(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3598742号
(P3598742)
(45)発行日　平成16年12月8日(2004.12.8)　(24)登録日　平成16年9月24日(2004.9.24)

(51)Int.Cl.[7]　F I
G06F 17/30　G06F 17/30 210A
G06F 17/30 210D
G06F 17/30 140

請求項の数 15　(全 46 頁)

(21)出願番号　特願平9-166100
(22)出願日　平成9年6月23日(1997.6.23)
(65)公開番号　特開平10-207911
(43)公開日　平成10年8月7日(1998.8.7)
審査請求日　平成13年7月10日(2001.7.10)
(31)優先権主張番号　特願平8-313761
(32)優先日　平成8年11月25日(1996.11.25)
(33)優先権主張国　日本国(JP)

(73)特許権者　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目17番22号
(74)代理人　100092152
弁理士　服部　毅巖
(72)発明者　沼田　賢一
神奈川県足柄上郡中井町境430　グリーンテクなかい　富士ゼロックス株式会社内

審査官　深津　始

(56)参考文献　特開平3－174653（JP，A）
特開平8－249356（JP，A）

最終頁に続く

(54)【発明の名称】文書検索装置及び文書検索方法

(57)【特許請求の範囲】
【請求項１】
ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を検索し、構成要素毎に分類して表示する文書検索装置において、
前記文書を蓄積する文書蓄積手段と、
分類しようとする前記構成要素の分類単位を指定する分類単位指定手段と、
前記文書蓄積手段から文書を読み込み、その論理構造を解析する論理構造解析手段と、
前記分類単位指定手段にて指定された分類単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する基本ベクトル生成手段と、
前記分類単位の構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しからキーワードを抽出し、見出しベクトルを生成する見出しベクトル生成手段と、
前記基本ベクトルと、当該基本ベクトルの生成元となった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しから生成された前記見出しベクトルとを合成し、合成ベクトルを生成するベクトル合成手段と、
前記合成ベクトルを前記分類単位の構成要素と対応付けて保持する合成ベクトル保持手段と、
前記合成ベクトルの類似度に基づいて、前記文書の構成要素を分類する分類手段と、
前記分類手段による分類の結果を表示する表示手段と、
を有することを特徴とする文書検索装置。

【請求項２】
前記分類手段による分類の結果を保持する分類結果保持手段と、前記分類の結果に対し再分類を指示する再分類指示手段と、をさらに有することを特徴とする請求項１記載の文書検索装置。
【請求項３】
前記分類手段による分類の結果から、内容を表示するための構成要素を選択する表示要素選択部と、選択された前記構成要素よりも論理構造的に上位の構成要素の単位から表示させたい構成要素の表示単位を選択する表示範囲選択部と、前記表示要素選択部にて選択された前記構成要素から前記文書の木構造上の根に至る経路上に位置する前記表示単位の構成要素の下位の全ての構成要素を前記文書蓄積手段から抽出し、前記表示手段に表示させる表示文書抽出部と、を含む検索文書抽出手段を、さらに有することを特徴とする請求項１記載の文書検索装置。
【請求項４】
ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を、構成要素毎に検索して表示する文書検索装置において、
前記文書を蓄積する文書蓄積手段と、
検索しようとする前記構成要素の検索単位を指定する検索単位指定手段と、
前記文書蓄積手段から文書を読み込み、その論理構造を解析する論理構造解析手段と、
前記検索単位指定手段にて指定された検索単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する基本ベクトル生成手段と、
前記検索単位の構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しからキーワードを抽出し、見出しベクトルを生成する見出しベクトル生成手段と、
前記基本ベクトルと、当該基本ベクトルの生成元となった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しから生成された前記見出しベクトルとを合成し、合成ベクトルを生成するベクトル合成手段と、
前記合成ベクトルを前記検索単位の構成要素と対応付けて保持する合成ベクトル保持手段と、
キーワード列または自然言語文により構成される問合せを入力する問合せ入力手段と、
問合せから問合せベクトルを生成する問合せベクトル生成手段と、
前記問合せベクトルと前記合成ベクトルとの類似度に基づいて、前記検索単位の構成要素を検索して、類似度の高い合成ベクトルに対応する構成要素を抽出する検索手段と、
前記検索手段による検索の結果を表示する表示手段と、
を有することを特徴とする文書検索装置。
【請求項５】
前記検索手段による検索の結果から、内容を表示するための構成要素を選択する表示要素選択部と、選択された前記構成要素よりも論理構造的に上位の構成要素の単位から表示させたい構成要素の表示単位を選択する表示範囲選択部と、前記表示要素選択部にて選択された前記構成要素から前記文書の木構造上の根に至る経路上に位置する前記表示単位の構成要素の下位の全ての構成要素を前記文書蓄積手段から抽出し、前記表示手段に表示させる表示文書抽出部と、を含む検索文書抽出手段を、さらに有することを特徴とする請求項４記載の文書検索装置。
【請求項６】
ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を検索し、構成要素毎に分類して表示する文書検索装置において、
前記文書を蓄積する文書蓄積手段と、
分類しようとする前記構成要素の分類単位を指定する分類単位指定手段と、
前記文書蓄積手段から文書を読み込み、その論理構造を解析する論理構造解析手段と、
前記分類単位指定手段にて指定された分類単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する基本ベクトル生成手段と、

前記基本ベクトル生成手段にて生成された基本ベクトルのそれぞれに対して、その元になった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素それぞれに対応して、その構成要素の下位の全ての構成要素のうち当該経路上に位置しない構成要素の内容から要素ベクトルを生成し、生成した当該要素ベクトルを合成して内容ベクトルを生成する内容ベクトル生成手段と、
前記基本ベクトルと、当該基本ベクトルの生成元となった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素に応じて生成された前記内容ベクトルとを合成し、合成ベクトルを生成するベクトル合成手段と、
前記合成ベクトルを前記分類単位の構成要素と対応付けて保持する合成ベクトル保持手段と、
前記合成ベクトルの類似度に基づいて、前記文書の構成要素を分類する分類手段と、
前記分類手段による分類の結果を表示する表示手段と、
を有することを特徴とする文書検索装置。

【請求項７】
前記内容ベクトル生成手段は、前記要素ベクトルが見出しであるか否かによって、前記各要素ベクトルを合成する際に、前記各要素ベクトルに異なる重み付けを行うことを特徴とする請求項６記載の文書検索装置。

【請求項８】
前記内容ベクトル生成手段は、前記基本ベクトルと前記要素ベクトルとの類似度に応じて、前記各要素ベクトルを合成する際に、前記各要素ベクトルに異なる重み付けを行うことを特徴とする請求項６記載の文書検索装置。

【請求項９】
前記分類手段による分類の結果を保持する分類結果保持手段と、前記分類の結果に対し再分類を指示する再分類指示手段と、をさらに有することを特徴とする請求項６記載の文書検索装置。

【請求項１０】
前記分類手段による分類の結果から、内容を表示するための構成要素を選択する表示要素選択部と、選択された前記構成要素よりも論理構造的に上位の構成要素の単位から表示させたい構成要素の表示単位を選択する表示範囲選択部と、前記表示要素選択部にて選択された前記構成要素から前記文書の木構造上の根に至る経路上に位置する前記表示単位の構成要素の下位の全ての構成要素を前記文書蓄積手段から抽出し、前記表示手段に表示させる表示文書抽出部と、を含む検索文書抽出手段を、さらに有することを特徴とする請求項６記載の文書検索装置。

【請求項１１】
ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を、構成要素毎に検索して表示する文書検索装置において、
前記文書を蓄積する文書蓄積手段と、
検索しようとする前記構成要素の検索単位を指定する検索単位指定手段と、
前記文書蓄積手段から文書を読み込み、その論理構造を解析する論理構造解析手段と、
前記検索単位指定手段にて指定された検索単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する基本ベクトル生成手段と、
前記基本ベクトル生成手段にて生成された基本ベクトルのそれぞれに対して、その元になった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素それぞれに対応して、その構成要素の下位の全ての構成要素のうち当該経路上に位置しない構成要素の内容から要素ベクトルを生成し、生成した当該要素ベクトルを合成して内容ベクトルを生成する内容ベクトル生成手段と、
前記基本ベクトルと、当該基本ベクトルの生成元となった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素に応じて生成された前記内容ベクトルとを合成し、合成ベクトルを生成するベクトル合成手段と、
前記合成ベクトルを前記検索単位の構成要素と対応付けて保持する合成ベクトル保持手段

と、
キーワード列または自然言語文により構成される問合せを入力する問合せ入力手段と、
問合せから問合せベクトルを生成する問合せベクトル生成手段と、
前記問合せベクトルと前記合成ベクトルとの類似度に基づいて、前記検索単位の構成要素を検索して、類似度の高い合成ベクトルに対応する構成要素を抽出する検索手段と、
前記検索手段による検索の結果を表示する表示手段と、
を有することを特徴とする文書検索装置。
【請求項１２】
前記内容ベクトル生成手段は、前記要素ベクトルが見出しであるか否かによって、前記各要素ベクトルを合成する際に、前記各要素ベクトルに異なる重み付けを行うことを特徴とする請求項１１記載の文書検索装置。
【請求項１３】
前記内容ベクトル生成手段は、前記基本ベクトルと前記要素ベクトルとの類似度に応じて、前記各要素ベクトルを合成する際に、前記各要素ベクトルに異なる重み付けを行うことを特徴とする請求項１１記載の文書検索装置。
【請求項１４】
前記検索手段による検索の結果から、内容を表示するための構成要素を選択する表示要素選択部と、選択された前記構成要素よりも論理構造的に上位の構成要素の単位から表示させたい構成要素の表示単位を選択する表示範囲選択部と、前記表示要素選択部にて選択された前記構成要素から前記文書の木構造上の根に至る経路上に位置する前記表示単位の構成要素の下位の全ての構成要素を前記文書蓄積手段から抽出し、前記表示手段に表示させる表示文書抽出部と、を含む検索文書抽出手段を、さらに有することを特徴とする請求項１１記載の文書検索装置。
【請求項１５】
コンピュータにより、ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を検索し、構成要素毎に分類して表示する文書検索方法において、
文書蓄積手段が、前記文書を蓄積するステップと、
分類単位指定手段が、分類しようとする前記構成要素の分類単位を指定するステップと、
論理構造解析手段が、前記蓄積された文書を読み込み、その論理構造を解析するステップと、
基本ベクトル生成手段が、前記指定された分類単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成するステップと、
見出しベクトル生成手段が、前記分類単位の構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しからキーワードを抽出し、見出しベクトルを生成するステップと、
ベクトル合成手段が、前記基本ベクトルと、当該基本ベクトルの生成元となった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しから生成された前記見出しベクトルとを合成し、合成ベクトルを生成するステップと、
合成ベクトル保持手段が、前記合成ベクトルを前記分類単位の構成要素と対応付けて保持するステップと、
分類手段が、前記合成ベクトルの類似度に基づいて、前記文書の構成要素を分類するステップと、
表示手段が、分類の結果を表示するステップと、
を有することを特徴とする文書検索方法。
【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及び文書検索方法に関し、特にある構成要素が他の構成要素を含んでいくような木構造の構造を持った文書群を検索し、構成要素毎に分類して表示する文書検索装置及び文書検索方法に関する。

【０００２】
【従来の技術】
通常、文書の特徴を捉えるためには、文書内容からキーワードを抽出する方法が一般的である。キーワードは、形態素解析や、予め用意された単語リストとの照合等によって抽出される。
【０００３】
抽出されたキーワードを利用して、文書の検索を自動的に行うシステムも考えられている。キーワードを利用して文書自動検索システムを構成する場合には、ベクトル空間モデルと呼ばれる手法を利用することがある。この手法では、文書と文書検索の問合せ（以下、単に問合せという）とをキーワードの重みを要素としたベクトルで表現する。そして、文書のベクトルと問合せのベクトルとの類似度を計算し、類似度の高い文書から順に文書検索の結果として出力する。この際、予めベクトルの類似度の高い文書同士を、カテゴリーと呼ばれる集合に分類しておき、検索速度を上げることも考えられている。
【０００４】
なお、キーワードに対応する重みには、そのキーワードが文書中で用いられていれば正の値を、用いられていなければ０を設定する。また重みとして設定する値には、対応するキーワードの文書中の出現頻度ＴＦ（Ｔｅｒｍ　Ｆｒｅｑｕｅｎｃｙ）とカテゴリー内での分散度ＩＤＦ（Ｉｎｖｅｒｓｅ　Ｄｏｃｕｍｅｎｔ　Ｆｒｅｑｕｅｎｃｙ）とを用いる場合が多い。
【０００５】
このような文書自動検索システムでは、抽出されたキーワードから、文書と問合せとの類似度や文書同士の類似度、文書とカテゴリーとの類似度等を計算している。類似度の計算方法には、共通したキーワードの数が多いものを類似度が高いものとする単純な方法や、上記のベクトル空間モデルのように、キーワードの出現頻度および分散に基づいて重み付けを行ったベクトルの類似度計算を行う方法など、さまざまなものがある。なお、ベクトルの類似度計算の方法には、ベクトルの内積やコサイン係数を利用する。このような文書自動検索システムでは検索の単位を文書としているものが多い。
【０００６】
しかし、検索の単位を文書とした場合、検索結果として得られるのは文書全体であり、検索者はその文書がどんなに長くとも、その全体に目を通さなければ必要な部分を見つけることはできない。また逆に、文章中に複数の話題が展開されている場合に、そこに含まれる話題に関する問合せを行っても、文書全体と問合せとの類似度が低いために、検索洩れとなることがある。
【０００７】
これらの問題点について、ここで具体例をあげて説明する。
図２９は、自動検索の対象文書の例を示した図である。
ここでは、４つの文書３０１、３０２、３０３、３０４と各々の内容から抽出されたキーワード集合とを共に示す。但し、ここでは説明を簡単にするために、実際の文書よりも出現するキーワード数を少なくしてある。
【０００８】
４つの文書３０１、３０２、３０３、３０４は、キーワードの共有の度合に基づいて、２つのカテゴリー３１０と３２０とに分類できる。カテゴリー３１０は、文書３０１と文書３０２との共通のキーワード「情報、空間、可視化、構造、アーキテクチャ、実験」から、おそらく「情報空間の可視化を行うシステムについて記述したもの」と推定できる。同様に、カテゴリー３２０は、文書３０３と文書３０４との共通のキーワード「文書、類似度、ベクトル、分類、実験、評価、適合率」から、「文書を類似度に基づいて分類するシステムについて記述したもの」と推定できる。
【０００９】
図３０は、図２９に示した各文書を、その段落構成と共に示したものである。各文書は複数の段落に分かれており、各段落ごとにキーワードが抽出されている。なお、図２９に示

した各文書に対応するキーワード集合は、各文書の全段落のキーワード集合の論理和をとったものである。

【００１０】

ここで、利用者が「文書検索」について興味を持っているとする。図３０に示した文書の中で「文書検索」について述べていると思われる箇所は、文書３０２の第２段落と、文書３０３の第２段落である。

【００１１】

しかし、図２９に示した分類では文書３０２と文書３０３とは、それぞれ別のカテゴリー３１０、３２０に分類されている。その上、カテゴリー３１０の主題は「情報空間の可視化」であり、カテゴリー３２０の主題は「文書の分類」であり、どちらにも「文書検索」との関連性はない。これでは、カテゴリー３１０、３２０に「文書検索」について述べている文書が分類されていると推測することは難しい。

【００１２】

つまり、検索の単位が文書である場合、文書中に複数の話題が展開されていても、その情報が文書全体の主題の中に埋もれてしまう。そのため、せっかく関連性のある話題が蓄積されていても、必要なときに抽出できないという問題が生じてしまう。

【００１３】

文書自動検索システムにおける上記の問題を解決する方法として、文書を章、節、段落などの論理的な構成要素に分割し、分割された構成要素を単位として、検索を行う方法がある。

【００１４】

例えば、文書から章見出しと段落を抽出し、問合せと章見出しとの類似度、および問合せと段落の類似度を各々計算し、２つの類似度を加算して、これを問合せと章全体との類似度とし、類似度の高い順に章単位で検索結果として出力する方法がある（特開平４－８４２７１号公報「文書内情報検索装置」）。この方法では問合せに関連する単語が章見出しと段落との両方に含まれている章を、一方にしか含まれない章よりも、高い候補順位で検索できる。

【００１５】

【発明が解決しようとする課題】

しかし、この方法では文書を章と段落のみからなるものと仮定しており、それ以上の複雑な階層構造を持つ文書については、考慮されていない。また、各章がそれぞれ独立の情報として扱われており、その章の内容が文書内に占める位置、すなわち文脈については考慮されない。

【００１６】

図３１は、表題、章、節、段落などの論理構成要素を持つ文書の一例を示す図である。図において、段落Ｐ２は、「『情報検索の技術動向』という表題の文書中の、『自然言語処理の利用』について記述された章の、特に『文書構造解析』について記述された節の中にある」という文脈（文章背景）を持っている。

【００１７】

しかし、上記の方法でこの段落を検索する場合、類似度計算の対象となるのはキーワード「文、意味、役割、検索」および３章２節の見出しから抽出されるキーワード「文書、構造、解析」のみであって、上記文脈は何ら考慮されない。従って、段落Ｐ２は自然言語処理を利用した技術について述べているものとは判断されず、結果として検索されないことになる。

【００１８】

また、文脈を考慮せずに構成要素単位でカテゴリーへの分類を行っても、同様の問題が生じる。図３１において段落Ｐ２の記述に基づいて、段落を単位に分類しても、段落Ｐ２が「自然言語処理の中の文書構造解析に関わるもの」であることは、分類の結果には反映されない。

【００１９】

つまり、文書の構成要素は、それのみで必要な情報を十分に含んでいるとは限らない。
本発明はこのような点に鑑みてなされたものであり、文書全体の文脈も考慮に入れた上で構成要素単位の検索を可能とする文書検索装置及び文書検索方法を提供することを目的とする。

【００２０】

【課題を解決するための手段】

本発明では上記課題を解決するために、ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を検索し、構成要素毎に分類して表示する文書検索装置において、前記文書を蓄積する文書蓄積手段と、分類しようとする前記構成要素の分類単位を指定する分類単位指定手段と、前記文書蓄積手段から文書を読み込み、その論理構造を解析する論理構造解析手段と、前記分類単位指定手段にて指定された分類単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する基本ベクトル生成手段と、前記分類単位の構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しからキーワードを抽出し、見出しベクトルを生成する見出しベクトル生成手段と、前記基本ベクトルと、当該基本ベクトルの生成元となった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しから生成された前記見出しベクトルとを合成し、合成ベクトルを生成するベクトル合成手段と、前記合成ベクトルを前記分類単位の構成要素と対応付けて保持する合成ベクトル保持手段と、前記合成ベクトルの類似度に基づいて、前記文書の構成要素を分類する分類手段と、前記分類手段による分類の結果を表示する表示手段と、を有することを特徴とする文書検索装置が提供される。

【００２１】

このような構成の文書検索装置で、ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を検索し、構成要素毎に分類して表示する場合、文書蓄積手段が文書を蓄積する。分類単位指定手段にて分類しようとする構成要素の分類単位が指定されると、論理構造解析手段は、文書蓄積手段から文書を読み込み、その論理構造を解析する。基本ベクトル生成手段は、分類単位指定手段にて指定された分類単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する。また、見出しベクトル生成手段は、分類単位の構成要素から文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しからキーワードを抽出して見出しベクトルを生成する。ベクトル合成手段は、基本ベクトルと、基本ベクトルの生成元となった構成要素から文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しから生成された見出しベクトルとを合成して、合成ベクトルを生成する。合成ベクトル保持手段は、生成された合成ベクトルを、分類単位の各構成要素と対応付けて保持する。分類手段は、合成ベクトルの類似度に基づいて文書の構成要素を分類する。そして、表示手段は、分類の結果を表示する。

【００２２】

このように、本発明の文書検索装置では、検索対象とする文書の分類単位毎の内容と、その分類単位よりも上位に配置された構成要素の見出しとをキーワードにて示し、それらのキーワードから生成したベクトルの類似度に基づいて、検索対象文書の構成要素を分類する。従って、類似度が高く、相互に関連性のある構成要素を文章全体の見出しの情報も生かして絞り込み、表示させることができる。

【００２３】

また、本発明では上記課題を解決するために、ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を構成要素毎に検索して表示する文書検索装置において、前記文書を蓄積する文書蓄積手段と、検索しようとする前記構成要素の検索単位を指定する検索単位指定手段と、前記文書蓄積手段から文書を読み込み、その論理構造を解析する論理構造解析手段と、前記検索単位指定手段にて指定された検索単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する基本ベクトル生成手段と、前記検索単位の構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しからキーワードを抽出し、見出しベクトルを生成する見出しベクトル生成手段

と、前記基本ベクトルと、当該基本ベクトルの生成元となった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しから生成された前記見出しベクトルとを合成し、合成ベクトルを生成するベクトル合成手段と、前記合成ベクトルを前記検索単位の構成要素と対応付けて保持する合成ベクトル保持手段と、キーワード列または自然言語文により構成される問合せを入力する問合せ入力手段と、問合せから問合せベクトルを生成する問合せベクトル生成手段と、前記問合せベクトルと前記合成ベクトルとの類似度に基づいて、前記検索単位の構成要素を検索して、類似度の高い合成ベクトルに対応する構成要素を抽出する検索手段と、前記検索手段による検索の結果を表示する表示手段と、を有することを特徴とする文書検索装置が提供される。

【００２４】

このような構成の文書検索装置で、ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を構成要素毎に検索して表示する場合、文書蓄積手段が文書を蓄積する。検索単位指定手段にて検索しようとする構成要素の検索単位が指定されると、論理構造解析手段は、文書蓄積手段から文書を読み込み、その論理構造を解析する。基本ベクトル生成手段は、検索単位指定手段にて指定された検索単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する。見出しベクトル生成手段は、検索単位の構成要素から文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しからキーワードを抽出し、見出しベクトルを生成する。ベクトル合成手段は、基本ベクトルと、基本ベクトルの生成元となった構成要素から文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しから生成された見出しベクトルとを合成し、合成ベクトルを生成する。合成ベクトル保持手段は、合成ベクトルを検索単位の構成要素と対応付けて保持する。問合せ入力手段は、キーワード列または自然言語文により構成される問合せを入力する。問合せベクトル生成手段は、問合せから問合せベクトルを生成する。検索手段は、問合せベクトルと合成ベクトルとの類似度に基づいて、検索単位の構成要素を検索して、類似度の高い合成ベクトルに対応する構成要素を抽出する。そして、表示手段が検索手段による検索の結果を表示する。

【００２５】

このように、本発明の文書検索装置では、検索対象とする文書の検索単位毎の内容と、その検索単位よりも上位に配置された構成要素の見出しとをキーワードにて示し、それらのキーワードから生成した合成ベクトルと、問合せから生成した問合せベクトルと、の類似度に基づいて構成要素を検索する。従って、問合せと同一キーワードを含む確率の高い構成要素を、文章全体の見出しの情報も生かして検索することができる。

【００２６】

また、本発明では上記課題を解決するために、ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を検索し、構成要素毎に分類して表示する文書検索装置において、前記文書を蓄積する文書蓄積手段と、分類しようとする前記構成要素の分類単位を指定する分類単位指定手段と、前記文書蓄積手段から文書を読み込み、その論理構造を解析する論理構造解析手段と、前記分類単位指定手段にて指定された分類単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する基本ベクトル生成手段と、前記基本ベクトル生成手段にて生成された基本ベクトルのそれぞれに対して、その元になった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素それぞれに対応して、その構成要素の下位の全ての構成要素のうち当該経路上に位置しない構成要素の内容から要素ベクトルを生成し、生成した当該要素ベクトルを合成して内容ベクトルを生成する内容ベクトル生成手段と、前記基本ベクトルと、当該基本ベクトルの生成元となった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素に応じて生成された前記内容ベクトルとを合成し、合成ベクトルを生成するベクトル合成手段と、前記合成ベクトルを前記分類単位の構成要素と対応付けて保持する合成ベクトル保持手段と、前記合成ベクトルの類似度に基づいて、前記文書の構成要素を分類する分類手段と、前記分類手段による分類の結果を表示する表示手段と、を有することを特徴とする文書検索装置が提供される。

【００２７】
このような構成の文書検索装置で、ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を検索し、構成要素毎に分類して表示する場合、文書蓄積手段が文書を蓄積する。分類単位指定手段が、分類しようとする構成要素の分類単位を指定すると、論理構造解析手段が、文書蓄積手段から文書を読み込んでその論理構造を解析する。基本ベクトル生成手段は、分類単位指定手段にて指定された分類単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する。また、内容ベクトル生成手段は、基本ベクトル生成手段にて生成された基本ベクトルのそれぞれに対して、その元になった構成要素から文書の木構造上の根に至る経路上に位置する全ての構成要素それぞれに対応して、その構成要素の下位の全ての構成要素のうち経路上に位置しない構成要素の内容から要素ベクトルを生成し、生成した要素ベクトルを合成して内容ベクトルを生成する。ベクトル合成手段は、基本ベクトルの生成元となった構成要素から文書の木構造上の根に至る経路上に位置する全ての構成要素に応じて生成された基本ベクトルと内容ベクトルとを合成し、合成ベクトルを生成する。合成ベクトル保持手段は、合成ベクトルを分類単位の構成要素と対応付けて保持する。分類手段は、合成ベクトルの類似度に基づいて、文書の構成要素を分類する。そして表示手段が、分類手段による分類の結果を表示する。

【００２８】
このように、本発明の文書検索装置では、検索対象とする文書の分類単位毎の内容と、その分類単位以外のすべての構成要素の内容とをキーワードにて示し、それらのキーワードから生成したベクトルの類似度に基づいて検索対象文書の構成要素を分類する。従って、類似度が高く、相互に関連性のある構成要素を文章全体の内容の情報も生かして絞り込み、表示させることができる。

【００２９】
また、本発明では上記課題を解決するために、ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を、構成要素毎に検索して表示する文書検索装置において、前記文書を蓄積する文書蓄積手段と、検索しようとする前記構成要素の検索単位を指定する検索単位指定手段と、前記文書蓄積手段から文書を読み込み、その論理構造を解析する論理構造解析手段と、前記検索単位指定手段にて指定された検索単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する基本ベクトル生成手段と、前記基本ベクトル生成手段にて生成された基本ベクトルのそれぞれに対して、その元になった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素それぞれに対応して、その構成要素の下位の全ての構成要素のうち当該経路上に位置しない構成要素の内容から要素ベクトルを生成し、生成した当該要素ベクトルを合成して内容ベクトルを生成する内容ベクトル生成手段と、前記基本ベクトルと、当該基本ベクトルの生成元となった構成要素から前記文書の木構造上の根に至る経路上に位置する全ての構成要素に応じて生成された前記内容ベクトルとを合成し、合成ベクトルを生成するベクトル合成手段と、前記合成ベクトルを前記検索単位の構成要素と対応付けて保持する合成ベクトル保持手段と、キーワード列または自然言語文により構成される問合せを入力する問合せ入力手段と、問合せから問合せベクトルを生成する問合せベクトル生成手段と、前記問合せベクトルと前記合成ベクトルとの類似度に基づいて、前記検索単位の構成要素を検索して、類似度の高い合成ベクトルに対応する構成要素を抽出する検索手段と、前記検索手段による検索の結果を表示する表示手段と、を有することを特徴とする文書検索装置が提供される。

【００３０】
このような構成の文書検索装置で、ある構成要素が他の構成要素を含んでいくような木構造の論理構造を持った文書群を、構成要素毎に検索して表示する場合、文書蓄積手段が文書を蓄積する。検索単位指定手段が、検索しようとする構成要素の検索単位を指定すると、論理構造解析手段が文書蓄積手段から文書を読み込み、その論理構造を解析する。基本ベクトル生成手段は、検索単位指定手段にて指定された検索単位の各構成要素の内容からキーワードを抽出し、基本ベクトルを生成する。内容ベクトル生成手段は、基本ベクトル

生成手段にて生成された基本ベクトルのそれぞれに対して、その元になった構成要素から文書の木構造上の根に至る経路上に位置する全ての構成要素それぞれに対応して、その構成要素の下位の全ての構成要素のうち経路上に位置しない構成要素の内容から要素ベクトルを生成し、生成した要素ベクトルを合成して内容ベクトルを生成する。ベクトル合成手段は、基本ベクトルと、基本ベクトルの生成元となった構成要素から文書の木構造上の根に至る経路上に位置する全ての構成要素に応じて生成された内容ベクトルとを合成する。問合せ入力手段は、キーワード列または自然言語文により構成される問合せを入力する。問合せベクトル生成手段は、問合せから問合せベクトルを生成する。検索手段は、問合せベクトルと合成ベクトルとの類似度に基づいて、検索単位の構成要素を検索して、類似度の高い合成ベクトルに対応する構成要素を抽出する。そして、表示手段が、検索手段による検索の結果を表示する。

【００３１】

このように、本発明の文書検索装置では、検索対象とする文書の検索単位毎の内容と、その検索単位以外の内容とをキーワードにて示し、それらのキーワードから生成した合成ベクトルと、問合せから生成した問合せベクトルと、の類似度に基づいて構成要素を検索する。従って、問合せと同一キーワードを含む確率の高い構成要素を、文書全体の内容の情報も生かして検索することができる。

【００３２】

【発明の実施の形態】

以下、本発明の実施の形態を図面を参照して説明する。

図１は、本発明の文書検索装置の第１の実施の形態の原理構成を示した図である。

【００３３】

文書検索装置の第１の実施の形態は、文書蓄積部１と、分類単位指定部２と、論理構造解析部３と、基本ベクトル生成部４と、見出しベクトル生成部５と、ベクトル合成部６と、合成ベクトル保持部７と、分類部８と、分類結果保持部９と、表示部１０と、再分類指示部１１と、検索文書抽出部１２と、から構成されている。また、検索文書抽出部１２は、表示要素選択部１２ａと、表示範囲選択部１２ｂと、表示文書抽出部１２ｃと、から構成されている。

【００３４】

文書蓄積部１は、検索の母集合となる文書群を蓄積している。ここではその蓄積の形態は特に限定しない。分類単位指定部２は、分類単位を「文書」にするか、「章」にするか、「節」にするか、「段落」にするか等の指定を行う。論理構造解析部３は、文書蓄積部１から文書を読み込み、読み込んだ文書の論理構造を解析する。基本ベクトル生成部４は、論理構造解析部３にて解析された文書の論理構造を、分類単位指定部２にて指定された分類単位で分割して各分類単位毎にキーワードを抽出し、このキーワードに基づいて基本ベクトルを生成する。

【００３５】

見出しベクトル生成部５は、論理構造解析部３にて解析された文書の論理構造において、基本ベクトル生成部４にて基本ベクトル生成の対象となった分類単位の構成要素よりも上位に配置されている全構成要素の見出しからキーワードを抽出し、このキーワードに基づいて見出しベクトルを生成する。なお、この見出しベクトル生成については、後に例を挙げて説明する。

【００３６】

ベクトル合成部６は、基本ベクトル生成部４にて生成された基本ベクトルと、見出しベクトル生成部５にて生成された見出しベクトルとを合成して、合成ベクトルを生成する。合成ベクトル保持部７は、基本ベクトル生成部４にて基本ベクトル生成の対象となった分類単位の構成要素と、ベクトル合成部５にて生成された合成ベクトルとを、対応付けて保持する。

【００３７】

分類部８は、合成ベクトル保持部７に保持された合成ベクトルの類似度に基づいて、検索

の母集合となる文書を１つ以上のカテゴリーに分類する。この分類は、分類単位指定部２にて指定された単位毎に行われる。分類結果保持部９は、分類部８にて行われた分類の結果を保持する。表示部１０は、分類部８にて行われた分類の結果、すなわち、１つ以上のカテゴリーと、各カテゴリーに含まれる分類単位に分割された文書とを表示する。再分類指示部１１は、分類部８にて行われた分類結果から、再分類の必要なカテゴリーを選択し、再分類を指示する。

【００３８】

検索文書抽出部１２は、分類部８にて分類されたカテゴリーから、必要なものを選択し、表示部１０に表示させる。すなわち、表示要素選択部１２ａが、表示部１０にて通常表示されるカテゴリーおよび各カテゴリーに含まれる分類単位に分割された文書から、抽出表示の対象とする要素を選択する。表示範囲選択部１２ｂが、抽出表示の範囲を選択する。そして、表示文書抽出部１２ｃが、表示範囲選択部１２ｂにて選択された抽出表示範囲の文書を、文書蓄積部１から抽出し、表示部１０に表示させる。

【００３９】

次に、このような構成の文書検索装置の第１の実施の形態における文書検索の手順を説明する。

図２は、文書検索装置の第１の実施の形態における文書検索の全体手順を示すフローチャートである。

【００４０】

本発明の文書検索装置の第１の実施の形態において文書検索を行う場合には、まず分類単位指定部２にて分類単位の指定が行われる（ステップＳ１）。次に、論理構造解析部３が文書蓄積部１から検索対象である文書を１つ読み込み（ステップＳ２）、読み込んだ文書の論理構造を解析する（ステップＳ３）。基本ベクトル生成部４は、論理構造解析部３にて解析された論理構造の分類単位の構成要素毎に、基本ベクトルを生成する（ステップＳ４）。また、見出しベクトル生成部５も基本ベクトル生成部４と同様、分類単位の構成要素毎に見出しベクトルを生成する（ステップＳ５）。

【００４１】

ベクトル合成部６は基本ベクトルと見出しベクトルとを合成して、合成ベクトルを生成する（ステップＳ６）。合成ベクトル保持部７はベクトル生成の対象となった分類単位の構成要素と合成ベクトルとを、対応付けて保持する（ステップＳ７）。ここで、文書蓄積部１に、検索対象であり、まだステップＳ２～ステップＳ７の処理の対象となっていない文書が残っているか否かが判断される（ステップＳ８）。そして、文書が残っていればステップＳ２に再度進み、文書が残っていなければステップＳ９に進む。文書が残っていないということは、検索対象である文書すべてに関して、ベクトルが生成されたということである。分類部８は、合成ベクトル保持部７に保持された合成ベクトルを読み込んで（ステップＳ９）文書を分類し、その結果を表示部１０に表示させる（ステップＳ１０）。

【００４２】

次に、このフローチャートのステップＳ１０にあたる分類および結果表示の処理について、順を追って説明する。

図３は、文書検索装置の第１の実施の形態における文書の分類および結果表示の手順を示すフローチャートである。

【００４３】

分類部８は、図２のステップＳ９にて読み込んだ各文書の分類単位の構成要素毎の合成ベクトルの類似度に基づいて、すべての検索対象文書を１つ以上のカテゴリーに分類する（ステップＳ１１）。分類結果保持部９は、ステップＳ１１にて行われた分類の結果を保持する（ステップＳ１２）。また、表示部１０は、ステップＳ１１にて行われた分類の結果を表示する（ステップＳ１３）。

【００４４】

次に、表示された分類結果から、再分類が必要か否か判断される（ステップＳ１４）。ここで、再分類が必要であると判断されると、再分類指示部１１にて再分類の必要なカテゴ

リーの選択が行われる（ステップＳ１５）。この選択は複数であってもよく、再分類が必要であるとされたカテゴリーに対しては、再度ステップＳ１１～ステップＳ１４の処理が繰り返される。

【００４５】

ステップＳ１４における判断にて再分類の必要がないと判断された場合には、抽出表示を行うか否かの判断が行われる（ステップＳ１６）。抽出表示を行う場合には、表示要素選択部１２ａにて抽出表示の対象となる要素が選択され（ステップＳ１７）、表示範囲選択部１２ｂにて抽出表示の範囲が選択される（ステップＳ１８）。そして表示文書抽出部１２ｃにて文書蓄積部１から抽出表示範囲の文書が抽出され（ステップＳ１９）、表示部１０に表示される（ステップＳ２０）。

【００４６】

この時点で、抽出表示を行わなかった場合には表示部１０に分類結果が表示される。また、抽出表示を行った場合には、表示部１０に抽出文書が表示される。そして、いずれにしても、表示部１０の表示から、この文書検索処理を終了するか否かが判断され（ステップＳ２１）、終了しない場合は、再度ステップＳ１３以降の処理へ進む。

【００４７】

ここで、本発明の文書検索装置の第１の実施の形態にて文書検索を行った場合に、どのように分類および検索が行われるのかを、図２および図３のフローチャートに沿って具体的な例を挙げて説明する。

【００４８】

図４は、図１に示した文書蓄積部１に蓄積される文書の一例を示す図である。図に示した文書１００は、「情報検索の技術動向」という表題を持つ文書である。この「文書」は、それぞれが「見出し」を持つ複数の「章」から構成されており、１章の見出しは「はじめに」、２章の見出しは「ＸＸＸ」、３章の見出しは「自然言語処理の利用」、４章の見出しは「ＺＺＺ」、・・・である。また、「章」は、それぞれが「見出し」を持つ複数の「節」から構成されている。３章１節の見出しは「ＹＹＹ」であり、３章２節の見出しは「文書構造解析」である。そして「節」は、さらに複数の「段落」から構成されている。３章２節の段落Ｐ１はキーワード（文脈、テキスト、概念、検索）で示される内容を持ち、段落Ｐ２はキーワード（文、意味、役割、検索）で示される内容を持ち、段落Ｐ３はキーワード（検索、精度、統計、情報）で示される内容を持っている。

【００４９】

分類単位指定部２にて分類単位を「段落」とし（図２のステップＳ１）、論理構造解析部３に図４に示した文書１００が検索対象として読み込まれると（図２のステップＳ２）、文書１００の論理構造が解析され、木構造として表現される（図２のステップＳ３）。

【００５０】

図５は、図４に示した文書の論理構造を木構造にて示した図である。

ここでは、文書の構成要素を矩形で示している。なお、図中左に記載されている構成要素ほど、上位に配置されていることを示す。また、この木構造に沿って、特定の構成要素からその上位および下位の任意の構成要素へ到達できる。

【００５１】

基本ベクトル生成部４は、まず文書１００を、分類単位である「段落」毎に読み込み、各「段落」の内容からキーワードを抽出する。

この抽出には、形態素解析を利用して構成要素内の文章からキーワードとなる単語を抽出する方法や、予めキーワードの候補となる単語リストを用意しておき、構成要素内の文章に含まれた単語を照合して、一致した単語をキーワードとして抽出する方法等、様々な方法が考えられる。なお、本発明の文書検索におけるキーワードの抽出には上記のような従来技術を利用するものとし、特に方法の限定はしない。

【００５２】

次に基本ベクトル生成部４は、次式（１）で示される計算を行って、抽出したキーワードｋ個を基に、各「段落」に対するｔ次元の基本ベクトルを生成する（図２のステップＳ４

）。ここで次元数ｔは、分類に利用するキーワードの総数である。この次元数ｔの算出方法はキーワードの抽出方法によって異なり、キーワードの抽出に単語リストを使用する場合には、その単語リストに含まれる単語の総数がｔとなる。

【００５３】

ある構成要素P_iの基本ベクトルA_{pi}は次式（１）で表現される。

【００５４】

【数１】

$$A_{pi}=(p_{i1},p_{i2},\cdots,p_{it}) \quad \cdots\cdots(1)$$

ここでp_{ik}は、構成要素P_iにおけるキーワードｋの値を表わす。なお本実施の形態では、構成要素P_i中に単語ｋが出現していれば$p_{ik}=1$、出現していなければ$p_{ik}=0$とする。このキーワードｋの値には、単語ｋの出現頻度や分散の度合に応じて重み付けを行っても良い。

【００５５】

ここで、文書１００に則し具体的に基本ベクトルの例を挙げる。単語リストとして「情報、検索、技術、動向、自然、言語、処理、利用、文書、構造、解析、役割、文、意味」を利用すると、キーワード（文脈、テキスト、概念、検索）を含む段落Ｐ１の基本ベクトルA_{P1}は、次式（２）で表現される。

【００５６】

【数２】

$$A_{P1}=(0,1,0,0,0,0,0,0,0,0,0,0,0,0) \quad \cdots\cdots(2)$$

同様に、キーワード（文、意味、役割、検索）を含む段落Ｐ２の基本ベクトルA_{P2}は、次式（３）で表現される。

【００５７】

【数３】

$$A_{P2}=(0,1,0,0,0,0,0,0,0,0,0,1,1,1) \quad \cdots\cdots(3)$$

また、キーワード（検索、精度、統計、情報）を含む段落Ｐ３の基本ベクトルA_{P3}は、次式（４）で、表現される。

【００５８】

【数４】

$$AP3=(1,1,0,0,0,0,0,0,0,0,0,0,0,0) \quad \cdots\cdots(4)$$

見出しベクトル生成部５は、論理構造解析部３において抽出された全論理構成要素のうち、各「段落」から文書の木構造の根に至る経路上にある全ての構成要素の「見出し」（「文書」の「表題」、「章」の「見出し」、「節」の「見出し」）を読み込む。そして、各「見出し」に関するｔ次元の見出しベクトルを生成する（図２のステップＳ５）。ここで例えば、「文書の表題」から得られる見出しベクトルＢは、次式（５）で表現される。

【００５９】

【数５】

$$B=(B_1,B_2,\cdots,B_k) \quad \cdots\cdots(5)$$

ここでB_kは、「文書の表題」におけるキーワードｋの値を表わす。また、「章の見出し」から得られる見出しベクトルＣは、次式（６）で表現される。

【００６０】

【数６】

$$C=(C_1,C_2,\cdots,C_k) \quad \cdots\cdots(6)$$

ここでC_kは、「章の見出し」におけるキーワードｋの値を表わす。また同様に、「節の見出し」から得られる見出しベクトルＤは、次式（７）で表現される。

【００６１】

【数７】

$$D=(D_1,D_2,\cdots,D_k) \quad \cdots(7)$$

ここでD_kは、「節の見出し」におけるキーワードｋの値を表わす。ここでは、論理構成要素として「文書」、「章」、「節」、「段落」のみを挙げたが、これ以外に構成要素を持つような文書でも、式（５）～式（７）と同様に見出しベクトルを生成する。なお、見出しベクトルを求める式（５）～式（７）は、本質的には式（１）と同一のものである。

【００６２】

ここで，文書１００に則し具体的に見出しベクトルの例を挙げると、キーワード（情報、検索、技術、動向）を含む文書の見出しベクトルB_{100}は、次式（８）で表現される。

【００６３】

【数８】

$$B_{100} = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \quad \cdots\cdots (8)$$

同様に、キーワード（自然、言語、処理、利用）を含む３章の見出しベクトルC_{100}は次式（９）で表現される。

【００６４】

【数９】

$$C_{100} = (0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0) \quad \cdots\cdots (9)$$

またキーワード（文書、構造、解析）を含む３章２節の見出しベクトルD_{100}は次式（１０）で、表現される。

【００６５】

【数１０】

$$D_{100} = (0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0) \quad \cdots\cdots (10)$$

このようにして、分類単位の構成要素に対する基本ベクトルおよび見出しベクトルが生成されると、ベクトル合成部６はこれらのベクトルを合成し、合成ベクトルを生成する（図２のステップＳ６）。ここで、ある章の下のある節の下のある段落の合成ベクトルＳは、次式（１１）で表わされる。

【００６６】

【数１１】

$$S = A + w1\ D + w2\ C + w3\ B \quad \cdots\cdots (11)$$

ここで、式（１１）において、Ａは基本ベクトル、ＤはＡを生成した構成要素を包含する節の見出しベクトル、ＣはＤを生成した節を包含する章の見出しベクトル、Ｂは文書の表題ベクトルを表わす。また、ｗ１ 、ｗ２ 、ｗ３ は、それぞれのベクトルに対する重み付けを表わす。

【００６７】

重み付けには、０より大きく１以下の任意の値をとることができる。一般に、文書が論理的な階層構造を持つとき、上位の階層は下位の階層よりも広い概念を表す。従って、本実施の形態では、重み付けw_rを次式（１２）にて算出する。

【００６８】

【数１２】

$$w_r = 1/(r+1) \quad \cdots\cdots (12)$$

この式（１２）において、ｒは分類単位「段落」からの距離を表わす。すなわち、「節」、「章」、「文書」の順に重み付けが小さくなり、節見出しの重みは１／２、章見出しの重みは１／３、表題の重みは１／４になる。このように下位概念に重きを置いた分類は、検索者の要求が具体的である場合に適している。

【００６９】

図６は、図４に示した文書１００の段落Ｐ２に対する合成ベクトルの生成を示す表である。

図において、見出しベクトルの欄に記載されている値は、既に重み付けされた値である。ここでは、節見出しへの重み付けを０．６、章見出しへの重み付けを０．４、表題への重み付けを０．２としている。ベクトル合成部６にて合成される段落Ｐ２に対する合成ベクトルV_{P2}は、次式（１３）にて表わされる。

【００７０】

【数１３】

$$V_{P2} = (0.2,\ 1.2,\ 0.2,\ 0.2,\ 0.4,\ 0.4,\ 0.4,\ 0.4,\ 0.6,\ 0.6,\ 0.6,\ 0.6,\ 1.0,\ 1.0,\ 1.0) \quad \cdots\cdots\cdots (13)$$

【００７１】

このようにして分類単位の構成要素に対する合成ベクトルが生成されると、合成ベクトル保持部７は、分類単位の構成要素と合成ベクトルとを対応付けて保持する（図２のステップＳ７）。

【００７２】

図７は、図４に示した文書１００に関して、分類単位「段落」に対応する合成ベクトル保持の様子を示した図である。なお、ここでは先に求めた段落Ｐ２に対する合成ベクトルについてのみ具体例を示しておく。

【００７３】

この時点で文書蓄積部１に検索の対象とすべき文書がまだ残っていれば、以上と同様の処理を再度行う（図２ステップＳ８の分岐にてステップＳ２へ進む）が、ここでは図４に示した文書１００のみを検索の対象として説明を続ける。

【００７４】

検索対象である文書すべての分類単位の構成要素に対する合成ベクトルの生成、保持が終了すると、分類部８は合成ベクトル保持部７に保持された合成ベクトルを、すべて読み込む（図２のステップＳ９）。そして分類部８は、読み込んだ合成ベクトルの類似度に基づいて、検索対象である文書すべてを、分類単位毎に、１つ以上のカテゴリーに分類する（図３のステップＳ１１）。

【００７５】

ここで、ベクトルの類似度の計算方法としては、様々なものが知られている。ベクトルの内積を利用する方法や、コサイン係数に基づく方法等、本実施の形態では、これらの従来技術を任意に適用すればよい。また分類の方法としては、多変量解析における種々のクラスタリング手法が利用できる。

【００７６】

分類の結果は、分類結果保持部９に保持され（図３のステップＳ１２）、表示部１０に表示される（図３のステップＳ１３）。

ここで、図４に示した文書１００に則して、分類結果を表示する表示部１０の表示画面の一例を挙げてみる。

【００７７】

図８は、分類結果の表示画面の一例を示す図である。

表示画面２００には、分類結果であるカテゴリー群の表示を行うカテゴリー表示ウィンドウ２１０、再分類を指示する分類ボタン２０１、検索文書の抽出表示を行う際に表示範囲の選択を行う範囲選択ボタン２０２、および検索文書の抽出表示を指示する文書取り出しボタン２０３、が設けられている。

【００７８】

ここで、カテゴリー表示ウィンドウ２１０には、分類の結果であるカテゴリー２１１、２１２、・・・が表示されている。各カテゴリーには、再分類を行う際カテゴリーの選択に使用する選択用チェックボックス２１１ａ、２１２ａ、・・・が設けられている。また、各カテゴリーには、そのカテゴリーに分類された分類単位の構成要素を表示する要素表示

欄２１１ｂ、２１１ｃ、・・・、２１２ｂ、２１２ｃ、・・・、・・・が設けられている。なお、カテゴリー表示ウィンドウ２１０およびカテゴリー２１１、２１２、・・・には、必要に応じてスクロールバーが設けられる。

【００７９】

要素表示欄２１１ｂ、２１２ｂ、・・・に分類単位の構成要素を表示する方法には様々な方法が考えられる。各分類単位の構成要素の先頭から一定長の文字列を表示させる方法や、各分類単位の構成要素から抽出されたキーワード群を表示させる方法等、分類単位の構成要素を一意に識別する方法であれば、どのような方法でもよい。なお、図に示した表示画面２００では、各分類単位の構成要素の先頭から一定長の文字列を表示させる方法を示している。また、表示画面２００の範囲選択ボタン２０２は「文書」、「章」、「節」、「段落」から範囲を選択できるように構成されているが、さらに複雑な論理構成を持つ文書を検索対象とする場合には、その構成に基づいて範囲を選択できるように構成される。

【００８０】

このように表示された分類の結果から、カテゴリーの再分類が必要か否か判断される（図３のステップＳ１４）。分類が必要と判断された場合には、分類結果として得られたカテゴリー群から再分類の対象とすべき１つ以上のカテゴリーが選択される（図３のステップＳ１５）。

【００８１】

ここで、図８に示した表示画面にて再分類を指示する方法を説明する。

図９は、図８に示した表示画面において、再分類を指示する様子を示した図である。

【００８２】

カテゴリーの選択は、選択用チェックボックス２１１ａ、２１２ａ、・・・をマウス等のポインティングデバイスで選択することによって行われる。図は、選択用チェックボックス２１１ａが選択された様子を示しており、選択用チェックボックス２１１ａには選択されたことを示すチェックが入っている。この状態で、再分類を指示する分類ボタン２０１をマウス等のポインティングデバイスにて選択すると、カテゴリー２１１を対象に再度分類部８による分類が行われて、更に１つ以上のカテゴリーに分類される。この分類の結果は分類結果保持部９に保持され、表示部１０に表示される。

【００８３】

本実施の形態の文書検索装置では、表示された分類結果に再分類が必要でない時や、再分類が必要か否かを表示画面からだけでは判断できない時には、分類単位の構成要素を抽出して表示させることができる（図３のステップＳ１６）。抽出表示を行う場合には、まず抽出表示をさせたい分類単位の構成要素を表示要素として選択し（図３のステップＳ１７）、次に抽出表示をさせたい範囲を選択する（図３のステップＳ１８）。

【００８４】

図１０は、図８に示した表示画面において、抽出表示を指示する様子を示した図である。表示要素および表示範囲の選択は、マウス等のポインティングデバイスにて行う。図は、表示要素として要素表示欄２１１ｂが選択され、表示範囲として範囲選択ボタン２０２の「節」が選択され、選択された２箇所が反転等により確認できるようになっている様子を示している。つまり、表示画面２００のカテゴリー表示ウィンドウ２１０に示された、要素表示欄２１１ｂ、２１１ｃ、・・・、２１２ｂ、２１２ｃ、・・・、・・・は、図１に示した表示要素選択部１２ａに対応している。同様に、表示画面２００の範囲選択ボタン２０２は、図１に示した表示範囲選択部１２ｂに対応している。また、取り出しボタン２０３は、図１に示した表示文書抽出部１２ｃに対応している。従って、この状態で取り出しボタン２０３がマウス等のポインティングデバイスにて選択されると、文書蓄積部１に蓄積された文書から、選択範囲および選択要素が抽出される（図３のステップＳ１９）。そして抽出された文書は、表示部１０に表示される（図３のステップＳ２０）。

【００８５】

図１１は、抽出文書の表示画面の一例である。

表示画面２２０には、選択された表示範囲の構成要素名を表示する範囲名表示欄２２１と

、選択され、抽出された表示範囲の文書を表示する範囲内容表示欄２２２と、選択要素を拡大表示する抽出要素表示欄２２３とが設けられている。　ここでは、図１０に示した抽出表示の指示通りに、図４に示した文書１００の、３章２節の段落Ｐ２を抽出要素として、また、この段落Ｐ２を含む節を抽出範囲として、表示している。なお、表示画面２２０には、必要に応じてスクロールバーが設けられる。

【００８６】

文書の抽出表示をしてみて、さらに分類が必要であると判断されれば、処理は再び分類部８による文書の分類へ進み、その必要がなければ、この文書検索処理は終了である（図３のステップＳ２１）。

【００８７】

このように、本発明の文書検索装置の第１の実施の形態では、蓄積された文書を任意の分類単位で分割し、その際、分類単位の上位に位置する構成要素の見出しをも考慮に入れた上で、類似度の高いもの同士で分類し、カテゴリーを生成して必要な話題を含む分類単位の構成要素のみを絞り込む。その後、必要に応じて詳細な内容を表示させることができるため、多量の文書や巨大な文書から、必要な話題のみを簡単に抽出することが可能である。図１１では抽出要素を拡大表示させたが、単に矩形で囲んだり、下線を引いたり、といった方法で表示範囲との区別がつくようにしてもよい。

【００８８】

次に、本発明の文書検索装置の第２の実施の形態について説明する。

第２の実施の形態では、第１の実施の形態で示した文書検索装置にて階層的な論理構造の明示されていない文書を対象とした文書検索を行う。この実施の形態の装置構成は図１に示した第１の実施の形態の装置構成と同一である。また、文書検索処理は図２及び図３に示したフローチャートと同様の処理手順にて実現される。

【００８９】

具体的な文書群として、例えば、WWW（Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂ）で公開されているＨＴＭＬ（Ｈｙｐｅｒ　Ｔｅｘｔ　Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ）文書を文書蓄積部１に格納して、分類および検索の対象とすることができる。ＨＴＭＬは、構造化文書の国際規格であるＳＧＭＬ（Ｓｔａｎｄａｒｄ　Ｇｅｎｅｒａｌｉｚｅｄ　Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ；　ＩＳＯ８８７９）　の応用のひとつであり、文書タイプ定義ＤＴＤ（Ｄｏｃｕｍｅｎｔ　Ｔｙｐｅ　Ｄｅｆｉｎｉｔｉｏｎ）によって、文書構造が規定されている。

【００９０】

ＨＴＭＬのＤＴＤは階層的な論理構造を明示的に規定できるものではないが、本実施の形態では見出しの大きさを疑似的に階層の深さとみなすことにより、階層的な論理構造を持つものとして解析する。また、文書中に出現する章番号、節番号により、階層的な構造を取り出すことができる。

【００９１】

ここで、本発明の第２の実施の形態の文書検索装置にて、ＨＴＭＬ文書を分類する場合の処理について具体例をあげて説明する。なお、説明に必要な文書検索装置の構成要素及び文書検索処理のステップ等は、第１の実施の形態の説明に使用した図１～図３の符号を流用する。

【００９２】

図１２は、図１に示した文書蓄積部１に蓄積されるＨＴＭＬ文書の一例を示す図である。図に示した文書１１０において、文書中の各要素は、その開始を示すタグと終了を示すタグによって囲まれている。ある要素Ａについて、開始タグは＜Ａ＞　、終了タグは＜／Ａ＞で示される。従って、文書１１０はまず、文書の開始を示すタグ＜ＨＴＭＬ＞と、文書の終了を示すタグ＜／ＨＴＭＬ＞によって囲まれている。

【００９３】

また、文書１１０はヘッダ要素（ＨＥＡＤ）と本体要素（ＢＯＤＹ）からなり、ヘッダ要素は表題要素（ＴＩＴＬＥ）を包含している。そして、本体要素は、見出し要素（Ｈ２お

よびＨ３）と段落要素（Ｐ）の並びを包含している。なお、ＨＴＭＬのＤＴＤでは見出し要素としてＨ１～Ｈ６の６種類が規定されており、Ｈ１が最も重要度の高い見出しとして、Ｈ６が最も重要度の低い見出しとして規定される。

【００９４】

一般的に階層的な見出しを構成する場合、重要度の高い見出しほど上位階層の見出しとして記述される。従って、ここでは、要素Ｈ２を章見出し、要素Ｈ３を節見出しとして論理構造を解析する。また、見出しとともに記述される章番号や節番号から論理構造を解析する。

【００９５】

分類単位指定部２にて分類単位を「段落」とし（　図２のステップＳ１）、論理構造解析部３に図１２に示した文書１１０が読み込まれると（　図２のステップＳ２）、文書の論理構造が解析され、表題要素（ＴＩＴＬＥ）、見出し要素（Ｈ２およびＨ３）、段落要素（Ｐ）が取り出され、木構造として表現される（　図２のステップＳ３）。

【００９６】

図１３は、図１２に示した文書の論理構造を木構造にて示した図である。

基本ベクトル生成部４は、まず文書１１０を、分類単位である「段落」毎に読み込み、各「段落」内容からキーワードを抽出する。ここでキーワードの抽出方法は従来技術を利用するものとし、特に一つの方法には限定しない。

【００９７】

次に基本ベクトル生成部４は、前述の式（１）で示される計算を行って、各段落に対するｔ次元の基本ベクトルを生成する（　図２のステップＳ４）。

ここで、文書１１０のように文書の階層的な論理構造が明示されていない場合、式（１）におけるp_{ik}の値に単語ｋの出現頻度及び分散の度合に応じて重み付けを行い、基本ベクトルのノルムが１になるように正規化を行う必要がある。基本ベクトルのノルムが１になるように正規化することにより、「段落」と「見出し」のテキスト量の違いによる影響を排除することができる。

【００９８】

例えば、段落数の合計がＮ個である文書群を対象とした場合、ｉ番目（ｉ＝１，・・・，Ｎ）の段落のベクトル$A_{Pi}=(p_{i1}, p_{i2}, \cdots, p_{it})$のｊ番目（ｊ＝１，・・・，ｔ）の要素の値$P_{ij}$は、出現頻度（ＴＦ）と分散（ＩＤＦ）を用いて、次式（１４）、（１５）にて表される。

【００９９】

【数１４】

$$P_{ij}=f_{ij}\cdot g_j \qquad \cdots\cdots(14)$$

【０１００】

【数１５】

$$g_j=\mathrm{Log}(N/d_j) \qquad \cdots\cdots(15)$$

ただし、f_{ij}は、ｊ番目の単語のｉ番目の段落における出現頻度でＴＦに対応する。またg_jはＩＤＦに対応し、d_jはｊ番目の単語が出現する段落の数を表す。ここで、A_{Pi}をノルムが１になるように正規化したA'_{Pi}は、次式（１６）により計算される。

【０１０１】

【数１６】

$$A'_{Pi}=\frac{A_{Pi}}{\sqrt{\sum_{j=1}^{t}P_{ij}^{2}}} \qquad \cdots\cdots(16)$$

【０１０２】
本実施の形態では、基本ベクトルとしてこのＡ’$_{Pi}$を用いる。
ここで例えば、図１２に示す文書１１０の２章３節の段落Ｐ１８を対象に基本ベクトルを生成する場合、段落Ｐ１８からはキーワードとして（ＮＩＳＴ、現在、コンピュータ、セキュリティ、ハンドブック、作成、．．．）が抽出される。そして、各キーワードの出現頻度（ＴＦ）と分散（ＩＤＦ）が計算され、ノルムが１になるように正規化された基本ベクトルＡ’$_{P18}$が生成される。
【０１０３】
次に、見出しベクトル生成部５は、論理構造解析部３において抽出された全論理構成要素のうち、各「段落」より上位にあるすべての「見出し」（文書の「表題」、「章見出し」「節見出し」）を読み込む。そして、各「見出し」に関するｔ次元の見出しベクトルを生成する（図２のステップＳ５）。なお、ここでも、各見出しベクトルにも基本ベクトル同様、単語の出現頻度と分散を用いて重み付けを行い、見出しベクトルのノルムが１になるように正規化を行う。
【０１０４】
図１２に示す文書１１０の段落Ｐ１８に関しては、上位の見出しとして文書の表題「Ｎａｔｉｏｎａｌ　Ｉｎｆｏｒｍａｔｉｏｎ　Ｉｎｆｒａｓｔｒｕｃｔｕｒｅ　（ＮＩＩ）」、２章の見出しである「ネットワークセキュリティのための方策」、２章３節の見出しである「ネットワークセキュリティ管理」からそれぞれキーワードが抽出され、各キーワードの出現頻度と分散が計算され、ノルムが１になるように正規化された見出しベクトルが生成される。
【０１０５】
ベクトル合成部６は、図２のステップＳ４及びステップＳ５で生成された基本ベクトルと見出しベクトルとを合成する（図２のステップＳ６）。なお、合成ベクトルの算出に関しては前述の式（１１）を、式（１１）における各ベクトルに対する重み付けに関しては、前述の式（１２）を、それぞれ流用する。
【０１０６】
分類単位の構成要素に対する合成ベクトルが生成されると、合成ベクトル保持部７は、分類単位の構成要素と合成ベクトルを対応付けて保持する（図２のステップＳ７）。この時点で文書蓄積部１に分類の対象とすべき文書がまだ残っていれば、以上と同様の処理を行う（図２のステップＳ８の分岐にてステップＳ２へ進む）。分類対象である文書すべての分類単位の構成要素に対する合成ベクトルの生成、保持が終了すると、分類部８は合成ベクトル保持部７に保持された合成ベクトルを、すべて読み込む（図２のステップＳ９）。そして分類部８は、読み込んだ合成ベクトルの類似度に基づいて、分類対象である文書すべてを、分類単位毎に、１つ以上のカテゴリーに分類し（図３のステップＳ１１）、分類結果保持部９に分類結果を保持する（図３のステップＳ１２）。
【０１０７】
なお、ここではベクトルの類似度の計算方法としてコサイン係数に基づく方法を用いる。コサイン係数に基づくベクトルの類似度の計算方法については、後述する式（２０）にて述べる。また、分類の方法として、公知技術であるｋ平均クラスタリングを用いる。
【０１０８】
ここで、本実施の形態にて実際に文書を分類した実験の結果を説明する。
対象とする文書群として、ＷＷＷで公開されているＨＴＭＬ文書３３件（社団法人日本電子工業振興協会が発行している「電子工業月報」の「ニューヨーク駐在員報告」９４年６月号から９７年２月号）を利用、１６３５段落に分割して、文書の分類を行った。
【０１０９】
図１４は、実験対象とする文書群を人手により分類した結果を示す分類表である。
実験の対象とした３３件の文書はすべて米国の情報産業の動向に関するもので、図１４に示した分類表の各カテゴリには、そのカテゴリに含まれる段落群の内容から判断して適切なトピックが付与されている。なお、カテゴリは全部で２７個あり、各カテゴリは最少で

８個、最多で１６４個の段落を含んでいる。また、これ以降この分類表の示す分類結果を、標準セットと呼ぶ。

【０１１０】

実験では、上述のＨＴＭＬ文書３３件を図１に示した文書蓄積部１に蓄積し、図２及び図３のフローチャートに示す処理手順に従って処理した結果を標準セットと比較した。

【０１１１】

ここで、ｋ平均クラスタリングについて説明する。なお、これ以降、「クラスタ」は「カテゴリー」と同義であるとする。

ｋ平均クラスタリングでは、生成するクラスタの数ｋをパラメータとして与える。ｋ個のクラスタ中心の初期値は、標本ベクトルの中から適当に選ばれる。

【０１１２】

ここでは人手による分類と比較するため、クラスタ数ｋを図１４に示した標準セットのカテゴリー数と同数、すなわちｋ＝２７とする。また、ｋ個のクラスタ中心の初期値は、図１４に示した標準セットの各カテゴリーのトピックに対して、最も類似度の高い段落を選ぶ。

【０１１３】

各カテゴリーのトピックと、対象とする「段落」との類似度は、トピックから生成したベクトルと、各「段落」の基本ベクトルとの類似度計算、例えばコサイン係数を用いる方法等によって計算できる。

【０１１４】

これにより、図２および図３のフローチャートに示す処理手順に沿って処理した結果のクラスタ数は、標準セットのカテゴリー数と揃えることができる。また、標準セットの各カテゴリーのトピックに対応するクラスタが生成されることが期待できる。

【０１１５】

なお、ここでは標準セットの各カテゴリーのトピックに対して、最も類似度の高い「段落」の基本ベクトルをクラスタ中心の初期値として選択したが、これは人手による分類との比較を目的としたものであり、実用上は、クラスタ中心の初期値をどのように決めても良い。

【０１１６】

標準セットのカテゴリー集合を｛C_1，C_2，・・・，C_m｝、本実施の形態における分類の結果のクラスタ集合を｛C'_1，C'_2，・・・，C'_m｝とし、クラスタＡの要素数をｎ（Ａ）、クラスタＡとクラスタＢに共通する要素の数をｎ（Ａ∩Ｂ）で表す。また、対象となる構成要素の総数をＮとする。

【０１１７】

このとき、標準セットに対する正解率Ｓは次式（１７）により求められる。

【０１１８】

【数１７】

$$S = \frac{\sum_{i=1}^{m} n(C_i \cap C_i')}{N} \qquad \cdots\cdots\cdots(17)$$

【０１１９】

上記の式（１７）から、この実験結果である正解率Ｓを算出すると、７０．６％となった。これに対し、図２のステップＳ５およびステップＳ６を省略した場合、つまり基本ベクトルのみを合成ベクトル保持部７に保持するようにした場合、分類結果の正解率Ｓは、５５．６％となる。すなわち、基本ベクトルに見出しベクトルを合成して分類を行うことにより、正解率Ｓは１５．０％向上した。

【０１２０】

以上説明したように、本発明の文書検索装置の第２の実施の形態では、蓄積された文書を任意の分類単位で分割し、その際、分類単位の構成要素の上位に位置する見出しを、そのテキストの長さを含めて考慮に入れた上で、類似度の高いもの同士で分類し、カテゴリーを生成して必要な話題を含む分類単位の構成要素のみを絞り込む。　その後、必要に応じて詳細な内容を表示させることができるため、多量の文書や巨大な文書から、必要な話題のみを簡単に抽出することが可能である。

【０１２１】

次に、本発明の文書検索装置の第３の実施の形態を説明する。

図１５は、本発明の文書検索装置の第３の実施の形態の原理構成を示した図である。

文書検索装置の第３の実施の形態は、文書蓄積部２１と、検索単位指定部２２と、論理構造解析部２３と、基本ベクトル生成部２４と、見出しベクトル生成部２５と、ベクトル合成部２６と、合成ベクトル保持部２７と、問合せ入力部２８と、問合せベクトル生成部２９と、検索部３０と、表示部３１と、検索文書抽出部３２と、から構成されている。

【０１２２】

文書蓄積部２１、論理構造解析部２３およびベクトル合成部２６は、図１に示した第１の実施の形態の文書蓄積部１、論理構造解析部３及びベクトル合成部６と同一の構成要素であるので、説明は省略する。

【０１２３】

また、表示部３１および検索文書抽出部３２も、図１に示した第１の実施の形態の表示部１０および検索文書抽出部１２と同一の構成要素であるので、説明を省略する。検索文書抽出部３２を構成する表示要素選択部３２ａ、表示範囲選択部３２ｂおよび表示文書抽出部３２ｃは、図１に示した検索文書抽出部１２を構成する表示要素選択部１２ａ、表示範囲選択部１２ｂおよび表示文書抽出部３２ｃと同一の構成要素である。

【０１２４】

検索単位指定部２２は、検索単位を「文書」にするか、「章」にするか、「節」にするか、「段落」にするか等の指定を行う。基本ベクトル生成部２４は、論理構造解析部２３にて解析された文書の論理構造を、検索単位指定部２２にて指定された検索単位で分割して検索単位の各構成要素毎にキーワードを抽出し、このキーワードに基づいて基本ベクトルを生成する。

【０１２５】

見出しベクトル生成部２５は、論理構造解析部２３にて解析された文書の論理構造において、基本ベクトル生成部２４にて基本ベクトル生成の対象となった検索単位の構成要素から文書の木構造上の根に至る経路上に位置するすべての構成要素の見出しからキーワードを抽出し、このキーワードに基づいて見出しベクトルを生成する。合成ベクトル保持部２７は、基本ベクトル生成部２４にて基本ベクトル生成の対象となった検索単位の構成要素と、ベクトル合成部２５にて生成された合成ベクトルとを、対応付けて保持する。

【０１２６】

問合せ入力部２８は、キーワード列または自然言語文により構成される検索の問合せの入力を受け付ける。問合せベクトル生成部２９は、問合せ入力部２８にて受け付けた問合せから、問合せベクトルを生成する。そして、検索部３０は、問合せベクトル生成部２９にて生成された問合せベクトルと、合成ベクトル保持部２７にて保持されている合成ベクトルとを読み込む。そして、問合せベクトルと合成ベクトルとの類似度を算出し、類似度の高い合成ベクトルと対応する検索単位の構成要素を検索する。

【０１２７】

次に、このような構成の文書検索装置の第３の実施の形態における文書検索の手順を説明する。

図１６は、文書検索装置の第３の実施の形態における文書検索の準備を行う手順を示すフローチャートである。

【０１２８】

本発明の文書検索装置の第３の実施の形態において文書検索を行う場合には、まず検索単

位指定部２２にて検索単位の指定が行われる（ステップＳ３１）。次に、論理構造解析部２３が文書蓄積部２１から検索対象である文書を１つ読み込み（ステップＳ３２）、読み込んだ文書の論理構造を解析する（ステップＳ３３）。基本ベクトル生成部２４は、論理構造解析部２３にて解析された論理構造の検索単位の構成要素毎に、基本ベクトルを生成する（ステップＳ３４）。また、見出しベクトル生成部２５も基本ベクトル生成部２４と同様、検索単位の構成要素毎に見出しベクトルを生成する（ステップＳ３５）。

【０１２９】

ベクトル合成部２６は、基本ベクトルと見出しベクトルとを合成して、合成ベクトルを生成する（ステップＳ３６）。合成ベクトル保持部２７は、ベクトル生成の対象となった検索単位の構成要素と合成ベクトルとを、対応付けて保持する（ステップＳ３７）。ここで、文書蓄積部２１に、検索対象であり、まだステップＳ３２～ステップＳ３７の処理の対象となっていない文書が残っているか否かが判断される（ステップＳ３８）。そして、文書が残っていればステップＳ３２に再度進み、文書が残っていなければこのフローチャートの処理を終了する。ステップＳ３８の判断において文書が残っていないということは、検索対象である文書すべてに関して、ベクトルが生成されたということである。そして、図１６に示した文書検索の準備が終了した後に実際に文書検索を行う。

【０１３０】

図１７は、文書検索装置の第３の実施の形態において、文書検索準備終了後に行う文書検索処理の手順を示したフローチャートである。

本発明の文書検索装置の第３の実施の形態において文書検索を行う場合、問合せ入力部２８にて、問合せを入力する（ステップＳ４１）。問合せベクトル生成部２９は、入力された問合せがキーワード列であればそこから、自然言語文であればその内容からキーワードを抽出して、問合せベクトルを生成する（ステップＳ４２）。

【０１３１】

問合せベクトルが生成されると、検索部３０は、問合せベクトル生成部２９から問合せベクトルを、合成ベクトル保持部２７から検索対象である文書に対する合成ベクトルすべてを、読み込む（ステップＳ４３）。そして問合せベクトルと合成ベクトル群との比較を行い（ステップＳ４４）、問合せベクトルとの類似度の高い合成ベクトルから順に、対応する検索単位の構成要素とともに表示部３１に表示させる（ステップＳ４５）。

【０１３２】

ここまでの処理で、文書検索自体は一応終了であるが、ここで、文書検索の結果を基に、検索した文書の抽出表示を行うか否かの判断が行われる（ステップＳ４６）。抽出表示を行う場合には、表示要素選択部３２ａにて抽出表示の対象となる構成要素が選択され（ステップＳ４７）、表示範囲選択部３２ｂにて抽出表示の範囲が選択される（ステップＳ４８）。そして表示文書抽出部３２ｃにて文書蓄積部２１から抽出表示範囲の文書が抽出され（ステップＳ４９）、表示部３１に表示される（ステップＳ５０）。

【０１３３】

この時点で、抽出表示を行わなかった場合、表示部３１には文書検索の結果が表示されている。また、抽出表示を行った場合、表示部３１には抽出文書が表示されている。そして、いずれにしても、表示部３１の表示から、この文書検索処理を終了するか否かが判断され（ステップＳ５１）、終了しない場合には、再度ステップＳ４６以降の処理へ進み、別の構成要素に対する抽出表示を行うことが可能である。

【０１３４】

ここで、本実施の形態の文書検索装置にて文書検索を行う場合に、問合せを入力したり、検索結果を表示したりする入出力画面の一例を挙げておく。

図１８は、本発明の文書検索装置の第３の実施の形態に係る入出力画面の一例を示した図である。

【０１３５】

入出力画面２３０には、問合せの入力を受け付ける問合せ入力欄２３１、検索開始を指示する検索ボタン２３２、検索文書の抽出表示を行う際に表示範囲の選択を行う範囲選択ボ

タン２３３、検索文書の抽出表示を指示する表示ボタン２３４、および検索結果を表示する検索結果表示欄２３５が設けられている。

【０１３６】

ここで、問合せ入力欄２３１は問合せ入力部２８に対応している。また、検索ボタン２３２は検索部３０に、範囲選択ボタン２３３は表示範囲選択部３２ｂに、表示ボタン２３４は表示文書抽出部３２ｃに、検索結果表示欄２３５は表示部３１に、それぞれ対応している。

【０１３７】

次に、本発明の文書検索装置の第３の実施の形態で文書検索を行った場合に、どのように処理が行われるのかを、図１６のフローチャートに沿って、具体的な例を挙げて説明する。

【０１３８】

図１９は、文書蓄積部２１に蓄積される文書の一例を示す図である。

文書１２０は、「情報検索の技術動向」という表題を持つ文書である。文書全体は複数の「章」から構成されており、１章の見出しは「背景」、２章の見出しは「ベクトル空間モデル」、３章の見出しは「自然言語処理の利用」、４章の見出しは「ユーザインタフェース」、・・・である。また、章の中には複数の「節」から構成されているものがある。３章１節の見出しは「シソーラス」であり、３章２節の見出しは「文書構造解析」である。章や節はさらに複数の「段落」から構成されている。１章の段落Ｐ２１はキーワード（インターネット、情報、アクセス）で示される内容を持ち、段落Ｐ２２はキーワード（マルチメディア、テキスト、検索）で示される内容を持っている。また、２章の段落Ｐ２３はキーワード（部分、照合、検索）で示される内容を持ち、段落Ｐ２４はキーワード（単語、重み、ベクトル）で示される内容を持ち、段落Ｐ２５はキーワード（頻度、分散、類似度）で示される内容を持っている。更に、３章１節の段落Ｐ２６はキーワード（語彙、辞書、検索）で示される内容を持ち、３章２節の段落Ｐ２７はキーワード（文脈、テキスト、概念、検索）で示される内容を持ち、段落Ｐ２８はキーワード（文、意味、役割、検索）で示される内容を持ち、段落Ｐ２９はキーワード（検索、精度、統計、情報）で示される内容を持っている。そして、４章の段落Ｐ３０はキーワード（利用者、インタラクション、ブラウジング）で示される内容を持っている。

【０１３９】

このような文書１２０を含んだ検索対象に文書検索を行う際には、まず図１６に示したフローチャートの処理に沿って、検索単位の全ての構成要素毎に合成ベクトルを生成する。

【０１４０】

ここで検索単位として「段落」を指定するとする。その後、図１８に示した入出力画面２３０の問合せ入力欄２３１に、「問合せ」として、例えば「自然言語処理を利用した情報検索」を入力し、検索ボタン２３２をマウスなどのポインティングデバイスにて選択する（図１７のステップＳ４１）。

【０１４１】

すると、問合せベクトル生成部２９が、入力された問合せからキーワードを抽出して問合せベクトルを生成する（図１７のステップＳ４２）。検索部３０は、作成された問合せベクトルを問合せベクトル生成部２９から、保持されている合成ベクトル群を合成ベクトル保持部２７から読み込む（図１７のステップＳ４３）。それから、読み込んだ問合せベクトルと合成ベクトルとを比較する（図１７のステップＳ４４）。

【０１４２】

図２０は、図１９に示した文書１２０の段落毎の合成ベクトルと問合せベクトルとの比較の様子を示した表である。

本実施の形態では、問合せベクトルと合成ベクトル群との類似度は、一致しているキーワードの数値の合計から算出され、表１２１が生成される。表１２１によれば、「情報」と「検索」とをキーワードとして段落内に有し、「自然」「言語」「処理」「利用」を上位構造である章（３章）の見出しとして有し、「情報」と「検索」とをやはり上位構造であ

る文書の見出しとして有する段落Ｐ２９が、問合せ「自然言語処理を利用した情報検索」と最も類似度が高いことになる。また、段落Ｐ２６、Ｐ２７、Ｐ２８の３つは、段落Ｐ２９の次に類似度が高いことになる。

【０１４３】

このようにして類似度を求めた検索部３０は、類似度の高い順に、検索結果表示欄２３５に検索単位を表示する（図１７のステップＳ４５）。

図２１は、検索結果の表示された入出力画面の一例を示す図である。

【０１４４】

図において、問合せ入力欄２３１には、問合せ「自然言語処理を利用した情報検索」が入力されている。また検索結果表示欄２３５には、類似度の高い順に、検索単位の構成要素を一意に特定する情報（ここでは、キーワード）と類似度とが、表示されている。検索単位の構成要素を一意に特定する情報としてはここに示すキーワードを利用する方法の他に、文字列を先頭から一定の長さだけ抽出する等の方法を採ってもよい。また、検索対象となる文書が多ければ多い程、検索結果も多くなるので、実用に際しては一定の類似度を閾値として設け、閾値以下の類似度の検索単位の構成要素は表示しないようにする。

【０１４５】

このようにして検索した結果から、抽出表示を行うことができる。抽出表示したい要素（検索単位）が、検索結果表示欄２３５からマウス等のポインティングデバイスにて選択され（図１７のステップＳ４７）、同様に範囲選択ボタン２３３から任意の表示範囲が選択され（図１７のステップＳ４８）、表示ボタン２３４が選択されると、検索文書抽出部３２の表示文書抽出部３２ｃによって、対象となる文書が文書蓄積部２１から抽出され（図１７のステップＳ４９）、表示部３１に表示される（図１７のステップＳ５０）。この抽出表示の様子は、第１の実施の形態で行った抽出表示の様子と全く同じである。

【０１４６】

このように、本発明の文書検索装置の第３の実施の形態では、蓄積された文書を任意の検索単位で分割し、各検索単位の構成要素に含まれるキーワードに対応して生成される合成ベクトルと、問合せに含まれるキーワードに対応して生成される問合せベクトルとの類似度を計算し、類似度の高いものから順に表示させる。その後、必要に応じて詳細な内容を表示させることができるため、多量の文書や巨大な文書から、必要な話題のみを簡単に抽出することが可能である。

【０１４７】

次に、本発明の文書検索装置の第４の実施の形態を説明する。

図２２は、本発明の文書検索装置の第４の実施の形態の原理構成を示した図である。なお、本発明の文書検索装置の第４の実施の形態の構成は、第１の実施の形態の構成と基本的に同一である。よって、同一構成要素には同一符号を付して詳しい説明を省略する。

【０１４８】

文書検索装置の第４の実施の形態は、文書蓄積部１と、分類単位指定部２と、論理構造解析部３と、基本ベクトル生成部４と、内容ベクトル生成部４５と、ベクトル合成部４６と、合成ベクトル保持部７と、分類部８と、分類結果保持部９と、表示部１０と、再分類指示部１１と、検索文書抽出部１２と、から構成されている。

【０１４９】

内容ベクトル生成部４５は、論理構造解析部３にて解析された文書の論理構造において、基本ベクトル生成部４にて基本ベクトル生成の対象となった分類単位の構成要素以外の構成要素の内容からキーワードを抽出し、内容ベクトルを生成する。この内容ベクトル生成については、後に例を挙げて詳しく説明する。

【０１５０】

ベクトル合成部４６は、基本ベクトル生成部４にて生成された基本ベクトルと、内容ベクトル生成部４５にて生成された内容ベクトルとを合成して、合成ベクトルを生成する。ベクトル合成部４６の処理は、第１の実施の形態のベクトル合成部６の処理と、基本的には同一のものである。

【０１５１】
このような構成の文書検索装置における文書検索の手順は第１の実施の形態における文書検索の手順とほぼ同じで、図２、図３に示したフローチャートに沿って処理が行われる。但し、図２に示したステップＳ５の見出しベクトルの生成の代わりに、内容ベクトルの生成が行われる。以下、この内容ベクトルの生成について説明を行う。
【０１５２】
図２３は、文書検索装置の第４の実施の形態における文書検索中、内容ベクトルを生成する際の手順を示すフローチャートである。
内容ベクトル生成部４５は、まず基本ベクトル生成部４で基本ベクトル生成の対象となっている分類単位を「Ｘ」として読み込む（ステップＳ６１）。次に、「Ｘ」に、親に当たる構成要素が存在するか否か判断する（ステップＳ６２）。「Ｘ」に、親に当たる構成要素が存在しなければ、このフローチャートに示す処理は終了となる。また、「Ｘ」に、親に当たる構成要素が存在すれば、その構成要素を「Ｙ」として読み込む（ステップＳ６３）。それから、「Ｙ」に、まだベクトル生成の対象となっていない子構成要素があるか否か判断する（ステップＳ６４）。ここで言うベクトル生成では、生成されるベクトルが基本ベクトルか、要素ベクトル（内容ベクトルを構成するために生成される複数のベクトルを、要素ベクトルと呼ぶ）かの違いは問わない。
【０１５３】
まだベクトル生成の対象となっていない子構成要素が存在すれば、その構成要素を「Ｚ」として読み込み（ステップＳ６５）、「Ｚ」を対象とした要素ベクトルを生成する（ステップＳ６６）。なお、この「Ｚ」を対象とした要素ベクトル生成の手順については、後に詳しく説明する。
【０１５４】
「Ｚ」を対象とした要素ベクトル生成が終了した場合には再度ステップＳ６４に進む。ステップＳ６４において、「Ｙ」にはもう、まだベクトル生成の対象となっていない子構成要素はない、と判断された場合、ステップＳ６４～ステップＳ６６の処理で生成された要素ベクトルを、「Ｙ」の内容ベクトルとして合成する（ステップＳ６７）。構成要素「Ｙ」に対する内容ベクトル生成が終了したので、「Ｙ」を「Ｘ」として（ステップＳ６８）、再度ステップＳ６１に進む。
【０１５５】
次に、このフローチャートのステップＳ６６にあたる、要素ベクトル生成の手順を説明する。
図２４は、文書検索装置の第４の実施の形態における文書検索中、内容ベクトル生成のために、要素ベクトルを生成する手順を示すフローチャートである。
【０１５６】
内容ベクトル生成部４５は、まず要素ベクトル生成の対象となっている要素が、段落あるいは見出しであるか否か判断する（ステップＳ７１）。そして、要素ベクトル生成の対象となっている要素が、段落あるいは見出しである場合、その内容に含まれるキーワードを抽出して要素ベクトルを生成して（ステップＳ７２）、このフローチャートに示す処理は終了となる。また、ステップＳ７１の判断において、要素ベクトル生成の対象となっている要素が段落でも見出しでもなかった場合、その要素に、まだベクトル生成の対象となっていない子構成要素が存在するか否か判断する（ステップＳ７３）。ここで言うベクトル生成では、生成されるベクトルが基本ベクトルか、要素ベクトルか、の違いは問わない。
【０１５７】
まだベクトル生成の対象となっていない子構成要素が存在すれば、その子構成要素を読み込む（ステップＳ７４）。そして、読み込んだ子構成要素を対象として、要素ベクトルの生成を行う（ステップＳ７５）、すなわちこのフローチャートに示すステップＳ７１～終了までの処理対象とする。
【０１５８】
子構成要素を対象とした要素ベクトルの生成が終了したら、再度ステップＳ７３へ進む。

ステップＳ７３において、まだベクトル生成の対象となっていない子構成要素が存在しない、と判断された場合、ステップＳ７３～ステップＳ７５の処理で生成された要素ベクトルを合成し（ステップＳ７６）、このフローチャートに示す処理は終了となる。

【０１５９】

ここで、本発明の文書検索装置の第４の実施の形態で文書検索を行った場合に、どのように処理が行われるのかを、図２、図３および図２３、図２４のフローチャートに沿って図１９に示した文書１２０を例に具体的に説明する。

【０１６０】

分類単位指定部２にて検索単位が「段落」とされ（図２のステップＳ１）、論理構造解析部３に図１９に示した文書１２０が検索対象として読み込まれると（図２のステップＳ２）、文書１２０の論理構造が解析され、木構造として表現される（図２のステップＳ３）。

【０１６１】

図２５は、図１９に示した文書の論理構造を木構造にて示した図である。

基本ベクトル生成部４は、文書１２０を検索単位である「段落」毎に読み込み、各「段落」の内容からキーワードを抽出する。ここで、キーワードの抽出には従来技術を利用するものとし、特に方法の限定はしない。

【０１６２】

次に基本ベクトル生成部４は、前述の式（１）で示される計算を行って、抽出したキーワードｋ個を基に、各「段落」に対するｔ次元の基本ベクトルを生成する（図２のステップＳ４）。ここで次元数ｔは、分類に利用するキーワードの総数である。この次元数ｔの算出方法はキーワードの抽出方法によって異なり、キーワードの抽出に単語リストを使用する場合には、その単語リストに含まれる単語の総数がｔとなる。

【０１６３】

内容ベクトル生成部４５は、論理構造解析部３において抽出された全論理構成要素のうち、各「段落」より上位にある全ての構成要素を読み込む。そして、各構成要素に関するｔ次元の内容ベクトルを生成する。

【０１６４】

ここで例えば、基本ベクトル生成部４で、文書１２０の３章２節の、段落Ｐ２８を対象に基本ベクトルF_{p28}が生成されているとする。この場合、どのようにして基本ベクトルF_{P28}に対応した内容ベクトルが生成されるか、図２３～図２４のフローチャートに沿って説明する。

【０１６５】

内容ベクトル生成部４５は、内容ベクトル生成対象「Ｘ」として、文書１２０の３章２節段落Ｐ２８を読み込む（図２３のステップＳ６１）。「Ｘ」に親要素が存在するか否かを、図２５に示した木構造から判断すると（図２３のステップＳ６２）、「節」が存在している。そこでこの文書１２０の３章の２節を、「Ｙ」とする（図２３のステップＳ６３）。「Ｙ」に未処理の子構成要素が存在しているか否かを判断すると（図２３のステップＳ６４）、まず「見出し」が存在している。そこでこの文書１２０の３章２節の「見出し」を、「Ｚ」とし（図２３のステップＳ６５）、「Ｚ」の要素ベクトルを生成する。「Ｚ」の要素ベクトルを生成するにあたって、この要素「Ｚ」が段落、もしくは見出しであるか否かを判断すると（図２４のステップＳ７１）、「Ｚ」は「見出し」である。従って、この要素「Ｚ」の内容、すなわち文字列「文書構造解析」からキーワードを抽出して、要素ベクトルG_1を生成する（図２４のステップＳ７２）。なお、以後頻出する要素ベクトルの生成には、基本ベクトルの生成式である前述の式（１）をそのまま流用する。

【０１６６】

再度「Ｙ」に未処理の子構成要素が存在しているか否かを判断すると（図２３のステップＳ６４）、「段落Ｐ２７」が存在している。そこでこの文書１２０の３章２節の「段落Ｐ２７」を、新たな「Ｚ」とし（図２３のステップＳ６５）、「Ｚ」の要素ベクトルを生成する。「Ｚ」の要素ベクトルを生成するにあたって、この要素「Ｚ」が段落、もしくは見

出しであるか否かを判断すると（図２４のステップＳ７１）、「Ｚ」は「段落」である。従って、この要素「Ｚ」からキーワード（文脈、テキスト、概念、検索）を抽出して、要素ベクトルG_2を生成する（図２４のステップＳ７２）。

【０１６７】

同様に、文書１２０の３章２節の「段落Ｐ２９」に対する要素ベクトルG_3も生成する。なお、文書１２０の３章２節の「段落Ｐ２８」は既に基本ベクトルの生成対象となっているので、要素ベクトル生成の対象からは外される。「段落Ｐ２９」に対する要素ベクトル生成が終了すると、「Ｙ」には、もう未処理の子構成要素は存在しない、との判断がなされる（図２３のステップＳ６４）。そこで、ここまでで生成した要素ベクトルG_1～G_3を合成し、Ｙ（文書１２０の３章２節）に対する内容ベクトルG_{P28}を生成する（図２３のステップＳ６７）。また、内容ベクトルG_{P28}の生成が終了したので、新たな内容ベクトル生成対象「Ｘ」として「Ｙ」を定義する（図２３のステップＳ６８）。

【０１６８】

内容ベクトル生成部４５は、内容ベクトル生成対象「Ｘ」として、文書１２０の３章２節を読み込む（図２３のステップＳ６１）。「Ｘ」に親要素が存在するか否かを、図２５に示した木構造から判断すると（図２３のステップＳ６２）、「章」が存在している。そこでこの文書１２０の３章を、Ｙとする（図２３のステップＳ６３）。「Ｙ」に未処理の子構成要素が存在しているか否かを判断すると（図２３のステップＳ６４）、まず「見出し」が存在している。そこでこの文書１２０の３章の「見出し」を、「Ｚ」とし（図２３のステップＳ６５）、「Ｚ」の要素ベクトルを生成する。「Ｚ」の要素ベクトルを生成するにあたって、この要素「Ｚ」が段落、もしくは見出しであるか否かを判断すると（図２４のステップＳ７１）、「Ｚ」は「見出し」である。従って、この要素「Ｚ」の内容、すなわち文字列「自然言語処理の利用」からキーワードを抽出して、要素ベクトルH_1を生成する（図２４のステップＳ７２）。

【０１６９】

再度「Ｙ」に未処理の子構成要素が存在しているか否かを判断すると（図２３のステップＳ６４）、「１節」が存在している。そこでこの文書１２０の３章の「１節」を、新たな「Ｚ」とし（図２３のステップＳ６５）、「Ｚ」の要素ベクトルを生成する。「Ｚ」の要素ベクトルを生成するにあたって、この要素「Ｚ」が段落、もしくは見出しであるか否かを判断すると（図２４のステップＳ７１）、「Ｚ」は見出しでも、段落でもない。また、この要素「Ｚ」が未処理の子構成要素を持つか否か判断すると（図２４のステップＳ７３）、まず「見出し」が存在している。

【０１７０】

そこで、この文書１２０の３章１節の「見出し」を読み込み（図２４のステップＳ７４）、この要素に対する要素ベクトルを生成する（図２４のステップＳ７５）。すなわち、この要素が段落、もしくは見出しであるか否か判断し（図２４のステップＳ７１）、「見出し」であるので、この要素の内容である文字列「シソーラス」からキーワードを抽出して、要素ベクトルH_{21}を生成する（図２４のステップＳ７２）。

【０１７１】

再度「Ｚ」が未処理の子構成要素を持つか否か判断すると（図２４のステップＳ７３）、「段落Ｐ２６」が存在している。そこで、この文書１２０の３章１節の「段落Ｐ２６」を読み込み（図２４のステップＳ７４）、この要素に対する要素ベクトルを生成する（図２４のステップＳ７５）。すなわち、この要素が段落、もしくは見出しであるか否か判断し（図２４のステップＳ７１）、「段落」であるので、この要素からキーワード（語彙、辞書、検索）を抽出して、要素ベクトルH_{22}を生成する（図２４のステップＳ７２）。

【０１７２】

再度「Ｚ」が未処理の子構成要素を持つか否か判断すると（図２４のステップＳ７３）、もう未処理の子構成要素は存在しないので、ここまでで生成した要素ベクトルH_{21}とH_{22}とを合成し、要素「Ｚ」（文書１２０の３章１節）の要素ベクトルH_2を生成する

（図２４のステップＳ７６）。
【０１７３】
それから再度「Ｙ」に未処理の子構成要素が存在しているか否かを判断すると（図２３のステップＳ６４）、もう未処理の子構成要素は存在しない。従って、ここまでで生成した要素ベクトルH_1　とH_2　とを合成し、「Ｙ」（文書１２０の３章）に対する内容ベクトルH_{P28}　を生成する（図２３のステップＳ６７）。また、内容ベクトルH_{P28}　の生成の終了が終了したので、新たな内容ベクトル生成対象「Ｘ」として「Ｙ」を定義する（図２３のステップＳ６８）。
【０１７４】
内容ベクトル生成部４５は、内容ベクトル生成対象「Ｘ」として、文書１２０の３章を読み込む（図２３のステップＳ６１）。「Ｘ」に親要素が存在するか否かを、図２５に示した木構造から判断すると（図２３のステップＳ６２）、「文書」が存在している。そこでこの文書１２０を「Ｙ」とし（図２３のステップＳ６３）、ここまでの説明と同様の手順で「文書１２０」に対する内容ベクトルJ_{P28}　を生成する。新たな内容ベクトル生成対象「Ｘ」として「文書」を定義すると（図２３のステップＳ６８）、「Ｘ」には親要素が存在しない。よって、この内容ベクトル生成処理は終了する。
【０１７５】
ここまでの処理をまとめると、分類単位の構成要素であり、基本ベクトルF_{P28}　生成の対象とされた「段落Ｐ２８」に対し、「文書１２０の３章２節」の内容ベクトルG_{P28}　と、「文書１２０の３章」の内容ベクトルH_{P28}　と、「文書１２０」の内容ベクトルJ_{P28}　とが生成されている。
【０１７６】
図２の、文書検索全体のフローチャートに戻って説明を続けると、ベクトル合成部４６により、ここまでに生成された基本ベクトルと内容ベクトルとが合成される（図２のステップＳ６）。ここで、ある章の下のある節の下のある段落に対する合成ベクトルＴは、次式（１８）で表わされる。
【０１７７】
【数１８】
$$T = F + w_1 G + w_2 H + w_3 J \quad \cdots\cdots (18)$$
この式（１８）において、Ｆは基本ベクトル、Ｇは節の内容ベクトル、Ｈは章の内容ベクトル、Ｊは文書の内容ベクトルを表わす。また、w_1　、w_2　、w_3　は前述の式（１２）にて算出される重み付けである。このように下位概念に重きを置いた検索は、検索者の要求が具体的である場合に適している。
【０１７８】
具体例として先に説明した基本ベクトルF_{P28}　、内容ベクトルG_{P28}　、H_{P28}　、J_{P28}　を式（１８）に適用すると、合成ベクトルT_{P28}　は、下記の式（１９）にて表わすことができる。
【０１７９】
【数１９】
$$T_{P28} = F_{P28} + w_1 G_{P28} + w_2 H_{P28} + w_3 J_{P28} \quad \cdots\cdots (19)$$
なお、内容ベクトルＧ、Ｈ、Ｊには、０より大きく１以下の任意の値ｗｃ等を乗じて、さらに重み付けをすることができる。
【０１８０】
このようにして分類単位の構成要素に対する合成ベクトルが生成されると、合成ベクトル保持部７は、分類単位の構成要素と合成ベクトルとを対応付けて保持する（図２のステップＳ７）。この時点で文書蓄積部１に検索の対象とすべき文書がまだ残っていれば、以上と同様の処理を再度行う（図２のステップＳ８の分岐にてステップＳ２へ進む）が、ここでは図１９に示した文書１２０のみを検索の対象として説明を続ける。
【０１８１】

検索対象である文書すべての分類単位の構成要素に対する合成ベクトルの生成、保持が終了すると、分類部８は合成ベクトル保持部７に保持された合成ベクトルを、すべて読み込む（図２のステップＳ９）。そして分類部８は、読み込んだ合成ベクトルの類似度に基づいて、検索対象である文書すべてを、分類単位毎に、１つ以上のカテゴリーに分類し、表示する（図２のステップＳ１０）。ここで、文書の分類および結果の表示は、第１の実施の形態と同様、図３に示したフローチャートに沿って処理するものとする。

【０１８２】

以上説明したように、本発明の文書検索装置の第４の実施の形態では、蓄積された文書を任意の分類単位で分割し、その際、分類単位の構成要素の上位に位置する構成要素の内容をも考慮に入れた上で、類似度の高いもの同士で分類し、カテゴリーを生成して必要な話題を含む構成要素のみを絞り込む。その後、必要に応じて詳細な内容を表示させることができるため、多量の文書や巨大な文書から、必要な話題のみを簡単に抽出することが可能である。

【０１８３】

なお、本実施の形態では、「見出し」の要素ベクトルと、「段落」の要素ベクトルとを同価値として内容ベクトルを合成したが、検索の趣旨や文書の様子によって、「見出し」の要素ベクトルへの重み付けと「段落」の要素ベクトルへの重み付けとを変えてもよい。例えば、「見出し」の要素ベクトルへの重み付けを、「段落」の要素ベクトルへの重み付けよりも重くすれば、「段落」の内容よりも「見出し」の内容を重視した文書検索が可能となる。

【０１８４】

また、「見出し」あるいは「段落」に対する要素ベクトルを生成する際に、既に生成されている基本ベクトルとの類似度を計算し、高い類似度を持つ要素ベクトルへの重み付けを重くするようにしてもよい。例えば、従来知られているようにコサイン係数を利用すると、次式（２０）にてｔ次元のベクトルＰ＝（p_1，・・・，p_t），Ｑ＝（q_t，・・・，q_t）の類似度ｓｉｍ（Ｐ，Ｑ）（０＜ｓｉｍ（Ｐ，Ｑ）＜１）が計算できる。

【０１８５】

【数２０】

$$sim(P,Q) = \frac{\sum_{i=1}^{t} p_i q_i}{\sqrt{\sum_{i=1}^{t} p_i^{\,2}}\sqrt{\sum_{i=1}^{t} q_i^{\,2}}} \qquad \cdots\cdots\cdots (20)$$

【０１８６】

この類似度ｓｉｍ（Ｐ、Ｑ）を重み付けに利用すれば、分類が個々の検索単位の内容に特に着目して行われるようになるため、文書検索の精度を高めることが可能となる。

【０１８７】

次に、本発明の文書検索装置の第５の実施の形態を説明する。

第５の実施の形態では、第４の実施の形態で示した文書検索装置にて、階層的な論理構造の明示されていない文書を対象とした文書検索を行う。この実施の形態の装置構成は図２２に示した第４の実施の形態の装置構成と同一である。また、文書検索処理は図２および図３に示したフローチャートと同様の処理手順にて実現される。但し、図２に示したステップＳ５の見出しベクトルの生成の代わりに、内容ベクトルの生成が行われる。以下、この内容ベクトルの生成について説明する。

【０１８８】

図２６は、文書検索装置の第５の実施の形態における文書検索中、内容ベクトルを生成す

る際の手順を示すフローチャートである。
内容ベクトル生成部４５は、まず論理構造解析部３にて解析された文書の論理構造から構成要素をひとつ読み込む（ステップＳ８１）。次に、読み込んだ構成要素が図２のステップＳ４で基本ベクトルの生成の対象となった構成要素であるか否か判断する（ステップＳ８２）。読み込んだ構成要素が、図２のステップＳ５で基本ベクトルの生成の対象となった構成要素であれば、この構成要素に対しては処理を行わず、次の構成要素を読み込むために再度ステップＳ８１へ進む。ステップＳ８１で読み込んだ構成要素が、基本ベクトルの生成の対象となった構成要素ではない場合、その構成要素が段落、もしくは見出しであるか否か判断する（ステップＳ８３）。
【０１８９】
構成要素が段落でも見出しでもない場合、この構成要素に対しては処理を行わず、次の構成要素を読み込むために再度ステップＳ８１へ進む。ステップＳ８１で読み込んだ構成要素が、段落、もしくは見出しである場合、この構成要素に対する要素ベクトルを生成する（ステップＳ８４）。次に、図２のステップＳ４で基本ベクトルの生成対象となった構成要素と、ステップＳ８４で要素ベクトルを生成した構成要素との距離を計算する（ステップＳ８５）。
【０１９０】
ここで、構成要素間の距離は、図２のステップＳ３で論理構造を解析した結果として得られる木構造に基づいて計算する。この計算方法について、図を示して説明する。
【０１９１】
図２７は、図３１に示した文書の論理構造を木構造にて表現し、構成要素間の距離を示した図である。
木構造上で、あるノードから目的のノードへ辿る際に経由する枝の数を、ノード間の距離という。構成要素間の距離を木構造上でのノード間の距離と定義すると、この図２７において、段落Ｐ２から、３．２節の「見出し」までの距離は２である。また、段落Ｐ２から第３章の「見出し」までの距離は３、段落Ｐ２から文書の「表題」までの距離は４となる。さらに、段落Ｐ２から第１章の「見出し」までの距離は５となる。
【０１９２】
図２６のフローチャートに戻って説明を続けると、ステップＳ８５にて構成要素間の距離が計算されると、この距離に基づいて、ステップＳ８４で生成された要素ベクトルに対して重み付けを行う（ステップＳ８６）。重み付けの方法については、後に例を挙げて詳しく説明する。最後に、全ての構成要素を読み込んだか否かを判断し（ステップＳ８７）、まだ読み込んでいない構成要素があれば、ステップＳ８１へ戻って、次の構成要素に対する処理を続行する。また、全ての構成要素を読み込んでいれば、ここまでに生成された全ての要素ベクトルを合成して内容ベクトルを生成し（ステップＳ８８）、処理を終了する。
【０１９３】
ここで、本発明の第５の実施の形態にて文書検索を行った場合、どのように分類および検索が行われるか、図２、図３および図２６のフローチャートに沿って具体的な例を挙げて説明する。
【０１９４】
具体的な文書群として、例えば、ＷＷＷで公開されているＨＴＭＬ文書を文書蓄積部１に格納して、分類の対象とすることができる。本実施の形態では、本発明の第２の実施の形態において説明した、図１２に示した文書を、文書蓄積部１に蓄積されるＨＴＭＬ文書の例として用いる。
【０１９５】
分類単位指定部２にて分類単位を「段落」とし（図２のステップＳ１）、論理構造解析部３に図１２に示した文書が読み込まれると（図２のステップＳ２）、文書の論理構造が解析され、表題要素（ＴＩＴＬＥ）、見出し要素（Ｈ２およびＨ３）、段落要素（Ｐ）が取り出され、木構造として表現される（図２のステップＳ３）。

【０１９６】
基本ベクトル生成部４は、図１２に示した文書を、分類単位である「段落」毎に読み込み、各「段落」内容からキーワードを抽出する。キーワードの抽出方法としては、従来技術を利用するものとし、特に一つの方法には限定しない。次に基本ベクトル生成部４は、式（１）で示される計算を行って、各段落に対するｔ次元の基本ベクトルを生成する（図２のステップＳ４）。
【０１９７】
なお、本実施の形態では、前述の式（１）におけるp_{ik}の値に単語ｋの出現頻度及び分散の度合に応じて重み付けを行い、基本ベクトルのノルムが１になるように正規化を行う。基本ベクトルのノルムが１になるように正規化することにより、各構成要素のテキスト量の違いによる影響を排除することができる。
【０１９８】
単語の出現頻度、分散の度合に応じた重み付けの方法、および基本ベクトルのノルムが１になるように正規化する方法については、本発明の第２の実施の形態において説明した方法と同一の方法とする。
【０１９９】
ここで例えば、図１２に示した文書１１０の２章３節の「段落Ｐ１８」を対象に基本ベクトルを生成する場合、基本ベクトル生成部４では、まず、「段落Ｐ１８」からキーワードとして（ＮＩＳＴ、現在、コンピュータ、セキュリティ、ハンドブック、作成、．．．）が抽出され、各キーワードの出現頻度（ＴＦ）と分散（ＩＤＦ）が計算され、ノルムが１になるように正規化された基本ベクトルL_{P18}が生成される。
【０２００】
内容ベクトル生成部４５は、論理構造解析部３において抽出された全論理構成要素から、まず「表題」を読み込む（ステップＳ８１）。構成要素が基本ベクトルの生成対象であるか否かを判断すると（ステップＳ８２）、ここでは「段落Ｐ１８」を基本ベクトルの生成対象としているので、「表題」は基本ベクトルの生成対象ではない。
【０２０１】
次に、構成要素が段落、もしくは見出しであるか否かを判断すると（ステップＳ８３）、「表題」は最上位の見出しである。従って、「表題」からキーワードを抽出して、要素ベクトルM_1を生成する。なお、以後頻出する要素ベクトルの生成方法は、基本ベクトルの生成方法と同じであるとする。
【０２０２】
ここで、基本ベクトルの生成の対象となった「段落Ｐ１８」と、要素ベクトルを生成した「表題」との距離を、図１３に示した木構造に沿って計算すると（ステップＳ８５）、構成要素間の距離は４である。構成要素間の距離が計算されると、「表題」の要素ベクトルM_1に対して、重み付けがなされる（ステップＳ８６）。
【０２０３】
本実施の形態では、構成要素間の距離ｄについて単調減少する重み関数α（ｄ）を、重み付けに利用する。すなわち、基本ベクトルの生成対象である構成要素から遠い距離にある構成要素ほど、要素ベクトルの値が小さくなるように重み付けが行われる。なお、$0 < \alpha(d) \leqq 1$とする。　例えば、「表題」の要素ベクトルM_1に対しては、α（４）で重み付けを行う。
【０２０４】
要素ベクトルへの重み付けの処理が終了すると、内容ベクトル生成部４５は、文書中の全ての構成要素を読み込んだか否かを判断する（ステップＳ８７）。図１２に示した文書では、「表題」の次に「段落Ｐ１１」が存在するので、ステップＳ８１へ戻り、「段落Ｐ１１」が読み込まれる。
【０２０５】
構成要素が基本ベクトルの生成対象であるか否かを判断すると（ステップＳ８２）、「段落Ｐ１１」は基本ベクトルの生成対象ではない。次に、構成要素が段落、もしくは見出し

であるか否かを判断すると（ステップＳ８３）、「段落Ｐ１１」は段落である。

【０２０６】

したがって、「段落Ｐ１１」からキーワードを抽出して、要素ベクトルM_2　を生成する。次に、基本ベクトルの生成の対象となった「段落Ｐ１８」と、要素ベクトルを生成した「段落Ｐ１１」との距離を、図１３に示した木構造に沿って計算すると（ステップＳ８５）、構成要素間の距離は４である。構成要素間の距離が計算されると、「段落Ｐ１１」の要素ベクトルＭ２に対して、重み付けがなされる（ステップＳ８６）。「段落Ｐ１１」の要素ベクトルM_2　に対しては、α（４）で重み付けを行う。

【０２０７】

要素ベクトルへの重み付けの処理が終了すると、内容ベクトル生成部４５は、文書中の全ての構成要素を読み込んだか否かを判断する（ステップＳ８７）。図１２に示した文書では、「段落Ｐ１１」の次に「章」が存在するので、ステップＳ８１へ戻り、「章」が読み込まれる。

【０２０８】

以下同様に、文書中の全ての構成要素が読み込まれ、その構成要素が基本ベクトルの生成対象ではなく、かつ、段落もしくは見出しである場合には、要素ベクトルが生成され、基本ベクトルの生成対象である「段落Ｐ１８」との距離ｄに基づいて、重み付け関数α（ｄ）で重み付けが行われる。すなわち、図１３に示した文書の論理構造のうち、構成要素「章」と「節」については、要素ベクトルは生成されない。また、基本ベクトルの生成対象である「段落Ｐ１８」についても、要素ベクトルは生成されない。

【０２０９】

ここまでの処理をまとめると、分類単位の構成要素であり、基本ベクトルの生成の対象とされた「段落Ｐ１８」に対し、「表題」、１章の「見出し」、１章１節と１章２節の「見出し」、２章の「見出し」、２章１節から２章４節の「見出し」、「段落Ｐ１１」～「段落Ｐ１７」、および「段落Ｐ１９」の、計１７個の要素ベクトルM_1　～M_{17}が生成され、各要素ベクトルは、「段落　Ｐ１８」との距離ｄに応じて、関数α（ｄ）で重み付けされている。

【０２１０】

全ての構成要素が読み込まれると、ここまでに生成された全ての要素ベクトルを合成して内容ベクトルを生成する（ステップＳ８８）。基本ベクトルの生成の対象となった段落に対し、内容ベクトルＭは、次式（２１）にて表される。

【０２１１】

【数２１】

$$M = \sum_{i=1}^{n} \alpha(d_i) M_i \qquad \cdots\cdots(21)$$

【０２１２】

ここで、ｎは要素ベクトルの生成対象となった構成要素の数を表し、d_i　は基本ベクトルの生成の対象となった構成要素と、要素ベクトルM_i　の生成の対象となった構成要素との距離を表す。

【０２１３】

具体例として、「段落Ｐ１８」を基本ベクトルの生成対象とした場合、生成される内容ベクトルM_{P18}　は、次式で表される。

【０２１４】

【数２２】

$$M_{P18} = \sum_{i=1}^{17} \alpha(d_i) M_i \qquad \cdots\cdots(22)$$

【０２１５】
ここで、d_i は「段落Ｐ１８」と要素ベクトルM_i の生成の対象となった構成要素との距離を表す。図２の文書検索全体のフローチャートに戻って説明を続けると、ベクトル合成部４６により、ここまでに生成された基本ベクトルと内容ベクトルとが合成される（図２のステップＳ６）。
【０２１６】
すなわち、「段落Ｐ１８」に対する合成ベクトルU_{P18} は、次式（２３）にて表される。
【０２１７】
【数２３】
$$U_{P18} = L_{P18} + M_{P18} \quad \cdots\cdots (23)$$
このようにして分類単位の構成要素に対する合成ベクトルが生成されると、合成ベクトル保持部７は、分類単位の構成要素と合成ベクトルとを対応付けて保持する（図２のステップＳ７）。
【０２１８】
この時点で文書蓄積部１に分類の対象とすべき文書がまだ残っていれば、以上と同様の処理を再度行う（図２のステップＳ８の分岐にてステップＳ２へ進む）が、ここでは、図１２に示した文書のみを対象として説明を続ける。
【０２１９】
分類対象である文書すべての分類単位の構成要素に対する合成ベクトルの生成、保持が終了すると、分類部８は合成ベクトル保持部７に保持された合成ベクトルを、すべて読み込む（図２のステップＳ９）。
【０２２０】
そして分類部８は、読み込んだ合成ベクトルの類似度に基づいて、分類対象である文書すべてを、分類単位毎に、１つ以上のカテゴリーに分類し、分類結果保持部９に分類結果を保持する（図３のステップＳ１２）。本実施の形態では、ベクトルの類似度の計算方法として、コサイン係数に基づく方法、すなわち前述の式（１６）を用いる。また、分類の方法として、ｋ平均クラスタリングを用いる。
【０２２１】
ここで、図１４に示した標準セットと、本実施の形態にて文書を分類した結果とを比較する実験を行った。なお、図３のステップＳ１１にて行われる分類の方法としては第２の実施の形態にて説明したｋ平均クラスタリングを採用した。
【０２２２】
重み付け関数α（ｄ）＝　１／ｄとし、正解率Ｓを算出する式（１７）に基づいて、本実施の形態の正解率Ｓを算出すると、６１．２％となった。これに対し、図２のステップＳ５およびステップＳ６を省略した場合、すなわち、基本ベクトルのみを合成ベクトル保持部７に保持するようにした場合、分類結果の正解率Ｓは、５５．６％となる。すなわち、基本ベクトルに内容ベクトルを合成して分類を行うことにより、正解率は５．６％向上した。
【０２２３】
なお、この実験では要素ベクトルに重み付けを行う際に重み付け関数α（ｄ）＝　１／ｄとしたが、α（ｄ）＝　１／　（ｎｄ）とした場合、正解率Ｓは、６７．０％となった。ただし、ｎは、文書中のすべての構成要素の数を表す。
【０２２４】
このように、前述のα（ｄ）＝　１／ｄを、さらに文書中のすべての構成要素数で除することにより、文書毎の構成要素の総数の差異による影響を排除して、より高精度な分類が可能になる。
【０２２５】
また、要素ベクトルに重み付けを行う際に、既に生成されている基本ベクトルとの類似度を計算し、高い類似度を持つ要素ベクトルへの重み付けを重くするようにしてもよい。べ

クトルの類似度の計算は、例えば、式（１６）に示したものを利用すればよく、重み付け関数α（ｄ）＝　１／ｄに式（１６）に示した類似度を乗じて実験を行うと、正解率Ｓは７２．８％となった。

【０２２６】

このように、基本ベクトルと要素ベクトルの類似度を重み付けとして加味することにより、分類の精度をさらに高めることが可能となる。

次に、本発明の文書検索装置の第６の実施の形態を説明する。

【０２２７】

図２８は、本発明の文書検索装置の第６の実施の形態の原理構成を示した図である。なお、本発明の文書検索装置の第６の実施の形態の構成は、第３の実施の形態の構成と第４の実施の構成を組み合わせたものである。よって、同一構成要素には同一符号を付して詳しい説明を省略する。

【０２２８】

文書検索装置の第６の実施の形態は、文書蓄積部２１と、検索単位指定部２２と、論理構造解析部２３と、基本ベクトル生成部２４と、内容ベクトル生成部４５と、ベクトル合成部４６と、合成ベクトル保持部２７と、問合せ入力部２８と、問合せベクトル生成部２９と、検索部３０と、表示部３１と、検索文書抽出部３２と、から構成されている。ここで、内容ベクトル生成部４５と、ベクトル合成部４６とは第４の実施の形態の構成要素と、それ以外は第３の実施の形態の構成要素と、同一のものである。

【０２２９】

このような構成の文書検索装置にて文書検索が行われる際には、まず図１６に示したフローチャートの手順にて準備が行われるが、本実施の形態では、図１６のステップＳ３５の見出しベクトル生成の代わりに、前述の図２３～図２４にて説明した内容ベクトル生成が行われる。

【０２３０】

すなわち、まず検索単位指定部２２にて検索単位の指定が行われる（図１６のステップＳ３１）。次に、論理構造解析部２３が文書蓄積部２１から検索対象である文書を１つ読み込み（図１６のステップＳ３２）、読み込んだ文書の論理構造を解析する（図１６のステップＳ３３）。基本ベクトル生成部２４は、論理構造解析部２３にて解析された論理構造の検索単位の構成要素毎に、基本ベクトルを生成する（ステップＳ３４）。ここで内容ベクトル生成部４５は、図２３～図２４にて説明した手順にて、内容ベクトルを生成する。また、ベクトル合成部４６は、基本ベクトルと内容ベクトルとを合成して、合成ベクトルを生成する（図１６のステップＳ３６）。

【０２３１】

合成ベクトル保持部２７は、ベクトル生成の対象となった検索単位の構成要素と合成ベクトルとを、対応付けて保持する（ステップＳ３７）。ここで、文書蓄積部２１に、検索対象であり、まだステップＳ３２～ステップＳ３７の処理の対象となっていない文書が残っているか否かが判断される（ステップＳ３８）。そして、文書が残っていればステップＳ３２に再度進み、文書が残っていなければこのフローチャートの処理を終了する。ステップＳ３８の判断において文書が残っていないということは、検索対象である文書すべてに関して、ベクトルが生成されたということである。

【０２３２】

以上の文書検索準備が終了した後に、実際に文書検索が行われるが、この文書検索は、図１７に示したフローチャートの手順にて行われる。

すなわち、まず問合せ入力部２８にて、問合せが入力される（図１７のステップＳ４１）。問合せベクトル生成部２９は、入力された問合せがキーワード列であればそこから、自然言語文であればその内容からキーワードを抽出して、問合せベクトルを生成する（図１７のステップＳ４２）。

【０２３３】

問合せベクトルが生成されると、検索部３０は、問合せベクトル生成部２９から問合せベ

クトルを、合成ベクトル保持部２７から検索単位の構成要素に対する合成ベクトルすべてを、読み込む（図１７のステップＳ４３）。そして問合せベクトルと合成ベクトル群との比較を行い（図１７のステップＳ４４）、問合せベクトルとの類似度の高い合成ベクトルから順に、対応する検索単位の構成要素とともに表示部３１に表示させる（図１７のステップＳ４５）。

【０２３４】

ここまでの処理で文書検索自体は一応終了であるが、ここで文書検索の結果を基に、検索した文書の抽出表示を行うか否かの判断が行われる（図１７のステップＳ４６）。抽出表示を行う場合には、表示要素選択部３２ａにて抽出表示の対象となる構成要素が選択され（図１７のステップＳ４７）、表示範囲選択部３２ｂにて抽出表示の範囲が選択される（図１７のステップＳ４８）。そして表示文書抽出部３２ｃにて文書蓄積部２１から抽出表示範囲の文書が抽出され（図１７のステップＳ４９）、表示部３１に表示される（図１７のステップＳ５０）。

【０２３５】

この時点で、抽出表示を行わなかった場合、表示部３１には文書検索の結果が表示されている。また、抽出表示を行った場合、表示部３１には抽出文書が表示されている。そしていずれにしても、表示部３１の表示から、この文書検索処理を終了してもよいか否かが判断され（図１７のステップＳ５１）、終了しない場合は、再度ステップＳ４６以降の処理へ進み、別の構成要素に対する抽出表示を行うことが可能である。

【０２３６】

このように、本発明の文書検索装置の第６の実施の形態では、蓄積された文書を任意の検索単位で分割し、その際、検索単位の構成要素と各検索単位の構成要素の上位に位置する構成要素の内容とに含まれるキーワードに対応して生成される合成ベクトルと、問合せに含まれるキーワードに対応して生成される問合せベクトルとの類似度を計算し、類似度の高いものから順に表示させる。その後、必要に応じて詳細な内容を表示させることができるため、多量の文書や巨大な文書から、必要な話題のみを簡単に抽出することが可能である。

【０２３７】

なお、本実施の形態では、「見出し」の要素ベクトルと、「段落」の要素ベクトルとを同価値として内容ベクトルを合成したが、検索の趣旨や文書の様子によって、「見出し」の要素ベクトルへの重み付けと「段落」の要素ベクトルへの重み付けとを変えてもよい。例えば、「見出し」の要素ベクトルへの重み付けを、「段落」の要素ベクトルへの重み付けよりも重くすれば、「段落」の内容よりも「見出し」の内容を重視した文書検索が可能となる。

【０２３８】

また、「見出し」あるいは「段落」に対する要素ベクトルを生成する際に、既に生成されている基本ベクトルとの類似度を計算し、高い類似度を持つ要素ベクトルへの重み付けを重くするようにしてもよい。例えば、従来知られているようにコサイン係数を利用して前述の式（２０）にてｔ次元のベクトルＰ＝（p_1，・・・，p_t），Ｑ＝（q_1，・・・，q_t）の類似度ｓｉｍ（Ｐ，Ｑ）（０＜ｓｉｍ（Ｐ，Ｑ）＜１）を計算し、この類似度ｓｉｍ（Ｐ，Ｑ）を重み付けに利用すれば、分類が個々の検索単位の内容に特に着目して行われるようになる。これにより、文書検索の精度を高めることが可能となる。

【０２３９】

【発明の効果】

以上説明したように本発明では、文書全体の文脈を考慮に入れて、文書の構成要素の検索を行うために、検索単位もしくは分類単位として指定された構成要素が持つキーワードから、基本ベクトルを生成する基本ベクトル生成手段を設けた。そして、指定された構成要素から文書の木構造上の根に至る経路上に位置する全ての構成要素の見出しから生成した見出しベクトルや、指定された構成要素以外の構成要素が持つキーワードから生成した内容ベクトルと、基本ベクトルとを合成して合成ベクトルを生成する合成ベクトル生成部と

を設けた。そして、さらに、この合成ベクトルの類似度を利用して文書の構成要素を分類、表示する分類手段を設けた。またはこの分類手段の代わりに、問合せから生成する問合せベクトルと、合成ベクトルとを比較して類似度の高いものを検索する検索手段を設けた。

【０２４０】

これにより、文書全体の文脈を考慮に入れた上で構成要素単位の話題の検索が可能となる。また、見出しベクトルや内容ベクトルを基本ベクトルと合成する際に任意の重み付けを行うことにより、検索意図や検索対象文書の性質に沿った検索を行うことができる。

【図面の簡単な説明】

【図１】本発明の文書検索装置の第１の実施の形態の原理構成を示した図である。

【図２】文書検索装置の第１の実施の形態における文書検索の全体手順を示すフローチャートである。

【図３】文書検索装置の第１の実施の形態における文書の分類および結果表示の手順を示すフローチャートである。

【図４】図１に示した文書蓄積部に蓄積される文書の一例を示す図である。

【図５】図４に示した文書の論理構造を木構造にて示した図である。

【図６】図４に示した文書の段落Ｐ２に対する合成ベクトルの生成を示す表である。

【図７】図４に示した文書に関して、分類単位「段落」に対応する合成ベクトル保持の様子を示した図である。

【図８】分類結果の表示画面の一例を示す図である。

【図９】図８に示した表示画面において、再分類を指示する様子を示した図である。

【図１０】図８に示した表示画面において、抽出表示を指示する様子を示した図である。

【図１１】抽出文書の表示画面の一例である。

【図１２】図１に示した文書蓄積部に蓄積されるＨＴＭＬ文書の一例を示す図である。

【図１３】図１２に示した文書の論理構造を木構造にて示した図である。

【図１４】実験対象とする文書群を人手により分類した結果を示す分類表である。

【図１５】本発明の文書検索装置の第３の実施の形態の原理構成を示した図である。

【図１６】文書検索装置の第３の実施の形態における文書検索の準備を行う手順を示すフローチャートである。

【図１７】文書検索装置の第３の実施の形態において文書検索準備終了後に行う文書検索処理の手順を示すフローチャートである。

【図１８】本発明の文書検索装置の第３の実施の形態に係る入出力画面の一例を示した図である。

【図１９】文書蓄積部に蓄積される文書の一例を示す図である。

【図２０】図１９に示した文書の段落毎の合成ベクトルと問合せベクトルとの比較の様子を示した表である。

【図２１】検索結果の表示された入出力画面の一例を示す図である。

【図２２】本発明の文書検索装置の第４の実施の形態の原理構成を示した図である。

【図２３】文書検索装置の第４の実施の形態における文書検索中、内容ベクトルを生成する際の手順を示すフローチャートである。

【図２４】文書検索装置の第４の実施の形態における文書検索中、内容ベクトルを生成するために、要素ベクトルを生成する手順を示すフローチャートである。

【図２５】図１９に示した文書の論理構造を木構造にて示した図である。

【図２６】文書検索装置の第５の実施の形態における文書検索中、内容ベクトルを生成する際の手順を示すフローチャートである。

【図２７】図３１に示した文書の論理構造を木構造にて表現し、構成要素間の距離を示した図である。

【図２８】本発明の文書検索装置の第６の実施の形態の原理構成を示した図である。

【図２９】自動検索の対象文書の例を示した図である。

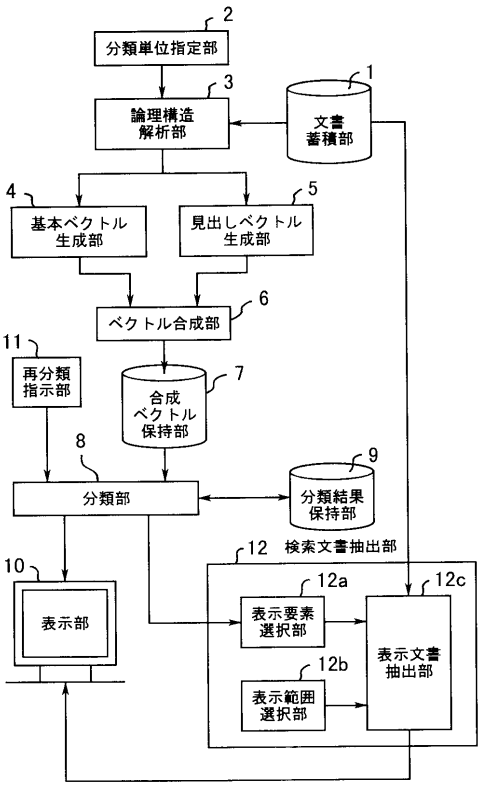
【図３０】図２９に示した各文書を、その段落構成と共に示したものである。

【図３１】表題、章、節、段落などの論理構成要素を持つ文書の一例を示す図である。

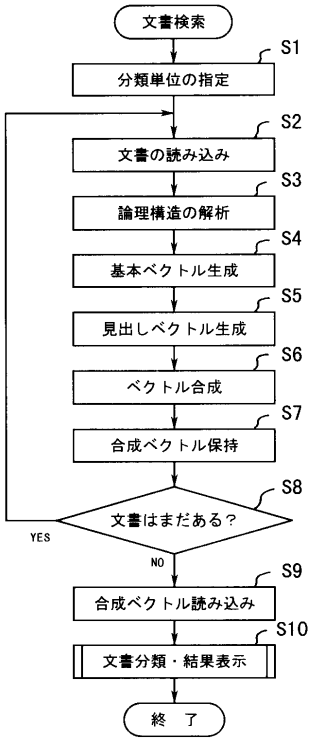
【符号の説明】

１　文書蓄積部
２　分類単位指定部
３　論理構造解析部
４　基本ベクトル生成部
５　見出しベクトル生成部
６　ベクトル合成部
７　合成ベクトル保持部
８　分類部
９　分類結果保持部
１０　表示部
１１　再分類指示部
１２　検索文書抽出部
１２ａ　表示要素選択部
１２ｂ　表示範囲選択部
１２ｃ　表示文書抽出部

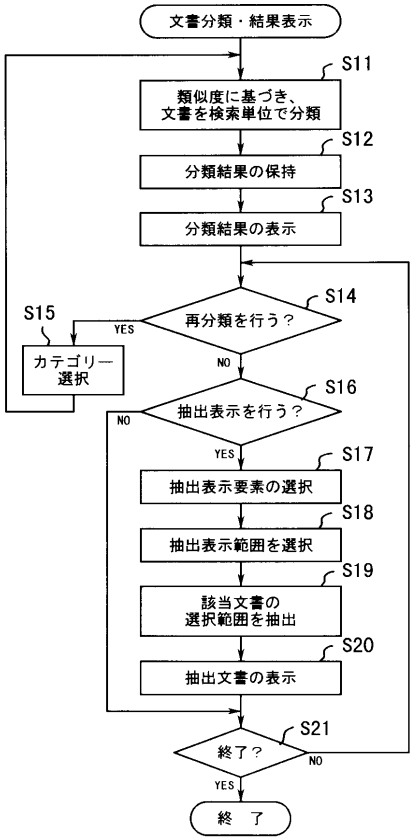
【図１】



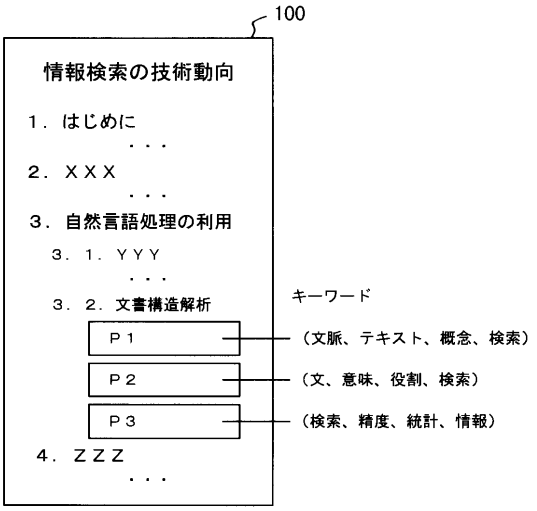
【図２】



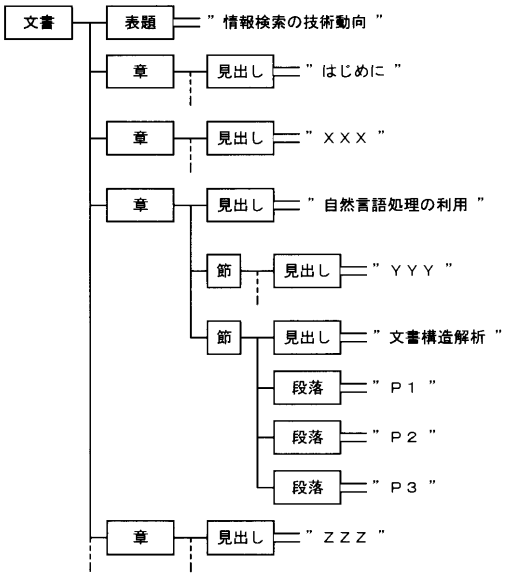
【図３】



【図４】



【図５】



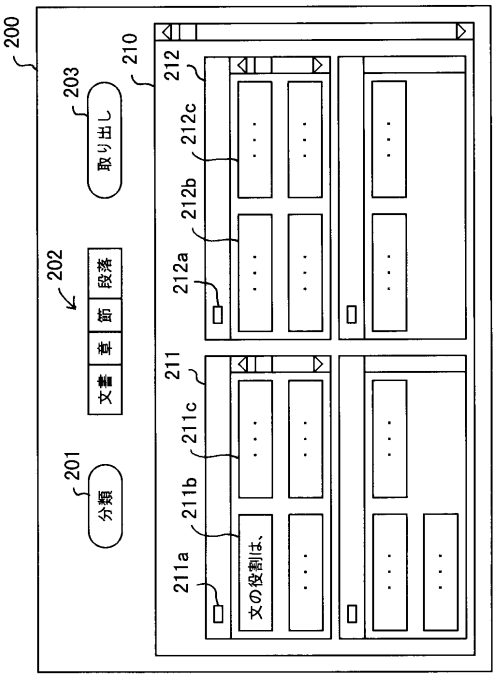
【図６】

ベクトル / キーワード	A_{P2}	w_1D	w_2C	w_3B	V_{P2}
情報				0.2	0.2
検索	1.0			0.2	1.2
技術				0.2	0.2
動向				0.2	0.2
自然			0.4		0.4
言語			0.4		0.4
処理			0.4		0.4
利用			0.4		0.4
文書		0.6			0.6
構造		0.6			0.6
解析		0.6			0.6
役割	1.0				1.0
文	1.0				1.0
意味	1.0				1.0

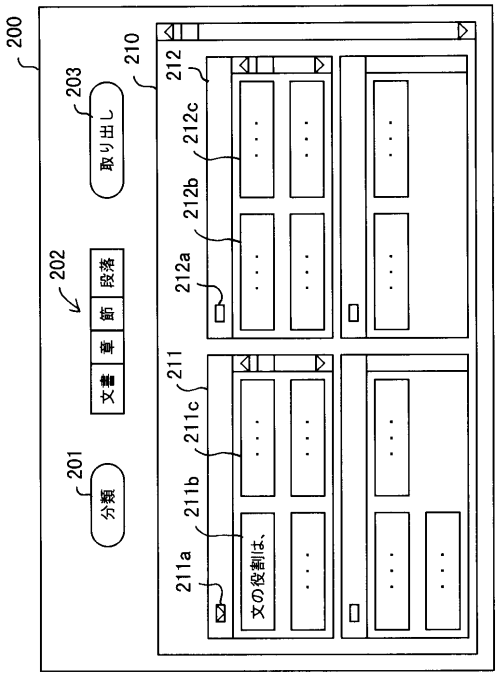
【図７】

構成要素ID	合成ベクトル
Ｐ１	$(P_{11}, P_{12}, \cdots, P_{1t})$
Ｐ２	$(P_{21}, P_{22}, \cdots, P_{2t})$ = (0.2, 1.2, 0.2, 0.2, 0.4, 0.4, 0.4, 0.4, 0.6, 0.6, 0.6, 0.6, 1.0, 1.0, 1.0)
Ｐ３	$(P_{31}, P_{32}, \cdots, P_{3t})$
……	……

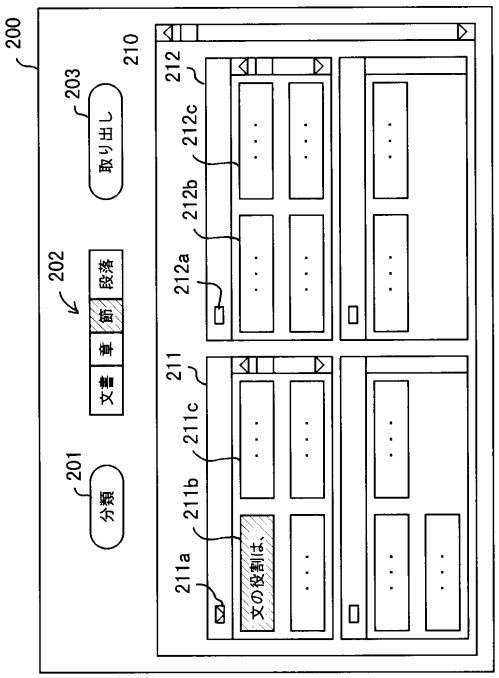
【図８】



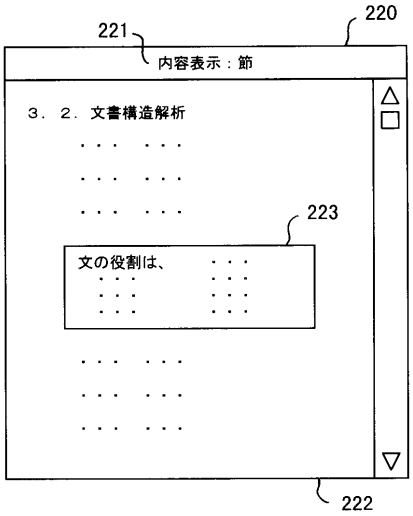
【図９】



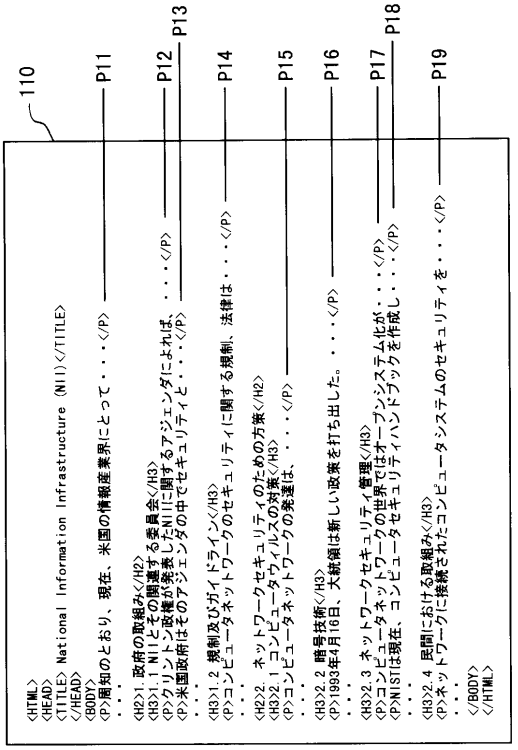
【図１０】



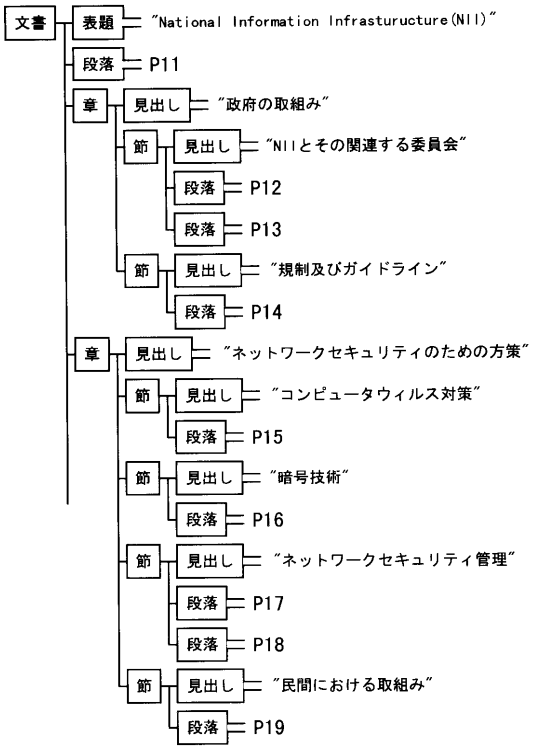
【図１１】



【図１２】



【図１３】

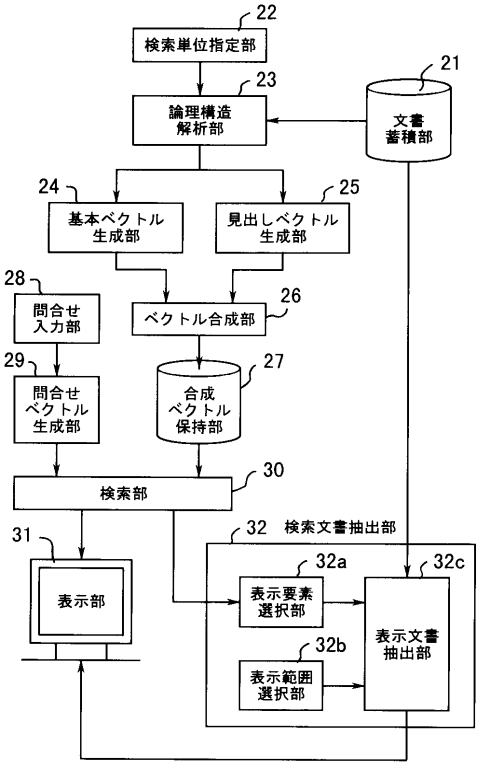


【図１４】

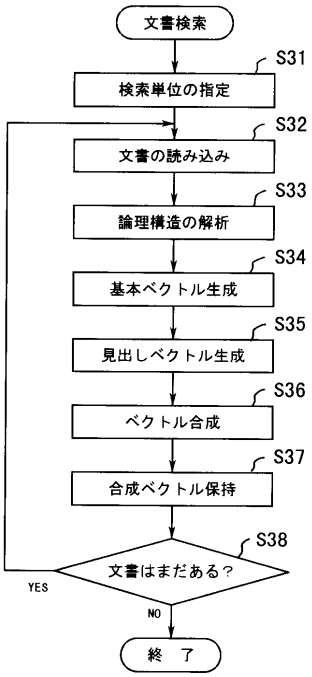
クラスタNo.	トピック	段落数
1	ＮＩＩ	７０
2	セキュリティ	７３
3	情報公開	８
4	暗号技術	１６４
5	インターネットの技術的背景	１０５
6	インターネット接続サービス	６７
7	フラットパネルディスプレイ	６３
8	ＨＰＣＣ計画	１３６
9	ＮＳＦｎｅｔ	２４
１０	教育	９４
１１	ＥＤＩ	９８
１２	エレクトロニック・コマース	９５
１３	ビデオ・オン・デマンド	５５
１４	マルチメディア	３１

クラスタNo.	トピック	段落数
１５	パソコン	３４
１６	企業動向	６８
１７	ＭＰＵ	２０
１８	マイクロソフト	１３０
１９	ＢＢＳ	４８
２０	法律	２７
２１	イントラネット	９８
２２	インターネット電話	１７
２３	インターネットビジネス	３３
２４	ＣＡＬＳ	３３
２５	モデム	２０
２６	アップル	１０
２７	ＮＣ	９
２８	分類不能	５

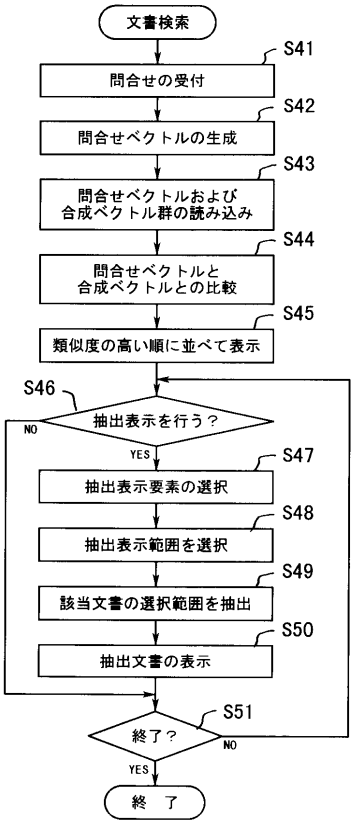
【図１５】



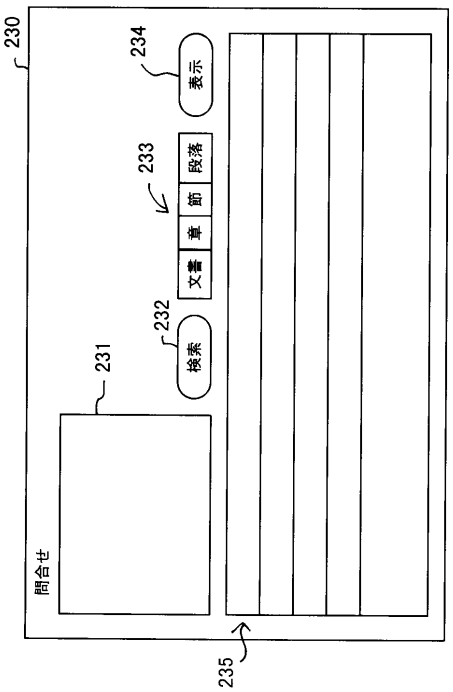
【図１６】



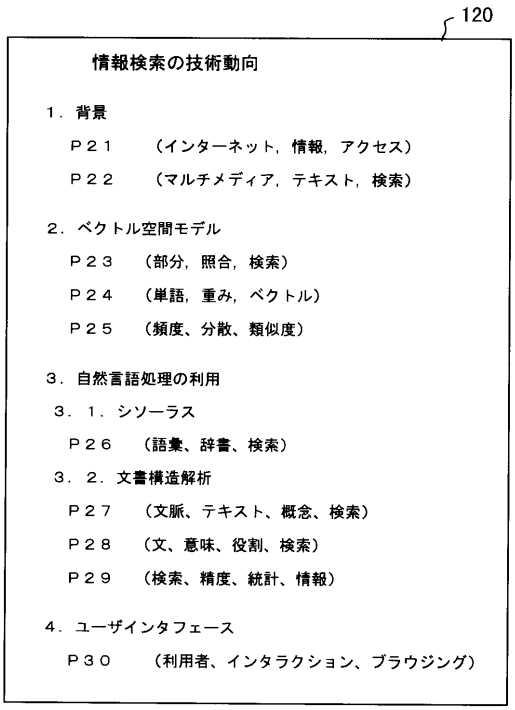
【図１７】



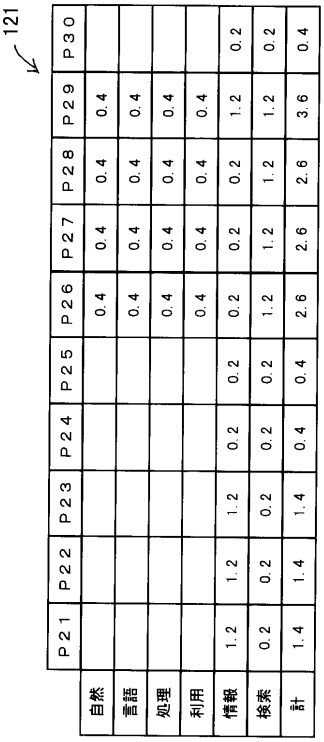
【図１８】



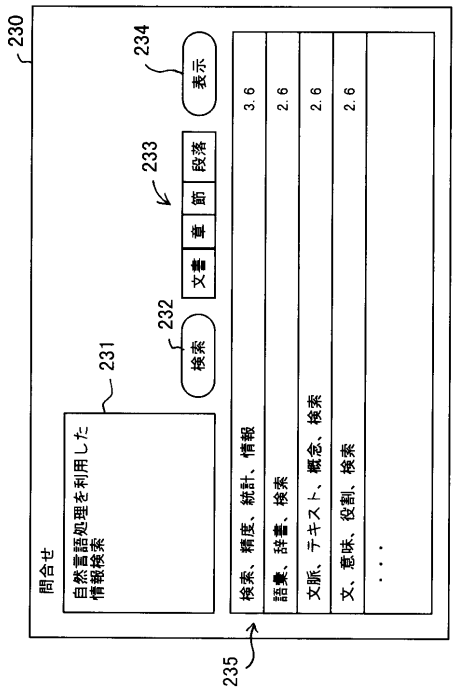
【図１９】



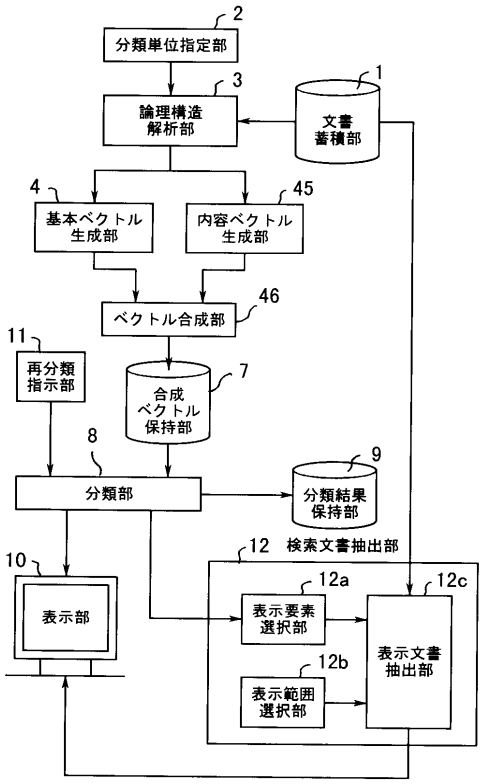
【図２０】



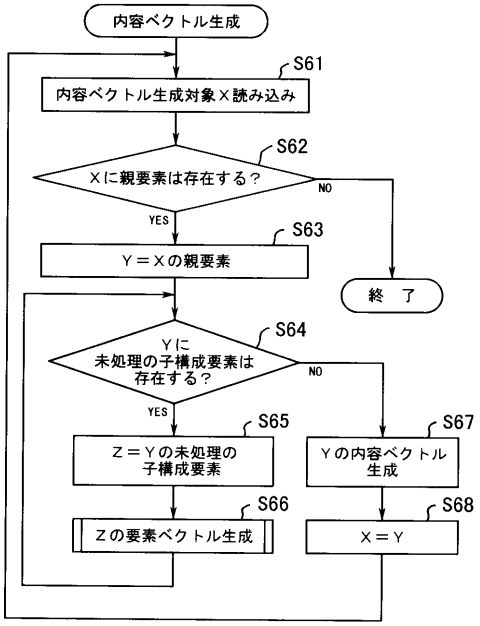
【図２１】



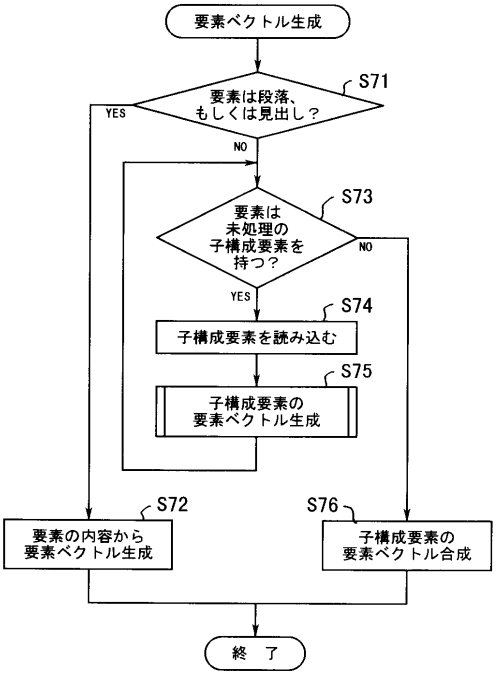
【図２２】



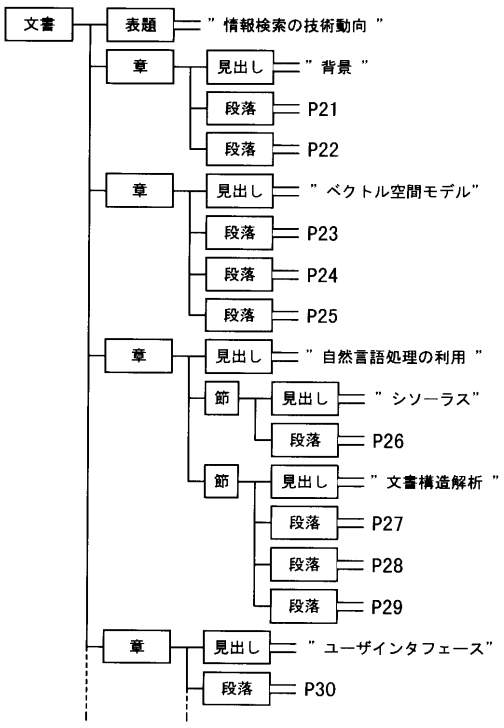
【図２３】



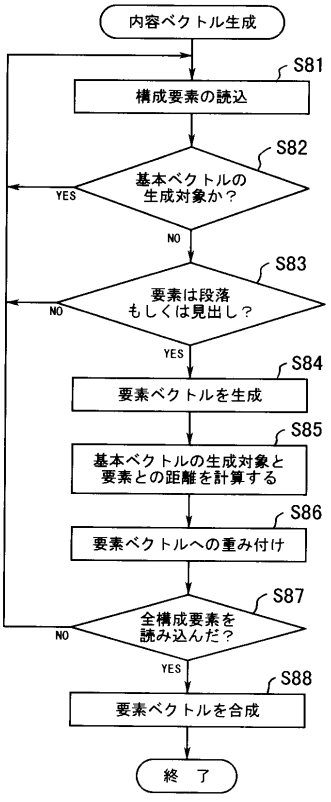
【図２４】



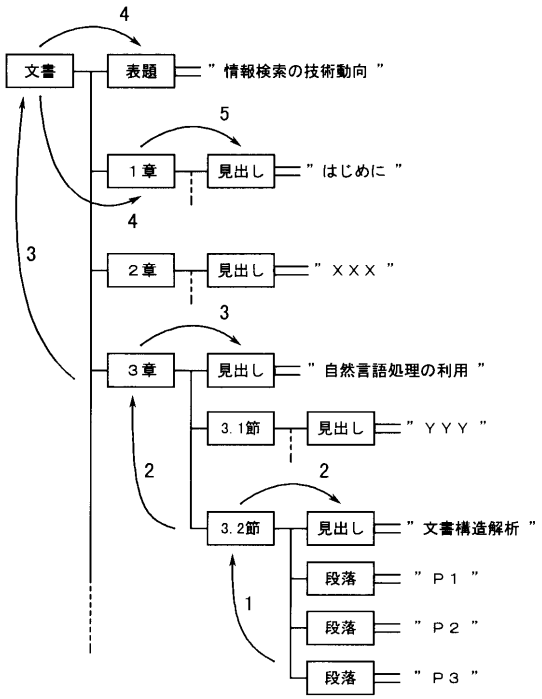
【図２５】



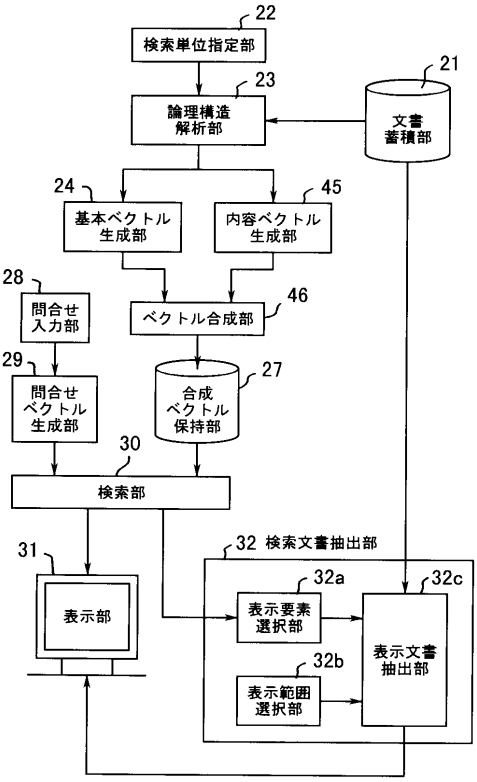
【図２６】



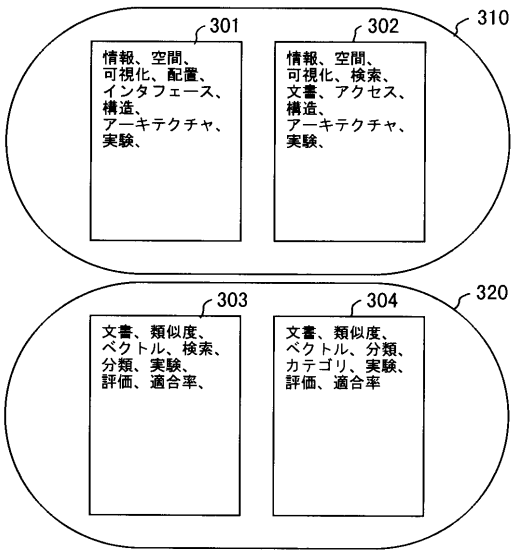
【図２７】



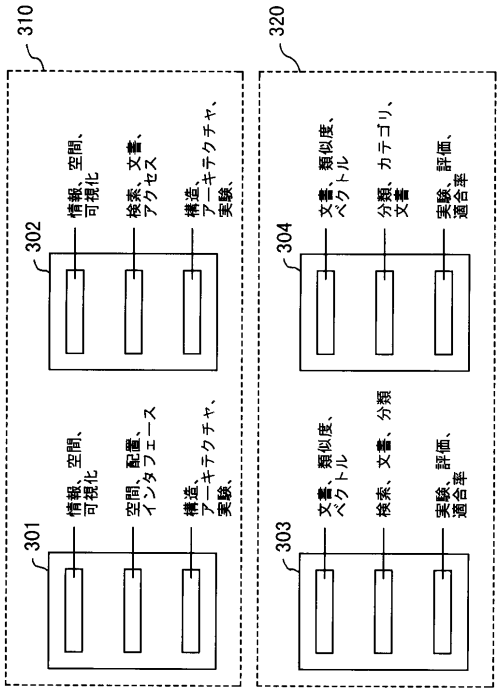
【図２８】



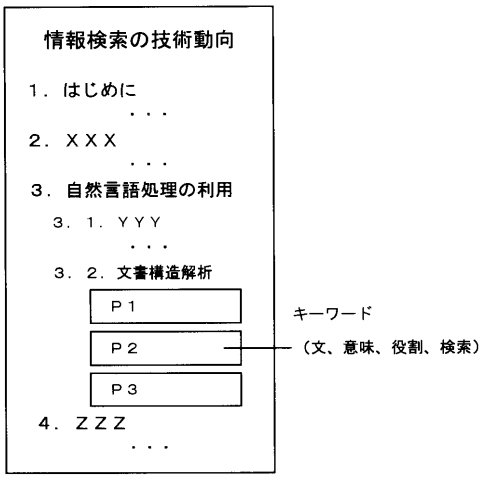
【図２９】



【図３０】



【図３１】


情報検索の技術動向
１．はじめに
・・・
２．ＸＸＸ
・・・
３．自然言語処理の利用
３．１．ＹＹＹ
・・・
３．２．文書構造解析
Ｐ１
Ｐ２
Ｐ３
キーワード
（文、意味、役割、検索）
４．ＺＺＺ
・・・

フロントページの続き

(58)調査した分野(Int.Cl.[7], ＤＢ名)
G06F 17/30 - G06F 17/30 419
JICSTファイル(JOIS)