(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4803709号
(P4803709)
(45)発行日　平成23年10月26日(2011.10.26)　(24)登録日　平成23年8月19日(2011.8.19)

(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 22　(全 18 頁)

(21)出願番号　特願2005-203157(P2005-203157)
(22)出願日　平成17年7月12日(2005.7.12)
(65)公開番号　特開2007-25788(P2007-25788A)
(43)公開日　平成19年2月1日(2007.2.1)
審査請求日　平成20年6月11日(2008.6.11)

特許法第３０条第１項適用　進藤三佳、内元清貴、井佐原均，「コーパス・シソーラスに基づいた英語形容詞の意味拡張の調査・分析」，第１９回ことば工学研究会資料，ｐ．２３－３３，社団法人人工知能学会，２００５年３月５日発行　進藤三佳、内元清貴、井佐原均，「感覚・知覚領域に起源を持つ英語形容詞の意味拡張の調査・分析」，言語処理学会第１１回年次大会　発表論文集，ｐ１１０７－１１１０，言語処理学会，２００５年３月１５日発行

(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　100130498
弁理士　佐野　禎哉
(72)発明者　内元　清貴
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　進藤　三佳
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　井佐原　均
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】単語用法差異情報取得プログラム及び同装置

(57)【特許請求の範囲】
【請求項１】
同一又は類似の意味を有する複数のターゲット単語について、例文データベースであるコーパス、及び語と語の上位下位概念の関係が記述されたデータベースであるシソーラスを検索可能に備え又は接続したコンピュータに、各ターゲット単語の用法の違いに関する情報を抽出し出力させるためのプログラムであって、当該コンピュータに、
複数のターゲット単語の入力を受け付けるターゲット単語入力ステップと、
前記コーパスにアクセスして、前記ターゲット単語入力ステップで受け付けた各ターゲット単語で検索して当該ターゲット単語をそれぞれ含む文データを抽出する文抽出ステップと、
前記文抽出ステップで抽出した文データをそれぞれ構文解析し、各文データに含まれるターゲット単語と文法的関係にある名詞を抽出する名詞抽出ステップと、
前記シソーラスにアクセスして、前記名詞抽出ステップで抽出した名詞で検索し、この名詞及びその上位概念を表すノードを抽出するとともに、それらノードと、これらノード同士の上位下位概念のつながりを表すリンクとから構成される有向グラフを、対応するターゲット単語ごとに生成する有向グラフ生成ステップと、
前記有向グラフ生成ステップで生成した各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する差異抽出ステップと、
前記差異抽出ステップで抽出した有向グラフの差異を、前記ターゲット単語の用法の違いに関する情報として出力する差異出力ステップと、

を実行させることを特徴とする単語用法差異情報取得プログラム。
【請求項２】
前記差異抽出ステップにおいて、前記コンピュータに、
各有向グラフにおいて、同一のノード又は同一のノード及びリンクを有する部分を共有化させて各有向グラフを重ね合わせることによって、異なるノードを抽出する処理を実行させる、請求項１に記載の単語用法差異情報取得プログラム。
【請求項３】
前記ターゲット単語入力ステップで受け付けたターゲット単語が３つ以上の場合、
前記差異抽出ステップにおいて、前記コンピュータに、
特定の一のターゲット単語以外の複数のターゲット単語について前記有向グラフ生成ステップで生成した各有向グラフを合成して共通の有向グラフを生成し、この共通の有向グラフと前記特定の一のターゲット単語の有向グラフとを比較して、これら有向グラフ間において異なるノードを抽出し、この工程を各ターゲット単語ごとに繰り返し行う処理を実行させる、請求項１又は２に記載の単語用法差異情報取得プログラム。
【請求項４】
前記コンピュータに、
前記名詞抽出ステップにおいて、各文データに含まれるターゲット単語と文法的関係にある名詞を、当該名詞が前記ターゲット単語と共に文データに出現する頻度に関するデータと併せて抽出する処理を実行させ、
前記有向グラフ生成ステップにおいて、生成する有向グラフの各ノードに前記頻度に関するデータによる重み付けする処理を実行させ、
前記差異抽出ステップにおいて、前記有向グラフ生成ステップで生成した重み付けが施された有向グラフを利用して、各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する処理を実行させる、請求項１乃至３の何れかに記載の単語用法差異情報取得プログラム。
【請求項５】
前記名詞抽出ステップにおいて、頻度に関するデータとして対応するターゲット単語に対して抽出された全名詞に占める当該名詞の頻度の割合を表す頻度比率を適用し、
前記コンピュータに、
前記有向グラフ生成ステップにおいて、生成する有向グラフにおいて前記名詞に対応するノードに前記頻度を付与するとともに当該ノードの上位概念のノードにその下位のノードの頻度の合計値を付与し、全ノードに個々の頻度を正規化した頻度比率を付与することで、有向グラフにこの頻度比率に基づく重み付けする処理を実行させ、
前記差異抽出ステップにおいて、前記有向グラフ生成ステップで生成した重み付けが施された比較対象となる２つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを前記異なるノードである差異ノードに組み入れ、当該差異ノードを抽出する処理を実行させる、請求項４に記載の単語用法差異情報取得プログラム。
【請求項６】
前記差異抽出ステップにおいて、前記コンピュータに、
比較対象となる２つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを暫定的に差異ノードとして有向グラフの差異部分に組み入れ、当該差異部分のうち各最上位ノードを各ターゲット単語について頻度比率が大きい方から順に所定数ずつ抽出し、その抽出したノードのうち共通するノードの割合を算出する工程を、前記頻度比率を逓減させながら繰り返すことで、各々の工程で得られた共通するノードの割合が一定値以上である場合、その共通するノードの割合を前回の工程で得られた共通するノードの割合と比較して、その比較した値が一定値以上である場合に、当該工程で暫定的に差異ノードと決定したノードを差異ノードとして決定し、当該差異ノードを抽出する処理を実行させる、請求項５に記載の単語用法差異情報取得プログラム。

【請求項７】
前記頻度に関するデータとして、前記頻度比率に代えて、頻度の値自体を適用している請求項５又は６に記載の単語用法差異情報取得プログラム。
【請求項８】
前記コンピュータに、
前記差異抽出ステップにおいて、抽出した異なるノードを、頻度に基づく重みの大きい方から順に所定数のノードをさらに抽出する処理を実行させ、
前記差異出力ステップにおいて、前記所定数のノードを前記用法の違いに関する情報として出力する処理を実行させる、請求項４乃至７の何れかに記載の単語用法差異情報取得プログラム。
【請求項９】
前記コンピュータに、
前記差異出力ステップにおいて、前記異なるノードのうち最上位のノードを前記用法の違いに関する情報として出力する処理を実行させる、請求項１乃至８の何れかに記載の単語用法差異情報取得プログラム。
【請求項１０】
前記コンピュータに、
前記差異出力ステップにおいて、異なるノードのうち最上位のノードに加えて又はそれに代えて共通のノードの最下位のノードを前記用法の違いに関する情報として出力する処理を実行させる、請求項１乃至８の何れかに記載の単語用法差異情報取得プログラム。
【請求項１１】
前記ターゲット単語入力ステップにおいて入力受付可能なターゲット単語の品詞を、形容詞又は動詞に制限している、請求項１乃至１０の何れかに記載の単語用法差異情報取得プログラム。
【請求項１２】
プログラムに従って作動するコンピュータにより構成され、入力された同一又は類似の意味を有する複数のターゲット単語について、各ターゲット単語の用法の違いに関する情報を抽出し出力させる単語用法差異情報取得装置であって、前記コンピュータは、例文データベースであるコーパス、及び語と語の上位下位概念の関係が記述されたデータベースであるシソーラスを検索可能に備え又は接続しており、
複数のターゲット単語の入力を受け付けるターゲット単語入力手段と、
前記コーパスにアクセスして、前記ターゲット単語入力手段で受け付けた各ターゲット単語で検索して当該ターゲット単語をそれぞれ含む文データを抽出する文抽出手段と、
前記文抽出手段で抽出した文データをそれぞれ構文解析し、各文データに含まれるターゲット単語と文法的関係にある名詞を抽出する名詞抽出手段と、
前記シソーラスにアクセスして、前記名詞抽出手段で抽出した名詞で検索し、この名詞及びその上位概念を表すノードを抽出するとともに、それらノードと、これらノード同士の上位下位概念のつながりを表すリンクとから構成される有向グラフを、対応するターゲット単語ごとに生成する有向グラフ生成手段と、
前記有向グラフ生成手段で生成した各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する差異抽出手段と、
前記差異抽出手段で抽出した有向グラフの差異を、前記ターゲット単語の用法の違いに関する情報として出力する差異出力手段と、
を具備してなることを特徴とする単語用法差異情報取得装置。
【請求項１３】
前記差異抽出手段において、
各有向グラフにおいて、同一のノード又は同一のノード及びリンクを有する部分を共有化させて各有向グラフを重ね合わせることによって、異なるノードを抽出する処理を実行する、請求項１２に記載の単語用法差異情報取得装置。
【請求項１４】

前記ターゲット単語入力手段で受け付けたターゲット単語が３つ以上の場合、
前記差異抽出手段において、特定の一のターゲット単語以外の複数のターゲット単語について前記有向グラフ生成ステップで生成した各有向グラフを合成して共通の有向グラフを生成し、この共通の有向グラフと前記特定の一のターゲット単語の有向グラフとを比較して、これら有向グラフ間において異なるノードを抽出し、この工程を各ターゲット単語ごとに繰り返し行う処理を実行する、請求項１２又は１３に記載の単語用法差異情報取得装置。
【請求項１５】
前記名詞抽出手段において、各文データに含まれるターゲット単語と文法的関係にある名詞を、当該名詞が前記ターゲット単語と共に文データに出現する頻度に関するデータと併せて抽出する処理を実行し、
前記有向グラフ生成手段において、生成する有向グラフの各ノードに前記頻度に関するデータによる重み付けする処理を実行し、
前記差異抽出手段において、前記有向グラフ生成手段で生成した重み付けが施された有向グラフを利用して、各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する処理を実行する、請求項１２乃至１４の何れかに記載の単語用法差異情報取得装置。
【請求項１６】
前記名詞抽出手段において、頻度に関するデータとして対応するターゲット単語に対して抽出された全名詞に占める当該名詞の頻度の割合を表す頻度比率を適用し、
前記有向グラフ生成手段において、生成する有向グラフにおいて前記名詞に対応するノードに前記頻度を付与するとともに当該ノードの上位概念のノードにその下位のノードの頻度の合計値を付与し、全ノードに個々の頻度を正規化した頻度比率を付与することで、有向グラフにこの頻度比率に基づく重み付けする処理を実行し、
前記差異抽出手段において、前記有向グラフ生成手段で生成した重み付けが施された比較対象となる２つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを前記異なるノードである差異ノードに組み入れ、当該差異ノードを抽出する処理を実行する、請求項１５に記載の単語用法差異情報取得装置。
【請求項１７】
前記差異抽出手段において、比較対象となる２つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを暫定的に差異ノードとして有向グラフの差異部分に組み入れ、当該差異部分のうち各最上位ノードを各ターゲット単語について頻度比率が大きい方から順に所定数ずつ抽出し、その抽出したノードのうち共通するノードの割合を算出する工程を、前記頻度比率を逓減させながら繰り返すことで、各々の工程で得られた共通するノードの割合が一定値以上である場合、その共通するノードの割合を前回の工程で得られた共通するノードの割合と比較して、その比較した値が一定値以上である場合に、当該工程で暫定的に差異ノードと決定したノードを差異ノードとして決定し、当該差異ノードを抽出する処理を実行する、請求項１６に記載の単語用法差異情報取得装置。
【請求項１８】
前記頻度に関するデータとして、前記頻度比率に代えて、頻度の値自体を適用している請求項１６又は１７に記載の単語用法差異情報取得装置。
【請求項１９】
前記差異抽出手段において、抽出した異なるノードを、頻度に基づく重みの大きい方から順に所定数のノードをさらに抽出する処理を実行し、
前記差異出力手段において、前記所定数のノードを前記用法の違いに関する情報として出力する処理を実行する、請求項１５乃至１８の何れかに記載の単語用法差異情報取得装置。
【請求項２０】

前記差異出力手段において、前記異なるノードのうち最上位のノードを前記用法の違いに関する情報として出力する処理を実行する、請求項１２乃至１９の何れかに記載の単語用法差異情報取得装置。
【請求項２１】
前記差異出力手段において、異なるノードのうち最上位のノードに加えて又はそれに代えて共通のノードの最下位のノードを前記用法の違いに関する情報として出力する処理を実行する、請求項１２乃至１９の何れかに記載の単語用法差異情報取得装置。
【請求項２２】
前記ターゲット単語入力手段において入力受付可能なターゲット単語の品詞を、形容詞又は動詞に制限している、請求項１２乃至２１の何れかに記載の単語用法差異情報取得装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、複数の類義語の用法の相違を自動的に解析するためのプログラム及び装置に関するものである。
【背景技術】
【０００２】
複数の同義語や類義語を文中（発話文、記述文とも）で正確に使い分けることは、その言語を外国語として学習する者が難しいと感じるだけでなく、その言語を母国語として日常的に使っている者にとっても存外難しいものである。近年では、ワードプロセッシングソフトウェア（ワープロソフト）、外国語学習ソフトウェア翻訳ソフトウェア等が日常的に使用されるようになっており、これらソフトウェアには様々な入力・編集・出力支援機能が備えられていることがあるが、ユーザが類義語を用例の違いによって的確に使い分けられるように自動的に峻別したり指摘したりすることは実現されていない。
【０００３】
一つの試みとして、単語同士の共起の度合いによって、その単語がどのような単語と共に使われやすいか、という言語学的な研究がなされている（非特許文献１参照）。この研究では、入力文を構文解析し、一文に出現する構文的に関係のある単語同士について、その際に偶然性を排除する処理を施したうえで共起のスコアを計測し、そのスコアをソートすることによって高スコアの単語同士は構文的に関係が深いと推定される、というものである。この場合、複数の類義語をターゲットとしてそれぞれ共起スコアが高い単語を抽出すれば、どの単語にはどのような用例があるかを推測することは可能である。
【非特許文献１】Stefan Th. Griesand Anatol Stefanowitsch, “Extending collostructional analysys：A corpus-based perspective on alternations”, International Journal ofCorpus Linguistics, 9:1, 2004
【発明の開示】
【発明が解決しようとする課題】
【０００４】
しかしながら、上記文献に記載の方法では、具体的な単語同士が一文に出現しているか否かという情報のみを以て、その単語同士に関係があると推定しているため、共起相手の単語が異なれば、ターゲットとする単語を用いることが正しい用法であるのかは不明である。換言すれば、上記文献に記載の方法では、その文においてターゲットとする単語をその類義語に置換しても正しい用法であるといえるのか、どのような単語をターゲットとすれば正しい用法となるのか判断することができない。そのため、如何なる文であれば複数の同義語のうちどれを用いるべきか、というような精度の高い情報を得ることが求められている。
【０００５】
本発明は、このような課題を解決するためになされたものであり、同義語や類義語についてどのような意味の語と一緒に使われることが多いかという用法の違いに関する汎用性

の高い情報を、高精度で自動的に得られるようにすることを目的としている。
【課題を解決するための手段】
【０００６】
すなわち本発明の単語用法差異情報取得プログラムは、同一又は類似の意味を有する複数のターゲット単語について、例文データベースであるコーパス、及び語と語の上位下位概念の関係が記述されたデータベースであるシソーラスを検索可能に備え又は接続したコンピュータに、各ターゲット単語の用法の違いに関する情報を抽出し出力させるためのプログラムである。そして、このコンピュータに、複数のターゲット単語の入力を受け付けるターゲット単語入力ステップと、コーパスにアクセスして、ターゲット単語入力ステップで受け付けた各ターゲット単語で検索して当該ターゲット単語をそれぞれ含む文データを抽出する文抽出ステップと、文抽出ステップで抽出した文データをそれぞれ構文解析し、各文データに含まれるターゲット単語と文法的関係にある名詞を抽出する名詞抽出ステップと、シソーラスにアクセスして、名詞抽出ステップで抽出した名詞で検索し、この名詞及びその上位概念を表すノードを抽出するとともに、それらノードと、これらノード同士の上位下位概念のつながりを表すリンクとから構成される有向グラフを、対応するターゲット単語ごとに生成する有向グラフ生成ステップと、有向グラフ生成ステップで生成した各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する差異抽出ステップと、前記差異抽出ステップで抽出した有向グラフの差異を、前記ターゲット単語の用法の違いに関する情報として出力する差異出力ステップと、を実行させることを特徴としている。
【０００７】
ここで、前記入力ステップで入力されるターゲット単語の数は、２つ以上であれば特に制限されることはないが、本発明の趣旨からそれらの単語は何れも同一又は類似の意味を有している同義語又は類義語であるとする。前記名詞抽出ステップにおける構文解析処理は、一般に知られている形態素解析ソフトウェアや構文解析ソフトウェアの機能を利用することができる。また、同ステップにおいて抽出される名詞が、「ターゲット単語と文法的関係にある」とは、その名詞とターゲット単語とが、修飾・被修飾関係や述語・項関係（例えば主述関係等）を意味し、上記構文解析処理によって得られる情報からこれらの関係を特定することができる。さらに、前記差異抽出ステップにおいて抽出される「異なるターゲット単語の有向グラフ間において異なるノード」とは、１つ以上のノードであればその数は特に制限のない限り限定されず、比較される有向グラフ同士で異なると判断されるノードのうち最上位のノードのみを示す場合があり得、その最上位のノード以下のノードを含む場合もあり得る。このようなプログラムは、単独のプログラムとして利用できることはもちろんであるが、他のプログラム、例えば文書入力プログラムや翻訳プログラム、言語学習プログラム等の一部に組み込んで利用することもできる。なお、入力される複数のターゲット単語は、同一であっても許容される。すなわち、入力される複数の同義語又は類義語には、同一の単語も含まれる。例えば、同一のターゲット単語が入力された場合、それらターゲット単語同士で、修飾・被修飾の関係にある名詞と主述関係にある名詞とを比較して、どのような意味の名詞と文法的関係があるときにはどちらの関係を使いやすいか、といった用法の違いを得ることも可能である。このことは、例えば、ある形容詞がターゲット単語である場合、意味的関係にある名詞が人ならば叙述用法が使われることが多い、というような情報を得ることも有用であることを意味する。
【０００８】
このような本発明のプログラムを利用することで、これまでは得られなかった同義語や類義語の用法の違いに関する情報を、各ターゲット単語間の有向グラフ上におけるノードの差異として得ることができ、単語の正しい用法を学習したり、入力文の自動校正に利用するなどの場面で役立てることができる。特にコーパスやシソーラスは、一般に知られているものを適宜利用することができるが、それらの規模がある程度大きいほど、また格納されるデータの分野に偏りが少ないほど、出力されるターゲット単語の用法の違いに関する情報の信頼度を向上することができる。なお、差異出力ステップにおける情報の出力は

、ディスプレイ等の表示装置に情報を表示させる態様、プリンタ等の印刷装置に情報を印刷させる態様、他のコンピュータに情報を送信する態様など、種々の態様を適宜選択することができる。

【０００９】

上述のような本発明のプログラムは、例えば、前記差異抽出ステップにおいて、コンピュータに、各有向グラフにおいて、同一のノード又は同一のノード及びリンクを有する部分を共有化させて各有向グラフを重ね合わせることによって、異なるノードを抽出する処理を実行させるものとすることができる。このようにすることで、ターゲット単語ごとに生成された有向グラフの異なる部分のノードを容易に得ることができる。また同時に、共通する部分のノードを得ることも可能となる。

【００１０】

特に、ターゲット単語入力ステップで受け付けたターゲット単語が３つ以上の場合は、前記差異抽出ステップにおいて、コンピュータに、特定の一のターゲット単語以外の複数のターゲット単語について前記有向グラフ生成ステップで生成した各有向グラフを合成して共通の有向グラフを生成し、この共通の有向グラフと前記特定の一のターゲット単語の有向グラフとを比較して、これら有向グラフ間において異なるノードを抽出し、この工程を各ターゲット単語ごとに繰り返し行う処理を実行させることによって、３つ以上の有向グラフの比較を容易なものとすることができる。

【００１１】

また、上述した本発明の何れかの態様においては、コンピュータに、前記名詞抽出ステップにおいて、各文データに含まれるターゲット単語と文法的関係にある名詞を、当該名詞が前記ターゲット単語と共に文データに出現する頻度に関するデータと併せて抽出する処理を実行させ、前記有向グラフ生成ステップにおいて、生成する有向グラフの各ノードに前記頻度に関するデータによる重み付けする処理を実行させ、前記差異抽出ステップにおいて、前記有向グラフ生成ステップで生成した重み付けが施された有向グラフを利用して、各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する処理を実行させるようにすることが好適である。このような重み付き有向グラフを用いれば、得られる情報に優劣を付けて適宜必要なものだけを取り出すことを容易にすることができる。

【００１２】

特にこの場合、前記名詞抽出ステップにおいて、頻度に関するデータとして対応するターゲット単語に対して抽出された全名詞に占める当該名詞の頻度の割合を表す頻度比率を適用することが適している。こうした場合、コンピュータに、前記有向グラフ生成ステップにおいて、生成する有向グラフにおいて名詞に対応するノードに頻度を付与するとともに当該ノードの上位概念のノードにその下位のノードの頻度の合計値を付与し、全ノードに個々の頻度を正規化した頻度比率を付与することで、有向グラフにこの頻度比率に基づく重み付けする処理を実行させ、差異抽出ステップにおいて、前記有向グラフ生成ステップで生成した重み付けが施された比較対象となる２つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを前記異なるノードである差異ノードに組み入れ、当該差異ノードを抽出する処理を実行させるとよい。このようにすることで、上記の重み付けを、頻度比率によって信頼性の高いものとすることが可能となる。

【００１３】

さらにこの場合には、前記差異抽出ステップにおいて、前記コンピュータに、比較対象となる２つの有向グラフにおいて同一のノードの頻度比率の比を各々算出し、この比の値が所定値以上であればそのノードを暫定的に差異ノードとして有向グラフの差異部分に組み入れ、当該差異部分のうち各最上位ノードを各ターゲット単語について頻度比率が大きい方から順に所定数ずつ抽出し、その抽出したノードのうち共通するノードの割合を算出する工程を、前記頻度比率を逓減させながら繰り返すことで、各々の工程で得られた共通するノードの割合が一定値以上である場合、その共通するノードの割合を前回の工程で得

られた共通するノードの割合と比較して、その比較した値が一定値以上である場合に、当該工程で暫定的に差異ノードと決定したノードを差異ノードとして決定し、当該差異ノードを抽出する処理を実行させることも可能である。このようにすることで、得られる膨大な差異に関する情報から、その差異が急激に変化するノードを特定して、より適切な差異点を見つけ出し、同義語や類義語の用法が大きく異なるノードを適切に特定することができる。なお、頻度比率の逓減処理は、頻度比率が０になるまで適宜の数値ごとに行ってもよいし、機械学習を導入して入力された単語によって低減率を適宜調整しても良い。また、共通するノードの割合と前回の工程で得られた共通するノードの割合の比較処理は、両者の差やその絶対値によって行ってもよいし、比によって行ってもよい。

【００１４】

なお、上記の重み付け以降の処理においては、頻度に関するデータとして、頻度比率に代えて、頻度の値自体を適用してもよいのはいうまでもない。

【００１５】

さらに、頻度（頻度比率を適用する場合を含む）に基づく処理においては、コンピュータに、前記差異抽出ステップにおいて、抽出した異なるノードを、頻度に基づく重みの大きい方から順に所定数のノード（異なるノード全部 or 最上位の異なるノードに限定　可）をさらに抽出する処理を実行させ、前記差異出力ステップにおいて、前記所定数のノードを前記用法の違いに関する情報として出力する処理を実行させるようにしてもよい。ここで、頻度に基づく重みの大きい方から順に所定数のノードを抽出する際には、異なるノード全部を抽出しても良いし、異なるノードのうち最上位の異なるノードに限定して抽出することも妨げない。

【００１６】

以上の発明の各態様においては、コンピュータに、前記差異出力ステップにおいて、前記異なるノードのうち最上位のノードを前記用法の違いに関する情報として出力する処理を実行させる場合に、用法が異なると判断される最上位の概念を特定して、ユーザに理解しやすい有用な情報を得ることができる。

【００１７】

また同様に、コンピュータに、前記差異出力ステップにおいて、異なるノードのうち最上位のノードに加えて又はそれに代えて共通のノードの最下位のノードを前記用法の違いに関する情報として出力する処理を実行させることも可能であり、この場合は、用法が異なる部分と共通する部分の境界を適切にえることができる。なおこの場合、共通するノードのうち最下位のノードを上述のように頻度で重み付けして所定数だけ出力しても良いし、当該最下位のノードを全て出力しても構わない。

【００１８】

以上に述べた本発明においては、特に前記ターゲット単語入力ステップにおいて入力受付可能なターゲット単語の品詞を、同義語や類義語の使い分けが難しい形容詞又は動詞に制限している場合に、特に有用である。

【００１９】

また、本発明に係る単語用法差異情報取得装置は、上述したようなプログラムに従って作動するコンピュータにより構成され、入力された同一又は類似の意味を有する複数のターゲット単語について、各ターゲット単語の用法の違いに関する情報を抽出し出力させる単語用法差異情報取得装置である。このコンピュータは、例文データベースであるコーパス、及び語と語の上位下位概念の関係が記述されたデータベースであるシソーラスを検索可能に備え又は接続してある。そして、複数のターゲット単語の入力を受け付けるターゲット単語入力手段と、コーパスにアクセスして、ターゲット単語入力手段で受け付けた各ターゲット単語で検索して当該ターゲット単語をそれぞれ含む文データを抽出する文抽出手段と、文抽出手段で抽出した文データをそれぞれ構文解析し、各文データに含まれるターゲット単語と文法的関係にある名詞を抽出する名詞抽出手段と、シソーラスにアクセスして、前記名詞抽出手段で抽出した名詞で検索し、この名詞及びその上位概念を表すノードを抽出するとともに、それらノードと、これらノード同士の上位下位概念のつながりを

表すリンクとから構成される有向グラフを、対応するターゲット単語ごとに生成する有向グラフ生成手段と、有向グラフ生成手段で生成した各有向グラフを比較して、異なるターゲット単語の有向グラフ間において異なるノードを抽出する差異抽出手段と、差異抽出手段で抽出した有向グラフの差異を、前記ターゲット単語の用法の違いに関する情報として出力する差異出力手段と、を具備してなることを特徴とするものであり、上述したプログラムにより動作するコンピュータとして、上述と同様の作用効果が得られる。

【００２０】

このような装置は、一般的には、コンピュータが備えるハードディスクデバイス等の記憶装置に前記プログラムを格納しておき、必要に応じてメモリに当該プログラムを読み出してやＣＰＵ（中央演算装置）による処理を行い、各種出入力デバイスを作動させる態様が採用されるが、例えばネットワークサーバに前記プログラムを格納しておいて、そのサーバにアクセスした端末コンピュータがプログラムに従って作動するような態様を採用しても良い。

【００２１】

以下、詳述しないが、上記の各態様のプログラムによって作動する本発明の単語用法差異情報取得装置は、対応するプログラムと基本的には同様の作用効果が得られるものである。

【発明の効果】

【００２２】

本発明によれば、単に同義語や類義語がどのような単語と共起して文中で用いられるかという情報を得るのではなく、複数の同義語や類義語についてどの語がどのような意味を表す文において用いるのが適切か、という単語間での用法の違いを自動的に精度良く得ることが可能である。したがって、個々の単語に特化した情報ではなく、ターゲットとした単語が用いられる文について汎用的な情報が得られることとなり、その応用範囲も広いといえる。そして、本発明を単独で用いる場合には、同義語や類義語の用法の違いに関する情報が得られるので、語学学習や単語の用法チェックに利用することができ、また例えば本発明のプログラム又は装置を、外国語学習ソフトウェア、文書入力ソフトウェア、翻訳ソフトウェアやそれらを組み込んだ装置に用いる場合には、単に文法的な誤りを抽出、指摘する程度にとどまらず、用例に基づいた入力・編集支援として活用することが可能である。

【発明を実施するための最良の形態】

【００２３】

以下、本発明の一実施形態を、図面を参照して説明する。

【００２４】

この実施形態は、本発明の単語用法差異情報取得プログラム（以下、「本プログラム」と呼ぶ場合がある）に従って作動する単語用法差異情報取得装置Ａ（以下、「本装置Ａ」と呼ぶ場合がある）である。本装置Ａを実現するコンピュータは、一例としてごく一般的なパーソナルコンピュータで足りる。そこで本実施形態では、当該コンピュータとして、汎用パーソナルコンピュータを採用している。

【００２５】

斯かるパーソナルコンピュータのハードウェア構成はごく一般的なものであるため詳述しないが、通常は、中央演算装置（ＣＰＵ）等のプロセッサ、メインメモリ（ＲＡＭ）、ハードディスクドライブ（ＨＤＤ）等の補助記憶デバイス等を通信線（バス線等）で接続した態様で備えており、これらがシステムコントローラやＩ／Ｏコントローラ等に制御されて連携して動作する。さらにパーソナルコンピュータは、入力デバイスとしてキーボードやマウス等のポインティングデバイス、出力デバイスの一形態として情報を画像や映像として表示するディスプレイ及びグラフィックチップ等の表示制御デバイス、外部機器とのデータ授受を行うための通信デバイス等を備えているものとする。そして、通常はＨＤＤ等にオペレーションソフトウェア（ＯＳ）に加えて本プログラムがインストールされており、当該パーソナルコンピュータを本装置Ａとして機能させる場合には、適宜ＲＡＭに

読み出した本プログラムに基づくＣＰＵの処理により各部を動作させる。この処理工程で生成されるデータや外部から入力又は取得されたデータは、一時的にＲＡＭ等に蓄えられて以降の処理に利用される。

【００２６】

特に本実施形態では、前記パーソナルコンピュータに、例文データベースであるコーパスＤＢ１と、語と語の上位下位概念の関係が記述されたデータベースであるシソーラスＤＢ２を、前記通信デバイスを通じて接続しており（図１参照）、必要に応じてこれらにアクセスして情報を検索し読み出してくることができるようにしている。ここで、本実施形態では、コーパスＤＢ１の一例として、大規模コーパスとしてよく知られている「ＢＮＣ；（The British National Corpus ; URL： http://www.natcorp.ox.ac.uk/）」をインターネットを通じて利用し、シソーラスＤＢ２の一例として、語の意味的な上位下位関係をネットワーク構造のデータベースとして持つ大規模シソーラスとして知られる「ＷｏｒｄＮｅｔ２．０」を利用するものとするが、本発明にはこれら以外のコーパスやシソーラスを用いることができるのはいうまでもない。

【００２７】

さて、本装置Ａは、本プログラムに従って作動することで、図１に概略的な機能構成図を示すように、ターゲット単語入力手段１、文抽出手段２、名詞抽出手段３、有向グラフ生成手段４、差異抽出手段５、差異出力手段６の各機能を有することになる。以下、本装置Ａにおける各手段の機能と処理手順を、適宜具体例を示しながら説明する。特に具体例としては、説明の理解を容易にするため、入力されるターゲット単語として、英語の「ｋｅｅｎ」と「ｅａｇｅｒ」の２単語を採用する。両単語は、辞書においてほぼ同義の意味が記載されており、（「ｋｅｅｎ：ｖｅｒｙ ｉｎｔｅｒｅｓｔｅｄ，ｅａｇｅｒ ｏｒ ｗａｎｔｉｎｇ （ｔｏ ｄｏ） ｓｏｍｅｔｈｉｎｇ ｖｅｒｙ ｍｕｃｈ」，「ｅａｇｅｒ:ｗａｎｔｉｎｇ ｍｕｃｈ ｔｏ ｄｏ ｏｒ ｈａｖｅ ｅｓｐ． ｓｏｍｅ－ｔｈｉｎｇ ｉｎｔｅｒｅｓｔｉｎｇ ｏｒ ｅｎｊｏｙａｂｌｅ」，ＣＩＤＥ（The Cambridge International Dictionary of English）による）、共に「強く欲求する」という意味を表す形容詞であり、両者の文の意味に応じた正確な使い分けは難しく、特に英語を母国語としない者にとってはその困難性は顕著である。なお、以下の説明で言及する図面のうち、図２，図３，図５，図７は、本装置Ａによる情報処理手順を示す概略的なフローチャートである。

【００２８】

本装置Ａによる大まかな情報処理手順は、図２に示すようなものである。すなわち、ユーザによるキーボード等の入力デバイスの操作により、ターゲット単語入力手段１がターゲット単語（例えば上記「ｋｅｅｎ」と「ｅａｇｅｒ」）の入力を受け付ける（Ｓ１）。そして、文抽出手段２がコーパスＤＢ１からそのターゲット単語を含む文データを抽出する（Ｓ２）。次に、名詞抽出手段３が、各文データを構文解析し、ターゲット単語と文法的関係にある名詞を抽出する（Ｓ３）。さらに、有向グラフ生成手段が、各ターゲット単語ごとに、それらの名詞をキーワードとしてシソーラスＤＢ２を検索して有向グラフを生成する（Ｓ４）。そして、差異抽出手段５が、各有向グラフ（ターゲット単語の数に対応する。上記例のようにターゲット単語が２つの場合は２つの有向グラフ）を比較し、各ターゲット単語の用法の違いに関する情報を抽出し（Ｓ５）、最後に差異出力手段６が、抽出されたこの用法の違いに関する情報を、前記ディスプレイに表示するなどして出力する（Ｓ６）。以下に、各手段におけるデータ処理の内容等を詳細に説明する。

【００２９】

まず、ターゲット単語入力手段１は、入力を受け付け可能な単語の品詞を、形容詞又は動詞に制限しているものとする。これを実現するべくパーソナルコンピュータにターゲット単語入力手段１の機能を付与するには、例えば本プログラムの一部に、品詞解析プログラムを組み込んでおくことができる。すなわち、入力された単語で辞書データベース等を検索し、その単語の品詞が形容詞又は動詞であれば入力を受け付け、それら以外であれば入力の受け付けを拒否するようにすることができる。

【００３０】
また、文抽出手段２では、入力を受け付けたターゲット単語でコーパスＤＢ１を検索して、ターゲット単語を含む文データを全て抽出するものとする。これは、ターゲット単語を含む限定用法、叙述用法の両方の文についてできるだけ数多く且つ万遍なく収集することで、最終的に出力する「用法の違いに関する情報」の信頼性を向上させるためである。
【００３１】
名詞抽出手段３による処理工程Ｓ３では、具体的には例えば図３に示すような処理が実行される。なお、以下の工程は、各ターゲット単語ごとに且つ各文データについて実行される。すなわち、まず、コーパスＤＢ１から抽出された文データについて、まず文データを形態素解析する（Ｓ３１）。この形態素解析処理では、具体的には、ａ．文データを単語毎に区切り、ｂ．各単語から原形を抽出して、ｃ．各単語に品詞データを付与する。次に、形態素解析された文データを構文解析することにより、文データが有する文の統語構造を取得し（Ｓ３２）、ターゲット単語と所定の文法関係（修飾・被修飾関係、又は主述関係等の述語・項関係）にある名詞データを特定し、これを抽出する（Ｓ３３）。以上の処理を各ターゲット単語ごとに且つ各文データについて実行することで、抽出された名詞データには、各ターゲット単語ごとに、当該名詞が文データに出現した回数を示す頻度データが付与される。
【００３２】
ここで構文解析処理について詳述すると、単語に関する品詞データの他にも、シソーラス（上記シソーラスＤＢ２と同じでなくてもよい）を検索して得られる上位概念の情報、格フレーム辞書等の辞書を検索して得られる述語と項との関係を表すデータ（例えば動詞の場合、その目的格にはどのような意味の名詞を取り得るか、という情報）、必要に応じて設定された処理ルールや機械学習により設定された処理ルール、適宜のデータベースから抽出した文法データ（例えば、S->NP VPのような句構造文法に基づく文法規則）や共起データ等のデータを適宜利用して処理することで、名詞句や動詞句を特定したり、句と句の間に文法関係があるものを特定したりすることで、文の統語構造を取得する。斯かる構文解析処理には、一般に知られている適宜の構文解析プログラムを利用することができるが、本実施形態では、一例として、各語彙間の文法関係を解析できるパーサ（Parser）であるＲＡＳＰ（Robust Accurate Statistical Parsing）を利用するものとする。図４に、コーパスＤＢ１から抽出された「ｋｅｅｎ」と「ｅａｇｅｒ」をそれぞれ含む文例（限定的用法、叙述用法（主格補語）の各々１つずつ）と、これら文例のＲＡＳＰによる出力例を示す。ＲＡＳＰによる出力では、形態素解析結果（上段括弧内）と、その文の統語構造（下段太字部分）とが得られる。形態素解析結果では、単語（原形）が｜｜で区分けされており、各単語について、当該文中で出現した順番（数字）と品詞データが付与されている。文の統語構造は、文法関係が|ncmod|（限定用法、修飾・被修飾関係）や|ncsubj|-|xcomp|（叙述用法、主語・述語関係）として表され、それらの右側に当該関係にある単語が並べられる（当然ながら、一方はターゲット単語である）。
【００３３】
特に本実施形態では、ＲＡＳＰによる構文解析処理のルールとして、次のような設定を行うものとする。すなわち、
『|ncmod|については、ＲＡＳＰにより|ncmod|として抽出されるもののうち、形容詞と名詞の関係のみ、|ncmod|の関係として抽出した。ここで、形容詞は品詞がJJで始まる単語とし、名詞は品詞がNN、PN、VVG、PPのいずれかで始まる単語とする。また、品詞がNPで始まる単語は複合語の一部であることが多いため、品詞NPで始まる単語が、形容詞と|ncmod|の関係にあり、かつ、名詞とも|ncmod|の関係にある場合には、その品詞NPで始まる単語を介して形容詞と名詞の間にも|ncmod|の関係があるものとし、その形容詞と名詞を|ncmod|の関係にあるものとして抽出する。形容詞と名詞が前置詞で関係付けられている場合は、前置詞が「of」の場合のみ|ncmod|の関係として抽出する。|ncsubj|-|xcomp|については、『名詞と|ncsubj|の関係にある単語が、形容詞と|xcomp|の関係にある場合、その名詞と形容詞を|ncsubj|-|xcomp|の関係にあるものとして抽出する。ただし、このとき名

詞と形容詞を仲介している単語は、次の単語に限定する。
＜be, find, become, make, seem, appear, feel,
look, sound, smell, taste, remain, keep, stay, come, end-up, get, go, grow,
prove, turn, turn out, wind-up, burn, lie, loom, play, plead, rest, stand,
stand-up, blush, fall, fall-down, freeze, run, slam, spring, wax＞
ただし、形容詞と名詞が前置詞で関係付けられている場合は抽出しない。また、形容詞の右が「that」の場合(it is [形容詞] that ...の場合)も抽出しない。』。なお、以上の他の文法関係の例には、|dobj|-|xcomp|（叙述用法、動詞・目的格関係）を挙げることができる。その場合のＲＡＳＰにおける構文解析処理のルールとしては、次のようなものとすることができる。すなわち、『名詞と|dobj|の関係にある単語が、形容詞と|xcomp|の関係にある場合、その名詞と形容詞を|dobj|-|xcomp|の関係にあるものとして抽出する。品詞NPで始まる単語が、ある単語と|dobj|の関係にあり、かつ、名詞とも|ncmod|の関係にある場合には、その品詞NPで始まる単語を|ncmod|の関係にある名詞と置き換えたのち、上記の|dobj|-|xcomp|の関係を抽出する。ただし、このとき名詞と形容詞を仲介している単語は、次の単語に限定する。
＜hold, keep, leave, call, confess, profess,
pronounce, report, like, prefer, want, wish, believe, consider, deem, find,
hold, imagine, judge, presume, rate, reckon, suppose, think, drive, get, make,
prove, render, send, turn, certify, declare, proclaim＞
ただし、形容詞と名詞が前置詞で関係付けられている場合は抽出しない。また、形容詞の右がthatの場合(it is JJ that ...の場合)も抽出しない。』

【００３４】

次に、有向グラフ生成手段４において検索するシソーラスＤＢ２について簡単に説明する。本実施形態で利用するシソーラスＤＢ２である「ＷｏｒｄＮｅｔ２．０」は、ネットワーク構造をもつ有向グラフ状のデータ構造を有しており、この有向グラフの各ノードは、単語あるいは意味カテゴリ（概念）を表す。ノード間に意味的な上位下位の関係がある場合には、そのノード間にリンクが張られている。全てのノードは少なくとも一つの他のノードとの間にリンクが張られており、最上位のノードは「Ｒｏｏｔ」、最下位のノードは単語である（図６参照、最下位のノードには@が付してある）。

【００３５】

さて、有向グラフ生成手段４による処理工程Ｓ４では、具体的には例えば図５に示すような手順で処理が実行される。まず、シソーラスＤＢ２にアクセスして、上述した名詞抽出手段３で抽出されたターゲット単語と対をなす各名詞データで検索し、当該名詞がネットワーク（有向グラフ）上でどのノードの位置に現れるかを特定する（Ｓ４１）。そして、特定したノードとＲＯＯＴノード（最上位）とを結ぶ最短のパス（リンク及びノードを介した経路）を抽出する（Ｓ４２）。次に、ターゲット単語毎に、抽出された全名詞データについて前工程で抽出したパスを重ね合わせることで、有向グラフを生成する（Ｓ４３）。この工程で生成された有向グラフは、元のシソーラスＤＢ２の有向グラフの一部を切り取った形になっており、これがターゲット単語の意味拡張を表していると考えると、パスを重ね合わせたときの各ノードの重なりが多いほどそのノードの持つ意味カテゴリがコーパスＤＢ１に高い頻度で出現していると考えることができる。そこで、次の工程として、各ノードの重なりの数をＲＯＯＴノードでの重なりの数で割ることによって、各ノードの頻度比率（出現比率）を算出する（正規化）（Ｓ４４）。ＲＯＯＴノードの頻度比率は１であり、各単語のノードの頻度比率は、コーパスＤＢ１での当該単語の出現比率を同じターゲット単語と共起して抽出された全名詞の出現頻度で割り算した値となる。以上の処理を経て、各ノードに頻度比率が付与された有向グラフ（以下、「重み付き有向グラフ」と呼ぶ）が生成される（Ｓ４５）。すなわち、この重み付き有向グラフは、ターゲット単語の意味拡張をコーパスＤＢ１での共起した名詞の頻度で重み付けされたものである。図６に、ターゲット単語「ｋｅｅｎ」及び「ｅａｇｅｒ」について生成された重み付き有向グラフの例を示す。

【００３６】
次に、差異抽出手段５による処理工程Ｓ５では、具体的には例えば図７に示すような手順で処理が実行される。まず、工程Ｓ４５で生成された対比されるべき２つの重み付き有向グラフから、同じノードを抽出し、両重み付き有向グラフを重ね合わせる（Ｓ５１）。この処理により、両重み付き有向グラフ間で共通するノードと異なるノードが大まかに分けられる。図８に、この重ね合わせ処理のイメージ図を示す。同図中、○印がノードを示し、２つのノード間を結ぶ直線がリンクを示している。ここで、共通するノードのうち有向グラフ上で意味的に最下位にあるノード（ボトムノード）と、差異のノードのうち有向グラフ上で意味的に最上位にあるノード（トップノード）の境界が、２つのターゲット単語の意味拡張の共通部分と差異部分の境目であり、二つの語彙の意味的な違いを顕著に表している部分であるといえる。２つのターゲット単語の用法の違いを粗く分けるときには、前記トップノード又はそれ以下を、差異ノードに決定して差異抽出手段５による処理工程Ｓ５を終了することもできるが、共通するノードであっても頻度比率の違いが著しい場合があり、特にその違いが顕著な場合は差異ノードとして扱う方が好ましいといえる場合がある。

【００３７】
そこで、次の工程として、それら同じノードについて頻度比率の比Ｃを計算する（Ｓ５２）。これら同じノードは通常、それぞれの重み付き有向グラフにおいて相互に異なる頻度比率を有している。ただし、この比Ｃの計算では、その値が１以下となるように、頻度比率の値が大きい方を分母とする。次に、算出された比Ｃを所定の閾値Ｃｘと比較することで、その比Ｃの算出に用いられたノードを、暫定的な差異ノードと共通ノードの何れかに分ける。ここで閾値Ｃｘは、初期値において１以下の適宜の正数を採用することができるが、ボトムノード近辺における頻度比率の比の値を採用することが好ましい。具体的な処理としては、比Ｃが閾値Ｃｘ以上の場合は、それらのノードを暫定的な差異ノード、それ以外の場合（比Ｃが閾値Ｃｘ未満）は当該ノードを暫定的な共通ノードとする（Ｓ５３）。この処理で暫定的な差異ノードとしたノードについては、工程Ｓ５１で元々異なるノードであった部分に組み入れて、それら差異ノードからなる有向グラフからトップノードを全て抽出する（Ｓ５４）。そして、各ターゲット単語について、それらトップノードの集合を頻度比率の高いものから順にソートし（Ｓ５５）、上位所定数（Ｎ個）のトップノードを抽出する（Ｓ５６）。この抽出数Ｎは任意の値であるが、通常は１０から１００の間の適宜の値を採用すればよく、好ましくは２０前後の値で十分であり、本実施形態ではＮ＝２０とする。次に、これらＮ個のノードのうち、両ターゲット単語の有向グラフにおいて共通するノードの値P_c^nが所定値Ｃｐ以上か否かを判定する（Ｓ５７）。このＣｐの値には、適宜の値を採用することができるが、好ましくは０．３から０．７の間の値とするのがよく、さらに好ましくは０．５前後であり、本実施形態ではＣｐ＝０．５とする。ここで、共通するノードの値P_c^nが所定値Ｃｐ未満であれば、一旦、比Ｃの値が０か否かを判断し（Ｓ５７ａ）、Ｃ＝０でなければ（Ｓ５７；Ｎｏ）、閾値Ｃｘの値を一定値下げて、工程Ｓ５３に戻る（Ｓ５７ｂ）。ここで、工程Ｓ５７ｂにおける閾値Ｃｘの変化値は、適宜の値（例えば０．１）を採用することができる。他方、Ｃ＝０であれば（Ｓ５７；Ｙｅｓ）、共通するノードの値P_c^nが初めてＣｐ以上となる比Ｃの値を採用して（Ｓ５７ｃ）、本工程Ｓ５の処理を終了する。また、共通するノードの値P_c^nが所定値Ｃｐ以上の場合は（Ｓ５７；Ｙｅｓ）、共通するノードの値P_c^nと前回の工程Ｓ５３から工程Ｓ５７の処理サイクルにおいて得られた共通するノードの値P_c^{n-1}とを比較し、その比較した値が一定値Ｑ以上か否か、すなわちその比較した値が急激に変化したか否かを判定する（Ｓ５８）（但し、１回目の処理サイクルは除く）。この比較処理は、両方の値の差（差の絶対値；$|P_c^n - P_c^{n-1}|$）を採用してもよいし、両方の値の比（P_c^n / P_c^{n-1}）を採用してもよい。本実施形態では差の絶対値を採用しており、判断基準となる値Ｑを４／２０（０．２）としている。但しこの値Ｑは任意であり、例えば機械学習により適宜最適化を図ることができる。そして、この比較処理の結果の値が一定値Ｑ未満であれば（Ｓ５８；Ｎｏ）、工程Ｓ５７に戻るが、一定値Ｑ以上であれば（Ｓ５８；Ｙｅｓ）、共通するノー

ドであっても頻度比率の違いが著しいと考えることができるので、工程Ｓ５３で暫定的な差異ノードとしたノードを差異ノードに決定し、他方、工程Ｓ５３暫定的な共通ノードとしたノードを共通ノードに決定する（Ｓ５９）。以上を以て、差異抽出手段５による処理工程Ｓ５を終了する。

【００３８】

以上のようにして、対比されるべき２つの重み付き有向グラフで共通ノードと差異ノードに分けられることになる。なお、次の差異出力手段６による処理に備えて、比較されるターゲット単語の用法の違いに関する情報として、工程Ｓ５９で決定した差異ノードのトップノードのみを全て又は一部（頻度比率が上位の一部又はランダムに一部）抽出しておいたり、このトップノードを含む全差異ノードを抽出しておいたり、トップノードを含む上位所定数（又はランダムに所定数）の差異ノードを抽出しておいたり、トップノードとそれから有向グラフ上を辿ることができる最下位の差異ノード（名詞を示すノード）を抽出しておくことができる。また、差異ノードのトップノードを決定することで、その直上位のノードを共通ノードの最下位ノード（ボトムノード）として抽出することができるので、これを用法の違いに関する情報（当該情報の裏返しの意味として、比較されるターゲット単語の意味の共通する最下限概念）又はその一部として利用することができる。本実施形態では、ターゲット単語の用法の違いに関する情報として、頻度比率が高い方から所定数の差異ノードのトップノードと、各トップノードから辿れる最下位の単語ノードの所定数、またそれらに加えて共通ノードのボトムノードの頻度比率が低い方から所定数とその直上位のノードを抽出しておき、差異出力手段６により、これらのノードを整理して出力するものとする。

【００３９】

図９にターゲット単語が「ｋｅｅｎ」と「ｅａｇｅｒ」の場合における画面又は印刷出力例を示す。同図上欄は「ｋｅｅｎ」と「ｅａｇｅｒ」の共通ノードを示し、下欄は差異ノードを示す。上欄中、右欄は共通ノードのボトムノード（単語）であり、左欄はその直上位概念のノードである。下欄中、左欄は「ｋｅｅｎ」と「ｅａｇｅｒ」の別を示し、それらのそれぞれについて中欄は差異ノードのトップノード、右欄は各トップノードの下位に当たる単語ノードである。このような出力例からユーザは、「ｋｅｅｎ」については、競争を前提とした活動（sportsman, player）や専門的な職業に従事する主体（gardener）や熱狂的で欲求の程度が極端に強い主体（supporter, fan）等が出現するような意味の文で用いられる傾向が強く、本来の意味である「痛み、刃の鋭さ、獰猛さ」といったものに含まれる「はなはだしさやネガティブ性」を温存しているのに対して、「ｅａｇｅｒ」については、好んである行動をしたがるという自然でポジティブな主体（audience, buyer, volunteer）等が出現するような意味の文で用いられる傾向が強く、刺激の甚だしさよりも味覚等の感覚が持つ「自然な欲求」を温存している、という２つの単語の用法の違いが分かる。

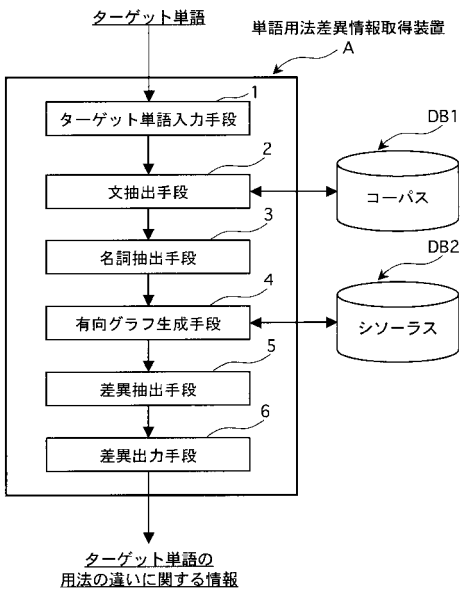
【００４０】

なお、本発明は上述した実施形態に限定されるものではない。例えば、入力されるターゲット単語は３つ以上とすることが可能である。その場合、例えば各ターゲット単語について有向グラフを生成しておき、１つ以外の有向グラフを全て重ね合わせて有向グラフを生成し、それと残りの有向グラフを比較することで、あるターゲット単語と他のターゲット単語の用法の違いに関する情報を得ることができる。また、全てのターゲット単語の有向グラフを全て重ね合わせることだけでも、各ターゲット単語の用法の違いの概略的な情報を得ることができる。その他、有向グラフに頻度による重み付けを行わない態様や、頻度比率の変わりに頻度の値そのものを利用する態様などの種々の変更が可能である。また、本発明のプログラムや装置を、他の文書入力プログラム、言語学習プログラム、翻訳プログラム等の各種プログラムやそれらに従って作動する装置の一部に応用することも可能である。さらに、各部の具体的構成や処理工程についても上記実施形態に限られるものではなく、本発明の趣旨を逸脱しない範囲で種々変形が可能である。

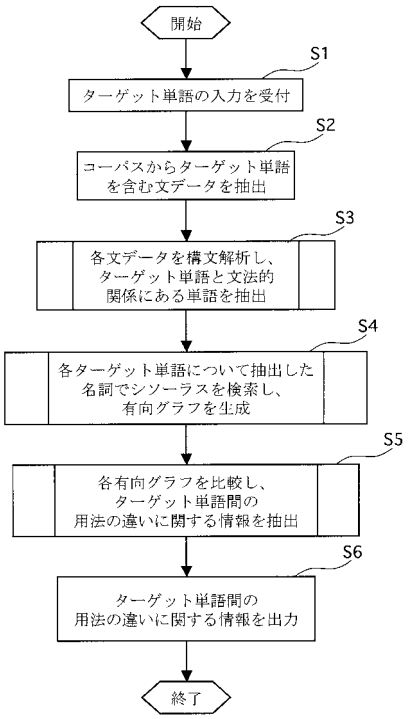
【図面の簡単な説明】

【００４１】
【図１】本発明の一実施形態に係る単語用法差異情報取得装置の概略的な機能構成図
【図２】同装置の単語用法差異情報取得プログラムに基づく概略的な全処理工程図
【図３】図２の処理工程の一部（Ｓ３）を詳細に示す処理工程図
【図４】同装置に入力されるターゲット単語に基づいてコーパスから抽出される文例とその構文解析例を示す図
【図５】図２の処理工程の一部（Ｓ４）を詳細に示す処理工程図
【図６】図２の処理工程Ｓ４により生成される重み付き有向グラフ例を示す図
【図７】図２の処理工程の一部（Ｓ５）を詳細に示す処理工程図
【図８】図７の処理工程Ｓ５においてなされる有向グラフの重ね合わせ処理を示すイメージ図
【図９】図２の処理工程Ｓ６による出力例を示す図
【符号の説明】
【００４２】
Ａ…単語用法差異情報取得装置
ＤＢ１…コーパス
ＤＢ２…シソーラス
１…ターゲット単語入力手段
２…文抽出手段
３…名詞抽出手段
４…有向グラフ生成手段
５…差異抽出手段
６…差異出力手段

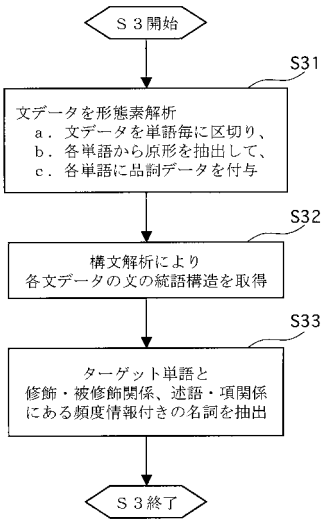
【図１】



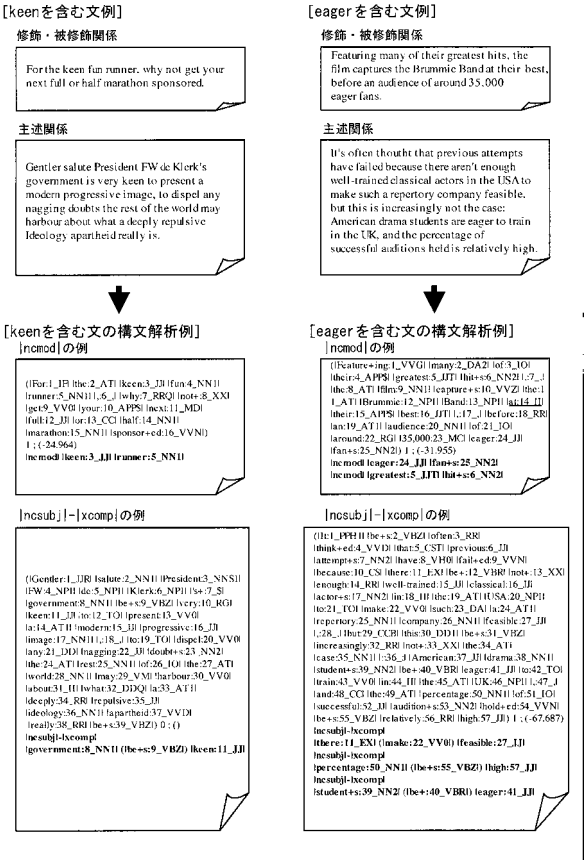
【図２】



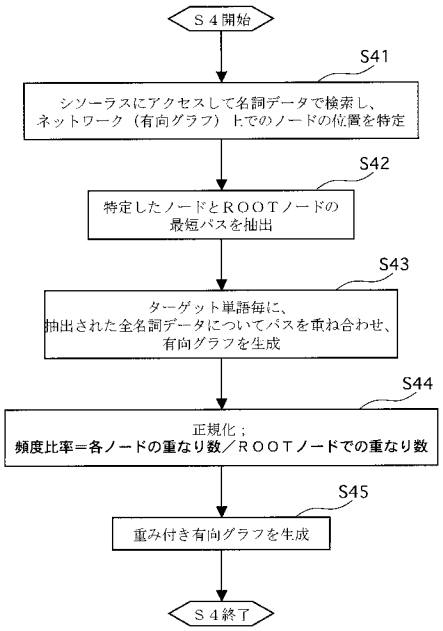
【図３】



【図４】

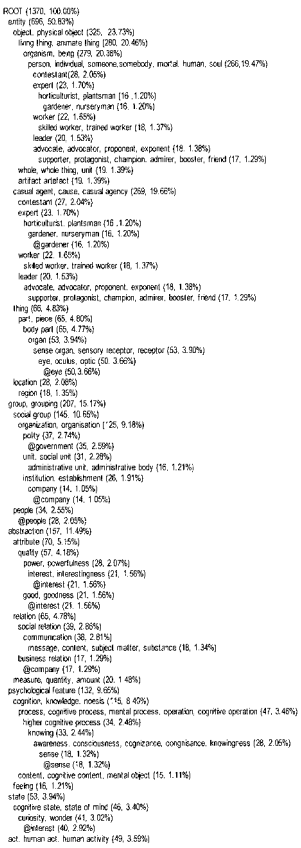


【図５】

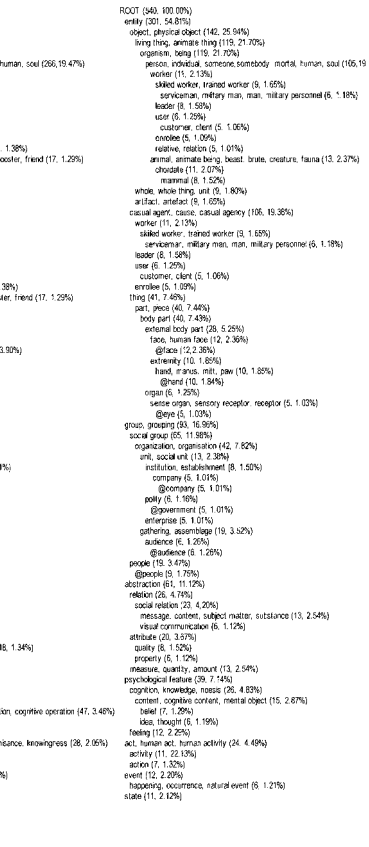


【図６】

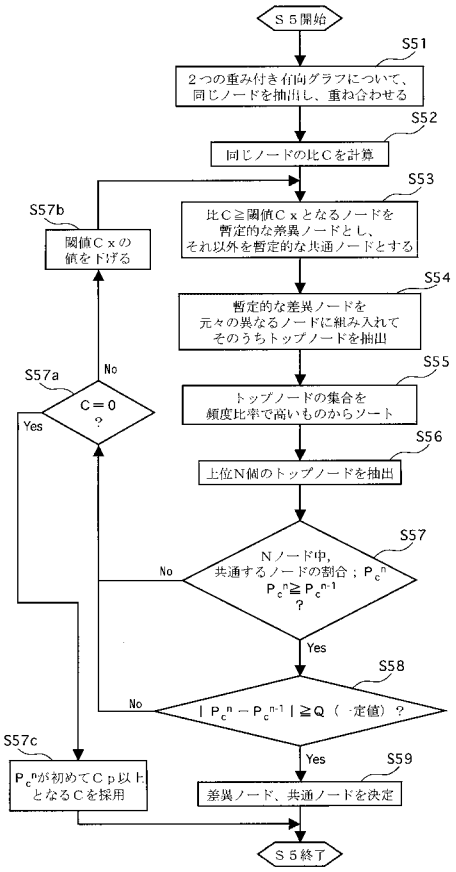
＜ｋｅｅｎの重み付き有向グラフ例＞



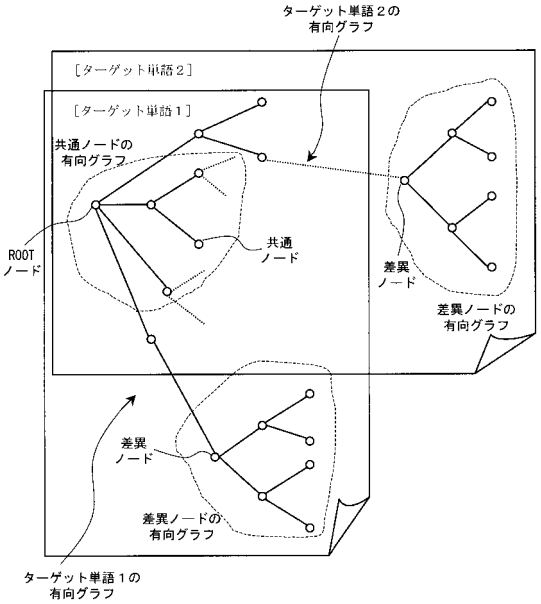
＜ｅａｇｅｒの重み付き有向グラフ例＞



【図７】



【図８】



【図９】

［keenとeagerの差異ノードと共通ノードの出力例］

Common nodes and nouns included in each node		
PEOPLE(keen: 2.6%, eager 3.5%))		people(keen: 28, eager: 9)
INSTITUTION(keen: 1.9%, eager 1.5%)		company(keen: 14, eager: 5)
Different nodes and nouns included in each node		
KEEN	CONTESTANT (4.1%)	sportsman(21), player(17), golfer(13), rival(4) etc.
	EXPERT (3.4%)	gardener(33), observer (6) etc.
	ADVOCATE (2.8%)	supporter(27), fan(8), enthusiast(3), admirer(2) etc.
	PEER, COMPEER (1.8%)	member(16), colleague(2), participant(2) etc.
	REGION (1.4%)	country(5), germany(2), russia(2) etc.
	GATHERING (3.5%),	audience(8), crowd(2) etc.
EAGER	USER (2.5%)	buyer(6), customer(5), consumer(2) etc.
	ANIMAL, BEAST (2.4%)	dog(4), horse(2), elephant(1) etc.
	SERVICEMAN (2.4%)	volunteer(4), man(3) etc.
	ENROLEE (2.4%)	student(10), pupil(3), undergraduate(1) etc.

フロントページの続き

審査官　成瀬　博之

(56)参考文献　特開２００４－２７２６７８（ＪＰ，Ａ）
清水正勝　他２名，新聞コーパスの調査に基づくフランス語人称代名詞の使い分け基準について（“ｏｎ”と“ｌ’ｏｎ”を例に），情報処理学会研究報告，日本，社団法人情報処理学会，２００３年１０月２４日，Vol.2003,No.107(2003-CH-60)，9-16頁

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２０－１７／３０