

(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11月19日(19.11.2020)



(10) 国際公開番号

WO 2020/230278 A1

(51) 国際特許分類:
G06N 20/00 (2019.01) G06F 16/215 (2019.01)

(21) 国際出願番号: PCT/JP2019/019191

(22) 国際出願日: 2019年5月14日(14.05.2019)

(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: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).

(72) 発明者: 長 裕敏(CHO, Hirotoshi); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT 知的財産センタ内 Tokyo (JP). 山室 健(YAMAMURO, Takeshi); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT 知的財産センタ内 Tokyo (JP). 内山 寛之(UCHIYAMA, Hiroyuki); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT 知的財産センタ内 Tokyo (JP).

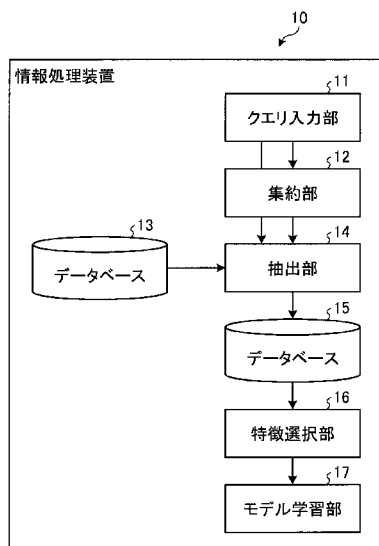
(74) 代理人: 特許業務法人酒井国際特許事務所 (SAKAI INTERNATIONAL PATENT OFFICE); 〒1000013 東京都千代田区霞が関3丁目8番1号 虎の門三井ビルディング Tokyo (JP).

(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, 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,

(54) Title: INFORMATION PROCESSING DEVICE, EXTRACTION METHOD, AND EXTRACTION PROGRAM

(54) 発明の名称: 情報処理装置、抽出方法および抽出プログラム



- 10 Information processing device
- 11 Query input unit
- 12 Summarizing unit
- 13, 15 Database
- 14 Extracting unit
- 16 Feature selecting unit
- 17 Model learning unit

(57) Abstract: This information processing device (10) has: a summarizing unit (12) that, for each record in a dimension table, counts the number of records in a table in which are summarized the records in which feature quantities other than master keys overlap or are similar; and an extracting unit (14) that, when the ratio of the number of records in a factor table with respect to the number of records counted by the summarizing unit (12) exceeds a predetermined threshold, extracts analysis data which includes external keys of the summarized table as representative values of feature groups of the dimension table.

(57) 要約: 情報処理装置 (10) は、ディメンションテーブルにおける各レコードについて、主キー以外の特徴量が重複または類似しているレコードを集約したテーブルのレコード数をカウントする集約部 (12) と、集約部 (12) によってカウントされたレコード数に対するファクトテーブルのレコード数の比率が所定の閾値を超える場合には、集約したテーブルの外部キーをディメンションテーブルの特徴群の代表値として含む分析用データを抽出する抽出部 (14) とを有する。

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)）

明 細 書

発明の名称： 情報処理装置、抽出方法および抽出プログラム

技術分野

[0001] 本発明は、情報処理装置、抽出方法および抽出プログラムに関する。

背景技術

[0002] 従来、機械学習では分析用データを抽出するための前処理として、データベースに格納された複数テーブルを主キーと外部キーを用いて結合処理を行い単一テーブルに変換する。また、機械学習モデルの予測精度向上のために抽出した全特微量から予測したい特微量に強い関連がある特微量の組み合わせを選択する特徴選択を行う。これらの処理によって生じる遅延は、データサイズとともに増大し、大規模データを用いたデータ分析においては無視できないものとなっている。

[0003] スタースキーマのデータ構造を持つデータベースにおいて、ファクトテーブルが持つ外部キーを、その外部キーが参照するディメンションテーブル内の全特微量の代表とみなすことでテーブルの結合処理を一部省略する手法がある（例えば、非特許文献1、2参照）。手法が適用可能かの判定は、Tuple Ratio (TR) というファクトテーブルのレコード数 n_s と外部キーが指すレコードの総数、すなわちディメンションテーブルのレコード数 n_r の比を指標値として行う。 n_s が n_r より十分大きい（例えば、 $n_s / n_r > 20$ ）場合、テーブルの結合を省略しても機械学習モデルの精度が低下する可能性は限りなく低いため、外部キーをディメンションテーブルの特徴群の代表値として特微量を削減することで、機械学習の前処理で特に時間がかかる特徴選択処理時間の削減が望める。

先行技術文献

非特許文献

[0004] 非特許文献1：A. Kumar, J. Naughton, J. M. Patel, and X. Zhu.
"To Join or Not to Join? Thinking Twice about Joins before

e Feature Selection” In SIGMOD, 2016. pp. 19-34.

非特許文献2: V. Shah, A. Kumar, and X. Zhu. ” Are Key-Foreign Key Joins Safe to Avoid when Learning High-Capacity Classifiers?” PVLDB, vol.11, no.3, pp.366-379, 2017.

発明の概要

発明が解決しようとする課題

[0005] しかしながら、従来の手法では、特徴選択にかかる時間を削減することができない場合があるという課題があった。例えば、従来の手法では、機械学習モデルの精度を保ちつつ、結合処理を省略するためには、 $n_s / n_r > 20$ の条件を満たす必要があるため、手法が適用できないケースが存在する。

課題を解決するための手段

[0006] 上述した課題を解決し、目的を達成するために、本発明の情報処理装置は、ディメンションテーブルにおける各レコードについて、主キー以外の特徴量が重複または類似しているレコードを集約したテーブルのレコード数をカウントする集約部と、前記集約部によってカウントされたレコード数に対するファクトテーブルのレコード数の比率が所定の閾値を超える場合には、前記集約したテーブルの外部キーを前記ディメンションテーブルの特徴群の代表値として含む分析用データを抽出する抽出部とを有することを特徴とする。

発明の効果

[0007] 本発明によれば、より多くのケースに対して特徴選択にかかる時間を削減することができるという効果を奏する。

図面の簡単な説明

[0008] [図1]図1は、第1の実施形態に係る情報処理装置の構成例を示すブロック図である。

[図2]図2は、映画の評価予測データセットの一例を示す図である。

[図3]図3は、集約したテーブルのレコード数を算出する処理例を説明する図

である。

[図4]図4は、集約したテーブルのレコード数を算出する処理例を説明する図である。

[図5]図5は、分析用データセットを抽出する処理例を説明する図である。

[図6]図6は、第1の実施形態に係る情報処理装置における処理の流れの一例を示すフローチャートである。

[図7]図7は、抽出プログラムを実行するコンピュータを示す図である。

発明を実施するための形態

[0009] 以下に、本願に係る情報処理装置、抽出方法および抽出プログラムの実施の形態を図面に基づいて詳細に説明する。なお、この実施の形態により本願に係る情報処理装置、抽出方法および抽出プログラムが限定されるものではない。

[0010] [分析用データ抽出処理]

ここでは、まず、分析用データを抽出する処理についての前提となる技術を説明する。一般に予測モデル構築のワークフローでは、分析用データを抽出するためにデータベース上に格納された複数のテーブルに分散して保存されているデータを結合して単一のテーブルにまとめるデータ抽出処理を行う。Sをファクトテーブル、Yを予測したい特徴量、 X_S をSの特徴群、kをSが持つ外部キーの総数、 FK_i を i ($1 \leq i \leq k$) 番目の外部キー、 R_i を i 番目のディメンションテーブル、 RID_i を R_i の主キー、 X_{R_i} を R_i の特徴群とすると、情報処理装置は、下記表1のようなSQLクエリをデータベースに入力して分析用データを抽出する。

[0011] [表1]

(表1)

SQL 1 分析用データ取得クエリ例
SELECT Y, X_S , X_{R_1} , ..., X_{R_k}
FROM S, R_1 , ..., R_k
WHERE $S.FK_1 = R_1.RID_1, \dots, S.FK_k = R_k.RID_k,$

[0012] その後、情報処理装置は、取得した分析用データT ($Y, X_S, X_{R_1}, \dots,$

X_{Rk}) に対して、特徴選択を行い X_S と X_R の内から Y に強い関連がある特徴量の組み合わせを抽出することで予測精度を高める。特徴選択の実現手法の例としてラッパー法とフィルター法の二つがあげられる。

[0013] ラッパー法は、 X_S 、 X_R の部分集合を使って実際にモデルを作り、交差確認等で求めた汎化誤差を最小にする特徴群の部分集合を選択する手法である。また、フィルター法は、 Y と各 X_S 、 X_R との情報ゲインといった特徴量の良さの基準で部分集合を選択する手法である。一般に予測精度が高いラッパー法を用いることが多いが、特徴量の数が多いと部分集合の組み合わせ爆発が起こるため計算コストが大きく予測モデル構築ワークフローの中でも特にボトルネックになる。

[0014] なお、上記した非特許文献 1 の手法では、 FK_i を X_{Ri} の代表値として使用することで生じるモデル精度低下リスクを定量化している。これにより、精度への影響がすくない結合処理を省略することで特徴量を削減し、特徴選択における特徴組み合わせの計算コストを削減することでワークフローの高速化を可能にしている。 S と R_i における X_{Ri} を FK_i に変換した時のモデル精度低下のリスク指標値 $\text{Tuple ratio}(TR_i)$ は、 n_S を S のレコード数、 n_{R_i} を R_i のレコード数とした時、 $TR_i = n_S / n_{R_i}$ で表される（導出詳細は非特許文献 1 参照）。

[0015] この手法を用いることで、下記 (1) 式の時、 X_{Ri} を FK_i に変換することで高速化が可能だが、 n_S が n_{Ri} より十分大きくないと適用することができず、手法が適用できないケースが多く存在する。

[0016] [数1]

$$1/TR_i \approx 0 (TR_i > 20) \quad \dots(1)$$

[0017] 以下で説明する本実施形態では、情報処理装置が、 TR_i を用いた手法において、 X_{Ri} に対して重複あるいは類似したレコードを集約し、 R_i が保持する意味情報を保持しつつ、 n_R を削減する。 R'_i を R_i に対して X_{Ri} をキーに集約してレコードの削減を行ったディメンションテーブル、 n'_{Ri} を R'_i のレコード数、 FK'_i を集約された R'_i の各グループを参照する外部キーとす

る。ここで $TR'_i = n_s / n'_{Ri}$ とした時、特徴変換判定を下記（２）式のよう
に定義する。

[0018] [数2]

$$1/TR'_i \approx 0 (TR'_i > 20) \quad \dots(2)$$

[0019] 情報処理装置が、特徴変換判定の条件を満たした時に FK'_i を X_{Ri} の代表
値として特徴変換を行う。ここで、 $n'_{Ri} \leq n_{Ri}$ なので TR_i の適用範囲が増
え、より多くのケースに対して精度を担保したまま特徴選択にかかる時間を
削減することができる。つまり、 n_s が n'_{Ri} より十分大きい（例えば、 n_s
/ $n'_{Ri} > 20$ ）場合、テーブルの結合を省略しても機械学習モデルの精度
が低下する可能性は限りなく低いため、外部キーをディメンションテーブル
の特徴群の代表値として特徴量を削減することで、機械学習の前処理で特に
時間がかかる特徴選択処理時間の削減が望める。

[0020] [第１の実施形態]

以下の実施の形態では、第１の実施形態に係る情報処理装置１０の構成、
情報処理装置１０の処理の流れを順に説明し、最後に第１の実施形態による
効果を説明する。

[0021] [情報処理装置の構成]

まず、図１を用いて、本実施形態の情報処理装置１０を含むシステムの構
成例を説明する。図１は、第１の実施形態に係る情報処理装置の構成例を示
すブロック図である。なお、図１に示す構成は一例にすぎず、具体的な構成
は特に限定されない。

[0022] 図１に例示するように、情報処理装置１０は、クエリ入力部１１、集約部
１２、データベース１３、抽出部１４、データベース１５、特徴選択部１６
およびモデル学習部１７を有する。

[0023] クエリ入力部１１、集約部１２、抽出部１４、特徴選択部１６およびモデ
ル学習部１７は、各種のプログラムが動作することにより各種の処理部とし
て機能する。例えば、クエリ入力部１１、集約部１２、抽出部１４、特徴選
択部１６およびモデル学習部１７は、ＣＰＵ（Central Processing Unit）

、MPU (Micro Processing Unit) 等の電子回路や、ASIC (Application Specific Integrated Circuit)、FPGA (Field Programmable Gate Array) 等の集積回路である。

[0024] また、データベース13、15は、例えば、HDD (Hard Disk Drive)、SSD (Solid State Drive)、光ディスク等の記憶装置である。データベース13、15は、RAM (Random Access Memory)、フラッシュメモリ、NVS RAM (Non Volatile Static Random Access Memory) 等のデータを書き換え可能な半導体メモリであってもよい。

[0025] クエリ入力部11は、分析用データ取得のためのクエリの入力を受け付け、受け付けたクエリをそのまま集約部12および抽出部14に出力する。例えば、クエリ入力部11は、分析用データ取得のためのクエリとして、前述の表1に例示するようなクエリの入力を受け付け、集約部12および抽出部14に出力する。

[0026] 集約部12は、ディメンションテーブルにおける各レコードについて、主キー以外の特徴量が重複または類似しているレコードを集約したテーブルのレコード数をカウントする。具体的には、集約部12は、クエリ入力部11によって分析用データを取得するためのクエリの入力が受け付けられた場合に、主キー以外の特徴量が同一のレコードまたは類似のレコードを除去したテーブルのレコード数をカウントする。

[0027] 例えば、集約部12は、クエリ入力部11に入力されたクエリ中で S と R_i に対する結合処理、すなわちWhere句に $S.FK_i = R_i.RID_i$ が含まれる場合、各 R_i に対してSELECT句に含まれていた X_{R_i} を集約キーとして重複または、類似レコードの集約を行う。そして、集約部12は、集約したテーブル R'_i のレコード数 n'_{R_i} を算出して抽出部14に出力する。下記表2のSQL2に、Distinctを利用して R_i の重複レコードを除去した時の n'_{R_i} 算出クエリを示す。また、下記表3のSQL3に、Similarity Group by (下記の参考文献参考) を利用して R_i の類似レコードを除去した時の n'_{R_i} 算出クエリを示す。

参考文献：M. Tang, R. Y. Tahboub, W. G. Aref, M. J. Atallah, Q. M. Malluhi, M. Ouzzani, and Y. N. Silva. "Similarity group-by operators for multi-dimensional relational data." IEEE TKDE, pp.510-523, 2016.

[0028] [表2]

(表2)

SQL 2 重複レコードを除去した n_{R_i}' 算出クエリ例
SELECT COUNT(DISTINCT X_{R_i}) FROM R_i

[0029] [表3]

(表3)

SQL 3 類似レコードを除去した n_{R_i}' 算出クエリ例
SELECT COUNT(*) FROM(SELECT X_{R_i} FROM R_i GROUP BY X_{R_i} DISTANCE-TO-ALL L2 WITHIN ϵ ON-OVERLAP JOIN-ANY) AS DUMMY

[0030] 抽出部14は、集約部12によってカウントされたレコード数に対するファクトテーブルのレコード数の比率が所定の閾値を超える場合には、集約したテーブルの外部キーをディメンションテーブルの特徴群の代表値として含む分析用データセットを抽出する。また、抽出部14は、集約部12によってカウントされたレコード数に対するファクトテーブルのレコード数の比率が所定の閾値を超えない場合には、クエリ入力部11が受け付けたクエリをそのまま用いて、複数テーブルを結合し、分析用データセットを抽出する。

[0031] 例えば、抽出部14は、まず、データベース13からファクトテーブルのレコード数 n_s を取得する。データベース13は、ファクトテーブルのレコード数 n_s を記憶する。そして、抽出部14は、集約したテーブル R_i' のレコード数 n_{R_i}' とファクトテーブルのレコード数 n_s から $TR_i' = n_s / n_{R_i}'$ を算出し、算出した TR_i' が20を超えているか否かを判定する特徴変換判定を行う。

[0032] 抽出部14は、いずれの TR_i' も条件を満たさない場合は、クエリ入力部

11が受け付けた分析用データ取得のためのクエリをそのままデータベース15に入力し、分析用データを抽出する。データベース15は、ファクトテーブルおよびディメンションテーブルを記憶する。また、抽出部14は、条件を満たした TR'_i が一つでもある場合はクエリ入力部11で入力されたクエリ中に含まれる X_{R_i} を FK'_i に変換して抽出する特徴量を削減したクエリを出力する。

[0033] 下記の表4のSQL4に、重複レコードを除去し、1, ..., kの全 R_i について条件を満たした時にSQL1の各 X_{R_i} を FK'_i に変換したクエリ例を示す。また、表5のSQL5に、類似レコードを除去し、1, ..., kの全 R_i について条件を満たした時にSQL1の各 X_{R_i} を FK'_i に変換したクエリ例を示す。

[0034] [表4]

(表4)

SQL4 重複レコードを除去した R'_i に対して特徴変換した分析用データ抽出クエリ例
SELECT Y, X_S , FK'_1 , ..., FK'_k FROM $S, R'_1, \dots, R'_k$ WHERE $S.FK_1 = R'_1.RID_1, \dots, S.FK_k = R'_k.RID_k$, R'_i : SELECT RID_i , FIRST_VALUE(RID_i) OVER (PARTITION BY X_{R_i}) AS FK'_i FROM R_i

[0035] [表5]

(表5)

SQL5 類似レコードを除去した R'_i に対して特徴変換した分析用データ抽出クエリ例
SELECT Y, X_S , FK'_1 , ..., FK'_k FROM $S, R'_1, \dots, R'_k$, WHERE $S.FK_1 = R'_1.RID_1, \dots, S.FK_k = R'_k.RID_k$, R'_i : SELECT UNNEST(ARRAY(SELECT UNNEST(ARRAY_AGG(RID_i)))) as RID_i , min(RID_i) as FK'_i FROM R_i GROUP BY X_{R_i} DISTANCE-TO-ALL L2 WITHIN ϵ ON-OVERLAP JOIN-ANY

[0036] 特徴選択部16は、抽出部14によって抽出された分析用データに対して

所定の特徴選択手法を用いて、特徴群の部分集合を選択する。例えば、特徴選択部 16 は、分析用データを取得し、分析用データに対してラッパー法やフィルター法を用いて特徴選択を行い、特徴群の中から汎化誤差を最小にするような特徴群の部分集合を選択してモデル学習部 17 に出力する。

[0037] モデル学習部 17 は、特徴選択部 16 によって選択された特徴群の部分集合のデータを用いて、機械学習モデルを生成する。例えば、モデル学習部 17 は、最良の特徴群の部分集合のデータを用いて、ロジスティック回帰や SVM (Support Vector Machine) といった機械学習モデルを生成して出力する。

[0038] [分析用データの抽出処理例]

ここで、図 2 のデータセットに対して情報処理装置 10 を適用した時の分析用データを抽出するまでの一連の処理について説明する。図 2 は、映画の評価予測データセットの一例を示す図である。ここでは、データセットから映画の評価 (Rating) を R_1 と R_2 の特徴群から予測するモデルを作成する場合の例を説明する。

[0039] まず、クエリ入力部 11 は、分析用データ取得のために、図 2 のデータセットを結合して予測したい特徴量 Y と予測に用いる特徴群 X を取得するクエリ SQL 6 (下記の表 6 参照) を入力として受け付けそのまま集約部 12 および抽出部 14 に出力する。SQL 6 の各変数を SQL 1 の変数で言い換えるとそれぞれ $k=2$ 、 $Y = \{Rating\}$ 、 $X_S = \{\phi\}$ 、 $X_{R_1} = \{Year, Genre\}$ 、 $X_{R_2} = \{Sex, Age, Income\}$ 、 $RID_1 = \{MovieID\}$ 、 $RID_2 = \{UserID\}$ となる。

[0040] [表 6]

(表 6)

SQL 6 分析用データ取得クエリ
SELECT Rating, Year, Genre, Sex, Age, Income
FROM S, R_1 , R_2
WHERE S.MovieID = R_1 .MovieID, S.UserID = R_2 .UserID,

[0041] ここでは、集約部 12 が、 R_1 に対しては重複レコード、 R_2 に対しては類

似レコードの集約処理を行い重複、類似レコードの削減を行う場合の例について説明する。下記の表 7 における SQL 7 に `Distinct` を利用して R_1 の重複レコードを除去した時の n'_{R_1} 算出クエリを例示する。図 3 に例示するように、集約部 12 は、この n'_{R_1} 算出クエリを用いて、 $n'_{R_1} = 6$ を算出する。

[0042] [表7]

(表7)

SQL 7 重複レコードを除去した n'_{R_1} 算出クエリ
<code>SELECT COUNT(DISTINCT ROW(Year, Genre)) FROM R₁</code>

[0043] また、下記の表 8 における SQL 8 に `Similarity Group by` を利用して R_2 の類似レコードを除去した時の n'_{R_2} 算出クエリを例示する。図 4 に例示するように、集約部 12 は、この n'_{R_2} 算出クエリを用いて、 $n'_{R_2} = 4$ を算出する。図 3 および図 4 は、集約したテーブルのレコード数を算出する処理例を説明する図である。

[0044] [表8]

(表8)

SQL 8 類似レコードを除去した n'_{R_2} 算出クエリ
<code>SELECT COUNT(*) FROM(</code> <code>(SELECT Sex, Age, Income</code> <code>FROM RangeNormalizedR₂</code> <code>GROUP BY Sex, Age, Income DISTANCE-TO-ALL L2 WITHIN 0.3</code> <code>ON-OVERLAP JOIN-ANY) AS R'₂</code> <code>RangeNormalizedR₂:</code> <code>SELECT Sex,(Age-(SELECT MIN(Age) FROM R₂))/((SELECT MAX(age) FROM</code> <code>R₂)-(SELECT MIN(Age) FROM R₂)) AS Age,(INCOME-(SELECT MIN(INCOME)</code> <code>FROM R₂))/((SELECT MAX(INCOME) FROM R₂)-(SELECT MIN(INCOME)</code> <code>FROM R₂)) as Income FROM R₂</code>

[0045] そして、抽出部 14 は、例えば、データベース 13 から $n_s = 150$ を取得し、 $TR'_{1'}$ 、 $TR'_{2'}$ とともに特徴変換判定の条件を満たすか判定する。上記の例では、抽出部 14 は、 $TR'_{1'}$ を算出するために、「 $TR'_{1'} = n_s / n'_{R_1}$ 」の式に従い、 $150 / 6 = 25$ を算出する。そして、抽出部 14 は、 TR

$\hat{r}_1 = 25$ であるため、所定の閾値である 20 を超えており、特徴変換判定の条件を満たしていると判定する。また、抽出部 14 は、 $TR \hat{r}_2$ を算出するために、「 $TR \hat{r}_2 = n_s / n_{R1}$ 」の式に従い、 $150 / 4 = 37.5$ を算出する。そして、抽出部 14 は、 $TR \hat{r}_2 = 37.5$ であるため、所定の閾値である 20 を超えており、特徴変換判定の条件を満たしていると判定する。

[0046] このため、抽出部 14 は、 $X_{R1} = \{Year, Genre\}$ 、 $X_{R2} = \{Sex, Age, Income\}$ を $FK \hat{r}_1 = \{MovieID \hat{r}_1\}$ 、 $FK \hat{r}_1 = \{UserID \hat{r}_1\}$ にそれぞれ変換して、抽出する特徴量を削減したクエリをデータベース 15 に出力する。下記の表 9 に変換後のクエリ例を示す。

[0047] [表9]

(表9)

SQL 9 特徴変換した分析用データ抽出クエリ
<pre> SELECT Rating, MovieID', UserID' FROM S, (SELECT MovieID, FIRST_VALUE(MovieID) OVER (PARTITION BY Year, Genre) AS MovieID' FROM R₁) AS R₁', (SELECT UNNEST(ARRAY(SELECT UNNEST(ARRAY_AGG(UsrID)))) AS UserID, MIN(UserID) AS UserID' FROM RangeNormalizedR₂ GROUP BY Sex, Age, Income DISTANCE-TO-ALL L2 WITHIN 0.3 ON-OVERLAP JOIN-ANY) AS R₂' WHERE S.MovieID = R₁'.MovieID, S.UserID = R₂'.UserID RangeNormalizedR₂:SQL8 参照 </pre>

[0048] そして、抽出部 14 は、データベース 15 から図 5 のように $\{Rating, UserID, MovieID\}$ の特徴群を持つ分析用データセットを抽出して特徴選択部 16 に出力する。これは従来取得するはずであった特徴群である $\{Rating, Year, Genre, Sex, Age, Income\}$ より特徴量を削減できているため、後段の特徴選択処理の処理時間を大きく削減することができる。図 5 は、分析用データセットを抽出する処理例を説明する図である。

[0049] [情報処理装置の処理手順]

次に、図 6 を用いて、第 1 の実施形態に係る情報処理装置 10 による処理

手順の例を説明する。図6は、第1の実施形態に係る情報処理装置における処理の流れの一例を示すフローチャートである。

- [0050] まず、図6を用いて、クエリ入力部11が分析用データ取得のためのクエリの入力を受け付けると（ステップS101肯定）、集約部12は、ディメンションテーブルにおける各レコードについて、主キー以外の特徴量が重複または類似しているレコードを集約したテーブルのレコード数をカウントする（ステップS102）。例えば、集約部12は、クエリ入力部11に入力されたクエリ中でSと R_i に対する結合処理、すなわちWhere句に $S.FK_i = R_i.RID_i$ が含まれる場合、各 R_i に対してSELECT句に含まれていた X_{R_i} を集約キーとして重複または、類似レコードの集約を行い、集約したテーブル R'_i のレコード数 n'_{R_i} をカウントする。
- [0051] そして、抽出部14は、集約部12によって集約されたテーブルのレコード数に対するファクトテーブルのレコード数の比率を算出し（ステップS103）、比率が20を超えているかを判定する（ステップS104）。つまり、抽出部14は、集約したテーブル R'_i のレコード数 n'_{R_i} とファクトテーブルのレコード数 n_s から $TR'_i = n_s / n'_{R_i}$ を算出し、算出した TR'_i が20を超えているか否かを判定する特徴変換判定を行う。
- [0052] この結果、抽出部14は、比率が20を超えていると判定した場合には（ステップS104肯定）、集約したディメンションテーブルの外部キーをディメンションテーブルの特徴群の代表値として含む分析用データを抽出する（ステップS105）。また、抽出部14は、比率が20を超えていないと判定した場合には（ステップS104否定）、複数テーブルを結合し、分析用データを抽出する（ステップS106）。つまり、抽出部14は、クエリ入力部11が受け付けた分析用データ取得のためのクエリをそのままデータベース15に入力し、分析用データを抽出する。
- [0053] そして、特徴選択部16は、分析用データに対して特徴選択を行い、特徴群の部分集合を選択する（ステップS107）。例えば、特徴選択部16は、分析用データを取得し、分析用データに対してラッパー法やフィルター法

を用いて特徴選択を行い、特徴群の中から汎化誤差を最小にするような特徴群の部分集合を選択する。

[0054] その後、モデル学習部 17 は、特徴群の部分集合のデータを用いて、機械学習モデルを生成する（ステップ S 108）。例えば、モデル学習部 17 は、最良の特徴群の部分集合のデータを用いて、ロジスティック回帰や SVM (Support Vector Machine) といった機械学習モデルを生成する。

[0055] [第 1 の実施形態の効果]

このように、第 1 の実施形態に係る情報処理装置 10 は、ディメンションテーブルにおける各レコードについて、主キー以外の特徴量が重複または類似しているレコードを集約したテーブルのレコード数をカウントする集約部 12 と、集約部 12 によってカウントされたレコード数に対するファクトテーブルのレコード数の比率が所定の閾値を超える場合には、集約したテーブルの外部キーをディメンションテーブルの特徴群の代表値として含む分析用データを抽出する抽出部 14 とを有する。このため、第 1 の実施形態に係る情報処理装置 10 では、より多くのケースに対して特徴選択にかかる時間を削減することが可能である。

[0056] つまり、情報処理装置 10 は、集約したテーブル R'_i のレコード数 n'_{R_i} をカウントしている。 $n'_{R_i} \leq n_{R_i}$ である。このため、情報処理装置 10 では、TR を用いた手法の適用範囲が増え、より多くのケースに対して精度を担保したまま特徴選択にかかる時間を削減することができる。また、第 1 の実施形態に係る情報処理装置 10 では、特徴量を削減できるため、特徴選択処理の処理時間を削減することが可能である。

[0057] (システム構成等)

また、図示した各装置の各構成要素は機能概念的なものであり、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、各装置の分散・統合の具体的形態は図示のものに限られず、その全部または一部を、各種の負荷や使用状況などに応じて、任意の単位で機能的または物理的に分散・統合して構成することができる。さらに、各装置にて行なわれる各処理機

能は、その全部または任意の一部が、CPUおよび当該CPUにて解析実行されるプログラムにて実現され、あるいは、ワイヤードロジックによるハードウェアとして実現され得る。

[0058] また、本実施の形態において説明した各処理のうち、自動的におこなわれるものとして説明した処理の全部または一部を手動的におこなうこともでき、あるいは、手動的におこなわれるものとして説明した処理の全部または一部を公知の方法で自動的におこなうこともできる。この他、上記文書中や図面中で示した処理手順、制御手順、具体的名称、各種のデータやパラメータを含む情報については、特記する場合を除いて任意に変更することができる。

[0059] (プログラム)

また、上記実施形態において説明した情報処理装置が実行する処理をコンピュータが実行可能な言語で記述したプログラムを作成することもできる。例えば、実施形態に係る情報処理装置10が実行する処理をコンピュータが実行可能な言語で記述した抽出プログラムを作成することもできる。この場合、コンピュータが抽出プログラムを実行することにより、上記実施形態と同様の効果を得ることができる。さらに、かかる抽出プログラムをコンピュータ読み取り可能な記録媒体に記録して、この記録媒体に記録された抽出プログラムをコンピュータに読み込ませて実行することにより上記実施形態と同様の処理を実現してもよい。

[0060] 図7は、抽出プログラムを実行するコンピュータを示す図である。図7に例示するように、コンピュータ1000は、例えば、メモリ1010と、CPU1020と、ハードディスクドライブインタフェース1030と、ディスクドライブインタフェース1040と、シリアルポートインタフェース1050と、ビデオアダプタ1060と、ネットワークインタフェース1070とを有し、これらの各部はバス1080によって接続される。

[0061] メモリ1010は、図7に例示するように、ROM (Read Only Memory) 1011及びRAM1012を含む。ROM1011は、例えば、BIOS

S (Basic Input Output System) 等のブートプログラムを記憶する。ハードディスクドライバインタフェース 1030 は、図 7 に例示するように、ハードディスクドライブ 1090 に接続される。ディスクドライバインタフェース 1040 は、図 7 に例示するように、ディスクドライブ 1100 に接続される。例えば磁気ディスクや光ディスク等の着脱可能な記憶媒体が、ディスクドライブ 1100 に挿入される。シリアルポートインタフェース 1050 は、図 7 に例示するように、例えばマウス 1110、キーボード 1120 に接続される。ビデオアダプタ 1060 は、図 7 に例示するように、例えばディスプレイ 1130 に接続される。

[0062] ここで、図 7 に例示するように、ハードディスクドライブ 1090 は、例えば、OS 1091、アプリケーションプログラム 1092、プログラムモジュール 1093、プログラムデータ 1094 を記憶する。すなわち、上記の、抽出プログラムは、コンピュータ 1000 によって実行される指令が記述されたプログラムモジュールとして、例えばハードディスクドライブ 1090 に記憶される。

[0063] また、上記実施形態で説明した各種データは、プログラムデータとして、例えばメモリ 1010 やハードディスクドライブ 1090 に記憶される。そして、CPU 1020 が、メモリ 1010 やハードディスクドライブ 1090 に記憶されたプログラムモジュール 1093 やプログラムデータ 1094 を必要に応じて RAM 1012 に読み出し、各種処理手順を実行する。

[0064] なお、抽出プログラムに係るプログラムモジュール 1093 やプログラムデータ 1094 は、ハードディスクドライブ 1090 に記憶される場合に限られず、例えば着脱可能な記憶媒体に記憶され、ディスクドライブ等を介して CPU 1020 によって読み出されてもよい。あるいは、抽出プログラムに係るプログラムモジュール 1093 やプログラムデータ 1094 は、ネットワーク (LAN (Local Area Network)、WAN (Wide Area Network) 等) を介して接続された他のコンピュータに記憶され、ネットワークインタフェース 1070 を介して CPU 1020 によって読み出されてもよい。

符号の説明

- [0065] 1 0 情報処理装置
- 1 1 クエリ入力部
- 1 2 集約部
- 1 3、1 5 データベース
- 1 4 抽出部
- 1 6 特徴選択部
- 1 7 モデル学習部

請求の範囲

- [請求項1] ディメンションテーブルにおける各レコードについて、主キー以外の特徴量が重複または類似しているレコードを集約したテーブルのレコード数をカウントする集約部と、
- 前記集約部によってカウントされたレコード数に対するファクトテーブルのレコード数の比率が所定の閾値を超える場合には、前記集約したテーブルの外部キーを前記ディメンションテーブルの特徴群の代表値として含む分析用データを抽出する抽出部と
- を有することを特徴とする情報処理装置。
- [請求項2] 前記集約部は、前記分析用データを取得するためのクエリの入力が受け付けられた場合に、主キー以外の特徴量が同一のレコードまたは類似のレコードを除去したテーブルのレコード数をカウントし、
- 前記抽出部は、前記集約部によってカウントされたレコード数に対するファクトテーブルのレコード数の比率が所定の閾値を超えない場合には、前記クエリを用いて、前記分析用データを抽出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] 前記抽出部によって抽出された分析用データに対して所定の特徴選択手法を用いて、特徴群の部分集合を選択する特徴選択部と、
- 前記特徴選択部によって選択された特徴群の部分集合のデータを用いて、学習モデルを生成するモデル学習部と
- をさらに有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4] 情報処理装置によって実行される抽出方法であって、
- ディメンションテーブルにおける各レコードについて、主キー以外の特徴量が重複または類似しているレコードを集約したテーブルのレコード数をカウントする集約工程と、
- 前記集約工程によってカウントされたレコード数に対するファクトテーブルのレコード数の比率が所定の閾値を超える場合には、前記集約したテーブルの外部キーを前記ディメンションテーブルの特徴群の

代表値として含む分析用データを抽出する抽出工程と
を含むことを特徴とする抽出方法。

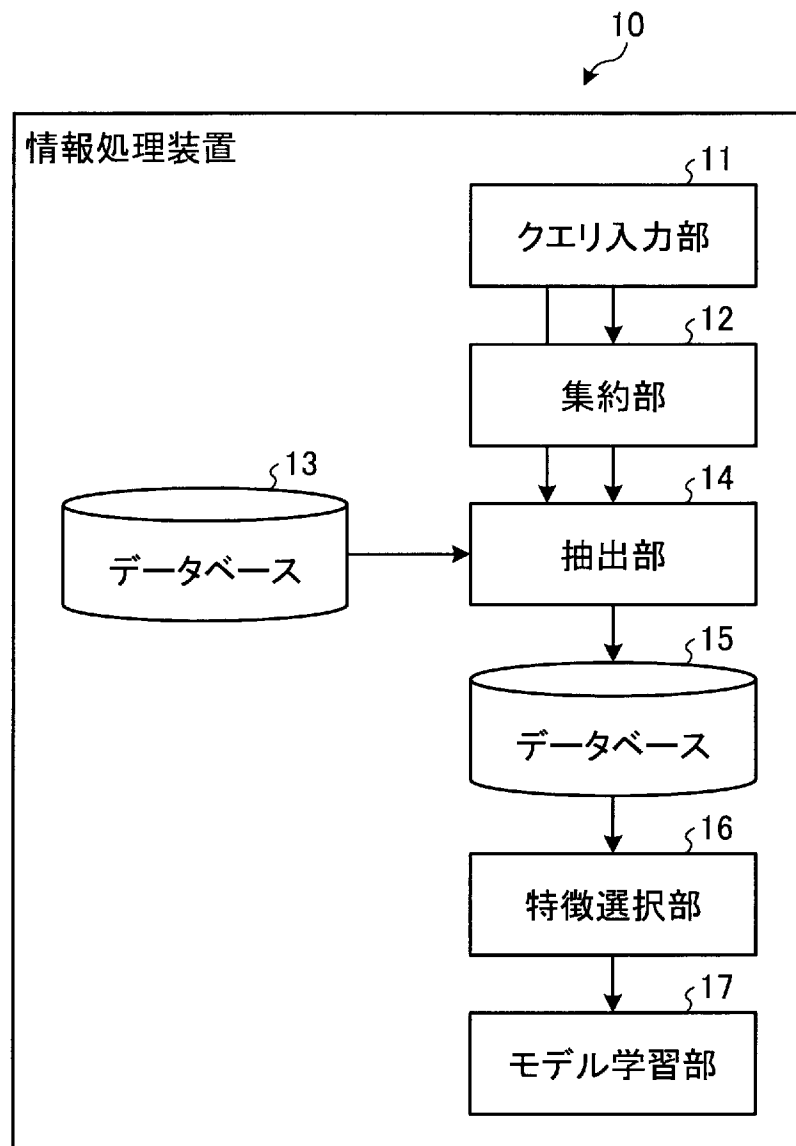
[請求項5]

ディメンションテーブルにおける各レコードについて、主キー以外の特徴量が重複または類似しているレコードを集約したテーブルのレコード数をカウントする集約ステップと、

前記集約ステップによってカウントされたレコード数に対するファクトテーブルのレコード数の比率が所定の閾値を超える場合には、前記集約したテーブルの外部キーを前記ディメンションテーブルの特徴群の代表値として含む分析用データを抽出する抽出ステップと

をコンピュータに実行させることを特徴とする抽出プログラム。

[図1]



[図2]

評価テーブルS

ID	Rating	MovieID	UserID
1	5	1	1
2	4	1	2
3	5	1	3
4	3	1	6
5	4	1	7
...	...	...	...
146	1	9	9
147	2	10	1
148	3	10	4
149	5	10	5
150	5	10	10

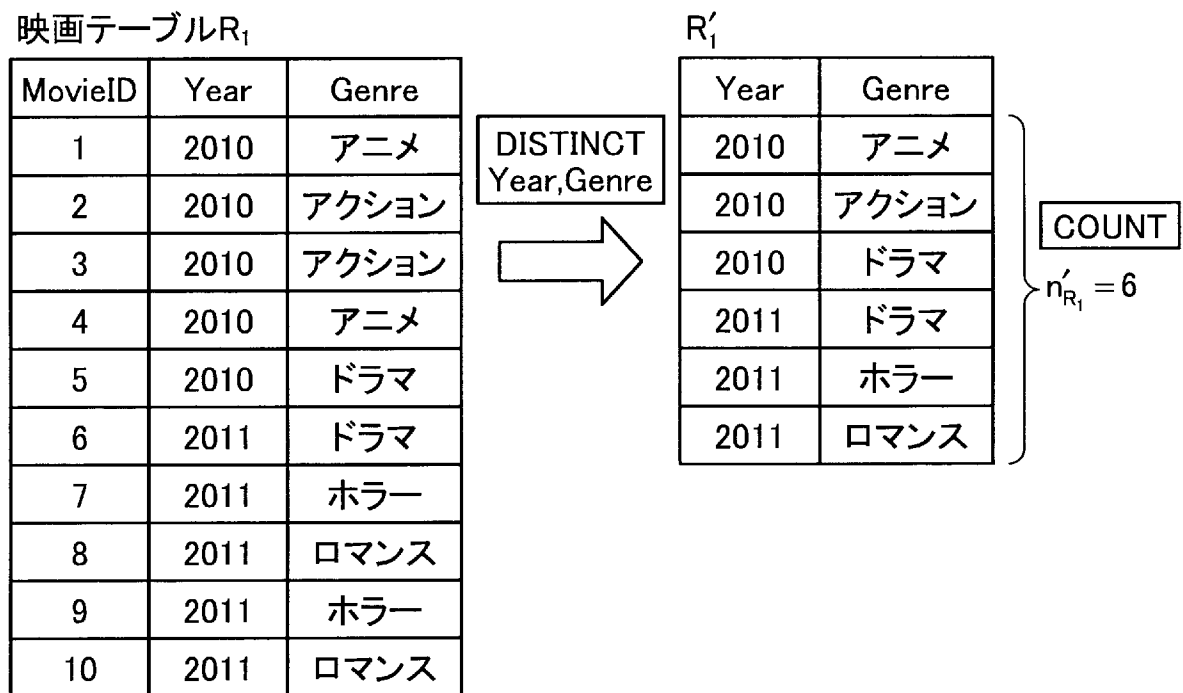
映画データベースR₁

MovieID	Year	Genre
1	2010	アニメ
2	2010	アクション
3	2010	アクション
4	2010	アニメ
5	2010	ドラマ
6	2011	ドラマ
7	2011	ホラー
8	2011	ロマンス
9	2011	ホラー
10	2011	ロマンス

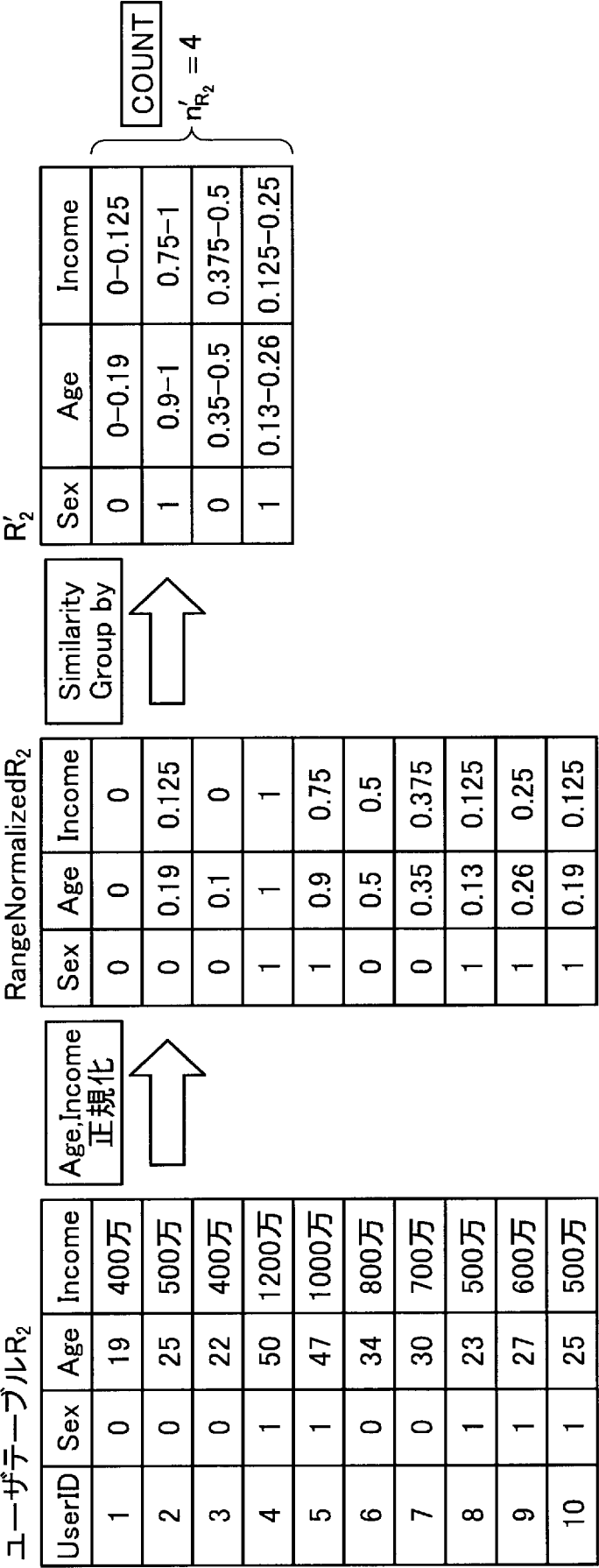
ユーザテーブルR₂

UserID	Sex	Age	Income
1	0	19	400万
2	0	25	500万
3	0	22	400万
4	1	50	1200万
5	1	47	1000万
6	0	34	800万
7	0	30	700万
8	1	23	500万
9	1	27	600万
10	1	25	500万

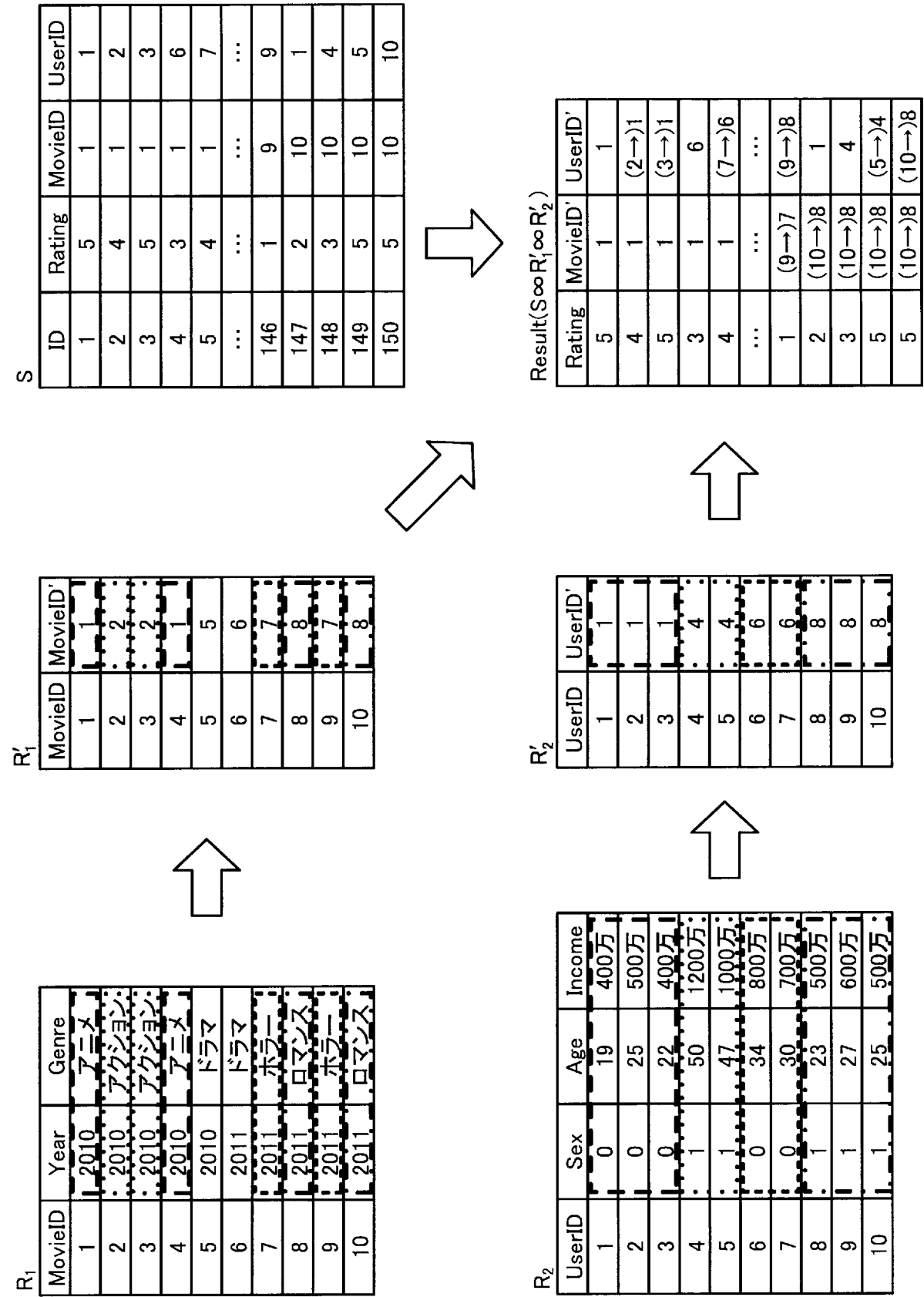
[図3]



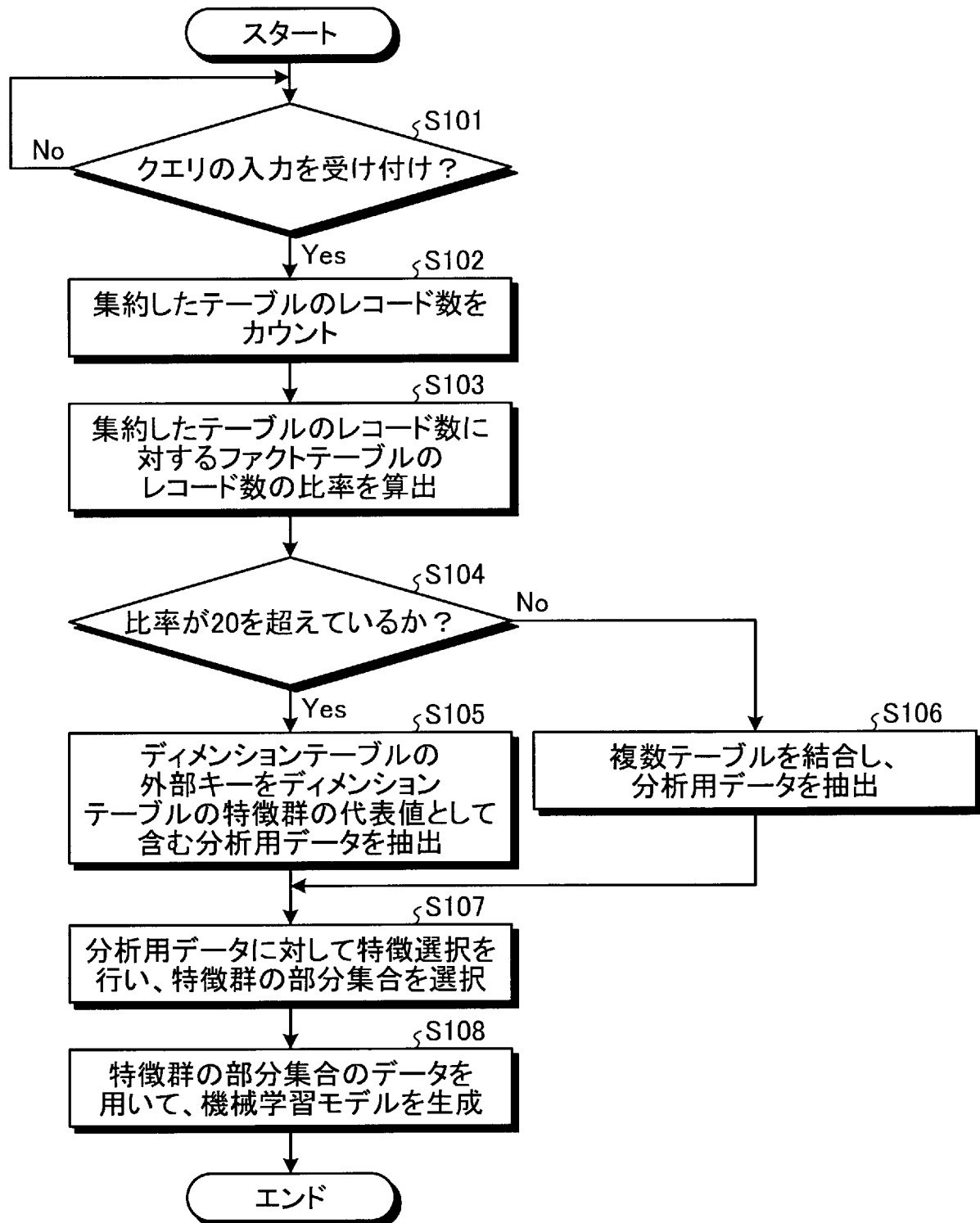
[図4]



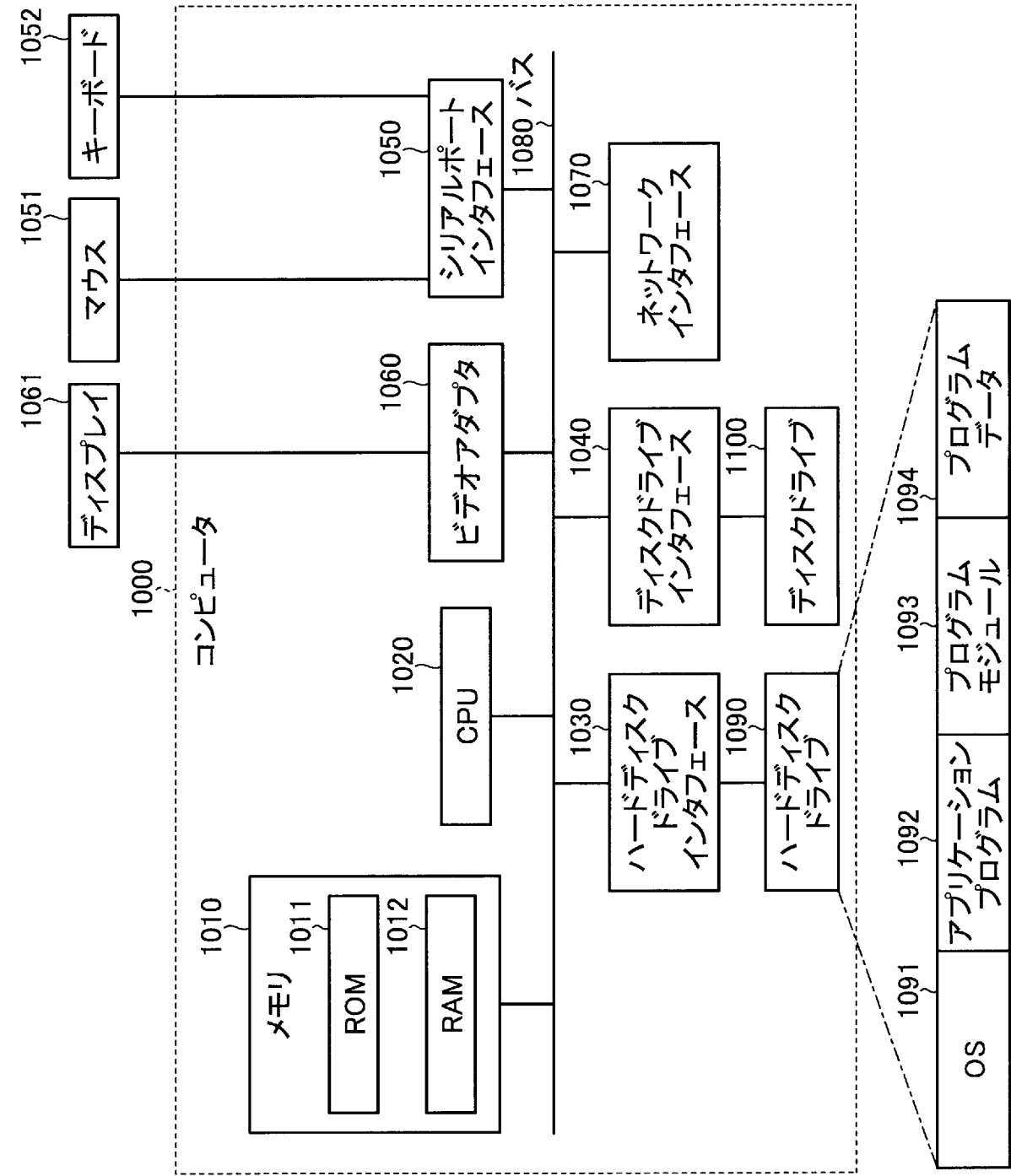
[図5]



[図6]



[図7]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/019191

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06N20/00 (2019.01) i, G06F16/215 (2019.01) 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. G06N3/00-99/00, G06F16/00-16/958

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.
Y	志村薫, 外 2 名, RDB の構造を考慮したデータベースからの学習手法について, 第 80 回 (平成 30 年) 全国大会講演論文集 (1), 13 March 2018, pp. 1-457, 1-458, particularly, p. 1-457, left column, line 1 to right column, line 20, fig. 1, non-official translation (SHIMURA, Minoru and 2 others, A learning method from database considering the structure of RDB, Proceedings of the 80th (2018) National Convention of IPSJ (1))	1-5
Y	JP 2018-106563 A (PANASONIC INTELLECTUAL PROPERTY MANAGEMENT CO., LTD.) 05 July 2018, paragraph [0083], fig. 5-7 (Family: none)	1-5



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
02.08.2019

Date of mailing of the international search report
13.08.2019

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/019191

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2016/0308898 A1 (TEEPLE, D. J. W.) 20 October 2016, paragraph [0058] (Family: none)	1-5
Y	RASCHKA, S. 著, 株式会社クイープ訳, Python 機械学習プログラミング達人データサイエンティストによる理論と実践, ISBN 978-4-8443-8060-3, 第1版第6刷, 東京: 株式会社インプレス, 11 April 2017, pp. 107-119, (written by RASCHKA, S., translated by QUIPU, Tokyo: IMPRESS CORPORATION), non-official translation (Theory and practice by python machine learning programming expert data scientist, first edition, 6th printing)	3
Y	BALDI, P. et al 著, 水田正弘ほか訳, 確率モデルによる Web データ解析法データマイニング技法から e-コマースまで, ISBN 978-4-627-09531-1, 第1版第1刷, 東京: 森北出版株式会社, 31 May 2007, pp. 112-114, (written by BALDI, P. et al., translated by MIZUTA, Masahiro et al., Tokyo: MORIKITA PUBLISHING CO., LTD.), non-official translation (A web data analysis using stochastic models from data mining techniques to e-commerce, first edition, first printing)	3

A. 発明の属する分野の分類 (国際特許分類 (I P C))

Int.Cl. G06N20/00(2019.01) i, G06F16/215(2019.01) 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 (I P C))

Int.Cl. G06N3/00-99/00
G06F16/00-16/958

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	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 1 9 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 9 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 9 年

国際調査で利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	志村薫、外 2 名, RDB の構造を考慮したデータベースからの学習手法について、第 80 回 (平成 30 年) 全国大会講演論文集 (1), 2018. 03. 13, pp. 1-457-1-458, 特に 1-457 ページ左欄第 1 行~右欄第 20 行, 図 1	1-5
Y	JP 2018-106563 A (パナソニック IP マネジメント株式会社) 2018. 07. 05, 段落[0083], 第 5-7 図 (ファミリーなし)	1-5
Y	US 2016/0308898 A1 (TEEPLE, David James Wayne) 2016. 10. 20, 段落[0058] (ファミリーなし)	1-5

☒ 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

0 2 . 0 8 . 2 0 1 9

国際調査報告の発送日

1 3 . 0 8 . 2 0 1 9

国際調査機関の名称及びあて先

日本国特許庁 (I S A / J P)
郵便番号 1 0 0 - 8 9 1 5
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)

北元 健太

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 4 5

5 B

3 8 5 6

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	RASCHKA, Sebastian 著, 株式会社クイープ訳, Python 機械学習プログラミング 達人データサイエンティストによる理論と実践, ISBN 978-4-8443-8060-3, 第1版第6刷, 東京: 株式会社インプレス, 2017.04.11, pp.107-119	3
Y	BALDI, Pierre et al 著, 水田正弘ほか訳, 確率モデルによる Web データ解析法 データマイニング技法から e-コマースまで, ISBN 978-4-627-09531-1, 第1版第1刷, 東京: 森北出版株式会社, 2007.05.31, pp.112-114	3