(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2018-142188
(P2018-142188A)
(43)公開日　平成30年9月13日(2018.9.13)

| (51)Int.Cl. | | F I | | テーマコード（参考） |
|---|---|---|---|---|
| G06F 17/30 | (2006.01) | G06F 17/30 | 210A | |
| | | G06F 17/30 | 170A | |

審査請求 未請求　請求項の数 5　ＯＬ　（全 16 頁）

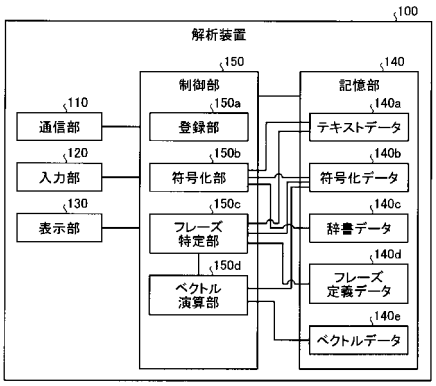
(21)出願番号　特願2017-36288(P2017-36288)
(22)出願日　平成29年2月28日(2017.2.28)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　110002147
特許業務法人酒井国際特許事務所
(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　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
(72)発明者　吉田　裕之
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

(54)【発明の名称】解析プログラム、解析方法および解析装置

(57)【要約】
【課題】解析対象の文書に対する解析速度および解析精度の向上を図る。
【解決手段】解析装置１００は、解析対象の文書を単語単位で符号化した、複数の符号化単語を生成する。解析装置１００は、符号化単語、または、複数の符号化単語の組み合わせに割り当てられた符号化フレーズに関する、解析対象の文書の出現状況に応じてそれぞれ生成されたベクトル値を、符号化単語、または、符号化フレーズに割り当てる。
【選択図】図２

本実施例1に係る解析装置の構成を示す機能ブロック図



【特許請求の範囲】
【請求項１】
コンピュータに、
解析対象の文書を単語単位で符号化した、複数の符号化単語を生成し、
前記符号化単語、または、前記複数の符号化単語の組み合わせに割り当てられた符号化フレーズに関する、前記解析対象の文書の出現状況に応じてそれぞれ生成されたベクトル値を、前記符号化単語、または、前記符号化フレーズに割り当てる
処理を実行させることを特徴とする解析プログラム。
【請求項２】
第１符号化フレーズの前後に位置する前記符号化単語または他の第２符号化フレーズに対してSkip-gramによるモデル化を行い、前記第１符号化フレーズの前後に、前記符号化単語または前記第２符号化フレーズが出現する確率を特定することで、前記第１符号化フレーズのベクトル値を算出することを特徴とする請求項１に記載の解析プログラム。
【請求項３】
フレーズを構成する第１単語と第２単語との組を定義したテーブルを基にして、前記解析対象の文書に含まれる第１単語を特定し、特定した第１単語と同じ文中に第２単語が含まれる場合には、特定した第１単語と同じ文中に含まれる第２単語を削除し、特定した第１単語の直後に前記第２単語を配置する処理を更にコンピュータに実行させることを特徴とする請求項１または２に記載の解析プログラム。
【請求項４】
コンピュータが実行する解析方法であって、
解析対象の文書を単語単位で符号化した、複数の符号化単語を生成し、
前記符号化単語、または、前記複数の符号化単語の組み合わせに割り当てられた符号化フレーズに関する、前記解析対象の文書の出現状況に応じてそれぞれ生成されたベクトル値を、前記符号化単語、または、前記符号化フレーズに割り当てる
処理を実行することを特徴とする解析方法。
【請求項５】
解析対象の文書を単語単位で符号化した、複数の符号化単語を生成する符号化部と、
前記符号化単語、または、前記複数の符号化単語の組み合わせに割り当てられた符号化フレーズに関する、前記解析対象の文書の出現状況に応じてそれぞれ生成されたベクトル値を、前記符号化単語、または、前記符号化フレーズに割り当てる演算部と
を有することを特徴とする解析装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、解析プログラム等に関する。
【背景技術】
【０００２】
従来、文書を分散表現する手法として解析対象の文書を構成する形態素それぞれに基づいて、文書からベクトルを生成するWord2Vec技術が存在する。例えば、Word2Vec技術では、ある単語（形態素）と、ある単語に隣接する他の単語との関係に基づいて、各単語のベクトル値を算出する処理を行う。
【０００３】
ここで、Word2Vec技術等により文書をベクトルを用いて分散表現する場合に、解析対象となる文書に含まれる「the」、「a」等の冠詞、「on」、「of」等の前置詞等の高頻度の単語の影響が過大となる。このため、Word2Vec技術では、高頻度の単語をストップワードとして文書から排除した後に、ベクトルによる分散表現を生成する。
【０００４】
例えば、解析対象の文書「He takes care of his daughter」をWord2Vec技術では、ストップワードとなる「of」を除外した後、「He takes care his daughter」に含

まれる単語それぞれをベクトル化する。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２００６－４８６８５号公報
【特許文献２】特開２００９－１５１７５７号公報
【非特許文献】
【０００６】
【非特許文献１】Distributed Representations of Words and Phrases and their Compositionality,Tomas Mikolov et. al, pp. 3111-3119, Advances in Neural Information Processing Systems 26,2013,Curran Associates,Inc.
【発明の概要】
【発明が解決しようとする課題】
【０００７】
しかしながら、上述した従来技術では、解析対象の文書に対する解析速度および解析精度が低いという問題がある。
【０００８】
たとえば、Word2Vec技術により、ストップワードとして除外される「the」、「a」等の冠詞、「on」、「of」等の前置詞等は、特定の文字列において存在の有無により意味が変化する場合がある。具体的には、「take care of」の「of」、「the Japanese」の「the」は、存在の有無により意味が変化するため、かかる「of」、「the」等を除外してベクトル化を行うと、本来の文書の意味が変わった状態のベクトル化がなされることから、生成されたベクトルを用いた解析の精度が低下する場合がある。
【０００９】
また、従来のWord2Vec技術で用いられる解析において、ストップワードを含めて計算量が過大とならない解析手法は知られておらず、適切な計算時間により、目的の精度を得ることができない。
【００１０】
１つの側面では、本発明は、解析対象の文書に対する解析速度および解析精度の向上を図ることができる解析プログラム、解析方法および解析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】
【００１１】
第１の案では、コンピュータに下記の処理を実行させる。コンピュータは、解析対象の文書を単語単位で符号化した、複数の符号化単語を生成する。コンピュータは、符号化単語、または、複数の符号化単語の組み合わせに割り当てられた符号化フレーズに関する、解析対象の文書の出現状況に応じてそれぞれ生成されたベクトル値を、符号化単語、または、符号化フレーズに割り当てる。
【発明の効果】
【００１２】
符号化したフレーズに対しても、値を割り当てたベクトルを生成することにより、解析対象の文書に対する解析速度および解析精度の向上を図ることができる。
【図面の簡単な説明】
【００１３】
【図１】図１は、本実施例１に係る解析装置の処理の一例を説明するための図である。
【図２】図２は、本実施例１に係る解析装置の構成を示す機能ブロック図である。
【図３】図３は、フレーズ定義データのデータ構造の一例を示す図である。
【図４】図４は、本実施例１に係る解析装置の処理手順を示すフローチャートである。
【図５】図５は、本実施例２に係る解析装置の処理の一例を説明するための図である。
【図６】図６は、本実施例２に係る解析装置の構成を示す機能ブロック図である。

【図７】図７は、離間フレーズテーブルのデータ構造の一例を示す図である。
【図８】図８は、本実施例２に係る解析装置の処理手順を示すフローチャートである。
【図９】図９は、オートマトンを用いたベクトル演算の一例を説明するための図である。
【図１０】図１０は、解析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】
【００１４】
以下に、本願の開示する解析プログラム、解析方法および解析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】
【００１５】
図１は、本実施例１に係る解析装置の処理の一例を説明するための図である。ここでは、解析装置が、テキストデータ１０ａを符号化した符号化データ１０ｂをベクトル化する場合について説明する。図１に示す例では、テキストデータ１０ａを「Every day we take care of our dauthter」とする。
【００１６】
解析装置は、所定のフレーズを定義したフレーズ定義テーブルと、テキストデータ１０ａとを比較して、テキストデータ１０ａに含まれる各単語のうち、所定のフレーズを構成する複数の単語の組を特定する。図１に示す例では、「take care of」が所定のフレーズとして特定される。
【００１７】
解析装置は、単語とコードとを対応付けた辞書データと、テキストデータ１０ａに含まれる単語とを比較することで、テキストデータ１０ａの単語を単語単位で符号化することで、符号化データ１０ｂを生成する。例えば、解析装置は、単語「Every」をコードＡ１、単語「day」をコードＡ２、単語「we」をコードＡ３、単語「take」をコードＡ４に符号化する。解析装置は、単語「care」をコードＡ５、単語「of」をコードＡ６、単語「our」をコードＡ７、単語「daughter」をコードＡ８に符号化する。
【００１８】
解析装置は、複数のコードのうち、所定のフレーズを構成する単語のコードの組と、所定のフレーズを構成する単語に対応しないコードを特定する。以下の説明では、適宜、所定のフレーズを構成する単語のコードの組を、「符号化フレーズ」と表記する。所定のフレーズを構成する単語に対応しないコードを、「符号化単語」と表記する。図１に示す例では、コードＡ４～コードＡ６の組が、符号化フレーズ１５となる。他のコードＡ１～Ａ３、Ａ７，Ａ８は、それぞれ符号化単語となる。
【００１９】
解析装置は、符号化単語および符号化フレーズの出現状況に応じて、符号化単語および符号化フレーズのベクトル値をそれぞれ算出することで、符号化データ１０ｂをベクトル化する。
【００２０】
解析装置が、符号化単語のベクトル値を算出する処理について説明する。ベクトル値の算出対象となる符号化単語を、対象単語と表記する。解析装置は、符号化データ１０ｂ上において、対象単語の前方２つの符号化単語または符号化フレーズと、対象単語の後方２つの符号化単語に対して、Skip-gramによるモデル化を行い、対象単語の前後に符号化単語または符号化フレーズが出現する確率を特定することで、対象単語のベクトル値を算出する。
【００２１】
例えば、解析装置は、コードＡ３のベクトル値を算出する場合には、コードＡ１、コードＡ２、符号化フレーズ１５、コードＡ７に対して、Skip-gramによるモデル化を行う。解析装置は、コードＡ３の前後に、コードＡ１、コードＡ２、符号化フレーズ１５、コードＡ７が出現する確率を特定することで、コードＡ３のベクトル値を算出する。解析装置

は、コードＡ１、Ａ２、Ａ７、Ａ８についても同様の処理を実行することで、各コードのベクトル値を算出する。

【００２２】

解析装置が、符号化フレーズのベクトル値を算出する処理について説明する。ベクトル値の算出対象となる符号化フレーズを、対象フレーズと表記する。解析装置は、符号化データ１０ｂ上において、対象フレーズの前方２つの符号化単語または符号化フレーズと、対象フレーズの後方２つの符号化単語に対して、Skip-gramによるモデル化を行い、対象フレーズの前後に符号化単語または符号化フレーズが出現する確率を特定することで、対象フレーズのベクトル値を算出する。

【００２３】

例えば、解析装置は、符号化フレーズ１５のベクトル値を算出する場合には、コードＡ２、コードＡ３、コードＡ７、コードＡ８に対して、Skip-gramによるモデル化を行う。解析装置は、符号化フレーズ１５の前後に、コードＡ２、コードＡ３、コードＡ７、コードＡ８が出現する確率を特定することで、符号化フレーズ１５のベクトル値を算出する。解析装置は、他の符号化フレーズについても同様の処理を実行することで、各コードのベクトル値を算出する。

【００２４】

解析装置は、各符号単語、各符号化フレーズに対応するベクトル値を割り当てることで、符号化データ１０ｂをベクトル化する。

【００２５】

上記に記載したように、本実施例１に係る解析装置によれば、テキストデータ１０ａに含まれる単語からストップワードを除外すること無く、単語単位で符号化を行うことで、符号化データ１０ｂを生成する。解析装置は、符号化単語のうち、所定のフレーズを構成する符号化単語を符号化フレーズとしてまとめ、符号化単語および符号化フレーズの出現状況に応じて、符号化データ１０ｂをベクトル化する。このように、解析装置は、ストップワードを除外しないため、符号化データ１０ｂをベクトル化する際の解析精度を向上させることができる。また、解析装置が利用する符号化では、符号化データ１０ｂが単語単位で符号化されているため、Zipによる符号化と異なり、符号化したままで各単語を区別できるため、復号化を行うことなく、文書に含まれる単語のベクトル値を算出でき、解析速度を向上できる。

【００２６】

図２は、本実施例１に係る解析装置の構成を示す機能ブロック図である。図２に示すように、この解析装置１００は、通信部１１０と、入力部１２０と、表示部１３０と、記憶部１４０と、制御部１５０とを有する。

【００２７】

通信部１１０は、ネットワークを介して他の外部装置と通信を実行する処理部である。例えば、解析装置１００は、後述するテキストデータ１４０ａ、辞書データ１４０ｃ、フレーズ定義データ１４０ｄ等を、通信部１１０を介して、受信しても良い。

【００２８】

入力部１２０は、解析装置１００に各種の情報を入力する入力装置である。例えば、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。

【００２９】

表示部１３０は、制御部１５０から出力される各種の情報を表示する表示装置である。例えば、表示部１３０は、液晶ディスプレイやタッチパネル等に対応する。

【００３０】

記憶部１４０は、テキストデータ１４０ａと、符号化データ１４０ｂと、辞書データ１４０ｃと、フレーズ定義データ１４０ｄと、ベクトルデータ１４０ｅとを有する。記憶部１４０は、ＲＡＭ（Random Access Memory）、ＲＯＭ（Read Only Memory）、フラッシュメモリ（Flash Memory）などの半導体メモリ素子や、ＨＤＤ（Hard Disk Drive）などの記憶装置に対応する。

【００３１】
テキストデータ１４０ａは、複数の単語を含む文字列データである。テキストデータ１４０ａの一例は、図１に示したテキストデータ１０ａとなる。
【００３２】
符号化データ１４０ｂは、テキストデータ１４０ａに含まれる各単語を単語単位で符号化したデータである。符号化データ１４０ｂの一例は、図１に示した符号化データ１０ｂとなる。
【００３３】
辞書データ１４０ｃは、単語と、単語に対応するコードとを対応付けるデータである。
【００３４】
フレーズ定義データ１４０ｄは、フレーズを構成する複数の単語の組み合わせを定義したデータである。図３は、フレーズ定義データのデータ構造の一例を示す図である。図３に示すように、フレーズ定義データ１４０ｄには、各種のフレーズが定義されている。図３に示すフレーズは一例であり、他のフレーズも含まれている。
【００３５】
図３では一例として、フレーズを、符号化前の単語の組で定義しているがこれに限定されるものではない。例えば、フレーズ定義データ１４０ｄは、フレーズを構成する単語を、符号化後のコードによって定義しても良い。すなわち「in front of」であれば、「（inのコード） （frontのコード） （ofのコード）」によって、定義することができる。
【００３６】
ベクトルデータ１４０ｅは、符号化データ１４０ｂに含まれる各符号化単語、各符号化フレーズに割り当てられたベクトル値を示す情報であり、符号化データ１４０ｂをベクトル化した情報である。
【００３７】
制御部１５０は、登録部１５０ａと、符号化部１５０ｂと、フレーズ特定部１５０ｃと、ベクトル演算部１５０ｄとを有する。制御部１５０は、ＣＰＵ（Central Processing Unit）やＭＰＵ（Micro Processing Unit）などによって実現できる。また、制御部１５０は、ＡＳＩＣ（Application Specific Integrated Circuit）やＦＰＧＡ（Field Programmable Gate Array）などのハードワイヤードロジックによっても実現できる。
【００３８】
登録部１５０ａは、通信部１１０または入力部１２０を介して、各種の情報を受け付けた場合に、受け付けた情報を記憶部１４０に登録する処理部である。例えば、登録部１５０ａは、テキストデータ１４０ａ、辞書データ１４０ｃ、フレーズ定義データ１４０ｄを受け付けた場合には、受け付けたテキストデータ１４０ａ、辞書データ１４０ｃ、フレーズ定義データ１４０ｄを記憶部１４０に登録する。
【００３９】
符号化部１５０ｂは、テキストデータ１４０ａを符号化する処理部である。符号化部１５０ｂは、テキストデータ１４０ａに含まれる各単語と、辞書データ１４０ｃとを比較して、辞書データ１４０ｃにヒットした単語を、単語単位で符号化する処理を繰り返し実行することで、符号化データ１４０ａを生成する。
【００４０】
フレーズ特定部１５０ｃは、フレーズ定義データ１４０ｄを基にして、符号化データ１４０ｂに含まれる各符号化単語のうち、符号化フレーズを構成する符号化単語の組を特定する処理部である。フレーズ特定部１５０ｃは、符号化データ１４０ｂのコードのうち、符号化フレーズを構成する符号化単語の組の情報を、ベクトル演算部１５０ｄに出力する。
【００４１】
フレーズ定義データ１４０ｄに定義されたフレーズが、符号化前の単語の組で定義され

ている場合の処理について説明する。フレーズ特定部１５０ｃは、テキストデータ１４０ａと、フレーズ定義データ１４０ｄとを比較することで、テキストデータ１４０ａに含まれるフレーズを構成する単語の組を特定する。フレーズ特定部１５０ｃは、特定したフレーズを構成する単語の組に対応する各コードを、辞書データ１４０ｃを基にして特定する。フレーズ特定部１５０ｃは、特定した各コードと、符号化データ１４０ｂとを比較して、符号化データ１４０ｂに存在する符号化フレーズを特定する。

【００４２】

フレーズ定義データ１４０ｄに定義されたフレーズが、符号化後のコードの組で定義されている場合の処理について説明する。フレーズ特定部１５０ｃは、符号化データ１４０ｂと、フレーズ定義データ１４０ｄとを比較することで、符号化データ１４０ｂに含まれる符号化フレーズを特定する。

【００４３】

ベクトル演算部１５０ｄは、符号化データ１４０ｂに含まれる符号化単語および符号化フレーズの出現状況に応じて、符号化単語および符号化フレーズのベクトル値をそれぞれ算出し、割り当てることで、ベクトルデータ１４０ｅを生成する処理部である。ベクトル演算部１５０ｄは、演算部の一例である。

【００４４】

ベクトル演算部１５０ｄは、フレーズ特定部１５０ｃから特定結果を受け付け、符号化データ１４０ｂに含まれる各符号化単語のうち、符号化フレーズに対応する符号化単語と、符号化フレーズに対応しない符号化単語とを区別する。ベクトル演算部１５０ｄは、符号化フレーズのベクトル値と、符号化フレーズに対応しない符号化単語のベクトル値とを算出する。

【００４５】

ベクトル演算部１５０ｄが、符号化単語のベクトル値を算出する処理について説明する。ベクトル値の算出対象となる符号化単語を、対象単語と表記する。ベクトル演算部１５０ｄは、符号化データ１４０ｂ上において、対象単語の前方２つの符号化単語または符号化フレーズと、対象単語の後方２つの符号化単語に対して、Skip-gramによるモデル化を行い、対象単語の前後に符号化単語または符号化フレーズが出現する確率を特定することで、対象単語のベクトル値を算出する。

【００４６】

図１を用いて説明すると、ベクトル演算部１５０ｄは、コードＡ３のベクトル値を算出する場合には、コードＡ１、コードＡ２、符号化フレーズ１５、コードＡ７に対して、Skip-gramによるモデル化を行う。ベクトル演算部１５０ｄは、コードＡ３の前後に、コードＡ１、コードＡ２、符号化フレーズ１５、コードＡ７が出現する確率を特定することで、コードＡ３のベクトル値を算出する。解析装置は、コードＡ１、Ａ２、Ａ７、Ａ８についても同様の処理を実行することで、各コードのベクトル値を算出する。

【００４７】

ベクトル演算部１５０ｄが、符号化フレーズのベクトル値を算出する処理について説明する。ベクトル値の算出対象となる符号化フレーズを、対象フレーズと表記する。ベクトル演算部１５０ｄは、符号化データ１４０ｂ上において、対象フレーズの前方２つの符号化単語または符号化フレーズと、対象フレーズの後方２つの符号化単語に対して、Skip-gramによるモデル化を行い、対象フレーズの前後に符号化単語または符号化フレーズが出現する確率を特定することで、対象フレーズのベクトル値を算出する。

【００４８】

図１を用いて説明すると、ベクトル演算部１５０ｄは、符号化フレーズ１５のベクトル値を算出する場合には、コードＡ２、コードＡ３、コードＡ７、コードＡ８に対して、Skip-gramによるモデル化を行う。ベクトル演算部１５０ｄは、符号化フレーズ１５の前後に、コードＡ２、コードＡ３、コードＡ７、コードＡ８が出現する確率を特定することで、符号化フレーズ１５のベクトル値を算出する。ベクトル演算部１５０ｄは、他の符号化フレーズについても同様の処理を実行することで、各コードのベクトル値を算出する。

【００４９】
ここで、ベクトル演算部１５０ｄが、Skip-gramによるモデル化を行い、対象単語（対象フレーズ）のベクトル値を算出する処理は、例えば、文献（Tomas Mikolov他、「Distributed Representations of Words and Phrases and their Compositionality」）に記載した技術を利用する。
【００５０】
図４は、本実施例１に係る解析装置の処理手順を示すフローチャートである。図４に示すように、解析装置１００の符号化部１５０ｂは、テキストデータ１４０ａを読み込む（ステップＳ１０１）。符号化部１５０ｂは、テキストデータ１４０ａを、辞書データ１４０ｃを基にして、単語単位に符号化することで、符号化データ１４０ｂを生成する（ステップＳ１０２）。
【００５１】
解析装置１００のフレーズ特定部１５０ｃは、フレーズ定義データ１４０ｄを基にして、符号化データ１４０ｂに含まれる各コードのうち、符号化単語（符号化フレーズに含まれない符号化単語）と、符号化フレーズとを特定する（ステップＳ１０３）。
【００５２】
解析装置１００のベクトル演算部１５０ｄは、符号化フレーズに含まれない符号化単語について、符号化単語の出現状況に応じて、ベクトル値を算出する（ステップＳ１０４）。ベクトル演算部１５０ｄは、符号化フレーズについて、符号化フレーズの出現状況に応じて、ベクトル値を算出する（ステップＳ１０５）。ベクトル演算部１５０ｄは、符号化データ１４０ｂに対するベクトルデータ１４０ｅを生成する（ステップＳ１０６）。
【００５３】
上記に記載したように、解析装置１００によれば、テキストデータ１４０ａに含まれる単語からストップワードを除外すること無く、単語単位で符号化を行うことで、符号化データ１４０ｂを生成する。解析装置１００は、符号化単語のうち、所定のフレーズを構成する符号化単語を符号化フレーズとしてまとめ、符号化単語および符号化フレーズの出現状況に応じて、符号化データ１４０ｂをベクトル化する。このように、解析装置１００は、ストップワードを除外しないため、符号化データ１４０ｂをベクトル化する際の解析精度を向上させることができる。また、解析装置１００が利用する符号化では、符号化データ１４０ｂが単語単位で符号化されているため、Zipによる符号化と異なり、符号化したままで各単語を区別できるため、復号化を行うことなく、文書に含まれる単語のベクトル値を算出でき、解析速度を向上できる。
【実施例２】
【００５４】
図５は、本実施例２に係る解析装置の処理の一例を説明するための図である。ここでは、解析装置が解析するテキストデータ２０ａを「We take lunch out」とする。テキストデータ２０ａに含まれる「take、out」はフレーズに対応する単語の組であるが、各単語が離れているため、実施例１で説明した解析装置１００が利用するフレーズ定義データ１４０ｄにヒットせず、符号化フレーズのベクトル値を算出できない場合がある。以下の説明では、それぞれが離間した単語により構成されるフレーズを「離間フレーズ」と表記する。
【００５５】
本実施例２に係る解析装置は、テキストデータ２０ａを走査して、離間フレーズを検出した場合には、離間フレーズを構成する各単語が連続するように、テキストデータ２０ａの単語を並び変えることで、テキストデータ２１ａを生成する。例えば、図５に示す例では、解析装置１００は、テキストデータ２０ａの「out」を削除し、削除した「out」を「take」の直後に配置することで、テキストデータ２１ａを生成する。解析装置は、テキストデータ２１ａ（テキストデータ２１ａを符号化した符号化データ）に基づいて、テキストデータ２１ａをベクトル化する。解析装置がテキストデータ２１ａに基づいて、テキストデータ２１ａをベクトル化する処理は、上述した実施例１の処理と同様である。

【００５６】
上記処理を実行することで、本実施例２に係る解析装置によれば、離間フレーズがテキストデータ２０ａに存在する場合でも、離間フレーズを符号化フレーズとして特定することができる。このため、離間フレーズを構成する各単語のコード毎のベクトル値を算出することを抑止して、解析精度を向上させることができる。
【００５７】
図６は、本実施例２に係る解析装置の構成を示す機能ブロック図である。図６に示すように、この解析装置２００は、通信部２１０と、入力部２２０と、表示部２３０と、記憶部２４０と、制御部２５０とを有する。このうち、入力部２２０、表示部２３０に関する説明は、図２で説明した入力部１２０、表示部１３０に関する説明と同様であるため、説明を省略する。
【００５８】
通信部２１０は、ネットワークを介して他の外部装置と通信を実行する処理部である。例えば、解析装置２００は、後述するテキストデータ２４０ａ、離間フレーズテーブル２４０ｃ、辞書データ２４０ｄ、フレーズ定義データ２４０ｅ等を、通信部２１０を介して、受信しても良い。
【００５９】
記憶部２４０は、テキストデータ２４０ａと、テキストデータ２４１ａと、符号化データ２４０ｂと、離間フレーズテーブル２４０ｃと、辞書データ２４０ｄと、フレーズ定義データ２４０ｅと、ベクトルデータ２４０ｆとを有する。記憶部２４０は、ＲＡＭ、ＲＯＭ、フラッシュメモリなどの半導体メモリ素子や、ＨＤＤなどの記憶装置に対応する。
【００６０】
テキストデータ２４０ａは、複数の単語を含む文字列データである。テキストデータ２４０ａの一例は、図５に示したテキストデータ２０ａとなる。テキストデータ２４１ａは、図５で説明したように、離間フレーズの単語が連続するように単語が並び変えられテキストデータ２１ａに対応するデータである。
【００６１】
符号化データ２４０ｂは、テキストデータ２４１ａに含まれる各単語を単語単位で符号化したデータである。
【００６２】
離間フレーズテーブル２４０ｃは、離間フレーズに関する情報を定義したテーブルである。図７は、離間フレーズテーブルのデータ構造の一例を示す図である。図７に示すように、この離間フレーズテーブル２４０ｃは、主単語、副単語、フレーズを対応付ける。主単語は、離間フレーズのうち、最初に現れる単語である。副単語は、離間フレーズのうち、主単語の次に現れる単語である。フレーズは、離間フレーズを構成する単語を連続して並べたものである。
【００６３】
辞書データ２４０ｄ、フレーズ定義データ２４０ｅに関する説明は、図２で説明した辞書データ１４０ｃ、フレーズ定義データ１４０ｄに関する説明と同様である。
【００６４】
ベクトルデータ２４０ｆは、符号化データ２４０ｂに含まれる各符号化単語、各符号化フレーズに割り当てられたベクトル値を示す情報であり、符号化データ２４０ｂをベクトル化した情報である。
【００６５】
制御部２５０は、登録部２５０ａと、離間フレーズ処理部２５０ｂと、符号化部２５０ｃと、フレーズ特定部２５０ｄと、ベクトル演算部２５０ｅとを有する。制御部２５０は、ＣＰＵやＭＰＵなどによって実現できる。また、制御部２５０は、ＡＳＩＣやＦＰＧＡなどのハードワイヤードロジックによっても実現できる。
【００６６】
登録部２５０ａは、通信部２１０または入力部２２０を介して、各種の情報を受け付け

た場合に、受け付けた情報を記憶部２４０に登録する処理部である。例えば、登録部２５０ａは、テキストデータ２４０ａ、離間フレーズテーブル２４０ｃ、辞書データ２４０ｄ、フレーズ定義データ２４０ｅを受け付けた場合には、受け付けた各データを記憶部２４０に登録する。

【００６７】

離間フレーズ処理部２５０ｂは、テキストデータ２４０ａと、離間フレーズテーブル２４０ｃとを比較して、テキストデータ２４０ａに含まれる離間フレーズを特定する。離間フレーズ処理部２５０ｂは、特定した離間フレーズの単語が連続するように並び変えを行うことで、テキストデータ２４１ａを生成する。以下において、離間フレーズ処理部２５０ｂの処理の一例について説明する。

【００６８】

離間フレーズ処理部２５０ｂは、テキストデータ２４０ａと、離間フレーズテーブル２４０ｃとを比較して、離間フレーズテーブル２４０ｃの主単語にヒットする単語を、テキストデータ２４０ａから特定する。離間フレーズ処理部２５０ｂは、主単語にヒットする単語が存在する場合には、ヒットした単語と同じ文中で、ヒットした単語から後ろ方向に、所定語数未満の位置に、副単語（主単語に対応する副単語）にヒットするか否かを判定する。離間フレーズテーブル２４０ｃは、主単語および副単語にヒットした場合には、係る主単語、副単語を離間フレーズとして特定する。

【００６９】

例えば、離間フレーズ処理部２５０ｂは、テキストデータ２４０ａと、離間フレーズテーブル２４０ｃとを比較し、主単語「take」がヒットしたものとする。離間フレーズ処理部２５０ｂは、ヒットした主単語「take」と同じ文中で、「take」から後ろ方向に所定語数未満の位置に、副単語「out」が存在する場合には、離間した「take」、「out」を離間フレーズであると特定する。

【００７０】

離間フレーズ処理部２５０ｂは、離間フレーズを特定すると、離間フレーズの副単語を削除し、主単語の直後に副単語を配置する処理を実行する。離間フレーズ処理部２５０ｂは、各離間フレーズについて、上記処理を繰り返し実行することで、テキストデータ２４１ａを生成する。

【００７１】

なお、離間フレーズ処理部２５０ｂは、主単語をフレーズに置き換え、副単語を削除することで、主単語と副単語とが連続するように置き換えを行っても良い。

【００７２】

符号化部２５０ｃは、テキストデータ２４１ａを符号化する処理部である。符号化部２５０ｃは、テキストデータ２４１ａに含まれる各単語と、辞書データ２４０ｄとを比較して、辞書データ２４０ｄにヒットした単語を、単語単位で符号化する処理を繰り返し実行することで、符号化データ２４０ｂを生成する。

【００７３】

フレーズ特定部２５０ｄは、フレーズ定義データ２４０ｅを基にして、符号化データ２４０ｂに含まれる各符号化単語のうち、符号化フレーズを構成する符号化単語の組を特定する処理部である。フレーズ特定部２５０ｄは、符号化データ２４０ｂのコードのうち、符号化フレーズを構成する符号化単語の組の情報を、ベクトル演算部２５０ｅに出力する。フレーズ特定部２５０ｄに関するその他の処理は、図２に示したフレーズ特定部１５０ｃの処理と同様である。

【００７４】

ベクトル演算部２５０ｅは、符号化データ２４０ｂに含まれる符号化単語および符号化フレーズの出現状況に応じて、符号化単語および符号化フレーズのベクトル値をそれぞれ算出し、割り当てることで、ベクトルデータ２４０ｆを生成する処理部である。ベクトル演算部２５０ｅに関するその他の処理は、図２で説明したベクトル演算部１５０ｅに関する処理と同様である。

【００７５】
図８は、本実施例２に係る解析装置の処理手順を示すフローチャートである。図８に示すように、解析装置２００の離間フレーズ処理部２５０ｂは、テキストデータ２４０ａを読み込む（ステップＳ２０１）。離間フレーズ処理部２５０ｂは、離間フレーズテーブル２４０ｃとテキストデータ２４０ａとを比較して離間フレーズを特定する（ステップＳ２０２）。
【００７６】
離間フレーズ処理部２５０ｂは、離間フレーズにヒットしない場合には（ステップＳ２０３，Ｎｏ）、ステップＳ２０５に移行する。離間フレーズ処理部２５０ｂは、離間フレーズにヒットした場合には（ステップＳ２０３，Ｙｅｓ）、ステップＳ２０４に移行する。離間フレーズ処理部２５０ｂは、離間フレーズに対応する副単語を移動し、主単語の直後に副単語を配置する（ステップＳ２０４）。
【００７７】
解析装置２００の符号化部２５０ｃは、テキストデータ２４１ａを、辞書データ２４０ｄを基にして、単語単位に符号化することで、符号化データ２４０ｂを生成する（ステップＳ２０５）。
【００７８】
解析装置２００のフレーズ特定部２５０ｄは、フレーズ定義データ２４０ｅを基にして、符号化データ２４０ｂに含まれる各コードのうち、符号化単語（符号化フレーズに含まれない符号化単語）と、符号化フレーズとを特定する（ステップＳ２０６）。
【００７９】
解析装置２００のベクトル演算部２５０ｅは、符号化フレーズに含まれない符号化単語について、符号化単語の出現状況に応じて、ベクトル値を算出する（ステップＳ２０７）。ベクトル演算部２５０ｅは、符号化フレーズについて、符号化フレーズの出現状況に応じて、ベクトル値を算出する（ステップＳ２０８）。ベクトル演算部２５０ｅは、符号化データ２４０ｂに対するベクトルデータ２４０ｆを生成する（ステップＳ２０９）。
【００８０】
上記に記載したように、解析装置２００によれば、離間フレーズがテキストデータ２０ａに存在する場合でも、離間フレーズを符号化フレーズとして特定することができる。このため、離間フレーズを構成する各単語のコード毎のベクトル値を算出することを抑止して、解析精度を向上させることができる。
【００８１】
ところで、本実施例で説明した解析装置１００（２００）は、各符号化単語とベクトル値との関係が既知の場合には、各符号化単語とベクトル値とを対応付けたオートマトンを用いて、テキストデータのベクトル化を行ってもよい。
【００８２】
図９は、オートマトンを用いたベクトル演算の一例を説明するための図である。図９に示す例では、テキストデータを「Every day we take care of our daughter」とする。解析装置１００は、図１と同様にして、テキストデータ１０ａに含まれる各単語を、単語単位で符号化することで、符号化データ１０ｂを生成する。
【００８３】
続いて、解析装置１００は、符号化データ１０ａとオートマトン５０とを比較して、各符号化単語をベクトル値に変換する。ここで、オートマトン５０は、各符号化単語とベクトル値とを対応付けた情報である。なお、オートマトン５０は、ストップワード等に対応する符号化単語に対応するベクトル値を「０」に設定しておく。
【００８４】
図９に示す例では、オートマトン５０による変換により、コードＡ１～Ａ５、コードＡ６～Ａ８が、ベクトル値Ｖ１～Ｖ７に変換される。なお、コードＡ６は、ステップワードに対応する符号化単語であるため、ベクトル値は０に設定される。
【００８５】

上記のように、オートマトン５０を利用して、符号化データ１０ｂをベクトル化することで、ベクトル化する処理を高速化することが可能となる。

【００８６】

次に、上記実施例に示した解析装置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１０は、解析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。

【００８７】

図１０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、記憶媒体からプログラム等を読み取る読み取り装置３０４と、ネットワークを介して他のコンピュータとの間でデータの授受を行う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。

【００８８】

ハードディスク装置３０７は、離間フレーズ処理プログラム３０７ａ、符号化プログラム３０７ｂ、フレーズ特定プログラム３０７ｃ、ベクトル演算プログラム３０７ｄを有する。ＣＰＵ３０１は、離間フレーズ処理プログラム３０７ａ、符号化プログラム３０７ｂ、フレーズ特定プログラム３０７ｃ、ベクトル演算プログラム３０７ｄを読み出してＲＡＭ３０６に展開する。

【００８９】

離間フレーズ処理プログラム３０７ａは、離間フレーズ処理プロセス３０６ａとして機能する。符号化プログラム３０７ｂは、符号化プロセス３０６ｂとして機能する。フレーズ特定プログラム３０７ｃは、フレーズ特定プロセス３０６ｃとして機能する。ベクトル演算プログラム３０７ｄは、ベクトル演算プロセス３０６ｄとして機能する。

【００９０】

離間フレーズ処理プロセス３０６ａの処理は、離間フレーズ処理部２５０ｂの処理に対応する。符号化プロセス３０６ｂの処理は、符号化部１５０ｂ、２５０ｃの処理に対応する。フレーズ特定プロセス３０６ｃの処理は、フレーズ特定部１５０ｃ、２５０ｄの処理に対応する。ベクトル演算プロセス３０６ｄの処理は、ベクトル演算部１５０ｄ、２５０ｅの処理に対応する。

【００９１】

なお、各プログラム３０７ａ～３０７ｄ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ディスク、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～３０７ｄを読み出して実行するようにしても良い。

【００９２】

以上の各実施例を含む実施形態に関し、さらに以下の付記を開示する。

【００９３】

（付記１）コンピュータに、

解析対象の文書を単語単位で符号化した、複数の符号化単語を生成し、

前記符号化単語、または、前記複数の符号化単語の組み合わせに割り当てられた符号化フレーズに関する、前記解析対象の文書の出現状況に応じてそれぞれ生成されたベクトル値を、前記符号化単語、または、前記符号化フレーズに割り当てる

処理を実行させることを特徴とする解析プログラム。

【００９４】

（付記２）第１符号化フレーズの前後に位置する前記符号化単語または他の第２符号化フレーズに対してSkip-gramによるモデル化を行い、前記第１符号化フレーズの前後に、前記符号化単語または前記第２符号化フレーズが出現する確率を特定することで、前記第１

符号化フレーズのベクトル値を算出することを特徴とする付記１に記載の解析プログラム。

【００９５】

（付記３）フレーズを構成する第１単語と第２単語との組を定義したテーブルを基にして、前記解析対象の文書に含まれる第１単語を特定し、特定した第１単語と同じ文中に第２単語が含まれる場合には、特定した第１単語と同じ文中に含まれる第２単語を削除し、特定した第１単語の直後に前記第２単語を配置する処理を更にコンピュータに実行させることを特徴とする付記１または２に記載の解析プログラム。

【００９６】

（付記４）コンピュータが実行する解析方法であって、

解析対象の文書を単語単位で符号化した、複数の符号化単語を生成し、

前記符号化単語、または、前記複数の符号化単語の組み合わせに割り当てられた符号化フレーズに関する、前記解析対象の文書の出現状況に応じてそれぞれ生成されたベクトル値を、前記符号化単語、または、前記符号化フレーズに割り当てる

処理を実行することを特徴とする解析方法。

【００９７】

（付記５）第１符号化フレーズの前後に位置する前記符号化単語または他の第２符号化フレーズに対してSkip-gramによるモデル化を行い、前記第１符号化フレーズの前後に、前記符号化単語または前記第２符号化フレーズが出現する確率を特定することで、前記第１符号化フレーズのベクトル値を算出することを特徴とする付記４に記載の解析方法。

【００９８】

（付記６）フレーズを構成する第１単語と第２単語との組を定義したテーブルを基にして、前記解析対象の文書に含まれる第１単語を特定し、特定した第１単語と同じ文中に第２単語が含まれる場合には、特定した第１単語と同じ文中に含まれる第２単語を削除し、特定した第１単語の直後に前記第２単語を配置する処理を更にコンピュータに実行させることを特徴とする付記４または５に記載の解析方法。

【００９９】

（付記７）解析対象の文書を単語単位で符号化した、複数の符号化単語を生成する符号化部と、

前記符号化単語、または、前記複数の符号化単語の組み合わせに割り当てられた符号化フレーズに関する、前記解析対象の文書の出現状況に応じてそれぞれ生成されたベクトル値を、前記符号化単語、または、前記符号化フレーズに割り当てる演算部と

を有することを特徴とする解析装置。

【０１００】

（付記８）演算部は、第１符号化フレーズの前後に位置する前記符号化単語または他の第２符号化フレーズに対してSkip-gramによるモデル化を行い、前記第１符号化フレーズの前後に、前記符号化単語または前記第２符号化フレーズが出現する確率を特定することで、前記第１符号化フレーズのベクトル値を算出することを特徴とする付記７に記載の解析装置。

【０１０１】

（付記９）フレーズを構成する第１単語と第２単語との組を定義したテーブルを基にして、前記解析対象の文書に含まれる第１単語を特定し、特定した第１単語と同じ文中に第２単語が含まれる場合には、特定した第１単語と同じ文中に含まれる第２単語を削除し、特定した第１単語の直後に前記第２単語を配置する離間フレーズ処理部を更に有することを特徴とする付記７または８に記載の解析装置。

【符号の説明】

【０１０２】

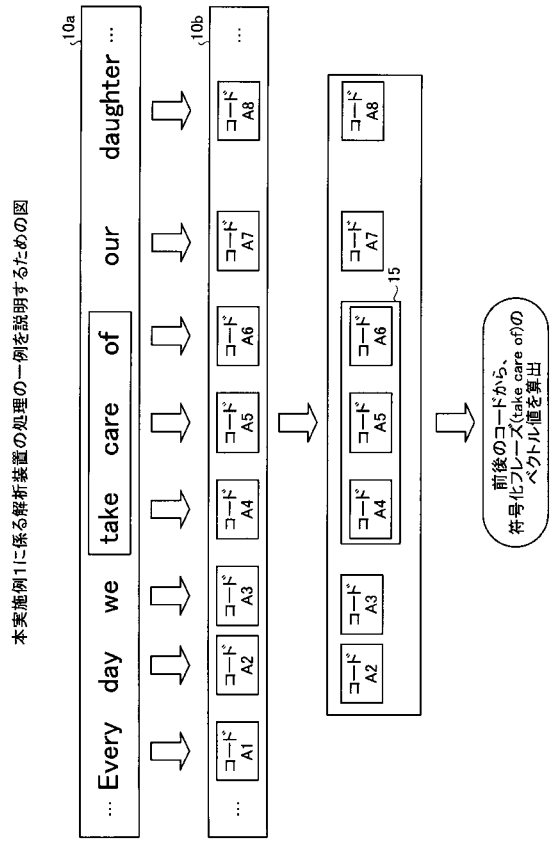
１００、２００　解析装置

１１０、２１０　通信部

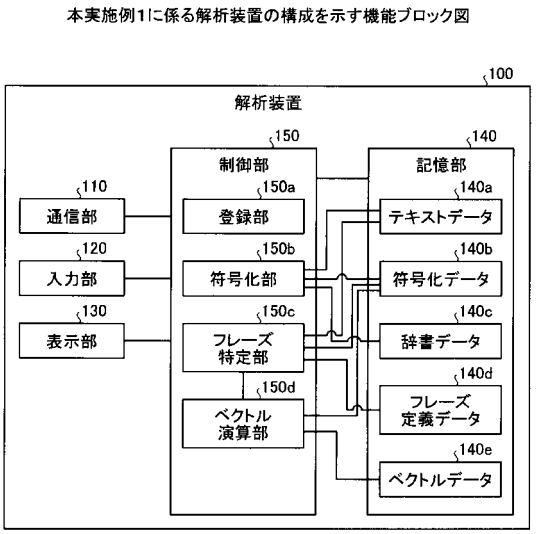
１２０、２２０　入力部

１３０、２３０　表示部
１４０、２４０　記憶部
１５０、２５０　制御部

【図１】




【図２】

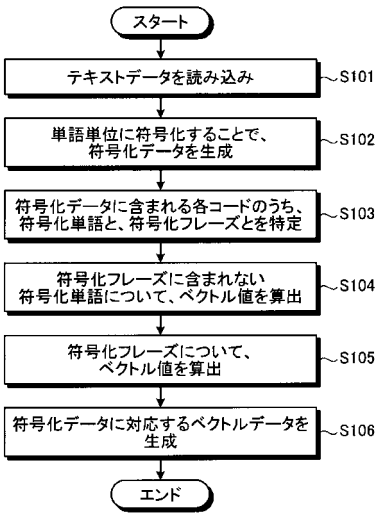


【図３】

フレーズ定義データのデータ構造の一例を示す図

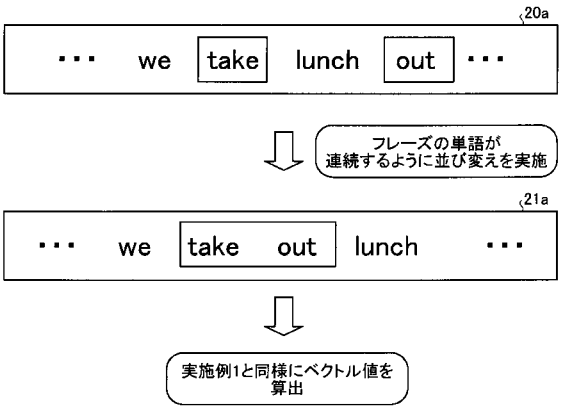

【図４】

本実施例1に係る解析装置の処理手順を示すフローチャート



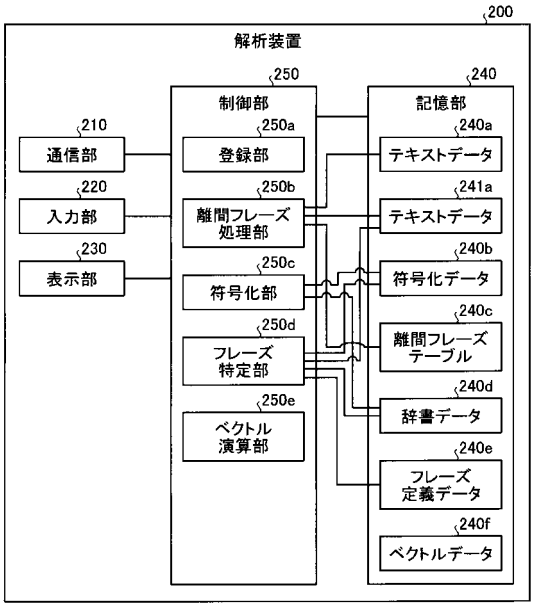

【図５】

本実施例2に係る解析装置の処理の一例を説明するための図




【図６】

本実施例2に係る解析装置の構成を示す機能ブロック図



【図７】

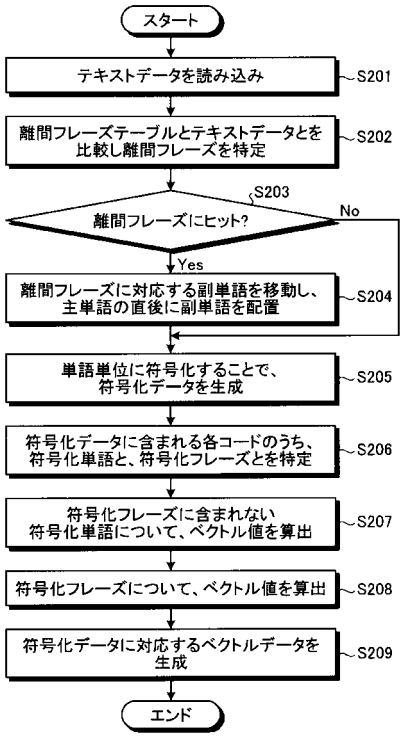
離間フレーズテーブルのデータ構造の一例を示す図

240c

| 主単語 | 副単語 | フレーズ |
|---|---|---|
| … | … | … |
| figure | out | figure out |
| take | out | take out |
| turn | off | turn off |
| put | out | put out |
| make | out | make out |
| … | … | … |

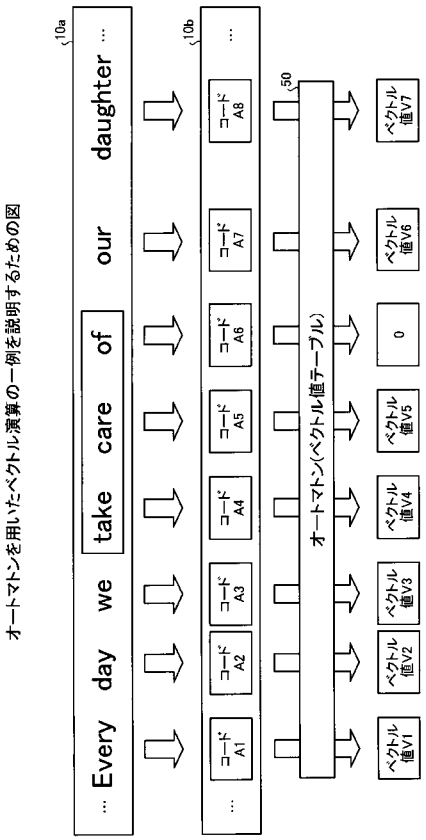
【図８】

本実施例2に係る解析装置の処理手順を示すフローチャート




【図９】

オートマトンを用いたベクトル演算の一例を説明するための図




【図１０】

解析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