

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号  
特許第7643580号  
(P7643580)

(45)発行日 令和7年3月11日(2025.3.11)

(24)登録日 令和7年3月3日(2025.3.3)

(51)国際特許分類

F I

G 0 6 F 16/31 (2019.01)

G 0 6 F 16/31

G 0 6 F 16/35 (2025.01)

G 0 6 F 16/35

請求項の数 5 (全22頁)

|             |                             |          |                                |
|-------------|-----------------------------|----------|--------------------------------|
| (21)出願番号    | 特願2023-552424(P2023-552424) | (73)特許権者 | 000005223                      |
| (86)(22)出願日 | 令和3年10月4日(2021.10.4)        |          | 富士通株式会社                        |
| (86)国際出願番号  | PCT/JP2021/036696           |          | 神奈川県川崎市中原区上小田中4丁目1番1号          |
| (87)国際公開番号  | WO2023/058099               | (74)代理人  | 110002147                      |
| (87)国際公開日   | 令和5年4月13日(2023.4.13)        |          | 弁理士法人酒井国際特許事務所                 |
| 審査請求日       | 令和5年12月21日(2023.12.21)      | (72)発明者  | 片岡 正弘                          |
|             |                             |          | 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 |
|             |                             | (72)発明者  | 永浦 良平                          |
|             |                             |          | 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 |
|             |                             | (72)発明者  | ミヨ ウェイタント                      |
|             |                             |          | 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 |
|             |                             |          | 最終頁に続く                         |

(54)【発明の名称】 処理方法、処理プログラムおよび情報処理装置

(57)【特許請求の範囲】

【請求項1】

データベースに登録する複数の文字または文字列を受け付けると、前記複数の文字または文字列にそれぞれ応じた複数のベクトルを算出し、

算出した前記複数のベクトルを、ベクトル間距離により複数のクラスタに分類し、

前記複数のクラスタそれぞれについて、各クラスタに含まれる一又は複数のベクトルを代表する代表ベクトルと、前記各クラスタの前記代表ベクトルから前記各クラスタに含まれる前記一又は複数のベクトルの大きさに応じた距離を求め、

前記各クラスタを識別するクラスタ識別情報に対して、算出した前記代表ベクトルと、求めた前記距離とを対応付けた辞書情報を記憶部に記憶し、

前記複数のベクトルに対応するクラスタ識別情報と、前記データベースに登録された複数の文字または文字列の位置情報とを対応付けたインデックス情報を生成する、

処理をコンピュータが実行する処理方法。

【請求項2】

前記複数のベクトルと、前記辞書情報の前記代表ベクトルおよび前記距離とを基にして、前記複数のベクトルに対応するクラスタ識別情報を特定する処理を更にコンピュータが実行することを特徴とする請求項1に記載の処理方法。

【請求項3】

検索クエリを受け付けた場合に、前記検索クエリのベクトルと、前記辞書情報の代表ベクトルおよび距離とを基にして、前記検索クエリのベクトルに対応するクラスタ識別情報

を特定し、前記検索クエリのベクトルに対応するクラスタ識別情報と、前記インデックス情報とを基にして、検索クエリに対応する複数の文字または文字列の位置情報を特定する処理を更にコンピュータが実行することを特徴とする請求項 1 に記載の処理方法。

【請求項 4】

データベースに登録する複数の文字または文字列を受け付けると、前記複数の文字または文字列にそれぞれ応じた複数のベクトルを算出し、

算出した前記複数のベクトルを、ベクトル間距離により複数のクラスタに分類し、

前記複数のクラスタそれぞれについて、各クラスタに含まれる一又は複数のベクトルを代表する代表ベクトルと、前記各クラスタの前記代表ベクトルから前記各クラスタに含まれる前記一又は複数のベクトルの大きさに応じた距離を求め、

前記各クラスタを識別するクラスタ識別情報に対して、算出した前記代表ベクトルと、求めた前記距離とを対応付けた辞書情報を記憶部に記憶し、

前記複数のベクトルに対応するクラスタ識別情報と、前記データベースに登録された複数の文字または文字列の位置情報とを対応付けたインデックス情報を生成する、

処理をコンピュータに実行させることを特徴とする処理プログラム。

【請求項 5】

データベースに登録する複数の文字または文字列を受け付けると、前記複数の文字または文字列にそれぞれ応じた複数のベクトルを算出し、

算出した前記複数のベクトルを、ベクトル間距離により複数のクラスタに分類し、

前記複数のクラスタそれぞれについて、各クラスタに含まれる一又は複数のベクトルを代表する代表ベクトルと、前記各クラスタの前記代表ベクトルから前記各クラスタに含まれる前記一又は複数のベクトルの大きさに応じた距離を求め、

前記各クラスタを識別するクラスタ識別情報に対して、算出した前記代表ベクトルと、求めた前記距離とを対応付けた辞書情報を記憶部に記憶し、

前記複数のベクトルに対応するクラスタ識別情報と、前記データベースに登録された複数の文字または文字列の位置情報とを対応付けたインデックス情報を生成する、

処理を実行する制御部を有する情報処理装置。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、処理方法等に関する。

【背景技術】

【0002】

DB (Data Base) にはテキストなどの膨大な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、入力クエリを受け付けた場合に、事前に設定しておいた転置インデックスを用いて、データ検索を実行する。事前準備において、従来技術では、各テキストが DB に登録されると、形態素解析し、その各単語と、該当するテキストの DB 上の位置情報とをそれぞれ対応付けることで、転置インデックスを生成している。係る処理は、新規にテキストが DB に登録される度に、繰り返し実行される。

【先行技術文献】

【特許文献】

【0004】

【文献】特開 2017-111479 号公報

【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、上述した従来技術では、多次元のベクトルに対応した AI (Artificial

10

20

30

40

50

Intelligence) 分析に向け、データの登録処理の負荷が大きいという問題がある。

【0006】

AI分析では、文字列で構成されるテキストや有機化合物の化学構造式、ゲノムの塩基配列などのデータに登録において、意味を持つ単位の単語や官能基、文や官能基一次構造などに対し、ベクトルの割り当てや積算の必要性が発生する。しかし、テキストを例にすると、単語の種類は一般的に約百万種類であり、DBに登録された全てのテキストの各単語や文のベクトルとその位置を転置インデックスに対応付けて登録することは困難である。このため、転置インデックスに登録するベクトルの数を絞り込む必要があるが、膨大な数のベクトルから、検索処理に有効なベクトルを同定する処理に対して多くの時間を要する。

10

【0007】

1つの側面では、本発明は、データの登録処理の負荷を軽減することができる処理方法、処理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】

【0008】

第1の案では、コンピュータが次の処理を実行する。コンピュータは、データベースに登録する複数の文字または文字列を受け付けると、複数の文字または文字列にそれぞれ応じた複数のベクトルを算出する。コンピュータは、算出した複数のベクトルを、ベクトル間距離により複数のクラスタに分類する。コンピュータは、複数のクラスタそれぞれについて、各クラスタに含まれる一又は複数のベクトルを代表する代表ベクトルと、各クラスタの代表ベクトルから各クラスタに含まれる一又は複数のベクトルの大きさに応じた距離を求める。コンピュータは、各クラスタを識別するクラスタ識別情報に対して、算出した代表ベクトルと、求めた距離とを対応付けた辞書情報を記憶部に記憶する。

20

【発明の効果】

【0009】

データの登録処理の負荷を軽減することができる。

【図面の簡単な説明】

【0010】

【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である。

30

【図2】図2は、文ベクトル辞書のデータ構造の一例を示す図である。

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

【図5】図5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6】図6は、本実施例1に係るクラスターテーブルのデータ構造の一例を示す図である。

【図7】図7は、クラスターデータの一例を示す図である。

【図8】図8は、単語ベクトル辞書のデータ構造の一例を示す図である。

【図9】図9は、文ベクトルを算出する処理を説明するための図(1)である。

40

【図10】図10は、文ベクトルを算出する処理を説明するための図(2)である。

【図11】図11は、クラスターIDを特定する処理を説明するための図(1)である。

【図12】図12は、クラスターIDを特定する処理を説明するための図(2)である。

【図13】図13は、準備処理の処理手順を示すフローチャートである。

【図14】図14は、検索処理の処理手順を示すフローチャートである。

【図15】図15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16】図16は、第2単語転置インデックスの一例を示す図である。

【図17】図17は、文転置インデックスTSのその他の適用例を説明するための図である。

【図18】図18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

50

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】

【0011】

以下に、本願の開示する処理方法、処理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】

【0012】

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、登録フェーズの処理、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とを順に説明する。

【0013】

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である。本実施例1では一例として、複数の単語を有する文を、テキストとして説明を行う。図1の説明では、文1、文2、文3、文4を用いて説明を行う。なお、準備フェーズの処理は多量の文が含まれる母集団を対象に実施されるものとする。

【0014】

文1を「馬は人参が好きです。」とする。文2を「人参は馬の好物です。」とする。文3を「今日の天気は快晴です。」とする。文4を「桜の花が咲きました。」とする。なお、文1と文2は、主語（S）と目的語（O）が置換されているが、同じ意味を持つ文である。

【0015】

情報処理装置は、DBに登録する複数の文の情報を受け付けると、文のベクトルをそれぞれ算出する。情報処理装置は、文に含まれる各単語のベクトルを積算することで、文のベクトルを算出する。以下の説明では、文のベクトルを「文ベクトル」と表記する。各単語のベクトルは、単語ベクトル辞書に定義される。単語のベクトルを「単語ベクトル」と表記する。

【0016】

文1の文ベクトルを「SV1」とする。文2の文ベクトルを「SV2」とする。文3の文ベクトルを「SV3」とする。文4の文ベクトルを「SV4」とする。本実施例1では一例として、文ベクトルの次元を7次元として説明を行う。文ベクトルの第n成分のベクトルを文ベクトル（n）と表記する（ $n = 1 \sim 7$ ）。

【0017】

情報処理装置は、各文の文ベクトルを基にしてクラスタリングを実行することで、各文を複数のクラスタに分類する。クラスタは、7次元の特徴量空間1に配置される。各クラスタには、クラスタを一意に識別するクラスタID（identification）が設定される。図1に示す例では、文1および文2は、クラスタID「Cr1」のクラスタに分類される。文3は、クラスタID「Cr2」のクラスタに分類される。文4は、クラスタID「Cr3」のクラスタに分類される。

【0018】

以下の説明では、クラスタID「Cr1」～「Cr3」のクラスタをそれぞれ、クラスタCr1～Cr3と表記する。図1では、クラスタCr1～Cr3を示すが、特徴量空間1には、他のクラスタが含まれていてもよい。

【0019】

情報処理装置は、クラスタリングの結果を基にして、文ベクトル辞書を生成する。図2は、文ベクトル辞書のデータ構造の一例を示す図である。図2に示すように、文ベクトル辞書50は、クラスタIDと、代表ベクトルと、径とを対応付ける。

【0020】

クラスタIDは、クラスタを一意に識別する情報である。代表ベクトルは、同一のクラスタに含まれる文ベクトルを代表するベクトルである。代表ベクトルは、代表ベクトル（1）～（7）を有する。代表ベクトル（n）は、同一のクラスタに属する文ベクトル（n

10

20

30

40

50

）の中心（重心）に対応するベクトルである（ $n = 1 \sim 7$ ）。径は、クラスタの球（立方体）の径（辺の長さ）などを示す。

【0021】

図3は、本実施例1に係る情報処理装置の登録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、DB141に登録する入力テキスト2を受け付けると、入力テキスト2に含まれる各文と、文ベクトル辞書50とを基にして、文転置インデックスTSを生成する。

【0022】

図3で説明する例では、入力テキスト2を「子供と牧場に行った。ポニーは人参が好きだ。乗馬のご褒美に人参を与えた」とする。説明の便宜上、入力テキスト2に含まれる文のうち、「子供と牧場に行った」を文2Aと表記する。「ポニーは人参が好きだ。」を文2Bと表記する。「乗馬のご褒美に人参を与えた」を文2Cと表記する。

10

【0023】

文2B「ポニーは人参が好きだ。」を用いて説明を行う。情報処理装置は、文2Bに含まれる単語の単語ベクトルを積算することで、文ベクトル「SV2B」を算出する。情報処理装置は、文2Bの文ベクトルと、文ベクトル辞書50の代表ベクトルとを比較して、文2Bが属するクラスタIDを特定する。

【0024】

たとえば、情報処理装置は、文ベクトルと代表ベクトルとの類似度（コサイン類似度、あるいは距離）を算出し、文ベクトルとの類似度が最大となる代表ベクトルを特定し、特定した代表ベクトルのクラスタIDを、文2Bの属するクラスタIDとする。

20

【0025】

情報処理装置は、文2Bの文ベクトル「SV2B」と、クラスタCr1の代表ベクトルとの類似度が最大となる場合に、文2Bの属するクラスタIDを「Cr1」と特定する。

【0026】

情報処理装置は、DB141に、入力テキスト2をレコードの単位で登録する。情報処理装置は、クラスタIDと、レコードポインタと、位置ポインタとを対応付けて、文転置インデックスTSに登録する。レコードポインタは、該当するクラスタに属する文を有するテキストが登録されたDB141上のレコードの位置であって、DB141の先頭から位置を示す。位置ポインタは、該当するクラスタに属する文が登録されたDB141の該当レコード上の位置であって、レコードポインタからの相対的な位置を示す。

30

【0027】

たとえば、文2Bの文ベクトル「SV2B」がクラスタCr1に属する場合には、文転置インデックスTSにおいて、レコードポインタ（1）には、入力テキスト2が登録されたDB141の該当レコードの先頭位置（DB141の先頭からの位置）が設定される。位置ポインタ（1）には、DB141の該当レコードに登録された文2Bの位置（該当レコードの先頭からの相対的な位置）が設定される。情報処理装置は、同一のクラスタIDに対応するレコードポインタ（m）、位置ポインタ（m）に、既に他の情報が登録されている場合には、レコードポインタ（m+1）、位置ポインタ（m+1）に、上記の情報を登録する。mは、自然数に対応する。

40

【0028】

情報処理装置は、文2A、文2Cについても、文2Bと同様にして、文ベクトルを基にして、クラスタIDを特定し、文転置インデックスTSに対して、クラスタIDと、レコードポインタと、位置ポインタとを対応付けて登録する。

【0029】

情報処理装置は、DB141のレコードに登録する他の入力テキストを取得した場合も、上記処理を繰り返し実行し、文転置インデックスTSを更新する。

【0030】

上記のように、本実施例1に係る情報処理装置は、入力テキスト2をDB141のレコードに登録する場合に、文ベクトル辞書50を用いて、入力テキストの文がいずれのクラ

50

スタIDに属するかを特定し、特定した結果を基にして、文転置インデックスTSを生成する。このように、予め、文の各クラスタの代表ベクトルや径が格納された、文ベクトル辞書50を用いることで、文とクラスタIDとを容易に紐づけることができるため、登録処理の負荷を軽減することができる。

【0031】

続いて、検索フェーズの処理について説明する。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、検索クエリを受け付けると、検索クエリと、文転置インデックスTSを基にして、文を検索する。

【0032】

図4では、検索クエリ4を「ポニーは人参が好きだ。」として説明を行う。情報処理装置は、検索クエリ4に含まれる単語の単語ベクトルを積算することで、文ベクトル「QV4」を算出する。情報処理装置は、検索クエリ4の文ベクトルと、文ベクトル辞書50の代表ベクトルとを比較して、検索クエリ4が属するクラスタIDを特定する。

【0033】

たとえば、情報処理装置は、検索クエリ4の文ベクトル「QV4」と代表ベクトルとの類似度（コサイン類似度、あるいは距離）を算出する。情報処理装置は、文ベクトル「QV4」との類似度が最大となる代表ベクトルを特定し、特定した代表ベクトルのクラスタIDを、検索クエリ4の属するクラスタIDとする。ここでは、検索クエリ4の属するクラスタIDを「Cr1」として説明を行う。

【0034】

情報処理装置は、検索クエリ4の属するクラスタID「Cr1」と、文転置インデックスTSとを比較して、検索クエリ4に対応する文をDB141から検索する。たとえば、情報処理装置は、クラスタID「Cr1」の行に設定されたレコードポインタ（1）、位置ポインタ（1）に示される、文「ポニー人参が好きだ。」を抽出する。

【0035】

情報処理装置は、クラスタID「Cr1」の行に設定されたレコードポインタ（2）、位置ポインタ（2）に示される、文「馬は人参が好きだ。」を抽出する。情報処理装置は、クラスタID「Cr1」の行に設定されたレコードポインタ（3）、位置ポインタ（3）に示される、文「Hoses likes carrots」を抽出する。

【0036】

説明を省略するが、情報処理装置は、クラスタID「Cr1」の行に設定されたレコードポインタ（m）、位置ポインタ（m）に示される文を抽出する。

【0037】

上記処理のように、情報処理装置は、検索クエリ4の属するクラスタIDと、文転置インデックスTSとを用いることで、検索クエリ4に対応する文を効率的に検索することができる。図4で説明した例では、情報処理装置は、検索クエリ4に対応する文「ポニー人参が好きだ。」、文「馬は人参が好きだ。」、文「Hoses likes carrots」を検索する。

【0038】

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5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。

【0039】

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。

【0040】

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、検索クエリ等を入力してもよい。

10

20

30

40

50

## 【0041】

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、検索クエリに対応する検索結果が、表示部130に表示される。

## 【0042】

記憶部140は、テキストデータ40と、クラスタテーブル41と、単語ベクトル辞書45と、文ベクトル辞書50と、DB141と、文転置インデックスTSとを有する。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。

10

## 【0043】

テキストデータ40は、文ベクトル辞書50を生成するために用いられる情報である。たとえば、テキストデータ40は、青空文庫等に含まれる文、ネットワーク上に公開されている論文等が含まれる。テキストデータ40に含まれる文は、図1で説明した文1～文4等に対応する。

## 【0044】

クラスタテーブル41は、クラスタに関する情報を保持する。図6は、本実施例1に係るクラスタテーブルのデータ構造の一例を示す図である。図6に示すように、このクラスタテーブル41は、クラスタIDと、クラスタデータとを対応付ける。クラスタIDは、クラスタを一意に識別する情報である。

20

## 【0045】

クラスタデータは、該当するクラスタに含まれる文ベクトルの情報等を有する。図7は、クラスタデータの一例を示す図である。図7では、クラスタID「Cr1」のクラスタデータについて説明する。図7に示すように、クラスタデータ5には、領域5a、5b、テーブル5c、5dを有する。領域5aには、クラスタID「Cr1」が設定される。領域5bには、クラスタ（クラスタCr1）の径が設定される。

## 【0046】

テーブル5cには、クラスタ（クラスタCr1）の代表ベクトル(1)～(7)が設定される。テーブル5dには、クラスタCr1に属する文ベクトルが登録される。代表ベクトル(1)～(7)は、テーブル5dに登録された文ベクトルを基に算出される。代表ベクトル(n)は、テーブル5dの各ベクトルの第n成分のベクトルの中心を示すベクトルとなる。

30

## 【0047】

図示を省略するが、図6に示した他のクラスタデータについても、図7で説明した情報と同様に情報を有しているものとする。

## 【0048】

図5の説明に戻る。単語ベクトル辞書45は、単語に割り当てられた符号、単語ベクトルを定義するテーブルである。図8は、単語ベクトル辞書のデータ構造の一例を示す図である。図8に示すように、この単語ベクトル辞書45は、符号、単語、単語ベクトル(1)～(7)を有する。符号は、単語に割り当てられる符号(Code)である。単語は、文字列に含まれる単語である。単語ベクトル(1)～(7)は、単語に割り当てられたベクトルである。単語ベクトルの第n成分を単語ベクトル(n)と表記する(n=1～7)。

40

## 【0049】

文ベクトル辞書50は、図2で説明した文ベクトル辞書50に対応する。図2で説明したように、文ベクトル辞書50には、クラスタIDと、代表ベクトルと、径とを対応付ける。

## 【0050】

DB141は、入力テキスト等を登録する。

## 【0051】

50

文転置インデックスTSは、図3、図4で説明した文転置インデックスTSに対応する。文転置インデックスTSは、クラスタIDと、レコードポインタおよび位置ポインタの組とを対応付ける。一つのクラスタIDに、複数のレコードポインタおよび位置ポインタの組が登録される場合もある。

#### 【0052】

図5の説明に戻る。制御部150は、取得部151と、前処理部152と、検索部153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。

10

#### 【0053】

取得部151は、通信部110、入力部120を介して、各種のデータを取得する。たとえば、取得部151は、テキストデータ40を取得し、記憶部140に登録する。

#### 【0054】

前処理部152は、テキストデータ40を基にして、文転置インデックスTSを生成する。前処理部152は、文ベクトルを算出する処理、文ベクトル辞書を生成する処理、文転置インデックスTSを生成する処理を順に実行する。

#### 【0055】

前処理部152が実行する文ベクトルを算出する処理について説明する。図9および図10は、文ベクトルを算出する処理を説明するための図である。図9について説明する。図9では、前処理部152が、テキストデータ40に含まれる文1「馬は人参が好きです。」を抽出し、文1の文ベクトルを算出する場合について説明する。

20

#### 【0056】

前処理部152は、文1「馬は人参が好きです。」に対して、形態素解析を実行することで、複数の単語に分解する。分解した各単語には、「△(スペース)」を付与する。たとえば、文1「馬は人参が好きです。」は、「馬△」、「は△」、「人参△」、「が△」、「好き△」、「です△」、「。△」に分割する。

#### 【0057】

前処理部152は、分割した各単語と、単語ベクトル辞書45とを比較することで、各単語に対応する符号を特定し、単語と置き換える。たとえば、各単語「馬△」、「は△」、「人参△」、「が△」、「好き△」、「です△」、「。△」は、それぞれ、「C1」、「C2」、「C3」、「C4」、「C5」、「C6」、「C7」に置き換えられる。

30

#### 【0058】

図10の説明に移行する。前処理部152は、単語ベクトル辞書45と、各符号とを基にして、符号に割り当てられた単語ベクトル(1)～(7)を特定する。たとえば、符号「C1」の単語ベクトル(1)～(7)は、 $wv1-1 \sim 1-7$ とする。符号「C2」の単語ベクトル(1)～(7)は、 $wv2-1 \sim 2-7$ とする。符号「C3」の単語ベクトル(1)～(7)は、 $wv3-1 \sim 3-7$ とする。

#### 【0059】

符号「C4」の単語ベクトル(1)～(7)は、 $wv4-1 \sim 4-7$ とする。符号「C5」の単語ベクトル(1)～(7)は、 $wv5-1 \sim 5-7$ とする。符号「C6」の単語ベクトル(1)～(7)は、 $wv6-1 \sim 6-7$ とする。符号「C7」の単語ベクトル(1)～(7)は、 $wv7-1 \sim 7-7$ とする。

40

#### 【0060】

前処理部152は、要素毎に単語ベクトルを積算することで、文1の文ベクトルSV1を算出する。たとえば、前処理部152は、各単語ベクトル(1)となる $wv1-1 \sim 7-1$ を積算することで、文ベクトルSV1の第1成分「SV1-1」を算出する。前処理部152は、各単語ベクトル(2)となる $wv1-2 \sim 7-2$ を積算することで、文ベクトルSV1の第2成分「SV1-2」を算出する。各単語ベクトル(3)となる $wv1-3 \sim 7-3$ を積算することで、文ベクトルSV1の第3成分「SV1-3」を算出する。

50

## 【0061】

前処理部152は、各単語ベクトル(4)となる $wv1-4 \sim 7-4$ を積算することで、文ベクトルSV1の第4成分「SV1-4」を算出する。前処理部152は、各単語ベクトル(5)となる $wv1-5 \sim 7-5$ を積算することで、文ベクトルSV1の第5成分「SV1-5」を算出する。前処理部152は、各単語ベクトル(6)となる $wv1-6 \sim 7-6$ を積算することで、文ベクトルSV1の第6成分「SV1-6」を算出する。前処理部152は、各単語ベクトル(7)となる $wv1-7 \sim 7-7$ を積算することで、文ベクトルSV1の第7成分「SV1-7」を算出する。

## 【0062】

前処理部152は、テキストデータ40に含まれる各文についても、上記処理を繰り返し実行することで、各文の文ベクトルを算出する。

10

## 【0063】

続いて、前処理部152が実行する文ベクトル辞書50を生成する処理について説明する。前処理部152が文ベクトル辞書50を生成する処理は、図1、図2で説明した処理に対応する。前処理部152は、上記処理により算出した文ベクトルを基にしてクラスタリングを実行することで、各文を複数のクラスタに分類する。

## 【0064】

前処理部152は、各クラスタについて、クラスタデータを生成し、クラスタIDとクラスタデータとを対応付けて、クラスタテーブル41に登録する。ここでは、図7で説明したクラスタCr1(クラスタID「Cr1」)のクラスタデータを生成する場合について説明する。前処理部152は、特徴量空間上において、クラスタCr1に属する文ベクトルをテーブル5dに設定する。

20

## 【0065】

前処理部152は、テーブル5dに設定した各文ベクトルの各成分の中心を算出することで、代表ベクトル(1)～(7)を算出し、テーブル5cに設定する。前処理部152は、特徴量空間上において、クラスタCr1に属する全ての文ベクトルを包含する最小の球を設定する。前処理部152は、設定した球の径を、領域5dに登録する。

## 【0066】

前処理部152は、他のクラスタについても上記処理を繰り返し実行することで、各クラスタのクラスタデータを生成し、クラスタテーブル41に登録する。

30

## 【0067】

前処理部152は、クラスタテーブル41から、クラスタIDと、代表ベクトル(1)～(7)と、径とを抽出し、文ベクトル辞書50に登録する。

## 【0068】

続いて、前処理部152が実行する文転置インデックスTSを生成する処理について説明する。前処理部152が文転置インデックスTSを生成する処理は、図3で説明した処理に対応する。前処理部152は、DB141に登録する入力テキスト2を受け付けると、入力テキスト2に含まれる各文と、文ベクトル辞書50とを基にして、文転置インデックスTSを生成する。

## 【0069】

図3の文2B「ポニーは人参が好きだ。」を用いて説明を行う。前処理部152は、文2Bに含まれる単語の単語ベクトルを積算することで、文ベクトル「SV2B」を算出する。前処理部152は、上記で説明した文ベクトルを算出する処理と同様にして、文2Bの文ベクトル「SV2B」を算出する。前処理部152は、文2Bの文ベクトルと、文ベクトル辞書50の代表ベクトルとを比較して、文2Bが属するクラスタIDを特定する。

40

## 【0070】

たとえば、前処理部152は、文ベクトル「SV2B」と代表ベクトルとの類似度(コサイン類似度、あるいは距離)を算出し、文ベクトル「SV2B」との類似度が最大となる代表ベクトルを特定する。前処理部152は、特定した代表ベクトルのクラスタIDを、文2Bの属するクラスタIDとする。

50

## 【0071】

前処理部152は、文2Bの文ベクトル「SV2B」と、クラスタCr1の代表ベクトルとの類似度が最大となる場合に、文2Bの属するクラスタIDを「Cr1」と特定する。

## 【0072】

図11および図12は、クラスタIDを特定するための処理を説明するための図である。図11において、特徴量空間1には、クラスタCr1, Cr2, Cr3が含まれる。たとえば、図11に示すように、前処理部152は、文2Bの文ベクトルSV2Bが、クラスタCr1の領域内に含まれる場合には、文2Bの属するクラスタIDを「Cr1」と特定する。

## 【0073】

図12において、特徴量空間1には、クラスタCr1, Cr2, Cr3が含まれる。たとえば、図12に示すように、前処理部152は、文2Bの文ベクトルSV2Bが、クラスタCr1, Cr2, Cr3の領域内に含まれない場合には、次の処理を実行する。前処理部152は、文ベクトルSV2BとクラスタCr1の表面との距離l1、文ベクトルSV2BとクラスタCr2の表面との距離l2、文ベクトルSV2BとクラスタCr3の表面との距離l3をそれぞれ算出する。

## 【0074】

前処理部152は、距離l1～l3のうち、最も短い距離に対応するクラスタのクラスタIDを、文2Bの属するクラスタIDとして特定する。たとえば、前処理部152は、距離l1～l3のうち、距離l2が最も短い距離の場合には、文2Bの属するクラスタIDを「Cr2」とする。

## 【0075】

前処理部152は、DB141に、入力テキスト2を登録する。前処理部152は、クラスタIDと、レコードポイントと、位置ポイントとを対応付けて、文転置インデックスTSに登録する。

## 【0076】

たとえば、文2Bの文ベクトル「SV2B」がクラスタCr1に属する場合には、文転置インデックスTSにおいて、レコードポイント(1)には、DB141に登録された入力テキスト2の先頭位置(DB141の先頭からの位置)が設定される。位置ポイント(1)には、DB141に登録された文2Bの位置(レコードポイントからの相対的な位置)が設定される。前処理部152は、同一のクラスタIDに対応するレコードポイント(m)、位置ポイント(m)に、既に他の情報が登録されている場合には、レコードポイント(m+1)、位置ポイント(m+1)に、上記の情報を登録する。mは、自然数に対応する。

## 【0077】

前処理部152は、文2A、文2Cについても、文2Bと同様にして、文ベクトルを基にして、クラスタIDを特定し、文転置インデックスTSに対して、クラスタIDと、レコードポイントと、位置ポイントとを対応付けて登録する。

## 【0078】

前処理部152は、DB141のレコードに登録する他の入力テキストを取得した場合も、上記処理を繰り返し実行し、文転置インデックスTSを更新する。また、前処理部152は、文2Bと同様にして、テキストデータ40の各文を、DB141のレコードに登録し、文転置インデックスTSを更新してもよい。

## 【0079】

図5の説明に戻る。検索部153は、検索クエリを受け付けた場合に、検索クエリと、文転置インデックスTSとを基にして、検索クエリに対応する文を、DB141の各レコードから検索する。検索部153の処理は、図4で説明した処理に対応する。

## 【0080】

検索部153が、検索クエリ4「ポニーは人参が好きだ。」を取得した場合の処理について説明する。検索部153は、検索クエリ4に含まれる単語の単語ベクトルを積算する

10

20

30

40

50

ことで、文ベクトル「Q V 4」を算出する。検索部 1 5 3 が、検索クエリ 4 の文ベクトルを算出する処理は、前処理部 1 5 2 が文ベクトルを算出する処理と同様である。

【0081】

検索部 1 5 3 は、検索クエリ 4 の文ベクトル「Q V 4」と、文ベクトル辞書 5 0 の代表ベクトルとを比較して、検索クエリ 4 が属するクラスタ ID を特定する。

【0082】

たとえば、検索部 1 5 3 は、検索クエリ 4 の文ベクトル「Q V 4」と代表ベクトルとの類似度（コサイン類似度、あるいは距離）を算出する。検索部 1 5 3 は、文ベクトル「Q V 4」との類似度が最大となる代表ベクトルを特定し、特定した代表ベクトルのクラスタ ID を、検索クエリ 4 の属するクラスタ ID とする。ここでは、検索クエリ 4 の属するクラスタ ID を「C r 1」として説明を行う。

10

【0083】

検索部 1 5 3 は、検索クエリ 4 の属するクラスタ ID 「C r 1」と、文転置インデックス T S とを比較して、検索クエリ 4 に対応する文を D B 1 4 1 から検索する。たとえば、検索部 1 5 3 は、クラスタ ID 「C r 1」の行に設定されたレコードポインタ（1）、位置ポインタ（1）に示される、文「ポニー人参が好きだ。」を抽出する。

【0084】

検索部 1 5 3 は、クラスタ ID 「C r 1」の行に設定されたレコードポインタ（2）、位置ポインタ（2）に示される、文「馬は人参が好きだ。」を抽出する。検索部 1 5 3 は、クラスタ ID 「C r 1」の行に設定されたレコードポインタ（3）、位置ポインタ（3）に示される、文「Hoses likes carrots」を抽出する。

20

【0085】

説明を省略するが、検索部 1 5 3 は、クラスタ ID 「C r 1」の行に設定されたレコードポインタ（m）、位置ポインタ（m）に示される文を抽出する。

【0086】

検索部 1 5 3 は、検索結果を、表示部 1 3 0 に出力して表示させる。検索部 1 5 3 は、検索クエリ 4 を受け付ける度に、上記処理を繰り返し実行する。

【0087】

次に、情報処理装置 1 0 0 の処理手順の一例について説明する。図 1 3 は、準備処理の処理手順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 1 0 0 の取得部 1 5 1 は、テキストデータ 4 0 を取得し、記憶部 1 4 0 に登録する（ステップ S 1 0 1）。

30

【0088】

情報処理装置 1 0 0 の前処理部 1 5 2 は、テキストデータ 4 0 に含まれる各文の文ベクトルを算出する（ステップ S 1 0 2）。前処理部 1 5 2 は、複数の文ベクトルに対してクラスタリングを実行する（ステップ S 1 0 3）。

【0089】

前処理部 1 5 2 は、クラスタリング結果を基にして、クラスタテーブル 4 1 を生成する（ステップ S 1 0 4）。前処理部 1 5 2 は、クラスタテーブル 4 1 を基にして、文ベクトル辞書を生成する（ステップ S 1 0 5）。

40

【0090】

前処理部 1 5 2 は、D B 1 4 1 に登録する入力テキストを取得する（ステップ S 1 0 6）。前処理部 1 5 2 は、入力テキストの文ベクトルを算出する（ステップ S 1 0 7）。前処理部 1 5 2 は、入力テキストの文ベクトルと、文ベクトル辞書 5 0 とを基にして、入力テキストの文に対応するクラスタ ID を特定する（ステップ S 1 0 8）。

【0091】

前処理部 1 5 2 は、入力テキストを D B 1 4 1 のレコードに登録する（ステップ S 1 0 9）。前処理部 1 5 2 は、クラスタ ID と、レコードポインタと、位置ポインタとを対応付けて、文転置インデックス T S に登録する（ステップ S 1 1 0）。

【0092】

50

前処理部 152 は、処理を継続する場合には（ステップ S 111, Yes）、ステップ S 106 に移行する。一方、前処理部 152 は、処理を継続しない場合には（ステップ S 111, No）、処理を終了する。

【0093】

図 14 は、検索処理の処理手順を示すフローチャートである。図 14 に示すように、情報処理装置 100 の取得部 151 は、検索クエリを取得する（ステップ S 201）。

【0094】

情報処理装置 100 の検索部 153 は、検索クエリの文ベクトルを算出する（ステップ S 202）。検索部 153 は、検索クエリの文ベクトルと、文ベクトル辞書 50 とを基にして、検索クエリのクラスタ ID を特定する（ステップ S 203）。

【0095】

検索部 153 は、検索クエリのクラスタ ID と、文転置インデックスとを基にして、DB 141 のレコードから文を検索する（ステップ S 204）。検索部 153 は、検索結果を表示部 130 に出力する（ステップ S 205）。

【0096】

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、テキストデータ 40 に含まれる複数の文の文ベクトルを算出し、クラスタリングを実行することで、複数のクラスタに関する情報を生成し、文ベクトル辞書 50 を生成する。文ベクトル辞書 50 では、クラスタ ID と代表ベクトルおよび径とを対応付けている。たとえば、情報処理装置 100 は、新規の入力テキストを DB 141 のレコードに登録する場合に、入力テキストの文ベクトルと、文ベクトル辞書 50 の代表ベクトルおよび径とを比較することで、入力テキストの文ベクトルの属するクラスタ ID を容易に割り当てることができる。

【0097】

情報処理装置 100 は、入力テキスト 2 を DB 141 のレコードに登録する場合に、文ベクトル辞書 50 を用いて、入力テキストの文がいずれのクラスタ ID に属するかを特定し、特定したクラスタ ID と、格納されたレコードと文の位置情報とを対応付けることで、文転置インデックス TS を生成する。このように、文ベクトル辞書 50 を用いることで、文とクラスタ ID とを容易に紐づけることができるため、文転置インデックス TS に関する登録処理の負荷を軽減することができる。

【0098】

情報処理装置 100 は、検索クエリを受け付けた場合に、検索クエリと文ベクトル辞書とを基にして、検索クエリに対応するクラスタ ID を特定し、特定したクラスタ ID と、文転置インデックス TS とを基にして、検索クエリに対応する文の位置を特定する。これによって、検索クエリに対応する文を効率的に検索することができる。たとえば、クラスタ ID に対して、複数の位置情報（レコードポインタ、位置ポインタ）が対応付けられている場合には、一度に、複数の位置情報を特定し、検索クエリに対応する文を効率的に検索することができる。

【実施例 2】

【0099】

上述した実施例 1 では、文ベクトルの粒度でクラスタリングを実行し、文転置インデックスを生成していたがこれに限定されるものではなく、単語ベクトルの粒度で転置インデックス（後述する第 2 単語転置インデックス WS 2 に相当）を生成してもよい。

【0100】

図 15 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 15 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。

【0101】

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。

10

20

30

40

50

## 【0102】

記憶部240は、テキストデータ40と、クラスタテーブル41と、第1単語ベクトル辞書45Aと、第2単語ベクトル辞書45Bと、文ベクトル辞書50と、DB141と、文転置インデックスTSとを有する。記憶部240は、第1単語転置インデックスWS1と、第2単語転置インデックスWS2とを有する。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。

## 【0103】

テキストデータ40は、文ベクトル辞書50を生成するために用いられる情報である。テキストデータ40に関する説明は、実施例1と同様である。

10

## 【0104】

クラスタテーブル41は、クラスタに関する情報を保持する。クラスタテーブル41は、クラスタIDと、クラスタデータとを対応付ける。クラスタテーブル41は、文ベクトルの粒度でクラスタリングした結果に基づくクラスタデータと、単語ベクトルの粒度でクラスタリングした結果に基づくクラスタデータとを有する。

## 【0105】

第1単語ベクトル辞書45Aは、単語に割り当てられた符号、単語ベクトルを定義するテーブルである。第1単語ベクトル辞書45Aは、図8で説明した単語ベクトル辞書45に対応する。

## 【0106】

第2単語ベクトル辞書45Bは、単語ベクトルの粒度でクラスタリングした結果に基づく、クラスタIDと、代表ベクトルと、径とを対応付ける。

20

## 【0107】

文ベクトル辞書50は、文ベクトルの粒度でクラスタリングした結果に基づく、クラスタIDと、代表ベクトルと、径とを対応付ける。文ベクトル辞書50に関する説明は、実施例1と同様である。

## 【0108】

DB141のレコードには、入力テキスト等を登録する。

## 【0109】

文転置インデックスTSは、文ベクトルが属するクラスタIDと、レコードポインタおよび位置ポインタの組とを対応付ける。文転置インデックスTSの説明は、実施例1と同様である。

30

## 【0110】

第1単語転置インデックスWS1は、単語と、位置情報（レコードポインタおよび位置ポインタ）とを対応付ける。

## 【0111】

第2単語転置インデックスWS2は、単語と、位置情報（レコードポインタおよび位置ポインタ）とを対応付ける。図16は、第2単語転置インデックスの一例を示す図である。レコードポインタは、該当するクラスタに属する単語を有するテキストが登録されたDB141のレコード上の位置であって、DB141の先頭から位置を示す。位置ポインタは、該当するクラスタに属する単語が登録されたDB141のレコード上の位置であって、当該レコードの先頭からの相対的な位置を示す。

40

## 【0112】

たとえば、単語「人参」の単語ベクトルがクラスタwCr1に属する場合には、第2単語転置インデックスWS2において、レコードポインタ（1）には、DB141のレコードに登録された入力テキスト2の先頭位置（DB141の先頭からの位置）が設定される。位置ポインタ（1）には、DB141に登録された単語「人参」の位置（当該レコードの先頭からの相対的な位置）が設定される。

## 【0113】

図15の説明に戻る。制御部250は、取得部251と、前処理部252と、検索部2

50

53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150は、例えばASICやFPGA等の集積回路により実行されてもよい。

【0114】

取得部251は、通信部210、入力部220を介して、各種のデータを取得する。たとえば、取得部151は、テキストデータ40を取得し、記憶部240に登録する。

【0115】

前処理部252は、テキストデータを基にして、文転置インデックスTS、第1単語インデックスWS1、第2単語転置インデックスWS2を生成する。前処理部252が、文転置インデックスTSを生成する処理は、実施例1と同様である。

【0116】

前処理部252が、第1単語転置インデックスWS1を生成する処理について説明する。前処理部252は、テキストデータ40に含まれる文を抽出し、形態素解析を実行することで、複数の単語に分割する。前処理部252は、単語と、単語に対応する位置情報（レコードポイントおよび位置ポイント）とを対応付けて、第1単語転置インデックスWS1に登録する。位置情報は、DB141に登録されたテキストデータ40に基づく位置情報とする。前処理部252は、テキストデータ40に含まれる各単語について、上記処理を繰り返し実行する。

【0117】

前処理部252が、第2単語転置インデックスWS2を生成する処理について説明する。前処理部252は、テキストデータ40に含まれる文を抽出し、形態素解析を実行することで、複数の単語に分割する。前処理部252は、分割した各単語と、第1単語ベクトル辞書45Aとを比較することで、各単語に対応する符号を特定し、単語と置き換える。前処理部252は、第1単語ベクトル辞書45Aと、各符号とを基にして、符号に割り当てられた単語ベクトル(1)～(7)を特定する。

【0118】

前処理部252は、単語ベクトルの粒度で、クラスタリングを実行し、各クラスタについて、クラスタデータを生成し、クラスタIDとクラスタデータとを対応付けて、クラスタテーブル41に登録する。前処理部252は、単語ベクトルの粒度で行ったクラスタ結果と、実施例1で説明した文ベクトルの粒度で行ったクラスタ結果とを区別して、クラスタテーブル41に登録する。

【0119】

単語ベクトルの粒度で実行したクラスタデータには、クラスタID、代表ベクトル(1)～(7)、径が含まれる。単語ベクトルの粒度で求められる代表ベクトル(1)～(7)は、各単語ベクトルの各成分の中心を算出することで求められる。前処理部252は、クラスタIDと、代表ベクトル(1)～(7)、径を、第2単語ベクトル辞書45Bに登録する。

【0120】

前処理部252は、DB141に登録する入力テキスト2を受け付けると、入力テキスト2に含まれる各単語と、第1単語ベクトル辞書45A、第2単語ベクトル辞書45Bとを基にして、第2単語転置インデックスWS2を生成する。

【0121】

前処理部252は、入力テキスト2に含まれる単語を選択し、選択した単語と、第1単語ベクトル辞書45Aの単語ベクトルとを比較して、単語ベクトルを特定する。前処理部252は、単語ベクトルと、第2単語ベクトル辞書45Bとを比較して、単語ベクトルが属するクラスタIDを特定する。

【0122】

前処理部252は、DB141に、入力テキスト2に登録する。前処理部252は、単語ベクトルの属するクラスタIDと、レコードポイントと、位置ポイントとを対応付けて、第2単語転置インデックスWS2に登録する。

【0123】

10

20

30

40

50

検索部 253 は、検索クエリを受け付けた場合に、検索クエリと、文転置インデックス TS、第 2 単語転置インデックス WS2 とを基にして、検索クエリに対応する文または単語を、DB141 から検索する。検索クエリが「文」で指定されている場合には、実施例 1 で説明した検索部 153 と同様の処理を実行する。

#### 【0124】

検索部 253 は、検索クエリが「単語」で指定されている場合には、検索クエリの単語と、第 1 単語ベクトル辞書 45A とを基にして単語ベクトルを特定し、単語ベクトルと第 2 単語ベクトル辞書 45B とを基にして、単語の属するクラスター ID を特定する。検索部 253 は、検索クエリの単語の属するクラスター ID と、第 2 単語転置インデックス WS2 とを比較して、検索クエリに対応する単語（あるいは、単語を含む文）を DB141 から検索する。検索部 253 は、検索結果を、表示部 230 に出力して表示する。

10

#### 【0125】

検索部 253 は、検索クエリが「文」で設定されている場合でも、文を複数の単語に分割し、各単語について、上記処理を実行することで、検索を行ってもよい。

#### 【0126】

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、テキストデータ 40 に含まれる複数の単語の単語ベクトルを算出し、クラスタリングを実行することで、複数のクラスターに関する情報を生成し、第 2 単語ベクトル辞書 45B に単語の属するクラスター ID を設定する。たとえば、情報処理装置 200 は、新規の入力テキストを DB141 に登録する場合に、入力テキストの単語ベクトルと、第 2 単語ベクトル辞書 45B とを比較することで、入力テキストの単語ベクトルの属するクラスター ID を容易に割り当てることができる。

20

#### 【0127】

ところで、上記実施例で説明した文転置インデックス TS をその他の処理に適用することで、その他の処理で要する演算コスト等を削減することができる。

#### 【0128】

図 17 は、文転置インデックス TS のその他の適用例を説明するための図である。たとえば、日本語文を英文に翻訳する学習モデル M1 の機械学習を行う場合、教師データ T1 を用いる。学習モデル M1 は、CNN (Convolutional Neural Network) や RNN (Recurrent Neural Network)、オートエンコーダ (Autoencoder) 等に対応する。教師データ T1 は、日本語文を入力データとし、英文を正解データとする教師データである。たとえば、文ベクトル辞書 50 または、第 2 単語ベクトル辞書 45B を用いて、教師データ T1 の日本語文および英文をクラスター ID に変換し、クラスター ID を用いて、機械学習を実行してもよい。これにより、類似する文同士（同一のクラスターに属する文）には、同一のクラスター ID が割り振られ、文の種類が削減（縮退）するため、機械学習において、SoftMax 関数などの演算量を大幅に削減することができる。

30

#### 【0129】

なお、実施例 1、および実施例 2 において、文字列で構成されるテキストを例として、説明したが、プログラムや有機化合物、ゲノムなどでも良い。単語に相当し、意味を持つ、命令・変数（プログラム）や官能基（有機化合物）、タンパク質基本構造 kmer（ゲノム）などであっても良い。また、文に相当し、行（プログラム）や官能基一次構造（有機化合物）、タンパク質一次構造（ゲノム）などであっても良い。

40

#### 【0130】

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 18 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。

#### 【0131】

図 18 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 CPU 301 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

50

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 RAM 306 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。

【0132】

ハードディスク装置 307 は、取得プログラム 307a、前処理プログラム 307b、検索プログラム 307c を有する。また、CPU 301 は、各プログラム 307a ～ 307c を読み出して RAM 306 に展開する。

【0133】

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

10

【0134】

取得プロセス 306a は、取得部 151、252 の処理に対応する。前処理プロセス 306b の処理は、前処理部 152 (252) の処理に対応する。検索プロセス 306c の処理は、検索部 153 (253) の処理に対応する。

【0135】

なお、各プログラム 307a ～ 307c 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 ～ 307c を読み出して実行するようにしてもよい。

20

【符号の説明】

【0136】

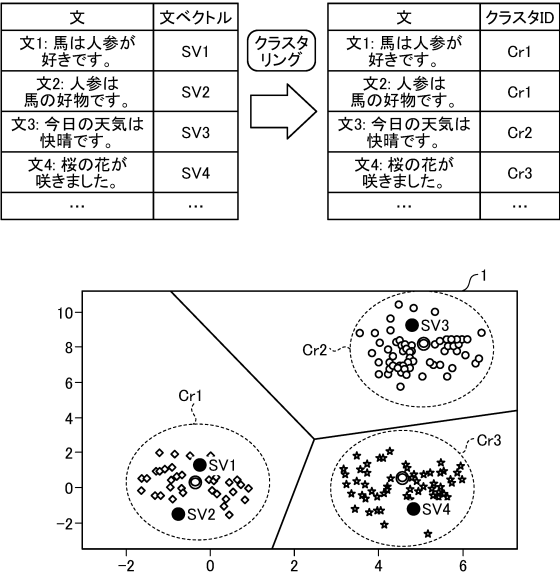
|          |        |
|----------|--------|
| 100, 200 | 情報処理装置 |
| 110, 210 | 通信部    |
| 120, 220 | 入力部    |
| 130, 230 | 表示部    |
| 140, 240 | 記憶部    |
| 141 DB   |        |
| 150, 250 | 制御部    |
| 151, 251 | 取得部    |
| 152, 252 | 前処理部   |
| 153, 253 | 検索部    |

30

40

50

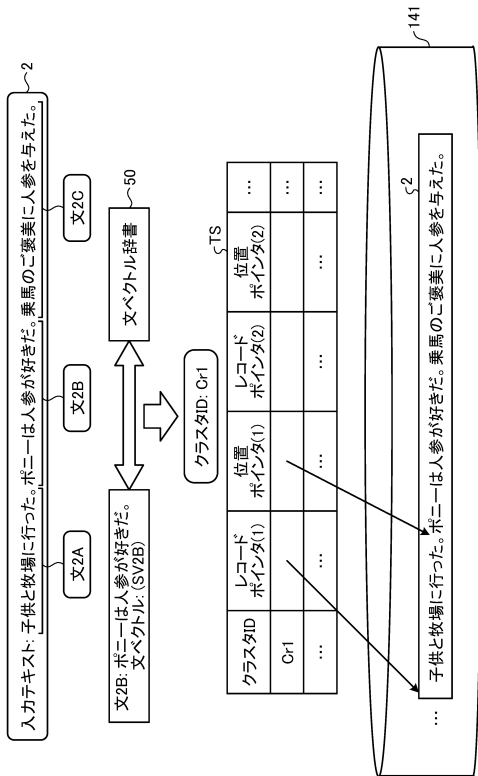
【図面】  
【図 1】



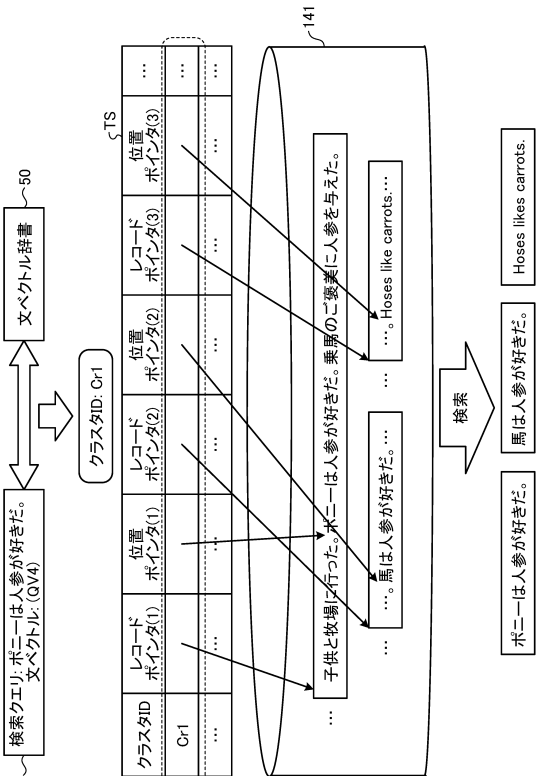
【図 2】

| クラスタ<br>ID | 代表ベクトル        |               |               |               |               |               |               | 径      |
|------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|--------|
|            | 代表<br>ベクトル(1) | 代表<br>ベクトル(2) | 代表<br>ベクトル(3) | 代表<br>ベクトル(4) | 代表<br>ベクトル(5) | 代表<br>ベクトル(6) | 代表<br>ベクトル(7) |        |
| Cr1        | 0.011         | 0.012         | 0.013         | 0.014         | 0.015         | 0.016         | 0.017         | 0.0152 |
| Cr2        | 0.021         | 0.022         | 0.023         | 0.024         | 0.025         | 0.026         | 0.027         | ...    |
| Cr3        | 0.031         | 0.032         | 0.033         | 0.034         | 0.034         | 0.036         | 0.037         | ...    |
| Cr4        | 0.041         | 0.042         | 0.043         | 0.044         | 0.045         | 0.046         | 0.047         | ...    |
| ...        | ...           | ...           | ...           | ...           | ...           | ...           | ...           | ...    |

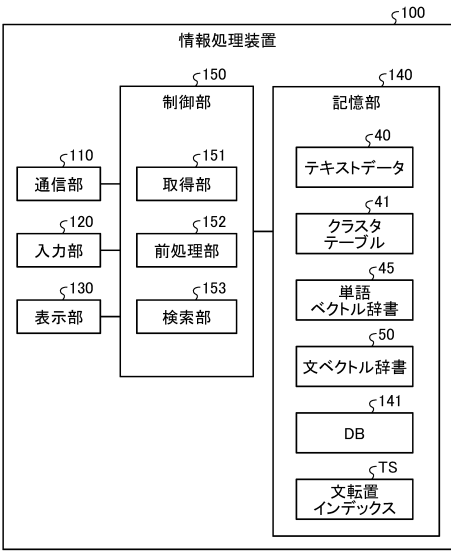
【図 3】



【図 4】



【図 5】



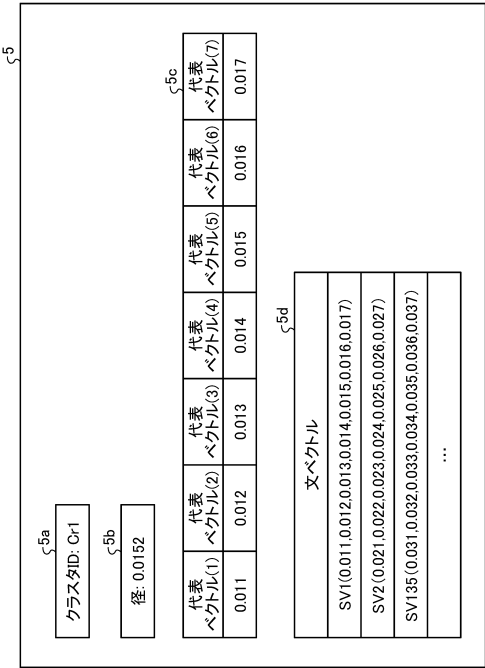
【図 6】

| クラスターデータ 41 |                       |
|-------------|-----------------------|
| クラスターID     | クラスターデータ              |
| Cr1         | クラスターID「Cr1」のクラスターデータ |
| Cr2         | クラスターID「Cr2」のクラスターデータ |
| Cr3         | クラスターID「Cr3」のクラスターデータ |
| Cr4         | クラスターID「Cr4」のクラスターデータ |
| ...         | ...                   |

10

20

【図 7】



【図 8】

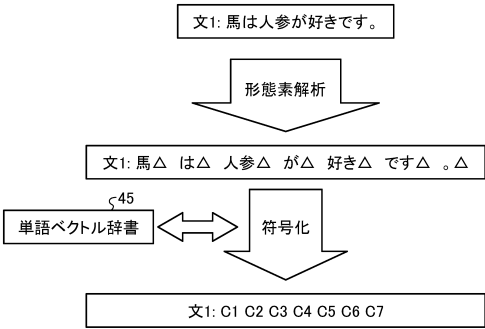
| 符号  | 単語  | 単語ベクトル(1) | 単語ベクトル(2) | 単語ベクトル(3) | 単語ベクトル(4) | 単語ベクトル(5) | 単語ベクトル(6) | 単語ベクトル(7) |
|-----|-----|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| C1  | 馬   | ww1-1     | ww1-2     | ww1-3     | ww1-4     | ww1-5     | ww1-6     | ww1-7     |
| C2  | は   | ww2-1     | ww2-2     | ww2-3     | ww2-4     | ww2-5     | ww2-6     | ww2-7     |
| C3  | 人參  | ww3-1     | ww3-2     | ww3-3     | ww3-4     | ww3-5     | ww3-6     | ww3-7     |
| C4  | が   | ww4-1     | ww4-2     | ww4-3     | ww4-4     | ww4-5     | ww4-6     | ww4-7     |
| C5  | 好き  | ww5-1     | ww5-2     | ww5-3     | ww5-4     | ww5-5     | ww5-6     | ww5-7     |
| C6  | です  | ww6-1     | ww6-2     | ww6-3     | ww6-4     | ww6-5     | ww6-6     | ww6-7     |
| C7  | 。   | ww7-1     | ww7-2     | ww7-3     | ww7-4     | ww7-5     | ww7-6     | ww7-7     |
| C8  | の   | ww8-1     | ww8-2     | ww8-3     | ww8-4     | ww8-5     | ww8-6     | ww8-7     |
| C9  | 好物  | ww9-1     | ww9-2     | ww9-3     | ww9-4     | ww9-5     | ww9-6     | ww9-7     |
| C10 | 今日  | ww10-1    | ww10-2    | ww10-3    | ww10-4    | ww10-5    | ww10-6    | ww10-7    |
| C11 | 天気  | ww11-1    | ww11-2    | ww11-3    | ww11-4    | ww11-5    | ww11-6    | ww11-7    |
| C12 | 快晴  | ww12-1    | ww12-2    | ww12-3    | ww12-4    | ww12-5    | ww12-6    | ww12-7    |
| ... | ... | ...       | ...       | ...       | ...       | ...       | ...       | ...       |

30

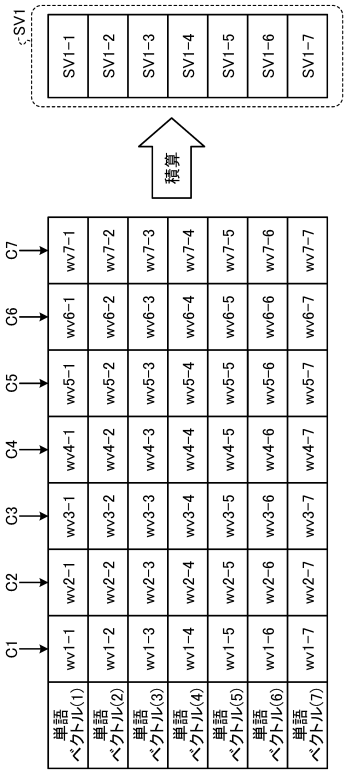
40

50

【図 9】



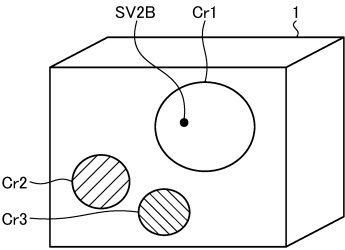
【図 10】



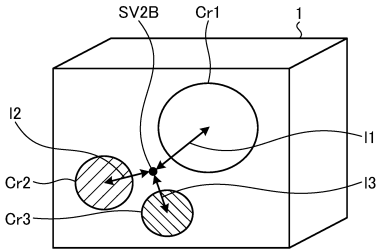
10

20

【図 11】



【図 12】

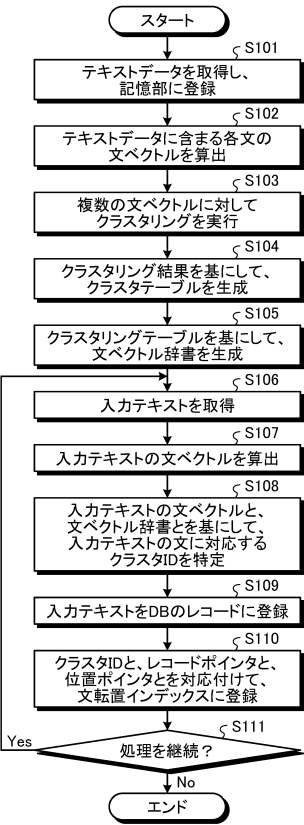


30

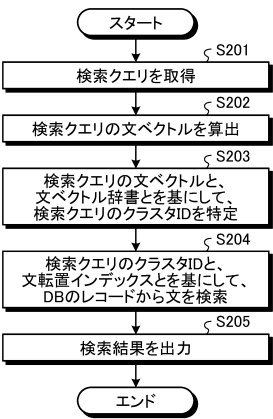
40

50

【図 1 3】



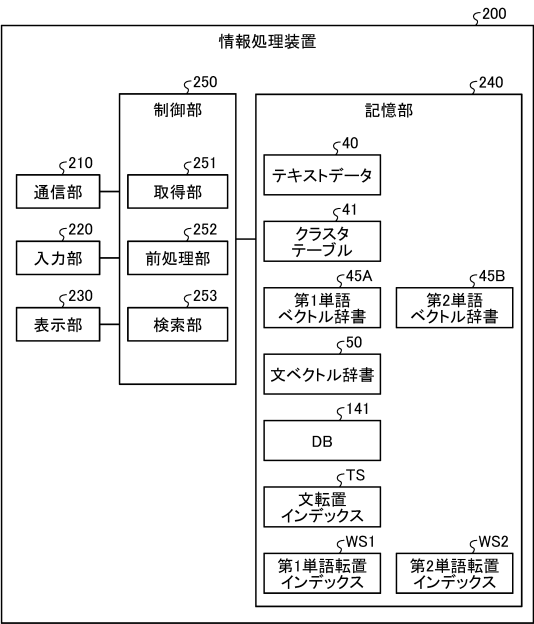
【図 1 4】



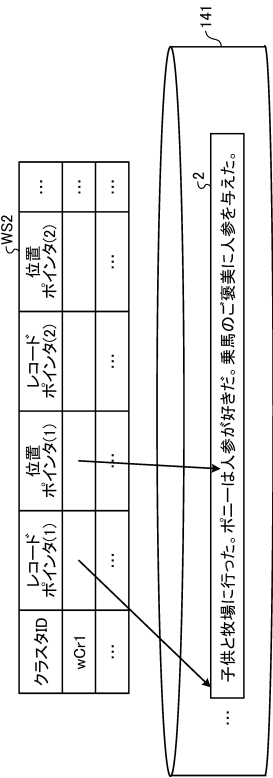
10

20

【図 1 5】



【図 1 6】

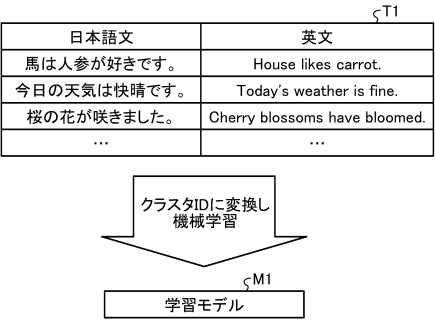


30

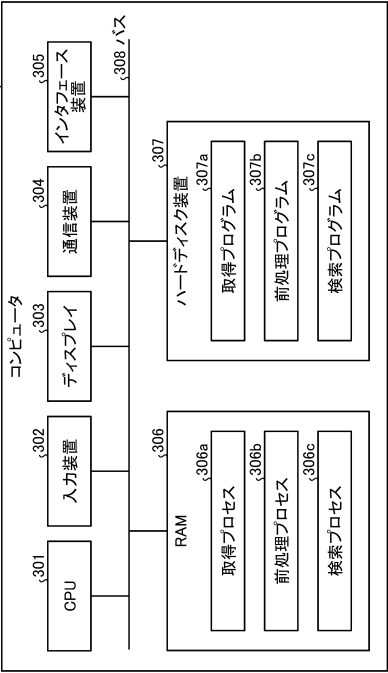
40

50

【図 17】



【図 18】



10

20

30

40

50

---

フロントページの続き

(72)発明者 尾上 聡

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

審査官 齊藤 貴孝

(56)参考文献 米国特許出願公開第2012/0209847 (US, A1)

中国特許出願公開第112256880 (CN, A)

米国特許第11036764 (US, B1)

特開2004-078446 (JP, A)

(58)調査した分野 (Int.Cl., DB名)

G06F 16/00-16/958