(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4362492号
(P4362492)
(45)発行日　平成21年11月11日(2009.11.11)　(24)登録日　平成21年8月21日(2009.8.21)

(51)Int.Cl.　F I
*G06F 17/30 (2006.01)*　G06F 17/30 210A
G06F 17/30 210D
G06F 17/30 350C

請求項の数 12　(全 34 頁)

(21)出願番号　特願2006-99401(P2006-99401)
(22)出願日　平成18年3月31日(2006.3.31)
(65)公開番号　特開2007-272699(P2007-272699A)
(43)公開日　平成19年10月18日(2007.10.18)
審査請求日　平成18年3月31日(2006.3.31)

(73)特許権者　504202472
大学共同利用機関法人情報・システム研究機構
東京都港区南麻布四丁目６番７号
(73)特許権者　504402647
有限会社エクセリードテクノロジー
東京都杉並区松庵３丁目２０番地１１号グレイス松庵２０２号
(74)代理人　100083806
弁理士　三好　秀和
(74)代理人　100101247
弁理士　高橋　俊一
(74)代理人　100109380
弁理士　小西　恵

最終頁に続く

(54)【発明の名称】文書インデキシング装置、文書検索装置、文書分類装置、並びにその方法及びプログラム

(57)【特許請求の範囲】
【請求項１】
入力された日本語文書テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、漢字文字列及びカタカナ文字列をそれぞれ抽出する文字コード識別部と、
抽出された前記漢字文字列及び前記カタカナ文字列のうち、２文字以上連続する文字列の出現回数をカウントする文字列出現回数カウント部と、
前記出現回数がカウントされた文字列のうち、前記入力された日本語文書テキスト内で、第１の所定比率或いは第１の所定出現回数以上の出現頻度を有する漢字文字列を、前記日本語文書テキスト内で、前記第１の所定比率より大きい第２の所定比率或いは前記第１の所定出現回数より小さい第２の所定出現回数以上の出現頻度を有するカタカナ文字列を、それぞれキーワードとして抽出するキーワード生成部と、
前記キーワードのそれぞれについて、前記入力された日本語文書テキスト内で、前記キーワードの出現回数と、当該キーワードと同一文字種別に属する抽出されたキーワードの最小出現回数との差分を重みとして算出する重み算出部と、
前記重みを前記キーワードに付加して得られる重み付きキーワードと前記入力された日本語文書テキストとを対応付けるキーワード管理部と、
前記対応付けられた重み付きキーワード及び前記日本語文書テキストとを格納する文書格納部と、
入力されたキーワードと、前記文書格納部に格納された日本語文書テキストに対応付け

られた重み付きキーワードとを比較し、前記入力されたキーワードと少なくとも部分的に一致する重み付きキーワードを識別し、前記日本語文書テキストについて、前記一致するキーワードに付加された重みの総和を一致度として得、該一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストを選択して、クライアント装置に送出するキーワード一致度算出部とを具備する
ことを特徴とする文書インデキシングサーバ装置。
【請求項２】
前記文字コード識別部は、さらに、前記入力された日本語テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、アルファベット文字列を抽出し、
前記文字列出現回数カウント部は、さらに、抽出された前記アルファベット文字列のうち、２文字以上連続する文字列の出現回数をカウントし、
前記キーワード生成部は、さらに、前記出現回数がカウントされた文字列のうち、前記入力された日本語文書テキスト内で、前記第１の所定比率より大きい第３の所定比率或いは前記第１の所定出現回数より小さい第３の所定出現回数以上の出現頻度を有するアルファベット文字列をキーワードとして得る
ことを特徴とする請求項１に記載の文書インデキシングサーバ装置。
【請求項３】
上記文書インデキシングサーバ装置は、さらに、
クライアント装置から前記重みが付加されたキーワードを受信する受信部を具備し、
前記キーワード一致度算出部は、前記日本語文書テキストについて、前記受信されたキーワードに付加された第１の重みと前記一致するキーワードに付加された第２の重みとの積を総和して一致度として得る
ことを特徴とする請求項２に記載の文書インデキシングサーバ装置。
【請求項４】
上記文書インデキシングサーバ装置は、さらに、
前記一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストから、前記受信されたキーワードを含む文章のみを抽出して要約を作成し、前記クライアント装置に送出する要約生成部を具備する
ことを特徴とする請求項１又は２に記載の文書インデキシングサーバ装置。
【請求項５】
上記文書インデキシングサーバ装置は、さらに、
当該日本語文書テキストに対応付けられた重み付きキーワードと、文書格納部に格納される他の日本語文書テキストに対応付けられた重み付きキーワードとを比較することにより、当該日本語文書テキストのキーワードに付加された第１の重みと他の日本語文書テキストのキーワードに付加された第２の重みとの積を総和して文書間一致度を算出し、算出された文書間一致度のうち所定の閾値以上の文書間一致度を、すべての日本語文書テキストの組み合わせについて記述する一致度マトリクスを生成する一致度マトリクス生成部と、
この一致度マトリクスを参照することにより、当該日本語文書テキストから他の日本語文書テキストへの前記所定の閾値以上の文書間一致度を示す有向グラフを形成する有向グラフ形成部と、
形成された有向グラフを順次辿って相互に到達可能な関係を有する複数の日本語文書テキストを、１つの日本語文書テキスト群に分類する分類部とを具備する
ことを特徴とする請求項１ないし４のいずれか記載の文書インデキシングサーバ装置。
【請求項６】
キーワードを入力するキーワード入力部と、
入力されたキーワードをサーバ装置に送信するキーワード送信部と、
入力されたキーワードの送信に応答して、前記サーバ装置から、検索結果として日本語文書テキストを受信して表示出力する文書表示部とを具備し、

受信される前記日本語文書テキストは、入力されたキーワードと、前記サーバ装置の文書格納部に格納された日本語文書テキストに対応付けられた重み付きキーワードとを比較し、前記入力されたキーワードと少なくとも部分的に一致する重み付きキーワードを識別し、前記日本語文書テキストについて、前記一致するキーワードに付加された重みの総和を一致度として得、該一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストであり、

前記日本語文書テキストのキーワードに付加される重みは、前記日本語文書テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、漢字文字列及びカタカナ文字列をそれぞれ抽出し、抽出された前記漢字文字列及び前記カタカナ文字列のうち、２文字以上連続する文字列の出現回数をカウントし、前記出現回数がカウントされた文字列のうち、前記日本語文書テキスト内で、第１の所定比率或いは第１の所定出現回数以上の出現頻度を有する漢字文字列を、前記日本語文書テキスト内で、前記第１の所定比率より大きい第２の所定比率或いは前記第１の所定出現回数より小さい第２の所定出現回数以上の出現頻度を有するカタカナ文字列を、それぞれキーワードとして得、前記キーワードのそれぞれについて、前記日本語文書テキスト内で、前記キーワードの出現回数と、当該キーワードと同一文字種別に属する抽出されたキーワードの最小出現回数との差分を重みとして算出することにより得られるものである

ことを特徴とする文書インデキシングクライアント装置。

【請求項７】

上記文書インデキシングクライアント装置は、さらに、

前記キーワードの送信に応答して、前記サーバ装置から、検索結果である日本語文書テキストの要約テキストを受信すると共に提示する要約提示部を具備し、

前記要約テキストは、前記送信されたキーワードを含む文章のみからなる

ことを特徴とする請求項６に記載の文書インデキシングクライアント装置。

【請求項８】

上記文書インデキシングクライアント装置は、さらに、

前記キーワードの送信に応答して、前記サーバ装置から、複数の日本語文書テキストの分類を受信する受信部と、

受信された複数の分類から、１又は複数の分類の選択入力を促す分類選択入力部とを具備する

ことを特徴とする請求項６又は７に記載の文書インデキシングクライアント装置。

【請求項９】

文字コード識別部と、文字列出現回数カウント部と、キーワード生成部と、重み算出部と、キーワード管理部と、文書格納部と、キーワード一致度算出部を備える文書インデキシングサーバ装置により実行される文書インデキシング方法であって、

前記文字コード識別部により、入力された日本語文書テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、漢字文字列及びカタカナ文字列をそれぞれ抽出するステップと、

前記文字列出現回数カウント部により、抽出された前記漢字文字列及び前記カタカナ文字列のうち、２文字以上連続する文字列の出現回数をカウントするステップと、

前記キーワード生成部により、前記出現回数がカウントされた文字列のうち、前記入力された日本語文書テキスト内で、第１の所定比率或いは第１の所定出現回数以上の出現頻度を有する漢字文字列を、前記日本語文書テキスト内で、前記第１の所定比率より大きい第２の所定比率或いは前記第１の所定出現回数より小さい第２の所定出現回数以上の出現頻度を有するカタカナ文字列を、それぞれキーワードとして抽出するステップと、

前記重み算出部により、前記キーワードのそれぞれについて、前記入力された日本語文書テキスト内で、前記キーワードの出現回数と、当該キーワードと同一文字種別に属する抽出されたキーワードの最小出現回数との差分を重みとして算出するステップと、

前記キーワード管理部により、前記得られたキーワードと前記入力された日本語文書テキストとを対応付けるステップと、

前記文書格納部により、前記対応付けられたキーワード及び前記日本語文書テキストとを格納するステップと、
前記キーワード一致度算出部により、入力されたキーワードと、前記文書格納部に格納された日本語文書テキストに対応付けられた重み付きキーワードとを比較し、前期入力されたキーワードと少なくとも部分的に一致する重み付きキーワードを識別し、前記日本語文書テキストについて、前記一致するキーワードに付加された重みの総和を一致度として得、該一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストを選択して、クライアント装置に送出するステップとを含む
ことを特徴とする文書インデキシング方法。
【請求項１０】
キーワード入力部と、キーワード送信部と、文書表示部とを備える文書インデキシングクライアント装置により実行される文書インデキシング方法であって、
前記キーワード入力部により、キーワードを入力するステップと、
前記キーワード送信部により、入力されたキーワードをサーバ装置に送信するステップと、
前記文書表示部により、入力されたキーワードの送信に応答して、前記サーバ装置から、検索結果として日本語文書テキストを受信して表示出力するステップとを含み、
受信される前記日本語文書テキストは、入力されたキーワードと、前記サーバ装置の文書格納部に格納された日本語文書テキストに対応付けられた重み付きキーワードとを比較し、前記入力されたキーワードと少なくとも部分的に一致する重み付きキーワードを識別し、前記日本語文書テキストについて、前記一致するキーワードに付加された重みの総和を一致度として得、該一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストであり、
前記日本語文書テキストのキーワードに付加される重みは、前記日本語文書テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、漢字文字列及びカタカナ文字列をそれぞれ抽出し、抽出された前記漢字文字列及び前記カタカナ文字列のうち、２文字以上連続する文字列の出現回数をカウントし、前記出現回数がカウントされた文字列のうち、前記日本語文書テキスト内で、第１の所定比率或いは第１の所定出現回数以上の出現頻度を有する漢字文字列を、前記日本語文書テキスト内で、前記第１の所定比率より大きい第２の所定比率或いは前記第１の所定出現回数より小さい第２の所定出現回数以上の出現頻度を有するカタカナ文字列を、それぞれキーワードとして得、前記キーワードのそれぞれについて、前記日本語文書テキスト内で、前記キーワードの出現回数と、当該キーワードと同一文字種別に属する抽出されたキーワードの最小出現回数との差分を重みとして算出することにより得られるものである
ことを特徴とする文書インデキシング方法。
【請求項１１】
文書インデキシング処理を、文字コード識別部と、文字列出現回数カウント部と、キーワード生成部と、重み算出部と、キーワード管理部と、文書格納部と、キーワード一致度算出部を備える文書インデキシングサーバ装置として動作するコンピュータに実行させるための文書インデキシングプログラムであって、該プログラムは、前記コンピュータに、
前記文字コード識別部により、入力された日本語文書テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、漢字文字列及びカタカナ文字列をそれぞれ抽出する処理と、
前記文字列出現回数カウント部により、抽出された前記漢字文字列及び前記カタカナ文字列のうち、２文字以上連続する文字列の出現回数をカウントする処理と、
前記キーワード生成部により、前記出現回数がカウントされた文字列のうち、前記入力された日本語文書テキスト内で、第１の所定比率或いは第１の所定出現回数以上の出現頻度を有する漢字文字列を、前記日本語文書テキスト内で、前記第１の所定比率より大きい第２の所定比率或いは前記第１の所定出現回数より小さい第２の所定出現回数以上の出現頻度を有するカタカナ文字列を、それぞれキーワードとして抽出する処理と、

前記重み算出部により、前記キーワードのそれぞれについて、前記入力された日本語文書テキスト内で、前記キーワードの出現回数と、当該キーワードと同一文字種別に属する抽出されたキーワードの最小出現回数との差分を重みとして算出する処理と、
前記キーワード管理部により、前記得られたキーワードと前記入力された日本語文書テキストとを対応付ける処理と、
前記文書格納部により、前記対応付けられたキーワード及び前記日本語文書テキストとを格納する処理と、
前記キーワード一致度算出部により、入力されたキーワードと、前記文書格納部に格納された日本語文書テキストに対応付けられた重み付きキーワードとを比較し、前期入力されたキーワードと少なくとも部分的に一致する重み付きキーワードを識別し、前記日本語文書テキストについて、前記一致するキーワードに付加された重みの総和を一致度として得、該一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストを選択して、クライアント装置に送出する処理とを含む処理を実行させるためのものである
ことを特徴とする文書インデキシングプログラム。
【請求項１２】
文書インデキシング処理を、キーワード入力部と、キーワード送信部と、文書表示部とを備える文書インデキシングクライアント装置として動作するコンピュータに実行させるための文書インデキシングプログラムであって、該プログラムは、前記コンピュータに、
前記キーワード入力部により、キーワードを入力する処理と、
前記キーワード送信部により、入力されたキーワードをサーバ装置に送信する処理と、
前記文書表示部により、入力されたキーワードの送信に応答して、前記サーバ装置から、検索結果として日本語文書テキストを受信して表示出力する処理とを含む処理を実行させるためのものであり、
受信される前記日本語文書テキストは、入力されたキーワードと、前記サーバ装置の文書格納部に格納された日本語文書テキストに対応付けられた重み付きキーワードとを比較し、前記入力されたキーワードと少なくとも部分的に一致する重み付きキーワードを識別し、前記日本語文書テキストについて、前記一致するキーワードに付加された重みの総和を一致度として得、該一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストであり、
前記日本語文書テキストのキーワードに付加される重みは、前記日本語文書テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、漢字文字列及びカタカナ文字列をそれぞれ抽出し、抽出された前記漢字文字列及び前記カタカナ文字列のうち、２文字以上連続する文字列の出現回数をカウントし、前記出現回数がカウントされた文字列のうち、前記日本語文書テキスト内で、第１の所定比率或いは第１の所定出現回数以上の出現頻度を有する漢字文字列を、前記日本語文書テキスト内で、前記第１の所定比率より大きい第２の所定比率或いは前記第１の所定出現回数より小さい第２の所定出現回数以上の出現頻度を有するカタカナ文字列を、それぞれキーワードとして得、前記キーワードのそれぞれについて、前記日本語文書テキスト内で、前記キーワードの出現回数と、当該キーワードと同一文字種別に属する抽出されたキーワードの最小出現回数との差分を重みとして算出することにより得られるものである
ことを特徴とする文書インデキシングプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、文書インデキシング装置、文書検索装置、文書分類装置、並びにその方法及びプログラムに関する。より詳しくは、文書テキストを管理する文書管理サーバと、該文書テキストを検索及び提示する文書検索装置に実装される、インターネットなどの通信回線を介したデジタル化された文書テキストの検索システムにおいて、大量の文書テキスト、殊に既存の日本語文書テキストに対してキーワードを簡易且つ自動的に抽出し、該文書テキストに対して付与し、及び付与されたキーワードに基づいて、自由キーワードによる

文書テキスト検索を容易化すると共に、文書テキストを簡易且つ自動的に分類するための技術に関する。

【背景技術】

【０００２】

近年、企業から、及び個人からの、双方向の情報収集及び情報発信が容易化され、その結果、大量の文書データによる知識集積が進展している。これら大量の文書データは、例えば、インターネット上のＷｅｂページからのダウンロードにより、企業内イントラネット上のファイルサーバ中或いはパーソナルコンピュータのハードディスク上への蓄積により、又はメールソフトの受発信済みデータとして得られる。文書データを格納するためのファイル形式は、テキストファイル、アプリケーションが直接アクセスするネイティブファイル、或いは例えばＸＭＬ、ＨＴＭＬ等により構造化されたテキストファイルであってよく、或いはテキストデータを抽出可能なＰＤＦファイルであってもよい。

【０００３】

グループウェア等のソフトウエアが、ネットワークを構成するいずれかのコンピュータに導入され、該コンピュータがファイルサーバを構成すれば、このファイルサーバが保存する共有文書データを含む各種データは、ネットワークに接続される各コンピュータ、すなわちクライアント端末からアクセス可能となる。このグループウェアには、クライアント端末からの要求に応じて、ファイルサーバに蓄積された文書データから、所望の文書データを検索させる機能が備えられる。このグループウェア等が提供する検索機能を利用することにより、利用者は、クライアント端末を介して、ファイルサーバが管理する大量の文書データから、所望の文書データを検索する利用形態が一般的である。或いは、利用者は、自身が管理するパーソナルコンピュータ内やＷｅｂページから、これらが提供する検索ツールを用いて、所望の文書データを検索することができる。

【０００４】

ところで、従来における文書データの検索手法は、テキスト全文検索処理が未だ一般的であるが、このテキスト全文検索には、文書データの数や各文書データの容量に比例して、検索時間も長くなり、しばしば実用的検索時間によっては所望の文書データが検索されないという問題がある。

【０００５】

この検索時間を短縮するため、文書データをデータベースに登録する者が、登録すべき文書データに対するキーワード等のメタデータ付与を登録の際に行なう手法が利用されている。所与の文書データを形態素に区切り、その動詞、助詞、助動詞、名詞等の品詞種別を認識して形態素と品詞の対応付けを行なう形態素解析エンジンを介して文書データにインデキシングを行なう手法もまた、利用されている。これらの手法は、Ｗｅｂページ上で既知である多数のサーチエンジンにも同様に実装されている。

【０００６】

しかしながら、文書データをデータベースに登録する際に、こうしたインデキシングが行なわれておらず、従って検索しても見つけられない状態にある文書データが既に大量に存在する。こうした文書データをファイルサーバに保存しても、このファイルサーバがゴミ箱化していくだけで、保存された文書データを再利用することはできない。情報化社会における情報の再利用、再活用を促進するためには、文書データの保存時に、その文書内容に効率的に且つ自動的にインデキシングを施し、このインデキシングをコンピュータに把握させることが要請される。

【０００７】

すなわち、デジタル文書コンテンツのさらなる流通を促進するためには、大量に存在し、且つ、経済分野、技術分野や芸術分野等、多様なカテゴリーにそれぞれ属する文書テキストに対して、簡易且つ自動的にキーワードを付与し、及びキーワードが付与された文書テキストの類似性を評価し、文書テキストを高精度に細分類することが、文書コンテンツ検索可能性の向上に資する。

【０００８】

特許文献１は、キーワード抽出対象である文書テキストから、形態素解析により名詞を選択し、選択された名詞ごとに、文書テキスト内出現頻度を求め、同時に全文検索（フルテキストサーチ）により文書データベース全体中での出現頻度を求めて、入力テキスト文書内での出現頻度／文書データベース全体中での出現頻度、を当該名詞の重要度として算出し、該重要度の高い名詞をキーワードとして抽出する技術を開示する。特許文献１において、入力テキスト文書内での出現頻度を、文書データベース全体中での出現頻度で除するのは、文書データベースに格納された文書テキストの多くにおいて出現する名詞を、不要語としてキーワードから除外することを意図しており（例えば、特許文書における「特許」、「発明」等の名詞は文書データベース全体に亘って出現頻度が高く、個々の特許文書を識別するためには有用でない名詞である。）、従って、文書データベースに格納される文書テキストが、例えば経済分野や技術分野等、特定の大項目に含まれるような一定の均質性を備えていることを前提とする。

【０００９】

一方、特許文献２は、キーワード抽出対象である文書テキストから、文書テキスト中の隣接する少なくとも２個以上の語が、漢字、カタカナ、アルファベット、長音又は数字の任意の組み合わせにある場合に、その連続する語をキーワード候補として抽出し、抽出されたキーワード候補ごとに、同義語辞書や用語辞書等を参照して、同義関係となるキーワード候補及び後方部分一致関係となるキーワード候補をそれぞれ取り纏め、取り纏められたキーワード群について出現頻度を算出することにより、キーワードを抽出する技術を開示する。

【００１０】

また、出願人らは、すでに特許文献３において、文書テキストへの自動的キーワード付与の技術を提案している。

【特許文献１】特開２０００－７６２５４

【特許文献２】特開平６－１８７３７３

【特許文献３】特願２００５－３１９４５４

【発明の開示】

【発明が解決しようとする課題】

【００１１】

しかしながら、特許文献１に開示された技術では、文書テキストからのキーワード抽出を、形態素解析エンジンを用いて名詞を抽出することにより行なうものであり、この形態素解析エンジン用辞書に存在していないキーワードを認識することはできないし、文書データベース全体における名詞の出現頻度を、重要度算出のための係数として利用するため、異なるカテゴリーに属する文書を保有し、文書間の均質性を欠く文書データベースの場合には、キーワード抽出の精度が低下する。

【００１２】

また、特許文献２に開示された技術でも、意味的な関連を有する複数のキーワードを１つのキーワード群に取り纏めるために、辞書に依存して同義語の判定及び後方部分一致の判定を行なうものであり、これらの判定用に予め辞書を定義しなければならない。

【００１３】

例えば情報通信の技術分野等、殊に変革の激しい分野において顕著であるが、カタカナ語やアルファベットで記述される多くの新たな略語が導入される場合、辞書がこれらの略語等の新たな用語に迅速に追随していくのは非常に困難である。キーワード抽出のため参照される辞書は、時代と共に古くなるとの内在的欠点を有し、この辞書の更新を随時行なわない限り、実用的な精度でキーワードの抽出を実現することはできない。

【００１４】

もとより、テキストデータの全文検索は、非常に高負荷処理であって、実用的検索時間内には所望の文書データを検索することは著しく困難である。しかるに、この検索時間を短縮化するには、人手を介在させて、文書データ登録時にキーワードを抽出し、このキーワード群を検索時に参照される辞書として生成する、或いはＸＭＬ方式等によるメタデー

タ作成を行なうという登録時の処理を要し、こうした登録時の人手による処理は文書データのファイルサーバ等への自動登録を阻害するとともに、既に蓄積されている膨大な文書データを再利用することを実質的に不可能とする。

【００１５】

本発明は、上記課題に鑑みてされたものであり、その目的は、所与の文書テキストデータに対して、簡易且つ自動的にインデキシングを行い、辞書或いは人的ノウハウのいずれにも依存することなく、キーワードメタデータを簡易且つ低コストで自動発生させ、利用者の文書テキストデータ検索を容易化することの可能な文書インデキシング装置、文書検索装置、文書分類装置、並びにその方法及びプログラムを提供することにある。

【００１６】

また、本発明の他の目的は、利用者が入力した自由キーワードに基づく文書テキストのフリーワード検索において、自動生成されたキーワード及びその出現頻度を利用して、入力フリーワードと文書テキストとの間の一致を判定することにより、簡易且つ高精度に、目的とする文書テキストを検索結果として得ることのできる文書インデキシング装置、文書検索装置、文書分類装置、並びにその方法及びプログラムを提供することにある。

【００１７】

さらに、本発明の他の目的は、文書テキストから自動抽出されたキーワード及びその出現頻度を利用した文書テキスト間の一致度判定に基づいて、大量の文書テキストを、簡易且つ自動的に、相互に類似する文書テキスト群に分類することにある。

【課題を解決するための手段】

【００１８】

本発明に係るキーワード自動抽出の原理は、文書テキストデータ、特に２バイト以上の文字コード体系（例えば、Ｓ－ＪＩＳやＵｎｉｃｏｄｅ等）を有する例えば日本語文書テキストデータから、各文字に割り当てられた文字コードを用いて文書テキスト中の文字種別、例えば漢字及びカタカナを識別し、識別された文字種別ごとに区切られた文字列から、文書テキスト内における出現頻度の高い文字列を自動認識し、出現頻度の高い文字列をキーワードとして抽出することによって、文書テキストデータに自動的にインデキシングを行なうものである。

【００１９】

ここで、出現頻度とは、入力文書テキスト内でカウントされるキーワード（同種文字列）の出現回数を示し、文字コード種別ごと（漢字、カタカナ、アルファベット、ひらかな、数字等）にカウントされる。

【００２０】

さらに、本発明においては、抽出されたキーワードについてカウントされた出現頻度のみから得られた「重み」を、当該キーワードの重要度を示す指標として、当該キーワードに付加して、「重み付きキーワード」とする。

【００２１】

本発明によりインデキシング可能な文書テキストデータは、２バイト以上のコードで記述される例えば日本語文書テキストデータが好適であるが、別コード領域の文字（例えば、漢字、ひらかな、カタカナ、アルファベット等）が混在して文章が記述される文書テキストデータであればよく、その入力ファイル形式は、テキストファイルの他、アプリケーションが直接アクセスするネイティブファイルや、例えばＸＭＬ、ＨＴＭＬ等により構造化されたテキストファイルであってよく、或いはテキストデータを抽出可能なＰＤＦファイルであってもよい。

【００２２】

また、識別されるべき文字種別は、漢字、カタカナに加えて、あるいはこれらに替えて、必要に応じ、ひらかな、アルファベット等であってよい。

【００２３】

本発明において抽出されるキーワードの数は、好適には、例えば１０ないし１００など２桁以上の数としてよい。従来、人手でキーワードを付与する場合には、１つの文書テキ

ストに対して最大限１０個以下の数のキーワードが、キーワード付与のコスト及びキーワード提示時の一覧性の双方を考慮した場合、実用的であったが、本発明においては、キーワードは専らコンピュータが自動的に付与し、これを利用する処理もコンピュータ内部で実行されるものであることを考慮して、好適には、最終的に１つの文書テキストから自動抽出されるキーワードの数には一切制限を設けなくてよい。単純に、文字コードの相違のみで入力文字列を区別して、文字コード体系の切れ目の前後で入力文字列を分離し、それぞれの文字種ごとに別キーワードとしてカウントし、結果として１つの文書テキストから所定の出現頻度の閾値を上回るキーワードが多数抽出された場合にも、不要語を除外したり、意味解析等により複数のキーワードを取り纏める或いは重要度を判断する等の付加的処理を設けない。例えば、本発明においては、「野球」と「野球選手」とは両者とも出現頻度がある程度高い場合には、異なるキーワードとして抽出される。文字種別を跨って、１つのキーワードが抽出されることはない。

【００２４】

本発明によれば、文章テキストデータの登録時に、予め登録者によるキーワード付与や辞書登録を要することがなく、またこの辞書を用いた意味認識、形態素解析等の高負荷の処理を要することがない。このため、文書テキストデータの登録時におけるインデキシングが完全に自動化され、登録された文書テキストデータの利用者による検索が容易化する。特に、すでに蓄積されている大量の文書テキストデータに自動的にインデキシングすることが可能となるので、既存文書データの再利用に資する。さらに、文書の意味認識を必要としないので、新たな語彙が生じた場合にあっても、本発明に係るインデキシングシステムをメンテナンスする必要は生じ得ない。

【００２５】

また、本発明に係るフリーワード検索の原理は、上記のキーワード自動抽出処理において抽出されたキーワードごとに、そのキーワードの文書テキスト内での出現頻度のみから算出する値を当該キーワードの重みとし、抽出されたキーワードに重みを付加して記憶する。この重み付きキーワードに基づいて、利用者から入力された自由キーワードごとに、好適には、格納蓄積された文書テキスト中で入力自由キーワードに一致している重み付きキーワードに付与された「重み」を、文書テキストごとに総和して得られる「一致度」を算出し、一致度の高い文書テキスト、或いは所定値以上の一致度が算出された文書テキストを、フリーワード検索結果として送出する。入力キーワードに重みが付加されていた場合には、文書テキスト中で一致したキーワードごとに、入力キーワードに付与された重みと、入力キーワードと一致した文書テキストのキーワードに付与された重みとの積の総和を、文書テキストごとに算出して「一致度」としてもよい。

【００２６】

この「一致度」とは、入力された自由キーワードないし文書テキスト内のキーワードのそれぞれについて、対象文書テキスト内で入力キーワードと一致するキーワードの重みを、加算して得られる値であり、好適には、この「一致度」の算出における「一致」とは、１対のキーワードが、完全に、又は部分的に一致する文字列を有することをいう。

【００２７】

本発明によれば、上記のキーワード自動抽出処理において抽出されたキーワード及びその出現頻度のみに基づいて、自由キーワードによる簡易且つ高精度の文書テキスト検索が実現される。

【００２８】

また、本発明に係る文書テキスト分類の原理は、上記の「一致度」すなわち、キーワード自動抽出処理において抽出されたキーワードとその出現頻度のみから得られる指標に基づいて、ある文書テキストから他の文書テキストへの一致度及びその逆方向での一致度をそれぞれ算出し、所定の閾値以上の一致度（例えば相互に０でない一致度）を有する文書テキスト間のリンクで有向パスを形成し、この有向パスを順次辿って相互に行き着くことができる関係（以下において、「双方向に連結している関係」として参照される。）を有する複数の文書テキストを、相互に類似する文書テキスト群として、１つの文書テキスト

群に分類する。

【００２９】

さらに好適には、１つの文書テキスト群に分類された複数の文書テキスト同士の１対の有向パス相互間の相違に基づいて、単一の文書テキストのみを介してチェイン状に連結される関係を検出することにより、１つの分類を分割して、複数の細分類に細分割してもよい。さらに、格納される文書テキスト数が非常に多い場合には、検索キーワードを利用者に入力させ、入力された検索キーワードによりまず分類を選択させ、利用者に選択された分類に属する文書テキスト群のみを検索対象として、キーワード検索を実行してもよい。

【００３０】

本発明によれば、上記のキーワード自動抽出処理において抽出されたキーワード及びその出現頻度のみから得られる一致度のみに基づいて、多数の文書テキストを、簡易且つ高精度で、相互に高い関連性を有する文書テキスト群に自動的に分類することができる。殊に、異なるカテゴリーに属する文書テキストを保有する、文書間の均質性を欠く文書データベースをキーワード抽出対象とした場合にあっても、キーワード抽出及びこれを用いた分類の精度が低下することがない。

【００３１】

本発明のある特徴によれば、入力された日本語文書テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、漢字文字列及びカタカナ文字列をそれぞれ抽出する文字コード識別部と、抽出された前記漢字文字列及び前記カタカナ文字列のうち、２文字以上連続する文字列の出現頻度をカウントする文字列出現頻度カウント部と、前記出現頻度がカウントされた文字列のうち、前記入力された日本語文書テキスト内で第１の所定比率或いは所定出現回数以上の出現頻度を有する漢字文字列を、前記日本語文書テキスト内で第２の所定比率或いは所定出現回数以上の出現頻度を有するカタカナ文字列を、それぞれキーワードとして得るキーワード生成部と、前記キーワードのそれぞれについて、その出現頻度から、前記日本語文書テキスト内で同じ文字種別に属するキーワードについて算出された最小出現頻度を基準として、得られる値を重みとして算出する重み算出部と、前記重みを前記キーワードに付加して得られる重み付きキーワードと前記入力された日本語文書テキストとを対応付けるキーワード管理部と、前記対応付けられた重み付きキーワード及び前記日本語文書テキストとを格納する文書格納部とを具備することを特徴とする文書インデキシングサーバ装置が提供される。

【００３２】

前記文字コード識別部は、さらに、前記入力された日本語文書テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、アルファベット文字列を抽出し、前記文字列出現頻度カウント部は、さらに、抽出された前記アルファベット文字列のうち、２文字以上連続する文字列の出現頻度をカウントし、前記キーワード生成部は、さらに、前記出現頻度がカウントされた文字列のうち、前記入力された日本語文書テキスト内で第３の所定比率或いは所定出現回数以上の出現頻度を有するアルファベット文字列をキーワードとして得てよい。

【００３３】

本発明の他の特徴によれば、上記文書インデキシングサーバ装置に、さらに、入力されたキーワードと、前記文書格納部に格納された日本語文書テキストに対応付けられた重み付きキーワードとを比較し、前記入力されたキーワードと少なくとも部分的に一致する重み付きキーワードを識別し、前記日本語文書テキストについて、前記一致するキーワードに付加された重みの総和を一致度として得、該一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストを選択して、クライアント装置に送出するキーワード一致度算出部を具備してなることを特徴とする文書検索サーバ装置が提供される。

【００３４】

上記文書検索サーバ装置は、さらに、クライアント装置から前記重みが付加されたキーワードを受信する受信部を具備し、前記キーワード一致度算出部は、前記日本語文書テキストについて、前記受信されたキーワードに付加された第１の重みと前記一致するキーワ

ードに付加された第２の重みとの積を総和して一致度として得てよい。
【００３５】
上記文書検索サーバ装置は、さらに、前記一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストから、前記受信されたキーワードを含む文章のみを抽出して要約を生成して、前記クライアント装置に送出する要約生成部を具備してよい。
【００３６】
本発明の他の特徴によれば、上記文書インデキシングサーバ装置に、さらに、日本語文書テキスト間で、該日本語文書テキストに対応付けられた重み付きキーワードを比較し、相互に少なくとも部分的に一致する重み付きキーワードを識別し、各日本語文書テキストについて、自日本語文書テキストのキーワードに付加された第１の重みと前記一致するキーワードに付加された第２の重みとの積を総和して一致度として得るキーワード一致度算出部と、相互に所定の閾値以上の一致度を有する日本語文書テキスト間で形成される１対の有向パスによって連結される日本語文書テキストのすべてを、１つの分類とする分類部とを具備してなることを特徴とする文書分類サーバ装置が提供される。
【００３７】
上記文書分類サーバ装置は、さらに、前記１対の有向パスによって連結される日本語文書テキスト群を、１つのノードに縮退し、縮退されたノードを、前記分類部への入力としてよい。
【００３８】
上記文書分類サーバ装置は、さらに、前記１対の有向パスの一方が、他方と異なる日本語文書テキストのリンクを通過することを検出し、検出された有向パスによって連結される日本語文書テキスト群のみを抽出して、１つの細分類とする細分類部を具備してよい。
【００３９】
上記文書分類サーバ装置は、さらに、クライアント装置から日本語文書テキストの分類を識別する情報を受信する第２の受信部を具備し、前記キーワード一致度算出部は、識別された分類に属する日本語文書テキストのみを、前記入力されたキーワードによる検索対象としてよい。
【００４０】
本発明の他の特徴によれば、自由キーワードを入力し、入力された自由キーワードをサーバ装置に送信する自由キーワード入力部と、入力された自由キーワードを前記サーバ装置に送信するキーワード送信部と、入力された自由キーワードの送信に応答して、前記サーバ装置から、日本語文書テキストを受信し、検索結果として表示出力する文書表示部とを具備し、前記受信される日本語文書テキストは、前記入力された自由キーワードと、前記サーバ装置上で格納された日本語文書テキストに対応付けられた重み付きキーワードとを比較し、前記入力されたキーワードと少なくとも部分的に一致するキーワードを識別し、前記日本語テキストについて、前記入力された自由キーワードに付加された第１の重みと前記一致するキーワードに付加された第２の重みとの積を総和して得られた一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストであり、前記一致度は、前記日本語文書テキスト内での前記重み付きキーワードの出現頻度のみに基づいて、算出されることを特徴とする文書検索クライアント装置が提供される。
【００４１】
上記文書検索クライアント装置は、さらに、前記自由キーワードの送信に応じて、前記サーバ装置から、検索結果である日本語文書テキストの要約テキストを受信すると共に、提示する要約提示部を具備し、前記要約テキストは、前記送信された自由キーワードを含む文章のみからなってよい。
【００４２】
上記文書検索クライアント装置は、さらに、前記自由キーワードの送信に応じて、前記サーバ装置から、複数の日本語文書テキストの分類を受信する受信部と、前記受信された複数の分類から、１又は複数の分類の選択入力を促す分類選択入力部とを具備してよい。
【００４３】

本発明の他の特徴によれば、入力された日本語文書テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、漢字文字列及びカタカナ文字列をそれぞれ抽出するステップと、抽出された前記漢字文字列及び前記カタカナ文字列のうち、２文字以上連続する文字列の出現頻度をカウントするステップと、前記出現頻度がカウントされた文字列のうち、前記入力された日本語文書テキスト内で第１の所定比率或いは所定出現回数以上の出現頻度を有する漢字文字列を、前記日本語文書テキスト内で第２の所定比率或いは所定出現回数以上の出現頻度を有するカタカナ文字列を、それぞれキーワードとして得るステップと、前記キーワードのそれぞれについて、その出現頻度から、前記日本語文書テキスト内で同じ文字種別に属するキーワードについて算出された最小出現頻度を基準として、得られる値を重みとして算出するステップと、前記重みを前記キーワードに付加して得られる重み付きキーワードと前記入力された日本語文書テキストとを対応付けるステップと、前記対応付けられた重み付きキーワード及び前記日本語文書テキストとを格納するステップとを含むことを特徴とする文書インデキシング処理をコンピュータに実行させるための方法が提供される。

【００４４】

本発明の他の特徴によれば、自由キーワードを入力し、入力された自由キーワードをサーバ装置に送信するステップと、入力された自由キーワードを前記サーバ装置に送信するステップと、　入力された自由キーワードの送信に応答して、前記サーバ装置から、日本語文書テキストを受信し、検索結果として表示出力するステップとを含み、前記受信される日本語文書テキストは、前記入力された自由キーワードと、前記サーバ装置上で格納された日本語文書テキストに対応付けられた重み付きキーワードとを比較し、前記入力されたキーワードと少なくとも部分的に一致するキーワードを識別し、前記日本語テキストについて、前記入力された自由キーワードに付加された第１の重みと前記一致するキーワードに付加された第２の重みとの積を総和して得られた一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストであり、前記一致度は、前記日本語文書テキスト内での前記重み付きキーワードの出現頻度のみに基づいて、算出されることを特徴とする文書インデキシング処理をコンピュータに実行させるための方法が提供される。

【００４５】

本発明の他の特徴によれば、文書インデキシング処理をコンピュータに実行させるための文書インデキシングプログラムであって、該プログラムは、前記コンピュータに、入力された日本語文書テキストから、該テキストを構成する文字の文字種別を文字コードによって識別することにより、漢字文字列及びカタカナ文字列をそれぞれ抽出する処理と、抽出された前記漢字文字列及び前記カタカナ文字列のうち、２文字以上連続する文字列の出現頻度をカウントする処理と、前記出現頻度がカウントされた文字列のうち、前記入力された日本語文書テキスト内で第１の所定比率或いは所定出現回数以上の出現頻度を有する漢字文字列を、前記日本語文書テキスト内で第２の所定比率或いは所定出現回数以上の出現頻度を有するカタカナ文字列を、それぞれキーワードとして得る処理と、前記キーワードのそれぞれについて、その出現頻度から、前記日本語文書テキスト内で同じ文字種別に属するキーワードについて算出された最小出現頻度を基準として、得られる値を重みとして算出する処理と、前記重みを前記キーワードに付加して得られる重み付きキーワードと前記入力された日本語文書テキストとを対応付ける処理と、前記対応付けられた重み付きキーワード及び前記日本語文書テキストとを格納する処理とを含む処理を実行させるためのものであることを特徴とする文書インデキシングプログラムが提供される。

【００４６】

本発明の他の特徴によれば、文書インデキシング処理をコンピュータに実行させるための文書インデキシングプログラムであって、該プログラムは、前記コンピュータに、自由キーワードを入力し、入力された自由キーワードをサーバ装置に送信する処理と、入力された自由キーワードを前記サーバ装置に送信する処理と、入力された自由キーワードの送信に応答して、前記サーバ装置から、日本語文書テキストを受信し、検索結果として表示出力する処理とを含む処理を実行させるためのものであり、前記受信される日本語文書テ

キストは、前記入力された自由キーワードと、前記サーバ装置上で格納された日本語文書テキストに対応付けられた重み付きキーワードとを比較し、前記入力されたキーワードと少なくとも部分的に一致するキーワードを識別し、前記日本語テキストについて、前記入力された自由キーワードに付加された第１の重みと前記一致するキーワードに付加された第２の重みとの積を総和して得られた一致度が最大になるか又は前記一致度が所定値以上である日本語文書テキストであり、

前記一致度は、前記日本語文書テキスト内での前記重み付きキーワードの出現頻度のみに基づいて、算出されることを特徴とする文書インデキシングプログラムが提供される。

【発明の効果】

【００４７】

本発明によれば、文章テキストデータの登録時に、文書テキストデータから重要キーワードを文字コードのみに基づいて判別することにより自動的にインデキシングを実行する。このため、予め登録者によるキーワード付与や辞書登録を要することがなく、またこの辞書を用いた意味認識、形態素解析等の処理を要することがない。従って、文書テキストデータに簡易且つ自動的にインデキシングすることができ、登録された文書テキストデータの利用者による検索が容易化する。特に、すでに蓄積されている大量の文書テキストデータに自動的にインデキシングすることが可能となるので、既存文書データの再利用に資する。さらに、文書の意味認識を必要としないので、新たな語彙が生じた場合にあっても、本発明に係るインデキシングシステムをメンテナンスする必要は生じ得ないという利点が得られる。

【００４８】

さらに、本発明によれば、抽出されたキーワードについてカウントされた出現頻度のみから得られた「重み」を、当該キーワードの重要度を示す指標として、当該キーワードに付加して、「重み付きキーワード」とし、この「重み付きキーワード」に基づいて、すなわち、キーワード自動抽出処理において抽出されたキーワード及びその出現頻度のみに基づいて、入力キーワードと文書テキストとの一致度を算出するので、自由キーワードによる簡易且つ高精度の文書テキスト検索が可能となる。

【００４９】

さらに、キーワード自動抽出処理において抽出されたキーワード及びその出現頻度のみから得られる一致度のみに基づいて、文書間の一致度を算出し、この一致度の有向性に基づいて文書テキスト同士の関連性を評価するので、多数の文書テキストを、簡易且つ高精度で、相互に高い関連性を有する文書テキスト群に自動的に分類することが可能となる。殊に、異なるカテゴリーに属する文書テキストを保有し、文書間の均質性を欠く文書データベースをキーワード抽出対象とした場合にあっても、キーワード抽出及びこれを用いた分類の精度が低下することがないという利点が得られる。

【００５０】

従って、利用者による文書データ検索における利便性が向上するとともに、蓄積された大量の既存文書データの再利用が促進される。

【発明を実施するための最良の形態】

【００５１】

以下、図面を参照して、本発明の実施の形態を説明する。

【００５２】

第１の実施形態

＜第１の実施形態の構成＞

図１は、本発明の第１の実施形態に係る文書管理サーバ１及びクライアントコンピュータ２を具備する、文書テキストに自動的に重み付きキーワードを付与する文書インデキシングシステムの一構成例を示す。

【００５３】

文書管理サーバ１は、インデキシングされるべき検索対象の文書テキストデータを格納する外部記憶装置である文書データベース１１と、インデキシングされるべき検索対象の

文書テキストデータを入力する文書入力部１２と、入力された文書テキストデータからキーワードを自動抽出し、該キーワードごとにその「重み」（抽出されたキーワードの出現頻度から一意に算出される値）を対応付けるキーワード自動抽出部１３と、インデキシングされた文書テキストデータと抽出された重み付きキーワードとの対応付け及び記憶保持を管理する重み付きキーワード管理部１４と、キーワードが付与された文書データを外部記憶装置であるキーワード付与文書データベース１６に格納すると共に、入力された管理キーワードと一致するキーワードが付与された文書テキストデータをキーワード付与文書データベース１６から検索する文書格納部１５と、クライアントコンピュータ２からのキーワード入力を受け付け、重み付きキーワード管理部１４を介して入力キーワードに一致する重み付きキーワードを含む文書テキストのそれぞれについて、後述する計算方法により得られる「一致度」を算出し、文書格納部１５を介して、最も「一致度」の大きい１つ又は複数の文書テキストを読み出すよう文書格納部１５に指示するキーワード一致度算出部１７と、検索された文書テキストデータをクライアントコンピュータ２に出力制御する文書送信管理部１８とを具備する。なお、本明細書において「重み付きキーワード」とは、キーワード自動抽出部１３により入力文書テキストデータから抽出され、キーワード付与文書データベース１６に該文書テキストデータと対応付けて記憶されるキーワードであって、該キーワードの文書テキスト中での出現頻度から後述する算出方法により一意に得られる値である「重み」が付加されたキーワードを意味する。また、当然ながら、本実施形態は、文書入力部１２に入力される入力手段を文書データベース１１に限定するものではない。この入力手段は、文書データベース１１の他、直接文書データの入力を受け付ける手段の他、例えばＣＤ－ＲＯＭ、ＤＶＤ、ＭＯ等任意の外部記録媒体に記録された文書データを読み込み、入力として受け付けてもよい。

【００５４】

キーワード自動抽出部１３は、より詳細には、入力文書テキストデータの各文字の文字コードを文字種別ごと分類するコード別文字分類部１３１と、漢字に分類された文字列から連続する漢字文字列の出現頻度をカウントする漢字ラン出願頻度カウンタ１３２と、連続する漢字文字列のそれぞれの出現頻度に基づいて漢字キーワードを抽出する漢字キーワード抽出部１３３と、カタカナに分類された文字列から連続するカタカナ文字列の出現頻度をカウントするカタカナラン出現頻度カウンタ１３４と、連続するカタカナ文字列のそれぞれの出現頻度に基づいてカタカナキーワードを抽出するカタカナキーワード抽出部１３５と、アルファベットに分類された文字列から連続するアルファベット文字列の出現頻度をカウントするアルファベットラン出現頻度カウンタ１３７と、連続するアルファベット文字列のそれぞれの出現頻度に基づいてアルファベットキーワードを抽出するアルファベットキーワード抽出部１３８と、抽出された漢字キーワード、カタカナキーワード及びアルファベットキーワードを入力文書テキストデータに対応付けて重み付きキーワード管理部１４に出力する文書・キーワード群対応付け部１３６とを具備する。

【００５５】

文書管理サーバ１と、クライアントコンピュータ２とは、例えばインターネットやＬＡＮなどのネットワーク３を介して、相互に接続される。或いは代替的に、図１における文書管理サーバ１とクライアントコンピュータ２との機能を一体とし、１つのコンピュータに実装してもよい。

【００５６】

一方、クライアントコンピュータ２は、入力装置からの自由キーワード、或いは一覧提示された重み付きキーワードからのキーワードの選択入力を受け付けるキーワード入力部２３と、入力自由キーワードをキーワード一致度算出部１７に送出するキーワード送付管理部２４と、文書送信管理部１８から受信される自由キーワードに対応付けられた文書テキストデータを受信する文書受信管理部２５と、受信された文書テキストデータを利用者に提示するディスプレイ部２６とを具備する。クライアントコンピュータ２は、さらに、文書管理サーバ１上のキーワード付与文書データベース１６に格納されている文書テキストに対応付けられた重み付きキーワードの一覧を受信し、クライアントコンピュータ２上

に提示制御する重みつつきキーワード一覧提示部を具備してもよい。

【００５７】

なお、本実施形態は、利用者がキーワード入力部２３を介して行なう入力方式及び手段を特に限定するものではない。これら入力手段は、利用者からの直接入力を受け付けてもよく、あるいは例えばＵＳＢメモリやＩＣカードなどに例示される外部記録媒体に記憶されたシーケンスを入力として受け付けてもよく、また任意のファイルとして予め格納されたデータを入力として受け付けてもよい。

【００５８】

さらに、図１においては、クライアントコンピュータ２において、自由キーワードの入力を受け付け、文書管理サーバ１に送信し、検索された文書テキストデータを、同じクライアントコンピュータ２において受信及び提示する構成が図示されるが、これに替えて、自由キーワード入力を受け付け、文書管理サーバ１に送信する要求入力端末と、文書管理サーバ１から送信される文書テキストデータを受信及び表示出力する文書提示端末とが異なるコンピュータ装置であってもよい。要求入力端末としては、例えば、携帯電話や携帯情報端末（ＰＤＡ）を用いて入力を受け付けてよく、あるいはネットワーク接続可能なＩＣカードリーダなどを用いてＩＣカードからのシーケンスを受け付け、他のクライアントコンピュータにおいて、文章テキストデータを受信して表示出力してもよい。

【００５９】

＜第１の実施形態における重み付きキーワード自動抽出処理＞

１．キーワード抽出処理詳細

図１を参照し、文書管理サーバ１により管理されるべき文書テキストデータは、好適には文書テキストデータの登録時に、まず文書管理サーバ１の文書入力部１２に入力され、キーワード自動抽出部１３に受け渡される。キーワード自動抽出部１３内のコード別文字分類部１３１において、まず入力文書中の漢字のみが、漢字の連続性を維持したまま抽出される。

【００６０】

第１の実施形態において、漢字、及び後述するカタカナは、いずれも文字コードのレベルで識別される。このため、文字種別の識別のために特別な処理を必要としない。文字には、それぞれ対応する文字コードが定義されており、例えば日本語を扱う上での文字コードの規格には、ＪＩＳやＵｎｉｃｏｄｅなど複数存在する。どの文字コード規格においても、漢字、カタカナ、ひらかな、アルファベットはそれぞれ特定のコード領域内にまとまった状態で収納されている。例えば、Ｕｎｉｃｏｄｅの場合、漢字（ＣＪＫ Ｕｎｉｆｉｅｄ Ｉｄｅｏｇｒａｐｈｓ）はＵ＋４Ｅ００～Ｕ＋９ＦＢＦ、カタカナはＵ＋３０Ａ０～Ｕ＋３０ＦＦ、アルファベット（Ｃ０ Ｃｏｎｔｒｏｌｓ ａｎｄ Ｂａｓｉｃ Ｌａｔｉｎ）はＵ＋００００～Ｕ＋００７Ｆのコード領域で定義されるため、入力文字がこれらのコード領域のいずれに該当するかだけを識別すれば足りる。

【００６１】

各文字の文字コードを識別して、現在の文字種別が変化するごとに、文字列を区切って切り出すことにより、漢字の連続性を維持したまま抽出された文字列は、漢字ラン出現頻度カウンタ１３２に入力され、この漢字ラン出現頻度カウンタ１３２は、入力文書テキスト全体に対する連続する漢字の組み合わせ、すなわち連続する漢字文字列の出現頻度をカウントする。本明細書において、このような連続する漢字文字列を、「漢字ラン」と称する。例えば、「彼は病気勝ちだったにもかかわらず、前向きに生き、トランジスタ工学の大いなる発展と、トランジスタ産業の育成に大きな功績を上げた。」という文書がキーワード自動抽出部１３に入力されたと仮定すると、漢字ラン出現頻度カウンタ１３２は、「彼」、「病気勝」、「前向」、「生」、「工学」、「進展」、「産業」、「育成」、「大」、「功績」、「上」がそれぞれ漢字ランである。このようなランに属する文字数を、以下「ラン長」と称する。上記の例では、ラン「大」のみが出現頻度２であり、他のランはすべて出現頻度１である。漢字ラン出現頻度カウンタ１３２に入力される漢字が、１字で孤立したもの、すなわち文書テキスト中で前後には漢字以外の文字種別の字が配列されて

いる漢字は、ラン長１のランとして、同種の文字種別に属する連続する文字はその最大長の組み合わせを１つのランとして取り出す。

【００６２】

すなわち、ランとは、連続する同一種類に属する文字列の最大長のもので、$C_i$と呼び、漢字ランをK、カタカナランをH、アルファベットランをRとすると、

$$C_i \in \{K, H, R\} \quad (1)$$

であり、各ランの文字数をラン長と呼ぶ。漢字ランは単純に連続する漢字列であるが、カタカナランではスペース、なか点「・」、ー（長音記号）、半角・全角の区別は無視してラン長を得る。アルファベットランも同様に、なか点、スペース、大文字・小文字の区別、半角・全角の区別を無視する。好適には、これらのランのうち、漢字とカタカナランはラン長２以上のもの、アルファベットランはラン長３以上のもののみの出現頻度がカウントされる。

【００６３】

漢字ラン出現頻度カウンタ１３２は、これらのランの出現頻度をカウントする。より詳細には、漢字ラン出現頻度カウンタ１３２は、一時記憶領域においてテーブル等を備え、新たなランを取り出すと、このランがテーブルに存在するか否かを判定し、存在すれば当該ランのカウントを＋１とし、存在しなければ新たなランとしてテーブルに新たなエントリーを追加する。

【００６４】

次に、漢字キーワード抽出部１３３は、漢字ラン出現頻度カウンタ１３２が出力する漢字ランの中から、ラン長２以上の漢字ランのみを取り出して、その出現頻度を再カウントし、その出現頻度がラン長２以上の漢字ラン全数に対して例えば１．５％を超える漢字ランを漢字キーワードとして抽出する。或いは代替的に、この入力文書テキスト中の漢字ラン全数に対するあるランの出現頻度の閾値を可変に構成し、例えば１％から２％の範囲に設定してもよい。さらに代替的に、入力文書テキストの文書種別（例えば、小説、学術論文、口頭による演説・講演等）に応じて、出現頻度の閾値を再帰的に最適化し、１％未満或いは２％を超える範囲に設定可能としてもよい。

【００６５】

この出現頻度を、漢字キーワード抽出の閾値として利用する場合には、代替的に、入力日本語文字テキスト中の漢字文字列の総数、日本語文字テキスト文書全体の文字数、または日本語テキスト全体の中の漢字の総数のいずれかを分母とし、これに対しての当該漢字文字列の出現比率を算出してもよい。

【００６６】

２．キーワードの重み算出処理詳細

第１の実施形態においては、抽出されたキーワードについてカウントされた出現頻度が、当該キーワードの重要度の指標となり得るとの知見に基づき、キーワード抽出後に、当該キーワードの「重み」を算出する。抽出されたキーワードｋｗ直後に、括弧付きで頻度（ｎ）データを付加し、ｋｗ（ｎ）とする。この（ｎ）データは、同じ文字種別のキーワードとして抽出されたもののうち、最も出現頻度が低いものをｋｗ（１）と表現し、これより出現頻度が高いキーワードには、括弧内に、実際の出現頻度－（最小の出現頻度－１）を付加する。このキーワードｋｗ（ｎ）を、重み付きキーワードといい、特に断らない限り、本明細書において単に「キーワード」と言った場合には、「重み付きキーワード」を示すものとする。

【００６７】

カタカナラン出現頻度カウンタ１３４は、文書テキストデータ中のカタカナラン、すなわち連続するカタカナのみの文字列の、それぞれの出現頻度を、漢字ラン出現頻度カウンタ１３２と同様の手法でカウントする。上記の文書例では、「トランジスタ」のみがカタカタランであり、その出現頻度は２である。

【００６８】

カタカナキーワード抽出部１３５は、カタカナラン出現頻度カウンタ１３４が出力する

カタカナランの中から、ラン長２以上のカタカナラン全数に対して出現頻度が例えば５％を超えるカタカナランをカタカナキーワードとして抽出する。或いは代替的に、この入力文書テキスト中のカタカナラン全数に対するあるランの出現頻度の閾値を可変に構成し、例えば３％から７％の範囲に設定してもよい。さらに代替的に、入力文書テキストの文書種別（例えば、小説、学術論文、口頭による演説・講演等）に応じて、出現頻度の閾値を再帰的に最適化し、３％未満或いは７％を超える範囲に設定可能としてもよい。

【００６９】

この出現頻度を、カタカナキーワード抽出の閾値として利用する場合には、代替的に、入力日本語文字テキスト中のカタカナ文字列の総数、日本語文字テキスト文書全体の文字数、または日本語テキスト全体の中のカタカナの総数のいずれかを分母とし、これに対しての当該カタカナ文字列の出現比率を算出してよい。後述するアルファベットの場合も同様に出現比率を算出することができる。

【００７０】

アルファベットラン出現頻度カウンタ１３７は、文書テキストデータ中のアルファベットラン、すなわち連続するアルファベットのみの文字列の、それぞれの出現頻度を、漢字ラン出現頻度カウンタ１３２と同様の手法でカウントする。

【００７１】

アルファベットキーワード抽出部１３８は、アルファベットラン出現頻度カウンタ１３７が出力するアルファベットランの中から、ラン長２以上のアルファベットラン全数に対して、カタカナランと同様、出現頻度が例えば５％を超えるアルファベットランをアルファベットキーワードとして抽出する。或いは代替的に、この入力文書テキスト中のアルファベットラン全数に対するあるランの出現頻度の閾値を可変に構成し、例えば３％から７％の範囲に設定してもよい。さらに代替的に、入力文書テキストの文書種別（例えば、小説、学術論文、口頭による演説・講演等）に応じて、出現頻度の閾値を再帰的に最適化し、３％未満或いは７％を超える範囲に設定可能としてもよい。

【００７２】

なお、カタカナランについては、カタカナ文字列に含まれる「・」（なか点）、「ー」（長音記号）、スペースとカタカナ文字列の最後の「ー」（長音記号）は無視するものとする。アルファベットランについては、アルファベット文字列に挟まれる「・」（なか点）、スペースは無視し、大文字と小文字は同一文字と見做すものとする。

【００７３】

なお、キーワードとして抽出するか否かは、好適には、当該文字種別に属するランの全数に対する当該ランの出現頻度の比率により決定され、例えば、好適には、漢字は１．５％以上、カタカナは５％以上、アルファベットは５％以上に、キーワード抽出の閾値が設定されてよい。この閾値を大きくすれば抽出されるキーワード数が減少し、逆に閾値を小さくすれば抽出されるキーワード数が増加する。例えば、漢字の場合、ラン長２であり、且つ出現数１の漢字ランが多いため、閾値を小さく設定することが好ましい。カタカナの場合、カタカナは１つの単語として抽出されやすく、一般的にはそのランの種類も少ないので、５％と漢字の場合より閾値を大きく設定することが好ましい。アルファベットの場合、カタカナと同様、雑音が少なく、１つの単語として抽出されやすいため、５％と設定することが好ましく、殊に例えば学術的文書の場合に有効である。しかしながら、これらの閾値は、キーワードの具体的用途や、文書テキストの種別等に応じて可変であり、キーワード抽出の閾値にいかなる具体的数値を設定するかは、本発明の要旨の画定には影響しない。

【００７４】

次に、文書・キーワード群対応付け部１３６は、入力された文書テキストデータと、漢字キーワード抽出部１３３、カタカナキーワード抽出部１３５及びアルファベットキーワード抽出部１３８により抽出された重み付けキーワード群とを対応付ける。

【００７５】

重み付けキーワード管理部１４は、入力された文書テキストデータに対応付けられた重

み付けキーワード群を、文書格納部１５は、入力された文書テキストデータ自体を、それぞれキーワード付与文書データベース１６に格納する。或いは代替的に、重み付けキーワード管理部１４は、キーワード付与文書データベース１６以外の別の記憶媒体に、好適にはより高速なアクセス速度が保証される一時記憶媒体等の記憶媒体に、文書テキストデータに対応付けられた重み付けキーワード群を保持記憶してもよい。

【００７６】

図３Ａ及び図３Ｂは、例示的に、朝日新聞の社説における漢字ラン及びカタカナランの出現頻度を示す。図３Ａにおいて、０５０７０６＿１（２００５年７月６日の社説）の例では、「首相（出現頻度７、出現比率４．７％）」、「反対（出現頻度６、出現比率４．０％）」、「党内（出現頻度５、出現比率３．３％）」、「法案（出現頻度３、出現比率２．０％）」、「派閥（出現頻度５、出現比率２．０％）」、「執行部（出現頻度３、出現比率２．０％）」「自民党（出現頻度３、出現比率２．０％）」、などが漢字キーワードとして抽出される。図３Ｂにおいて、０５０７０９＿２（２００５年７月９日の社説）の例では、漢字キーワードとして「組織（出現頻度３、出現比率２．６％）」「世界（出現頻度３、出現比率２．６％）」、「犯行（出現頻度２、出現比率１．７％）」、「宗教（出現頻度２、出現比率１．７％）」、「国際（出現頻度２、出現比率１．７％）」、「攻撃（出現頻度２、出現比率１．７％）」、「寛容（出現頻度２、出現比率１．７％）」、が抽出され、一方「テロ（出現頻度１７、出現比率３８．６％）」、「イスラム（出現頻度６、出現比率１３．６％）」、「ロンドン（出現頻度３、出現比率６．８％）」、「イラク（出現頻度３、出現比率６．８％）」、「イラク（出現頻度２、出現比率４．５％）」「アルカイダ（出現頻度２、出現比率４．５％）」などのカタカナキーワードも多く抽出されることが理解される。

【００７７】

図２は、朝日新聞の社説のうち２００５年７月６日から１０月１７までの２００件の社説を入力文書テキストとして、これらのタイトルを除く本文テキスト中の漢字ラン、カタカナラン、及びアルファベットランの出現頻度をカウントした結果を示す。社説は、総文字数が１，５００文字程度の比較的短い文書であり、漢字ランについては出現頻度２以上又は出現比率１％以上を閾値とすれば適当であり、カタカナラン及びアルファベットランについては出現するものすべて（出現頻度１以上）を取り出すのが適当であることが理解される。なお、図２中出現率とは、出現頻度を全文字数で除算して得られる値を％表示したものである。より長文の文書テキスト、例えば、論文や特許明細書の場合には、出現頻度に加え、或いはこれに替えて、出現頻度の比率を閾値として用い、例えばカタカナラン及びアルファベットランについては、出現頻度の比率５％以上を閾値としてよい。

【００７８】

変形例として、キーワード抽出のための、出現頻度の閾値の初期値を、上記の値より大きく設定し、該当するランが存在するにもかかわらずキーワードとして抽出されない場合には、キーワードとして抽出されるまでこの閾値を小さくなるよう調整してもよい。さらに、小さく調整した後もキーワードが抽出されない場合には、出現頻度の比率に替えて、或いは出現頻度の比率と共に、出現頻度の上位から所定番目までのランをキーワードとして抽出するよう構成してもよい。

【００７９】

＜第１の実施形態における自由キーワードによる文書テキストデータ検索処理＞

１．自由キーワード送信処理

図１を参照し、文書を検索しようとする利用者は、クライアントコンピュータ２のキーワード入力部２３に、自由キーワードを入力する。このキーワード入力は、例えばキーボード或いはマウス等の任意のポインティングデバイスを使用して行なわれてよい。この入力されるキーワードは、重み付きキーワードであってもよく、代替的に重みが付加されないキーワードであってもよい。入力された自由キーワードは、ディスプレイ部２６を介してディスプレイモニター上に表示出力されると共に、キーワード送付管理部２４に受け渡される。文書管理サーバ１において、キーワード送付管理部２４は、受け渡された自由キ

ーワードを、ネットワーク３を介して文書管理サーバ１のキーワード一致算出部１７に送出する。キーワード一致算出部１７は、クライアントコンピュータ２から受信されたキーワードと、重み付きキーワード管理部１４が管理する重み付きキーワードとの一致度を後述のとおり順次算出し、一致度が高い重み付きキーワードを選別する。選別された重み付きキーワードに対応付けられた文書テキストは、文書格納部１５を介してキーワード付与データベース１６から読み出され、文書送信管理部１８により、クライアントコンピュータ２の文書受信管理部２５に送出される。クライアントコンピュータ２において、文書受信管理部２５により受信されたフリーワード検索結果である１つ又は複数の文書テキストは、ディスプレイ部２６を介して利用者に提示出力される。

【００８０】

２．一致度算出処理

図４において、例示的に、新聞社説を入力文書テキストとした場合に、出現頻度２以上の漢字ラン、出現頻度１以上のカタカナラン及びアルファベットランが、重み付きキーワードとして抽出されている。図４の括弧内の数値は、漢字キーワードの重みは、出現頻度２の漢字ラン（抽出されるキーワードのうち最小の出現頻度の漢字ラン）を重み１とし、出現頻度３以上の漢字ランをその出現頻度から１を減じた値を重みとして、示されている。カタカナラン及びアルファベットランの重みは、出現頻度そのままを重みとして示されている。

【００８１】

ここで、利用者が、自由キーワード「総選挙（２）、投票（３）、政党（１）、郵政民営化（２）、有権者（１）、政策（１）」を投入したと仮定すると、キーワードの一致度は、次のとおり算出される。投入された自由キーワードＸに一致するキーワードとは、

ａ）Ｘに完全一致するキーワード、又は、

ｂ）Ｘを含むキーワードか或いはＸが相手のキーワードの１つを包含している場合の当該キーワード、又は

ｃ）Ｘの長さ２以上の連続するランを含むキーワード、のいずれかである。

【００８２】

ｂ）の前段の場合、Ｘを含む最も短い相手のキーワードを、ｂ）の後段の場合、Ｘが包含する最も長い相手のキーワードを、それぞれＸに一致するキーワードと見なす。候補が複数得られた場合には、重みが最も高いものを選択してよい。ｃ）の場合、Ｘの最も長い部分を共有するキーワードを、Ｘに一致するキーワードと見なす。

【００８３】

すなわち、文書ＡのキーワードＡｋｗＸ（$n_k$）が、文書ＢのキーワードＢｋｗＸ（$n_k$）に「一致する」とは、次の３通りのいずれかの場合である（一致度算出のための優先度の順に示す）。文書Ａを上記の自由キーワードとした場合も同様である。

【００８４】

ａ）Ｘに完全に一致するキーワードが文書Ｂにある。

【００８５】

$$AkwX(n_k) = BkwX(n_k) \qquad (2)$$

ｂ）文書ＡのキーワードＡｋｗＸ（$n_k$）を含むキーワードが文書ＢのキーワードＢ$w_1$ｋｗＸ$w_2$にあるか、又は文書ＡのキーワードＡ$w_1$ｋｗＸ$w_2$（$n_k$）Ｘが、ＢｋｗＸ（$n_k$）のキーワードの１つを包含している。ここで、$w_1$及び$w_2$は、１以上のラン長を持つ文字列である。前者の場合、ＡｋｗＸ（$n_k$）を含む最も短いＢのキーワードを、後者の場合、Ａ$w_1$ｋｗＸ$w_2$（$n_k$）が包含する最も長いＢのキーワードを、ＡｋｗＸ（$n_k$）に一致するＢのキーワードとする。候補が複数あるときは、重みが最も高いものとする。

【００８６】

$$AkwX(n_k) \in Bw_1kwXw_2(nk) \qquad (3)$$

$$Aw_1kwXw_2(n_k) \supset BkwX(n_k) \qquad (4)$$

ｃ）Ａ文書のキーワードｋｗの長さ２（アルファベットの場合は３）以上の連続する部

分$w_1 k w X w_2$を含むキーワードが文書Ｂにある。このとき文書Ａの$A w_1 k w X w_2$の最も長い部分を共有する文書Ｂのキーワード$B w_1 k w X w_2$を、ｋｗＸに一致するキーワードとする。$w_1$及び$w_2$は、１以上のラン長を持つ。

【００８７】

$$A w_1 k w X w_2 (n_k) \in B w_1 k w X w_2 (n_k) \quad (5)$$

これらのいずれかに該当するキーワードが文書Ｂにない場合、Ｘに一致するキーワードは文書Ｂにないものとする。

【００８８】

次に、文書Ａの文書Ｂに対する「一致度」とは、

ａ）通常の一致度：文書Ｂに一致するキーワードを有する文書Ａのキーワード数、又は、

ｂ）重み付き一致度ｃｎ：文書Ｂの一致するキーワードの重み$n_k m$と対応する文書Ａの重み$n_k m$の積を、文書Ａのキーワード全てで総和して得られる数、のいずれかである。

【数１】

$$Cn = n_k n \times n_k m \quad (6)$$

$$\sum_{k=1}^{n} cn_k = cn_1 + cn_2 + cn_3 + \cdots + cn_n \quad (7)$$

【００８９】

なお、本明細書において、特に断らない場合は、「重み付き一致度」を単に「一致度」という。

【００９０】

図４に戻り、入力自由キーワードＸに一致するキーワードが、上記のとおり得られた後、入力自由キーワードＸの、検索対象文書テキストに対する一致度とは、例えば、「相手の一致するキーワードの重みと、対応する入力自由キーワードＸの重みの積を、投入されたキーワードすべてについて総和して得られる数値」であり、検索対象文書テキストごとに算出される。図４の第１行目の例においては、「有権者」と「政策」とが、入力自由キーワードと完全一致し、「郵政民営化」が、「郵政民営化法案」に包含されている。投入された自由キーワードに重みが付加されていない場合には、すべての自由キーワードの重みを１と見なし、一致度は、１×３（「有権者」）＋１×２（「政策」）＋１×１（「郵政民営化」）＝６と算出される。重みつき自由キーワードが投入された場合には、「郵政民営化」が２×１となるので、１×３（「有権者」）＋１×２（「政策」）＋２×１（「郵政民営化」）＝７と算出される。図４の２行目以降の例においては、同様に、重み付き自由キーワードが投入された場合には、３行目の例が、一致度３、４行目の例が一致度７、５行目の例が一致度６と算出され、他の例は一致度０と算出される。投入された自由キーワードに重みが付加されていない場合には、３行目の例が一致度２、４行目の例が一致度６、５行目の例が一致度３と算出される。従って、一致度が最も大きいものは、自由キーワードに重みが付加されているか否かにかかわらず、１行目の例及び４行目の例となり、この２つの文書テキストが、キーワード付与文書データベース１６から読み出されて、クライアントコンピュータ２の文書受信管理部２５に送出される。

【００９１】

上記の例では、最大の一致度が算出された文書テキストはすべてキーワード付与文書データベース１６から読み出されたが、読み出される文書テキストが多すぎる場合、クライアントコンピュータ２に送出されるデータ量が膨大となることが懸念され、この場合、変形例として、検索結果の候補文書テキストの要約のみを、まずクライアントコンピュータ２に送出し、ディスプレイ部２６を介して利用者に提示してよい。要約の作成は、投入された自由キーワードは、キーワードを投入した利用者の関心の度合いを示すものであると

の知見に従い、投入自由キーワードと一致するキーワードを含む文章だけを取り出すことにより行なう。図４の例においては、１行目の例では、一致したキーワードが存在し、そのうち、「有権者」の出現回数は４回、「政策」の出現回数は３回、「郵政民営化」の出現回数は２回であるため、要約として抽出される文章の数は、最大９であり、１つの文章にこれらのキーワードが重複して記述されている場合には、要約として抽出される文章の数は減少することが理解される。この変形例においては、このように作成された要約がまずクライアントコンピュータ２に送出されて、ディスプレイ部２６を介して利用者に提示され、利用者が提示された要約を閲覧することにより、所望の文書テキストを選択し、選択された文書テキストの送信要求が文書管理サーバ１に送信されて、キーワード付与文書データベース１６から選択された文書テキストの本文が読み出され、クライアントコンピュータ２に送出される。これにより、文書管理サーバ１からクライアントコンピュータ２に対して送出されるデータの通信コストを抑制することができる。

【００９２】

なお、第１の実施形態では、一例として、まず漢字キーワードを抽出し、次いでカタカナキーワード、アルファベットキーワードを抽出するキーワード自動抽出処理を開示したが、代替的に、これらの処理を並列に同時実行してもよく、カタカナキーワード抽出処理を先行して実行してもよく、或いはこれらのうちいずれか１種類を必要に応じて実行してもよい。さらに、アルファベットキーワード抽出処理を、漢字キーワード抽出処理及びカタカナキーワード抽出処理と共に実行してもよく、この場合いずれの文字種別のキーワード抽出から順次実行してもよく、或いはすべてのキーワード抽出処理を並列に同時実行してもよい。これらの変形例のいずれも本発明の開示に含まれることは言うまでもない。

【００９３】

＜本実施形態に係る文書インデキシングシステムのハードウエア構成＞

図１３は、第１の実施形態に係る文書管理サーバ１及び／又はクライアントコンピュータ２のハードウエア構成の一例を示すブロック図である。図１１に示されるコンピュータ装置１１０である文書管理サーバ１及び／又はクライアントコンピュータ２において、ＣＰＵ１１１は、ＲＯＭ１１４および／またはハードディスクドライブ１１６に格納されたプログラムに従い、ＲＡＭ１１５を一次記憶用ワークメモリとして利用して、システム全体を制御する。さらに、ＣＰＵ１１１は、マウス１１２ａまたはキーボード１１２を介して入力される利用者の指示に従い、ハードディスクドライブ１１６に格納されたプログラムに基づき、第１の実施形態に係る文書インデキシング処理、フリーワード文書検索処理を実行する。ディスプレイインタフェイス１１３には、ＣＲＴやＬＣＤなどのディスプレイが接続され、ＣＰＵ１１１が実行する文書インデキシング処理、フリーワード文書検索処理の入力待ち受け画面、処理経過や処理結果、検索結果である文章テキストデータなどが表示される。リムーバブルメディアドライブ１１７は、主に、リムーバブルメディアからハードディスクドライブ１１６へファイルを書き込んだり、ハードディスクドライブ１１６から読み出したファイルをリムーバブルメディアへ書き込む場合に利用される。リムーバブルメディアとしては、フロッピディスク(ＦＤ)、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＣＤ－Ｒ／Ｗ、ＤＶＤ－ＲＯＭ、ＤＶＤ－Ｒ、ＤＶＤ－Ｒ／Ｗ、ＤＶＤ－ＲＡＭやＭＯ、あるいはメモリカード、ＣＦカード、スマートメディア、ＳＤカード、メモリスティックなどが利用可能である。

【００９４】

プリンタインタフェイス１１８には、レーザビームプリンタやインクジェットプリンタなどのプリンタが接続される。ネットワークインタフェイス１１９は、コンピュータ装置をネットワークへ接続するためのインターフェースである。

【００９５】

なお、第１の実施形態に係る文書管理サーバ１及び／又はクライアントコンピュータ２における入力手段は、マウス１１２ａあるいはキーボード１１２に限定されることなく、任意のポインティングデバイス、例えばトラックボール、トラックパッド、タブレットなどを適宜用いることができる。携帯情報端末を上記各実施形態に係るクライアントコンピ

ュータ２として用いる場合には、入力部をボタンやモードダイヤル等で構成してもよい。
【００９６】
また、図１１に示した第１の実施形態に係る文書管理サーバ１及び／又はクライアントコンピュータ２のハードウエア構成は一例に過ぎず、その他の任意のハードウエア構成を用いることができることはいうまでもない。
【００９７】
殊に、第１の実施形態に係る文書インデキシング処理、フリーワード文書検索処理の全部又は一部は、上記コンピュータ端末装置１００あるいはＰＤＡ等の携帯情報端末装置等によって実現されてもよく、コンピュータ端末装置等とサーバ装置とをＢｌｕｅｔｏｏｔｈ（登録商標）等の無線、あるいはインターネット（ＴＣＰ／ＩＰ）、公共電話網（ＰＳＴＮ）、統合サービス・ディジタル網（ＩＳＤＮ）等の有線通信回線で相互接続した、インターネットあるいは任意の周知のローカル・エリア・ネットワーク（ＬＡＮ）またはワイド・エリア・ネットワーク（ＷＡＮ）からなるネットワークシステムによってコンテンツ提示処理が実現されてもよい。例えば、ＰＤＡ等の携帯情報端末装置が自由キーワードの検索要求を文書管理サーバ１に対して送信し、文書管理サーバ１は、所定の或いは要求された識別子のクライアントコンピュータ２に対して、文書テキストデータを配信してもよい。
【００９８】
以上のとおり、第１の実施形態によれば、文章テキストデータの登録時に、文書テキストデータから重要キーワードを文字コードのみに基づいて判別することにより自動的にインデキシングを実行する。このため、予め登録者によるキーワード付与や辞書登録を要することがなく、またこの辞書を用いた意味認識、形態素解析等の処理を要することがない。従って、文書テキストデータに簡易且つ自動的にインデキシングすることができ、登録された文書テキストデータの利用者による検索が容易化する。特に、すでに蓄積されている大量の文書テキストデータに自動的にインデキシングすることが可能となるので、既存文書データの再利用に資する。さらに、文書の意味認識を必要としないので、新たな語彙が生じた場合にあっても、本発明に係るインデキシングシステムをメンテナンスする必要は生じ得ないという利点が得られる。
【００９９】
さらに、キーワード抽出の際にカウントされる出現頻度のみから得た「重み」をキーワードに付加して重み付きキーワードとし、このキーワードの「重み」を当該キーワードの重要度の指標と捉えて、これに基づき入力自由キーワードと文書テキストとの一致度を算出する。このため、フリーワードによる文書検索であっても、文書検索が高精度で行なえるという利点が得られる。
【０１００】
第２の実施形態
図５ないし図１２を参照して、本発明の第２の実施形態を、第１の実施形態と相違する点についてのみ説明する。第２の実施形態は、第１の実施形態により自動抽出された重み付きキーワードに基づいて、文書テキストを自動分類する。
【０１０１】
＜第２の実施形態の構成＞
図６は、本発明の第２の実施形態に係る文書管理サーバ１を具備する文書分類システムの一構成例を示す。
【０１０２】
文書管理サーバ１は、第１の実施形態と同様、インデキシングされるべき検索対象の文書テキストデータを入力する文書入力部１２と、入力された文書テキストデータからキーワードを自動抽出するキーワード自動抽出部１３と、インデキシングされた文書テキストデータと抽出された重み付きキーワードとの対応付け及び記憶保持を管理する重み付きキーワード管理部１４と、重み付きキーワードが付与された文書データを外部記憶装置であるキーワード付与文書データベース１６に格納する文書格納部１５と、文書間の一致度を

算出するキーワード一致算出部１７と、検索された文書テキストデータをクライアントコンピュータ２に出力制御する文書送信管理部１８とを具備する。或いは代替的に、キーワード自動抽出部１３を、文書キーワード抽出装置である別体のコンピュータに実装し、この文書キーワード抽出装置により抽出されたキーワード群と、これらに対応付けられた文書テキストとの対が、文書管理サーバ１の文書入力部１２に入力されるよう構成されてもよい。第２の実施形態において、文書分類システムは、さらに、第２の実施形態はさらに、第１の実施形態と同様のクライアントシステムを備えてよい。

【０１０３】

第２の実施形態に係る文書管理サーバ１は、さらに、所定値（以下に説明する「分類精度」）以下の一致度を０で置き換えた後、複数の文書間で算出された一致度をすべての文書テキストの組み合わせについて記述する一致度マトリクスを生成する一致度マトリクス生成部６３と、この一致度マトリクスを用いて、文書間に構成されるループを検出する有向ループ検出部６４と、文書間のすべての双方向連結成分を検出する双方向連結成分検出部６５と、検出された双方向連結成分ごとに、ループを構成するノード（文書テキスト）間のパスを検出し、制限付き双方向連結成分（部分グラフ）に分割する制限付き連結成分分割部６６と、分割された制限付き双方向連結成分（部分グラフ）に属する文書テキスト群を、１つの再分類に分類する細分類付与部６７とを具備する。

【０１０４】

なお、第２の実施形態に係るキーワード一致度算出部１７は、重み付きキーワード管理部１４が管理する重み付きキーワードを読み出し、読み出された重み付きキーワードのすべての組み合わせについて、一致するキーワードを検出し、キーワードの一致が検出された複数の文書間での一致度を、上記のとおり算出する。

【０１０５】

＜第２の実施形態における自動分類処理詳細＞

１．有向グラフの形成

図５は、例示的に、一致度マトリクス生成部６３が生成する一致度マトリクスを示す。図５において、新聞社説でビジネスのカテゴリーに分類される社説１５件（「ビ１」、「ビ２」、・・・、「ビ１５」）と、スポーツのカテゴリーに分類される社説８件（「ス１」、「ス２」、・・・「ス８」）とのすべての組み合わせについて、文書間一致度算出部６２が、上記のとおりに算出するキーワードの一致度が表により示される。キーワード自動抽出部１３により、第１の実施形態において説明された方法で、それぞれの社説の重み付きキーワードが抽出され、図５の「数」の欄には、各社説が保有している自動抽出されたキーワード数が記述され、図５のマトリクスの交点には、Ｘ軸上の社説に属するキーワードが、Ｙ軸上の社説に対して有する一致度が、記述される。

【０１０６】

第１の実施形態においては、クライアントコンピュータ２から入力される自由キーワードに対して、キーワード付与文書データベース１６に格納された文書テキストごとの一致度が算出されたが、第２の実施形態に係るキーワード一致検出部６１及び文書間一致度算出部６３は、キーワード付与データベース１６に格納されたすべての文書テキストの組み合わせについて、キーワードの一致を検出し、文書間の一致度が上記のとおり算出される。従って、図５の横軸上の社説に属するキーワードが、第１の実施形態においてクライアントコンピュータ２から受信される自由キーワードに相当する。図５から理解されるとおり、ある社説ＡのＢに対する一致度は、必ずしも社説ＢのＡに対する一致度と一致しない。従って、図５のマトリクスは、Ｘ軸ノードからＹ軸ノードまでの重み（Ｘ軸ノードからＹ軸ノードに対する一致度）付きのリンクを複数含む有向グラフと考えることができる。すなわち、図５の横軸上の２３の社説のそれぞれをノードとすると、社説（すなわち、ノード）Ａから社説Ｂに向かう有向リンクの重みは、ノードＡのノードＢに対する一致度として得られる。ノードＡからノードＢへの交点に記述される一致度が０の場合には、ノードＡからノードＢに到達する有向リンクがないことになる。

【０１０７】

このように定義される有向グラフにおいて、ノードＡからノードＢにリンクを辿って到達でき、逆にノードＢからノードＡへもリンクを辿って到達できる場合、ノードＡとノードＢとは、「双方向に連結している」という。この場合、ノードＡからノードＢへリンクを辿って到達できるので、ノードＡからノードＢに達する有向パスが存在し、逆に、ノードＢからノードＡへリンクを辿って到達できるので、ノードＢからノードＡに達する有向パスも存在し、従って、双方向に連結する２つのノードＡ，Ｂは、相互に有向パスで連結される。ここで、パスとは、リンクで繋がったノード列をいう。この互いに双方向に連結するノードの最大の集合を、「双方向連結成分」と呼ぶ。図５のマトリクスから得られる有向グラフは、孤立したノードを含めて、いくつかの双方向成分に分割することができる。

【０１０８】

第２の実施形態においては、図５に例示される文書間の一致度マトリクスにおいて、ノイズを排除するため、所定値以下の一致度を、０で置き換え、この所定値は、分類のためのノイズを排除する目的で用いられる閾値であることから、「分類精度値」という。図５に示されるように、所定値以下の一致度を０で置き換えた一致度マトリクス上で、有向グラフを描き、これを以下説明するように、有向ループを検出し、検出された有向ループを縮退することにより、双方向連結成分に分割する。すなわち、双方向連結成分とは、相互に所定の分類精度以上の一致度を有する関係に立つノードの最大集合である。

【０１０９】

図７は、有向ループ検出部６４が実行する有向ループ検出処理の詳細を示すフローチャートである。第２の実施形態においては、有向ループの検出及びその縮退処理により、双方向連結成分が検出される。図７において、まず一致度マトリクス上の１つのノードを選択し（ステップＳ７０１）、カウンタｉを０に初期化し（ステップＳ７０２）、選択されたノードにマーク$n_i$を付け、マークｋ又はマークｐが付与されていない出リンク（出力先ノードまでのリンク）を探索する（ステップＳ７０３）。このとき、１つ前のノードに戻る出リンク以外を優先させる。１つ前に戻る出リンクしかない場合は、ノード数２のループしかないことになる。出リンクがある場合には、マーク$n_i$が付けられたノードに戻り、１つのループが見つかったことになり、出リンクにマークｋを付ける（ステップＳ７０４）。次のノードにマーク$n_i$が付けられていない場合には（ステップＳ７０５Ｎ）、ノードにマークＰが付けられていれば（ステップＳ７０６Ｙ）、ステップＳ７１４に進み、ノードにマークＰが付けられていなければ（ステップＳ７０６Ｎ）、カウンタｉをインクリメントして（ステップＳ７０７）、ステップＳ７０３に戻る。ステップＳ７０５において、次のノードにマーク$n_i$が付けられている場合には（ステップＳ７０５Ｙ）、ノードＮにマークｍを付け、マークｋが付けられたリンクを辿り、マークｍを付けていく（ステップＳ７０８）。マークｍが付けられたノードが、有向ループを構成するノードとなる（ステップＳ７０９）。ここで、ノードに付けられるマークｍは検出されたループを辿るためのループマークであり、リンクに付けられるマークｋとは、ループを構成するリンクであることを示す「ループ内出リンクマーク」であり、共に、見つかったループを辿るための操作で用いる。

【０１１０】

図７のステップＳ７０３において、出リンクがない場合には、ステップ７１により規定される、単方向の有向パスに入った場合にそこから抜け出すための手順に進む。より詳細には、ステップＳ７０３において出リンクがない場合には、ステップＳ７１０に進み、カウンタｉが０の場合（ステップＳ７１０Ｙ）、マークＰを付けて終了し（ステップＳ７１１）、カウンタｉが０でない場合（ステップＳ７１０Ｎ）、マークＰを付け（ステップＳ７１２）、さらに、カウンタｉをデクリメントし（ステップＳ７１３）、ステップＳ７１４に進む。ステップＳ７１４において、マークｋのリンクを逆に辿り、マーク$n_i$が付けられたノードに戻って、マークｋ、マークｐのない出リンクを見つけ、見つけられた出リンクのマークｋをマークｐに置き換える（ステップＳ７１４）。ステップＳ７１４で、出リンクがある場合には、ステップＳ７０４に戻り（ステップＳ７１５）、出リンクがない

場合には、カウンタｉが０でない場合、ステップＳ７１２に戻り（ステップＳ７１７）、カウンタｉが０であればマークｎ$_i$が付けられたノードにマークＰを付ける（ステップＳ７１８）。マークＰが付けられたノードは、孤立ノードとなる（ステップＳ７１９）。ステップＳ７１に規定される処理により、単方向の有向リンクにしか属さず、ループを形成しないノードが検出される。マークＰが付けられたノードを、孤立ノードと呼ぶ。ループを形成するノードは、１対の有向パスにより形成されるループによって双方向に連結されており、他方、孤立ノードは、単方向にしか連結されていない。

【０１１１】

２．有向ループの縮退処理

図８は、双方向連結成分検出部６５が実行する双方向連結成分の検出処理の詳細を示すフローチャートである。双方向連結成分は、図７に示される処理により検出された有向ループに属するノードを、１つのノードに縮退させることにより、検出される。図８において、まず、図７に示される処理により検出されたループ中の、ノードＮに対して、マークｍが付けられたノードを、次々に縮退させ、この縮退したノードにマークＳを付ける（ステップＳ８０１）。縮退により、縮退されるノード間のリンクは、見かけ上なくなり、それ以外のノードとの間のリンクの重みは、縮退されたノードとそれ以外の外部のノードとの間のリンクの重み（一致度）を加算して得られる。図８において、ノードの縮退後、再度図７に示す処理を適用して、縮退されたノードを含むグラフに対して、有向ループの検出操作を繰り返すが、このとき最初に選択するノードは、マークＳ又はマークＰが付けられていないノードを優先して選択する。マークＳ或いはマークＰが付けられていないノードがなくなると、次にマークＳが付けられたノード間のループを検出し、そのループに属するマークＳのノードを縮退させる。すなわち、マークＳ又はマークＰが付いていないノードがある場合には（ステップＳ８０２Ｙ）、その１つのノードをとり、ノードＮとし（ステップＳ８０３）、図７のステップＳ７０２に戻る（ステップＳ８０４）。一方、マークＳ又はマークＰが付いていないノードがない場合には（ステップＳ８０２Ｎ）、マークＳの付けられたノードと、これらのノード間のリンクからなる部分グラフで、ループを検出する（ステップＳ８０５）。ループが検出された場合、このループに属するノードを縮退させ、縮退されたノードにマークＳを付ける（ステップＳ８０６）。ループがなくなった場合、ループを包含しないグラフにまで縮退されたことになり、このときに、最終的に、マークＳが付けられたノードが、双方向連結成分に相当する。それぞれのノードを、縮退前のノード群に復元する。（ステップＳ８０７）。マークＳが付けられた部分だけを部分グラフとして縮退前に復元することにより、１つの双方向連結成分が取り出せる。マークＰが付けられたノードは、孤立ノードである。

【０１１２】

３．双方向連結成分におけるチェイン状連結の回避処理

第２の実施形態においては、図７及び図８の処理により得られた双方向連結成分におけるチェイン状連結（以下に説明される）を回避して、相互により関連性の高い文書テキスト群のみを１分類とする再分類を実現するため、以下のとおり、パスの検出と制限付き連結成分検出とを実行する。

【０１１３】

図９Ａ及び図９Ｂは、例示的に、縮退後のノードにより構成されるグラフを示す。図９Ａ及び図９Ｂにおいて、「ス１」のノードは、「ス１」－「ス２」－「ス３」－「ス４」－「ス８」－「ス１」からなるループを縮退させて、改めて「ス１」と設定したノードとする。図９Ａに示されるグラフは、図５において分類精度（すなわち、一致するキーワードの重みを目標文書について総和して得られる一致度の閾値）を９とした場合、及び図９Ｂに示されるグラフは、同様に分類精度を１８とした場合を示す。図９Ａに示すとおり、分類精度を９とした場合には、スポーツ社説以外に、３つのビジネス社説が、１つの分類に分類されるが、同じスポーツ社説でも、「ス５」の「大相撲－国際化は面白い」だけが、図９Ａの分類から除かれる。図５中、横軸「ス５」の列において、「ス５」の一致度は、「ビ１」に対して１、「ビ１０」に対して２、それ以外に対しては０であり、分類精度

を９と設定すれば、一致度１や一致度２は０と置き換えられるから、「ス５」は孤立ノードとなる。分類精度を１８に上げた場合には、図９Ｂに示される分類からは、ビジネス社説のすべてが除かれ、「ス５」以外のすべてのスポーツ社説のみから構成される分類となる。図９Ｂに示す分類精度１８の場合は、図４に示すとおり、「ス７」以外のスポーツ社説のすべては野球を、「ス７」は五輪を、「ス１」が野球の五輪問題を、それぞれテーマとしているため、「ス１」が「ス７」をその他のノードに連結する構造になる。「ス６」は、野球をテーマとするが、「ス８」（縮退後に「ス１」に含まれた）とだけ連結される。図９Ａに示す分類精度９の場合は、「ス７」が一致キーワード「ＮＨＫ」と「必要」とにより「ビ１０」に、「ス２」（縮退後に「ス１」に含まれた）が一致キーワード「改革」で「ビ７」に、それぞれ連結する構造になる。「ビ１５」は、「ス１」及び「ス２」の２つのノードに対してリンクを持つ。

【０１１４】

このように、比較的類似する社説同士が、１つに分類されているものの、図９Ａにおいて、「ス６」、「ス７」、「ビ７」、「ビ１０」等は、縮退からの各ノードの復元後にも、単一のノード（縮退復元後の「ス１」内のうち１つだけのノード）のみを介して、他のノードに連結される。これらのノードが、２つの主テーマを有すると仮定すると、２つのテーマのそれぞれに関するキーワードにより、ノードがチェイン状に次々連結されていき、チェインの一端のノードは、他端のノードとかけ離れたテーマとなる虞がある。第２の実施形態においては、こうした不都合を回避するため、図１０に示すパス検出処理、及び図１１に示す制限付き連結成分検出処理により、１つの分類を、複数の部分に細分類する。

【０１１５】

双方向連結成分中で、２つのノード間の１対の有向パスのうち、少なくとも一方向の有向パスが、他の有向パスと異なり、第３のノード経由で連結されている場合、「制限付きで双方向に連結されている」といい、あらゆる２つのノード間の有向パスの対が、制限付きで双方向に連結されている場合、これに属するノードの最大の集合を、「制限付き連結成分」という。図８の処理により得られた分類は、単なる双方向の「連結成分」であり、これをさらに「制限付き連結成分」に分割すれば、それぞれの「制限付き連結成分」は、これに属するノード間が相互により類似する関係を有する分類となることが期待される。

【０１１６】

２つのノードが、制限付きで連結していれば、両者の間には、双方向に互いに異なるノードを含む有向パスが存在する。この「パス」とは、リンクで連結されるノード列をいい、１対の有向パスのうち、一方向の有向パスでは１つのノード列を介して連結し、他方向の有向パスでは別のノード列を介して連結することができる。従って、双方向に互いに異なるノード列を経由する１対の有向パスを検出すれば、ただ１つのノードのみを介して連結されているノード間のパスを排除することが可能となる（例えば、図９Ａにおける「ス６」、「ス７」、「ビ７」、「ビ１０」を分類から切り離すことができる。）。

【０１１７】

図１０は、この「双方向連結成分」を「制限付連結成分」に分割する処理の詳細を示す。まず、図８において取り出された双方向連結成分の中で、図７の処理を適用して、ノード数３以上のループを検出する。ノード数２までのループは、２つのノードからなる連結成分であるから、分類の最小単位を構成するものとし、以下で検出された１つの「制限付き連結成分」と同様に１つの細分類とする。図１０において、ノード数３以上のループに属するすべてのノードに、マークｑを付け、このノードを１つ取り出してノードＸとし、このノードＸの出リンクの１つを取り上げ、このノードＸの相手ノードにマークｂを付ける（ステップＳ１００１）。マークｂが付けられたノードからの出リンクで、相手ノードにマークｂのないものを見つけ（ステップＳ１００２）、見つからない場合には、１つ前のノードに戻り（ステップＳ１００３）、ステップＳ１００２に戻る。一方、マークｂが付けられたノードからの出リンクで、相手ノードにマークｂのないものが見つかった場合には、そのノードにマークｑが付けられているか否かを判断し、そのノードにマークｑが

付けられていない場合には（ステップＳ１００４Ｎ）、そのノードにマークｂを付けて（ステップＳ１００５）、次のノードに進み（ステップＳ１００６）、ステップＳ１００２に戻る。一方、ステップＳ１００４において、当該ノードにマークｑが付けられている場合には（ステップＳ１００４Ｙ）、そのノードがＸである場合には（ステップＳ１００７）、１つ前のノードに戻って（ステップＳ１００９）、ステップＳ１００２に戻り（ステップＳ１０１０）、そのノードがＸでない場合には（ステップＳ１００７Ｎ）、そのパス（即ち、辿ってきたノード列）上のすべてのノードにマークｑを付け（ステップＳ１００８）、１つ前のノードに戻って（ステップＳ１００９）、ステップＳ１００２に戻る（ステップＳ１０１０）。すなわち、ノードＸからの出リンクを取り上げ、相手ノードにマークｑを付け、マークｂがなければマークｂを付ける。このマークｂが付けられたノードからの出リンクに対して、同じ操作を繰り返す。もしマークｂが付けられたノードであれば、そこで操作を止め、別の出リンクの操作に進む。ノードＸの全ての出リンクに対して、この操作が終了すると、ループ上の次のノードをＸとしてこの操作をする。ループ上のすべてのノードに対してこの操作を行なった結果、マークｑが付けられたノードが、制限付き連結成分を構成するノードとなる。これを取り出すとき、これに含まれないノードへのリンクを持つノードがあれば、このノードを２つに分割し、一方を残して他方を制限付き連結成分として取り出す。この制限付き連結成分を取り出す際に、これに含まれないノードへのリンクを持つノードを、「カットノード」といい、このカットノードは、２つ以上の制限付き連結成分に属しているため２つに分割される。残った部分に同じ操作を適用する。以上の手順で、双方向連結成分は、制限付き連結成分に分割される。

【０１１８】

図１１は、入力された文書テキスト群から、双方向連結成分を抽出し、さらに制限付き連結成分に分割することにより、文書テキストを細分類する方法を概観する。図１１において、まず、文書テキストごとに重み付きキーワードが抽出され（ステップＳ１１０１）、文書間のキーワード一致度が算出され、一致度マトリクスを参照して、有向グラフが生成される（ステップＳ１１０２）。この有向グラフ上で、ノード数３以上の有向ループが検出され、検出された有向ループに属さないノードには、孤立ノードを示すマークＰが付けられる（ステップＳ１１０３）。ループは、図８の手順により縮退され、縮退されたループにはマークＳが付けられる（ステップＳ１１０４）。この操作を、マークＰ又はマークＳが付かないノードがなくなり、かつ縮退されたグラフにループが存在しなくなるまで繰り返す（ステップＳ１１０５）。ステップＳ１１０３において、ループが存在しなくなった場合、縮退前のノード群に復元する（ステップＳ１１０６）。ここで縮退前の状態に復元されたノード群が、双方向連結成分を構成する。縮退前に復元された双方向連結成分に対して、図７の処理手順により、有向ループを１つ取り上げ、そのノード間の有向パスをすべて見つけ、見つけられた有向パス上のノードにマークｑをつける（ステップＳ１１０７）。マークｑが付けられたノードからなる部分グラフを抜き取って（ステップＳ１１０８）、ステップＳ１１０６に戻る。ステップＳ１１０６からステップＳ１１０８までの処理を、部分グラフが尽きるまで繰り返す。すなわち、ステップＳ１１０７において、すべての双方向連結成分について、見つけられた有向ループのすべてのノードから始まり、ループ上のいずれかのノードで終端するパスが、図１０の処理手順により網羅されるまで、ステップＳ１１０６からステップＳ１１０８までの処理を繰り返し（ステップＳ１１０９Ｎ）、すべての双方向連結成分についての処理が終了した時点で（ステップＳ１１０９Ｙ）、見つけられた有向パスを構成するノードと、元となったループを構成するノードの集合を抽出すると、これらのすべてが制限付き連結成分となり、この制限付き連結成分の１つが、１つの細分類に相当し、この制限付き連結成分であるノード群が分類される（ステップＳ１１１０）。

【０１１９】

上記の手順により、分類精度を１８とした場合には、図１２に示すように、図９Ｂ中の「ス６」及び「ス７」が分類から除かれて、野球をテーマとして社説のみ、すなわち「ス１」、「ス２」、「ス３」、「ス４」、「ス８」が、１つの分類に属する。分類精度を９

とした場合には、「ス６」、「ス７」に加えて、さらに「ビ７」、「ビ１０」が「ス１」の分類から除かれ、「ビ１５」を含む野球をテーマとする社説が「ス１」の分類に分類される。こうして、比較的類似性が高い、すなわち関連が深い文書群のみの分類に行き着くことができる。

【０１２０】

第２の実施形態においては、最終的に得られた細分類に属するノード（文書テキスト）に付与されたキーワードの集合を、当該細分類に付与する。例えば、キーワード付与文書データベース１６に格納される文書テキスト数が非常に多い場合に、クライアントコンピュータ２から受信される自由キーワードと最も一致する分類（細分類）をまず検索結果としてクライアントコンピュータ２上で一覧表示させ、選択された分類に属する文書テキストだけを検索対象文書テキストとして、上記のフリーワード検索処理を実行してもよい。このように構成すれば、文書検索の負荷を軽減することが可能となる。図１２において、分類全体を表すキーワードは、「ス１」、「ス２」、「ス３」、「ス４」、「ス８」からなる分類では、これらノード相互を連結するリンクに相当するキーワードの集合であり、「ビ１５」に対応するキーワードは、「ビ１５」と、「ス１」及び「ス２」を連結するリンクに相当するキーワードの集合である。

【０１２１】

第２の実施形態によれば、第１の実施形態により得られる機能に加え、さらに、文書テキストから自動抽出されたキーワード及びその出現頻度から得られるキーワードの重みのみを利用して、文書間の一致度を算出し、この一致度の有向性に基づいて文書テキスト同士の関連性を評価するので、大量の文書テキストが、簡易且つ高精度で類似する文書テキスト群に自動的に分類される。

【０１２２】

さらに、分類内に属する文書テキスト間リンクでのチェイン状連結を回避するので、異なるカテゴリーに属する文書テキストを入力としても、高精度の細分類が実現される。殊に、異なるカテゴリーに属する文書テキストを保有し、文書間の均質性を欠く文書データベースをキーワード抽出対象とした場合にあっても、キーワード抽出及びこれを用いた分類の精度が低下することがないという利点が得られる。

【０１２３】

従って、大量のデジタルコンテンツを、何らの事前定義を要することなく、自動的且つ実用的高精度に分類することが実現される。

【０１２４】

本発明の範囲は、図示され記載された例示的な実施形態に限定されるものではなく、本発明が目的とするものと均等な効果をもたらすすべての実施形態をも含み、その要旨を逸脱しない範囲で多様な改良ないし変更が可能である。例えば、インターネットでの情報提供ビジネス、パーソナルコンピュータのハードディスク内情報管理及びその情報分析、辞書機能の高度化等、多用な技術への応用が、簡易かつ安価に実現され、これにより、利用者の利便性が大幅に向上する。より具体的には、インターネット等のネットワーク環境下で情報を配信する仕組みを構築するコンテンツプロバイダ、データベース管理システム構築、パーソナルコンピュータ用データ管理ソフトウエア等の供給システムの構築を行なうためのサーバ、情報処理装置又は方法、並びにコンピュータプログラムとしての提供も可能となる。さらに、本発明の範囲は、請求項１により画される発明の特徴の組み合わせに限定されるものではなく、すべての開示されたそれぞれの特徴のうち特定の特徴のあらゆる所望する組み合わせによって画されうる。

【図面の簡単な説明】

【０１２５】

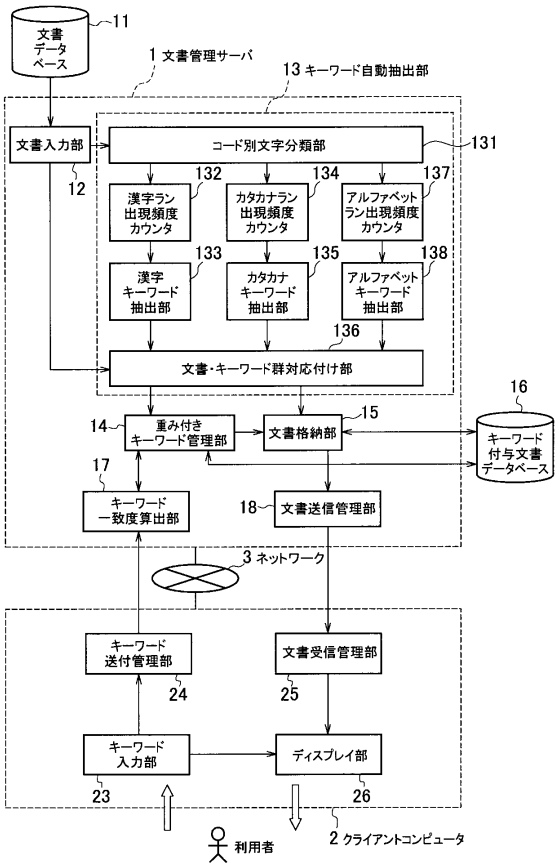
【図１】本発明の第１の実施形態に係る文書インデキシングシステムの機能構成の一例を示すブロック図である。

【図２】本発明の第１の実施形態に係る文書インデキシングシステムにより、新聞社説２００件から、漢字ラン、カタカナラン、アルファベットランの出現頻度をカウントして得

られた結果を非例示的に示す図である。
【図３Ａ】朝日新聞の社説２００５年７月６日を入力文書とした場合の第１の実施形態におけるラン出現頻度を示す図である。
【図３Ｂ】朝日新聞の社説２００５年７月９日を入力文書とした場合の第１の実施形態におけるラン出現頻度を示す図である。
【図４】新聞社説を入力文書テキストとした場合に、第１の実施形態において抽出されるキーワードとその重みを非例示的に示す図である。
【図５】本発明の第２の実施形態に係る一致度マトリクス生成部６３が生成する文書間一致度マトリクスを非例示的に示す図である。
【図６】本発明の第２の実施形態に係る文書管理サーバ１を具備する文書分類システムの一構成例を示すブロック図である。
【図７】本発明の第２の実施形態に係る有向ループ検出部６４が実行する有向ループ検出処理の詳細を示すフローチャートである。
【図８】本発明の第２の実施形態に係る双方向連結成分検出部６５が実行する双方向連結成分の検出処理の詳細を示すフローチャートである。
【図９Ａ】本発明の第２の実施形態における、縮退後のノードにより構成されるグラフを、分類精度９の場合で非例示的に示した模式図である。
【図９Ｂ】本発明の第２の実施形態における、縮退後のノードにより構成されるグラフを、分類精度１８の場合で非例示的に示した模式図である。
【図１０】本発明の第２の実施形態に係る制限付き連結成分分割部６６が実行するパス検出処理の詳細を示すフローチャートである。
【図１１】本発明の第２の実施形態において、入力された文書テキスト群から、双方向連結成分を抽出し、さらに制限付き連結成分に分割することにより実行される文書テキストの細分類方法を概観するフローチャートである。
【図１２】本発明の第２の実施形態により得られる細分類結果を非例示的に示す図である。
【図１３】本発明の各実施形態に係る文書管理サーバ及び／又はクライアントコンピュータのハードウエア構成の一例を示す図である。
【符号の説明】
【０１２６】
１　文書管理サーバ
２　クライアントコンピュータ
３　ネットワーク
１１　文書データベース
１２　文書入力部
１３　キーワード自動抽出部
１４　重み付きキーワード管理部
１５　文書格納部
１６　キーワード付与文書データベース
１７　キーワード一致度算出部
１８　文書送信管理部
２３　キーワード入力部
２４　キーワード送付管理部
２５　文書受信管理部
２６　ディスプレイ部
１３１　コード別文字分類部
１３２　漢字ラン出現頻度カウンタ
１３３　漢字キーワード抽出部
１３４　カタカナラン出現頻度カウンタ
１３５　カタカナキーワード抽出部

１３６　文書・キーワード群対応付け部
１３７　アルファベットラン出現頻度カウンタ
１３８　アルファベットキーワード抽出部

【図１】




【図２】

| | 漢字ラン | | カタカナラン | | アルファベットラン | |
|---|---|---|---|---|---|---|
| | 出現頻度 | 出現率 | 出現頻度 | 出現率 | 出現頻度 | 出現率 |
| 平均 | 17.36 | 1.53% | 4.97 | 0.44% | 0.21 | 0.02% |
| 最大 | 84 | 7.51% | 17 | 1.50% | 4 | 0.36% |
| 最小 | 6 | 0.04% | 0 | 0% | 0 | 0% |

【図３Ａ】

| | 050706_1 | | |
|---|---|---|---|
| 1 | 首相 | 7 | 4.7% |
| 2 | 反対 | 6 | 4.0% |
| 3 | 党内 | 5 | 3.3% |
| 4 | 法案 | 3 | 2.0% |
| 5 | 派閥 | 3 | 2.0% |
| 6 | 執行部 | 3 | 2.0% |
| 7 | 自民党 | 3 | 2.0% |
| 8 | 反発 | 2 | 1.3% |
| 9 | 反対派 | 2 | 1.3% |
| 10 | 衆院解散 | 2 | 1.3% |
| 11 | 参院 | 2 | 1.3% |
| 12 | 公明党 | 2 | 1.3% |
| 13 | 決定 | 2 | 1.3% |
| 14 | 議員 | 2 | 1.3% |
| 15 | 連立相手 | 1 | 0.7% |
| 16 | 連発 | 1 | 0.7% |
| 17 | 了承 | 1 | 0.7% |
| 18 | 離党 | 1 | 0.7% |
| 19 | 理解 | 1 | 0.7% |
| 20 | 利害調整 | 1 | 0.7% |
| 21 | 郵政民営化法案 | 1 | 0.7% |
| 22 | 郵政民営化 | 1 | 0.7% |
| 23 | 郵政法案可決 | 1 | 0.7% |
| 24 | 民営化 | 1 | 0.7% |
| 25 | 本来 | 1 | 0.7% |
| 26 | 本丸 | 1 | 0.7% |
| 27 | 法案審議 | 1 | 0.7% |
| 28 | 法案修正 | 1 | 0.7% |
| 29 | 方向 | 1 | 0.7% |
| 30 | 物語 | 1 | 0.7% |
| 31 | 副大臣 | 1 | 0.7% |
| 32 | 舞台 | 1 | 0.7% |
| 33 | 乖差 | 1 | 0.7% |
| 34 | 漂流 | 1 | 0.7% |
| 35 | 皮一枚 | 1 | 0.7% |
| 36 | 批判 | 1 | 0.7% |
| 37 | 否定 | 1 | 0.7% |
| 38 | 否決 | 1 | 0.7% |
| 39 | 反対票 | 1 | 0.7% |
| 40 | 反旗 | 1 | 0.7% |
| 41 | 薄笑 | 1 | 0.7% |
| 42 | 背水 | 1 | 0.7% |
| 43 | 年余 | 1 | 0.7% |
| 44 | 任期 | 1 | 0.7% |
| 45 | 二十数人 | 1 | 0.7% |
| 46 | 内心 | 1 | 0.7% |
| 47 | 内閣不信任案 | 1 | 0.7% |
| 48 | 内閣 | 1 | 0.7% |
| 49 | 特別委員会 | 1 | 0.7% |
| 50 | 得意 | 1 | 0.7% |
| 51 | その他の漢字 | 70 | 46.7% |
| 52 | 漢字ラン数 | 150 | |
| 53 | ポスト | 1 | 100.0% |
| 54 | カタカナラン数 | 1 | |

【図３Ｂ】

| | 050709_2 | | |
|---|---|---|---|
| 1 | 組織 | 3 | 2.6% |
| 2 | 世界 | 3 | 2.6% |
| 3 | 犯行 | 2 | 1.8% |
| 4 | 宗教 | 2 | 1.8% |
| 5 | 国際 | 2 | 1.8% |
| 6 | 攻撃 | 2 | 1.8% |
| 7 | 寛容 | 2 | 1.8% |
| 8 | 冷静 | 1 | 0.9% |
| 9 | 理由 | 1 | 0.9% |
| 10 | 明確 | 1 | 0.9% |
| 11 | 無差別 | 1 | 0.9% |
| 12 | 味方 | 1 | 0.9% |
| 13 | 万以上 | 1 | 0.9% |
| 14 | 亡命政権 | 1 | 0.9% |
| 15 | 亡命貴族 | 1 | 0.9% |
| 16 | 亡命 | 1 | 0.9% |
| 17 | 米政権 | 1 | 0.9% |
| 18 | 米軍 | 1 | 0.9% |
| 19 | 文明 | 1 | 0.9% |
| 20 | 武力 | 1 | 0.9% |
| 21 | 評価 | 1 | 0.9% |
| 22 | 非難 | 1 | 0.9% |
| 23 | 皮肉 | 1 | 0.9% |
| 24 | 卑劣 | 1 | 0.9% |
| 25 | 犯人 | 1 | 0.9% |
| 26 | 発言 | 1 | 0.9% |
| 27 | 爆弾 | 1 | 0.9% |
| 28 | 年前 | 1 | 0.9% |
| 29 | 任務 | 1 | 0.9% |
| 30 | 同様 | 1 | 0.9% |
| 31 | 同時爆破 | 1 | 0.9% |
| 32 | 同時 | 1 | 0.9% |
| 33 | 動揺 | 1 | 0.9% |
| 34 | 投資 | 1 | 0.9% |
| 35 | その他の漢字 | 71 | 62.3% |
| 36 | 漢字ラン数 | 114 | |
| 37 | テロ | 17 | 38.6% |
| 38 | イスラム | 6 | 13.6% |
| 39 | ロンドン | 3 | 6.8% |
| 40 | イラク | 3 | 6.8% |
| 41 | グループ | 2 | 4.5% |
| 42 | アルカイダ | 2 | 4.5% |
| 43 | マルクス | 1 | 2.3% |
| 44 | マドリード | 1 | 2.3% |
| 45 | ドイツ | 1 | 2.3% |
| 46 | ブレア | 1 | 2.3% |
| 47 | ブッシュ | 1 | 2.3% |
| 48 | フセイン | 1 | 2.3% |
| 49 | スコットランド | 1 | 2.3% |
| 50 | カール | 1 | 2.3% |
| 51 | アキレス | 1 | 2.3% |
| 52 | テロリスト | 1 | 2.3% |
| 53 | サミット | 1 | 2.3% |
| 54 | カタカナラン数 | 44 | |

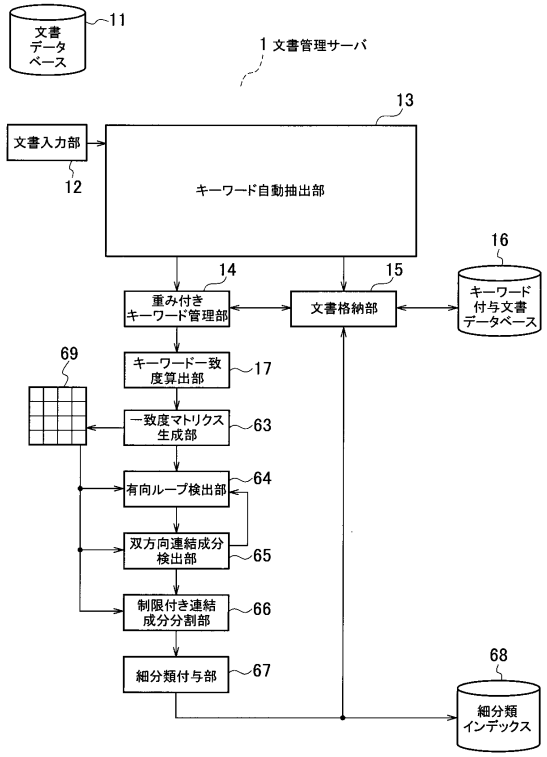
【図４】

| | タイトル | 漢字キーワード | カタカナキーワード |
|---|---|---|---|
| 1 | 「造反」処分__ご都合主義はだめだ | 自民党(9),反対(4),法案(3),首相(3),有権者(3),政策(2),実現(2),郵政民営化法案(1),決別,総選挙(1),反対派(1),政権維持(1),党本部(1),前議員(1),公認(1),参院(1),処分(1),争点(1) | スタイル(1),マニフェスト(1),リーダー(1),レッテル(1),アジア(1) |
| 2 | つくば新線__まちの未来が託された | 開業(3),首都圏(2),都心(2),沿線(2),自治体(2),秋葉原(1)初日(1),建設(1),鉄道建設(1),宅地開発(1),第三(1),沿線住宅(1),鉄道事業(1),万人(1),鉄道会社(1),安全(1) | キロ(2),バブル(2),セクター(2),エクスプレス(1),ニュータウン(1),ホーム(1),オーバーラン(1) |
| 3 | 税制改革__選択肢を明示して | 増税(5),消費税(3),所得税(3),負担(2),自民党(2),所得(2),選挙権(1),世代(1),有権者(1),実現(1),自民(1),年間(1),把握(1),民主党(1),財源(1),提案(1) | マニフェスト(3),サラリーマン(2),タブー(1),メド(1),カット(1),ドイツ(1),サービス(1) |
| 4 | きょう投票__どんな4年を選ぼうか | 政策(2),政党(2),面白(1),機会(1),有権者(1),総選挙(1),時代(1),小泉首相(1),民間(1),年金(1),問題(1),一番(1),期待(1) | フレーズ(1),マジック(1),イラク(1) |
| 5 | 郵政法案__民業圧迫を防げ | 政府(5),民主党(4),対案(3),民営化(3),政府案(3),金融(2),公社(2),貯金(1),保険(1),株式(1),資本関係(1),縮小(1),郵便事業(1),国会(1),職員(1),国会審議(1) | ビジネス(1),リストラ(1),グループ(1) |
| 6 | 揺れる阪神__あんじょう頼んまっせ | 村上氏(4),電鉄(2),村上(1),電鉄側(1),親会社(1),球界(1),球団経営(1),球団(1) | ファン(4),タイガースファン(3),ファンド(2),タイガース(2),グループ(2) |
| 7 | 平和賞__核拡散への警鐘だ | 施設(2),一方(1),実績(1),加盟国(1),番人(1),査察官(1),濃縮計画(1),核保有(1),査察(1),未申告(1),追加議定書(1),締約国(1),約束(1) | ノーベル(4),イラク(2),ウラン(2),イラン(2),ウィーン(1),モハメド・エルバラダイ(1),エネルギー(1),テロ(1),エルバラダイ(1),パキスタン(1),カーン(1) |

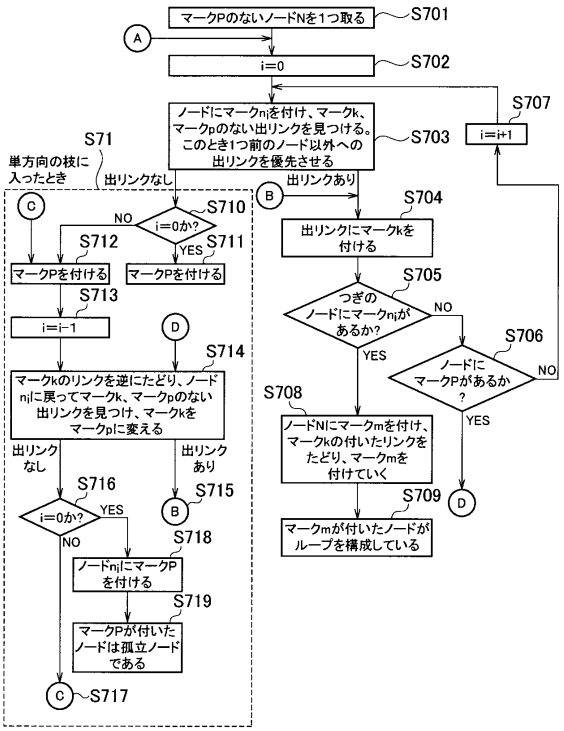
【図５】

| | | 数 | ビ1 | ビ2 | ビ3 | ビ4 | ビ5 | ビ6 | ビ7 | ビ8 | ビ9 | ビ10 | ビ11 | ビ12 | ビ13 | ビ14 | ビ15 | ス1 | ス2 | ス3 | ス4 | ス5 | ス6 | ス7 | ス8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 23 | 23 | 22 | 27 | 18 | 25 | 23 | 20 | 26 | 21 | 19 | 22 | 21 | 24 | 31 | 18 | 23 | 24 | 25 | 20 | 13 | 30 | 16 |
| ビ1 | 23 | — | 0 | 6 | 0 | 11 | 4 | 2 | 0 | 0 | 0 | 3 | 0 | 1 | 1 | 8 | 1 | 0 | 9 | 4 | 4 | 1 | 0 | 9 | 0 |
| ビ2 | 23 | 0 | — | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 4 | 0 |
| ビ3 | 22 | 6 | 0 | — | 0 | 5 | 0 | 39 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ4 | 27 | 11 | 0 | 0 | — | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ5 | 18 | 4 | 0 | 0 | 0 | — | 0 | 6 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 4 | 0 |
| ビ6 | 25 | 2 | 0 | 0 | 0 | 0 | — | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ7 | 23 | 0 | 0 | 0 | 0 | 0 | 0 | — | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ8 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | — | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ9 | 26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | — | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ10 | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | — | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ11 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | — | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ12 | 22 | 1 | 0 | 2 | 0 | 4 | 0 | 3 | 0 | 0 | 0 | 0 | — | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ13 | 21 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | — | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ14 | 24 | 8 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | — | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ビ15 | 31 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | — | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ス1 | 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | — | 4 | 0 | 0 | 0 | 0 | 6 | 0 |
| ス2 | 23 | 9 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | — | 0 | 0 | 0 | 0 | 6 | 0 |
| ス3 | 24 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | — | 0 | 0 | 0 | 6 | 0 |
| ス4 | 25 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | — | 0 | 0 | 6 | 0 |
| ス5 | 20 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | — | 0 | 6 | 0 |
| ス6 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | — | 6 | 0 |
| ス7 | 30 | 9 | 4 | 6 | 6 | 4 | 4 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | — | 0 |
| ス8 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 6 | — |

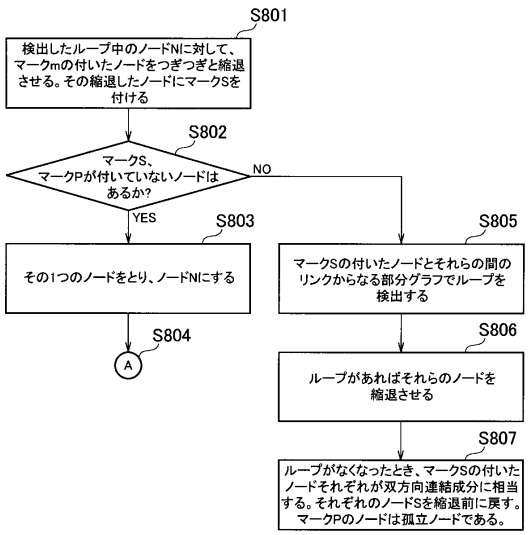
【図６】


文書データベース
11
1 文書管理サーバ
13
文書入力部
12
キーワード自動抽出部
16
キーワード付与文書データベース
14
15
重み付きキーワード管理部
文書格納部
69
キーワード一致度算出部
17
一致度マトリクス生成部
63
有向ループ検出部
64
双方向連結成分検出部
65
制限付き連結成分分割部
66
細分類付与部
67
68
細分類インデックス

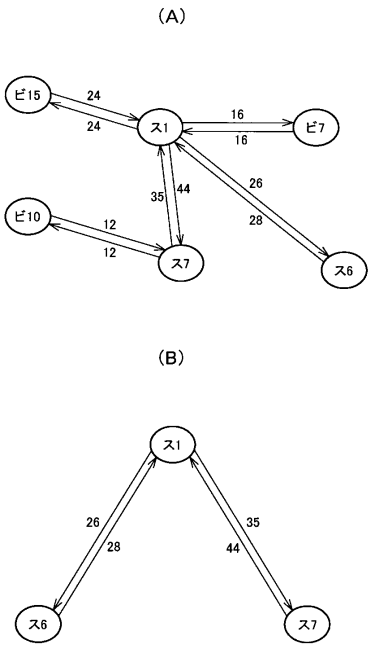
【図７】


マークPのないノードNを1つ取る
S701
A
i=0
S702
ノードにマークniを付け、マークk、マークpのない出リンクを見つける。このとき1つ前のノード以外への出リンクを優先させる
S703
i=i+1
S707
S71
単方向の枝に入ったとき
出リンクなし
出リンクあり
C
NO
i=0か?
S710
S712
S711
YES
マークPを付ける
マークPを付ける
S713
i=i-1
D
S714
マークkのリンクを逆にたどり、ノードniに戻ってマークk、マークpのない出リンクを見つけ、マークkをマークpに変える
出リンクなし
出リンクあり
S716
S715
i=0か?
YES
B
NO
S718
ノードniにマークPを付ける
S719
マークPが付いたノードは孤立ノードである
C
S717
B
S704
出リンクにマークkを付ける
S705
つぎのノードにマークniがあるか?
NO
YES
S706
ノードにマークPがあるか?
NO
YES
S708
ノードNにマークmを付け、マークkの付いたリンクをたどり、マークmを付けていく
S709
D
マークmが付いたノードがループを構成している

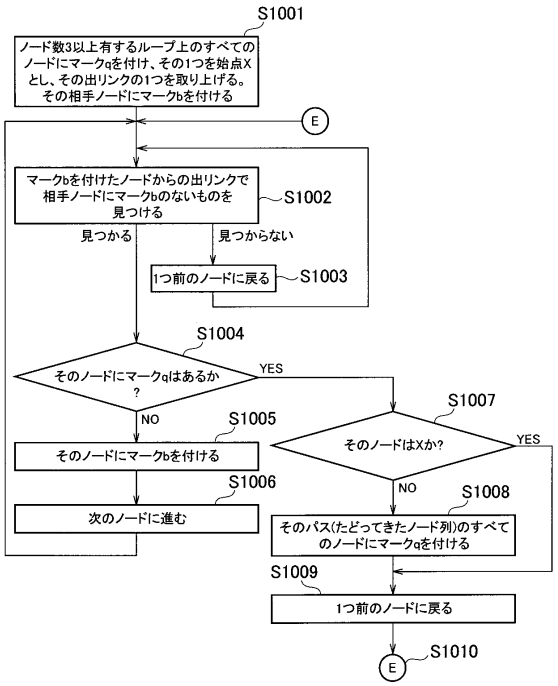
【図８】


S801
検出したループ中のノードNiに対して、マークmの付いたノードをつぎつぎと縮退させる。その縮退したノードにマークSを付ける
S802
マークS、マークPが付いていないノードはあるか?
NO
YES
S803
その1つのノードをとり、ノードNiにする
S804
A
S805
マークSの付いたノードとそれらの間のリンクからなる部分グラフでループを検出する
S806
ループがあればそれらのノードを縮退させる
S807
ループがなくなったとき、マークSの付いたノードそれぞれが双方向連結成分に相当する。それぞれのノードSを縮退前に戻す。マークPのノードは孤立ノードである。

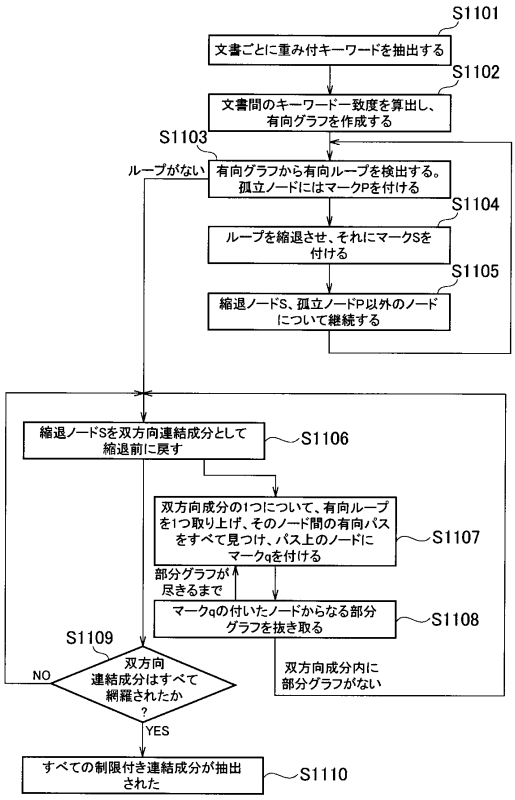
【図９】


(A)
ビ15
ス1
ビ7
ビ10
ス7
ス6
24
24
16
16
35
44
26
28
12
12
(B)
ス1
ス6
ス7
26
28
35
44

【図１０】



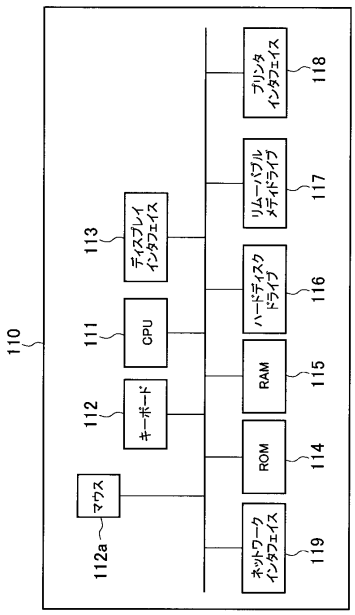

【図１１】




【図１２】

| 分類精度 | | 縮退したノードのタイトル | 分類全体を表すキーワード |
|---|---|---|---|
| 重み9 | ビ7 | 景気の行方ー回復の階段を上るには | 改革，必要，NHK |
| | ビ10 | NHK改革ー視聴者の声を上げよう | |
| | ビ15 | TBSと楽天　時代の波に乗れるか | 野球，経営，球団，プロ，経営者側，野球協約 |
| 重み18 | ス1 | 野球とソフト　五輪だけが舞台じゃない | プロ，選手，選手会，野球，有力選手，野球留学，野球留学生，野球部長，高校，高校生，制度，不祥事，部員，暴力，学校，出場，スポーツ，チーム |
| | ス2 | ドラフト改革ーファン無視の空振り | |
| | ス3 | 高校野球ーフェアプレーを忘れるな | |
| | ス4 | 高校野球ー後味の悪さはこれ限りに | |
| | ス8 | ロッテ優勝ー背番号26の勝利だ | |
| | ス6 | 揺れる阪神ーあんじょう頼んまっせ | ファン，球団，球団経営 |
| | ス7 | 五輪招致ーあの青空をもう一度？ | 五輪，世界，日本，必要 |

【図１３】



フロントページの続き

(72)発明者　曽根原　登
東京都千代田区一ツ橋２－１－２　大学共同利用機関法人　情報・システム研究機構　国立情報学研究所内
(72)発明者　釜江　尚彦
東京都千代田区一ツ橋２－１－２　大学共同利用機関法人　情報・システム研究機構　国立情報学研究所内
(72)発明者　沼田　秀穂
東京都杉並区松庵　３－２０－１１
(72)発明者　池田　佳代
東京都豊島区長崎　５－１８－８

審査官　野田　佳邦

(56)参考文献　特開平１１－１４３９０２（ＪＰ，Ａ）
特開平０８－１６１３４４（ＪＰ，Ａ）
特開平０６－１８７３７３（ＪＰ，Ａ）
特開平１１－０５３３８７（ＪＰ，Ａ）
特開昭６４－０２８７７０（ＪＰ，Ａ）
特開平１０－０１１４６０（ＪＰ，Ａ）
特開平１０－１８７７３６（ＪＰ，Ａ）
特開平１０－３０７８４０（ＪＰ，Ａ）
特開平１１－３３８８８３（ＪＰ，Ａ）
特開平１０－３２０４２１（ＪＰ，Ａ）
特開平０８－０３０６２７（ＪＰ，Ａ）
特開２００１－２４９９２２（ＪＰ，Ａ）
特開平０７－０６５０１８（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０