

(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
2022年12月22日(22.12.2022)



(10) 国際公開番号

WO 2022/264385 A1

- (51) 国際特許分類:
G06F 16/31 (2019.01) *G06F 16/35* (2019.01)
- (21) 国際出願番号: PCT/JP2021/023115
- (22) 国際出願日: 2021年6月17日(17.06.2021)
- (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: 片岡 正弘 (KATAOKA, Masahiro);
〒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). 尾上 聡 (ONOE, Satoshi); 〒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). 松村 量(MATSUMURA, Ryo); 〒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).

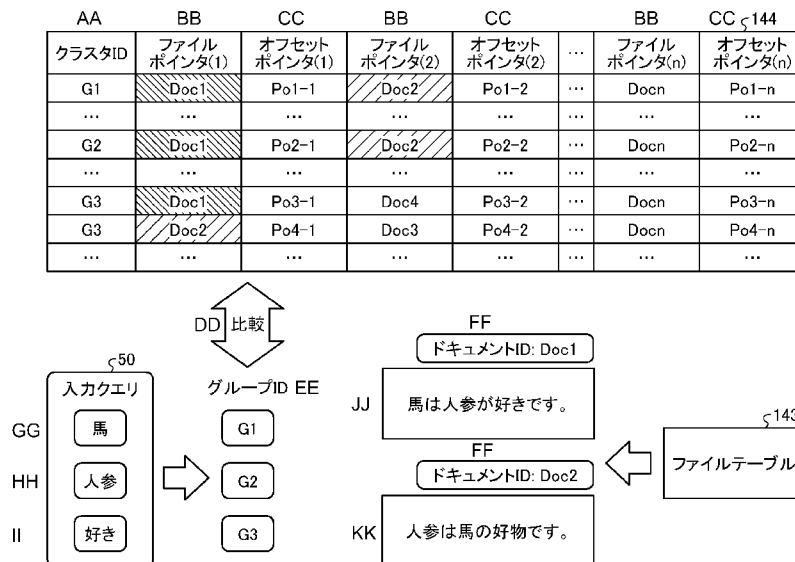
(74) 代理人: 弁理士法人 酒井国際特許事務所 (SAKAI INTERNATIONAL PATENT OFFICE); 〒1000013 東京都千代田区霞が関3丁目8番1号 虎の門三井ビルディング Tokyo (JP).

(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, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,

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

(54) 発明の名称: 検索方法、検索プログラムおよび情報処理装置

[図3]



(57) Abstract: This information processing device identifies, upon reception of character string data as a search condition, a cluster corresponding to the received character string data, such cluster being from among a plurality of clusters included in a plurality of pieces of character string data. The information processing device: consults a storage unit in which information pieces, respectively indicating a plurality of pieces of character string data included in a database to be searched, are stored in association with any of the plurality of clusters; and thereby identifies an information piece indicating character string data and associated with the identified cluster. The information processing device generates a search result according to the identified information indicating the character string data.

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,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, WS, 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))

(57) 要約: 情報処理装置は、検索条件として文字列データを受け付けると、複数の文字列データに含まれる複数のクラスタのうち、受け付けた文字列データに応じたクラスタを特定する。情報処理装置は、検索対象のデータベースに含まれる複数の文字列データそれぞれを示す情報を複数のクラスタのいずれかに対応付けて記憶する記憶部を参照して、特定したクラスタに対応付けられた文字列データを示す情報を特定する。情報処理装置は、特定した文字列データを示す情報に応じた検索結果を生成する。

明 細 書

発明の名称： 検索方法、検索プログラムおよび情報処理装置

技術分野

[0001] 本発明は、検索方法等に関する。

背景技術

[0002] DB (Data Base) には膨大なテキストデータが登録されており、かかるDBから、利用者に指定された入力クエリに類似するデータを適切に検索することが求められている。

[0003] 従来技術では、単語と単語のベクトルとを対応付けた単語ベクトル辞書を用いて、データ検索を行っている。たとえば、従来技術では、入力クエリに指定された単語と、単語ベクトルテーブルとを比較して、単語ベクトルを特定し、テキストデータに含まれる各単語のベクトルと比較を行って、入力クエリに対応する情報を検索する。

先行技術文献

特許文献

[0004] 特許文献1：特開2019-204362号公報

特許文献2：特開2018-180789号公報

特許文献3：特開2020-086999号公報

特許文献4：国際公開第2020/213158号

特許文献5：特開2009-217436号公報

発明の概要

発明が解決しようとする課題

[0005] 上述した従来技術では、入力クエリに指定された単語の単語ベクトルと、テキストデータに含まれる各単語のベクトルとを基にして、入力クエリに対応する情報を検索しているが、単語ベクトルが多次元の浮動小数点数値で表現されているため、転置インデックスのサイズが膨大となり、検索時間に時間を要してしまう。

[0006] 1つの側面では、本発明は、転置インデックスのサイズを抑制し、検索時間を短縮することができる検索方法、検索プログラムおよび情報処理装置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

[0007] 第1の案では、コンピュータに次の処理を実行させる。コンピュータは、検索条件として文字列データを受け付けると、複数の文字列データに含まれる複数のクラスタのうち、受け付けた文字列データに応じたクラスタを特定する。コンピュータは、検索対象のデータベースに含まれる複数の文字列データそれぞれを示す情報を複数のクラスタのいずれかに対応付けて記憶する記憶部を参照して、特定したクラスタに対応付けられた文字列データを示す情報を特定する。コンピュータは、特定した文字列データを示す情報に応じた検索結果を生成する。

発明の効果

[0008] 転置インデックスのサイズを抑制し、検索時間を短縮することができる。

図面の簡単な説明

[0009] [図1]図1は、本実施例1に係る情報処理装置の準備フェーズの処理を説明するための図(1)である。

[図2]図2は、本実施例1に係る情報処理装置の準備フェーズの処理を説明するための図(2)である。

[図3]図3は、本実施例1に係る情報処理装置の検索フェーズの処理を説明するための図である。

[図4]図4は、本実施例1に係る情報処理装置の構成を示す機能ブロック図である。

[図5]図5は、ファイルテーブルのデータ構造の一例を示す図である。

[図6]図6は、本実施例1に係る情報処理装置の準備フェーズの処理を示すフローチャートである。

[図7]図7は、本実施例1に係る情報処理装置の検索フェーズの処理を示すフローチャートである。

[図8]図8は、本実施例2に係る情報処理装置の準備フェーズの処理を説明するための図である。

[図9]図9は、本実施例2に係る情報処理装置の検索フェーズの処理を説明するための図(1)である。

[図10]図10は、本実施例2に係る情報処理装置の検索フェーズの処理を説明するための図(2)である。

[図11]図11は、本実施例2に係る情報処理装置の構成を示す機能ブロック図である。

[図12]図12は、本実施例2に係る情報処理装置の準備フェーズの処理を示すフローチャートである。

[図13]図13は、本実施例2に係る情報処理装置の検索フェーズの処理を示すフローチャートである。

[図14]図14は、実施例の情報処理装置と同様の機能を実現するコンピュータのハードウェア構成の一例を示す図である。

発明を実施するための形態

[0010] 以下に、本願の開示する検索方法、検索プログラムおよび情報処理装置の実施例を図面に基づいて詳細に説明する。なお、この実施例によりこの発明が限定されるものではない。

実施例 1

[0011] 本実施例1に係る情報処理装置は、準備フェーズの処理と、検索フェーズの処理を順に実行する。以下において、情報処理装置が実行する準備フェーズの処理と、検索フェーズの処理を順に説明する。

[0012] 図1は、本実施例1に係る情報処理装置の準備フェーズの処理を説明するための図(1)である。情報処理装置は、文字列と、文字列に対応するベクトルとを対応付けて保持するベクトル辞書141を有する。本実施例1では、文字列の一例として単語を用いて説明するが、これに限定されるものではなく、文字列は、文、項、文章であってもよい。以下の説明では、適宜、文字列を「単語」と表記する。

- [0013] 図1に示すベクトル辞書141の例では、単語「馬」にはベクトル「 WV_{α} 」が設定される。単語「人参」にはベクトル「 WV_{β} 」が設定される。単語「好き」にはベクトル「 WV_{γ} 」が設定される。単語「好物」にはベクトル「 WV_{η} 」が設定される。
- [0014] 情報処理装置は、ベクトル辞書141の各単語に設定されたベクトルに対して、クラスタリングを実行することで、類似する複数のベクトルを、同一のクラスタに分類する。情報処理装置は、クラスタリングとして、たとえば、k-means法を用いてもよい。同一のクラスタに分類されたベクトルに対応する単語には、同一のクラスタIDを割り当てる。情報処理装置は、係る処理を実行して、クラスタテーブル142を生成する。
- [0015] たとえば、クラスタリング結果5では、単語「馬」のベクトル「 WV_{α} 」は、クラスタID「G1」のクラスタに分類されている。単語「人参」のベクトル「 WV_{β} 」は、クラスタID「G2」のクラスタに分類されている。単語「好き」のベクトル「 WV_{γ} 」は、クラスタID「G3」のクラスタに分類されている。単語「好物」のベクトル「 WV_{η} 」は、クラスタID「G3」のクラスタに分類されている。
- [0016] クラスタテーブル142は、単語（文字列）と、クラスタIDと、ベクトルとを対応付ける。たとえば、クラスタリング結果5によって、単語「馬」には、クラスタID「G1」が対応付けられる。単語「人参」には、クラスタID「G2」が対応付けられる。単語「好き」および「好物」には、クラスタID「G3」が対応付けられる。
- [0017] 図2は、本実施例1に係る情報処理装置の準備フェーズの処理を説明するための図（2）である。情報処理装置は、クラスタテーブル142に含まれる各単語と、ファイルテーブル143とを比較して、転置インデックス144を生成する。
- [0018] ファイルテーブル143には、テキストデータを構成する複数のドキュメントファイルが含まれる。本実施例1では一例として、ドキュメントファイルには、文章の粒度の文字列が設定される。文章には、複数の項が含まれ、

各項には、複数の文が含まれる。各文には、複数の単語が含まれる。各ドキュメントファイルには、ドキュメントファイルを一意に識別するドキュメントIDが設定されている。

[0019] 情報処理装置は、クラスタテーブル142の単語と、ファイルテーブル143の各ドキュメントファイルの文字列とを比較して、単語がヒットする場合には、該当する単語のクラスタIDと、ファイルポインタと、オフセットポインタとを対応付けて、転置インデックス144に登録する。

[0020] ファイルポインタには、該当する単語を含むドキュメントファイルのドキュメントIDが設定される。オフセットポインタには、該当する単語を含むドキュメントファイルであって、かかるドキュメントファイル上における単語の位置情報（オフセット）が設定される。オフセットは、ファイルの先頭の単語から、該当する単語までを示す単語数である。なお、ファイルの先頭の文字から、該当する単語の先頭の文字を示す文字数であっても良い。

[0021] たとえば、クラスタID「G1」に属する単語「馬」が、ドキュメントID「Doc1」のドキュメントファイルに存在し、先頭の単語から単語「馬」までのオフセットを「Pos1」とする。この場合には、情報処理装置は、クラスタID「G1」、ファイルポインタ「Doc1」、オフセットポインタ「Pos1」を、転置インデックス144に登録する。なお、単語「馬」やクラスタID「G1」に属する単語が、複数のドキュメントファイルに存在する場合には、1個のクラスタID「G1」に、複数のファイルポインタとオフセットポインタとの組が、転置インデックス144に登録される。

[0022] 情報処理装置は、クラスタテーブル142に設定された各単語について上記処理を繰り返し実行することで、転置インデックス144を生成する。

[0023] 図3は、本実施例1に係る情報処理装置の検索フェーズの処理を説明するための図である。情報処理装置は、入力クエリ50を受け付ける。入力クエリ50には、複数の単語が含まれる。図3に示す例では、単語「馬」、「人参」、「好き」が含まれる場合を示す。情報処理装置は入力クエリ50の各

単語と、図 1、2 で示したクラスタテーブル 142 とを比較して、入力クエリ 50 の各単語に対応するクラスタ ID をそれぞれ特定する。たとえば、単語「馬」に対応するクラスタ ID は「G 1」、単語「人参」に対応するクラスタ ID は「G 2」、単語「好き」に対応するクラスタ ID は「G 3」となる。以下の説明では、入力クエリ 50 の各単語から特定したクラスタ ID を、「特定クラスタ ID」と表記する。

[0024] 情報処理装置は、特定クラスタ ID に対応する転置インデックス 144 のレコードのファイルポインタを基にして、全特定クラスタ ID に対応するドキュメント ID を特定する。ここでは、特定クラスタ ID を「G 1」、「G 2」、「G 3」とする。そうすると、ドキュメント ID「Doc 1」が、特定クラスタ ID を「G 1」、「G 2」、「G 3」に対応する。また、ドキュメント ID「Doc 2」が、特定クラスタ ID を「G 1」、「G 2」、「G 3」に対応する。

[0025] この場合、情報処理装置は、ドキュメント ID「Doc 1」のドキュメントファイルと、ドキュメント ID「Doc 2」のドキュメントファイルとを、ファイルテーブル 143 から抽出し、入力クエリ 50 の検索結果として、出力する。

[0026] 上記のように、本実施例 1 に係る情報処理装置は、準備フェーズにおいて、転置インデックス 144 を生成しておく。情報処理装置は、入力クエリ 50 を受け付けると、入力クエリ 50 に設定された単語に対応するクラスタ ID を特定する。情報処理装置は、転置インデックス 144 を参照し、特定クラスタ ID に対応付けられたドキュメント ID を基にして、検索結果を生成する。このように、入力クエリ 50 の各単語ベクトルに対応する転置インデックスに関して、単語のドキュメント ID とオフセットを単語ベクトルのクラスタ ID に対応付けることで、転置インデックスのサイズを抑制し、検索時間を短縮することができる。

[0027] 次に、本実施例 1 に係る情報処理装置の構成の一例について説明する。図 4 は、本実施例 1 に係る情報処理装置の構成を示す機能ブロック図である。

図4に示すように、この情報処理装置100は、通信部110、入力部120、表示部130、記憶部140、制御部150を有する。

[0028] 通信部110は、有線又は無線で外部装置等に接続され、外部装置等との間で情報の送受信を行う。たとえば、通信部110は、NIC (Network Interface Card) 等によって実現される。通信部110は、図示しないネットワークに接続されていてもよい。

[0029] 入力部120は、各種の情報を、情報処理装置100に入力する入力装置である。入力部120は、キーボードやマウス、タッチパネル等に対応する。たとえば、ユーザは、入力部120を操作して、入力クエリ等を入力してもよい。

[0030] 表示部130は、制御部150から出力される情報を表示する表示装置である。表示部130は、液晶ディスプレイ、有機EL (Electro Luminescence) ディスプレイ、タッチパネル等に対応する。たとえば、入力クエリに対応する検索結果が、表示部130に表示される。

[0031] 記憶部140は、ベクトル辞書141、クラスタテーブル142、ファイルテーブル143、転置インデックス144、入力クエリ50を有する。記憶部140は、たとえば、RAM (Random Access Memory)、フラッシュメモリ (Flash Memory) 等の半導体メモリ素子、または、ハードディスク、光ディスク等の記憶装置によって実現される。

[0032] ベクトル辞書141は、単語（文字列）とベクトルとを対応付けて保持する。ベクトル辞書141のデータ構造は、図1で示したベクトル辞書141に対応する。

[0033] 単語のベクトルは、事前に設定されているベクトルを用いてもよいし、ポアンカレエンベディング (Poincare Embeddings) と呼ばれる技術を用いて、ベクトルを割り当ててもよい。ポアンカレエンベディングは、たとえば、非特許文献「Valentin Khrulkov1 et al. 「Hyperbolic Image Embeddings」 Cornell University, 2019 April 3」等に記載された技術を用いればよい。

- [0034] ポアンカレエンベディングでは、ポアンカレ空間と呼ばれる空間に単語を埋め込み、埋め込んだ位置に応じたベクトルを割り当てる。また、類似する単語ほど、ポアンカレ空間において、近い位置に埋め込まれるという特徴がある。
- [0035] クラスタテーブル 142 は、単語の属するクラスタを特定する場合に用いられるテーブルである。クラスタテーブル 142 のデータ構造は、図 1 で示したクラスタテーブル 142 に対応する。図 1 で説明したように、クラスタテーブル 142 は、単語（文字列）と、クラスタ ID と、ベクトルとを対応付けるテーブルである。クラスタ ID は、単語が属するクラスタを一意に識別する情報である。
- [0036] ファイルテーブル 143 は、テキストデータを構成する複数のドキュメントファイルを有する。図 5 は、ファイルテーブルのデータ構造の一例を示す図である。図 5 に示すように、ファイルテーブル 143 は、テキストデータを構成する複数のドキュメント ID と、ドキュメントファイルとを対応付ける。ドキュメント ID は、ドキュメントファイルを一意に識別する情報である。ドキュメントファイルは、文章の粒度の文字列（テキストデータ）が設定される。
- [0037] 転置インデックス 144 は、クラスタ ID に属する単語が位置するドキュメントファイルおよび位置を特定するための情報である。転置インデックス 144 のデータ構造は、図 3 で説明した転置インデックス 144 に対応する。図 3 で説明したように、転置インデックス 144 は、クラスタ ID と、ファイルポインタおよびオフセットポインタの組（複数の組）とを対応付ける。ファイルポインタには、クラスタ ID に対応する単語が含まれるドキュメントファイルのドキュメント ID が設定される。オフセットポインタには、該当する単語を含むドキュメントファイルであって、かかるドキュメントファイル上における単語の位置情報（オフセット）が設定される。
- [0038] 入力クエリ 50 は、外部から指定される入力クエリである。入力クエリ 50 には、複数の単語が設定される。

- [0039] 図4の説明に戻る。制御部150は、前処理部151と、検索部152とを有する。制御部150は、たとえば、CPU (Central Processing Unit) やMPU (Micro Processing Unit)により実現される。また、制御部150は、例えばASIC (Application Specific Integrated Circuit) やFPGA (Field Programmable Gate Array) 等の集積回路により実行されてもよい。
- [0040] 前処理部151は、ベクトル辞書141を基にして、クラスタテーブル142を生成する処理、クラスタテーブル142およびファイルテーブル143を基にして、転置インデックス144を生成する処理を実行する。
- [0041] 前処理部151が、ベクトル辞書141を基にして、クラスタテーブル142を生成する処理について説明する。前処理部151の処理を、図1を用いて説明する。
- [0042] 前処理部151は、ベクトル辞書141の各単語に設定されたベクトルに対して、クラスタリングを実行することで、類似する複数のベクトルを、同一のクラスタに分類する。前処理部151は、クラスタリングとして、k-means法を用いてもよい。同一のクラスタに分類されたベクトルに対応する単語には、同一のクラスタIDを割り当てる。同一のクラスタに分類された各単語は、類似する単語であるといえる。
- [0043] 前処理部151は、単語と、クラスタIDと、ベクトルとを対応付けて、クラスタテーブル142に登録する。前処理部151は、係る処理を実行して、クラスタテーブル142を生成する。
- [0044] 前処理部151が、クラスタテーブル142およびファイルテーブル143を基にして、転置インデックス144を生成する処理について説明する。前処理部151の処理を、図2を用いて説明する。
- [0045] 前処理部151は、クラスタテーブル142に含まれる各単語と、ファイルテーブル143とを比較して、転置インデックス144を生成する。たとえば、前処理部151は、クラスタテーブル142の単語を選択し、選択した単語と、ファイルテーブル143の各ドキュメントファイルの文字列とを

比較して、単語がヒットする場合には、該当する単語のクラスターIDと、ファイルポインタと、オフセットポインタとを対応付けて、転置インデックス144に登録する。

[0046] 前処理部151は、クラスタテーブル142に設定された各単語について上記処理を繰り返し実行することで、転置インデックス144を生成する。

[0047] 検索部152は、入力クエリ50を受け付け、入力クエリ50に対応するドキュメントファイルを、ファイルテーブル143から検索する。以下において、検索部152の処理の一例について説明する。検索部152の処理を、図3を用いて説明する。

[0048] 検索部152は、入力クエリ50を受け付けると、入力クエリ50を記憶部140に登録する。検索部152は、入力クエリ50を、入力部120から受け付けてもよいし、通信部110を介して、外部装置から受け付けてもよい。検索部152は、入力クエリ50に含まれる各単語と、クラスタテーブル142とを基にして、各単語に対応するクラスターID（特定クラスターID）を特定する。たとえば、入力クエリ50に、単語「馬」、「人参」、「好き」が含まれている場合には、特定クラスターIDは、「G1」、「G2」、「G3」となる。

[0049] 検索部152は、特定クラスターIDと、転置インデックス144とを基にして、各特定クラスターIDに対応するドキュメントIDを特定する。たとえば、ドキュメントID「Doc1」が、特定クラスターIDを「G1」、「G2」、「G3」に対応する。また、ドキュメントID「Doc2」が、特定クラスターIDを「G1」、「G2」、「G3」に対応する。このため、検索部152は、ドキュメントID「Doc1」、「Doc2」を、各特定クラスターIDに対応するドキュメントIDとして特定する。

[0050] 検索部152は、特定したドキュメントIDに対応するドキュメントファイルを、ファイルテーブル143から抽出し、抽出したドキュメントファイルを、入力クエリ50の検索結果として生成する。検索部152は、検索結果を、表示部130に出力して表示させる。

- [0051] 次に、本実施例１に係る情報処理装置１００の処理手順の一例について説明する。図６は、本実施例１に係る情報処理装置の準備フェーズの処理を示すフローチャートである。図６に示すように、情報処理装置１００の前処理部１５１は、ベクトル辞書１４１の各単語のベクトルに対してクラスタリングを実行する（ステップＳ１０１）。
- [0052] 前処理部１５１は、クラスタリング結果を基にして、単語にクラスタＩＤを設定し、クラスタテーブル１４２を生成する（ステップＳ１０２）。
- [0053] 前処理部１５１は、クラスタテーブル１４２と、ファイルテーブル１４３とを基にして、転置インデックス１４４を生成する（ステップＳ１０３）。
- [0054] 図７は、本実施例１に係る情報処理装置の検索フェーズの処理を示すフローチャートである。図７に示すように、情報処理装置１００の検索部１５２は、入力クエリ５０を受け付け、記憶部１４０に登録する（ステップＳ２０１）。
- [0055] 検索部１５２は、入力クエリ５０に設定された単語を特定する（ステップＳ２０２）。検索部１５２は、クラスタテーブル１４２を基にして、入力クエリに設定された単語の特定クラスタＩＤを特定する（ステップＳ２０３）。
- [0056] 検索部１５２は、特定クラスタＩＤと、転置インデックス１４４とを基にして、特定クラスタＩＤに対応するドキュメントＩＤとオフセットを特定する（ステップＳ２０４）。
- [0057] 検索部１５２は、特定したドキュメントＩＤとオフセットに対応するドキュメントファイルを、ファイルテーブル１４３から抽出して、検索結果を生成する（ステップＳ２０５）。検索部１５２は、検索結果を出力する（ステップＳ２０６）。
- [0058] 次に、本実施例１に係る情報処理装置１００の効果について説明する。情報処理装置１００は、準備フェーズにおいて、転置インデックス１４４を生成しておく。情報処理装置１００は、入力クエリ５０を受け付けると、入力クエリ５０に設定された単語に対応するクラスタＩＤを特定する。情報処理

装置 100 は、転置インデックス 144 を参照し、特定クラスター ID に対応付けられたドキュメント ID を基にして、検索結果を生成する。このように、入力クエリ 50 の各単語に対応する転置インデックスに関して、単語のドキュメント ID とオフセットを単語ベクトルのクラスター ID に対応付けることで類似する単語を含めた検索を行うための転置インデックスのサイズを抑制し、検索時間を短縮することができる。なお、文字列で表現されたテキスト（単語）に加え、ソースプログラム（命令・変数）や有機化合物（官能基）、ゲノム（アミノ酸・タンパク質）、画像（アウトライン PostScript）などに対象を拡大することができる。

[0059] 情報処理装置 100 は、ベクトル辞書 141 に設定された各単語のベクトルに対してクラスタリングを実行し、同一のクラスターに含まれる単語に対して、同一のクラスター ID を設定し、クラスタテーブル 142 を生成する。これによって、類似する単語に同一のクラスター ID を割り当てることができ、かかるクラスター ID を用いることで、入力クエリ 50 に設定された単語に類似する文字列を省資源で高速に検索することができる。

[0060] 情報処理装置 100 は、クラスター ID と、ファイルポインタおよびオフセットポインタとを対応付けた転置インデックス 144 を生成する。これによって、クラスター ID に対応する単語を有するドキュメントファイルおよび単語の位置を効率的に特定することができる。なお、図 2 ではポインタ型の転置インデックスを示したが、ビットマップ型の転置インデックスでも良い。ファイルビットマップとオフセットビットマップのそれぞれが、ハッシュ化され、さらに転置インデックスのサイズを抑制することができる。

実施例 2

[0061] 本実施例 2 に係る情報処理装置について説明する。本実施例 2 に係る情報処理装置は、準備フェーズの処理と、検索フェーズの処理を順に実行する。以下において、情報処理装置が実行する準備フェーズの処理と、検索フェーズの処理を順に説明する。

[0062] 図 8 は、本実施例 2 に係る情報処理装置の準備フェーズの処理を説明する

ための図である。情報処理装置は、実施例 1 の情報処理装置 100 の方法に加えて、ベクトル辞書 241 の各単語に設定されたベクトルに対して、ソーラス辞書の概念分類番号（コンセプトインデックスとも呼ばれる）に準じてクラスタリング ID を設定し、同一のクラスタリング ID を持つ複数の単語ベクトルにポアンカレエンベディングを実行することで、各単語ベクトルを近似し、クラスタテーブル 242 a を生成することができる。これは単語ベクトルを概念番号に準じて、ポアンカレ空間に対応付けたと表現できる。たとえば、情報処理装置は、ベクトル辞書 241 の各単語にポアンカレエンベディングを実行して、各単語にベクトルを割り当てた後、クラスタリングを実行してもよい。

[0063] 情報処理装置は、クラスタテーブル 242 a を生成した後に、クラスタテーブル 242 a に設定された各ベクトルに関して、ベクトル空間 P 上の位置を特定し、同一のクラスタ ID に対応するベクトルを包含する最小の領域（たとえば、球状や立方体状の領域）を特定する。

[0064] たとえば、クラスタ ID 「G1」のベクトルを全て包含する最小の領域を「A1」とする。クラスタ ID 「G2」のベクトルを全て包含する最小の領域を「A2」とする。クラスタ ID 「G3」のベクトルを全て包含する最小の領域を「A3」とする。

[0065] 情報処理装置は、特定した各領域 A1、A2、A3 の中心座標、半径を特定し、該当するクラスタ ID に対応付けて、判定テーブル 242 b に登録する。たとえば、情報処理装置は、領域 A1 の中心座標が (x_1, y_1, z_1) であり、半径が「r1」である場合には、クラスタ ID 「G1」、中心座標 (x_1, y_1, z_1) 、半径「r1」を、判定テーブル 242 b に登録する。

[0066] 情報処理装置は、領域 A2 の中心座標が (x_2, y_2, z_2) であり、半径が「r2」である場合には、クラスタ ID 「G2」、中心座標 (x_2, y_2, z_2) 、半径「r2」を、判定テーブル 242 b に登録する。

[0067] 情報処理装置は、領域 A3 の中心座標が (x_3, y_3, z_3) であり、半

径が「 r_3 」である場合には、クラスターID「 G_3 」、中心座標（ x_3 、 y_3 、 z_3 ）、半径「 r_3 」を、判定テーブル242bに登録する。

[0068] 図9は、本実施例2に係る情報処理装置の検索フェーズの処理を説明するための図（1）である。情報処理装置は、入力クエリ60を受け付ける。入力クエリ60には、複数の単語が含まれる。図9に示す例では、入力クエリ60に、単語「馬」、「人参」、「好き」が含まれる場合を示す。

[0069] 情報処理装置は、ベクトル辞書241を用いて、入力クエリ60の各単語のベクトルを特定する。たとえば、単語「馬」のベクトルを「 WV_α 」、「人参」のベクトルを「 WV_β 」、「好き」のベクトルを「 WV_γ 」とする。

[0070] 情報処理装置は、判定テーブル242bから特定されるベクトル空間P上の領域A1、A2、A3と、入力クエリ60の各単語のベクトル「 WV_α 」、「 WV_β 」、「 WV_γ 」の位置とをそれぞれ比較し、単語のベクトルに対応するクラスターIDを特定する。

[0071] たとえば、単語「馬」のベクトル「 WV_α 」は、領域A1に含まれるため、単語「馬」のクラスターIDを「 G_1 」に設定する。単語「人参」のベクトル「 WV_β 」は、領域A2に含まれるため、単語「人参」のクラスターIDを「 G_2 」に設定する。単語「好き」のベクトル「 WV_γ 」は、領域A3に含まれるため、単語「好き」のクラスターIDを「 G_3 」に設定する。

[0072] 情報処理装置は、上記処理で求めたクラスターID（特定クラスターID）「 G_1 」、「 G_2 」、「 G_3 」と、転置インデックスとを基にして、検索処理を実行する。特定クラスターIDと、転置インデックスとを基にして、検索を行う処理は、情報処理装置100の処理と同様である。

[0073] 図10は、本実施例2に係る情報処理装置の検索フェーズの処理を説明するための図（2）である。情報処理装置は、入力クエリ60を受け付ける。入力クエリ60には、複数の単語が含まれる。図10に示す例では、入力クエリ60に、単語「馬」、「人参」、「お気に入り」が含まれる場合を示す。

[0074] 情報処理装置は、ベクトル辞書241を用いて、入力クエリ60の各単語

のベクトルを特定する。たとえば、単語「馬」のベクトルを「 WV_{α} 」、「人参」のベクトルを「 WV_{β} 」、「お気に入り」のベクトルを「 WV_{θ} 」とする。

[0075] 情報処理装置は、判定テーブル242bから特定されるベクトル空間P上の領域A1、A2、A3と、入力クエリ60の各単語のベクトル「 WV_{α} 」、「 WV_{β} 」、「 WV_{θ} 」の位置とをそれぞれ比較し、単語のベクトルに対応するクラスターIDを特定する。

[0076] たとえば、単語「馬」のベクトル「 WV_{α} 」は、領域A1に含まれるため、単語「馬」のクラスターIDを「G1」に設定する。単語「人参」のベクトル「 WV_{β} 」は、領域A2に含まれるため、単語「人参」のクラスターIDを「G2」に設定する。

[0077] ここで、単語「お気に入り」のベクトル「 WV_{θ} 」は、いずれの領域A1～A3にも含まれていない。この場合には、情報処理装置は、ベクトル「 WV_{θ} 」の位置と、領域A1～A3までの距離（表面までの距離）をそれぞれ比較し、ベクトル「 WV_{θ} 」の位置に最も近い領域A3を特定する。情報処理装置は、ベクトル「 WV_{θ} 」の位置に最も近い領域A3に対応するクラスターID「G3」を、単語「お気に入り」のクラスターIDに設定する。

[0078] 情報処理装置は、上記処理で求めたクラスターID（特定クラスターID）「G1」、「G2」、「G3」と、転置インデックスとを基にして、検索処理を実行する。特定クラスターIDと、転置インデックスとを基にして、検索を行う処理は、情報処理装置100の処理と同様である。

[0079] 上記のように、本実施例2に係る情報処理装置は、準備フェーズにおいて、ベクトル空間P上の領域の情報（中心座標および半径）と、クラスターIDとを対応付けた判定テーブル242bを生成しておく。情報処理装置は、入力クエリ60を受け付けると、入力クエリ60に設定された各単語について、ベクトル空間上の単語のベクトルの位置と、判定テーブル242bに基づく領域とを基にして、単語に対応するクラスターIDを特定し、実施例1と同様にして、検索結果を生成する。このように、入力クエリ60の各単語に対

応するベクトル空間上の位置から、クラスタIDを特定し、検索を行うことで、検索時間を短縮することができる。

[0080] 次に、本実施例2に係る情報処理装置の構成の一例について説明する。図11は、本実施例2に係る情報処理装置の構成を示す機能ブロック図である。図4に示すように、この情報処理装置200は、通信部210、入力部220、表示部230、記憶部240、制御部250を有する。

[0081] 通信部210、入力部220、表示部230に関する説明は、実施例1で説明した通信部110、入力部120、表示部130に関する説明と同様である。

[0082] 記憶部240は、ベクトル辞書241、クラスタテーブル242a、判定テーブル242b、ファイルテーブル243、転置インデックス244、入力クエリ60を有する。記憶部240は、たとえば、RAM、フラッシュメモリ等の半導体メモリ素子、または、ハードディスク、光ディスク等の記憶装置によって実現される。

[0083] ベクトル辞書241は、単語（文字列）とベクトルとを対応付けて保持する。ベクトル辞書241のデータ構造は、図8で示したベクトル辞書241に対応する。

[0084] クラスタテーブル242aは、単語の属するクラスタを特定する場合に用いられるテーブルである。クラスタテーブル242aのデータ構造は、図8で示したクラスタテーブル242aに対応する。図8で説明したように、クラスタテーブル242aは、単語（文字列）と、クラスタIDと、ベクトルとを対応付けるテーブルである。クラスタIDは、単語が属するクラスタを一意に識別する情報である。

[0085] 判定テーブル242bは、クラスタIDに対応するベクトル空間P上の領域A1、A2、A3の情報（中心座標、半径）を保持するテーブルである。判定テーブル242bのデータ構造は、図8で示した判定テーブル242bに対応する。

[0086] ファイルテーブル243は、複数のドキュメントファイルを有する。ファ

イルテーブル 243 のデータ構造は、実施例 1 で説明したファイルテーブル 143 と同様である。

[0087] 転置インデックス 244 は、クラスタ ID に属する単語が位置するドキュメントファイルおよび位置を特定するための情報である。転置インデックス 244 のデータ構造は、実施例 1 で説明した転置インデックス 144 と同様である。

[0088] 入力クエリ 60 は、外部から指定される入力クエリである。入力クエリ 60 には、複数の単語が設定される。

[0089] 図 11 の説明に戻る。制御部 250 は、前処理部 251 と、検索部 252 とを有する。制御部 250 は、たとえば、CPU や MPU により実現される。また、制御部 250 は、例えば ASIC や FPGA 等の集積回路により実行されてもよい。

[0090] 前処理部 251 は、ベクトル辞書 241 を基にして、クラスタテーブル 242 a を生成する処理、クラスタテーブル 242 a を基にして、判定テーブル 242 b を生成する処理、クラスタテーブル 242 およびファイルテーブル 243 を基にして、転置インデックス 144 を生成する処理を実行する。

[0091] 前処理部 251 が、ベクトル辞書 241 を基にして、クラスタテーブル 242 a を生成する処理は、実施例 1 で説明した、前処理部 151 の処理と同様である。

[0092] 前処理部 251 が、クラスタテーブル 242 a を基にして、判定テーブル 242 b を生成する処理について説明する。前処理部 251 の処理を、図 8 を用いて説明する。

[0093] 前処理部 251 は、クラスタテーブル 242 a を生成した後に、クラスタテーブル 242 a に設定された各ベクトルに関して、ベクトル空間 P 上の位置を特定し、同一のクラスタ ID に対応するベクトルを包含する最小の領域 A1、A2、A3 を特定する。

[0094] 前処理部 251 は、特定した各領域 A1、A2、A3 の中心座標、半径を特定し、該当するクラスタ ID と対応付けて、判定テーブル 242 b に登録

する。たとえば、図8で説明した例では、前処理部251は、クラスタID「G1」、中心座標(x_1 、 y_1 、 z_1)、半径「 r_1 」を、判定テーブル242bに登録する。前処理部251は、クラスタID「G2」、中心座標(x_2 、 y_2 、 z_2)、半径「 r_2 」を、判定テーブル242bに登録する。前処理部251は、クラスタID「G3」、中心座標(x_3 、 y_3 、 z_3)、半径「 r_3 」を、判定テーブル242bに登録する。

[0095] 前処理部251は、上記処理を実行することで、判定テーブル242bを生成する。

[0096] 前処理部251が、クラスタテーブル242およびファイルテーブル243を基にして、転置インデックス244を生成する処理は、実施例1で説明した、前処理部151の処理と同様である。

[0097] 検索部252は、入力クエリ60を受け付け、入力クエリ60に対応するドキュメントファイルを、ファイルテーブル243から検索する。以下において、検索部252の処理の一例について説明する。検索部252の処理を、図9、図10を用いて説明する。

[0098] 検索部252は、入力クエリ60を受け付けると、入力クエリ60を記憶部240に登録する。図9に示す例では、入力クエリ60に、単語「馬」、「人参」、「好き」が含まれる。

[0099] 検索部252は、ベクトル辞書241を用いて、入力クエリ60の各単語のベクトルを特定する。たとえば、単語「馬」のベクトルを「 WV_α 」、「人参」のベクトルを「 WV_β 」、「好き」のベクトルを「 WV_γ 」とする。

[0100] 検索部252は、判定テーブル242bから特定されるベクトル空間P上の領域A1、A2、A3と、入力クエリ60の各単語のベクトル「 WV_α 」、「 WV_β 」、「 WV_γ 」の位置とをそれぞれ比較し、単語のベクトルに対応するクラスタIDを特定する。

[0101] 検索部252は、単語「馬」のベクトル「 WV_α 」が、領域A1に含まれるため、単語「馬」のクラスタIDを「G1」に設定する。検索部252は、単語「人参」のベクトル「 WV_β 」が、領域A2に含まれるため、単語「

人参」のクラスタIDを「G2」に設定する。検索部252は、単語「好き」のベクトル「WV γ 」が、領域A3に含まれるため、単語「好き」のクラスタIDを「G3」に設定する。

[0102] 検索部252は、上記処理で求めたクラスタID（特定クラスタID）「G1」、「G2」、「G3」と、転置インデックス244とを基にして、検索処理を実行する。特定クラスタIDと、転置インデックス244とを基にして、検索を行う処理は、情報処理装置100の検索部152の処理と同様である。

[0103] 一方、図10に示す例では、入力クエリ60に、単語「ポニー」、「人参」、「好き」が含まれる。

[0104] 検索部252は、ベクトル辞書241を用いて、入力クエリ60の各単語のベクトルを特定する。たとえば、単語「ポニー」のベクトルを「WV θ 」、「人参」のベクトルを「WV β 」、「好き」のベクトルを「WV γ 」とする。

[0105] 検索部252は、判定テーブル242bから特定されるベクトル空間P上の領域A1、A2、A3と、入力クエリ60の各単語のベクトル「WV θ 」、「WV β 」、「WV γ 」の位置とをそれぞれ比較し、単語のベクトルに対応するクラスタIDを特定する。

[0106] 検索部252は、単語「馬」のベクトル「WV α 」が、領域A1に含まれるため、単語「馬」のクラスタIDを「G1」に設定する。検索部252は、単語「人参」のベクトル「WV β 」は、領域A2に含まれるため、単語「人参」のクラスタIDを「G2」に設定する。

[0107] ここで、固有名詞である単語「ポニー」はシソーラス辞書に含まれておらず、そのベクトル「WV θ 」は、いずれの領域A1～A3にも含まれていない。この場合には、検索部252は、単語「ポニー」が含まれる文や文章をもとにCBOW関数を用いてベクトルが算出される。算出されたベクトル「WV θ 」の位置と、領域A1～A3までの距離（表面までの距離）をそれぞれ比較し、ベクトル「WV θ 」の位置に最も近い領域A1を特定する。検索

部 2 5 2 は、ベクトル「WV θ 」の位置に最も近い領域 A 1 に対応するクラスター ID「G 1」を、単語「ポニー」のクラスター ID に設定する。

[0108] 検索部 2 5 2 は、上記処理で求めたクラスター ID（特定クラスター ID）「G 1」、「G 2」、「G 3」と、転置インデックス 2 4 4 とを基にして、検索処理を実行する。特定クラスター ID と、転置インデックス 2 4 4 とを基にして、検索を行う処理は、情報処理装置 1 0 0 の検索部 1 5 2 の処理と同様である。

[0109] 次に、本実施例 2 に係る情報処理装置 2 0 0 の処理手順の一例について説明する。図 1 2 は、本実施例 2 に係る情報処理装置の準備フェーズの処理を示すフローチャートである。図 1 2 に示すように、情報処理装置 2 0 0 の前処理部 2 5 1 は、ベクトル辞書 2 4 1 の各単語のベクトルに対してシソーラス辞典の概念分類番号に準じて、クラスター ID を設定する（ステップ S 3 0 1）。

[0110] 前処理部 2 5 1 は、クラスタリング結果を基にして、同一のクラスター ID を持つ複数の単語ベクトルに対してポアンカレエンベディングを実行し、クラスタテーブル 2 4 2 a を生成する（ステップ S 3 0 2）。前処理部 2 5 1 は、クラスタテーブル 2 4 2 a と、ファイルテーブル 2 4 3 とを基にして、転置インデックス 2 4 4 を生成する（ステップ S 3 0 3）。

[0111] 前処理部 2 5 1 は、クラスタテーブル 2 4 2 a に設定された各ベクトルに関して、ベクトル空間 P 上の位置を特定し、同一のクラスター ID に対応するベクトルを包含する最小の領域をそれぞれ特定する（ステップ S 3 0 4）。

[0112] 前処理部 2 5 1 は、クラスター ID と、領域の情報（中心座標、半径）との情報を判定テーブル 2 4 2 b に登録する（ステップ S 3 0 5）。

[0113] 図 1 3 は、本実施例 2 に係る情報処理装置の検索フェーズの処理を示すフローチャートである。図 1 3 に示すように、情報処理装置 2 0 0 の検索部 2 5 2 は、入力クエリ 6 0 を受け付け、記憶部 2 4 0 に登録する（ステップ S 4 0 1）。

[0114] 検索部 2 5 2 は、ベクトル辞書 2 4 1 を基にして、入力クエリ 6 0 に設定

された単語のベクトルを特定する（ステップS402）。検索部252は、ベクトル空間P上において、各単語のベクトルと、判定テーブル242bから特定されるクラスタIDに対応する各領域とを基にして、単語に対応する特定クラスタIDを特定する（ステップS403）。

[0115] 検索部252は、特定クラスタIDと、転置インデックス144とを基にして、特定クラスタIDに対応するドキュメントIDとオフセットを特定する（ステップS404）。

[0116] 検索部252は、特定したドキュメントIDとオフセットに対応するドキュメントファイルを、ファイルテーブル243から抽出して、検索結果を生成する（ステップS405）。検索部252は、検索結果を出力する（ステップS406）。

[0117] 次に、本実施例2に係る情報処理装置200の効果について説明する。情報処理装置200は、準備フェーズにおいて、ベクトル空間P上の領域の情報（中心座標および半径）と、クラスタIDとを対応付けた判定テーブル242bを生成しておく。情報処理装置200は、入力クエリ60を受け付けると、入力クエリ60に設定された各単語について、ベクトル空間上の単語のベクトルの位置と、判定テーブル242bに基づく領域とを基にして、単語に対応するクラスタIDを特定し、実施例1と同様にして、検索結果を生成する。このように、入力クエリ60の各単語に対応する転置インデックスに関して、単語のドキュメントIDとオフセットを単語ベクトルのクラスタIDに対応付けることで、類似する単語を含めた検索を行うための転置インデックスのサイズを抑制し、検索時間を短縮することができる。なお、文字列で表現されたテキスト（単語）に加え、ソースプログラム（命令・変数）や有機化合物（官能基）、ゲノム（アミノ酸・タンパク質）、画像（アウトラインPostScript）などに対象を拡大することができる。

[0118] 次に、上記実施例に示した情報処理装置100（200）と同様の機能を実現するコンピュータのハードウェア構成の一例について説明する。図14は、実施例の情報処理装置と同様の機能を実現するコンピュータのハードウ

エア構成の一例を示す図である。

[0119] 図14に示すように、コンピュータ300は、各種演算処理を実行するCPU301と、ユーザからのデータの入力を受け付ける入力装置302と、ディスプレイ303とを有する。また、コンピュータ300は、有線または無線ネットワークを介して、外部装置等との間でデータの授受を行う通信装置304と、インタフェース装置305とを有する。また、コンピュータ300は、各種情報を一時記憶するRAM306と、ハードディスク装置307とを有する。そして、各装置301～307は、バス308に接続される。

[0120] ハードディスク装置307は、前処理プログラム307a、検索プログラム307bを有する。また、CPU301は、各プログラム307a、307bを読み出してRAM306に展開する。

[0121] 前処理プログラム307aは、前処理プロセス306aとして機能する。検索プログラム307bは、検索プロセス306bとして機能する。

[0122] 前処理プロセス306aの処理は、前処理部151(251)の処理に対応する。検索プロセス306bの処理は、検索部152(252)の処理に対応する。

[0123] なお、各プログラム307a、307bについては、必ずしも最初からハードディスク装置307に記憶させておかなくても良い。例えば、コンピュータ300に挿入されるフレキシブルディスク(FD)、CD-ROM、DVD、光磁気ディスク、ICカードなどの「可搬用の物理媒体」に各プログラムを記憶させておく。そして、コンピュータ300が各プログラム307a、307bを読み出して実行するようにしてもよい。

符号の説明

[0124]	100, 200	情報処理装置
	110, 210	通信部
	120, 220	入力部
	130, 230	表示部

1 4 0, 2 4 0 記憶部

1 5 0, 2 5 0 制御部

請求の範囲

- [請求項1] 検索条件として文字列データを受け付けると、複数の文字列データに含まれる複数のクラスタのうち、受け付けた前記文字列データに応じたクラスタを特定し、
- 検索対象のデータベースに含まれる複数の文字列データそれぞれを示す情報を前記複数のクラスタのいずれかに対応付けて記憶する記憶部を参照して、特定した前記クラスタに対応付けられた文字列データを示す情報を特定し、
- 特定した前記文字列データを示す情報に応じた検索結果を生成する、
- 処理をコンピュータが実行することを特徴とする検索方法。
- [請求項2] 複数の文字列データに設定された複数のベクトルに対して、クラスタリングを実行することで、前記複数のベクトルを複数のクラスタに分類し、分類したクラスタを識別する情報と、クラスタに属するベクトルに対応する文字列データが含まれる前記データベースの位置情報とを対応付けたインデックス情報を生成して、前記記憶部に登録する処理を更に実行することを特徴とする請求項1に記載の検索方法。
- [請求項3] 前記文字列データを示す情報を特定する処理は、前記受け付けた文字列データに応じたクラスタと、前記インデックス情報とを基にして、前記受け付けた前記文字列データに応じたクラスタに対応する、前記受け付けた文字列データが含まれる前記データベースの位置情報を、前記文字列データを示す情報として特定することを特徴とする請求項2に記載の検索方法。
- [請求項4] 前記クラスタを特定する処理は、前記検索条件として受け付けた文字列データに関するベクトル空間上の位置と、事前に設定された複数の文字列データのベクトル空間上の領域とを基にして、前記検索条件として受け付けた文字列データのクラスタを特定することを特徴とする請求項1に記載の検索方法。

[請求項5] 検索条件として文字列データを受け付けると、複数の文字列データに含まれる複数のクラスタのうち、受け付けた前記文字列データに応じたクラスタを特定し、

 検索対象のデータベースに含まれる複数の文字列データそれぞれを示す情報を前記複数のクラスタのいずれかに対応付けて記憶する記憶部を参照して、特定した前記クラスタに対応付けられた文字列データを示す情報を特定し、

 特定した前記文字列データを示す情報に応じた検索結果を生成する、

 処理をコンピュータに実行させることを特徴とする検索プログラム。

[請求項6] 複数の文字列データに設定された複数のベクトルに対して、クラスタリングを実行することで、前記複数のベクトルを複数のクラスタに分類し、分類したクラスタを識別する情報と、クラスタに属するベクトルに対応する文字列データが含まれる前記データベースの位置情報とを対応付けたインデックス情報を生成して、前記記憶部に登録する処理を更に実行することを特徴とする請求項5に記載の検索プログラム。

[請求項7] 前記文字列データを示す情報を特定する処理は、前記受け付けた文字列データに応じたクラスタと、前記インデックス情報とを基にして、前記受け付けた前記文字列データに応じたクラスタに対応する、前記受け付けた文字列データが含まれる前記データベースの位置情報を、前記文字列データを示す情報として特定することを特徴とする請求項6に記載の検索プログラム。

[請求項8] 前記クラスタを特定する処理は、前記検索条件として受け付けた文字列データに関するベクトル空間上の位置と、事前に設定された複数の文字列データのベクトル空間上の領域とを基にして、前記検索条件として受け付けた文字列データのクラスタを特定することを特徴とす

る請求項5に記載の検索プログラム。

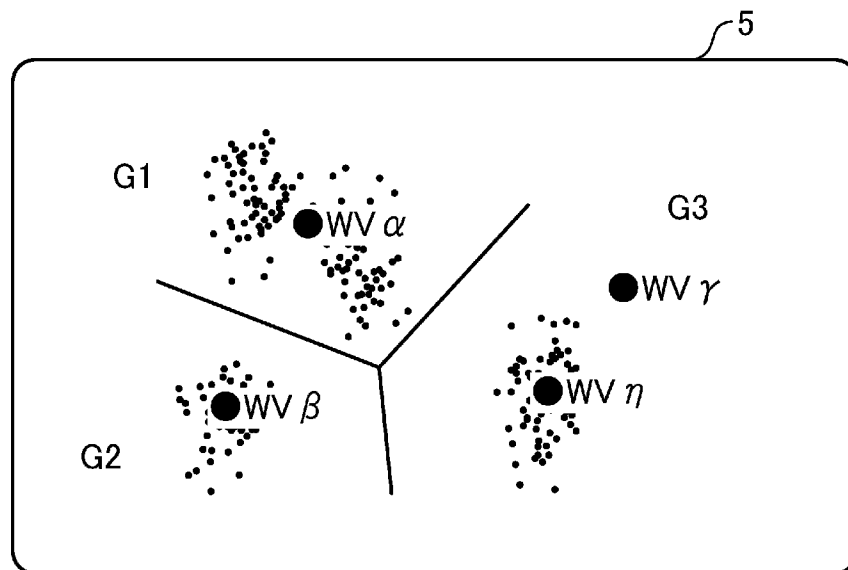
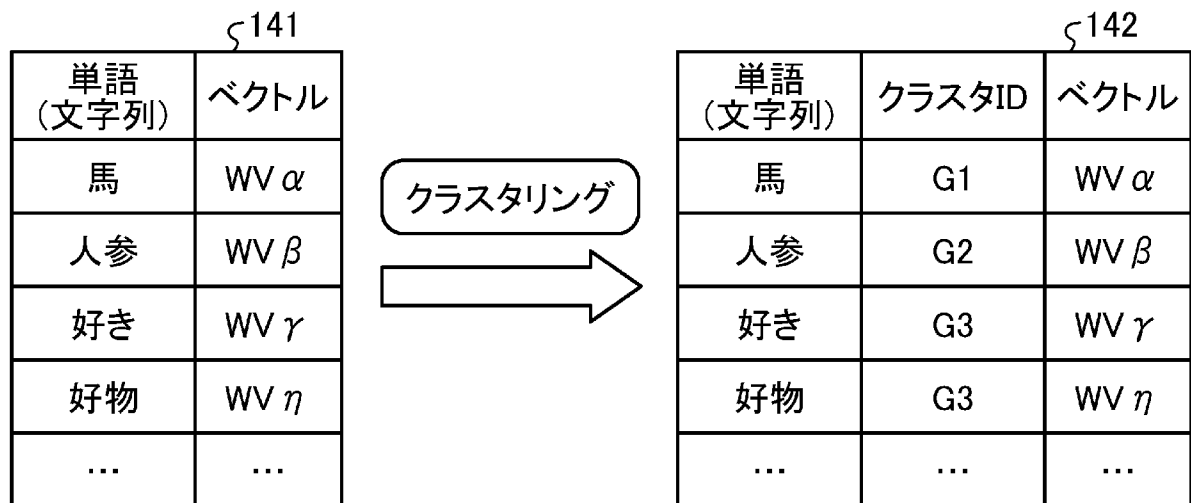
[請求項9] 検索条件として文字列データを受け付けると、複数の文字列データに含まれる複数のクラスタのうち、受け付けた前記文字列データに応じたクラスタを特定し、検索対象のデータベースに含まれる複数の文字列データそれぞれを示す情報を前記複数のクラスタのいずれかに対応付けて記憶する記憶部を参照して、特定した前記クラスタに対応付けられた文字列データを示す情報を特定し、特定した前記文字列データを示す情報に応じた検索結果を生成する検索部、
 を有することを特徴とする情報処理装置。

[請求項10] 複数の文字列データに設定された複数のベクトルに対して、クラスタリングを実行することで、前記複数のベクトルを複数のクラスタに分類し、分類したクラスタを識別する情報と、クラスタに属するベクトルに対応する文字列データが含まれる前記データベースの位置情報とを対応付けたインデックス情報を生成して、前記記憶部に登録する前処理部を更に有することを特徴とする請求項9に記載の情報処理装置。

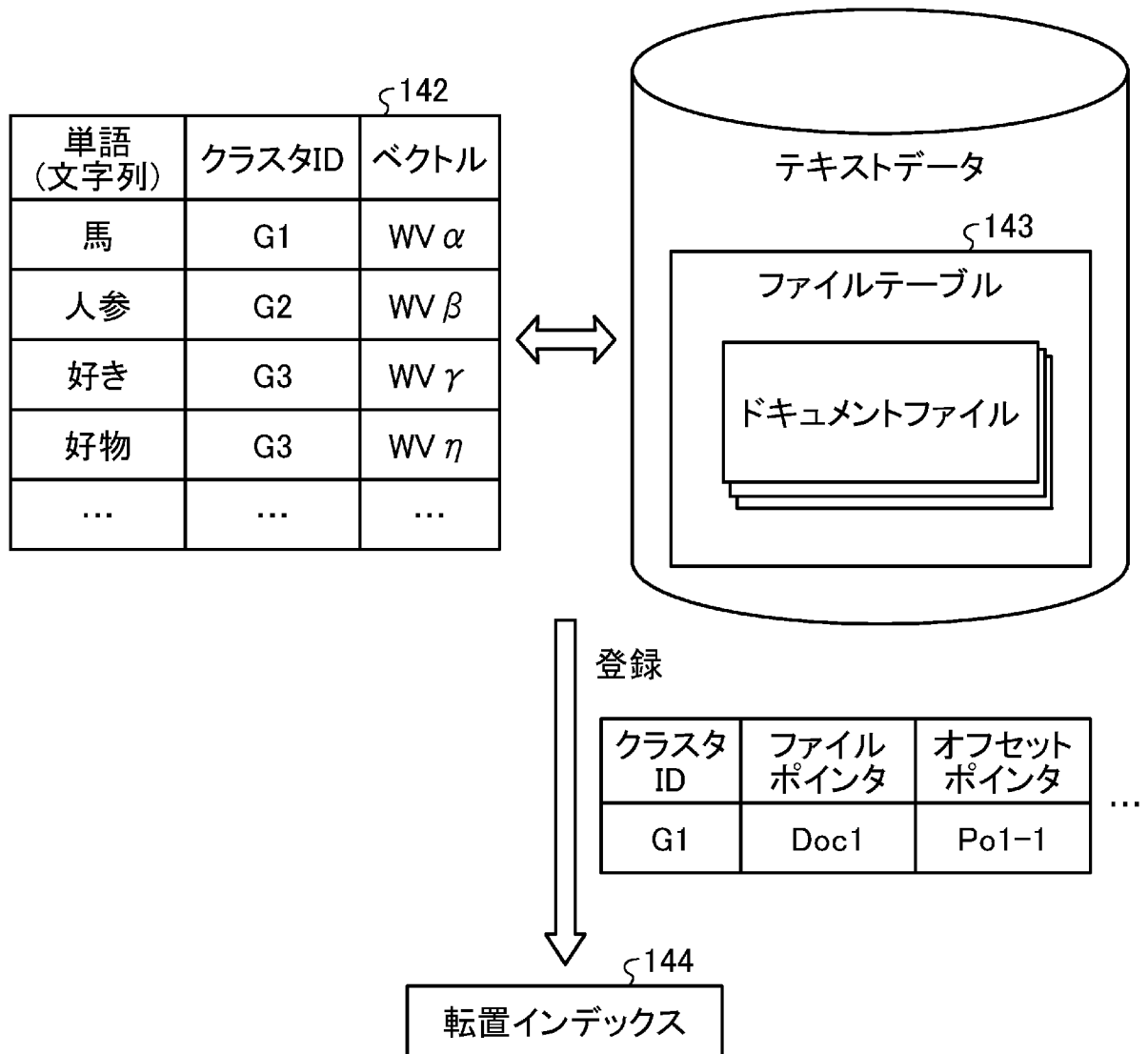
[請求項11] 前記検索部は、前記受け付けた文字列データに応じたクラスタと、前記インデックス情報とを基にして、前記受け付けた前記文字列データに応じたクラスタに対応する、前記受け付けた文字列データが含まれる前記データベースの位置情報を、前記文字列データを示す情報として特定することを特徴とする請求項2に記載の検索方法。

[請求項12] 前記検索部は、前記受け付けた文字列データに応じたクラスタと、前記インデックス情報とを基にして、前記受け付けた前記文字列データに応じたクラスタに対応する、前記受け付けた文字列データが含まれる前記データベースの位置情報を、前記文字列データを示す情報として特定することを特徴とする請求項10に記載の情報処理装置。

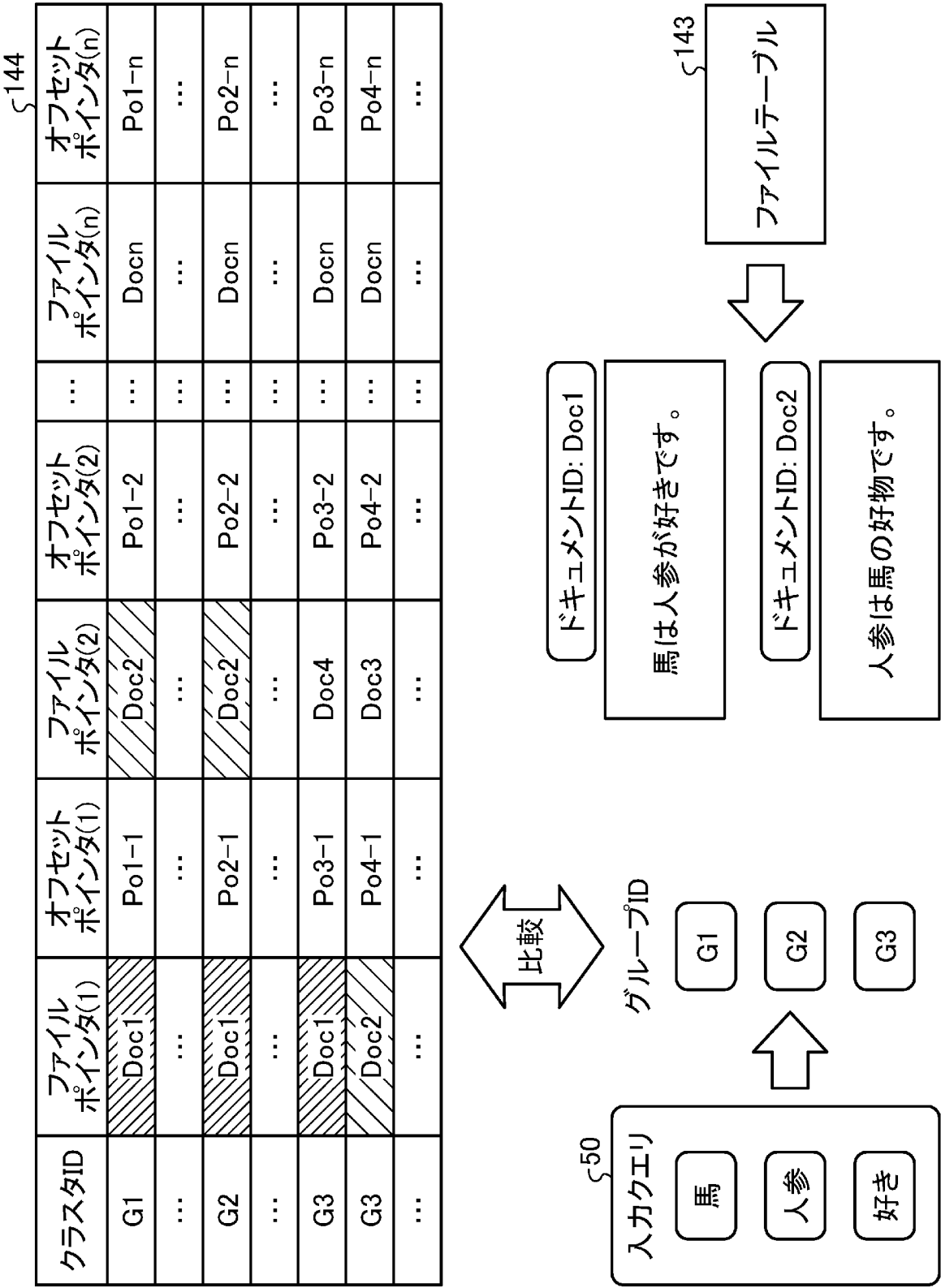
[図1]



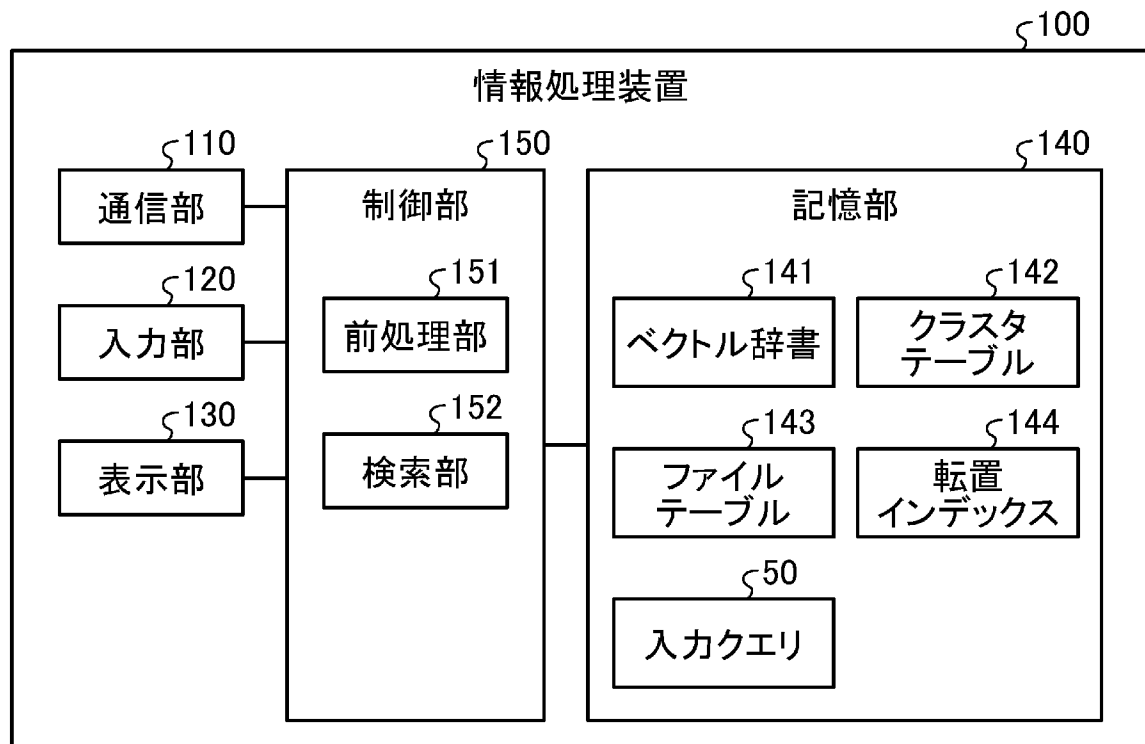
[図2]



[図3]



[図4]

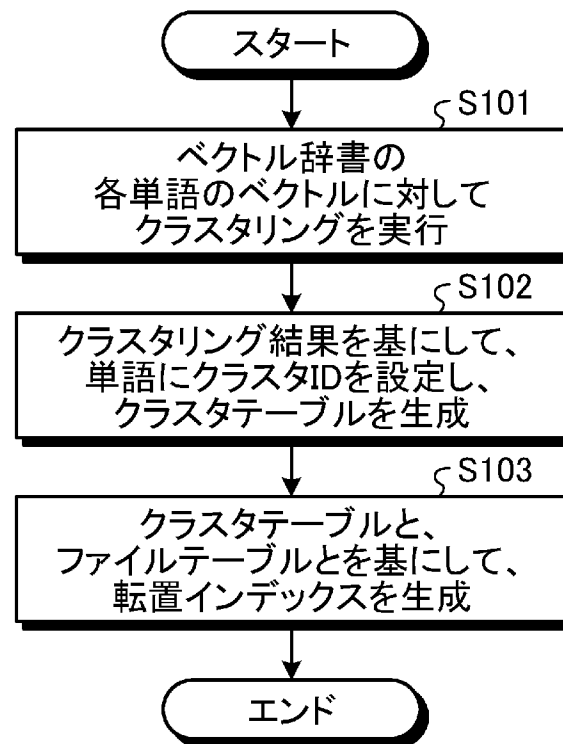


[図5]

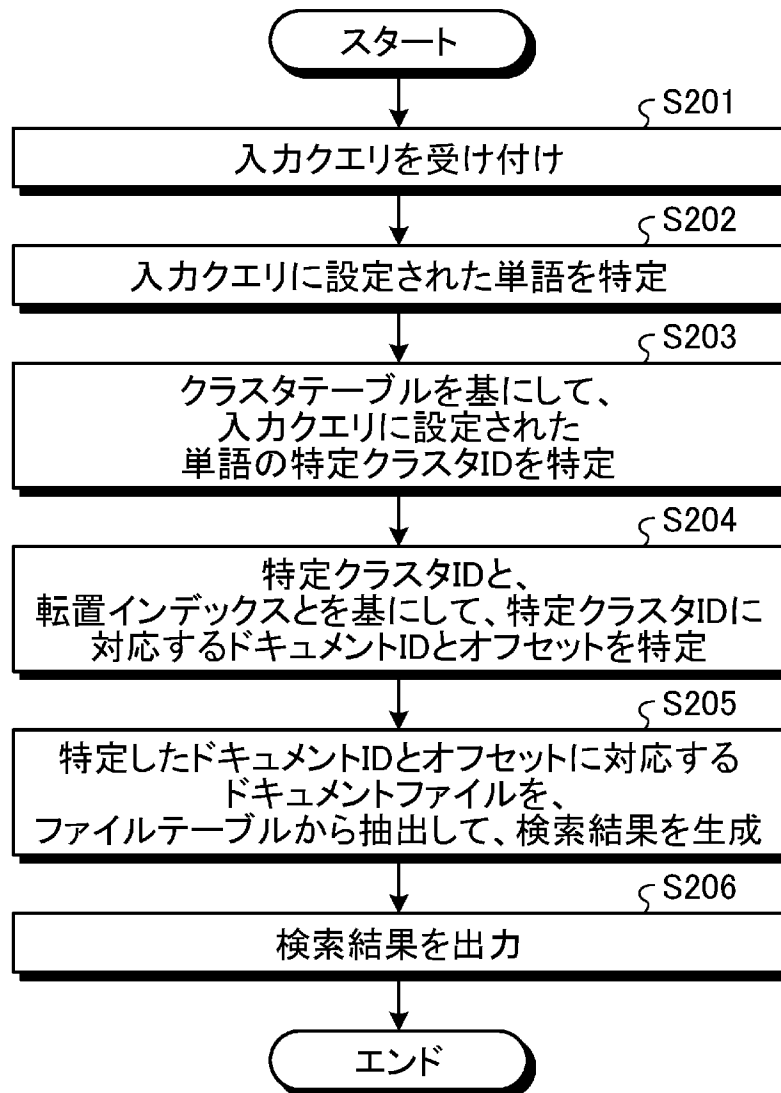
§143

ドキュメントID	ドキュメントファイル
Doc1	ドキュメントID「Doc1」に対応するドキュメントファイルのデータ
Doc2	ドキュメントID「Doc2」に対応するドキュメントファイルのデータ
Doc3	ドキュメントID「Doc3」に対応するドキュメントファイルのデータ
...	...

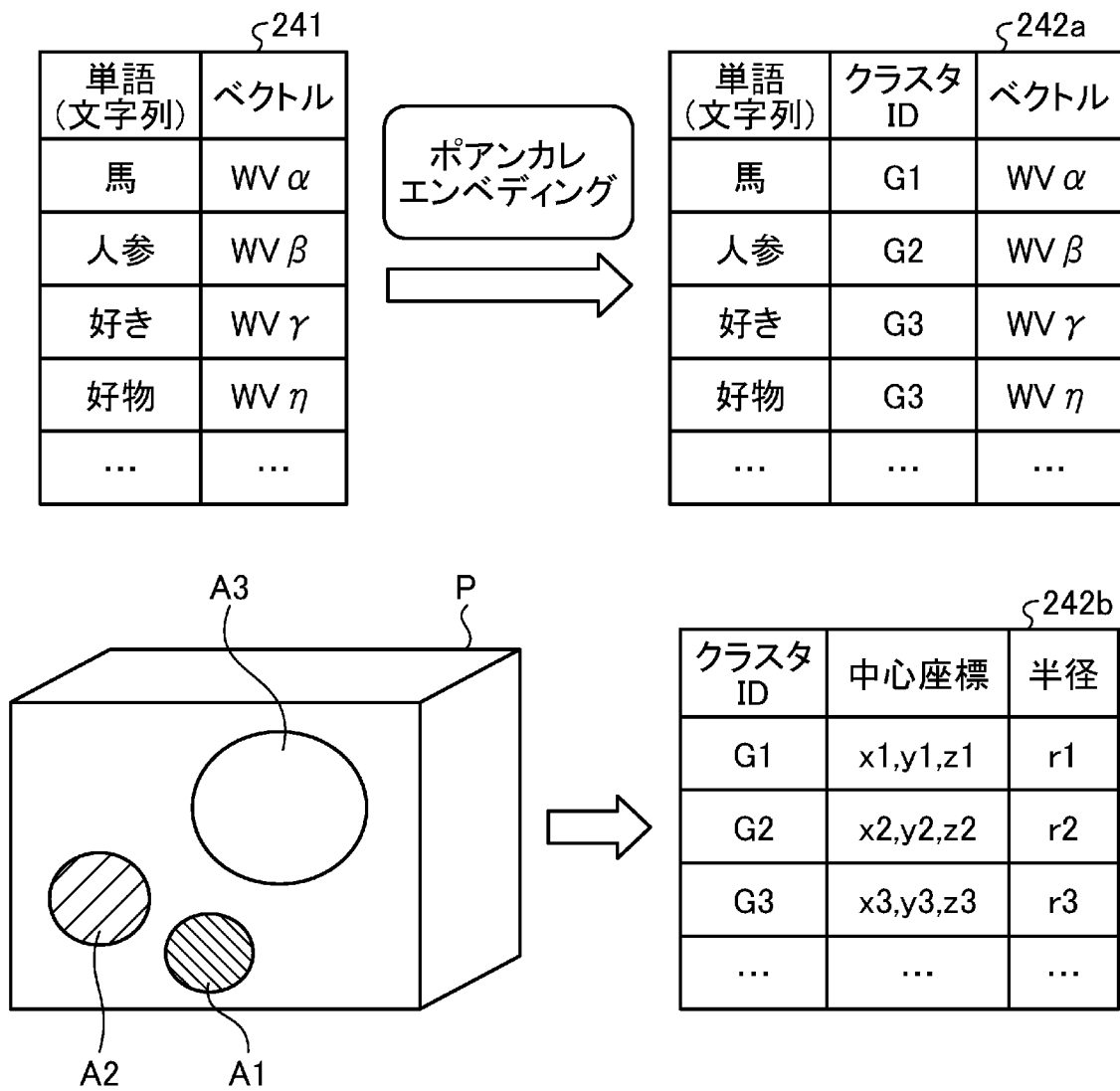
[図6]



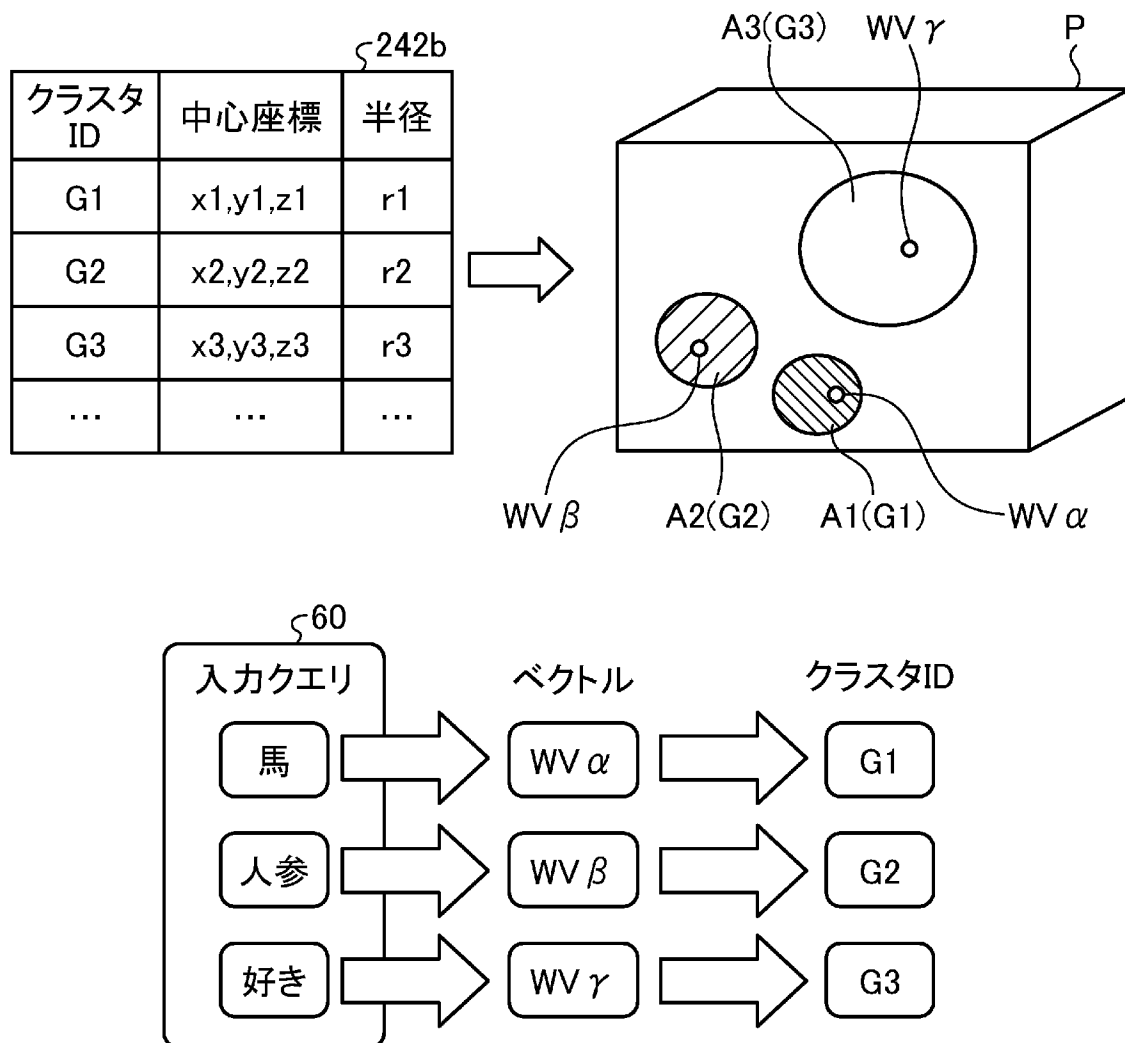
[図7]



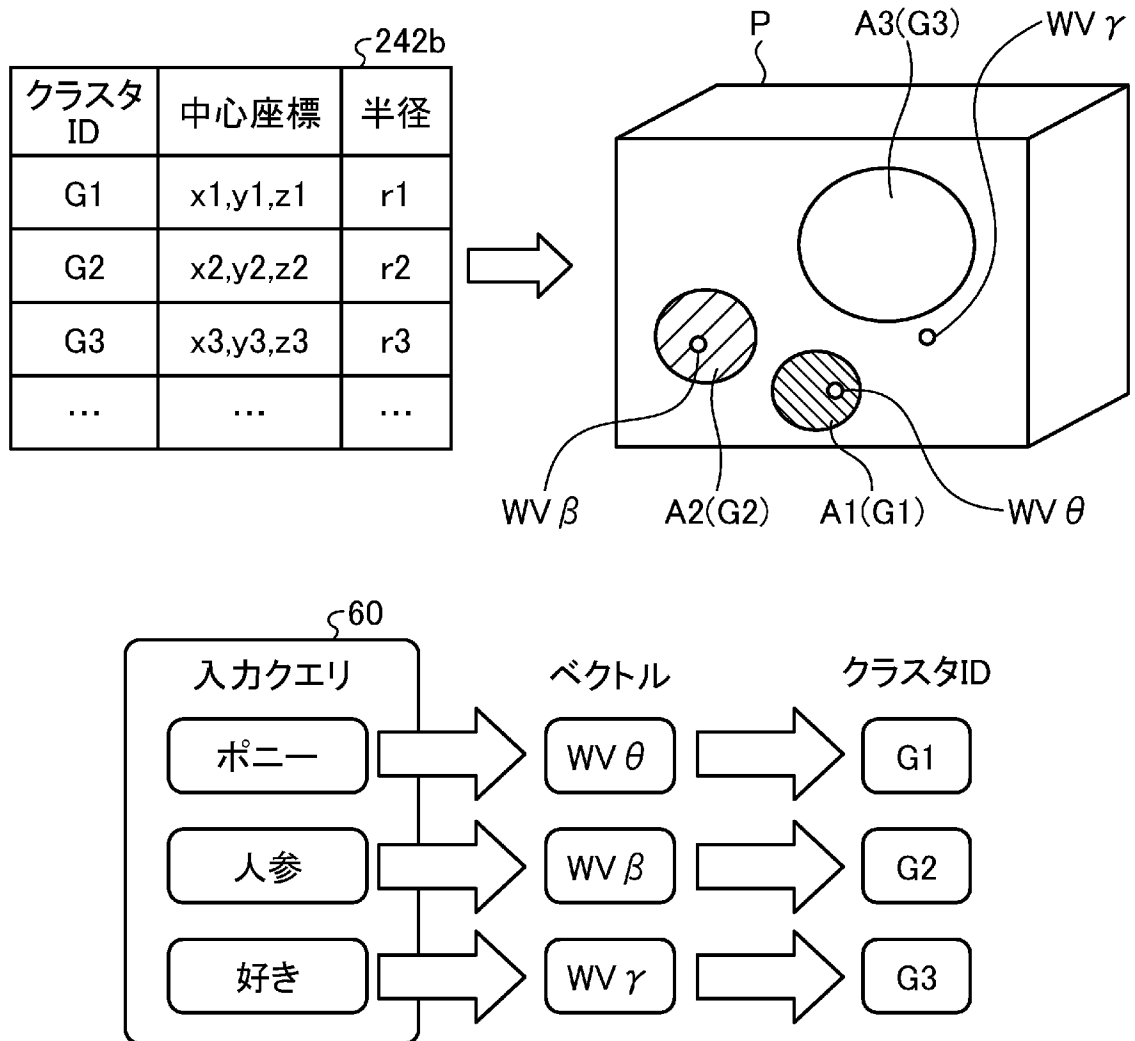
[図8]



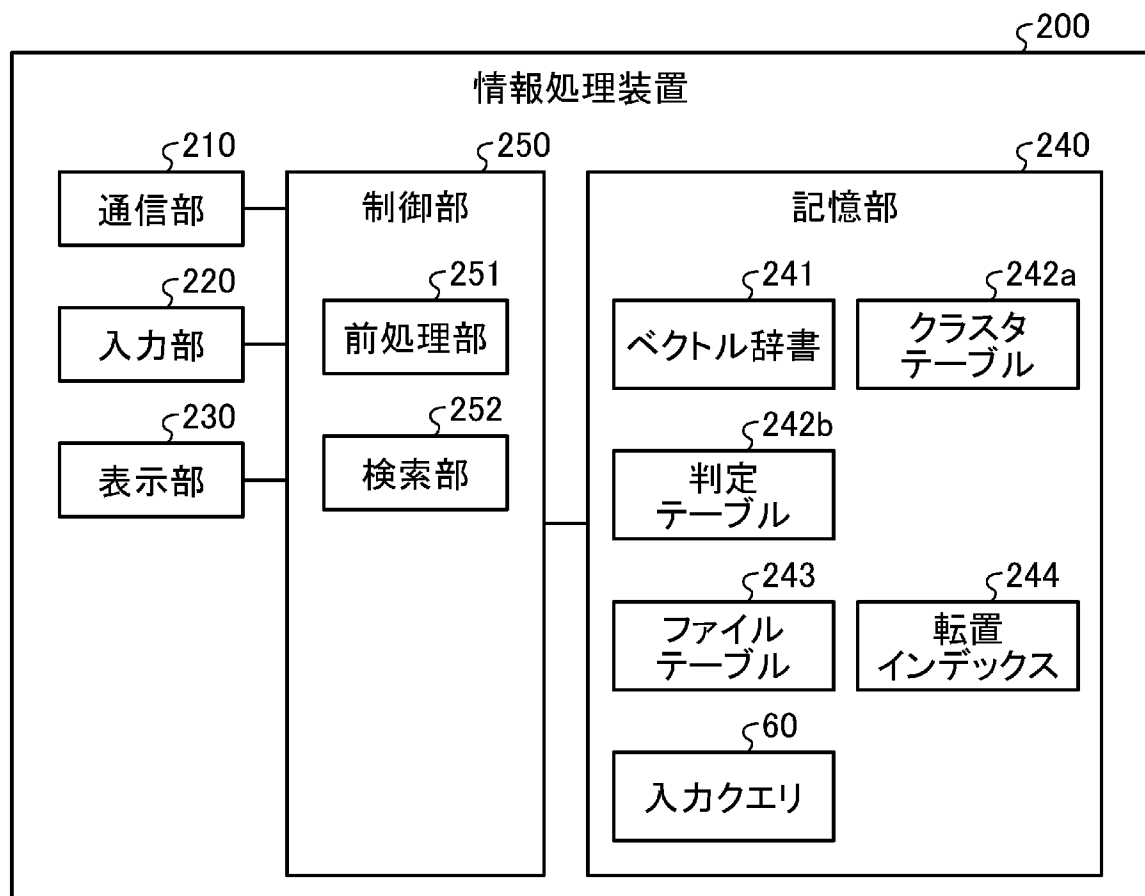
[図9]



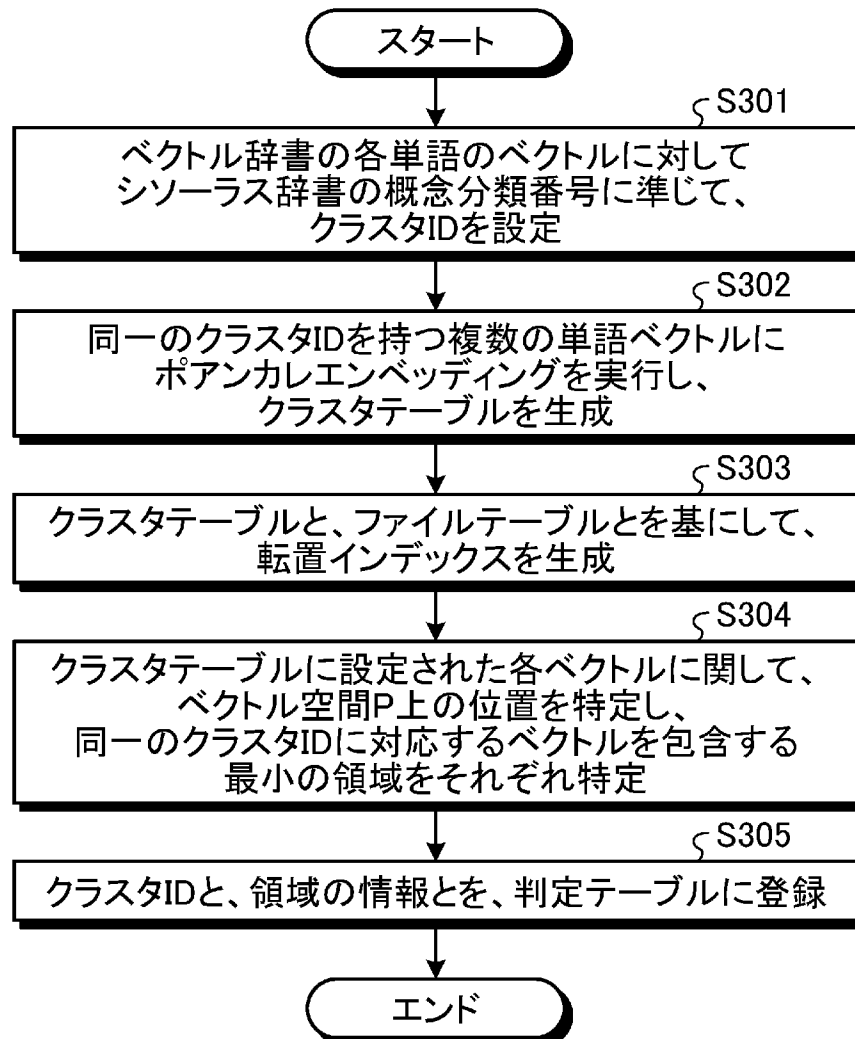
[図10]



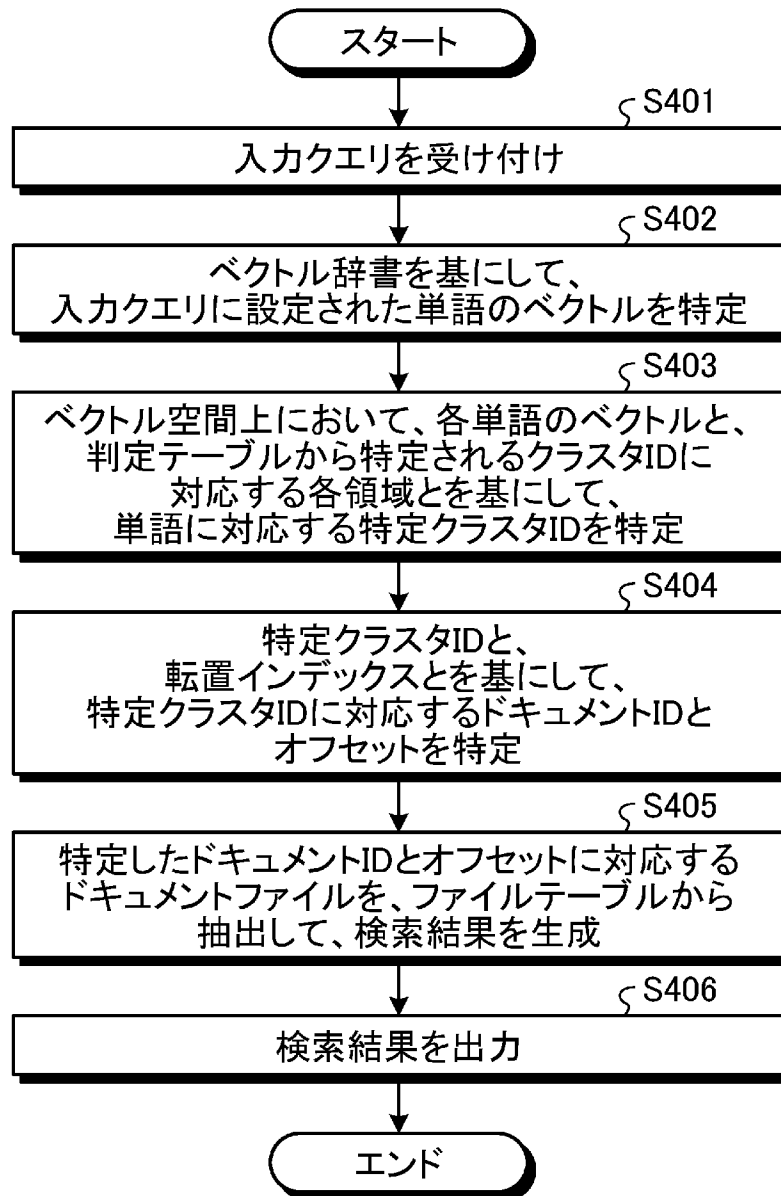
[図11]



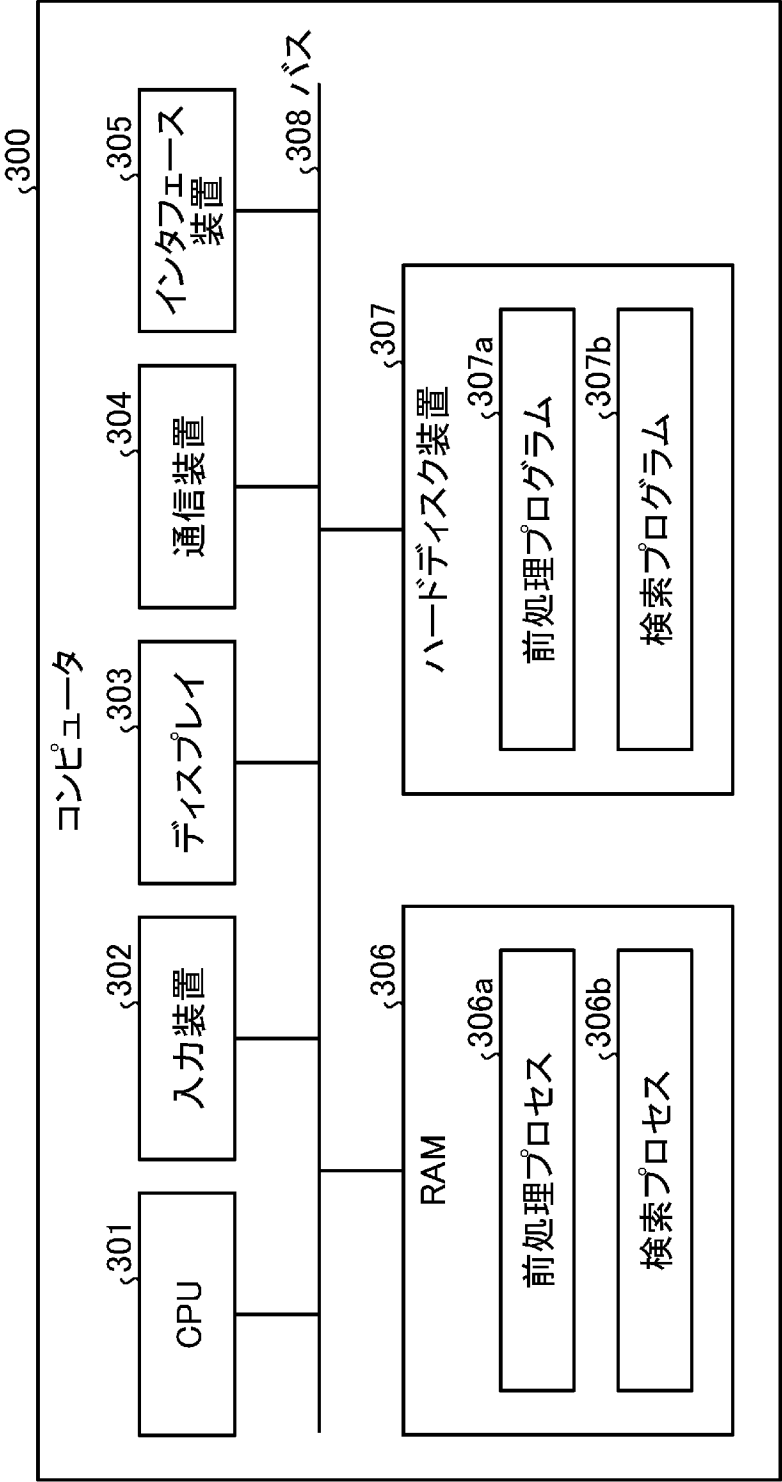
[図12]



[図13]



[図14]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2021/023115

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/31(2019.01)i; G06F 16/35(2019.01)i
FI: G06F16/31; G06F16/35

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)

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	1971-2021
Registered utility model specifications of Japan	1996-2021
Published registered utility model applications of Japan	1994-2021

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.
X	US 2018/0052908 A1 (EBAY INC.) 22 February 2018 (2018-02-22) paragraphs [0085]-[0086], [0090], fig. 8	1-12

☐ 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
10 August 2021 (10.08.2021)

Date of mailing of the international search report
17 August 2021 (17.08.2021)

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

Information on patent family members

International application No.

PCT/JP2021/023115

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
---	---------------------	---------------	---------------------

US 2018/0052908 A1	22 Feb. 2018	(Family: none)	
--------------------	--------------	----------------	--

A. 発明の属する分野の分類（国際特許分類（IPC）） G06F 16/31(2019.01)i; G06F 16/35(2019.01)i FI: G06F16/31; G06F16/35										
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）） 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										
<table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922 - 1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971 - 2021年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2021年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2021年</td> </tr> </table>			日本国実用新案公報	1922 - 1996年	日本国公開実用新案公報	1971 - 2021年	日本国実用新案登録公報	1996 - 2021年	日本国登録実用新案公報	1994 - 2021年
日本国実用新案公報	1922 - 1996年									
日本国公開実用新案公報	1971 - 2021年									
日本国実用新案登録公報	1996 - 2021年									
日本国登録実用新案公報	1994 - 2021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								
X	US 2018/0052908 A1 (EBAY INC.) 22.02.2018 (2018 - 02 - 22) 段落0085-0086, 0090, 図 8	1-12								
☐ 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。										
<table border="0"> <tr> <td style="vertical-align: top;"> * 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 </td> <td style="vertical-align: top;"> “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抵触するものではなく、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 </td> </tr> </table>			* 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抵触するものではなく、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						
* 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抵触するものではなく、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 10.08.2021	国際調査報告の発送日 17.08.2021									
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号	権限のある職員（特許庁審査官） 早川 学 5N 9652 電話番号 03-3581-1101 内線 3586									

PCT/JP2021/023115

様式 PCT/ISA/210 (パテントファミリー用別紙) (2015年1月)