

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2007-272699  
(P2007-272699A)

(43) 公開日 平成19年10月18日(2007.10.18)

(51) Int. Cl. F I テーマコード (参考)  
**G06F 17/30 (2006.01)** G06F 17/30 210A 5B075  
 G06F 17/30 170A

審査請求有 請求項の数 16 O L (全 34 頁)

(21) 出願番号	特願2006-99401 (P2006-99401)	(71) 出願人	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
(22) 出願日	平成18年3月31日 (2006.3.31)	(71) 出願人	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

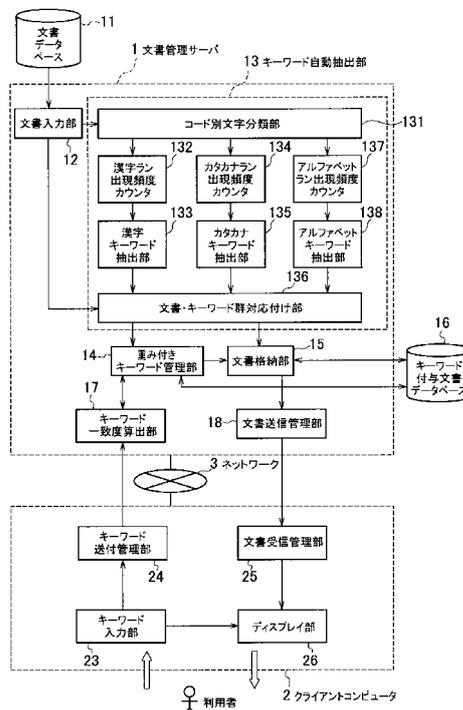
(54) 【発明の名称】 文書インデキシング装置、文書検索装置、文書分類装置、並びにその方法及びプログラム

(57) 【要約】

【課題】 大量の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、フリーワード文書テキスト検索、文書自動分類を容易化する。

【解決手段】 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(131)と、抽出された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(132、134)と、日本語文書テキスト内の各文字列の総数に対して所定比率以上の出現頻度を有する文字列をキーワードとして得、出現頻度のみから算出される重みを付加して重み付きキーワードを得るキーワード生成部(133、135)とを具備する。

【選択図】 図1



**【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。

**【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】**

請求項1に記載の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。

**【請求項 4】**

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る

ことを特徴とする請求項3に記載の文書検索サーバ装置。

**【請求項 5】**

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する

10

20

30

40

50

ことを特徴とする請求項 3 に記載の文書検索サーバ装置。

【請求項 6】

請求項 1 に記載の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

10

ことを特徴とする文書分類サーバ装置。

【請求項 7】

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する

ことを特徴とする請求項 6 に記載の文書分類サーバ装置。

【請求項 8】

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する

20

ことを特徴とする請求項 6 に記載の文書分類サーバ装置。

【請求項 9】

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する

ことを特徴とする請求項 6 に記載の文書分類サーバ装置。

30

【請求項 10】

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 응답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、

40

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。

【請求項 11】

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る

50

ことを特徴とする請求項 10 に記載の文書検索クライアント装置。

【請求項 12】

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する

ことを特徴とする請求項 10 に記載の文書検索クライアント装置。

【請求項 13】

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。

【請求項 14】

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回答して、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。

【請求項 15】

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

10

20

30

40

50

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

10

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16】

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、

20

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。

#### 【発明の詳細な説明】

#### 【技術分野】

30

#### 【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】

40

#### 【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】

50

グループウェア等のソフトウェアが、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。

10

## 【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。

20

## 【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。

30

## 【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。

40

## 【0009】

50

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従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、

10

20

30

40

50

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。

10

【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。

20

【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ファイルであってよい。

30

【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。

40

【0024】

50

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は、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化する。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。

10

20

**【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。

30

**【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ー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。

40

**【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

50

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は、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。

10

【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。

20

【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。

30

【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。

40

【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

50

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 응답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るとよい。

【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

10

20

30

40

50

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回答して、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。

10

【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。

20

30

【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回答して、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、

40

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

50

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は、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を得られる。

#### 【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は、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を得られる。

#### 【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

10

20

30

40

50

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 に该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現頻度カウンタ 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データ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

10

20

30

40

50

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カードリーダーなどを用いて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。

10

【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、ひらがな、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+9FFF、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。

20

【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ことにより、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、「彼は病氣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」、「病氣勝」、「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。

30

【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、

40

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

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「・」、-（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以上のものみの出現頻度がカウントされる。

【0063】

50

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、学术论文、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。

10

【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。

20

30

【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、学术论文、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。

40

【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】

50

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）、「-」（長音記号）、スペースとカタカナ文字列の最後の「-」（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 3 A 及び図 3 B 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 3 A 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 3 B において、050709\_2（2005 年 7 月 9 日の社説）

10

20

30

40

50

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部は、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

10

20

30

40

50

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。

10

【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。

20

【0084】

- a) Xに完全に一致するキーワードが文書Bにある。

【0085】

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

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 $Bw_1kwXw_2$ にあるか、又は文書Aのキーワード $Aw_1kwXw_2(n_k)$ 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、 $Aw_1kwXw_2(n_k)$ 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。

30

【0086】

$$AkwX(n_k) \quad Bw_1kwXw_2(n_k) \quad (3)$$

$$Aw_1kwXw_2(n_k) \quad BkwX(n_k) \quad (4)$$

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 $w_1kwXw_2$ を含むキーワードが文書Bにある。このとき文書Aの $Aw_1kwXw_2$ の最も長い部分を共有する文書Bのキーワード $Bw_1kwXw_2$ 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。

40

【0087】

$$Aw_1kwXw_2(n_k) \quad Bw_1kwXw_2(n_k) \quad (5)$$

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) 通常的一致度：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 $n_k$ と対応する文書Aの重み $n_k$ 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

50

。

## 【数 1】

$$C_n = n_k \times n_k \times m \quad (6)$$

$$\sum_{k=1}^n c_{n_k} = c_{n_1} + c_{n_2} + c_{n_3} + \dots + c_{n_n} \quad (7)$$

## 【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。 10

## 【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、 $1 \times 3$ （「有権者」） $+ 1 \times 2$ （「政策」） $+ 1 \times 1$ （「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  $2 \times 1$  となるので、 $1 \times 3$ （「有権者」） $+ 1 \times 2$ （「政策」） $+ 2 \times 1$ （「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。 20 30

## 【0091】

上記の例では、最大の一一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。 40

## 【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ェア構成>

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のハードウェ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ース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ース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ース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のハードウェア構成は一例に過ぎず、その他の任意のハードウェ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

10

20

30

40

50

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は、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を得られる。

10

#### 【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を得られる。

20

#### 【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。

30

#### 【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。

40

#### 【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

50

部 6 5 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 6 6 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 6 7 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 1 7 は、重み付きキーワード管理部 1 4 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】

10

< 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 6 3 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 6 2 が、上記のとおり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 1 3 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

20

#### 【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 1 6 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 6 1 及び文書間一致度算出部 6 3 は、キーワード付与データベース 1 6 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 から理解されたとおり、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 2 3 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。

30

#### 【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。

40

#### 【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

50

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<sub>i</sub>*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<sub>i</sub>*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<sub>i</sub>*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<sub>i</sub>*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<sub>i</sub>*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<sub>i</sub>*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

10

20

30

40

50

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、ループを含ま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。

10

20

#### 【0112】

### 3. 双方向連結成分におけるチェー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ーン状連結（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、「ピ1」に対して1、「ピ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「ピ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「ピ7」に、それぞれ連結する構造になる。「ピ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。

30

40

#### 【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

50

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、ノードがチェーン状に次々連結されていき、チェー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

10

20

30

40

50

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、「ピ15」に対応するキーワードは、「ピ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ーン状連結を回避するので、異なるカテゴリ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は、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、自動的に且つ実用的な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、多様な技術への応用が、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ェ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定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。

10

20

30

40

50

【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 6 5 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 9 A】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 9 B】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 1 8 の場合で非例示的に示した模式図である。

【図 1 0】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 6 6 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 1 1】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。 10

【図 1 2】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。

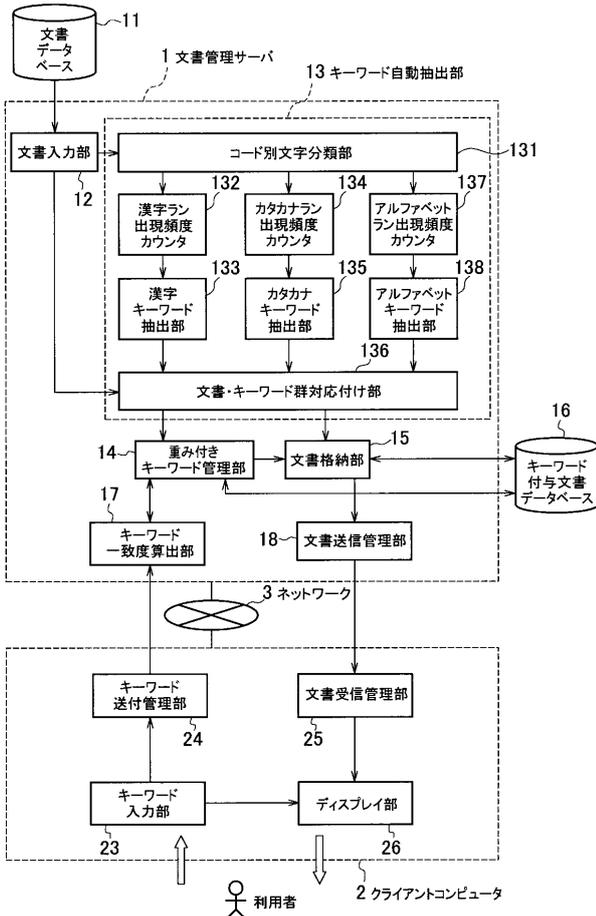
【図 1 3】本発明の各実施形態に係る文書管理サーバ及び / 又はクライアントコンピュータのハードウェア構成の一例を示す図である。

#### 【符号の説明】

#### 【 0 1 2 6 】

- |       |                   |    |
|-------|-------------------|----|
| 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            | 20 |
| 1 1   | 文書データベース          |    |
| 1 2   | 文書入力部             |    |
| 1 3   | キーワード自動抽出部        |    |
| 1 4   | 重み付きキーワード管理部      |    |
| 1 5   | 文書格納部             |    |
| 1 6   | キーワード付与文書データベース   |    |
| 1 7   | キーワード一致度算出部       |    |
| 1 8   | 文書送信管理部           |    |
| 2 3   | キーワード入力部          |    |
| 2 4   | キーワード送付管理部        | 30 |
| 2 5   | 文書受信管理部           |    |
| 2 6   | ディスプレイ部           |    |
| 1 3 1 | コード別文字分類部         |    |
| 1 3 2 | 漢字ラン出現頻度カウンタ      |    |
| 1 3 3 | 漢字キーワード抽出部        |    |
| 1 3 4 | カタカナラン出現頻度カウンタ    |    |
| 1 3 5 | カタカナキーワード抽出部      |    |
| 1 3 6 | 文書・キーワード群対応付け部    |    |
| 1 3 7 | アルファベットラン出現頻度カウンタ |    |
| 1 3 8 | アルファベットキーワード抽出部   | 40 |

【 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 3 A 】

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	難発	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	兼差	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	反発	1 0.7%
41	薄笑	1 0.7%
42	青水	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 3 B 】

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	動極	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	マドリッド	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	カーン	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

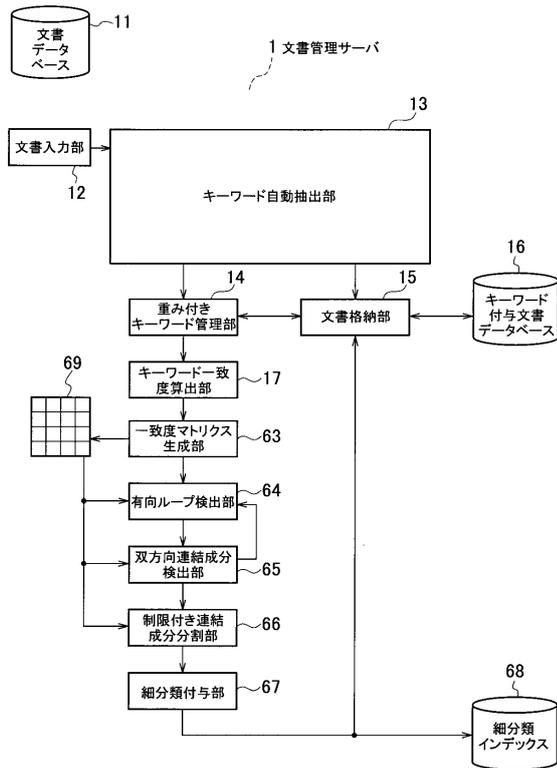
【 図 4 】

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 選択枝を明示 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 選ぶのか	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), 対策(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 あじょう 頼んませ	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)

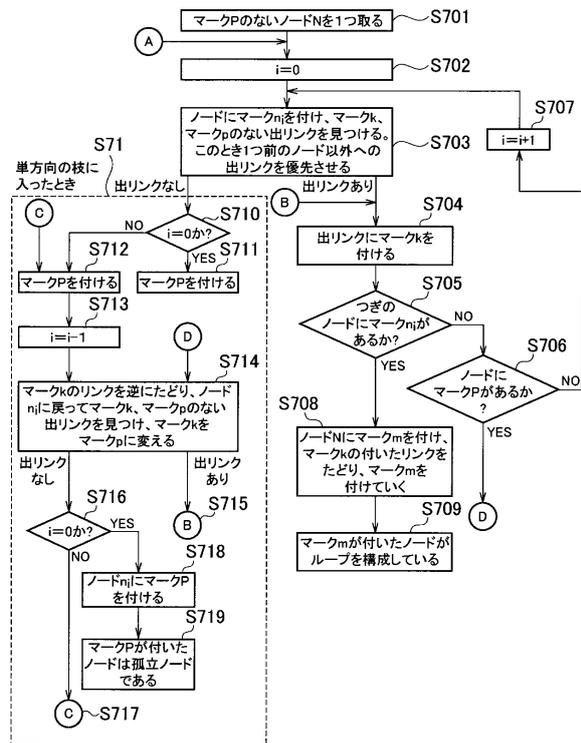
【 図 5 】

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100

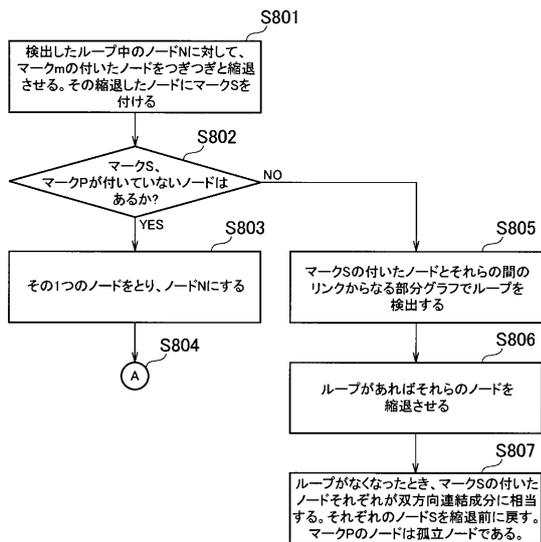
【 図 6 】



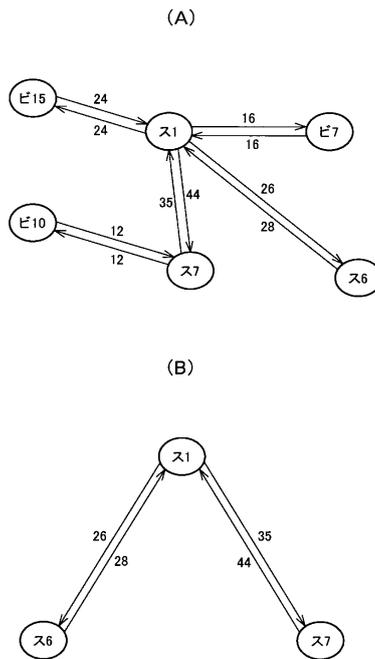
【 図 7 】



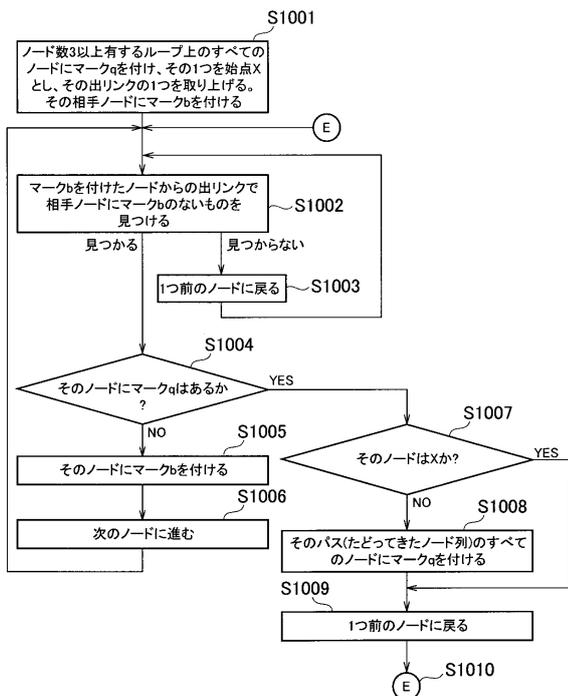
【 図 8 】



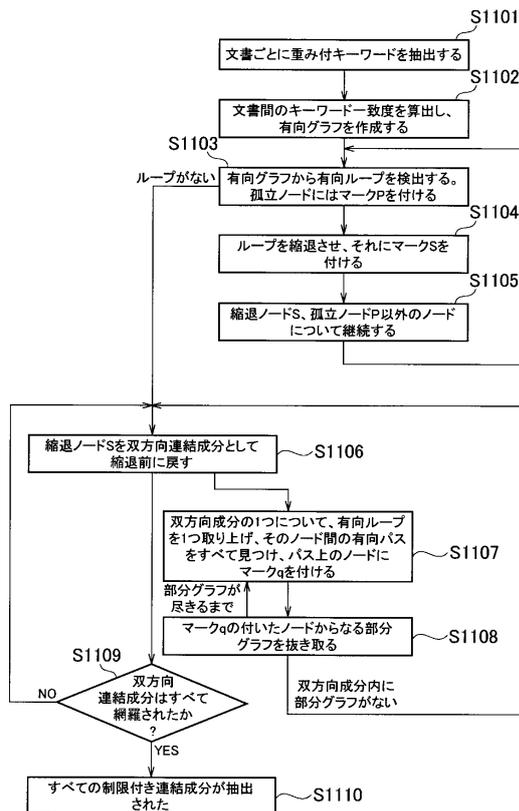
【 図 9 】



【 図 10 】



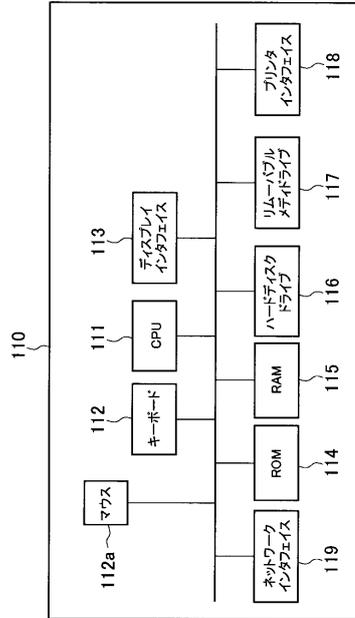
【 図 11 】



【 図 1 2 】

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	ピ7 景気の方へ回復の階段を上るには	改革, 必要, NHK	
	ピ10 NHK改革-視聴者の声を上げよう		
	ピ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, 野球留學生, 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-ファン無視の空振り		
	ス3 高校野球-フェアプレーを忘れるな		
	ス4 高校野球-後味の悪さはこれ限りに		
	ス8 ロッテ優勝-背番号26の勝利だ		
	ス6 揺れる阪神-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-あの青空をもう一度?		五輪, 世界, 日本, 必要

【 図 1 3 】



---

フロントページの続き

- (72)発明者 曾根原 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 - 2 0 - 1 1
- (72)発明者 池田 佳代  
東京都豊島区长崎 5 - 1 8 - 8
- Fターム(参考) 5B075 ND03 NK06 NK24