

(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2011-59917

(P2011-59917A)

(43) 公開日 平成23年3月24日(2011.3.24)

(51) Int.Cl.	F I	テーマコード (参考)
G06F 17/27 (2006.01)	G06F 17/27 Z	5B075
G06F 17/30 (2006.01)	G06F 17/30 220Z	5B091
	G06F 17/30 170A	

審査請求 未請求 請求項の数 13 O L (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)

(71) 出願人 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

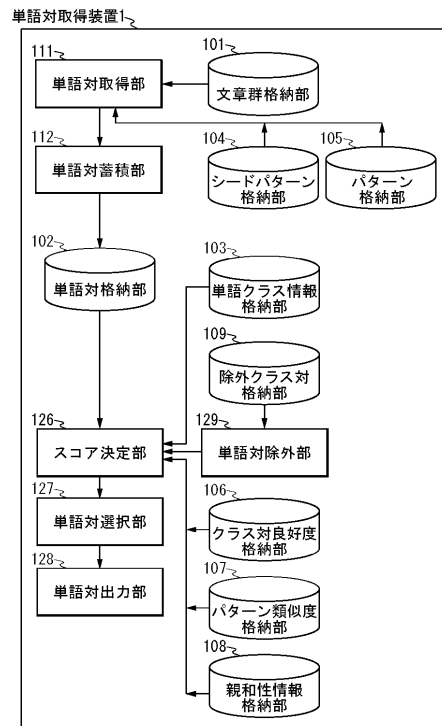
(54) 【発明の名称】 単語対取得装置、単語対取得方法、およびプログラム

(57) 【要約】

【課題】従来、所定の関係にある単語対を適切に取得できなかった。

【解決手段】単語のクラスを特定する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、1以上の文章群から、シードパターンと共に起る1以上の単語対を取得する単語対取得部と、1以上の単語対に対応するクラス対良好度を取得するクラス対良好度取得部と、クラス対良好度を用いて、各単語対のスコアを決定するスコア決定部と、スコアが予め決められた条件を満たすほど高い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により、単語対を適切に取得できる。

【選択図】図2



【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、

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。

10

【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。

20

【請求項 3】

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

30

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

40

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 から請求項 4 いずれか記載の単語対取得装置。

50

【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、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請求項 6 記載の単語対取得装置。

10

【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、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。

20

【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。

30

【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、

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、

40

50

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。

10

【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、

20

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。

30

【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）。

40

【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】

50

【0004】

しかしながら、従来の単語対取得装置においては、所定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。

10

【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。

20

【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。

30

【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。

40

【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

50

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

10

20

30

40

50

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に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。

10

【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

20

【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

30

【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】

40

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。

50

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よくよい。

【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

10

20

30

40

50

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103に蓄積する。

10

【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。

20

【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。

30

【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られる。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、と考慮して計算される。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。

40

【0042】

50

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（ f_1, f_2, \dots, f_n ）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（ $(f_1 + f_2 + \dots + f_n) / n$ ）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に起する 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

10

20

30

40

50

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 | 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を「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 | 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を「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

10

20

30

40

50

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（ $Cscore(C_i, C_j, P)$ ）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、 $Cscore(C_i, C_j, P)$ は、例えば、数式1における算出結果の逆数になる。

【数1】

$Cscore(c_i, c_j, P) =$

$$\begin{cases} \frac{\sum_{(n_i, n_j) \in c_i \times c_j} \|(n_i, P, n_j)\|}{\sum_{(n_i, n_j) \in c_i \times c_j} \|(n_i, *, n_j)\|} & \text{if condition } \alpha \text{ holds} \\ 0 & \text{otherwise} \end{cases}$$

【0059】

ここで、 n_i や n_j は名詞（単語）である。 c_i や c_j 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、 $\|(n_i, P, n_j)\|$ は、名詞 n_i と n_j が、シードパターンの集合と共起する頻度である。つまり、「 $\|(n_i, P, n_j)\| = \sum_{p \in P} \|(n_i, p, n_j)\|$ 」のことである。また、 $\|(n_i, *, n_j)\|$ は、名詞 n_i と n_j 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、「 $\|(n_i, *, n_j)\| = \sum_{m \in M} \|(n_i, m, n_j)\|$ 」である。よって、 $\|(n_i, *, n_j)\|$ は、名詞 n_i と n_j 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、 β 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、 β の例は、数式2である。数式2において、 n_i や n_j 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（ β ）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（ n_i や n_j ）のクラス対良好度は、0となる。

【数2】

$$\|\{(p \in P | \exists (n_i, n_j) \in c_i \times c_j, (n_i, p, n_j) \in M)\}\| \geq \beta$$

【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（ $Cscore(c_i, c_j, P)$ ）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。

10

20

30

40

50

【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。

10

【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。

20

【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、 $Para_1(p_{c_i \times c_j}, P)$ 、 $Para_2(p_{c_i \times c_j}, P)$ 、 $Para_3(p_{c_i \times c_j}, P)$ 、 $Para_4(p_{c_i \times c_j}, P)$ 、または $Para_5(p_{c_i \times c_j}, P)$ 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。

30

【数3】

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

【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、 $(Para_1(p_{c_i \times c_j}, P))$ である。また、「 $I(p_{c_i \times c_j})$ 」は、あるパターン p と、クラス c_i とクラス c_j に属する単語 n_i と n_j とが共起するインスタンスの集合を示す。「 $I(p_{c_i \times c_j})$ 」は、 $\{(n_i, n_j) \mid c_i \times c_j \mid (n_i, p, n_j) \in M\}$ である。また、「 $I(P_{c_i \times c_j})$ 」は、いずれかのシードパターンとクラス c_i とクラス c_j に属する単語 n_i と n_j とが共起するインスタンスを示す。そして、「 $I(P_{c_i \times c_j}) = \bigcup_{p \in P} I(p_{c_i \times c_j})$ 」である。そして、「 $\|I(p_{c_i \times c_j}) \cap I(P_{c_i \times c_j})\|$ 」は、「 $I(p_{c_i \times c_j})$ 」と「 $I(P_{c_i \times c_j})$ 」の重なりインスタンスの数である。また、「 $\|I(p_{c_i \times c_j}) \cup I(P_{c_i \times c_j})\|$ 」は、「 $I(p_{c_i \times c_j})$ 」と「 $I(P_{c_i \times c_j})$ 」の和集合のインスタンスの数である。

40

【0069】

$Para_1$ は、パターン p と共起するクラス c_i とクラス c_j に属する単語 n_i と n_j と、シードパターンの集合(P)と共起するクラス c_i とクラス c_j に属する単語 n_i と n_j とのJaccard係数として算出される。 $Para_1$ 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「 $p_{c_i \times c_j}$ 」の中から、適切なクラスを選択

50

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、「 $I(p_{c_i \times c_j})$ 」と「 $(P_{c_i \times c_j})$ 」の和集合のインスタンスの数が多きほど、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、「 $I(p_{c_i \times c_j})$ 」と「 $(P_{c_i \times c_j})$ 」との交わりがないパターン（ $||I(p_{c_i \times c_j}) \cap I(P_{c_i \times c_j})|| = 0$ 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 V_p 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 V_P 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、 $|V_p - V_P|$ （ V_p 、 V_P 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、 $V_p \cdot V_P / |V_p| / |V_P|$ により算出できる。なお、 $V_p \cdot V_P$ は、内積（ V_p 、 V_P の各ベクトルの要素の値の積の和）であり、 $|V_p|$ はベクトルの大きさ（ V_p 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 V_p とベクトル V_P 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】

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

【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「 $Para_1(p_{c_i \times c_j}, P)$ 」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、 $||I(p) \cap I(P)||$ は、「 $I(p)$ 」と「 $I(P)$ 」の重なりインスタンスの数である。また、 $||I(p) \cup I(P)||$ 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】

10

20

30

40

50

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、「 $I(p_{c_i \times c_j})$ 」と「 $(P_{c_i \times c_j})$ 」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】

$$\text{Para}_3(p_{c_i \times c_j}, P) = \frac{2 \cdot \|I(p_{c_i \times c_j}) \cap I(P_{c_i \times c_j})\|}{\|I(p_{c_i \times c_j})\| + \|I(P_{c_i \times c_j})\|}$$

10

【0078】

数式 5 において、 $\|I(p_{c_i \times c_j})\| + \|I(P_{c_i \times c_j})\|$ は、あるパターン p と、クラス c_i とクラス c_j に属する単語 n_i と n_j 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 c_i とクラス c_j に属する単語 n_i と n_j 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、 Para_3 の分母の $\|I(p_{c_i \times c_j})\| + \|I(P_{c_i \times c_j})\|$ を、 $\|I(p_{c_i \times c_j})\| \times \|I(P_{c_i \times c_j})\|$ などと変形しても良い。また、 Para_3 の分母について、 $\|I(p_{c_i \times c_j})\|$ と $\|I(P_{c_i \times c_j})\|$ 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、 $\|I(p_{c_i \times c_j})\|$ と $\|I(P_{c_i \times c_j})\|$ 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、 $\|I(p_{c_i \times c_j})\| \cdot \|I(P_{c_i \times c_j})\|$ をパラメータとする増加関数であれば良い。

20

【数 6】

$$\text{Para}_4(p_{c_i \times c_j}, P) = \frac{\|I(p_{c_i \times c_j}) \cap I(P_{c_i \times c_j})\|}{\max(\|I(p_{c_i \times c_j})\|, \|I(P_{c_i \times c_j})\|)}$$

【0079】

数式 6 において、 $\max(\|I(p_{c_i \times c_j})\|, \|I(P_{c_i \times c_j})\|)$ は、クラス c_i とクラス c_j に属する単語 n_i と n_j 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 c_i とクラス c_j に属する単語 n_i と n_j 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、 $\|I(p_{c_i \times c_j})\| \cdot \|I(P_{c_i \times c_j})\|$ をパラメータとする増加関数であれば良い。

30

【数 7】

$$\text{Para}_5(p_{c_i \times c_j}, P) = \frac{1}{2} (D_{KL}(p_1 \parallel \frac{p_1+p_2}{2}) + D_{KL}(p_2 \parallel \frac{p_1+p_2}{2}))$$

【0080】

また、数式 7 において、 $D_{KL}(p_1 \parallel p_2)$ は、数式 8 のように示される。数式 8 における $D_{KL}(p_1 \parallel p_2)$ は、確率分布 p_1 と p_2 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。

40

【数 8】

$$D_{KL}(p_1 \parallel p_2) = \sum_{(n_i, n_j)} p_1(n_i, n_j) \log_2 \frac{p_1(n_i, n_j)}{p_2(n_i, n_j)}$$

50

【0081】

数式7、8において、 p_1 と p_2 は、クラス対 $c_i \times c_j$ に属する単語対 (n_i, n_j) と、 $p_{c_i \times c_j}$ とが共起する確率分布である。 p_2 は、クラス対 $c_i \times c_j$ に属する単語対 (n_i, n_j) 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 V_p を p に対して作成する。そして、この各ベクトルの要素 (n_i, n_j) の値が、 $p_1(n_i, n_j)$ 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 V_P を P に対して作成する。そして、この各ベクトルの要素 (n_i, n_j) の値が、 $p_2(n_i, n_j)$ 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

10

20

30

40

50

、小さな値となる。

【数 9】

$$\text{Assoc}_1(n, p, n') = 1$$

【数 10】

$$\text{Assoc}_2(n, p, n') = \log \frac{\| (n, p, n') \|}{\| (n, *, n') \| \| (*, p, *) \|}$$

【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。

10

【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。

20

【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。

30

【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。

40

【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

50

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】

$$\text{Score}(n_i, n_j, P) = \max_{\substack{c_i \in \text{classes}(n_i), c_j \in \text{classes}(n_j), \\ (n_i, p, n_j) \in M}} \{ \text{CScore}(c_i, c_j, P) \cdot \text{Para}(p_{c_i} \times c_j, P) \cdot \text{Assoc}(n_i, p, n_j) \}$$

10

【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 Para 1 から Para 5 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 Assoc 1 または Assoc 2 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。

20

【数 12】

$$\text{Score} = \text{CScore} \cdot \text{Para}_1 \cdot \text{Assoc}_1$$

【数 13】

$$\text{Score} = \text{CScore} \cdot \text{Para}_1 \cdot \text{Assoc}_2$$

【数 14】

$$\text{Score} = \text{CScore} \cdot \text{Para}_2 \cdot \text{Assoc}_2$$

30

【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、 $\text{Cscore}^2 \times \text{Para} \times \text{Assoc}$ 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。

40

【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

50

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るても含まないと考えるても良い。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。

10

【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。

20

【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、2.5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。

30

【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。

40

【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】

50

(ステップ S 6 0 5) スコア決定部 1 2 6 は、ステップ S 6 0 4 で取得したクラス対が、除外クラス対格納部 1 0 9 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 S 6 1 7 に行き、除外クラス対でなければステップ S 6 0 6 に行く。

【 0 1 1 2 】

(ステップ S 6 0 6) スコア決定部 1 2 6 は、クラス対良好度格納部 1 0 6 から、ステップ S 6 0 4 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。

【 0 1 1 3 】

(ステップ S 6 0 7) スコア決定部 1 2 6 は、カウンタ j に 1 を代入する。

【 0 1 1 4 】

(ステップ S 6 0 8) スコア決定部 1 2 6 は、 j 番目のパターンが、パターン格納部 1 0 5 または、シードパターン格納部 1 0 4 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 S 6 0 9 に行き、存在しなければステップ S 6 1 5 に行く。

【 0 1 1 5 】

(ステップ S 6 0 9) スコア決定部 1 2 6 は、 j 番目のパターンをパターン格納部 1 0 5 または、シードパターン格納部 1 0 4 から取得する。

【 0 1 1 6 】

(ステップ S 6 1 0) スコア決定部 1 2 6 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 1 0 7 から取得する。

【 0 1 1 7 】

(ステップ S 6 1 1) スコア決定部 1 2 6 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 1 0 8 から取得する。

【 0 1 1 8 】

(ステップ S 6 1 2) スコア決定部 1 2 6 は、ステップ S 6 0 6 で取得したクラス対良好度、ステップ S 6 1 0 で取得した類似度、およびステップ S 6 1 1 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。

【 0 1 1 9 】

(ステップ S 6 1 3) スコア決定部 1 2 6 は、ステップ S 6 1 2 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。

【 0 1 2 0 】

(ステップ S 6 1 4) スコア決定部 1 2 6 は、カウンタ j を 1、インクリメントする。ステップ S 6 0 8 に戻る。

【 0 1 2 1 】

(ステップ S 6 1 5) スコア決定部 1 2 6 は、ステップ S 6 1 3 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。

【 0 1 2 2 】

(ステップ S 6 1 6) スコア決定部 1 2 6 は、ステップ S 6 1 5 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。

【 0 1 2 3 】

(ステップ S 6 1 7) スコア決定部 1 2 6 は、カウンタ i を 1、インクリメントする。ステップ S 6 0 2 に戻る。

【 0 1 2 4 】

(ステップ S 6 1 8) 単語対選択部 1 2 7 は、ステップ S 6 1 6 で蓄積したスコアをキーとして、単語対をソートする。

【 0 1 2 5 】

(ステップ S 6 1 9) 単語対選択部 1 2 7 は、ステップ S 6 1 8 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。

【 0 1 2 6 】

(ステップ S 6 2 0) 単語対出力部 1 2 8 は、ステップ S 6 1 9 で取得された 1 以上の単語対を出力し、処理を終了する。

10

20

30

40

50

【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。

10

【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「C₂₉₀」および「C₄₇₁」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。

20

【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。

30

【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。

40

【数15】

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

50

【 0 1 3 5 】

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、 $|I(p) \cap I(P)|$ 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、 $|I(p) \cup I(P)|$ は、「 $I(p)$ 」と「 $I(P)$ 」の和集合のインスタンスの数である。

【 0 1 3 6 】

また、4つのベースライン方法のうち第三の方法、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。

10

【 0 1 3 7 】

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。

【 0 1 3 8 】

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。

【 0 1 3 9 】

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、二つ目は、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。

20

30

（実験 1）

【 0 1 4 0 】

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。

【 0 1 4 1 】

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。

40

50

【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のタプルからランダムに100のタ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。

10

(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。

20

【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、Espresso(寛大な(lenient)ケース)では、50%程度の精度であり、CD-IIIの方法はEspressoと比較して、30%以上も優れていた。

30

【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。

40

(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,

50

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。

10

【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。

20

【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。

30

【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

40

50

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。

10

【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「 $P(\langle v, rel \rangle | n)$ 」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、 $\langle v, rel \rangle$ の組の発生の条件付き確率「 $P(\langle v, rel \rangle | n)$ 」は、以下の数式16により算出できる。「 $P(\langle v, rel \rangle | n)$ 」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 $\langle v, rel \rangle$ と共起する」とする。

20

30

【数16】

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

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、「 $P(\langle v, rel \rangle | n) = (f(\langle v, rel, n \rangle) + 1) / (f(\langle v, rel, n \rangle) + 1)$ 」でも良い。

40

【0161】

数式16において、「 $f(\langle v, rel, n \rangle)$ 」は、 $\langle v, rel, n \rangle$ の出現頻度である。また、Dは、 $\{\langle v, rel \rangle | f(\langle v, rel, n \rangle) > 0\}$ として定義されるセットである。また、「 $f(\langle v, rel, n \rangle) = 0$ 」の場合、「 $P(\langle v, rel \rangle | n)$ 」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 $\langle v, rel, n \rangle$ の組の出現確率を算出する。

【数 17】

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

【0163】

数式 17 において、「a」は $\langle v, rel \rangle$ の組および「n」の隠れクラスを示す。数式 17 において、確率「 $P(\langle v, rel \rangle | a)$ 」、「 $P(n | a)$ 」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。

10

【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（「 $P(\langle v, rel \rangle | a)$ 」、「 $P(n | a)$ 」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「 $P(\langle v, rel \rangle | a)$ 」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、「 $P(\langle v, rel \rangle | a)$ 」、「 $P(n | a)$ 」および「 $P(a)$ 」が更新される。

【0165】

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

20

【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 $\langle v, rel \rangle$ 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。

30

【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。

40

【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

50

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。

10

【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。

20

【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。

30

【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。

40

【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】

50

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。

10

【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。

20

【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、MPU 3413 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 ROM 3415 と、MPU 3413 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 RAM 3416 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。

30

【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-ROM 3501、または FD 3502 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 RAM 3416 にロードされる。プログラムは、CD-ROM 3501、FD 3502 またはネットワークから直接、ロードされても良い。

40

【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

50

。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】

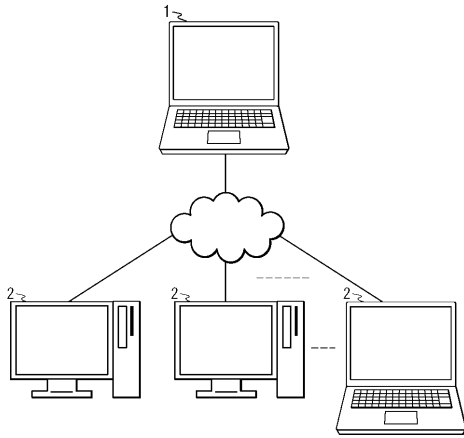
以上のように、本発明にかかる単語対取得装置は、所定の関係にある単語対を適切に取得できるという効果を有し、単語対取得装置等として有用である。

【符号の説明】

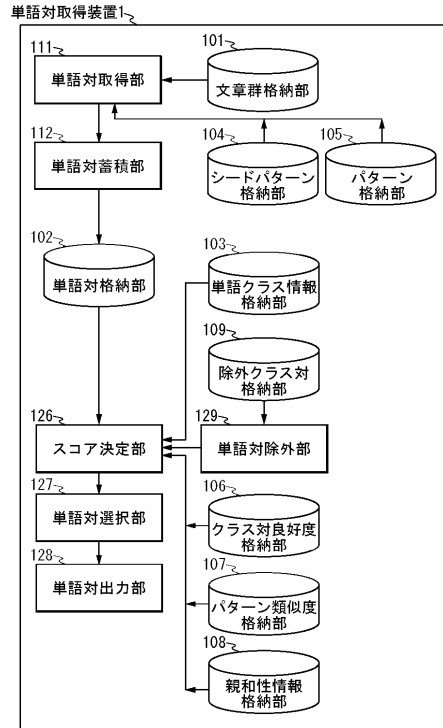
【0189】

1	単語対取得装置	
101	文章群格納部	20
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	30
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	40
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	50

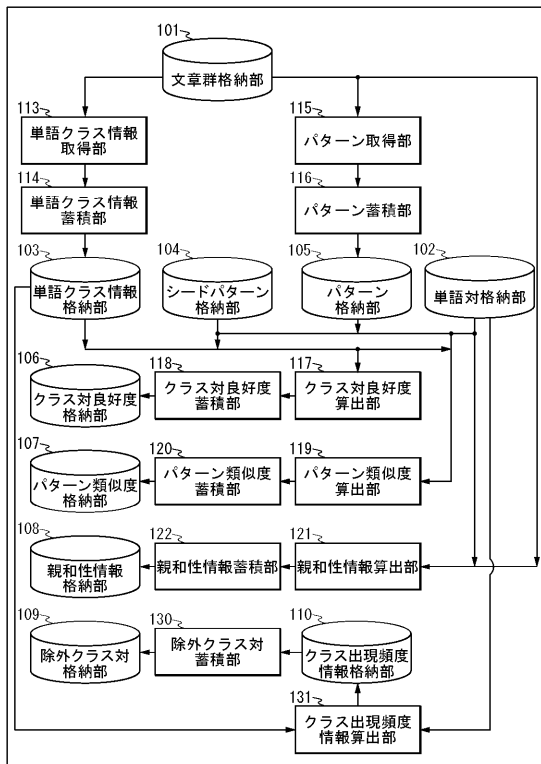
【 図 1 】



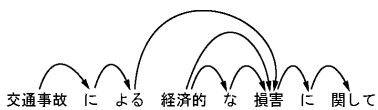
【 図 2 】



【 図 3 】



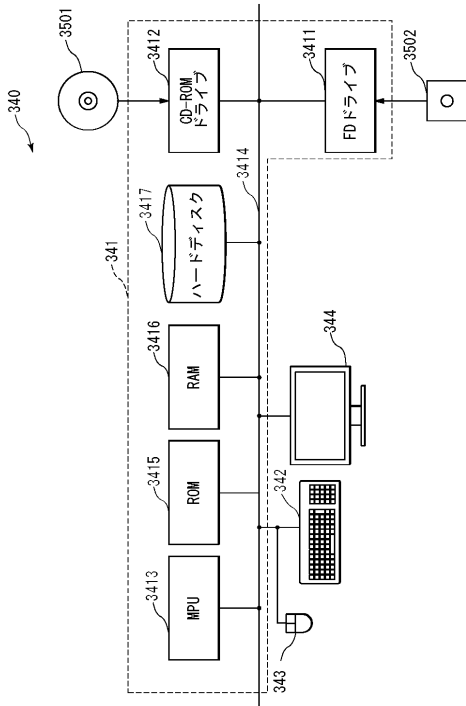
【 図 4 】



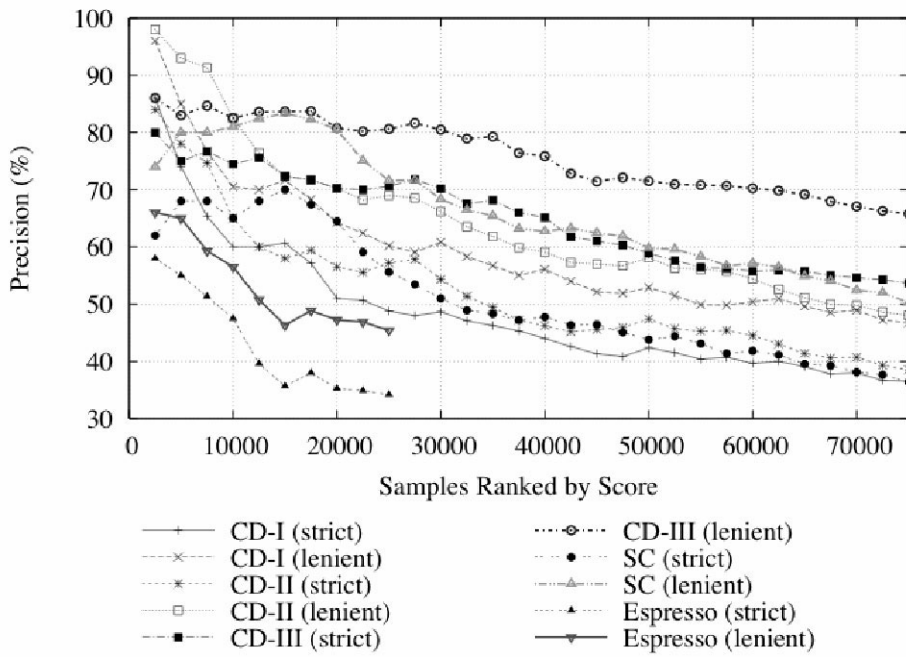
【 図 5 】



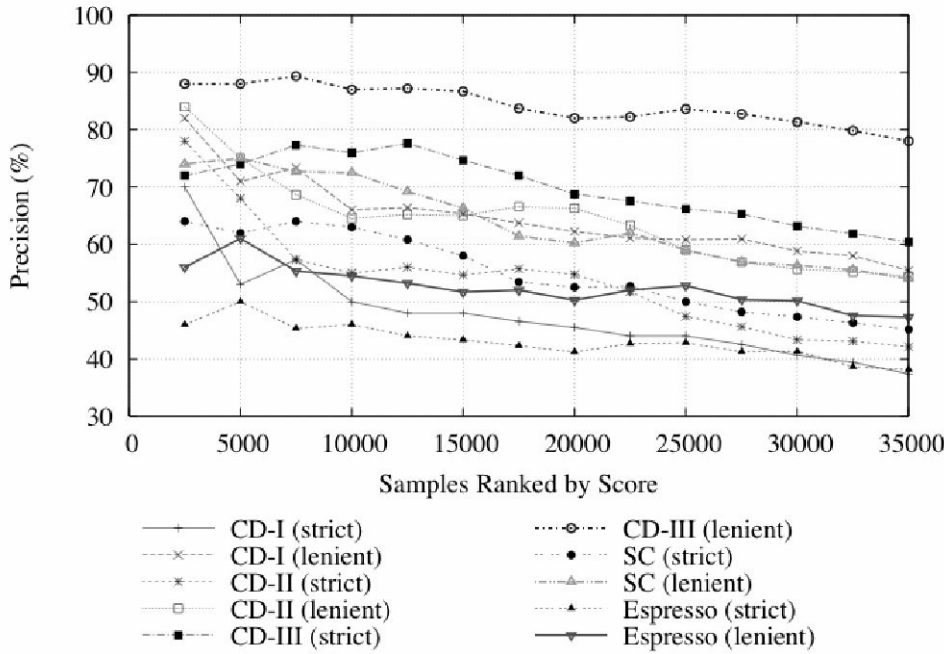
【 図 17 】



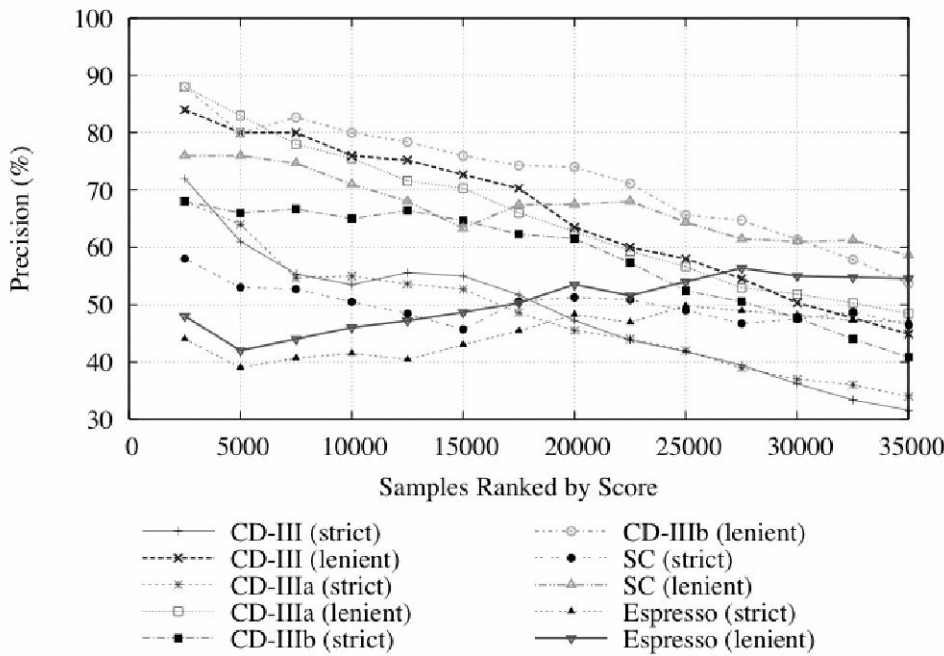
【 図 10 】



【 図 1 2 】



【 図 1 4 】



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

Fターム(参考) 5B075 ND03 NK32

5B091 AA15 CA12 CC05 CC16