(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2013-20439
(P2013-20439A)
(43)公開日　平成25年1月31日(2013.1.31)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	320D	5B091
G06F 17/27	(2006.01)	G06F	17/30	350C	
		G06F	17/27	Z	

審査請求 未請求　請求項の数 20　ＯＬ　（全 26 頁）

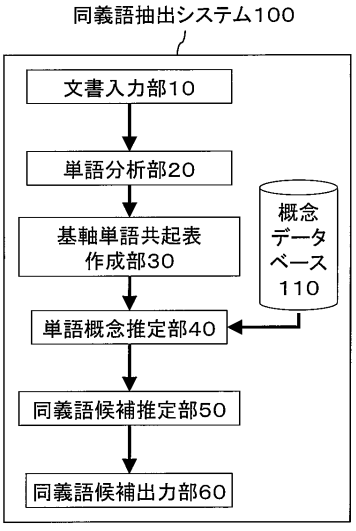
(21)出願番号　特願2011-153084 (P2011-153084)
(22)出願日　平成23年7月11日 (2011.7.11)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(71)出願人　504139662
国立大学法人名古屋大学
愛知県名古屋市千種区不老町１番
(74)代理人　100077838
弁理士　池田　憲保
(74)代理人　100082924
弁理士　福田　修一
(74)代理人　100129023
弁理士　佐々木　敬
(72)発明者　平尾　英司
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】同義語抽出システム、方法およびプログラム

(57)【要約】　（修正有）
【課題】情報システム構築に関する提案書や仕様書等、所定の案件に関する文書で、意義は同じで語形が異なる同義語のある文章の曖昧さを改善する。
【解決手段】文章に使用されている各単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行う単語分析部と、任意の単語を基軸単語として選択し、基軸単語と共起関係にある共起語とその共起数に基づく基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する基軸単語共起表作成部と、単語の一般概念情報を概念データベースに問い合わせ、各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する単語概念推定部と、各基軸単語概念ベクトル間の類似性を判定し、類似性が高い基軸単語の組合せを同義語候補として抽出する同義語候補推定部と、同義語候補を出力する同義語候補出力部とを備える。
【選択図】図１



【特許請求の範囲】
【請求項１】
対象とする文書もしくは文書群の入力を受け付ける文書入力部と、文書もしくは文書群を構成する文章に使用されている各単語の抽出および単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行う単語分析部と、任意の単語を基軸単語として選択し、単語毎の単語情報に基づき、任意の範囲および条件で基軸単語と共起関係にある共起語とその共起数に基づく基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する基軸単語共起表作成部と、単語の概念分類、同義語、類義語、用法といった単語の一般概念を体系付けた一般概念情報を収集して蓄積し、特定の単語に関する問い合わせに対し、単語の意味や用法に関連する一般概念情報を検索し応答する概念データベースと、基軸単語共起表の各共起語の一般概念情報を概念データベースに問い合わせ、任意の範囲内で基軸単語共起表における各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する単語概念推定部と、各基軸単語に対応する基軸単語概念ベクトル間の類似性を所定の判定基準によって判定し、基軸単語共起ベクトルの意味的な類似性が高い基軸単語の組合せを同義語候補として抽出する同義語候補推定部と、同義語候補を出力する同義語候補出力部と、を備えたことを特徴とする同義語抽出システム。
【請求項２】
対象とする文書もしくは文書群の入力を受け付ける文書入力部と、単語の概念分類、同義語、類義語、用法といった単語の一般概念を体系付けた一般概念情報を収集して蓄積し、特定の単語に関する問い合わせに対し、単語の意味や用法に関連する一般概念情報を検索し応答する概念データベースと、文書もしくは文書群を構成する文章に使用されている各単語の抽出および単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行い、概念データベースに抽出された各単語で一般概念情報の登録が無く、かつ文字数が２文字以上の単語を複合語として抽出し、複合語を構成するあらゆる部分文字列について、一般概念情報の登録がある部分文字列を複合語の有意構成語として抽出し、登録が無い部分文字列を不明構成語として抽出し、さらに有意構成語の一般概念情報を取得する単語分析部と、任意の単語を基軸単語として選択し、単語毎の単語情報に基づき、任意の範囲および条件で基軸単語と共起関係にある共起語とその共起数に基づく基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する基軸単語共起表作成部と、各単語の単語情報および複合語に基づき、任意の範囲および条件で複合語と共起する単語を複合語共起語として、複合語毎に複合語共起語の種類と共起数をまとめた複合語共起表を作成し、複合共起表と構成語に基づき、前記複合語共起表から同じ構成語を含む部分一致複合語の複合語共起語からなる複合語共起ベクトルを抽出し、構成語別に部分一致複合語共起表を作成し、部分一致複合語共起表の複合語共起ベクトルから得られる共起ベクトル空間における各部分一致複合語間の集約度を構成語支配度として算出する構成語支配度算出部と、複合語毎の各構成語の一般概念情報に基づき複合語が関連する概念をまとめた複合語概念構成表を作成し、各構成語支配度で複合語毎の各概念の重み付け係数を算出し、複合語概念構成表の対応する箇所に重み付け係数を登録することで、複合語概念配分表を作成し、重み付けされた複数の概念の合成概念として未知の複合語の概念を推定する複合語概念配分推定部と、基軸単語共起表の基軸単語共起ベクトルの各複合語共起語の内で複合語になっている共起語について、構成語毎の概念に置き換えることで、合成概念に変換し、基軸単語共起表の各共起語の一般概念情報を概念データベースに問い合わせ、任意の範囲内で基軸単語共起表における各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する単語概念推定部と、各基軸単語に対応する基軸単語概念ベクトル間の類似性を所定の判定基準によって判定し、基軸単語共起ベクトルの意味的な類似性が高い基軸単語の組合せを同義語候補として抽出する同義語候補推定部と、同義語候補を出力する同義語候補出力部と、を備えたことを特徴とする同義語抽出システム。
【請求項３】

前記概念データベースは、単語を分類体系付けて記憶しており、単語間の同義関係、類義関係、上位/下位関係、部分/全体関係について、一般概念情報として取得できるシソーラスであり、前記同義語候補推定部の前記判定基準が、シソーラスに基づく各階層での各基軸単語間の非類似度を算出し、より詳細な分類での非類似度ほど重視するように重み付けした非類似度指標が任意の閾値より小さい概念ベクトルを持つ基軸単語の組合せとする、ことを特徴とする請求項１又は２に記載の同義語抽出システム。

【請求項４】

前記基軸単語共起表作成部が、品詞が動詞であれば係り受け関係が有る単語、名詞であれば同一段落内の単語のように品詞毎に共起と見なす範囲をおよび条件を変えて共起語の抽出および共起数の算出を行う、ことを特徴とする請求項１乃至３のいずれか１項に記載の同義語抽出システム。

【請求項５】

前記単語分析部が、複合語を構成する部分文字列の内、概念データベースに概念情報の登録がある部分文字列の組合せパターンが複数考えられる場合は、不明構成語の文字数が最も少なくなる組合せパターンを判定し、その組合せパターンでの有意構成語、不明構成語を抽出することを特徴とする請求項２乃至４のいずれか１項に記載の同義語抽出システム。

【請求項６】

構成語支配度算出部が、品詞が動詞であれば係り受け関係が有る単語、名詞であれば同一段落内の単語のように品詞毎に共起と見なす範囲をおよび条件を変えて複合語共起語の抽出および複合語共起数の算出を行う、ことを特徴とする請求項２乃至５のいずれか１項に記載の同義語抽出システム。

【請求項７】

前記構成語支配度算出部における部分一致複合語間の集約度が、各部分一致複合語に対応するベクトル間の散らばりの小ささを表す指標として、ばらつきを示す指標と単調減少の関係にある関数で算出される、ことを特徴とする請求項２乃至６のいずれか１項に記載の同義語抽出システム。

【請求項８】

前記構成語支配度算出部における部分一致複合語間の集約度が、共起語の品詞によって重み付けを行ったベクトル空間に基づいて算出される、ことを特徴とする請求項２乃至７のいずれか１項に記載の同義語抽出システム。

【請求項９】

前記複合語概念配分推定部が、複合語の各構成語の構成語支配度を複合語毎の構成語支配度の総和で除すことで、正規化した重み付け係数を算出する、ことを特徴とする請求項２乃至８のいずれか１項に記載の同義語抽出システム。

【請求項１０】

対象とする文書もしくは文書群の入力を受け付ける文書受付工程と、文書もしくは文書群を構成する文章に使用されている各単語の抽出および単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行う単語情報抽出工程と、任意の単語を基軸単語として選択し、単語毎の単語情報に基づき、任意の範囲および条件で基軸単語と共起関係にある共起語とその共起数に基づく基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する基軸単語共起表作成工程と、単語の概念分類、同義語、類義語、用法といった単語の一般概念を体系付けた一般概念情報を収集して蓄積し、特定の単語に関する問い合わせに対し、単語の意味や用法に関連する一般概念情報を検索し応答する概念データベースに基軸単語共起表の各共起語の一般概念情報を問い合わせ、任意の範囲内で基軸単語共起表における各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する単語概念推定工程と、各基軸単語に対応する基軸単語概念ベクトル間の類似性を所定の判定基準によって判定し、基軸単語共起ベクトルの意味的な類似性が高い基軸単語の組合せを同義語候補として抽出する同義語候補推定工程と、同義語候補を出力する同義語候補出

力工程と、を含むことを特徴とする要求文書分析方法。
【請求項１１】
対象とする文書もしくは文書群の入力を受け付ける文書受付工程と、文書もしくは文書群を構成する文章に使用されている各単語の抽出および単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行い、単語の概念分類、同義語、類義語、用法といった単語の一般概念を体系付けた一般概念情報を収集して蓄積し、特定の単語に関する問い合わせに対し、単語の意味や用法に関連する一般概念情報を検索し応答する概念データベースに抽出された各単語で一般概念情報の登録が無く、かつ文字数が２文字以上の単語を複合語として抽出し、複合語を構成するあらゆる部分文字列について、一般概念情報の登録がある部分文字列を複合語の有意構成語として抽出し、登録が無い部分文字列を不明構成語として抽出し、さらに有意構成語の一般概念情報を取得する単語分析工程と、任意の単語を基軸単語として選択し、単語毎の単語情報に基づき、任意の範囲および条件で基軸単語と共起関係にある共起語とその共起数に基づく基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する基軸単語共起表作成工程と、各単語の単語情報および複合語に基づき、任意の範囲および条件で複合語と共起する単語を複合語共起語として、複合語毎に複合語共起語の種類と共起数をまとめた複合語共起表を作成し、複合共起表と構成語に基づき、前記複合語共起表から同じ構成語を含む部分一致複合語の複合語共起語からなる複合語共起ベクトルを抽出し、構成語別に部分一致複合語共起表を作成し、部分一致複合語共起表の複合語共起ベクトルから得られる共起ベクトル空間における各部分一致複合語間の集約度を構成語支配度として算出する構成語支配度算出工程と、複合語毎の各構成語の一般概念情報に基づき複合語が関連する概念をまとめた複合語概念構成表を作成し、各構成語支配度で複合語毎の各概念の重み付け係数を算出し、複合語概念構成表の対応する箇所に重み付け係数を登録することで、複合語概念配分表を作成し、重み付けされた複数の概念の合成概念として未知の複合語の概念を推定する複合語概念配分推定工程と、基軸単語共起表の基軸単語共起ベクトルの各複合語共起語の内で複合語になっている共起語について、構成語毎の概念に置き換えることで、合成概念に変換し、基軸単語共起表の各共起語の一般概念情報を概念データベースに問い合わせ、任意の範囲内で基軸単語共起表における各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する単語概念推定工程と、各基軸単語に対応する基軸単語概念ベクトル間の類似性を所定の判定基準によって判定し、基軸単語共起ベクトルの意味的な類似性が高い基軸単語の組合せを同義語候補として抽出する同義語候補推定部と、同義語候補を出力する同義語候補出力工程と、を含むことを特徴とする要求文書分析方法。
【請求項１２】
前記概念データベースは、単語を分類体系付けて記憶しており、単語間の同義関係、類義関係、上位/下位関係、部分/全体関係について、一般概念情報として取得できるシソーラスであり、前記同義語候補推定部の前記判定基準が、シソーラスに基づく各階層での各基軸単語間の非類似度を算出し、より詳細な分類での非類似度ほど重視するように重み付けした非類似度指標が任意の閾値より小さい概念ベクトルを持つ基軸単語の組合せとする、ことを特徴とする請求項１０又は１１に記載の要求文書分析方法。
【請求項１３】
前記基軸単語共起表作成工程が、品詞が動詞であれば係り受け関係が有る単語、名詞であれば同一段落内の単語のように品詞毎に共起と見なす範囲をおよび条件を変えて共起語の抽出および共起数の算出を行う、ことを特徴とする請求項１０乃至１２のいずれか１項に記載の要求文書分析方法。
【請求項１４】
前記単語分析工程が、複合語を構成する部分文字列の内、概念データベースに概念情報の登録がある部分文字列の組合せパターンが複数考えられる場合は、不明構成語の文字数が最も少なくなる組合せパターンを判定し、その組合せパターンでの有意構成語、不明構成語を抽出することを特徴とする請求項１１乃至１３のいずれか１項に記載の要求文書分

析方法。
【請求項１５】
構成語支配度算出工程が、品詞が動詞であれば係り受け関係が有る単語、名詞であれば同一段落内の単語のように品詞毎に共起と見なす範囲をおよび条件を変えて複合語共起語の抽出および複合語共起数の算出を行う、ことを特徴とする請求項１１乃至１４のいずれか１項に記載の要求文書分析方法。
【請求項１６】
前記構成語支配度算出工程における部分一致複合語間の集約度が、各部分一致複合語に対応するベクトル間の散らばりの小ささを表す指標として、ばらつきを示す指標と単調減少の関係にある関数で算出される、ことを特徴とする請求項１１乃至１５のいずれか１項に記載の要求文書分析方法。
【請求項１７】
前記構成語支配度算出工程における部分一致複合語間の集約度が、共起語の品詞によって重み付けを行ったベクトル空間に基づいて算出される、ことを特徴とする請求項１１乃至１６のいずれか１項に記載の要求文書分析方法。
【請求項１８】
前記複合語概念推定工程が、複合語の各構成語の構成語支配度を複合語毎の構成語支配度の総和で除すことで、正規化した重み付け係数を算出する、ことを特徴とする請求項１１乃至１７のいずれか１項に記載の要求文書分析方法。
【請求項１９】
請求項１０乃至１８のいずれか一項に記載の要求文書分析方法をコンピュータによって実現するためのプログラム。
【請求項２０】
請求項１９に記載のプログラムを記録したコンピュータ読み取り可能な記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、同義語抽出システム、方法およびプログラムに関し、特に、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書内から、意義は同じで語形が異なっている同義語を抽出する同義語抽出システム、方法およびプログラムに関する。
【背景技術】
【０００２】
近年、情報処理装置を用いて、自然言語で書かれた文書を分析して、その文書の意味や意義を自動抽出するシステムが開発されている。そのなかで、文章中の同義語の取り扱いが問題になることがある。
同義語抽出システムに関する技術の一例が、特許文献１に「類似表現抽出装置」として記載されている。この特許文献１に開示された類似表現抽出装置は、データ記憶部、単語グループ記憶部、シソーラス記憶部、文書入力部、単語グループ作成処理部、評価調整処理部から構成されている。このような構成を有する類似表現抽出装置は、次のように動作する。
【０００３】
すなわち、文書入力部は、入力インタフェースとして電子文書の入力を受け付ける。単語グループ作成処理部は、前記文書入力部で入力された電子文書内の文を形態素解析し、得られた形態素解析結果を前記データ記憶部に書き込み、前記データ記憶部内の形態素解析結果を構文解析し、構文解析結果として得られた文脈情報を前記データ記憶部に書き込み、前記データ記憶部内の文脈情報から２文節の係り受けの組を含む共起表現を抽出し、この共起表現を前記データ記憶部に書き込み、前記データ記憶部内の共起表現のうち、所定の品詞の組合せの２文節からなる共起表現に基づいて、この共起表現における一方の単語毎に、他方の単語との共起頻度と、前記電子文書内の単語との共起頻度とからなる単語

属性値を算出し、前記単語属性値を前記一方の単語に関連付けることにより、当該単語毎に単語ベクトルを作成し、この単語ベクトルを前記データ記憶部に書き込み、前記データ記憶部内の各単語ベクトル間の単語類似度を計算し、得られた単語類似度を、当該計算に用いた各単語ベクトルに関連付けて前記データ記憶部に書き込み、前記データ記憶部内の単語類似度に基づいて、教師なし学習手法により、前記単語類似度の算出に用いた各単語ベクトルが示す各単語を同一の単語グループに分類し、当該分類された各単語を含む単語グループを前記単語グループ記憶部に書き込む。さらに、評価調整処理部は、前記シソーラス記憶部内のシソーラス情報に含まれる表現のうち、前記入力された電子文書に含まれる表現を学習データとして生成し、前記生成された学習データに基づいて当該学習データ間の類似度を計算し、この類似度により学習データを含む学習データグループを作成し、前記学習データグループの個数に対し、前記単語グループ記憶部内の単語グループの個数を一致させるように、当該単語グループを統合し、前記統合された単語グループ毎に、前記学習データグループ内の学習データを含む度合を示す大域評価値を計算し、この大域評価値を前記データ記憶部に書き込み、前記統合された単語グループ毎に、単語グループ内の各単語を示す各単語ベクトルに関連する単語類似度の分散を計算し、得られた分散を局所評価値として前記データ記憶部に書き込み、前記大域評価値及び前記局所評価値に基づいて、これら両評価値の和を上限値にするように、前記データ記憶部内の単語グループの境界を調整し、前記調整された単語グループ内の各単語を前記類似表現として抽出し、当該抽出した類似表現の各単語を出力する。なお、データ記憶部は、単語グループ作成処理部、評価調整処理部から読出／書込可能な記憶装置であり、処理前後のデータ等が適宜記憶される。単語グループ記憶部は、単語グループ作成処理部、評価調整処理部から読出／書込可能な記憶装置であり、類似表現の各単語からなる単語グループが記憶される。シソーラス記憶部は、評価調整処理部から読出／書込可能な記憶装置であり、予めシソーラス情報が記憶されている。このような構成により、文書中の単語について、共起の頻度による単語類似度に基づく単語グループと、シソーラスでの距離などに基づく学習データグループを作成し、学習データグループの個数と構成単語に単語グループの個数および構成単語を一致させるように単語グループの境界を調整することで類似表現の各単語を抽出している。

【０００４】

さらに、同義語抽出システムに関する技術の他の例が、特許文献２に「辞書生成装置」として記載されている。この特許文献２に開示されたソフトウェアの辞書生成装置では、次のように動作する。

【０００５】

入力部は学習用の文書の入力を受け付ける。次に、単語分割部は、入力した文書中のテキストを単語に分割する。さらに、共起頻度表生成部は、文書中の所定の範囲内に出現する単語の頻度統計を収集する。シソーラス頻度表変換部は、辞書の類義関連性をカスタマイズするためのシソーラス情報を仮想的な頻度表に変換する。頻度表統合部は、上記共起頻度表と仮想頻度表を統合する。関連性学習部は、共起頻度表をもとに単語間の関連性を学習し、共起頻度表を圧縮して概念辞書を作成する。このような構成により、辞書の類義関連性をカスタマイズするためのシソーラス情報を仮想的な頻度表に変換することにより、共起頻度表に存在しない単語の頻度情報を補完し、関連性学習処理を行うことで、元の単語量での共起頻度表では取得できなかった潜在的な関連性を辞書に取り込むことを実現している。

【０００６】

また、同義語抽出システムに関する技術の他の例が、特許文献３に「共起行列生成装置」として記載されている。この特許文献３に開示された共起行列生成装置では、次のように動作する。

【０００７】

第１クラスタリング手段は、第１共起行列を入力とし、第１共起行列の行ベクトルの集合をＮ´個のクラスタにクラスタリングし、各クラスタに、Ｎ´個の成分番号のうちの一

つを、クラスタによって成分番号が異なるように付与し、各行ベクトルに対応する単語に、該行ベクトルが属するクラスタの成分番号を対応付ける。第２共起行列生成手段は、形態素解析結果と成分番号付単語集合を入力とし、形態素解析結果中の単語の異なりの集合と、Ｎ´個の成分番号との間で、各行が単語に対応し、各列が成分番号に対応しているような第２共起行列を生成し、該テキストの所定の範囲において、任意の単語Ａと、該範囲中の単語に対応付けられた成分番号Ｂとが共起する頻度を、該テキスト中の全ての所定の範囲にわたって加算した値を、該単語Ａと該成分番号Ｂに対応する第２共起行列の要素とする。第３共起行列生成手段は、任意の単語と任意の成分番号に対し、第１共起行列と第２共起行列の対応する要素を、線形結合した値を、対応する要素とする第３共起行列を生成する。このような構成により、概念語間の類似性を考慮した品質の高さを備え、なおかつ、概念語間の識別性も備えた共起行列を生成することができる。また、典型的には単語・成分番号間共起に基づく手法によって生成した単語・成分番号間共起行列を第１共起行列とし、各概念語を共起語とみなして、共起語に対応する共起ベクトルの集合のクラスタリングにより、共起語群をクラスタリングする。これにより、同一カテゴリに属する共起語群は、一つのクラスタを形成する。次に、概念語の集合と、各クラスタに対応付けられた成分番号の集合との間の共起により、第２共起行列を生成する。このように、クラスタに対応付けられた成分番号との共起頻度をとることにより、同一のカテゴリに属する個々の共起語との共起頻度は、対応する成分番号との共起頻度に含まれるため、共起ベクトルが、より適切なものとなる。これにより、意味の近い概念語間の類似性が高まる。

【先行技術文献】

【特許文献】

【０００８】

【特許文献１】特開２０１０－１５２５６１号公報
【特許文献２】特開２００５－２５０７６２号公報
【特許文献３】特開２０１１－６５３１７号公報

【発明の概要】

【発明が解決しようとする課題】

【０００９】

このような技術の第一の課題は、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は同じで語形が異なっている同義語の抽出に、上記技術による同義語の抽出方法を適用すると、同義語の抽出率が低くなってしまうことである。その理由は、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は同じで語形が異なっている同義語のある文書の多くは、文章量が限られているため任意の単語に対する共起語として同一の単語が出現する可能性が低く、特許文献１の手法で用いられているような共起語の類似性で単語の類似判定を行うことが難しいためである。

【００１０】

また、上記技術の第二の課題は、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は同じで語形が異なっている同義語の抽出に、上記技術による同義語の抽出方法を適用すると、その所定の案件に関する文書から意義は同じで語形が異なっている同義語を抽出することができないことである。その理由は、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は同じで語形が異なっている同義語は、事前にその同義関係を把握することが難しく、特許文献２の手法で用いられているようなカスタマイズされたシソーラスを準備することが困難であるためである。

【００１１】

さらに、先に述べた技術の第三の課題は、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は同じで語形が異なっている同義語の抽出に、上記技術による同義語の抽出方法を適用すると、その所定の案件に関する文書から意義は同じで語形が異なっている同義語を正確に抽出できないことである。その理由は、目的

とする出現頻度の低い単語について引用明細の手法を適用した場合、その単語の成分番号のベクトルは疎なものとなるため、成分番号のベクトルの類似性で行った単語のクラスタリング結果が不正確なものとなり、結果として得られる共起行列も不正確なものとなってしまうためである。

【００１２】

本発明の目的は、上記課題に鑑み、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は同じで語形が異なっている同義語を抽出する、同義語抽出システム、方法およびプログラムを提供することにある。

【課題を解決するための手段】

【００１３】

本発明に係る同義語抽出システムは、対象とする文書もしくは文書群の入力を受け付ける文書入力部と、文書もしくは文書群を構成する文章に使用されている各単語の抽出および単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行う単語分析部と、任意の単語を基軸単語として選択し、単語毎の単語情報に基づき、任意の範囲および条件で基軸単語と共起関係にある共起語とその共起数に基づく基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する基軸単語共起表作成部と、単語の概念分類、同義語、類義語、用法といった単語の一般概念を体系付けた一般概念情報を収集して蓄積し、特定の単語に関する問い合わせに対し、単語の意味や用法に関連する一般概念情報を検索し応答する概念データベースと、基軸単語共起表の各共起語の一般概念情報を概念データベースに問い合わせ、任意の範囲内で基軸単語共起表における各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する単語概念推定部と、各基軸単語に対応する基軸単語概念ベクトル間の類似性を所定の判定基準によって判定し、基軸単語共起ベクトルの意味的な類似性が高い基軸単語の組合せを同義語候補として抽出する同義語候補推定部と、同義語候補を出力する同義語候補出力部と、を備える。

【００１４】

また、本発明の他の形態に係る同義語抽出システムは、対象とする文書もしくは文書群の入力を受け付ける文書入力部と、単語の概念分類、同義語、類義語、用法といった単語の一般概念を体系付けた一般概念情報を収集して蓄積し、特定の単語に関する問い合わせに対し、単語の意味や用法に関連する一般概念情報を検索し応答する概念データベースと、文書もしくは文書群を構成する文章に使用されている各単語の抽出および単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行い、概念データベースに抽出された各単語で一般概念情報の登録が無く、かつ文字数が２文字以上の単語を複合語として抽出し、複合語を構成するあらゆる部分文字列について、一般概念情報の登録がある部分文字列を複合語の有意構成語として抽出し、登録が無い部分文字列を不明構成語として抽出し、さらに有意構成語の一般概念情報を取得する単語分析部と、任意の単語を基軸単語として選択し、単語毎の単語情報に基づき、任意の範囲および条件で基軸単語と共起関係にある共起語とその共起数に基づく基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する基軸単語共起表作成部と、各単語の単語情報および複合語に基づき、任意の範囲および条件で複合語と共起する単語を複合語共起語として、複合語毎に複合語共起語の種類と共起数をまとめた複合語共起表を作成し、複合共起表と構成語に基づき、上記複合語共起表から同じ構成語を含む部分一致複合語の複合語共起語からなる複合語共起ベクトルを抽出し、構成語別に部分一致複合語共起表を作成し、部分一致複合語共起表の複合語共起ベクトルから得られる共起ベクトル空間における各部分一致複合語間の集約度を構成語支配度として算出する構成語支配度算出部と、複合語毎の各構成語の一般概念情報に基づき複合語が関連する概念をまとめた複合語概念構成表を作成する複合語概念配分推定部と、各構成語支配度で複合語毎の各概念の重み付け係数を算出し、複合語概念構成表の対応する箇所に重み付け係数を登録することで、複合語概念配分表を作成し、重み付けされた複数の概念の合成概念として未知の複合語の概念を推定する複合語概念配分推定部と、基軸単語共起表の基軸単語共起ベクトルの各複合語共

起語の内で複合語になっている共起語について、構成語毎の概念に置き換えることで、合成概念に変換し、基軸単語共起表の各共起語の一般概念情報を概念データベースに問い合わせ、任意の範囲内で基軸単語共起表における各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する単語概念推定部と、各基軸単語に対応する基軸単語概念ベクトル間の類似性を所定の判定基準によって判定し、基軸単語共起ベクトルの意味的な類似性が高い基軸単語の組合せを同義語候補として抽出する同義語候補推定部と、同義語候補を出力する同義語候補出力部と、を備える。

【００１５】

また、本発明に係る要求文書分析方法は、対象とする文書もしくは文書群の入力を受け付ける文書受付工程と、文書もしくは文書群を構成する文章に使用されている各単語の抽出および単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行う単語情報抽出工程と、任意の単語を基軸単語として選択し、単語毎の単語情報に基づき、任意の範囲および条件で基軸単語と共起関係にある共起語とその共起数に基づく基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する基軸単語共起表作成工程と、単語の概念分類、同義語、類義語、用法といった単語の一般概念を体系付けた一般概念情報を収集して蓄積し、特定の単語に関する問い合わせに対し、単語の意味や用法に関連する一般概念情報を検索し応答する概念データベースに基軸単語共起表の各共起語の一般概念情報を問い合わせ、任意の範囲内で基軸単語共起表における各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する単語概念推定工程と、各基軸単語に対応する基軸単語概念ベクトル間の類似性を所定の判定基準によって判定し、基軸単語共起ベクトルの意味的な類似性が高い基軸単語の組合せを同義語候補として抽出する同義語候補推定工程と、同義語候補を出力する同義語候補出力工程と、を含む。

【００１６】

また、本発明の他の形態に係る要求文書分析方法は、対象とする文書もしくは文書群の入力を受け付ける文書受付工程と、文書もしくは文書群を構成する文章に使用されている各単語の抽出および単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行い、単語の概念分類、同義語、類義語、用法といった単語の一般概念を体系付けた一般概念情報を収集して蓄積し、特定の単語に関する問い合わせに対し、単語の意味や用法に関連する一般概念情報を検索し応答する概念データベースに抽出された各単語で一般概念情報の登録が無く、かつ文字数が２文字以上の単語を複合語として抽出し、複合語を構成するあらゆる部分文字列について、一般概念情報の登録がある部分文字列を複合語の有意構成語として抽出し、登録が無い部分文字列を不明構成語として抽出し、さらに有意構成語の一般概念情報を取得する単語分析工程と、任意の単語を基軸単語として選択し、単語毎の単語情報に基づき、任意の範囲および条件で基軸単語と共起関係にある共起語とその共起数に基づく基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する基軸単語共起表作成工程と、各単語の単語情報および複合語に基づき、任意の範囲および条件で複合語と共起する単語を複合語共起語として、複合語毎に複合語共起語の種類と共起数をまとめた複合語共起表を作成し、複合共起表と構成語に基づき、上記複合語共起表から同じ構成語を含む部分一致複合語の複合語共起語からなる複合語共起ベクトルを抽出し、構成語別に部分一致複合語共起表を作成し、部分一致複合語共起表の複合語共起ベクトルから得られる共起ベクトル空間における各部分一致複合語間の集約度を構成語支配度として算出する構成語支配度算出工程と、複合語毎の各構成語の一般概念情報に基づき複合語が関連する概念をまとめた複合語概念構成表を作成し、各構成語支配度で複合語毎の各概念の重み付け係数を算出し、複合語概念構成表の対応する箇所に重み付け係数を登録することで、複合語概念配分表を作成し、重み付けされた複数の概念の合成概念として未知の複合語の概念を推定する複合語概念配分推定工程と、基軸単語共起表の基軸単語共起ベクトルの各複合語共起語の内で複合語になっている共起語について、構成語毎の概念に置き換えることで、合成概念に変換し、基軸単語共起表の各共

起語の一般概念情報を概念データベースに問い合わせ、任意の範囲内で基軸単語共起表における各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する単語概念推定工程と、各基軸単語に対応する基軸単語概念ベクトル間の類似性を所定の判定基準によって判定し、基軸単語共起ベクトルの意味的な類似性が高い基軸単語の組合せを同義語候補として抽出する同義語候補推定部と、同義語候補を出力する同義語候補出力工程と、を含む。

【発明の効果】

【００１７】

本発明によれば、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は同じで語形が異なっている同義語を抽出する、同義語抽出システム、方法およびプログラムを提供できる。

【図面の簡単な説明】

【００１８】

【図１】本発明の第１の実施形態に係る同義語抽出システムの構成を示すブロック図である。

【図２】図１に示した同義語抽出システムの動作例を示すシーケンス図である。

【図３】本発明の第２の実施形態に係る同義語抽出システムの構成を示すブロック図である。

【図４】図３に示した同義語抽出システムの動作例を示すシーケンス図である。

【図５】本発明の第１の実施例に係る同義語抽出システムの構成を示すブロック図である。

【図６】基軸単語共起表ＳＶの一部の例を示す説明図である。

【図７】インターネット・サーバＺ内に保存されたシソーラスの一般概念情報Ｃｇの分類体系の例を示す説明図である。

【図８】大分類の基軸単語概念表ＳＣ１の一部の例を示す説明図である。

【図９】中分類の基軸単語概念表ＳＣ２の一部の例を示す説明図である。

【図１０】小分類の基軸単語概念表ＳＣ３の一部の例を示す説明図である。

【図１１】本発明の第２の実施例に係る同義語抽出システムの構成を示すブロック図である。

【図１２】構成語「システム」を含む部分一致複合語共起表ＶＵｘの一部の例を示す説明図である。

【図１３】構成語「変更」を含む部分一致複合語共起表ＶＵｘの一部の例を示す説明図である。

【図１４】複合語「システム変更」に関する複合語概念配分表Ｔｅの一部の例を示す説明図である。

【図１５】複合語を考慮した大分類の基軸単語概念表ＳＣ１の一部の例を示す説明図である。

【図１６】複合語を考慮した中分類の基軸単語概念表ＳＣ２の一部の例を示す説明図である。

【図１７】複合語を考慮した小分類の基軸単語概念表ＳＣ３の一部の例を示す説明図である。

【発明を実施するための形態】

【００１９】

［実施形態１］

最初、本発明の第１の実施形態について、図面を参照して詳細に説明する。

【００２０】

図１は、本発明の第１の実施形態に係る同義語抽出システム１００の構成を示すブロック図である。

【００２１】

図１を参照すると、本発明の第１の実施形態に係る同義語抽出システム１００は、基本

的に電子機器内もしくはサーバと電子機器およびこれらを相互に接続するインターネット等の情報通信ネットワークからなるシステム内に、少なくとも、文書入力部１０、単語分析部２０、基軸単語共起表作成部３０、単語概念推定部４０、同義語候補推定部５０、同義語候補出力部６０、概念データベース１１０と、を含む。

【００２２】

図示の同義語抽出システム１００は、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は同じで語形が異なっている同義語のある文書の同義語抽出システムである。

【００２３】

電子機器で同義語抽出システムを構成する場合、同義語抽出システム１００は、プログラム制御により動作するコンピュータで実現可能である。図示はしないが、この種のコンピュータは、周知のように、データを入力する入力装置と、データ処理装置と、データ処理装置での処理結果を出力する出力装置と、種々のデータベースとして働く補助記憶装置とを備えている。そして、データ処理装置は、プログラムを記憶するリードオンリメモリ（ＲＯＭ）と、データを一時的に記憶するワークエリアとして使用されるランダムアクセスメモリ（ＲＡＭ）と、ＲＯＭに記憶されたプログラムに従って、ＲＡＭに記憶されているデータを処理する中央処理装置（ＣＰＵ）とから構成される。

【００２４】

この場合、データ処理装置が、文書入力部１０、単語分析部２０、基軸単語共起表作成部３０、単語概念推定部４０、同義語候補推定部５０として働き、補助記憶装置が概念データベース１１０として動作し、出力装置が同義語候補出力部６０として働く。

【００２５】

次に、同義語抽出システム１００を構成する各構成要素の動作について説明する。

【００２６】

文書入力部１０は、同義語を抽出する対象とする文書もしくは文書群の入力を受け付ける。

【００２７】

単語分析部２０は、文書もしくは文書群を構成する各文章に形態素解析や構文解析を適用することで、各文章に使用されている全単語の抽出および単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行う。ここで、単語は名詞、動詞、形容詞など単独で意味をなす自立語に限定しても良い。上記単語情報には必要に応じて単語間の係り受け関係などを含めても良い。

【００２８】

基軸単語共起表作成部３０は、単語分析部２０で抽出された各文章に使用されている任意の単語を基軸単語として順次選択し、単語毎の単語情報などを用いて任意の基軸単語共起判定ルールで基軸単語と共起関係とみなされる共起語とその共起数とで表される基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する。ここで、上記基軸単語共起判定ルールとしては１文、１段落内の全文章、目次上の同一項目内での全文章、文書全体、存在する文書名や目次上の項目名など、文書の特徴に合わせて共起語と見なす範囲を設定して良く、１文内での共起する動詞、および目次上の同一項目内の文章内の名詞のように品詞毎に共起とみなす範囲を変えても良い。さらに、単語情報に単語間の係り受け関係が含まれる場合は、係り受け関係のある単語かどうかを上記基軸単語共起判定ルールとして利用しても良い。また、共起数は共起回数でも良いが、共起回数を基軸単語毎の全共起語数で除した頻度などでも良い。また、基軸単語共起語とその共起数について、抽出元とする所定文書について、重要度や確度、文書間の親子関係などに基づく重み付けを行なうようにしても良い。また、基軸単語共起表とは各行が各基軸単語に、各列が各共起語に対応している行列で、基軸単語に対する共起語の共起数が表の各値として登録されたものである。なお、基軸単語は相互的なもので、先に基軸単語として選択された単語であっても、後に他の単語を基軸単語とみなす場合は共起語として扱う。

【００２９】

概念データベース１１０は、収集された単語の概念分類および一般的な同義語、類義語、用法などの一般概念情報を蓄積し、特定の単語に関する問い合わせに対し、単語の意味や用法に関連する一般概念情報を検索し応答するデータベースである。概念データベース１１０は、単語の上位/下位関係、部分/全体関係、同義関係、類義関係などによって単語を分類し、体系づけたシソーラスなどが相当する。なお、概念データベース１１０として、インターネット上のデータベースを使用することとしてもよい。

【００３０】

単語概念推定部４０は、基軸単語共起表の基軸単語共起ベクトルの各共起語のそれぞれについて、概念データベース１１０に一般概念情報を問い合わせ、任意の範囲内で基軸単語共起表における各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する。概念への変換で異なる共起語が同じ概念となる場合はそれぞれの共起語を合流し、共起数の和を対応箇所へ登録する。また、概念データベース１１０として大分類、中分類、小分類のような複数の階層での概念が一般概念情報として登録されたシソーラスを用いる場合、階層毎に概念表を作成し、大分類など広い概念での基軸単語概念表で異なる共起語が同じ概念となる場合は、それぞれの共起語を合流し、共起数の和を対応箇所へ登録する。他に、概念データベース１１０として同義語を含む類義語群が一般概念情報として登録された類語辞書を用いた場合、共起語を対応する類義語群の各類義語に変換し、各類義語の共起数として対応する共起語の共起数を割り当て、同一の基軸単語の共起語に関して変換された類義語毎の共起数の延べ数を基軸単語概念ベクトルとして算出しても良い。なお、概念データベース１１０に共起語に対応する概念が無い場合、上記共起語を概念に変換せず、共起語の単語をそのまま概念として扱い残す。

【００３１】

同義語候補推定部５０は、各基軸単語に対応する概念ベクトル間の類似性を所定の判定基準によって判定し、基軸単語の共起ベクトルの意味的な類似性が高く、同義語の可能性が想定される基軸単語の組合せを同義語候補として抽出する。ここで、類似性の判定を行う「判定基準」は共起語の意味的な類似性を判断する基準であれば良い。例えば、各基軸単語に対応する概念ベクトル間のコサイン距離やユークリッド距離などを非類似度として、これらの距離が任意の閾値より小さい概念ベクトルを持つ基軸単語の組合せとする方法などで良い。或いは、概念データベース１１０として複数の階層での概念が一般概念情報として登録されたシソーラスを用いて概念ベクトルを作成した場合、各階層での非類似度を算出し、小分類などより詳細な深い分類での非類似度ほど重視するように重み付けした非類似度指標が任意の閾値より小さい概念ベクトルを持つ基軸単語の組合せとする方法などで良い。

【００３２】

同義語候補出力部６０は、同義語候補推定部５０で抽出した同義語候補を出力する。ここで、出力形態は、所要の形態で出力すればよく、文書内における同義語候補の組合せを色分けや太字による強調などで明示することで、文書全体を出力する形態などが適当である。他にも、出力形態としては、同義語候補の組合せを抽出した表などの形態であって良い。また、出力形態としては、同義語候補とされた基軸単語を主ノード、その共起語を中間ノード、概念を端ノードとして関係をリンクで結んだグラフを表示し、同義語候補とされた基軸単語を最短で繋ぐリンクを色分けして強調するなどの形態であって良い。また、出力形態としては、同義語候補を抽出する際に用いた非類似度などで同義語間に定量的な同義度を付加し、同義度が任意に設定された閾値より大きい同義語のみに表示を限定しても良い。もしくは、出力形態としては、同義語候補間の同義度によって色分けや太字による強調もしくはグラフの単語の文字の大きさなどに強弱を与えるなどしても良い。また、各出力形態を選択できるようにして、ベースとなる表示形態から必要に応じて表やグラフに移行できるようにしてもよい。また、必要に応じて動詞や名詞などを選択的に出力するようにしてもよい。

【００３３】

次に、図１及び図２のシーケンス図を参照して、本発明の第１の実施形態に係る同義語抽出システム１００の全体の動作について詳細に説明する。なお、図２に示すシーケンス図及び以下の説明は処理例であり、適宜求める処理に応じて処理順等を入れ替えたり処理を戻したり繰り返したりすることを行ってもよい。

【００３４】

文書入力部１０は、対象とする文書もしくは文書群の入力を受け付ける（図２のステップＡ１）。

【００３５】

単語分析部２０は、文書もしくは文書群を構成する各文章に形態素解析や構文解析を適用することで、各文章に使用されている全単語の抽出および単語毎の品詞や格、組み合される助詞、単語間の係り受け関係に関する単語情報の抽出を行う（ステップＡ２）。

【００３６】

基軸単語共起表作成部３０は、単語分析部２０で抽出された各文章に使用されている任意の単語を基軸単語として選択し、単語毎の単語情報に基づき、所定の基軸単語共起判定ルールで基軸単語と共起関係とみなされる共起語とその共起数とで表される基軸単語共起ベクトルを全基軸単語についてまとめた基軸単語共起表を作成する（ステップＡ３）。

【００３７】

概念データベース１１０は、収集蓄積されている単語の概念分類および同義語、類義語、用法などの一般概念情報から、特定の単語に関する問い合わせに対して、適宜、単語の意味や用法に関連する一般概念情報を検索し応答する（ステップＡ４）。

【００３８】

単語概念推定部４０は、基軸単語共起表の基軸単語共起ベクトルの各共起語のそれぞれについて、概念データベース１１０に一般概念情報から概念分類や代表的な同義語や類義語などの概念に相当する情報を問い合わせ、任意の範囲内で基軸単語共起表における各基軸単語共起ベクトルの各共起語を概念に変換した基軸単語概念ベクトルを全基軸単語についてまとめた基軸単語概念表を作成する（ステップＡ５）。

【００３９】

同義語候補推定部５０は、各基軸単語に対応する基軸単語概念ベクトル間の類似性を所定の判定基準によって判定し、基軸単語共起ベクトルの意味的な類似性が高く、同義語の可能性が想定される基軸単語の組合せを同義語候補として順次抽出する（ステップＡ６）。

【００４０】

同義語候補出力部６０は、同義語候補推定部５０で抽出できた同義語候補を出力する（ステップＡ７）。

【００４１】

次に、本発明の第１の実施形態に係る同義語抽出システム１００の効果について説明する。

【００４２】

本第１の実施形態では、文書内もしくは文書群内の基軸単語共起ベクトルを基軸単語概念ベクトルに変換することによって、意味的には類似するが単語としては一致しない共起語も考慮して同義語候補を抽出するように構成しているため、各単語の出現回数が少なく基軸単語共起ベクトルが疎行列で類似の判定が困難な文章量の少ない条件でも類似性の評価が可能になり、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は同じで語形が異なっている同義語を精度よく抽出できる。

【００４３】

尚、上記本発明の第１の実施形態に係る同義語抽出システム１００は、同義語抽出方法として実現され得る。また、上記本発明の第１の実施形態に係る同義語抽出システム１００は、同義語抽出プログラムによりコンピュータによって実行させるようにしても良い。

【００４４】

［実施形態２］

次に、本発明の第２の実施形態について、図面を参照して詳細に説明する。

【００４５】

図３は、本発明の第３の実施形態に係る同義語抽出システム１００Ａの構成を示すブロック図である。

【００４６】

図３を参照すると、本発明の第２の実施形態に係る同義語抽出システム１００Ａは、構成語支配度算出部３５と、複合語概念配分推定部３６と、を更に含むと共に、後述するように単語分析部と単語概念推定部の動作が相違する点を除いて、図１に示した第１の実施形態に係る同義語抽出システム１００と同様の構成を有し、動作をする。したがって、単語分析部に２０Ａの参照符号を、単語概念推定部に４０Ａの参照符号を付してある。

【００４７】

図示の同義語抽出システム１００Ａを上述したコンピュータで実現した場合、データ処理装置が、文書入力部１０、単語分析部２０Ａ、構成語支配度算出部３５、複合語概念配分推定部３６、基軸単語共起表作成部３０、単語概念推定部４０Ａ、同義語候補推定部５０として働き、補助記憶装置が概念データベース１１０として動作し、出力装置が同義語候補出力部６０として働く。

【００４８】

単語分析部２０Ａが単語の中の複合語および複合語の構成語、構成語に対応する概念を取得し、構成語支配度算出部３５が、複合語の構成語毎の構成語支配度を算出し、複合語概念配分推定部３６が、構成語支配度に基づき複合語の構成語毎の概念に重み付けを行った複合語概念配分表を作成し、単語概念推定部４０Ａが、基軸単語の共起語を概念に変換する際に、共起語の中の複合語について複合語概念配分表に基づく変換を行う。

【００４９】

次に、同義語抽出システム１００Ａを構成する各構成要素の動作について説明する。

【００５０】

単語分析部２０Ａは、図１に示した単語分析部２０の動作に加え、抽出された各単語の一般概念情報を概念データベース１１０に問い合わせ、概念データベース１１０に登録が無く、かつ文字数が２文字以上の単語を複合語として抽出する点で、図１に示した単語分析部２０と異なる。さらに単語分析部２０Ａは、複合語を構成するあらゆる部分文字列について、概念データベース１１０に一般概念情報を問い合わせ、一般概念情報の登録がある部分文字列を複合語の有意構成語として抽出し、抽出した有意構成語を元の複合語から分離した場合に概念データベース１１０に一般概念情報の登録が無い部分文字列が残る場合は不明構成語として抽出し、さらに有意構成語の一般概念情報を取得する点で、図１に示した単語分析部２０と異なる。なお複合語を構成する部分文字列の内、概念データベース１１０に一般概念情報の登録がある部分文字列の組合せパターンが複数考えられる場合は、任意の構成語分離ルールに基づいて最適な組合せパターンを判定し、その組合せパターンでの有意構成語、不明構成語を抽出する。ここで、構成語分離ルールとしては、不明構成語の文字数が最も少なくなるパターンを優先するルールや、入力された文書中に単独の単語として出現する頻度が高い有意構成語を優先するルール、一般の文書中に単独の単語として出現する頻度が高い有意構成語を優先するルール、およびこれらを組合せたルールなどが有効である。また、入力された文書中に含まれる他の複合語に共通して使用されている文字列が所定頻度以上に使用されている場合にはその文字列を除いた残りの文字列について、有意構成語として優先するルールを用いてもよい。なお、一般概念情報とはシソーラスにおける分類や、単語の意味を直接的に表すキーワード、類語の集合などが考えられる。なお、以下で単に構成語と記載した場合は有意構成語と不明構成語を含む。

【００５１】

構成語支配度算出部３５は、単語分析部２０Ａで抽出された各文章に使用されている単語の単語情報および複合語に基づき、任意の複合語共起判定ルールで複合語と共起関係とみなされる単語を複合語共起語として、複合語毎に複合語共起語とその共起数を抽出し、これらをまとめることで複合語共起表を作成する。ここで、上記複合語共起判定ルールと

しては１文、１段落内の全文章、目次上の同一項目内での全文章、文書全体、文章のスタイル、文章群の中での位置付けなど、文書の特徴に合わせて複合語共起語と見なす範囲を設定して良い。例えば、品詞が動詞であれば１文内での共起、名詞であれば目次上の同一項目内での全文章内共起のように品詞毎に文書群の範囲を変えるようにすれば良い。また、共起数は共起回数でも良いが、共起回数を複合語毎の全共起語数で除した頻度などでも良い。さらに、単語情報に単語間の係り受け関係が含まれる場合は、係り受け関係のある単語かどうかを上記範囲および条件として利用しても良い。また、複合語共起表とは各行が各複合語に、各列が各複合語共起語に対応している行列で、複合語に対する複合語共起語の共起数が表の各値として登録されたものである。さらに、構成語支配度算出部３５は、複合共起表と単語分析部２０Ａで抽出された構成語に基づき、上記複合語共起表から同じ構成語を含む部分一致複合語の複合語共起語からなる複合語共起ベクトルを抽出し、構成語別に部分一致複合語共起表を作成する。そして、部分一致複合語共起表の複合語共起ベクトルから得られる共起ベクトル空間における各部分一致複合語間の集約度を構成語支配度として算出する。ここで、共起ベクトル空間は各ベクトルを対等としても良いが、複合語共起語の品詞によって重み付けを行ったベクトル空間に変換しても良い。また、各部分一致複合語間の集約度とは各部分一致複合語に対応するベクトル間の散らばりの小ささを表す指標であればどのような算出方法によっても良い。例えば分散や標準偏差、変動係数などの一般に統計で用いられるばらつきを示す指標と単調減少の関係にある関数であればよく、分散の逆数や変動係数の逆数などが適している。

【００５２】

複合語概念配分推定部３６は、複合語毎に単語分析部２０Ａで概念データベース１１０から取得した各構成語の一般概念情報に基づき複合語が関連する概念をまとめた複合語概念構成表を作成する。複合語概念構成表とは各行が各複合語に、各列が複合語の各構成語の概念に対応した行列で、複合語と概念との間に構成語を介した関連があるかどうかの有無が登録されたものである。なお、複合語の構成語に不明構成語が含まれる場合、不明構成語自体を概念として新たに列を加える。さらに、複合語概念配分推定部３６は、構成語支配度算出部３５で算出した各構成語支配度で複合語毎の各概念の重み付け係数を算出し、複合語概念構成表の対応する箇所に重み付け係数を登録することで、複合語概念配分表を作成し、重み付けされた複数の概念の合成概念として未知の複合語の概念を推定する。ここで、上記重み付け係数の算出方法としては、各構成語の構成語支配度を複合語毎の構成語支配度の総和で除すことで正規化した値を指標とする方法などが有効である。

【００５３】

単語概念推定部４０Ａは、上記説明した単語概念推定部４０の動作に加え、基軸単語共起表作成部３０で作成された基軸単語共起表の基軸単語共起ベクトルの各共起語の内で複合語になっている共起語について、複合語概念配分推定部３６で作成した複合語概念配分表に基づき、構成語別の概念の共起数に重み付けした合成概念へ変換した結果を、基軸単語概念ベクトルに反映し基軸単語概念表を作成する点で、図１に示した単語概念推定部４０と異なる。

【００５４】

それ以外の文書入力部１０、基軸単語共起表作成部３０、同義語候補推定部５０、同義語候補出力部６０、概念データベース１１０の構成と機能は、図１に示した第１の実施形態のそれらとそれぞれ同じであるので、説明を省略する。

【００５５】

次に、図３及び図４のシーケンス図を参照して、本発明の第２の実施形態に係る同義語抽出システム１００Ａの全体の動作について詳細に説明する。なお、図４に示すシーケンス図および以下の説明は処理例であり、第１の実施形態と同様に処理順等を入れ替えたり処理を戻したりすることを行ってもよい。

【００５６】

上述した第１の実施形態の動作と比較すると、以下に説明する本第２の実施形態の動作は、次の動作が加わっている点で異なる。

【００５７】
すなわち、単語分析部２０Ａは、図１に示した単語分析部２０の動作（ステップＡ２）に加え、抽出された各単語の一般概念情報を概念データベース１１０に問い合わせ、概念データベース１１０に登録が無く、かつ文字数が２文字以上の単語を複合語として抽出する（ステップＢ１）。
【００５８】
さらに単語分析部２０Ａは、複合語を構成するあらゆる部分文字列について、概念データベース１１０に一般概念情報を問い合わせ、一般概念情報の登録がある部分文字列を複合語の有意構成語として抽出し、抽出した有意構成語を元の複合語から分離した場合に概念データベース１１０に一般概念情報の登録が無い部分文字列が残る場合は不明構成語として抽出し、さらに有意構成語の一般概念情報を取得する（ステップＢ２）。
【００５９】
次に構成語支配度算出部３５は、単語分析部２０Ａで抽出された各文章に使用されている単語の単語情報、および複合語に基づき、任意の複合語共起判定ルールで複合語と共起関係とみなされる単語を複合語共起語として、複合語毎に複合語共起語とその共起数を抽出し、これらをまとめることで複合語共起表を作成する（ステップＢ３）。
【００６０】
さらに構成語支配度算出部３５は、複合共起表と単語分析部２０Ａで抽出された構成語に基づき、上記複合語共起表から同じ構成語を含む部分一致複合語の複合語共起語からなる複合語共起ベクトルを抽出し、構成語別に部分一致複合語共起表を作成し、部分一致複合語共起表の複合語共起ベクトルから得られる共起ベクトル空間における各部分一致複合語間の集約度を構成語支配度として算出する（ステップＢ４）。
【００６１】
次に複合語概念配分推定部３６は、複合語毎に単語分析部２０Ａで概念データベース１１０から取得した各構成語の一般概念情報に基づき複合語が関連する概念をまとめた複合語概念構成表を作成する（ステップＢ５）。
【００６２】
さらに複合語概念配分推定部３６は、構成語支配度算出部３５で算出した各構成語支配度で複合語毎の各概念の重み付け係数を算出し、複合語概念構成表の対応する箇所に重み付け係数を登録することで、複合語概念配分表を作成し、重み付けされた複数の概念の合成概念として未知の複合語の概念を推定する（ステップＢ６）。
【００６３】
単語概念推定部４０Ａは、図１に示した単語概念推定部４０の動作（ステップＡ５）に加え、基軸単語共起表作成部３０で作成された基軸単語共起表の基軸単語共起ベクトルの各複合語共起語の内で複合語になっている共起語について、複合語概念配分推定部３６で作成した複合語概念配分表に基づき、構成語別の概念の共起数に重み付けした合成概念へ変換した結果を、基軸単語概念ベクトルに反映し基軸単語概念表を作成する（ステップＡ５‘）。
【００６４】
他のステップの動作は、上述した第１の実施形態における動作と同一であるので、それらの説明については省略する。
【００６５】
次に、本発明の第２の実施形態の効果について説明する。
【００６６】
第２の実施形態では、第１の実施の形態の効果に加え、共起語の中の複合語について構成語毎の構成語支配度を算出し、構成語支配度に基づき重み付けを行った概念に変換する。これによって、シソーラスなどに一般概念情報の登録が無い複合語なども考慮して同義語候補を抽出するように構成できるため、基軸単語共起ベクトルから基軸単語概念ベクトルへの変換の障害となる、独自の複合語の多い文章群でも類似性の評価が可能になり、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は

同じで語形が異なっている同義語をより精度よく抽出できる。
【００６７】
尚、上記本発明の第２の実施形態に係る同義語抽出システム１００Ａは、同義語抽出方法として実現され得る。また、上記本発明の第１の実施形態に係る同義語抽出システム１００Ａは、同義語抽出プログラムによりコンピュータによって実行させるようにしても良い。
【実施例１】
【００６８】
次に、図５を参照して、具体的な第１の実施例を用いて、本発明の第１の実施形態に係る同義語抽出システム１００の動作について説明する。
【００６９】
本第１の実施例では、次のことを目的としている。
【００７０】
先ず、同義語抽出システム１００は、情報システム構築に関する提案書や仕様書等といった一般的な意味と異なった概念を示す意味としても使用される同義語を含む文書Ｄ内に含まれる所定の案件に関する文書から意義は同じで語形が異なっている同義語候補Ａを推定する。そして、同義語抽出システム１００は、推定結果を出力することで、未登録の用語に関する用語集の作成や語の統一を支援する。また、本第１の実施例では、同義語抽出システム１００は、図５に示されるように、文書解析システムＹと、インターネット・サーバＺとで構成されるものとする。
【００７１】
文書解析システムＹは、分析実施者Ｂの持つＰＣ端末上で動作し、入力部及び出力部を介して、分析実施者Ｂが同義語を抽出したい文書群を構成する文章の入力と、同義語候補Ａの提示を実現する。
【００７２】
インターネット・サーバＺは、通信ネットワークを介して文書解析システムＹを実装した分析実施者Ｂの持つＰＣ端末と接続されている。インターネット・サーバＺは、文書解析システムＹからの単語の意味などの概念情報の問い合わせに対し、単語の概念分類や一般的な同義語や類義語、用法に関連する一般概念情報Ｃｇの検索を可能にする装置である。
【００７３】
図５と図１との対応関係について説明する。
【００７４】
文書入力部１０と、単語分析部２０と、基軸単語共起表作成部３０と、単語概念推定部４０と、同義語候補推定部５０とは、文書解析システムＹ内に含まれている。同義語候補出力部６０は、ＰＣ端末の出力部として動作する。概念データベース１１０はインターネット・サーバＺ内に含まれている。
【００７５】
この様な手段を備えた文書解析システムＹ、インターネット・サーバＺは以下のような動作をする。
【００７６】
文書解析システムＹは、入力部から、分析実施者Ｂが特定の案件に関する文書から意義は同じで語形が異なっている同義語候補Ａを推定したい文書群を構成する文書Ｄの入力を受け付ける。そして、文書解析システムＹは、文書Ｄを構成する文章毎に形態素解析および構文解析を適用し、文書を構成する単語に分解し、各単語の品詞とその係り受け関係を解析することで、名詞および、動詞、形容詞、形容動詞を単語Ｗとして抽出する。なお、動詞の内でサ行変格活用に属する動詞は活用部分を除去しいわゆるサ変名詞化したものを動詞として抽出する。
【００７７】
さらに文書解析システムＹは、文書Ｄに含まれる単語Ｗの内で名詞を基軸単語Ｓとし、

各基軸単語Ｓｉ（ｉ＝１、２、・・・、ｎ）について、特定の基軸単語Ｓｉと係り受け関係にある動詞と形容詞と形容動詞、および目次上の同一項目内の文章内で共起する名詞を、共起語Ｖｊ（ｊ＝１、２、・・・、ｍ）として抽出し、基軸単語Ｓｉに対する各共起語Ｖｉｊの共起回数を共起数Ｎｉｊとして集計し、全ての基軸単語Ｓに対する各共起語Ｖについて表形式にまとめた基軸単語共起表ＳＶを作成する。なお、基軸単語共起表ＳＶの基軸単語Ｓｉに対する各共起語Ｖｊの共起数Ｎｉｊをまとめたデータセットを基軸単語共起ベクトルＮｉと呼ぶ。例えば、文書Ｄから、基軸単語Ｓとして「演算システム」、「分析機能」、・・・などの単語が、共起語Ｖとして「利用」、「操作」、「構築」、「改善」、「システム変更」、「メカニズム」、「瞬時」、「短期」、「稼働」、「高速処理」、・・・などの単語が抽出された場合、基軸単語共起表ＳＶは図６のような、各行に基軸単語Ｓを各列に共起語Ｖを配置し、その共起数Ｎｉｊを記載した表になる。また、図６の基軸単語Ｓｉの行のデータセットが基軸単語共起ベクトルＮｉに相当し、「演算システム」の基軸単語共起ベクトルＮｉは｛０、３、２、０、４、０、１、０、３、０、・・・｝のように表される。なお、基軸単語Ｓと共起語Ｖはいずれも名詞を含むため、先に基軸単語として選択された単語も、他の単語が基軸単語の場合は共起語として扱い、相互で重複して登録する。

【００７８】

インターネット・サーバＺは、単語の一般的な上位/下位関係、部分/全体関係、同義関係、類義関係などによって単語を分類し、体系づけたシソーラスの一般概念情報Ｃｇを蓄積する。また、インターネット・サーバＺは、任意の単語の情報を抽出する検索エンジンなどの機能も提供することで、文書解析システムＹからの問い合わせに応じて、問い合わせ対象の単語の一般的な概念分類として大分類、中分類、小分類を一般概念情報Ｃｇとして抽出し、提示する。

【００７９】

文書解析システムＹは、基軸単語共起表ＳＶの各共起語Ｖｊのそれぞれの一般概念情報Ｃｇについてインターネット・サーバＺに問い合わせを行うことで、インターネット・サーバＺ内に保存されたシソーラスの一般概念情報Ｃｇの分類体系から、各共起語Ｖｊが属する大分類の共起語概念Ｃ１ｖｊと、中分類の共起語概念Ｃ２ｖｊと、小分類の共起語概念Ｃ３ｖｊとを抽出し、基軸単語共起表ＳＶにおける共起語Ｖｊを共起語概念Ｃ１ｖｊに変換し、同じ概念となる共起語Ｖｉをまとめ、共起数Ｎｉｊの和を対応箇所へ登録した、大分類の基軸単語概念表ＳＣ１、基軸単語共起表ＳＶにおける共起語Ｖｊを共起語概念Ｃ２ｖｊに変換し、同じ概念となる共起語Ｖｉをまとめ、共起数Ｎｉｊの和を対応箇所へ登録した、中分類の基軸単語概念表ＳＣ２、基軸単語共起表ＳＶにおける共起語Ｖｊを共起語概念Ｃ３ｖｊに変換し、同じ概念となる共起語Ｖｉをまとめ、共起数Ｎｉｊの和を対応箇所へ登録した、小分類の基軸単語概念表ＳＣ３を作成する。なお、大分類の基軸単語概念表ＳＣ１の基軸単語Ｓｉに対する各共起語概念Ｃ１ｖｊの共起数Ｎｃ１ｉｊをまとめたデータセットを大分類基軸単語概念ベクトルＮｃ１ｉと呼び、中分類の基軸単語概念表ＳＣ２の基軸単語Ｓｉに対する各共起語概念Ｃ２ｖｊの共起数Ｎｃ２ｉｊをまとめたデータセットを中分類基軸単語概念ベクトルＮｃ２ｉと呼び、小分類の基軸単語概念表ＳＣ３の基軸単語Ｓｉに対する各共起語概念Ｃ３ｖｊの共起数Ｎｃ３ｉｊをまとめたデータセットを小分類基軸単語概念ベクトルＮｃ３ｉと呼ぶ。例えば、図６の基軸単語共起表ＳＶにおける各共起語Ｖｊについて、図７のような共起語概念Ｃ１ｖｊ、共起語概念Ｃ２ｖｊ、共起語概念Ｃ３ｖｊが抽出された場合、大分類の基軸単語概念表ＳＣ１は図８、中分類の基軸単語概念表ＳＣ２は図９、小分類の基軸単語概念表ＳＣ３は図１０のような各行に基軸単語Ｓを各列に共起語概念Ｃ１ｖｊを配置した表となる。基軸単語概念表ＳＣ１、ＳＣ２、ＳＣ３の各共起数は大分類の基軸単語概念表ＳＣ１を例とすると、共起語Ｖの内で「利用」、「操作」、「構築」、「改善」、「稼働」の共起語概念Ｃ１ｖｊは「人間活動」で共通のため、これらの共起語における共起数を同一の基軸単語「演算システム」に関して足し合わせた「８」がＮｃ１ｉｊとなる。同様に共起語Ｖの内で「メカニズム」、「瞬時」、「短期」の共起語概念Ｃ１ｖｊは「抽象」で共通のため、これらの共起語における共

起数を基軸単語「演算システム」に関して足し合わせた「１」がＮｃ１ｉｊとなる。なお、インターネット・サーバＺに一般概念情報Ｃｇの登録が無い「システム変更」、「高速処理」などの複合語は、共起語の単語をそのまま仮の概念として残す。図８より、基軸単語「演算システム」の大分類基軸単語概念ベクトルＮｃ１ｉは｛８、４、１、０、・・・｝のように表される。

【００８０】

次に文書解析システムＹは、基軸単語Ｓｐに対応する大分類基軸単語概念ベクトルＮｃ１ｐと基軸単語Ｓｑに対応する大分類基軸単語概念ベクトルＮｃ１ｑの間のコサイン距離ｄｃ１ｐｑと、中分類基軸単語概念ベクトルＮｃ２ｐとＮｃ２ｑの間のコサイン距離ｄｃ２ｐｑと、小分類基軸単語概念ベクトルＮｃ３ｐとＮｃ３ｑの間のコサイン距離ｄｃ３ｐｑとを算出し、以下の（１）式によりそれぞれの分類重み付け係数β１、β２、β３（β１<β２<β３）を掛けた和を基軸単語間距離ｄｐｑとして算出し、基軸単語間距離ｄｐｑが任意の判定閾値Ｔより小さい基軸単語Ｓｐと基軸単語Ｓｑの組合せを、基軸単語の共起ベクトルの意味的な類似性が高く、同義語の可能性が想定される基軸単語の組合せである同義語候補Ａとして抽出する。この処理を全ての基軸単語Ｓｉの組合せについて行う。

【００８１】

ｄｐｑ＝β１×ｄｃ１ｐｑ＋β２×ｄｃ２ｐｑ＋β３×ｄｃ３ｐｑ・・・（１）式

【００８２】

例えば、図８～１０の例では基軸単語「演算システム」と「分析機能」のコサイン距離は、ｄｃ１ｐｑ＝０．２６、ｄｃ２ｐｑ＝０．５７、ｄｃ３ｐｑ＝０．６８となり、分類重み付け係数をβ１＝０．００９、β２＝０．０９、β３＝０．９、判定閾値Ｔ＝０．７とすると、基軸単語間距離ｄｐｑ＝０．６７で判定閾値Ｔより小さくなるので、「演算システム」と「分析機能」はこの文章内では同義語である可能性があると判定される。また、基軸単語間距離ｄｐｑ＝０．６７は、図６に基づく基軸単語共起ベクトルＮｉ間の距離０．８７よりも小さく、概念情報に変換して意味を考慮することで「演算システム」と「分析機能」との同義性が分かりやすくなることが分かる。

【００８３】

さらに文書解析システムＹは、同義語候補Ａａ｛Ｓｐ、Ｓｑ｝について、要求文書Ｄで該当する同義語候補Ａａ｛Ｓｐ、Ｓｑ｝を色分けもしくは太字による強調などの加工を行い、加工後の要求文書Ｄを、出力部から出力する。

【実施例２】

【００８４】

次に、図９を参照して、具体的な第２の実施例を用いて、本発明の第２の実施形態に係る同義語抽出システム１００Ａの動作を説明する。

【００８５】

本第２の実施例では、同義語抽出システム１００Ａは、図１１に示されるように、インターネット・サーバＺ’を利用するものとする。

【００８６】

文書解析システムＹａは、分析実施者Ｂの持つＰＣ端末上で動作し、入力部及び出力部を介して、分析実施者Ｂが同義語を抽出したい文書群を構成する文章の入力と、同義語候補Ａの提示を実現する。

【００８７】

インターネット・サーバＺ’は、既存のシソーラスを提供するサーバであり、通信ネットワークを介して文書解析システムＹａを実装した分析実施者Ｂの持つＰＣ端末と接続されている。インターネット・サーバＺ’は、文書解析システムＹａからの単語の意味情報の問い合わせに対し、単語の概念分類や一般的な同義語や類義語、用法に関連する一般概念情報Ｃｇの検索を可能にする装置である。

【００８８】

本第２の実施例では、第１の実施例の動作に加え、文書解析システムＹａが構成語支配度算出部３５と、複合語概念配分推定部３６と、を更に含む。

【００８９】
すなわち、図１１と図３との対応関係は次のように成る。
【００９０】
文書入力部１０と、単語分析部２０Ａと、構成語支配度算出部３５と、複合語概念配分推定部３６と、基軸単語共起表作成部３０と、単語概念推定部４０Ａと、同義語候補推定部５０とは、文書解析システムＹａ内に含まれている。同義語候補出力部６０は、ＰＣ端末の出力部として動作する。概念データベース１１０はインターネット・サーバＺ’内に含まれている。
【００９１】
この様な構成を含めた文書解析システムＹａは、上述した第１の実施例に対して、以下のような動作を加える。
【００９２】
文書解析システムＹａは、各共起語Ｖｊのそれぞれの一般概念情報Ｃｇをインターネット・サーバＺ’に問い合わせることで、インターネット・サーバＺ’内に保存されたシソーラスに、各共起語Ｖｊの一般概念情報Ｃｇが登録されているかどうかを検索し、シソーラスに一般概念情報Ｃｇの登録が無く、かつ文字数が２文字以上の単語を複合語Ｖｍｅ（ｅ＝１、２、・・・、ｈ）として抽出する。例えば「高速処理」という単語がシソーラスに登録されていない場合は、２文字以上であるため複合語として抽出する。
【００９３】
さらに文書解析システムＹａは、複合語Ｖｍｅ毎に複合語Ｖｍｅの文字列をあらゆるパターンで分離し、分離した全ての部分文字列について、インターネット・サーバＺ’内に保存されたシソーラスに一般概念情報Ｃｇが登録されているかどうかを検索する。そして、一般概念情報の登録がない部分文字列の文字数が最も少なくなるパターンでの、部分文字列を複合語Ｖｍｅの構成語Ｐｅｋ（ｋ＝１、２、・・・、ｌ）として処理し、構成語Ｐｅｋの内、一般概念情報Ｃｇの登録が有る部分文字列は有意構成語Ｐａｅｋ、登録が無い部分文字列は不明構成語Ｐｂｅｋとして、それぞれ複合語毎に抽出する。先の「高速処理」という複合語の例では、｛「高」、「速処理」｝、｛「高速」、「処理」｝、｛「高速処」、「理」｝が分離可能な文字列として想定され、「速処理」と「高速処」がシソーラスに登録されていない場合は、「高」、「高速」、「処理」、「理」が有意構成語Ｐａｅｋの候補、「速処理」、「高速処」が不明構成語Ｐｂｅｋの候補となるが、一般概念情報Ｃｇの登録がない部分文字列の文字数が最も少ない｛「高速」、「処理」｝の組合せが複合語「高速処理」の有意構成語として選択される。さらに文書解析システムＹａは、インターネット・サーバＺ’内に保存されたシソーラスに一般概念情報Ｃｇから、有意構成語Ｐａｅｋが属する大分類の構成語概念Ｃａ１ｅｋと、中分類の構成語概念Ｃａ２ｅｋと、小分類の構成語概念Ｃａ３ｅｋとを取得する。
【００９４】
文書解析システムＹａは、「構築する情報システムの機能」など文書Ｄで一定の範囲の内容に言及している文章群として分析者Ｂが指定した段落の文章内で複合語Ｖｍｅと共起する名詞、および複合語Ｖｍｅに係る動詞と形容詞、形容動詞をｓ個の複合語共起語Ｕｍｅｒ（ｒ＝１、２、・・・、ｓ）として、複合語Ｖｍｅ毎に複合語共起語Ｕｍｅｒと、共起と見なした範囲内での共起回数Ｍｅｒを抽出し、各行を各複合語Ｖｍｅに、各列を各複合語共起語Ｕｍｅｒに対応させ、複合語Ｖｍｅに対する複合語共起語Ｕｍｅｒの共起回数Ｍｅｒを各値として登録した疎行列からなる複合語共起表ＶＵｍを作成する。さらに、文書解析システムＹａは、上記複合語共起表ＶＵｍの各構成語Ｐｅｋ別に、同じ構成語Ｐｘ（ｘ＝１、２、・・・、ｔ）を含むｔ個の複合語Ｖｍｘの行成分（Ｍｘ１，Ｍｘ２，Ｍｘ３，・，・，・，Ｍｘｓ）を抽出し、各行成分を各複合語Ｖｍｘに、各列を各複合語共起語Ｕｍｘｒに対応させ、複合語Ｖｍｘに対する複合語共起語Ｕｍｘｒの共起回数Ｍｘｒを各値として登録した疎行列からなる部分一致複合語共起表ＶＵｘを作成する。例えば「システム」という構成語を含む部分一致複合語共起表としては図１２、「変更」という構成語を含む部分一致複合語共起表としては図１３のような表が作成される。さらに、文書解

析システムＹａは、以下の数１のように、部分一致複合語共起表ＶＵｘの複合語共起語Ｕｍｘｒ毎のデータ列（Ｍ１ｒ，Ｍ２ｒ，Ｍ３ｒ，・，・，・，Ｍｔｒ）で分散σｘｒを算出し、全複合語共起語Ｕｍｘｒの分散σｘｒの平均値の平方根の逆数を構成語Ｐｘの構成語支配度Ｇｘとして算出する。

【００９５】

【数１】

$$Gx=\sqrt{\frac{S}{\sqrt{\sum_{r=1}^{S}\left(\frac{\sum_{x=1}^{t}\left(M_{xr}-\frac{\sum_{x=1}^{t}M_{xr}}{t}\right)^{2}}{t}\right)}}} \qquad \cdots \quad (2)\text{式}$$

【００９６】

なお、複合語Ｖｍｅの構成語に不明構成語Ｐｂｅｋが有る場合は、不明構成語Ｐｂｅｋの文字列を新概念Ｃｂｅｋとする。

【００９７】

文書解析システムＹａは、複合語Ｖｍｅ毎の各構成語Ｐｅｋに対応する各構成語支配度Ｇｘｅｋの値を構成語支配度Ｇｘｅｋの総和で除すことで正規化した概念重み付け係数αｅｋを算出する。さらに文書解析システムＹａは、複合語Ｖｍｅ毎に大分類の構成語概念Ｃａ１ｅｋと、中分類の構成語概念Ｃａ２ｅｋと、小分類の構成語概念Ｃａ３ｅｋおよび新概念Ｃｂｅｋに基づき、複合語概念配分表Ｔｅを作成する。複合語概念配分表Ｔｅは、複合語Ｖｍｅ毎に作られ、各構成語Ｐｅｋに対応する、大分類の構成語概念Ｃａ１ｅｋと中分類の構成語概念Ｃａ２ｅｋと小分類の構成語概念Ｃａ３ｅｋと新概念Ｃｂｅｋ、および概念重み付け係数αｅｋを登録した表である。例えば、複合語「システム変更」に関して、構成語「システム」の構成語支配度Ｇｘが１．４７でシソーラスでの概念が「装置」、構成語「変更」の構成語支配度Ｇｘが２．２１でシソーラスでの概念が「修正」であった場合、複合語概念配分表Ｔｅは図１４のようになる。図１４は、複合語「システム変更」の概念を構成語「変更」と構成語「システム」の合成概念Ｃｅとして理解する場合、構成語「変更」の方が構成語「システム」よりも重要であることを示している。

【００９８】

文書解析システムＹａは、複合語Ｖｍｅが共起語Ｖｊの一つであるという観点から、基軸単語Ｓｉと共起した複合語Ｖｍｉｅを構成語Ｐｉｅｋに分解し、それぞれの構成語Ｐｉｅｋに対応する大分類の構成語概念Ｃａ１ｅｋを大分類の共起語概念Ｃ１ｖｅに、中分類の構成語概念Ｃａ２ｅｋを中分類の共起語概念Ｃ２ｖｅに、小分類の構成語概念Ｃａ３ｅｋと新概念Ｃｂｅｋとを小分類の共起語概念Ｃ３ｖｅに合流させる。さらに複合語概念配分表Ｔｅに基づき複合語Ｖｍｉｅの共起数Ｎｉｅに各構成語Ｐｉｅｋに対応する概念重み付け係数αｅｋを掛けた、重み付け共起数Ｎｉｅｋを算出し、大分類の基軸単語概念表ＳＣ１、中分類の基軸単語概念表ＳＣ２、小分類の基軸単語概念表ＳＣ３を作成する。例えば、図６の各共起語Ｖｉについて、図７のような共起語概念Ｃ１ｖｊ、共起語概念Ｃ２ｖｊ、共起語概念Ｃ３ｖｊが抽出された場合、複合語である「システム変更」と「高速処理」が「システム」と「変更」、および「高速」と「処理」という構成語に分離され、概念重み付け係数が図１４から「システム＝０．４」、「変更＝０．６」で、同様に「高速＝０．３」と「処理＝０．７」だった場合、重み付け共起数Ｎｉｅｋは「システム：１．６＝４×０．４」、「変更：２．４＝４×０．６」、「高速：１．２＝４×０．３」、「処理：２．８＝４×０．７」となり、「高速」の概念分類が「大分類：抽象」、「中分類：速度」、「小分類：速さ」で、「処理」の概念分類が「大分類：人間活動」、「中分類：動き」、「小分類：動作」であれば、大分類の基軸単語概念表ＳＣ１は図１５、中分類の

基軸単語概念表ＳＣ２は図１６、小分類の基軸単語概念表ＳＣ３は図１７のような表となる。図１５～１７の例では基軸単語「演算システム」と「分析機能」のコサイン距離は、ｄｃ１ｐｑ＝０．０３、ｄｃ２ｐｑ＝０．０３、ｄｃ３ｐｑ＝０．１５となり、分類重み付け係数をβ１＝０．００９、β２＝０．０９、β３＝０．９、判定閾値Ｔ＝０．７とすると、基軸単語間距離ｄｐｑ＝０．１４で判定閾値Ｔより小さくなるので、「演算システム」と「分析機能」はこの文章内では同義語である可能性があると判定される。また、基軸単語間距離ｄｐｑ＝０．１４は、図６に基づく基軸単語共起ベクトルＮｉ間の距離０．８７だけでなく、図８～１０に基づく基軸単語間距離ｄｐｑ＝０．６７に比べても小さく、複合語を考慮し、概念情報に変換して意味を考慮することで「演算システム」と「分析機能」との同義性がより分かりやすくなることが分かる。

【００９９】

他の動作は第一の実施例と同様である。

以上説明したように、本発明の同義語抽出システムによれば、情報システム構築に関する提案書や仕様書等といった、所定の案件に関する文書から意義は同じで語形が異なっている同義語のある文書について、その文書群で意義は同じで語形が異なる同義語を把握することが可能となり、誤解に基づく混乱や失敗などの削減につながる。その理由は、単語の類似性を共起語などの概念レベルでの一致具合で算出することで、特定の案件に関する文書群という限られた文書量の情報で、同一の共起語の使用が無くても、単語間の類似性を算出可能にしているためである。

また、本発明の具体的な構成は前述の実施の形態に限られるものではなく、この発明の要旨を逸脱しない範囲の変更があってもこの発明に含まれる。

【産業上の利用可能性】

【０１００】

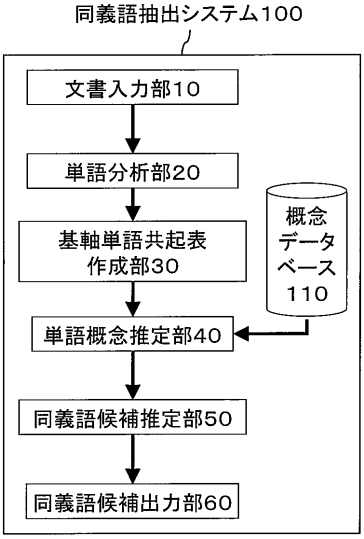
本発明によれば、ソフトウェアやシステムの開発における要件定義などの作業においてやり取りされる各種文書に関して、文書の曖昧さを除外することで文書の理解・作成・修正を支援することが可能になり、手戻りの減少や顧客満足の向上などシステム開発の効率化に関する用途に適用できる。また、同義語を精度よく抽出できるので、翻訳システムに用いて訳し分けに利用できる。

【符号の説明】

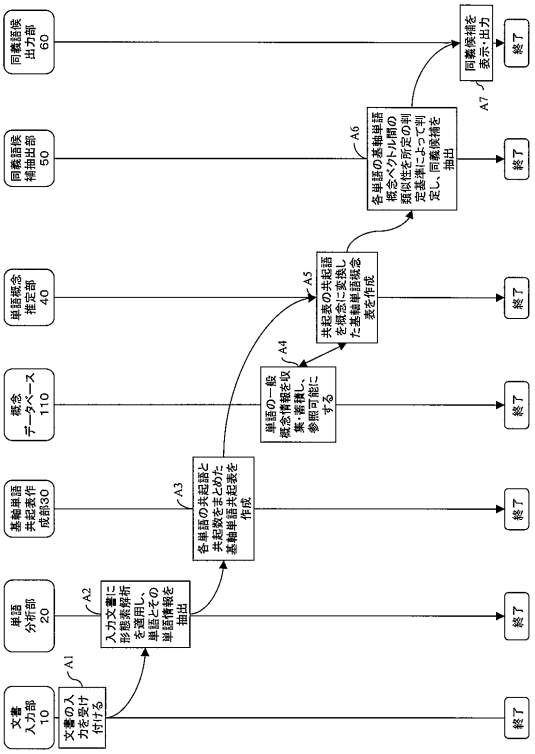
【０１０１】

１０　文書入力部
２０、２０Ａ　単語分析部
３０　基軸単語共起表作成部
３５　構成語支配度算出部
３６　複合語概念配分推定部
４０、４０Ａ　単語概念推定部
５０　同義語候補推定部
６０　同義語候補出力部
１００、１００Ａ　同義語抽出システム
１１０　概念データベース
Ｄ　文書
Ｙ、Ｙａ　文書解析システム
Ｚ、Ｚ’　インターネット・サーバ

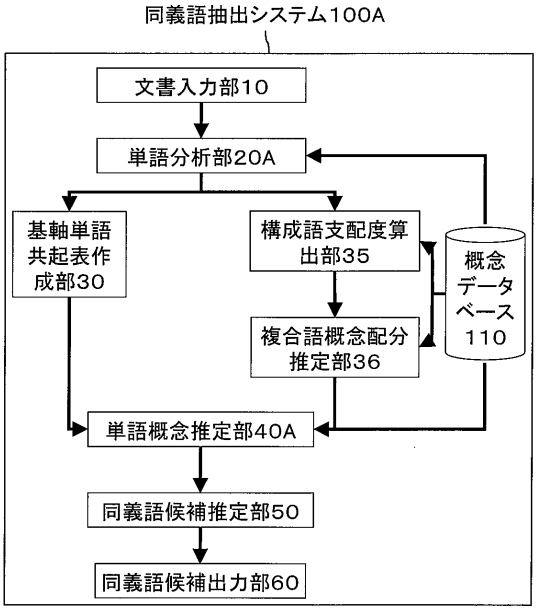
【図１】


同義語抽出システム100
文書入力部10
単語分析部20
基軸単語共起表作成部30
概念データベース110
単語概念推定部40
同義語候補推定部50
同義語候補出力部60

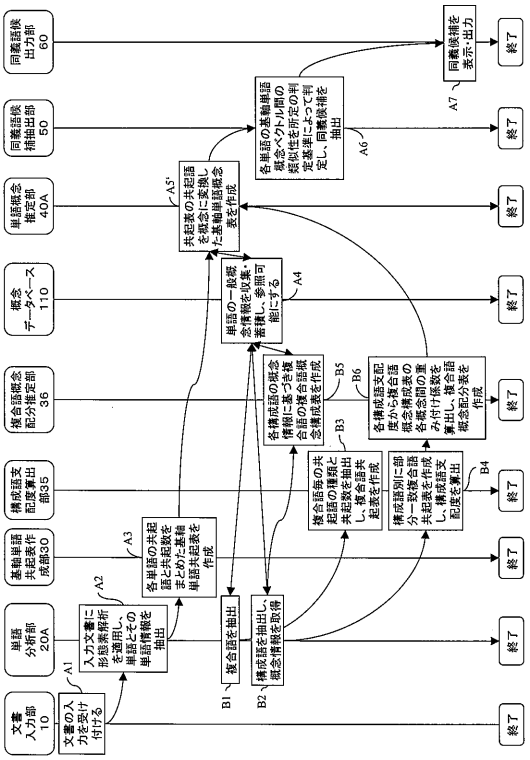
【図２】


文書入力部10
単語分析部20
基軸単語共起表作成部30
概念データベース110
単語概念推定部40
同義語候補抽出部50
同義語候出力部60
終了

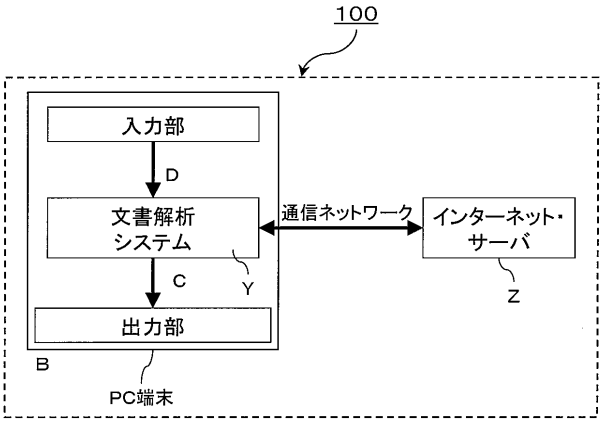
【図３】


同義語抽出システム100A
文書入力部10
単語分析部20A
基軸単語共起表作成部30
構成語支配度算出部35
概念データベース110
複合語概念配分推定部36
単語概念推定部40A
同義語候補推定部50
同義語候補出力部60

【図４】


文書入力部10
単語分析部20A
基軸単語共起表作成部30
構成語支配度算出部35
複合語概念配分推定部36
概念データベース110
単語概念推定部40A
同義語候補抽出部50
同義語候出力部60
終了

【図５】



【図６】

基軸単語S ＼ 共起語V	利用	操作	構築	改善	システム変更	メカニズム	瞬時	短期	稼働	高速処理	…
演算システム	0	3	2	0	4	0	1	0	3	0	…
分析機能	2	1	1	3	0	2	0	2	0	4	…
… … …	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	… … …

【図７】

共起語概念Cvj ＼ 共起語V	利用	操作	構築	改善	システム変更	メカニズム	瞬時	短期	稼働	高速処理	…
大分類(C1vj)	人間活動	人間活動	人間活動	人間活動	システム変更	抽象	抽象	抽象	人間活動	高速処理	…
中分類(C2vj)	仕事	仕事	作業	作業	システム変更	装置	時間	時間	動き	高速処理	…
小分類(C3vj)	使用	使用	作成	修正	システム変更	機構	短い	短い	動作	高速処理	…

【図８】

基軸単語S ＼ 共起語概念C1vj	人間活動（利用、操作、構築、改善、稼働を合流）	システム変更	抽象（メカニズム、瞬時、短期を合流）	高速処理	…
演算システム	8 =0+3+2+0+3	4	1 =0+1+0	0	…
分析機能	7 =2+1+1+3+0	0	4 =2+0+2	4	…
… … …	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	… … …

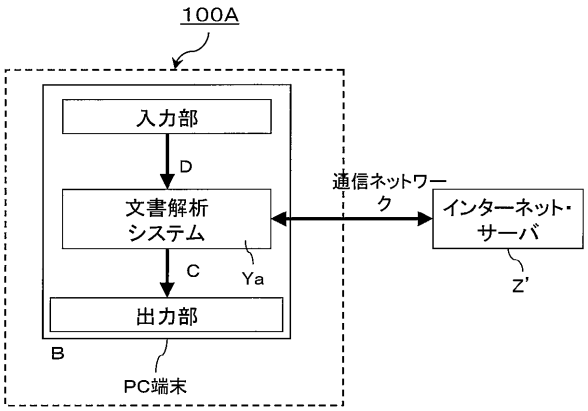
【図９】

基軸単語S ＼ 共起語概念C2vj	仕事（利用と操作を合流）	作業（構築と改善を合流）	システム変更	装置	時間（瞬時と短期を合流）	動き	高速処理	…
演算システム	3 =0+3	2 =2+0	4	0	1 =1+0	3	0	…
分析機能	3 =2+1	4 =1+3	0	2	2 =0+2	0	4	…
… … …	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	… … …

【図１０】

基軸単語S ＼ 共起語概念C3vj	使用（利用と操作を合流）	作成	修正	システム変更	機構	短い（瞬時と短期を合流）	動作	高速処理	…
演算システム	3 =0+3	2	0	4	0	1 =1+0	3	0	…
分析機能	3 =2+1	1	3	0	2	2 =0+2	0	4	…
… … …	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	… … …

【図１１】



【図１２】

複合語共起語 / 複合語	接続	予定	基盤	利用	更改	分離	ユーザ	ネットワーク	セグメント	評価	変更	実施	⋮
基準システム	1	1	3	1	2	2	1						・
情報システム			1										・
集積システム	1						1	3	1				・
評価システム										1	1		・
システム変更		1										1	・
・・・	・	・	・	・	・	・	・	・	・	・	・	・	・

【図１３】

複合語共起語 / 複合語	管理	伴う	設定	予定	検討	実施	・・・
変更内容	1	1		1	1	1	・
変更理由				1			・
システム変更				1		1	・
追加変更		1				1	・
変更禁止期間			1	1			・
組織変更	1	1				1	・
変更依頼					1	1	・
・・・	・	・	・	・	・	・	・

【図１４】

構成語P / 共起語概Caek	システム	変更
大分類(Ca1ek)	抽象	人間活動
中分類(Ca2ek)	装置	作業
小分類(Ca3ek)	装置	修正
概念重み付け係数αek	0.40＝1.47/(1.47+2.21)	0.60＝2.21/(1.47+2.21)

【図１５】

共起語概念C1vj / 基軸単語S	人間活動 (利用、操作、構築、改善、変更、稼働、処理を合流)	抽象 (システム、メカニズム、瞬時、短期、高速を合流)	・・・
演算システム	9.6 =0+3+2+0+2.4+3	2.6 =0+1.6+0+1+0	・・・
分析機能	9.8 =2+1+1+3+0+2.8	5.2 =2+0+2+1.2	・・・
・・・ ・・・ ・・・	・ ・ ・	・ ・ ・	・・・ ・・・ ・・・

【図１６】

共起語概念C2vj / 基軸単語S	使用 (利用と操作を合流)	修正 (「構築」と「改善」と構成語「変更」を合流)	装置 (「メカニズム」と構成語「システム」を合流)	短い (瞬時と短期を合流)	動作 (「稼働」と構成語「処理」を合流)	速さ (構成語「高速」)	・・・
演算システム	3 =0+3	4.4 =2+0+4×0.6	1.6 =4×0.4	1 =1+0	3	0	・・・
分析機能	3 =2+1	4 =1+3	2	2 =0+2	2.8 =4×0.7	1.2 =4×0.3	・・・
・・・ ・・・ ・・・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・・・ ・・・ ・・・

【図１７】

共起語概念C3vj / 基軸単語S	使用 (利用と操作を合流)	作成	修正 (「改善」と構成語「変更」を合流)	装置 (構成語「システム」)	機構	短い (瞬時と短期を合流)	動作 (「稼働」と構成語「処理」を合流)	速さ (構成語「高速」)	・・・
演算システム	3 =0+3	2	2.4 =0+4×0.6	1.6 =4×0.4	0	1 =1+0	3	0	・・・
分析機能	3 =2+1	1	3	0	2	2 =0+2	2.8 =4×0.7	1.2 =4×0.3	・・・
・・・ ・・・ ・・・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・・・ ・・・ ・・・

フロントページの続き

(72)発明者　古橋　武
愛知県名古屋市千種区不老町１番　国立大学法人名古屋大学内
(72)発明者　吉川　大弘
愛知県名古屋市千種区不老町１番　国立大学法人名古屋大学内

Ｆターム(参考) 5B091 AA15　AB11　AB17　CC04