(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5382651号
(P5382651)

(45)発行日　平成26年1月8日(2014.1.8)　(24)登録日　平成25年10月11日(2013.10.11)

(51)Int.Cl.　　F I
G06F 17/27 (2006.01)　　G06F 17/27 Z
G06F 17/30 (2006.01)　　G06F 17/30 170A

請求項の数 13　(全 35 頁)

(21)出願番号　特願2009-207944(P2009-207944)
(22)出願日　平成21年9月9日(2009.9.9)
(65)公開番号　特開2011-59917(P2011-59917A)
(43)公開日　平成23年3月24日(2011.3.24)
審査請求日　平成24年8月24日(2012.8.24)

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100115749
弁理士　谷川　英和
(72)発明者　ステイン　デ　サーガ
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　鳥澤　健太郎
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　風間　淳一
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

最終頁に続く

(54)【発明の名称】単語対取得装置、単語対取得方法、およびプログラム

(57)【特許請求の範囲】
【請求項１】
１以上の文章群を格納し得る文章群格納部と、
１以上の単語と当該１以上の単語が属するクラスを識別するクラス識別子とを対応づけて有する２以上の単語クラス情報を格納し得る単語クラス情報格納部と、
２つのクラスの良さを示す指標であり、当該２つのクラスに属する単語対が、所定の関係を有する２つの単語対を取得するために利用するパターンであるシードパターンと良く共起する程度を示すクラス対良好度を格納し得るクラス対良好度格納部と、
２つの単語である単語対を１以上格納し得る単語対格納部と、
前記単語対格納部に格納されている１以上の単語対が有する各単語が属する２つのクラスのクラス対良好度を前記クラス対良好度格納部から取得するクラス対良好度取得部と、
前記クラス対良好度取得部が取得したクラス対良好度を用いて、前記単語対格納部の各単語対のスコアを決定するスコア決定部と、
前記スコア決定部が決定したスコアが予め決められた条件を満たすほど、スコアが高い１以上の単語対を取得する単語対選択部と、
前記単語対選択部が取得した１以上の単語対を出力する単語対出力部とを具備する単語対取得装置。
【請求項２】
所定の関係を有する２つの単語対を取得するために利用するパターンであるシードパターンを１以上格納し得るシードパターン格納部と、

２つの各クラスに属する単語対が、前記文章群格納部の１以上の文章群の中で、前記１以上のシードパターンと共起する回数または割合が多いほどクラス対良好度が大きくなるようにクラス対良好度を算出するクラス対良好度算出部とをさらに具備し、
前記クラス対良好度算出部が算出した２つのクラスのクラス対良好度は、前記クラス対良好度格納部に格納されているクラス対良好度である請求項１記載の単語対取得装置。
【請求項３】
所定の関係を有する２つの単語対を取得するために利用するパターンであるシードパターンではないパターンであり、前記所定の関係を有する２つの単語対を取得するために利用するパターンを１以上格納し得るパターン格納部と、
前記パターン格納部に格納されている１以上の各パターンと前記シードパターンとの類似度を、パターンごとに格納し得るパターン類似度格納部と、
前記シードパターン格納部に格納されている１以上のシードパターン、および前記パターン格納部に格納されている１以上のパターンのいずれかを取得し、前記文章群格納部に格納されている１以上の文章群から、前記シードパターンまたは前記パターンと共起する１以上の単語対を取得する単語対取得部とをさらに具備し、
前記スコア決定部は、
前記パターン類似度格納部に格納されている前記１以上の各パターンと前記シードパターンとの類似度をパラメータとする増加関数を用いて、前記単語対取得部が取得した各単語対のスコアを決定する請求項１または請求項２記載の単語対取得装置。
【請求項４】
前記１以上のシードパターンと共起する単語対に対応するクラス対と、前記パターン格納部に格納されている１以上の各パターンと共起する単語対に対応するクラス対との重なりが大きいほど、大きくなるように類似度を算出するパターン類似度算出部をさらに具備し、
前記パターン類似度算出部が算出した類似度は、前記パターン類似度格納部に格納されている類似度である請求項３記載の単語対取得装置。
【請求項５】
１以上の各単語対と１以上の各パターンとの親和性に関する情報であり、単語対とパターンと良く共起する程度を示す情報である親和性情報を格納し得る親和性情報格納部をさらに具備し、
前記スコア決定部は、
前記親和性情報格納部の親和性情報をもパラメータとする増加関数を用いて、前記単語対取得部が取得した各単語対のスコアを決定する請求項３または請求項４記載の単語対取得装置。
【請求項６】
前記単語対取得部が取得した１以上の単語対と、前記１以上の各パターンとが共起する回数または割合が多いほど、大きくなるように親和性情報を算出する親和性情報算出部をさらに具備し、
前記親和性情報格納部の親和性情報は、前記親和性情報算出部が算出した親和性情報である請求項５記載の単語対取得装置。
【請求項７】
前記スコア決定部は、
前記クラス対良好度、および前記シードパターンとパターンとの類似度、および前記親和性情報、の積が最も大きいシードパターンまたはパターンにおけるスコアを、各単語対のスコアとして決定する請求項６記載の単語対取得装置。
【請求項８】
前記文章群格納部に格納されている１以上の文章群の各文に対して、形態素解析および係り受け解析し、一番目に出現する第一の名詞または名詞句を起点として、二番目に出現する第二の名詞または名詞句を終点として、前記起点から前記終点までに至る形態素の繋がりをパターンとして取得し、または、前記起点からの形態素の繋がりと前記終点からの形

態素の繋がりが結ばれる形態素までをパターンとして取得するパターン取得部をさらに具備し、
前記パターン格納部のパターンは、前記パターン取得部が取得したパターンである請求項３から請求項７いずれか記載の単語対取得装置。
【請求項９】
最終的に出力しない単語対に対応するクラス対を識別する２つのクラス識別子である除外クラス対を１以上格納し得る除外クラス対格納部と、
前記１以上の除外クラス対に対応する単語対を出力する単語対から除外する単語対除外部とをさらに具備する請求項１から請求項８いずれか記載の単語対取得装置。
【請求項１０】
前記１以上の文章群における、各クラスに属する単語の平均出現頻度と、クラス識別子とを対に有するクラス出現頻度情報を、クラス毎に格納し得るクラス出現頻度情報格納部と、
前記平均出現頻度が予め決められた閾値以上の差を有する２つのクラスのクラス識別子を除外クラス対として、前記除外クラス対格納部に蓄積する除外クラス対蓄積部とをさらに具備する請求項９記載の単語対取得装置。
【請求項１１】
前記文章群格納部の１以上の文章群を用いて、同一の動詞、または同一の動詞と助詞と共起する回数または割合が多い単語を同一のクラスに属するように、１以上の単語クラス情報を取得する単語クラス情報取得部をさらに具備し、
前記単語クラス情報格納部の単語クラス情報は、前記単語クラス情報取得部が取得した単語クラス情報である請求項１から請求項１０いずれか記載の単語対取得装置。
【請求項１２】
記憶媒体に、
１以上の文章群を格納し、
１以上の単語と当該１以上の単語が属するクラスを識別するクラス識別子とを対応づけて有する２以上の単語クラス情報を格納し、
２つのクラスの良さを示す指標であり、当該２つのクラスに属する単語対が、所定の関係を有する２つの単語対を取得するために利用するパターンであるシードパターンと良く共起する程度を示すクラス対良好度を格納し、
所定の関係を有する２つの単語対を取得するために利用するパターンであるシードパターンを１以上格納しており、
単語対取得部、クラス対良好度取得部、スコア決定部、単語対選択部、および単語対出力部により実現される単語対取得方法であって、
前記単語対取得部により、前記記憶媒体に格納されている１以上のシードパターンのいずれかを取得し、前記記憶媒体に格納されている１以上の文章群から、前記取得したシードパターンと共起する１以上の単語対を取得する単語対取得ステップと、
前記クラス対良好度取得部により、前記単語対取得ステップで取得された１以上の単語対が有する各単語が属する２つのクラスのクラス対良好度を前記記憶媒体から取得するクラス対良好度取得ステップと、
前記スコア決定部により、前記クラス対良好度取得ステップで取得されたクラス対良好度を用いて、前記単語対取得ステップで取得された各単語対のスコアを決定するスコア決定ステップと、
前記単語対選択部により、前記スコア決定ステップで決定されたスコアが予め決められた条件を満たすほど、スコアが高い１以上の単語対を取得する単語対選択ステップと、
前記単語対出力部により、前記単語対選択ステップで取得された１以上の単語対を出力する単語対出力ステップとを具備する単語対取得方法。
【請求項１３】
記憶媒体に、
１以上の文章群を格納し、

１以上の単語と当該１以上の単語が属するクラスを識別するクラス識別子とを対応づけて有する２以上の単語クラス情報を格納し、
２つのクラスの良さを示す指標であり、当該２つのクラスに属する単語対が、所定の関係を有する２つの単語対を取得するために利用するパターンであるシードパターンと良く共起する程度を示すクラス対良好度を格納し、
所定の関係を有する２つの単語対を取得するために利用するパターンであるシードパターンを１以上格納しており、
コンピュータを、
前記記憶媒体に格納されている１以上のシードパターンのいずれかを取得し、前記記憶媒体に格納されている１以上の文章群から、前記取得したシードパターンと共起する１以上の単語対を取得する単語対取得部と、
前記単語対取得部が取得した１以上の単語対が有する各単語が属する２つのクラスのクラス対良好度を前記記憶媒体から取得するクラス対良好度取得部と、
前記クラス対良好度取得部が取得したクラス対良好度を用いて、前記単語対取得部が取得した各単語対のスコアを決定するスコア決定部と、
前記スコア決定部が決定したスコアが予め決められた条件を満たすほど、スコアが高い１以上の単語対を取得する単語対選択部と、
前記単語対選択部が取得した１以上の単語対を出力する単語対出力部として機能させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、所定の関係を有する２つの単語対を取得する単語対取得装置等に関するものである。

【背景技術】

【０００２】

従来、取り出したい単語対を少量与えて、当該単語対からパターンを取得する単語対取得装置があった。そして、従来の単語対取得装置は、その取得したパターンと共起する単語対を取得するものであった（例えば、非特許文献１参照）。

【先行技術文献】

【非特許文献】

【０００３】

【非特許文献１】P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations.In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLINGACL-06), pages 113-120, 2006.

【発明の概要】

【発明が解決しようとする課題】

【０００４】

しかしながら、従来の単語対取得装置においては、与えられた任意の関係にある単語対を適切に取得できなかった。

【課題を解決するための手段】

【０００５】

本第一の発明の単語対取得装置は、１以上の文章群を格納し得る文章群格納部と、１以上の単語と１以上の単語が属するクラスを識別するクラス識別子とを対応づけて有する２以上の単語クラス情報を格納し得る単語クラス情報格納部と、２つのクラスの良さを示す指標であるクラス対良好度を格納し得るクラス対良好度格納部と、２つの単語である単語

対を１以上格納し得る単語対格納部と、単語対格納部に格納されている１以上の単語対が有する各単語が属する２つのクラスのクラス対良好度をクラス対良好度格納部から取得するクラス対良好度取得部と、クラス対良好度取得部が取得したクラス対良好度を用いて、単語対格納部の各単語対のスコアを決定するスコア決定部と、スコア決定部が決定したスコアが予め決められた条件を満たすほど、スコアが高い１以上の単語対を取得する単語対選択部と、単語対選択部が取得した１以上の単語対を出力する単語対出力部とを具備する単語対取得装置である。

【０００６】

かかる構成により、クラス対良好度を用いて、所定の関係にある単語対を適切に取得できる。

【０００７】

また、本第二の発明の単語対取得装置は、第一の発明に対して、所定の関係を有する２つの単語対を取得するために利用するパターンであるシードパターンを１以上格納し得るシードパターン格納部と、２つの各クラスに属する単語対が、文章群格納部の１以上の文章群の中で、１以上のシードパターンと共起する回数または割合が多いほどクラス対良好度が大きくなるようにクラス対良好度を算出するクラス対良好度算出部とをさらに具備し、クラス対良好度算出部が算出した２つのクラスのクラス対良好度は、クラス対良好度格納部に格納されているクラス対良好度である単語対取得装置である。

【０００８】

かかる構成により、クラス対良好度が適切に算出でき、そのクラス対良好度を用いて、所定の関係にある単語対を適切に取得できる。

【０００９】

また、本第三の発明の単語対取得装置は、第一または第二の発明に対して、シードパターンではないパターンであり、所定の関係を有する２つの単語対を取得するために利用するパターンを１以上格納し得るパターン格納部と、パターン格納部に格納されている１以上の各パターンとシードパターンとの類似度を、パターンごとに格納し得るパターン類似度格納部と、シードパターン格納部に格納されている１以上のシードパターン、およびパターン格納部に格納されている１以上のパターンのいずれかを取得し、文章群格納部に格納されている１以上の文章群から、シードパターンまたはパターンと共起する１以上の単語対を取得する単語対取得部とをさらに具備し、スコア決定部は、パターン類似度格納部に格納されている１以上の各パターンとシードパターンとの類似度をも用いて、単語対取得部が取得した各単語対のスコアを決定する単語対取得装置である。

【００１０】

かかる構成により、シードパターンとパターンとの類似度を用いて、所定の関係にある単語対をさらに適切に取得できる。

【００１１】

また、本第四の発明の単語対取得装置は、第三の発明に対して、１以上のシードパターンと共起する単語対に対応するクラス対と、パターン格納部に格納されている１以上の各パターンと共起する単語対に対応するクラス対との重なりが大きいほど、大きくなるように類似度を算出するパターン類似度算出部をさらに具備し、パターン類似度算出部が算出した類似度は、パターン類似度格納部に格納されている類似度である単語対取得装置である。

【００１２】

かかる構成により、シードパターンとパターンとの類似度を適切に算出でき、その類似度を用いて、所定の関係にある単語対をさらに適切に取得できる。

【００１３】

また、本第五の発明の単語対取得装置は、第一から第四いずれかの発明に対して、１以上の各単語対と１以上の各パターンとの親和性に関する情報である親和性情報を格納し得る親和性情報格納部をさらに具備し、スコア決定部は、親和性情報格納部の親和性情報をも用いて、単語対取得部が取得した各単語対のスコアを決定する単語対取得装置である。

【００１４】
かかる構成により、パターンと単語対の親和性を用いて、所定の関係にある単語対をさらに適切に取得できる。
【００１５】
また、本第六の発明の単語対取得装置は、第五の発明に対して、単語対取得部が取得した１以上の単語対と、１以上の各パターンとが共起する回数または割合が多いほど、大きくなるように親和性情報を算出する親和性情報算出部をさらに具備し、親和性情報格納部の親和性情報は、親和性情報算出部が算出した親和性情報である単語対取得装置である。
【００１６】
かかる構成により、パターンと単語対の親和性を適切に算出でき、その親和性を用いて、所定の関係にある単語対をさらに適切に取得できる。
【００１７】
また、本第七の発明の単語対取得装置は、第六の発明に対して、スコア決定部は、クラス対良好度、シードパターンとパターンとの類似度、および親和性情報との積が最も大きいシードパターンまたはパターンにおけるスコアを、各単語対のスコアとして決定する単語対取得装置である。
【００１８】
かかる構成により、単語対のスコアを精度高く算出でき、その結果、所定の関係にある単語対を極めて適切に取得できる。
【００１９】
また、本第八の発明の単語対取得装置は、第三から第七いずれかの発明に対して、文章群格納部に格納されている１以上の文章群の各文に対して、形態素解析および係り受け解析し、第一の名詞または名詞句を起点として、第二の名詞または名詞句を終点として、起点から終点までに至る形態素の繋がりをパターンとして取得し、または、起点からの形態素の繋がりと終点からの形態素の繋がりが結ばれる形態素までをパターンとして取得するパターン取得部をさらに具備し、パターン格納部のパターンは、パターン取得部が取得したパターンである単語対取得装置である。
【００２０】
かかる構成により、文章群から適切にパターンを取得でき、そのパターンを用いて、所定の関係にある単語対を適切に取得できる。
【００２１】
また、本第九の発明の単語対取得装置は、第一から第八いずれかの発明に対して、最終的に出力しない単語対に対応するクラス対を識別する２つのクラス識別子である除外クラス対を１以上格納し得る除外クラス対格納部と、１以上の除外クラス対に対応する単語対を出力する単語対から除外する単語対除外部とをさらに具備する単語対取得装置である。
【００２２】
かかる構成により、不適切な単語対を出力する可能性を低くでき、その結果、所定の関係にある単語対をより適切に取得できる。
【００２３】
また、本第十の発明の単語対取得装置は、第九の発明に対して、１以上の文章群における、各クラスに属する単語の平均出現頻度と、クラス識別子とを対に有するクラス出現頻度情報を、クラス毎に格納し得るクラス出現頻度情報格納部と、平均出現頻度が予め決められた閾値以上の差を有する２つのクラスのクラス識別子を除外クラス対として、除外クラス対格納部に蓄積する除外クラス対蓄積部とをさらに具備する単語対取得装置である。
【００２４】
かかる構成により、不適切な単語対を出力する可能性を非常に低くでき、その結果、所定の関係にある単語対をより適切に取得できる。
【００２５】
また、本第十一の発明の単語対取得装置は、第一から第十のいずれかの発明に対して、文章群格納部の１以上の文章群を用いて、同一の動詞、または同一の動詞と助詞と共起す

る回数または割合が多い単語を同一のクラスに属するように、１以上の単語クラス情報を取得する単語クラス情報取得部をさらに具備し、単語クラス情報格納部の単語クラス情報は、単語クラス情報取得部が取得した単語クラス情報である単語対取得装置である。

【００２６】

かかる構成により、単語クラス情報をより適切に取得できる。

【発明の効果】

【００２７】

本発明による単語対取得装置によれば、所定の関係にある単語対を適切に取得できる。

【図面の簡単な説明】

【００２８】

【図１】実施の形態１における単語対取得装置１を含む単語取得システムの概念図

【図２】同単語対取得装置の、単語対を取得する処理を行う構成要素に着目したブロック図

【図３】同単語対取得装置の、単語対を取得する処理を行う前の環境整備を行う構成要素に着目したブロック図

【図４】同文字列の係り受け解析の結果を示す図

【図５】同文字列の係り受け解析の結果を示す図

【図６】同単語対取得装置の動作について説明するフローチャート

【図７】同単語クラス情報管理表を示す図

【図８】同クラス出現頻度情報管理表を示す図

【図９】同単語対等の出力例を示す図

【図１０】同実験１における各方法の精度を示すグラフ

【図１１】同単語対等の出力例を示す図

【図１２】同実験２における各方法の精度を示すグラフ

【図１３】同単語対等の出力例を示す図

【図１４】同実験３における各方法の精度を示すグラフ

【図１５】同確率分布管理表を示す図

【図１６】同コンピュータシステムの概観図

【図１７】同コンピュータシステムのブロック図

【発明を実施するための形態】

【００２９】

以下、単語対取得装置等の実施形態について図面を参照して説明する。なお、実施の形態において同じ符号を付した構成要素は同様の動作を行うので、再度の説明を省略する場合がある。

【００３０】

（実施の形態１）

【００３１】

本実施の形態において、所定の関係を有する２つの単語対を取得する単語対取得装置について説明する。本単語対取得装置は、単語対が属するクラス対の良さ（後述するクラス対良好度）を指標として、単語対を選択する。また、本単語対取得装置は、単語対を取り出す際に利用するパターンの良さ（後述する類似度）を指標として、単語対を選択する。さらに、本単語対取得装置は、パターンと単語対の親和性（後述する親和性情報）を用いて、単語対を選択する。

【００３２】

図１は、本実施の形態における単語対取得装置１を含む単語取得システムの概念図である。単語取得システムは、単語対取得装置１と、１以上の文章群格納装置２とを含む。文章群格納装置２は、文章群を格納しているサーバ装置である。文章群格納装置２は、例えば、ウェブ上のサーバ装置であり、１以上のウェブページを格納している。かかる場合、文章群は、ウェブページである。また、単語対取得装置１は、１以上の文章群格納装置２から、文章群を取得し、当該文章群を少なくとも一時的に格納している。

【００３３】
図２および図３は、本実施の形態における単語対取得装置１のブロック図である。図２は、単語対取得装置１の構成要素のうちの、主として、単語対を取得する処理を行う構成要素に着目したブロック図である。図３は、単語対取得装置１の構成要素のうちの、主として、単語対を取得する処理を行う前の環境整備を行う構成要素に着目したブロック図である。ただし、図２、図３は、単語対取得装置１を分離した構成の一例に過ぎない。
【００３４】
単語対取得装置１は、文章群格納部１０１、単語対格納部１０２、単語クラス情報格納部１０３、シードパターン格納部１０４、パターン格納部１０５、クラス対良好度格納部１０６、パターン類似度格納部１０７、親和性情報格納部１０８、除外クラス対格納部１０９、クラス出現頻度情報格納部１１０、単語対取得部１１１、単語対蓄積部１１２、単語クラス情報取得部１１３、単語クラス情報蓄積部１１４、パターン取得部１１５、パターン蓄積部１１６、クラス対良好度算出部１１７、クラス対良好度蓄積部１１８、パターン類似度算出部１１９、パターン類似度蓄積部１２０、親和性情報算出部１２１、親和性情報蓄積部１２２、クラス対良好度取得部１２３、パターン類似度取得部１２４、親和性情報取得部１２５、スコア決定部１２６、単語対選択部１２７、単語対出力部１２８、単語対除外部１２９、除外クラス対蓄積部１３０、クラス出現頻度情報算出部１３１を備える。
【００３５】
文章群格納部１０１は、１以上の文章群を格納し得る。文章群とは、例えば、ウェブページである。ただし、文章群は何でも良い。文章群は、テキストデータ、所定のデータベースなどでも良く、その構造も問わない。文章群格納部１０１の文章群は、通信手段や放送受信手段などで取得した文章群であることは好適である。文章群格納部１０１は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。文章群格納部１０１に文章群が記憶される過程は問わない。例えば、記録媒体を介して文章群が文章群格納部１０１で記憶されるようになってもよく、通信回線等を介して送信された文章群が文章群格納部１０１で記憶されるようになってもよく、あるいは、入力デバイスを介して入力された文章群が文章群格納部１０１で記憶されるようになってもよい。
【００３６】
単語対格納部１０２は、１以上の単語対を格納し得る。単語対とは、所定の関係を有する２つの単語である。単語とは、ここでは、通常、名詞や名詞句である。ただし、形容詞などの他の品詞を単語であると考えても良い。また、所定の関係とは、例えば、原因と結果の関係、原材料と製品の関係、現象とその現象の防止手段の関係などである。所定の関係が原因と結果の関係である場合、例えば、単語対は「ウィルス」と「風邪」などである。単語対格納部１０２は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。単語対格納部１０２に単語対が記憶される過程は問わない。ただし、通常、単語対取得部１１１が取得した単語対を、単語対蓄積部１１２が単語対格納部１０２に蓄積する。
【００３７】
単語クラス情報格納部１０３は、２以上の単語クラス情報を格納し得る。単語クラス情報は、１以上の単語と１以上の単語が属するクラスを識別するクラス識別子とを対応づけて有する情報である。クラスとは、同一の動詞と良く共起する単語（通常、名詞）を同一のクラスに属する単語とする。また、同一の動詞および助詞と良く共起する単語（通常、名詞）を同一のクラスに属する単語としても良い。ここで、良く共起する、とは、予め決められた回数（頻度）または割合以上、同一の動詞、または同一の動詞および助詞と共起する場合である。単語クラス情報は、クラス識別子と１以上の単語を識別する１以上の単語識別子とを有する情報でも良い。単語クラス情報格納部１０３は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。単語クラス情報格納部１０３に単語クラス情報が記憶される過程は問わない。ただし、通常、単語クラス情報取得部１１３が取得した単語クラス情報を、単語クラス情報蓄積部１１４が単語クラス情報格納部１

０３に蓄積する。
【００３８】
シードパターン格納部１０４は、１以上のシードパターンを格納し得る。シードパターンとは、所定の関係を有する２つの単語対を取得するために利用するパターンである。シードパターンは、予め与えられたパターンである。シードパターンは、単語対や新たなパターンを取得するための元になるパターンである。また、パターンとは、２つの単語と、表現パターンを含む文字列である。パターンは、例えば、「ＸはＹを引き起こす」「ＸによるＹ」などである。ここで、ＸとＹに置き換わる２つの単語が単語対である。つまり、ＸやＹは、いわゆる変数（文字列が入る）である。
なお、シードパターン格納部１０４に格納されているシードパターンは、例えば、１０や２０などのパターンである。シードパターン格納部１０４は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。
【００３９】
シードパターン格納部１０４にシードパターンが記憶される過程は問わない。ただし、シードパターンは、通常、ユーザの手入力により、シードパターン格納部１０４に蓄積される。
【００４０】
パターン格納部１０５は、１以上のパターンを格納し得る。パターンとは、シードパターンではないパターンであり、所定の関係を有する２つの単語対を取得するために利用するパターンである。ただし、パターンの中に、シードパターンを含んでも良い。パターン格納部１０５は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。パターン格納部１０５にパターンが記憶される過程は問わない。ただし、通常、パターン取得部１１５が取得したパターンを、パターン蓄積部１１６がパターン格納部１０５に蓄積する。なお、パターンも、ユーザの手作業により蓄積されても良い。
【００４１】
クラス対良好度格納部１０６は、２つのクラスの良さを示す指標であるクラス対良好度を格納し得る。ここで、２つのクラスをクラス対という。また、２つのクラスの良さを示す指標とは、２つのクラスに属する単語対がシードパターンと良く共起する程度である。２つのクラスに属する単語対がシードパターンと良く共起するほど、良いクラス対とする。クラス対良好度は、数値である。また、良いクラス対ほど、クラス対良好度が大きい値となる。クラス対良好度格納部１０６は、通常、２つのクラスのクラス識別子と、クラス対良好度とを対で有するクラス対良好度情報を１以上格納している。また、クラス対の悪さを示す指標を用いることも、クラス対良好度を用いることも同意義であると考える。クラス対良好度がクラス対の悪さを示す指標である場合、例えば、クラス対良好度が大きければ大きいほど、悪いクラス対である。なお、クラス対良好度がクラス対の悪さを示す指標である場合、後述する数式において、例えば、クラス対良好度は逆数である、と考えて計算される。クラス対良好度格納部１０６は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。クラス対良好度格納部１０６にクラス対良好度が記憶される過程は問わない。ただし、通常、クラス対良好度算出部１１７が算出したクラス対良好度を、クラス対良好度蓄積部１１８がクラス対良好度格納部１０６に蓄積する。
【００４２】
パターン類似度格納部１０７は、パターン格納部１０５に格納されている１以上の各パターンとシードパターンとの類似度を、パターンごとに格納し得る。パターン類似度格納部１０７は、例えば、パターンを識別するパターン識別子と類似度とを対応づけて有する。また、パターン類似度格納部１０７は、例えば、パターンと、類似度とを対応づけて有しても良い。パターンとシードパターンとの類似度の算出方法は問わない。類似度の具体的な算出方法は後述する。パターン類似度格納部１０７は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。パターン類似度格納部１０７に類似度が記憶される過程は問わない。ただし、通常、パターン類似度算出部１１９が算出した類似度を、パターン類似度蓄積部１２０がパターン類似度格納部１０７に蓄積する。

【００４３】
親和性情報格納部１０８は、１以上の各単語対と１以上の各パターンとの親和性に関する情報である親和性情報を格納し得る。親和性情報は、通常、単語対とパターンとの親和性の度合いを示す数値である。親和性情報が大きいほど、単語対とパターンとの親和性の度合いが高いことを示す。親和性情報格納部１０８は、例えば、パターン識別子またはパターンと、単語対または単語対の識別子（２つの単語識別子でも良い）と、親和性情報とを対応付けて有する。また、親和性情報は、単語対とパターンとの親和性が低い度合いでも良い。かかる場合、親和性情報が小さいほど、単語対とパターンとの親和性の度合いが高いことを示す。親和性情報格納部１０８は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。親和性情報格納部１０８に親和性情報が記憶される過程は問わない。ただし、通常、親和性情報算出部１２１が算出した親和性情報を、親和性情報蓄積部１２２が親和性情報格納部１０８に蓄積する。
【００４４】
除外クラス対格納部１０９は、除外クラス対を１以上格納し得る。除外クラス対とは、最終的に出力しない単語対に対応するクラス対を示す情報である。除外クラス対は、通常、２つのクラス識別子を有する情報である。ただし、除外クラス対は、単語対など、２つのクラス識別子が取得できる元の情報でも良い。除外クラス対格納部１０９は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。除外クラス対格納部１０９に除外クラス対が記憶される過程は問わない。ただし、通常、除外クラス対蓄積部１３０が取得した除外クラス対を除外クラス対格納部１０９に蓄積する。ただし、ユーザが手入力により、除外クラス対を除外クラス対格納部１０９に蓄積しても良い。
【００４５】
クラス出現頻度情報格納部１１０は、クラス出現頻度情報をクラス毎に格納し得る。クラス出現頻度情報とは、１以上の文章群内における、各クラスに属する単語の平均出現頻度と、クラス識別子とを対に有する情報である。平均出現頻度は、図示しないクラス出現頻度情報取得部が、例えば、以下の処理により、取得したものである。クラス出現頻度情報取得部は、各クラスに属するすべての単語の、１以上の文章群内における出現頻度（ｆ１，ｆ２，・・・，ｆｎ）を取得する。次に、クラス出現頻度情報取得部は、クラスごとに、クラス内のすべての単語の平均出現頻度（（ｆ１＋ｆ２＋・・・＋ｆｎ）／ｎ）を算出する。クラス出現頻度情報格納部１１０は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。クラス出現頻度情報格納部１１０にクラス出現頻度情報が記憶される過程は問わない。ただし、通常、クラス出現頻度情報算出部１３１が算出したクラス出現頻度情報をクラス出現頻度情報格納部１１０に蓄積する。ただし、ユーザが手入力により、クラス出現頻度情報をクラス出現頻度情報格納部１１０に蓄積しても良い。
【００４６】
単語対取得部１１１は、シードパターン格納部１０４に格納されている１以上のシードパターンのいずれかを取得し、文章群格納部１０１に格納されている１以上の文章群から、取得したシードパターンと共起する１以上の単語対を取得する。シードパターンなどのパターンと単語対が共起する、とは、文の中にパターン（単語対を除く文字列）が存在し、かつ、文の中に、単語対を構成する２つの単語が出現することである。例えば、パターンが「ＸはＹを引き起こす」である場合、単語「Ｘ」や「Ｙ」は、パターン「ＸはＹを引き起こす」と共起する、という。シードパターンが、「ＸはＹを引き起こす」である場合、単語対取得部１１１は、１以上の文章群の中の文に「ウィルスが風邪を引き起こす」から、単語対「ウィルス」と「風邪」を取得する。また、シードパターンが、「ＸによるＹ」であり、１以上の文章群の中の文が「交通事故による経済的な損害に関して」である場合、単語対取得部１１１は、以下のように処理して、単語対「交通事故」と「損害」を取得する。つまり、単語対取得部１１１は、「交通事故による経済的な損害に関して」に「による」が存在することをパターンマッチングなどの言語処理技術により認識する。次に、単語対取得部１１１は、１以上の文章群の中の文「交通事故による経済的な損害に関し

て」を形態素解析し、「交通事故｜に｜よる｜経済的｜な｜損害｜に｜関して」、および各形態素の品詞を得る。そして、次に、単語対取得部１１１は、係り受け解析して、図４に示すような形態素間の係り受けの情報（矢印の情報）を得る。そして、単語対取得部１１１は、「による」に繋がる名詞「交通事故」と、「による」から繋がる名詞「損害」を取得する。この「交通事故」と「損害」が単語対である。なお、上記の形態素解析を行う技術として、ＪＵＭＡＮ（ＵＲＬ：http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html　参照）や、ＣｈａＳｅｎ（ＵＲＬ：http://chasen.naist.jp/hiki/ChaSen/　参照）などが存在し、公知技術である。また、係り受け解析を行う技術として、日本語構文解析システムＫＮＰ（ＵＲＬ：http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html　参照）などがあり、公知技術である。

【００４７】

単語対取得部１１１は、シードパターン格納部１０４に格納されている１以上のシードパターン、およびパターン格納部１０５に格納されている１以上のパターンのいずれか（通常、すべて）を用いて、単語対を取得することはさらに好適である。つまり、単語対取得部１１１は、１以上のシードパターンと１以上のパターンのいずれかを、順次、取得し、文章群格納部１０１に格納されている１以上の文章群から、シードパターンまたはパターンと共起する１以上の単語対を取得することはさらに好適である。

【００４８】

また、単語対取得部１１１は、シードパターンやパターンを用いずに単語対を取得しても良い。つまり、単語対取得部１１１は、１以上の文章群の中の各文から、２つの単語（通常、名詞）の対を取得しても良い。かかる場合、単語対取得部１１１は、１文の中に共起する１以上の単語対を取得することとなる。

【００４９】

単語対取得部１１１は、通常、ＭＰＵやメモリ等から実現され得る。単語対取得部１１１の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００５０】

単語対蓄積部１１２は、単語対取得部１１１が取得した１以上の単語対を、単語対格納部１０２に蓄積する。単語対蓄積部１１２は、通常、ＭＰＵやメモリ等から実現され得る。単語対蓄積部１１２の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００５１】

単語クラス情報取得部１１３は、文章群格納部１０１の１以上の文章群を用いて、１以上の単語クラス情報を取得する。単語クラス情報取得部１１３は、例えば、１以上の文章群の中の各文を形態素解析し、すべての動詞と助詞との組（すべての動詞でも良い）を取得する。また、単語クラス情報取得部１１３は、例えば、１以上の文章群の中の各文を形態素解析し、すべての名詞（名詞句を含む）を取得する。そして、単語クラス情報取得部１１３は、各名詞が、各動詞と助詞の組（または、各動詞）と共起する回数または割合を名詞ごとに算出する。次に、単語クラス情報取得部１１３は、名詞ごとに、各動詞と助詞の組（または、各動詞）と共起する回数または割合を要素に持つベクトルを取得する。次に、単語クラス情報取得部１１３は、名詞ごとのベクトルが予め決められた以上に類似する名詞の集合を一つのクラスに属するものとして、単語クラス情報を取得する。なお、単語クラス情報は、１以上の単語とクラス識別子とを有する情報である。また、クラスの数は、例えば、数百、または数千などの多い数である。

【００５２】

単語クラス情報取得部１１３は、通常、ＭＰＵやメモリ等から実現され得る。単語クラス情報取得部１１３の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００５３】
単語クラス情報蓄積部１１４は、単語クラス情報取得部１１３が取得した２以上の単語クラス情報を単語クラス情報格納部１０３に蓄積する。単語クラス情報蓄積部１１４は、通常、ＭＰＵやメモリ等から実現され得る。単語クラス情報蓄積部１１４の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。
【００５４】
パターン取得部１１５は、文章群格納部１０１に格納されている１以上の文章群の各文から、パターンを取得する。具体的には、例えば、パターン取得部１１５は、文章群格納部１０１に格納されている１以上の文章群の各文に対して、形態素解析および係り受け解析し、第一の名詞（名詞句を含む）を起点として、第二の名詞を終点として、起点から終点までに至る形態素の繋がりをパターンとして取得する。または、パターン取得部１１５は、起点からの形態素の繋がりと終点からの形態素の繋がりが結ばれる形態素までをパターンとして取得することはさらに好適である。例えば、１以上の文章群の文が「交通事故による経済的な損害に関して」である場合、パターン取得部１１５は、当該文を形態素解析し、「交通事故｜に｜よる｜経済的｜な｜損害｜に｜関して」を得る。また、形態素解析により、パターン取得部１１５は、第一の名詞「交通事故」と第二の名詞「損害」が名詞であることを検出する。そして、係り受け解析により、パターン取得部１１５は、図４の係り受けの情報を得る。次に、パターン取得部１１５は、第一の名詞「交通事故」を起点として、第二の名詞「損害」を終点として、起点から終点までに至る形態素の繋がり「ＸによるＹ」をパターンとして取得する。なお、ここで、第二の名詞「損害」に繋がる形態素群「経済的な」は、パターンから消去される。また、例えば、１以上の文章群の文が「交通事故による経済の損害に関して」である場合、パターン取得部１１５は、当該文を形態素解析し、「交通事故｜に｜よる｜経済｜の｜損害｜に｜関して」を得る。パターン取得部１１５は、第一の名詞「交通事故」と第二の名詞「経済」と第三の名詞「損害」が名詞であることを検出する。そして、係り受け解析により、パターン取得部１１５は、図５の係り受けの情報を得る。次に、パターン取得部１１５は、第一の名詞「交通事故」である起点からの形態素の繋がりと、第二の名詞「経済」である終点からの形態素の繋がりが結ばれる形態素「損害」までをパターンとして取得する。ここで、パターン取得部１１５は、「ＸによるＹの損害」をパターンとして取得する。
【００５５】
また、パターン取得部１１５は、与えられた２つの名詞（単語対）を用いて、パターンを取得しても良い。つまり、例えば、２つの名詞「交通事故」と「損害」とが与えられた時に、パターン取得部１１５は、「交通事故による経済的な損害に関して」に「交通事故」と「損害」とが含まれることを検知する。そして、パターン取得部１１５は、「交通事故による経済的な損害に関して」を形態素解析し、かつ、係り受け解析し、図４の係り受けの情報を得る。次に、パターン取得部１１５は、第一の名詞「交通事故」を起点として、第二の名詞「損害」を終点として、起点から終点までに至る形態素の繋がり「ＸによるＹ」をパターンとして取得する。
【００５６】
パターン取得部１１５は、通常、ＭＰＵやメモリ等から実現され得る。パターン取得部１１５の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。
【００５７】
パターン蓄積部１１６は、パターン取得部１１５が取得した１以上のパターンをパターン格納部１０５に蓄積する。パターン蓄積部１１６は、通常、ＭＰＵやメモリ等から実現され得る。パターン蓄積部１１６の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。
【００５８】

クラス対良好度算出部１１７は、２つの各クラスに属する単語対が、文章群格納部１０１の１以上の文章群の中で、１以上のシードパターンと共起する回数または割合が多いほどクラス対良好度が大きくなるようにクラス対良好度を算出する。クラス対良好度（ＣＳｃｏｒｅ（Ｃｉ，Ｃｊ，Ｐ））は、例えば、以下の数式１により算出される。なお、数式１において、クラス対良好度はクラス対の良さを示すものとしているので、クラス対良好度がクラス対の悪さを示す指標である場合、ＣＳｃｏｒｅ（Ｃｉ，Ｃｊ，Ｐ）は、例えば、数式１における算出結果の逆数になる。

【数１】

$$CScore(ci,cj,P)=\begin{cases}\dfrac{\sum_{(ni,nj)\in ci\times cj}\|(ni,P,nj)\|}{\sum_{(ni,nj)\in ci\times cj}\|(ni,*,nj)\|} & \text{if condition } \alpha \text{ holds}\\ 0 & \text{otherwise}\end{cases}$$

【００５９】

ここで、ｎｉやｎｊは名詞（単語）である。ｃｉやｃｊはクラスである。また、Ｐは、シードパターンの集合である。＊は、任意のパターンを示す。そして、｜｜（ｎｉ，Ｐ，ｎｊ）｜｜は、名詞ｎｉとｎｊが、シードパターンの集合と共起する頻度である。つまり、「｜｜（ｎｉ，Ｐ，ｎｊ）｜｜＝Σ$_{p\in P}$｜｜（ｎｉ，Ｐ，ｎｊ）｜｜」のことである。また、｜｜（ｎｉ，＊，ｎｊ）｜｜は、名詞ｎｉとｎｊが、１以上の文章群（Ｍ）の中で、任意のパターンと共起する頻度である。つまり、「｜｜（ｎｉ，＊，ｎｊ）｜｜＝Σ$_{(ni,p,nj)\in M}$｜｜（ｎｉ，ｐ，ｎｊ）｜｜」である。よって、｜｜（ｎｉ，＊，ｎｊ）｜｜は、名詞ｎｉとｎｊが、単に共起する頻度と等しい。

【００６０】

また、αは、条件を示す。また、αは、所定数の異なるシードパターンと共起しなければならないという条件である。また、αの例は、数式２である。数式２において、ｎｉやｎｊが、β（例えば、３）以上の異なるシードパターンと共起することを示す。つまり、数式２が条件（α）である場合、２以下のシードパターンとしか共起しない単語対（ｎｉやｎｊ）のクラス対良好度は、０となる。

【数２】

$$\|\{(p\in P\mid \exists (ni,nj)\in ci\times cj,(ni,p,nj)\in M\}\|\geq\beta$$

【００６１】

数式２において、Ｍは、１以上の文章群である。

【００６２】

また、数式１において、クラス対良好度（ＣＳｃｏｒｅ（ｃｉ，ｃｊ，Ｐ））は、２つの各クラスに属する単語が、１以上のシードパターンと共起する回数または割合が多いほど、その度合いが大きくなるような算出式の一例である。また、数式１において、２つの各クラスに属する単語が、シードパターン以外のパターンと共起する回数が多いほど、その度合いが小さくなるような算出式の一例である。

【００６３】

なお、クラス対良好度の代わりに、クラス対が良好でない度合いを用いた場合は、２つの各クラスに属する単語が、１以上のシードパターンと共起する回数または割合が多いほど、その度合いが小さくなるように算出される。この場合も、２つの各クラスに属する単語が、１以上のシードパターンと共起する回数または割合が多いほど、クラス対良好度が大きくなるように、クラス対良好度を算出することと同意義ととらえる。

【００６４】

クラス対良好度算出部１１７は、通常、ＭＰＵやメモリ等から実現され得る。クラス対良好度算出部１１７の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアは

ＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００６５】

クラス対良好度蓄積部１１８は、クラス対良好度算出部１１７が算出したクラス対良好度を、クラス対良好度格納部１０６に蓄積する。クラス対良好度蓄積部１１８は、通常、ＭＰＵやメモリ等から実現され得る。クラス対良好度蓄積部１１８の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００６６】

パターン類似度算出部１１９は、１以上のシードパターンと、パターン格納部１０５に格納されている各パターンとの類似度を算出する。パターン類似度算出部１１９は、通常、１以上のシードパターンと共起する単語対に対応するクラス対と、１以上の各パターンと共起する単語対に対応するクラス対との重なりが大きいほど、シードパターンとパターンとの類似度が大きくなるように、類似度を算出する。

【００６７】

パターン類似度算出部１１９は、シードパターンとパターンとの類似度を、例えば、数式３、数式４、数式５、数式６、または数式７により算出する。つまり、類似度は、$Para_1(p_{ci\times cj}, P)$、$Para_2(p_{ci\times cj}, P)$、$Para_3(p_{ci\times cj}, P)$、$Para_4(p_{ci\times cj}, P)$、または$Para_5(p_{ci\times cj}, P)$などである。数式３から７において、Ｐは、シードパターンの集合であり、ｐは、いずれかのパターンである。通常、ｐは、シードパターンでも良い。

【数３】

$$Para_1(p_{ci\times cj}, P) = \frac{||I(p_{ci\times cj}) \cap I(P_{ci\times cj})||}{||I(p_{ci\times cj}) \cup I(P_{ci\times cj})||}$$

【００６８】

数式３において、シードパターンとパターンとの類似度は、（$Para_1(p_{ci\times cj}, P)$）である。また、「$I(p_{ci\times cj})$」は、あるパターンｐと、クラスｃｉとクラスｃｊに属する単語ｎｉとｎｊとが共起するインスタンスの集合を示す。「$I(p_{ci\times cj})$」は、{（ｎｉ，ｎｊ）∈ｃｉ×ｃｊ｜（ｎｉ，ｐ，ｎｊ）∈Ｍ}である。また、「$(P_{ci\times cj})$」は、いずれかのシードパターンとクラスｃｉとクラスｃｊに属する単語ｎｉとｎｊとが共起するインスタンスを示す。そして、「$I(P_{ci\times cj}) = \cup_{p\in P} I(p_{ci\times cj})$」である。そして、｜｜$I(p_{ci\times cj}) \cap I(P_{ci\times cj})$｜｜は、「$I(p_{ci\times cj})$」と「$(P_{ci\times cj})$」の重なりのインスタンスの数である。また、｜｜$I(p_{ci\times cj}) \cup I(P_{ci\times cj})$｜｜は、「$I(p_{ci\times cj})$」と「$(P_{ci\times cj})$」の和集合のインスタンスの数である。

【００６９】

Ｐａｒａ１は、パターンｐと共起するクラスｃｉとクラスｃｊに属する単語ｎｉとｎｊと、シードパターンの集合（Ｐ）と共起するクラスｃｉとクラスｃｊに属する単語ｎｉとｎｊとのJaccard係数として算出される。Ｐａｒａ１によって、クラス依存であり、パターンｐを用いて生成される多くの単語対「$p_{ci\times cj}$」の中から、適切なクラスを選択できることとなる。

【００７０】

数式３を用いれば、１以上のシードパターンと共起する単語対に対応するクラス対と、１以上の各パターンと共起する単語対に対応するクラス対との重なりが大きいほど、シードパターンとパターンとの類似度が大きくなるように、類似度が算出される。また、数式３を用いれば、「$I(p_{ci\times cj})$」と「$(P_{ci\times cj})$」の和集合のインスタンスの数が多いほど、類似度が小さくなるように、類似度が算出される。

【００７１】

また、パターン類似度算出部１１９は、例えば、数式３を用いて、シードパターンとパ

ターンとの類似度を算出する場合、「Ｉ（$p_{ci \times cj}$）」と「（$P_{ci \times cj}$）」との交わりがないパターン（||Ｉ（$p_{ci \times cj}$）∩Ｉ（$P_{ci \times cj}$）||＝０のパターン）を除くことは好適である。かかることにより、処理の高速化が図れる。

【００７２】

また、数式３の変形として、以下のようにシードパターンとパターンとの類似度（$Para_3'$）を算出しても良い。

【００７３】

つまり、ｐと共起する単語対をベクトルの要素、パターンｐと共起する単語対の個数をその単語対のベクトルの要素の値とするベクトルＶｐをｐに対して構成する。そして、シードパターンＰと共起する単語対をベクトルの次元、Ｐと共起する単語対の個数をその単語対のベクトルの次元の値とするベクトルＶＰをＰに対して構成する。ただし、シードパターンＰは集合であるので、Ｐの各ｐに対して、ベクトルを作り、そのベクトルの和を、Ｐのベクトルとする。

【００７４】

そして、これらのベクトルの距離、または角度を算出する。距離は、|Ｖｐ－ＶＰ|（Ｖｐ，ＶＰの各ベクトルの要素の値の差の二乗の和の平方根）により算出できる。角度は、Ｖｐ・ＶＰ／|Ｖｐ|／|ＶＰ|により算出できる。なお、Ｖｐ・ＶＰは、内積（Ｖｐ，ＶＰの各ベクトルの要素の値の積の和）であり、|Ｖｐ|はベクトルの大きさ（Ｖｐのベクトルの要素の値の二乗の和の平方根）である。

【００７５】

これは、ベクトルＶｐとベクトルＶＰの類似度が大きいほど、シードパターンとパターンとの類似度が大きくなることであり、言い換えれば、上述したように、１以上の各パターンと共起する単語対に対応するクラス対との重なりが大きいほど、シードパターンとパターンとの類似度が大きくなる、ということである。

【数４】

$$Para_2(p_{ci \times cj}, P) = Para_1(p_{ci \times cj}, P) \cdot \frac{\| I(p) \cap I(P) \|}{\| I(p) \cup I(P) \|}$$

【００７６】

数式４において、クラスに独立なパターンも、類似度の算出に取り入れている。また、数式４は、数式３の「$Para_1$（$p_{ci \times cj}$，Ｐ）」を用いた変形例である。希なクラスの結合は、少しのインスタンスのみを含んでいるという問題（希薄性問題という。）がある。数式４は、この希薄性問題を解決するものである。数式４における「Ｉ（ｐ）」は、文章群（Ｍ）において、パターンｐと共起する単語対のインスタンスの集合である。「Ｉ（Ｐ）」は、シードパターンＰと共起する単語対のインスタンスの集合である。そして、||Ｉ（ｐ）∩Ｉ（Ｐ）||は、「Ｉ（ｐ）」と「Ｉ（Ｐ）」の重なりのインスタンスの数である。また、||Ｉ（ｐ）∪Ｉ（Ｐ）||は、「Ｉ（ｐ）」と「Ｉ（Ｐ）」の和集合のインスタンスの数である。なお、数式４は、クラス対の中のＪａｃｃａｒｄ係数の補足となる。つまり、数式４において、クラスに含まれる単語対に限定せず、すべての単語対に関して計算されている。

【００７７】

数式４も数式３と同様に、１以上のシードパターンと共起する単語対に対応するクラス対と、１以上の各パターンと共起する単語対に対応するクラス対との重なりが大きいほど、シードパターンとパターンとの類似度が大きくなるように、類似度が算出される。また、「Ｉ（$p_{ci \times cj}$）」と「（$P_{ci \times cj}$）」の和集合のインスタンスの数が多いほど、類似度が小さくなるように、類似度が算出される。また、数式４を用いれば、「Ｉ（ｐ）」と「Ｉ（Ｐ）」の重なりのインスタンスの数が多いほど、シードパターンとパターンとの類似度が大きくなるように、類似度が算出される。さらに、数式４を用いれば、「Ｉ（ｐ）」と「Ｉ（Ｐ）」の和集合のインスタンスの数が多いほど、類似度が小さくなるように、類似度が算出される。

【数５】

$$Para_3(p_{ci\times cj},P)=\frac{2\cdot\|I(p_{ci\times cj})\cap I(P_{ci\times cj})\|}{\|I(p_{ci\times cj})\|+\|I(P_{ci\times cj})\|}$$

【００７８】

数式５において、｜｜Ｉ（$p_{ci\times cj}$）｜｜＋｜｜Ｉ（$P_{ci\times cj}$）｜｜は、あるパターンｐと、クラスｃｉとクラスｃｊに属する単語ｎｉとｎｊとが共起するインスタンスの集合の数と、シードパターンＰとクラスｃｉとクラスｃｊに属する単語ｎｉとｎｊとが共起するインスタンスの集合の数との和である。なお、数式の変形として、Ｐａｒａ３の分母の｜｜Ｉ（$p_{ci\times cj}$）｜｜＋｜｜Ｉ（$P_{ci\times cj}$）｜｜を、｜｜Ｉ（$p_{ci\times cj}$）｜｜×｜｜Ｉ（$P_{ci\times cj}$）｜｜などと変形しても良い。また、Ｐａｒａ３の分母について、｜｜Ｉ（$p_{ci\times cj}$）｜｜と｜｜Ｉ（$P_{ci\times cj}$）｜｜の重み付けを行って、和算または積算を行っても良い。つまり、数式５は、｜｜Ｉ（$p_{ci\times cj}$）｜｜と｜｜Ｉ（$P_{ci\times cj}$）｜｜とをパラメータとする減少関数であれば良い。また、数式５は、｜｜Ｉ（$p_{ci\times cj}$）∩Ｉ（$P_{ci\times cj}$）｜｜をパラメータとする増加関数であれば良い。

【数６】

$$Para_4(p_{ci\times cj},P)=\frac{\|I(p_{ci\times cj})\cap I(P_{ci\times cj})\|}{\max(\|I(p_{ci\times cj})\|,\|I(P_{ci\times cj})\|)}$$

【００７９】

数式６において、ｍａｘ（｜｜Ｉ（$p_{ci\times cj}$）｜｜，｜｜（$P_{ci\times cj}$）｜｜）は、クラスｃｉとクラスｃｊに属する単語ｎｉとｎｊとが共起するインスタンスの集合の数と、シードパターンＰとクラスｃｉとクラスｃｊに属する単語ｎｉとｎｊとが共起するインスタンスの集合の数とのうちの大きい方の数である。数式６において、｜｜Ｉ（$p_{ci\times cj}$）∩Ｉ（$P_{ci\times cj}$）｜｜をパラメータとする増加関数であれば良い。

【数７】

$$Para_5(p_{ci\times cj},P)=\frac{1}{2}\left(D_{KL}\left(p_1\,\|\,\frac{p_1+p_2}{2}\right)+D_{KL}\left(p_2\,\|\,\frac{p_1+p_2}{2}\right)\right)$$

【００８０】

また、数式７において、D_{KL}（p_1｜｜p_2）は、数式８のように示される。数式８におけるD_{KL}（p_1｜｜p_2）は、確率分布ｐ１とｐ２とのKullback-Leiblerダイバージェンス（ＫＬダイバージェンスとも言う。）である。Kullback-Leiblerダイバージェンスについては、「風間淳一，De Saeger，Stijn，鳥澤健太郎，村田真樹「係り受けの確率的クラスタリングを用いた大規模類似語リストの作成」言語処理学会第15回年次大会(NLP 2009)」等に説明されている。Kullback-Leiblerダイバージェンスは、公知であるので、詳細な説明を省略する。

【数８】

$$D_{KL}(p_1\,\|\,p_2)=\sum_{(ni,nj)}p_1(ni,nj)\log_2\frac{p_1(ni,nj)}{p_2(ni,nj)}$$

【００８１】

数式７、８において、p_1とp_2は、クラス対ｃｉ×ｃｊに属する単語対（ｎｉ、ｎｊ）と、$p_{ci\times cj}$とが共起する確率分布である。p_2は、クラス対ｃｉ×ｃｊに属する単語対（ｎｉ、ｎｊ）と、Ｐとが共起する確率分布である。

【００８２】

また、パターンｐと共起する単語対をベクトルの次元、ｐと共起する単語対の個数をｐの総出現数で割った値を、その単語対のベクトルの次元の値とするベクトルＶｐをｐに対して作成する。そして、この各ベクトルの要素（ｎｉ，ｎｊ）の値が、ｐ１（ｎｉ，ｎｊ）である。

【００８３】
また、シードパターンＰと共起する単語対をベクトルの次元、Ｐと共起する単語対の個数をＰの総出現数で割った値を、その単語対のベクトルの次元の値とするベクトルＶＰをＰに対して作成する。そして、この各ベクトルの要素（ｎｉ，ｎｊ）の値が、ｐ２（ｎｉ，ｎｊ）である。
【００８４】
なお、ＫＬダイバージェンスも、ベクトル同士の類似度が大きいものをとる指標である。つまり、ＫＬダイバージェンスにおいて、例えば、p_1とp_2が同じである場合、ＤＫＬのp_1/p_2が１になり、$\log_2 p_1/p_2$が０になり、ＫＬダイバージェンスも０になる。また、p_1とp_2とが異なる値の場合、最終的なＫＬダイバージェンスの値は正の値となる。
【００８５】
パターン類似度算出部１１９は、通常、ＭＰＵやメモリ等から実現され得る。パターン類似度算出部１１９の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。
【００８６】
パターン類似度蓄積部１２０は、パターン類似度算出部１１９が算出したパターン類似度を、パターンごとに、パターン類似度格納部１０７に蓄積する。
【００８７】
パターン類似度蓄積部１２０は、通常、ＭＰＵやメモリ等から実現され得る。パターン類似度蓄積部１２０の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。
【００８８】
親和性情報算出部１２１は、単語対とパターンとの親和性情報を算出する。親和性情報算出部１２１は、単語対取得部１１１が取得した１以上の単語対と、１以上の各パターンとが共起する回数または割合が多いほど、大きくなるように親和性情報を算出することは好適である。親和性情報算出部１２１は、例えば、数式９、または数式１０により、単語対とパターンとの親和性を算出する。
【００８９】
数式９において、親和性情報（Ａｓｓｏｃ１）は、定数の１であるので、後述するスコア算出において、親和性情報が考慮されないことと同様である。
【００９０】
数式１０において、｜｜（ｎ，ｐ，ｎ'）｜｜は、単語対（ｎ，ｎ'）と、パターンｐとが共起する頻度である。つまり、親和性情報算出部１２１は、かかる頻度が大きいほど、大きくなるように親和性情報を算出することとなる。また、｜｜（ｎ，＊，ｎ'）｜｜は、単語対（ｎ，ｎ'）が任意のパターンと共起する（つまり、単語対（ｎ，ｎ'）の出現の）頻度である。さらに、｜｜（＊，ｐ，＊）｜｜は、パターンｐの出現頻度である。つまり、親和性情報は、単語対（ｎ，ｎ'）が任意のパターンと共起する頻度が高ければ高いほど、小さな値となる。また、親和性情報は、パターンｐの出現頻度が高ければ高いほど、小さな値となる。
【数９】

$$Assoc_1(n,p,n')=1$$

【数１０】

$$Assoc_2(n,p.n')=\log\frac{\|(n,p.n')\|}{\|(n,*,n')\|\|(*,p,*)\|}$$

【００９１】
親和性情報算出部１２１は、通常、ＭＰＵやメモリ等から実現され得る。親和性情報算

出部１２１の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００９２】

親和性情報蓄積部１２２は、親和性情報算出部１２１が算出した親和性情報を、親和性情報格納部１０８に蓄積する。親和性情報蓄積部１２２は、通常、単語対とパターンと親和性情報とを対応付けて、親和性情報格納部１０８に蓄積する。親和性情報蓄積部１２２は、通常、ＭＰＵやメモリ等から実現され得る。親和性情報蓄積部１２２の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００９３】

クラス対良好度取得部１２３は、単語対取得部１１１が取得した１以上の単語対が有する各単語が属する２つのクラスのクラス対良好度をクラス対良好度格納部１０６から取得する。ここで、通常、クラス対良好度取得部１２３は、２つのクラスの２つのクラス識別子であるクラス識別子対（２つのクラス識別子）を単語クラス情報格納部１０３から取得し、該クラス識別子対に対応するクラス対良好度をクラス対良好度格納部１０６から取得する。クラス対良好度取得部１２３は、通常、ＭＰＵやメモリ等から実現され得る。クラス対良好度取得部１２３の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００９４】

パターン類似度取得部１２４は、シードパターンとパターンとの類似度を、パターン類似度格納部１０７から取得する。パターン類似度取得部１２４は、例えば、スコア算出対象のパターンを識別するパターン識別子に対応する類似度を、パターン類似度格納部１０７から取得する。パターン類似度取得部１２４は、通常、ＭＰＵやメモリ等から実現され得る。パターン類似度取得部１２４の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００９５】

親和性情報取得部１２５は、親和性情報を、親和性情報格納部１０８から取得する。親和性情報取得部１２５は、例えば、スコア算出対象のパターンおよびスコア算出対象の単語対に対応する親和性情報を、親和性情報格納部１０８から取得する。親和性情報取得部１２５は、通常、ＭＰＵやメモリ等から実現され得る。親和性情報取得部１２５の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００９６】

スコア決定部１２６は、クラス対良好度取得部１２３が取得したクラス対良好度を用いて、単語対取得部１１１が取得した各単語対のスコアを決定する。スコア決定部１２６は、クラス対良好度を増加関数とする演算式により、スコアを決定する。また、スコア決定部１２６は、パターン類似度格納部１０７に格納されている１以上の各パターンとシードパターンとの類似度をも用いて、単語対取得部１１１が取得した各単語対のスコアを決定することは好適である。かかる場合、スコア決定部１２６は、類似度を増加関数とする演算式により、スコアを決定する。また、スコア決定部１２６は、親和性情報格納部１０８の親和性情報をも用いて、単語対取得部１１１が取得した各単語対のスコアを決定することは好適である。かかる場合、スコア決定部１２６は、親和性情報を増加関数とする演算式により、スコアを決定する。

【００９７】

また、スコア決定部１２６は、数式１１に示すように、クラス対良好度、シードパターンとパターンとの類似度、および親和性情報との積が最も大きいシードパターンまたはパターンにおけるスコアを、各単語対のスコアとして決定することは好適である。

【数１１】

$$Score(ni,nj,P)=\max_{ci\in classes(ni),cj\in classes(nj),(ni,p,nj)\in M}\{CScore(ci,cj,P)\cdot Para(p_{ci\times cj},P)\cdot Assoc(ni,p,nj)\}$$

【００９８】

また、スコア決定部１２６は、例えば、数式１１におけるＰａｒａは、上述したＰａｒａ１からＰａｒａ５のいずれかが適用できる。また、スコア決定部１２６は、数式１１におけるＡｓｓｏｃは、上述したＡｓｓｏｃ１またはＡｓｓｏｃ２のいずれかが適用できる。つまり、数式１１は、さらに具体的には、以下の数式１２、または数式１３、または数式１４等でも良い。数式１２から数式１４において、引数、および演算子「ｍａｘ」は省略されている。なお、数式１２により、スコアを算出する方法を、Ｃｌａｓｓ　Ｄｅｐｅｎｄｅｎｔ　Ｉ（ＣＤ－Ｉ）という。また、数式１３により、スコアを算出する方法を、Ｃｌａｓｓ　Ｄｅｐｅｎｄｅｎｔ　ＩＩ（ＣＤ－ＩＩ）という。さらに、数式１４により、スコアを算出する方法を、Ｃｌａｓｓ　Ｄｅｐｅｎｄｅｎｔ　ＩＩＩ（ＣＤ－ＩＩＩ）という。

【数１２】

$$Score=CScore\cdot Para_1\cdot Assoc_1$$

【数１３】

$$Score=CScore\cdot Para_1\cdot Assoc_2$$

【数１４】

$$Score=CScore\cdot Para_2\cdot Assoc_2$$

【００９９】

数式１１から数式１４において、スコアは、Ｃｓｃｏｒｅ、Ｐａｒａ、およびＡｓｓｏｃの３つの値の積により算出された。ただし、スコアは、３つの値の和で算出されても良いし、スコアは、$Cscore^2\times Para\times Assoc$により算出されても良い。つまり、スコアは、Ｃｓｃｏｒｅ、Ｐａｒａ、およびＡｓｓｏｃをパラメータとして算出されれば良い。また、スコアは、通常、Ｃｓｃｏｒｅが大きいほど大きな値となり、Ｐａｒａが大きいほど大きな値となり、Ａｓｓｏｃが大きいほど大きな値となる。

【０１００】

スコア決定部１２６は、通常、ＭＰＵやメモリ等から実現され得る。スコア決定部１２６の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０１０１】

単語対選択部１２７は、スコア決定部１２６が決定したスコアが予め決められた条件を満たすほど、スコアが高い１以上の単語対を取得する。単語対選択部１２７は、通常、スコアにより単語対をソートして、例えば、スコアが閾値以上の単語対を取得する。または、単語対選択部１２７は、スコアにより単語対をソートして、スコアが上位から所定数（例えば、１０００）の単語対を取得するなどしても良い。また、単語対選択部１２７は、スコアにより単語対をソートして、例えば、単語対出力部１２８が、スコアの上位から降順に、すべての単語対を出力するようにしても良い。かかる場合も、単語対選択部１２７は、１以上の単語対を取得し、単語対出力部１２８は、１以上の単語対を出力したこととなる。

【０１０２】

単語対選択部１２７は、通常、ＭＰＵやメモリ等から実現され得る。単語対選択部１２７の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０１０３】

単語対出力部１２８は、単語対選択部１２７が取得した１以上の単語対を出力する。こ

こで、出力とは、ディスプレイへの表示、プロジェクターを用いた投影、プリンタへの印字、音出力、外部の装置への送信、記録媒体への蓄積、他の処理装置や他のプログラムなどへの処理結果の引渡しなどを含む概念である。単語対出力部１２８は、ディスプレイやスピーカー等の出力デバイスを含むと考えても含まないと考えても良い。単語対出力部１２８は、出力デバイスのドライバーソフトまたは、出力デバイスのドライバーソフトと出力デバイス等で実現され得る。

【０１０４】

単語対除外部１２９は、除外クラス対格納部１０９に格納されている１以上のいずれかの除外クラス対に対応する単語対を出力する単語対から除外する。また、単語対出力部１２８は、単語対除外部１２９が除外した単語対について、通常、出力しない。ここで、除外するとは、通常、削除する、意味である。ただし、除外するとは、スコアを低くすることや、当該単語対の順位を下げる（例えば、最下位にする）などのことも含んでも良い。単語対除外部１２９は、通常、ＭＰＵやメモリ等から実現され得る。単語対除外部１２９の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０１０５】

除外クラス対蓄積部１３０は、平均出現頻度が予め決められた閾値以上の差を有する２つのクラスのクラス識別子を除外クラス対として、除外クラス対格納部１０９に蓄積する。閾値は、例えば、２５倍である。除外クラス対蓄積部１３０は、通常、ＭＰＵやメモリ等から実現され得る。除外クラス対蓄積部１３０の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０１０６】

次に、単語対取得装置１の動作について、図６のフローチャートを用いて説明する。図６のフローチャートにおいて、単語対格納部１０２、単語クラス情報格納部１０３、シードパターン格納部１０４、パターン格納部１０５、クラス対良好度格納部１０６、パターン類似度格納部１０７、親和性情報格納部１０８、および除外クラス対格納部１０９の構成要素の中に、各構成要素が格納し得る情報が格納された後、所定の関係にある単語対を出力する処理について説明する。

【０１０７】

（ステップＳ６０１）スコア決定部１２６は、カウンタｉに１を代入する。

【０１０８】

（ステップＳ６０２）スコア決定部１２６は、単語対格納部１０２（または、単語対取得部１１１が取得した単語対のうち）に、ｉ番目の単語対が存在するか否かを判断する。ｉ番目の単語対が存在すればステップＳ６０３に行き、ｉ番目の単語対が存在しなければステップＳ６１８に行く。

【０１０９】

（ステップＳ６０３）スコア決定部１２６は、ｉ番目の単語対を取得する。

【０１１０】

（ステップＳ６０４）スコア決定部１２６は、ｉ番目の単語対に対するクラス対を取得する。ここで、クラス対とは、２つのクラス識別子であっても良い。

【０１１１】

（ステップＳ６０５）スコア決定部１２６は、ステップＳ６０４で取得したクラス対が、除外クラス対格納部１０９に格納されている除外クラス対であるか否かを判断する。除外クラス対であればステップＳ６１７に行き、除外クラス対でなければステップＳ６０６に行く。

【０１１２】

（ステップＳ６０６）スコア決定部１２６は、クラス対良好度格納部１０６から、ステップＳ６０４で取得したクラス対に対応するクラス対良好度を取得する。

【０１１３】

（ステップＳ６０７）スコア決定部１２６は、カウンタｊに１を代入する。
【０１１４】
（ステップＳ６０８）スコア決定部１２６は、ｊ番目のパターンが、パターン格納部１０５または、シードパターン格納部１０４に存在するか否かを判断する。ｊ番目のパターンが存在すればステップＳ６０９に行き、存在しなければステップＳ６１５に行く。
【０１１５】
（ステップＳ６０９）スコア決定部１２６は、ｊ番目のパターンをパターン格納部１０５または、シードパターン格納部１０４から取得する。
【０１１６】
（ステップＳ６１０）スコア決定部１２６は、ｊ番目のパターンに対応する類似度を、パターン類似度格納部１０７から取得する。
【０１１７】
（ステップＳ６１１）スコア決定部１２６は、ｉ番目の単語対、およびｊ番目のパターンに対応する親和性情報を、親和性情報格納部１０８から取得する。
【０１１８】
（ステップＳ６１２）スコア決定部１２６は、ステップＳ６０６で取得したクラス対良好度、ステップＳ６１０で取得した類似度、およびステップＳ６１１で取得した親和性情報を用いて、ｉ番目の単語対、およびｊ番目のパターンに対応するスコアを算出する。
【０１１９】
（ステップＳ６１３）スコア決定部１２６は、ステップＳ６１２で算出したｊ番目のパターンに対応するスコアを、バッファに一時蓄積する。
【０１２０】
（ステップＳ６１４）スコア決定部１２６は、カウンタｊを１、インクリメントする。ステップＳ６０８に戻る。
【０１２１】
（ステップＳ６１５）スコア決定部１２６は、ステップＳ６１３でバッファに一時蓄積したスコアの中で、最大のスコアを取得する。
【０１２２】
（ステップＳ６１６）スコア決定部１２６は、ステップＳ６１５で取得したスコアを、ｉ番目の単語対と対応付けて蓄積する。
【０１２３】
（ステップＳ６１７）スコア決定部１２６は、カウンタｉを１、インクリメントする。ステップＳ６０２に戻る。
【０１２４】
（ステップＳ６１８）単語対選択部１２７は、ステップＳ６１６で蓄積したスコアをキーとして、単語対をソートする。
【０１２５】
（ステップＳ６１９）単語対選択部１２７は、ステップＳ６１８でソートした単語対のうち、予め決められた条件を満たすほど、スコアが高い１以上の単語対を取得する。
【０１２６】
（ステップＳ６２０）単語対出力部１２８は、ステップＳ６１９で取得された１以上の単語対を出力し、処理を終了する。
【０１２７】
なお、図６のフローチャートにおいて説明しなかったが、単語対格納部１０２の単語対、単語クラス情報格納部１０３の単語クラス情報、シードパターン格納部１０４のシードパターン、パターン格納部１０５のパターン、クラス対良好度格納部１０６のクラス対良好度、パターン類似度格納部１０７の類似度、親和性情報格納部１０８の親和性情報、および除外クラス対格納部１０９の除外クラス対について、それぞれ上述した処理により、格納される。
【０１２８】

また、図６のフローチャートにおいて、除外クラス対に対応する単語対を処理から除くことは、ステップＳ６０５において、行われた。しかし、除外クラス対に属する単語対を、出力する単語対から除く処理は、他のタイミング（例えば、出力する直前など）でも良い。

【０１２９】

以下、本実施の形態における単語対取得装置１の具体的な動作について説明する。　今、文章群格納部１０１は、ウェブ上の１以上のウェブページを格納している。また、単語対格納部１０２は、１以上のウェブページから取得した名詞である単語の対を多数格納している。

【０１３０】

また、単語クラス情報格納部１０３は、例えば、図７に示すような単語クラス情報管理表を保持している。図７に示す単語クラス情報管理表は、クラス識別子「C_{290}」および「C_{471}」の単語クラス情報のみを示している。また、一の単語が複数のクラスに属することもあり得る。なお、本単語クラス情報管理表は、例えば、単語クラス情報取得部１１３が上述した処理により、取得した情報である。

【０１３１】

また、クラス出現頻度情報格納部１１０は、図８に示すクラス出現頻度情報管理表を保持している。クラス出現頻度情報管理表は、「クラス」と「平均出現頻度」とを有するレコードである。「クラス」は、クラス識別子が属性値となる。「平均出現頻度」は、クラス識別子で識別されるクラスに属する単語対の平均出現頻度が属性値となる。

【０１３２】

かかる状況において、３つの実験を行った。実験１は、原因と結果の単語対を取得する実験である。実験２は、製品と材料の単語対を取得する実験である。実験３は、現象と防止手段を取得する実験である。

【０１３３】

また、３つの実験において、４つのベースライン方法と、本願の単語対取得装置１による方法とを比較する実験を行った。４つのベースライン方法のうちの第一の方法は、Ｅｓｐｒｅｓｓｏ（ＥＳＰ）と呼ばれる方法である（非特許文献１参照）。ＥＳＰは、上述したように、取り出したい単語対を少量与えて、当該単語対からパターンを取得する。そして、従来の単語対取得装置は、その取得したパターンと共起する単語対を取得するものである。また、ＥＳＰは、反復するブートストラップ方法である。

【０１３４】

また、４つのベースライン方法のうちの第二の方法は、単語対取得装置１とは異なり、クラスを用いない方法である。つまり、この第二の方法は、Ｓｉｎｇｌｅ　Ｃｌａｓｓ（ＳＣ）と呼び、数式１５により、単語対のスコアが算出される。

【数１５】

$$\mathrm{Score}(n,n',P)=\max_{(n,p,n')\in M}\left\{\frac{\|I(p)\cap I(P)\|}{\|I(p)\cup I(P)\|}\cdot \mathrm{Assoc}_2(n,p,n')\right\}$$

【０１３５】

数式１５において、「Ｉ（ｐ）」は、パターンｐと共起する単語対のインスタンス、「Ｉ（Ｐ）」は、シードパターンＰと共起する単語対のインスタンスである。また、｜｜Ｉ（ｐ）∩Ｉ（Ｐ）｜｜は、「Ｉ（ｐ）」と「Ｉ（Ｐ）」の重なり（差集合）のインスタンスの数である。また、｜｜Ｉ（ｐ）∪Ｉ（Ｐ）｜｜は、「Ｉ（ｐ）」と「Ｉ（Ｐ）」の和集合のインスタンスの数である。

【０１３６】

また、４つのベースライン方法のうちの第三の方法、および第四の方法は、ランダムベ

ースラインメソッドである。第三の方法は、「Ｒ－Ｉ」という。Ｒ－Ｉは、１以上の文章群から、パターンｐと共起する単語対を取得する方法である。第四の方法は、「Ｒ－ＩＩ」という。Ｒ－ＩＩは、１以上の文章群から、シードパターンＰと共起する単語対を取得する方法である。

【０１３７】

また、単語対取得装置１による方法とは、上述したＣＤ－Ｉ、ＣＤ－ＩＩ、ＣＤ－ＩＩＩの３つである。

【０１３８】

また、３つの各実験で、すべての方法に与えるシードパターンは同じである。ただし、当然ながら、３つの実験で利用するシードパターンは異なる。そして、３名の判断者が、各方法が出力した単語対が正しいか否かを判断した。

【０１３９】

また、各方法において、出力であるランク付けされた単語対の集合を、セグメントに分割した。セグメントとは、例えば、上位５０００、上位５０００から１５０００、上位１５０００から３５０００、および上位３５０００から７５０００である。そして、各方法において、各セグメントから、ランダムに１００のサンプル（単語対）を取得した。そして、すべてのセグメントにおいて、単語対が所定の関係の単語対である正解率（精度）を算出した。なお、実験において、２つの評価基準を適用した。一つ目は、３名が正解とした場合のみ、単語対を正解とする「厳しい（strict）」判断、２つ目は、過半数（２名）が正解とした場合も、単語対を正解とする「寛大な（lenient）」判断である。また、評価のために、５００の単語対のストップワードリストを使用した。このようにすることで、各方法の出力から代名詞の対、名詞化の対、およびストップワードの対を除外できた。（実験１）

【０１４０】

実験１は、原因と結果の単語対を取得する実験である。実験１において、シードパターン格納部１０４には、例えば、「ＸはＹを引き起こす」「ＸがＹの原因となる」などの２０のシードパターンを格納した。

【０１４１】

単語対取得装置１のＣＤ－ＩＩＩによる方法では、図９に示すような単語対等の出力が得られた。図９において、クラス対、ランク、および単語対を示す。ランクは、スコアにより付けられた順位である。ＣＤ－ＩＩＩを用いた場合、予期しない単語対が、Ｗｅｂ（１以上の文章群）から取得できた。これは、Ｗｅｂから、知らない、かつ有用な単語対（結果と原因の単語対）が取得できることを示している。図１０は、実験１における各方法の精度を示すグラフである。図１０において、横軸（Ｓａｍｐｌｅｓ　Ｒａｎｋｅｄ　ｂｙ　Ｓｃｏｒｅ）は、スコアにより、ランク付けしたサンプル（単語対）を示し、縦軸（Ｐｒｅｃｉｓｉｏｎ（％））は、単語対の正解率（精度）を示す。図１０によれば、単語対取得装置１のＣＤ－ＩＩＩの方法（寛大な（lenient）ケース）において、トップ６０，０００の単語対の精度は７０％以上であり、トップ３０，０００の単語対の精度は８０％以上である。これは、ＥＳＰやＳＣと比較して、ＣＤ－ＩＩＩの精度が極めて高いことを示す。さらに、トップ５０００のセグメントにおいて、ＣＤ－ＩＩ（寛大な（lenient）ケース）は、９３％程度の精度を達成している。以上により、ＣＤ－ＩＩやＣＤ－ＩＩＩのクラスを用いた方法は、極めて効果的であることが言える。

【０１４２】

なお、「ＸはＹを引き起こす」「ＸがＹの原因となる」というシードパターンを用いた場合、単語対取得装置１において、３２，２１３の単語対（２０，６８７のストップワードを除く）が取得できた。また、１，２８２のクラスに単語が分類できた。なお、ストップワードとは、出力から除外すべきワードである。

【０１４３】

また、８回の反復を行ったＥｓｐｒｅｓｓｏでは、「ＸによるＹ」のパターンを用いて、１，５２０，６６２の単語対が取得できた。

【０１４４】
　また、Ｒ－Ｉ（完全にランダムなベースライン方法）では、１００のランダムに取得した（ｎ，ｐ，ｎ'）タップルから、原因と結果の関係を有する単語対を取得できなかった。また、Ｒ－ＩＩの方法では、シードパターンと共起する２０，６７８のタップルからランダムに１００のタップルを選択したところ、厳しい（strict）ケースで４６％の精度、寛大な（lenient）ケースで７１％の精度であった。これらは、いずれも、単語対取得装置１の方法よりも悪いことが分かった。
（実験２）
【０１４５】
　実験２は、製品と材料の単語対を取得する実験である。実験２において、シードパターン格納部１０４には、例えば、「ＹはＸにより作られる」「ＸはＹの材料である」などの１４のシードパターンを格納した。
【０１４６】
　そして、単語対取得装置１を用いて、例えば、図１１に示すような単語対等の出力が得られた。また、単語対取得装置１の単語対取得部１１１は、１１，４７１の単語対（８，６３３のストップワードを取り除いた後）を取得できた。また、単語対取得装置１は、６２０のクラス対を取得した。
【０１４７】
　また、図１２は、実験２における各方法の精度を示すグラフである。図１２において、単語対取得装置１におけるＣＤ－ＩＩＩの方法（寛大な（lenient）ケース）は、トップ３０，０００のサンプル（セグメント）において、８０％以上の精度であることを示している。一方、Ｅｓｐｒｅｓｓ（寛大な（lenient）ケース）では、５０％程度の精度であり、ＣＤ－ＩＩＩの方法はＥｓｐｒｅｓｓと比較して、３０％以上も優れていた。
【０１４８】
　また、クラス依存の方法であるＣＤ－Ｉ、ＣＤ－ＩＩでは、上位のランクのセグメントを見れば、非常に良い結果を示している。ただし、ＣＤ－Ｉ、ＣＤ－ＩＩでは、下位のランクでは、急激に精度が低下している。なお、Ｅｓｐｒｅｓｓｏ（寛大な（lenient）ケース）は、ＣＤ－Ｉ（寛大な（lenient）ケース）の低ランク（３０，０００程度）に対しては優位である。
【０１４９】
　さらに、Ｒ－Ｉでは、正しい単語対を取得できなかった。また、Ｒ－ＩＩでは、厳しい（strict）ケースで５９％の精度、寛大な（lenient）ケースでは７２％の精度で、単語対を取得した。
（実験３）
【０１５０】
　実験３は、現象と防止手段の単語対を取得する実験である。実験３において、シードパターン格納部１０４には、例えば、「ＸによりＹを防ぐ」「Ｙを防止するＸ」などの２０のシードパターンを格納した。
【０１５１】
　単語対取得装置１による方法では、例えば、図１３に示すような単語対等の出力が得られた。また、単語対取得装置１の単語対取得部１１１は、１８，５１２の単語対（９，９４６のストップワードを取り除いた後）を取得できた。また、単語対取得装置１は、１，１６１のクラス対を取得した。
【０１５２】
　また、図１４は、実験３における各方法の精度を示すグラフである。図１２において、単語対取得装置１におけるＣＤ－ＩＩＩの方法は、トップの２つのセグメント（トップ５，０００、および5,０００から１５，０００）において、ＥｓｐｒｅｓｓｏとＳＣと比較して、優れていることが分かる。この実験３では、単語対取得装置１による方法のうち、ＣＤ－ＩＩＩのみ評価した。また、実験３において、ＣＤ－ＩＩＩの方法を、拡張した方法（ＣＤ－ＩＩＩａ、ＣＤ－ＩＩＩｂ）をも用いて、単語対取得装置１を評価した。拡

張した方法を用いたのは、シードパターンを含むパターンにより取得された単語対の中には、具体的な防止手段より、その自体を防止する行為を示す単語が含まれていたからである。例えば、単語対取得装置１が取得した単語対には、パターン「Ｙを防止するＸ」に対応して「空腹を防止する手段」や「漏れを防止するメカニズム」の中の「空腹」と「手段」、「漏れ」と「メカニズム」など単語対があった。「手段」や「メカニズム」などは、不適切であるとして、除外するようにした。これは、上述した除外クラス対格納部１０９の除外クラス対を用いて、採用しない単語対を決定することなどである。

【０１５３】

ＣＤ－ＩＩＩａは、ＣＤ－ＩＩＩと似ているが、除外クラス対格納部１０９の除外クラス対に対応する単語対を除く点が異なる。ＣＤ－ＩＩＩａにおいて、平均出現頻度が予め決められた閾値以上の差を有する２つのクラスのクラス識別子を除外クラス対としている。ここで、閾値は、２５倍である。除外クラス対を利用するのは、非常に出現頻度の高い単語が属するクラスに属する単語は、所定の関係にある良好な単語対を構成する単語になりにくい、と考えられるからである。なお、図８の平均出現頻度管理表を用いた場合、クラス９とクラス４９の平均出現頻度の差は、２５倍以上（約１３５倍）であり、クラス９とクラス４９は、除外クラス対となる。

【０１５４】

また、ＣＤ－ＩＩＩｂも、ＣＤ－ＩＩＩａと同様に、ＣＤ－ＩＩＩと似ているが、除外クラス対格納部１０９の除外クラス対に対応する単語対を除く点が異なる。ＣＤ－ＩＩＩｂにおいて、手作業で９つの除外クラス対を与えている。実験者が、ＣＤ－ＩＩＩの出力をチェックし、９つの除外クラス対を決定した。図１４によれば、概ね、ＣＤ－ＩＩＩｂが良好な結果を示している。

【０１５５】

さらに、Ｒ－Ｉでは、厳しい（strict）および寛大な（lenient）ケースともに、１００サンプルの中からは、正しい単語対を取得できなかった。また、Ｒ－ＩＩでは、厳しい（strict）ケースで５９％の精度、寛大な（lenient）ケースでは６８％の精度で、単語対を取得した。

【０１５６】

以上、本実施の形態によれば、所定の関係にある単語対を、精度高く取得できる。

【０１５７】

なお、本実施の形態において、クラス対良好度のみを利用して、単語対のスコアを算出しても良い。かかる場合、例えば、スコアは、クラス対良好度と一致しても良い。そして、この単語対取得装置１は、１以上の文章群を格納し得る文章群格納部と、１以上の単語と当該１以上の単語が属するクラスを識別するクラス識別子とを対応づけて有する２以上の単語クラス情報を格納し得る単語クラス情報格納部と、２つのクラスの良さを示す指標であるクラス対良好度を格納し得るクラス対良好度格納部と、所定の関係を有する２つの単語対を取得するために利用するパターンであるシードパターンを１以上格納し得るシードパターン格納部と、前記シードパターン格納部に格納されている１以上のシードパターンのいずれかを取得し、前記文章群格納部に格納されている１以上の文章群から、前記取得したシードパターンと共起する１以上の単語対を取得する単語対取得部と、前記単語対取得部が取得した１以上の単語対が有する各単語が属する２つのクラスのクラス対良好度を前記クラス対良好度格納部から取得するクラス対良好度取得部と、前記クラス対良好度取得部が取得したクラス対良好度を用いて、前記単語対取得部が取得した各単語対のスコアを決定するスコア決定部と、前記スコア決定部が決定したスコアが予め決められた条件を満たすほど、スコアが高い１以上の単語対を取得する単語対選択部と、前記単語対選択部が取得した１以上の単語対を出力する単語対出力部とを具備する単語対取得装置である。

【０１５８】

また、本実施の形態において、単語クラス情報格納部１０３は、単語ごとに、確率分布情報を格納していても良い。確率分布情報とは、用語が、１以上の各クラスに属する確率

の分布（集合）の情報である。確率分布情報は、ベクトルを構成し得る。クラスとは、名詞を１以上有する情報群、または、名詞を抽象化したものを１以上有する情報群である。クラスとは、例えば、同じ動詞、または同じ動詞と助詞の組と共起しやすい名詞の集合である。クラスは、適宜、隠れクラスという。なお、かかる場合、単語クラス情報格納部１０３が有する単語クラス情報は、図１５のようになる。図１５は、確率分布管理表である。なお、確率分布管理表の各データは、単語毎に各クラスに属する確率を有するベクトルであるが、このベクトルも、１以上の単語と当該１以上の単語が属するクラスのクラス識別子とを対応づけて有する単語クラス情報の一種である、と言える。なお、図１５において、クラス識別子は、ベクトル内の要素番号で決まる。

【０１５９】

そして、単語クラス情報取得部１１３は、文章群格納部１０１の１以上の文章群を用いて、図１５のような確率分布管理表を構築しても良い。つまり、例えば、１，０００，０００の名詞句と、１００，０００の動詞と助詞のセットを用いて、確率「Ｐ（＜ｖ，ｒｅｌ＞｜ｎ）」を１以上の文章群（Ｓｈｉｎｚａｔｏらが発表した以下のウェブコーパス「K. Shinzato, D. Kawahara, C. Hashimoto and S. Kurohashi. 2008. A Large-Scale Web Data Collection as A Natural Language Processing Infrastructure. In the 6th International Conference on Language Resources and Evaluation (LREC).」）から取得する。なお、＜ｖ，ｒｅｌ＞の組の発生の条件付き確率「Ｐ（＜ｖ，ｒｅｌ＞｜ｎ）」は、以下の数式１６により算出できる。「Ｐ（＜ｖ，ｒｅｌ＞｜ｎ）」は、名詞ｎの文法的なコンテキストの確率分布である。なお、ｖは動詞、ｒｅｌは助詞、ｎは名詞（名詞句を含む）である。なお、名詞は単語に相当する。また、名詞ｎと助詞ｒｅｌからなる文節が、動詞ｖを含む文節を修飾するときに、「名詞ｎが＜ｖ，ｒｅｌ＞と共起する」とする。

【数１６】

$$P(\langle v,rel\rangle|n)=\frac{\log(f(\langle v,rel,n\rangle))+1}{\sum_{\langle v,rel\rangle\in D}\log(f(\langle v,rel,n\rangle)+1} \quad \text{if } f(\langle v,rel,n\rangle)>0,$$

【０１６０】

また、数式１６において、ｌｏｇを使っているが、ｌｏｇを使わなくても良い。よって、数式１６は、「Ｐ（＜ｖ，ｒｅｌ＞｜ｎ）＝（ｆ（＜ｖ，ｒｅｌ，ｎ＞）＋１）／（Σｆ（＜ｖ，ｒｅｌ，ｎ＞）＋１）でも良い。

【０１６１】

数式１６において、「ｆ（＜ｖ，ｒｅｌ，ｎ＞）」は、＜ｖ，ｒｅｌ，ｎ＞の出現頻度である。また、Ｄは、｛＜ｖ，ｒｅｌ＞｜ｆ（＜ｖ，ｒｅｌ，ｎ＞）＞０｝として定義されるセットである。また、「ｆ（＜ｖ，ｒｅｌ，ｎ＞）＝０」の場合、「Ｐ（＜ｖ，ｒｅｌ＞｜ｎ）」は、「０」である。

【０１６２】

また、単語クラス情報取得部１１３は、「EM-based clustering」というクラス分類方法により、名詞を分類しても良い。つまり、単語クラス情報取得部１１３は、以下の数式１７で示される＜ｖ，ｒｅｌ，ｎ＞の組の出現確率を算出する。

【数１７】

$$P(\langle v,rel,n\rangle) =_{def} \sum_{a\in A} P(\langle v,rel\rangle|a)P(n|a)P(a),$$

【０１６３】

数式１７において、「ａ」は＜ｖ，ｒｅｌ＞の組および「ｎ」の隠れクラスを示す。数式１７において、確率「Ｐ（＜ｖ，ｒｅｌ＞｜ａ）」、「Ｐ（ｎ｜ａ）」および「Ｐ（ａ）」が直接的に算出できない。隠れクラス「ａ」が与えられたコーパスから取得できない

からである。

【０１６４】

「EM-based clustering」は、与えられたコーパス（１以上の文章群）から、これらの確率（「Ｐ（＜ｖ，ｒｅｌ＞｜ａ）」、「Ｐ（ｎ｜ａ）」および「Ｐ（ａ）」）を推定する。「EM-based clustering」は「Ｅステップ」と「Ｍステップ」の２つのステップからなる。「Ｅステップ」において、確率「Ｐ（＜ｖ，ｒｅｌ＞｜ａ）」が算出される。「Ｍステップ」において、「Ｅステップ」における結果を用いて、最大尤度になるまで、「Ｐ（＜ｖ，ｒｅｌ＞｜ａ）」、「Ｐ（ｎ｜ａ）」および「Ｐ（ａ）」が更新される。

【０１６５】

以上の処理により、各＜ｖ，ｒｅｌ＞、ｎ、およびａに対して、確率「Ｐ（＜ｖ，ｒｅｌ＞｜ａ）」、「Ｐ（ｎ｜ａ）」および「Ｐ（ａ）」が算出される。

【０１６６】

そして、「Ｐ（ａ｜ｎ）」は、以下の数式１８により算出される。

【数１８】

$$P(a|n)=\frac{P(n|a)P(a)}{\sum_{a\in A}P(n|a)P(a)}.$$

【０１６７】

「Ｐ（ａ｜ｎ）」は、ｎのクラスを決定するために用いられる。例えば、最大の「Ｐ（ａ｜ｎ）」を有するクラスが、ｎが属するクラスである。類似する＜ｖ，ｒｅｌ＞の組と共起する名詞句は、同じクラスに属する傾向がある。

【０１６８】

また、本明細書で記載したクラス対に関して、以下のように絞り込むことは好適である。つまり、図示しない手段または上述したいずれかの構成要素（発明特定事項）により、文章群格納部１０１の文章群から、シードパターンと共起する単語対を取り出し、当該単語対が、予め決められた数（閾値は予め格納されている）以上存在するクラス対に限定する。そして、限定されたクラス対、または限定されたクラス対の単語対を用いて、上述した処理（単語対の取得処理や、クラス対良好度の算出や、パターンの類似度の算出や、スコアの算出など）が行われる。かかることにより、処理の高速化が図れる。

【０１６９】

また、本明細書で記載した各種の数式は、技術的思想を反映する範囲で、多少の変形を加えても良いことは言うまでもない。

【０１７０】

さらに、本実施の形態における処理は、ソフトウェアで実現しても良い。そして、このソフトウェアをソフトウェアダウンロード等により配布しても良い。また、このソフトウェアをＣＤ－ＲＯＭなどの記録媒体に記録して流布しても良い。なお、このことは、本明細書における他の実施の形態においても該当する。なお、本実施の形態における情報処理装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、記憶媒体に、１以上の文章群を格納し、１以上の単語と当該１以上の単語が属するクラスを識別するクラス識別子とを対応づけて有する２以上の単語クラス情報を格納し、２つのクラスの良さを示す指標であるクラス対良好度を格納し、所定の関係を有する２つの単語対を取得するために利用するパターンであるシードパターンを１以上格納しており、コンピュータを、前記記憶媒体に格納されている１以上のシードパターンのいずれかを取得し、前記記憶媒体に格納されている１以上の文章群から、前記取得したシードパターンと共起する１以上の単語対を取得する単語対取得部と、前記単語対取得部が取得した１以上の単語対が有する各単語が属する２つのクラスのクラス対良好度を前記記憶媒体から取得するクラス対良好度取得部と、前記クラス対良好度取得部が取得したクラス対良好度を用いて、前記単語対取得部が取得した各単語対のスコアを決定するスコア決定部と、前記スコア決定部が決定したスコアが予め決められた条件を満たすほど、スコアが高い１以上の単語対を取得する単語対選択部と、前記単語対選択部が取得した１以上の単語対を出

力する単語対出力部として機能させるためのプログラム、である。
【０１７１】
また、上記プログラムにおいて、コンピュータを、２つの各クラスに属する単語対が、前記文章群格納部の１以上の文章群の中で、前記１以上のシードパターンと共起する回数または割合が多いほどクラス対良好度が大きくなるようにクラス対良好度を算出するクラス対良好度算出部として、さらに機能させ、
前記クラス対良好度算出部が算出した２つのクラスのクラス対良好度は、記憶媒体に格納されているクラス対良好度であることは好適である。
【０１７２】
また、上記プログラムにおいて、記憶媒体に、シードパターンではないパターンであり、前記所定の関係を有する２つの単語対を取得するために利用する１以上のパターン、および、前記１以上の各パターンと前記シードパターンとの類似度を、パターンごとにさらに格納し、前記単語対取得部は、前記記憶媒体に格納されている１以上のシードパターン、および前記記憶媒体に格納されている１以上のパターンのいずれかを取得し、前記記憶媒体に格納されている１以上の文章群から、前記シードパターンまたは前記パターンと共起する１以上の単語対を取得し、前記スコア決定部は、前記記憶媒体に格納されている前記１以上の各パターンと前記シードパターンとの類似度をも用いて、前記単語対取得部が取得した各単語対のスコアを決定するものとして、コンピュータを機能させるプログラムであることは好適である。
【０１７３】
また、上記プログラムにおいて、コンピュータを、前記１以上のシードパターンと共起する単語対に対応するクラス対と、前記記憶媒体に格納されている１以上の各パターンと共起する単語対に対応するクラス対との重なりが大きいほど、大きくなるように類似度を算出するパターン類似度算出部をさらに具備し、前記パターン類似度算出部が算出した類似度は、前記記憶媒体に格納されている類似度であることは好適である。
【０１７４】
また、上記プログラムにおいて、記憶媒体に、1以上の各単語対と1以上の各パターンとの親和性に関する情報である親和性情報を、さらに格納し、前記スコア決定部は、前記記憶媒体の親和性情報をも用いて、前記単語対取得部が取得した各単語対のスコアを決定するものとして、コンピュータを機能させるプログラムであることは好適である。
【０１７５】
また、上記プログラムにおいて、コンピュータを、前記単語対取得部が取得した１以上の単語対と、前記１以上の各パターンとが共起する回数または割合が多いほど、大きくなるように親和性情報を算出する親和性情報算出部としてさらに機能させ、前記記憶媒体の親和性情報は、前記親和性情報算出部が算出した親和性情報であることは好適である。
【０１７６】
また、上記プログラムにおいて、前記スコア決定部は、前記クラス対良好度、前記シードパターンとパターンとの類似度、および前記親和性情報との積が最も大きいシードパターンまたはパターンにおけるスコアを、各単語対のスコアとして決定するものとして、コンピュータを機能させるプログラムであることは好適である。
【０１７７】
また、上記プログラムにおいて、コンピュータを、前記記憶媒体に格納されている１以上の文章群の各文に対して、形態素解析および係り受け解析し、第一の名詞または名詞句を起点として、第二の名詞または名詞句を終点として、前記起点から前記終点までに至る形態素の繋がりをパターンとして取得し、または、前記起点からの形態素の繋がりと前記終点からの形態素の繋がりが結ばれる形態素までをパターンとして取得するパターン取得部としてさらに機能させ、前記記憶媒体のパターンは、前記パターン取得部が取得したパターンであることは好適である。
【０１７８】
また、上記プログラムにおいて、記憶媒体に、最終的に出力しない単語対に対応するク

ラス対を識別する２つのクラス識別子である除外クラス対を１以上格納し、コンピュータを、前記１以上の除外クラス対に対応する単語対を出力する単語対から除外する単語対除外部としてさらに機能させることは好適である。

【０１７９】

また、上記プログラムにおいて、記憶媒体に、前記１以上の文章群における、各クラスに属する単語の平均出現頻度と、クラス識別子とを対に有するクラス出現頻度情報を、クラス毎に格納し、コンピュータを、前記平均出現頻度が予め決められた閾値以上の差を有する２つのクラスのクラス識別子を除外クラス対として、前記除外クラス対格納部に蓄積する除外クラス対蓄積部としてさらに機能させることは好適である。

【０１８０】

また、図１６は、本明細書で述べたプログラムを実行して、上述した実施の形態の単語対取得装置１等を実現するコンピュータの外観を示す。上述の実施の形態は、コンピュータハードウェア及びその上で実行されるコンピュータプログラムで実現され得る。図１６は、このコンピュータシステム３４０の概観図であり、図１７は、コンピュータシステム３４０の内部構成を示す図である。

【０１８１】

図１６において、コンピュータシステム３４０は、ＦＤドライブ３４１１、ＣＤ－ＲＯＭドライブ３４１２を含むコンピュータ３４１と、キーボード３４２と、マウス３４３と、モニタ３４４とを含む。

【０１８２】

図１７において、コンピュータ３４１は、ＦＤドライブ３４１１、ＣＤ－ＲＯＭドライブ３４１２に加えて、ＭＰＵ３４１３と、ＣＤ－ＲＯＭドライブ３４１２及びＦＤドライブ３４１１に接続されたバス３４１４と、ブートアッププログラム等のプログラムを記憶するためのＲＯＭ３４１５と、ＭＰＵ３４１３に接続され、アプリケーションプログラムの命令を一時的に記憶するとともに一時記憶空間を提供するためのＲＡＭ３４１６と、アプリケーションプログラム、システムプログラム、及びデータを記憶するためのハードディスク３４１７とを含む。ここでは、図示しないが、コンピュータ３４１は、さらに、ＬＡＮへの接続を提供するネットワークカードを含んでも良い。

【０１８３】

コンピュータシステム３４０に、上述した実施の形態の単語対取得装置１等の機能を実行させるプログラムは、ＣＤ－ＲＯＭ３５０１、またはＦＤ３５０２に記憶されて、ＣＤ－ＲＯＭドライブ３４１２またはＦＤドライブ３４１１に挿入され、さらにハードディスク３４１７に転送されても良い。これに代えて、プログラムは、図示しないネットワークを介してコンピュータ３４１に送信され、ハードディスク３４１７に記憶されても良い。プログラムは実行の際にＲＡＭ３４１６にロードされる。プログラムは、ＣＤ－ＲＯＭ３５０１、ＦＤ３５０２またはネットワークから直接、ロードされても良い。

【０１８４】

プログラムは、コンピュータ３４１に、上述した実施の形態の単語対取得装置１等の機能を実行させるオペレーティングシステム（ＯＳ）、またはサードパーティープログラム等は、必ずしも含まなくても良い。プログラムは、制御された態様で適切な機能（モジュール）を呼び出し、所望の結果が得られるようにする命令の部分のみを含んでいれば良い。コンピュータシステム３４０がどのように動作するかは周知であり、詳細な説明は省略する。

【０１８５】

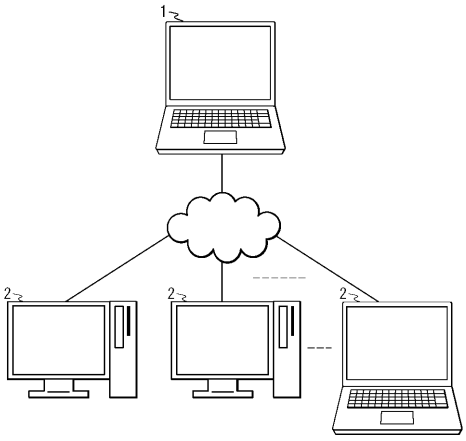
また、上記プログラムを実行するコンピュータは、単数であってもよく、複数であってもよい。すなわち、集中処理を行ってもよく、あるいは分散処理を行ってもよい。

【０１８６】

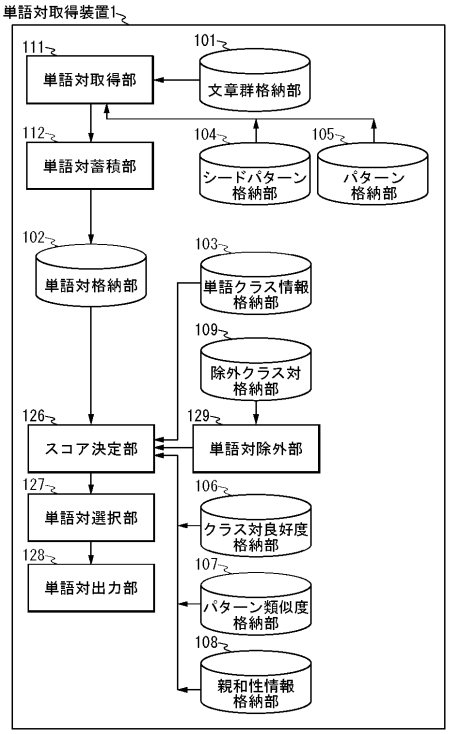
また、上記各実施の形態において、各処理（各機能）は、単一の装置（システム）によって集中処理されることによって実現されてもよく、あるいは、複数の装置によって分散処理されることによって実現されてもよい。

【０１８７】
本発明は、以上の実施の形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。
【産業上の利用可能性】
【０１８８】
以上のように、本発明にかかる単語対取得装置は、所定の関係にある単語対を適切に取得できるという効果を有し、単語対取得装置等として有用である。
【符号の説明】
【０１８９】
１　単語対取得装置
１０１　文章群格納部
１０２　単語対格納部
１０３　単語クラス情報格納部
１０４　シードパターン格納部
１０５　パターン格納部
１０６　クラス対良好度格納部
１０７　パターン類似度格納部
１０８　親和性情報格納部
１０９　除外クラス対格納部
１１０　クラス出現頻度情報格納部
１１１　単語対取得部
１１２　単語対蓄積部
１１３　単語クラス情報取得部
１１４　単語クラス情報蓄積部
１１５　パターン取得部
１１６　パターン蓄積部
１１７　クラス対良好度算出部
１１８　クラス対良好度蓄積部
１１９　パターン類似度算出部
１２０　パターン類似度蓄積部
１２１　親和性情報算出部
１２２　親和性情報蓄積部
１２３　クラス対良好度取得部
１２４　パターン類似度取得部
１２５　親和性情報取得部
１２６　スコア決定部
１２７　単語対選択部
１２８　単語対出力部
１２９　単語対除外部
１３０　除外クラス対蓄積部
１３１　クラス出現頻度情報算出部

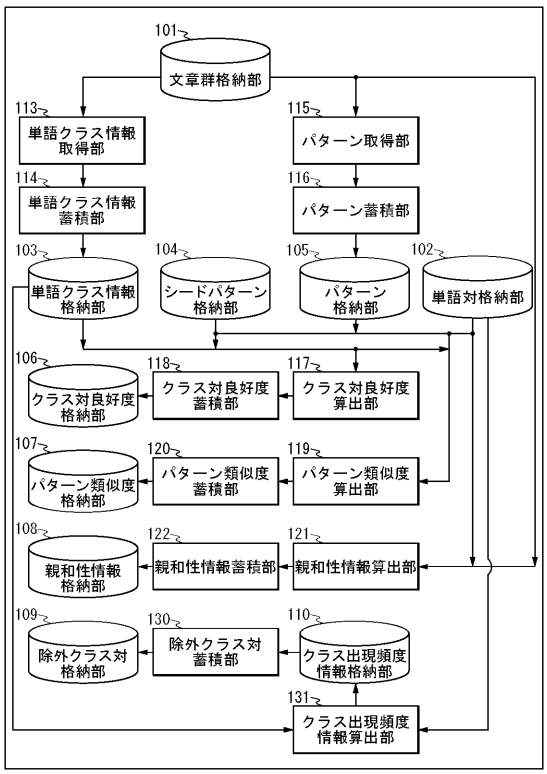
【図１】


1
2
2
2

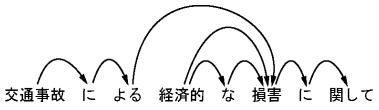
【図２】


単語対取得装置1
111
単語対取得部
101
文章群格納部
112
単語対蓄積部
104
シードパターン格納部
105
パターン格納部
102
単語対格納部
103
単語クラス情報格納部
109
除外クラス対格納部
126
スコア決定部
129
単語対除外部
127
単語対選択部
106
クラス対良好度格納部
128
単語対出力部
107
パターン類似度格納部
108
親和性情報格納部

【図３】


101
文章群格納部
113
単語クラス情報取得部
115
パターン取得部
114
単語クラス情報蓄積部
116
パターン蓄積部
103
単語クラス情報格納部
104
シードパターン格納部
105
パターン格納部
102
単語対格納部
106
クラス対良好度格納部
118
クラス対良好度蓄積部
117
クラス対良好度算出部
107
パターン類似度格納部
120
パターン類似度蓄積部
119
パターン類似度算出部
108
親和性情報格納部
122
親和性情報蓄積部
121
親和性情報算出部
109
除外クラス対格納部
130
除外クラス対蓄積部
110
クラス出現頻度情報格納部
131
クラス出現頻度情報算出部

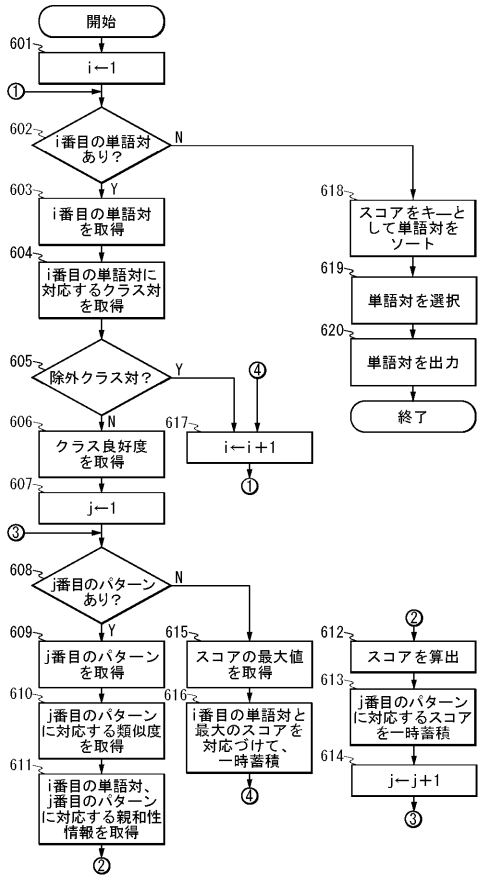
【図４】


交通事故
に
よる
経済的
な
損害
に
関して

【図５】


交通事故
に
よる
経済
の
損害
に
関して

【図６】



【図７】

C_{290}	C_{471}
歯周病	窒素酸化物
アレルギー	イオウ酸化物
ニキビ	プリオン
炎症	NO_X
⋮	⋮

【図８】

クラス	平均出現頻度
9	3581266.93
49	26505.42
385	1865229.69
494	4707922.42
118	21214.89
205	7521190.63
⋮	⋮

【図９】

クラス対	ランク	単語対
$C_{471}\times C_{290}$	22	チロシナーゼ ― そばかす
$C_{468}\times C_{290}$	62	かび ― におい
$C_{468}\times C_{290}$	274	ダニ ― 皮膚トラブル
$C_{471}\times C_{290}$	394	残留塩素 ― かゆみ
$C_{475}\times C_{1}$	5889	日本酒 ― 肥満
$C_{290}\times C_{290}$	6523	虫歯 ― 口臭
$C_{471}\times C_{1}$	17135	タウリン ― 動脈硬化

【図１１】

クラス対	ランク	単語対
$C_{176}\times C_{475}$	614	麦芽 ― ウイスキー
$C_{176}\times C_{227}$	1128	コラーゲン ― ゼリー
$C_{176}\times C_{227}$	5032	カカオ豆 ― チョコ
$C_{252}\times C_{270}$	34971	さとうきび ― 自動車燃料

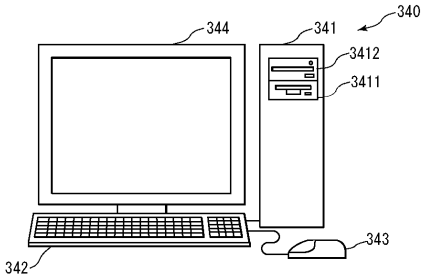
【図１３】

クラス対	ランク	単語対
$C_{172}\times C_{49}$	820	サーモスタット ― 加熱
$C_{333}\times C_{49}$	831	発砲スチロール ― 蒸発
$C_{212}\times C_{191}$	11856	手段 ― 洪水被害
$C_{240}\times C_{191}$	17627	会議 ― 事故

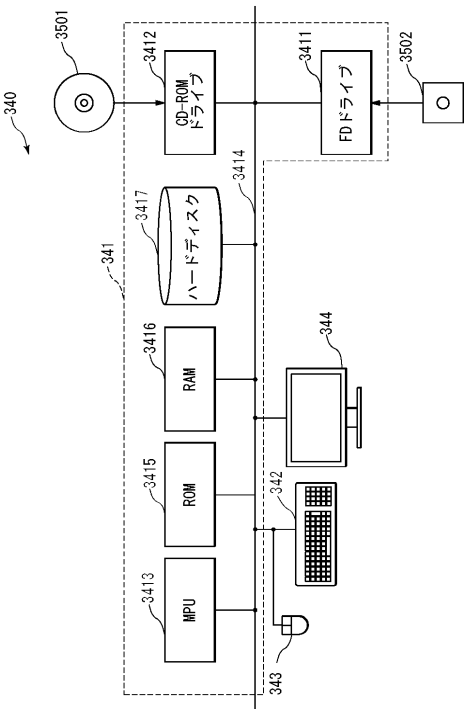
【図１５】

用語	確率分布情報
自動車	(0.1, 0.05, 0, 0.2, ・・・・・)
小型車	(0.08, 0.03, 0.02, 0.3, ・・・・・)
ハイブリッド車	(0.07, 0.04, 0.1, 0.15, ・・・・・)
⋮	⋮

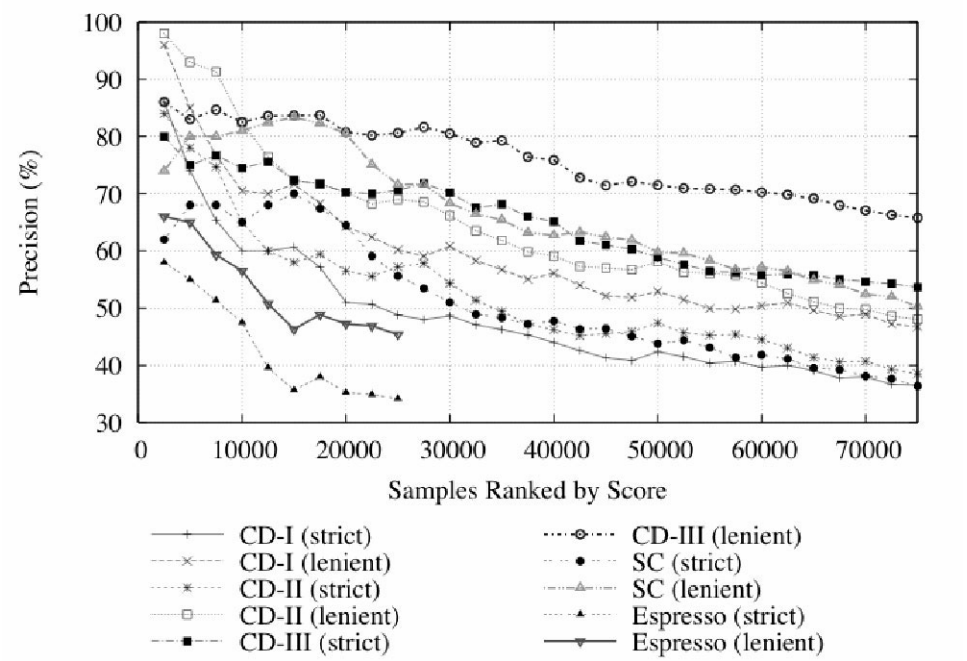
【図１６】



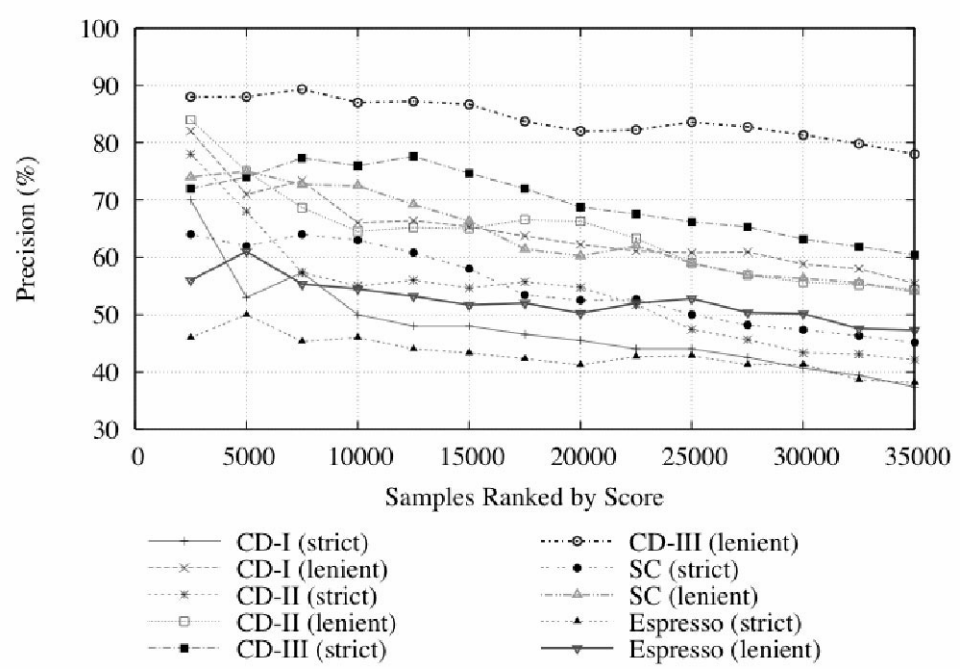
【図１７】


3501
3412
CD-ROM ドライブ
3411
FD ドライブ
3502
340
3414
3417
ハードディスク
341
3416
RAM
344
3415
ROM
342
3413
MPU
343

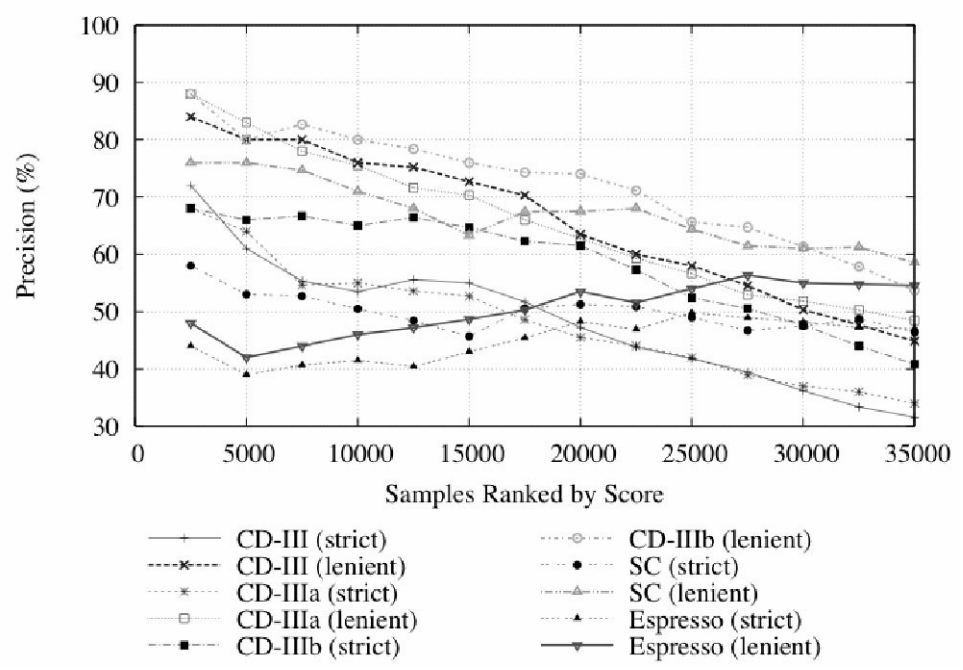
【図１０】


Precision (%)
100
90
80
70
60
50
40
30
0
10000
20000
30000
40000
50000
60000
70000
Samples Ranked by Score
CD-I (strict)
CD-I (lenient)
CD-II (strict)
CD-II (lenient)
CD-III (strict)
CD-III (lenient)
SC (strict)
SC (lenient)
Espresso (strict)
Espresso (lenient)

【図１２】


Precision (%)
100
90
80
70
60
50
40
30
0
5000
10000
15000
20000
25000
30000
35000
Samples Ranked by Score
CD-I (strict)
CD-I (lenient)
CD-II (strict)
CD-II (lenient)
CD-III (strict)
CD-III (lenient)
SC (strict)
SC (lenient)
Espresso (strict)
Espresso (lenient)

【図１４】


Precision (%)
100
90
80
70
60
50
40
30
0
5000
10000
15000
20000
25000
30000
35000
Samples Ranked by Score
CD-III (strict)
CD-III (lenient)
CD-IIIa (strict)
CD-IIIa (lenient)
CD-IIIb (strict)
CD-IIIb (lenient)
SC (strict)
SC (lenient)
Espresso (strict)
Espresso (lenient)

フロントページの続き

(72)発明者　黒田　航
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

審査官　成瀬　博之

(56)参考文献　特開２００９－２６５８８９（ＪＰ，Ａ）
特開２００３－２５６４４７（ＪＰ，Ａ）
麻野間直樹　他１名，未解析コーパスからの依存関係単語対の収集，第６１回（平成１２年後期）全国大会講演論文集（２），日本，社団法人　情報処理学会，２０００年１０月　3日，２－１４１～2－１４２

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２－１７／３０