係る構成により、各カラムの尺度水準を自動判定し、各カラムの尺度水準と、適用可能な演算をユーザに提示しながら、データ加工を進めることが可能となる。これにより、データ分析の知識のない非専門家でも、データの性質を把握しながら、容易にデータ分析を行うことができる。

実施例 3

[0123] 本発明のデータ自動加工システムを利用したデータ自動解析システムの例について説明する。

[0124] 図１７は、本実施例のデータ自動解析システムの構成図を表した図である。データ自動解析システム１８０１は、センサなどで取得したビックデータである入力データ１８０２を受け付け、データについて解析を行い、出力データ１８０３を出力する。データ自動解析システム１８０１は、データ前処理部１８０４、加工データデータベース１８０５、およびデータ解析部１８０６を備えている。

[0125] データ前処理部１８０４は、入力データ１８０２を受け付け、データ解析に適したデータに加工し、加工データベース１８０５に格納する。データ前処理部１８０４は、実施例１で説明したデータ自動加工システムを内部に有しており、入力データのカラムごとに尺度水準を判定し、適用可能演算を各カラムに施してデータを加工する。

[0126] データ解析部１８０６は、加工データベース１８０５に格納されているデータを基に、相関分析、回帰分析、またはクラスタリングなど既知の機械学習や統計分析の処理を行い、データを解析する。そして解析して得た結果を

出力データ１８０３として出力部(図示せず)により出力する。

[0127] データ自動解析システム１８０１を実現するハードウェア構成については、実施例１と同様に、図１１で示したものである。

[0128] このように、実施例３に表すデータ自動解析システム１８０１は、数字、文字、および符号に関するデータを受け付ける受付部１９０４と、データについて、データの型を判定するデータ型判定部１９０５と、データが数字型である場合に、前データの分布に基づいてデータの尺度水準を判定する尺度水準判定部１９０６と、尺度水準に基づいてデータを加工するデータ加工部１９０８と、加工部によって加工されたデータを解析するデータ解析部１８０６と、解析部によって解析されたデータを出力する出力部と、を有するデータ自動解析システム。係る構成により、データの前処理に関するユーザの負担を軽減させることができ、データ自動解析システムの前処理を容易にすることができる。

符号の説明

[0129] １００ 入力テーブル、 １０１ 作業者ＩＤ、 １０２ 処理数、
１０３ 商品ＩＤ、 １０４ 作業ＩＤ、 １１１ 作業者ＩＤ
１１２ 処理数、 １１３ 商品ＩＤ、 ４００ 入力テーブル
４０１ 処理数、 ４０２ 商品ＩＤ、 ４０３ 優先度
４０４ キー部、４０５ バリュー部、 ４１０ 尺度水準テーブル
４１１ 処理数、 ４１２ 商品ＩＤ、 ４１３ 優先度
４１４ キー部、 ４１５ バリュー部
４２０ 適用可能加工演算テーブル、 ４２１ 処理数
４２２ 商品ＩＤ、 ４２３ 優先度、 ４２４ キー部
４２５ バリュー部、 ４３０ 加工データテーブル、 ４３１ 処理数
４３２ 商品ＩＤ、 ４３３ 優先度、 ４３４ キー部
４３５ バリュー部、 ５０１ データ受付ステップ
５０２、５０７ カラム数繰り返し、 ５０３ データの型判定ステップ
５０４ 正規表現合致判定ステップ、 ５０５ データ分布判定ステップ

５０６　尺度水準判定ステップ、　５０８　尺度水準格納ステップ

６０１　連続性判定ステップ

６０２　中心性及び単調減少性判定ステップ

６０３　平滑性判定ステップ、　６０４　等分散性判定ステップ

６０５　名義尺度判定、　６０６　間隔尺度判定

６０７　比例尺度判定、　７０１　日付表現及び時刻表現判定ステップ

７０２　時間表現判定ステップ

７０３　リスト表現及び単調変化判定ステップ、　７０４　名義尺度判定

７０５　順序尺度判定、　７０６　比例尺度判定

７０７　間隔尺度判定、　８０１　尺度水準受付ステップ

８０２、８０４　カラム数繰り返し、　８０３　加工演算抽出ステップ

８０５　加工演算送信ステップ、　８１０　演算タイプ指定ステップ

９０１　適用可能加工演算格納テーブル、　９０２　尺度水準

９０３　演算タイプ、　９０４　適用可能加工演算

１００１　加工演算受付ステップ

１００２、１００５　カラム数繰り返し

１００３　演算ロバスト性判定ステップ

１００４　最適加工演算選択ステップ

１００６　最適加工演算送信ステップ

１１０１、１１０７　適用可能演算数繰り返し

１１０２　N分割ステップ、　１１０３、１１０５　分割データ数繰り返し

１１０４　演算適用ステップ、　１１０６　分散計算ステップ

１１０８　分散値最小演算選択ステップ、　１２０１　ＣＰＵ

１２０２　ＲＯＭ、　１２０３　ＲＡＭ、　１２０４　キーボード

１２０５　表示装置、　１２０６　ＨＤＤ、　１２０７　プリンタ

１２０８　マウス、　１２０９　バス、　１２１０　ＤＢ、

１２１１　ネットワーク

１３０１～１３０３　名義尺度を持つデータ分布のヒストグラム例図

１３０４～１３０６　比例尺度を持つデータ分布のヒストグラム例図

１３０７、１３０８　間隔尺度を持つデータ分布のヒストグラム例図

１４１０　等分散性を持つデータ分布のヒストグラム例図

１４２０　等分散性を持たないデータ分布のヒストグラム例図

１５１０、１５２０、１５３０　単調変化するデータ分布のヒストグラム例図

１６１０　入力テーブル、　１６２０　平均と分散を求めた後のテーブル

１７０１　入力テーブル表示部、　１７０２　演算タイプ選択部

１７０６　尺度水準結果表示部、　１７０７　尺度水準選択部

１７０８　適用可能演算表示部、　１７０９　適用演算選択部

１７１０　データ加工結果表示部、　１８０１　データ自動解析システム

１８０２　入力データ、　１８０３　出力データ

１８０４　データ前処理部、　１８０５　加工データ

１８０６　データ解析部、　１９０１　データ自動加工システム

１９０２　入力データ、　１９０３　出力データベース

１９０４　データ受付部、　１９０５　データ型判定部

１９０６　尺度水準判定部、　１９０７　尺度水準データベース

１９０８　データ加工部、　２００１　加工演算判定部

２００２　加工演算データベース、　２００３　表示部

２１０１　加工演算選択部。

請求の範囲

[請求項1] 数字、文字、および符号に関するデータを受け付ける受付部と、

前記データについて、前記データの型を判定するデータ型判定部と、

前記データが数字型である場合に、前記データの分布に基づいて前記データの尺度水準を判定する尺度水準判定部と、

前記尺度水準に基づいて前記データを加工するデータ加工部と、を有するデータ自動加工システム。

[請求項2] 請求項１に記載のデータ自動加工システムであって、

前記データの分布は、前記データの値と前記データの値の出現頻度に基づいたデータの頻度分布であることを特徴とするデータ自動加工システム。

[請求項3] 請求項２に記載のデータ自動加工システムであって、

前記尺度水準判定部は、前記データの値と前記データの値の出現頻度からなるヒストグラムの形状を基に前記データの尺度水準を判定することを特徴とするデータ自動加工システム。

[請求項4] 請求項２に記載のデータ自動加工システムであって、

前記尺度水準判定部は、前記データの頻度分布について連続性を有しているか判定し、前記データの頻度分布が連続性を有していないと判定した場合に前記データを名義尺度と判定することを特徴とするデータ自動加工システム。

[請求項5] 請求項２に記載のデータ自動加工システムであって、

前記尺度水準判定部は、前記データの頻度分布について連続性、中心性、単調減少性、および等分散性を有しているか判定し、前記データの頻度分布が連続性を有しており、中心性を有する若しくは単調減少性を有しており、かつ等分散性を有していない場合に前記データを比例尺度と判定することを特徴とするデータ自動加工システム。

[請求項6] 請求項１に記載のデータ自動加工システムであって、

前記尺度水準判定部は、前記データが文字列型である場合に、前記データと所定の正規表現との合致有無に基づいて前記データの尺度水準を判定することを特徴とするデータ自動加工システム。

[請求項7] 請求項６に記載のデータ自動加工システムであって、

前記尺度水準判定部は、前記データがリスト表現の正規表現と合致し、かつ、前記データの値と前記データの値の出現頻度に基づいたデータの頻度分布が単調変化を示している場合に前記データを順序尺度と判定することを特徴とするデータ自動加工システム。

[請求項8] 請求項６に記載のデータ自動加工システムであって、

前記尺度水準判定部は、前記データが時刻表現の正規表現および時間表現の正規表現と合致した場合に、前記データの値と前記データの値の出現頻度に基づいたデータの頻度分布について等分散性を有していると判定された場合に前記データが時刻表現であると判定し、前記データの値と前記データの値の出現頻度に基づいたデータの頻度分布について等分散性を有していないと判定された場合に前記データが時間表現であると判定し、前記データの尺度水準を判定することを特徴とするデータ自動加工システム。

[請求項9] 請求項１に記載のデータ自動加工システムであって、

数字、文字、及び符号に関する各データに対して、前記各データの尺度水準と紐づけて前記各データに適用可能な加工演算を判定する加工演算判定部と、

前記適用可能な加工演算を画面に表示する表示部と、を有することを特徴とするデータ自動加工システム。

[請求項10] 請求項１に記載のデータ自動加工システムであって、

数字、文字、及び符号に関する各データに対して、前記各データの尺度水準と紐づけて前記各データに適用可能な加工演算を判定する加工演算判定部と、

前記適用可能な加工演算のうち、演算適用後の前記各データの値の

ばらつきが最も小さい加工演算を選択する最適加工演算選択部を有し、

前記データ加工部は、前記各データの値のばらつきが最も小さい加工演算を適用することにより前記データを加工することを特徴とするデータ自動加工システム。

[請求項11] 数字、文字、および符号に関するデータを入力とするデータ自動加工方法であって、

前記データを受け付ける受付ステップと、

前記データについて、前記データの型を判定するデータ型判定ステップと、

前記データが数字型である場合に、前記データの分布に基づいて前記データの尺度水準を判定する尺度水準判定ステップと、

前記尺度水準に基づいて前記データを加工するデータ加工ステップと、を有することを特徴とするデータ自動加工方法。

[請求項12] 請求項１１に記載のデータ自動加工方法であって、

前記データの分布は、前記データの値と前記データの値の出現頻度に基づいたデータの頻度分布であることを特徴とするデータ自動加工方法。

[請求項13] 数字、文字、および符号に関するデータを受け付ける受付部と、

前記データについて、データの型を判定するデータ型判定部と、

前記データが数字型である場合に、前記データの分布に基づいて前記データの尺度水準を判定する尺度水準判定部と、

前記尺度水準に基づいて前記データを加工するデータ加工部と、

前記加工部によって加工されたデータを解析するデータ解析部と、

前記解析部によって解析されたデータを出力する出力部と、を有するデータ自動解析システム。

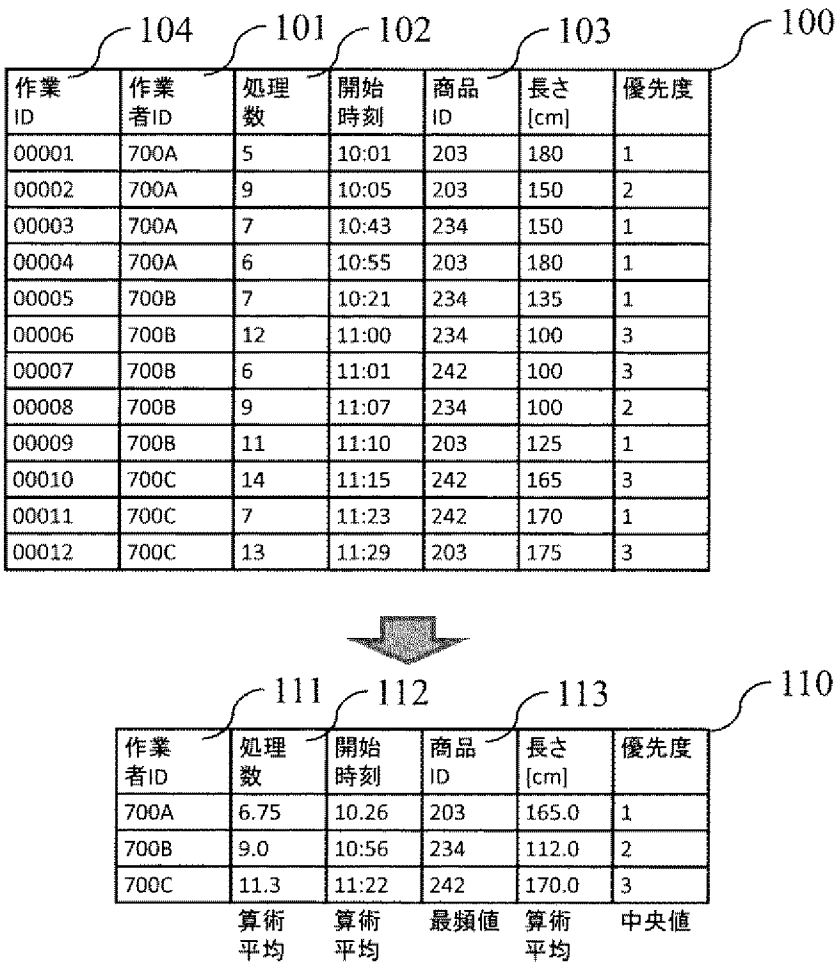
[請求項14] 請求項１３に記載のデータ自動解析システムであって、

前記データの分布は、前記データの値と前記データの値の出現頻度

に基づいたデータの頻度分布であることを特徴とするデータ自動解析システム。

[図1]

図1


104
101
102
103
100
作業ID
作業者ID
処理数
開始時刻
商品ID
長さ[cm]
優先度
00001 700A 5 10:01 203 180 1
00002 700A 9 10:05 203 150 2
00003 700A 7 10:43 234 150 1
00004 700A 6 10:55 203 180 1
00005 700B 7 10:21 234 135 1
00006 700B 12 11:00 234 100 3
00007 700B 6 11:01 242 100 3
00008 700B 9 11:07 234 100 2
00009 700B 11 11:10 203 125 1
00010 700C 14 11:15 242 165 3
00011 700C 7 11:23 242 170 1
00012 700C 13 11:29 203 175 3
111
112
113
110
作業者ID
処理数
開始時刻
商品ID
長さ[cm]
優先度
700A 6.75 10.26 203 165.0 1
700B 9.0 10:56 234 112.0 2
700C 11.3 11:22 242 170.0 3
算術平均
算術平均
最頻値
算術平均
中央値

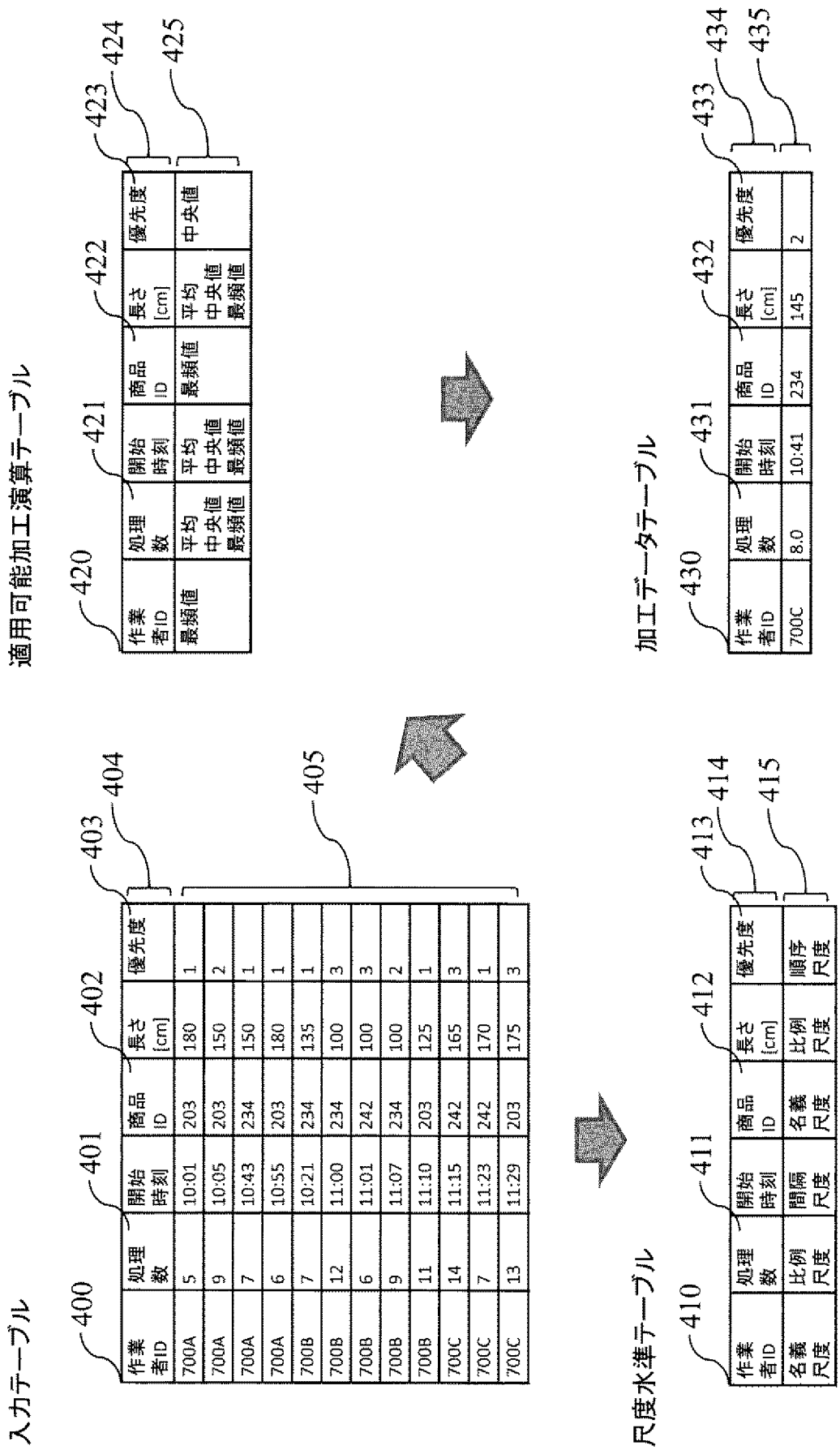
[図2]

図2

	概要	データの例	意味のあるデータ加工演算 (情報が多い、ロバスト性の高い順)
比例 尺度	比率が等しい 0に意味あり	作業効率[個/秒]:(0.353, 1.246...) 作業時間:(14:00, 30:00...)(840, 1800...) 作業数:(1, 5, 10...) 数量:(2, 40, 210...)	[正規化]基準値正規化、レンジ正規化 [量子化]個数等分割、値域等分割 [代表値]算術平均、幾何平均、中央値、最頻値 [散布度]変動率、分散、レンジ、ユニーク数
順序 尺度	大小・前後関係に意味あり 差に意味はない	従業員レベル:(5, 4, 3, 2, 1)(A, B, C, D, E) オーダー順:(1, 2, 3, 4, 5)	[正規化]--- [量子化]個数等分割 [代表値]中央値、最頻値 [散布度]ユニーク数
間隔 尺度	間隔が等しい 差に意味あり 0に意味はない	開始時刻:(13:15:00...)(131500...)(AM 1:15...) X座標:(0, 1, 2 ...) 日付:(2013/11/23,...)(20131123, ...)	[正規化]レンジ正規化 [量子化]個数等分割、値域等分割 [代表値]算術平均、中央値、最頻値 [散布度]分散、レンジ、ユニーク数
名義 尺度	他のデータとの区別のみに意味あり	従業員ID:(86100, 787813...)(62100, 11922...) 商品コード:(4329471, 4432719...)(12342, 15421...) 従業員名:(平山淳一、日立太郎....) 在庫フラグ:(0, 1)(T, F)	[正規化]--- [量子化]個数等分割 [代表値]最頻値 [散布度]ユニーク数

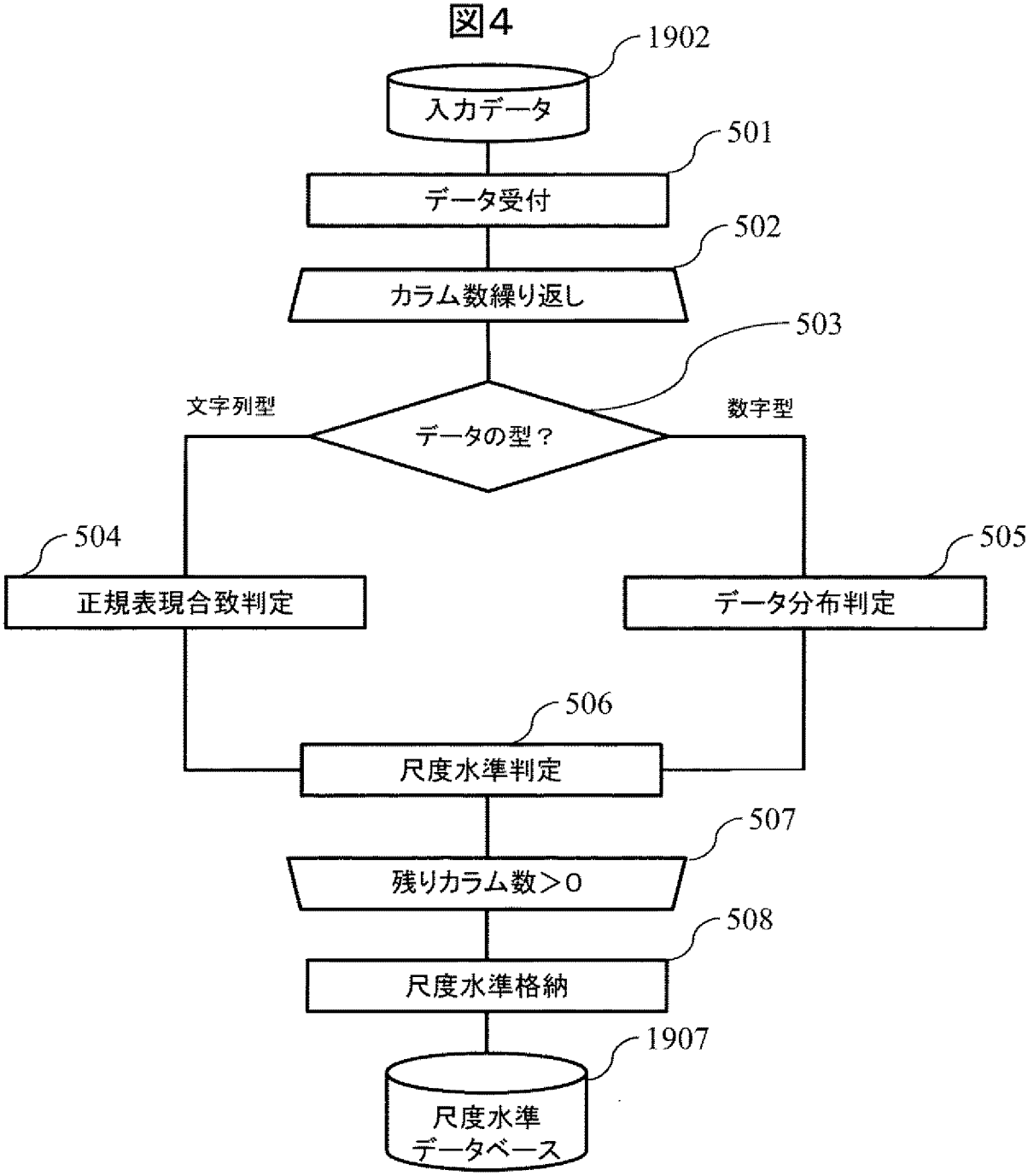
[図3]

図3


入力テーブル
400
401
402
403
404
405
作業者ID
処理数
開始時刻
商品ID
長さ[cm]
優先度
700A 5 10:01 203 180 1
700A 9 10:05 203 150 2
700A 7 10:43 234 150 1
700A 6 10:55 203 180 1
700B 7 10:21 234 135 1
700B 12 11:00 234 100 3
700B 6 11:01 242 100 3
700B 9 11:07 234 100 2
700B 11 11:10 203 125 1
700C 14 11:15 242 165 3
700C 7 11:23 242 170 1
700C 13 11:29 203 175 3
適用可能加工演算テーブル
420
421
422
423
424
425
作業者ID 最頻値
処理数 平均 中央値 最頻値
開始時刻 平均 中央値 最頻値
商品ID 最頻値
長さ[cm] 平均 中央値 最頻値
優先度 中央値
加工データテーブル
430
431
432
433
434
435
作業者ID 700C
処理数 8.0
開始時刻 10:41
商品ID 234
長さ[cm] 145
優先度 2
尺度水準テーブル
410
411
412
413
414
415
作業者ID 名義尺度
処理数 比例尺度
開始時刻 間隔尺度
商品ID 名義尺度
長さ[cm] 比例尺度
優先度 順序尺度

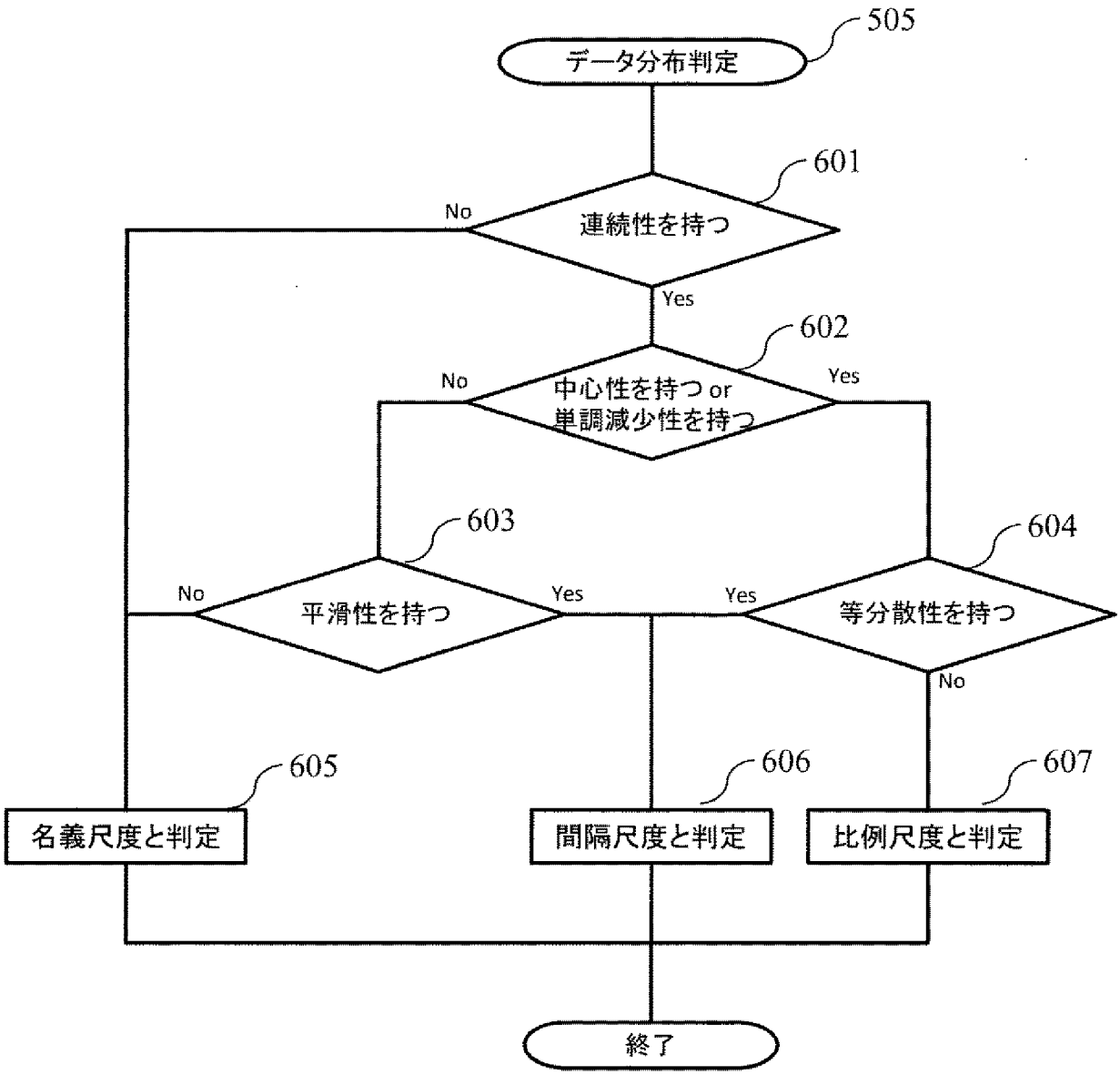
[図4]

図4


1902
入力データ
501
データ受付
502
カラム数繰り返し
503
文字列型
データの型？
数字型
504
正規表現合致判定
505
データ分布判定
506
尺度水準判定
507
残りカラム数＞0
508
尺度水準格納
1907
尺度水準
データベース

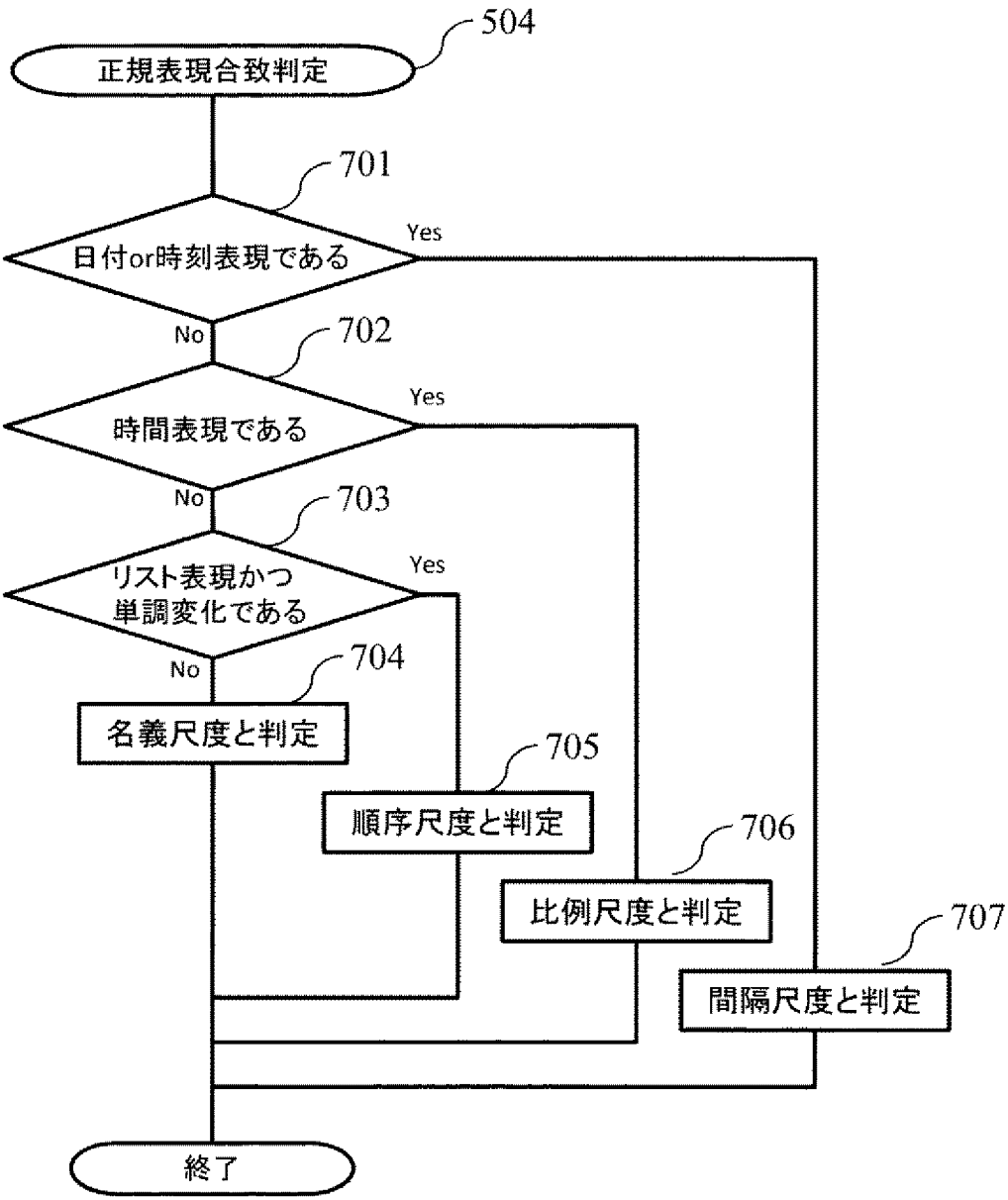
[図5]

図5


505
データ分布判定
601
連続性を持つ
No
Yes
602
中心性を持つ or
単調減少性を持つ
No
Yes
603
平滑性を持つ
No
Yes
604
等分散性を持つ
Yes
No
605
名義尺度と判定
606
間隔尺度と判定
607
比例尺度と判定
終了

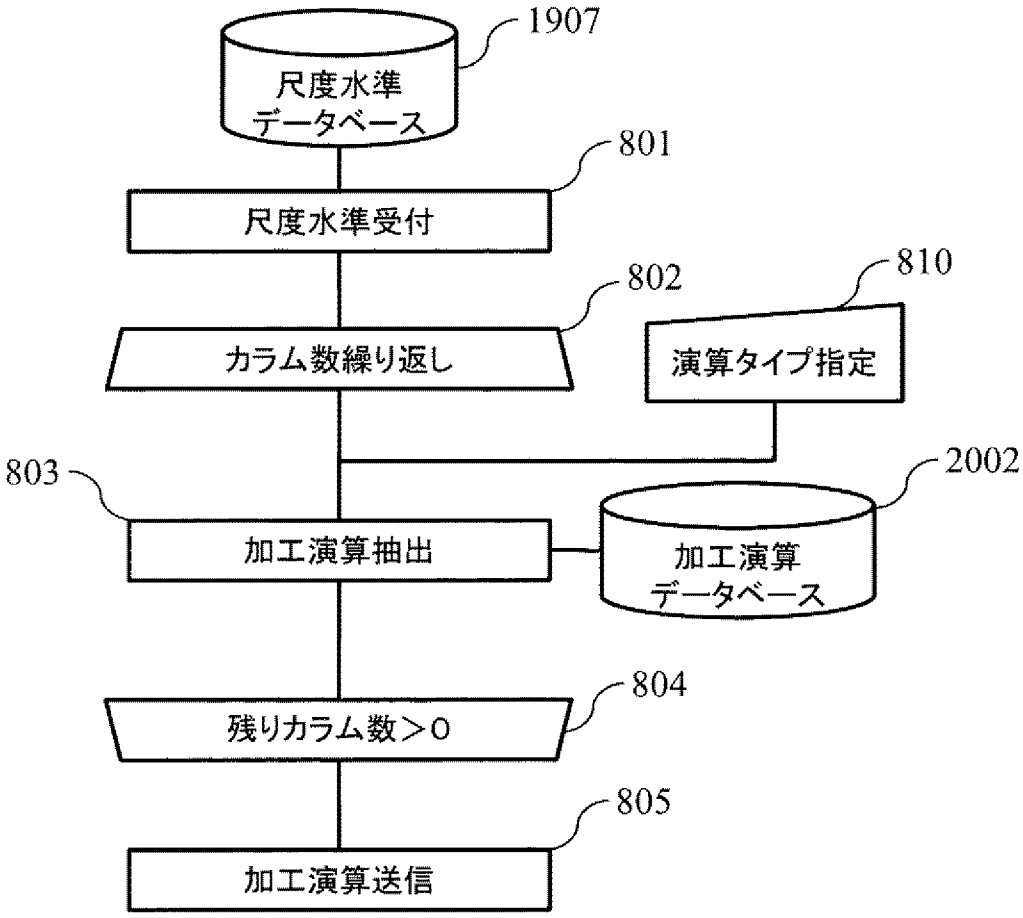
[図6]

図6


504
正規表現合致判定
701
日付or時刻表現である
Yes
No
702
時間表現である
Yes
No
703
リスト表現かつ
単調変化である
Yes
No
704
名義尺度と判定
705
順序尺度と判定
706
比例尺度と判定
707
間隔尺度と判定
終了

[図7]

図7


1907
尺度水準
データベース
801
尺度水準受付
802
カラム数繰り返し
810
演算タイプ指定
803
加工演算抽出
2002
加工演算
データベース
804
残りカラム数＞0
805
加工演算送信

[図8]

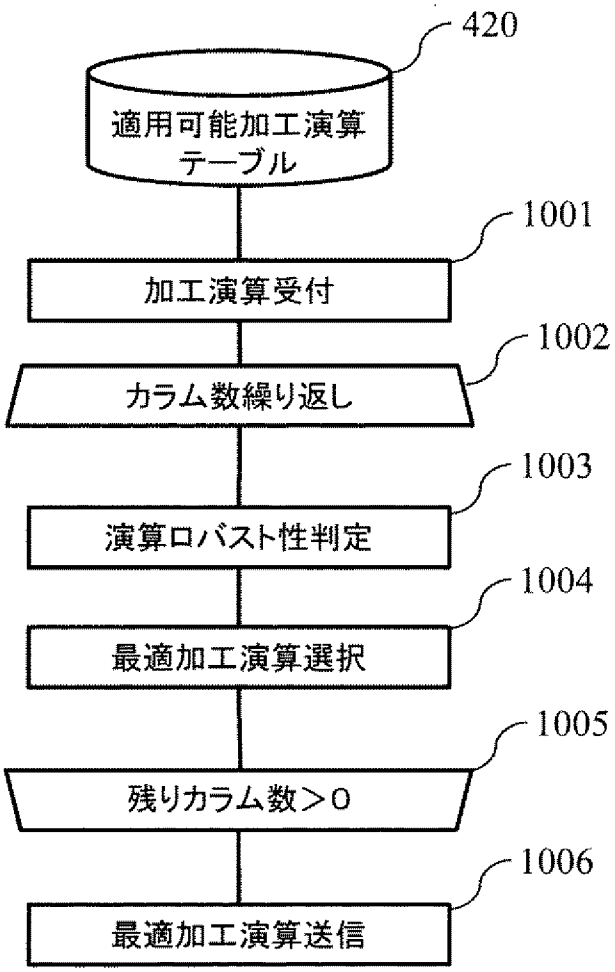
図8

902 903 904 901

尺度水準	演算タイプ	適用可能加工演算
比例尺度	正規化	基準値正規化、レンジ正規化
	量子化	個数等分割、値域等分割
	代表値	算術平均、幾何平均、中央値、最頻値
	散布度	変動率、分散、レンジ、ユニーク数
間隔尺度	正規化	レンジ正規化
	量子化	個数等分割、値域等分割
	代表値	算術平均、中央値、最頻値
	散布度	分散、レンジ、ユニーク数
順序尺度	正規化	ーー
	量子化	個数等分割
	代表値	中央値、最頻値
	散布度	ユニーク数
名義尺度	正規化	ーー
	量子化	個数等分割
	代表値	最頻値
	散布度	ユニーク数

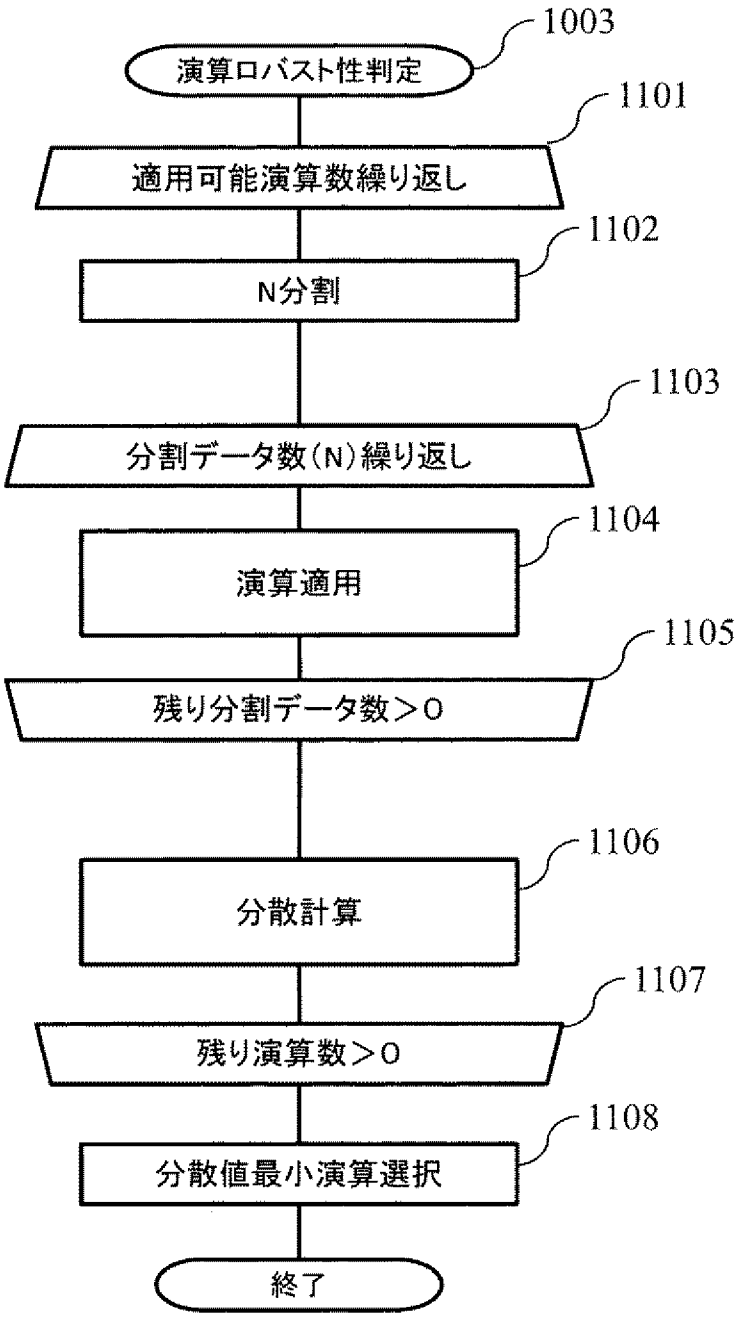
[図9]

図9


420
適用可能加工演算テーブル
1001
加工演算受付
1002
カラム数繰り返し
1003
演算ロバスト性判定
1004
最適加工演算選択
1005
残りカラム数＞0
1006
最適加工演算送信

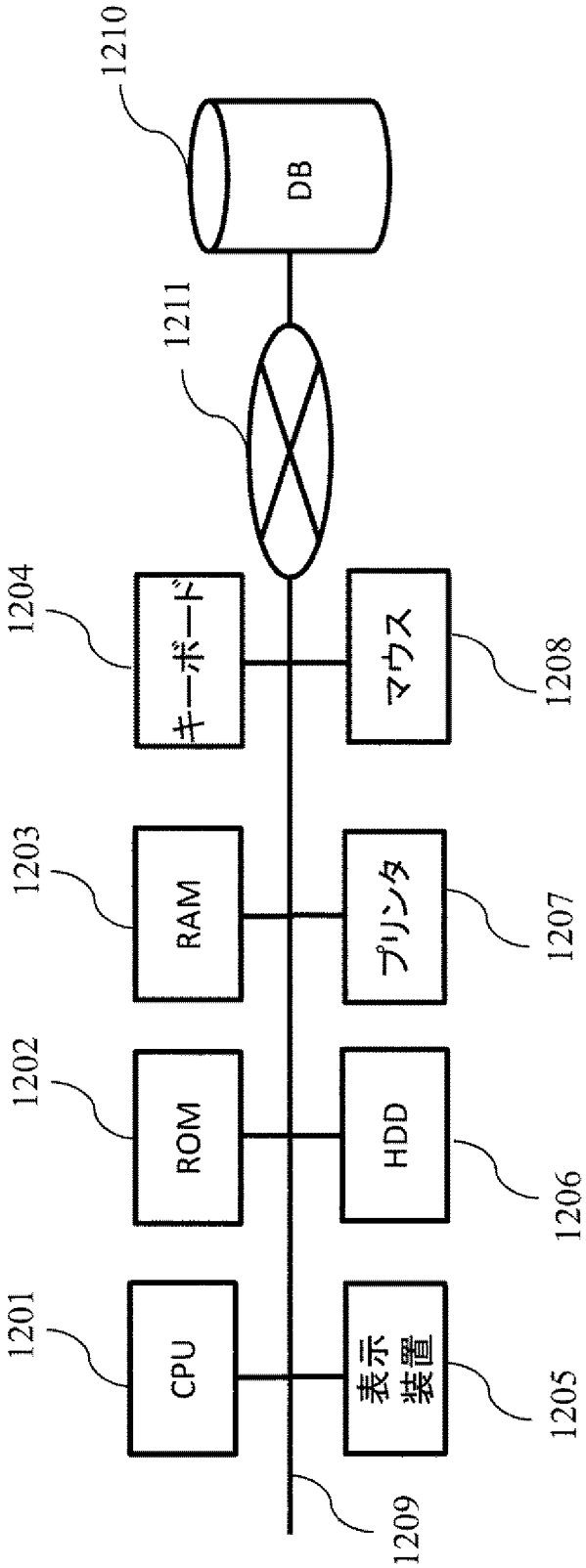
[図10]

図10


1003
演算ロバスト性判定
1101
適用可能演算数繰り返し
1102
N分割
1103
分割データ数(N)繰り返し
1104
演算適用
1105
残り分割データ数＞0
1106
分散計算
1107
残り演算数＞0
1108
分散値最小演算選択
終了

[図11]

図11


1210
DB
1211
1204
キーボード
マウス
1208
1203
RAM
プリンタ
1207
1202
ROM
HDD
1206
1201
CPU
表示装置
1205
1209

[図12]

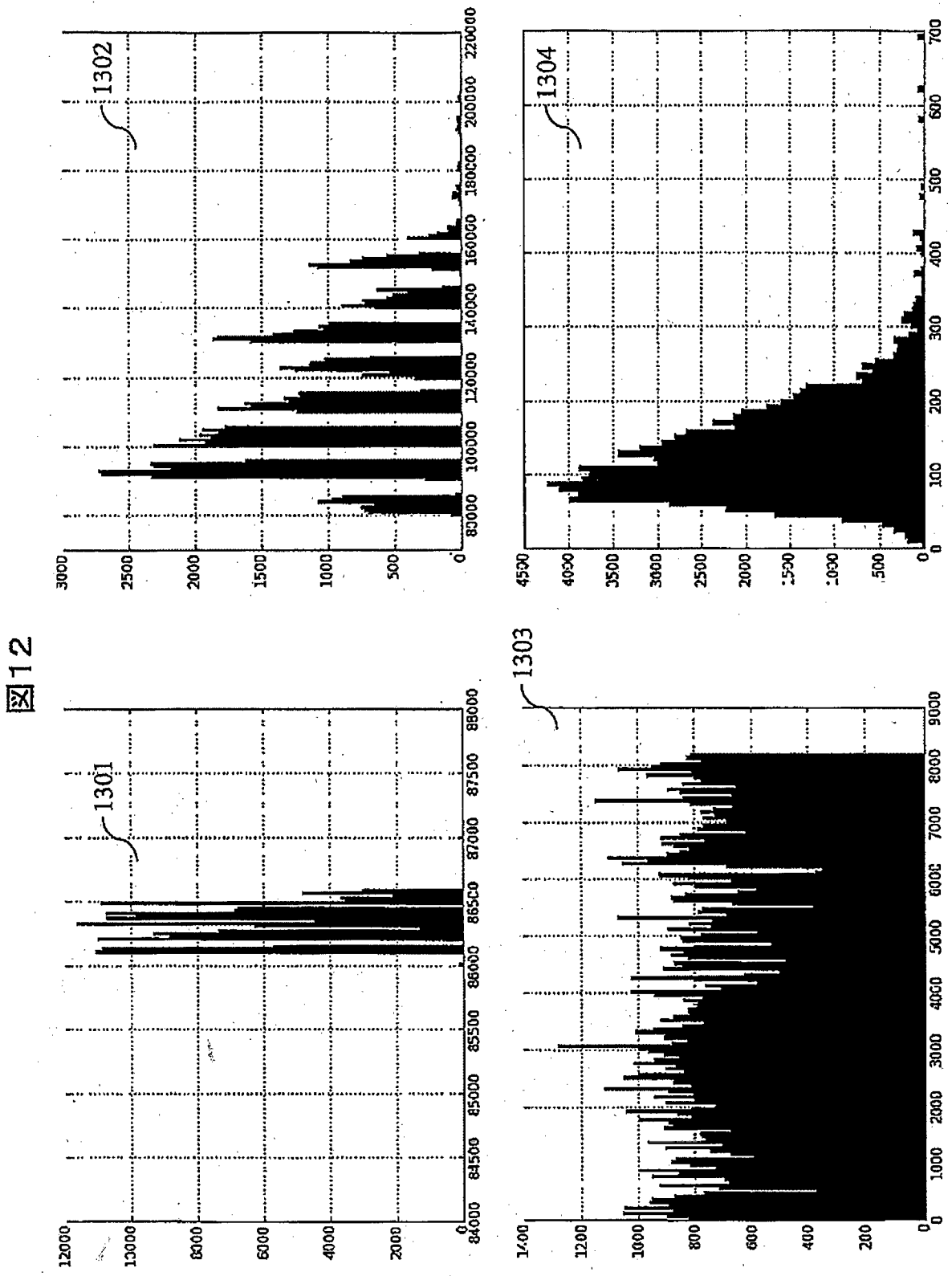
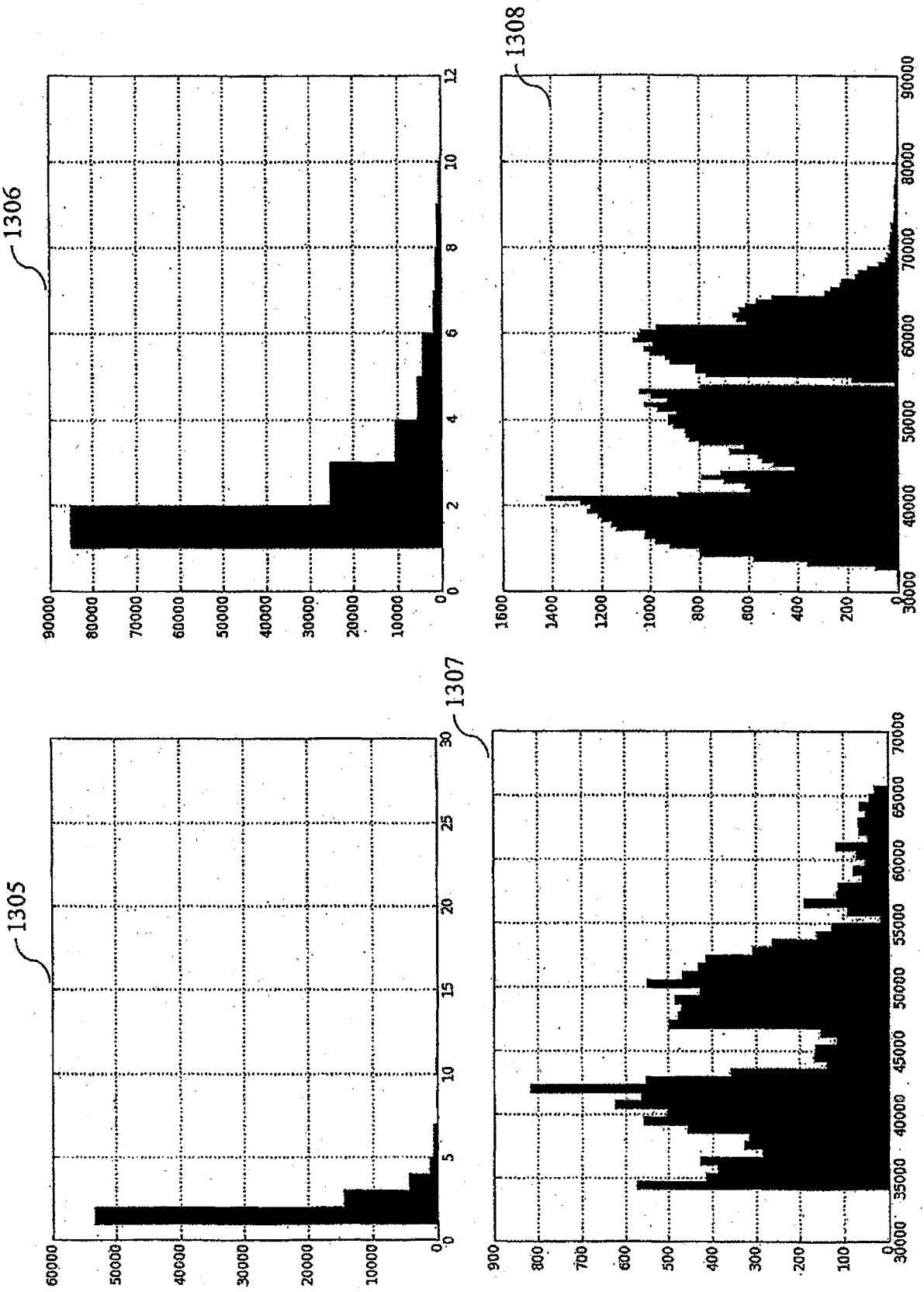


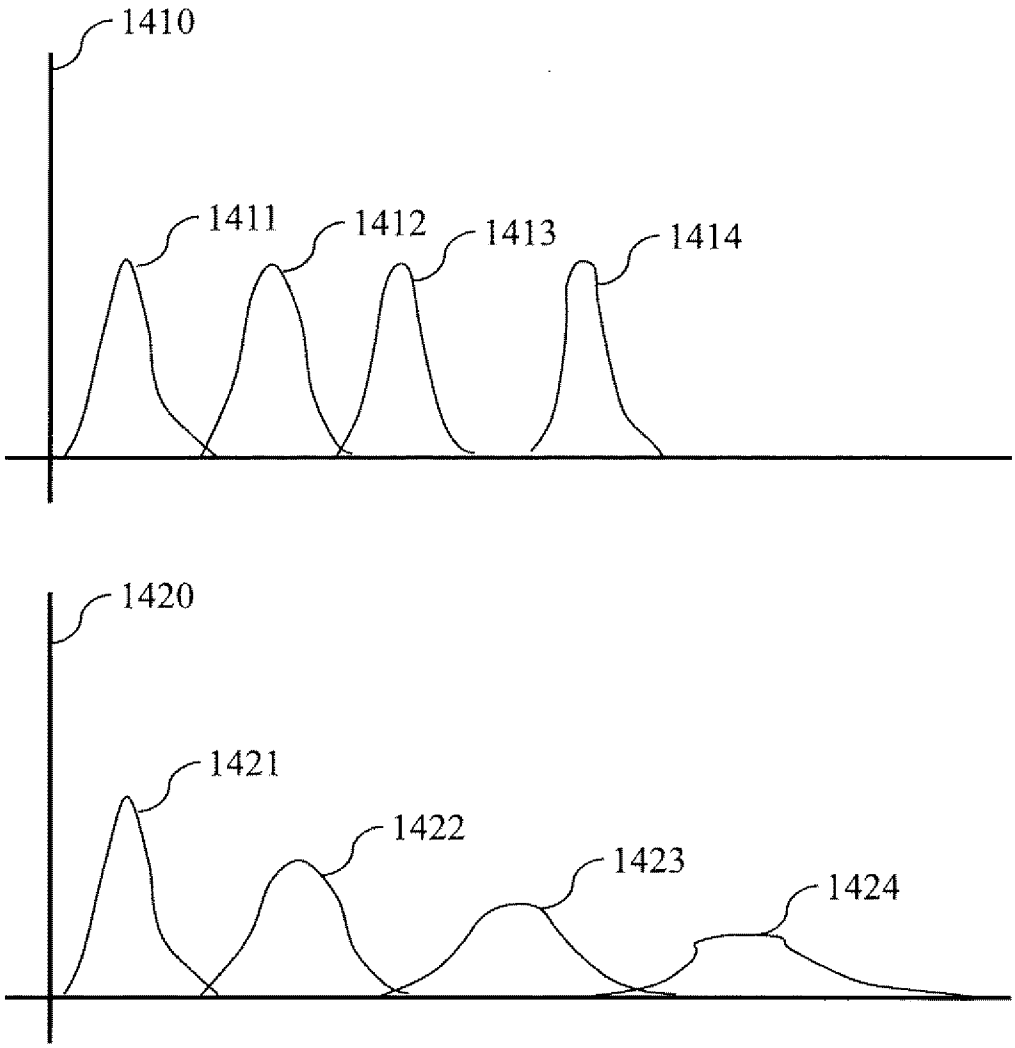
図12

[図12] つづき



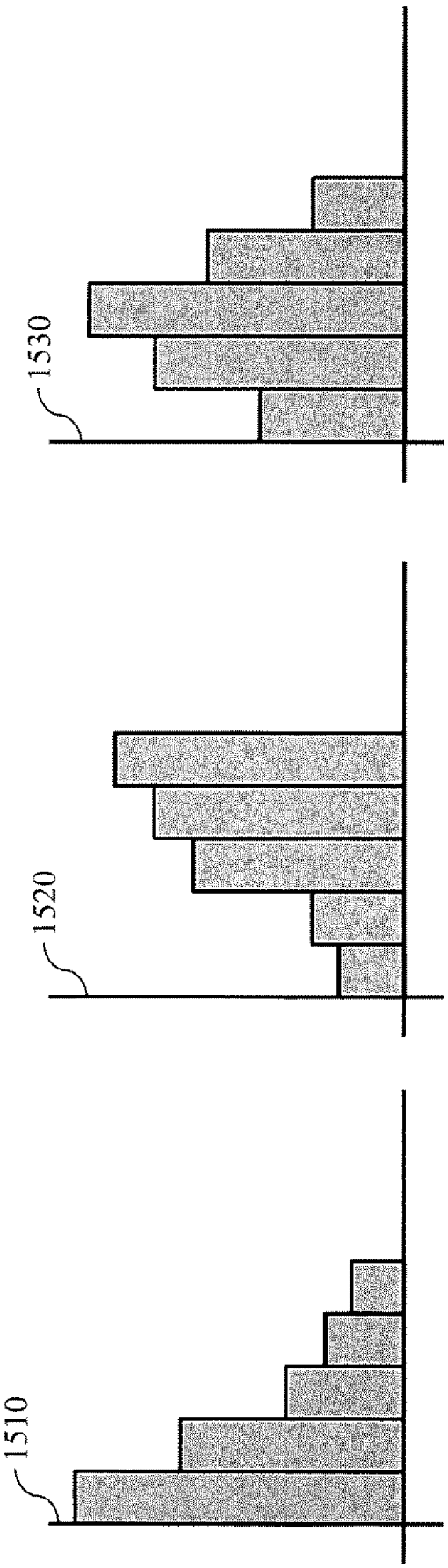
[図13]

図13



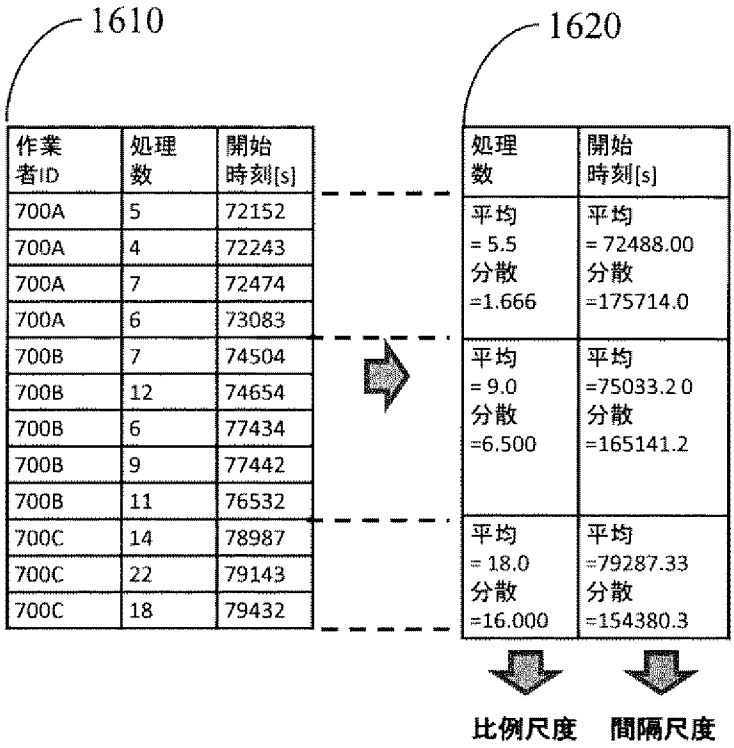
[図14]

図14



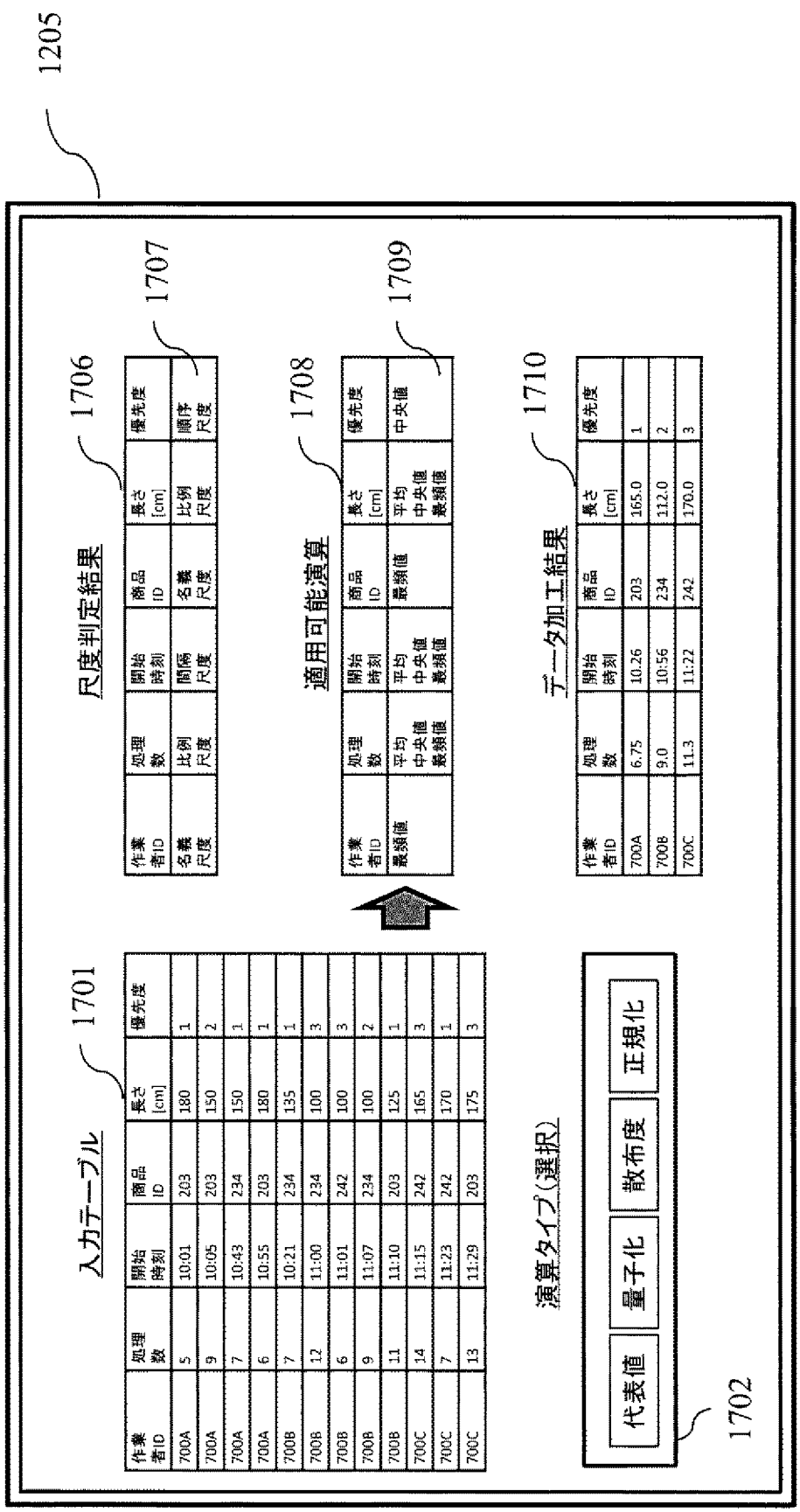
[図15]

図15


1610
1620
作業者ID
処理数
開始時刻[s]
700A 5 72152
700A 4 72243
700A 7 72474
700A 6 73083
700B 7 74504
700B 12 74654
700B 6 77434
700B 9 77442
700B 11 76532
700C 14 78987
700C 22 79143
700C 18 79432
処理数
開始時刻[s]
平均 = 5.5 分散 =1.666
平均 = 72488.00 分散 =175714.0
平均 = 9.0 分散 =6.500
平均 =75033.2 0 分散 =165141.2
平均 = 18.0 分散 =16.000
平均 =79287.33 分散 =154380.3
比例尺度
間隔尺度

[図16]

図16


1205
入力テーブル
1701
作業者ID | 処理数 | 開始時刻 | 商品ID | 長さ[cm] | 優先度
700A | 5 | 10:01 | 203 | 180 | 1
700A | 9 | 10:05 | 203 | 150 | 2
700A | 7 | 10:43 | 234 | 150 | 1
700A | 6 | 10:55 | 203 | 180 | 1
700B | 7 | 10:21 | 234 | 135 | 1
700B | 12 | 11:00 | 234 | 100 | 3
700B | 6 | 11:01 | 242 | 100 | 3
700B | 9 | 11:07 | 234 | 100 | 2
700B | 11 | 11:10 | 203 | 125 | 1
700C | 14 | 11:15 | 242 | 165 | 3
700C | 7 | 11:23 | 242 | 170 | 1
700C | 13 | 11:29 | 203 | 175 | 3
演算タイプ(選択)
代表値
量子化
散布度
正規化
1702
尺度判定結果
1706
作業者ID | 処理数 | 開始時刻 | 商品ID | 長さ[cm] | 優先度
名義尺度 | 比例尺度 | 間隔尺度 | 名義尺度 | 比例尺度 | 順序尺度
1707
適用可能演算
1708
作業者ID | 処理数 | 開始時刻 | 商品ID | 長さ[cm] | 優先度
最頻値 | 平均 中央値 最頻値 | 平均 中央値 最頻値 | 最頻値 | 平均 中央値 最頻値 | 中央値
1709
データ加工結果
1710
作業者ID | 処理数 | 開始時刻 | 商品ID | 長さ[cm] | 優先度
700A | 6.75 | 10:26 | 203 | 165.0 | 1
700B | 9.0 | 10:56 | 234 | 112.0 | 2
700C | 11.3 | 11:22 | 242 | 170.0 | 3

[図17]

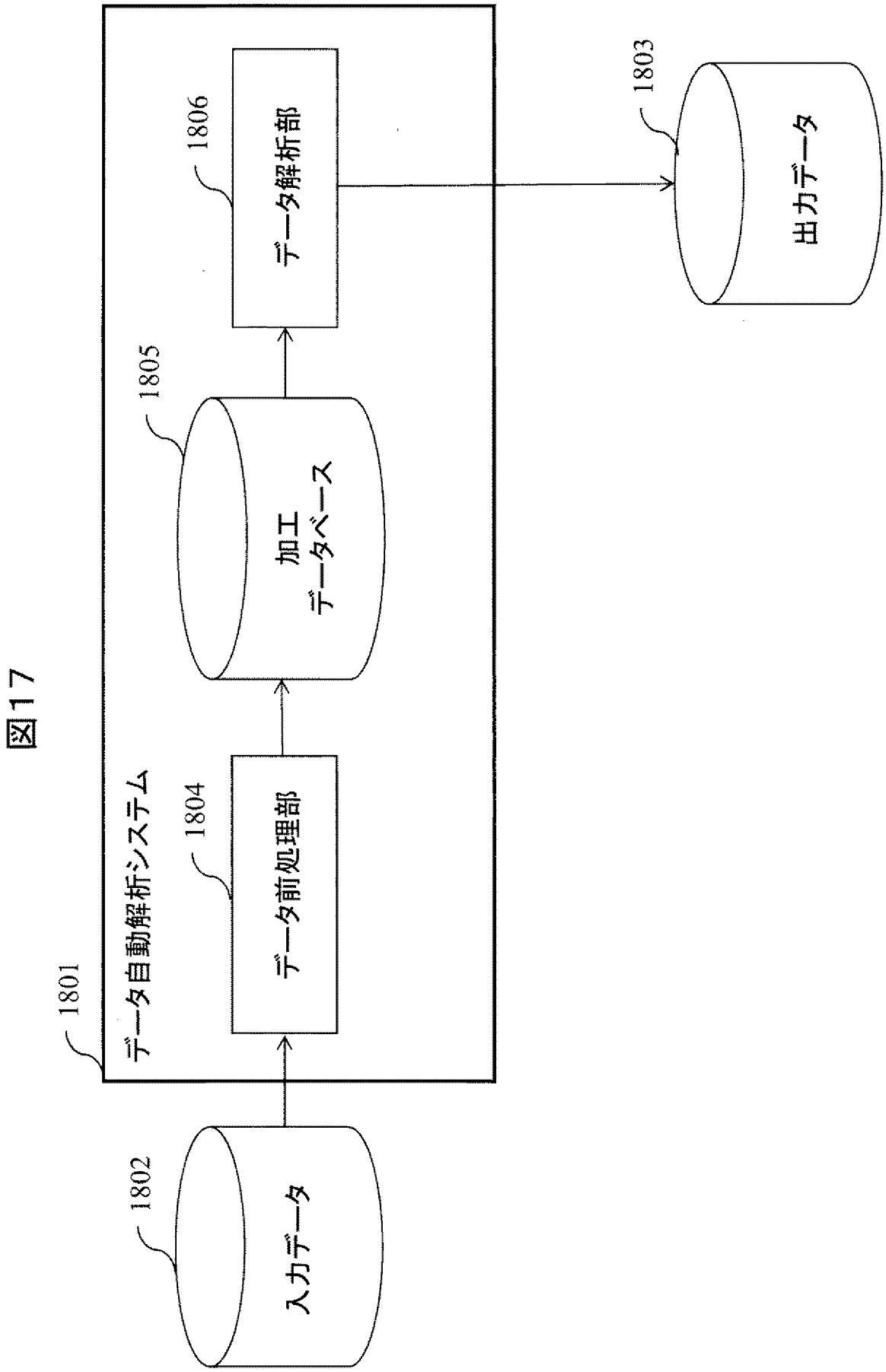
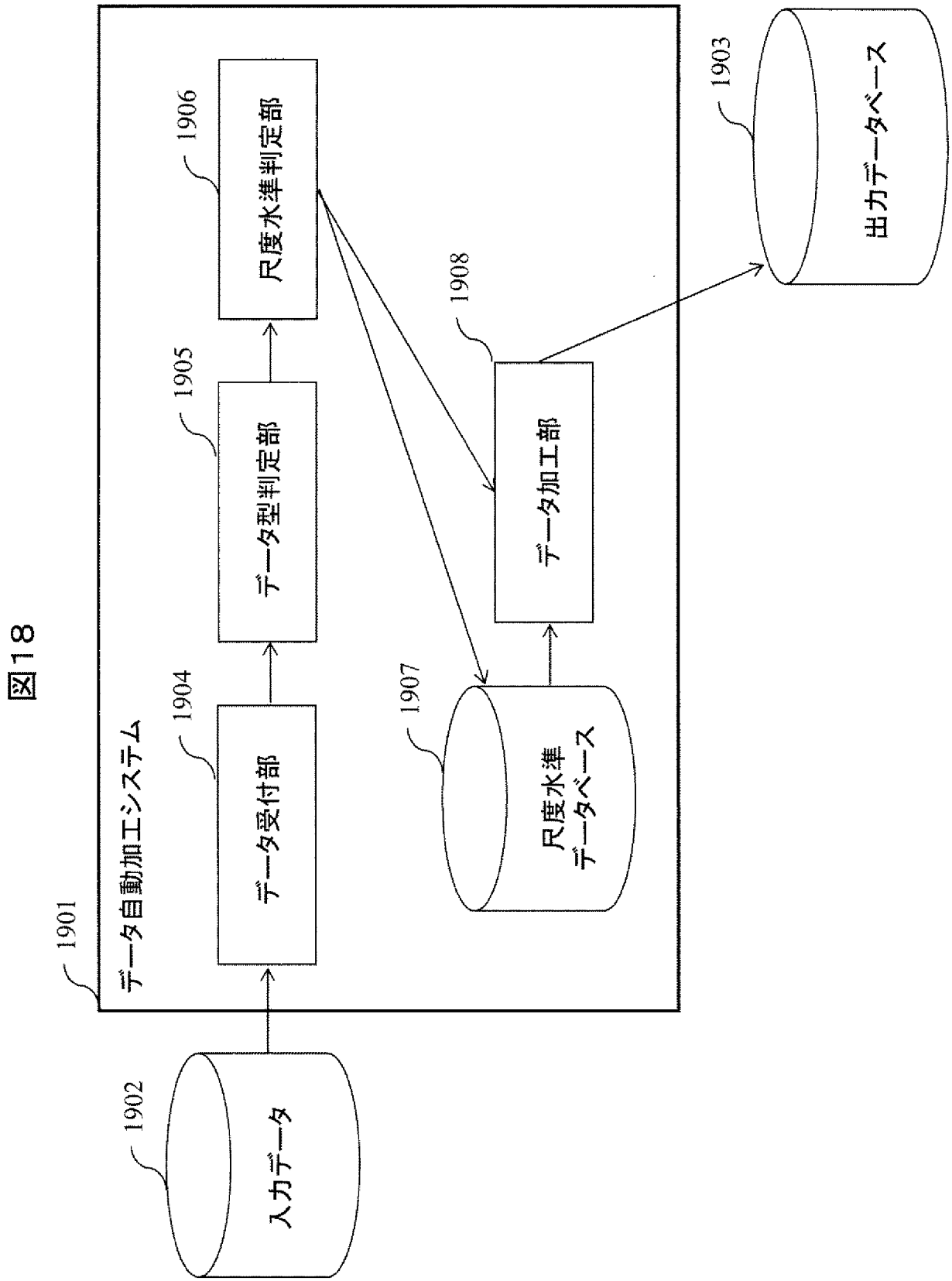

1801
データ自動解析システム
1802
入力データ
1804
データ前処理部
1805
加工
データベース
1806
データ解析部
1803
出力データ

図17

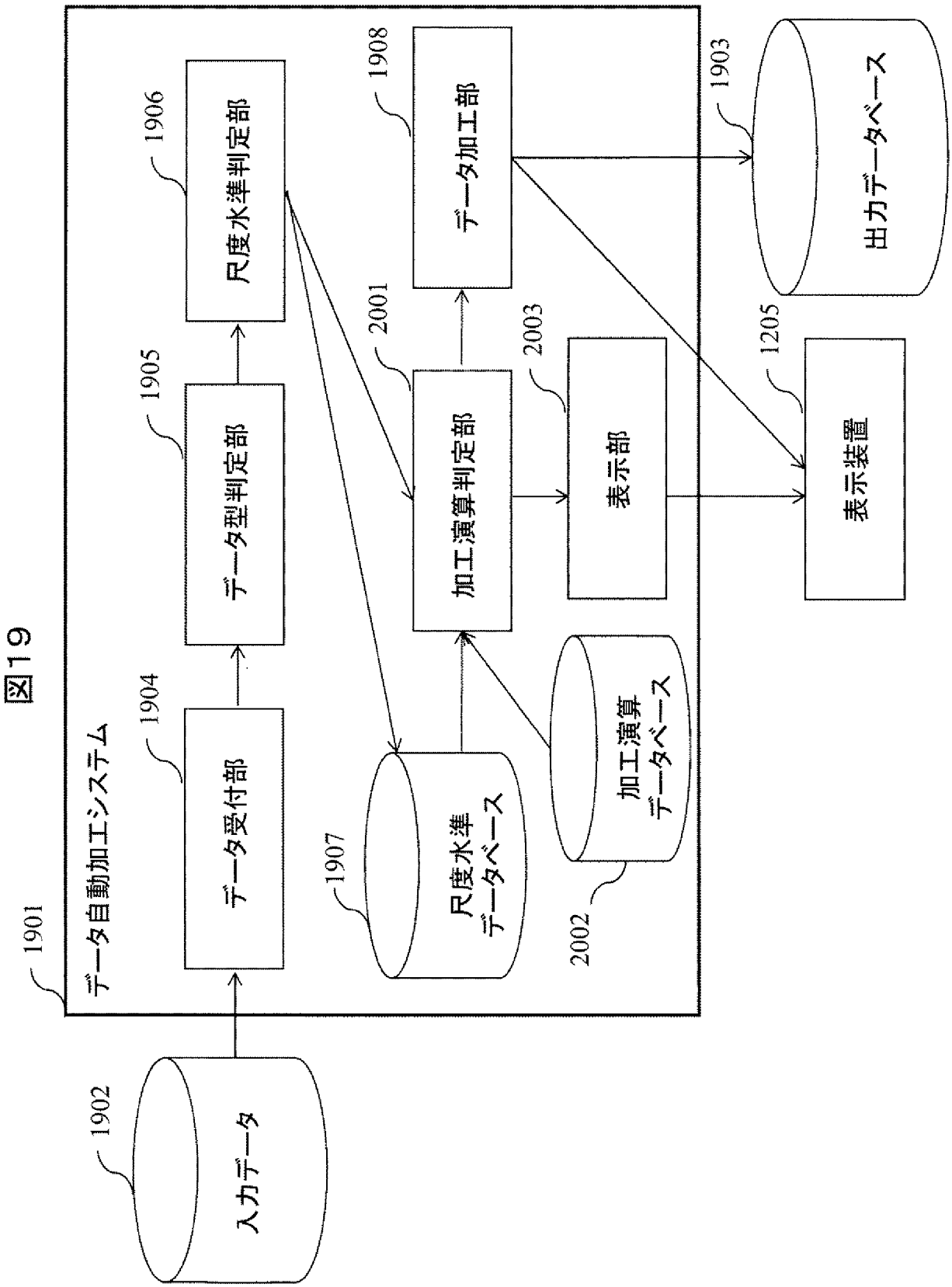
[図18]

図18


1901
データ自動加工システム
1902
入力データ
1903
出力データベース
1904
データ受付部
1905
データ型判定部
1906
尺度水準判定部
1907
尺度水準
データベース
1908
データ加工部

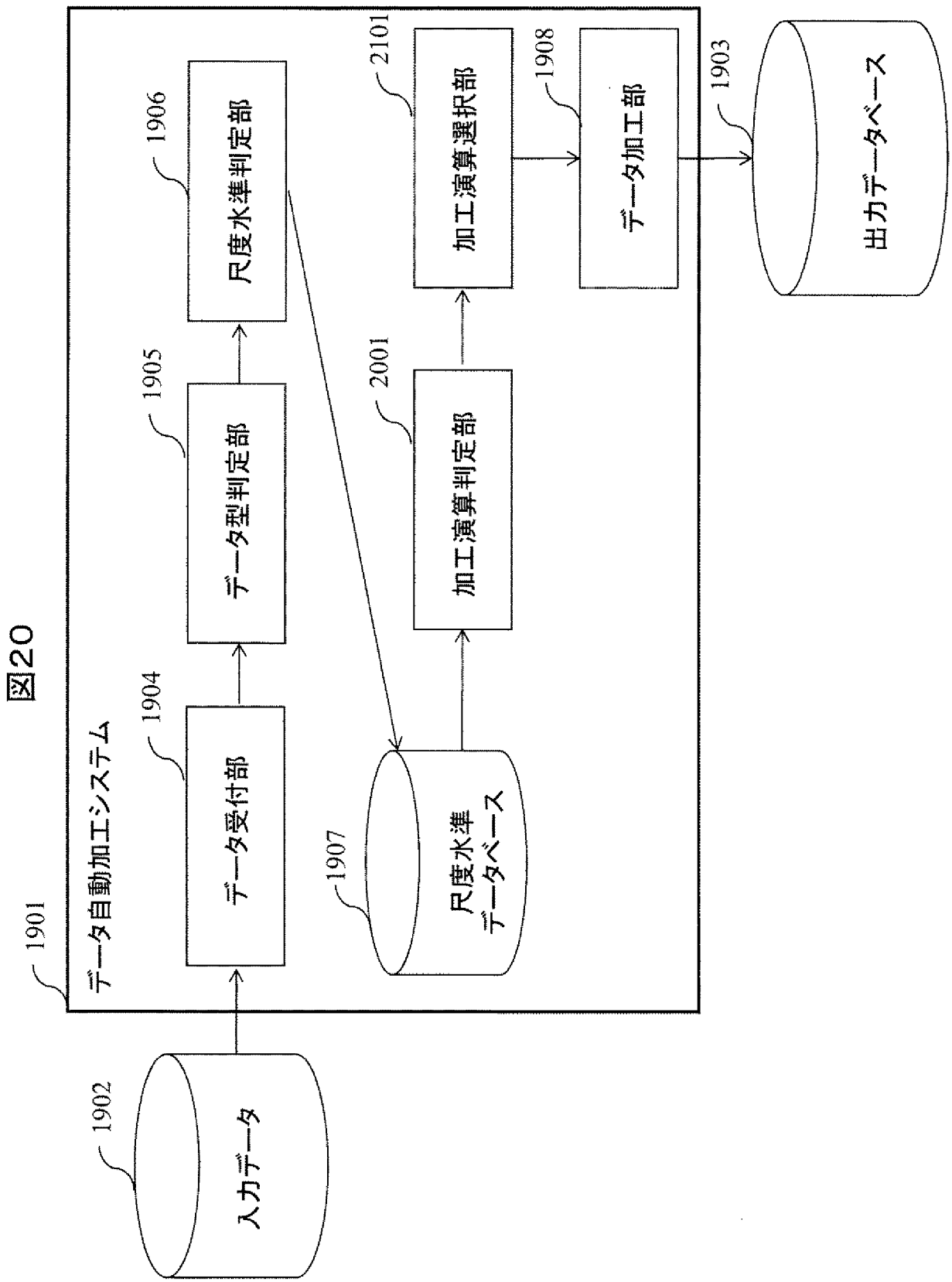
[図19]

図19


データ自動加工システム
1901
入力データ
1902
データ受付部
1904
データ型判定部
1905
尺度水準判定部
1906
尺度水準
データベース
1907
データ加工部
1908
出力データベース
1903
加工演算判定部
2001
加工演算
データベース
2002
表示部
2003
表示装置
1205

[図20]

図20


データ自動加工システム
1901
入力データ
1902
データ受付部
1904
データ型判定部
1905
尺度水準判定部
1906
尺度水準
データベース
1907
加工演算判定部
2001
加工演算選択部
2101
データ加工部
1908
出力データベース
1903

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2015/061778

A. CLASSIFICATION OF SUBJECT MATTER
G06Q50/10(2012.01)i, *G06F17/30*(2006.01)i, *G06F19/00*(2011.01)i

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)
G06Q50/10, G06F17/30, G06F19/00

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

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2015
Kokai Jitsuyo Shinan Koho	1971-2015	Toroku Jitsuyo Shinan Koho	1994-2015

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	JP 2015-32013 A (KDDI Corp.), 16 February 2015 (16.02.2015), particularly, paragraphs [0023] to [0025] (Family: none)	1-7,9-14 8
Y A	JP 2000-353163 A (Justsystem Corp.), 19 December 2000 (19.12.2000), particularly, paragraphs [0032], [0039] (Family: none)	1-7,9-14 8
Y A	written by KELLY Diane, translated by Hideo JOHO, Methods for Evaluating Interactive Information Retrieval Systems with Users, Maruzen Publishing Co., Ltd., 20 April 2013 (20.04.2013), pages 45 to 47	4-5 1-3,6-14

☒ Further documents are listed in the continuation of Box C. ☐ 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
06 July 2015 (06.07.15)	14 July 2015 (14.07.15)

Name and mailing address of the ISA/	Authorized officer
Japan Patent Office 3-4-3,Kasumigaseki,Chiyoda-ku, Tokyo 100-8915,Japan	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2015/061778

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	written by Berry, Michael J.A., translated by Atsushi EHARA, Mastering Data Mining <Riron-hen>, 1st edition, Kaibundo Shuppan Kabushiki Kaisha, 30 October 2002 (30.10.2002), pages 204 to 209	6-7 1-5,8-14

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１５／０６１７７８

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06Q50/10(2012.01)i, G06F17/30(2006.01)i, G06F19/00(2011.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06Q50/10, G06F17/30, G06F19/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１５年
日本国実用新案登録公報	１９９６－２０１５年
日本国登録実用新案公報	１９９４－２０１５年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Y	JP 2015-32013 A（ＫＤＤＩ株式会社）	1-7,9-14
A	2015.02.16, 特に段落 23-25（ファミリーなし）	8
Y	JP 2000-353163 A（株式会社ジャストシステム）	1-7,9-14
A	2000.12.19, 特に段落 32,39（ファミリーなし）	8
Y	KELLY Diane 著, 上保秀夫訳, インタラクティブ情報検索システムの評価, 丸善出版株式会社,	4-5
A	2013.04.20, pp. 45-47	1-3,6-14

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの

「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「Ｏ」口頭による開示、使用、展示等に言及する文献

「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
０６．０７．２０１５	１４．０７．２０１５

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｌ　４５３４
大野　朋也
電話番号　０３－３５８１－１１０１　内線　３５６２

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Y	ベリー　マイケルＪ．Ａ．著，江原淳訳，マスタリング・データマイニング　＜理論編＞，第１版，海文堂出版株式会社，2002.10.30，pp.204-209	6-7
A		1-5,8-14