(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2009-122760
(P2009-122760A)
(43)公開日　平成21年6月4日(2009.6.4)

(51)Int.Cl.			F I			テーマコード（参考）
G06T	1/00	(2006.01)	G06T	1/00	200A	5B050
G06F	3/048	(2006.01)	G06F	3/048	656C	5E501

審査請求 未請求　請求項の数 16　ＯＬ　（全 19 頁）

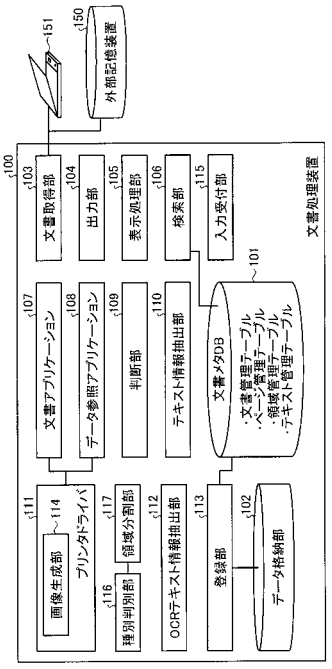
(21)出願番号　特願2007-293394 (P2007-293394)
(22)出願日　平成19年11月12日 (2007.11.12)

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(74)代理人　100089118
弁理士　酒井　宏明
(72)発明者　岩崎　雅二郎
東京都大田区中馬込１丁目３番６号　株式会社リコー内

Ｆターム(参考)　5B050 BA10 BA16 BA20 CA07 DA10 EA06 GA08
5E501 AA01 AA14 AC15 AC33 BA05 CA03 CA04

(54)【発明の名称】文書処理装置、文書処理方法及び文書処理プログラム

(57)【要約】
【課題】文字の抽出精度を向上させる。
【解決手段】少なくとも２以上のアプリケーションで作成された文書データを取得する文書取得部と、文書データから、文書画像データを生成する画像生成部と、文書情報を、アプリケーション別に領域として分割する領域分割部と、分割された領域毎に、文字コードを抽出可能であるか否か判断する判断部と、文字コードを抽出可能と判断された領域から、文字コード情報を抽出するテキスト情報抽出部と、文字コードが抽出できないと判断された場合、文書画像に対して文字認識処理を施して得られた文字コードを第２の文字情報として抽出する文字認識テキスト情報抽出部と、第１の文字情報と、第２の文字情報と、文書データと、文書画像データと、を対応付けて記憶する文書メタＤＢと、を備える。
【選択図】　図１



【特許請求の範囲】
【請求項１】
少なくとも２以上のアプリケーションで作成された文書情報を取得する文書情報取得手段と、
取得した前記文書情報から、文書画像を生成する画像生成手段と、
取得した前記文書情報を、アプリケーション別に領域として分割する領域分割手段と、
分割された前記領域毎に、文字コードを抽出可能であるか否か判断する判断手段と、
前記文字コードを抽出可能と判断された前記領域から、第１の文字情報を抽出する第１の文字情報抽出手段と、
前記文字コードが抽出できないと判断された場合、前記文書画像に対して文字認識処理を施して得られた文字コードを第２の文字列情報として抽出する第２の文字情報抽出手段と、
前記第１の文字情報と、前記第２の文字情報と、前記文書情報及び前記文書画像の少なくともいずれか一方と、を対応付けて記憶する記憶手段と、
を備えることを特徴とする文書処理装置。
【請求項２】
検索条件に基づいて、前記記憶手段の前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上に対して検索を行う検索手段と、
検索条件を満たした、前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上と、前記記憶手段で対応付けられている前記文書情報及び前記文書画像の少なくともいずれか一方を出力する出力手段と、
をさらに備えることを特徴とする請求項１に記載の文書処理装置。
【請求項３】
前記第１の文字情報抽出手段は、さらに前記第１の文字情報の位置を示す第１の位置情報を抽出し、
前記記憶手段は、さらに前記第１の位置情報を対応付けて記憶すること、
を特徴とする請求項１又は２に記載の文書処理装置。
【請求項４】
前記出力処理手段は、前記第１の文字情報を、前記記憶手段で対応付けられた前記第１の位置情報に配置して出力すること、
を特徴とする請求項３に記載の文書処理装置。
【請求項５】
前記第２の文字情報抽出手段は、さらに前記第２の文字情報の位置を示す第２の位置情報を抽出し、
前記記憶手段は、さらに前記第２の位置情報を対応付けて記憶すること、
を特徴とする請求項１乃至４のいずれか一つに記載の文書処理装置。
【請求項６】
前記表示処理手段は、前記第２の文字情報を、前記記憶手段で対応付けられた前記第２の位置情報に配置して出力すること、
を特徴とする請求項５に記載の文書処理装置。
【請求項７】
前記取得した文書情報を読み込むアプリケーションをさらに備え、
前記画像生成手段は、前記文書処理装置内の印刷ドライバ内に含まれ、前記文書情報を読み込んだアプリケーションによる出力命令に従って前記文書画像データを生成すること、
を特徴とする請求項１乃至６のいずれか一つに記載の文書処理装置。
【請求項８】
文書情報取得手段が、少なくとも２以上のアプリケーションで作成された文書情報を取得する文書情報取得ステップと、
画像生成手段が、取得した前記文書情報から、文書画像を生成する画像生成ステップと

、
領域分割手段が、取得した前記文書情報を、アプリケーション別に領域として分割する領域分割ステップと、
判断手段が、分割された前記領域毎に、文字コードを抽出可能であるか否か判断する判断ステップと、
第１の文字情報抽出手段が、前記文字コードを抽出可能と判断された前記領域から、第１の文字情報を抽出する第１の文字情報抽出ステップと、
第２の文字情報抽出手段が、前記文字コードが抽出できないと判断された場合、前記文書画像に対して文字認識処理を施して得られた文字コードを第２の文字列情報として抽出する第２の文字情報抽出ステップと、
前記第１の文字情報と、前記第２の文字情報と、前記文書情報及び前記文書画像の少なくともいずれか一方と、を対応付けて記憶手段に記憶する記憶ステップと、
を有することを特徴とする文書処理方法。
【請求項９】
検索手段が、検索条件に基づいて、前記記憶手段の前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上に対して検索を行う検索ステップと、
出力手段が、検索条件を満たした、前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上と、前記記憶手段で対応付けられている前記文書情報及び前記文書画像の少なくともいずれか一方を出力する出力ステップと、
をさらに有することを特徴とする請求項８に記載の文書処理方法。
【請求項１０】
前記第１の文字情報抽出ステップは、さらに前記第１の文字情報の位置を示す第１の位置情報を抽出し、
前記記憶ステップは、さらに前記第１の位置情報を対応付けて記憶すること、
を特徴とする請求項８又は９に記載の文書処理方法。
【請求項１１】
前記出力処理ステップは、前記第１の文字情報を、前記記憶手段で対応付けられた前記第１の位置情報に配置して出力すること、
を特徴とする請求項１０に記載の文書処理方法。
【請求項１２】
前記第２の文字情報抽出ステップは、さらに前記第２の文字情報の位置を示す第２の位置情報を抽出し、
前記記憶ステップは、さらに前記第２の位置情報を対応付けて記憶すること、
を特徴とする請求項８乃至１１のいずれか一つに記載の文書処理方法。
【請求項１３】
前記表示処理ステップは、前記第２の文字情報を、前記記憶手段で対応付けられた前記第２の位置情報に配置して出力すること、
を特徴とする請求項１２に記載の文書処理方法。
【請求項１４】
アプリケーションが、前記取得した文書情報を読み込む読込ステップをさらに有し、
前記画像生成ステップは、前記文書処理装置内の印刷ドライバ内に含まれた前記画像生成手段が、前記文書情報を読み込んだアプリケーションによる出力命令に従って前記文書画像データを生成すること、
を特徴とする請求項８乃至１３のいずれか一つに記載の文書処理方法。
【請求項１５】
コンピュータを、
少なくとも２以上のアプリケーションで作成された文書情報を取得する文書情報取得手段と、
取得した前記文書情報から、文書画像を生成する画像生成手段と、
取得した前記文書情報を、アプリケーション別に領域として分割する領域分割手段と、

分割された前記領域毎に、文字コードを抽出可能であるか否か判断する判断手段と、
前記文字コードを抽出可能と判断された前記領域から、第１の文字情報を抽出する第１の文字情報抽出手段と、
前記文字コードが抽出できないと判断された場合、前記文書画像に対して文字認識処理を施して得られた文字コードを第２の文字列情報として抽出する第２の文字情報抽出手段と、
前記第１の文字情報と、前記第２の文字情報と、前記文書情報及び前記文書画像の少なくともいずれか一方と、を対応付けて記憶手段に記憶させる登録手段と、
として機能させることを特徴とする文書処理プログラム。
【請求項１６】
検索条件に基づいて、前記記憶手段の前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上に対して検索を行う検索手段と、
検索条件を満たした、前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上と、前記記憶手段で対応付けられている前記文書情報及び前記文書画像の少なくともいずれか一方を出力する出力手段と、
をさらに機能させることを特徴とする請求項１５に記載の文書処理プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、文書処理装置、文書処理方法及び文書処理プログラムに関するものである。
【背景技術】
【０００２】
近年、コンピュータ関連技術の向上、ネットワーク環境が整備によって文書の電子化が進んでいる。これによりオフィスのペーパレス化が促進されている。
【０００３】
そして、文書の電子化が進むにつれ、電子化された文書データを一括管理し、検索を行えるようにしたいという要求がある。
【０００４】
そこで、特許文献１は、描画コードから、テキスト情報、すなわち文字コードを抽出し、抽出したテキスト情報と、描画コードから生成された文書画像データとを対応付けている。この描画コードから文書画像データを生成している以上、文字コード等を含む中間データであると考えられる。このため、描画コードから容易に文字コードを抽出している。
【０００５】
【特許文献１】特開平８－２１２３３１号公報
【発明の開示】
【発明が解決しようとする課題】
【０００６】
ところで、文書データに画像データとして図や表を入れたりする場合も多い。このような図や表は画像データ等で貼り付けられている。また、図や表に限らず、ＨＴＭＬで表現されるＷｅｂのページの場合には見栄えを重視して文字も画像として入れる場合が頻繁に見受けられる。
【０００７】
しかしながら、引用文献１に記載された技術では、描画コードから文字コードを抽出するのでは、図や表を示す画像データが埋め込まれている場合に抽出できないという問題がある。
【０００８】
一方で、文書データから生成した文書画像データに対して、文字認識を掛けて、図や表から画像データを抽出するのでは文字の抽出精度が落ちるという問題がある。
【０００９】
本発明は、上記に鑑みてなされたものであって、文字の抽出精度を向上させた文書処理

装置、文書処理方法及び文書処理プログラムを提供することを目的とする。
【課題を解決するための手段】
【００１０】
上述した課題を解決し、目的を達成するために、請求項１にかかる発明は、少なくとも２以上のアプリケーションで作成された文書情報を取得する文書情報取得手段と、取得した前記文書情報から、文書画像を生成する画像生成手段と、取得した前記文書情報を、アプリケーション別に領域として分割する領域分割手段と、分割された前記領域毎に、文字コードを抽出可能であるか否か判断する判断手段と、前記文字コードを抽出可能と判断された前記領域から、第１の文字情報を抽出する第１の文字情報抽出手段と、前記文字コードが抽出できないと判断された領域に含まれる第２の文字情報を、前記文書画像に対して文字認識処理を施して得られた文字コードから抽出する第２の文字情報抽出手段と、前記第１の文字情報と、前記第２の文字情報と、前記文書情報及び前記文書画像の少なくともいずれか一方と、を対応付けて記憶する記憶手段と、を備えることを特徴とする。
【００１１】
また、請求項２にかかる発明は、請求項１にかかる発明において、検索条件に基づいて、前記記憶手段の前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上に対して検索を行う検索手段と、検索条件を満たした、前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上と、前記記憶手段で対応付けられている前記文書情報及び前記文書画像の少なくともいずれか一方を出力する出力手段と、をさらに備えることを特徴とする。
【００１２】
また、請求項３にかかる発明は、請求項１又は２にかかる発明において、前記第１の文字情報抽出手段は、さらに前記第１の文字情報の位置を示す第１の位置情報を抽出し、前記記憶手段は、さらに前記第１の位置情報を対応付けて記憶すること、を特徴とする。
【００１３】
また、請求項４にかかる発明は、請求項３にかかる発明において、前記出力処理手段は、前記第１の文字情報を、前記記憶手段で対応付けられた前記第１の位置情報に配置して出力すること、を特徴とする。
【００１４】
また、請求項５にかかる発明は、請求項１乃至４のいずれか一つにかかる発明において、前記第２の文字情報抽出手段は、さらに前記第２の文字情報の位置を示す第２の位置情報を抽出し、前記記憶手段は、さらに前記第２の位置情報を対応付けて記憶すること、を特徴とする。
【００１５】
また、請求項６にかかる発明は、請求項５にかかる発明において、前記表示処理手段は、前記第２の文字情報を、前記記憶手段で対応付けられた前記第２の位置情報に配置して出力すること、を特徴とする。
【００１６】
また、請求項７にかかる発明は、請求項１乃至６のいずれか一つにかかる発明において、前記取得した文書情報を読み込むアプリケーションをさらに備え、前記画像生成手段は、前記文書処理装置内の印刷ドライバ内に含まれ、前記文書情報を読み込んだアプリケーションによる出力命令に従って前記文書画像データを生成すること、を特徴とする。
【００１７】
また、請求項８にかかる発明は、文書情報取得手段が、少なくとも２以上のアプリケーションで作成された文書情報を取得する文書情報取得ステップと、画像生成手段が、取得した前記文書情報から、文書画像を生成する画像生成ステップと、領域分割手段が、取得した前記文書情報を、アプリケーション別に領域として分割する領域分割ステップと、判断手段が、分割された前記領域毎に、文字コードを抽出可能であるか否か判断する判断ステップと、第１の文字情報抽出手段が、前記文字コードを抽出可能と判断された前記領域から、第１の文字情報を抽出する第１の文字情報抽出ステップと、第２の文字情報抽出手

段が、前記文字コードが抽出できないと判断された領域に含まれる第２の文字情報を、前記文書画像に対して文字認識処理を施して得られた文字コードから抽出する第２の文字情報抽出ステップと、前記第１の文字情報と、前記第２の文字情報と、前記文書情報及び前記文書画像の少なくともいずれか一方と、を対応付けて記憶手段に記憶する記憶ステップと、を有することを特徴とする。

【００１８】

また、請求項９にかかる発明は、請求項８にかかる発明において、検索手段が、検索条件に基づいて、前記記憶手段の前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上に対して検索を行う検索ステップと、出力手段が、検索条件を満たした、前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上と、前記記憶手段で対応付けられている前記文書情報及び前記文書画像の少なくともいずれか一方を出力する出力ステップと、をさらに有することを特徴とする。

【００１９】

また、請求項１０にかかる発明は、請求項８又は９にかかる発明において、前記第１の文字情報抽出ステップは、さらに前記第１の文字情報の位置を示す第１の位置情報を抽出し、前記記憶ステップは、さらに前記第１の位置情報を対応付けて記憶すること、を特徴とする。

【００２０】

また、請求項１１にかかる発明は、請求項１０にかかる発明において、前記出力処理ステップは、前記第１の文字情報を、前記記憶手段で対応付けられた前記第１の位置情報に配置して出力すること、を特徴とする。

【００２１】

また、請求項１２にかかる発明は、請求項８乃至１１のいずれか一つにかかる発明において、前記第２の文字情報抽出ステップは、さらに前記第２の文字情報の位置を示す第２の位置情報を抽出し、前記記憶ステップは、さらに前記第２の位置情報を対応付けて記憶すること、を特徴とする。

【００２２】

また、請求項１３にかかる発明は、請求項１２にかかる発明において、前記表示処理ステップは、前記第２の文字情報を、前記記憶手段で対応付けられた前記第２の位置情報に配置して出力すること、を特徴とする。

【００２３】

また、請求項１４にかかる発明は、請求項８乃至１３のいずれか一つにかかる発明において、アプリケーションが、前記取得した文書情報を読み込む読込ステップをさらに有し、前記画像生成ステップは、前記文書処理装置内の印刷ドライバ内に含まれた前記画像生成手段が、前記文書情報を読み込んだアプリケーションによる出力命令に従って前記文書画像データを生成すること、を特徴とする。

【００２４】

また、請求項１５にかかる発明は、コンピュータを、少なくとも２以上のアプリケーションで作成された文書情報を取得する文書情報取得手段と、取得した前記文書情報から、文書画像を生成する画像生成手段と、取得した前記文書情報を、アプリケーション別に領域として分割する領域分割手段と、分割された前記領域毎に、文字コードを抽出可能であるか否か判断する判断手段と、前記文字コードを抽出可能と判断された前記領域から、第１の文字情報を抽出する第１の文字情報抽出手段と、前記文字コードが抽出できないと判断された領域に含まれる第２の文字情報を、前記文書画像に対して文字認識処理を施して得られた文字コードから抽出する第２の文字情報抽出手段と、前記第１の文字情報と、前記第２の文字情報と、前記文書情報及び前記文書画像の少なくともいずれか一方と、を対応付けて記憶手段に記憶させる登録手段と、として機能させることを特徴とする。

【００２５】

また、請求項１６にかかる発明は、請求項１５にかかる発明において、検索条件に基づいて、前記記憶手段の前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上に

対して検索を行う検索手段と、検索条件を満たした、前記第１の文字情報及び前記第２の文字情報のいずれか一つ以上と、前記記憶手段で対応付けられている前記文書情報及び前記文書画像の少なくともいずれか一方を出力する出力手段と、をさらに機能させることを特徴とする。

【発明の効果】

【００２６】

本願発明によれば、文書情報を作成したアプリケーションに拘わらず、文字情報を抽出可能にすると共に、抽出される文字情報の精度を向上させるという効果を奏する。

【発明を実施するための最良の形態】

【００２７】

以下に添付図面を参照して、この発明にかかる文書処理装置、文書処理方法及び文書処理プログラムの最良な実施の形態を詳細に説明する。

【００２８】

（第１の実施の形態）

図１は、第１の実施の形態にかかる文書処理装置１００の構成を示すブロック図である。本図に示すように文書処理装置１００の内部は、文書メタＤＢ１０１と、データ格納部１０２と、文書取得部１０３と、出力部１０４と、表示処理部１０５と、検索部１０６と、文書アプリケーション１０７と、データ参照アプリケーション１０８と、判断部１０９と、テキスト情報抽出部１１０と、プリンタドライバ１１１と、文字認識テキスト情報抽出部１１２と、登録部１１３と、入力受付部１１５と、種別判別部１１６と、領域分割部１１７と、を備え、取得した文書データを解析した後、当該文書データの管理を行う。また、当該文書データの検索を可能とする。

【００２９】

文書データとは、任意のアプリケーションで作成された、文字コード等を含む電子文書と、ページ毎に画像として表されている文書画像データとを含むものとする。文書画像データとしては、例えば紙文書をスキャナで読み込んだデータ等とする。本実施の形態にかかる文書データは、特に、画像データやオブジェクトを含む文書データとし、換言すれば文書作成用アプリケーション（例えば、文書アプリケーション１０７）、並びに画像作成アプリケーション及び表作成アプリケーションの少なくとも一方という、複数のアプリケーションで作成されたデータを意味する。

【００３０】

データ格納部１０２は、取得した文書データや、当該文書データから生成されたページ画像データを格納する。当該文書データに関する詳細な情報は、後述する文書メタデータベース１０１で管理されている。

【００３１】

文書メタデータベース１０１は、文書管理テーブルと、ページ管理テーブルと、領域管理テーブルと、テキスト管理テーブルとを有している。

【００３２】

図２は、文書管理テーブルのテーブル構造を示した図である。本図に示すように、文書管理テーブルは、文書ＩＤと、タイトルと、作成更新日と、ページ数と、ファイルフォーマットと、ファイルパスと、ファイル名とを対応付けて保持する。

【００３３】

文書ＩＤは、文書データ毎に付与されたユニークなＩＤであり、これにより文書データを特定できる。タイトルは文書データのタイトルである。作成更新日は、文書データの作成日又は最終更新日を保持する。ページ数は文書データのページ数を保持している。ファイルフォーマットは、文書データ毎のフォーマットを保持している。これにより、管理している文書データが、アプリケーションで作成された電子文書、Ｗｅｂページ及び文書画像データ等のうちいずれかのフォーマットであるか特定することができる。

【００３４】

ファイルパスは、文書データが格納された場所を示している。そして、ファイル名は、

文書データのファイル名を示している。

【００３５】

図３は、ページ管理テーブルのテーブル構造を示した図である。本図に示すように、ページ管理テーブルは、ページＩＤと、文書ＩＤと、ページ番号と、ページ画像データパスとを対応付けて保持している。次に各フィールドについて説明する。

【００３６】

ページＩＤは、文書データを構成するページ毎に付与されたユニークなＩＤであり、このＩＤにより、文書データのページを一意に特定できる。文書ＩＤは、当該ページを含んでいる文書データを特定するＩＤとする。ページ番号は、当該ページを含んでいる文書データ中における、当該ページのページ番号とする。ページ画像データパスは、当該ページを表したページ画像データが格納されている場所を保持する。

【００３７】

図４は、領域管理テーブルのテーブル構造を示した図である。本図に示すように、領域管理テーブルは、領域ＩＤと、文書ＩＤと、ページＩＤと、領域座標と、種別と、タイトルと、テキストと、を対応付けて保持している。次に各フィールドについて説明する。

【００３８】

領域管理テーブルの領域ＩＤは、文書データに含まれるページを分割した領域毎に付与されたユニークなＩＤであり、このＩＤにより、文書データの各ページに含まれている領域を特定できる。文書ＩＤとページＩＤは、当該領域を含んでいる文書データ及びページを特定するＩＤとする。領域座標は、当該領域を特定する座標を保持し、本実施の形態では左上の頂点座標と右下の頂点座標を保持することで当該領域を特定する。

【００３９】

領域管理テーブルの種別は、当該領域のデータの種別を特定する情報を保持する。データの種別としては、例えばテキスト、画像（画像作成アプリケーションで作成された画像データ）、図（組織図、フローチャート、ガントチャート等の図作成用アプリケーションで作成されたオブジェクト等も含む）、写真、表（例えば表作成アプリケーションで作成されたオブジェクト等も含む）、グラフ（円グラフ、棒グラフなど表計算アプリケーションで作成されたオブジェクト等も含む）等とする。これら種別を、文書データを構成する文書構成要素とする。

【００４０】

タイトルは、当該領域を示すタイトルを保持する。テキストは、当該領域に含まれていたテキスト情報を保持する。このテキスト情報は、当該領域の種別がテキストの場合に限るものではなく、画像等であっても文字認識処理を実行して抽出したテキスト情報を保持する。当該フィールドに対して文字列を検索キーとして検索を行うことで、当該文字列を含む領域を特定できる。

【００４１】

図５は、テキスト管理テーブルのテーブル構造を示した図である。本図に示すように、テキスト管理テーブルは、文書ＩＤと、ページＩＤと、領域ＩＤと、文字情報と、開始位置座標と、字間と、行間と、文字揃えと、文字組と、フォントサイズとを、を対応付けて保持している。文書ＩＤ、ページＩＤ及び領域ＩＤにより、どの文書のどのページのどの領域のテキストを管理しているか特定できる。文字情報は、当該領域から抽出された一つ又は複数の文字コードを格納している。開始位置座標は、抽出元の領域内において、抽出された文字コードの位置座標（例えば左上の頂点座標）を表している。フォントサイズは抽出された文字コードのフォントサイズを示している。

【００４２】

上述したデータ格納部１０２及び文書メタデータベース１０１は、ＨＤＤ（Ｈａｒｄ Ｄｉｓｋ Ｄｒｉｖｅ）、光ディスク、メモリカード、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）などの一般的に利用されているあらゆる記憶装置により構成することができる。

【００４３】

文書取得部１０３は、文書データを取得する。図６は、本実施の形態で処理対象となる文書データを示す図である。図６に示すように、文書データは、表６０１が画像データとして含むとともに、縦組みの文章６０３も画像として含むものとする。これら表６０１、文章６０３は、文書処理装置１００においてそれぞれ別領域として扱われる。つまり、文書データに含まれる画像データやオブジェクトはこれらを作成したアプリケーションに応じて分割される。

【００４４】

また、テキストの領域についても、領域６０２など破線で分けられた領域毎に別領域として扱われる。取得した文書データは、フォーマット形式に応じて適切なアプリケーション、又は当該アプリケーションのＡＰＩを呼び出して、読み込みが行われるものとする。以下にアプリケーションの例について説明する。

【００４５】

文書アプリケーション１０７は、第１のフォーマット形式の文書データを読み込み可能なアプリケーションとする。当該文書アプリケーション１０７では、第１のフォーマット形式の文書データの参照、編集が可能とする。

【００４６】

データ参照アプリケーション１０８は、第２のフォーマット形式の文書データを読み込み可能なアプリケーションとする。データ参照アプリケーション１０８では、第２のフォーマット形式の文書データを参照のみ可能とする。

【００４７】

なお、文書アプリケーション１０７では、第２のフォーマット形式の文書データを認識できず、データ参照アプリケーション１０８では、第１のフォーマット形式の文書データを認識できないものとする。

【００４８】

プリンタドライバ１１１は、画像生成部１１４を備え、文書データの出力命令を受け付ける。本実施の形態においては、出力命令を受け付けた場合、画像生成部１１４を呼び出して、ページ画像データの生成を行う。

【００４９】

画像生成部１１４は、プリンタドライバ１１１が文書データの出力命令を受け付けた場合に、当該文書データの全表示領域を表すページ画像データを生成する。本実施の形態にかかる画像生成部１１４は、文書データのページ毎にページ画像データを生成する。

【００５０】

領域分割部１１７は、文書データの各ページを、当該ページに含まれるオブジェクト及び画像データ別（つまりデータを作成したアプリケーション別）、並びに文書に含まれる段落若しくはコラム別に、領域として分割する。ページに含まれる領域を分割する手法については、周知の手法を問わず、どのような手法を用いても良い。

【００５１】

種別判別部１１６は、文書データを構成する各領域に対して、当該領域を表す種別を判別する。本実施の形態にかかる種別判別部１１６は、種別判別の対象となる領域から特徴量を抽出し、抽出された特徴量に基づいてパターン認識処理を行うことで、領域の種別を判断する。その際に用いるパターン認識手法としては、どのような手法を用いても良いが、例えばニューラルネットやサポートベクターマシン手法を用いてもよい。これらニューラルネットやサポートベクターマシン手法を用いることで、学習用のデータセットを作成し学習させることで、より精度の高い領域の識別の判断を行うことができる。

【００５２】

判断部１０９は、文書データのページ毎に画像データのみ含むページであるか否か判断する。また、判断部１０９は、当該ページが画像データ以外の文書構成要素を含むページと判断した場合、当該ページを構成する領域毎に、判別された種別が文字コードを抽出可能な種別であるか否か判断する。文字コードが抽出可能な種別としては、テキスト要素の他に、例えば表計算アプリケーションで作成されたオブジェクトを表す文書構成要素など

が考えられる。

【００５３】

テキスト情報抽出部１１０は、種別がテキスト要素等の文字コードが抽出可能と判断された領域から、テキスト情報、すなわち文字コード群を抽出する。なお、本実施の形態は、テキスト情報抽出部１１０がテキスト情報を抽出する領域として、テキスト要素以外で文字コードが抽出可能な種別としては、文字コードが抽出可能なデータ、例えば任意のアプリケーションで作成されたオブジェクトデータ等が考えられる。

【００５４】

文字認識テキスト情報抽出部１１２は、画像生成部１１４により生成されたページ画像データの範囲のうち、判断部１０９により種別がテキスト要素等ではないと判断された領域に対応する範囲に対して文字認識処理を実行して得られた文字コードを、当該領域のテキスト情報として抽出する。この文字認識処理は、ＯＣＲ（Optical Character Recognition）に用いられている文字認識処理とする。

【００５５】

このように、文字認識テキスト情報抽出部１１２は、上述した文字認識処理を行う前に、予め文書データの領域と、生成されたページ画像データの範囲とを対応付けておく。これにより、文字コードを抽出できない領域から、テキスト情報を抽出することが可能となる。なお、範囲と領域とを対応付ける手法については、周知の手法を問わず、どのような手法を用いても良い。

【００５６】

つまり、本実施の形態にかかる文書処理装置１００では、テキスト情報抽出部１１０でテキスト情報を抽出すると共に、当該テキスト情報抽出部１１０で抽出できない領域については、当該領域に対応するページ画像データの表示領域に対して文字認識処理を実行して、テキスト情報を抽出する。これにより、文書データに含まれるテキスト情報は全て抽出できる。ところで、文字認識処理によるテキスト情報の抽出は、文字認識に誤りが生じる可能性がある。そこで、本実施の形態では、文字コードを抽出可能な領域については、文字認識処理を実行せずに、テキスト情報を抽出することとした。これにより、文書データから生成された画像データの全領域に文字認識処理を実行してテキスト情報を抽出した場合より、テキスト情報の抽出精度を向上させることができる。

【００５７】

登録部１１３は、文書メタデータベース１０１に対して、文書データ、当該文書データのページ及び領域に関する情報の登録、並びに文書データ及び画像データをデータ格納部１０２に格納する処理を行う。なお、詳細な登録手順については後述する。

【００５８】

また、上述した種別判別部１１６、判断部１０９、文字認識テキスト情報抽出部１１２、テキスト情報抽出部１１０、登録部１１３は、文書アプリケーション１０７やデータ参照アプリケーション１０８に組み込まれるプラグイン形式のプログラムでも良いし、別アプリケーションとして作成しても良い。

【００５９】

入力受付部１１５は、図示しない入力デバイスから情報の入力を受け付ける。例えば、入力受付部１１５は、文書データの検索要求や、その際の文書データの検索条件等の入力を受け付ける。

【００６０】

検索部１０６は、入力された検索条件に従って、文書管理テーブル、ページ管理テーブル、領域管理テーブル及びテキスト管理テーブルに対して検索を行う。また、検索部１０６は、領域管理テーブル及びテキスト管理テーブルに対して検索を行うことで、テキスト情報抽出部１１０により抽出されたテキスト情報、及び文字認識テキスト情報抽出部１１２により抽出されたテキスト情報のいずれか一つ以上に検索を行うことができる。

【００６１】

表示処理部１０５は、図示しない表示装置に対して、任意の情報の表示処理を行う。例

えば、例えば、表示処理部１０５は、検索部１０６の検索結果として、検索条件に一致するテキスト情報を含むページ画像データ等の一覧の表示処理を行う。なお、本実施の形態とは異なるが、表示処理部１０５が、検索条件に一致するテキスト情報を含む文書データを表示しても良い。この後、入力受付部１１５が、表示されたページ画像データの一覧から、利用したい文書データのページ画像データの選択を受け付ける。

【００６２】

出力部１０４は、利用者により選択されたページ画像データを含む文書データを、当該文書データの読み込み可能なアプリケーション（例えば文書アプリケーション１０７）に出力する。例えば、入力受付部１１５が、表示処理部１０５により表示処理されたページの画像データに対する選択を受け付けた場合に、出力部１０４が、選択を受け付けた画像データをページとして含む文書データの出力を行う等が考えられる。これにより、利用者は選択した文書を利用することができる。

【００６３】

次に、以上のように構成された本実施の形態にかかる文書処理装置１００における文書を取得してから文書メタデータベース１０１に登録するまでの処理について説明する。図７は、本実施の形態にかかる文書処理装置１００における上述した処理の手順を示すフローチャートである。なお、当該フローチャートでは、一般的となっている文書画像認識技術を最大限利用しているものとする。

【００６４】

まず、文書取得部１０３が、スキャナ１５１や外部記憶装置１５０等の外部環境から、文書データを取得する（ステップＳ７０１）。

【００６５】

そして、文書取得部１０３が、取得した文書データを認識可能なアプリケーション（例えば、文書アプリケーション１０７など）のＡＰＩを呼び出して、当該文書データの読み込み処理を行う（ステップＳ７０２）。これにより、文書データの内容を把握することができる。

【００６６】

次に、登録部１１３が、取得した文書データを、データ格納部１０２に格納すると共に、文書管理テーブルに文書データに関する情報の登録を行う（ステップＳ７０３）。文書データに関する情報としては、文書管理テーブルの各フィールドに示されたタイトル、作成更新日、ページ数、ファイルフォーマット、当該文書データの格納先となるファイルパス、当該文書データのファイル名とする。

【００６７】

そして、判断部１０９が、当該文書データに含まれるページにおいて、全領域が画像データであるか否か判断する（ステップＳ７０４）。なお、ステップＳ７０３は、最初は１ページ目について判断を行い、当該ステップが繰り返される度に次ページを判断するように処理を行う。

【００６８】

次に、当該ページの全領域が画像データと判断された場合（ステップＳ７０４：Ｙｅｓ）、登録部１１３が、当該ページを表す画像データと共に、当該ページの関する情報を対応付けて、ページ管理テーブルに登録する（ステップＳ７０５）。ページに関する情報としては、当該ページを含む文書の文書ＩＤと、当該ページのページ番号、当該ページを著す画像の格納先の画像データパスとする。なお、当該ページを表す画像データは、登録部１１３により、データ格納部に格納されるものとする。

【００６９】

そして、領域分割部１１７が、当該ページ画像を領域毎に分割する（ステップＳ７０６）。なお、ページに含まれる領域の分割は、空白領域の幅に基づいて分割するなどの周知の手法を問わず、あらゆる手法で分割することができる。当該分割により、領域毎の領域座標を特定できる。

【００７０】

次に、種別判別部１１６が、当該ページの分割された領域毎に、当該領域から抽出された特徴量に基づいて、当該領域の種別を判別する（ステップＳ７０７）。この領域からの特徴量の抽出は、種別判別部１１６が行うものとする。

【００７１】

そして、種別の判別が行われた後、文字認識テキスト情報抽出部１１２は、画像データの当該領域に対して文字認識処理を実行して、当該領域に含まれているテキスト情報、すなわち文字コード群を抽出する（ステップＳ７０８）。なお、文字認識テキスト情報抽出部１１２は、文字認識処理により文字コードを抽出する際、各文字コードの位置座標も取得する。なお、位置座標のみならず、適切なフォントサイズも取得する。なお、複数文字を一レコードに登録する場合、字間、行間、文字揃え等も取得する。

【００７２】

次に、登録部１１３が、当該領域に関する情報を領域管理テーブルに登録すると共に、当該領域から抽出されたテキストに関する情報を、テキスト管理テーブルに登録する（ステップＳ７０９）。当該領域に関する情報としては、当該領域を含むページのページＩＤ、当該ページを含む文書ＩＤ、当該領域の領域座標、当該領域の種別、テキストなどが考えられる。

【００７３】

なお、テキスト管理テーブルに文字コードを登録する場合、テキストに関する情報として、抽出先の文書ＩＤ、ページＩＤ及び領域ＩＤと共に、当該領域ＩＤから抽出された文字コード（文字情報）と、当該文字コードの開始位置座標と、フォントサイズとを対応付けて登録する。開始位置座標とは、当該文字コードを含む矩形の左上座標とする。また、テキスト管理テーブルには、一文字単位でレコードとして登録しても良いし、複数の文字列をレコードに登録しても良い。複数の文字列をレコードに格納する場合、それぞれの文字列の位置座標を特定するための情報、例えば字間、行間、文字揃え、文字組などを対応付けて登録する。これにより、領域に対して文字の配置を再現することが可能となる。

【００７４】

次に、判断部１０９が、当該ページに含まれる全ての領域についてステップＳ７０６～Ｓ７０８の処理を行ったか否か判断する（ステップＳ７１０）。処理を行っていない領域があると判断した場合（ステップＳ７１０：Ｎｏ）、処理を行っていない領域に対する種別の判別から開始する（ステップＳ７０７）。

【００７５】

そして、判断部１０９が、処理を行ったと判断した場合（ステップＳ７１０：Ｙｅｓ）、ステップＳ７２０へと処理を進ませる。

【００７６】

一方、判断部１０９が、当該文書データに含まれるページにおいて、全領域が画像データではないと判断した場合（ステップＳ７０４：Ｎｏ）、文書データを読み込んだアプリケーションのＡＰＩが、プリンタドライバ１１１に対して、当該ページの出力命令を行う。

【００７７】

そして、プリンタドライバ１１１は、当該出力命令に従って、画像生成部１１４が、当該ページの画像データを生成する（ステップＳ７１１）。

【００７８】

次に、登録部１１３が、生成された当該ページを表す画像データと共に、当該ページの関する情報を対応付けて、ページ管理テーブルに登録する（ステップＳ７１２）。

【００７９】

領域分割部１１７が、当該ページを領域毎に分割する（ステップＳ７１３）。なお、ページに含まれる領域の分割は、空白領域の幅に基づいて分割するなどの周知の手法を問わず、あらゆる手法で分割することができる。当該分割により、領域毎の領域座標を特定できる。

【００８０】

その後、種別判別部１１６が、分割された領域毎に、当該領域から抽出された特徴量に基づいて、当該領域の種別を判別する（ステップＳ７１４）。

【００８１】

次に、判断部１０９が、判別された種別に基づいて、当該領域がテキスト要素又は文字コードが抽出可能なオブジェクトであるか否か判断する（ステップＳ７１５）。

【００８２】

そして、判断部１０９が、当該領域がテキスト要素又は文字コードが抽出可能なオブジェクトであると判断した場合（ステップＳ７１４：Ｙｅｓ）、テキスト情報抽出部１１０が、当該領域に含まれているテキスト情報、すなわち文字コードと文字コードの位置座標を抽出する（ステップＳ７１６）。なお、文字コードの位置座標とは、当該領域を含むページが印刷された場合に位置する当該文字コードの位置座標とする。この位置座標は、例えば、生成されたページ画像データから特定することができる。

【００８３】

一方、判断部１０９が、当該領域がテキスト要素又は文字コードが抽出可能なオブジェクトではないと判断した場合（ステップＳ７１４：Ｎｏ）、文字認識テキスト情報抽出部１１２が、画像データの当該領域に対して文字認識処理を実行して、当該領域に含まれているテキスト情報、すなわち文字コード群を抽出する（ステップＳ７１７）。なお、文字認識テキスト情報抽出部１１２は、文字認識処理を実行して文字コードを抽出する際、各文字コードの位置座標も取得する。

【００８４】

その後、登録部１１３が、当該領域に関する情報を領域管理テーブルに登録すると共に、当該領域から抽出されたテキスト情報を、テキスト管理テーブルに登録する（ステップＳ７１８）。

【００８５】

そして、判断部１０９が、当該ページに含まれる全領域について、ステップＳ７１５～Ｓ７１８までの処理を行ったか否か判断する（ステップＳ７１９）。処理を行っていない領域があると判断した場合（ステップＳ７１９：Ｎｏ）、処理を行っていない領域に対する種別の判別から開始する（ステップＳ７１４）。

【００８６】

一方、判断部１０９が、ページの全領域に対して処理を行ったと判断した場合（ステップＳ７１９：Ｙｅｓ）、ステップＳ７２０へと処理を進ませる。

【００８７】

次に、判断部１０９は、文書データに含まれる全ページに対して、ステップＳ７０４～Ｓ７１９までの処理を行ったか否か判断する（ステップＳ７２０）。行っていないと判断した場合（ステップＳ７２０：Ｎｏ）、再びステップＳ７０４から処理を開始する（ステップＳ７０４）。

【００８８】

一方、判断部１０９は、文書データに含まれる全ページに対して、ステップＳ７０４～Ｓ７１９までの処理を行ったと判断した場合（ステップＳ７２０：Ｙｅｓ）、処理を終了する。

【００８９】

上述した処理手順では、電子文書を処理する場合について具体的に説明したが、文書画像データの場合も同様に処理することができる。この場合、ステップＳ７０４で常にＹｅｓと判断される。

【００９０】

上述した処理手順により、文書データに関する情報が文書管理テーブルに登録されると共に、当該文書データに含まれる各ページ、各領域に関する情報が、ページ管理テーブル及び領域管理テーブルに登録される。さらには、各領域に含まれているテキスト情報が、テキスト管理テーブルに、位置座標と対応付けて登録されることになる。このように、当該領域がテキスト要素であるか否かにかかわらず登録されることになる。

【００９１】
上述した処理手順では、画像データの解析手法と、文書データの解析手法を組み合わせることで容易に精度よく、当該文書データに含まれているテキスト情報を抽出することができる。
【００９２】
また、現在、文書データのフォーマット形式は、アプリケーション毎に様々な形式のフォーマットがある。そして、これら様々な形式のフォーマットに対応するために、本実施の形態では、上述した文書メタデータベース１０１に、フォーマット形式によらず全ての文書データの画像データ及びテキスト情報を登録することとした。これにより、文書フォーマットに関係なく、文書データの検索を行うことができる。
【００９３】
また、文書データの画像データを生成する際に、プリンタドライバ内の画像生成部１１４を用いることとした。これにより、文書データのフォーマット形式にかかわらず、文書データの画像データを生成できる。
【００９４】
次に、以上のように構成された本実施の形態にかかる文書処理装置１００における文書データの検索処理について説明する。図８は、本実施の形態にかかる文書処理装置１００における上述した処理の手順を示すフローチャートである。
【００９５】
まず、入力受付部１１５が、利用者が操作する入力デバイスによる、検索条件の入力を受け付ける（ステップＳ８０１）。
【００９６】
次に、検索部１０６が、入力された検索条件に従って、文書メタデータベース１０１に対して検索を行う（ステップＳ８０２）。この検索としては、例えば、利用者により入力された文字列を検索キーとして、領域管理テーブルのテキストフィールドや、テキスト管理テーブルの文字情報フィールドに対して検索を行うことが考えられる。または、ページ管理テーブルや、文書管理テーブルに対して検索条件を設定して検索を行ってもよい。そして、検出された領域、テキスト情報等からページＩＤ、文書ＩＤ等を特定できる。
【００９７】
そこで、表示処理部１０５が、検索条件を満足する領域やテキスト情報を含むページ画像データの一覧を、表示装置に表示する（ステップＳ８０３）。なお、当該一覧に表示されたページ画像データ毎に、ページＩＤが対応付けられているものとする。これにより、利用者が、ページ画像データを選択した場合に、ページＩＤが入力されることになる。
【００９８】
次に、入力受付部１１５が、表示されたページ画像データの一覧から、利用者が選択したページ画像データを示すページＩＤの入力を受け付ける（ステップＳ８０４）。
【００９９】
そして、出力部１０４が、利用者が選択したページ画像データを示すページＩＤから、当該ページを含む文書データを特定し、当該文書データを読み込み可能なアプリケーションに出力する（ステップＳ８０５）。
【０１００】
上述した処理手順により、利用者は検索条件に該当する文書データを利用することが可能となる。また、文書メタデータベース１０１のテーブル群に対して検索を行うので、文書データのフォーマット形式を意識することなく、検索を行うことを可能としている。
【０１０１】
また、本実施の形態においては、検索終了後に検索条件に一致するページ画像データを表示したが、ページ画像データに制限するものではなく、文書データを表示しても良い。
【０１０２】
ところで、従来、任意のアプリケーションで作成された文書データに画像データが埋め込まれている場合、当該画像データに含まれている文字列を検索できないという問題が生

じていた。そこで本実施の形態にかかる文書処理装置１００では、文書データに画像データが含まれている場合であっても適切に抽出して、文書メタデータベース１０１に登録するので、検索可能となる。

【０１０３】

上述した実施の形態にかかる文書処理装置１００によれば、文書データの各ページの領域に応じて、テキストの抽出処理を変更しているので、解析精度を向上させることができる。また、一般的な文字認識処理（ＯＣＲ）と、文書データのテキスト要素又はオブジェクトからの文字コードの抽出を組み合わせることで、簡易に精度よい解析が可能となる。

【０１０４】

また、本実施の形態にかかる文書処理装置１００によれば、文書データの画像データで表現された図や表に対して、文字認識処理を実行することで、テキスト情報を抽出している。これにより、文書データに図や表が画像データとして貼り付けられている場合でも、解析が可能となる。さらに図や表に含まれているテキスト情報で検索することが可能となる。

【０１０５】

さらに、文書処理装置１００では、文書画像データであっても、テキスト情報の抽出が可能である。このように文書画像データ、文書データを問わず検索対象とすることが可能となる。

【０１０６】

上述した文書処理装置１００では、画像データを生成し、当該画像データからテキスト情報を抽出することで、当該領域に含まれているオブジェクトのフォーマットと依存せず解析が可能である。

【０１０７】

上述した文書処理装置１００では、文書管理テーブル、ページ管理テーブル、領域管理テーブル及びテキスト管理テーブルに対して検索を行うことができるので、文書、ページ領域、テキスト情報単位での検索が可能となる。

【０１０８】

（変形例）

また、上述した各実施の形態に限定されるものではなく、以下に例示するような種々の変形が可能である。

【０１０９】

上述した第１の実施の形態にかかる文書処理装置１００では、検索結果としてページ画像データの一覧を表示していた。しかしながら、ページ画像データの一覧の表示に制限するものではない。

【０１１０】

例えば、表示処理部１０５が、ページ画像データの代わりに、領域管理テーブル及びテキスト管理テーブルが保持する情報を利用して各ページを再現して表示しても良い。具体的には、領域管理テーブル及びテキスト管理テーブルが保持する情報を取得して、検索された領域を含むページについて、当該ページを構成する文字コードを、当該文字コードと対応付けられた位置座標に配置することで、当該ページを再現できる。

【０１１１】

上述した処理手順により、文書データのフォーマット形式にかかわらず、当該文書データのページを再現して表示することができる。さらに、ページ画像データではなく文字コードを用いてページを再現するので、文書処理装置１００の処理負担を軽減させることができる。

【０１１２】

図９は、文書処理装置１００の機能を実現するためのプログラムを実行したＰＣのハードウェア構成を示した図である。本実施の形態の文書処理装置１００は、ＣＰＵ（Central Processing Unit）９０１と、ＲＯＭ（Read Only Memory）９０２やＲＡＭ（Random Access Memory）９０３と、ＨＤＤ（Hard Disk Drive）、ＣＤ（Compact Disk）

ドライブ装置等の外部記憶装置９０５と、ディスプレイ装置等の表示装置９０６と、キーボードやマウス等の入力デバイス９０７と、通信Ｉ／Ｆ９０４と、これらを接続するバス９０８を備えており、通常のコンピュータを利用したハードウェア構成となっている。

【０１１３】

本実施形態の文書処理装置１００で実行される文書処理プログラムは、インストール可能な形式又は実行可能な形式のファイルでＣＤ－ＲＯＭ、フレキシブルディスク（ＦＤ）、ＣＤ－Ｒ、ＤＶＤ（Digital Versatile Disk）等のコンピュータで読み取り可能な記録媒体に記録されて提供される。

【０１１４】

また、本実施形態の文書処理装置１００で実行される文書処理プログラムを、インターネット等のネットワークに接続されたコンピュータ上に格納し、ネットワーク経由でダウンロードさせることにより提供するように構成しても良い。また、本実施形態の文書処理装置１００で実行される文書処理プログラムをインターネット等のネットワーク経由で提供または配布するように構成しても良い。

【０１１５】

また、本実施形態の文書処理装置１００で実行される文書処理プログラムを、ＲＯＭ等に予め組み込んで提供するように構成してもよい。

【０１１６】

上述した実施の形態の文書処理装置１００で実行される文書処理プログラムは、文書処理装置１００において上記記録媒体から読み出して実行することによりＲＡＭ９０２上にロードされ、上記ソフトウェア構成で説明した各部がＲＡＭ９０３上に生成されるようになっている。

【産業上の利用可能性】

【０１１７】

以上のように、本発明にかかる文書処理装置、文書処理方法及び文書処理プログラムは、文書データの処理に有用であり、特に、画像を含む文書データを検索可能に格納する技術に適している。

【図面の簡単な説明】

【０１１８】

【図１】実施の形態にかかる文書処理装置の構成を示すブロック図である。

【図２】文書管理テーブルのテーブル構造を示した図である。

【図３】ページ管理テーブルのテーブル構造を示した図である。

【図４】領域管理テーブルのテーブル構造を示した図である。

【図５】テキスト管理テーブルのテーブル構造を示した図である。

【図６】実施の形態で処理対象となる文書データを示す図である。

【図７】実施の形態にかかる文書処理装置における、文書を取得してから文書メタデータベースに登録するまでの処理の手順を示すフローチャートである。

【図８】本実施の形態にかかる文書処理装置における、文書データの検索処理の手順を示すフローチャートである。

【図９】本実施の形態にかかる文書処理装置の機能を実現するためのプログラムを実行したＰＣのハードウェア構成を示した図である。

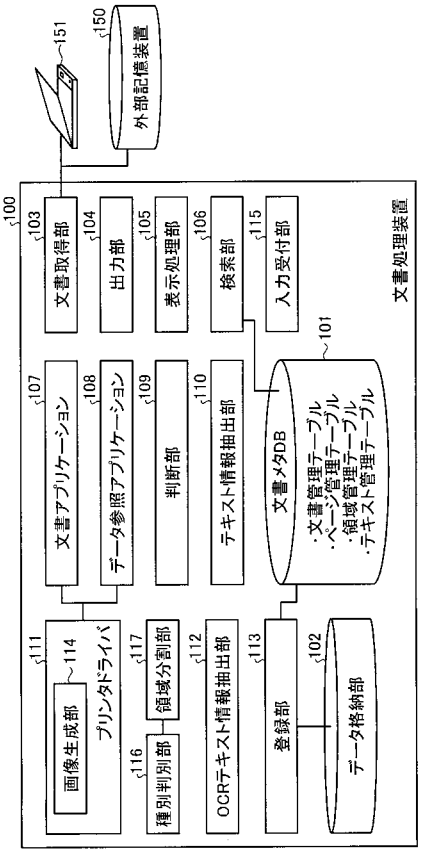
【符号の説明】

【０１１９】

１００　文書処理装置
１０１　文書メタデータベース
１０２　データ格納部
１０３　文書取得部
１０４　出力部
１０５　表示処理部
１０６　検索部

１０７　文書アプリケーション
１０８　データ参照アプリケーション
１０９　判断部
１１０　テキスト情報抽出部
１１１　プリンタドライバ
１１２　文字認識テキスト情報抽出部
１１３　登録部
１１４　画像生成部
１１５　入力受付部
１１６　種別判別部
１５０　外部記憶装置
１５１　スキャナ
９０１　ＣＰＵ
９０２　ＲＯＭ
９０３　ＲＡＭ
９０４　通信Ｉ／Ｆ
９０５　外部記憶装置
９０６　表示装置
９０７　入力デバイス
９０８　バス

【図１】



【図２】

文書ID	タイトル	作成更新日	ページ数	ファイルフォーマット	ファイルパス	ファイル名
DOC0001	イメージについて	2005/11/19	22	doc	/doc/image.doc	image.doc
DOC0002	文書の書き方	2007/1/14	4	tiff	/doc/doc.tiff	doc.tiff
：	：	：	：	：	：	：

【図３】

ページID	文書ID	ページ番号	画像データパス
P000001	DOC0001	1	……
：	：	：	：

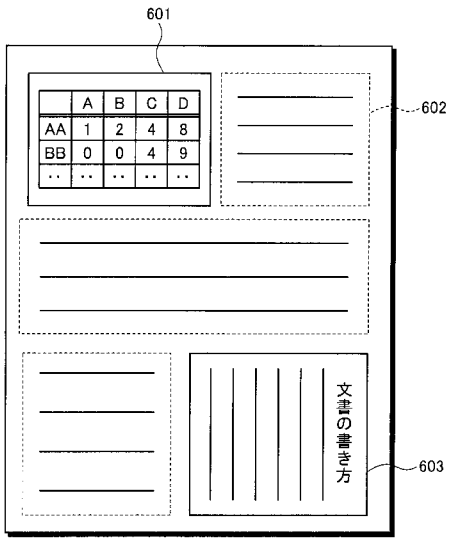
【図４】

領域ID	文書ID	ページID	領域座標	種別	タイトル	テキスト
R000001	DOC0001	P000001	(0,0)-(500,200)	テキスト	イメージ	今…
：	：	：	：	：	：	：

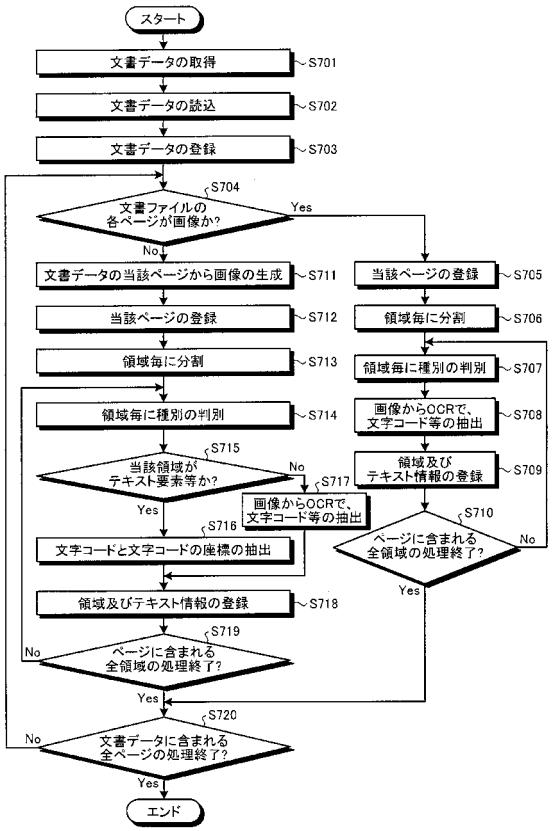
【図５】

文書ID	ページID	領域ID	文字情報	開始位置座標	字間	行間	文字揃え	文字組	フォントサイズ
DOC0001	P000001	R000001	今……	10,15	0	4	左寄せ	縦組み	21p
DOC0002	P002001	R002001	秋	30,20	－	－	－	－	21p
：	：	：	：	：	：	：	：	：	：

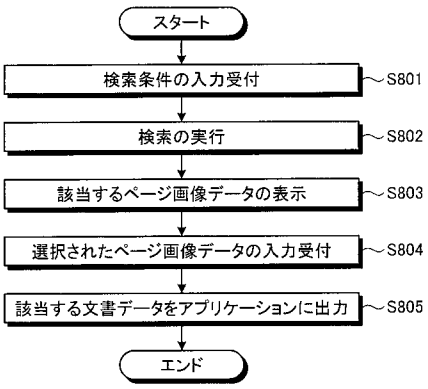
【図６】



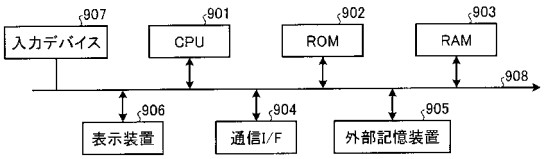
【図７】



【図８】


スタート
検索条件の入力受付
S801
検索の実行
S802
該当するページ画像データの表示
S803
選択されたページ画像データの入力受付
S804
該当する文書データをアプリケーションに出力
S805
エンド

【図９】


907
入力デバイス
901
CPU
902
ROM
903
RAM
908
906
表示装置
904
通信I/F
905
外部記憶装置