(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5167546号
(P5167546)
(45)発行日　平成25年3月21日(2013.3.21)　(24)登録日　平成25年1月11日(2013.1.11)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 350C
G06F 17/30 170A

請求項の数 21　(全 66 頁)

(21)出願番号　特願2008-530812(P2008-530812)
(86)(22)出願日　平成19年3月16日(2007.3.16)
(86)国際出願番号　PCT/JP2007/055448
(87)国際公開番号　WO2008/023470
(87)国際公開日　平成20年2月28日(2008.2.28)
審査請求日　平成22年3月9日(2010.3.9)
(31)優先権主張番号　特願2006-224563(P2006-224563)
(32)優先日　平成18年8月21日(2006.8.21)
(33)優先権主張国　日本国(JP)

(73)特許権者　504132272
国立大学法人京都大学
京都府京都市左京区吉田本町３６番地１
(74)代理人　100078868
弁理士　河野　登夫
(74)代理人　100114557
弁理士　河野　英仁
(72)発明者　白松　俊
京都府京都市左京区吉田本町　京都大学大学院情報学研究科内
(72)発明者　駒谷　和範
京都府京都市左京区吉田本町　京都大学大学院情報学研究科内

最終頁に続く

(54)【発明の名称】文単位検索方法、文単位検索装置、コンピュータプログラム、記録媒体及び文書記憶装置

(57)【特許請求の範囲】
【請求項１】
自然言語からなる複数の文書データが記憶されている文書集合を用い、該文書集合から取得した文書データを一又は複数の文からなる文単位に分別しておく一方、言葉を順次受け付け、受け付けた言葉に基づいて前記文書集合から分別してある文単位を検索する文単位検索方法において、
文書データ中に連なる文単位夫々に、該文単位及び先行文脈に基づき求められる該文単位での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付けて予め記憶しておくステップと、
言葉を受け付ける都度、該言葉に、該言葉及び先行文脈に基づき求められる該言葉での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付けるステップと、
受け付けた言葉に対応付けた重み付き単語群と類似する重み付き単語群が対応付けて記録されている文単位を、前記文書集合から抽出する類似文単位抽出ステップと、
抽出した文単位を出力するステップと
を含むことを特徴とする文単位検索方法。
【請求項２】
前記類似文単位抽出ステップは、
受け付けた言葉に対応付けた重み付き単語群の内の複数の単語の重み値の分布と、予め分別された文単位に対応付けられている重み付き単語群の内の複数の単語の重み値の分布

とが、所定の条件を満たすか否かを判断するステップと、
所定の条件を満たすと判断された重み付き単語群が対応付けられている文単位を抽出するステップと
を含むことを特徴とする請求項１に記載の文単位検索方法。
【請求項３】
前記類似文単位抽出ステップは、
予め分別された文単位から、受け付けた言葉に対応付けた重み付き単語群と同一の単語を含む単語群が対応付けられた文単位を抽出するステップと、
受け付けた言葉と抽出した文単位とで、対応付けられた単語群の内の同一の単語毎に重み値の差分を算出するステップと、
抽出した文単位に、算出した差分が小さい順に優先順位を付与するステップと
を含み、
抽出した文単位を、優先順位に基づいて出力する
ことを特徴とする請求項１又は２に記載の文単位検索方法。
【請求項４】
前記重み付き単語群を、各単語を１次元とし、単語毎に付与される重み値の大きさを各単語に対応する次元方向の要素として持つ多次元ベクトルとして算出するステップを含み、
前記類似文単位抽出ステップは、
分別した文単位毎に記憶してある前記多次元ベクトルと、受け付けた言葉に対応付けた前記多次元ベクトルとの距離を算出するステップと、
文単位に、算出した距離が短い順に優先順位を付与するステップと
を含み、
付与された優先順位に従って出力する
ことを特徴とする請求項１又は２に記載の文単位検索方法。
【請求項５】
文単位又は受け付けた言葉に重み付き単語群を対応付ける際、
各単語が、前記文単位又は前記言葉よりも後続の文単位又は言葉に出現する又は参照される参照確率を算出する参照確率算出ステップを含み、
算出した参照確率を各単語の重み値として付与する
ことを特徴とする請求項１乃至４のいずれかに記載の文単位検索方法。
【請求項６】
前記参照確率算出ステップは、
前記各単語が先行の文単位を含む複数の文単位に出現するパターン、又は前記単語を先行の文単位から参照するパターンを含む特徴パターンを特定するステップと、
前記文書集合から取得された文書データ中で、前記特徴パターンと同一の特徴パターンが特定される単語が、後続の文単位で出現する又は参照される割合を算出するステップと
を含み、
算出した割合を参照確率とする
ことを特徴とする請求項５に記載の文単位検索方法。
【請求項７】
前記文書集合から抽出される単語毎に、該単語が先行の文単位を含む複数の文単位に出現するパターン、又は前記単語を先行の文単位から参照するパターンを含む特徴パターンを特定する特定ステップと、
特定した特徴パターンと同一の特徴パターンが特定される単語が、前記文書データ中で後続の文単位で出現したか又は参照されたかを判定する判定ステップと、
特定した特徴パターンと、該特徴パターンで特定される単語に対して判定した結果との回帰分析を行って前記参照確率に対する前記特徴パターンの回帰係数を算出する回帰ステップと
を含み、

文単位に重み付き単語群を対応付けて記憶しておく際、又は受け付けた言葉に重み付き単語群を対応付ける際、
前記参照確率算出ステップは、
前記文単位又は言葉毎に、該文単位又は言葉での単語の特徴パターンを特定し、
特定した特徴パターンに対する前記回帰係数を使用して参照確率を算出する
ことを特徴とする請求項５に記載の文単位検索方法。
【請求項８】
文単位に対しては、書き言葉からなる第１文書集合から取得された文書データ中で前記割合を算出し、
受け付けた言葉に対しては、話し言葉からなる第２文書集合から取得された文書データ中で前記割合を算出する
ことを特徴とする請求項６に記載の文単位検索方法。
【請求項９】
書き言葉からなる第１文書集合及び話し言葉からなる第２文書集合夫々について、
前記特定ステップ、前記判定ステップ及び前記回帰ステップを実行しておき、
前記参照確率算出ステップは、
前記文単位で特定した単語の特徴パターンに対しては、第１文書集合について実行した前記回帰ステップにより算出された回帰係数を使用して参照確率を算出し、
前記受け付けた言葉で特定した単語の特徴パターンに対しては、第２文書集合について実行した前記回帰ステップで算出された回帰係数を使用して参照確率を算出する
ことを特徴とする請求項７に記載の文単位検索方法。
【請求項１０】
前記特徴パターンは、
前記単語を先行の文単位又は言葉から参照している場合の前記先行の文単位又は言葉から前記単語が含まれる文単位又は言葉までの、文単位又は言葉の数、
前記単語が出現又は参照されている直近の先行の文単位又は言葉における前記単語の係り受け情報、
前記単語が含まれる文単位又は言葉までに出現した又は参照された回数、
前記単語が出現又は参照されている直近の先行の文単位又は言葉における前記単語の名詞区別、
前記単語が出現又は参照されている直近の先行の文単位又は言葉中で前記単語が主題であるか否か、
前記単語が出現又は参照されている直近の先行の文単位又は言葉中で前記単語が主語であるか否か、
前記単語が含まれる文単位又は言葉における人称、
及び、
前記単語が含まれる文単位又は言葉における品詞情報、
の内の一又は複数を含む情報で特定される
ことを特徴とする請求項６乃至９のいずれかに記載の文単位検索方法。
【請求項１１】
前記特徴パターンは、
前記単語を先行の文単位又は言葉から参照している場合の前記先行の文単位又は言葉から前記単語が含まれる文単位又は言葉までに対応する時間、
前記単語が出現又は参照されている直近の先行の文単位又は言葉中で前記単語に対応する発話速度、
及び、
前記単語が出現又は参照されている直近の先行の文単位又は言葉中で前記単語に対応する音声の周波数
の内の一又は複数を含む情報で特定される
ことを特徴とする請求項６乃至１０のいずれかに記載の文単位検索方法。

【請求項１２】
前記文章集合から抽出される単語の内の一の単語について、
前記分別された文単位に対応付けられている重み付き単語群の内から、前記一の単語が含まれる単語群であり、且つ前記一の単語の重み値が所定値以上である単語群を抽出する第１ステップと、
該第１ステップで抽出した単語群の各単語の重み値を単語毎に統合した値を、前記一の単語の各単語への関連度として付与した関連単語群を作成する第２ステップと、
作成した関連単語群を前記一の単語に対応付けて記憶する第３ステップと、
前記抽出された単語夫々について前記第１ステップ乃至第３ステップを予め実行するステップと、
文単位毎又は受け付けた言葉毎に対応付けられた重み付き単語群の各単語の重み値夫々を、各単語に対応付けて記憶されている前記関連単語群の各単語の関連度を使用して付与し直す関連度付加ステップと
を含むことを特徴とする請求項１乃至１１のいずれかに記載の文単位検索方法。
【請求項１３】
前記第２ステップは、
前記抽出した単語群について、各単語群に含まれる各単語の重み値に、前記一の単語の重み値で重み付けした総和を算出するステップと、
算出した総和を平均化するステップと、
作成する関連単語群の各単語の前記関連度として、各単語の重み値の平均化された総和を付与するステップと
を含むことを特徴とする請求項１２に記載の文単位検索方法。
【請求項１４】
前記関連度付加ステップは、
文単位毎又は受け付けた言葉毎に対応付けられた重み付き単語群の各単語について、
各単語に対応付けて記憶されている前記関連単語群に含まれる各単語の関連度を、前記重み付き単語群の各単語の重み値に乗算するステップと、
乗算結果に基づいて前記重み付き単語群の各単語の重み値として付与し直すステップと
を含むことを特徴とする請求項１２又は１３に記載の文単位検索方法。
【請求項１５】
各単語夫々についての前記関連単語群を、各単語を１次元とし、単語毎に付与される関連度の大きさを各単語に対応する次元方向の要素として持つ多次元の関連度ベクトルとして算出するステップと
を含み、
前記関連度付加ステップは、
分別した文単位毎に記憶してある前記多次元ベクトルを、各単語の関連度ベクトルの列によって変換する
ことを特徴とする請求項１２乃至１４のいずれかに記載の文単位検索方法。
【請求項１６】
自然言語からなる複数の文書データが記憶されている文書集合を用い、言葉を受け付け、受け付けた言葉に基づいて前記文書集合を検索する文単位検索方法において、
前記文書集合から得られる文書データを一又は複数の文からなる文単位に分別しておくステップ、
分別した文単位毎に、該文単位に出現する単語、又は、文書データ中の先行の文単位から参照する単語を抽出するステップ、
前記文単位に対して抽出した単語毎に、各文単位における特徴を特定して記憶しておくステップ、
分別した文単位毎に、該文単位に対して抽出した単語が該文単位及び先行の文単位で出現する場合の前記特徴の組み合わせのパターン、又は先行の文単位から参照する場合の参照のパターンを含む特徴パターンを特定するステップ、

特定した特徴パターンと、該特徴パターンで特定された単語が後続の文単位で出現又は参照されたか否かとを記憶しておくステップ、
前記文書集合から得られる文書中の文単位全体に対し、一の特徴パターンで特定される単語が後続の文単位で出現又は参照される参照確率の回帰分析を行って特徴パターンに対応する回帰係数を得る回帰学習を実行するステップ、
分別した文単位毎に、
文書データ中で先行の文単位から各文単位に至るまでに抽出された各単語について、前記文単位で特定される特徴パターンに対応する前記回帰係数を使用し、前記単語の前記参照確率を算出するステップ、
算出した参照確率を夫々付与した重み付き単語群を対応付けて予め記憶しておくステップ、
言葉を受け付けた場合、受け付けた順に言葉を記憶するステップ、
言葉を受け付けた場合、
受け付けた言葉に出現する単語又は前記言葉よりも先に受け付けた言葉から参照する単語を抽出するステップ、
抽出した各単語の前記受け付けた言葉における特徴を特定するステップ、
先に受け付けた言葉で出現する場合の特徴の組み合わせのパターン、又は先に受け付けた言葉から参照する場合の参照のパターンを含む特徴パターンを特定するステップ、
特定された特徴パターンに対応する前記回帰係数を使用して、前記単語の前記参照確率を算出するステップ、
算出した参照確率を夫々付与した重み付き単語群を前記言葉に対応付けるステップ、
前記受け付けた言葉と、予め分別されてある文単位とで、対応付けられている重み付き単語群の内の同一の単語毎に付与されている参照確率の差分を算出するステップ、
予め分別されてある文単位に、前記参照確率の差分が小さい順に優先順位を付与するステップ、及び、
前記文単位を付与された優先順位に基づいて出力するステップ
を含むことを特徴とする文単位検索方法。

【請求項１７】
自然言語からなる複数の文書データが記憶されている文書集合から文書データを取得する手段と、言葉を順次受け付ける手段とを備え、受け付けた言葉に基づいて前記文書集合を検索する文単位検索装置において、
取得した文書データを一又は複数の文からなる文単位に分別する手段と、
取得した文書データ中に連なる文単位夫々に、該文単位及び先行文脈に基づき求められる該文単位での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付けて記憶する手段と、
言葉を受け付けた場合に受け付けた順に記憶する手段と、
新たに言葉を受け付ける都度、該言葉に、該言葉及び該先行文脈に基づき求められる該言葉での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付ける手段と、
予め分別された文単位から、受け付けた言葉に対応付けた重み付き単語群と類似する重み付き単語群が対応付けて記録されている文単位を抽出する手段と、
抽出した文単位を出力する手段と
を備えることを特徴とする文単位検索装置。

【請求項１８】
自然言語からなる複数の文書データが記憶されている文書集合から、文書データを取得することが可能であるコンピュータを、言葉を順次受け付ける手段と、受け付けた言葉に基づいて前記文書集合を検索する手段として機能させることができるコンピュータプログラムにおいて、
取得した文書データを一又は複数の文からなる文単位に分別する手段、
取得した文書データ中に連なる文単位夫々に、該文単位及び先行文脈に基づき求められ

る該文単位での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付けて記憶する手段、
言葉を受け付けた場合に受け付けた順に記憶する手段、
新たに言葉を受け付ける都度、該言葉に、該言葉及び先行文脈に基づき求められる該言葉での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付ける手段、及び、
予め分別された文単位から、受け付けた言葉に対応付けた重み付き単語群と類似する重み付き単語群が対応付けて記録されている文単位を抽出する手段
として機能させることを特徴とするコンピュータプログラム。
【請求項１９】
請求項１８に記載のコンピュータプログラムを記録した、コンピュータで読み取り可能な記録媒体。
【請求項２０】
自然言語からなる複数の文書データを記憶する手段と、記憶した文書データを、文書データの先頭から順に一又は複数の文からなる文単位に分別する手段とを備え、分別した文単位毎に、該文単位に出現する単語又は先行する文単位から参照する単語が抽出してあり、分別した文単位毎に抽出した単語が記憶してある文書記憶装置において、
文書データ中に連なる文単位毎に、複数の単語が、該文単位よりも後続の文単位に出現するか又は参照される参照確率を算出する手段と、
前記文単位夫々に、該文単位での顕現性を表わす重み値として前記参照確率が付与された前記複数の単語からなる重み付き単語群を対応付けて記憶する手段と
を備えることを特徴とする文書記憶装置。
【請求項２１】
抽出されてある単語の内の一の単語について、
文単位夫々に対応付けられている重み付き単語群の内から、前記一の単語が含まれる単語群であり、且つ前記一の単語の重み値が所定値以上である単語群を抽出する抽出手段と、
該抽出手段が抽出した単語群の各単語の重み値を単語毎に統合した値を、前記一の単語の各単語への関連度として付与した関連単語群を作成する作成手段と、
作成した関連単語群を前記一の単語に対応付けて記憶する記憶手段と
を備え、
前記抽出されてある単語夫々について前記抽出手段、前記作成手段及び前記記憶手段の処理を実行するようにしてあり、各単語に対応付けて夫々の関連単語群を記憶するようにしてあること
を特徴とする請求項２０に記載の文書記憶装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、検索のためにユーザから受け付けたテキスト、音声等の言葉に基づいて、多数の文書データ記憶されている文書集合からの検索を行う検索方法に関する。特に、文脈の流れの中で意味が動的に変化する文書中の意味のまとまりの単位である文単位から、受け付けた言葉と意味合いが類似する文単位を直接的に検索することができる文単位検索方法、文単位検索装置、コンピュータを前記文単位検索装置として機能させるコンピュータプログラム、該コンピュータプログラムを記録したコンピュータ読み取り可能な記録媒体、及び文書記憶装置に関する。
【背景技術】
【０００２】
インターネット上で提供される各種サービスには、ユーザによって入力されたキーワード又は文に基づいて、インターネットで公開されている文書から関連する文書を検索し、一覧にして出力する文書検索サービスがある。

【０００３】
従来の文書検索サービスには、以下のようなものがある。インターネットで公開されている文書を自動的に集めて記憶し、夫々の文書毎に、文書中に出現する単語を文書中での出現確率と共に記憶しておき、キーワード又は文等の言葉を受け付けた場合に、記憶した文書集合から受け付けたキーワード又は文に含まれる単語の出現確率の高い順に優先順位を付与して文書を抽出し、抽出した文書から、当該単語が含まれる文又は段落を出力する。

【０００４】
文書検索サービスを利用するユーザは、知りたい情報を検索するために関連するキーワードを自分で考える必要がある。最近の文書検索サービスでは、自然文を入力文として受け付け、入力文を形態素解析し、入力文のキーワードを識別して検索要求を自動的に作成することができる場合もある。

【０００５】
また、文書検索サービスでは通常、自然文の入力を受け付ける場合でも、入力文に含まれる単語を抽出し、抽出した単語が含まれている文書を検索結果として出力する。したがって、ユーザは、目的の検索結果を得るために入力するキーワードに関連するキーワード又は入力するキーワードの意味付けが変化する単語を更に入力して絞込みをさせる必要があった。例えば、単に「大統領」では、どの国の大統領なのかは不明であるため、「大統領、アメリカ」とキーワードを付加する必要がある。更にアメリカの大統領の何を調べたいかによって、「大統領、アメリカ、出身」、「大統領、アメリカ、政策」等、検索結果を得やすくするための情報を考える必要がある。

【０００６】
したがって、ユーザが得たいと考える検索結果を実際に得るためには、ユーザはキーワードの組み合わせを考え、何回か試行することが必要になる。例えば、ユーザが「アメリカの大統領は、他の国との間で経済面の問題が発生した場合どのような対策をとるのか」という情報を知りたい場合であっても、「アメリカ、大統領、経済」では検索結果が大量に出力され、大量に出力された検索結果からユーザは文書を選択しなければならない。そこで例えば、「政策」というキーワードを付加して絞込み、「アメリカ、大統領、経済、政策」というキーワードを入力する。この場合、「政策」という言葉が意味の広い上位概念であっても、「政策」というキーワード自体で絞込みをすることになるため、内容としては経済政策についての論述が記載された文書も、「政策」という言葉の出現頻度が低い文書は漏れてしまうことがある。このように、ユーザが検索の目的を達するためのキーワードを考えて試行することで検索結果を得るのは難しい。付加的な情報を入力する度に、本来の検索の目的から、検索結果の内容が離れていく場合もある。

【０００７】
また、上述の例でユーザが知りたいのは、経済面での政策であって、しかも国際的な政策についてである。ユーザの入力が自然文によるものであっても、「アメリカ、大統領、他の国、経済、問題、発生、場合、対策」の単語の何れの単語が一番重要であるのかは、人間が読む場合は把握できるが、装置又はコンピュータが扱う情報量として定量的に表現することは難しい。したがって、キーワードは全て含んでいるものの、「アメリカの経済の問題と他国の大統領の対策」とについて論述された文書が出力されることも想定できる。

【０００８】
さらに、検索対象である文書が非常に長い場合は、その文書の中で文脈が動的に変化しているにも拘わらず、その文書を一単位として出現する単語に基づいた検索がされる。したがって、アメリカの大統領の歴史と、他の国の大統領の歴史と、各国の経済のしくみと、各国での失業対策についての内容とが章に分けられて記載されている文書が存在する場合、検索のキーワードをほとんど含むために検索結果として出力される。実際にはそれらの章が文脈的に繋がっていない場合でも、キーワードを含む文又は段落を夫々部分的に抽出した結果が出力されてしまう。そのため、その抽出された部分に至るまでの先行文脈の

影響を含む意味と、ユーザの意識の上での検索意図とが、意味的にマッチするか否かは量り得ない。

【０００９】

一方、検索対象である文書に、検索のために入力したキーワードは頻繁に出現してはいないにも拘わらず、入力したキーワードが文脈上重要な意味を持って含まれている場合がある。例えば、主題となる単語ほど指示代名詞又はゼロ代名詞で表現される。したがって、知りたい情報を検索するユーザは、検索のために入力したキーワードが指示代名詞又はゼロ代名詞で表現されている文又は段落こそ、検索結果として得たい情報である場合が考えられる。しかしながら、実際の出現頻度で検索結果に優先順位を付与する場合、ユーザが入力したキーワードの出現頻度が低いために絞込みによって候補から除かれ、検索結果として出力されない。

【００１０】

そこで、文書中の単語を抽出し、当該単語の品詞情報、単語間の係り受け情報、更に指示代名詞又はゼロ代名詞と照応関係にある単語を明示した情報を、文書を形態素解析等により解析した結果に付加して記憶させておき、記憶させた情報に基づいて装置又はコンピュータによる文書の検索、質問応答、機械翻訳を実現する技術が提案されている（非特許文献１）。

【００１１】

単語間の係り受け又は照応等の関係は、自然文であるがために文節の順序が複雑であり、人間が読む場合は意味を判別できても機械的に認識することが難しい。そこで、非特許文献１に記載されている技術では、単語間の係り受け又は照応等の関係をタグによって文又は句毎の情報として文書データに付加して記憶しておく。また、日本語の場合は特に、主語が省略されている文が多いので、機械的に翻訳する際に主語の補完が必要である。そこで非特許文献１に記載されている技術では、文毎に主語又はゼロ代名詞等の補完情報を付加する。これにより、当該情報が付加された文書を利用することによって正確に機械翻訳することが可能となる。文中で省略された単語、又は指示代名詞若しくはゼロ代名詞で表されている単語も、例えば文書を検索する場合の出現頻度の算出等の応用技術に利用することができる。

【非特許文献１】橋田浩一「大域文書修飾」人工知能学会全国大会（第１１回）論文集ｐｐ．６２－６３（１９９７）

【発明の開示】

【発明が解決しようとする課題】

【００１２】

文章を書く時、又は発話する時の、その各文又は各発話夫々におけるユーザの注目対象（重点対象）は、会話や文章の文脈の流れに従って動的に変化する。つまり、会話や文章における単語への注目度合いを表す重みは、動的に変化する。よって、会話や文章に関連する情報を検索するサービスを実現するためには、文脈に応じた単語の重みの動的変化を追跡する必要がある。

【００１３】

しかしながら、従来の文書検索サービスでは、検索のために入力された単語の出現頻度の高い文書を抽出し、抽出した文書から、当該単語を含む文又は段落を抽出して出力するため、当該単語のその文又は段落の文脈で動的に変わる重みについては考慮されずに検索される。したがって、出現頻度に基づく検索では、確かに検索のために入力された単語を含んではいるものの、文脈上当該単語がユーザが考えるように使用されていない場合があり、ユーザの検索目的を達成することができるとは限らない。各単語の文脈上の意味における各文での重み、即ち文脈上注目されているか否かについては特定できない。したがって、入力したキーワードをユーザの考える意味合い通りに使用した文又は段落を出力することはできない。

【００１４】

また、非特許文献１の技術では、品詞情報等の文法に照らして識別が可能な情報を自動

的に解析し、指示代名詞又はゼロ代名詞等の補完、照応又は係り受けについての情報を文書に付加することができる。当該情報の付加により、参照されている名詞を出現頻度として利用することができるので、文又は段落等での単語間の関係は付加された情報により解析が可能である。しかしながら、各単語の文又は段落での注目されている度合い、即ち顕現性は、定量的に測ることはできない。

【００１５】

非特許文献１の技術は、自然文による質問に対して当該質問文で省略されている単語等を考慮してコンピュータに応答させる質問応答の実現へ応用が可能である。しかし、複数のユーザによる対話の文脈上の意味を定量的な値として算出し、第三者の発話としてユーザの対話の文脈に沿った発話を生成し、提示することを可能にするのは容易でない。

【００１６】

また、従来の文書検索サービスでは、文書中に出現する頻度が少ない場合でも文脈上深く関連する背景知識を表わすような単語を考慮して検索することはできなかった。したがって、検索するユーザが意識しているが検索のために入力された単語としては現れていない単語を、同様に連想させる文又は段落を直接的に出力することはできなかった。

【００１７】

本発明は斯かる事情に鑑みてなされたものであり、一又は複数の文からなる文単位毎に、その文単位での単語の顕現性を表わす重み値が夫々付与された重み付き単語群を対応付けて記憶しておき、検索のために受け付けた言葉についても、その言葉での重み値が付与された重み付き単語群を対応付け、重み付き単語群が類似する文単位を抽出して出力する構成とする。受け付けた言葉から、ユーザの意識にある先の言葉からの文脈が反映された意味を表わす情報を自動的に生成し、文脈の流れの中で意味が動的に変化する文書中の文単位の内から、受け付けた言葉から生成された情報が表わす文脈上の意味のまとまりが類似する文単位を直接的に検索することができる文単位検索方法、文単位検索装置、コンピュータを前記文単位検索装置として機能させるコンピュータプログラム、及び該コンピュータプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを目的とする。

【００１８】

本発明の目的は、文単位又は受け付ける言葉に対応付けられる重み付き単語群中の各単語の顕現性を表わす重み値を、後続の文単位又は言葉で出現する確率又は参照される確率として算出することにより、文脈の流れの中にある文単位又は言葉夫々で時系列に変化する単語の顕現性を定量的に表わして用いることができる文単位検索方法及び文書記憶装置を提供することにある。

【００１９】

また、本発明の目的は、関連する単語への関連度を定量的に算出し、各文単位又は言葉における各単語の顕現性に関連度を反映させることにより、ユーザから発せされる言葉又は筆記された文章には出現していない場合でも、ユーザが言葉を発しているとき又は筆記しているときに意識している単語を連想させる文単位をも効果的に検索することができる文単位検索方法及び文書記憶装置を提供することにある。

【課題を解決するための手段】

【００２０】

第１発明に係る文単位検索方法は、自然言語からなる複数の文書データが記憶されている文書集合を用い、該文書集合から取得した文書データを一又は複数の文からなる文単位に分別しておく一方、言葉を順次受け付け、受け付けた言葉に基づいて前記文書集合から分別してある文単位を検索する文単位検索方法において、文書データ中に連なる文単位夫々に、該文単位及び先行文脈に基づき求められる該文単位での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付けて予め記憶しておくステップと、言葉を受け付ける都度、該言葉に、該言葉及び先行文脈に基づき求められる該言葉での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付けるステップと、受け付けた言葉に対応付けた重み付き単語群と類似する重み付き単語群が対応付け

て記録されている文単位を、前記文書集合から抽出する類似文単位抽出ステップと、抽出した文単位を出力するステップとを含むことを特徴とする。

【００２１】

第２発明に係る文単位検索方法は、前記類似文単位抽出ステップは、受け付けた言葉に対応付けた重み付き単語群の内の複数の単語の重み値の分布と、予め分別された文単位に対応付けられている重み付き単語群の内の複数の単語の重み値の分布とが、所定の条件を満たすか否かを判断するステップと、所定の条件を満たすと判断された重み付き単語群が対応付けられている文単位を抽出するステップとを含むことを特徴とする。

【００２２】

第３発明に係る文単位検索方法は、前記類似文単位抽出ステップは、予め分別された文単位から、受け付けた言葉に対応付けた重み付き単語群と同一の単語を含む単語群が対応付けられた文単位を抽出するステップと、受け付けた言葉と抽出した文単位とで、対応付けられた単語群の内の同一の単語毎に重み値の差分を算出するステップと、抽出した文単位に、算出した差分が小さい順に優先順位を付与するステップとを含み、抽出した文単位を、優先順位に基づいて出力することを特徴とする。

【００２３】

第４発明に係る文単位検索方法は、前記重み付き単語群を、各単語を１次元とし、単語毎に付与される重み値の大きさを各単語に対応する次元方向の要素として持つ多次元ベクトルとして算出するステップを含み、前記類似文単位抽出ステップは、分別した文単位毎に記憶してある前記多次元ベクトルと、受け付けた言葉に対応付けた前記多次元ベクトルとの距離を算出するステップと、文単位に、算出した距離が短い順に優先順位を付与するステップとを含み、付与された優先順位に従って出力することを特徴とする。

【００２４】

第５発明に係る文単位検索方法は、文単位又は受け付けた言葉に重み付き単語群を対応付ける際、各単語が、前記文単位又は前記言葉よりも後続の文単位又は言葉に出現する又は参照される参照確率を算出する参照確率算出ステップを含み、算出した参照確率を各単語の重み値として付与することを特徴とする。

【００２５】

第６発明に係る文単位検索方法は、前記参照確率算出ステップは、前記各単語が先行の文単位を含む複数の文単位に出現するパターン、又は前記単語を先行の文単位から参照するパターンを含む特徴パターンを特定するステップと、前記文書集合から取得された文書データ中で、前記特徴パターンと同一の特徴パターンが特定される単語が、後続の文単位で出現する又は参照される割合を算出するステップとを含み、算出した割合を参照確率とすることを特徴とする。

【００２６】

第７発明に係る文単位検索方法は、前記文書集合から抽出される単語毎に、該単語が先行の文単位を含む複数の文単位に出現するパターン、又は前記単語を先行の文単位から参照するパターンを含む特徴パターンを特定する特定ステップと、特定した特徴パターンと同一の特徴パターンが特定される単語が、前記文書データ中で後続の文単位で出現したか又は参照されたかを判定する判定ステップと、特定した特徴パターンと、該特徴パターンで特定される単語に対して判定した結果との回帰分析を行って前記参照確率に対する前記特徴パターンの回帰係数を算出する回帰ステップとを含み、文単位に重み付き単語群を対応付けて記憶しておく際、又は受け付けた言葉に重み付き単語群を対応付ける際、前記参照確率算出ステップは、前記文単位又は言葉毎に、該文単位又は言葉での単語の特徴パターンを特定し、特定した特徴パターンに対する前記回帰係数を使用して参照確率を算出することを特徴とする。

【００２７】

第８発明に係る文単位検索方法は、文単位に対しては、書き言葉からなる第１文書集合から取得された文書データ中で前記割合を算出し、受け付けた言葉に対しては、話し言葉からなる第２文書集合から取得された文書データ中で前記割合を算出することを特徴とす

る。

【００２８】

第９発明に係る文単位検索方法は、書き言葉からなる第１文書集合及び話し言葉からなる第２文書集合夫々について、前記特定ステップ、前記判定ステップ及び前記回帰ステップを実行しておき、前記参照確率算出ステップは、前記文単位で特定した単語の特徴パターンに対しては、第１文書集合について実行した前記回帰ステップにより算出された回帰係数を使用して参照確率を算出し、前記受け付けた言葉で特定した単語の特徴パターンに対しては、第２文書集合について実行した前記回帰ステップで算出された回帰係数を使用して参照確率を算出することを特徴とする。

【００２９】

第１０発明に係る文単位検索方法は、前記特徴パターンは、前記単語を先行の文単位又は言葉から参照している場合の前記先行の文単位又は言葉から前記単語が含まれる文単位又は言葉までの、文単位又は言葉の数、前記単語が出現又は参照されている直近の先行の文単位又は言葉における前記単語の係り受け情報、前記単語が含まれる文単位又は言葉までに出現した又は参照された回数、前記単語が出現又は参照されている直近の先行の文単位又は言葉における前記単語の名詞区別、前記単語が出現又は参照されている直近の先行の文単位又は言葉中で前記単語が主題であるか否か、前記単語が出現又は参照されている直近の先行の文単位又は言葉中で前記単語が主語であるか否か、前記単語が含まれる文単位又は言葉における人称、及び、前記単語が含まれる文単位又は言葉における品詞情報、の内の一又は複数を含む情報で特定されることを特徴とする。

【００３０】

第１１発明に係る文単位検索方法は、前記特徴パターンは、前記単語を先行の文単位又は言葉から参照している場合の前記先行の文単位又は言葉から前記単語が含まれる文単位又は言葉までに対応する時間、前記単語が出現又は参照されている直近の先行の文単位又は言葉中で前記単語に対応する発話速度、及び、前記単語が出現又は参照されている直近の先行の文単位又は言葉中で前記単語に対応する音声の周波数の内の一又は複数を含む情報で特定されることを特徴とする。

【００３１】

第１２発明に係る文単位検索方法は、前記文章集合から抽出される単語の内の一の単語について、前記分別された文単位に対応付けられている重み付き単語群の内から、前記一の単語が含まれる単語群であり、且つ前記一の単語の重み値が所定値以上である単語群を抽出する第１ステップと、該第１ステップで抽出した単語群の各単語の重み値を単語毎に統合した値を、前記一の単語の各単語への関連度として付与した関連単語群を作成する第２ステップと、作成した関連単語群を前記一の単語に対応付けて記憶する第３ステップと、前記抽出された単語夫々について前記第１ステップ乃至第３ステップを予め実行するステップと、文単位毎又は受け付けた言葉毎に対応付けられた重み付き単語群の各単語の重み値夫々を、各単語に対応付けて記憶されている前記関連単語群の各単語の関連度を使用して付与し直す関連度付加ステップとを含むことを特徴とする。

【００３２】

第１３発明に係る文単位検索方法は、前記第２ステップは、前記抽出した単語群について、各単語群に含まれる各単語の重み値に、前記一の単語の重み値で重み付けした総和を算出するステップと、算出した総和を平均化するステップと、作成する関連単語群の各単語の前記関連度として、各単語の重み値の平均化された総和を付与するステップとを含むことを特徴とする。

【００３３】

第１４発明に係る文単位検索方法は、前記関連度付加ステップは、文単位毎又は受け付けた言葉毎に対応付けられた重み付き単語群の各単語について、各単語に対応付けて記憶されている前記関連単語群に含まれる各単語の関連度を、前記重み付き単語群の各単語の重み値に乗算するステップと、乗算結果に基づいて前記重み付き単語群の各単語の重み値として付与し直すステップとを含むことを特徴とする。

【００３４】
第１５発明に係る文単位検索方法は、各単語夫々についての前記関連単語群を、各単語を１次元とし、単語毎に付与される関連度の大きさを各単語に対応する次元方向の要素として持つ多次元の関連度ベクトルとして算出するステップとを含み、前記関連度付加ステップは、分別した文単位毎に記憶してある前記多次元ベクトルを、各単語の関連度ベクトルの列によって変換することを特徴とする。
【００３５】
第１６発明に係る文単位検索方法は、自然言語からなる複数の文書データが記憶されている文書集合を用い、言葉を受け付け、受け付けた言葉に基づいて前記文書集合を検索する文単位検索方法において、前記文書集合から得られる文書データを一又は複数の文からなる文単位に分別しておくステップ、分別した文単位毎に、該文単位に出現する単語、又は、文書データ中の先行の文単位から参照する単語を抽出するステップ、前記文単位に対して抽出した単語毎に、各文単位における特徴を特定して記憶しておくステップ、分別した文単位毎に、該文単位に対して抽出した単語が該文単位及び先行の文単位で出現する場合の前記特徴の組み合わせのパターン、又は先行の文単位から参照する場合の参照のパターンを含む特徴パターンを特定するステップ、特定した特徴パターンと、該特徴パターンで特定された単語が後続の文単位で出現又は参照されたか否かとを記憶しておくステップ、前記文書集合から得られる文書中の文単位全体に対し、一の特徴パターンで特定される単語が後続の文単位で出現又は参照される参照確率の回帰分析を行って特徴パターンに対応する回帰係数を得る回帰学習を実行するステップ、分別した文単位毎に、文書データ中で先行の文単位から各文単位に至るまでに抽出された各単語について、前記文単位で特定される特徴パターンに対応する前記回帰係数を使用し、前記単語の前記参照確率を算出するステップ、算出した参照確率を夫々付与した重み付き単語群を対応付けて予め記憶しておくステップ、言葉を受け付けた場合、受け付けた順に言葉を記憶するステップ、言葉を受け付けた場合、受け付けた言葉に出現する単語又は前記言葉よりも先に受け付けた言葉から参照する単語を抽出するステップ、抽出した各単語の前記受け付けた言葉における特徴を特定するステップ、先に受け付けた言葉で出現する場合の特徴の組み合わせのパターン、又は先に受け付けた言葉から参照する場合の参照のパターンを含む特徴パターンを特定するステップ、特定された特徴パターンに対応する前記回帰係数を使用して、前記単語の前記参照確率を算出するステップ、算出した参照確率を夫々付与した重み付き単語群を前記言葉に対応付けるステップ、前記受け付けた言葉と、予め分別されてある文単位とで、対応付けられている重み付き単語群の内の同一の単語毎に付与されている参照確率の差分を算出するステップ、予め分別されてある文単位に、前記参照確率の差分が小さい順に優先順位を付与するステップ、及び、前記文単位を付与された優先順位に基づいて出力するステップを含むことを特徴とする。
【００３６】
第１７発明に係る文単位検索装置は、自然言語からなる複数の文書データが記憶されている文書集合から文書データを取得する手段と、言葉を順次受け付ける手段とを備え、受け付けた言葉に基づいて前記文書集合を検索する文単位検索装置において、取得した文書データを一又は複数の文からなる文単位に分別する手段と、取得した文書データ中に連なる文単位夫々に、該文単位及び先行文脈に基づき求められる該文単位での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付けて記憶する手段と、言葉を受け付けた場合に受け付けた順に記憶する手段と、新たに言葉を受け付ける都度、該言葉に、該言葉及び該先行文脈に基づき求められる該言葉での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付ける手段と、予め分別された文単位から、受け付けた言葉に対応付けた重み付き単語群と類似する重み付き単語群が対応付けて記録されている文単位を抽出する手段と、抽出した文単位を出力する手段とを備えることを特徴とする。
【００３７】
第１８発明に係るコンピュータプログラムは、自然言語からなる複数の文書データが記

憶されている文書集合から、文書データを取得することが可能であるコンピュータを、言葉を順次受け付ける手段と、受け付けた言葉に基づいて前記文書集合を検索する手段として機能させることができるコンピュータプログラムにおいて、取得した文書データを一又は複数の文からなる文単位に分別する手段、取得した文書データ中に連なる文単位夫々に、該文単位及び先行文脈に基づき求められる該文単位での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付けて記憶する手段、言葉を受け付けた場合に受け付けた順に記憶する手段、新たに言葉を受け付ける都度、該言葉に、該言葉及び先行文脈に基づき求められる該言葉での顕現性を表わす重み値が付与された複数の単語からなる重み付き単語群を対応付ける手段、及び、予め分別された文単位から、受け付けた言葉に対応付けた重み付き単語群と類似する重み付き単語群が対応付けて記録されている文単位を抽出する手段として機能させることを特徴とする。

【００３８】

第１９発明に係るコンピュータで読み取り可能な記録媒体には、第１８発明のコンピュータプログラムが記録されていることを特徴とする。

【００３９】

第２０発明に係る文書記憶装置は、自然言語からなる複数の文書データを記憶する手段と、記憶した文書データを、文書データの先頭から順に一又は複数の文からなる文単位に分別する手段とを備え、分別した文単位毎に、該文単位に出現する単語又は先行する文単位から参照する単語が抽出してあり、分別した文単位毎に抽出した単語が記憶してある文書記憶装置において、文書データ中に連なる文単位毎に、複数の単語が、該文単位よりも後続の文単位に出現するか又は参照される参照確率を算出する手段と、前記文単位夫々に、該文単位での顕現性を表わす重み値として前記参照確率が付与された前記複数の単語からなる重み付き単語群を対応付けて記憶する手段とを備えることを特徴とする。

【００４０】

第２１発明に係る文書記憶装置は、抽出されてある単語の内の一の単語について、文単位夫々に対応付けられている重み付き単語群の内から、前記一の単語が含まれる単語群であり、且つ前記一の単語の重み値が所定値以上である単語群を抽出する抽出手段と、該抽出手段が抽出した単語群の各単語の重み値を単語毎に統合した値を、前記一の単語の各単語への関連度として付与した関連単語群を作成する作成手段と、作成した関連単語群を前記一の単語に対応付けて記憶する記憶手段とを備え、前記抽出されてある単語夫々について前記抽出手段、前記作成手段及び前記記憶手段の処理を実行するようにしてあり、各単語に対応付けて夫々の関連単語群を記憶するようにしてあることを特徴とする。

【００４１】

第１発明、第１７発明、第１８発明及び第１９発明では、自然言語からなる文書データが記録された文書集合から文書データが取得され、取得された文書データは更に一又は複数の文である文単位に分別される。文単位毎に、文書集合中で出現する各単語についてその文単位での重み値が付与され、重み値が付与された単語の重み付き単語群が文単位に対応付けて記憶される。言葉を受け付けた場合、受け付けた言葉についてもその言葉での重み値が付与された単語の重み付き単語群が対応付けられる。予め分別されている文単位から、受け付けた言葉に対応付けられた重み付き単語群と類似する重み付き単語群が対応付けられている文単位が抽出され、出力される。

【００４２】

第２発明では、第１発明において類似する重み付き単語群が対応付けられている文単位を抽出する際、予め文単位に対応付けて記憶されている重み付き単語群の内の複数の単語の重み値の分布が、受け付けた言葉に対応付けられた重み付き単語群の内の複数の単語の重み値の分布と所定の条件を満たすか否かの判断により類似するか否かが判定され、類似すると判定された重み付き単語群が対応付けられている文単位が抽出される。

【００４３】

第３発明では、第１発明又は第２発明において類似する重み付き単語群が対応付けられている文単位を抽出する際、重み付き単語群に同一の単語が含まれる文単位が抽出され、

その同一の単語に付与されている重み値の差分が小さい順に優先順位が付与される。

【００４４】

第４発明では、第１発明における重み付き単語群は、各単語を１次元とし、単語毎に付与される重み値の大きさを各単語に対応する次元方向の要素として持つ多次元ベクトルとして得られる。重み付き単語群が類似するか否かの判定を、重み付き単語群同士、即ち多次元ベクトル間の距離が短いか否かで判定される。抽出された文単位は、多次元ベクトル間の距離が短い順、即ち重み付き単語群同士が類似する順に出力される。

【００４５】

第５発明では、第１発明乃至第４発明において各単語に付与される重み値として、各単語が夫々、後続の文単位又は言葉に出現する又は参照される参照確率が算出されて付与される。

【００４６】

第６発明では、第５発明において算出される参照確率は、各単語に対して特定される先行の文単位から各文単位に至るまでの出現のパターン、又は先行の文単位からの参照のパターンを含む特徴パターンと同一の特徴パターンが特定される単語が、文書集合中で後続の文単位でさらに出現する又は参照される割合として算出される。

【００４７】

第７発明では、文書集合から抽出される各単語に対し特定される特徴パターンと、その特徴パターンが特定される単語が文書集合中の文書中の後続の文単位で出現したか又は参照されたかの判定結果とが回帰分析され、単語が後続の文単位で出現又は参照される参照確率に対する特徴パターンの回帰係数が算出される。第５発明において算出される参照確率は、単語毎に夫々の特徴パターンが特定され、その特徴パターンと回帰係数とから算出される。

【００４８】

第８発明及び第９発明では、文書集合が書き言葉からなる第１文書集合と、話し言葉からなる第２文書集合とに分けられて用いられる。文単位に対応付けられる重み付き単語群の各単語へ付与する参照確率は、第１文書集合に基づいて算出され、受け付けた言葉に対応付けられる重み付き単語群の各言葉へ付与する参照確率は、第２文書集合に基づいて算出される。

【００４９】

第１０発明では、第６発明乃至第９発明において参照確率を算出する際に、各単語の特徴パターンを特定するための特徴として、先行の文単位又は言葉で出現又は参照している場合の現在の文単位又は言葉に至るまでの数、出現又は参照した場合の単語の係り受け情報、出現した回数又は参照された回数、単語の名詞区別、単語が主題であるか、単語が主語であるか、単語の人称、単語の品詞情報等の情報が定量的に扱われる。

【００５０】

第１１発明では、第６発明乃至第１０発明において参照確率を算出する際に、各単語の特徴パターンを特定するための特徴として、先行の文単位又は言葉で出現又は参照している場合に先行の文単位又は言葉からの時間、出現又は参照した場合のその単語に相当する音声の発話速度、音声の周波数の高低の情報が定量的に扱われる。

【００５１】

第１２発明では、第１発明乃至第１１発明において、文書集合から抽出される単語の内の一の単語について、その単語の重み値が所定値以上の重み付き単語群が抽出される。その一の単語について抽出された複数の重み付き単語群の各単語の重み値を単語毎に統合した一の重み付き単語群が関連単語群として作成される。作成された関連単語群の各単語の関連度は、一の単語に所定値以上の重み値が付与されている場合の各単語の重み値への関連の深さを表わしている。文書集合から抽出される単語夫々に対して関連単語群が生成され記憶される。各文単位又は言葉に対応付けられた重み付き単語群の各単語の重み値が、夫々の単語に対応付けられた関連単語群の各単語の関連度を使用して付与し直される。

【００５２】

第１３発明では、第１２発明において一の単語に対する関連単語群が作成される際、一の単語の重み値が所定値以上である重み付き単語群として抽出された単語群が、その重み付き単語群での前記一の単語に対する重み値によって重み付けされた総和が算出される。総和は平均化され、各単語について平均化された重み値の総和が関連単語群の各単語の関連度として付与される。

【００５３】

第１４発明では、前記１２発明又は第１３発明で記憶される関連単語群の各単語の関連度が、文単位毎又は受け付けた言葉毎に対応付けられた重み付き単語群の各単語の重み値に乗算され、乗算結果が重み付き単語群の各単語の重み値として付与し直される。重み付き単語群の内の一の単語に注目した場合、一の単語に対応付けられた関連単語群の各単語の関連度が使用される。重み付き単語群の内の一の単語以外の各単語の重み値と、前記一の単語に対応付けられた関連単語群の各単語の関連度とが乗算されることにより、関連度の高い他の単語の重み値からの前記一の単語の重み値への影響が加味される。

【００５４】

第１５発明では、第１２発明乃至第１４発明における関連単語群は、各単語を１次元とし、単語毎に付与される関連度の大きさを各単語に対応する次元方向の要素として持つ多次元の関連度ベクトルとして得られる。各文単位又は言葉に対応付けられた多次元ベクトルは、各単語に対する関連語ベクトルの列からなる行列で変換される。即ち、多次元ベクトルは単語の各１次元間の距離が関連度が高い単語の次元間ほど距離が短い斜交座標系における多次元ベクトルで表現される。したがって、重み付き単語群を表現する多次元ベクトルは、それに含まれる単語と関連度が高い単語軸方向に回転され、関連度が高い単語を含む多次元ベクトル間の距離はより短くなる。

【００５５】

第１６発明では、文書集合から取得された文書データを更に分別した文単位毎に、文単位又は先行の文単位から参照する単語が抽出され、各単語に対して各文単位における特徴が特定され、先行の文単位から各文単位に至るまでの特徴の組み合わせのパターン、又は各単語の先行の文単位からの参照のパターンを含む特徴パターンが特定される。特定された特徴パターンによる参照確率の回帰学習に基づいて、抽出された各単語の参照確率が算出され、重み付き単語群として予め文単位毎に記憶される。受け付けた言葉に対しても先行の言葉に基づいた特徴パターンが特定されて各単語の参照確率が算出され、重み付き単語群が対応付けられる。予め記憶してある文単位は、受け付けた言葉の重み付き単語群と同一の単語の参照確率の差分が小さい順に優先順位が付与されて出力される。

【００５６】

第２０発明では、文書集合から取得された文書データを更に分別した文単位毎に、その文単位での単語の重みが付与された重み付き単語群が対応付けられて記憶される。

【００５７】

第２１発明では、第１２発明で文書から抽出されてある単語夫々について作成された関連単語群が記憶される。

【発明の効果】

【００５８】

本発明による場合、文書集合から取得した文書データ中の一又は複数の文からなる文単位毎に、複数の単語夫々の当該文単位での重み値を付与した重み付き単語群が対応付けられて記憶される。重み値付き単語群は、各文単位での各単語の重み値の組であり、文単位毎の意味のまとまりを示す情報として推定することができる。各重み値に先行の文単位から続く文脈が反映された値が付与されていることにより、分別された連なる文単位中の各文単位での重み付き単語群は、文書全体での意味のまとまりと異なり、文書中にある先行の文から続く文脈の流れの中で、動的に時系列的に変化していく意味のまとまりとして捉えることができる。検索のために入力される言葉での重み値が付与された重み付き単語群と類似する重み付き単語群が対応付けられる文単位が抽出されることにより、文書全体ではなく、単語の顕現性、即ち意味のまとまりが類似する文単位を直接的に検索することが

できる。

【００５９】

また、重み付き単語群が類似するか否かは、受け付けた言葉の重み付き単語群の内の複数の単語の重み値の分布と、予め記憶してある重み付き単語群の内の複数の単語の重み値の分布とを比較した場合に、分布同士が類似であると判断できる所定の条件を満たすとき、記憶してある重み付き単語群が受け付けた言葉の重み付き単語群と類似するということができる。例えば、重み付き単語群同士が類似していると判断できる所定の条件を、各単語の重み値の分布が相似であるといえる条件とした場合、重み付き単語群が類似しているということができる。つまり、一方の重み付き単語群において一の単語の重み値の他の単語の重み値に対する比率が、他方の重み付き単語群における一の単語の重み値の他の単語の重み値に対する比率にも保存される場合、それらの重み付き単語群同士は類似していると判断することができる。また、所定の条件を、例えば、一又は複数の単語に注目した場合にその単語の重み値がいずれも所定値以上であるか否かに設定することで判断することもできる。また、受け付けた言葉に対応付けた重み付き単語群と、予め分別された文単位に対応付けられている重み付き単語群と比較した場合に、同一の単語の重み値の差分が小さいか否かにより類似するか否かを判断することもできる。

【００６０】

また、重み付き単語群を、各単語を１次元として、各単語の文単位又は言葉での重み値を各次元成分に対する要素として持つ多次元ベクトルとして表現することにより、文単位又は言葉毎の意味のまとまりを定量的なベクトルとして扱うことができる。また、文単位又は言葉毎の意味のまとまりを定量的な多次元ベクトルとして扱うことにより、ベクトル演算が可能なコンピュータを利用して、受け付けた言葉に対応付けられたベクトルと記憶してある文単位毎に対応付けられたベクトルとの距離を算出することによって類似する文単位を直接的に抽出することができる。さらに、多次元ベクトルとして表現することによって、受け付けた言葉、又は予め分別された文単位の多次元ベクトルが満たす条件を、多次元空間上のどの空間に相当するか否かによって設定することができ、類似する文単位を直接的に抽出することができる。

【００６１】

なお、ここでいう文書集合は、いわゆる書き言葉からなる文書データの集合に限らない。したがって、それらを分別した文単位も書き言葉からなる文単位とは限らない。文書データは既に記憶されてあるデータを意味してリアルタイムに受け付ける言葉と区別するものであり、話し言葉による対話が順に書下された文書データでもよい。

【００６２】

また、受け付ける言葉は、検索の目的で入力される単語、文章等に限らず、例えばユーザ同士の対話中の音声を含む各発話でもよい。各発話での重み値が付与された重み付き単語群に基づいて文単位を抽出するので、対話中で発話毎に意味が動的に、時系列的に変化していくことを考慮した意味のまとまりを発話毎に推定することができる。したがって、各発話に対して推定される意味のまとまりに類似する文単位を抽出して提示することが可能になる。

【００６３】

さらに、本発明による場合、重み付き単語群の各単語の重み値を、後続の文単位又は言葉でも出現又は参照される参照確率として付与することにより、各単語の重み値を注目されている度合い、即ち顕現性を示す定量的な値で表わすことができる。文脈上のその文単位において重要な注目されている単語は、継続して出現又は参照される確率が高いと考えられる。したがって、参照確率はその文単位における各単語の注目されている度合い、即ち顕現性を示すということができる。

【００６４】

また、各文単位で実際に出現することなしに指示代名詞又はゼロ代名詞で表わされる単語、又は指示代名詞又はゼロ代名詞でも表わされていない単語であっても、文単位又は言葉に実際に出現していない単語であっても後続の文単位又は言葉で出現又は参照される単

語は、その文単位又は言葉での顕現性が高いと考えられる。各文単位を基準とした先行の複数の文単位での単語の特徴パターンに基づいて参照確率を算出するので、実際に出現していない単語であっても、顕現性の高さをより正しく定量的に表わすことができる。

【００６５】

さらに、言葉を音声で受け付けた場合は、言葉が発声されたときの声の特徴、即ち話す速度、声調からも、その言葉に含まれる単語がその言葉で重みを持っているのか否かを定量的に特徴づけて各単語の顕現性の高さを表わすことができる。

【００６６】

さらに、本発明による場合、検索結果として出力する文単位が書き言葉である場合は、書き言葉からなる文書集合に基づいて参照確率を算出し、受け付けた言葉が話し言葉である場合は、話し言葉からなる文書集合に基づいて参照確率を学習、算出する。これにより、書き言葉と話し言葉とで異なる特徴を踏まえて、より意味合いが似た文単位を出力することができる。

【００６７】

また、本発明による場合、単語毎に各単語からの関連度を定量的に算出して記憶しておく。重み付き単語群の内の各単語の重み値を、他の単語の重み値と、各単語からの前記一への単語の関連度とに基づいて算出し直す。これにより、一の単語の重み値に対し、他の単語の内の一の単語に対する関連度が高い単語の重み値の影響を反映させることができる。つまり、一の単語に対する関連度が高い単語の重み値が高い場合は、一の単語の重み値が高くなることを再現することができる。

【００６８】

一の単語に対する関連語群を関連度ベクトルとして表現し、重み付き単語群を多次元ベクトルで表現した場合に各単語に対する関連度ベクトルの列からなる行列で多次元ベクトルを変換することにより、関連度の強い単語を含む重み付き単語群を表現する多次元ベクトル間の距離が短くなる。

【００６９】

これにより、重み付き単語群の内の一の単語以外の単語の内、前記一の単語への関連度が高い単語の重み値の影響を、前記一の単語の重み値に反映することができる。各文単位又は言葉での各単語の顕現性に関連度を反映させて、受け付けた言葉に表れていない場合であってもユーザに意識されている単語を連想させる文単位を効果的に検索することができる等の優れた効果を奏する。

【図面の簡単な説明】

【００７０】

【図１】本発明に係る文単位検索方法の概要を示す説明図である。

【図２】実施の形態１における文単位検索装置を用いた検索システムの構成を示すブロック図である。

【図３】実施の形態１における文単位検索装置のＣＰＵが、取得した文書データに対する形態素解析及び統語解析処理の解析結果からタグ付け及び単語抽出を行い記憶する処理手順を示すフローチャートである。

【図４】実施の形態１における文書記憶手段で記憶される文書データの内容の一例を示す説明図である。

【図５】実施の形態１における文単位検索装置のＣＰＵが、形態素解析及び統語解析した結果を付与して文書記憶手段に記憶させる文書データの一例を示す説明図である。

【図６】実施の形態１における文単位検索装置のＣＰＵが取得した全文書データから抽出した単語のリストの例を示す説明図である。

【図７】実施の形態１における文単位検索装置のＣＰＵが、文書記憶手段で記憶しているタグ付け済み文書データからサンプルを抽出し、回帰分析を行って参照確率を算出するための回帰式を推定する処理手順を示すフローチャートである。

【図８】実施の形態１における文書記憶手段で記憶された文書データ中の文で特定される特徴パターンの例を示す説明図である。

【図９】実施の形態１における文単位検索装置のＣＰＵが、文書記憶手段で記憶しているタグ付け済みの文書データの文毎に単語の参照確率を算出し、記憶する処理手順を示すフローチャートである。
【図１０】実施の形態１における文単位検索装置のＣＰＵが、文書記憶手段で記憶しているタグ付け済みの文書データの文毎に単語の参照確率を算出し、記憶する処理手順を示すフローチャートである。
【図１１】実施の形態１における文単位検索装置のＣＰＵが、文書データに示される文書を文毎に分別した一例を示す説明図である。
【図１２】実施の形態１における文単位検索装置のＣＰＵが、参照確率を算出した結果を付与して文書記憶手段に記憶させる文書データの一例を示す説明図である。
【図１３】実施の形態１における文単位検索装置のＣＰＵが、文単位毎に算出した重み付き単語群を索引付けして記憶した場合のデータベースの内容例を示す説明図である。
【図１４】文単位検索装置のＣＰＵにより文毎に記憶される単語及び該単語に対して算出された参照確率の組が、文が続くにつれてどのように変化するかを示す説明図である。
【図１５】実施の形態１における文単位検索装置及び受付装置の検索処理の処理手順を示すフローチャートである。
【図１６】実施の形態１における文単位検索装置及び受付装置の検索処理の処理手順を示すフローチャートである。
【図１７】実施の形態１における文単位検索装置及び受付装置の検索処理の処理手順を示すフローチャートである。
【図１８】実施の形態１における文単位検索装置のＣＰＵが、受付装置から受信したテキストデータに対して特定した特徴パターンの例を示す説明図である。
【図１９】実施の形態２における文単位検索装置及び受付装置の検索処理の処理手順を示すフローチャートである。
【図２０】実施の形態３における本発明の検索方法に関わる、一の単語と関連の深い単語の顕現性の影響の概要を示す説明図である。
【図２１】実施の形態３における文単位検索装置のＣＰＵが関連語群を作成する処理手順を示すフローチャートである。
【図２２】実施の形態３における文単位検索装置のＣＰＵが関連語群を作成する処理手順を示すフローチャートである。
【図２３】実施の形態３における文単位検索装置のＣＰＵによって関連語群が作成される場合の、各処理の過程での重み付き単語群の例を示す説明図である。
【図２４】実施の形態３における文単位検索装置のＣＰＵが、各文単位に対応付けられて記憶されている重み付き単語群の各単語の重み値を算出し直す処理手順を示すフローチャートである。
【図２５】実施の形態３における文単位検索装置のＣＰＵが、各文単位に対応付けられて記憶されている重み付き単語群の各単語の重み値を算出し直す処理手順の詳細を示すフローチャートである。
【図２６】実施の形態３における文単位検索装置のＣＰＵによって算出された各単語の顕現性を表わす重み値の内容例を示す説明図である。
【図２７】実施の形態３における文単位検索装置及び受付装置の検索処理の処理手順を示すフローチャートである。
【図２８】実施の形態３における文単位検索装置及び受付装置の検索処理の処理手順を示すフローチャートである。
【図２９】本発明の文単位検索方法を文単位検索装置で実施する場合の構成を示すブロック図である。
【符号の説明】
【００７１】
１　文単位検索装置
１１　ＣＰＵ

１３　記憶手段
１５　通信手段
１６　文書集合接続手段
１７　補助記憶手段
１８　可搬型記録媒体
１Ｐ　制御プログラム
２　文書記憶手段
４　受付装置

【発明を実施するための最良の形態】

【００７２】

以下本発明をその実施の形態を示す図面に基づき具体的に説明する。

【００７３】

図１は、本発明に係る文単位検索方法の概要を示す説明図である。図１中の１００は、複数の文書データが記憶されている文書集合を表わしており、文書集合１００から取得される一の文書１０１は、一又は複数の文からなる文単位S_1，…，S_i，S_{i+1}，…で構成されている。文単位S_1，…，S_i，S_{i+1}，…は、文書１０１の先頭から順に文脈の流れに沿い、時系列的に変遷する意味合いを有して連なっている。図１中の２００は、ユーザＡとユーザＢとの会話を表わしている。ユーザＡとユーザＢの会話２００は、上から下へ時系列に連なるユーザＡ及びユーザＢからの発話U_{j-3}，…，U_j の集合である。会話は、発話U_{j-3}，U_{j-2}，U_{j-1}，U_j の順になされている。なお、ユーザＡとユーザＢとを区別せずに単に連続する発話の集合として捉えてもよい。

【００７４】

本発明に係る文単位検索方法は、文単位又は言葉をユーザが筆記又は発話した時点での各単語への注目度合いを定量的な重み値として表わして各単語に付与し、時系列に連続する文単位又は言葉毎に変遷していく各単語への注目度合いを反映した重み付き単語群を各文単位における文脈上の意味合いを表わす指標として用いることにより、同様の文脈上の意味合いを有する文単位を直接的に検索し、出力することを目的としている。

【００７５】

図１の説明図で示す例での会話２００は、ユーザＡとユーザＢとの間でなされている京都への旅行についての会話である。会話２００中の発話U_{j-3}では「京都」「旅行」が現れ、文脈の流れは「京都の旅行」である。発話U_{j-2}では、「京都」、「旅行」は現れていないが「“京都への旅行の”時期」についての発話であり、「京都」「旅行」「時期」について注目がされている。U_{j-1}では「暑い」が現れている。U_{j-1}では「京都」、「旅行」は現れていないが、「“京都は”暑い」のであり、依然「京都」は文脈上の意味に対して重みを持っている。さらにユーザＡとユーザＢとの間では、U_{j-1}の発話の時点では、「旅行」よりも「京都」及び「時期」が注目されており、ユーザＡとユーザＢとは文脈上の意味合いが変遷していることを共通して認識できるはずである。さらに、発話U_j の中で「有名」「祭」が現れている。このU_j の発話の時点だけを考えれば、「京都」「旅行」「時期」「暑い」という単語は現れていない。しかし、少なくともユーザＡにとっては、発話U_j は文脈上「夏」の「京都」の「祭」についての意味合いを有している。したがって、発話U_j の時点でも、依然として「京都」は文脈上の意味合いに対して重みを持っている。なお、発話U_j を発したユーザＡは少なくとも、祭に相当する単語として「祇園祭」などを想起しているはずである。

【００７６】

これに対し、文書集合１００中の文書１０１には京都の旅行記が記されている。その中の文単位S_i は、「７月」の「京都」といえば「祇園祭」という意味合いを有している。即ち、文単位S_i は、『「夏」の「７月」の「京都」の「祭」といえば』、「祇園祭」であるという意味合いを有している。つまり、発話U_j と、文単位S_i とは、共通して「夏」「京都」「祭」に重みを有しており、文脈上の意味合いが類似している。このように、本発明に係る文単位検索方法では、発話U_j の際にユーザが意識している、先行の発話か

らの文脈上の意味のまとまりを推定し、類似する文脈上の意味合いを有する文単位S_kを直接的に検索して出力することを目的としている。

【００７７】

本発明に係る文単位検索方法を実施するコンピュータシステムを実現した場合、連続する発話を受け付け、それらの言葉の文脈上の意味と類似する文単位を文書集合から抽出するのみならず、ユーザＡとユーザＢとの会話中に、コンピュータシステムが発話毎に関連する情報を提示して会話に参入する鼎談が可能になる。また、コンピュータシステムがユーザＡとユーザＢとの会話を支援することも可能になる。図１の説明図の例で、会話１００のユーザＡによる発話Ｕｊ の次に、コンピュータシステムによって「７月の京都といえば祇園祭です。」等の音声の出力がされた場合は、ユーザＡとユーザＢとコンピュータシステムとの間での鼎談が実現することになる。また、ユーザＡとユーザＢとの会話が続かなくなった場合に、コンピュータシステムによって「７月の京都といえば祇園祭」等の情報の提示がされることで、ユーザＡとユーザＢとの会話への支援も実現する。

【００７８】

そこで、このような文脈上の意味が類似する文単位を文書集合から検索することを実現するために、本発明に係る文単位検索方法をコンピュータ装置に実施させる。この場合、コンピュータ装置には、予め文書集合の文書データを夫々文単位に分別しておく処理、及び分別した文単位に各文単位の文脈上の意味を表わす定量的な情報を記憶させておく処理を含む事前処理が必要になる。さらに、コンピュータ装置が発話を受け付けた場合、その発話の会話の流れ上の意味を表わす定量的な情報を求める処理、及び、発話に対して求めた情報に基づいて意味が類似する文単位を抽出して検索結果として出力する処理を含む検索処理が必要になる。

【００７９】

したがって、以下に説明する実施の形態１乃至３では、本発明に係る文単位検索方法をコンピュータ装置に実施させるために必要なハードウェア構成についてまず説明する。さらにコンピュータ装置による処理を、事前処理と検索処理とを区別して段階的に説明する。具体的には、各実施の形態において、

「１．ハードウェアの構成及びシステムの概要」、

事前処理として

「２．文書データの取得及び自然言語解析」、及び

「３．文書データの文毎の意味のまとまりの定量化」、

次に

「４．検索処理」

の順に説明する。

【００８０】

なお、以下に説明する実施の形態１乃至３では、本発明に係る文単位検索方法を実施する例として、文書データの文書集合を記憶しておくハードウェアと、発話を受け付けるコンピュータ装置と、文書集合が記憶されたハードウェア及び発話を受け付けるコンピュータ装置に接続して検索処理を実行するコンピュータ装置とで構成される検索システムを挙げて説明する。

【００８１】

また、以下に示す例では主に、文書集合が日本語の自然文からなる場合について各処理、具体例を示している。しかしながら、本発明の文単位検索方法は、日本語のみならず、他の言語にも適用することができることは勿論である。この場合、言語解析（形態素解析、統語解析）等の言語毎に特有の文法上の取り扱い等は、その言語毎に最適な方法を用いる。

【００８２】

（実施の形態１）

１．ハードウェアの構成及びシステムの概要

図２は、実施の形態１における文単位検索装置１を用いた検索システムの構成を示すブ

ロック図である。検索システムは、文書データからの検索処理を実行する文単位検索装置１と、自然言語からなる文書データを記憶する文書記憶手段２と、インターネット等のパケット交換網３と、ユーザから入力されるキーワード又は音声等の言葉を受け付ける受付装置４，４，…とで構成される。文単位検索装置１は、ＰＣ（Ｐｅｒｓｏｎａｌ Ｃｏｍｐｕｔｅｒ）であり、自然言語からなる文書データを記憶する文書記憶手段２と接続される。また、受付装置４，４，…もＰＣであり、文単位検索装置１は、パケット交換網３を介して受付装置４，４，…と接続され通信が可能である。

【００８３】

実施の形態１の検索システムでは、文単位検索装置１は、検索の対象である文単位を含む文書データを文書記憶手段２に予め記憶しておく。文単位検索装置１は、文書記憶手段２に記憶した文書データを、予め文単位に分別し、後に検索処理が可能なように各文単位に文脈上の意味を表わす定量的な情報を記憶させておく。また、受付装置４，４，…は、受け付けた言葉をコンピュータで処理可能なテキストデータ又は音声データに変換し、パケット交換網３を介して当該データを文単位検索装置１へ送信する。文単位検索装置１が、受信した言葉のデータに基づいて文書記憶手段２に記憶した文書データから一又は複数の文からなる文単位を抽出し、抽出した文単位をパケット交換網３を介して受付装置４，４，…へ出力することで文単位の検索を実現する。

【００８４】

文単位検索装置１は、少なくとも、各種ハードウェアを制御するＣＰＵ１１と、各種ハードウェア間を接続する内部バス１２と、不揮発性のメモリからなる記憶手段１３と、揮発性のメモリからなる一時記憶領域１４と、パケット交換網３と接続するための通信手段１５と、文書記憶手段２と接続するための文書集合接続手段１６と、ＤＶＤ、ＣＤ－ＲＯＭ等の可搬型記録媒体１８を用いる補助記憶手段１７とを備える。

【００８５】

記憶手段１３には、ＤＶＤ、ＣＤ－ＲＯＭ等の可搬型記録媒体１８から取得した、ＰＣが本発明に係る文単位検索装置１として動作するための制御プログラム１Ｐが記憶されている。ＣＰＵ１１は、制御プログラム１Ｐを記憶手段１３から読み出して実行すると共に、内部バス１２を介して各種ハードウェアを制御する。一時記憶領域１４は、ＣＰＵ１１の演算処理によって一時的に発生する情報が記憶される。

【００８６】

ＣＰＵ１１は、受付装置４，４，…から送信される言葉のデータを通信手段１５を介して受信したことを検知し、受信した言葉のデータに基づいて処理を実行し、検索処理を行う。また、ＣＰＵ１１は、文書集合接続手段１６を介して文書記憶手段２で記憶している文書データを取得し、且つ、文書集合接続手段１６を介して文書データを文書記憶手段２に記憶させることが可能である。

【００８７】

ＤＶＤ、ＣＤ－ＲＯＭ等の可搬型記録媒体１８から補助記憶手段１７を介して取得した、記憶手段１３に記憶されている制御プログラム１Ｐでは更に、記憶手段１３で記憶している辞書情報に基づいて文字列で表された文書データを形態素解析及び統語解析等の自然言語解析をＣＰＵ１１に実行させることができるようにしてある。

【００８８】

受付装置４，４，…は、少なくとも、各種ハードウェアを制御するＣＰＵ４１と、各種ハードウェア間を接続する内部バス４２と、不揮発性メモリからなる記憶手段４３と、揮発性メモリからなる一時記憶領域４４と、マウス又はキーボード等の操作手段４５と、モニタ等の表示手段４６と、マイク及びスピーカ等の音声入出力手段４７と、パケット交換網３へ接続するための通信手段４８とを備える。

【００８９】

記憶手段４３には、ＰＣが受付装置４，４，…として動作するための処理プログラム等が記憶されている。ＣＰＵ４１は、処理プログラムを記憶手段４３から読み出して実行すると共に、内部バス４２を介して各種ハードウェアを制御する。一時記憶領域４４は、Ｃ

ＰＵ４１の演算処理によって一時的に発生する情報が記憶される。

【００９０】

ＣＰＵ４１は、ユーザからの文字列入力操作を操作手段４５を介して検知し、入力された文字列を一時記憶領域４４に記憶することができる。ＣＰＵ４１は、ユーザから入力された音声を音声入出力手段４７を介して検知し、記憶手段４３に記憶された音声認識のためのプログラムを読み出して実行することによって入力された音声をテキストデータに変換することができる。また、ＣＰＵ４１は、ユーザから入力された音声を音声入出力手段４７により、コンピュータで処理可能な音声データとして入力することができる。

【００９１】

また、ＣＰＵ４１は、ユーザからの文字列入力操作又は音声入力を検知することで得られたテキスト又は音声の言葉のデータを通信手段４８を介して文単位検索装置１へ送信する。

【００９２】

なお、ＣＰＵ４１は、音声データをテキストデータに変換して送信してもよく、その場合は、ＣＰＵ４１は、音声認識によって得られる音声データの特徴、例えば各単語に相当する音素が発声された時の速度、単語に相当する音素の周波数等のデータを共に送信してもよい。また、ＣＰＵ４１は、各単語に相当する音声データ間の時間差についても記憶しておき、以前に受け付けた言葉にその単語が含まれていた時点との時間差も共に文単位検索装置１へ送信してもよい。

【００９３】

２．文書データの取得及び自然言語解析

上述のように構成される検索システムにおいて、文単位検索装置１はまず、事前処理として文書集合を用意して、後に各文書データに含まれる文単位毎の意味のまとまりを表わすことができるようにしておく処理を行なう。「２．文書データの取得及び自然言語解析」では、文単位検索装置１が文書記憶手段２に文書データを記憶しておき、各文書データを言語解析して一又は複数の文からなる文単位に分別し、さらに文単位毎に文法的な特徴を解析し、文書記憶手段２に文単位毎に記憶しておく処理について説明する。なお、実施の形態１では、文単位検索装置１は文単位を一の文とした場合について説明する。

【００９４】

文単位検索装置１のＣＰＵ１１は、検索の対象である文単位を含む文書データを文書記憶手段２に予め記憶しておく。文単位検索装置１のＣＰＵ１１は、通信手段１５及びパケット交換網３を介して取得可能な文書データをＷｅｂクローリングにより取得し、文書集合接続手段１６を介して文書記憶手段２に記憶する。文単位検索装置１のＣＰＵ１１は、取得して文書集合接続手段１６を介して文書記憶手段２に記憶してある文書データを文単位に分別し、夫々言語解析（形態素解析及び統語解析）を行い、その結果を文単位毎に対応付けて記憶する処理を行なう。

【００９５】

以下に、文単位検索装置１のＣＰＵ１１が、文書データを取得し、取得した文書データに対して形態素解析及び統語解析の自然言語解析をして、文単位毎に記憶する処理手順について説明する。図３は、実施の形態１における文単位検索装置１のＣＰＵ１１が、取得した文書データに対する形態素解析及び統語解析処理の解析結果からタグ付け及び単語抽出を行い記憶する処理手順を示すフローチャートである。図３のフローチャートに示す処理は、文単位毎にその文単位に出現する単語又は先行の文単位から参照する単語を抽出する処理と、各文単位における各単語の特徴を特定して記憶しておく処理に対応する。

【００９６】

ＣＰＵ１１は、Ｗｅｂクローリングを開始すると文書データを取得したか否か判断する（ステップＳ１１）。ＣＰＵ１１が文書データを取得していないと判断した場合は（Ｓ１１：ＮＯ）、ＣＰＵ１１は処理をステップＳ１１へ戻し、文書データを取得するまで待機する。ＣＰＵ１１が文書データを取得したと判断した場合は（Ｓ１１：ＹＥＳ）、ＣＰＵ１１は、取得した文書データから一文毎の読み出しを試み、読み出しが成功したか否かを

判断する（ステップＳ１２）。

【００９７】

ＣＰＵ１１が、読み出し箇所が文書データの終端に至っておらず、文の読み出しが成功したと判断した場合は（Ｓ１２：ＹＥＳ）、読み出した文の形態素解析及び統語解析を行う（ステップＳ１３）。

【００９８】

ＣＰＵ１１は、形態素解析及び統語解析の結果から、解析した文に出現する単語及び当該文で先行の文から参照する単語を抽出し、リストに記憶する（ステップＳ１４）。更に、ＣＰＵ１１は、後述で説明するように解析結果からタグを生成し（ステップＳ１５）、読み出した文にタグを付加して、文書集合接続手段１６を介して文書記憶手段２に記憶させる（ステップＳ１６）。

【００９９】

一方、ＣＰＵ１１が、読み出し箇所が文書データの終端に至っており、文の読み出しが失敗したと判断した場合は（Ｓ１２：ＮＯ）、取得した文書データに対する処理を終了する。

【０１００】

上述の処理を、文書データを取得する都度に行い、タグ付け済みの文書データを文書記憶手段２に記憶しておく。

【０１０１】

次に、文単位検索装置１のＣＰＵ１１による上述の処理の詳細を、具体例を挙げて説明する。

【０１０２】

図４は、実施の形態１における文書記憶手段２で記憶される文書データの内容の一例を示す説明図である。文書記憶手段２で記憶される文書データは、文単位検索装置１のＣＰＵ１１が通信手段１５を介して、パケット交換網３に接続され公開されているＷｅｂサーバから取得されたＨＴＭＬ（ＨｙｐｅｒＴｅｘｔ　Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ）等のテキストデータをもとに記憶される。図４に示す一例も、インターネットで公開されたＷｅｂページ（http://ja.wikipedia.org/wiki/祭より抜粋）より取得することができたＨＴＭＬデータの文書である。以下、この文書例を使用して文書の解析及び検索等について説明する。

【０１０３】

文単位検索装置１のＣＰＵ１１は、図３のフローチャートに示したステップＳ１２の文の読み出しの処理において、取得した文書データ中の文字列を「文」の言語単位（文単位）に分別する。分別する方法として例えば、ＣＰＵ１１は、日本語からなる文書データである場合、句点「。」を表す文字列によって、又は、英語からなる文書データである場合はピリオド「．」を表す文字列によって分別してもよい。

【０１０４】

次に、図３のフローチャートに示した文単位検索装置１のＣＰＵ１１によるステップＳ１３の形態素解析及び統語解析の処理の詳細を説明する。

【０１０５】

文単位検索装置１のＣＰＵ１１は、「文」の言語単位に対して辞書情報に基づいた形態素解析を行い、文の最小構成単位である形態素を同定して形態素の構造を解析する。例えば、図４に示した文書データでは、ＣＰＵ１１は、記憶手段１３の辞書情報に基づいて、「祭」「神霊」等の名詞、「九州」等の固有名詞、「祀る」等の動詞、「と」「は」等の助詞、「、」「。」等の記号等を示す文字列と照合することで形態素を同定する。形態素解析の手法については今日では種々の手法が提案されており、本発明では当該形態素解析の手法を限定するものではない。

【０１０６】

さらに、文単位検索装置１のＣＰＵ１１は、同定した形態素毎にその品詞情報（名詞、助詞、形容詞、動詞、副詞等）と、日本語文である場合は日本語の文法、英文である場合

は英語の文法に基づく品詞間の結束性を統計的に求めた文法情報とに基づいて形態素間の文法的関係を抽出する統語解析を行う。例えば、文法を木構造に当てはめて形態素の品詞情報から木構造に従って形態素間の関係を抽出することができる。解析対象が（形容詞＋名詞＋助詞＋名詞）である場合、まず解析対象が名詞であるか否かを判断する。名詞でないと判断した場合は次に、当該解析対象が（形容詞＋名詞）に当てはまるか否かを判断する。したがって、当該解析対象の先頭の形態素が形容詞句であるか否かを判断する。先頭の形態素が形容詞であると判断した場合は、当該形容詞が後続する名詞を修飾する当該解析対象の中で一番大きな修飾語であると判断される。つまり（形容詞＋（名詞））という関係が抽出される。

【０１０７】

次に、残りの解析対象が（名詞）であるか否かを判断する。複数の形態素からなり、名詞ではないと判断した場合は、当該残りの解析対象が（形容詞＋名詞）に当てはまるか否かを判断する。したがって、残りの解析対象の先頭の形態素が形容詞であるか否かを判断する。残りの解析対象の先頭の形態素が形容詞でないと判断した場合は、（形容詞＋名詞）の形容詞の部分を（名詞＋助詞）に展開し、残りの解析対象が（（名詞＋助詞）＋名詞）に当てはまるか否かを判断する。残りの解析対象が（（名詞＋助詞）＋名詞）に当てはまると判断した場合は、当該解析対象（形容詞＋名詞＋助詞＋名詞）の形態素間の文法的関係は[形容詞＋｛（名詞＋助詞）＋名詞｝]であると抽出することができる。統語解析の方法についてもこのような方法を基礎とする手法に限らず、形態素解析の手法同様に今日では種々の手法が提案されており本発明では当該統語解析の手法を限定するものではない。

【０１０８】

実施の形態１では、一例として形態素解析及び統語解析についてｃｈａｓｅｎ（http://chasen.org）及びＣａｂｏＣｈａ（工藤　拓、松本　裕治「チャンキングの段階適用による日本語係り受け解析」情報処理学会論文誌Ｖｏｌ．６、Ｎｏ．４３、ｐｐ．１８３４－１８４２（２００２）、http://chasen.org/~taku/software/cabocha参照）にて開示された技術に基づいて行う。他にＫＮＰ（Ｋｕｒｏｈａｓｈｉ－Ｎａｇａｏ　Ｐａｒｓｅｒ）（黒橋　禎夫、長尾　眞「並列構造の検出に基づく長い日本語文の構造解析」自然言語処理Ｖｏｌ．１、Ｎｏ．１、ｐｐ．３５－５７（１９９４））で開示されている技術に基づいて解析するのでもよい。

【０１０９】

文単位検索装置１のＣＰＵ１１は、解析した形態素及び形態素間の文法的関係を、ＸＭＬ（ｅＸｔｅｎｓｉｂｌｅ　Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ）に基づくタグで表した文書データを生成して文書記憶手段２に記憶させる。本発明が利用する形態素解析及び統語解析の自然言語解析方法（ｃｈａｓｅｎ、ＣａｂｏＣｈａ）では入力された文字列を形態素解析し、さらに統語解析して各形態素の品詞情報、形態素の係り先を示す情報等を分別した形態素毎に出力するようにしてある。文単位検索装置１の記憶手段１３に記憶されている制御プログラム１Ｐでは、当該自然言語解析方法を文単位検索装置１のＣＰＵ１１に実行させることができるように構成されている。

【０１１０】

本発明が利用する形態素解析及び統語解析では、例えば、図４に示した「九州地方北部では、秋に行われるものに対して（お）くんちと称する場合もある。」という文の文字列に対しまず文節番号が付される。（０：九州地方北部では、／１：秋に行われるものに対して（お）くんちと称する場合も／２：ある。）さらに各文節で形態素に分別され、形態素毎の品詞情報、形態素の基本形情報、発音情報等が付加される。文節番号０の文節は、（０：九州（名詞＋固有名詞＋地域＋一般、九州、キュウシュウ）／地方（名詞＋一般、地方、チホウ）／北部（名詞＋一般、北部、ホクブ）／で（助詞＋格助詞＋一般、で、デ）／は（助詞＋係助詞、は、ハ）／、（記号＋読点））と形態素の同定及び情報の付加が行われる。「九州」という形態素は名詞であって固有名詞であり、地域を示す名詞でもあり、一般名詞として使用されることもある。また基本形は「九州」であり、「キュウシュ

ウ」と発音することを判別することができる。他の文節も同様である。また、係り受け情報は例えば、（０　２，１　２，２　－１）と文節間の係り受け関係が判別可能なように取得できる。この例では、文節番号０の文節は文節番号２の文節を係り先とし、文節番号１の文節は文節番号２の文節を係り先とすることが判別できる。また、文節番号２の文節は係り先がないことを係り先を－１とすることで判別できる。

【０１１１】

図５は、実施の形態１における文単位検索装置１のＣＰＵ１１が、形態素解析及び統語解析した結果を付与して文書記憶手段２に記憶させる文書データの一例を示す説明図である。図４に示した内容の文書データに対して図３のフローチャートに示した処理手順が実行されたことにより文書記憶手段２に記憶された文書データの例に相当する。

【０１１２】

図５に示すように、文単位検索装置１のＣＰＵ１１により、図４に示した内容の文書の一部が固有名詞、名詞、助詞、動詞等の形態素に分別され、形態素間の文法的関係性はタグの入れ子によって表されている。図５に示す例は、ＧＤＡ（Ｇｌｏｂａｌ　Ｄｏｃｕｍｅｎｔ　Ａｎｎｏｔａｔｉｏｎ；http://i-content.org/gda参照）で提案されている規則に則ったタグ付け手法に従ったものである。本発明では当該規則に従うことを限定するものではない。また、形態素の情報及び形態素間の係り受けの情報をコンピュータが情報処理によって識別できるようにすることができればＸＭＬのタグ付けによる方法には限らない。

【０１１３】

ＧＤＡに基づくタグ付けは基本的に＜タグ名　属性名＝“属性値”＞で表される。図５に示される例では、＜ｓｕ＞で示されるタグは、文（Ｓｅｎｔｅｎｔｉａｌ　ｕｎｉｔ）を表すタグである。図５に示した例では、「九州地方北部では、秋に行われるものに対して（お）くんちと称する場合もある。」の文は、「九州地方北部では」「、」「秋に行われるものに対して（お）くんちと称する場合も」「ある」「。」の三つの文節と句読点との単位を有していることがタグによって判別できる。＜ａｄ＞で示されるタグは、終助詞以外の助詞（ｐａｒｔｉｃｌｅ）、副詞（ａｄｖｅｒｂ）、連体詞などを示すタグであるが、文節０の「九州地方北部では」も全体で副詞的な役割を果たすことを示すことができる。＜ｎ＞で示されるタグは、名詞（ｎｏｕｎ）を示す。＜ｖ＞で示されるタグは、動詞（ｖｅｒｂ）を示す。また、図５に示したタグの他に形容詞（ａｄｊｅｃｔｉｖｅ）を示す＜ａｊ＞タグ等がある。

【０１１４】

属性名ｓｙｎで表される属性は、当該属性が付与されているタグで挟まれた文節又は語等の言語単位間の係り受け関係を示す。属性値ｆ（ｆｏｒｗａｒｄ；前向き）が付与されている文では、当該文を構成する言語単位は一番近い後続の言語単位に係ることを示す。したがって、原則では文節０の「九州地方北部では」は、文節１の「秋に行われるものに対して（お）くんちと称する場合も」へ係り、文節１の「秋に行われるものに対して（お）くんちと称する場合も」は文節２の「ある」に係る。

【０１１５】

しかし統語解析により、文節０の「九州地方北部では」は文節２の「ある」に係り、文節１の「秋に行われるものに対して（お）くんちと称する場合も」は文節２の「ある」に係ることが判別できているため、上述原則はあてはまらない。したがって、係り受けの受ける側ではない「句」（ｐｈｒａｓｅ）であることを示す“ｐ”を各タグに付加することで、係り受けの関係を示すことができる。例えば、＜ａｄｐ＞で示されるタグは、タグ＜ａｄ＞に、句であることを示す“ｐ”が組み合わさったものである。＜ａｄｐ＞タグではさまれた文節は副詞句であって、係り受けの受ける側の文節ではないことを示す。したがって、図５に示した例では、文節１の「秋に行われるものに対して（お）くんちと称する場合も」は、副詞句であって受ける側の文節ではないため、文節０の「九州地方北部では」は、文節１の「秋に行われるものに対して（お）くんちと称する場合も」へ係らずに「ある」に係ることが示される。その他、“ｐ”は「句」であることを明示するために付加

される。

【０１１６】

また、＜ｎ＞で示すタグについても、＜ｎｐ＞とすることで係り受けの受ける側の語ではないことを示すことができる。「九州地方北部」は、「九州」「地方」「北部」と夫々＜ｎ＞で挟まれる形態素に分別でき、「九州」は「地方」に、「地方」は「北部」に係るため“ｐ”は不要である。一方、「催事（催し、イベント）、フェスティバルのこと」では、「催事（催し、イベント）」は「フェスティバル」に係らず「の」に係るため、「フェスティバル」を挟むタグを＜ｎｐ＞とすることで、係り受けの関係を示すことができる。

【０１１７】

なお、「九州」のような場所を表す固有名詞、又は「太郎」のような人の名前を表す固有名詞は、夫々＜ｐｌａｃｅｎａｍｅ＞＜ｐｅｒｎａｍｅ＞のタグによって示すことができる。

【０１１８】

指示代名詞、ゼロ代名詞等の先行する語又は文から参照する形態素については、照応関係を表す属性を用いて表すことができる。ＧＤＡでは、属性名ｉｄを用いて指示代名詞、ゼロ代名詞が先行の語又は文の何れの語を示すかをあらわすことができる。例えば、「右側にボタンがあるので、それを押してください。」という文に対して、人間がこれを読む場合は「それ」が「ボタン」を指すことを自然に補完することができる。しかし、コンピュータで処理する場合は、辞書情報との照合によって「それ」が指示代名詞であることを同定することはできるが、何を示しているかを判別することはできない。そこでＧＤＡでは、「それ」が示す「ボタン」にｉｄ属性を付加し、さらに、ｉｄ属性で示された形態素との等価（ｅｑｕａｌ）関係を示す属性名ｅｑにより、「それ」＝「ボタン」を示すことができる。具体的には「右側にボタンがあるので、それを押してください。」に対し、「右側に＜ｎｐ　ｉｄ＝“Ｂｔｎ”＞ボタン＜／ｎｐ＞があるので、＜ｎｐ　ｅｑ＝“Ｂｔｎ”＞それ＜／ｎｐ＞を押してください。」とすることで（他のタグは省略）、「それ」＝「ボタン」の関係を示すことができる。

【０１１９】

ゼロ代名詞に対しては、ｅｑ属性を付加できる代名詞そのものがない。したがって、「それ」＝「ボタン」を動作の対象とする「押し」という動詞に、対象を明示する情報を付加することで、ゼロ代名詞が表す対象を示すことができる。そこで、タグではさんだ形態素の動作の対象（ｏｂｊｅｃｔ）を示す属性名ｏｂｊにより、「押し」という動作の対象が「ボタン」であることを示すことができる。具体的には、「右側にボタンがあるので、押してください。」という文に対し、「右側に＜ｎｐ　ｉｄ＝“Ｂｔｎ”＞ボタン＜／ｎｐ＞があるので、＜ｖ　ｏｂｊ＝“Ｂｔｎ”＞押し＜／ｖ＞てください。」とすることで、省略された対象との関係を明示することができる。

【０１２０】

また、参照される語と参照する語とが離れている場合であっても、上述のｉｄ属性、ｅｑ属性、ｏｂｊ属性によってその照応関係を示すことができる。例えば、「右側に＜ｎｐ　ｉｄ＝“Ｂｔｎ”＞ボタン＜／ｎｐ＞があります。」「＜ｎｐ　ｅｑ＝“Ｂｔｎ”＞それ＜／ｎｐ＞には×のマークがついています。」「停止する際に＜ｖ　ｏｂｊ＝“Ｂｔｎ”＞押し＜／ｖ＞てください。」とすることによって、第２文の「それ」が「ボタン」を示すこと、及び第３文の「押し」の対象が「ボタン」であることを示すことができる。

【０１２１】

また、各形態素を挟む＜ｎ＞＜ａｄ＞＜ｖ＞等ののタグの属性情報には、形態素（ｍｏｒｐｈｅｍｅ）解析の結果を示す情報が属性名ｍｐｈで付加される。属性値は、形態素解析によって取得できた形態素の品詞情報、基本形情報、発音情報等を示す。具体的には、属性名ｍｐｈに対し、付加情報、品詞情報、活用形情報、基本形情報、及び発音情報を属性値とし、ｍｐｈ＝“付加情報；品詞情報；活用形情報；基本形情報；発音情報”と表す。図５に示した例において「九州」は、品詞情報を名詞＋固有名詞＋地域＋一般で分類す

ることができ、基本形は九州であり「キュウシュウ」と発音することが＜ｍｐｈ＞タグによって明示される。なお、本発明では、形態素解析及び統語解析をｃｈａｓｅｎで提示される方法に基づいて行っているため、形態素の付加情報としてｃｈａｓｅｎという識別情報が付加されている。

【０１２２】

上述のように、文単位検索装置１のＣＰＵ１１はＷｅｂクローリングによって取得した文書データに対し、形態素解析及び統語解析の結果をＧＤＡの規則に則ってタグ付けし、タグ付けした結果であるＸＭＬデータを文書集合接続手段１６を介して文書記憶手段２に記憶させる。文書データをＸＭＬデータで記憶しておくことにより、文単位検索装置１のＣＰＵ１１は当該文書データのタグを文字列解析によって識別し、タグに付加された属性情報を識別することによって各形態素の情報及び文法的関係を特定することができる。

【０１２３】

さらに文単位検索装置１のＣＰＵ１１は、Ｗｅｂクローリングによって取得した文書データを形態素解析する際に、取得した全文書データに出現する単語を抽出して識別番号を割り振りリストで記憶手段１３に記憶する。図６は、実施の形態１における文単位検索装置１のＣＰＵ１１が取得した全文書データから抽出した単語のリストの例を示す説明図である。図６の説明図に示す例では、３１２４５個の単語がリストとして挙げられている。なお、記憶される単語からは、「こと」、「もの」などのありふれた単語は除かれる。接続詞又は冠詞同様一般的すぎる言葉であり、頻繁に出現するにも拘わらず、その単語自体は意味をなさないために検索処理に負担がかかり、検索対象として不適切であるからである。

【０１２４】

３．文書データの文毎の意味のまとまりの定量化

３－１．文毎の意味のまとまりの定義

次に、文単位検索装置１のＣＰＵ１１は、文書記憶手段２で記憶した文書データ中の一文毎に当該文の意味のまとまりを定量的に表す情報を特定する。文の意味のまとまりを定量的に表す情報とは、ユーザが当該文を使用（発話、筆記、聴取又は読解）するときに、ユーザが注目している単語群と、ユーザが各単語に注目する度合い、即ち顕現性（ｓａｌｉｅｎｃｅ）を定量的に示す値（単語の重み値）とで表す。

【０１２５】

各単語の文中での顕現性は、従来の検索サービスによってされてきた出現頻度によって定量化することもできる。しかしながら、出現頻度は文書、又は文書集合全体を母体として求めるものである。したがって、文書毎に各単語の出現頻度を算出することで、文書全体の意味のまとまりを定量的に表すことはできても、文書中での流れに応じて一文毎に動的に変化する文脈を反映した意味のまとまりを表すことはできない。

【０１２６】

また、単語の文中での顕現性は、先行する文での当該単語の注目度、現在の文での当該単語の注目度の遷移をその単語の使用のされ方で文法的に区別して表すことができる。つまり、先行する文で主題（主語）であった単語が現在の文でも主題（主語）である場合は、現在の文で当該単語は一番注目されている顕現性の高い単語であるといえる。これに対し先行する文では出現していないが現在の文で主題（主語）である単語は、現在の文で注目されているものの、前述の主題として使用され続ける場合に比べて顕現性は低いといえる。この顕現性の定式化は、中心化理論（Grosz et al., 1995、Nariyama, 2002、Poesio et al., 2004）として研究が続けられている。

【０１２７】

中心化理論による定式化では、各単語の顕現性をコンピュータ等で定量的に計算するための特徴量として表わされていない。各単語の遷移の仕方が中心化理論で定義される遷移の仕方の何れに属するか否かが判別できるに過ぎない。そこで本発明では各単語の各文での顕現性を定量的に算出する。

【０１２８】

実施の形態１では、単語毎に各文単位での参照確率を算出し、算出した参照確率を各単語の文単位での顕現性を表わす重み値として付与する。

【０１２９】

なぜなら、単語が当該文で注目されているほど、継続して後続の文でも出現又は参照される確率が高いことから、後続の文で出現する確率又は後続の文から参照される確率を参照確率とし、当該単語の顕現性と捉えることができるからである。また、単語が後続の文で出現又は参照される参照確率は、定量的に扱うことが困難な単語の意味を特徴とするのではなく、文単位検索装置１による情報処理によって解析可能な、単語が出現するパターン又は参照するパターンを含む特徴パターンを特定し、特定した特徴パターンと同一の特徴パターンで出現又は参照される単語が実際に後続の文で出現又は参照される割合が参照確率として算出される。

【０１３０】

以下、単語毎の参照確率を各単語の文単位での重み値とし、夫々の重み値が付与された当該文での単語の集合を重み付き単語群という。各文単位の意味のまとまりは、参照確率という定量的な重み値が付与された重み付き単語群で表わすことができる。

【０１３１】

３－２．回帰モデル学習

参照確率の算出は、特定した特徴パターンと同一の特徴パターンが出現した数に対して、同一の特徴パターンのうち当該単語が実際に後続の文で出現又は参照される割合をその参照確率として求める。この際、特定した特徴パターンと同一の特徴パターンが夫々の特徴パターン毎に多量に且つほぼ同数で出現する場合は、統計的に問題なく参照確率を算出することができる。しかし、実際に同一の特徴パターンが出現する数は限られ、信頼に足り得る参照確率を算出するには膨大な文書データが必要となる。したがって、後続の文で出現又は参照されるか否かをその事象の発生の要因である単語の特徴パターンから予測するための回帰式を、特徴パターンと実際に後続の文で出現又は参照されたかの事象とで回帰モデル学習をすることによって求める。

【０１３２】

以下、回帰モデル学習のためのサンプルである特徴パターンに対する「３－２－１．特徴パターンの特定」と特徴パターンを用いた「３－２－２．回帰式の学習」とに段階を分けて説明する。

【０１３３】

３－２－１．特徴パターンの特定

文書記憶手段２に記憶してある文書データ中の文は＜ｓｕ＞で示すタグで挟まれ、当該文で出現する単語、若しくは文の中の指示代名詞又はゼロ代名詞と照応関係にある単語は、タグの属性情報により特定することが可能である。そこで、本発明の文単位検索装置１では、文書記憶手段２で記憶した文書データに対し、特徴パターンを以下のように特定する。

【０１３４】

文書データ中の一の文ｓと、当該文書データ中での一の文に対する先行する文に含まれる単語ｗの対をサンプル（ｓ，ｗ）とする。当該サンプルに対する特徴パターンｆ（ｓ，ｗ）は、以下の特徴量によって特定される。文ｓと、文ｓより先行する文のうち単語ｗが、最近に出現又は参照された文ｓとの距離（文の数）の特徴量（ｄｉｓｔ）、文ｓより先行する文で単語ｗが、最近に出現又は参照された場合、単語ｗが係っている助詞の特徴量（ｇｒａｍ）、及び文ｓより先行する文で単語ｗが出現又は参照された数（ｃｈａｉｎ）の特徴量等を例として挙げることができる。なお、特徴量はこれに限らず、単語ｗが最近のトピックを示す単語であるか否か、又は単語ｗが一人称であるか否か等でもよい。

【０１３５】

文書記憶手段２で記憶した文書データには形態素解析及び統語解析の結果がＧＤＡに則ったタグによって記述されているため、文書データの文字列解析によってタグ＜ｓｕ＞で区切られる文の分別及び計数、各文内のタグで示される品詞情報による助詞の特定、指示

代名詞又はゼロ代名詞で参照するものも含んだ単語の出現回数の計数が可能である。したがって、文単位検索装置１のＣＰＵ１１は、ＧＤＡに則ったタグ及びその属性値を解析することで各サンプルに対する特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎを特定することができる。

【０１３６】

文単位検索装置１のＣＰＵ１１が、文書記憶手段２で記憶しているタグ付け済みの文書データに対しサンプルを抽出し、抽出したサンプルに対して特徴量を求めて特徴パターンを特定し、抽出したサンプルの特徴パターンから参照確率を算出するため回帰式を回帰分析により推定する処理手順について説明する。図７は、実施の形態１における文単位検索装置１のＣＰＵ１１が、文書記憶手段２で記憶しているタグ付け済み文書データからサンプルを抽出し、回帰分析を行って参照確率を算出するための回帰式を推定する処理手順を示すフローチャートである。図７のフローチャートに示す処理は、分別した文単位毎に特徴パターンを特定する処理、及び、特徴パターンと、特定された単語が後続の文単位で出現又は参照されたか否かの判定結果とに基づいて参照確率を算出するための回帰学習を実行する処理に対応する。

【０１３７】

文単位検索装置１のＣＰＵ１１は、文書記憶手段２から文書集合接続手段１６を介してタグ付け済みの文書データを取得する（ステップＳ２１）。ＣＰＵ１１は、取得した文書データに付加されたタグ＜ｓｕ＞を文字列解析によって識別して文に分別する（ステップＳ２２）。次にＣＰＵ１１は、文を示す＜ｓｕ＞内の各タグを文字列解析によって識別し、文に対し当該文で出現する単語又は参照される単語を対応付けてサンプルを抽出する（ステップＳ２３）。抽出したサンプルに対し、タグを文字列解析によって識別してｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎからなる特徴パターンを特定する（ステップＳ２４）。

【０１３８】

ＣＰＵ１１は、分別した文が取得した文書データの終端であるか否かを判断し（ステップＳ２５）、ＣＰＵ１１が、分別した文が文書データの終端でないと判断した場合は（Ｓ２５：ＮＯ）、ＣＰＵ１１は処理をステップＳ２２に戻し、後続の文について＜ｓｕ＞タグを識別することで分別する処理を継続する。分別した文が取得した文書データの終端であるか否かは、例えば現在分別した文を挟む＜ｓｕ＞＜／ｓｕ＞の後に、＜ｓｕ＞タグが後続するかしないかを判断し、後続しないと判断した場合は終端であると判断することができる。

【０１３９】

一方、ＣＰＵ１１が文書データの終端であると判断した場合は（Ｓ２５：ＹＥＳ）、ＣＰＵ１１は、所定の数のサンプルの抽出が終了したか否かを判断する（ステップＳ２６）。ＣＰＵ１１がサンプルの抽出が終了していないと判断した場合は（Ｓ２６：ＮＯ）、ＣＰＵ１１は、処理をステップＳ２１へ戻し、異なるタグ付け済みの文書データを取得し、サンプルの抽出を継続する。

【０１４０】

ＣＰＵ１１がサンプルの抽出が終了したと判断した場合は（Ｓ２６：ＹＥＳ）、ＣＰＵ１１は、抽出したサンプルに対して回帰分析を行い、各特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎに対する回帰式の回帰係数を推定し（ステップＳ２７）、処理を終了する。

【０１４１】

次に、文単位検索装置１のＣＰＵ１１による上述の処理の詳細を、具体例を挙げて説明する。

【０１４２】

図８は、実施の形態１における文書記憶手段２で記憶された文書データ中の文で特定される特徴パターンの例を示す説明図である。図８に示す文s_iでの、当該文s_iと、先行する文に含まれる単語「太郎君」とのサンプル（s_i，太郎君）の特徴パターンｆ（s_i，太郎君）は以下のようにして特定される。現在の文s_iと、先行する文のうち最近に、単語「太郎君」が出現又は参照された文s_{i-1}との距離の特徴量（ｄｉｓｔ）は、s_iの

直後に続く文s_{i+1}までの文の数２であるためｄｉｓｔ＝２である。また、最近「太郎君」が出現又は参照されたs_{i-1}での単語「太郎君」（彼で参照）が係っている助詞は「は」であるため、ｇｒａｍ＝ハである。更に、文s_iより先行の文s_{i-2}，s_{i-1}で単語「太郎君」が出現又は参照されたためｃｈａｉｎ＝２である。したがって、特徴パターンはｆ（s_i，太郎君）＝（ｄｉｓｔ＝２，ｇｒａｍ＝ハ，ｃｈａｉｎ＝２）と特定される。英語の場合、ｇｒａｍは前置詞によって特定される。

【０１４３】

上述のように、文書データ中の文からサンプル（ｓ，ｗ）を抽出し、抽出した全サンプルに対して特徴パターンｆ（ｓ，ｗ）を特定する。

【０１４４】

３－２－２．回帰式の学習

次に、図７のフローチャートに示したステップＳ２７の回帰分析について、詳細な処理を説明する。

【０１４５】

実施の形態１では、Ｌｏｇｉｓｔｉｃ　Ｒｅｇｒｅｓｓｉｏｎモデルに基づいて回帰分析を行う。回帰分析はこれに限らず、ｋＮＮ（ｋ－Ｎｅａｒｅｓｔ　Ｎｅｉｇｈｂｏｒｓ）平滑化＋Ｓｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｒｅｇｒｅｓｓｉｏｎ（ＳＶＲ）モデルなど、他の回帰分析の手法を使用してもよい。

【０１４６】

ｋＮＮ平滑化＋ＳＶＲモデルを使用する場合、扱うことのできる特徴パターンの特徴量として、次の８要素を使用して回帰モデルの学習ができる。８要素とは、前述のｄｉｓｔ、ｇｒａｍ、ｃｈａｉｎに加えて、以下の５要素を特徴量として扱うことができる。一つは、先行の文単位の内で単語ｗを参照した場合の名詞の種別（ｅｘｐ，代名詞：１／非代名詞：０）でもよい。また、他の一つは、その単語ｗが先行の文単位において出現又は参照されている場合に主題であるか否か（ｌａｓｔ＿ｔｏｐｉｃ，ｙｅｓ：１／ｎｏ：０）でもよい。他の一つは単語ｗが先行の文単位において出現又は参照されている場合に主語であるか否か（ｌａｓｔ＿ｓｂｊ，ｙｅｓ：１／ｎｏ：０）でもよい。他の一つは、サンプル（ｓ，ｗ）において、単語ｗが一人称であるか否か（ｐ１，ｙｅｓ：１／ｎｏ：０）でもよい。他の一つは、単語ｗが出現又は参照されている直近の先行の文単位での単語ｗの品詞情報（ｐｏｓ，名詞：１、動詞：２、等）でもよい。さらに他の一つは、単語ｗが文書中のタイトル又は見出しで参照されているか否か（ｉｎ_ｈｅａｄｅｒ、ｙｅｓ：１／ｎｏ：０）でもよい。さらに、音声データに基づいて回帰分析する場合、８要素の内の１つとして、単語の直近の参照箇所の発話時刻からの秒数（ｔｉｍｅ＿ｄｉｓｔ）、単語の直近の参照箇所を含む文節の１音節あたりの発話速度（の話者平均に対する比）（ｓｙｌｌａｂｌｅ＿ｓｐｅｅｄ）、単語の直近の参照箇所を含む文節の，最低発話音高と最高発話音高の周波数比（ｐｉｔｃｈ＿ｆｌｕｃｔ）の内のいずれか一又は複数を使用することができる。音声データの特徴量についても回帰分析することにより、後述するように文単位検索装置１のＣＰＵ１１が言葉のデータとして音声データを受信した場合に、その特徴量から参照確率を算出することができる。

【０１４７】

このように、ｋＮＮ平滑化＋ＳＶＲモデルを使用する場合、より詳細な特徴量に基づいて参照確率を算出することができ、より緻密な参照確率を算出ことができる。

【０１４８】

本実施の形態１では、文s_iの後続の文s_{i+1}で単語ｗが実際に出現又は参照されたか否かを被説明変数、サンプル（s_i，ｗ）に対して特定された特徴パターンのｄｉｓｔ、ｇｒａｍ、ｃｈａｉｎを特徴量とし、全サンプル（ｓ，ｗ）に対して、Ｌｏｇｉｓｔｉｃ　Ｒｅｇｒｅｓｓｉｏｎモデルにより回帰分析する。これにより、ｄｉｓｔ、ｇｒａｍ、ｃｈａｉｎという特徴量が与えられた場合に、s_{i+1}で単語ｗが出現又は参照される確率Ｐｒ（s_{i+1}，ｗ）を算出するための回帰式を得ることができる。

【０１４９】

Ｌｏｇｉｓｔｉｃ Ｒｅｇｒｅｓｓｉｏｎモデルで求められる確率は、一般的に、説明変数（特徴量）ｘ１，ｘ２，…，ｘｎに対して以下の式（１）で求められる。

【０１５０】

【数１】

$$\Pr = \frac{1}{1+\exp(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n)} \quad \cdots (1)$$

【０１５１】

式（１）のパラメータ（回帰係数）b_0，b_1，…，b_nは、学習するサンプルから最尤法によって推定する。本発明で算出する文ｓでの単語ｗの参照確率の回帰分析とは、被説明変数を、後続の文s_{i+1}で出現又は参照されないサンプルは０、出現又は参照されるサンプルは１とし、説明変数を特徴量であるｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎとし、抽出したサンプルを学習して、以下の式（２）のパラメータ（回帰係数）b_0，b_1，b_2，b_3を推定することを指す。

【０１５２】

【数２】

$$\Pr = \frac{1}{1+\exp(b_0 + b_1 dist + b_2 gram + b_3 chain)} \quad \cdots (2)$$

【０１５３】

抽出したサンプルから学習したパラメータ（回帰係数）は、例えばb_0＝－１．４２５、b_1＝－０．５６４、b_2＝１１．０３６、b_3＝３．１１５と推定される（１００００サンプルから回帰分析）。この場合、これらのパラメータを当てはめた式（３）が参照確率を求めるための回帰式である。

【０１５４】

【数３】

$$\Pr = \frac{1}{1+\exp(-1.425 - 0.564 \times dist + 11.036 \times gram + 3.115 \times chain)} \quad \cdots (3)$$

【０１５５】

推定されるパラメータ（回帰係数）b_0，b_1，b_2，b_3の値は、文書記憶手段２で記憶する文書データによって異なる。例えば、文書記憶手段２で記憶する文書データが書き言葉である新聞記事のみからなる場合と話し言葉である発話を文書データに変換したもののみからなる場合とでは、夫々推定されるパラメータは異なる。また、書き言葉として同種の新聞記事のみからなる文書データに対しても、その文書データの量、文書データの文書の内容によって推定されるパラメータの値b_0，b_1，b_2，b_3は異なる。そこで本発明では、話し言葉での回帰分析のために、書き言葉と話し言葉とで区別して文書データを記憶しておき、話し言葉からなる文書データに対しても回帰分析によってパラメータ

を推定し、参照確率を算出するための回帰式を記憶しておく。なお、受付装置４，４，…で受け付ける言葉が、音声入力された発話ではなく文字入力によって書き言葉からなる文章を入力したものに限定されている場合は、話し言葉と書き言葉とで文書データを区別せずに文書記憶手段２で記憶する構成としてもよい。

【０１５６】

以上の回帰分析により、式（３）の回帰式の特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎに対するパラメータが求められる。したがって、文単位検索装置１のＣＰＵ１１が文単位の各単語の特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎからなる特徴パターンを特定することにより、当該特徴パターンを有する単語の参照確率を算出することができる。

【０１５７】

３－３．文単位毎の顕現性の定量化

回帰分析により回帰式が得られたため、文単位検索装置１のＣＰＵ１１は、文単位毎に抽出された単語毎に特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎを特定することにより、単語毎の参照確率を算出することができる。そこで、文単位検索装置１のＣＰＵ１１は、文書記憶手段２で記憶しているタグ付け済みの文書データを取得して文毎に分別し、当該文で出現する単語又は参照する単語に対して特徴パターンを特定し参照確率を算出する。これにより、先行する文の文脈上の意味が反映された文毎の意味のまとまりを定量的に表すことができる。

【０１５８】

文単位検索装置１のＣＰＵ１１が回帰分析後に、文書記憶手段２で記憶している文書データの文毎に、単語及び単語毎の参照確率（重み付き単語群）を算出する処理について以下に説明する。

【０１５９】

文単位検索装置１のＣＰＵ１１は、文書記憶手段２で記憶している文書データを取得して、文書データに含まれる文毎にその文と先行の文とにおける各単語の文法的な特徴パターンを特定し、特定した特徴パターンと回帰式とに基づいて文毎に各単語の参照確率を算出して予め記憶する。

【０１６０】

文単位検索装置１のＣＰＵ１１は、各単語と夫々の単語の参照確率との組（重み付き単語群）を各文単位毎に対応付けて記憶しておく。即ちＣＰＵ１１は、文書集合から取得する全文書の全文について記憶する処理を行なう。一方、ＣＰＵ１１は、後の検索処理において、全文書の全文の内の、受け付けた言葉と文脈上の意味が類似する文を抽出する。したがって、この場合、全文書の全文を一つ一つ読み出して夫々に対応付けられている各文の文脈上の意味を表わす重み付き単語群を読み出すのでは処理の負荷が大きい。

【０１６１】

そこで、文単位検索装置１のＣＰＵ１１は、各文に対して先行の文の文脈上の意味を表わした重み付き単語群を、後の処理で全文書の全文を一つ一つ読み出すことなしに抽出する処理を可能にするために、各文毎に算出した重み付き単語群をデータベース化して索引付けしておく処理を行なう。

【０１６２】

図９及び図１０は、実施の形態１における文単位検索装置１のＣＰＵ１１が、文書記憶手段２で記憶しているタグ付け済みの文書データの文毎に単語の参照確率を算出し、記憶する処理手順を示すフローチャートである。図９及び図１０のフローチャートに示す処理は、文単位毎に、各単語に対して特定した特徴パターンと、特徴パターンに対応する回帰係数とを使用して参照確率を算出する処理、算出した参照確率を単語との組で予め記憶しておく処理に対応する。

【０１６３】

文単位検索装置１のＣＰＵ１１は、文書記憶手段２から文書集合接続手段１６を介してタグ付け済みの文書データを取得する（ステップＳ３０１）。ＣＰＵ１１は、取得した文書データに付加されたタグ＜ｓｕ＞を文字列解析によって識別して文に分別する（ステッ

プＳ３０２）。次にＣＰＵ１１は、文を示す＜ｓｕ＞内の各タグを文字列解析によって識別し、文に対し、当該文で出現する単語又は参照される単語を抽出し（ステップＳ３０３）、当該文書データについて参照確率の算出を行う間は、抽出した単語を一時記憶領域１４で記憶する（ステップＳ３０４）。

【０１６４】

ＣＰＵ１１は、一時記憶領域１４に記憶した、当該文を含む文書データについての単語に対し、単語に付加されたタグを文字列解析によって識別してｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎからなる特徴パターンを特定する（ステップＳ３０５）。次にＣＰＵ１１は、特定した特徴パターンの各特徴量を式（３）に代入し参照確率を算出する（ステップＳ３０６）。

【０１６５】

ＣＰＵ１１は、文に対する各単語の参照確率を、一時記憶領域１４で記憶している全単語に対して算出したか否かを判断する（ステップＳ３０７）。ＣＰＵ１１が全単語に対して参照確率を算出していないと判断した場合は（Ｓ３０７：ＮＯ）、ＣＰＵ１１は、処理をステップＳ３０５に戻し、他の単語についての特徴パターンの特定及び参照確率の算出を継続する。一方、ＣＰＵ１１が全単語に対して参照確率を算出したと判断した場合は（Ｓ３０７：ＹＥＳ）、ＣＰＵ１１は、一時記憶領域１４で記憶している単語及び各単語に対して算出した参照確率の組（重み付き単語群）をｓａｌｉｅｎｃｅ属性を付加して記憶する（ステップＳ３０８）。この際、ＣＰＵ１１は参照確率を所定の値で絞込み、参照確率が所定の値未満である単語については記憶しない。

【０１６６】

次に、ＣＰＵ１１は、現在の文に対して付加した単語及び各単語の参照確率の組（重み付き単語群）を後に抽出することができるように、索引付けして重み付き単語群のデータベースに記憶する（ステップＳ３０９）。ＣＰＵ１１はデータベースを記憶手段１３に記憶してもよいし、文書集合接続手段１６を介して文書記憶手段２に記憶してもよい。なお、ＣＰＵ１１は、索引付けの処理の１つとして以下のような処理を実行する。

【０１６７】

ＣＰＵ１１は例えば、ステップＳ３０８で得られた重み付き単語群の内の、一の単語の参照確率に注目し、一の単語の参照確率が所定値以上であるか否かを判定する。次に、ＣＰＵ１１は重み付き単語群の内の、他の一の単語の参照確率が所定値以上であるか否かを判定する。ＣＰＵ１１は、算出した重み付き単語群を、一の単語の参照確率が所定値以上のグループ、一の単語の参照確率が所定未満のグループのいずれに属するか、さらに一の単語の参照確率が所定値以上のグループに属する場合は、さらに他の単語の参照確率が所定値以上のグループ、他の単語の参照確率が所定値未満のグループのいずれに属するかを判定しておく。ＣＰＵ１１は、このような処理を繰り返して算出した重み付き単語群がいずれのグループに属するかを判定し、属するグループの識別情報に対応付けて記憶しておく。この索引付けの処理は例えば、k-d tree探索アルゴリズムを適用することができる。

【０１６８】

ＣＰＵ１１は、ステップＳ３０１で取得した文書データ中の全文について各文毎に重み付き単語群を対応付ける処理を終了したか否かを判断する（ステップＳ３１０）。ＣＰＵ１１は、文書データ中の全文について各文毎に重み付き単語群を対応付ける処理を終了したか否かを以下のように判断する。例えば、現在の文を挟む＜ｓｕ＞＜／ｓｕ＞の後に、＜ｓｕ＞タグが後続するか否かを判断し、後続しないと判断した場合は終端であると判断することができる。ＣＰＵ１１がステップＳ３０１で取得した文書データ中の全文について各文毎に重み付き単語群を対応付ける処理を終了していないと判断した場合は（Ｓ３１０：ＮＯ）、ＣＰＵ１１は、処理をステップＳ３０２に戻し、次の文に対して処理を継続する。一方、ＣＰＵ１１がステップＳ３０１で取得した文書データ中の全文について各文毎に重み付き単語群を対応付ける処理を終了したと判断した場合は（Ｓ３１０：ＹＥＳ）、ＣＰＵ１１は、文書データで抽出されて一時記憶領域１４に記憶していた単語を消去する（ステップＳ３１１）。

【０１６９】
ＣＰＵ１１は、全文書データについて、単語及び単語の参照確率をｓａｌｉｅｎｃｅ属性によって記憶する処理を終了したか否かを判断する（ステップＳ３１２）。ＣＰＵ１１が全文書データについて、単語及び単語の参照確率をｓａｌｉｅｎｃｅ属性によって記憶する処理を終了していないと判断した場合は（Ｓ３１２：ＮＯ）、ＣＰＵ１１は、処理をステップＳ３０１へ戻し、別の文書データを取得して処理を継続する。ＣＰＵ１１が全文書データについて、単語及び単語の参照確率をｓａｌｉｅｎｃｅ属性によって記憶する処理を終了したと判断した場合は（Ｓ３１２：ＹＥＳ）、ＣＰＵ１１は、単語の参照確率を算出して予め記憶する処理を終了する。
【０１７０】
次に、文単位検索装置１のＣＰＵ１１が図９及び図１０のフローチャートに示した処理を図５に示した文書データに対して行った場合について具体的に説明する。
【０１７１】
図１１は、実施の形態１における文単位検索装置１のＣＰＵ１１が、文書データに示される文書を文毎に分別した一例を示す説明図である。
【０１７２】
文単位検索装置１のＣＰＵ１１は、ステップＳ３０１及びステップＳ３０２の処理により、文書記憶手段２で記憶している文書データから、＜ｓｕ＞タグを識別して文毎に分別する。図１１に示す例では、文はs_1「祭とは、神霊などを祀る儀式。」、s_2「祭礼、祭祀とも呼ばれる。」、s_3「九州地方北部では、秋に行われるものに対して（お）くんちと称する場合もある。」に分別される。文単位検索装置１のＣＰＵ１１によるステップＳ３０３の処理により、文s_1，s_2，s_3から抽出される単語は、単語のリストに記憶された単語と一致する「祭」、「神霊」、「儀式」、「祭礼」、「祭祀」、「九州」、「九州地方」、「九州地方北部」、「秋」、「くんち」、「場合」である（図６参照）。
【０１７３】
文単位検索装置１のＣＰＵ１１は、ステップＳ３０５の処理により、各単語群の文s_3での顕現性（参照確率）を定量的に求めるために、各単語群の特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎからなる特徴パターンを特定する。例えば、文s_3での「九州」（識別番号：９７１４）（図６参照）の特徴パターンは以下のように特定される。
【０１７４】
図１１の説明図に示すように、文s_3での「九州」のｄｉｓｔは、最近出現した文s_3と、後続の文s_4との距離１によりｄｉｓｔ＝１である。また、文s_3での「九州」のｇｒａｍは、最近「九州」が出現した文s_3では「九州」が係るのは助詞ではなく「地方」へ係るために名詞接続と特定できｇｒａｍ＝名詞接続である。文s_3での「九州」のｃｈａｉｎは、s_1からs_3まで「九州」が出現した回数は一回であるのでｃｈａｉｎ＝１である。したがって、特徴パターンｆ（s_3、九州）＝（ｄｉｓｔ＝１，ｇｒａｍ＝名詞接続，ｃｈａｉｎ＝１）と特定される。したがって、文単位検索装置１のＣＰＵ１１は、図９及び図１０のフローチャートのステップＳ３０６の処理により、式（３）に特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎの値を代入して参照確率を算出する。
【０１７５】
ここで、ｇｒａｍで表される特徴量の代入値は、文書記憶手段２で記憶した文書データからサンプル（ｓ，ｗ）を抽出し、夫々に対して算出した単語ｗの参照確率をｇｒａｍ毎に平均値を算出し代入値とする。例えば、抽出したサンプル（ｓ，ｗ）のうち、ｇｒａｍ＝ハを有する単語に対して算出した参照確率の平均値が特徴量ｇｒａｍが「ハ」である場合に代入する値である。実施の形態１では、例として、ｇｒａｍ＝ハの場合はｇｒａｍ＝０．０５４０、ｇｒａｍ＝ガの場合はｇｒａｍ＝０．０２８８、ｇｒａｍ＝ノの場合はｇｒａｍ＝０．０１９８、ｇｒａｍ＝ヲの場合はｇｒａｍ＝０．０１７９、ｇｒａｍ＝ニである場合はｇｒａｍ＝０．０１２４、ｇｒａｍ＝名詞接続である場合は、ｇｒａｍ＝０．００３５２が算出される。
【０１７６】

なお、単語が、助詞「ハ」に係る場合、助詞「ガ」に係る場合、助詞「ノ」に係る場合、助詞「ヲ」に係る場合での、当該単語が後続の文で出現する参照確率の平均値は、「ハ」（主題）「ガ」（主語）「ノ」「ヲ」（目的語）の順に高く、当該文での中心であるか否かを示す中心化理論で定式化している主題＞主語＞目的語…の序列とほぼ整合する。

【０１７７】

文$ｓ_3$での「九州」の参照確率（文$ｓ_4$で「九州」が出現又は参照される確率）は、特定した特徴量に基づいて以下式（４）のように算出される。

【０１７８】

【数４】

$$\mathrm{Pr}=\frac{1}{1+\exp(-1.425-0.564\times 1+11.036\times 0.00352+3.115\times 1)}$$
$$=0.238$$

・・・（４）

【０１７９】

式（４）に示したように、文$ｓ_3$での「九州」の参照確率は０．２３８と算出される。算出された参照確率は文$ｓ_3$に対して記憶される。文単位検索装置１のＣＰＵ１１は、文$ｓ_3$に対し単語をリストで記憶した識別番号で表し、参照確率を対応付けて記憶する。本発明では、文の単位を区切る＜ｓｕ＞タグに対して属性名ｓａｌｉｅｎｃｅを定義し、属性値は単語の識別番号及び参照確率の組を羅列したものと定義して以下のように文毎に単語及び該単語の参照確率（重み付き単語群）を記憶する。

【０１８０】

＜ｓｕ　ｓａｌｉｅｎｃｅ＝“単語$_1$の識別番号：単語$_1$の参照確率　単語$_2$の識別番号：単語$_2$の参照確率　単語$_3$の識別番号：単語$_3$の参照確率…”＞…＜／ｓｕ＞

【０１８１】

図１２は、実施の形態１における文単位検索装置１のＣＰＵ１１が、参照確率を算出した結果を付与して文書記憶手段２に記憶させる文書データの一例を示す説明図である。文$ｓ_3$では「九州」（９７１４）の参照確率（文$ｓ_3$での重み値。以下同様）が０．２３８、「九州地方北部」（９７１６）の参照確率が０．１１５９、…と記憶され、後続の文$ｓ_4$では「九州」（９７１４）の参照確率が０．２３８、「祭」（２２９５３）の参照確率が０．１８３６、…と記憶される。文毎に異なる単語及び参照確率の組（重み付き単語群）が記憶され、文毎の意味のまとまりを表す情報として検索に使用することができる。文$ｓ_3$及び文$ｓ_4$で、「九州」（９７１６）は、同値の参照確率が算出されているが、文$ｓ_5$，文$ｓ_6$，…と続く毎に、九州地方に限らない「祭」についての記述が続く場合は「九州」の参照確率は次第に低下していくと考えられる。

【０１８２】

図１３は、実施の形態１における文単位検索装置１のＣＰＵ１１が、文単位毎に算出した重み付き単語群を索引付けして記憶した場合のデータベースの内容例を示す説明図である。なお、図１３の内容例は、図１２の内容例に示した文$ｓ_4$に対応付けられる重み付き単語群が、図９及び図１０のフローチャートに示したＣＰＵ１１のステップＳ３０９によって索引付けされたデータに相当する。

【０１８３】

図１３に示すように、ＣＰＵ１１は重み付き単語群を、いずれのグループに属するかを示す情報（k-d tree ノードID）に対応付けて記憶しておく。さらにその際、ＣＰＵ１１

は、その重み付き単語群がいずれの文書データの文単位に対応付けられているかを特定できるよう、タグ付け済み文書データのファイル名及び文書データ中の位置（タグ情報）を記憶しておく。これにより、後の処理で受け付けた言葉に対して求めた重み付き単語群と類似する重み付き単語群が対応付けられている文単位を抽出することが容易になる。

【０１８４】

図１４は、文単位検索装置１のＣＰＵ１１により文毎に記憶される単語及び該単語に対して算出された参照確率の組が、文が続くにつれてどのように変化するかを示す説明図である。図１４では、文s_1、文s_2、文s_3、文s_4と続くにつれて、時系列で文脈が動的に変化することに応じて、夫々の文で顕現性の高い単語が夫々異なることが判る。

【０１８５】

４．検索処理

４－１．ユーザから入力された言葉の受け付け

次に、実施の形態１における検索処理について説明する。検索処理は、受付装置４，４，…でユーザから入力されるキーワード又は音声等の言葉を受け付けたことを起点として開始する。

【０１８６】

受付装置４のＣＰＵ４１は、操作手段４５を介してユーザが入力する文字列を検知して一時記憶領域４４に記憶する処理、又は音声入出力手段４７を介してユーザが入力する音声を検知して文字列に変換し一時記憶領域４４に記憶する処理が可能である。また、受付装置４のＣＰＵ４１はユーザが入力する文字列を解析して一文一文に分別する機能を有する。例えば、日本語の場合は句点「。」、英語の場合はピリオド「．」等の所定の文字を識別して分別するのでもよい。また、Ｅｎｔｅｒキーが押下されたことを操作手段４５を介して検知する都度、Ｅｎｔｅｒキーが入力されるまでの文字列を一文と分別するのでもよい。ユーザからの音声入力に対しては、例えば、音声認識機能によって音声を文字列に変換し、変換した文字列から文字列解析によって文に分別してもよいし、無音を検出したところで文に分別してもよい。受付装置４のＣＰＵ４１は、分別した一文一文をテキストデータとして通信手段４８を介して文単位検索装置１へ送信する。

【０１８７】

４－２．受け付けた言葉に対する意味のまとまりの定量化

次に、文単位検索装置１のＣＰＵ１１が、受付装置４，４，…で受け付けた言葉を示すテキストデータを受信した場合に、文書記憶手段２で記憶している文書中の文を検索する処理について説明する。受け付けた言葉を示すテキストデータに対しても、意味のまとまりの定量化、即ち当該テキストデータの単語抽出及び単語の参照確率の算出を行う。これにより、ユーザが言葉を入力するときにユーザの潜在的な意識にある先行の言葉からの流れに応じた文脈を反映した意味のまとまりを表わす情報を、後述する検索処理における検索要求として自動的に作成することができる。

【０１８８】

文単位検索装置１のＣＰＵ１１は、ユーザから受け付けた言葉を示すテキストデータをパケット交換網３及び通信手段１５を介して受付装置４，４，…から受信した場合、一時記憶領域１４に受信した順にテキストデータを記憶すると共に、受信したテキストデータで示される文に対して形態素解析及び統語解析を行う。また、受信したテキストデータで示された文ｓと、文ｓより以前に受信したテキストデータで示された文に出現した単語ｗとの対（ｓ，ｗ）に対し、特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎで表される特徴パターンｆ（ｓ，ｗ）を特定する。

【０１８９】

文単位検索装置１のＣＰＵ１１は、受信したテキストデータの文ｓでの単語ｗの特徴パターンｆ（ｓ，ｗ）を特定した場合、特定した特徴パターンと先に得られた回帰式とに基づいて参照確率を算出する。文単位検索装置１のＣＰＵ１１は、各単語について参照確率を算出し、各単語と各単語について算出した参照確率とを用いて、既に文単位に対応付けて記憶してある重み付き単語群、即ち各単語と各単語の参照確率との組と比較する処理を

おこなって文単位の検索を行う。

【０１９０】

なお、文単位検索装置１のＣＰＵ１１は、受付装置４，４，…からテキストデータのみならず、ユーザから入力された発話の音声データも受信することが可能である。この場合、音声データをテキストデータと同様に音声データに表わされている単語の文法上の特徴パターンを特定することにより、同様の処理を行なう。また、音声データの場合は音声データで得られる特徴を、その単語の顕現性が高いか否かを判断するための特徴量として扱うことも可能である。例えば、ＣＰＵ１１は、単語が出現又は参照された場合に、先行の言葉で出現又は参照されてからの時間差を一つの特徴量として扱うことができる。またＣＰＵ１１は、その単語が出現又は参照された直近の先行の言葉中で、その単語が発声されたときの発話速度及び／又は音声の周波数を他の特徴量として扱うことができる。これらは、テキストデータに変換された後では検知することができない、時間情報又は単語にこめられた感情を定量的に表わす情報である。

【０１９１】

受付装置４がユーザから入力された言葉を受け付けて文単位検索装置１へ送信し、文単位検索装置１のＣＰＵ１１が受付装置４から受信したテキストデータに基づいて文書記憶手段２で記憶している文書データから検索を行う処理手順についてフローチャートを用いて説明する。図１５、図１６、及び図１７は、実施の形態１における文単位検索装置１及び受付装置４の検索処理の処理手順を示すフローチャートである。

【０１９２】

受付装置４のＣＰＵ４１は、ユーザによる文字列入力操作を操作手段４５を介して検知したか否か、又はユーザによる音声入力を音声入出力手段４７を介して検知したか否かを判断する（ステップＳ４０１）。ＣＰＵ４１がユーザによる文字列入力操作又は音声入力を検知していないと判断した場合は（Ｓ４０１：ＮＯ）、ＣＰＵ４１は、処理をステップＳ４０１へ戻し、ユーザによる文字列入力操作又は音声入力を検知するまで待機する。

【０１９３】

一方、受付装置４のＣＰＵ４１がユーザによる文字列入力操作又は音声入力を検知したと判断した場合は（Ｓ４０１：ＹＥＳ）、受付装置４のＣＰＵ４１は、入力された文字列又は音声入力を変換した文字列から、入力された言葉を一文に分別して一時記憶領域４４に記憶し（ステップＳ４０２）、ユーザから入力された言葉をパケット交換網３を介して文単位検索装置１へ送信する（ステップＳ４０３）。

【０１９４】

文単位検索装置１のＣＰＵ１１は、受付装置４から、ユーザによって入力された言葉を受信し（ステップＳ４０４）、ＣＰＵ１１は、受信した言葉を文として一時記憶領域１４に受信順にテキストデータで記憶する（ステップＳ４０５）。このとき、テキストデータ毎に文識別番号を付加して記憶してもよい。

【０１９５】

ＣＰＵ１１は、記憶したテキストデータを形態素解析及び統語解析し（ステップＳ４０６）、解析によって抽出された単語を一時記憶領域１４に記憶する（ステップＳ４０７）。このときＣＰＵ１１は、リストに記憶してある単語と照合し、リストの識別番号で単語を記憶する。

【０１９６】

なお、文単位検索装置１のステップＳ４０７における処理により、一時記憶領域１４には、一連として入力された言葉（発話）の中で一度は出現又は参照された単語が記憶されることになる。なお、ステップＳ４０７における単語の抽出は必ずしも行わなくてもよい。その場合は、リストに記憶してある全単語に対し、後述する特徴パターンの特定の処理を行う。

【０１９７】

ＣＰＵ１１は、一時記憶領域１４に記憶している単語夫々に対し、過去に受信して記憶してあるテキストデータ及びステップＳ４０６の形態素解析及び統語解析の結果に基づい

て、特徴パターンを特定する（ステップＳ４０８）。ＣＰＵ１１は、特定した特徴パターンの特徴量を、予め話し言葉について回帰分析して求めた参照確率を算出するための回帰式に代入し、単語毎に参照確率を算出する（ステップＳ４０９）。ＣＰＵ１１は、一時記憶領域１４で記憶している全単語について参照確率を算出したか否かを判断する（ステップＳ４１０）。ＣＰＵ１１が記憶している全単語について参照確率を算出していないと判断した場合は（Ｓ４１０：ＮＯ）、処理をステップＳ４０８へ戻し、別の単語について特徴パターンの特定及び参照確率の算出の処理を行う。

【０１９８】

ＣＰＵ１１が記憶している全単語について参照確率を算出したと判断した場合は（Ｓ４１０：ＹＥＳ）、一時記憶領域１４に夫々参照確率を算出して記憶している全単語に対し、所定値以上の参照確率が算出された単語に絞り込む（ステップＳ４１１）。参照確率が極端に低い単語を除去することにより、後の演算によるＣＰＵ１１自身への負荷を低減させるためである。ＣＰＵ１１は、受け付けた言葉に対して絞り込まれた単語及び単語の参照確率に基づいて以下のような検索処理を行う。

【０１９９】

これまでの処理により、受け付けた言葉に対し、以前に受け付けた言葉から続く流れ上の意味のまとまりを定量的に表わす単語と単語の参照確率の組（重み付き単語群）を検索要求として生成することができた。以下の検索処理（一点鎖線で囲まれたステップＳ４１２からステップＳ４１６まで）は、受け付けた言葉に対して得られた重み付き単語群と、予め記憶してある文単位の重み付き単語群とを比較し、夫々の重み付き単語群の内の複数の単語の重み値の分布が類似するか否かによって、言葉と文とで意味が類似するか否かを判定し、類似する文を抽出する処理の一例である。

【０２００】

ＣＰＵ１１は、記憶手段１３又は文書記憶手段２のデータベースから、各文に対応付けられて記憶されている単語と単語の参照確率との組（以下重み付き単語群という）を読み出す（ステップＳ４１２）。

【０２０１】

このとき、ＣＰＵ１１は、ある程度類似する重み付き単語群を絞り込んで読み出すことができるように、ステップＳ４１１までの処理で得られた受け付けた言葉に対応付けられる重み付き単語群が、データベースに記憶してある重み付き単語群同様にいずれのグループに属するかを判定する。ＣＰＵ１１は、受け付けた言葉に対応付けられた重み付き単語群が属するグループの重み付き単語群をデータベースから読み出す。これにより、全く類似しない重み付き単語群と比較することを回避し、ある程度類似する重み付き単語群を絞り込んで抽出することができる。

【０２０２】

次にＣＰＵ１１は、ステップＳ４１２で読み出した重み付き単語群から、受け付けた言葉の重み付き単語群と同一の単語を含む重み付き単語群を抽出する（ステップＳ４１３）。ＣＰＵ１１は、抽出した文と同一の単語夫々について、参照確率の差分を算出する（ステップＳ４１４）。ＣＰＵ１１は、同一の単語の数の多い順及び同一の単語の参照確率の差分が小さい順に、抽出した重み付き単語群に類似度を付与し（ステップＳ４１５）、抽出した重み付き単語群が対応付けられている文を文書集合の文書データから読み出す（ステップＳ４１６）。このとき、ＣＰＵ１１は、類似度が所定値以上の重み付き単語群のみに対応する文を読み出してもよい。ＣＰＵ１１は、抽出した文を類似度でソートする（ステップＳ４１７）。

【０２０３】

上述のステップＳ４１２からステップＳ４１７までの処理により、受け付けた言葉に対して得られた重み付き単語群の内の複数の単語の重み値の分布と、類似する重み値の分布を有する重み付き単語群が対応付けられた文を抽出することができる。

【０２０４】

次にＣＰＵ１１は、各文を表すテキストデータを検索結果のテキストデータとして受付

装置４へ通信手段１５を介して送信する（ステップＳ４１８）。

【０２０５】

受付装置４のＣＰＵ４１は、検索結果のテキストデータを通信手段４８を介して受信し（ステップＳ４１９）、受信したテキストデータを表示手段４６を介してモニタ等に表示し（ステップＳ４２０）、処理を終了する。

【０２０６】

受付装置４のＣＰＵ４１は、ユーザからの言葉の入力を検知する都度、一文に分別したテキストデータ又は音声データを文単位検索装置１へ送信する。文単位検索装置１のＣＰＵ１１は、受付装置４からテキストデータ又は音声データ、音声データと共に送信される情報を受信する都度、単語及び単語毎の参照確率を算出して、ユーザから受け付けた言葉に対し、先行の言葉からの流れが反映された意味のまとまりを表わす情報、即ち重み付き単語群を検索要求として作成する。文単位検索装置１のＣＰＵ１１は、受け付けた言葉に対して作成した検索要求（重み付き単語群）に基づいて記憶している文書データから文単位を抽出し、検索結果としてテキストデータを送信する。

【０２０７】

実施の形態１における受付装置４のＣＰＵ４１は、検索結果のテキストデータを受信する都度、モニタ等に表示する。したがって、受付装置４ではユーザから言葉が入力される都度、当該言葉と意味のまとまりが類似するテキストデータが検索結果として表示される。

【０２０８】

なお、受付装置４は、必ずしもユーザから言葉が入力される都度毎回テキストデータを送信し、検索結果を受け付けて表示する構成としなくともよい。例えば、所定の期間中に入力された複数の言葉に相当するテキストデータ又は音声データを文単位検索装置１へ送信し、複数の言葉に対応する検索結果を受け付けて表示する構成でもよい。

【０２０９】

図１５、図１６及び図１７のフローチャートに示した文単位検索装置１のＣＰＵ１１による処理の詳細を具体例を挙げて以下に説明する。

【０２１０】

図１８は、実施の形態１における文単位検索装置１のＣＰＵ１１が、受付装置４から受信したテキストデータに対して特定した特徴パターンの例を示す説明図である。図１８中の文単位S_{i-2}，文単位S_{i-1}，文単位S_iは夫々、受信した各テキストデータで示される文である。

【０２１１】

図１８中の文単位S_iでの、当該文単位s_i及び先行する文単位に含まれる単語「おくんち」とのサンプル（s_i，おくんち）の特徴パターンは以下のようにして特定される。現在の文s_i及び先行する文のうち、単語「おくんち」が最近出現又は参照された文s_{i-2}との距離の特徴量（ｄｉｓｔ）は、ｄｉｓｔ＝３である。また、単語「おくんち」が最近出現又は参照されたs_{i-2}での「おくんち」が係っている格助詞は「って」であるため、ｇｒａｍ＝ッテである。更に、文s_iより先行の文s_{i-2}で単語「おくんち」が出現又は参照されたためｃｈａｉｎ＝１である。したがって、特徴パターンはｆ（s_i，おくんち）＝（ｄｉｓｔ＝３，ｇｒａｍ＝ッテ，ｃｈａｉｎ＝１）と特定される。英語の場合、ｇｒａｍは前置詞によって特定される。

【０２１２】

文単位検索装置１では、話し言葉についても文書記憶手段２で記憶している文書データについて回帰分析を行い、特徴パターンを特定した場合に特徴量を代入することで参照確率を算出することができる回帰式が予め導出されている。したがって、文単位検索装置１のＣＰＵ１１は、文s_iの「おくんち」に対して、特定した特徴パターンの特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎに基づいて参照確率を算出することができる。更に、文単位検索装置１のＣＰＵ１１は、文s_iについて過去に出現又は参照された単語も含めて参照確率を算出し、単語と単語の参照確率とを求める。文単位検索装置１のＣＰＵ１１は、求め

た単語と参照確率とに基づいて、文書記憶手段２で記憶してあるｓａｌｉｅｎｃｅ属性を予め記憶してある文単位から同一の単語の参照確率が所定の値以上である文単位を直接的に抽出する。文単位検索装置１のＣＰＵ１１は、抽出した文を示すテキストデータを通信手段１５を介して受付装置４へ送信する。

【０２１３】

このような文単位検索装置１のＣＰＵ１１の処理により、受信したテキストデータが表す言葉の意味のまとまりを当該言葉毎に単語及び単語の参照確率（重み値）で表すことができる。また、予め文書記憶手段２で記憶してある文書データの各文についても、意味のまとまりを表す単語及び単語の参照確率（重み付き単語群）が記憶されるので、ユーザから受け付けた言葉に対し、抽出された単語の参照確率が類似するか否かによって意味のまとまりが類似する文を直接的に検索することができる。

【０２１４】

（実施の形態２）

実施の形態２では、事前処理の段階で文書記憶手段２で記憶した文書データの文毎に、抽出した単語と単語毎に算出した参照確率との組（重み付き単語群）を顕現性ベクトルとして扱う。さらに、受け付けた言葉に対して算出する単語と単語毎に算出した参照確率との組（重み付き単語群）も顕現性ベクトルとして扱う。そして検索処理の段階においては、実施の形態１に示したように、受け付けた言葉の重み付き単語群の内の複数の単語の重み値の分布と、予め文毎に対応付けてある重み付き単語群の内の複数の単語の重み値の分布とが類似する条件にあるか否かを、同一の単語が記憶されており、同一の単語の差分が小さいか否かで判断した。これに対し、実施の形態２では、夫々の重み付き単語群を顕現性ベクトルで表わし、類似する条件にあるか否かを顕現性ベクトル間の距離の短さによって判断する。

【０２１５】

実施の形態２における、本発明に係る文単位検索装置１を用いた検索システムの「１．ハードウェアの構成及び概要」、及び「２．文書データの取得及び自然言語解析」については、実施の形態１と同様であるため説明を省略する。「３．文書データの文毎の意味のまとまりの定量化」、及び「４．検索処理」について以下に説明するが、実施の形態１と同一の符号を用いて説明する。なお、「３．文書データの文毎の意味のまとまりの定量化」、及び「４．検索処理」についても、実施の形態１と共通する点については詳細な説明を省略する。

【０２１６】

３．文書データの文毎の意味のまとまりの定量化

３－１．文毎の意味のまとまりの定義

実施の形態２では、実施の形態１と同様に文毎の意味のまとまりを定量的に表す情報は、ユーザが当該文を使用（発話、筆記、聴取又は読解）するときに、ユーザが注目している単語群と、ユーザが各単語に注目する度合い、即ち顕現性（ｓａｌｉｅｎｃｅ）を定量的に示す値（単語の重み値）とで表す。また、実施の形態１と同様に、顕現性を定量的に示す重み値として後続の文で出現する又は参照される確率を示す参照確率を使用する。

【０２１７】

３－２．回帰モデル学習

実施の形態２でも、参照確率については実施の形態１の３－２．回帰モデル学習と同様に、文書記憶手段２で記憶している文書データのサンプルに対する回帰分析によって得られる回帰係数を含む回帰式を用いて算出する。

【０２１８】

３－３．文単位毎の顕現性の定量化

実施の形態２でも、文単位検索装置１のＣＰＵ１１は、回帰分析によって得られた回帰係数を含む回帰式を使用して、抽出された単語毎に特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎを特定することで単語毎の参照確率を算出することができる。ここで、単語毎の参照確率をその単語の重み値として付与した重み付き単語群が得られる。実施の形態２では、文

毎の意味のまとまりを表わす重み付き単語群は、単語を夫々一次元とし、単語毎に算出した参照確率を各単語に対応する次元成分の要素として持つ顕現性ベクトルとして扱う。つまり、文書記憶手段２で記憶される文書データ中の文の意味のまとまりは、文書記憶手段２で記憶される文書データから抽出し、図６に示すリストに記憶している３１２４５次元の多次元空間におけるベクトルで表すことができる。

【０２１９】

したがって、（あい，あいだ，あいまい，…，Ｚ，Ｚくん）という単語群からなる３１２４５次元の基底空間に対し、図１１に示した文$ｓ_3$の顕現性ベクトルｖ（$ｓ_3$）は、文$ｓ_3$での９７１４番目の「九州」次元に対応する要素が参照確率の大きさ（重み値）０．２３８で表され、また、９７１６番目の「九州地方北部」次元に対応する要素が参照確率の大きさ０．１１５９で表されるので、（０，０，…，０．２３８，０，０．１１５９，…，０）と３１２４５次元のベクトルで表現して扱うことができる。

【０２２０】

なお、実施の形態２において文単位検索装置１のＣＰＵ１１が参照確率を算出した結果を付与して文書記憶手段２に記憶させる文書データは、実施の形態１の図１１の説明図に示した文書データと同様である。即ち、文書記憶手段２に記憶される文書データには、次元の番号及び次元成分の要素である参照確率の値が記憶される。実施の形態２における文単位検索装置１のＣＰＵ１１が、文書記憶手段２で記憶しているタグ付け済みの文書データの文毎に単語の参照確率を算出し、文毎に対応付けてデータベースに記憶する処理手順は、実施の形態１と同様であるため説明を省く。

【０２２１】

４．検索処理

次に、実施の形態２における検索処理について説明する。「４－１．ユーザから入力された言葉の受け付け」については、受付装置４のＣＰＵ４１が行う処理については実施の形態１と同様である。

【０２２２】

４－２．受け付けた言葉に対する意味のまとまりの定量化

文単位検索装置１のＣＰＵ１１が、受付装置４で受け付けた言葉を示すテキストデータを受信した場合に、文書記憶手段２で記憶している文書中の文を検索する処理について説明する。文単位検索装置１のＣＰＵ１１は、受け付けた言葉を示すテキストデータに対しても、受け付けた言葉の文脈上の意味のまとまりを単語の多次元空間における方向性を示す顕現性ベクトルで表す。

【０２２３】

文単位検索装置１のＣＰＵ１１は、実施の形態１での処理同様に、受付装置４から受信したテキストデータに対してリストに記憶された３１２４５次元の単語に対する特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎで表される特徴パターンを特定する。なお、過去に一連として受信したテキストデータで出現していない単語については、対応する次元成分の要素を０として特徴パターンの特定を省く。

【０２２４】

特徴パターンを表す特徴量ｄｉｓｔ，ｇｒａｍ，ｃｈａｉｎから、回帰式に基づいて次元成分の要素としての参照確率を夫々算出することができる。したがって、文単位検索装置１のＣＰＵ１１は、テキストデータを受信する都度、受信したテキストデータで示される言葉のそれまでの文脈上の意味のまとまりを表わす顕現性ベクトルを算出することができる。

【０２２５】

文単位検索装置１のＣＰＵ１１は、受け付けた言葉に対して算出した顕現性ベクトルと、文書記憶手段２で記憶してある、ｓａｌｉｅｎｃｅ属性を予め付加した文の顕現性ベクトルとの距離をベクトル演算によって直接算出し、距離が短い文を抽出する。図６の各単語を１次元とした場合の３１２４５次元の多次元空間の中で意味のまとまりの方向性が類似する文を検索することができる。文単位検索装置１のＣＰＵ１１は、抽出した文を示す

テキストデータを、通信手段１５を介して受付装置４へ送信する。ベクトル演算を扱うことが可能なコンピュータを用いる場合は、文毎の意味のまとまりを顕現性ベクトルで表して直接的に演算をすることができる。

【０２２６】

文単位検索装置１のＣＰＵ１１が、受付装置４で検索要求の言葉を示すテキストデータを受信し、受信したテキストデータに基づいて文書記憶手段２で記憶している文書データから顕現性ベクトルを用いて検索を行う処理手順について説明する。図１９は、実施の形態２における文単位検索装置１及び受付装置４の検索処理の処理手順を示すフローチャートである。なお、図１９のフローチャートに示す処理手順では、実施の形態１における図１５、図１６及び図１７のフローチャートに示した検索処理の処理手順と同一の処理については、各ステップに同一の符号を用いて詳細な説明を省略する。

【０２２７】

図１９のフローチャートに示す処理手順の内、一点鎖線で囲まれた各ステップＳ５０１からステップＳ５０６までの処理が、実施の形態１における図１５、図１６及び図１７のフローチャートに示した処理手順と異なる。実施の形態１におけるステップＳ４１２からステップＳ４１６までの処理の代わりに、実施の形態２における文単位検索装置１のＣＰＵ１１により実行されるステップＳ５０１からステップＳ５０６までの処理について、以下に説明する。

【０２２８】

文単位検索装置１のＣＰＵ１１は、一時記憶領域１４に夫々参照確率を算出して記憶している全単語に対し、所定値以上の参照確率が算出された単語に絞り込み（ステップＳ４１１）、絞り込まれた各単語と、算出された各単語の参照確率とに基づいて受け付けた言葉の顕現性ベクトルを算出する（ステップＳ５０１）。

【０２２９】

ステップＳ５０１までの処理により、受け付けた言葉に対し、以前に受け付けた言葉から続く流れ上の意味のまとまりを定量的に表わす顕現性ベクトルを検索要求として生成することができた。以下の処理は、受け付けた言葉に対して得られた顕現性ベクトルと、予め記憶してある文毎の顕現性ベクトルとを比較し、夫々の顕現性ベクトルが表わす各単語の重み値の分布が類似するか否かを判定する処理の一例である。

【０２３０】

ＣＰＵ１１は、データベースに記憶してある重み付き単語群即ち顕現性ベクトルを読み出す（ステップＳ５０２）。このとき、ステップＳ４１１までの処理で得られた受け付けた言葉に対応付けられる顕現性ベクトルが、データベースに記憶してある顕現性ベクトル同様にいずれのグループに属するかを判定する。ＣＰＵ１１は、受け付けた言葉に対応付けられた顕現性ベクトルが属するグループの顕現性ベクトルをデータベースから読み出す。これにより、各単語の重み値の分布が類似する顕現性ベクトルをある程度絞り込んで抽出することができる。

【０２３１】

ＣＰＵ１１は、受け付けた言葉に対応付けた顕現性ベクトルと読み出した顕現性ベクトルとの距離を算出する（ステップＳ５０３）。ＣＰＵ１１は、読み出した顕現性ベクトルを、算出した距離が所定値未満である顕現性ベクトルに絞り込み（ステップＳ５０４）、絞り込まれた顕現性ベクトルが対応付けられて記憶されている文を読み出す（ステップＳ５０５）。ＣＰＵ１１は、読み出した文に算出した距離が短い順に類似度を付与する（ステップＳ５０６）。

【０２３２】

実施の形態２における文単位検索装置１のＣＰＵ１１によるステップＳ５０１からステップＳ５０６までの処理により、受け付けた言葉と文脈上の意味合いが類似する文が抽出される。

【０２３３】

その後の抽出された文に対するステップＳ４１７以降の処理は実施の形態１と同様であ

る。

【０２３４】

なお、上述の処理手順の内の、ＣＰＵ１１が受け付けた言葉に対応付けた顕現性ベクトルと、読み出した顕現性ベクトルとの距離を算出するステップＳ５０３の処理は、具体的には以下のように算出する。受け付けた言葉U_iに対応付けた顕現性ベクトルがｖ（u_i）と表わされ、読み出した顕現性ベクトルがｖ（s_i）と表わされる場合、ＣＰＵ１１は以下に示す式（５）のように、コサイン距離を算出する。

【０２３５】

【数５】

$$\frac{\vec{v}(s_i)\cdot\vec{v}(u_i)}{|\vec{v}(s_i)||\vec{v}(u_i)|} = \frac{\sum_{k=1}^{31245}\Pr(w_k\,|\,\mathrm{pre}(s_i))\Pr(w_k\,|\,\mathrm{pre}(u_i))}{\sqrt{\sum_{k=1}^{31245}\Pr(w_k\,|\,\mathrm{pre}(s_i))^2}\sqrt{\sum_{k=1}^{31245}\Pr(w_k\,|\,\mathrm{pre}(u_i))^2}} \quad \cdots (5)$$

【０２３６】

ただし、式（５）に示したように距離を算出した場合、言葉の顕現性ベクトルｖ（u_i）と、読み出した顕現性ベクトルｖ（s_i）とが近いほど、算出したコサイン距離の値は大きくなる。したがって、ＣＰＵ１１はステップＳ５０６において、算出したコサイン距離が大きい順に類似度を付与する。

【０２３７】

このような文単位検索装置１のＣＰＵ１１及び受付装置４のＣＰＵ４１の処理により、受け付けた言葉の意味のまとまりを、当該言葉毎に各単語の参照確率を要素とした顕現性ベクトルで表すことができる。また、予め文書記憶手段２で記憶してある文書データの各文についても、意味のまとまりを表す各単語の参照確率を要素とした顕現性ベクトルが記憶してあるため、単語の多次元空間での方向性を表す顕現性ベクトル間の距離によって、意味のまとまりが類似する文を直接的に検索することができる。

【０２３８】

（実施の形態３）

実施の形態１又は２では、事前処理の段階の「３．文書データの文単位毎の意味のまとまりの定量化」を行なう処理の中で、重み付き単語群として当該単語と単語の参照確率との組、又は顕現性ベクトルを文単位毎に対応付けて記憶しておいた。また、その後の「４．検索の処理」でも「４－２．受け付けた言葉に対する意味のまとまりの定量化」の処理の中で、重み付き単語群として単語と単語の参照確率との組、又は顕現性ベクトルを求めて受け付けた言葉に対応付けた。これに対し、実施の形態３では、文単位又は言葉毎に対応付けた重み付き単語群（単語と単語の参照確率との組、又は顕現性ベクトル）に対し、各単語の顕現性を表わす重み値を、単語に関連の深い他の単語からの連想を加味して算出し直す処理を実行する。

【０２３９】

具体的に連想とは、文単位毎に対応付けられている重み付き単語群の内のある単語が、その文単位又は先行の文単位に出現していない場合であっても、その単語と関連の深い単語の顕現性が高い場合はその単語もその文単位で注目されているはずであることをいう。

したがって、一の単語が注目されている時に同時に注目されやすい単語を関連語とする。そして、各単語の顕現性を表わす重み値に、関連の深い単語の顕現性からの影響を反映させる。

【０２４０】

図２０は、実施の形態３における本発明の検索方法に関わる、一の単語と関連の深い単語の顕現性の影響の概要を示す説明図である。図２０の説明図は、一又は複数のユーザ間の会話の例を表わしている。会話は発話U_1，U_2，U_3，U_4の集合であり、U_1，U_2，U_3，U_4の順になされている。

【０２４１】

ここで、発話U_1，U_2，U_3，U_4にはいずれにも「大阪」は出現していない。また、U_1よりも先行の発話で「大阪」が出現しており、発話U_1，U_2，U_3，U_4夫々での「大阪」の顕現性がゼロではなく、ある程度の高さを有していたとしても、その後「大阪」は出現していないので、発話U_4の時点で「大阪」の顕現性を現す参照確率を定量的に算出した場合、その値が低下している可能性がある。

【０２４２】

しかしながら、「大阪」という単語がそれまでの文単位又は言葉に出現していない場合であっても、発話U_1、U_3には単語「アメリカ村」及び「ミナミ」が出現している。したがって、「アメリカ村」及び「ミナミ」は、発話U_4の時点で参照確率を夫々算出した場合、その値は高いはずである。「アメリカ村」も「ミナミ」も、「大阪」の代表的な繁華街であるから、発話U_4で「大阪」の単語が出現又は参照していなくとも、「アメリカ」又は「ミナミ」が出現していることによって、関連の深い「大阪」の顕現性は本来、高くなるはずである。したがって、図２０の例では、発話U_4における「大阪」の顕現性を現す参照確率は、高い値を有しているはずである。

【０２４３】

そこで、実施の形態３では、文単位又は言葉毎に対応付けられる各単語の顕現性を表わす重み値を、関連する単語（関連語）の顕現性を考慮して算出し直す。

【０２４４】

参照確率を関連語の顕現性を考慮した重み値に算出し直すためにはまず、文単位検索装置１は、いずれの単語同士の関連が深いのかを表わす情報を先に取得しておく必要がある。そして次に、文単位毎に算出されている各単語の参照確率に、関連の深さを表わす関連度の影響を反映しておく。具体的には、例えば上述の例を用いた場合、「アメリカ村」の「大阪」への関連度を定量的に算出しておく。次に既に算出されている「アメリカ村」の参照確率へ、「大阪」への関連度の効果を反映させて、その文単位での「大阪」の顕現性を表わす重み値として算出し直して記憶しておく。

【０２４５】

そこで、実施の形態３ではまず、文単位検索装置１は、各単語の一の単語への関連度が重み値として付与された、一の単語に対する重み付き関連語群を作成する。具体的には、実施の形態１又は２において、「３－３．文単位毎の顕現性の定量化」の処理によって文単位毎に対応付けられて記憶されている重み付き単語群、即ち単語と単語の参照確率との組又は顕現性ベクトルを利用して、文単位検索装置１が各単語の重み付き関連語群を作成する。文単位検索装置１は、文書集合全体から抽出される各単語について、夫々の単語に対する重み付き関連語群を作成し、記憶しておく。

【０２４６】

そして次に、文単位検索装置１は、文単位毎に対応付けられて記憶されている重み付き単語群、即ち単語と単語の参照確率との組又は顕現性ベクトルの各単語の参照確率へ、各単語に関連が深い単語の参照確率からの影響を、関連度を利用して反映させ、各単語の重み値を算出し直して記憶する。

【０２４７】

さらに、文単位検索装置１は検索処理において、各言葉に対応付けた重み付き単語群、即ち単語と単語の参照確率との組又は顕現性ベクトルについても同様に関連度を利用して

各単語の重み値を算出し直す。文単位検索装置１は、受け付けた言葉に対応する単語と各単語に対して算出し直した重み値に基づいて、検索処理を行なう。

【０２４８】

以下に、文単位検索装置１のＣＰＵ１１が、各単語に対する重み付き関連語群の作成する処理について、「３－４．関連語群の作成」の節を追加して説明する。また、作成された関連語群を使用して、「３－３．文単位毎の顕現性の定量化」において算出した参照確率を関連を加味した重み値に算出し直す処理について、「３－５．連想の加味した意味のまとまりの定量化」の節を追加して説明する。「４－２．受け付けた言葉に対する意味のまとまりの定量化」において算出した参照確率を関連を加味した重み値に算出し直して検索を実行する処理について、「４－２’．受け付けた言葉に対する連想を加味した意味のまとまりの定量化」の節を設けて説明する。

【０２４９】

なお、実施の形態３における、本発明に係る文単位検索装置１を用いた検索システムの「１．ハードウェアの構成及び概要」、及び「２．文書データの取得及び自然言語解析」については、実施の形態１と同様であるため説明を省略する。「３．文書データの文毎の意味のまとまりの定量化」、及び「４．検索処理」について以下に説明するが、実施の形態１と同一の符号を用いて説明する。なお、「３．文書データの文毎の意味のまとまりの定量化」、及び「４．検索処理」についても、実施の形態１と共通する点については詳細な説明を省略する。

【０２５０】

３－４．関連語群の作成

関連語群は、図６で示した説明図で抽出されている全単語について一単語ずつ、文単位検索装置１によって以下の処理が行なわれることにより作成される。

【０２５１】

まず、文単位検索装置１は、「３－３．文単位毎の顕現性の定量化」で全ての文単位毎に対応付けられて記憶されている重み付き単語群から、一の単語の参照確率が所定値以上の重み付き単語群を抽出する。これは、上述のように関連語を、一の単語が注目されている時に同時に注目されやすい単語とするからであり、一の単語が注目されていない文単位が除去されるようにするためである。

【０２５２】

次に文単位検索装置１は、上述の処理で抽出された、一の単語の参照確率が所定値以上の重み付き単語群を統合する。具体的には、各重み付き単語群の各単語の参照確率に、その重み付き単語群に含まれる一の単語の参照確率による重み付けをして各単語の参照確率を平均化する。一の単語の参照確率による重み付けを行うのは、一の単語の参照確率がより高い重み付き単語群の各単語に対する参照確率を使用するためである。

【０２５３】

そして、全単語についての重み付き関連語群を同様に扱うため、重み付き関連語群の各単語の重み値を正規化する。

【０２５４】

以下に、本発明に係る文単位検索方法を実施する文単位検索装置１のＣＰＵ１１が、関連語群を作成する処理について説明する。図２１及び図２２は、実施の形態３における文単位検索装置１のＣＰＵ１１が関連語群を作成する処理手順を示すフローチャートである。図２１及び図２２のフローチャートに示す処理は、一の単語について、その重み値が所定値以上である単語群を抽出する処理、抽出した単語群の各単語の重み値を統合して関連度として各単語に付与した関連単語群を作成する処理、一の単語に対応付けて記憶しておく処理、各単語について各処理を実行する処理に対応する。

【０２５５】

文単位検索装置１のＣＰＵ１１は、記憶手段１３に記憶してあるリストから一の単語を選択する（ステップＳ６０１）。ＣＰＵ１１は、文書記憶手段２から文書集合接続手段１６を介してタグ付け済みの文書データを取得する（ステップＳ６０２）。ＣＰＵ１１は、

取得した文書データに付加されたタグ＜ｓｕ＞を文字列解析によって識別し、文単位を読み出す（ステップＳ６０３）。次にＣＰＵ１１は、＜ｓｕ＞内に記憶してあるｓａｌｉｅｎｃｅ属性を読み出し（ステップＳ６０４）、ｓａｌｉｅｎｃｅ属性に記憶してある単語及び単語の参照確率の組（重み付き単語群）の内、ステップＳ６０１で選択した一の単語の参照確率が所定値以上であるか否かを判断する（ステップＳ６０５）。

【０２５６】

ＣＰＵ１１が参照確率が所定値未満である（選択した一の単語が対応付けられていない）と判断した場合（Ｓ６０５：ＮＯ）、ＣＰＵ１１は、処理をステップＳ６０３へ戻して、後続の文単位を読み出し（Ｓ６０３）、ステップＳ６０４及びステップＳ６０５の処理を行なう。

【０２５７】

ＣＰＵ１１が参照確率が所定値以上であると判断した場合（Ｓ６０５：ＹＥＳ）、ＣＰＵ１１は、ステップＳ６０４でｓａｌｉｅｎｃｅ属性で読み出した重み付き単語群を一時記憶領域に記憶する（ステップＳ６０６）。

【０２５８】

ＣＰＵ１１は、ステップＳ６０２で取得した文書データの全文単位についてステップＳ６０４からステップＳ６０６までの処理を実行したか否かを判断する（ステップＳ６０７）。ＣＰＵ１１が全文単位について処理を実行していないと判断した場合（Ｓ６０７：ＮＯ）、ＣＰＵ１１は、処理をステップＳ６０３へ戻して、後続の文単位を読み出し（Ｓ６０３）、ステップＳ６０４からステップＳ６０６までの処理を実行する。

【０２５９】

ＣＰＵ１１が全文単位について処理を実行したと判断した場合（Ｓ６０７：ＹＥＳ）、ＣＰＵ１１は、全文書データについて、選択した一の単語の参照確率が所定値以上である重み付き単語群を抽出したか否かを判断する（ステップＳ６０８）。ＣＰＵ１１が全文書データについて選択した一の単語の参照確率が所定値以上である重み付き単語群を抽出していないと判断した場合（Ｓ６０８：ＮＯ）、ＣＰＵ１１は、処理をステップＳ６０２へ戻して次の文書データを取得して（Ｓ６０２）ステップＳ６０３からステップＳ６０７までの処理を実行する。

【０２６０】

ＣＰＵ１１が全文書データについて選択した一の単語の参照確率が所定値以上である重み付き単語群を抽出したと判断した場合（Ｓ６０８：ＹＥＳ）、ＣＰＵ１１は、ステップＳ６０６の処理によって抽出され、一時記憶領域１４に記憶してある重み付き単語群の集合を、夫々での一の単語の参照確率で重み付けした重み値の総和を夫々の単語に対して算出することにより作成する（ステップＳ６０９）。

【０２６１】

ＣＰＵ１１は、ステップＳ６０９において作成した一の単語の参照確率が所定値以上である重み付き単語群の総和、即ち総和された重み付き単語群の各単語の重み値を正規化する（ステップＳ６１０）。

【０２６２】

ＣＰＵ１１は、ステップＳ６１０で正規化された一の単語の参照確率が所定値以上である重み付き単語群を、各重み値を関連度とする関連語群としてステップＳ６０１で選択した一の単語に対応付けて記憶手段１３に、又は文書集合接続手段１６を介して文書記憶手段２に記憶する（ステップＳ６１１）。

【０２６３】

次に文単位検索装置１のＣＰＵ１１は、記憶手段１３に記憶してあるリストの全単語について関連語群を作成して記憶したか否かを判断する（ステップＳ６１２）。ＣＰＵ１１が全単語について関連語群を作成して記憶していないと判断した場合（Ｓ６１２：ＮＯ）、ＣＰＵ１１は、処理をステップＳ６０１へ戻して次の一の単語を選択し（Ｓ６０１）、選択した単語についてステップＳ６０２からステップＳ６１１までの処理を実行する。

【０２６４】

ＣＰＵ１１が全単語について関連語群を作成して記憶したと判断した場合（Ｓ６１２：ＹＥＳ）、ＣＰＵ１１は処理を終了する。

【０２６５】

なお、ステップＳ６０５において文単位検索装置１のＣＰＵ１１は、単純に、参照確率が所定値以上であるか否かを判断するのではなく、以下のような正規化処理を行ってから所定値との比較を行うようにしてもよい。例えば、文単位検索装置１のＣＰＵ１１は、文単位に対応付けられている各単語の参照確率の二乗の総和が「１」になるように、全参照確率の二乗和の二乗根で各参照確率を除算することによって正規化を行う。

【０２６６】

なお、ステップＳ６１０における正規化についても、各単語の重み値の二乗の総和が１になるように正規化する。例えば、文単位検索装置１のＣＰＵ１１は、全重み値の二乗和の二乗根により、各重み値を除算することによって正規化を行う。

【０２６７】

次に、実施の形態３における文単位検索装置１のＣＰＵ１１が、図２１及び図２２のフローチャートに示した処理を一の単語について行った場合に作成される関連語群の具体例を示す。

【０２６８】

図２３は、実施の形態３における文単位検索装置１のＣＰＵ１１によって関連語群が作成される場合の、各処理の過程での重み付き単語群の例を示す説明図である。なお、図２３の説明図に示す例は、文単位検索装置１のＣＰＵ１１によって、一の単語「アメリカ村」の参照確率が所定値（０．２）以上の重み付き単語群が抽出された場合の例である。図２３（ａ）は、図２１及び図２２のフローチャートに示したステップＳ６０５におけるＣＰＵ１１の処理により抽出されて、一時記憶領域１４に記憶されている重み付き単語群GW_1，GW_2，GW_3を示している。図２３（ｂ）は、同様にステップＳ６０７におけるＣＰＵ１１の処理により、一の単語の参照確率により重み付けされる重み付き単語群GW_1’，GW_2’，GW_3’を示している。図２３（ｃ）は、同様にステップＳ６０９におけるＣＰＵ１１の処理により、重み付けされて総和された重み付き単語群ＧＷ’’を示している。

【０２６９】

図２３（ａ）に示すように、一の単語「アメリカ村」の重み値（参照確率）が所定値０．２以上の重み付き単語群GW_1，GW_2，GW_3が抽出されている。

【０２７０】

図２３（ｂ）に示されている重み付き単語群GW_1’，GW_2’，GW_3’の、各単語の重み値には夫々の重み付き単語群中の一の単語「アメリカ村」の重み値（参照確率）が乗算されている。図２３（ａ）に示された単語群GW_1，GW_2，GW_3に対し、図２３（ｂ）に示された単語群GW_1’，GW_2’，GW_3’の各単語の重み値は、以下のようにして一の単語「アメリカ村」の重み値（参照確率）が乗算されている。例えば、重み付き単語群GW_1の各単語の重み値は、アメリカ村の重み値（参照確率）が０．６であるため、アメリカ村の参照確率で重み付けされて以下のようになる。

【０２７１】

単語群GW_1’：（秋：０（０．６×０），アメリカ村：０．３６（０．６×０．６），・・・，大熊座：０（０．６×０），大阪：０．１２（０．６×０．２），大鹿：０（０．６×０），・・・）

【０２７２】

つまり、一の単語「アメリカ村」の重み値が高いほど、他の単語の重み値の影響が反映される。

【０２７３】

図２３（ｃ）に示されている重み付き単語群ＧＷ’’の、各単語の重み値は、図２３（ｂ）に示したように夫々一の単語「アメリカ村」の重み値（参照確率）で重み付けされた重み値が単語毎に総和されている。図２３（ｃ）に示された単語群ＧＷ’’の各単語の重

み値は、図２３（ｂ）に示された単語群GW_1’，GW_2’，GW_3’以下のように総和される。

【０２７４】

単語群ＧＷ’’：（秋：０．０３（＝０＋０．０３＋０），アメリカ村：０．４９（＝０．３６＋０．０９＋０．０４），…，大熊座：０（＝０＋０＋０），大阪：０．２８（＝０．１２＋０．１２＋０．０．０４），大鹿：０（＝０＋０＋０），…）

【０２７５】

また、重み付けされて総和されることにより統合された重み付き単語群ＧＷ’’の各単語の重み値は、文単位検索装置１のＣＰＵ１１の処理により正規化される。

【０２７６】

正規化の処理についてはその方法は問わないが、例えば、文単位検索装置１のＣＰＵ１１は、各単語の重み値を二乗し、二乗した値の和の二乗根を算出し、各単語の重み値で割って、重み付き単語群ＧＷ’’の各単語の重み値を正規化するようにしてもよい。

【０２７７】

また、重み付けされて総和されることにより統合された重み付き単語群ＧＷ’’を、各単語を一次元とし、各単語の重み値を各次元方向の要素として多次元ベクトルである関連度ベクトルで表現した場合は、各重み値（要素）を多次元ベクトルのノルムで割ることにより、多次元ベクトルを正規化するようにしてもよい。このとき、ノルムはユークリッドノルムとは限らない。

【０２７８】

このように総和して正規化した結果の重み付き単語群が、文単位検索装置１のＣＰＵ１１により「アメリカ村」の関連語群として作成される。以下に示す例は、単語「アメリカ村」の関連語群の一例である。なお、各単語は、重み値の大きい順に列挙されている。

【０２７９】

関連語群（「アメリカ村」）＝（アメリカ村：０．６４７，アメリカ：０．３６９，大阪：０．２５８，村：０．１５９，防犯カメラ：０．１３９，カメラ：０．１３９，チェックアウト：０．１２９，アウト：０．１２９，中：０．１２８，女性：０．１２０，男：０．１０２，中央：０．０９８，犯行：０．０９２，人：０．０８７，たこ焼き：０．０８２，心斎橋：０．０７５，ミナミ：０．０７４，警察：０．０７３，時間：０．０７１，公園：０．０６５，昭和：０．０６４，今回：０．０６３，数：０．０６１，なんば：０．０６０，御津：０．０６０，ランドローバー（登録商標）：０．０５９，ローバー（登録商標）：０．０５９，名前：０．０５９，プラン：０．０５７，道頓堀：０．０５５，立川：０．０５５，ナンバー：０．０５４，西鉄：０．０５３，サツ：０．０５２，伊那：０．０５０，オリジナルステッカー：０．０４９，ステッカー：０．０４９，イン心斎橋：０．０４９，御堂筋線：０．０４９，…）

【０２８０】

なお上の例は、文書集合（ＧＤＡタグ付き毎日新聞コーパスhttp://www.gsk.or.jp/catalog.html参照）を使用して実際に作成した「アメリカ村」の関連語群である。

【０２８１】

上述の「アメリカ村」の関連語群の具体例に示したように、例えば、「アメリカ村」が注目されている場合、「大阪」は他の単語よりも注目される関連語であることを重み値によって定量的に表わすことができる。したがって、この関連語群の各単語の重み値は一の単語への関連度を表わしているということができる。上述の具体例では「アメリカ村」の「大阪」への関連度は、０．２５８である。

【０２８２】

以下、単語w_jに対して作成した関連語群の各重み値、即ち単語w_jの単語w_kへの関連度を$b_{j,k}$と表わす。一の単語w_jの関連語群はbw_j＝（w_1：$b_{j,1}$、w_2：$b_{j,2}$、・・・、w_n：$b_{j,n}$）と表わされる。なお、関連語群を関連度ベクトルとして表わす場合、bw_j＝（$b_{j,1}$、$b_{j,2}$、…、$b_{j,n}$）と表現する。

【０２８３】

文単位検索装置１のＣＰＵ１１は、上述のような処理を、図６の説明図に示した全単語について繰り返し行って各単語の関連単語群を作成し、文書記憶手段２又は文単位検索装置１の記憶手段１３に記憶しておく。このように、文書集合に出現する単語全てについて夫々関連度が定量的に算出されて付与された関連語群を作成して記憶しておくことにより、文単位毎の意味のまとまりを表わす重み付き単語群に対し、関連語の関連度による影響を反映させることができる。

【０２８４】

３－５．連想を加味した意味のまとまりの定量化

次に、文単位毎に記憶されている重み付き単語群、即ち単語と各単語の参照確率との組又は顕現性ベクトルに、作成された関連語群の各単語の関連度を反映させる。具体的には、文単位検索装置１は、既に算出されて記憶されている各単語の参照確率を読み出し、一の単語の重み値として、各単語の参照確率に各単語から一の単語への関連度を乗算した値を算出し直して記憶する。

【０２８５】

図２４は、実施の形態３における文単位検索装置１のＣＰＵ１１が、各文単位に対応付けられて記憶されている重み付き単語群の各単語の重み値を算出し直す処理手順を示すフローチャートである。図２４のフローチャートに示す処理は、文単位毎に対応付けられた重み付き単語群の各単語の重み値を、関連度を使用して付与し直す処理に対応する。

【０２８６】

文単位検索装置１のＣＰＵ１１は、文書記憶手段２から文書集合接続手段１６を介してタグ付け済みの文書データを取得する（ステップＳ７１）。ＣＰＵ１１は、取得した文書データに付加されたタグ＜ｓｕ＞を文字列解析によって識別し、文単位を読み出す（ステップＳ７２）。

【０２８７】

次にＣＰＵ１１は、＜ｓｕ＞内に記憶してあるｓａｌｉｅｎｃｅ属性を読み出し（ステップＳ７３）、ｓａｌｉｅｎｃｅ属性で対応付けて記憶してある単語及び単語の参照確率の組（重み付き単語群）の、各参照確率を関連語群を使用して連想を加味した重み値に算出し直す（ステップＳ７４）。ＣＰＵ１１は、各単語及び各単語についてステップＳ７４で算出し直した重み値の組である重み付き単語群（顕現性ベクトル）をｓａｌｉｅｎｃｅ属性を付加して記憶し直す（ステップＳ７５）。

【０２８８】

次にＣＰＵ１１は、ステップＳ７２で読み出した文単位が文書データの終端であるか否かを判断する（ステップＳ７６）。現在の文が取得した文書データの終端であるか否かは、現在の文を挟む＜ｓｕ＞＜／ｓｕ＞の後に、＜ｓｕ＞タグが後続するかしないかを判断し、後続しないと判断した場合は終端であると判断することができる。ＣＰＵ１１が文書データの終端でないと判断した場合は（Ｓ７６：ＮＯ）、ＣＰＵ１１は、処理をステップＳ７２に戻し、次の文単位に対して処理を継続する。一方、ＣＰＵ１１が文書データの終端であると判断した場合は（Ｓ７６：ＹＥＳ）、ＣＰＵ１１は、全文書データについて、重み付き単語群の各単語の重み値を算出し直してｓａｌｉｅｎｃｅ属性で対応付けて記憶する処理を終了したか否かを判断する（ステップＳ７７）。

【０２８９】

ＣＰＵ１１が全文書データについて、重み付き単語群の各単語の重み値を算出し直してｓａｌｉｅｎｃｅ属性によって記憶する処理を終了していないと判断した場合は（Ｓ７７：ＮＯ）、ＣＰＵ１１は、処理をステップＳ７１へ戻し、別の文書データを取得して処理を継続する。ＣＰＵ１１が全文書データについて、重み付き単語群の各単語の重み値を算出し直してｓａｌｉｅｎｃｅ属性によって記憶する処理を終了したと判断した場合は（Ｓ７７：ＹＥＳ）、ＣＰＵ１１は処理を終了する。

【０２９０】

なお、文単位検索装置１のＣＰＵ１１は、ステップＳ７４における各単語の重み値の算出し直しを以下のような処理を行なうことによって実現する。

【０２９１】
図２５は、実施の形態３における文単位検索装置１のＣＰＵ１１が、各文単位に対応付けられて記憶されている重み付き単語群の各単語の重み値を算出し直す処理手順の詳細を示すフローチャートである。図２５のフローチャートに示す処理は、各単語の関連度を重み付き単語群の重み値に乗算する処理、乗算した重み値に基づいて各単語の重み値を付与し直す処理に対応する。
【０２９２】
文単位検索装置１のＣＰＵ１１は、図２４のフローチャートのステップＳ７４で読み出したｓａｌｉｅｎｃｅ属性で対応付けて記憶してある重み付き単語群の各単語及び各単語の参照確率を読み出し、一時記憶領域１４に記憶しておく（ステップＳ８１）。ＣＰＵ１１は、各単語の内の一の単語を選択し（ステップＳ８２）、選択した一の単語の重み値について以下の処理を行なう。
【０２９３】
ＣＰＵ１１は、記憶手段１３又は文書記憶手段２に記憶してある各単語の関連度が付与された関連語群を読み出す（ステップＳ８３）。ＣＰＵ１１は、読み出した各単語の関連語群から、各単語から一の単語への関連度を取得する（ステップＳ８４）。ＣＰＵ１１は、取得した各単語から一の単語への関連度を一時記憶領域１４に記憶してある各単語の参照確率に夫々乗算し、和を算出する（ステップＳ８５）。
【０２９４】
ＣＰＵ１１によりステップＳ８５で算出された和が、一の単語について、関連語による連想が加味されて算出し直された顕現性を表わす重み値である。
【０２９５】
ＣＰＵ１１は、ステップＳ８１で一時記憶領域１４に記憶してある各単語全てについて、重み値を算出し直したか否かを判断する（ステップＳ８６）。ＣＰＵ１１が各単語全てについて重み値を算出し直していないと判断した場合（Ｓ８６：ＮＯ）、ＣＰＵ１１は、処理をステップＳ８２へ戻して、次の単語についてステップＳ８２からステップＳ８５までの重み値を算出し直す処理を実行する。ＣＰＵ１１が各単語全てについて重み値を算出し直したと判断した場合（Ｓ８６：ＹＥＳ）、ＣＰＵ１１は、処理を図２４のフローチャートのステップＳ７５へ戻す。
【０２９６】
なお、図２４のフローチャートの内のステップＳ７４及び図２５のフローチャートに示した文単位検索装置１のＣＰＵ１１による重み値を算出し直す処理は、実施の形態１における参照確率を算出して各文単位毎の顕現性を現す重み値として記憶する処理の中で実行してもよい。具体的には、図９のフローチャートに示した処理手順の内のステップＳ３０６とステップＳ３０７の処理の間にステップＳ７４及び図２５のフローチャートに示した処理を実行する構成でもよい。
【０２９７】
図２４及び図２５のフローチャートに示したＣＰＵ１１の処理手順において、文単位検索装置１のＣＰＵ１１が、各単語について算出した参照確率を連想を加味した重み値に算出し直す処理について、具体的な例を以下に示す。
【０２９８】
例えば、単語「アメリカ村」について作成した関連度群を使用する場合、文単位検索装置１により、ある文単位における「大阪」の顕現性を現す重み値を以下のように算出し直す。なお、「アメリカ村」について作成した関連度群の「大阪」への関連度は「０．３」であるとする。ある文単位に対応付けて記憶されている単語に「アメリカ村」が含まれており、「アメリカ村」の参照確率が０．４であり、「大阪」は含まれていない場合であっても、文単位検索装置１のＣＰＵ１１は、「アメリカ村」の参照確率０．４に、「アメリカ村」から「大阪」への関連度０．３を乗算して、その文単位における「大阪」の重み値は「０」ではなく「０．１２」に算出し直す。
【０２９９】

ここで、文脈連想を加味した単語w_kの各文s_iにおける顕現性を表わす重み値を、ｓａｌｉｅｎｃｅ（w_k｜ｐｒｅ（s_i））と表わす。また、単語w_kの各文s_iにおける参照確率をＰｒ（w_k｜ｐｒｅ（s_i））とする。この場合、単語w_jの単語w_kへの関連度を反映した場合、ｓａｌｉｅｎｃｅ（w_k｜ｐｒｅ（s_i））＝$b_{j,k}$×Ｐｒ（w_j｜ｐｒｅ（s_i））と算出し直される。なお、単語w_kへの関連度を有する単語w_jは他にも存在するので、全単語w_j（ｊ＝１，…，Ｎ）からの関連度の影響をも反映させて、文単位検索装置１は以下に示す式（６）のように各単語の重み値を算出し直す。

【０３００】

【数６】

$$salience(w_k \mid \mathrm{pre}(s_i)) = \sum_{j=1}^{N} b_{j,k} \times \Pr(w_j \mid \mathrm{pre}(s_i)) \quad \cdots \quad (6)$$

【０３０１】

したがって、文単位検索装置１のＣＰＵ１１は、以下に示す式（７）のように文単位Ｓにおける各単語w_k（ｋ＝１，…，Ｎ）の重み値を算出し直す。

【０３０２】

【数７】

$$\begin{aligned}\vec{V}(S_i) &= \left(salience(w_1 \mid \mathrm{pre}(s_i)), \cdots, salience(w_k \mid \mathrm{pre}(s_i)), \cdots, salience(w_N \mid \mathrm{pre}(s_i))\right)^T \\ &= \begin{bmatrix} b_{1,1} & \cdots & b_{N,1} \\ & \ddots & \\ \vdots & b_{j,k} & \vdots \\ & \ddots & \\ b_{1,N} & \cdots & b_{N,N} \end{bmatrix} \begin{bmatrix} \Pr(w_1 \mid \mathrm{pre}(s_i)) \\ \vdots \\ \Pr(w_j \mid \mathrm{pre}(s_i)) \\ \vdots \\ \Pr(w_N \mid \mathrm{pre}(s_i)) \end{bmatrix} \\ &= (\vec{b}_{w_1}, \cdots, \vec{b}_{w_k}, \cdots, \vec{b}_{w_N})\vec{v}(s_i) \end{aligned} \quad \cdots \quad (7)$$

【０３０３】

なお、式（７）の最終行の式は、実施の形態２に示したように、重み付き単語群、即ち単語と単語の参照確率との組を顕現性ベクトルｖ（s_i）として表現した場合に、ｓａｌｉｅｎｃｅ（w_k｜ｐｒｅ（s_i））をｋ番目の要素として有する連想を加味した後の顕現性ベクトルＶ（s_i）の各単語の重み値の算出の原理を表わす。

【０３０４】

この場合、各ｂｗ$_1$，…，ｂｗ$_N$は、全単語w_1，…，w_Nに対する関連語群をベクトルによって表現した関連度ベクトルである。

【０３０５】

重み付き単語群、即ち単語と単語の参照確率との組を多次元ベクトルｖ（s_i）で表現し、関連語群を関連度ベクトルｂｗ$_1$，…，ｂｗ$_N$で表現した場合、式（７）のように各単語の参照確率を、連想を加味した重み値に算出し直す処理は、以下のように解釈するこ

とができる。

【０３０６】

ｓａｌｉｅｎｃｅ（w_k｜ｐｒｅ（s_i））をｋ番目の要素として有する、連想を加味した顕現性ベクトルＶ（s_i）は、関連度ベクトルbw_1，…，bw_Nを基底とする斜交座標系における顕現性ベクトルｖ（s_i）であると解釈することができる。言い換えると、連想を加味した顕現性ベクトルＶ（s_i）は、参照確率をそのまま要素とする顕現性ベクトルｖ（s_i）を関連語軸方向へ回転させたものであると解釈することができる。

【０３０７】

関連度ベクトルbw_1，…，bw_Nを基底とする斜交座標系とは、連想を加味した各単語を１次元とした場合に、各基底ベクトル（各単語の次元方向に大きさ１のベクトル）は、夫々直行せず関連度が高い単語同士の基底ベクトル間の角度が小さくなるような座標系である。

【０３０８】

$b_{j,k}$を各要素とする変換行列を参照確率を要素とする顕現性ベクトルに乗算すると、関連する単語の次元方向に回転した顕現性ベクトルＶ（s_i）が得られると解釈することができる。

【０３０９】

したがって、文毎の意味のまとまりを表わす重み付き単語群を顕現性ベクトルで表現して記憶している場合、文単位検索装置１のＣＰＵ１１がその顕現性ベクトルを関連度ベクトルによって回転（変換）する処理を行なうことによって、文毎の意味のまとまりを連想が加味された顕現性ベクトルで表わして記憶しておくことができる。

【０３１０】

次に、上述のように定量的に関連度を表わした関連度群を使用して、各文単位の意味のまとまりを表わす各単語の重み値を連想を加味して算出し直す処理を実行した結果の具体例を以下に示す。図２６は、実施の形態３における文単位検索装置１のＣＰＵ１１によって算出された各単語の顕現性を表わす重み値の内容例を示す説明図である。図２６（ａ）に示した各文s_1，s_2に対する各単語の重み値は夫々、関連語群を使用して連想が加味される前の参照確率の値である。一方、図２６（ｂ）に示した各文s_1，s_2に対する各単語の重み値は、関連語群を使用して連想が加味された後の重み値である。

【０３１１】

なお、図２６に示す具体例は、日本語話し言葉コーパス（http://www.kokken.go.jp/katsudo/kenkyu_jyo/corpus/、ＣＳＪ／ｖｏｌ１７／Ｄ０３Ｆ００４０）より抽出した文単位の例である。

【０３１２】

図２６の内容例に示すように、図２６（ｂ）の文s_1における「大阪」の重み値は、図２６（ａ）の文s_1における「大阪」の参照確率の値０．３３３８と比較して、０．６２２９と高くなっている。また、図２６（ｂ）の文s_2における「大阪」の重み値は、図２６（ａ）の文s_2における参照確率の値０．３２０８と比較して、０．６６７５とさらに高くなっている。

【０３１３】

さらに、図２６（ａ）の参照確率の例では、文s_2における「大阪」の重み値は、文s_2に「アメリカ村」が出現しているにも拘わらず、その「大阪」の重み値への影響（励起）が考慮されていないために重み値が低下している。これに対し、図２６（ｂ）の連想を加味した後の重み値の例では、文s_2における「大阪」の重み値は、文s_2に「アメリカ村」が出現していることによって、出現していない「大阪」の顕現性を表わす重み値が高くなっている。「アメリカ村」と「大阪」との関連度の影響が反映されているからである。

【０３１４】

このように、文単位検索装置１が文単位毎に記憶している重み付き単語群に対し、参照確率という定量的な値を用いて関連度を表わした関連語群を用いて連想を加味することに

より、文単位で「アメリカ村」が注目されている場合の「大阪」の顕現性を、文単位又は言葉の書き手又は話し手の背景文脈により近づかせることができる。これにより、「大阪」の単語の顕現性を表わす重み値が低く算出されて、文単位の意味のまとまりが書き手又は話し手の実際の文脈と離れたように定量的に評価されてしまうことを回避することができる。

【０３１５】

４．検索処理

次に、実施の形態３における検索処理について説明する。「４－１．ユーザから入力された言葉の受け付け」については、受付装置４のＣＰＵ４１が行う処理については実施の形態１及び２と同様であるので、詳細な説明を省略する。

【０３１６】

４－２’．受け付けた言葉に対する連想を加味した意味のまとまりの定量化

次に、文単位検索装置１のＣＰＵ１１が、受付装置４，４，…で受け付けた言葉のデータを受信した場合に、文書記憶手段２で記憶している文書中の文を検索する処理について説明する。受け付けた言葉に対しても、意味のまとまりの定量化、即ち当該テキストデータの単語抽出及び単語の参照確率を算出し、さらに関連度を使用して重み値を算出し直す。

【０３１７】

実施の形態３では、文単位検索装置１のＣＰＵ１１は、受け付けた言葉の意味のまとまりを定量的に表わす単語と単語の参照確率との組又は顕現性ベクトル、即ち重み付き単語群に、関連語による連想を加味する。以下に、文単位検索装置１のＣＰＵ１１が受け付けた言葉に対応付けた重み付き単語群の各単語の重み値を連想を加味して算出し直し、算出し直した重み値に基づいて検索を実行する処理について説明する。

【０３１８】

図２７は、実施の形態３における文単位検索装置１及び受付装置４の検索処理の処理手順を示すフローチャートである。なお、図２７のフローチャートに示す処理手順では、実施の形態１における図１５、図１６及び図１７のフローチャートに示した検索処理の処理手順と同一の処理については各ステップに同一の符号を用いて詳細な説明を省略する。

【０３１９】

図２７のフローチャートに示す処理手順の内、二点鎖線で囲まれたステップＳ４００１の処理が、実施の形態１における図１５、図１６及び図１７のフローチャートに示した処理手順と異なる。即ち、ステップＳ４１１と、ステップＳ４１２との間に以下に説明するステップＳ４００１が追加されていることが異なる。

【０３２０】

以下に、実施の形態３において受け付けた言葉の意味のまとまりを表わす重み付き単語群を対応付け、予め記憶してある意味のまとまりが類似する文単位を抽出する検索処理について以下に説明する。

【０３２１】

ＣＰＵ１１は、一時記憶領域１４に夫々参照確率を算出して記憶している全単語に対し、所定値以上の参照確率が算出された単語に絞り込み（ステップＳ４１１）、ステップＳ４０８において算出した参照確率を、連想を加味した重み値に算出し直す（ステップＳ４００１）。ステップＳ４００１における、ＣＰＵ１１による連想を加味した重み値の算出し直しの処理は、図２５のフローチャートに示した処理と同様、単語を１つずつ選択し、選択した一の単語への各単語の関連度と各単語の参照確率とを乗算して算出する。

【０３２２】

それまでの処理により、受け付けた言葉に対し、以前に受け付けた言葉から続く流れ上の意味のまとまりを、連想を加味した上で定量的に表わす単語と単語の参照確率の組（重み付き単語群）を検索要求として生成することができた。

【０３２３】

ＣＰＵ１１はこの後、ステップＳ４００１で得られた連想が加味された重み付き単語群

に対し、各文毎に対応付けて記憶してある、連想が加味された重み付き単語群を読み出して、類似する文を抽出する処理を実行する。連想が加味された重み付き単語群についての以降の処理は実施の形態１と同様であるので詳細な説明を省略する。

【０３２４】

これにより、文単位検索装置１は、文書記憶手段２に記憶してある文書データから分別される文と受け付けた言葉とで、関連語を利用して連想を加味した意味のまとまりが類似しているか否かを判断し、類似すると判断された文を直接的に出力することができる。したがって、本発明の文単位検索方法を実施することにより、文脈上の意味のまとまりが類似する文単位を連想を加味して効果的に抽出し、直接的に出力することができる。

【０３２５】

なお、文単位検索装置１のＣＰＵ１１は、受け付けた言葉に対して重み付き単語群を対応付け、文毎に予め記憶してある重み付き単語群と類似しているか否かを判断する場合、図２７のフローチャートに示した処理手順のように、重み付き単語群が同一の単語を含んでいるか否かによって判断するとは限らない。さらに同一の単語に付与されている重み値の差分を算出し、算出した差分が小さい程類似すると判断するとは限らない。

【０３２６】

次に、文単位検索装置１のＣＰＵ１１が、受け付けた言葉と意味のまとまりが類似する文単位を抽出する処理を、意味のまとまりを顕現性ベクトル及び関連度ベクトルで表現し、ベクトル間の距離を算出することによって実現する場合について以下に説明する。

【０３２７】

図２８は、実施の形態３におけるベクトル表現を用いた場合の文単位検索装置１及び受付装置４の検索処理の処理手順を示すフローチャートである。なお、図２８のフローチャートに示す処理手順では、実施の形態１における図１５、図１６及び図１７のフローチャート、及び実施の形態２における図１９のフローチャートに示した検索処理の処理手順と同一の処理については各ステップに同一の符号を用いて詳細な説明を省略する。

【０３２８】

図２８のフローチャートに示す処理手順の内、一点鎖線で囲まれた各ステップＳ５０１からステップＳ５０６までの処理が、実施の形態１における図１５、図１６及び図１７のフローチャートに示した処理手順と異なる。実施の形態１におけるステップＳ４１２からステップＳ４１６までの処理の代わりに、実施の形態２における文単位検索装置１のＣＰＵ１１により実行されるステップＳ５０１からステップＳ５０６までの処理と同様の処理を行なう。図２８のフローチャートに示す処理手順の内、二点鎖線で囲まれたステップＳ５００１の処理が、実施の形態２における図１９のフローチャートに示した処理手順と異なる。即ち、ステップＳ５０１と、ステップＳ５０２との間に以下に説明するステップＳ５００１が追加されていることが異なる。

【０３２９】

文単位検索装置１のＣＰＵ１１は、ステップＳ５０１で算出した顕現性ベクトルを、関連語による連想を加味した顕現性ベクトルに算出し直す（ステップＳ５００１）。

【０３３０】

ＣＰＵ１１はこの後、ステップＳ５００１で得られた連想が加味された重み付き単語群に対し、各文毎に対応付けて記憶してある、連想が加味された顕現性ベクトルを読み出して、類似する文を抽出する処理を実行する。連想が加味された顕現性ベクトルを読み出して類似する文を抽出する処理は実施の形態２と同様であるので詳細な説明を省略する。

【０３３１】

なお、ＣＰＵ１１によるステップＳ５００１において、顕現性ベクトルを関連語による連想を加味した顕現性ベクトルに算出し直す処理は、ステップＳ５０１で算出した顕現性ベクトルを関連度ベクトル群（行列）で式（７）で示したように変換して（回転させて）算出する。具体的には、参照確率のみを要素とする多次元ベクトルｖ（s_i）に対して上述の連想を加味した顕現性ベクトルＶ（s_i）を算出する。

【０３３２】

なお、上述の図２８のフローチャートに示した処理手順の内の、ＣＰＵ１１が受け付けた言葉に対応付けた顕現性ベクトルと、読み出した顕現性ベクトルとの距離を算出するステップＳ５０３の処理は、実施の形態３では、具体的には以下のように算出する。受け付けた言葉U_iに対し連想が加味されて算出し直された顕現性ベクトルがＶ（u_i）と表わされ、読み出された、予め連想が加味されてある顕現性ベクトルがＶ（s_i）と表わされる場合、ＣＰＵ１１は以下に示す式（８）のように、コサイン距離を算出する。

【０３３３】

【数８】

$$\frac{\vec{V}(s_i)\cdot\vec{V}(u_i)}{|\vec{V}(s_i)||\vec{V}(u_i)|} = \frac{\sum_{k=1}^{N} salience(w_k|s_i)\,salience(w_k|u_i)}{\sqrt{\sum_{k=1}^{N} salience(w_k|s_i)^2}\sqrt{\sum_{k=1}^{N} salience(w_k|u_i)^2}} \cdots (8)$$

【０３３４】

ただし、式（８）に示したように距離を算出した場合、言葉の顕現性ベクトルＶ（u_i）と、読み出した顕現性ベクトルＶ（s_i）とが近いほど、算出したコサイン距離の値は大きくなる。したがって、ＣＰＵ１１はステップＳ５０６において、算出したコサイン距離が大きい順に類似度を付与する。

【０３３５】

文単位検索装置１のＣＰＵ１１による上述のような処理により、連想が加味された意味のまとまりを表わす顕現性ベクトル間の距離によって、意味のまとまりが類似する文単位を直接的に検索することができる。ベクトル表現を用いることにより、ＣＰＵ１１は、受け付けた言葉に対応付けられる連想が加味された重み付き単語群と、予め文に対応付けて記憶されている連想が加味された重み付き単語群とを一単語ずつ重み値を比較している処理を行なうことなしに、連想を加味した上で直接的に類似しているか否かを判断を行うことができる。

【０３３６】

また、実施の形態３における文単位検索装置１による場合、各文単位及び単語に対応付けられる顕現性ベクトルは、各単語に相当する次元間が直交しない関連度が高い単語の次元方向間の角度が小さくなるような斜交座標系で扱われる。このため、類似するか否かを判断する際にベクトル間の距離を比較した場合に、関連度が高い単語の次元方向に要素を有している場合は類似していると判断されるようになる。

【０３３７】

したがって、「大阪」の顕現性が高い文単位ｓが記憶されている場合、受け付けた言葉において例えば「オランダ村」の顕現性が高いときは、文単位ｓは受け付けた言葉に類似すると判断されない。しかし、受け付けた言葉において「アメリカ村」の顕現性が高いときは、受け付けた言葉において「大阪」の顕現性が励起されて高くなるので、文単位ｓはこの受け付けた言葉に類似すると判断される可能性が高くなる。

【０３３８】

これにより、受け付けた言葉に対し、連想を加味してより効果的に意味のまとまりが類似する文単位を検索して直接的に出力することができる。

【０３３９】

なお、実施の形態１乃至３では、検索結果として受信したテキストデータは、受付装置４が備える表示手段４６のモニタ等で表示する構成としたが、受信したテキストデータから音声に変換して、音声入出力手段４７のスピーカ等を介して出力する構成でもよい。これにより、ユーザは自分が音声入力した複数の言葉によって、又は他のユーザとの会話を音声入力することで、その会話の文脈と意味のまとまりが類似する文を検索結果として得ることができる。受け付けた言葉が話し言葉からなる場合に、発話では省略されている、ゼロ代名詞で表される単語をも含めた単語の顕現性が類似する文を直接的に検索結果として得ることができる。

【０３４０】

また、文単位検索装置１のＣＰＵ１１は、言葉のテキストデータを受信する都度、当該テキストデータに対して検索された文のうち、一番優先順位の高い文を表すテキストデータのみを受付装置４，４，…に送信する構成としてもよい。これにより、入力される言葉に対する検索結果を会話の第三者の発話として提示し、鼎談を実現することも可能である。

【０３４１】

なお、実施の形態１乃至実施の形態３では、文単位検索装置１は文毎に顕現性を示す情報を特定して記憶したが複数の文からなる段落（ｐａｒａｇｒａｐｈ）毎にタグ＜ｐ＞＜／ｐ＞で挟み、当該段落に対して特徴パターンを特定して顕現性を示す情報をｓａｌｉｅｎｃｅ属性によって記憶させ、段落を検索結果として出力する構成としてもよい。文又は段落に限らず、一定の意味のまとまりを表す単位であれば文節であっても構わない。話し言葉の場合は一文と識別できる文字列が非常に長くなることが考えられる。多数の文節から構成され、文節と文節は「～も」「～ので」等の接続助詞で続いているにも拘わらず、文脈が動的に変化していく場合は一文では意味がまとまっていないときがある。したがって、所定の文節の数を超えて構成される文の場合は、文節毎に一文であるとみなして処理を行う構成としてもよい。

【０３４２】

また、実施の形態１乃至実施の形態３では、話し言葉からなる文書データを書き言葉からなる文書データと区別して予め記憶しておく構成としたが、受信した言葉に対して各単語の特徴パターンを特定して参照確率を算出する都度、文書記憶手段２で記憶する構成としてもよい。この際、文単位検索装置１のＣＰＵ１１は、連続して受信した言葉が一連のものであるか否かの判断を当該言葉の送信元である受付装置４を識別する情報と、受付装置４がユーザの検索開始・完了操作を検知したことを示す情報とによってすることもできる。これにより、予め文書記憶手段２で記憶してある文書データのページに該当する単位で言葉を文書記憶手段２に記憶させることができる。

【０３４３】

なお、実施の形態１乃至実施の形態３では、文書データの取得とタグ付け、参照確率を求めるための回帰分析、更に言葉を受け付けた際の処理を文単位検索装置１が全て行う構成としたが、文単位検索装置と文書記憶装置とに分ける構成としてもよい。この場合は、文書記憶装置でＷｅｂクローリングを行って文書データを取得し、さらに形態素解析及び統語解析によってテキストデータにタグを付加して記憶しておく。また、文書記憶装置で記憶した文書データをもとに参照確率を算出するための式を回帰分析によって求め、求めた式を使用して、記憶した文書データに対して文毎の単語及び単語の参照確率を記憶する処理を予めしておく。文単位検索装置は、言葉を変換したテキストデータを受信した際に特徴パターンを特定し、文書記憶装置から参照確率を算出するための回帰式を取得して参照確率を算出して検索を行う。

【０３４４】

また、実施の形態１乃至実施の形態３では、ユーザからの文字列入力又は音声入力等の

言葉の入力は、受付装置４によってテキストデータに変換され、文単位検索装置１へ送信される構成とした。これに限らず、文単位検索装置１が、ユーザの文字列入力操作を受け付ける入出力手段、及びユーザの音声入力を受け付ける音声入力手段を備える構成でもよい。図２９は、本発明の文単位検索方法を文単位検索装置１で実施する場合の構成を示すブロック図である。この場合、文単位検索装置１は、ＣＰＵ１１、内部バス１２、記憶手段１３、一時記憶領域１４、文書集合接続手段１６及び補助記憶手段１７の他に、ユーザの操作を受けつけるマウス又はキーボード等の操作手段１４５、モニタ等の表示手段１４６及びマイク及びスピーカ等の音声入出力手段１４７を更に備える。

【０３４５】

図２９の構成図に示した構成の場合、文単位検索装置１のＣＰＵ１１は、音声入力手段から入力された音声の特徴を表わす、周波数又は会話速度等を検知し、発話における各単語の特徴パターンを特定することができる。各単語の文法的な特徴パターンは、入力された音声を音声認識によりテキストデータに変換して当該テキストデータに基づいて検索する構成としてもよい。

【０３４６】

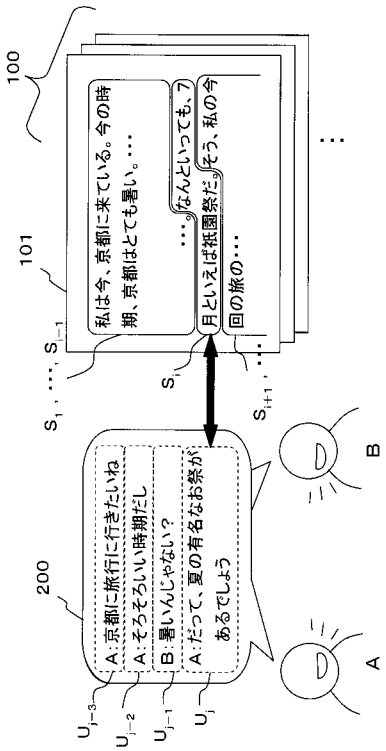
実施の形態１乃至実施の形態３では、受付装置４，４，…は、受け付けた文字列又は音声の言葉を一定の長さに区切ってデジタルデータに変換して送信するのみの装置として構成した。しかしながら、本発明の文単位検索方法を実施するためには、受付装置４，４，…が記憶手段４３に記憶しているプログラムを、受付装置４，４，…が受け付けた言葉を形態素解析及び統語解析、又は音素解析等の自然言語解析を実行することができるように構成してもよい。この場合、受付装置４，４，…のＣＰＵ４１は、受け付けた言葉における各単語の顕現性を表わす重み値を算出し、算出した重み付き単語群を検索要求として文単位検索装置１へ送信する構成でもよい。

【産業上の利用可能性】

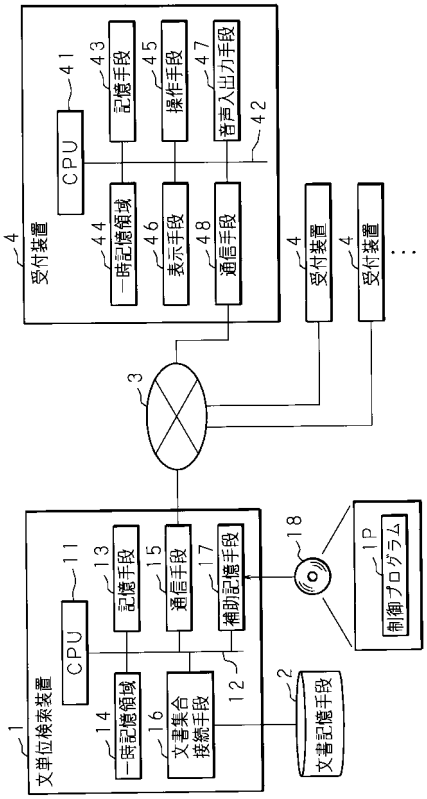
【０３４７】

本発明に係る文単位検索方法を、ユーザ間の会話を音声認識が可能なコンピュータ装置に実施させることにより、コンピュータ装置にユーザ間の会話に参加させて鼎談を実現する用途にも適用することが可能である。また、ユーザ間の会話又はチャットの文脈の流れに応じて切り替わる会話連動型広告の提示サービスを実現する用途にも適用可能である。会議中の文脈の流れに応じて、過去の議事録から類似関連する議事録を提示する会議支援サービスへの適用も可能である。さらに、執筆中の文章を言葉として受け付け、文脈の流れに応じて、関連する情報を提供する文章執筆支援サービスへの適用も可能である。

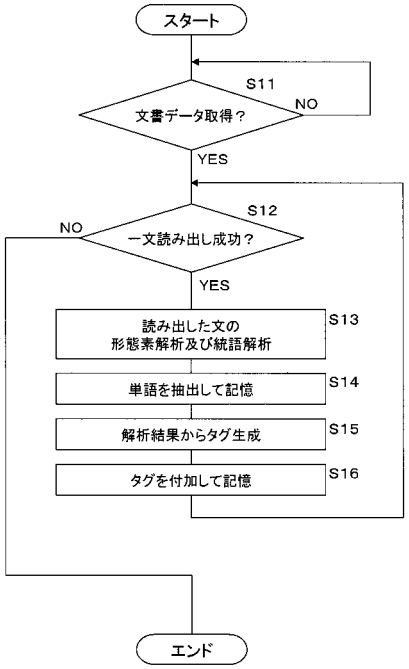
【図１】



【図２】



【図３】

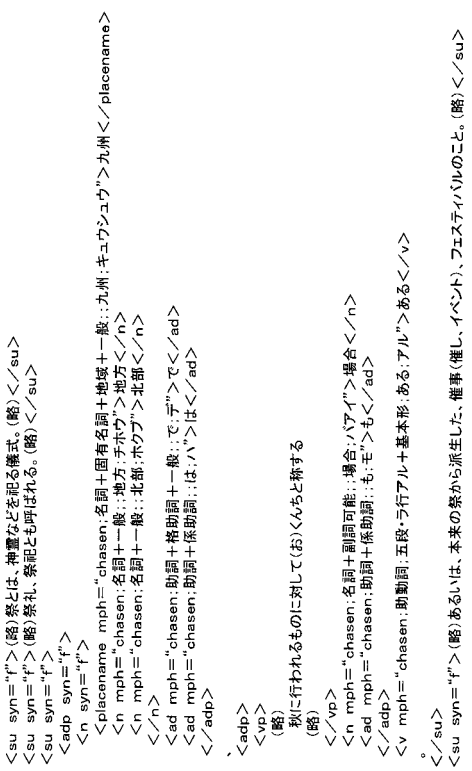


【図４】

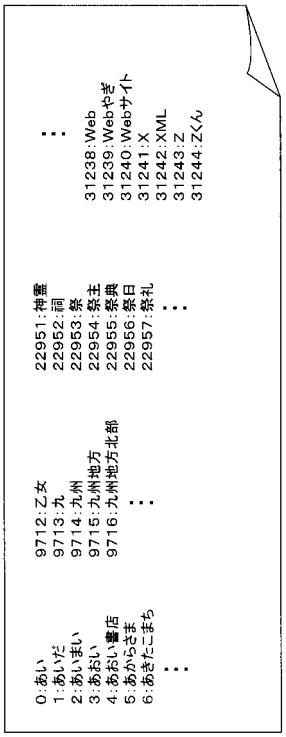
祭とは、神霊などを祀る儀式。祭礼、祭祀とも呼ばれる。
九州地方北部では、秋に行われるものに対して(お)くんちと称する場合もある。
あるいは、本来の祭から派生した、催事(催し、イベント)、フェスティバルのこと。

http://ja.wikipedia.org/wiki/祭より抜粋
(http://ja.wikipedia.org/wiki/%E7%A5%AD)

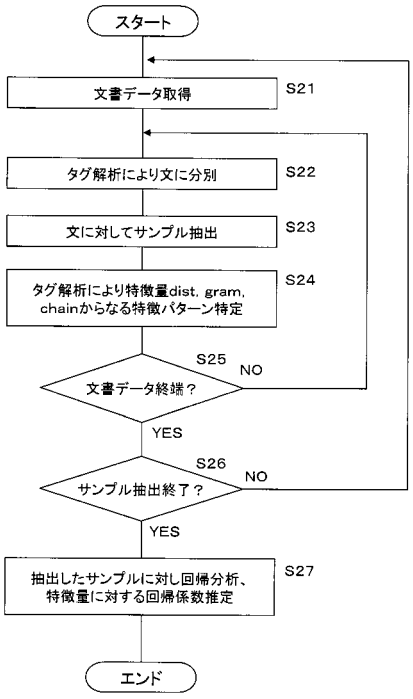
【図５】


<su syn="f">(略)祭とは、神霊などを祀る儀式。(略)</su>
<su syn="f">(略)祭礼、祭祀とも呼ばれる。(略)</su>
<su syn="f">
<adp syn="f">
<n syn="f">
<placename mph="chasen:名詞+固有名詞+地域+一般::九州:キュウシュウ">九州</placename>
<n mph="chasen:名詞+一般::地方:チホウ">地方</n>
<n mph="chasen:名詞+一般::北部:ホクブ">北部</n>
</n>
<ad mph="chasen:助詞+格助詞+一般::で:デ">で</ad>
<ad mph="chasen:助詞+係助詞::は:ハ">は</ad>
</adp>
、
<adp>
<vp>
(略)
秋に行われるものに対して(お)くんちと称する
(略)
</vp>
<n mph="chasen:名詞+副詞可能::場合:バアイ">場合</n>
<ad mph="chasen:助詞+係助詞::も:モ">も</ad>
</adp>
<v mph="chasen:助動詞:五段・ラ行アル+基本形:ある:アル">ある</v>
。
</su>
<su syn="f">(略)あるいは、本来の祭から派生した、催事(催し、イベント)、フェスティバルのこと。(略)</su>

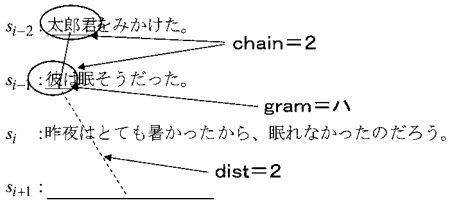
【図６】


0:あい
1:あいだ
2:あいまい
3:あおい
4:あおい書店
5:あからさま
6:あきたこまち
...
9712:乙女
9713:九
9714:九州
9715:九州地方
9716:九州地方北部
...
22951:神霊
22952:祠
22953:祭
22954:祭主
22955:祭典
22956:祭日
22957:祭礼
...
31238:Web
31239:Webやぎ
31240:Webサイト
31241:X
31242:XML
31243:Z
31244:Zくん

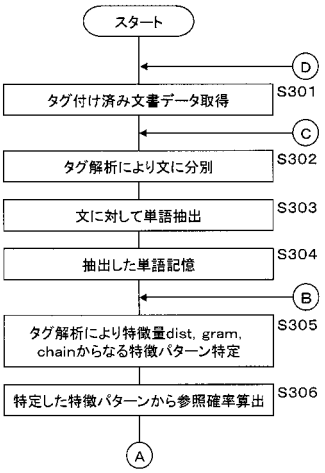
【図７】


スタート
文書データ取得
S21
タグ解析により文に分別
S22
文に対してサンプル抽出
S23
タグ解析により特徴量dist, gram, chainからなる特徴パターン特定
S24
S25
NO
文書データ終端？
YES
S26
NO
サンプル抽出終了？
YES
抽出したサンプルに対し回帰分析、特徴量に対する回帰係数推定
S27
エンド

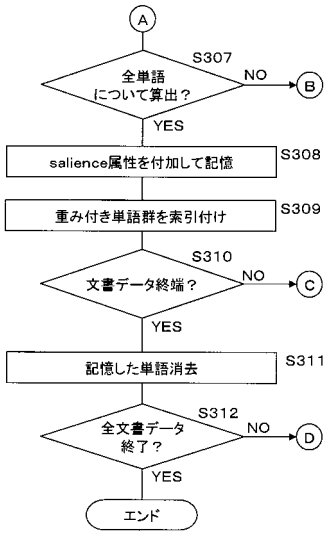
【図８】


s_{i-2} 太郎君をみかけた。
chain＝2
s_{i-1} :彼は眠そうだった。
gram＝ハ
s_i :昨夜はとても暑かったから、眠れなかったのだろう。
dist＝2
s_{i+1} :

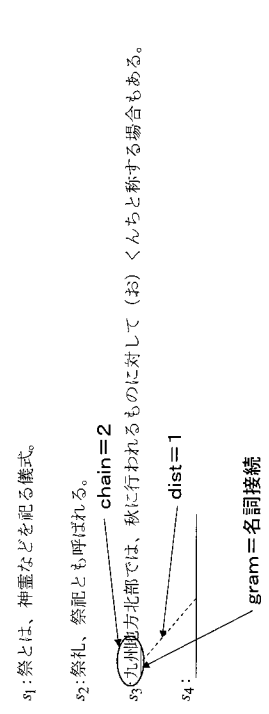
【図９】



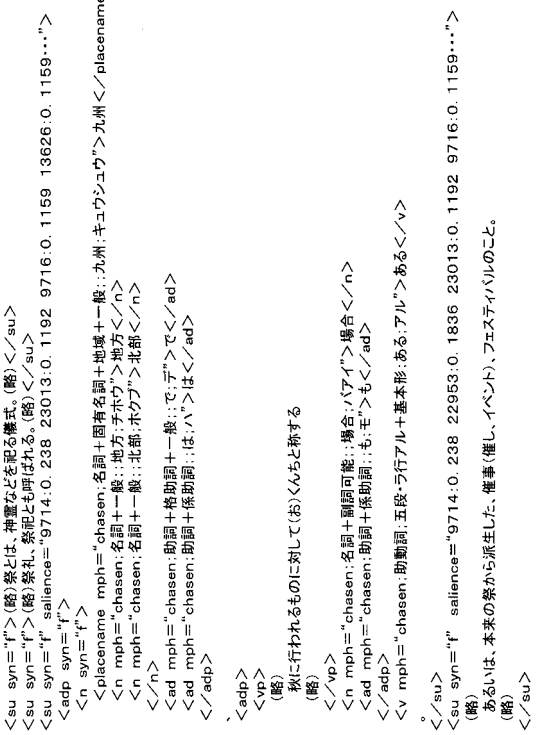
【図１０】



【図１１】



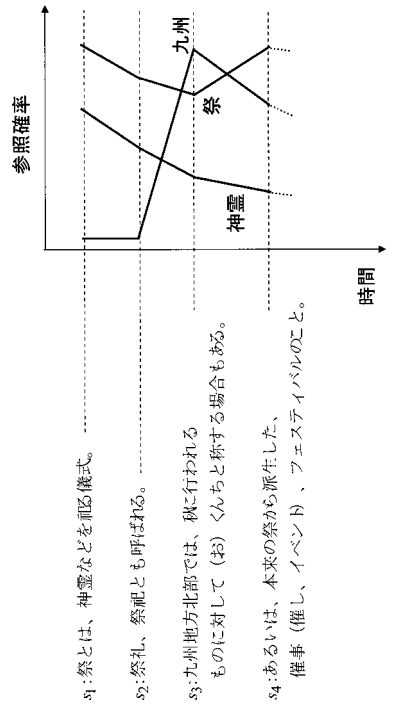
【図１２】



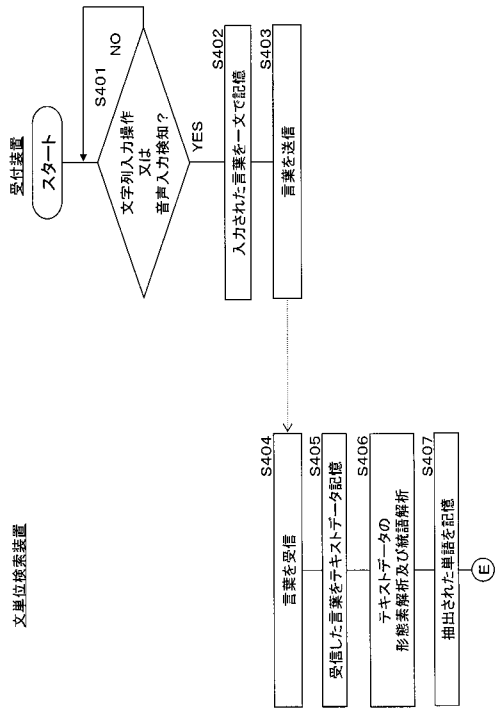
【図１３】

属するグループ (k－d treeノードID)	重み付き単語群	文書データ	文の位置
121	(9714:0.238 22953:0.1836 23013:0.1192 9716:0.1159 …)	"ja.wikipedia.org/wiki/%E7%A5%AD.gda"	"/html/body[1]/p[2]/su[8]"
121	…	…	…
…	…	…	…
122	…	…	…
…	…	…	…

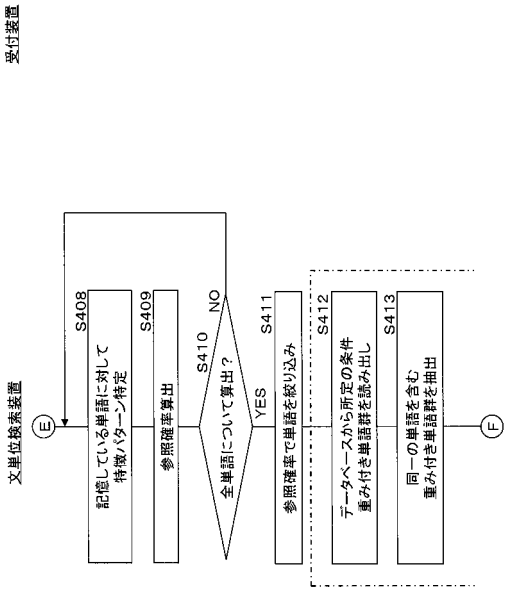
【図１４】



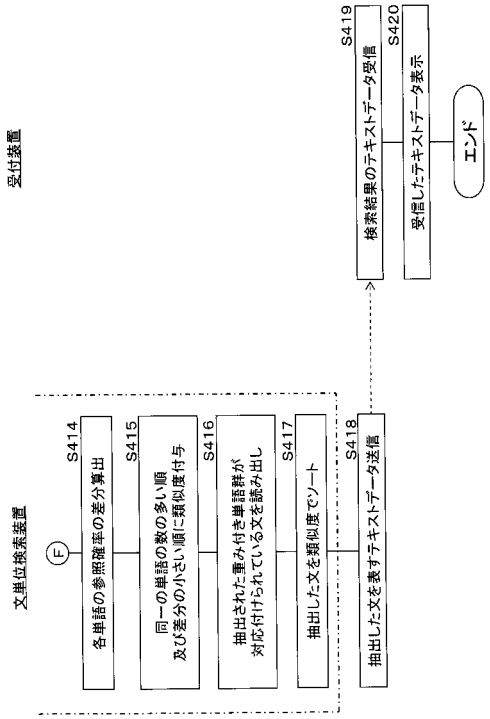
【図１５】



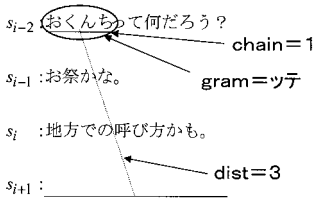
【図１６】



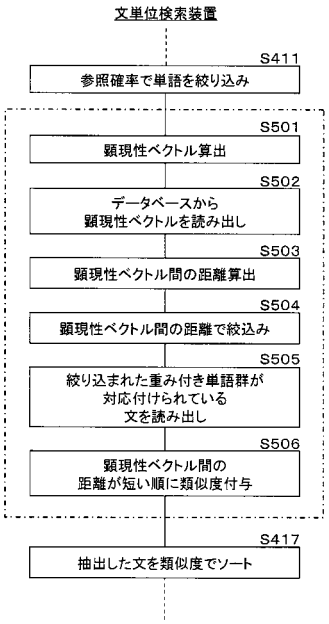
【図１７】



【図１８】



【図１９】



【図２０】

U_1:アメリカ村も面白いですよ

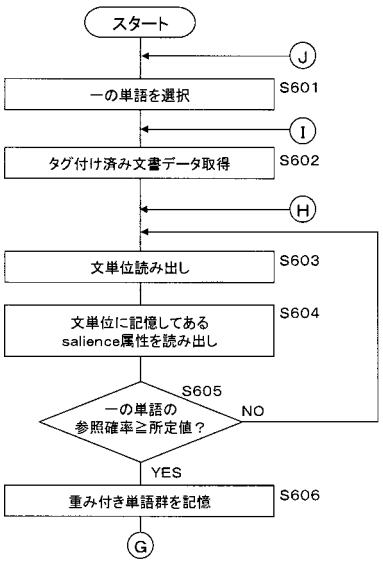
U_2:それは一体どこにあるんですか

U_3:アメリカ村もミナミのちょっと横なんですけど

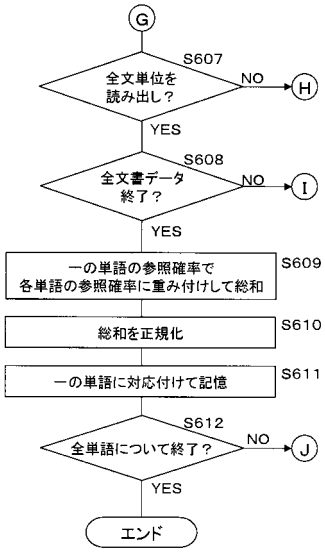
U_4:それは昔からある若者の町で

U_5:アメリカ村に有名な三角公園ていう公園があるんですね

【図２１】



【図２２】



【図２３】

(a)

GW_1：(秋:0, アメリカ村:0. 6, ・・・, 大熊座:0, 大阪:0. 2, 大鹿:0, ・・・)

GW_2：(秋:0. 1, アメリカ村:0. 3, ・・・, 大熊座:0, 大阪:0. 4, 大鹿:0, ・・・)

GW_3：(秋:0, アメリカ村:0. 2, ・・・, 大熊座:0, 大阪:0. 2, 大鹿:0, ・・・)

(b)

GW_1'：(秋:0, アメリカ村:0. 36, ・・・, 大熊座:0, 大阪:0. 12, 大鹿:0, ・・・)

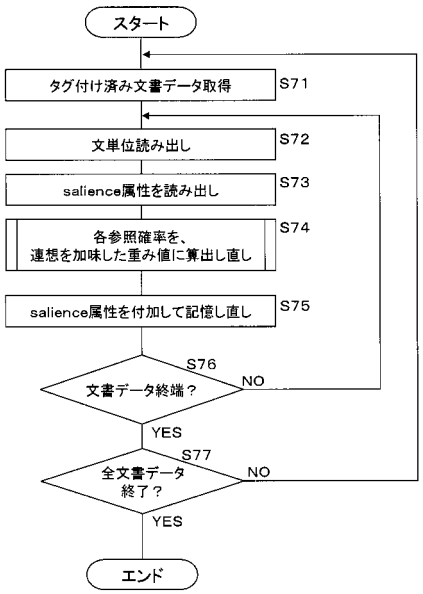
GW_2'：(秋:0. 03, アメリカ村:0. 09, ・・・, 大熊座:0, 大阪:0. 12, 大鹿:0, ・・・)

GW_3'：(秋:0, アメリカ村:0. 04, ・・・, 大熊座:0, 大阪:0. 04, 大鹿:0, ・・・)

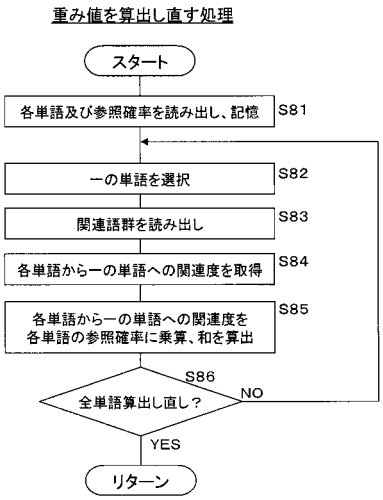
(c)

GW''：(秋:0. 03, アメリカ村:0. 49, ・・・, 大熊座:0, 大阪:0. 28, 大鹿:0, ・・・)

【図２４】



【図２５】



【図２６】

(a)

s_1: 若者の町で。

'大阪':0.3338 '東京':0.3338 '関西人':0.2978 'ミナミ':0.201 'キタ':0.1837 '火':0.1637 '関西':0.1291 'アメリカ':0.1265 '心斎橋':0.1265 '町':0.121 ...

s_2: アメリカ村に有名な三角公園ていう公園があるんですね。

'大阪':0.3208 '東京':0.3208 '関西人':0.2862 'アメリカ村':0.2021 'ミナミ':0.1932 '公園':0.1772 'キタ':0.1765 '火':0.1573 'アメリカ':0.1308 ...

(b)

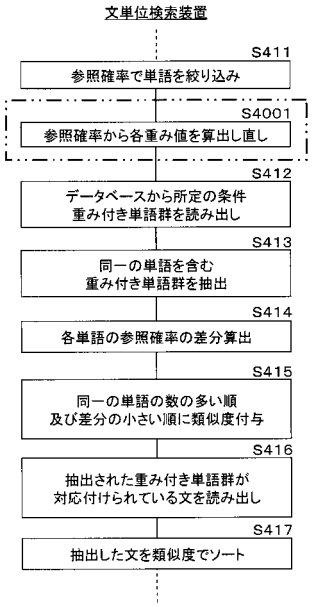
s_1: 若者の町で。

'大阪':0.6229 '人':0.5345 '東京':0.3287 '自分':0.2545 '関西':0.1914 'アメリカ':0.1858 '昔':0.1382 '今':0.1314 '関西人':0.1261 'キタ':0.1166 ...

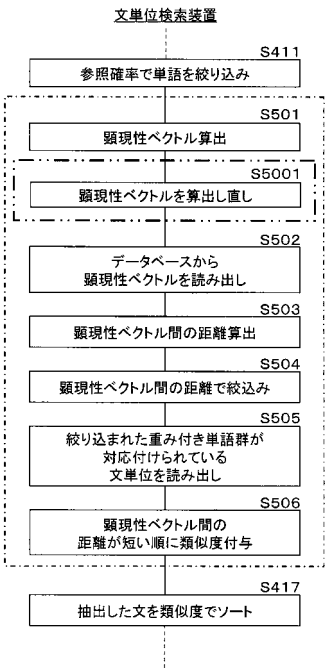
s_2: アメリカ村に有名な三角公園ていう公園があるんですね。

'大阪':0.6675 '人':0.5625 '東京':0.3182 'アメリカ':0.2604 '自分':0.2472 '関西':0.1889 '公園':0.1571 '今':0.1413 '昔':0.1374 'アメリカ村':0.1339

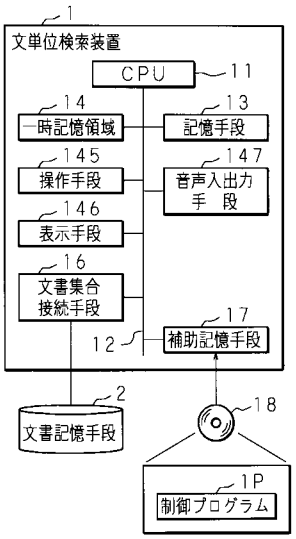
【図２７】



【図２８】



【図２９】


1
文単位検索装置
CPU
11
14
一時記憶領域
13
記憶手段
145
操作手段
147
音声入出力手段
146
表示手段
16
文書集合接続手段
17
12
補助記憶手段
2
文書記憶手段
18
1P
制御プログラム

フロントページの続き

(72)発明者　奥乃　博
京都府京都市左京区吉田本町　京都大学大学院情報学研究科内

審査官　吉田　誠

(56)参考文献　特開平０６－１６２０９２（ＪＰ，Ａ）
特開２００４－２３４１７５（ＪＰ，Ａ）
特開２００５－２５０７６２（ＪＰ，Ａ）
西脇　正通，関連記事を利用したテキストセグメンテーション，情報処理学会研究報告，日本，社団法人情報処理学会，２００２年１１月１３日，Ｖｏｌ．２００２　Ｎｏ．１０４，７９－８４ページ
徳永　健伸，言語と計算５　情報検索と言語処理，財団法人東京大学出版会，１９９９年１１月２５日，第1版，２６－５０ページ

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F　17/30