(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4370873号
(P4370873)
(45)発行日　平成21年11月25日(2009.11.25)　(24)登録日　平成21年9月11日(2009.9.11)

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

請求項の数 6　(全 12 頁)

(21)出願番号　特願2003-358081(P2003-358081)
(22)出願日　平成15年10月17日(2003.10.17)
(65)公開番号　特開2005-122550(P2005-122550A)
(43)公開日　平成17年5月12日(2005.5.12)
審査請求日　平成18年9月21日(2006.9.21)

(73)特許権者　000005496
富士ゼロックス株式会社
東京都港区赤坂九丁目７番３号
(74)代理人　100098084
弁理士　川▲崎▼　研二
(72)発明者　加藤　雅弘
神奈川県海老名市本郷２２７４番地　富士ゼロックス株式会社内

審査官　鈴木　和樹

最終頁に続く

(54)【発明の名称】文書分類装置、プログラムおよび文書分類方法

(57)【特許請求の範囲】
【請求項１】
文書の画像を表す画像データを取得する画像取得手段と、
前記画像データで表される画像を解析することによって、前記文書の各ページを構成する構成要素のレイアウトを表すレイアウト情報を取得し、前記構成要素のうち、各ページ内で文章が空間的に連続している領域である構成要素を、文章領域として抽出するレイアウト解析手段と、
前記文章領域に含まれる文字列を認識する文字認識手段と、
前記文字認識手段により認識された文字列から視覚的に強調された文字列を抽出し、抽出された文字列をキーワードとするキーワード抽出手段と、
前記レイアウト解析手段で抽出された複数の文章領域間の境界で各ページを分割し、各ページを根とし各文章領域を葉とし、前記境界の方向を基準とした木構造を用いて、各ページにおける複数の文章領域を階層的に表す構造データをページ毎に生成し、各文章領域に対応する特徴量をページ間で比較することで文章領域毎の包含関係を求め、前記包含関係に基づいて前記構造データにおける各文章領域の階層をページ間で調整する構造データ生成手段と、
前記構造データ生成手段で各文章領域の階層が調整された構造データに基づいて前記各キーワードの階層を求め、前記各キーワードの階層及び出現順を前記文書の論理構造として抽出する論理構造抽出手段と、
前記論理構造抽出手段で抽出された各文書の論理構造に含まれるキーワードの階層及び

出現順を比較することで前記各文書を分類して記憶する分類手段と
を有する文書分類装置。
【請求項２】
前記特徴量が、文章領域の位置、文章領域の大きさ、文字の大きさ、段組みの向き、１行あたりの平均的な文字数のうち少なくとも一つを含むことを特徴とする請求項１に記載の文書分類装置。
【請求項３】
前記構造データ生成手段が、垂直方向又は水平方向の少なくとも一方の境界で各ページを分割することを特徴とする請求項１に記載の文書分類装置。
【請求項４】
前記特徴量が、前記レイアウト解析手段により取得されたレイアウト情報を基に各文章領域に対応付けられる情報であることを特徴とする請求項１に記載の文書分類装置。
【請求項５】
コンピュータ装置を、
文書の画像を表す画像データを取得する画像取得手段と、
前記画像データで表される画像を解析することによって、前記文書の各ページを構成する構成要素のレイアウトを表すレイアウト情報を取得し、前記構成要素のうち、各ページ内で文章が空間的に連続している領域である構成要素を、文章領域として抽出するレイアウト解析手段と、
前記文章領域に含まれる文字列を認識する文字認識手段と、
前記文字認識手段により認識された文字列から視覚的に強調された文字列を抽出し、抽出された文字列をキーワードとするキーワード抽出手段と、
前記レイアウト解析手段で抽出された複数の文章領域間の境界で各ページを分割し、各ページを根とし各文章領域を葉とし、前記境界の方向を基準とした木構造を用いて、各ページにおける複数の文章領域を階層的に表す構造データをページ毎に生成し、各文章領域に対応する特徴量をページ間で比較することで文章領域毎の包含関係を求め、前記包含関係に基づいて前記構造データにおける各文章領域の階層をページ間で調整する構造データ生成手段と、
前記構造データ生成手段で各文章領域の階層が調整された構造データに基づいて前記各キーワードの階層を求め、前記各キーワードの階層及び出現順を前記文書の論理構造として抽出する論理構造抽出手段と、
前記論理構造抽出手段で抽出された各文書の論理構造に含まれるキーワードの階層及び出現順を比較することで前記各文書を分類して記憶する分類手段
として機能させるためのプログラム。
【請求項６】
ＣＰＵが文書の画像を表す画像データを取得する画像取得ステップと、
ＣＰＵが前記画像データで表される画像を解析することによって、前記文書の各ページを構成する構成要素のレイアウトを表すレイアウト情報を取得し、前記構成要素のうち、各ページ内で文章が空間的に連続している領域である構成要素を、文章領域として抽出するレイアウト解析ステップと、
ＣＰＵが前記文章領域に含まれる文字列を認識する文字認識ステップと、
ＣＰＵが前記文字認識ステップにより認識された文字列から視覚的に強調された文字列を抽出し、抽出された文字列をキーワードとするキーワード抽出ステップと、
ＣＰＵが前記レイアウト解析ステップで抽出された複数の文章領域間の境界で各ページを分割し、各ページを根とし各文章領域を葉とし、前記境界の方向を基準とした木構造を用いて、各ページにおける複数の文章領域を階層的に表す構造データをページ毎に生成し、各文章領域に対応する特徴量をページ間で比較することで文章領域毎の包含関係を求め、前記包含関係に基づいて前記構造データにおける各文章領域の階層をページ間で調整する構造データ生成ステップと、
ＣＰＵが前記構造データ生成ステップで各文章領域の階層が調整された構造データに基

づいて前記各キーワードの階層を求め、前記各キーワードの階層及び出現順を前記文書の論理構造として抽出する論理構造抽出ステップと、

ＣＰＵが前記論理構造抽出ステップで抽出された各文書の論理構造に含まれるキーワードの階層及び出現順を比較することで前記各文書を分類して画像蓄積部に記憶する分類ステップと

を有する文書分類方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、文書の画像を用いて文書を分類し、記憶する技術に関する。

【背景技術】

【０００２】

文書の種類を識別し、文書の種類を表す情報と文書画像とを対応付けてファイリングする文書管理システムが提案されている（例えば、特許文献１および２）。

特許文献１に記載の技術では、予め文書フォームとそれに対応する文書の種類が登録されており、入力された文書のフォームを識別し、そのフォームに対応する文書の種類を表す情報と当該文書画像とを対応付けて格納する。入力された文書のフォームに該当するフォームが登録されていない場合には、新たにそのフォームを登録する。これによって、予め登録されていないフォームを有する文書が入力された場合でも、その文書を分類してファイリングすることが可能になるとしている。

特許文献２に記載の技術では、入力された文書の画像から文章、図、表などの領域を抽出し、各領域のレイアウトを表す情報と当該文書画像とを対応付けて格納する。これによって、非定型、すなわちフォームによって識別できない文書であっても、そのレイアウトを表す情報によって分類してファイリングすることが可能になるとしている。

【特許文献１】特開２００２－２６９１２６号公報

【特許文献２】特開２００２－３４２３４３号公報

【発明の開示】

【発明が解決しようとする課題】

【０００３】

しかしながら、特許文献１および２の技術では、厳密な定型フォームを持たない文書の場合に登録されるフォームが際限なく増大してしまうおそれがある。例えば論文誌に掲載された論文のように書式が規定されてはいるものの、１件毎にページ数やレイアウトが異なる文書の場合がこれにあたる。

本発明は、上述した背景のもとになされたものであり、厳密な定型フォームではない文書を適切に分類することのできる技術の提供を目的とする。

【課題を解決するための手段】

【０００４】

上述の課題を解決するために、本発明は、文書の画像を表す画像データを取得する画像取得手段と、前記画像データで表される画像を解析することによって、前記文書の各ページを構成する構成要素のレイアウトを表すレイアウト情報を取得し、前記構成要素のうち、各ページ内で文章が空間的に連続している領域である構成要素を、文章領域として抽出するレイアウト解析手段と、前記文章領域に含まれる文字列を認識する文字認識手段と、前記文字認識手段により認識された文字列から視覚的に強調された文字列を抽出し、抽出された文字列をキーワードとするキーワード抽出手段と、前記レイアウト解析手段で抽出された複数の文章領域間の境界で各ページを分割し、各ページを根とし各文章領域を葉とし、前記境界の方向を基準とした木構造を用いて、各ページにおける複数の文章領域を階層的に表す構造データをページ毎に生成し、各文章領域に対応する特徴量をページ間で比較することで文章領域毎の包含関係を求め、前記包含関係に基づいて前記構造データにおける各文章領域の階層をページ間で調整する構造データ生成手段と、前記構造データ生成手段で各文章領域の階層が調整された構造データに基づいて前記各キーワードの階層を求

め、前記各キーワードの階層及び出現順を前記文書の論理構造として抽出する論理構造抽出手段と、前記論理構造抽出手段で抽出された各文書の論理構造に含まれるキーワードの階層及び出現順を比較することで前記各文書を分類して記憶する分類手段とを有する文書分類装置を提供する。

【０００５】

また、本発明は、コンピュータ装置を、文書の画像を表す画像データを取得する画像取得手段と、前記画像データで表される画像を解析することによって、前記文書の各ページを構成する構成要素のレイアウトを表すレイアウト情報を取得し、前記構成要素のうち、各ページ内で文章が空間的に連続している領域である構成要素を、文章領域として抽出するレイアウト解析手段と、前記文章領域に含まれる文字列を認識する文字認識手段と、前記文字認識手段により認識された文字列から視覚的に強調された文字列を抽出し、抽出された文字列をキーワードとするキーワード抽出手段と、前記レイアウト解析手段で抽出された複数の文章領域間の境界で各ページを分割し、各ページを根とし各文章領域を葉とし、前記境界の方向を基準とした木構造を用いて、各ページにおける複数の文章領域を階層的に表す構造データをページ毎に生成し、各文章領域に対応する特徴量をページ間で比較することで文章領域毎の包含関係を求め、前記包含関係に基づいて前記構造データにおける各文章領域の階層をページ間で調整する構造データ生成手段と、前記構造データ生成手段で各文章領域の階層が調整された構造データに基づいて前記各キーワードの階層を求め、前記各キーワードの階層及び出現順を前記文書の論理構造として抽出する論理構造抽出手段と、前記論理構造抽出手段で抽出された各文書の論理構造に含まれるキーワードの階層及び出現順を比較することで前記各文書を分類して記憶する分類手段として機能させるためのプログラムを提供する。

【０００６】

また、本発明は、ＣＰＵが文書の画像を表す画像データを取得する画像取得ステップと、ＣＰＵが前記画像データで表される画像を解析することによって、前記文書の各ページを構成する構成要素のレイアウトを表すレイアウト情報を取得し、前記構成要素のうち、各ページ内で文章が空間的に連続している領域である構成要素を、文章領域として抽出するレイアウト解析ステップと、ＣＰＵが前記文章領域に含まれる文字列を認識する文字認識ステップと、ＣＰＵが前記文字認識ステップにより認識された文字列から視覚的に強調された文字列を抽出し、抽出された文字列をキーワードとするキーワード抽出ステップと、ＣＰＵが前記レイアウト解析ステップで抽出された複数の文章領域間の境界で各ページを分割し、各ページを根とし各文章領域を葉とし、前記境界の方向を基準とした木構造を用いて、各ページにおける複数の文章領域を階層的に表す構造データをページ毎に生成し、各文章領域に対応する特徴量をページ間で比較することで文章領域毎の包含関係を求め、前記包含関係に基づいて前記構造データにおける各文章領域の階層をページ間で調整する構造データ生成ステップと、ＣＰＵが前記構造データ生成ステップで各文章領域の階層が調整された構造データに基づいて前記各キーワードの階層を求め、前記各キーワードの階層及び出現順を前記文書の論理構造として抽出する論理構造抽出ステップと、ＣＰＵが前記論理構造抽出ステップで抽出された各文書の論理構造に含まれるキーワードの階層及び出現順を比較することで前記各文書を分類して画像蓄積部に記憶する分類ステップとを有する文書分類方法を提供する。

【発明の効果】

【０００８】

本発明によれば、厳密な定型フォームではない文書を適切に分類することができる。文章領域のレイアウト上の階層に基づいてキーワードの論理レベルを決定し、文書全体の論理構造を求め、異なる文書間で論理構造を比較するから、文書の論理構造に着目した分類が可能となる。また、文書のカテゴリーを表す分類情報を文書画像と対応付けて記憶するから、カテゴリー毎に文書を検索することが可能となる。

【発明を実施するための最良の形態】

【０００９】

以下、図面を参照して、本発明の実施の形態について説明する。

［構成］

図１は、文書分類装置１０のハードウェア構成を示す図である。ＲＯＭ（Read Only Memory）１０２には、プログラム１０Ｐが書き込まれている。ＣＰＵ（Central Processing Unit）１０１は、文書分類装置１０に電源（図示省略）が投入されると、ＲＯＭ１０２に書き込まれているプログラム１０Ｐを読み出し、ＲＡＭ（Random Access Memory）１０３をワークエリアとしてプログラム１０Ｐを実行する。ＣＰＵ１０１がプログラム１０Ｐを実行することによって、文書分類装置１０には、図８に示すモジュール群が仮想的に形成される。なお、外部の装置にプログラム１０Ｐを記憶させておき、通信網（図示省略）を介してプログラム１０ＰをダウンロードしてＲＯＭ１０２に記憶させることとしてもよい。

【００１０】

画像蓄積部１１７は、原稿の画像データを蓄積するハードディスクドライブである。画像処理部１１８はＣＰＵ１０１による制御の下で、画像蓄積部１１７に蓄積されている画像データを読み出し、各種の処理を行う。ＣＰＵ１０１、ＲＯＭ１０２，ＲＡＭ１０３、画像蓄積部１１７および画像処理部１１８はバス１１５に接続されている。

表示部１０５は、ＣＲＴ（Cathode Ray Tube）あるいは液晶パネルである。操作部１０７は、ポインティングデバイス（マウスあるいはデジタイザ）およびキーボードである。スキャナ１０９は、原稿を光学的に読み取り、画像信号を出力する。プリンタ１１１は、電子写真方式あるいはインクジェット方式のプリンタである。表示部１０５、操作部１０７、スキャナ１０９およびプリンタ１１１はそれぞれインターフェイス１０６，１０８，１１０，１１２を介してバス１１６に接続されており、バス１１６とバス１１５とはバスブリッジ１０４によって接続されている。バス１１６は、インターフェイス１１４を介してネットワーク１１３に接続されており、これによって文書分類装置１０と外部の装置との通信が可能となっている。

【００１１】

次に、ＣＰＵ１０１がプログラム１０Ｐを実行することによって文書分類装置１０に仮想的に形成されるモジュール群について、図８を用いて説明する。

画像取得手段２１は、文書の画像を表す画像データを取得する手段である。文書の画像を表す画像データとは、スキャナ１０９等の画像入力装置を用いて文書を走査することによって生成された画像データである。なお、画像データは、予め画像蓄積部１１７に格納されていてもよいし、外部の装置からネットワーク１１３を介して文書分類装置１０が受信することとしてもよい。

【００１２】

レイアウト解析手段２２は、画像データで表される画像を解析することによって文書の各ページを構成する構成要素のレイアウトを表すレイアウト情報を取得し、ページ内で文章が空間的に連続している文章領域を抽出する手段である。ここで構成要素とは、文章、図、表などである。本実施形態においては、文章が空間的に連続している領域を文章領域と呼ぶ。また、図によって占められる領域を図領域、表によって占められる領域を表領域と呼ぶ。画像データは、文書をスキャナ１０９で走査して得られた画素値によって構成されており、ページ記述言語などで記述された文書データが内包しているようなレイアウト情報を有していない。そのため、レイアウト解析手段では、既知のレイアウト解析技術を用いて、当該画像で表される文書のレイアウト情報を得る。レイアウト解析は、例えば特開２０００－９０１９４号公報に記載されている技術を用いて行う。この技術では、文書画像に含まれる文章が縦書きか横書きかを判定し、その判定結果によって文書を分割する境界を設定する。また、文書画像を構成する画素の投影分布を算出し、所定のしきい値に満たない頻度の区間を用いて分割境界を設定する。

レイアウト解析手段２２はこのようにして文書のレイアウト情報を取得し、各ページを文章領域、図領域および表領域の集合として認識する。

【００１３】

文字認識手段２３は、レイアウト解析手段２２により抽出された文章領域に含まれる文字列を認識する手段である。

キーワード抽出手段２４は、文字認識手段により認識された文字列から視覚的に強調された文字列を抽出し、抽出された文字列をキーワードとする手段である。キーワードの抽出は、例えば特開平９－２９７７６５号公報に記載されている方法を用いて行う。ここで、キーワードとは、何らかの方法により視覚的に強調されている文字列である。例えば、予め文字サイズの閾値を定めておき、この閾値を超える大きさの文字列を抽出する。あるいは、太字、斜体など、通常と異なるフォントを用いた文字列、枠で囲まれた文字列、下線を引かれた文字列などを抽出してもよい。

【００１４】

構造データ生成手段２５は、レイアウト解析手段２２で抽出された文章領域のレイアウト上の階層構造を表す構造データをページ毎に生成する手段である。構造データは、図４に示すように、各ページを根とする木構造によって形成されており、レイアウト解析により抽出された文章領域の各々は、構造データの葉の各々と対応付けられている。葉の各々には、レイアウト解析の際に得られたレイアウト情報、すなわち、各文章領域の位置、大きさ、行の向きなどの情報が対応付けられている。なお、図、表など文章以外の領域については、構造データに含めない。

【００１５】

論理構造抽出手段２６は、構造データ生成手段２５で生成された構造データとキーワード抽出手段２４で抽出されたキーワードとを用いて、文書の論理構造を抽出する手段である。

分類手段２７は、論理構造抽出手段２６で抽出された論理構造を用いて文書を分類して記憶する手段である。

なお、論理構造抽出手段２６および分類手段２７によって行われる処理については、動作の説明において詳述する。

【００１６】

［動作］

上記の構成からなる文書分類装置１０の動作について説明する。ただし、文書分類装置１０は、ハードウェアがソフトウェアを用いることによって動作する装置であるから、これ以降の説明においては、動作の主体を、仮想的に形成されるモジュールではなく、ハードウェアとする。

ここで、文書の例について説明する。図３は、文書分類装置１０による処理の対象とされる文書の例を示す図である。この例は、横書き２段組を基本とする書式に従って作成された文書（例えば、論文）の例である。

【００１７】

１ページ目と２ページ目とでは文書のレイアウトが異なることがわかる。１ページ目では、最上部に題目が大きな文字サイズで１行記載されており、その下に抄録が６行記載されている。さらにその下には、本文が２段組で左右の各段に１３行記載されている。２ページ目では、１ページ目の本文と同様に２段組で記載されており、同様のレイアウトのページがＬ－２ページ目まで続く。Ｌ－１ページ目では、ページ上方にグラフが挿入されている。Ｌページ目では、２ページ目と同様のレイアウトとなり、このレイアウトのページがＭ－３ページ目まで続く。Ｍ－２ページ目では、右の段の下部にクルマの図が挿入され、Ｍ－１ページ目では、ページ全体に表が記載されている。Ｍページ目では２ページ目と同様のレイアウトとなり、このレイアウトがＮ－１ページ目まで続く。そして、Ｎページ目では、右の段に著者の顔写真と紹介記事が記載されている。

【００１８】

図２は、ＣＰＵ１０１がプログラム１０Ｐを実行することによって行われる処理のフローを示す図である。ここでは、文書分類装置１０には電源が投入されており、ＣＰＵ１０１によってプログラム１０Ｐが実行されているものとする。

最初に、文書分類装置１０は画像データを取得する（ステップＳ０１）。ここでは、ま

ず、スキャナ１０９によって文書が読み取られ、文書の画像を表す画像信号が文書分類装置１０に送信される。文書分類装置１０は、スキャナ１０９から送信された画像信号を受信する。すると、ＣＰＵ１０１が、受信された画像信号に基づいて画像データを生成し、画像蓄積部１１７に格納する。そして、ＣＰＵ１０１は、画像蓄積部１１７に格納された画像データをＲＡＭ１０３上に展開する。

【００１９】

次に、ＣＰＵ１０１は、レイアウト解析を行って、文書を構成する各ページのレイアウトを表すレイアウト情報を取得する（ステップＳ０２）。図４は、図３に示した文書のレイアウト解析例（上段）およびレイアウト解析結果に基づいて作成された構造データの例（下段）を示す図である。各ページ中の矩形で囲まれた領域が文章領域、図領域、表領域を表しており、これらの中で番号が付されているものが文章領域である。本実施形態においては、文章領域のみを処理の対象とし、図領域および表領域は処理の対象としない。

次に、ＣＰＵ１０１は、文章領域に含まれる文字を認識するための処理を行い（ステップＳ０３）、認識された文字列からキーワードとなり得る文字列を抽出する（ステップＳ０４）。ここでは、特定の文字サイズ、特定のフォント、枠で囲まれた文字列、下線の引いてある文字列などを抽出する。

【００２０】

ステップＳ０３、ステップＳ０４の処理と並行して、ＣＰＵ１０１は、構造データの生成（ステップＳ０５）、構造データのレベル調整（ステップＳ０６）を行う。

まず、構造データの生成（ステップＳ０５）について説明する。図４に示すように、構造データは、ページの各々を根とする木構造を有している。木構造の階層は、以下のようにして決定される。１ページ目の例では、まず水平方向に３つの領域に分割された後、最も下の領域が垂直方向に２分割されたとみなされる。これを木構造で表すと、根の１つ下の階層には２つの葉と１つの中間接点が存在し、２つの葉は領域１と領域２に対応付けられる。中間接点の１つ下の階層にはさらに２つの葉が存在し、２つの葉は領域３と領域４に対応付けられる。

【００２１】

一方、Ｍ－２ページ目の例では、まず垂直方向に２つの領域に分割された後、右側の領域が水平方向に２分割されたものとみなされるが、本実施形態においては図領域および表領域を処理の対象としないため、領域２の下方に位置する図領域は無視される。従って、Ｍ－２ページ目では、領域１と領域２が等しい階層であるとみなされることとなる。Ｍ－１ページ目の表領域、Ｎページ目の図領域（顔写真）も同様に無視され、図４に示す構造データが得られる。

【００２２】

次に、レベル調整（ステップＳ０６）について説明する。図７は、構造データのレベル調整のフローを示す図である。まず、ステップＳ６０１では、文書の１ページ目の構造データを初期値とする。ステップＳ６０２では、２ページ目以降のページについて、直前のページとの間でノード間の対応付けを行う。各ノードには、当該ノードに対応する文章領域がレイアウトされている位置、領域の大きさ、領域内の平均的な文字サイズ、段組みが横方向の分割であるか縦方向の分割であるか、１行あたりの平均的な文字数といった、レイアウト解析の際に得られる種々の特徴量が対応付けられて記憶されている。ＣＰＵ１０１は、注目ページとその直前のページとの特徴量を比較し、領域毎の包含関係に基づいて、ノード間の対応付けを行う。

【００２３】

ここで、ノード間の対応付けについて説明する。図４によると、１ページ目の領域３が２ページ目の領域１に包含されている。同様に、１ページ目の領域４が２ページ目の領域２に包含されている（１対１の包含）。従って、２ページ目の領域１および２は、１ページ目の領域３および４と同等以上の階層に位置付けられることになる。しかし、１ページ目の領域１および２は、それぞれが２ページ目の領域１および２の一部を包含しているため、２ページ目の領域１および２は、１ページ目の領域１および２よりも下の階層に位置

付けられることとなる。よって、図５に示すように、２ページ目の領域１および２は、１ページ目の領域３および４と等しい階層（Level_2）に位置付けられる。

一方、Ｍページ目とＮページ目の例では、Ｍページ目の領域２がＮページ目の領域２、３および４を包含している（１対多の包含）。従って、Ｎページ目の領域２、３および４は、Ｍページ目の領域２よりも下の階層に位置付けられることとなる。よって、図５に示すように、Ｎページ目の領域２、３および４は、Ｍページ目の領域１および２の１つ下の階層（Level_3）に位置付けられる。

【００２４】

他方、上述のようなノード間の対応付けの結果、直前のページの構造データにおいてLevel_0を下位の階層へ変更する必要が生じた場合、すなわち、それ以前のページに存在しない上位のノードが注目ページに出現した場合（例えば、図４において１ページ目と２ページ目が入れ替わっている場合）には（ステップＳ６０３：ＹＥＳ）、ステップＳ６０４で構造データの階層の変更が必要な先頭ページまでさかのぼり、ステップＳ６０５で当該先頭ページ以降、未処理ページまでの構造データの階層を変更する。具体的には、仮想ノードを最上位に挿入し、全体の階層を下位にシフトさせる。一方、ステップＳ６０３で階層の変更が必要でないと判定された場合には、ステップＳ６０６に進み、注目ページの構造データの階層の変更が必要でないかどうかを判定し、変更が必要であるならばステップＳ６０７にて注目ページの構造データの階層を変更する。ステップＳ６０８では、全ページについてノード間の対応付けが完了したか否かを判定し、完了していないならばステップＳ６０２に戻る。

【００２５】

ノード間の対応付けが完了したならば、図２のステップＳ０７に進み、キーワードの階層を求める。このキーワードはステップＳ０４で文章領域から抽出されたキーワードである。このキーワードが属する文章領域の階層は、ステップＳ０６で求められた当該文章領域の階層である。

ステップＳ０８では、特定の階層のキーワードを用いて、文書の論理構造を抽出する。図６は、文書の論理構造を抽出した例を示す図である。この例では、文書Ａおよび文書Ｂはともに公開特許公報である。図６には、「要約」。「特許請求の範囲」、「技術分野」、…と続く特許公報の記載項目の見出しが出現順に示されている。公開特許公報は、帳票のように全ページにおいて文字枠が厳密に既定されているものではなく、ページ数やレイアウトが１件１件異なるものである。また、数式や表が記載された公報と記載されていない公報が存在する。また、各項目毎の文章の分量も１件毎に異なる。このような違いを含んだ文書の場合、従来知られているような厳密なフォーム認識では異なる種類の文書とみなされることになる。しかしながら、文書Ａと文書Ｂの論理構造に着目すれば、両者は明らかに同じ種類の文書である。このように、文書の論理構造を比較することによって、フォームが厳密には一致していない文書や、ページ数が１件毎に異なる文書であっても、同種の文書として分類することができる。ステップＳ０９では、このようにして分類された文書のカテゴリーを表す情報と当該文書の画像データとを対応付けて画像蓄積部１１７に格納する。

【００２６】

以上説明したように、本発明によれば、厳密な定型フォームではない文書を適切に分類することができる。文章領域のレイアウト上の階層に基づいてキーワードの論理レベルを決定し、文書全体の論理構造を求め、異なる文書間で論理構造を比較するから、文書の論理構造に着目した分類が可能となる。また、文書のカテゴリーを表す分類情報を文書画像と対応付けて記憶するから、カテゴリー毎に文書を検索することが可能となる。

【００２７】

［変形例］

以上説明した形態に限らず、本発明は種々の形態で実施可能である。例えば、上述の実施形態を以下のように変形した形態でも実施可能である。

スキャナが接続された１または複数のパーソナルコンピュータをネットワークを介して

文書分類装置１０に接続し、スキャナで読み込まれた文書の画像データを文書分類装置１０に送信し、文書分類装置１０において文書の分類および格納を行うようにしてもよい。このようにすれば、例えば、オフィス内の別々の場所に分散して保管されている文書を文書分類装置１０で集中管理することが可能となる。

文書の論理構造は、特定の階層のキーワードではなく、すべての階層のキーワードをその階層を表す情報とともに表したものであってもよい。

【図面の簡単な説明】

【００２８】

【図１】本発明の一実施形態に係る文書分類装置の構成を示す図である。

【図２】ＣＰＵがプログラムを実行することによって行われる処理のフローを示す図である。

【図３】文書分類装置による処理の対象とする文書の例を示す図である。

【図４】構造データの例を示す図である。

【図５】構造データのレベル調整の例を示す図である。

【図６】文書の論理構造の抽出例を示す図である。

【図７】構造データのレベル調整のフローを示す図である。

【図８】ＣＰＵがプログラムを実行することによって形成される仮想的モジュールを示す図である。

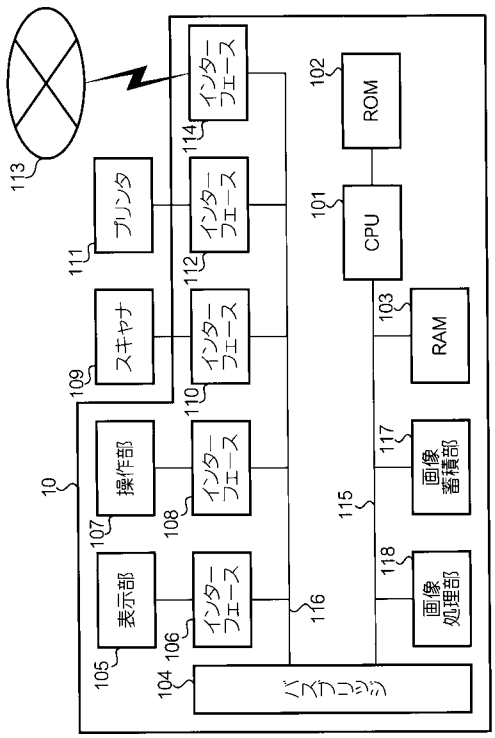
【符号の説明】

【００２９】

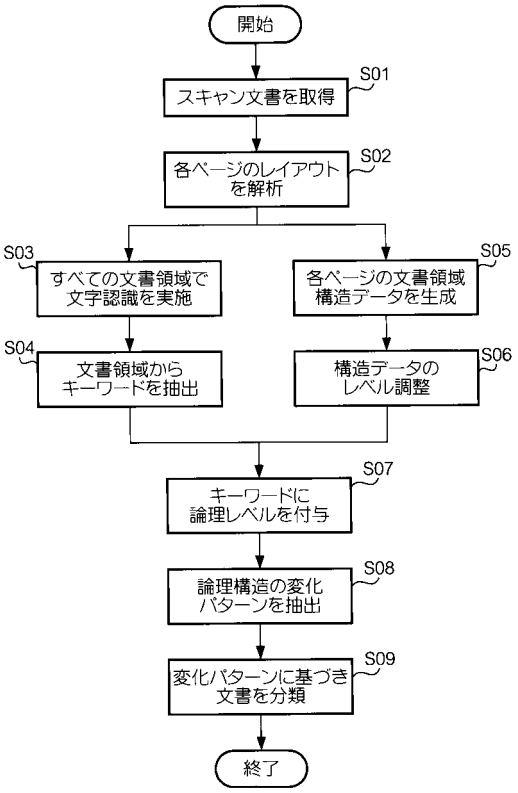
１０…文書分類装置、１０１…ＣＰＵ、１０２…ＲＯＭ，１０３…ＲＡＭ、１０５…表示部、１０７…操作部、１０９…スキャナ、１１１…プリンタ、１１３…ネットワーク、１１７…画像蓄積部、１１８…画像処理部、

２１…画像取得手段、２２…レイアウト解析手段、２３…文字認識手段、２４…キーワード抽出手段、２５…構造データ生成手段、２６…論理構造抽出手段、２７…分類手段。

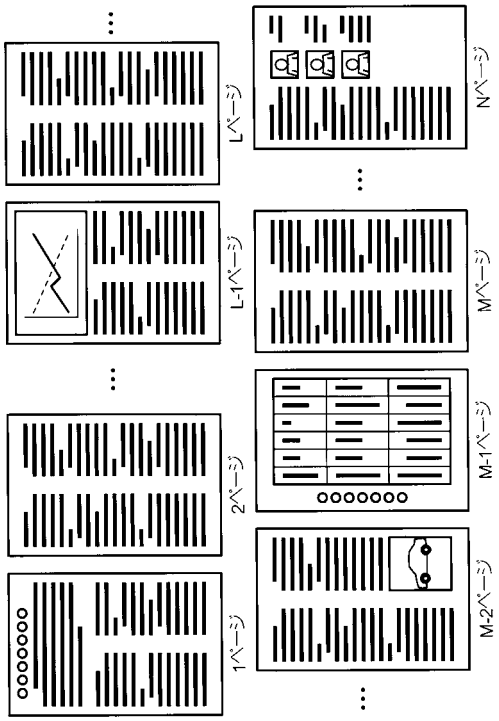
【図１】



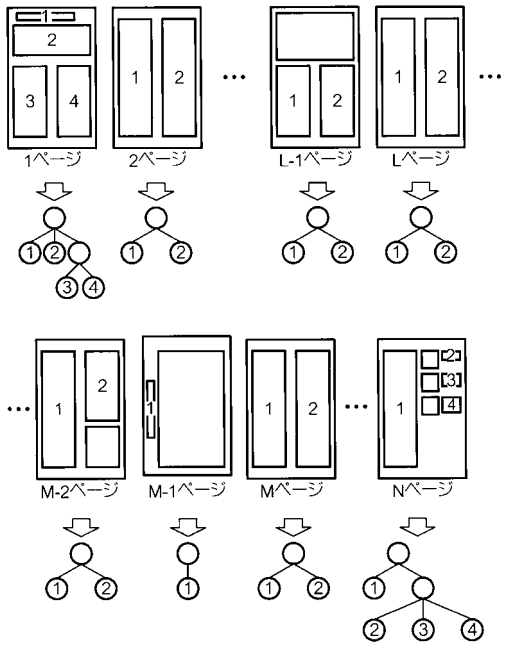
【図２】



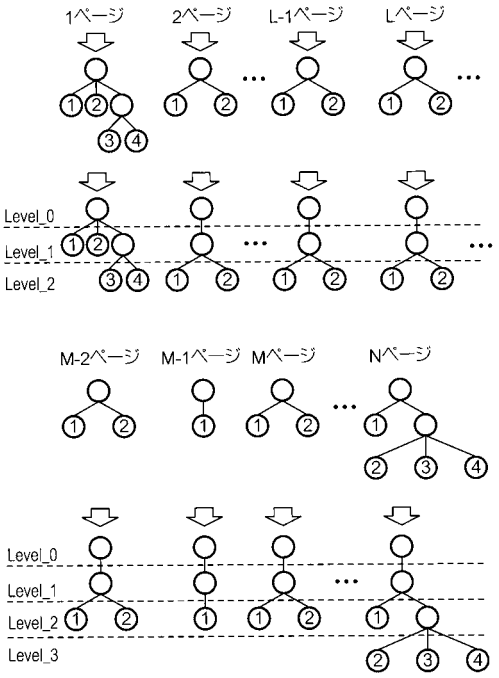
【図３】


1ページ
2ページ
L-1ページ
Lページ
M-2ページ
M-1ページ
Mページ
Nページ

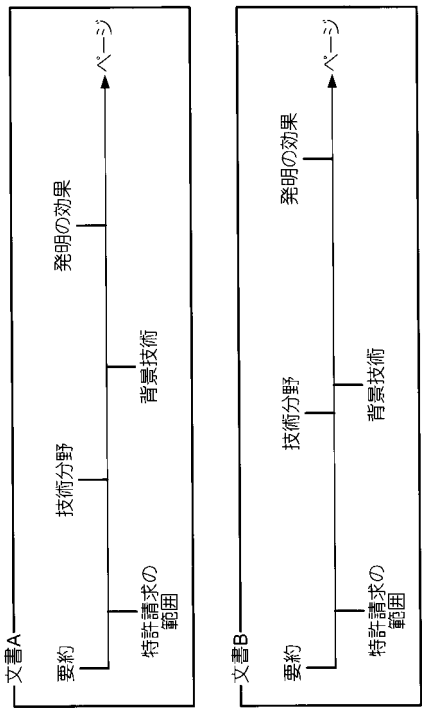
【図４】


1ページ
2ページ
L-1ページ
Lページ
M-2ページ
M-1ページ
Mページ
Nページ

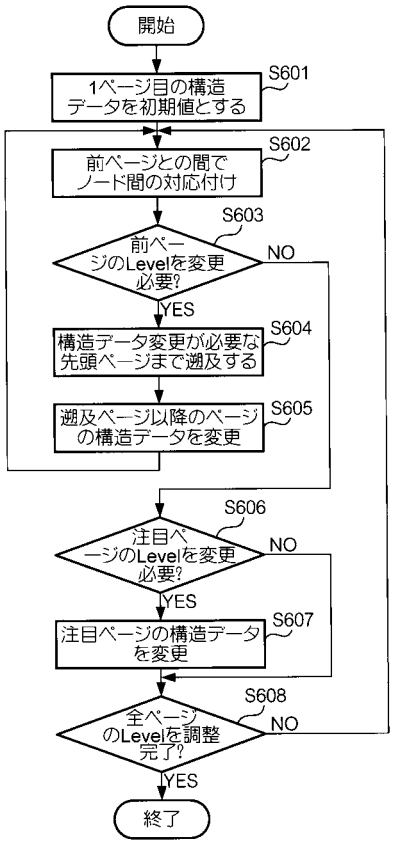
【図５】


1ページ
2ページ
L-1ページ
Lページ
Level_0
Level_1
Level_2
M-2ページ
M-1ページ
Mページ
Nページ
Level_0
Level_1
Level_2
Level_3

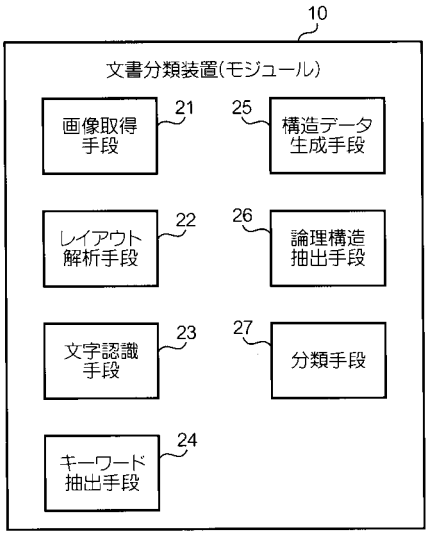
【図６】


文書A
要約
技術分野
発明の効果
特許請求の範囲
背景技術
ページ
文書B
要約
技術分野
発明の効果
特許請求の範囲
背景技術
ページ

【図７】


開始
1ページ目の構造データを初期値とする
S601
前ページとの間でノード間の対応付け
S602
前ページのLevelを変更必要?
S603
NO
YES
構造データ変更が必要な先頭ページまで遡及する
S604
遡及ページ以降のページの構造データを変更
S605
注目ページのLevelを変更必要?
S606
NO
YES
注目ページの構造データを変更
S607
全ページのLevelを調整完了?
S608
NO
YES
終了

【図８】


10
文書分類装置(モジュール)
画像取得手段
21
レイアウト解析手段
22
文字認識手段
23
キーワード抽出手段
24
構造データ生成手段
25
論理構造抽出手段
26
分類手段
27

フロントページの続き

(56)参考文献 特開平０６－２１４９８３（ＪＰ，Ａ）
特開２０００－０６７０８０（ＪＰ，Ａ）
特開２００３－２８８３３４（ＪＰ，Ａ）
特開２００２－３４２３４３（ＪＰ，Ａ）
特開平０９－２９７７６５（ＪＰ，Ａ）
特開平０１－１８３７８４（ＪＰ，Ａ）
黄瀬浩一，文書画像理解の目指すもの，電子情報通信学会技術研究報告（ＰＲＭＵ９７－２４０～２５０），日本，社団法人電子情報通信学会，１９９８年 2月２０日，第９７巻，第５５９号，ｐ．５５－６２

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０
Ｇ０６Ｆ １７／２１