

(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
2020年6月4日(04.06.2020)



(10) 国際公開番号

WO 2020/110239 A1

(51) 国際特許分類:
G06Q 10/10 (2012.01)

(21) 国際出願番号: PCT/JP2018/043831

(22) 国際出願日: 2018年11月28日(28.11.2018)

(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: 富士通株式会社(FUJITSU LIMITED)
[JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 Kanagawa (JP).

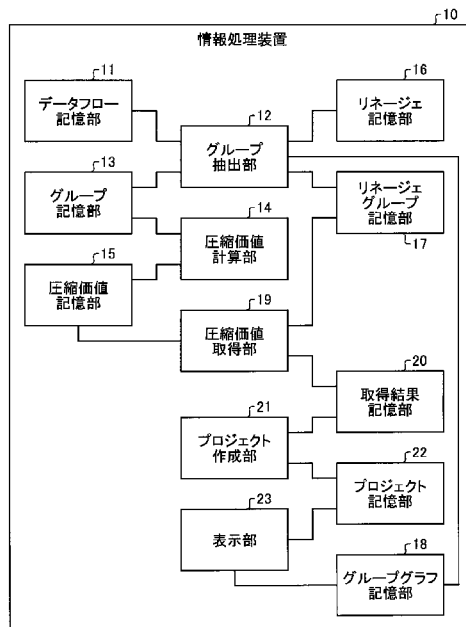
(72) 発明者: 北野 貴之 (KITANO, Takayuki);
〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内 Kanagawa (JP).

(74) 代理人: 特許業務法人酒井国際特許事務所 (SAKAI INTERNATIONAL PATENT OFFICE); 〒1000013 東京都千代田区霞が関3丁目8番1号 虎の門三井ビルディング 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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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,

(54) Title: INFORMATION PROCESSING DEVICE AND LINEAGE PROGRAM

(54) 発明の名称: 情報処理装置及びリネージュプログラム



- 10 Information processing device
- 11 Data flow storage unit
- 12 Group extraction unit
- 13 Group storage unit
- 14 Compression value calculation unit
- 15 Compression value storage unit
- 16 Lineage storage unit
- 17 Lineage group storage unit
- 18 Group graph storage unit
- 19 Compression value acquisition unit
- 20 Acquisition result storage unit
- 21 Project creation unit
- 22 Project storage unit
- 23 Display unit

(57) Abstract: In the present invention, a compression value storage unit stores a compression value for each group extracted from a plurality of data flows. A group extraction unit then extracts a group from a data flow of a lineage subject, and a compression value acquisition unit acquires the compression value of the group of the lineage subject from the compression value storage unit. A project creation unit then gathers into a project groups for which the compression value is greater than a prescribed value, said groups being gathered from the data flow of the lineage subject in order beginning

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).

添付公開書類:

- 一 国際調査報告 (条約第21条(3))

with the largest value, and compresses the data flow, and a display unit displays the compressed data flow.

(57) 要約: 複数のデータフローから抽出されたグループ毎に圧縮価値記憶部が圧縮価値を記憶する。そして、グループ抽出部がリネージュ対象のデータフローからグループを抽出し、圧縮価値取得部がリネージュ対象のグループの圧縮価値を圧縮価値記憶部から取得する。そして、プロジェクト作成部がリネージュ対象のデータフローから圧縮価値が所定値より大きいグループを大きい順にプロジェクトにまとめてデータフローを圧縮し、表示部が、圧縮されたデータフローを表示する。

明 細 書

発明の名称： 情報処理装置及びリネージュプログラム

技術分野

[0001] 本発明は、情報処理装置及びリネージュプログラムに関する。

背景技術

[0002] 近年、データの信頼性を判断するため、あるいは、データの使用先を把握するために、データの来歴と影響範囲をグラフ表示するリネージュ機能が使われている。図12は、リネージュ機能により表示されるデータフローの一例を示す図である。図12では、データはc s v (comma-separated values) ファイルであり、「品切れ. c s v」が対象として選択されている。楕円のアイコンはプロセス（処理）を表す。カードのアイコンはデータを表す。「P y t h o n」は、プログラミング言語であり、楕円の中の「P y t h o n」は、プロセスが「P y t h o n」で作成されていることを示す。

[0003] 「品切れ. c s v」より左が「品切れ. c s v」の来歴を示し、「品切れ. c s v」より右が「品切れ. c s v」の影響範囲を示す。すなわち、「売り上げ. c s v」と「商品マスター. c s v」から2つのプロセスにより「中間. c s v」が作成され、「中間. c s v」と「在庫. c s v」から1つのプロセスにより「品切れ. c s v」が作成される。また、「品切れ. c s v」から2つのプロセスにより「需要予測. c s v」が作成される。

[0004] なお、有向グラフに関する従来技術として、有向グラフに含まれるノードについて関心の指定を受け取って、関心のないノードを要約する計算を行う技術がある。また、メタデータオブジェクトを特徴づける要約データを計算し、メタデータオブジェクトをノードで表しノード間の関係を線で表す視覚的表現において、要約データをノードの近傍に重ねる技術がある。

先行技術文献

特許文献

[0005] 特許文献1：特表2017-530440号公報

特許文献2：特表2012-510688号公報

発明の概要

発明が解決しようとする課題

[0006] リネージュ機能で表示されるデータフローには、データとプロセスの数が多くなると全体像の把握が困難になるという問題がある。図13は、全体像の把握が困難なデータフローの一例を示す図である。図13では、対象として選択されたcsvファイルの来歴に含まれるプロセス及びcsvファイルの数が、全体像の把握が困難なほど多い。

[0007] 本発明は、1つの側面では、データフローの全体像の把握を容易にすることを目的とする。

課題を解決するための手段

[0008] 1つの態様では、情報処理装置は、圧縮価値記憶部と圧縮部と表示部とを有する。前記圧縮価値記憶部は、グループに含まれる複数の処理を1つのまとめ処理に圧縮する価値を示す圧縮価値をグループ毎に管理する圧縮価値情報を記憶する。ここで、グループは、データフローにおいて複数の処理と該複数の処理の先頭処理の入力データから最終処理の出力データまでのデータとを含む部分データフローである。前記圧縮部は、前記圧縮価値記憶部に記憶された圧縮価値情報に基づいて、リネージュ対象のデータフローにおいて圧縮価値が所定値より大きいグループの複数の処理を1つのまとめ処理にまとめる。前記表示部は、前記圧縮部により複数の処理が1つのまとめ処理にまとめられた圧縮データフローを表示する。

発明の効果

[0009] 本発明は、1つの側面では、データフローの全体像の把握を容易にすることができる。

図面の簡単な説明

[0010] [図1A]図1Aは、圧縮価値の計算に用いられる複数のデータフローを示す図である。

[図1B]図1 Bは、データフローから抽出されたグループの例を示す図である。

[図1C]図1 Cは、異なるグループとして判定される例を示す図である。

[図1D]図1 Dは、異なるグループとして判定される他の例を示す図である。

[図1E]図1 Eは、リネージュ対象のデータフローのグループの例を示す第1の図である。

[図1F]図1 Fは、リネージュ対象のデータフローのグループの例を示す第2の図である。

[図1G]図1 Gは、リネージュ対象のデータフローのグループの例を示す第3の図である。

[図1H]図1 Hは、リネージュ対象のデータフローのグループの例を示す第4の図である。

[図1I]図1 Iは、複数のプロセスを1つのプロジェクトにまとめる例を示す第1の図である。

[図1J]図1 Jは、複数のプロセスを1つのプロジェクトにまとめる例を示す第2の図である。

[図1K]図1 Kは、複数のプロセスを1つのプロジェクトにまとめる例を示す第3の図である。

[図1L]図1 Lは、図1 Eに示したデータフローの圧縮データフローを示す図である。

[図2]図2は、実施例に係る情報処理装置の機能構成を示す図である。

[図3]図3は、データフロー記憶部の一例を示す図である。

[図4]図4は、グループ記憶部の一例を示す図である。

[図5]図5は、圧縮価値記憶部の一例を示す図である。

[図6]図6は、リネージュ記憶部の一例を示す図である。

[図7]図7は、グループグラフ記憶部の一例を示す図である。

[図8]図8は、取得結果記憶部の一例を示す図である。

[図9]図9は、プロジェクト記憶部の一例を示す図である。

[図10]図10は、情報処理装置による処理のフローを示すフローチャートである。

[図11]図11は、実施例に係るリネージュプログラムを実行するコンピュータのハードウェア構成を示す図である。

[図12]図12は、リネージュ機能により表示されるデータフローの一例を示す図である。

[図13]図13は、全体像の把握が困難なデータフローの一例を示す図である。

発明を実施するための形態

[0011] 以下に、本願の開示する情報処理装置及びリネージュプログラムの実施例を図面に基づいて詳細に説明する。なお、この実施例は開示の技術を限定するものではない。

実施例

[0012] まず、実施例に係る情報処理装置がリネージュ機能により表示するデータフローを図1A～図1Lを用いて説明する。実施例に係る情報処理装置は、図1A～図1Dに示すように、複数のデータフローを用いて、複数のプロセスを含む部分データフローの圧縮価値を計算して、圧縮価値記憶部に記憶する。ここで、圧縮価値とは、部分データフローの複数のプロセスをまとめて1つに圧縮する価値を示す値である。なお、圧縮価値の詳細については後述する。そして、実施例に係る情報処理装置は、図1E～図1Lに示すように、リネージュ対象のデータフローを圧縮価値に基づいて圧縮して、圧縮データフローを表示する。

[0013] 図1Aは、圧縮価値の計算に用いられる複数のデータフローを示す図である。ここでは、データフローA～データフローDで表される4つのデータフローが圧縮価値の計算に用いられる。実施例に係る情報処理装置は、データフローAにおいて、「Data2.csv」と「Data1.csv」の統計的な差異として「行数の減少」を特定する。また、実施例に係る情報処理装置は、データフローAにおいて、「Data3.csv」と「Data2

、 csv」の統計的な差異として「値の数の増加」を特定する。統計的な差異としては、他に「行数の増加」、「値の数の減少」、「値の範囲の減少」、「値の範囲の増加」、「値の種類の減少」、「値の種類の増加」等がある。実施例に係る情報処理装置は、これらの統計的な差異を、入力データと出力データを比較することで特定する。

[0014] そして、実施例に係る情報処理装置は、「Data2.csv」と「Data1.csv」の統計的な差異「行数の減少」を生み出すプロセス「Python1」のアルゴリズムとして「削除」を特定する。特定されたアルゴリズムは、プロセスの下に表示される。統計的な差異「行数の減少」を生み出すプロセスのアルゴリズムとしては、「削除」の他に「外れ値除外」がある。「削除」であるか「外れ値除外」であるかは、入力データと出力データを比較することで特定される。また、実施例に係る情報処理装置は、「Data3.csv」と「Data2.csv」の統計的な差異「値の数の増加」を生み出すプロセス「Python2」のアルゴリズムとして「補間」を特定する。

[0015] 同様に、実施例に係る情報処理装置は、「Data4.csv」と「Data3.csv」の統計的な差異を生み出すプロセス「Python3」のアルゴリズムとして「正規化」を特定する。また、実施例に係る情報処理装置は、「Data5.csv」と「Data4.csv」の統計的な差異を生み出すプロセス「Python4」のアルゴリズムとして「重回帰分析」を特定する。

[0016] ただし、「重回帰分析」については、入力データと出力データを比較することで特定されるのではなく、データフロー作成時に部品として用意されたプロセスに付加されたアルゴリズムとして特定される。このように、実施例に係る情報処理装置は、入力データと出力データの差異を生み出すプロセスのアルゴリズムを、入力データと出力データを比較することで特定する以外に、部品として用意されたプロセスに付加されたアルゴリズムを用いて特定する。

[0017] また、実施例に係る情報処理装置は、データフローBにおいて、アルゴリズムとして「名寄せ」を特定する。また、実施例に係る情報処理装置は、データフローCにおいて、「独自アルゴリズム」を特定する。ここで、「独自アルゴリズム」は、入力データと出力データを比較することで特定されず、部品として用意されたプロセスでない場合に、プロセスに付加されるアルゴリズムである。

[0018] 実施例に係る情報処理装置は、2つ以上のプロセスと2つ以上のプロセスの先頭のプロセスの入力データから最後のプロセスの出力データまでのデータとを含む部分データフローをグループとして全てのデータフローから全て抽出する。そして、実施例に係る情報処理装置は、異なるデータフローに含まれる2つのグループについて、統計的な差異とアルゴリズムを特定し、対応する統計的な差異と、対応するアルゴリズムが一致するか否かを判定する。そして、対応する統計的な差異と、対応するアルゴリズムが一致する場合に、実施例に係る情報処理装置は、2つのグループは同一であると判定し、グループの類似度に1を加える。また、実施例に係る情報処理装置は、2つのグループが同一であるか否かの判定をグループの全ての組み合わせについて行う。

[0019] 例えば、実施例に係る情報処理装置は、図1Bに示すように、データフローAから、「Data1.csv→Python1→Data2.csv→Python2→Data3.csv」をグループA1として抽出する。ここで、「グループA1」は、グループを識別するグループ番号が「A1」であるグループである。また、実施例に係る情報処理装置は、データフローBから、「Data1.csv→Python1→Data2.csv→Python2→Data3.csv」をグループB1として抽出する。

[0020] そして、実施例に係る情報処理装置は、グループA1において、「Data2.csv」と「Data1.csv」の統計的な差異として「行数の減少」を特定する。また、実施例に係る情報処理装置は、グループA1において、「Data3.csv」と「Data2.csv」の統計的な差異とし

て「値の数の増加」を特定する。また、実施例に係る情報処理装置は、統計的な差異「行数の減少」を生み出すアルゴリズムとして「削除」を特定し、統計的な差異「値の数の増加」を生み出すアルゴリズムとして「補間」を特定する。

[0021] 同様に、実施例に係る情報処理装置は、グループB1において、「Data2.csv」と「Data1.csv」の統計的な差異として「行数の減少」を特定する。また、実施例に係る情報処理装置は、グループB1において、「Data3.csv」と「Data2.csv」の統計的な差異として「値の数の増加」を特定する。また、実施例に係る情報処理装置は、統計的な差異「行数の減少」を生み出すアルゴリズムとして「削除」を特定し、統計的な差異「値の数の増加」を生み出すアルゴリズムとして「補間」を特定する。

[0022] グループA1とグループB1では、対応する統計的な差異が「行数の減少」と「値の数の増加」で同じであり、対応するアルゴリズムも「削除」と「補間」で同じである。したがって、実施例に係る情報処理装置は、アルゴリズムが「削除→補間」であり、統計的な差異が「行数の減少→値の数の増加」で表されるグループの類似度に1を加える。

[0023] 図1Cは、異なるグループとして判定される例を示す図である。図1Cにおいて、グループA1とグループB2はアルゴリズムの数が異なるので、異なるグループと判定される。図1Dは、異なるグループとして判定される他の例を示す図である。図1Dにおいて、データフローCから抽出されたグループC6とデータフローDから抽出されたグループD3は、少なくともアルゴリズムが異なるので、異なるグループと判定される。

[0024] 図1Bに示したグループA1を「削除」→「補間」で表すと、データフローAからは、「削除」を先頭のアルゴリズムとするグループとして、「削除」→「補間」→「正規化」、「削除」→「補間」→「正規化」→「重回帰分析」が抽出される。また、「補間」を先頭のアルゴリズムとするグループとして、「補間」→「正規化」、「補間」→「正規化」→「重回帰分析」が抽

出される。また、「正規化」を先頭のアルゴリズムとするグループとして、「正規化」→「重回帰分析」が抽出される。したがって、データフロー Aからは、6 個のグループが抽出される。同様に、データフロー Bからは、「削除」→「補間」のグループ B 1 等の 15 個のグループが抽出される。

[0025] 実施例に係る情報処理装置は、グループの組み合わせの全てについて、297 回判定を行い、グループ「削除→補間」の類似度を 5 とする。また、実施例に係る情報処理装置は、グループ「正常化→重回帰分析」の類似度を 1、グループ「削除→補間→名寄せ」の類似度を 1、グループ「補間→名寄せ」の類似度を 1、その他のグループの類似度を 0 とする。

[0026] そして、実施例に係る情報処理装置は、圧縮価値を類似度とアルゴリズムの数に基づいてグループ毎に計算する。例えば、実施例に係る情報処理装置は、 $\text{圧縮価値} = \text{類似度} \times \text{係数}^5$ 、により計算する。ここで、係数はアルゴリズム数に基づく値であり、例えば、アルゴリズム数である。「 $^$ 」はべき乗を示す。

[0027] 類似度が大きいことはグループの出現頻度が高いことを示し、グループの圧縮価値は大きい。また、アルゴリズム数が多いグループは圧縮率が大きいので、グループの圧縮価値は大きい。

[0028] 実施例に係る情報処理装置は、例えば、グループ「削除→補間」の圧縮価値を $5 \times 2^5 = 160$ 、グループ「正常化→重回帰分析」の圧縮価値を $1 \times 2^5 = 32$ 、グループ「削除→補間→名寄せ」の圧縮価値を $1 \times 3^5 = 243$ とする。同様に、実施例に係る情報処理装置は、例えば、グループ「補間→名寄せ」の類似度を $1 \times 2^5 = 32$ とし、その他のグループの圧縮価値を 0 とする。

[0029] そして、実施例に係る情報処理装置は、計算した圧縮価値をグループに対応付けて圧縮価値記憶部に記憶する。なお、ここでは、4 つのデータフローから圧縮価値を計算したが、実施例に係る情報処理装置は、より多くのデータフローからグループの圧縮価値を計算し、様々な部分データフローの圧縮価値を記憶する。

[0030] そして、実施例に係る情報処理装置は、リネージュ対象のデータフローから全てのグループを抽出し、各グループの統計的な差異とアルゴリズムを特定する。そして、実施例に係る情報処理装置は、各グループの圧縮価値を圧縮価値記憶部を参照して取得する。

[0031] 図1E～図1Hは、リネージュ対象のデータフローのグループの例を示す図である。図1Eの例では、実施例に係る情報処理装置は、「Data2.csv」と「Data1.csv」の統計的な差異として「行数の減少」を特定し、「Data3.csv」と「Data2.csv」の統計的な差異として「値の数の増加」を特定する。そして、実施例に係る情報処理装置は、統計的な差異「行数の減少」を生み出すアルゴリズムとして「削除」を特定し、統計的な差異「値の数の増加」を生み出すアルゴリズムとして「補間」を特定する。そして、実施例に係る情報処理装置は、グループ「削除→補間」の圧縮価値として160を取得する。

[0032] また、図1Fの例では、実施例に係る情報処理装置は、「Data4.csv」と「Data3.csv」の統計的な差異として「値の種類の減少」を特定する。そして、実施例に係る情報処理装置は、統計的な差異「値の種類の減少」を生み出すアルゴリズムとして「名寄せ」を特定する。そして、実施例に係る情報処理装置は、グループ「削除→補間→名寄せ」の圧縮価値として243を取得する。

[0033] また、図1Gの例では、実施例に係る情報処理装置は、「Data7.csv」と「Data4.csv」及び「Data6.csv」との統計的な差異として「値の範囲の変更」を特定する。また、実施例に係る情報処理装置は、「Data6.csv」と「Data5.csv」の統計的な差異として「独自」を特定する。そして、実施例に係る情報処理装置は、統計的な差異「値の範囲の変更」を生み出すアルゴリズムとして「正規化」を特定し、統計的な差異「独自」を生み出すアルゴリズムとして「独自アルゴリズム」を特定する。そして、実施例に係る情報処理装置は、グループに「独自アルゴリズム」が含まれているので、圧縮価値を0とする。

[0034] また、図1 Hの例では、実施例に係る情報処理装置は、「Data 8. csv」と「Data 7. csv」の統計的な差異として「流用部品（重回帰分析）」を特定する。そして、実施例に係る情報処理装置は、統計的な差異「流用部品（重回帰分析）」を生み出すアルゴリズムとして「重回帰分析」を特定する。そして、実施例に係る情報処理装置は、グループ「正規化→重回帰分析」について圧縮価値として32を取得する。

[0035] そして、実施例に係る情報処理装置は、圧縮価値が大きいグループから順番に複数のプロセスを1つのまとめプロセスとしてのプロジェクトにまとめる。そして、実施例に係る情報処理装置は、グループに含まれる先頭のプロセスの入力データをプロジェクトの入力データとし、最後のプロセスの出力データをプロジェクトの出力データとする。図1 I～図1 Kは、複数のプロセスを1つのプロジェクトにまとめる例を示す図である。

[0036] 図1 Iでは、圧縮価値が243で最大のグループ「削除→補間→名寄せ」が「Project 1」にまとめられる。また、先頭の「削除」の入力データ「Data 1. csv」が「Project 1」の入力データとされ、最後の「名寄せ」の出力データ「Data 4. csv」が「Project 1」の出力データとされる。なお、図1 Iにおいて、「グループNo. 2」は、圧縮価値記憶部においてグループ「削除→補間→名寄せ」を識別する番号が「2」であることを示す。

[0037] 図1 Jでは、次に圧縮価値（160）が大きいグループ「削除→補間」が「Project 2」にまとめられる。また、先頭の「削除」の入力データ「Data 1. csv」が「Project 2」の入力データとされ、最後の「補間」の出力データ「Data 3. csv」が「Project 2」の出力データとされる。

[0038] 図1 Kでは、次に圧縮価値（32）が大きいグループ「正規化→重回帰分析」が「Project 3」にまとめられる。また、先頭の「正常化」の入力データ「Data 4. csv」と「Data 6. csv」が「Project 3」の入力データとされ、最後の「重回帰分析」の出力データ「Data

a 8. csv」が「Project 3」の出力データとされる。

[0039] そして、実施例に係る情報処理装置は、リネージュ対象のデータフローをプロジェクトに置き換えられた圧縮データフローとして表示する。図1Lは、図1Eに示したデータフローの圧縮データフローを示す図である。図1Lに示すように、まとめられた処理のアルゴリズムは、プロジェクトの下にまとめて表示される。図1Eのデータフローと比較して、図1Lの圧縮データフロー1では、プロセス数及びデータ数が少なくなり、全体の把握が容易になっている。

[0040] また、図1Lにおいて、ユーザによりプロジェクトがマウスで選択されると、実施例に係る情報処理装置は、プロジェクトにまとめられる前のプロセス又はプロジェクトを吹き出しで表示する。また、吹き出しの中のプロジェクトがマウスで選択されると、実施例に係る情報処理装置は、吹き出し中のプロジェクトにまとめられる前のプロセス又はプロジェクトを別の吹き出しで表示する。

[0041] このように、プロジェクトにまとめられる前のプロセス又はプロジェクトを階層的に表示することで、実施例に係る情報処理装置は、データフローの全体から詳細への把握を支援することができる。

[0042] 次に、実施例に係る情報処理装置の機能構成について説明する。図2は、実施例に係る情報処理装置の機能構成を示す図である。図2に示すように、実施例に係る情報処理装置10は、データフロー記憶部11と、グループ抽出部12と、グループ記憶部13と、圧縮価値計算部14と、圧縮価値記憶部15とを有する。また、実施例に係る情報処理装置10は、リネージュ記憶部16と、リネージュグループ記憶部17と、グループグラフ記憶部18と、圧縮価値取得部19と、取得結果記憶部20と、プロジェクト作成部21と、プロジェクト記憶部22と、表示部23とを有する。

[0043] データフロー記憶部11は、複数のデータフローのグラフ構造の情報を記憶する。情報処理装置10は、例えば、ユーザがマウスを用いて行った指示を受け付けてファイルからデータフローのグラフ構造の情報を読み出してデ

ータフロー記憶部 11 に格納したり追加したりする。

[0044] 図 3 は、データフロー記憶部 11 の一例を示す図である。図 3 に示すように、データフロー記憶部 11 は、データフローを識別するデータフロー名とデータフローのグラフ構造の情報を対応付けて記憶する。データフロー記憶部 11 は、例えば、データフロー A について、「Data 1. csv → Python 1 → Data 2. csv」、「Data 2. csv → Python 2 → Data 3. csv」を記憶する。また、データフロー記憶部 11 は、データフロー A について、「Data 3. csv → Python 3 → Data 4. csv」、「Data 4. csv → Python 4 → Data 5. csv」を記憶する。

[0045] グループ抽出部 12 は、データフロー記憶部 11 が記憶する情報を用いて全てのグループを抽出し、各グループについて、統計的な差異とアルゴリズムを特定して、グループ記憶部 13 に格納する。

[0046] グループ記憶部 13 は、グループの情報を記憶する。図 4 は、グループ記憶部 13 の一例を示す図である。図 4 に示すように、グループ記憶部 13 は、グループを識別するグループ No. に対応付けて、アルゴリズムを識別するアルゴリズム名と統計的な差異とを記憶する。例えば、グループ記憶部 13 は、グループ A 1 について、アルゴリズム名として「削除 → 補間」を記憶し、統計的な差異として「行数の減少 → 値の数の増加」を記憶する。

[0047] 圧縮価値計算部 14 は、グループの類似度を計算し、類似度をアルゴリズムの数に基づいて、 $\text{圧縮価値} = \text{類似度} \times \text{係数}^5$ を計算し、圧縮価値記憶部 15 に格納する。

[0048] 圧縮価値記憶部 15 は、圧縮価値をグループ毎に管理する圧縮価値情報を記憶する。図 5 は、圧縮価値記憶部 15 の一例を示す図である。図 5 に示すように、圧縮価値記憶部 15 は、アルゴリズム名と、統計的な差異と、アルゴリズム数と、類似度と、圧縮価値をグループ毎に記憶する。例えばアルゴリズム名が「削除 → 補間」であり、統計的な差異が「行数の減少 → 値の数の増加」であるグループのアルゴリズム数は「2」であり、類似度は「5」で

あり、圧縮価値は「160」である。

[0049] リネージュ記憶部16は、リネージュ対象のデータフローのグラフ構造の情報を記憶する。情報処理装置10は、例えば、ユーザがマウスを用いて行った指示を受け付けてファイルからリネージュ対象のデータフローのグラフ構造の情報を読み出してリネージュ記憶部16に格納する。

[0050] 図6は、リネージュ記憶部16の一例を示す図である。図6に示すように、リネージュ記憶部16は、リネージュ対象のデータフローのグラフ構造の要素を識別する番号であるNo.と要素のグラフ構造とを対応付けて記憶する。ここで、要素は1つのプロセスとその入力データ及び出力データのグラフ構造である。例えば、識別する番号が「1」である要素のグラフ構造は「Data1.csv→Python1→Data2.csv」である。

[0051] グループ抽出部12は、リネージュ記憶部16が記憶する情報を用いてリネージュ対象のデータフローから全てのグループを抽出し、各グループについて、統計的な差異とアルゴリズムを特定して、リネージュグループ記憶部17に格納する。リネージュグループ記憶部17は、リネージュ対象のデータフローから抽出されたグループの情報を記憶する。また、グループ抽出部12は、抽出したグループのグラフ構造を作成し、グループグラフ記憶部18に格納する。

[0052] グループグラフ記憶部18は、リネージュ対象のデータフローのグループのグラフ構造の情報を記憶する。図7は、グループグラフ記憶部18の一例を示す図である。図7に示すように、グループグラフ記憶部18は、グループNo.とグループのグラフ構造を対応付けて記憶する。例えば、識別する番号が「1」であるグループのグラフ構造は「Data1.csv→Python1→Data2.csv」と「Data2.csv→Python2→Data3.csv」である。

[0053] 圧縮価値取得部19は、リネージュグループ記憶部17が情報を記憶するグループについて、圧縮価値記憶部15から圧縮価値を取得し、グループの情報とともに取得結果記憶部20に格納する。圧縮価値取得部19は、圧縮

価値記憶部 15 に圧縮価値が記憶されていないグループについては、圧縮価値を 0 として取得結果記憶部 20 に格納する。

[0054] 取得結果記憶部 20 は、リネージュグループ記憶部 17 が記憶する情報に圧縮価値が付加された情報をグループ毎に記憶する。図 8 は、取得結果記憶部 20 の一例を示す図である。図 8 に示すように、取得結果記憶部 20 は、グループを識別するグループ No. とアルゴリズム名と統計的な差異と圧縮価値をグループ毎に記憶する。例えば、取得結果記憶部 20 は、「1」で識別されるグループについて、アルゴリズム名として「削除→補間」を記憶し、統計的な差異として「行数の減少→値の数の増加」を記憶し、圧縮価値として「160」を記憶する。

[0055] プロジェクト作成部 21 は、取得結果記憶部 20 から圧縮価値が 0 でないグループを圧縮価値が大きい順に特定し、特定したグループの複数のプロセスを 1 つのプロジェクトにまとめて、プロジェクトの情報をプロジェクト記憶部 22 に格納する。ここで、プロジェクトは、複数のプロセスがまとめられたプロセスである。プロジェクトの入力データは、まとめられた複数のプロセスの先頭のプロセスの入力データであり、プロジェクトの出力データは、まとめられた複数のプロセスの最後のプロセスの出力データである。なお、プロジェクト作成部 21 は、圧縮価値が所定の値以上のグループをプロジェクトにまとめてもよい。

[0056] プロジェクト記憶部 22 は、プロジェクトの情報を記憶する。プロジェクトの情報には、プロジェクトにまとめられる前のグループの情報が含まれる。図 9 は、プロジェクト記憶部 22 の一例を示す図である。図 9 に示すように、プロジェクト記憶部 22 は、グループ No. とアルゴリズム名と統計的な差異と圧縮価値とプロジェクト名と階層をプロジェクト毎に記憶する。

[0057] グループ No. とアルゴリズム名と統計的な差異と圧縮価値は、まとめられる前のグループの情報である。プロジェクト名は、プロジェクトを識別する名前である。階層は、プロジェクトの階層である。他のプロジェクトに含まれないプロジェクトの階層は「1」である。階層が「1」のプロジェクト

に含まれ、他のプロジェクトに含まれないプロジェクトの階層は「2」である。一般に、階層が n のプロジェクトは、階層が「1」、階層が「2」、 $\dots$ 、階層が「 $n-1$ 」のプロジェクトに含まれる。

[0058] 例えば、プロジェクト記憶部22は、「Project 1」で識別されるプロジェクトについて、「2」で識別されるグループの情報と階層が「1」であることを記憶する。

[0059] 表示部23は、プロジェクト記憶部22とグループグラフ記憶部18を参照し、リネージュ対象のデータフローについて、最上位階層のプロジェクトを含むデータフローを圧縮データフロー1として表示する。ここで、最上位階層のプロジェクトとは、階層が「1」のプロジェクトである。

[0060] また、表示部23は、プロセスアイコンにはアルゴリズム名を添えて表示する。また、表示部23は、まとめられたプロセスのアルゴリズム名をプロジェクトアイコンに添えて表示する。

[0061] そして、表示部23は、プロジェクトアイコンに合わせられたマウスカーソルを検出すると、プロジェクトにまとめられたプロセス又は次の下の階層のプロジェクトを含む部分データフローを表示する。また、表示部23は、部分データフローに含まれるプロジェクトアイコンに合わせられたマウスカーソルを検出すると、プロジェクトにまとめられたプロセス又は次の下の階層のプロジェクトを含む部分データフローを表示する。一般に、表示部23は、階層が「 n 」のプロジェクトのアイコンにマウスカーソルが合わせられると、階層が「 n 」のプロジェクトにまとめられた階層が「 $n+1$ 」のプロジェクトを含む部分データフローを表示する。

[0062] 次に、情報処理装置10による処理のフローについて説明する。図10は、情報処理装置10による処理のフローを示すフローチャートである。図10において、ステップS1～ステップS6は、グループの圧縮価値を算出する処理であり、ステップS7～ステップS14は、圧縮価値に基づいてリネージュ対象のデータフローを表示する処理である。

[0063] 図10に示すように、情報処理装置10は、2つのデータフローの連続す

る部分をグルーピングする（ステップS 1）。ここで、グループには、2つ以上のプロセスと2つ以上のプロセスの先頭のプロセスの入力データから最後のプロセスの出力データまでのデータとが含まれる。

[0064] そして、情報処理装置10は、2つのグループに含まれる全プロセスについて、入力データと出力データの統計的な差異を特定する（ステップS 2）。そして、情報処理装置10は、2つのグループに含まれる全プロセスについて、差異を生み出しうるアルゴリズムの中から、該当するアルゴリズムを探してアルゴリズム名を特定する（ステップS 3）。

[0065] そして、情報処理装置10は、2つのグループの対応する統計的な差異及びアルゴリズムが一致していれば、グループの類似度を+1する（ステップS 4）。なお、対応する統計的な差異及びアルゴリズムが一致すれば、アルゴリズムの数も一致する。そして、情報処理装置10は、全てのデータフローと全てのグルーピングの組み合わせで類似度を求めたか否かを判定し（ステップS 5）、類似度を求めている組み合わせがある場合には、ステップS 1に戻る。

[0066] 一方、全てのデータフローと全てのグルーピングの組み合わせで類似度を求めた場合には、情報処理装置10は、類似度にアルゴリズム数に関する係数を掛けて圧縮価値を求める（ステップS 6）。圧縮価値は圧縮価値記憶部15に記憶され、リネージュ対象のデータフローを表示する場合に利用される。

[0067] リネージュ対象のデータフローを表示する処理では、情報処理装置10は、リネージュ対象のデータフローの連続する部分をグルーピングする（ステップS 7）。そして、グループに含まれるプロセスについて、情報処理装置10は、入力データと出力データの統計的な差異を求め、差異を生み出すアルゴリズムの名前を特定する（ステップS 8）。

[0068] そして、情報処理装置10は、統計的な差異とアルゴリズムに基づいて、圧縮価値記憶部15から、グループの圧縮価値を取得し（ステップS 9）、全てのグループで圧縮価値を取得したか否かを判定する（ステップS 10）

。そして、圧縮価値を取得していないグループがある場合には、情報処理装置 10 は、ステップ S 7 に戻る。

[0069] 一方、全てのグループで圧縮価値を取得した場合には、情報処理装置 10 は、圧縮価値が大きい順にグループをプロジェクトとして階層的にまとめ上げる（ステップ S 11）。そして、情報処理装置 10 は、リネージュ対象のデータフローについて、最上位階層のプロジェクトを含む圧縮データフロー 1 を表示する（ステップ S 12）。そして、情報処理装置 10 は、プロジェクトアイコンに合わせられたマウ斯卡ーソルを検出すると、プロジェクトの中身の詳細な部分データフローを表示する（ステップ S 13）。

[0070] そして、情報処理装置 10 は、終了か否かを判定し（ステップ S 14）、終了でない場合には、ステップ S 13 に戻り、終了と判定した場合には、表示処理を終了する。

[0071] このように、情報処理装置 10 は、圧縮価値に基づいてプロジェクトを作成し、プロジェクトにまとめられた部分データフローを階層的に表示するので、データフローの段階的な把握を支援することができる。

[0072] 上述してきたように、実施例では、圧縮価値をグループ毎に管理する圧縮価値情報を圧縮価値記憶部 15 が記憶する。そして、グループ抽出部 12 がリネージュ対象のデータフローからグループを抽出し、圧縮価値取得部 19 がリネージュ対象のグループの圧縮価値を圧縮価値記憶部 15 から取得する。そして、プロジェクト作成部 21 がリネージュ対象のデータフローから圧縮価値が 0 より大きいグループを大きい順にプロジェクトにまとめてデータフローを圧縮し、表示部 23 が、圧縮されたデータフローを表示する。したがって、情報処理装置 10 は、データフローの全体像の把握を容易にすることができる。

[0073] また、実施例では、表示部 23 は、プロジェクトアイコンにマウ斯卡ーソルが合わされたことを検出すると、プロジェクトにまとめられる前のグループの部分データフローを階層的に表示する。したがって、情報処理装置 10 は、データフローの一部の詳細の把握を容易にすることができる。

[0074] また、実施例では、グループ抽出部 1 2 が複数のデータフローからグループを抽出し、圧縮価値計算部 1 4 が他のグループとの類似度とグループに含まれるプロセスの数に基づいて、グループ毎に圧縮価値を計算して圧縮価値記憶部 1 5 に格納する。したがって、情報処理装置 1 0 は、出現頻度が高くプロセス数の多い部分データフローを圧縮することができ、適切な圧縮を行うことができる。

[0075] また、実施例では、圧縮価値計算部 1 4 は、プロセスの入力データと出力データの統計的な差異、及び、統計的な差異を生み出したアルゴリズムに基づいて、類似度を計算するので、グループ間の類似度を適切に計算することができる。

[0076] なお、実施例では、情報処理装置 1 0 について説明したが、情報処理装置 1 0 が有する構成をソフトウェアによって実現することで、同様の機能を有するリネージュプログラムを得ることができる。そこで、リネージュプログラムを実行するコンピュータについて説明する。

[0077] 図 1 1 は、実施例に係るリネージュプログラムを実行するコンピュータのハードウェア構成を示す図である。図 1 1 に示すように、コンピュータ 5 0 は、メインメモリ 5 1 と、CPU (Central Processing Unit) 5 2 と、LAN (Local Area Network) インタフェース 5 3 と、HDD (Hard Disk Drive) 5 4 とを有する。また、コンピュータ 5 0 は、スーパー I/O (Input Output) 5 5 と、DVI (Digital Visual Interface) 5 6 と、ODD (Optical Disk Drive) 5 7 とを有する。

[0078] メインメモリ 5 1 は、プログラムやプログラムの実行途中結果等を記憶するメモリである。CPU 5 2 は、メインメモリ 5 1 からプログラムを読み出して実行する中央処理装置である。CPU 5 2 は、メモリコントローラを有するチップセットを含む。

[0079] LAN インタフェース 5 3 は、コンピュータ 5 0 を LAN 経由で他のコンピュータに接続するためのインタフェースである。HDD 5 4 は、プログラムやデータを格納するディスク装置であり、スーパー I/O 5 5 は、マウスや

キーボード等の入力装置を接続するためのインタフェースである。D V I 5 6は、液晶表示装置を接続するインタフェースであり、O D D 5 7は、D V D、C D - Rの読み書きを行う装置である。

[0080] L A Nインタフェース5 3は、P C Iエクスプレス (P C I e) によりC P U 5 2に接続され、H D D 5 4及びO D D 5 7は、S A T A (Serial Advanced Technology Attachment) によりC P U 5 2に接続される。スーパーI O 5 5は、L P C (Low Pin Count) によりC P U 5 2に接続される。

[0081] そして、コンピュータ5 0において実行されるリネージュプログラムは、コンピュータ5 0により読み出し可能な記録媒体の一例であるC D - Rに記憶され、O D D 5 7によってC D - Rから読み出されてコンピュータ5 0にインストールされる。あるいは、リネージュプログラムは、L A Nインタフェース5 3を介して接続された他のコンピュータシステムのデータベース等に記憶され、これらのデータベースから読み出されてコンピュータ5 0にインストールされる。そして、インストールされたりネージュプログラムは、H D D 5 4に記憶され、メインメモリ5 1に読み出されてC P U 5 2によって実行される。

[0082] また、実施例では、プロセスが処理を行う場合について説明したが、データフローに含まれる処理はタスク、モジュール等の他の処理でもよい。

符号の説明

- [0083]
- | | |
|-----|--------------|
| 1 | 圧縮データフロー |
| 1 0 | 情報処理装置 |
| 1 1 | データフロー記憶部 |
| 1 2 | グループ抽出部 |
| 1 3 | グループ記憶部 |
| 1 4 | 圧縮価値計算部 |
| 1 5 | 圧縮価値記憶部 |
| 1 6 | リネージュ記憶部 |
| 1 7 | リネージュグループ記憶部 |

- 1 8 グループグラフ記憶部
- 1 9 圧縮価値取得部
- 2 0 取得結果記憶部
- 2 1 プロジェクト作成部
- 2 2 プロジェクト記憶部
- 2 3 表示部
- 5 0 コンピュータ
- 5 1 メインメモリ
- 5 2 C P U
- 5 3 L A N インタフェース
- 5 4 H D D
- 5 5 スーパー I O
- 5 6 D V I
- 5 7 O D D

請求の範囲

- [請求項1] データフローにおいて複数の処理と該複数の処理の先頭処理の入力データから最終処理の出力データまでのデータとを含む部分データフローをグループとして、グループに含まれる複数の処理を1つのまとめ処理に圧縮する価値を示す圧縮価値をグループ毎に管理する圧縮価値情報を記憶する圧縮価値記憶部と、
- 前記圧縮価値記憶部に記憶された圧縮価値情報に基づいて、リネージュ対象のデータフローにおいて圧縮価値が所定値より大きいグループの複数の処理を1つのまとめ処理にまとめる圧縮部と、
- 前記圧縮部により複数の処理が1つのまとめ処理にまとめられた圧縮データフローを表示する表示部と
- を有することを特徴とする情報処理装置。
- [請求項2] 前記表示部は、表示した圧縮データフローに含まれるまとめ処理の選択を受け付け、受け付けたまとめ処理について、前記圧縮部によりまとめられる前の処理又はまとめ処理を複数含む部分データフローを階層的に表示することを特徴とする請求項1に記載の情報処理装置。
- [請求項3] 複数のデータフローからグループを抽出してグループ毎に他のグループとの類似度を計算し、該類似度とグループに含まれる処理の数に基づいて前記圧縮価値を計算して前記圧縮価値記憶部に格納する圧縮価値計算部をさらに有することを特徴とする請求項1又は2に記載の情報処理装置。
- [請求項4] 前記圧縮価値計算部は、処理の入力データと出力データの統計的な差異、及び、該統計的な差異を生み出したアルゴリズムに基づいて、前記類似度を計算することを特徴とする請求項3に記載の情報処理装置。
- [請求項5] 前記圧縮価値計算部は、前記統計的な差異が特定できない処理を含むグループの圧縮価値を0とし、
- 前記圧縮部は、リネージュ対象のデータフローから圧縮価値が0よ

り大きいグループの複数の処理を1つのまとめ処理にまとめることを特徴とする請求項4に記載の情報処理装置。

[請求項6]

前記統計的な差異には、行数の減少、行数の増加、値の数の増加、値の数の減少、値の範囲の減少、値の範囲の増加、値の種類の減少、値の種類の増加が含まれ、前記アルゴリズムには、削除、補間、正規化、名寄せが含まれることを特徴とする請求項4に記載の情報処理装置。

[請求項7]

コンピュータに、

データフローにおいて複数の処理と該複数の処理の先頭処理の入力データから最終処理の出力データまでのデータとを含む部分データフローをグループとして、グループに含まれる複数の処理を1つのまとめ処理に圧縮する価値を示す圧縮価値をグループ毎に管理する圧縮価値情報を圧縮価値記憶部に記憶し、

前記圧縮価値記憶部に記憶された圧縮価値情報に基づいて、リネージュ対象のデータフローにおいて圧縮価値が所定値より大きいグループの複数の処理を1つのまとめ処理にまとめ、

リネージュ対象のデータフローにおいて圧縮価値が所定値より大きいグループの複数の処理が1つのまとめ処理にまとめられた圧縮データフローを表示する

処理を実行させることを特徴とするリネージュプログラム。

[請求項8]

前記表示する処理は、表示した圧縮データフローに含まれるまとめ処理の選択を受け付け、受け付けたまとめ処理について、まとめられる前の処理又はまとめ処理を複数含む部分データフローを階層的に表示することを特徴とする請求項7に記載のリネージュプログラム。

[請求項9]

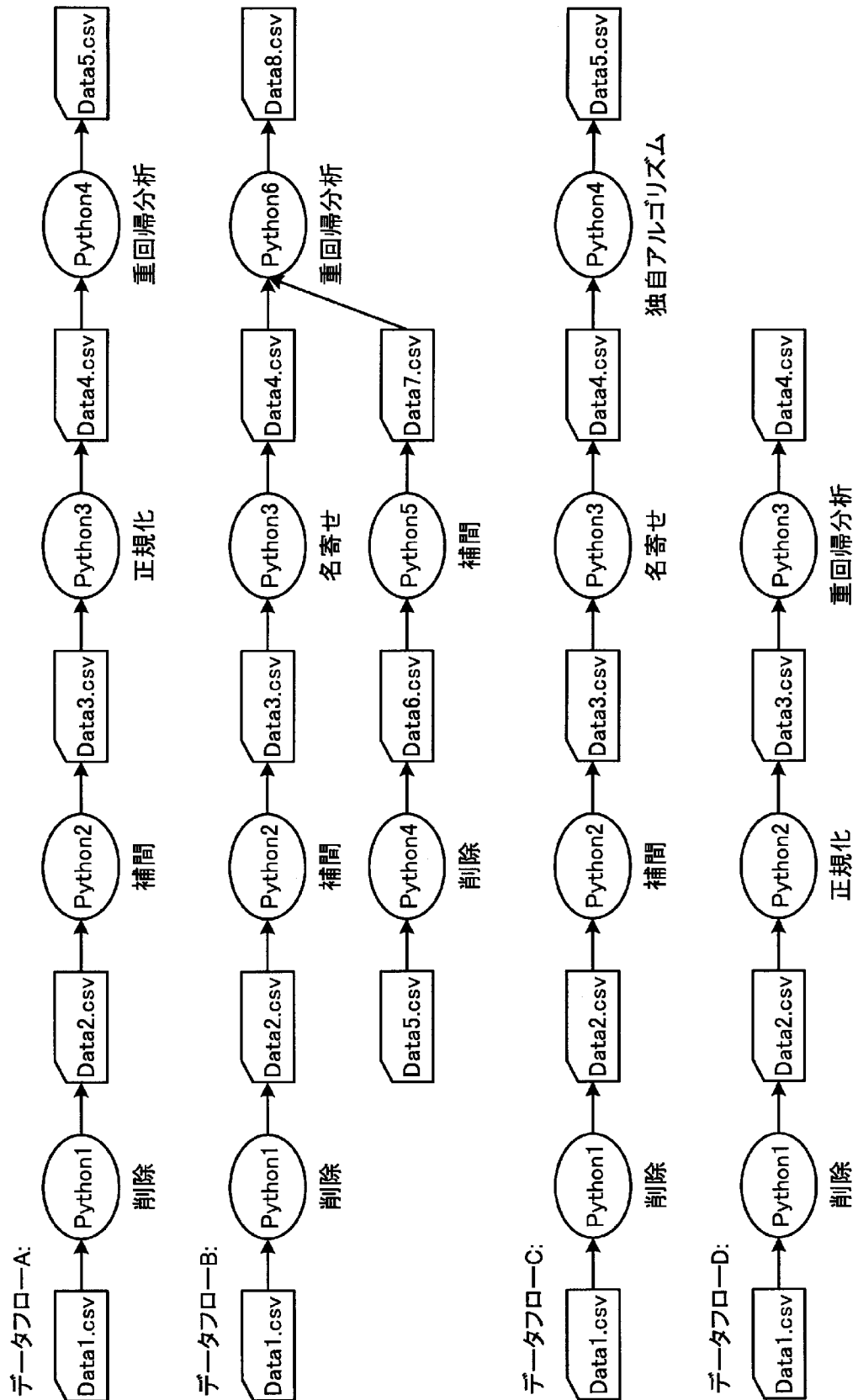
前記コンピュータに、

複数のデータフローからグループを抽出してグループ毎に他のグループとの類似度を計算し、該類似度とグループに含まれる処理の数に基づいて前記圧縮価値を計算して前記圧縮価値記憶部に格納する処理

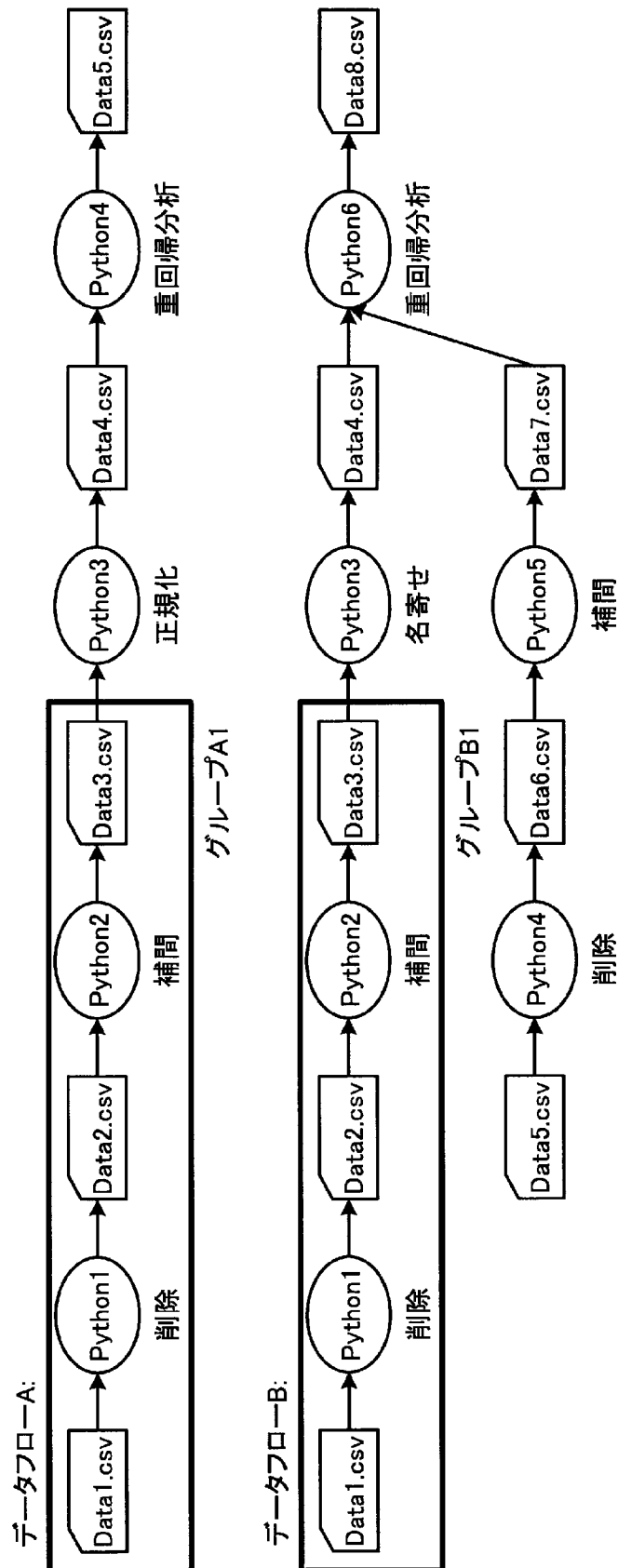
をさらに実行させることを特徴とする請求項 7 又は 8 に記載のリネージュプログラム。

[請求項10] 前記圧縮価値を計算する処理は、処理の入力データと出力データの統計的な差異、及び、該統計的な差異を生み出したアルゴリズムに基づいて、前記類似度を計算することを特徴とする請求項 9 に記載のリネージュプログラム。

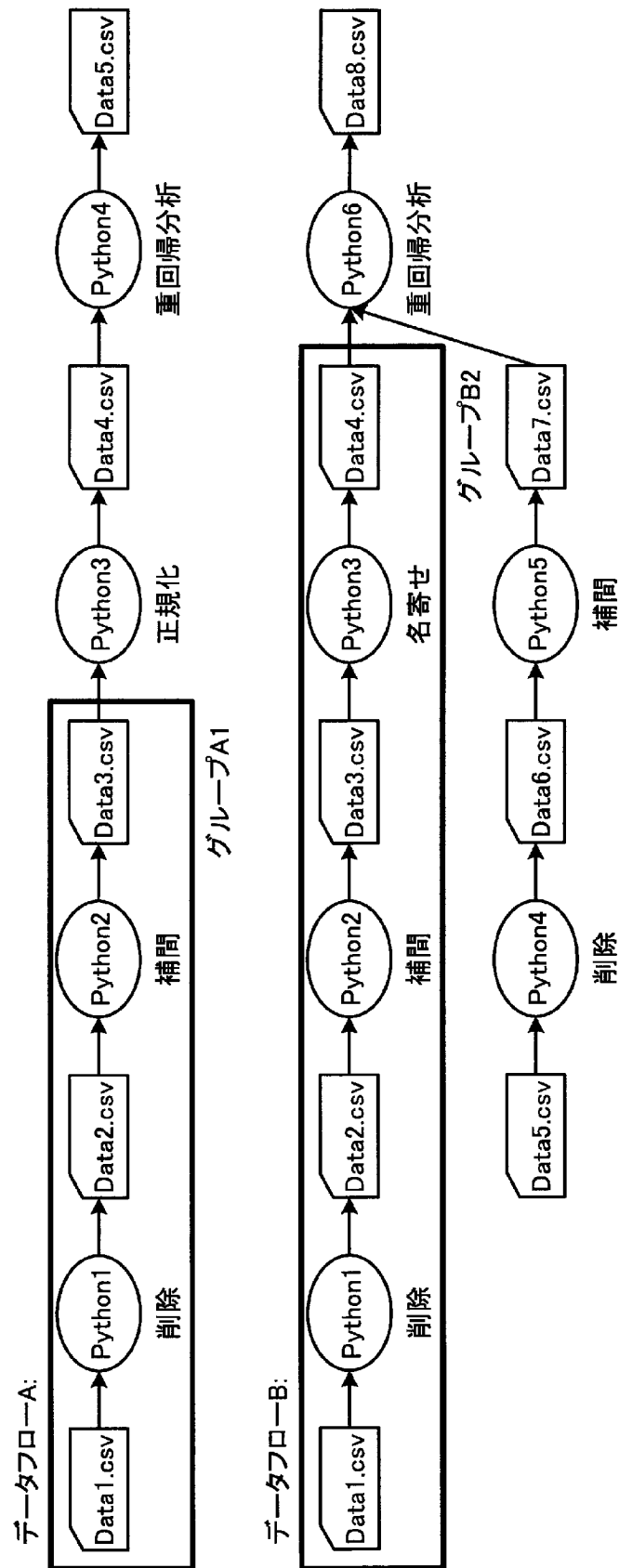
[図1A]



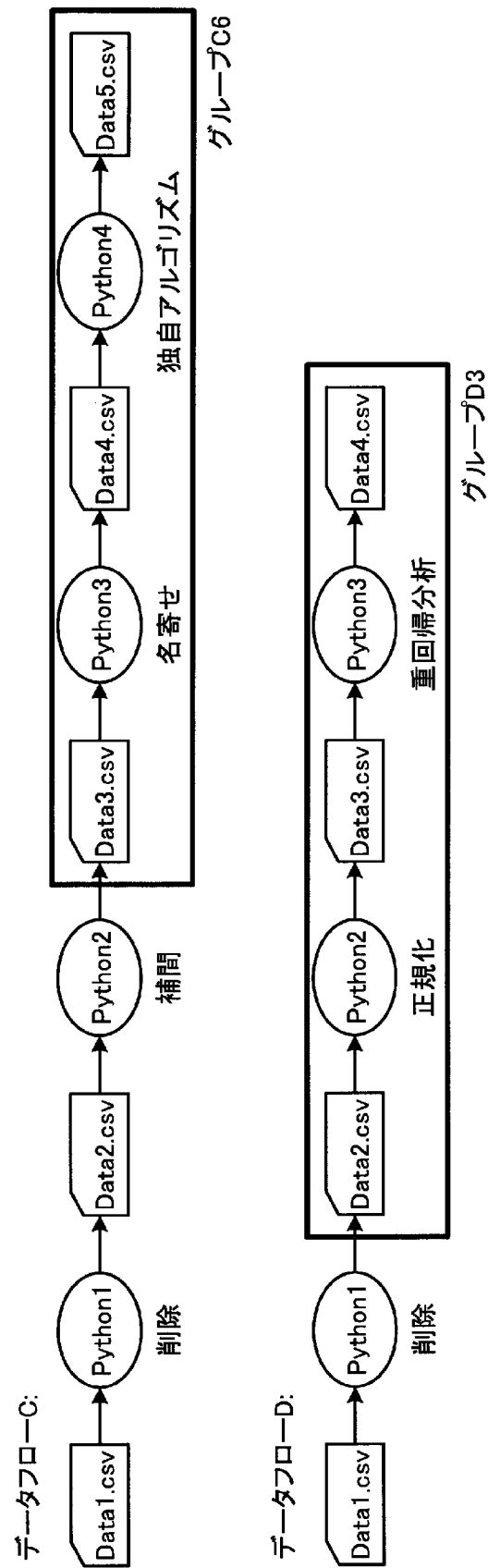
[図1B]



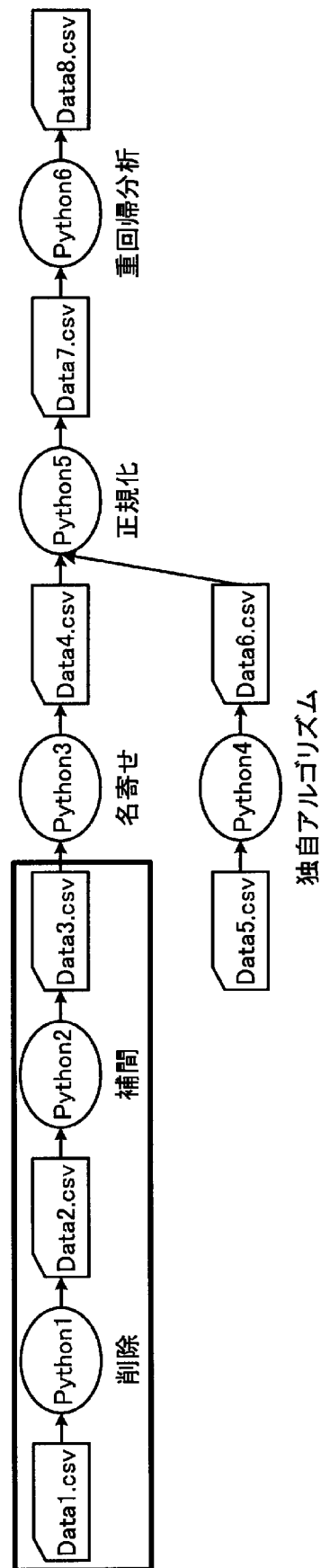
[図1C]



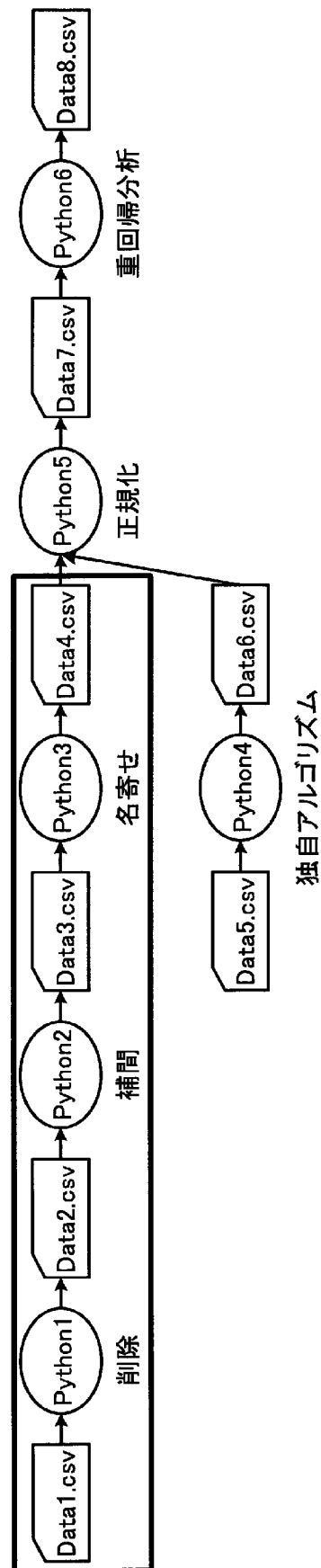
[図1D]



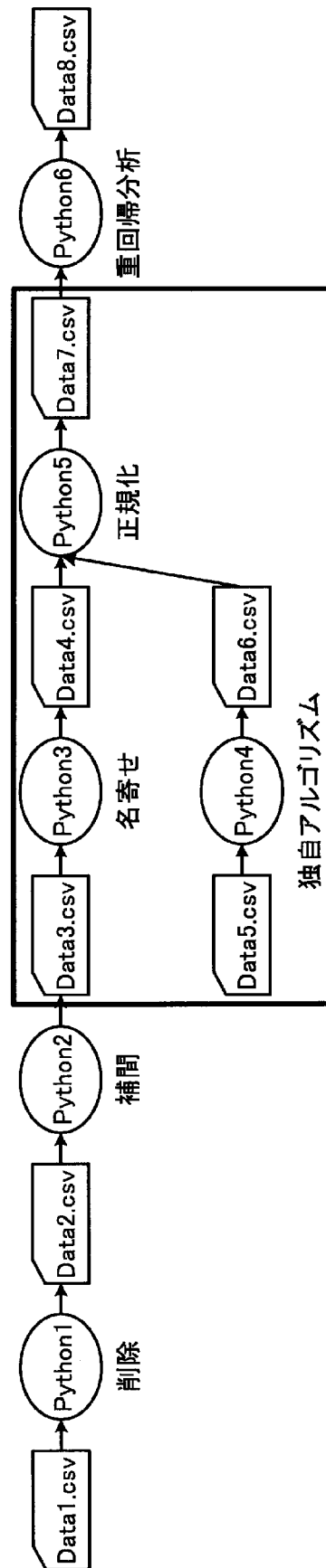
[図1E]



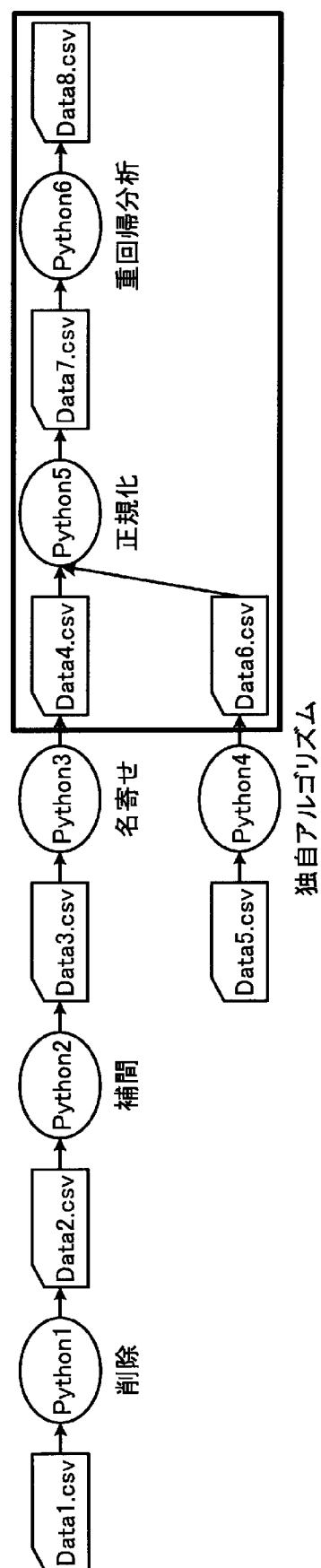
[図1F]



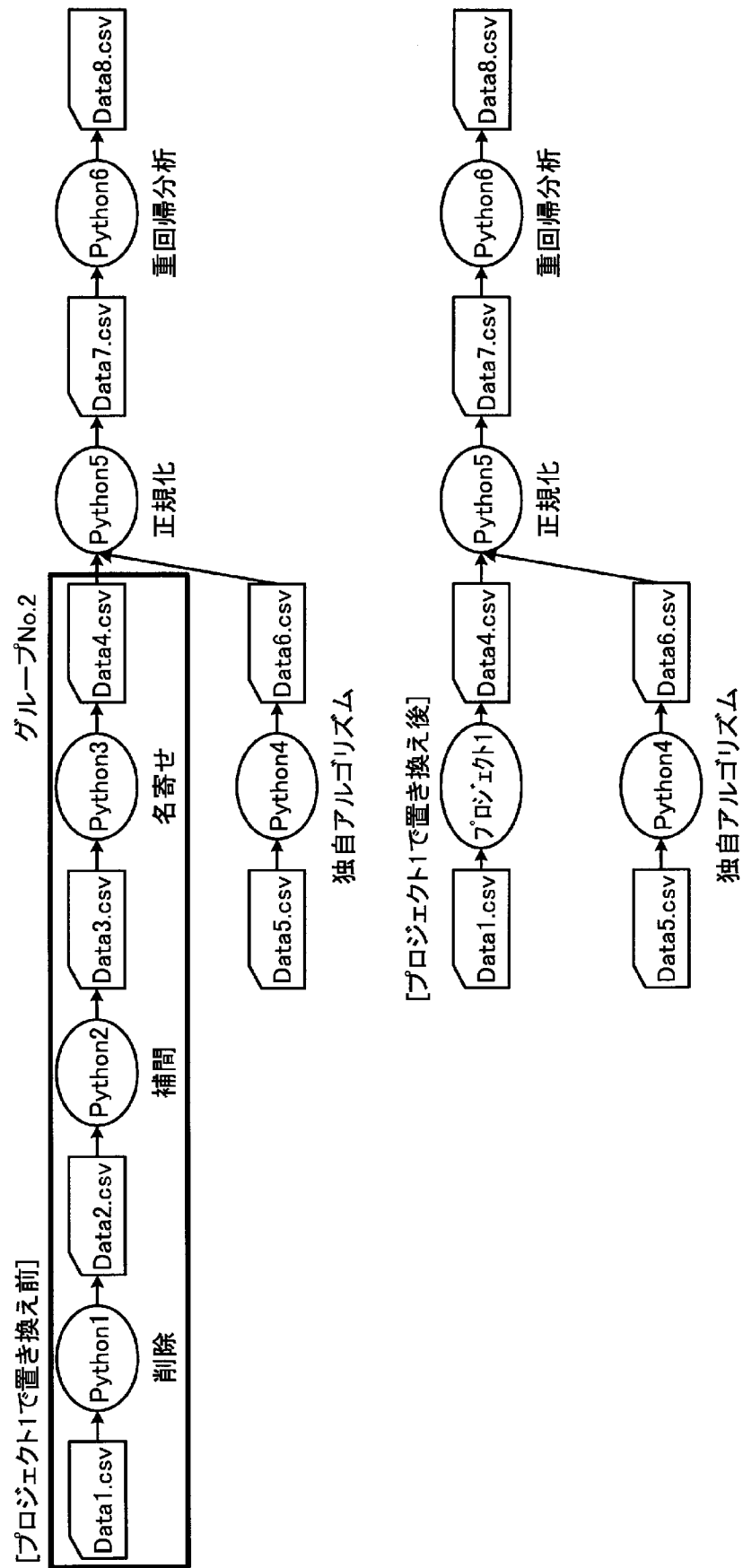
[図1G]



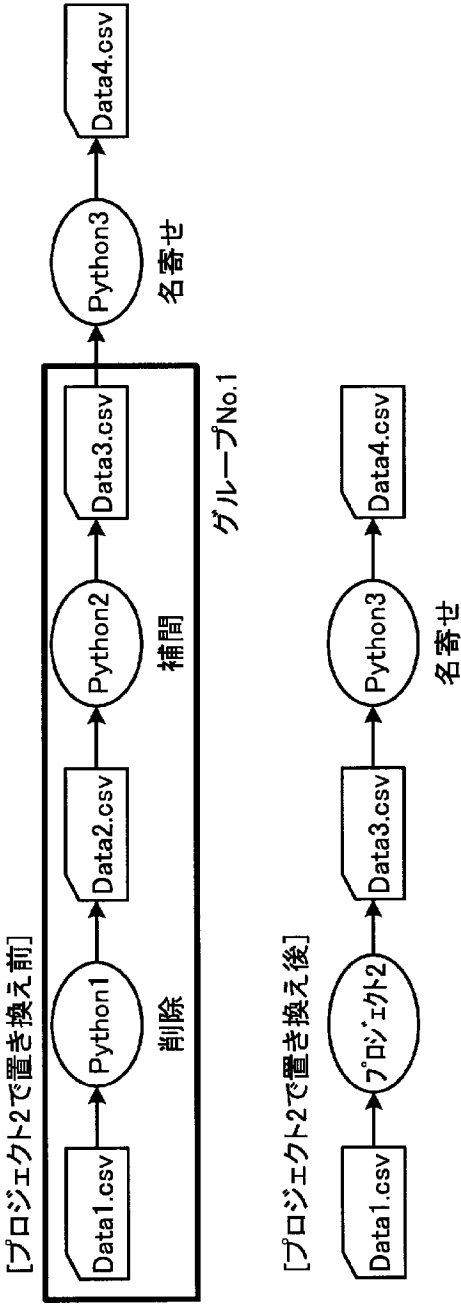
[図1H]



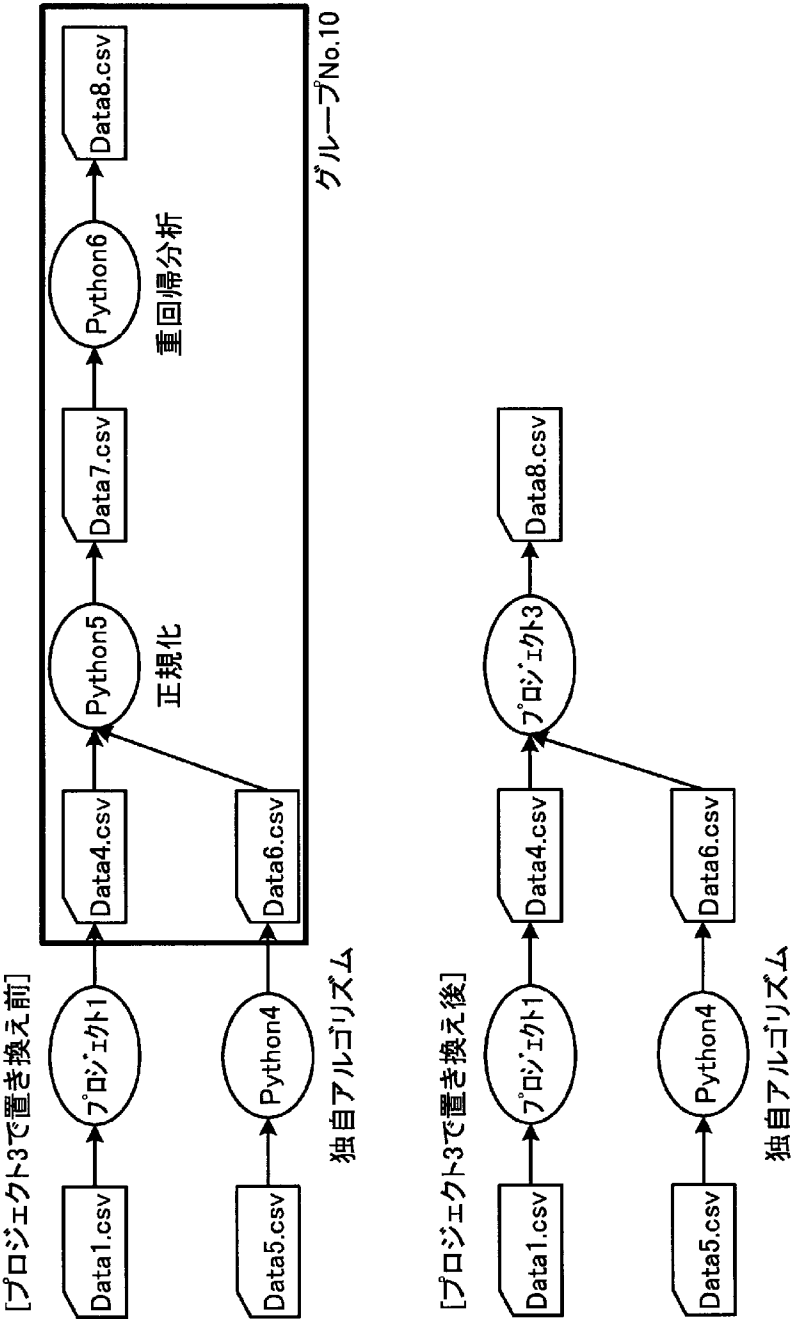
[図11]



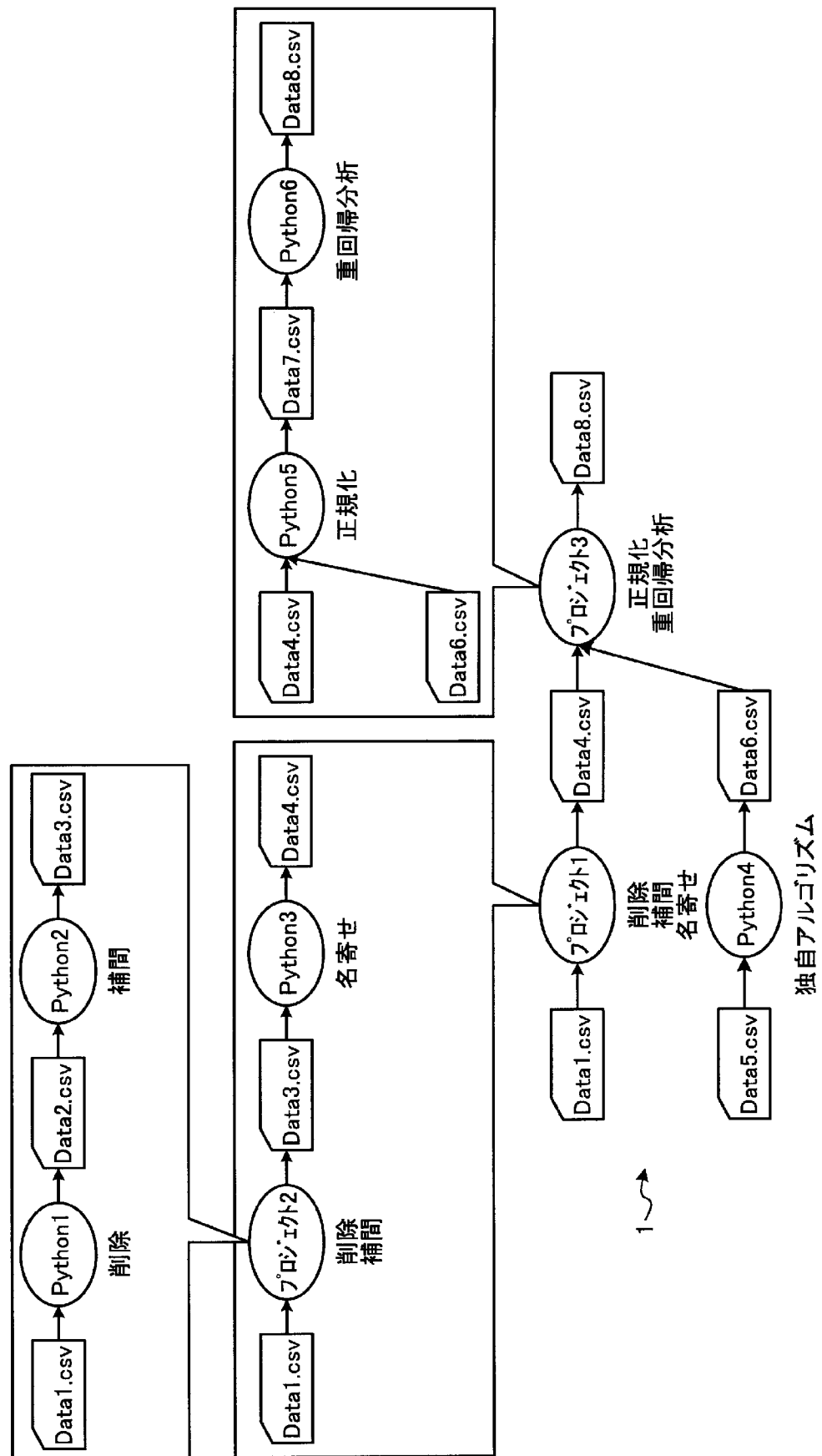
[図1J]



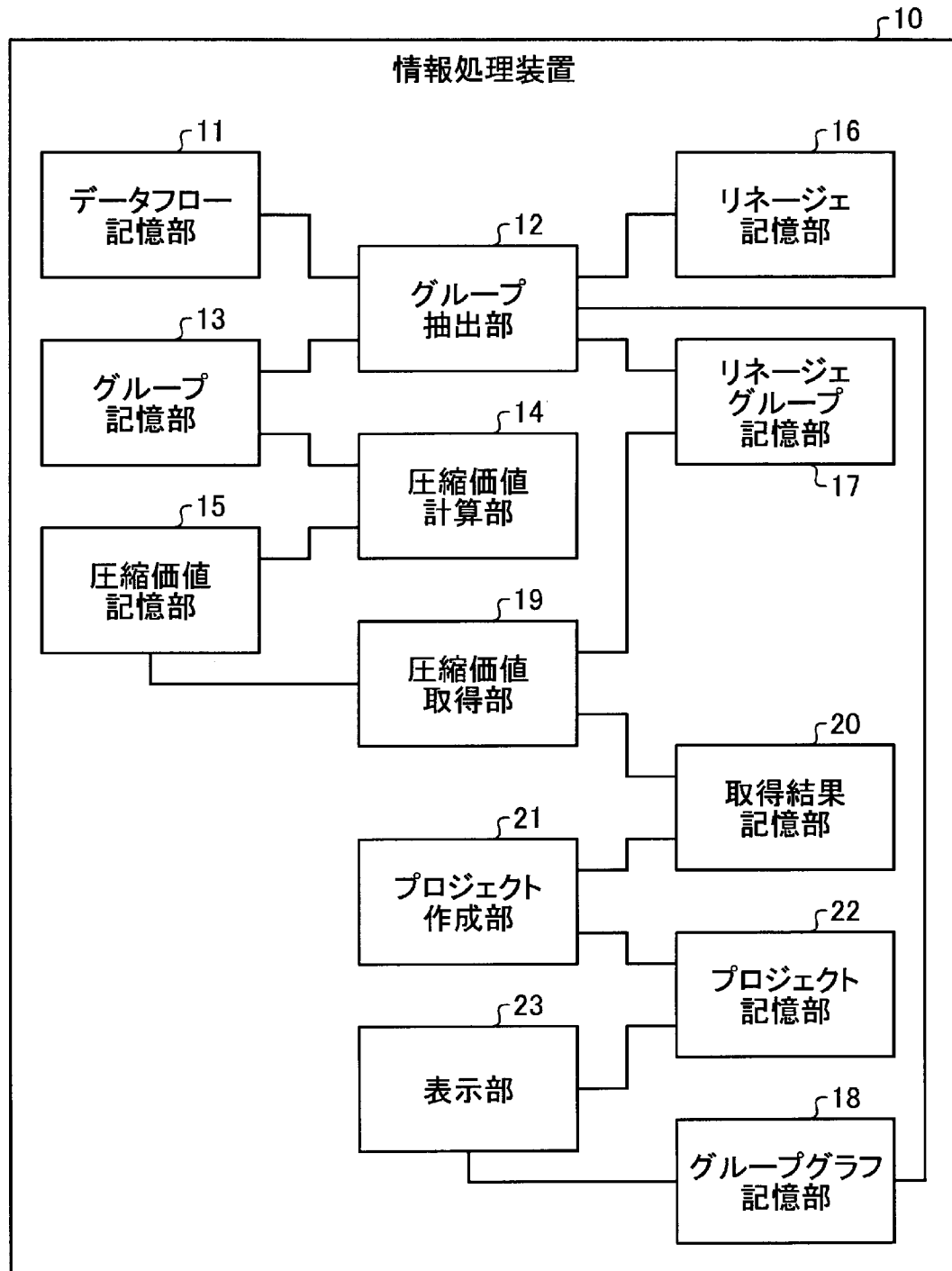
[図1K]



[図1L]



[図2]



[図3]

データフロー名	データフローのグラフ構造
データフローA	Data1.csv→Python1→Data2.csv
	Data2.csv→Python2→Data3.csv
	Data3.csv→Python3→Data4.csv
	Data4.csv→Python4→Data5.csv
データフローB	Data1.csv→Python1→Data2.csv
	Data2.csv→Python2→Data3.csv
	Data3.csv→Python3→Data4.csv
	Data5.csv→Python4→Data6.csv
	Data6.csv→Python5→Data7.csv
	Data4.csv&Data7.csv→Python6→Data8.csv
データフローC	Data1.csv→Python1→Data2.csv
	Data2.csv→Python2→Data3.csv
	Data3.csv→Python3→Data4.csv
	Data4.csv→Python4→Data5.csv
データフローD	Data1.csv→Python1→Data2.csv
	Data2.csv→Python2→Data3.csv
	Data3.csv→Python3→Data4.csv

[図4]

グループNo.	アルゴリズム名	統計的な差異
A1	削除→補間	行数の減少→値の数の増加
A2	削除→補間→正規化	行数の減少→値の数の増加→値の範囲の変更
...	...	...
A6	正規化→重回帰分析	値の範囲の変更→流用部品 (重回帰分析)
B1	削除→補間	行数の減少→値の数の増加
B2	削除→補間→名寄せ	行数の減少→値の数の増加→値の種類の減少
B3	削除→補間→名寄せ→重回帰分析	行数の減少→値の数の増加→値の種類の減少→流用部品 (重回帰分析)
B4	補間→名寄せ	値の数の増加→値の種類の減少
...	...	...
B15	削除→補間→名寄せ, 削除→補間, 重回帰分析	行数の減少→値の数の増加→値の種類の減少, 行数の減少→値の数の増加, 流用部品 (重回帰分析)
C1	削除→補間	行数の減少→値の数の増加
C2	削除→補間→名寄せ	行数の減少→値の数の増加→値の種類の減少
...	...	...
C6	名寄せ→独自アルゴリズム	値の種類の減少→独自アルゴリズム
D1	削除→正規化	行数の減少→値の範囲の変更
D2	削除→正規化→重回帰分析	行数の減少→値の範囲の変更→流用部品 (重回帰分析)
D3	正規化→重回帰分析	値の範囲の変更→流用部品 (重回帰分析)

[図5]

アルゴリズム名	統計的な差異	アルゴリズム数	類似度	圧縮価値
削除→補間	行数の減少→値の数の増加	2	5	160
削除→補間→正規化	行数の減少→値の数の増加→値の範囲の変更	3	0	0
...	...	...	0	0
正規化→重回帰分析	値の範囲の変更→流用部品 (重回帰分析)	2	1	32
削除→補間→名寄せ	行数の減少→値の数の増加→値の種類の減少	3	1	243
削除→補間→名寄せ→重回帰分析	行数の減少→値の数の増加→値の種類の減少→流用部品 (重回帰分析)	4	0	0
補間→名寄せ	値の数の増加→値の種類の減少	2	1	32
...	...	...	0	0

[図6]

No.	要素のグラフ構造
1	Data1.csv→Python1→Data2.csv
2	Data2.csv→Python2→Data3.csv
3	Data3.csv→Python3→Data4.csv
4	Data5.csv→Python4→Data6.csv
5	Data4.csv&Data6.csv→Python5→Data7.csv
6	Data7.csv→Python6→Data8.csv

[図7]

グループNo.	グループのグラフ構造
1	Data1.csv→Python1→Data2.csv, Data2.csv→Python2→Data3.csv
2	Data1.csv→Python1→Data2.csv, Data2.csv→Python2→Data3.csv, Data3.csv→Python3→Data4.csv
3	Data1.csv→Python1→Data2.csv, Data2.csv→Python2→Data3.csv, Data3.csv→Python3→Data4.csv, Data4.csv→Python5→Data7.csv
...	...
18	Data1.csv→Python1→Data2.csv, Data2.csv→Python2→Data3.csv, Data3.csv→Python3→Data4.csv, Data5.csv→Python4→Data6.csv, Data4.csv&Data6.csv→Python5→Data7.csv, Data7.csv→Python6→Data8.csv

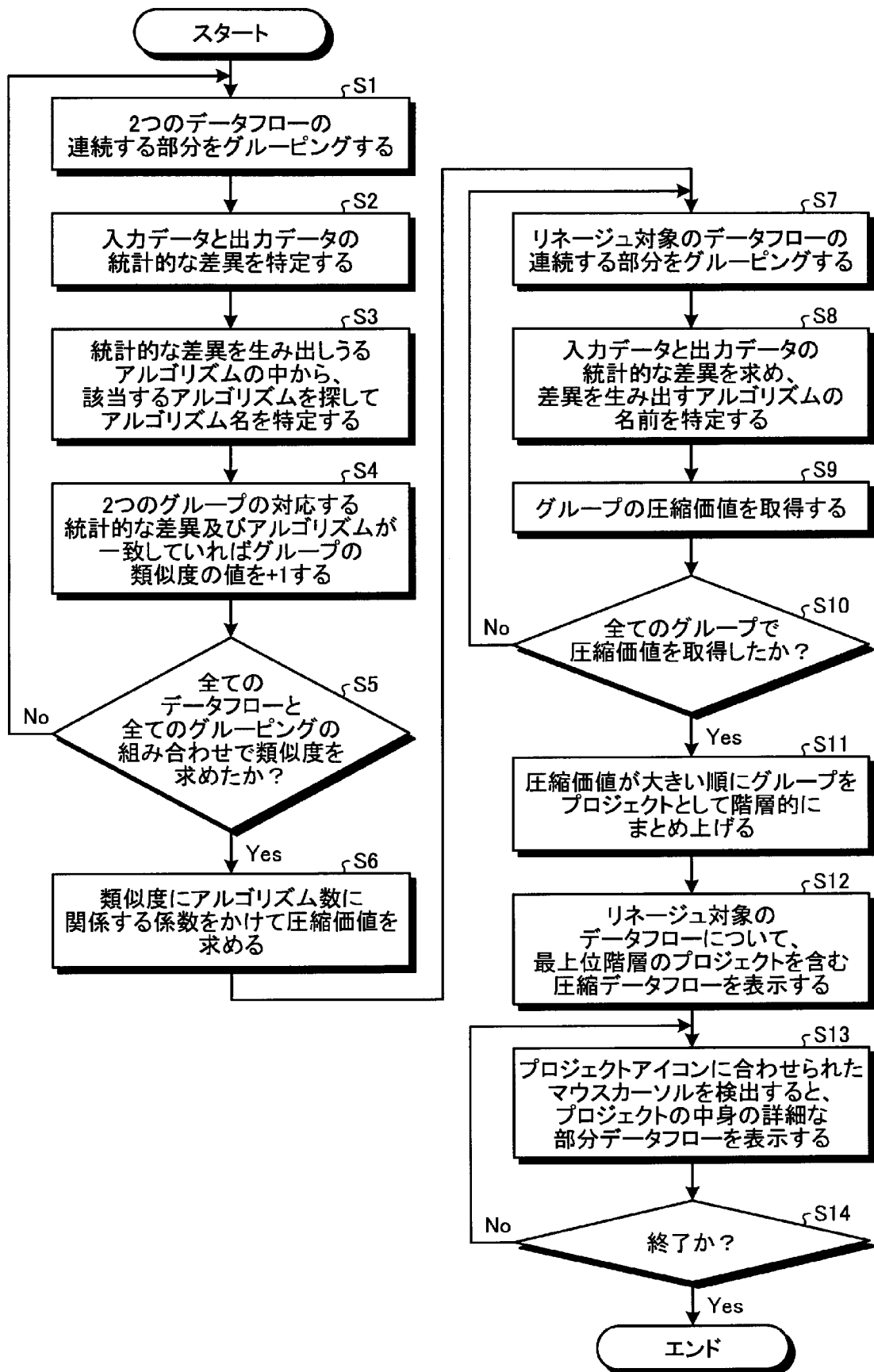
[図8]

グループNo.	アルゴリズム名	統計的な差異	圧縮価値
1	削除→補間	行数の減少→値の数の増加	160
2	削除→補間→名寄せ	行数の減少→値の数の増加→値の種類の減少	243
3	削除→補間→名寄せ→正規化	行数の減少→値の数の増加→値の種類の減少→値の範囲の変更	0
4	削除→補間→名寄せ→正規化→重回帰分析	行数の減少→値の数の増加→値の種類の減少→値の範囲の変更→流用部品 (重回帰分析)	0
5	補間→名寄せ	値の数の増加→値の種類の減少	32
6	補間→名寄せ→正規化	値の数の増加→値の種類の減少→値の範囲の変更	0
7	補間→名寄せ→正規化→重回帰分析	値の数の増加→値の種類の減少→値の範囲の変更→流用部品 (重回帰分析)	0
8	名寄せ→正規化	値の種類の減少→値の範囲の変更	0
9	名寄せ→正規化→重回帰分析	値の種類の減少→値の範囲の変更→流用部品(重回帰分析)	0
10	正規化→重回帰分析	値の範囲の変更→流用部品 (重回帰分析)	32
11~18	独自アルゴリズム→ α (例えば、正規化)	独自→ α (例えば、値の範囲の変更)	0

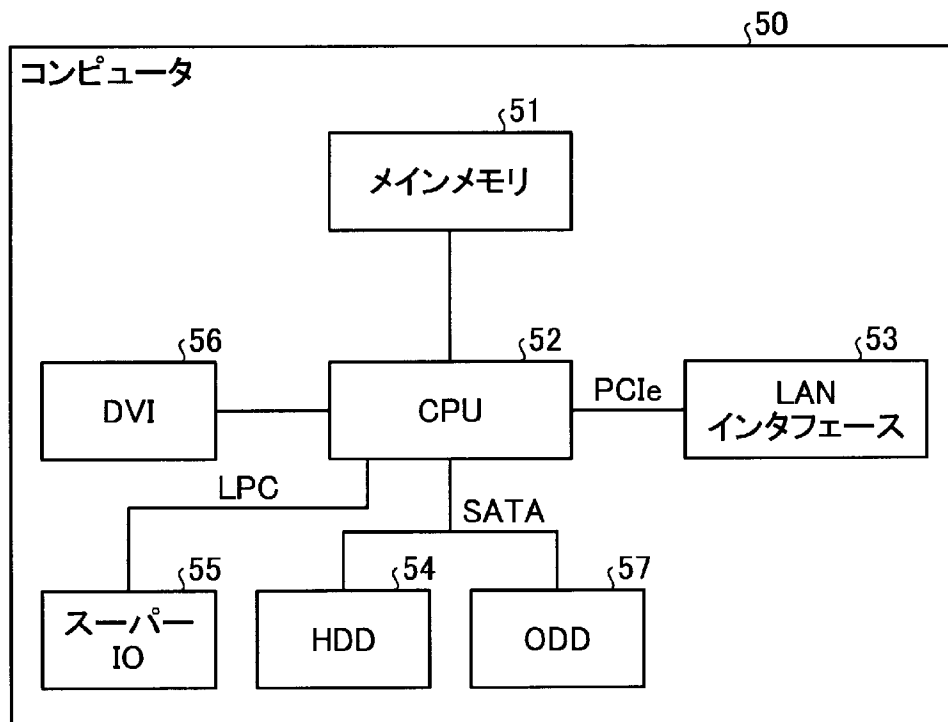
[図9]

グループNo.	アルゴリズム名	統計的な差異	圧縮価値	プロジェクト名	階層
2	削除→補間→名寄せ	行数の減少→値の数の増加→値の種類の減少	243	Project1	1
1	削除→補間	行数の減少→値の数の増加	160	Project2	2
5	補間→名寄せ	値の数の増加→値の種類の減少	32	-	-
10	正規化→重回帰分析	値の範囲の変更→流用部品（重回帰分析）	32	Project3	1

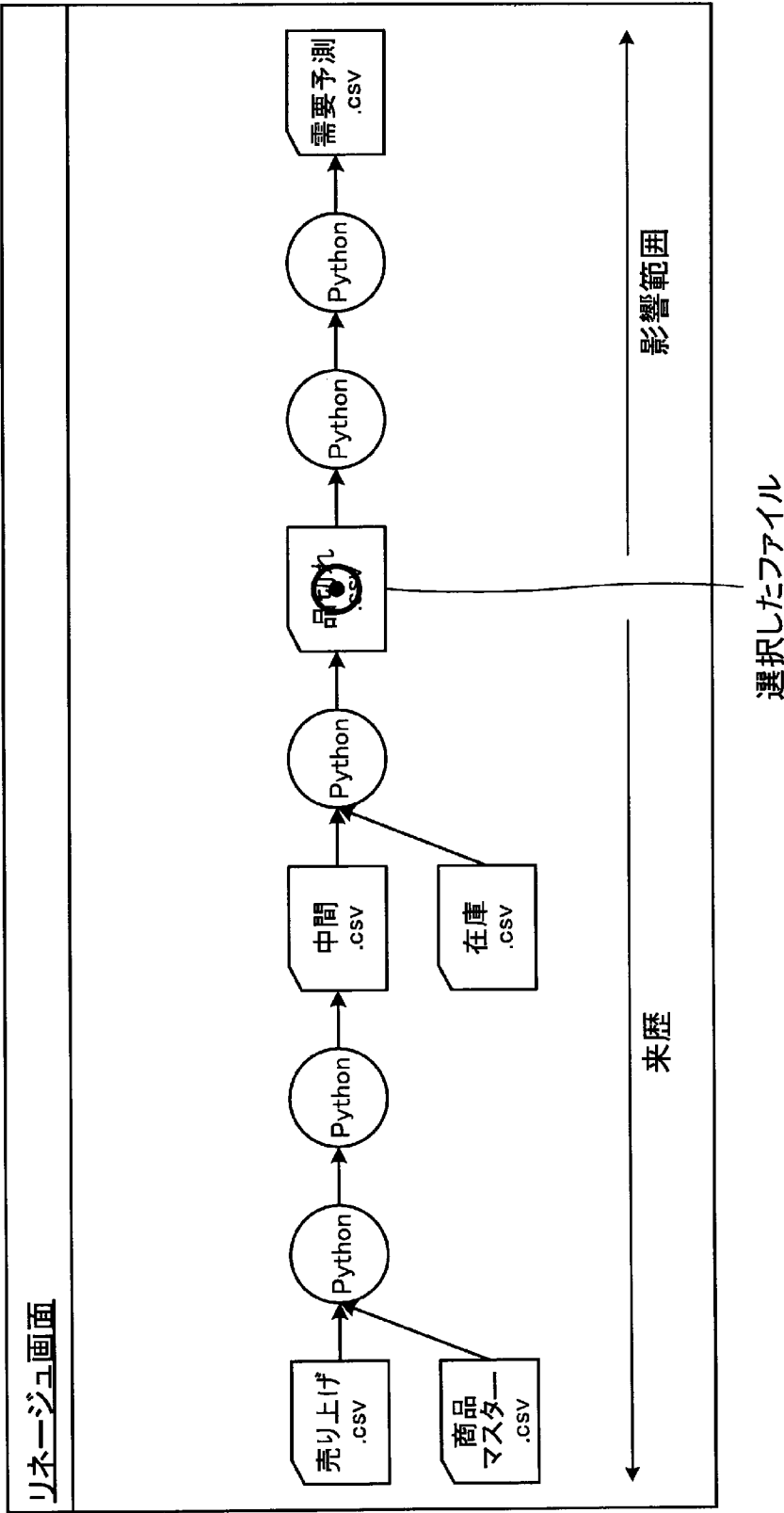
[図10]



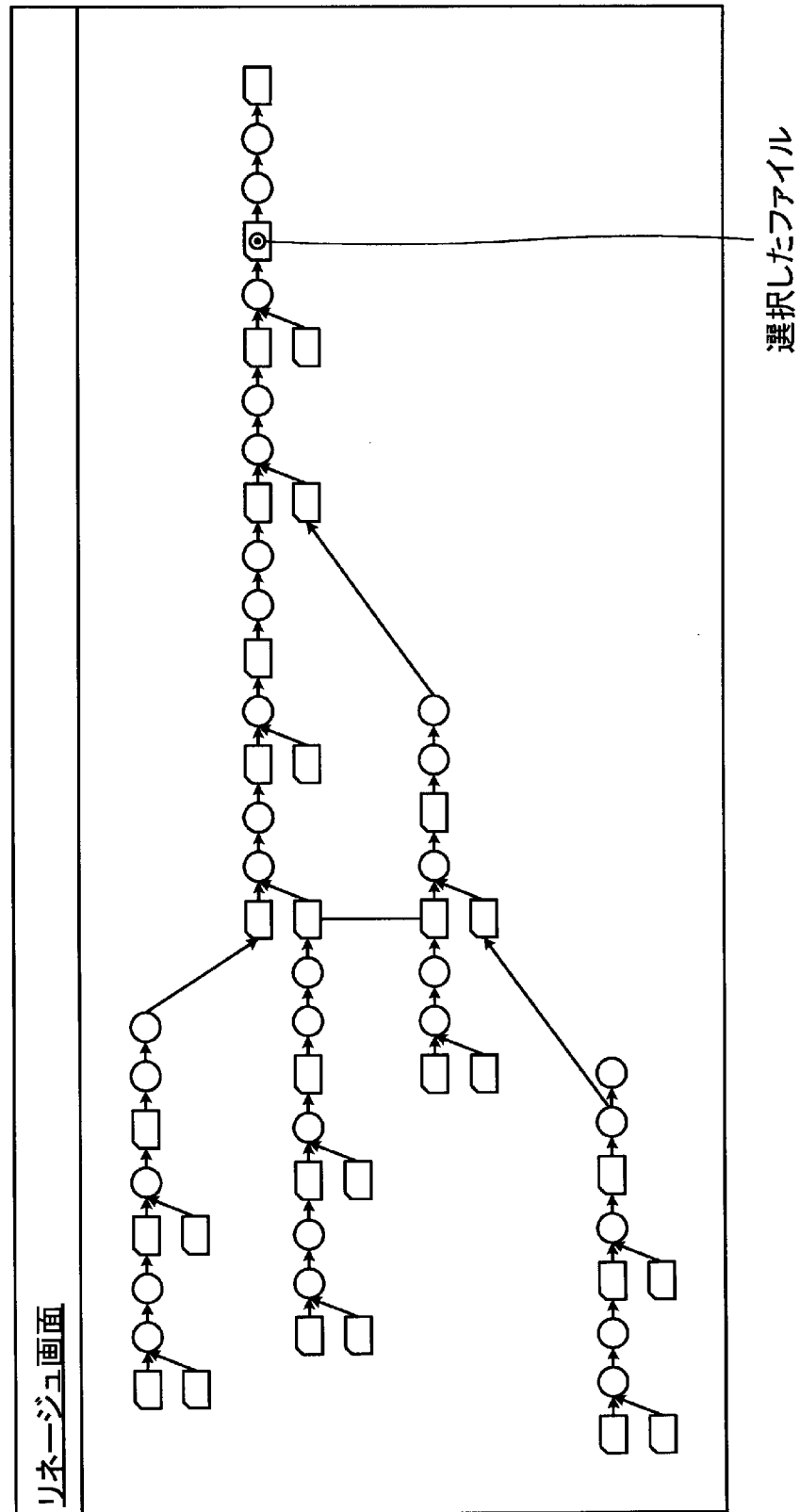
[図11]



[図12]



[図13]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2018/043831

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06Q10/10 (2012.01) 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)

Int.Cl. G06Q10/10

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

Published examined utility model applications of Japan 1922-1996

Published unexamined utility model applications of Japan 1971-2019

Registered utility model specifications of Japan 1996-2019

Published registered utility model applications of Japan 1994-2019

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.
A	JP 2005-122703 A (SAP AG) 12 May 2005, whole document & US 2005/0108217 A1	1-10
A	JP 2013-513143 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 18 April 2013, whole document & US 2011/0137922 A1 & CN 102640150 A	1-10
A	US 2007/0061353 A1 (BOBBIN, N. V. et al.) 15 March 2007, whole document (Family: none)	1-10

☐ Further documents are listed in the continuation of Box C.	☐ See patent family annex.
---	---

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 06.02.2019	Date of mailing of the international search report 19.02.2019
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Authorized officer Telephone No.

A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int.Cl. G06Q10/10(2012.01)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. G06Q10/10

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	1971-2019年
日本国実用新案登録公報	1996-2019年
日本国登録実用新案公報	1994-2019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
A	JP 2005-122703 A (エスエイピー エイジー) 2005.05.12, Whole Document & US 2005/0108217 A1	1-10
A	JP 2013-513143 A (インターナショナル・ビジネス・マシーンズ・ コーポレーション) 2013.04.18, Whole Document & US 2011/0137922 A1 & CN 102640150 A	1-10
A	US 2007/0061353 A1 (BOBBIN, Nathan Vernon et al.) 2007.03.15, Whole Document (ファミリーなし)	1-10

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	「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
「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	「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
「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)	「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
「O」 口頭による開示、使用、展示等に言及する文献	「&」 同一パテントファミリー文献
「P」 国際出願日前で、かつ優先権の主張の基礎となる出願	

国際調査を完了した日

06.02.2019

国際調査報告の発送日

19.02.2019

国際調査機関の名称及びあて先

日本国特許庁 (ISA/J P)

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)

原 忠

5 L

5880

電話番号 03-3581-1101 内線 3562