(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5708495号
(P5708495)
(45)発行日　平成27年4月30日(2015.4.30)　(24)登録日　平成27年3月13日(2015.3.13)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 320D
G06F 17/27 (2006.01)　G06F 17/30 210D
G06F 17/27 635

請求項の数 11　(全 23 頁)

(21)出願番号　特願2011-545194 (P2011-545194)
(86)(22)出願日　平成22年12月3日 (2010.12.3)
(86)国際出願番号　PCT/JP2010/071696
(87)国際公開番号　WO2011/070980
(87)国際公開日　平成23年6月16日 (2011.6.16)
審査請求日　平成25年11月8日 (2013.11.8)
(31)優先権主張番号　特願2009-282304 (P2009-282304)
(32)優先日　平成21年12月11日 (2009.12.11)
(33)優先権主張国　日本国(JP)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100095407
弁理士　木村　満
(72)発明者　水口　弘紀
日本国東京都港区芝五丁目７番１号　日本電気株式会社内
(72)発明者　久寿居　大
日本国東京都港区芝五丁目７番１号　日本電気株式会社内
(72)発明者　楠村　幸貴
日本国東京都港区芝五丁目７番１号　日本電気株式会社内

最終頁に続く

(54)【発明の名称】辞書作成装置、単語収集方法、及び、プログラム

(57)【特許請求の範囲】
【請求項１】
単語の入力を受け付け、入力された入力単語に関連する単語を文書データから出力し、以降は所定の条件に達するまで出力した単語を前記入力単語に追加し、該入力単語に関連する単語を文書データから出力することを繰り返していくことで単語を収集する辞書増殖処理における、入力単語と該入力単語によって出力された出力単語との入出力の過程を示す情報を記録する入出力過程記録手段と、
前記入出力過程記録手段に記録された情報に基づいて、前記入出力の過程における前記入力単語及び前記出力単語の類似度を用いて、前記辞書増殖処理で収集された単語を複数のクラスタに分類するクラスタ分類手段と、
前記入出力過程記録手段に記録された情報を参照し、前記クラスタ分類手段が分類したクラスタ毎に、該クラスタ内の単語が最初に入力を受け付けた入力単語からクラスタ内の各単語を出力するまでに要したターン数及び当該クラスタ内の各単語が最初に入力を受け付けた入力単語を出力するまでに要したターン数に基づいて、クラスタ内の単語が入力単語と同じ種類の単語であるか否かを判別する同種判別手段と、
前記辞書増殖処理で収集された単語と、該単語が属するクラスタと、該クラスタを構成する単語が最初に入力を受け付けた入力単語と同じ種類の単語であるか否かを示す情報と、を関連付けて出力する収集単語出力手段と、
を備えることを特徴とする辞書作成装置。
【請求項２】

単語の入力を受け付け、入力された入力単語に関連する単語を文書データから出力し、以降は所定の条件に達するまで出力した単語を前記入力単語に追加し、該入力単語に関連する単語を文書データから出力することを繰り返していくことで単語を収集する辞書増殖手段をさらに備える、

ことを特徴とする請求項１に記載の辞書作成装置。

【請求項３】

前記入出力過程記録手段は、複数回の入出力を繰り返した、入力単語と該入力単語によって出力された出力単語との入出力の過程を示す情報を記録する、

ことを特徴とする請求項１又は２に記載の辞書作成装置。

【請求項４】

前記クラスタ分類手段は、前記入出力過程記録手段に記録されている情報から、前記辞書増殖処理で収集した単語のうち共通の単語を入力にする単語同士、又は共通の単語を出力する単語同士ほどその値が大きくなる値を示す単語間の結束度を算出し、算出した結束度に基づいて、単語をクラスタに分類する、

ことを特徴とする請求項１乃至３の何れか１項に記載の辞書作成装置。

【請求項５】

前記同種判別手段は、前記入出力過程記録手段に記録されている情報に基づいて、クラスタ毎に、最初に入力を受け付けた入力単語から当該クラスタ内の各単語を出力するまでに要したターン数、及び当該クラスタ内の各単語が最初に入力を受け付けた入力単語を出力するまでに要したターン数を算出し、算出したターン数の平均値を用いて、当該クラスタ内の単語が最初に入力を受け付けた入力単語と同種であるか異種であるかの判別をする、

ことを特徴とする請求項１乃至４の何れか１項に記載の辞書作成装置。

【請求項６】

前記辞書増殖処理で収集された単語を種類毎に、複数の単語グループに分類して記憶する、単語グループ記憶手段と、

所定の条件を満たす一の単語グループのなかから所定数の単語を選択する単語選択手段と、をさらに備え、

前記単語選択手段が選択した単語を入力単語とした前記辞書増殖処理を実行し、

前記同種判別手段は、前記入出力過程記録手段に記録された情報に基づいて、前記クラスタ分類手段が分類したクラスタ毎に、該クラスタ内の単語が前記単語選択手段が選択した入力単語と同じ種類の単語であるか否かを判別する、

ことを特徴とする請求項１乃至５の何れか１項に記載の辞書作成装置。

【請求項７】

前記同種判別手段が判別した結果に基づいて、前記辞書増殖処理で収集された単語を前記単語グループ記憶手段に登録し、登録した単語グループのうち所定の条件を満たす単語グループがある場合に、前記単語選択手段に単語の選択を指示する再実行手段をさらに備え、

前記再実行手段は、収集単語を前記単語グループ記憶手段に登録する際、収集単語の属するクラスタが前記単語選択手段が選択した単語と同種の単語である場合には当該選択した単語と同じ単語グループに当該収集単語を登録し、異種であり且つ既に前記単語グループ記憶手段に記憶されている単語である場合には該記憶されている単語と同じ単語グループに収集単語を登録し、異種であり且つ未だ前記単語グループ記憶手段が記憶していない単語である場合には収集単語を新規の単語グループに登録する、

ことを特徴とする請求項６に記載の辞書作成装置。

【請求項８】

前記入出力過程記録手段に記録されている情報から算出された、前記辞書増殖処理で収集した単語のうち共通の単語を入力にする単語同士、又は共通の単語を出力する単語同士ほどその値が大きくなる値を示す単語間の結束度を記憶する結束度記憶手段をさらに備え、

前記単語選択手段は、前記一の単語グループ内の単語間の結束度に基づいて、所定数の単語を選択する、
ことを特徴とする請求項６又は７に記載の辞書作成装置。
【請求項９】
前記単語選択手段は、結束度の大きい順に単語を選択する割合、又は、結束度の小さい順に単語を選択する割合、が少なくとも予め設定されている条件情報に基づいて、所定数の単語を選択する、
ことを特徴とする請求項８に記載の辞書作成装置。
【請求項１０】
コンピュータが、単語の入力を受け付け、入力された入力単語に関連する単語を文書データから出力し、以降は所定の条件に達するまで出力した単語を前記入力単語に追加し、該入力単語に関連する単語を文書データから出力することを繰り返していくことで単語を収集した辞書増殖処理における入力単語と該入力単語によって出力された出力単語との入出力の過程を示す情報を記録する入出力過程記録ステップと、
コンピュータが、前記入出力過程記録ステップに記録された情報に基づいて、前記入出力の過程における前記入力単語及び前記出力単語の類似度を用いて、前記辞書増殖処理で収集された単語を複数のクラスタに分類するクラスタ分類ステップと、
コンピュータが、前記入出力過程記録ステップに記録された情報を参照し、前記クラスタ分類ステップが分類したクラスタ毎に、該クラスタ内の単語が最初に入力を受け付けた入力単語からクラスタ内の各単語を出力するまでに要したターン数及び当該クラスタ内の各単語が最初に入力を受け付けた入力単語を出力するまでに要したターン数に基づいて、クラスタ内の単語が入力単語と同じ種類の単語であるか否かを判別する同種判別ステップと、
コンピュータが、前記辞書増殖処理で収集された単語と、該単語が属するクラスタと、該クラスタを構成する単語が最初に入力を受け付けた入力単語と同じ種類の単語であるか否かを示す情報と、を関連付けて出力する収集単語出力ステップと、
を備えることを特徴とする単語収集方法。
【請求項１１】
コンピュータを、
単語の入力を受け付け、入力された入力単語に関連する単語を文書データから出力し、以降は所定の条件に達するまで出力した単語を前記入力単語に追加し、該入力単語に関連する単語を文書データから出力することを繰り返していくことで単語を収集する辞書増殖処理における、入力単語と該入力単語によって出力された出力単語との入出力の過程を示す情報を記録する入出力過程記録手段、
前記入出力過程記録手段に記録された情報に基づいて、前記入出力の過程における前記入力単語及び前記出力単語の類似度を用いて、前記辞書増殖処理で収集された単語を複数のクラスタに分類するクラスタ分類手段、
前記入出力過程記録手段に記録された情報を参照し、前記クラスタ分類手段が分類したクラスタ毎に、該クラスタ内の単語が最初に入力を受け付けた入力単語からクラスタ内の各単語を出力するまでに要したターン数及び当該クラスタ内の各単語が最初に入力を受け付けた入力単語を出力するまでに要したターン数に基づいて、クラスタ内の単語が入力単語と同じ種類の単語であるか否かを判別する同種判別手段、
前記辞書増殖処理で収集された単語と、該単語が属するクラスタと、該クラスタを構成する単語が最初に入力を受け付けた入力単語と同じ種類の単語であるか否かを示す情報と、を関連付けて出力する収集単語出力手段、
として機能させるプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、辞書作成装置、単語収集方法、及び、プログラムに関する。

【背景技術】
【０００２】
少数の同種の単語を入力に、文献データやＷｅｂページ等から多数の同種の単語を収集した辞書を作成する辞書作成の手法が知られている。なお、ここでいう辞書とは、共通の上位概念を持つ同種の単語の集合のことである。
【０００３】
上述した辞書作成の手法の一例が、非特許文献１に記載されている。この辞書作成の手法の概略を以下に示す。
【０００４】
まず、収集の元となる少量の単語を入力する。以下、はじめに入力した単語をシード単語と呼ぶ。次に、Ｗｅｂ検索エンジンを利用し、シード単語を含むＷｅｂページを収集する。次に、収集したＷｅｂページからシード単語とそれ以外の語を区切るパターンを作成する。そして、このパターンを使ってＷｅｂページから単語を抽出し、シード単語に追加する。なお、シード単語を入力してから単語が抽出されるまでをターンと呼ぶ。そして、単語が追加されたシード単語を用いて、さらにＷｅｂページを収集する。これを数ターン繰り返した後、抽出された単語をシード単語と同種の単語の集合（辞書）として出力する。
【０００５】
このような辞書作成の手法では、新たにシード単語に追加される単語が、シード単語と異なる種類の単語である場合がある。例えば、レストラン名のシード単語を入力して、レストラン名の辞書を作成する際に、同じ文献に掲載されており、かつ、パターンが似ているラーメン店名やうどん店名などの単語が、新たにシード単語に追加されてしまう等の場合である。
このような場合、その異なる種類の単語から、さらに異なる種類の単語が次々にシード単語に追加されてしまい、シード単語と異なる種類の単語が多く収集されてしまい、辞書の精度が悪化することが知られている。
【０００６】
このような事態を回避するために、各ターンで抽出される単語の信頼度を求め、特定の信頼度以上の単語のみをシード単語に追加して、次のターンで採用することが行われている。なお、この信頼度は、例えば、パターンの出現回数に基づく統計量や、パターンから検出された単語数に基づく統計量等が用いられる。非特許文献１では、信頼度として、単語のパターンによって抽出できたＷｅｂページの数を採用しており、抽出できたＷｅｂページの数が所定数以下の単語はシード単語に追加しないことで、異なる種類の単語が収集されることを防止している。
【先行技術文献】
【非特許文献】
【０００７】
【非特許文献１】水口弘紀、河合英紀、土田正明、久寿居大、Ｗｅｂ知識を利用したブートストラップによる辞書増殖手法、ＤＥＷＳ２００７、２００７
【発明の概要】
【発明が解決しようとする課題】
【０００８】
上述した信頼度を用いて辞書を作成をした場合、信頼度が低い異なる種類の単語（異種単語）は、収集対象から除外されてシードに追加されない。従って、ユーザは、どのような異種単語がシード単語から収集されているのかを全く知ることができず、異種単語を再利用してさらに異なるグループの単語を収集するような活用ができない。
【０００９】
本発明は、上記実情に鑑みてなされたものであり、どのような異種単語が収集されているのかをユーザに好適に出力することを可能にした辞書作成装置、単語収集方法、及び、プログラムを提供することを目的とする。

【課題を解決するための手段】
【００１０】
上記目的を達成するため、本発明の第１の観点に係る辞書作成装置は、
単語の入力を受け付け、入力された入力単語に関連する単語を文書データから出力し、以降は所定の条件に達するまで出力した単語を前記入力単語に追加し、該入力単語に関連する単語を文書データから出力することを繰り返していくことで単語を収集する辞書増殖処理における、入力単語と該入力単語によって出力された出力単語との入出力の過程を示す情報を記録する入出力過程記録手段と、
前記入出力過程記録手段に記録された情報に基づいて、前記入出力の過程における前記入力単語及び前記出力単語の類似度を用いて、前記辞書増殖処理で収集された単語を複数のクラスタに分類するクラスタ分類手段と、
前記入出力過程記録手段に記録された情報を参照し、前記クラスタ分類手段が分類したクラスタ毎に、該クラスタ内の単語が最初に入力を受け付けた入力単語からクラスタ内の各単語を出力するまでに要したターン数及び当該クラスタ内の各単語が最初に入力を受け付けた入力単語を出力するまでに要したターン数に基づいて、クラスタ内の単語が入力単語と同じ種類の単語であるか否かを判別する同種判別手段と、
前記辞書増殖処理で収集された単語と、該単語が属するクラスタと、該クラスタを構成する単語が最初に入力を受け付けた入力単語と同じ種類の単語であるか否かを示す情報と、を関連付けて出力する収集単語出力手段と、
を備えることを特徴とする。
また、本発明の第２の観点に係る単語収集方法は、
コンピュータが、単語の入力を受け付け、入力された入力単語に関連する単語を文書データから出力し、以降は所定の条件に達するまで出力した単語を前記入力単語に追加し、該入力単語に関連する単語を文書データから出力することを繰り返していくことで単語を収集した辞書増殖処理における入力単語と該入力単語によって出力された出力単語との入出力の過程を示す情報を記録する入出力過程記録ステップと、
コンピュータが、前記入出力過程記録ステップに記録された情報に基づいて、前記入出力の過程における前記入力単語及び前記出力単語の類似度を用いて、前記辞書増殖処理で収集された単語を複数のクラスタに分類するクラスタ分類ステップと、
コンピュータが、前記入出力過程記録ステップに記録された情報を参照し、前記クラスタ分類ステップが分類したクラスタ毎に、該クラスタ内の単語が最初に入力を受け付けた入力単語からクラスタ内の各単語を出力するまでに要したターン数及び当該クラスタ内の各単語が最初に入力を受け付けた入力単語を出力するまでに要したターン数に基づいて、クラスタ内の単語が入力単語と同じ種類の単語であるか否かを判別する同種判別ステップと、
コンピュータが、前記辞書増殖処理で収集された単語と、該単語が属するクラスタと、該クラスタを構成する単語が最初に入力を受け付けた入力単語と同じ種類の単語であるか否かを示す情報と、を関連付けて出力する収集単語出力ステップと、
を備えることを特徴とする。
また、本発明の第３の観点に係るプログラムは、
コンピュータを、
単語の入力を受け付け、入力された入力単語に関連する単語を文書データから出力し、以降は所定の条件に達するまで出力した単語を前記入力単語に追加し、該入力単語に関連する単語を文書データから出力することを繰り返していくことで単語を収集する辞書増殖処理における、入力単語と該入力単語によって出力された出力単語との入出力の過程を示す情報を記録する入出力過程記録手段、
前記入出力過程記録手段に記録された情報に基づいて、前記入出力の過程における前記入力単語及び前記出力単語の類似度を用いて、前記辞書増殖処理で収集された単語を複数のクラスタに分類するクラスタ分類手段、
前記入出力過程記録手段に記録された情報を参照し、前記クラスタ分類手段が分類した

クラスタ毎に、該クラスタ内の単語が最初に入力を受け付けた入力単語からクラスタ内の各単語を出力するまでに要したターン数及び当該クラスタ内の各単語が最初に入力を受け付けた入力単語を出力するまでに要したターン数に基づいて、クラスタ内の単語が入力単語と同じ種類の単語であるか否かを判別する同種判別手段、

前記辞書増殖処理で収集された単語と、該単語が属するクラスタと、該クラスタを構成する単語が最初に入力を受け付けた入力単語と同じ種類の単語であるか否かを示す情報と、を関連付けて出力する収集単語出力手段、

として機能させるプログラムである。

【発明の効果】

【００１１】

本発明によれば、辞書構築において収集された単語をクラスタリングし、クラスタ毎に、最初に入力した単語と同じ種類の単語であるか否かが判別される。従って、どのような異種単語が収集されているのかをユーザに好適に出力することができる。

【図面の簡単な説明】

【００１２】

【図１】本発明の第１実施形態に係る辞書作成装置の構成を示す図である。
【図２】収集過程記憶部に記憶される情報の構成例を示す図である。
【図３】収集単語記憶部に記憶される情報の構成例を示す図である。
【図４】辞書作成処理の動作を説明するためのフローチャートである。
【図５】辞書増殖処理の動作を説明するためのフローチャートである。
【図６】クラスタリング処理の動作を説明するためのフローチャートである。
【図７】単語間の入出力の関係を示したグラフである。
【図８】同種判別処理の動作を説明するためのフローチャートである。
【図９】本発明の第２実施形態に係る辞書作成装置の構成を示す図である。
【図１０】図１０（Ａ）及び図１０（Ｂ）は、単語グループ記憶部に記憶される情報の構成例を示す図である。
【図１１】辞書作成処理の動作を説明するためのフローチャートである。
【図１２】単語グループ更新処理の動作を説明するためのフローチャートである。
【図１３】本発明の第３実施形態に係る辞書作成装置の構成を示す図である。
【図１４】収集単語記憶部に記憶される情報の構成例を示す図である。
【図１５】各実施形態に係る辞書作成装置をコンピュータに実装する場合の、物理的な構成の一例を示すブロック図である。

【発明を実施するための形態】

【００１３】

以下、本発明の実施形態について、図面を参照しながら詳細に説明する。なお、本発明は下記の実施形態及び図面によって限定されるものではない。本発明の要旨を変更しない範囲で下記の実施形態及び図面に変更を加えることが出来るのはもちろんである。また、図中同一または相当部分には同一符号を付す。

また、本発明で辞書とは、共通の上位概念を持つ同種の単語の集合のことである。

【００１４】

（第１実施形態）

本発明の第１実施形態に係る辞書作成装置１００について説明する。辞書作成装置１００は、図1に示すように、入力部１０１と、辞書増殖部１０２と、クラスタリング部１０３と、種別判別部１０４と、出力部１０５と、文書記憶部１０６と、収集過程記憶部１０７と、収集単語記憶部１０８とを備える。

【００１５】

入力部１０１は、キーボードやマウスなどから構成される。ユーザは、入力部１０１を介して、辞書（同種の単語の集合）を作成するためのサンプルとなる単語（シード単語）を入力する。

【００１６】

辞書増殖部１０２は、非特許文献１に記載されているような従来の手法を用いて、シード単語と同種の単語を文書記憶部１０６に記憶されている文書内から収集する辞書増殖処理を行う。また、辞書増殖部１０２は、この辞書増殖処理において、どのような過程を経て単語が収集されたのかを示す情報を収集過程記憶部１０７に記録する。辞書増殖部１０２の行う辞書増殖処理の詳細については後述する。

【００１７】

クラスタリング部１０３は、収集過程記憶部１０７に記憶されている情報に基づいて、辞書増殖部１０２が収集した単語を複数のクラスタに分類（クラスタリング）する。クラスタリング部１０３の行う処理の詳細については後述する。

【００１８】

種別判別部１０４は、クラスタとそのクラスタに含まれる単語とを入力に、収集過程記憶部１０７に記憶されている情報を参照し、クラスタを構成する単語が、シード単語と同じ種類の単語であるか否かを判別する。種別判別部１０４の行う処理の詳細については後述する。

【００１９】

出力部１０５は、種々の情報を出力する。例えば、出力部１０５は、辞書増殖処理によって収集された単語を、分類されたクラスタ毎に、シード単語と異種か同種かを示す情報を付して出力（表示）する。

【００２０】

文書記憶部１０６は、辞書増殖部１０２による単語収集の対象となる各文書を定義するデータが記憶される。なお、各文書のデータにはＩＤ（文書ＩＤ）が付されている。

【００２１】

収集過程記憶部１０７には、辞書増殖処理において、どのような入出力の過程を経て単語が収集されたのかを示す情報が記録される。具体的には、図2に示すように、収集過程記憶部１０７には、辞書増殖処理におけるターン毎に、当該ターンのターン数と、当該ターンで入力された入力単語と、該入力単語から生成されたパターンによって出力された出力単語とが対応付けられて記録される。

例えば、図2の先頭のエントリから、辞書増殖処理の１ターン目に、「レストランＳ」から作成されたパターンにより「レストランＸ」が抽出されたことがわかる。

【００２２】

図1に戻り、収集単語記憶部１０８には、図3に示すように、収集された各単語と、各単語がどのクラスタに分類されているかを示すクラスタＩＤとが対応付けられて記憶される。また、各クラスタには、クラスタを構成する単語が、シード単語と同じ種類の単語であるのか（シード単語自体が当該クラスタに含まれる場合も同じ種類とする）、又は、異なる種類の単語であるのか、を示す情報が付与される。

例えば、図3から、「レストランＡ」と「レストランＢ」はクラスタ１に分類され、また、クラスタ１はシード単語と同じ種類の単語から構成されていることが分かる。同様に、「うどんＣ」と「うどんＤ」はクラスタ２に分類され、また、クラスタ２はシード単語と異なる種類の単語から構成されていることが分かる。

【００２３】

続いて、辞書作成装置１００で実施される処理の動作について説明する。

まず、ユーザは、入力部１０１を操作して、辞書（同種の単語の集合）を作成するためのサンプルとなる１乃至複数の単語（シード単語）を入力する。そして、入力したシード単語を元に、辞書を作成することを指示する。この指示操作に応じて、辞書作成装置１００は、図4に示す辞書作成処理を行う。

【００２４】

辞書作成処理が開始されると、まず、辞書増殖部１０２は、従来の手法で辞書増殖処理を行い、入力されたシード単語に関連する単語を収集する（ステップS100）。

【００２５】

辞書増殖処理（ステップS100）の詳細について、図5のフローチャートを参照して説明

する。辞書増殖処理が開始されると、まず、辞書増殖部１０２は、ユーザによって入力されたシード単語を収集単語記憶部１０８に登録する（ステップS101）。そして、辞書増殖部１０２は、ターン数を示すカウンタｉ（初期値０）を１インクリメントする（ステップS102）。

【００２６】

続いて、辞書増殖部１０２は、収集単語記憶部１０８に記憶されている単語のなかから所定数の単語をランダムに選択する（ステップS103）。そして、辞書増殖部１０２は、文書記憶部１０６に記憶されている文書のなかから、選択したシード単語が含まれている文書を検出する（ステップS104）。なお、ここでは、選択したシード単語を全て含む文書のみを検出してもよいし、選択したシード単語のうち所定数のシード単語を含む文書を検出してもよい。

【００２７】

続いて、辞書増殖部１０２は、検出した文書内における、ステップS103で選択したシード単語が出現する位置を特定し、シード単語とそれ以外の部分とを区切るパターンを作成する（ステップS105）。例えば、文書内でシード単語が出現する部分の前後の所定数の文字列を、パターンとして採用すればよい。

【００２８】

続いて、辞書増殖部１０２は、作成したパターンに合致する単語を、文書記憶部１０６に記憶されている文書から抽出する（ステップS106）。そして、辞書増殖部１０２は、抽出した単語を収集単語記憶部１０８に追加する（ステップS107）。

【００２９】

続いて、辞書増殖部１０２は、今回のターン数を示す情報（即ち、カウンタｉの値）と、ステップS103で選択した各単語（入力単語）と、入力単語から作成したパターンによりステップS106で抽出した単語（出力単語）とを対応付けて、収集過程記憶部１０７に記憶する（ステップS108）。

【００３０】

続いて、辞書増殖部１０２は、辞書増殖を終了させるための所定の終了条件を満たしているか否かを判別する（ステップS109）。終了条件としては、例えば、収集単語記憶部１０８に記憶した単語の数が所定数に達したか、又は、ターン数が所定数に達したか等の任意の条件を採用することが可能である。なお、後述するクラスタリング処理で収集した単語を適切にクラスタリングできるようにするために、ここでは、少なくとも２ターン以上は単語の収集を繰り返し実行するような終了条件を採用することが望ましい。

【００３１】

終了条件を満たしていないと判別した場合（ステップS109；No）、辞書増殖部１０２は、ステップS102～ステップS108を繰り返し、新たに単語が追加されたシード単語から単語を収集する処理を引き続き行う。

終了条件を満たしていると判別した場合（ステップS109；Yes）、辞書増殖部１０２は、辞書増殖処理を終了し処理をクラスタリング部１０３に移す。

【００３２】

図4に戻り、続いて、クラスタリング部１０３は、辞書増殖処理によって収集された単語をクラスタに分類するクラスタリング処理を行う（ステップS200）。

【００３３】

図6は、クラスタリング処理（ステップS200）の詳細を示すフローチャートである。クラスタリング処理が開始されると、まず、クラスタリング部１０３は、収集単語記憶部１０８から、未だ単語間の結束度を算出していない２つの単語を選択する（ステップS201）。

【００３４】

続いて、クラスタリング部１０３は、選択した２つの単語間の結束度を、収集過程記憶部１０７に記憶されている情報に基づいて算出する（ステップS202）。

【００３５】

なお、単語間の結束度とは、上述した辞書増殖処理において、共通の単語を入力にする単語同士、又は共通の単語を出力する単語同士ほど、その値が大きくなる指標のことである。例えば、２つの単語それぞれに入力される単語のうち共通の単語から２つの単語に入力される単語の割合と、２つの単語それぞれが出力する単語のうち２つの単語が共通の単語を出力する単語の割合と、の和を２つの単語間の結束度として算出することができる。

【００３６】

より具体的には、２つの単語a,b間の結束度をSim(a,b)とすると、以下の式により、結束度を算出することができる。

Sim(a,b)=Sim_in(a,b)+sim_out(a,b)

【００３７】

上式において、Sim_in(a,b)は、単語a,bそれぞれに入力される単語のうち共通の単語から入力される単語の割合を示す値である。Sim_in(a,b)=（単語aと単語bの両方に入力される共通の単語の数）／（（単語aに入力される単語の数）+（単語bに入力される単語の数））と求めることができる。

また、Sim_out(a,b)は、２つの単語a,bそれぞれが出力する単語のうち共通の単語を出力する単語の割合を示す値である。Sim_out(a,b)=（単語aと単語bの両方から主力された共通の単語の数）／（（単語aが出力した単語の数）+（単語bが出力した単語の数））と求めることができる。

【００３８】

続いて、クラスタリング部１０３は、収集単語記憶部１０８に記憶されているシード単語の全ての組で、結束度を算出したか否かを判別する（ステップS203）。

【００３９】

シード単語の全ての組で結束度を算出していない場合（ステップS203；No）、クラスタリング部１０３は、結束度未算出の２つのシード単語を選択して結束度を算出する処理（ステップS201、ステップS202）を繰り返す。

【００４０】

シード単語の全ての組で結束度を算出した場合（ステップS203；Yes）、クラスタリング部１０３は、算出した結束度を類似度として、最短距離法、最長距離法、および、群平均法などの公知のクラスタリング手法を用いてクラスタリングを行い、収集単語記憶部１０８に記憶されているシード単語を複数のクラスタに分類する（ステップS204）。

そして、クラスタリング部１０３は、クラスタリングした結果を記録する（ステップS205）。具体的には、クラスタリング部１０３は、収集単語記憶部１０８に記憶されている単語に、クラスタに分類した結果が反映されるようにクラスタＩＤを付与する。以上でクラスタリング処理は終了する。

【００４１】

このように、クラスタリング処理により、収集された単語間の結束度が算出され、算出された結束度に基づいて、収集単語が複数のクラスタに分類される。

【００４２】

ここで、上述したクラスタリング処理について、具体例を挙げて説明する。図7は、図2に示すような情報が収集過程記憶部１０７に記憶されている場合の、辞書増殖処理のターン１からターン３の単語間の入出力の関係をグラフで示した図である。この図において、各単語はノードで表され、入力単語から出力単語の方向にアーク（矢印）で結ばれる。例えば、図7より、単語「レストランＡ」は、ターン２に「レストランＸ」と「レストランＳ」から作成されたパターンにより抽出されたことがわかる。また、ターン３では、単語「レストランＡ」から作成されたパターンにより「レストランＥ」と「レストランＴ」とが抽出されたことがわかる。

【００４３】

ここで、「レストランＡ」と「レストランＢ」との間の結束度Sim(A,B)を算出する場合を考える。

「レストランＡ」に入力される単語は「レストランＸ」と「レストランＳ」であり、「

レストランＢ」に入力される単語は「レストランＳ」である。そして、このうち、「レストランＳ」が、「レストランＡ」と「レストランＢ」の両方に入力される。したがって、Sim_in(A,B)は、１／３となる。また、「レストランＡ」が出力する単語は「レストランＥ」と「レストランＴ」であり、「レストランＢ」が出力する単語は「レストランＴ」である。そして、このうち、「レストランＴ」が、「レストランＡ」と「レストランＢ」の両方から出力される。したがって、Sim_out(A,B)は、１／３となる。したがって、結束度Sim(A,B)=Sim_in(A,B)+Sim_out(A,B)=1/3+1/3=2/3と算出される。

【００４４】

同様に、他の単語間の結束度についても、以下のように算出される。

レストランＡとうどんＣとの間の結束度：Sim(A,C)=Sim_in(A,C)+Sim_out(A,C)=0+0=0

レストランＡとうどんＤとの間の結束度：Sim(A,D)=Sim_in(A,D)+Sim_out(A,D)=0+0=0

レストランＢとうどんＣとの間の結束度：Sim(B,C)=Sim_in(B,C)+Sim_out(B,C)=0+0=0

レストランＢとうどんＤとの間の結束度：Sim(B,D)=Sim_in(B,D)+Sim_out(B,D)=0+1/3=1/3

うどんＣとうどんＤとの間の結束度：Sim(C,D)=Sim_in(C,D)+Sim_out(C,D)=2/4+1/4=3/4

【００４５】

そして、これらの単語間の結束度を類似度として、公知のクラスタリングの手法を用いたクラスタリングがなされる。例えば、この結束度から、クラスタ１｛レストランＡ,レストランＢ｝、クラスタ２｛うどんＣ,うどんＤ｝の２つのクラスタが形成され、図3に示すように、収集単語記憶部１０８に記憶されている各単語に、クラスタＩＤが付与される。

【００４６】

図4に戻り、続いて、種別判別部１０４は、クラスタリング処理で分類したクラスタが、最初に入力された単語（シード単語）と同種の単語から構成されるか否かを判別する同種判別処理を行う（ステップS300）。

【００４７】

図8は、同種判別処理（ステップS300）の詳細を示すフローチャートである。同種判別処理が開始されると、まず、種別判別部１０４は、収集単語記憶部１０８から、同種判別を未だ行っていない１つのクラスタ、及び、当該クラスタに含まれる単語を選択する（ステップS301）。

【００４８】

続いて、種別判別部１０４は、収集過程記憶部１０７を参照して、選択したクラスタ内の単語が、最初に入力された単語（シード単語）と同種の単語であるか否かを判別する（ステップS302）。なお、この判別は、クラスタ内の各単語のシード単語までの近さに基づいて行えばよい。

具体的には、種別判別部１０４は、シード単語からクラスタ内の各単語を出力するまでに要したターン数や、クラスタ内の各単語がシード単語を出力するまでに要したターン数を算出し、算出したターン数に基づいて、同種か異種かの判別をすればよい。

【００４９】

続いて、種別判別部１０４は、判別結果を収集単語記憶部１０８に記録する（ステップS303）。

【００５０】

続いて、種別判別部１０４は、収集単語記憶部１０８に記憶されているクラスタ全てで、上述の同種判別を実施したか否かを判別する（ステップS304）。

【００５１】

同種判別未実施のクラスタがある場合（ステップS304；No）、種別判別部１０４は、そのクラスタを選択して同種判別をする処理（ステップS301～ステップS303）を繰り返す。

【００５２】

同種判別を未実施のクラスタがない場合（ステップS304；Yes）、同種判別処理は終了

する。

【００５３】

このように、同種判別処理が実施されることにより、クラスタ毎に、クラスタを構成する単語がシード単語と同じ種類の単語であるか異なる種類の単語であるかが判別される。

【００５４】

続いて、上述した同種判別処理について、具体例を挙げて説明する。

前提として、図7に示すような入出力関係が、図2に示す収集過程記憶部１０７に記憶されている情報から得られているものとする。また、「レストランＡ」と「レストランＢ」がクラスタ１、「うどんＣ」と「うどんＤ」がクラスタ２に分類されているものとする。また、同種判定に用いる閾値の値は0.6とする。なお、図7では、シード単語である「レストランＳ」と「レストランＴ」は、網掛けで示している。

【００５５】

まず、クラスタ１の同種判別について説明する。

クラスタ１内の単語「レストランＡ」は、「レストランＳ→レストランＡ」のルートにより、最短１ターンでシード単語「レストランＳ」から出力される。若しくは、「レストランＡ」は、「レストランＡ→レストランＴ」のルートにより、最短１ターンでシード単語「レストランＴ」を出力する。そのため、その最短のターン数１の逆数１を、「レストランＡ」のシード単語までの近さを表す値とする。

同様に、クラスタ１内の単語「レストランＢ」は、「レストランＳ→レストランＢ」のルートにより、最短１ターンでシード単語「レストランＳ」から出力される。若しくは、「レストランＢ」は、「レストランＢ→レストランＴ」のルートにより、最短１ターンでシード単語「レストランＴ」を出力する。そのため、その最短のターン数１の逆数１を、「レストランＢ」のシード単語までの近さを表す値とする。

したがって、クラスタ１全体でのシード単語までの近さは、「レストランＡ」と「レストランＢ」の近さの平均を取り１となる。この値は、閾値0.6以上であるため、クラスタ１は同種と判別され、その結果が収集単語記憶部１０８に記憶される。

【００５６】

続いて、クラスタ２の同種判別について説明する。

クラスタ２内の単語「うどんＣ」は、「レストランＳ→レストランＺ→うどんＣ」又は「レストランＴ→レストランW→うどんＣ」等のルートにより、最短２ターンでシード単語「レストランＳ」又は「レストランＴ」から出力される。そのため、その最短のターン数２の逆数0.5を、「うどんＣ」のシード単語までの近さを表す値とする。

同様に、クラスタ２内の単語「うどんＤ」は、「レストランＳ→レストランＺ→うどんＤ」又は「レストランＴ→レストランW→うどんＤ」等のルートにより、最短２ターンでシード単語「レストランＳ」又は「レストランＴ」から出力される。そのため、その最短のターン数２の逆数0.5を、「うどんＤ」のシード単語までの近さを表す値とする。

したがって、クラスタ２全体でのシード単語までの近さは、うどんＣとうどんＤの近さの平均を取り0.5となる。この値は、閾値0.6以下であるため、クラスタ２は異種と判別され、その結果が収集単語記憶部１０８に記憶される。

【００５７】

図4に戻り、続いて、出力部１０５は、収集単語記憶部１０８を参照して、収集され、クラスタに分類され、シード単語と同種か異種かを判別された単語を、それらの情報を関連付けて出力（表示）する（ステップS400）。例えば、出力部１０５は、「クラスタ１｛レストランＡ、レストランＢ｝：同種、クラスタ２｛うどんＣ、うどんＤ｝：異種」等と出力する。以上で、辞書作成処理は終了する。

【００５８】

このように、本実施形態では、辞書増殖処理によって収集された各単語は、クラスタに分類される。そして、各クラスタ毎に、シード単語と同じ種類の単語から構成されるか否かが判別されて出力される。従って、どのような異種の単語が収集されているのかをユーザに好適に出力することができる。

【００５９】
（第２実施形態）
第２実施形態に係る辞書作成装置２００は、図9に示すように、第１実施形態の辞書作成装置１００に、単語選択部２０１、再実行部２０２、および、単語グループ記憶部２０３が追加された構成である。なお、下記及び図面では、第１実施形態と同様のものについては、同一の符号を付す。また、第１実施形態と同様の構成要素の詳細な説明は、上記第１実施形態の説明に準じ、詳細な説明を省略する。
【００６０】
単語グループ記憶部２０３には、図10（Ａ）、図10（Ｂ）に示すように、収集した単語と、該単語が属するグループの識別情報であるグループ名とが対応付けられて記憶される。
【００６１】
単語選択部２０１は、単語グループ記憶部２０３を参照して、未収集のグループを１つ選択し、選択したグループから所定数の単語を選択する。そして、単語選択部２０１は、選択した単語をシード単語とした辞書増殖処理の実行を辞書増殖部１０２に指示する。
【００６２】
再実行部２０２は、収集され、クラスタに分類され、シード単語と同種か異種かを判別された単語にグループ名を付して単語グループ記憶部２０３に追加する。そして、再実行部２０２は、未だ収集を行っていないグループがある場合には、そのグループから単語を選択することを単語選択部２０１に指示をする。
【００６３】
なお、その他の各部（入力部１０１、辞書増殖部１０２、クラスタリング部１０３、種別判別部１０４、出力部１０５、文書記憶部１０６、収集過程記憶部１０７、収集単語記憶部１０８）は、第１実施形態と同様の処理を行うため、ここでは説明を省略する。但し、辞書増殖部１０２が単語収集の起点とするシード単語は、単語選択部２０１が選択した単語である。
【００６４】
続いて、辞書作成装置２００で実施される処理の動作について説明する。なお、予め、単語グループ記憶部２０３には、複数の単語が、グループ１として登録されている。また、このグループ１は、後述する収集未完グループであるとする。また、グループ１以外のグループは現時点では登録されていないものとする。
【００６５】
まず、ユーザは、入力部１０１を操作して、辞書を作成することを指示する。この指示操作に応じて、辞書作成装置２００は、図11に示す辞書作成処理を行う。
【００６６】
辞書作成処理が開始されると、単語選択部２０１は、単語グループ記憶部２０３を参照して、未収集のグループ（即ちグループ１）に含まれる単語のなかから、予め設定されている数の単語をシード単語として選択する（ステップS50）。
【００６７】
続いて、辞書増殖部１０２は、第１実施形態と同様に辞書増殖処理を行い、シード単語と同種の単語を収集する（ステップS100）。但し、ここでは、ステップS50で選択された単語をシード単語とする。
【００６８】
続いて、クラスタリング部１０３は、第１実施形態と同様にクラスタリング処理を行い、辞書増殖処理によって収集された単語をクラスタに分類する（ステップS200）。
【００６９】
続いて、種別判別部１０４は、第１実施形態と同様に同種判別処理を行い、クラスタが、シード単語と同種の単語から構成されるか否かを判別する（ステップS300）。
【００７０】
続いて、再実行部２０２は、シード単語と同種か異種かを判別されたクラスタ毎に、該

クラスタを構成する単語を単語グループ記憶部２０３に登録して、グルーピングする単語グループ更新処理を行う（ステップS330）。

【００７１】

図12に、単語グループ更新処理の詳細を示す。単語グループ更新処理が開始されると、まず、再実行部２０２は、上述のステップS200でクラスタリングしたクラスタのなかから未処理のクラスタを１つ選択する（ステップS331）。

【００７２】

続いて、再実行部２０２は、ステップS300の同種判別処理の結果を参照して、選択したクラスタがシード単語と同種の単語から構成されているか否かを判別する（ステップS332）。

【００７３】

シード単語と同種の場合（ステップS332；Yes）、再実行部２０２は、シード単語と同じグループ名を付して、選択したクラスタ内の単語を単語グループ記憶部２０３に登録する（ステップS333）。そして、ステップS337に処理を移す。

【００７４】

シード単語と異種の場合（ステップS332；No）、再実行部２０２は、単語グループ記憶部２０３を参照して、選択したクラスタ内の単語のなかに、既に単語グループ記憶部２０３に記憶されている単語（既存単語）があるか否かを判別する（ステップS334）。

【００７５】

既存単語があると判別された場合（ステップS334；Yes）、再実行部２０２は、その既存単語に付されているグループ名と同じグループ名を付して、選択したクラスタ内の単語を単語グループ記憶部２０３に登録する（ステップS335）。そして、ステップS337に処理を移す。

【００７６】

既存単語がないと判別された場合（ステップS334；No）、再実行部２０２は、新たに発行したグループ名を付して、選択したクラスタ内の単語を単語グループ記憶部２０３に登録する（ステップS336）。そして、ステップS337に処理を移す。

【００７７】

ステップS337では、再実行部２０２は、クラスタリングした全てのクラスタで、クラスタ内の単語を単語グループ記憶部２０３に登録する処理を行ったか否かを判別する。

【００７８】

未だ単語グループ記憶部２０３に登録する処理を行っていないクラスタがある場合（ステップS337；No）、再実行部２０２は、未処理のクラスタを選択して、クラスタ内の単語を単語グループ記憶部２０３に登録する一連の処理（ステップS331～ステップS336）を繰り返す。

【００７９】

全てのクラスタで、単語を単語グループ記憶部２０３に登録する処理を行った場合（ステップS337；Yes）、単語グループ更新処理は終了する。

【００８０】

図11に戻り、続いて、再実行部２０２は、単語収集が未だ完了していないグループ（以下、収集未完グループという）があるか否かを判別する（ステップS360）。

例えば、以下に示すａ）～ｄ）の何れかの条件を満たすグループを収集未完グループと判断すればよい。

ａ）グループ内の単語数が一定数以上に達していないグループ。

ｂ）グループ内の単語をシード単語とした辞書増殖処理を所定回数以上行っていないグループ。

ｃ）グループに新たに追加された単語が一定数以上あるグループ。

ｄ）ａ）～ｃ）を所定の重み付けを付した割合で組み合わせた条件に合致するグループ。

【００８１】

収集未完グループが有る場合（ステップS360；Yes）、再実行部２０２は、収集未完グループの１つからシード単語を選択することを単語選択部２０１に指示する。そして、シード単語から単語を収集して、クラスタリングし、シード単語と同種か異種かの判定を行い、グルーピングする処理が繰り返される（ステップS50～ステップS330）。

【００８２】

収集未完グループが無い場合（ステップS360；No）、出力部１０５は、収集した単語を出力する。但し、単語の属するクラスタ、および、そのクラスタがシード単語を同種であるか否かを示す情報に加えて、単語が属するグループ名を単語グループ記憶部２０３から取得する。そして、これらの情報を、収集した単語と関連付けて出力（表示）するものとする。以上で、辞書作成処理は終了する。

【００８３】

続いて、上述した辞書作成処理について、具体例を挙げて説明する。なお、前提として、図10（Ａ）に示すように、収集未完グループであるグループ１のみが、単語グループ記憶部２０３には記憶されているものとする。

【００８４】

従って、この状態で辞書作成処理が開始されると、まず、グループ１内の単語「レストランＳ」と「レストランＴ」が選択される（ステップS50）。続いて、この「レストランＳ」と「レストランＴ」とをシード単語とした辞書増殖処理が実行されて、単語が収集される（ステップS100）。そして、収集された単語は、その結束度に基づいてクラスタリングされ（ステップS200）、クラスタ毎に、シード単語「レストランＳ」「レストランＴ」と同種であるか否かが判別される（ステップS300）。ここでは、以下に示すようなクラスタ１～５が作成されたこととする。

・クラスタ１（同種）：「レストランＡ」「レストランＢ」
・クラスタ２（異種）：「うどんＣ」「うどんＤ」
・クラスタ３（同種）：「レストランＸ」「レストランＺ」「レストランＷ」
・クラスタ４（同種）：「レストランＳ」「レストランＴ」
・クラスタ５（異種）：「うどんＧ」「うどんＨ」

【００８５】

続いて、これらのクラスタ毎に、クラスタ内の単語をグループ化して単語グループ記憶部２０３に登録する単語グループ更新処理が実施される（ステップS330）。この場合、クラスタ１と、クラスタ３と、クラスタ４は、シード単語と同種と判定されているため、これらのクラスタ内の単語は、シード単語と同じグループ１の単語として単語グループ記憶部２０３に登録される（ステップS333）。

【００８６】

また、クラスタ２とクラスタ５は、シード単語と異種の単語であり、また、これらのクラスタ内の単語は未だ単語グループ記憶部２０３に記憶されていない。従って、クラスタ２とクラスタ５内の単語は、それぞれ、グループ２、グループ３の新規のグループ名を付されて、単語グループ記憶部２０３に登録される（ステップS336）。

【００８７】

そして、最終的には、図10（Ｂ）に示すように、クラスタ１～５内の単語がグループ名を付されて単語グループ記憶部２０３に登録される。

【００８８】

続いて、収集未完のグループがある場合には、そのグループ（即ち、グループ２かグループ３）のうちの１つを選択して、選択したグループ内の単語を新たにシード単語とした単語収集を行う一連の処理が繰り返される。

【００８９】

このように、第２実施形態では、異種単語がどの程度含まれているかだけでなく、同じような異種単語を新たなグループとて登録する。そして、そのグループ内の単語をシード単語として、さらに単語を収集することができる。これにより、初期に与えたシード単語と似ている単語も別グループとした単語収集を行うことができる。

【００９０】
（第３実施形態）
第２実施形態では、グループ内の単語から、ランダムに選択した所定数の単語をシード単語として辞書増殖を行った。そのため、少ない収集回数で多くの単語を取得したい場合、収集回数が多くなっても収集される単語がシード単語と類似する精度を高くしたい場合、などといった種々の場面に応じた適切な単語の収集ができない。本実施形態では、種々の場面に応じた適切な単語の収集を可能とすることを特徴とする。
【００９１】
第３実施形態に係る辞書作成装置３００は、図13に示すように、第２実施形態の辞書作成装置２００の単語選択部２０１が第二単語選択部３０１に置き換えられている。また、単語間結束度記憶部３０２が新たに追加されている。なお、下記及び図面では、第１実施形態、および、第２実施形態と同様のものについては、同一の符号を付す。また、第１実施形態、および、第２実施形態と同様の構成要素の詳細な説明は、上記第１実施形態、第２実施形態の説明に準じ、詳細な説明を省略する。
【００９２】
第二単語選択部３０１は、単語グループ記憶部２０３を参照して、未収集のグループを１つ選択し、選択したグループに含まれる単語から複数の単語を選択する。この際、第二単語選択部３０１は、単語間結束度記憶部３０２を参照して、結束度が所定の条件を満たす単語を優先的に選択する。
【００９３】
ここで、上記の所定の条件とは、例えば、「グループ内の単語のうち結束度の高い順に７５％、残りの２５％は結束度が低いものから順に選択する」などの条件である。結束度の高い単語のみを選択すると、頻繁に出現する単語のみが収集されるため、シード単語と類似の単語が収集される精度は高くなるが、収集される単語の数は少なくなり収集効率は悪化する。したがって、収集精度よりも収集効率を重視した単語収集を行いたい場合には、上記のような条件を採用することが望ましい。
また、収集効率よりも収集精度を重視した単語収集を行いたい場合には、「グループ内の単語のうち結束度の高い順に選択する」などの条件を採用することが望ましい。
なお、このような単語選択の条件を定義する条件情報が、予め、辞書作成装置３００の記憶部に記憶されているものとする。
【００９４】
単語間結束度記憶部３０２は、クラスタリング部１０３によって算出された単語間の結束度を記憶する。具体的には、図14に示すように、単語間結束度記憶部３０２には、２つの単語と、その２つの単語間の結束度とが対応付けられて記憶される。例えば、図14の先頭のエントリから、「レストランＳ」と「レストランＴ」との間の結束度は0.9とわかる。
【００９５】
なお、その他の各部（入力部１０１、辞書増殖部１０２、クラスタリング部１０３、種別判別部１０４、出力部１０５、文書記憶部１０６、収集過程記憶部１０７、収集単語記憶部１０８、再実行部２０２、単語グループ記憶部２０３）は、第２実施形態と同様の処理を行うため、ここでは説明を省略する。
【００９６】
続いて、辞書作成装置３００で実施される処理の動作について説明する。
なお、予め、収集の際に採用する結束度に関するグループから単語を選択するための条件が設定されているものとする。また、グループからは４つの単語を選択するものとする。
【００９７】
ユーザは、入力部１０１を操作して、辞書を作成することを指示する。この指示操作に応じて、辞書作成装置３００は、第２実施形態と同様の図11に示す辞書作成処理を行う。
【００９８】

まず、第二単語選択部３０１は、単語グループ記憶部２０３を参照して、未収集のグループを１つ選択し、単語間結束度記憶部３０２を参照して、所定の条件に基づいて、選択したグループ内の単語のうちから所定数（４つ）の単語をシード単語として選択する（ステップS50）。

【００９９】

例えば、「グループ内の単語のうち結束度の高い順に７５％、残りの２５％は結束度が低いものから順に選択する」条件が設定されている場合を考える。即ち、結束度の高い単語を３つ、結束度の低い単語を１つ選択することとなる。

この場合、第二単語選択部３０１は、まず、グループ内の単語のうち、単語間の結束度が最も高い２単語を選択する。次に、第二単語選択部３０１は、その２つの単語それぞれと結束度が最も高い単語を１つ選択する。そして、第二単語選択部３０１は、これら３つの単語それぞれと、結束度の低い単語を１つ選択する。

【０１００】

以降の処理は、第２実施形態と同様である。

即ち、辞書増殖部１０２は、第二単語選択部３０１によって選択された４つの単語をシード単語として、同種の単語を収集する辞書増殖処理を行う（ステップS100）。続いて、クラスタリング部１０３が、収集された単語をクラスタリングする（ステップS200）。なお、この際、クラスタリング部１０３は、クラスタリングするために算出した単語とその単語間の結束度とを、単語間結束度記憶部３０２に記録する。そして、種別判別部１０４が、クラスタ毎に、クラスタがシード単語と同種の単語から構成されるか否かを判別する（ステップS300）。そして、再実行部２０２が、収集した単語をグルーピングする（ステップS330）。そして、未収集のグループがある場合は（ステップS360；Yes）、未収集のグループからシード単語を選択して単語を収集する処理を繰り返し、未収集のグループがない場合は（ステップS360；No）、処理は終了する。

【０１０１】

このように、本実施形態では、グループ内の単語をランダムに選択するのではなく、単語間の結束度を考慮して単語を選択する。従って、種々の場面に対応した単語収集が可能となる。

【０１０２】

なお、本各実施形態は種々の変形、および、応用が可能である。

例えば、上記各実施形態では、文書記憶部１０６に記憶されている文書から単語を抽出したが、これに限らず、例えば、インターネット検索エンジンを用いて、インターネット上のＷｅｂページから、単語を抽出してもよい。

【０１０３】

図15は、本発明の各実施形態に係る辞書作成装置１００，２００，３００をコンピュータに実装する場合の、物理的な構成の一例を示すブロック図である。本発明の各実施形態に係る辞書作成装置１００，２００，３００は、一般的なコンピュータ装置と同様のハードウェア構成によって実現することができる。辞書作成装置１００，２００，３００は、制御部２１、主記憶部２２、外部記憶部２３、操作部２４、表示部２５および入出力部２６を備える。主記憶部２２、外部記憶部２３、操作部２４、表示部２５および入出力部２６はいずれも内部バス２０を介して制御部２１に接続されている。

【０１０４】

制御部２１はＣＰＵ（Central Processing Unit）等から構成され、外部記憶部２３に記憶されている制御プログラム３０に従って、前述した各実施形態における辞書作成処理を実行する。

【０１０５】

主記憶部２２はＲＡＭ（Random-Access Memory）等から構成され、外部記憶部２３に記憶されている制御プログラム３０をロードし、制御部２１の作業領域として用いられる。

【０１０６】

外部記憶部２３は、フラッシュメモリ、ハードディスク、ＤＶＤ－ＲＡＭ（Digital Ve

rsatile Disc Random-Access Memory）、ＤＶＤ－ＲＷ（Digital Versatile Disc ReWritable）等の不揮発性メモリから構成され、上述の処理を制御部２１に行わせるための制御プログラム３０を予め記憶する。また、外部記憶部２３は、制御部２１の指示に従って、この制御プログラム３０が記憶するデータを制御部２１に供給し、制御部２１から供給されたデータを記憶する。また、外部記憶部２３は、上述した各実施形態における文書記憶部１０６、収集過程記憶部１０７、収集単語記憶部１０８、単語グループ記憶部２０３、および、単語間結束度記憶部３０２を物理的に実現する。

【０１０７】

操作部２４はキーボードおよびマウスなどのポインティングデバイス等と、キーボードおよびポインティングデバイス等を内部バス２０に接続するインターフェース装置等から構成されている。操作部２４を介して、シード単語や辞書作成処理の開始の指示が制御部２１に供給される。

【０１０８】

表示部２５は、ＣＲＴ（Cathode Ray Tube）またはＬＣＤ（Liquid Crystal Display）などから構成され、種々の情報を表示する。例えば、表示部２５は、収集された各単語を、クラスタ毎に、シード単語と同種であるか異種であるかの情報を付して表示する。

【０１０９】

入出力部２６は、無線送受信機、無線モデムまたは網終端装置、およびそれらと接続するシリアルインタフェースまたはＬＡＮ（Local Area Network）インタフェース等から構成されている。例えば、入出力部２６を介して、インターネット上のWebページから単語を収集してもよい。

【０１１０】

図1、図9、および図13に示す辞書作成装置１００，２００，３００の辞書増殖部１０２、クラスタリング部１０３、種別判別部１０４、出力部１０５、単語選択部２０１、再実行部２０２、および、第二単語選択部３０１の処理は、制御プログラム３０が、制御部２１、主記憶部２２、外部記憶部２３、操作部２４、表示部２５および入出力部２６などを資源として用いて処理することによって実行する。

【０１１１】

なお、前記のハードウエェア構成やフローチャートは一例であり、任意に変更および修正が可能である。

【０１１２】

また、制御部２１、主記憶部２２、外部記憶部２３、操作部２４、入出力部２６および内部バス２０などから構成される辞書作成装置１００，２００，３００の処理を行う中心となる部分は、専用のシステムによらず、通常のコンピュータシステムを用いて実現可能である。たとえば、前記の動作を実行するためのコンピュータプログラムを、コンピュータが読み取り可能な記録媒体（フレキシブルディスク、ＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭ等）に格納して配布し、当該コンピュータプログラムをコンピュータにインストールすることにより、前記の処理を実行する辞書作成装置１００，２００，３００を構成してもよい。また、インターネット等の通信ネットワーク上のサーバ装置が有する記憶装置に当該コンピュータプログラムを格納しておき、通常のコンピュータシステムがダウンロード等することで辞書作成装置１００，２００，３００を構成してもよい。

【０１１３】

また、辞書作成装置１００，２００，３００の機能を、ＯＳ（オペレーティングシステム）とアプリケーションプログラムの分担、またはＯＳとアプリケーションプログラムとの協働により実現する場合などには、アプリケーションプログラム部分のみを記録媒体や記憶装置に格納してもよい。

【０１１４】

また、搬送波にコンピュータプログラムを重畳し、通信ネットワークを介して配信することも可能である。たとえば、通信ネットワーク上の掲示板(BBS, Bulletin Board System)に前記コンピュータプログラムを掲示し、ネットワークを介して前記コンピュータプロ

グラムを配信してもよい。そして、このコンピュータプログラムを起動し、ＯＳの制御下で、他のアプリケーションプログラムと同様に実行することにより、前記の処理を実行できるように構成してもよい。

【０１１５】

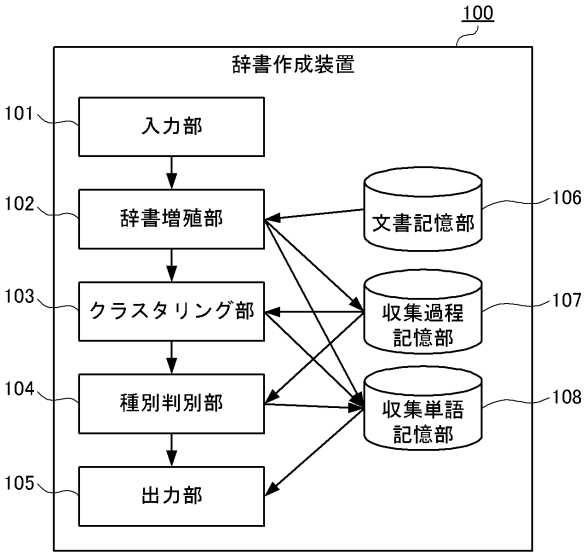
本発明は２００９年１２月１１日に出願された日本国特許出願２００９－２８２３０４号に基づく。本明細書中に日本国特許出願２００９－２８２３０４号の明細書、特許請求の範囲、図面全体を参照として取り込むものとする。

【符号の説明】

【０１１６】

１００　辞書作成装置
１０１　入力部
１０２　辞書増殖部
１０３　クラスタリング部
１０４　種別判別部
１０５　出力部
１０６　文書記憶部
１０７　収集過程記憶部
１０８　収集単語記憶部

【図１】



【図２】

収集過程記憶部

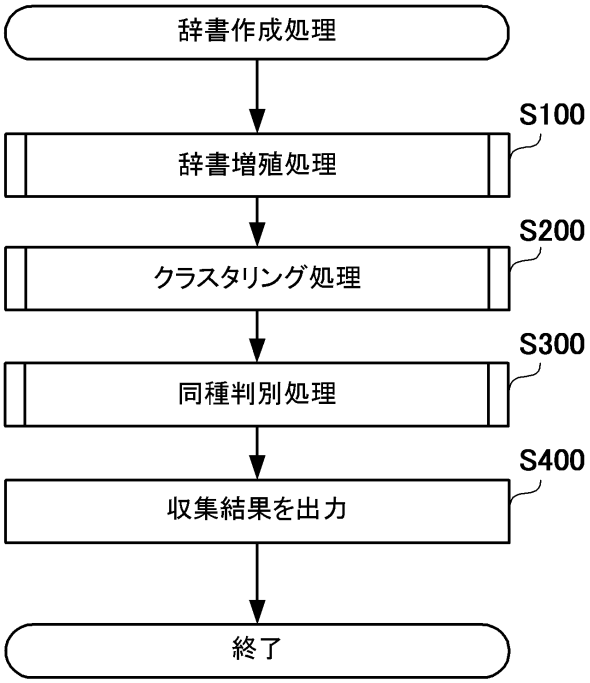
ターン数	入力単語	出力単語
1	レストランＳ	レストランＸ
1	レストランＳ	レストランＺ
1	レストランＳ	レストランＷ
1	レストランＴ	レストランＸ
1	レストランＴ	レストランＺ
1	レストランＴ	レストランＷ
2	レストランＸ	レストランＡ
2	レストランＳ	レストランＡ
2	レストランＳ	レストランＢ
2	レストランＺ	うどんＣ
2	レストランＷ	うどんＣ
2	レストランＺ	うどんＤ
2	レストランＷ	うどんＤ
3	レストランＡ	レストランＥ
3	レストランＡ	レストランＴ
3	レストランＢ	レストランＴ
3	うどんＣ	うどんＧ
3	うどんＣ	うどんＨ
3	うどんＤ	レストランＴ
3	うどんＤ	うどんＨ
…	…	…

【図３】

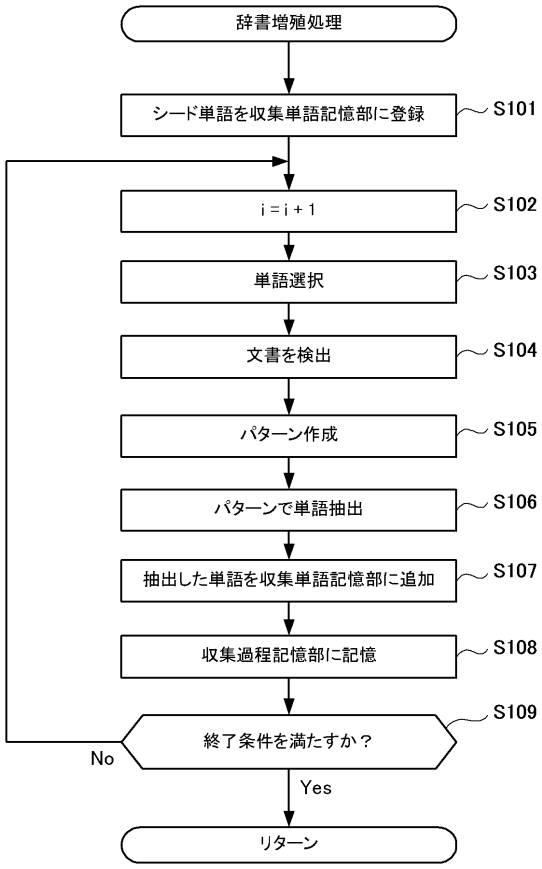
収集単語記憶部

収集単語	クラスタＩＤ	同種 or 異種？
レストランＡ	クラスタ１	同種
レストランＢ		
うどんＣ	クラスタ２	異種
うどんＤ		

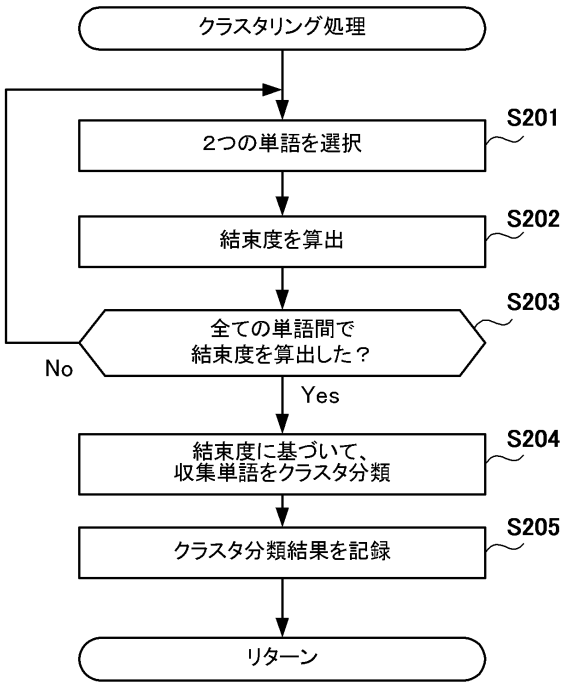
【図４】



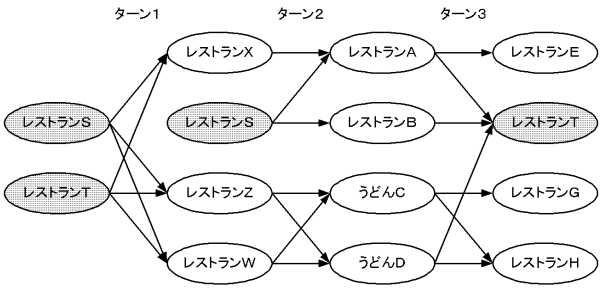
【図５】



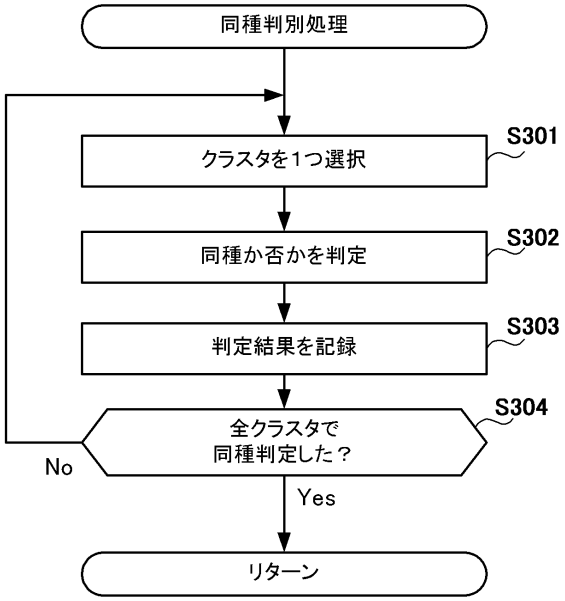
【図６】



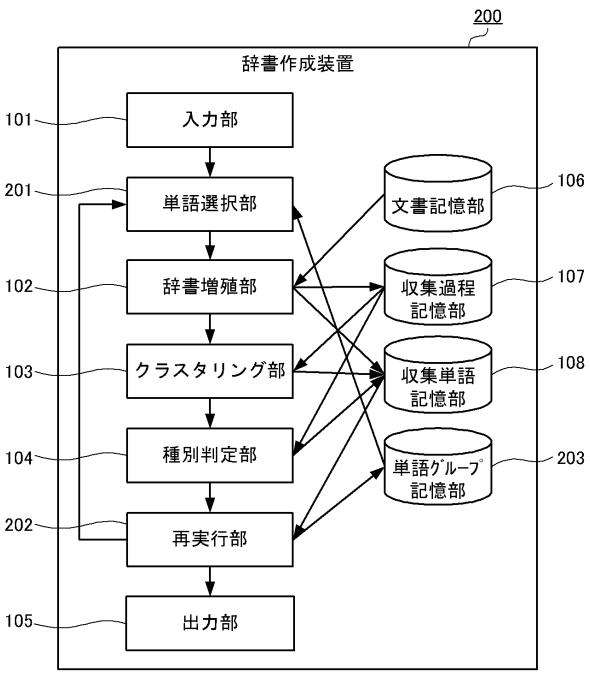
【図７】



【図８】



【図９】



【図１０】

(A)

単語グループ記憶部

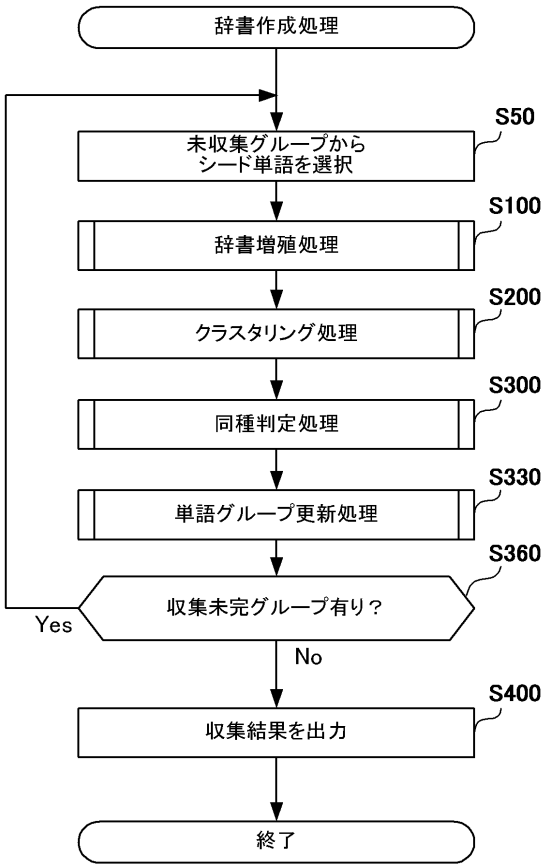
単語	グループ名
レストランS	グループ１
レストランT	

(B)

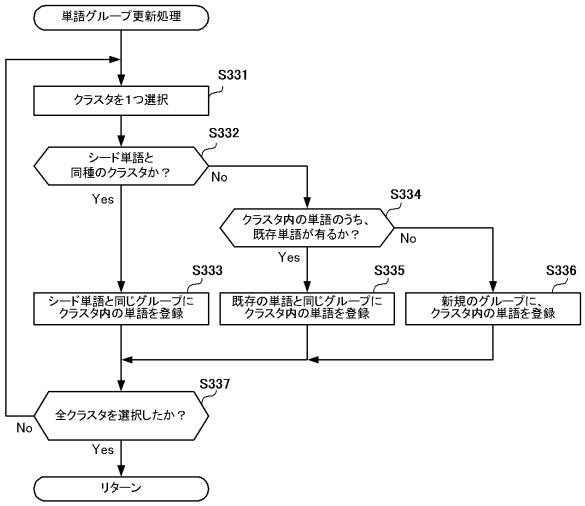
単語グループ記憶部

単語	グループ名
レストランS	グループ１
レストランT	
レストランX	
レストランW	
レストランZ	
レストランA	
レストランB	
うどんC	グループ２
うどんD	
うどんG	グループ３
うどんH	

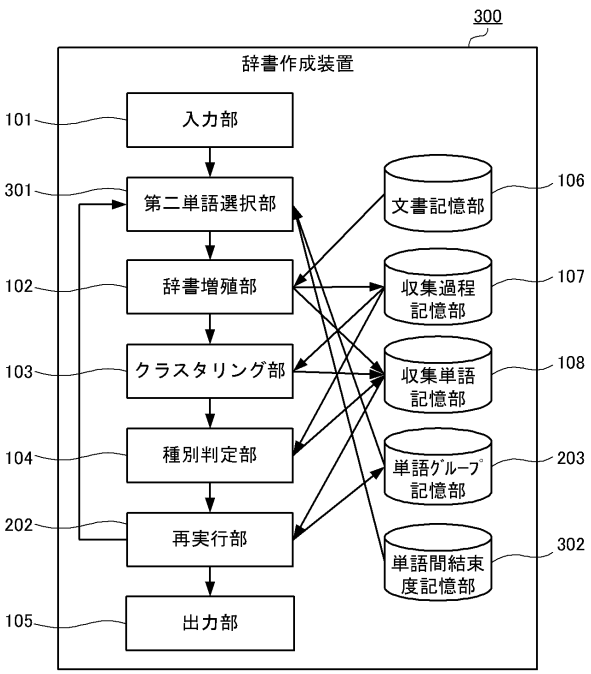
【図１１】



【図１２】



【図１３】

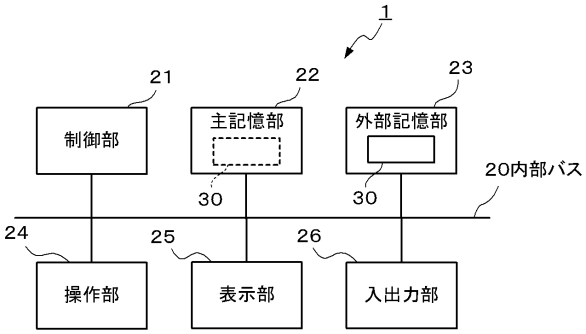


【図１４】

収集単語記憶部

単語１	単語２	結束度
レストランＳ	レストランＴ	0.9
レストランＡ	レストランＢ	0.8
レストランＸ	レストランＹ	0.8
レストランＷ	レストランＸ	0.9
うどんＣ	うどんＤ	0.75
うどんＤ	うどんＨ	0.8
レストランＳ	レストランＸ	0.5
…	…	…

【図１５】



フロントページの続き

審査官　松田　直也

(56)参考文献　特開２００７－２０７２１８（ＪＰ，Ａ）
水口　弘紀、河合　英紀、土田　正明、久寿居　大，Ｗｅｂ知識を利用したブートストラップによる辞書増殖手法，電子情報通信学会　第18回データ工学ワークショップ論文集，[online]，日本，電子情報通信学会，２００７年　6月　1日，[検索日　2014.6.3],インターネット，ＵＲＬ，http://www.ieice.org/~de/DEWS/DEWS2007/pdf/e8-5.pdf
大島　裕明、田中　克己，正解語ペア漸増による関連語取得のための両方向構文パターン発見，第1回データ工学と情報マネジメントに関するフォーラム　－ＤＥＩＭフォーラム－　論文集，[online]，日本，電子情報通信学会，２００９年　5月　9日，[検索日　2014.6.3]，インターネット，ＵＲＬ，http://db-event.jpn.org/deim2009/proceedings/files/B9-1.pdf

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０
Ｇ０６Ｆ　１７／２７