(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4208402号
(P4208402)

(45)発行日　平成21年1月14日(2009.1.14)　(24)登録日　平成20年10月31日(2008.10.31)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 320D
G06F 17/30 170A
G06F 17/30 210A

請求項の数 7　(全 14 頁)

(21)出願番号　特願2000-333510(P2000-333510)
(22)出願日　平成12年10月31日(2000.10.31)
(65)公開番号　特開2002-140355(P2002-140355A)
(43)公開日　平成14年5月17日(2002.5.17)
審査請求日　平成17年1月27日(2005.1.27)

(73)特許権者　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(74)代理人　100085660
弁理士　鈴木　均
(72)発明者　真野　博子
東京都大田区中馬込１丁目３番６号
株式会社　リコー内
(72)発明者　小川　泰嗣
東京都大田区中馬込１丁目３番６号
株式会社　リコー内

審査官　岩間　直純

最終頁に続く

(54)【発明の名称】文書検索装置、文書検索方法および記録媒体

(57)【特許請求の範囲】
【請求項１】
複数の文書の文書情報と、前記文書中に含まれる各単語の単語統計情報とを保持して構成される文書データベースと、
前記文書データベースからキーワードに適合する適合文書および適合しない非適合文書を選出する文書ランキング部と、
前記キーワードの関連語を選出する単語ランキング部と、
新しいキーワードを生成するキーワード生成部と、を備え、
前記文書ランキング部は、前記文書データベースから、装置に入力されたキーワードについて適合文書及び非適合文書を選出し、
前記単語ランキング部は、前記適合文書中の単単語について、前記文書ランキング部で選出した適合文書および非適合文書中の出現頻度と、前記文書データベースの検索対象文書中の出現頻度と、をもとに前記キーワードとの関連度を計算し、前記関連度の高い単単語を前記キーワードの関連語として選出し、さらに、前記文書ランキング部で選出した適合文書から連続した２つ以上の単語から構成される単語列を抽出し、前記単単語の関連語の関連度の中で最小のものに基づき前記単語列の関連度の下限値を計算するとともに、前記適合文書と非適合文書中の前記単語列の出現頻度をそれぞれ計算し、該出現頻度から暫定的な関連度を計算し、該暫定的な関連度と前記関連度の下限値とを比較し、前記暫定的な関連度のほうが前記関連度の下限値より小さいときは、前記単語列を関連語候補から外したうえで、単語列を選出して、前記選出された単単語及び前記選出された単語列を前記

キーワードの関連語とし、
前記キーワード生成部は、前記キーワードの関連語をもとの前記キーワードに追加して新しいキーワードとし、
前記文書ランキング部は、前記キーワード生成部で生成された新しいキーワードに適合する文書を検索するようにしたことを特徴とする文書検索装置。
【請求項２】
請求項１に記載の文書検索装置において、
前記単語ランキング部は、予め指定した不要語および記号からのみなる語を含む単語列をキーワードの関連語候補としないようにしたことを特徴とする文書検索装置。
【請求項３】
請求項１又は２に記載の文書検索装置において、
前記単語ランキング部は、前記関連語候補に残った単語列ついて、前記文書データベース中の文書に出現する単語列の実際の出現頻度を求め、該実際の出現頻度から前記単語列の関連度を計算する際に、単単語にくらべて前記文書ランキング部で選出された適合文書および非適合文書中の出現状況の影響する度合が高くなるように出現状況の比率を設定したことを特徴とする文書検索装置。
【請求項４】
請求項１乃至３のいずれか一項に記載の文書検索装置において、
前記単語ランキング部は、前記文書ランキング部で選出された適合文書から抽出した単語列の関連語候補から前記キーワードの関連語として選出するための関連度の下限を単単語にくらべて高く設定し、単語列が関連語として選ばれる数を抑えるようにしたことを特徴とする文書検索装置。
【請求項５】
請求項１乃至４のいずれか一項に記載の文書検索装置において、
前記文書ランキング部は、前記キーワード生成部で生成された新しいキーワードによって前記文書データベースを検索する際に、このキーワードに含まれる単単語と重複する単語を含む単語列については、その重みを下げて適合度を算出するようにしたことを特徴とする文書検索装置。
【請求項６】
コンピュータによって実行される、前記コンピュータに入力されたキーワードに適合する文書を複数の文書を保持する文書データベースから検索する文書検索方法において、
前記文書データベースからキーワードに適合する適合文書および適合しない非適合文書を選出する文書ランキング工程と、
前記キーワードの関連語を選出する単語ランキング工程と、
新しいキーワードを生成するキーワード生成工程と、を備え、
前記文書ランキング工程は、前記文書データベースから、前記コンピュータに入力されたキーワードについて適合文書及び非適合文書を選出し、
前記単語ランキング工程は、前記適合文書中の単単語について、前記文書ランキング工程で選出した適合文書および非適合文書中の出現頻度と、前記文書データベースの検索対象文書中の出現頻度と、をもとに前記キーワードとの関連度を計算し、前記関連度の高い単単語を前記キーワードの関連語として選出し、さらに、前記文書ランキング工程で選出した適合文書から連続した２つ以上の単語から構成される単語列を抽出し、前記単単語の関連語の関連度の中で最小のものに基づき前記単語列の関連度の下限値を計算するとともに、前記適合文書と非適合文書中の前記単語列の出現頻度をそれぞれ計算し、該出現頻度から暫定的な関連度を計算し、該暫定的な関連度と前記関連度の下限値とを比較し、前記暫定的な関連度のほうが前記関連度の下限値より小さいときは、前記単語列を関連語候補から外したうえで、単語列を選出して、前記選出された単単語及び前記選出された単語列を前記キーワードの関連語とし、
前記キーワード生成工程は、前記キーワードの関連語をもとの前記キーワードに追加して新しいキーワードとし、

前記文書ランキング工程は、前記キーワード生成工程で生成された新しいキーワードに適合する文書を検索するようにしたことを特徴とする文書検索方法。

【請求項７】

コンピュータに請求項６に記載の文書検索方法を実現させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書検索装置、文書検索方法および記録媒体に関し、より詳細には、与えられたキーワードに対して適合する文書を選択し、この適合文書から抽出したキーワードの関連語を付加したキーワードによって適合する文書を検索しなおす文書検索装置、文書検索方法および記録媒体に関する。

【０００２】

【従来の技術】

文書を多数集積している文書データベースからユーザの必要とする文書を探しだすには、ユーザが入力したキーワードを用いて一旦検索した後、そのキーワードに適合した文書中に出現する単語の中から入力したキーワードに関連した単語を選出し、はじめに入力したキーワードに追加し、再度、検索することで、よりユーザの求めるものに近いものを得る方法が知られている。

たとえば、キーワードの関連語を選出する方法として、適合文書中の各単語について、適合文書の中での出現状況などの統計情報を利用して、キーワードとの関連度を算出し、その値の大きい上位何単語かを選出する方法が提案されている（文献１：Robertson, S.E. "On term selection for query expansion" Journal of Documentation 46, Dec 1990,p359-364）。

【０００３】

次に、この従来の関連語抽出方法を説明する。

ユーザから入力されたキーワード中の各単語に対して単語の重要度に応じた重みを付与する。この単語の重みの計算式には、たとえば、確率モデルにもとづく Robertson の計算式（式１）が知られている（文献２：Robertson, S.E. and Walker, S. "On relevance weights with little relevance information," SIGIR 97, ACM Press, pp.16-24）。この文献２の技術においては、キーワード中の各単語の重みは、検索対象文書全体の中での各単語の出現状況 Wp、Wq に応じて付与される。

W（重み） ＝ Wp　Wq　　……　（式１）

ここで　Wp ＝ k4 + log(N / (N - n))

Wq ＝ log(n / (N - n))

N: 検索対象総文書数

n: 単語の出現する文書数

k4: 調整パラメータ

次に、キーワード中の各単語の重みをもとに、各文書の文書適合度を計算する。この文書適合度の計算式は、たとえば、文献２の計算式（式２）で求まる。

F（適合度） ＝ Σ(W × tf /(k1 + tf))　……　（式２）

ここで

W ：（式１）で求めた単語の重み

tf: 文書あたりの単語の出現数

k1: 調整パラメータ

各文書の文書適合度を求め、適合度の高い順に各文書を順序づけ、上位何件かを適合文書とみなし、下位何件かを非適合文書とみなす。

適合文書の選出後、適合文書中の不要語（たとえば冠詞の a など）を除いたすべての単語について、適合文書および非適合文書での出現状況、すなわちフィードバック情報を反映させて、それぞれの単語の重みを再計算する。

適合文書選出後の重みは、たとえば、文献２の計算式（式３）を用いて、検索対象文書全体での出現状況 Wp、Wq （（式１）のコメント参照）と適合文書／非適合文書の中での出現状況 WrとWs を比率 CpとCq で足し合わせて付与される。

W'（重み）＝(Cp・Wp+(1-Cp)・Wr)-(Cq・Wq+(1-Cq)・Ws) ……（式３）

ここで Wr = log((r + 0.5) / (R - r + 0.5))

Ws = log((s + 0.5) / (S - s + 0.5))

Cp = k5 / (k5 + √R)

Cq = k6 / (k6 + √S)

R: 適合文書数

r: 適合文書集合の中で単語の出現する文書数

S: 非適合文書数

s: 非適合文書集合の中で単語の出現する文書数

k5, k6: 調整パラメータ

【０００４】

さらに、この重みとフィードバック情報から適合文書中の不要語を除いた各単語について、キーワードとの関連度を求める。

関連度の算出方法としては、たとえば、Boughanem の計算式（式４）がある(文献３：Walker, S. et al., "Okapi at TREC-6:Automated ad hoc, VLC, routing, filtering and QSDR," The Sixth Test REtrieval Conference (TREC-6), 1996, NIST)。

関連度 = (r / R - α・s / S) × W' …… （式４）

ここで α: 調整パラメータ

このようにして、適合文書中の各単語について、キーワードとの関連度を求めて、関連度の高いものから順にキーワード関連語として選出し、入力したキーワードに追加して新しいキーワードを作成する。

この新しいキーワードを用いて、再度、適合文書を選出する。このとき、文書適合度の算出には、上記（式３）で求めた重みが使われる。

【０００５】

上記の従来の方法では、キーワードに対する関連語として選出されるのは、個々の単語（単単語）である。

しかし、キーワードに関連した言葉として追加すべきものは、単単語とはかぎらない。たとえば、キーワード「Microsoft」に関連する言葉としては、「Windows」の他にも単語二つから成る「Bill Gates」なども有効と考えられる。この場合、「Bill」と「Gates」を単語単位でばらばらにキーワードに追加するより「Bill Gates」とまとめて追加した方が、より大きな効果を期待できる。単語単位であると、たとえば、「Bill」に対して「Bill Clinton」等も一致してしまうからである。

したがって、キーワードの関連語として単単語を選出するだけでは、キーワードを補うのに充分ではなく、より大きな単位で、関連する語句を選出できることが望ましい。

この点に関して、たとえば、特開平１１－２５１０８号公報記載の技術では、検索された文書から関連語を選出する際に、特定の品詞に属する単語の組合せである「単語の組」も候補に加えることを提案している。

【０００６】

【発明が解決しようとする課題】

しかしながら、上記の従来技術では、以下のような問題がある。

（１）特定の品詞に属する単語の組合せを文書から抽出するには、形態素解析という手間がかかり、かつ、結果の信頼性が高いとは言えない処理を経なければならない。

（２）検索対象文書中の単単語だけでなく「単語の組」についても、文書内での出現状況などの統計情報を予め抽出しておかなければならない。

このため、関連語となりうる「単語の組」を網羅すると、その数は膨大となり、多くの記憶容量が必要になる。

（３）抽出した「単語の組」の候補から検索に有効な関連語を選出するときに、単単語の

選出用の関連度算出方法をそのまま適用している。
しかし、単単語と「単語の組」とでは、文書内における出現状況が大きく異なるものであり、これを考慮に入れず、単純に、単単語用の関連度算出方法をそのまま適用するのでは、検索に寄与しない「単語の組」が選ばれる公算が高い。
本発明は、上述の問題を解決するためのものであり、適合文書中から、検索キーワードに関連が高く検索に寄与する単語および単語の組合せを、記憶容量を増やしたり検索速度を著しく低下させることなく選び出す文書検索装置、文書検索方法および記録媒体を提供することを目的とする。

【０００７】

【課題を解決するための手段】

上記の問題を解決するために、請求項１記載の発明の文書検索装置は、複数の文書の文書情報と、前記文書中に含まれる各単語の単語統計情報とを保持して構成される文書データベースと、前記文書データベースからキーワードに適合する適合文書および適合しない非適合文書を選出する文書ランキング部と、前記キーワードの関連語を選出する単語ランキング部と、新しいキーワードを生成するキーワード生成部と、を備え、前記文書ランキング部は、前記文書データベースから、装置に入力されたキーワードについて適合文書及び非適合文書を選出し、前記単語ランキング部は、前記適合文書中の単単語について、前記文書ランキング部で選出した適合文書および非適合文書中の出現頻度と、前記文書データベースの検索対象文書中の出現頻度と、をもとに前記キーワードとの関連度を計算し、前記関連度の高い単単語を前記キーワードの関連語として選出し、さらに、前記文書ランキング部で選出した適合文書から連続した２つ以上の単語から構成される単語列を抽出し、前記単単語の関連語の関連度の中で最小のものに基づき前記単語列の関連度の下限値を計算するとともに、前記適合文書と非適合文書中の前記単語列の出現頻度をそれぞれ計算し、該出現頻度から暫定的な関連度を計算し、該暫定的な関連度と前記関連度の下限値とを比較し、前記暫定的な関連度のほうが前記関連度の下限値より小さいときは、前記単語列を関連語候補から外したうえで、単語列を選出して、前記選出された単単語及び前記選出された単語列を前記キーワードの関連語とし、前記キーワード生成部は、前記キーワードの関連語をもとの前記キーワードに追加して新しいキーワードとし、前記文書ランキング部は、前記キーワード生成部で生成された新しいキーワードに適合する文書を検索するようにしたことを特徴とする。

また、請求項２記載の発明の文書検索装置は、請求項１に記載の文書検索装置において、前記単語ランキング部は、予め指定した不要語および記号からのみなる語を含む単語列をキーワードの関連語候補としないようにしたことを特徴とする。

【０００８】

また、請求項３記載の発明の文書検索装置は、請求項１又は２に記載の文書検索装置において、前記単語ランキング部は、前記関連語候補に残った単語列ついて、前記文書データベース中の文書に出現する単語列の実際の出現頻度を求め、該実際の出現頻度から前記単語列の関連度を計算する際に、単単語にくらべて前記文書ランキング部で選出された適合文書および非適合文書中の出現状況の影響する度合が高くなるように出現状況の比率を設定したことを特徴とする。

また、請求項４記載の発明の文書検索装置は、請求項１乃至請求項３のいずれか一項に記載の文書検索装置において、前記単語ランキング部は、前記文書ランキング部で選出された適合文書から抽出した単語列の関連語候補から前記キーワードの関連語として選出するための関連度の下限を単単語にくらべて高く設定し、単語列が関連語として選ばれる数を抑えるようにしたことを特徴とする。

また、請求項５記載の発明の文書検索装置は、請求項１乃至請求項４のいずれか一項に記載の文書検索装置において、前記文書ランキング部は、前記キーワード生成部で生成された新しいキーワードによって前記文書データベースを検索する際に、このキーワードに含まれる単単語と重複する単語を含む単語列については、その重みを下げて適合度を算出するようにしたことを特徴とする。

【０００９】
また、請求項６記載の発明の文書検索方法は、コンピュータにより実行される、前記コンピュータに入力されたキーワードに適合する文書を複数の文書を保持する文書データベースから検索する文書検索方法において、前記文書データベースからキーワードに適合する適合文書および適合しない非適合文書を選出する文書ランキング工程と、前記キーワードの関連語を選出する単語ランキング工程と、新しいキーワードを生成するキーワード生成工程と、を備え、前記文書ランキング工程は、前記文書データベースから、前記コンピュータに入力されたキーワードについて適合文書及び非適合文書を選出し、前記単語ランキング工程は、前記適合文書中の単単語について、前記文書ランキング工程で選出した適合文書および非適合文書中の出現頻度と、前記文書データベースの検索対象文書中の出現頻度と、をもとに前記キーワードとの関連度を計算し、前記関連度の高い単単語を前記キーワードの関連語として選出し、さらに、前記文書ランキング工程で選出した適合文書から連続した２つ以上の単語から構成される単語列を抽出し、前記単単語の関連語の関連度の中で最小のものに基づき前記単語列の関連度の下限値を計算するとともに、前記適合文書と非適合文書中の前記単語列の出現頻度をそれぞれ計算し、該出現頻度から暫定的な関連度を計算し、該暫定的な関連度と前記関連度の下限値とを比較し、前記暫定的な関連度のほうが前記関連度の下限値より小さいときは、前記単語列を関連語候補から外したうえで、単語列を選出して、前記選出された単単語及び前記選出された単語列を前記キーワードの関連語とし、前記キーワード生成工程は、前記キーワードの関連語をもとの前記キーワードに追加して新しいキーワードとし、前記文書ランキング工程は、前記キーワード生成工程で生成された新しいキーワードに適合する文書を検索するようにしたことを特徴とする。
また、請求項７記載の発明は、コンピュータに請求項６に記載の文書検索方法を実現させる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特徴とする。
【００１０】
【発明の実施の形態】
以下に、図面を用いて本発明の実施の形態の構成および動作を詳細に述べる。
実施の形態の構成
図１は本発明に係る文書検索装置の構成例を示すブロック図である。
この実施の形態の文書検索装置は、キーワード入力部１１０ 、文書ランキング部１２０ 、単語ランキング部１３０ 、キーワード生成部１４０、文書出力部１５０、文書データベース１６０ より構成される。
キーワード入力部１１０は、ユーザがキーボード等により、文書データベース１６０中にある文書の特徴をあらわすキーワードとなる文字列を入力する。
文書ランキング部１２０は、キーワード入力部１１０から渡されたキーワードに対して、文書データベース１６０を検索し、適合する文書と適合しない文書とを選定する。また、文書ランキング部１２０は、キーワード生成部１４０で生成された新しいキーワードに対してもう一度適合する文書を選定する。
この選定された適合文書は、文書出力部１５０へ渡される。
【００１１】
単語ランキング部１３０は、文書ランキング部１２０で選定された適合文書の中から取り出された単語と入力されたキーワードとの間で計算される関連度に応じて関連語を選出し、キーワード生成部１４０へ渡す。
それらを入力したキーワードの関連語としてキーワードに追加し、その新しいキーワードを文書ランキング部１２０へ渡す。
キーワード生成部１４０は、単語ランキング部１３０から渡された関連語をもとのキーワードに追加して新しいキーワードを生成し、文書ランキング部１２０へ渡される。
文書出力部１４０は、文書ランキング部１２０で選出した適合文書をプリンタ、表示装置、記憶装置等へ出力するか、または、ネットワークを介して他のコンピュータ装置へ送信する。

文書データベース１６０は、検索対象となる文書を保持する文書情報と、その文書中に含まれている各単語の単語統計情報から構成される（図２参照）。
たとえば、文書情報には、各文書に対して次のような情報が保持される。
文書識別子（ＩＤ）、文書名、書誌事項（作成者、作成日、発行所等）、文書実体へのポインタ等
また、単語統計情報には、単語ごとに次のような統計情報を保持する。
単語の表記、この単語の文書データベース全体での出現頻度、単語出現情報等ここで単語出現情報としては、単語が出現する文書ごとに次の情報を保持する。
この単語が出現する文書の文書識別子、この文書に出現する単語出現頻度、この文書にこの単語が出現する出現位置の一覧等

【００１２】
（２）実施の形態の動作
次に、このように構成された本実施の形態の文書検索装置の動作について、図３のフローチャートに基いて説明する。
まず、キーボード等の入力装置からキーワードの文字列を入力する（ステップＳ１００）。
これにより、キーワード入力部１１０を構成する。
このキーワードは、たとえば、英語や日本語の単語や単語の組み合わせで構成し、必要に応じて単語の組み合わせは、単単語へ分解する。
この入力されたキーワード中のそれぞれの単語について、文書データベース１６０の単語統計情報を参照し、たとえば、上記（式１）を用いて単語の重要度に応じた重みを計算する（ステップＳ１１０）。
次に、検索対象である文書データベース１６０中のそれぞれの文書に対して、文書データベース１６０の単語統計情報とステップＳ１１０で計算されたキーワードの単語の重みとを参照し、その文書にキーワード中の単語がどのくらい含まれているかを示す適合度を、たとえば、上記（式２）を用いて計算し、文書一覧表を作成する（ステップＳ１２０）。
この文書一覧表を適合度をキーとして、降順に各文書を順序付け、その上位から所定の件数（たとえば、１０件程度）の文書を適合文書とみなし、下位から所定の件数（たとえば、５００件程度）の文書を非適合文書とみなす（ステップＳ１３０）。
あるいは、順序づけられた文書の一覧表（適合度、文書名や書誌事項等の一覧）をユーザに提示し、適合しているかどうか指示させ、適合していると指示された文書を適合文書とし、適合しないと指示された文書を非適合文書とするようにしてもよい。

【００１３】
ステップＳ１１０からステップＳ１３０までにより、文書ランキング部１２０を構成する。
ステップＳ１３０で選出した適合文書がユーザの所望した文書であるかどうかをユーザに指示させる（ステップＳ１４０）。
所望した文書でなければ、ステップＳ１５０へ進む。所望した文書であれば、ステップＳ１９０へ進む。
ステップＳ１３０で選出された適合文書を表示装置、プリンタや記憶装置等の出力装置へ、たとえば、ランク順に文書名や書誌事項等を一覧として出力したり、また、ネットワークで接続された他のコンピュータ装置へ送信することによってユーザに提示される（ステップＳ１９０）。
これにより、文書出力部１５０を構成する。
ステップＳ１３０で求めた適合文書中の単語を入力キーワードの関連語の候補となる関連語単語表として作成する。これは文書データベース１６０の単語統計情報に保持された適合文書に含まれる単語を取り出して作成される。このとき、予め用意された不要語表を参照して、これに登録されている単語は関連語単語表へは登録しない。
さらに、この関連語単語表に登録された単語ごとに、適合文書および非適合文書での出現状況を文書データベース１６０の単語統計情報から取り出し、たとえば、（式３）および

（式４）を使って、キーワードとの関連度を計算する。
この関連度の高いものから順に所定の数（たとえば、１０単語程度）だけ選択し、これを単単語のキーワード関連語として抽出する（ステップＳ１５０）。
また、所定の数の単語を選定したときの最小の関連度を記憶しておき、単語列の関連度の閾値計算に使う。
【００１４】
次に、文書ランキング部１２０で抽出された適合文書中の連続する２語以上からなる単語の組合せ（以下、これを単語列と呼ぶ）を適合文書の中から抽出し、関連語候補とする。
これら抽出された関連語候補の単語列中から、予め用意した不要語リストにある不要語を含む単語列や記号のみからなる語を含んでいる単語列を関連語候補から削除する。（ステップＳ１６０）。
以下の説明では、単語を２つ組合せたものを例として説明するが、３語以上の任意の数の組合せであっても同様に考えられる。
たとえば、入力されたキーワードが「Microsoft」であって、抽出された適合文書に次の文が含まれているとする。
Microsoft Chairman Bill Gates delivered a keynote address.
この場合、関連語候補として、以下の７つの単語列が抽出できる。
Microsoft Chairman
Chairman Bill
Bill Gates
Gates delivered
delivered a
a keynote
keynote address
ここで、不要語リストの不要語として「a」が登録されていれば、単語列「a keynote」および「delivered a」を削除する。
残った関連語候補の単語列について、入力されたキーワードと関連度の高いものをキーワード関連語として選出する（ステップＳ１７０）。
本発明では、単語列の関連度は、単単語の関連度算出に使用した計算式、たとえば、上記の（式３）および（式４）において、単語を単語列と置き換えて計算する。詳細は、後述の（３）にて説明する。
ステップＳ１５０からステップＳ１７０により、単語ランキング部１３０を構成する。
単単語の関連語（ステップＳ１５０）と単語列の関連語（ステップＳ１６０およびＳ１７０）をもとのキーワードに追加して新しいキーワードを作成する（ステップＳ１８０）。
これによりキーワード生成部１４０を構成する。
【００１５】
この新しいキーワードをステップＳ１１０からステップＳ１３０（文書ランキング部１２０）の処理と同様にして、再度、適合文書を選出する。
このとき、単語列の関連語を構成する単単語が、単単語としても関連語に重複して登録されている場合には、この単語列の関連語の（式１）による重み計算は、重みに所定の係数（たとえば、０．４から０．３程度）を乗じて重みを下げるようにして文書適合度（式２）を計算する。これは、単語列を含む文書には、同時に、その単語列を構成している単単語も含んでいるので、この含有関係を考慮にいれた重みとしたいためである。
本実施の形態の文書検索装置をこのような構成にすることによって、次のような効果がある。
・形態素解析のような重い処理に依らない方法で複数の単語の組合せを抽出することができる。
・時間を増やさずに検索のつど統計情報を収集することによって記憶容量を削減することができる。
・２語以上からなる単語の組合せを選出するのに適した関連度の算出方法を提案できた。

以上によって、検索に寄与する単語列をキーワードの関連語として選出することができるので、ユーザの所望する的確な文書を検索することができる。

【００１６】

（３）単語列のキーワード関連語の抽出

たとえば、（式３）および（式４）によって単語列のキーワードとの関連度を計算するためには、次の情報が必要となる。

（Ａ）文書データベース１６０の中の文書にこの単語列が出現する文書数

これにより（式３）のWp、Wq が求められる。

これは、文書データベース１６０の単語統計情報が単語ごとの出現状況データしか持っていないので、単語ごとの単語統計情報から文書中の単語の位置情報を得た上で、出現状況データを単語列用に統合する必要があり、これを文書データベース１６０中のすべての文書に対して処理するには、多大な処理時間が必要となる。

（Ｂ）適合文書および非適合文書にこの単語列が出現する文書数

これにより（式３）のWr、Ws が求められる。

これは、適合文書および非適合文書中の各文書の内容を走査して単語列が出現しているかどうか調べればよいが、これは図３のステップＳ１５０で単単語の出現する文書数を調べるのと同時に、単語統計情報から文書中の単語の位置情報を得て行なえば、処理時間への影響は少ない。

上記（Ａ）の計算時間を短縮するために、まず、単語列をキーワードの関連語として採用するための関連度の下限値（閾値）を決め、適合文書および非適合文書に単語列が出現する文書数を計算し、それぞれ（式４）に当てはめると、重みの下限値が決まってくる。

また、上記で求めた重みの下限値や適合文書および非適合文書に単語列が出現する文書数を（式３）に当てはめると、（式３）は文書データベース１６０の文書中に単語列が出現する出現頻度（以下、ｎとする）の単一変数の関数になっているので、これを解けば出現頻度を計算することができる。しかし、これを解かずとも次のように考えれば、その文書に単語列が出現するかどうかを最終的に走査することなく判断することができる。

【００１７】

（式３）は、検索対象文書中にこの単語列が出現する出現頻度（ｎ）の単調減少関数であるから、ｎ＝１のときが最大値を持つことになるので、各単語列でｎ＝１としたときの関連度（式４で計算される）が先に決めた関連度の下限値（閾値）より小さければ、当然のことに、ｎを実際に求めた関連度はさらに小さい値となることになる。

したがって、ｎ＝１として（式３）、（式４）から計算した単語列の関連度が先に決めた関連度の下限値（閾値）より小さい単語列は、関連語の候補からはずすことができる。これにより、検索対象の文書数より適合文書数や非適合文書数の方が小さいため、これらの適合文書や非適合文書での単語列の出現頻度を計算したとしても、全体の処理時間を大幅に削減することができる。

上記の単語列が関連語として採用される関連度の下限値（閾値）は、先に単単語の関連語を求めて記憶してある最小の関連度に一定の係数（たとえば、５以下程度の値とし、経験的には２．５から５位を採用する）を乗じた値とし、単単語が関連語として選出されるのに比べ、関連度の下限値（閾値）を高めに設定する。このように関連語に選出される単語列の数を減らしておけば、この後に、関連語を追加した新キーワードで再度、適合文書を検索するときに、処理時間が短くてすむ。

以上のことを考慮して、図４に示した手順で関連語候補の単語列からキーワード関連語を選出する。

【００１８】

単単語の関連語を選出したときに記憶した単単語の最小関連度に所定の係数を乗じて、単語列の関連度の下限値（閾値）を計算する（ステップＳ２００）。

各単語列に対して、各適合文書中に出現する出現頻度、および、各非適合文書中に出現する出現頻度をそれぞれ計算する（ステップＳ２１０）。

各単語列に対して、検索対象文書に出現する出現頻度（ｎ）を１とし、ステップＳ２１０

で計算した各適合文書中に出現する出現頻度、および、各非適合文書中に出現する出現頻度を用いて、（式３）と（式４）とから暫定的な単語列の関連度を計算し、この暫定的な関連度とステップＳ２０で計算した関連度の下限値（閾値）と比較する。暫定的な関連度の方が閾値より小さいときには、この単語列を関連語候補からはずす（ステップＳ２２０）。
残った関連語候補の単語列について、文書データベース１６０中の文書に出現する単語列の出現頻度を実際に求め、（式３）と（式４）によって単語列の関連度を計算する。
このとき（式３）によって、単単語にくらべて、適合文書の中での出現状況 Wr、Ws が重みに影響する度合が高くなるように、比率 Cp、Cq を予め設定しておき、単語列の重みを計算する。関連語としての単語列は、検索対象文書中に出現する頻度が少ないと考えた方が一般的であるため、検索対象文書集合の中での出現状況 Wp、Wq より、適合文書および非適合文書の中での出現状況 Wr、Ws の方を重みの判断基準としてよりふさわしいと考えられるからである。
この計算した単語列の関連度が、先に決めた関連度の下限値（閾値）より大きい場合、この単語列をキーワード関連語とする（ステップＳ２３０）。

【００１９】

＜コンピュータによる実施例＞

さらに、本発明は上記の実施の形態のみに限定されたものではない。たとえば、図１に示した文書検索装置は、図５のようなハードウェア構成を持つコンピュータ装置２００によっても実現が可能である。
即ち、コンピュータ装置２００は、キーボード、マウス、タッチパネル、スキャナ等により構成され、情報の入力に使用される入力装置１と、種々の出力情報や入力装置１からの入力された情報などを表示出力させる表示装置２と、種々のプログラムを動作させるＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ；中央処理ユニット）３と、プログラム自身を保持し、またそのプログラムがＣＰＵ３によって実行されるときに一時的に作成される情報等を保持するメモリ４と、本発明の文書検索装置で扱う文書データベース１６０およびプログラムやプログラム実行時の一時的な情報等を保持する記憶装置５と、プログラムやデータ等を記憶した記録媒体を装着してそれらを読み込み、メモリ４または記憶装置５へ格納するのに用いられる媒体駆動装置６と、ネットワーク９へ接続するためのインタフェースであるネットワーク接続装置７とから構成され、それらはバス８で接続されている。
また、ネットワーク９は、コンピュータ装置２００と他のコンピュータ装置２００とを結合するための伝送路であって、一般には、ケーブルで実現され、通信プロトコルにはＴＣＰ／ＩＰが使われる。但し、伝送路としてはケーブルだけではなく、それらの間の通信プロトコルが一致するものであれば無線、有線および放送波のいずれでもよく、たとえば、ＬＡＮ（Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）、ＷＡＮ（Ｗｉｄｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）、インターネット、アナログ電話網、デジタル電話網（ＩＳＤＮ：Ｉｎｔｅｇｒａｌ　Ｓｅｒｖｉｃｅ　Ｄｉｇｉｔａｌ　Ｎｅｔｗｏｒｋ）、ＰＨＳ（パーソナルハンディシステム）、携帯電話網、衛星通信網などを用いることができる。
このようなコンピュータ装置２００の構成において、図１に示した文書検索装置を構成する各機能をそれぞれプログラム化し、予めＣＤ－ＲＯＭ等の記録媒体に書き込んでおき、このＣＤ－ＲＯＭを各サイトのＣＤ－ＲＯＭドライブのような媒体駆動装置６を搭載したコンピュータ装置に装着して、これらのプログラムをそれぞれのコンピュータ装置のメモリ４あるいは記憶装置５に格納し、それを実行することによって、上記の実施の形態と同様な機能を実現することができる。

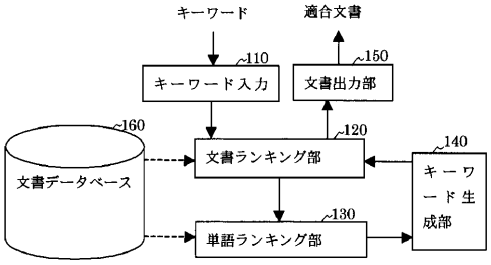
【００２０】

なお、記録媒体としては半導体媒体（たとえば、ＲＯＭ、ＩＣメモリカード等）、光媒体（たとえば、ＤＶＤ、ＭＯ、ＭＤ、ＣＤ－Ｒ等）、磁気媒体（たとえば、磁気テープ、フレキシブルディスク等）のいずれであってもよい。
また、コンピュータ装置２００のメモリ４へロードしたプログラムを実行することにより

前述した実施の形態の機能が実現されるだけでなく、そのプログラムの指示に基づき、オペレーティングシステム等が実際の処理の一部または全部を行い、その処理によって上述した実施の形態の機能が実現される場合も含まれる。
また、上述した実施の形態を実現するプログラムがＲＯＭ等のような半導体の記録媒体である場合には、媒体駆動装置６からではなく、直接、メモリ４へロードして実行される。
【００２１】
＜本発明のネットワーク環境での運用＞
図６は、本発明を有線または無線の通信ネットワークに接続して運用する形態の構成を示している。
たとえば、文書検索プログラムを保持するサーバー３００と複数のユーザが利用する端末３１０とをネットワーク９で接続する。
この場合、サーバー３００およびユーザの端末３１０は、図５に示した汎用のコンピュータ装置２００で構成される。
ユーザは、端末３１０からサーバー３００に対してログインしたり、文書検索のためのキーワードを入力し、サーバー３００の文書検索プログラムへ検索の実行を依頼する。サーバー３００の文書検索プログラムは指定されたキーワードに適合した検索結果を要求もとの端末３１０へ戻す。ユーザの端末３１０は、この検索結果を出力する。
このようにすることで、常に最新の文書検索プログラムを使えるという利点がある。
また、図６のようにサーバー３００と端末３１０とを有線または無線の通信ネットワークで接続した場合、サーバー３００の磁気ディスク等の記憶装置に本発明の機能を実現する文書検索プログラムを格納しておき、端末３１０に対してダウンロード等の形式で頒布することも可能である。
さらに、本発明の機能を実現する文書検索プログラムを媒体や放送波による配布で提供するようにしてもよい。
【００２２】
【発明の効果】
以上説明したように、本発明によれば、２語以上からなる関連語句を効率良く得ることができ、記憶容量を増やしたり検索速度を著しく低下させることなく、検索精度を向上させることができる。
【図面の簡単な説明】
【図１】本発明に係る文書検索装置の構成例を示すブロック図である。
【図２】図１中の文書データベースのデータ構造を説明するための図である。
【図３】図１に示す文書検索装置における処理の流れを説明するためのフローチャートである。
【図４】単語列から関連語候補を削除する処理の流れを説明するためのフローチャートである。
【図５】本発明に係る文書検索装置をコンピュータで実現するときのハードウェアの構成例を示す図である。
【図６】本発明をネットワーク環境で運用する場合を説明するための図である。
【符号の説明】
１１０　……　キーワード入力部
１２０　……　文書ランキング部
１３０　……　単語ランキング部
１４０　……　キーワード生成部
１５０　……　文書出力部
１６０　……　文書データベース
２００　……　コンピュータ装置
３００　……　サーバー
３１０　……　端末
１　……　入力装置

2 …… 表示装置
3 …… ＣＰＵ
4 …… メモリ
5 …… 記憶装置
6 …… 媒体駆動装置
7 …… ネットワーク接続装置
8 …… バス
9 …… ネットワーク

【図１】



【図２】

文書データベース１６０のデータ構造

文書情報

文書 ID	文書名	書誌事項	……	文書実体へのポインタ

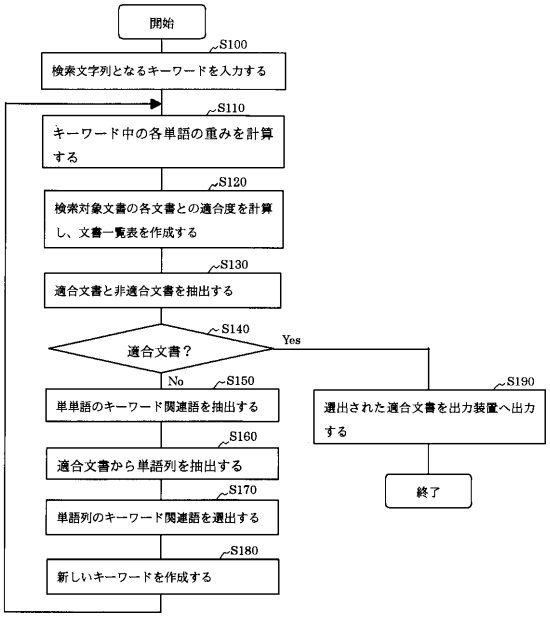
文書ファイル

文書データベース中の単語に関する統計情報

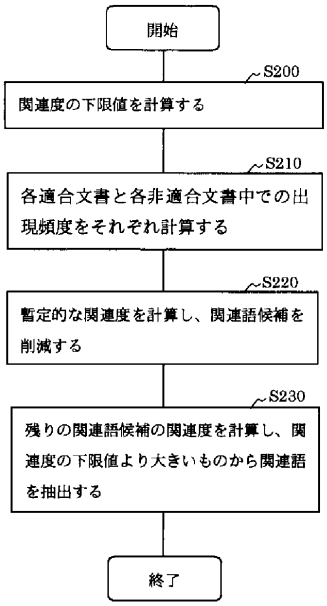
単語の表記	出現頻度	単語出現情報	……

（注）単語出現情報には、単語ごとに次の情報がある。
（１） この単語が出現する文書の文書識別子（文書ＩＤ）、
（２） この文書の中にこの単語が出現する頻度、
（３） この文書にこの単語が出現する出現位置の一覧

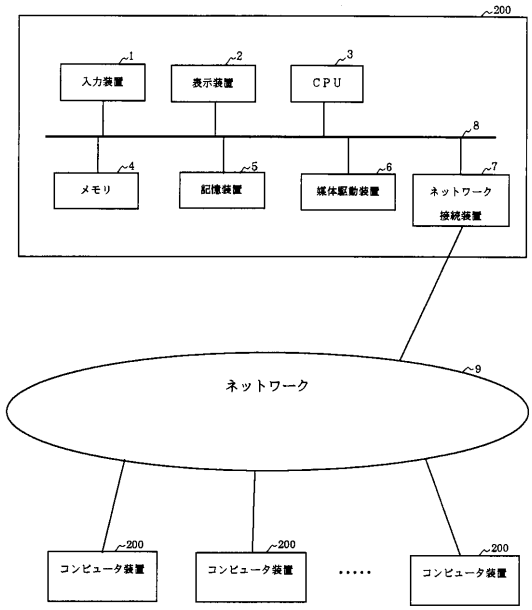
【図３】


開始
S100
検索文字列となるキーワードを入力する
S110
キーワード中の各単語の重みを計算する
S120
検索対象文書の各文書との適合度を計算し、文書一覧表を作成する
S130
適合文書と非適合文書を抽出する
S140
適合文書？
Yes
No
S150
単単語のキーワード関連語を抽出する
S160
適合文書から単語列を抽出する
S170
単語列のキーワード関連語を選出する
S180
新しいキーワードを作成する
S190
選出された適合文書を出力装置へ出力する
終了

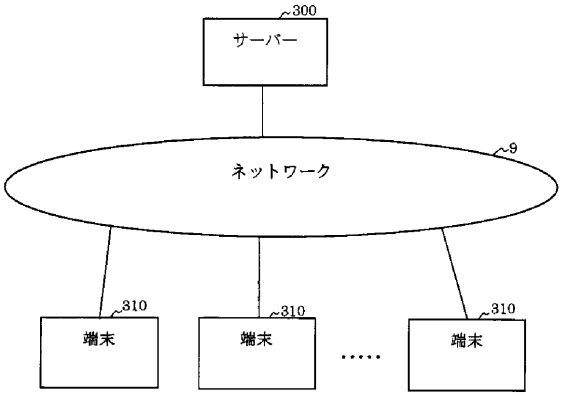
【図４】


開始
S200
関連度の下限値を計算する
S210
各適合文書と各非適合文書中での出現頻度をそれぞれ計算する
S220
暫定的な関連度を計算し、関連語候補を削減する
S230
残りの関連語候補の関連度を計算し、関連度の下限値より大きいものから関連語を抽出する
終了

【図５】


200
1
入力装置
2
表示装置
3
CPU
8
4
メモリ
5
記憶装置
6
媒体駆動装置
7
ネットワーク接続装置
9
ネットワーク
200
コンピュータ装置
200
コンピュータ装置
.....
200
コンピュータ装置

【図６】


300
サーバー
9
ネットワーク
310
端末
310
端末
.....
310
端末

フロントページの続き

(56)参考文献 特開平１１－３２８１８２（ＪＰ，Ａ）
特開平１１－０２５１０８（ＪＰ，Ａ）
特開平０９－０５４７７７（ＪＰ，Ａ）
特開平０３－１２５２６５（ＪＰ，Ａ）
ROBERTSON S.E., WALKER S., On relevance weights with little relevance information, SIGIR 97, 米国, ACM, １９９７年, pp. 16-24
WALKER S, et al., Okapi at TREC-6 Automatic ad hoc, VLC, routing, filetering and QSDR, The Sixth Test REtrieval Conference (TERC-6), １９９６年

(58)調査した分野(Int.Cl., ＤＢ名)
G06F 17/30