(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2017-146745
(P2017-146745A)
(43)公開日　平成29年8月24日(2017.8.24)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	210D
G06N 99/00	(2010.01)	G06F	17/30	170B
		G06F	17/30	220C
		G06N	99/00	153

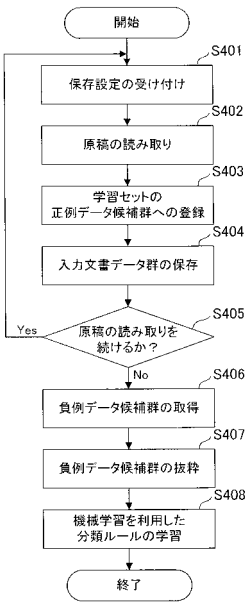
審査請求 未請求　請求項の数 10　ＯＬ　（全 18 頁）

(21)出願番号　特願2016-27352(P2016-27352)
(22)出願日　平成28年2月16日(2016.2.16)

(71)出願人　000001007
キヤノン株式会社
東京都大田区下丸子３丁目３０番２号
(74)代理人　100114775
弁理士　高岡　亮一
(74)代理人　100121511
弁理士　小田　直
(72)発明者　宮内　崇
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

(54)【発明の名称】情報処理装置、制御方法、情報処理システム、およびプログラム

(57)【要約】
【課題】高精度な分類ルールの構築を可能とする高品質な教師データを効率的に生成する情報処理装置を提供する。
【解決手段】ＭＦＰ１０１は、分類するクラスごとの画像データを正例データとして受け付け、受け付けた前記画像データに付与された文書ファイル情報に含まれる情報のうち少なくとも１つが一致する文書ファイル情報が付与された画像データを負例データとして取得し、前記正例データ及び負例データを用いて、画像データを種別ごとに分類するために用いる分類ルールを生成する。
【選択図】図４



【特許請求の範囲】
【請求項１】
分類する種別ごとの画像データを正例データとして受け付ける受付手段と、
受け付けた前記画像データに付与されたファイル情報に含まれる情報のうち少なくとも１つが一致するファイル情報が付与された画像データを負例データとして取得する取得手段と、
前記正例データ及び負例データを用いて、画像データを種別ごとに分類するために用いる分類ルールを生成する生成手段と、を備える
ことを特徴とする情報処理装置。
【請求項２】
前記ファイル情報は、少なくとも画像データのメタ情報を含む
ことを特徴とする請求項１に記載の情報処理装置。
【請求項３】
前記メタ情報は、画像データのタイトル、作成者名、ファイル形式、作成デバイス、生成日時、または当該画像データが含むキーワードのうち少なくとも１つを含む
ことを特徴とする請求項２に記載の情報処理装置。
【請求項４】
前記取得手段は、前記正例データに付与されたファイル情報に含まれる情報のうち、当該正例データにおいて共起性が高い情報を含むファイル情報が付与された画像データを負例データとして取得する
ことを特徴とする請求項１乃至３のいずれか１項に記載の情報処理装置。
【請求項５】
前記取得手段は、前記負例データとして取得した画像データのうち、前記正例データに付与されたファイル情報に含まれる情報との一致率が高いファイル情報が付与された画像データを前記負例データとして使用しない
ことを特徴とする請求項１乃至４のいずれか１項に記載の情報処理装置。
【請求項６】
前記取得手段は、前記負例データとして取得した画像データを、当該画像データのキーワードに基づき種別ごとに分類し、それぞれの種別において分類された画像データの個数が上限の値よりも多い場合は、当該分類された画像データの個数が当該上限の値以下となるように当該画像データを削除する
ことを特徴とする請求項１乃至５のいずれか１項に記載の情報処理装置。
【請求項７】
前記取得手段は、
受け付けた前記画像データに付与されたファイル情報に含まれる情報を項目ごとに表示する画面を有し、
前記画面において指定された項目ごとの値が当該画面において指定された条件を満たす画像データを前記負例データとして取得する
ことを特徴とする請求項１乃至６のいずれか１項に記載の情報処理装置。
【請求項８】
情報処理装置とサーバとを備えるシステムであって、
前記情報処理装置は、
分類する種別ごとの画像データを正例データとして受け付ける受付手段と、
受け付けた前記画像データに付与されたファイル情報に含まれる情報のうち少なくとも１つが一致するファイル情報が付与された画像データを、前記サーバから負例データとして取得する取得手段と、
前記正例データ及び負例データを用いて、画像データを種別ごとに分類するために用いる分類ルールを生成する生成手段と、を備え、
前記サーバは、
前記情報処理装置の要求に応じて、画像データを前記情報処理装置に送信する送信手

段を備える
ことを特徴とする情報処理システム。
【請求項９】
分類する種別ごとの画像データを正例データとして受け付ける受付工程と、
受け付けた前記画像データに付与されたファイル情報に含まれる情報のうち少なくとも１つが一致するファイル情報が付与された画像データを、負例データとして取得する取得工程と、
前記正例データ及び負例データを用いて、画像データを種別ごとに分類するために用いる分類ルールを生成する生成工程と、を備える
ことを特徴とする情報処理装置の制御方法。
【請求項１０】
請求項１乃至７のいずれか１項に記載の情報処理装置が備える各手段としてコンピュータを機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、情報処理装置、制御方法、情報処理システム、およびプログラムに関する。
【背景技術】
【０００２】
文書を扱うワークフローの効率化を実現する技術の１つとして、機械学習を利用した画像分類が提案されている。機械学習を利用した画像分類は、一般的に学習と分類（運用）の２つのプロセスを有し、画像データ群（教師データ、学習セット）を与えることで分類ルールを学習によって構築し、構築した分類ルールに基づいて入力画像を分類する。
【０００３】
データを複数の種別に分類するには、データが学習セットとして与えた種別のいずれであるかを分類できればよい場合が多いが、文書を扱う場合には、学習したいずれの種別でもない文書を「該当なし」と分類したいというニーズがある。例えば、ＭＦＰのスキャナによって大量の文書が読み込まれた際に、特定の種別の文書のみをあらかじめ指定されたフォルダに格納し、その他の種別の文書は「該当なし」に分類し、まとめて一か所のフォルダに格納するようなケースが考えられる。
【０００４】
機械学習では、学習セットとして与えられたデータに基づいて分類ルールを構築するため、学習セット内のデータは、運用時に入力されるデータと特徴量が近い方がよい。また、「該当なし」の分類を実現するには、本来分類したい種別のデータ（正例データ）に加えて、「その他」の種別であるデータ（負例データ）を用意した方がよく、負例データとしては、実際に分類時に入力される可能性の高い文書を用意することが望ましい。
【０００５】
しかし、ユーザが多くの種別の文書を扱っている場合に、本来分類したい種別のデータ（正例データ）以外の大量な文書データを負例データとして用意するのは、ユーザにとって大きな負担となってしまう。また、機械学習では正例データと負例データに同じ種別のデータが混在していると正しく分類ルールを構築することができない。そのため、初めて学習セットを用意する際だけでなく、正例データの種別を追加する度に、負例データの中に新しく追加した種別の正例データが混在していないかを確認する必要がある。
【０００６】
特許文献１は、正例の文書（正例データ）から特徴語を抽出し、ファイルサーバから取り出した負例候補文書から、当該正例の特徴語をなるべく含まず、かつ当該正例の特徴語以外の特徴語を多く含む文書を負例として選択する文書分類システムを開示している。
【先行技術文献】
【特許文献】
【０００７】

【特許文献１】特開２０１４－９６０８６号公報
【発明の概要】
【発明が解決しようとする課題】
【０００８】
しかしながら、特許文献１のように正例データと同じ種別である可能性の低い文書データを除くだけでは、効率よく高精度な分類器を構築することは困難である。一般に、学習セットのデータ量に応じて学習時間が増加する。このため、例えば、ユーザが用意したデータからその場で分類ルールを構築するシステムの場合には、学習セットを絞り込む必要がある。しかし、ファイルサーバからランダムに一定数のファイルを選ぶ等、学習に利用するデータを一律に削減してしまうと、実際に分類時に入力される可能性の高いデータも減り、分類精度が低下してしまう。
【０００９】
本発明は、高精度な分類ルールの構築を可能とする高品質な教師データを効率的に生成する情報処理装置の提供を目的とする。
【課題を解決するための手段】
【００１０】
本発明の一実施形態の情報処理装置は、分類する種別ごとの画像データを正例データとして受け付ける受付手段と、受け付けた前記画像データに付与されたファイル情報に含まれる情報のうち少なくとも１つが一致するファイル情報が付与された画像データを負例データとして取得する取得手段と、前記正例データ及び負例データを用いて、画像データを種別ごとに分類するために用いる分類ルールを生成する生成手段と、を備える。
【発明の効果】
【００１１】
本発明の情報処理装置によれば、高精度な分類ルールの構築を可能とする高品質な教師データを生成することができる。
【図面の簡単な説明】
【００１２】
【図１】第１実施形態における情報処理システム構成を示す図である。
【図２】ＭＦＰの構成例を示す図である。
【図３】サーバのハードウェア構成の一例を示す図である。
【図４】ＭＦＰが分類ルールを学習する処理を説明するためのフローチャートである。
【図５】負例データを構築する処理を示すフローチャートである。
【図６】文書ファイル情報の一例を示す図である。
【図７】負例データ候補群の絞り込み条件をユーザが編集する画面を示す図である。
【図８】文書ファイル情報による絞り込み結果の例を示す図である。
【図９】学習セットを用いた機械学習の一例を示す図である。
【図１０】特徴量の算出方法について説明する図である。
【図１１】画像データからパッチ画像を切り出す方法について説明する図である。
【図１２】学習セットを生成し、分類ルールを学習する処理を説明する図である。
【発明を実施するための形態】
【００１３】
以下、本発明を実施するための形態について図面などを参照して説明する。
（第１実施形態）
図１は、本実施形態における情報処理システム構成を示す図である。
第１実施形態における情報処理システムは、情報処理装置であるＭＦＰ１０１及びサーバ１０２を備える。
【００１４】
ＬＡＮ１０３には、ＭＦＰ１０１が接続されている。また、ＬＡＮ１０３は、インターネット１０４を経由してサービスを提供するサーバ１０２と接続されている。ＭＦＰ１０１及びサーバ１０２は、ＬＡＮ１０３を介して互いに接続されており、画像データや各種

情報の送受信を行う。なお、ＭＦＰ１０１とサーバ１０２とは、互いに接続され、画像データや各種情報を送受信できればよく、有線により直接接続されていてもよく、また、無線通信により接続されていてもよい。

【００１５】

サーバ１０２は、ＭＦＰ１０１から入力された画像データを格納し、ＭＦＰ１０１から指定された条件を満たす画像データをＭＦＰ１０１に送信するファイルサーバとして機能する。なお、本実施形態では、分類ルールを学習する際に使用する学習セットの生成や、当該学習セットを用いた分類ルールの構築はＭＦＰ１０１が実行するが、同様の処理をサーバ１０２が実行してもよい。

【００１６】

図２は、ＭＦＰ１０１の構成例を示す図である。

図２（Ａ）は、ＭＦＰ１０１のハードウェア構成の一例を示す図である。図２（Ａ）に示すように、ＭＦＰ１０１は、コントローラ２０、画像読取部２０１、画像出力部２０５、及び操作部２０７を備える。コントローラ２０は、装置制御部２００、画像処理部２０２、記憶部２０３、ＣＰＵ２０４、及びネットワークＩ／Ｆ部２０６を備える。

【００１７】

装置制御部２００は、ＭＦＰ１０１内およびネットワークＩ／Ｆ部２０６を経由した外部とのデータの受け渡しや、操作部２０７からの操作の受け付けを行う。画像読取部２０１は、原稿の画像を読み取り、画像データをコントローラ２０に出力する。画像処理部２０２は、画像読取部２０１や外部から入力される画像データを含む印刷情報を中間情報（以下「オブジェクト」と呼ぶ）に変換し、記憶部２０３のオブジェクトバッファに格納する。

【００１８】

オブジェクトは、テキスト、グラフィック、イメージの属性を持つ。さらに、オブジェクトバッファに格納したオブジェクトに基づきビットマップデータを生成し、記憶部２０３のバッファに格納する。その際、色変換処理、濃度調整処理、トナー総量制御処理、ビデオカウント処理、プリンタガンマ補正処理、ディザなどの疑似中間調処理を行う。記憶部２０３は、ＲＯＭ（Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）、ＨＤＤ（Ｈａｒｄ Ｄｉｓｋ Ｄｒｉｖｅ）などから構成される。

【００１９】

ＲＯＭは、ＣＰＵ２０４が実行する各種の制御プログラムや画像処理プログラムを格納する。ＲＡＭは、ＣＰＵ２０４がデータや各種情報を格納する参照領域や作業領域として用いられる。また、ＲＡＭおよびＨＤＤは、上述したオブジェクトバッファなどに用いられる。コントローラ２０は、ＲＡＭおよびＨＤＤ上で画像データを蓄積し、ページのソートや、ソートされた複数ページにわたる原稿を蓄積し、複数部プリント出力を行う。

【００２０】

なお、記憶部２０３を構成するＨＤＤは、ファイルサーバとして機能し、画像読取部２０１やネットワークＩ／Ｆ部２０６経由で入力された画像データが蓄積されているものとする。画像出力部２０５は、記録紙などの記録媒体にカラー画像を形成して出力する。ネットワークＩ／Ｆ部２０６は、ＭＦＰ１０１をＬＡＮ１０３に接続し、インターネット１０４や他の装置との間で各種情報を送受信する。操作部２０７は、タッチパネルや操作ボタンを備え、ユーザからの操作を受け付けて装置制御部２００へ該操作の情報を送信する。

【００２１】

図２（Ｂ）は、ＭＦＰ１０１の外観の一例を示す図である。画像読取部２０１は、複数の受光画素を有している。各受光画素の感度が夫々異なっていると、たとえ原稿上の各画素の濃度が同じであったとしても、各画素が夫々違う濃度であると認識されてしまう。そのため、画像読取部２０１では、最初に白板（一様に白い板）を露光走査し、露光走査して得られた反射光の量を電気信号に変換してコントローラに出力している。

【００２２】
なお、画像処理部２０２内には、各受光画素から得られた電気信号を元に、各受光画素の感度の違いを認識し、その違いを利用して、原稿上の画像をスキャンして得られた電気信号の値を補正する、公知のシェーディング補正処理部を有する。さらに、シェーディング補正部は、コントローラ内のＣＰＵ２０４からゲイン調整の情報を受取ると、当該情報に応じたゲイン調整を行う。
【００２３】
ゲイン調整は、原稿を露光走査して得られた電気信号の値を、どのように０～２５５の輝度信号値に割り付けるかを調整するために用いられる。このゲイン調整により、原稿を露光走査して得られた電気信号の値を高い輝度信号値に変換したり、低い輝度信号値に変換したりすることができるようになっている。すなわち、ゲイン調整により、読み取り信号のダイナミックレンジの調整が可能である。
【００２４】
続いて、この原稿上の画像をスキャンする構成について説明する。
画像読取部２０１は、原稿上の画像を露光走査して得られた反射光を受光画素に入力することで画像の情報を電気信号に変換する。さらに電気信号をレッドＲ，グリーンＧ，およびブルーＢの各色からなる輝度信号に変換し、当該輝度信号を画像としてコントローラ２０に対して出力する。
【００２５】
なお、原稿は原稿フィーダ２１１のトレイ２１２にセットされる。ユーザが操作部２０７から読み取り開始を指示すると、コントローラ２０から画像読取部２０１に原稿読み取り指示が与えられる。画像読取部２０１は、この指示を受けると原稿フィーダ２１１のトレイ２１２から原稿を１枚ずつフィードして、原稿の読み取り動作を行う。なお、原稿の読み取り方法は、原稿フィーダ２１１による自動送り方式に限られるものではなく、原稿を不図示のガラス面上に載置し露光部を移動させることで原稿の走査を行う方法であってもよい。
【００２６】
画像出力部２０５は、コントローラ２０から受取った画像を用紙上に形成する画像形成デバイスである。なお、本実施形態では、画像形成方式は、感光体ドラムや感光体ベルトを用いた電子写真方式であるが、これに限られるものではない。例えば、微少ノズルアレイからインクを吐出して用紙上に印字するインクジェット方式などであっても本発明は適用可能である。また、画像出力部２０５には、異なる用紙サイズまたは異なる用紙向きを選択可能とする複数の用紙カセット２１３乃至２１５が設けられている。排紙トレイ２１６には印字後の用紙が排出される。
【００２７】
図３は、サーバのハードウェア構成の一例を示す図である。
サーバ１０２は、ＣＰＵ３０１、ＲＡＭ３０２、ＲＯＭ３０３、ネットワークＩ／Ｆ部３０４、ＨＤＤ３０５、及びデータバス３０６を備える。ＣＰＵ３０１は、ＲＯＭ３０３に記憶された制御プログラムを読み出してＲＡＭ３０２にロードし、各種制御処理を実行する。ＲＡＭ３０２は、ＣＰＵ３０１の実行するプログラムや、ワークメモリ等の一時記憶領域として用いられる。
【００２８】
ネットワークＩ／Ｆ部３０４は、サーバ１０２をインターネット１０４に接続し、他の装置との間で各種情報を送受信する。ＨＤＤ３０５は、画像データや特徴量データ、各種プログラム等を格納する。ネットワークＩ／Ｆ部３０４を介してＭＦＰ１０１から受信した画像データは、データバス３０６を介してＣＰＵ３０１、ＲＡＭ３０２、及びＲＯＭ３０３に送受信される。
【００２９】
ＣＰＵ３０１がＲＯＭ３０３やＨＤＤ３０５に格納された画像処理プログラムを実行することによって、画像データに対する画像処理が実現される。また、ＨＤＤ３０５は、ネ

ットワークＩ／Ｆ部３０４を介してＭＦＰ１０１以外の外部装置からもデータの入力が可能であり、すでに文書の画像データを含む大量のファイルが格納されているものとする。

【００３０】

＜第１実施形態の詳細説明＞

図４は、学習セットを生成し、分類ルールを学習する処理を説明するフローチャートである。

図４に示す処理は、ＭＦＰ１０１およびサーバ１０２にて実行される。ＭＦＰ１０１において実行される処理は、ＣＰＵ２０４が記憶部２０３に格納されている処理プログラムをロードして実行することにより実現される。また、サーバ１０２において実行される処理は、ＣＰＵ３０１がＨＤＤ３０５に格納されている処理プログラムをＲＡＭ３０２にロードして実行することにより実現される。

【００３１】

なお、本実施形態では、ユーザがＭＦＰ１０１を用いて文書（原稿）をスキャンし、その種別毎に文書の画像データをサーバ１０２に格納するというワークフローの中で、同時にＭＦＰ１０１内で画像データの分類ルールを学習するシステムを想定している。このように、文書を扱うワークフローに機械学習を利用した分類ルールを応用すると、スキャナを備えたＭＦＰなどの入力機器から入力された文書の格納先や配布先の自動決定、ファイル名の自動付与などが可能になる。また、ユーザ毎に用意した文書から学習することで、個別にカスタマイズされた分類ルールを構築することも可能になる。

【００３２】

なお、文書のスキャン及びサーバへの格納と、分類ルールの学習を行うタイミングは上記のワークフローに限られるものではなく、文書のスキャン及びサーバへの格納と、分類ルールの学習が別々に実行されてもよい。第２実施形態では、すでにサーバ１０２に格納されたデータを分類ルールの学習時に取得する場合について説明する。また、分類ルールの学習は、データを読み込んだＭＦＰ１０１で必ずしも行う必要はなく、例えば画像データを格納したサーバで本実施形態の分類に係る処理を行ってもよい。

【００３３】

ステップＳ４０１において、ＭＦＰ１０１は、ユーザから操作部２０７経由で画像データの保存設定を受付ける。なお、画像データの保存設定は、ＭＦＰ１０１において読み込んだ画像データの保存先を示すフォルダのパスや、保存時のファイル名、ファイル形式などのことである。

【００３４】

ステップＳ４０２において、ＭＦＰ１０１は、操作部２０７からユーザの指示を受け付けると、原稿フィーダ２１１のトレイ２１２から原稿を１枚ずつフィードして、画像読取部２０１で原稿を読み取る。なお、本実施形態では、トレイ２１２にセットされる原稿は、同一種別の文書とする。また、同一種別の文書は、分類ルールにおいて同一のクラスに分類される文書とする。

【００３５】

ステップＳ４０３において、ＭＦＰ１０１は、ステップＳ４０２で画像読取部２０１が読み込んだ画像データ群を、記憶部２０３に学習セットの正例データ候補群として格納する。画像データ群を格納する際には、各画像データに文書ファイル情報を付与する。文書ファイル情報は、後述する負例データ候補群の絞り込みに利用する。文書ファイル情報としては、タイトルや作成者名、ファイル形式、作成ツール、作成デバイス、変換ツール、キーワード、生成日時、更新日時など、アプリケーションで電子ファイルを作成する際に付与される一般的なメタ情報を利用する。

【００３６】

キーワードとは、文書ファイルの特徴を表す文字列群であり、本実施形態では、原稿を読み込む際に文字認識を行い、その結果を利用する。例えば、タイトルとなる最初のページの上部中央や、ヘッダーやフッター、表内の項目など文書の特徴的な位置にある文字列、他の文字と比べてフォントの異なる文字列など、特徴的な文字列をキーワードとして利

用する。
【００３７】
　また、文書ファイル情報用のキーワード群と対応する項目とを辞書として保持しておき、文字認識を行った結果、辞書内のキーワードに当てはまる文字列が含まれる場合に、当該キーワードに対応する項目を文書ファイル情報のキーワードとして付与してもよい。文書ファイル情報用のキーワード群としては、「決裁書」や「申請書」、「注文書」といった一般的に利用される文書のタイトルや、企業名リストを利用する。
【００３８】
　なお、文書ファイル情報は、上記のようなメタ情報に限定されるものではなく、文字認識の過程等で得られる文字列の位置情報やフォントサイズなどの文書構造情報を用いてもよい。また、ＭＦＰ１０１での読み取り時に付与された読取解像度や色、割り当てなどのスキャン設定を用いてもよい。また、本実施形態では、文書ファイル情報と共に画像データ群が格納されるが、これに限定されるものではなく、例えば、読み込まれた画像データ群から算出される特徴量のデータを格納してもよい。
【００３９】
　ステップＳ４０４において、ＭＦＰ１０１は、ネットワークＩ／Ｆ部２０６を通じて画像読取部２０１で読み込まれた画像データ群をサーバ１０２に送信する。サーバ１０２は、ＬＡＮ１０３およびインターネット１０４を経由してＭＦＰ１０１から画像データ群を受信する。サーバ１０２のＣＰＵ３０１は、ステップＳ４０１において設定された画像データの保存設定に基づき、受け付けた画像データをＨＤＤ３０５に記録する。
【００４０】
　ステップＳ４０５において、ＭＦＰ１０１は、原稿の読み取りを続けるか否かの指示を、操作部２０７を介してユーザから受け付ける。原稿の読み取りを続ける場合には、処理はステップＳ４０１に戻る。原稿の読み取りを続けない場合には、処理はステップＳ４０６に進む。なお、原稿の読み取りを続けるか否かの判断は、上記の方法に限るものではない。例えば、ステップＳ４０１での原稿の読み取り回数をカウントし、あらかじめ操作部２０７を介してユーザによって設定された原稿の読み取り回数に達するまで原稿の読み取りを続けてもよい。
【００４１】
　ステップＳ４０６において、ＭＦＰ１０１は、記憶部２０３に格納されている文書の画像データおよび、インターネット１０４およびＬＡＮ１０３を経由してサーバ１０２から取得した文書の画像データを、負例データ候補群として記憶部２０３に格納する。ステップＳ４０７において、ＭＦＰ１０１は、ステップＳ４０６にて取得した負例データ候補群のファイルを抜粋する。負例データ候補群の抜粋処理の詳細については、図５を用いて後述する。
【００４２】
　ステップＳ４０８において、ＭＦＰ１０１は、ステップＳ４０３にて格納した正例データ候補群、およびステップＳ４０７にて格納した負例データ候補群を学習セットとして機械学習を利用した分類ルールの学習に用いる。本実施形態において、分類ルールの学習については、図９～１１を用いて後述する。
【００４３】
＜負例データ候補群の抜粋処理に係る詳細説明（ステップＳ４０７）＞
　運用時に入力される可能性の低い文書データは、運用時の分類精度に寄与しない無駄なデータとなってしまう。例えば、サーバからランダムに選ばれた５０個の文書データの中に使われていないデータが５個、別の業務で利用するデータが１０個含まれていた場合、分類ルールの構築に有効なデータが３５個となってしまう。このように、ランダムにデータを取得するだけでは、実際に分類時に入力される可能性の高い文書を減らしてしまう要因となる。
【００４４】
　また、データの冗長性を考慮していない場合も、実際に分類時に入力される可能性の高

い文書を減らしてしまう要因となる。例えば、負例データとして利用する文書データ５０個が、５種類各１０個の文書である場合と、５０種類各１個の文書である場合には、前者の方が分類時に入力される可能性の高い文書を減らしてしまう。本実施形態では、負例データ候補群の抜粋処理により、高精度な分類を可能とする負例データを取得することが可能となる。

【００４５】

図５は、負例データ候補群から負例データを構築する処理を示すフローチャートである。

詳細には、図５に示す処理は、分類ルールの構築に使用する学習セットの一部である負例データを、ステップＳ４０６にて取得した負例データ候補群から抜粋する処理である。図５に示す処理は、ＭＦＰ１０１のＣＰＵ２０４が、記憶部２０３に格納されている処理プログラムをロードして実行することで実現される。

【００４６】

ステップＳ５０１において、ＭＦＰ１０１は、ステップＳ４０３で記憶部２０３に記録された正例データ候補群から、ステップＳ４０３で付与された文書ファイル情報およびユーザの指示に基づき、負例データ候補群の絞り込み条件を取得する。ステップＳ５０２において、ＭＦＰ１０１は、ステップＳ５０１で取得した絞り込み条件に基づき、ステップＳ４０６で取得した負例データ候補群を絞り込む（抜粋する）。ステップＳ５０１およびステップＳ５０２の詳細については、図５～図８を用いて後述する。

【００４７】

ステップＳ５０３において、ＭＦＰ１０１は、ステップＳ５０２で抜粋した負例データ候補群から、冗長なデータを削減する。冗長なデータの特定には、例えば、文書ファイル情報の１つであるキーワードを特徴量としたクラスタリングを利用する。これは、同じキーワードで構成される文書は、同じ種別の文書である可能性が高いため、同じ種別の文書であると判定するためである。

【００４８】

なお、冗長なデータの特定は、上記の方法に限るものではない。例えば、キーワード以外の特徴量として文書構造情報に基づきタイトル文字列やタイトル文字列のフォントサイズ、タイトル文字列の位置等を特徴量としたクラスタリングを利用してもよく、また、それ以外の方法を用いてもよい。そして、同じ種別であると判定された文書が大量にある場合は、それらの中から一部を抜粋して、残りの文書は削除することにより冗長なデータを削減することができる。このとき、例えば、あらかじめ文書のデータ容量や個数の上限を決めておき、当該データ容量や個数が上限を超えた場合に、それらが上限の値以下となるように文書を削除すればよい。

【００４９】

ステップＳ５０４において、ＭＦＰ１０１は、ステップＳ５０３で冗長なデータが削減された負例データ候補群から正例データ候補群に含まれる種別の可能性がある文書を削除する。正例データ候補群に含まれる種別であるか否かの判定は、正例データの文書とキーワードが一致する確率（一致率）に基づいて行う。なお、正例データ候補群に含まれる種別であるか否かの判定は、上記の方法に限るものではない。

【００５０】

ここでの判定は、分類ルールを用いて「その他」に分類するか否かを判定する際の精度は必要としていない。文書構造情報の一致率や、画像特徴量の一致率を利用して、正例の種別であると疑わしい文書を削除できればよい。また、すでに分類ルールを一度構築しており、正例データの種別を追加する場合であれば、構築済みの分類ルールを適用して正例データの種別であるか否かを判定してもよい。

【００５１】

なお、本実施形態では、サーバ１０２から取得した画像データ群をＭＦＰ１０１が絞り込む処理を実行することにより負例データを作成したが、これに限られるものではない。例えば、図６を用いて説明する絞り込み条件に従って、サーバ１０２がデータの絞り込み

を行い、作成した負例データをＭＦＰ１０１に送信してもよい。
【００５２】
＜絞り込み条件の取得および絞り込み処理の詳細説明（ステップＳ５０１、Ｓ５０２）＞
絞り込み条件の取得および絞り込みの処理は、ＭＦＰ１０１のＣＰＵ２０４が実行する処理である。絞り込み条件の取得について、図６および図７を用いて説明する。
図６は、文書ファイル情報の一例を示す図である。正例データ候補群として与えられた３種別の文書に関して、文書ファイル情報を示している。図７は、正例データ候補群の文書ファイル情報による絞り込み条件をユーザが確認および編集するための画面の一例を示す図である。
【００５３】
図７の画面は、ラジオボタン７０１および７０２を有する。ラジオボタン７０１および７０２により、絞り込み条件を設定するか否かを切り替えることができる。ボタン７０３は、負例データ候補群の取得を指示（要求）するためのボタンであり、ラジオボタン７０１および７０２の状態に応じて取得する処理を切り替える。
【００５４】
具体的には、ラジオボタン７０１が選択されている場合には、条件式フィールド７０５および条件フィールド７０６において設定された内容に基づいて、記憶部２０３およびＨＤＤ３０５内の文書を絞り込んで取得する。ラジオボタン７０２が選択されている場合には、記憶部２０３およびＨＤＤ３０５内の文書を絞り込まずにそのまま取得する。ボタン７０４は、絞り込み条件の自動取得を指示するためのボタンである。
【００５５】
ボタン７０４によって絞り込み条件の自動取得が指示されると、ＭＦＰ１０１は、正例データ候補群の文書ファイル情報から絞り込み条件を取得して、条件式フィールド７０５および条件フィールド７０６に表示する。具体的には、条件式フィールド７０５および条件フィールド７０６には、図６に示した正例データ候補群の文書ファイル情報に基づいて、正例データ候補群の文書ファイル情報と１つでも共通の項目を含む文書が取得できる条件式が示される。条件フィールド７０６には、正例データ候補群の文書ファイル情報の各項目が条件として表示される。
【００５６】
また、条件式フィールド７０５には、条件フィールド７０６の各条件が、和集合を表す「＋」で結合された条件式が入力されている。すなわち、正例データ候補群の文書ファイル情報の各項目と１つでも共通の項目を含む文書が抽出される。なお、条件式の自動取得では、上記のように文書ファイル情報の各項目の和集合を抽出する方法に限られるものではない。例えば、正例データ候補群の間で、文書ファイル情報の共起性を計算し、共起性の高い文書ファイル情報の項目の組み合わせを含む文書を絞り込むように条件を表示してもよい。すなわち、正例データ候補群において付与されている頻度が高い文書ファイル情報の項目の組み合わせを使用して、文書を絞り込むようにしてもよい。
【００５７】
条件式フィールド７０５および条件フィールド７０６は、条件を表示するだけでなく、ユーザによる編集も受け付ける。ユーザは、ボタン７０４を用いて自動取得した条件を修正したい場合には編集すればよく、また、ユーザ所望の文書を絞り込むための条件を任意に設定することも可能である。ボタン７０９によって条件式フィールド７０５および条件フィールド７０６表示された条件をクリアすることも可能である。
【００５８】
また、図７に示す例では、条件フィールド７０６に条件番号７までの条件が一覧されているが、これらの数は可変であり、上限も現在表示されている１０個に限られるものではない。ボタン７１０によって、条件の追加が指示されると、条件の数（行数）を増やすことが可能である。また、条件式フィールド７０５において、条件フィールド７０６に表示されている条件番号と括弧や演算子を用いて多項演算のように条件式を入力することも可能である。例えば、和集合であれば「＋」の演算子で表記し、積集合であれば「＊」の演

算子で表記する。

【００５９】

また、条件式フィールド７０５および条件フィールド７０６で表現される絞り込み条件は、ファイルに保存または読み込みが可能である。ボタン７０７は、絞り込み条件をファイルに保存するためのボタンであり、ボタン７０７が押下されると条件式フィールド７０５および条件フィールド７０６に表示されている絞り込み条件がテキストファイル形式にて保存される。

【００６０】

また、ボタン７０８は、絞り込み条件をファイルから読み込むためのボタンであり、ボタン７０８が押下されるとファイルから読み込んだ絞り込み条件が、条件式フィールド７０５および条件フィールド７０６に表示される。なお、絞り込み条件を保存するファイルの形式は、テキストファイル形式に限られるものではなく、条件を表現することができれば特に限定されない。例えば、ＸＭＬ形式に保存してもよい。

【００６１】

図８は、記憶部２０３およびＨＤＤ３０５内の文書ファイルを、上記の絞り込み条件によって絞り込んだ結果の一例を示す図である。

図７に示した条件によって絞り込んだ場合に、負例データとして採用されるデータの１つがデータ８０１である。

【００６２】

文書ファイル情報の項目８０２（作成者名）、項目８０３（形式）、及び項目８０４（作成デバイス）が、それぞれ条件７１１、７１２、７１３と一致するため、採用される。一方、負例データとして採用されないデータの１つがデータ８０５である。文書ファイル情報の項目が、条件フィールド７０６に示す条件のいずれにも一致しないため、負例データとして採用されず、負例データ候補群から削除される。

【００６３】

＜機械学習を利用した分類ルールの構築の詳細説明（ステップＳ４０８）＞

次に、本実施形態で分類ルールの構築に利用する機械学習の手法について説明する。本実施形態では、機械学習の手法としてＲｅａｌ　ＡｄａＢｏｏｓｔと呼ばれる公知の手法を利用する。Ｒｅａｌ　ＡｄａＢｏｏｓｔは、大量の特徴量から、与えられた学習セットの分類に適した特徴量を選択して、その特徴量を組み合わせて分類器（分類ルール）を構成することが可能な手法である。

【００６４】

画像の分類時に大量の特徴量を利用すると、特徴量の計算負荷のためにパフォーマンスが低下する可能性がある。Ｒｅａｌ　ＡｄａＢｏｏｓｔのように、分類に適した特徴量を選択して、一部の特徴量だけを利用し、分類器を構成できることは、大きな利点である。ただし、Ｒｅａｌ　ＡｄａＢｏｏｓｔは、２クラス分類器であり、２種類のラベルがついたデータを分類するものである。つまり、このままでは、３種類以上の種別の画像データの分類には利用できない。

【００６５】

そこで、本実施形態では、２クラス分類器を多クラス分類器に拡張するＯＶＡ（Ｏｎｅ－Ｖｅｒｓｕｓ－Ａｌｌ）と呼ばれる公知の方法を利用する。ＯＶＡは、１つのクラス（対象クラス）とそれ以外のクラスを分類する分類器をクラスの数だけ作成し、それぞれの分類器の出力を、対象クラスの信頼度とする。すなわち、１つの分類器では、その分類器が分類するクラスに属するデータを正例データとし、それ以外のクラスに属するデータを負例データとして分類ルールを学習する。

【００６６】

各分類器は、その分類器が対象とする１つのクラスのデータが入力された場合に、出力する信頼度が高くなるように学習を行う。分類の際には、分類したいデータをすべての分類器に入力し、信頼度が最大であったクラスを分類先とする。また、すべての分類器の出力する信頼度が小さい場合や、複数の分類器が出力する信頼度が高くなった場合には、「

該当なし」や「不明」といった判定を行う。

【００６７】

図９は、学習セットを用いた機械学習の一例を示す図である。

この例では、学習セットとして、正例データ候補群の３つのクラス（種別）の文書（文書Ａ、文書Ｂ、文書Ｃ）および負例データ候補群の「その他」の文書（文書Ａ、文書Ｂ、文書Ｃではない文書）のそれぞれに対応する特徴量が用意されているものとする。この文書Ａ、文書Ｂ、文書Ｃの３種類のクラスを分類するために、ＯＶＡでは３種類の分類器を用意する。３種類の分類器はそれぞれ、文書Ａとそれ以外のクラスに文書を分類するための文書Ａ分類器、文書Ｂとそれ以外のクラスに文書を分類するための文書Ｂ分類器、文書Ｃとそれ以外のクラスに文書を分類するための文書Ｃ分類器である。

【００６８】

ここで、文書Ａ分類器を構築する方法について説明する。まず、ＭＦＰ１０１のＣＰＵ２０４は、分類ルールを学習するにあたって必要となる正例データおよび負例データを、学習セットの中から取得する。文書Ａ用分類器では、正例データは文書Ａのデータであり、負例データはそれ以外のクラスのデータである。したがって、ＣＰＵ２０４は、正例データ候補群の中から、文書Ａのラベルが付与された画像データを取得し、正例データとする。

【００６９】

また、ＣＰＵ２０４は、正例データ候補群の中から、文書Ａ以外（文書Ｂ、文書Ｃ）のラベルの付与された画像データを、負例データとして取得する。さらに、ＣＰＵ２０４は、負例データ候補群の中から、画像データを負例データとして取得する。このとき、負例データ候補群の中に、正例データである文書Ａのデータが混ざっている場合には、正しく分類ルールを学習することができない。このため、上記のステップＳ５０４の処理により文書Ａである可能性の高いものは取り除かれているものとする。

【００７０】

ＣＰＵ２０４は、取得した正例データおよび負例データの特徴量に基づき、ＲｅａｌＡｄａＢｏｏｓｔを利用して文書Ａ分類器を構築する。文書Ａ分類器では、文書Ａの特徴量が入力された場合に、大きい出力値（信頼度）が出力され、それ以外のクラスの文書の特徴量が入力された場合に、小さい出力値（信頼度）が出力される。文書Ｂ分類器、文書Ｃ分類器についても同様である。

【００７１】

なお、本実施形態で利用可能な機械学習の手法は、上記の手法に限定されるものではない。Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ ＭａｃｈｉｎｅやＲａｎｄｏｍ Ｆｏｒｅｓｔ等の公知の手法を利用してもよい。また、特徴量選択の枠組みが機械学習の手法に含まれていない場合に、分類時の分類速度を向上させたい場合には、主成分分析や判別分析を利用した特徴量選択等の公知の特徴量選択を行ってもよい。機器学習の手法が２クラス分類器である場合は、ＯＶＡ以外の、Ａｌｌ－Ｖｅｒｓｕｓ－Ａｌｌ（ＡＶＡ）やＥｒｒｏｒ－Ｃｏｒｒｅｃｔｉｎｇ Ｏｕｔｐｕｔ－Ｃｏｄｉｎｇ（ＥＣＯＣ）等の公知の手法を用いてもよい。

【００７２】

＜分類ルールの構築に利用する特徴量の詳細＞

本実施形態において分類ルールの構築に利用する特徴量について、図１０および図１１を用いて説明する。

【００７３】

図１０は、特徴量の算出方法について説明する図である。

本実施形態において特徴量は、入力画像１００１内から切り出されたパッチ画像１００２に対して勾配情報に基づき算出される９次元の特徴量である。ＭＦＰ１０１のＣＰＵ２０４は、パッチ画像１００２内の各画素について注目し、注目画素に隣接する画素の階調値から、勾配強度および勾配方向を算出する。

【００７４】

そして、ＣＰＵ２０４は、勾配強度に基づいてエッジ判定を行うことで、勾配強度が一定値以上の画素をエッジ画素、一定値より小さい画素を非エッジ画素と判定する。エッジ判定の結果、画素１００３は、非エッジ画素と判定された画素の一例であり、画素１００４は、エッジ画素と判定された画素の一例である。エッジ画素である画素１００４内の矢印は、勾配方向を表す。

【００７５】

勾配方向は、文字や罫線の線の方向を表現するため、１８０度回転した角度は同一方向とみなして、０～１８０度に正規化される。ＣＰＵ２０４は、エッジ画素群から勾配方向を２２．５度毎の８方向に量子化し、方向ごとの勾配強度積算値／パッチ画素数を計算して８ビンのヒストグラムを作成する。また、ＣＰＵ２０４は、非エッジ画素群から、非エッジ画素数／パッチ画素数を計算し、エッジ画素群から作成したヒストグラムと合わせて、１つのパッチ画像から９次元の特徴量を算出する。

【００７６】

エッジ画素と非エッジ画素を利用することで、罫線や文字の情報だけでなく、文書画像の大きな特徴である余白部分を表現することが可能になる。これまでの説明は、１つのパッチ画像１００２における特徴量の説明であるが、実際には、１つの入力画像から複数のパッチ画像を切り出して利用することにより、多数の特徴量を利用する。

【００７７】

図１１は、読み取った画像データからパッチ画像を切り出す方法について説明する図である。

ＣＰＵ２０４は、入力画像１１０１から余白をカットし、ノイズが表れやすい画像端１１０２を削除する。ＣＰＵ２０４は、余白カット後の画像１１０３を縮小することで、マルチスケール（複数の解像度の）画像を作成する。マルチスケールの画像を用意するのは、解像度ごとにエッジの構造が変わるためであり、画像読取部２０１の読取解像度や文書の解像度が多少異なっていても対応できるようにするためである。

【００７８】

画像１１０４は、余白カット後の画像１１０３を１／４に縮小した画像である。余白カット後の画像１１０３および縮小した画像１１０４から、パッチサイズと切り出し位置を変えながら、パッチ画像を切り出す。具体的には、まず、縮小した画像１１０４から、均等に１６分割して得られる１／１６サイズのパッチ画像１６枚と、均等に６４分割して得られる１／６４サイズのパッチ画像６４枚から、合計８０枚のパッチ画像を作成する。

【００７９】

また、余白カット後の画像１１０３から、同様に分割して８０枚のパッチ画像を作成することで、１枚の入力画像１１０１から、合計１６０枚のパッチ画像が得られる。各パッチ画像から９次元の特徴量を算出するため、１枚の入力画像１１０１から９×１６０＝１４４０次元の特徴量を算出することが可能となる。

【００８０】

なお、画像解像度、パッチサイズ、パッチ切り出し位置に関するパラメータは、上記の数字に限定されるものではない。また、算出する特徴量として、原稿の色の情報を利用するために、色ヒストグラムや色分散等を特徴量としてもよい。また、分類ルールの構築に利用する特徴量は、上記のような画像データに関する特徴量に限定されるものではない。例えば、負例データ候補群の絞り込みに利用するメタ情報や文書構造情報などの文書ファイル情報を利用してもよい。

【００８１】

また、本実施形態では、文書をＭＦＰ１０１により画像データとして読み込み、当該画像データを分類する場合について説明したが、これに限られるものではない。例えば、テキスト形式のデータに対しても、本発明の正例データを用いた負例データの絞り込みは適用可能である。

【００８２】

以上のように、本実施形態によれば、高精度な分類ルールの構築を可能とする高品質な

負例データを効率的に生成することができる。
【００８３】
（第２実施形態）
第１実施形態では、トレイ２１２にセットされ画像読取部２０１により一度に読み取られる原稿を正例データとして利用することを想定していた。これに対して、本実施形態では、トレイ２１２にセットされ画像読取部２０１により一度に読み取られる原稿に加え、すでにサーバ１０２上に格納された文書を正例データとして利用する場合を想定する。以下、第１実施形態との差分についてのみ説明する。
【００８４】
＜第２実施形態の詳細説明＞
図１２は、学習セットを生成し、分類ルールを学習する処理を説明するフローチャートである。
図１２に示す処理は、ＭＦＰ１０１およびサーバ１０２にて実行される。ＭＦＰ１０１において実行される処理は、ＣＰＵ２０４が記憶部２０３に格納されている処理プログラムをロードして実行することにより実現される。また、サーバ１０２において実行される処理は、ＣＰＵ３０１がＨＤＤ３０５に格納されている処理プログラムをＲＡＭ３０２にロードして実行することにより実現される。
【００８５】
なお、本実施形態では、ユーザがＭＦＰ１０１を用いて文書（原稿）をスキャンし、その種別毎に文書の画像データをサーバ１０２に保存するという業務フローの中で、同時にＭＦＰ１０１内で画像データの分類ルールを学習するシステムを想定している。さらに、本実施形態では、分類ルールの学習に利用する文書をサーバ１０２から取得することを想定している。
【００８６】
ステップＳ１２０１において、ＭＦＰ１０１は、正例データとして利用する文書を、原稿フィーダ２１１から読み込むか、サーバ１０２から選択するかを受け付ける。原稿フィーダ２１１から読み込む場合には、処理はステップＳ１２０２に進み、サーバ１０２から選択する場合には、処理はステップＳ１２０５に進む。ステップＳ１２０２およびステップＳ１２０３は、図４のステップＳ４０１およびステップＳ４０２と同様である。また、ステップＳ１２０４は、図４のステップＳ４０４と同様である。
【００８７】
ステップＳ１２０５において、ＭＦＰ１０１は、ユーザから操作部２０７経由でサーバ１０２のＨＤＤ３０５内のどの文書を利用するかの指示を受け付ける。サーバ１０２のＣＰＵ３０１は、ユーザの指示に基づきＨＤＤ３０５内の画像データ群を、インターネット１０４およびＬＡＮ１０３を経由してＭＦＰ１０１に送信する。ステップＳ１２０６において、ＭＦＰ１０１は、ステップＳ１２０３にて画像読取部２０１で読み込まれた画像データ群、または、ステップＳ１２０５にてサーバ１０２から受信した画像データ群を、記憶部２０３に学習セットの正例データ候補群として格納する。
【００８８】
格納する際には、各画像データに負例データ候補群の絞り込みにて利用する文書ファイル情報を付与する。画像読取部２０１で読み込まれた画像データ群に付与する文書ファイル情報は、図４のステップＳ４０３で付与する文書ファイル情報と同様である。一方、サーバ１０２から受信した画像データ群には、すでに文書ファイル情報が付与されている場合にはその文書ファイル情報を利用する。また、文書ファイル情報が不足している場合には不足している項目について、図４のステップＳ４０３で付与する文書ファイル情報と同様の文書ファイル類情報を付与する。
【００８９】
ステップＳ１２０７において、ＭＦＰ１０１は、正例データの登録を続けるか否かの指示を、操作部２０７を介してユーザから受け付ける。正例データの登録を続ける場合には、処理はステップＳ１２０１に戻る。正例データの登録を続けない場合には、処理はステ

ップＳ１２０８に進む。なお、正例データの登録を続けるか否かの判断は、上記の方法に限られるものではない。例えば、ステップＳ１２０６における正例データの登録数をカウントし、あらかじめ操作部２０７を介してユーザによって設定された正例データの登録数に達するまで正例データの登録を続けてもよい。ステップＳ１２０８～ステップＳ１２１０は、図４のステップＳ４０６～ステップＳ４０８と同様である。

【００９０】

このように、本実施形態によれば、負例データを作成する際に、ＭＦＰから入力された画像データ（正例データ）から得られる文書ファイル情報に加えて、サーバから取得された画像データにすでに付与されている文書ファイル情報を利用することができる。これにより、大量の文書の中からデータの容量を抑えつつ、分類時に入力される可能性の高いデータを負例データとして収集することができ、高精度の分類ルールを効率よく生成することが可能となる。

【００９１】

（その他の実施形態）

本発明は、上述の実施形態の１以上の機能を実現するプログラムを、ネットワーク又は記憶媒体を介してシステム又は装置に供給し、そのシステム又は装置のコンピュータにおける１つ以上のプロセッサーがプログラムを読出し実行する処理でも実現可能である。また、１以上の機能を実現する回路（例えば、ＡＳＩＣ）によっても実現可能である。

【００９２】

以上、本発明の好ましい実施形態について説明したが、本発明は、これらの実施形態に限定されず、その要旨の範囲内で種々の変形および変更が可能である。

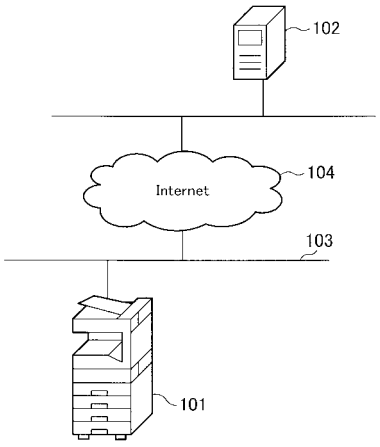
【符号の説明】

【００９３】

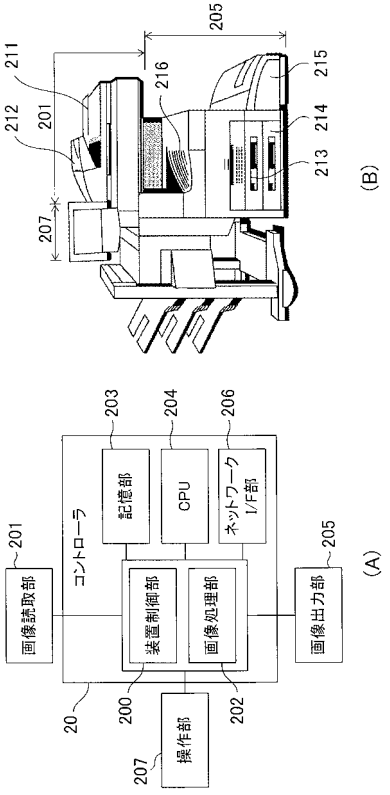
１０１　　ＭＦＰ

１０２　　サーバ

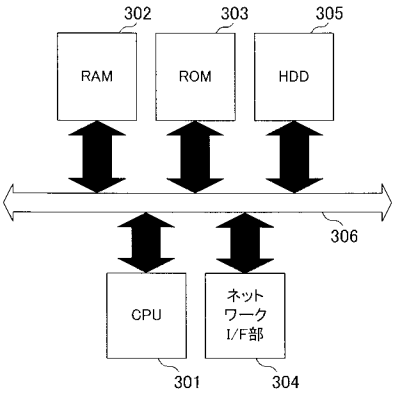
【図１】



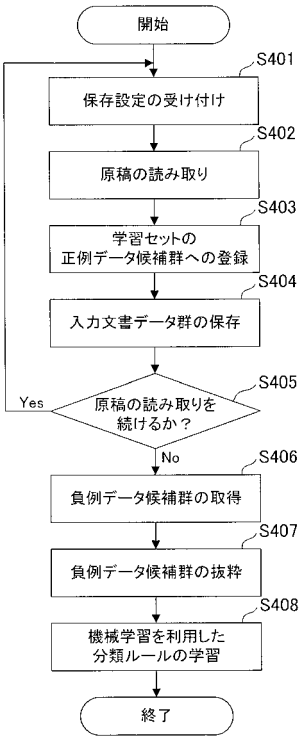
【図２】



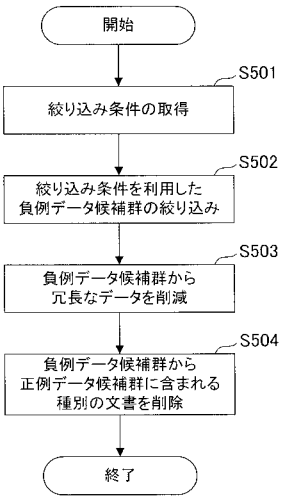
【図３】



【図４】



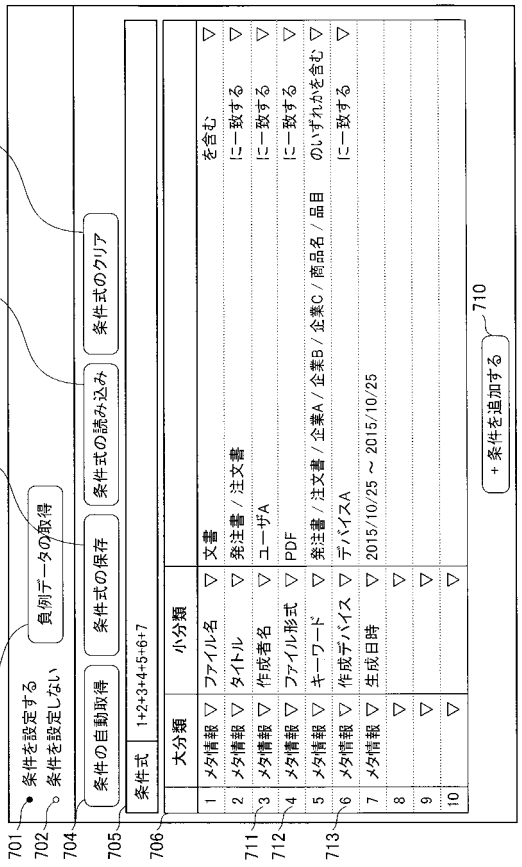
【図５】



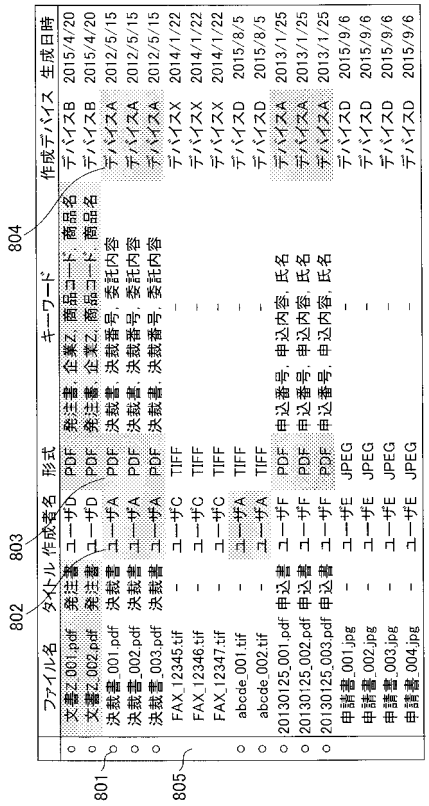
【図６】

クラス	ファイル名	タイトル	作成者名	形式	キーワード	作成デバイス	生成日時
文書A	文書A_001.pdf	発注書	ユーザA	PDF	発注書, 企業A, 商品コード, 商品名	デバイスA	2015/10/25
文書A	文書A_002.pdf	発注書	ユーザA	PDF	発注書, 企業A, 商品コード, 商品名	デバイスA	2015/10/25
文書A	文書A_003.pdf	発注書	ユーザA	PDF	発注書, 企業A, 商品コード, 商品名	デバイスA	2015/10/25
文書B	文書B_001.pdf	注文書	ユーザA	PDF	注文書, 企業B, 商品コード, 品目	デバイスA	2015/10/25
文書B	文書B_002.pdf	注文書	ユーザA	PDF	注文書, 企業B, 商品コード, 品目	デバイスA	2015/10/25
文書C	文書C_001.pdf	発注書	ユーザA	PDF	発注書, 企業C, 商品番号, 商品名	デバイスA	2015/10/25
文書C	文書C_002.pdf	発注書	ユーザA	PDF	発注書, 企業C, 商品番号, 商品名	デバイスA	2015/10/25
文書C	文書C_003.pdf	発注書	ユーザA	PDF	発注書, 企業C, 商品番号, 商品名	デバイスA	2015/10/25
文書C	文書C_004.pdf	発注書	ユーザA	PDF	発注書, 企業C, 商品番号, 商品名	デバイスA	2015/10/25

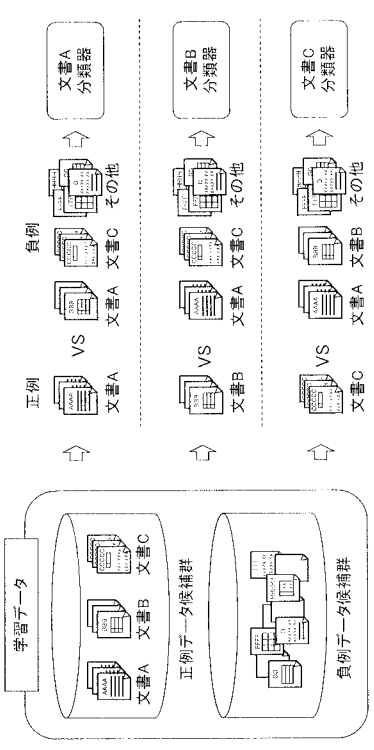
【図７】


条件を設定する
条件を設定しない
負例データの取得
条件の自動取得
条件式の保存
条件式の読み込み
条件式のクリア
条件式
1+2+3+4+5+6+7
大分類
小分類
メタ情報
ファイル名
文書
を含む
タイトル
発注書 / 注文書
に一致する
作成者名
ユーザA
ファイル形式
PDF
キーワード
発注書 / 注文書 / 企業A / 企業B / 企業C / 商品名 / 品目
のいずれかを含む
作成デバイス
デバイスA
生成日時
2015/10/25 ～ 2015/10/25
+ 条件を追加する
701
702
703
704
705
706
707
708
709
710
711
712
713

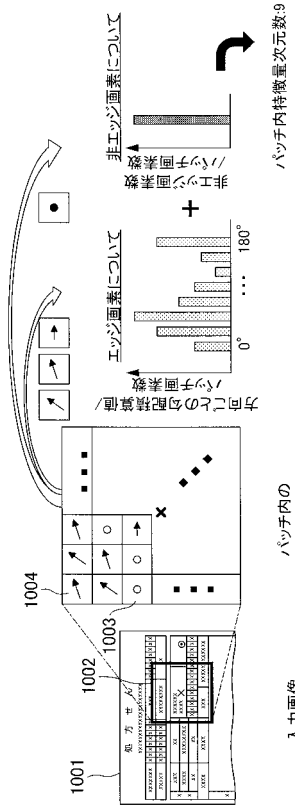
【図８】


ファイル名
タイトル
作成者名
形式
キーワード
作成デバイス
生成日時
文書Z_001.pdf 発注書 ユーザD PDF 発注書 企業Z, 商品コード, 商品名 デバイスB 2015/4/20
文書Z_002.pdf 発注書 ユーザD PDF 発注書 企業Z, 商品コード, 商品名 デバイスB 2015/4/20
決裁書_001.pdf 決裁書 ユーザA PDF 決裁書 決裁番号, 委託内容 デバイスA 2012/5/15
決裁書_002.pdf 決裁書 ユーザA PDF 決裁書 決裁番号, 委託内容 デバイスA 2012/5/15
決裁書_003.pdf 決裁書 ユーザA PDF 決裁書 決裁番号, 委託内容 デバイスA 2012/5/15
FAX_12345.tif - ユーザC TIFF - デバイスX 2014/1/22
FAX_12346.tif - ユーザC TIFF - デバイスX 2014/1/22
FAX_12347.tif - ユーザC TIFF - デバイスX 2014/1/22
abcde_001.tif - ユーザA TIFF - デバイスD 2015/8/5
abcde_002.tif - ユーザA TIFF - デバイスD 2015/8/5
20130125_001.pdf 申込書 ユーザF PDF 申込番号 申込内容, 氏名 デバイスA 2013/1/25
20130125_002.pdf 申込書 ユーザF PDF 申込番号 申込内容, 氏名 デバイスA 2013/1/25
20130125_003.pdf 申込書 ユーザF PDF 申込番号 申込内容, 氏名 デバイスA 2013/1/25
申請書_001.jpg - ユーザE JPEG - デバイスD 2015/9/6
申請書_002.jpg - ユーザE JPEG - デバイスD 2015/9/6
申請書_003.jpg - ユーザE JPEG - デバイスD 2015/9/6
申請書_004.jpg - ユーザE JPEG - デバイスD 2015/9/6
801
802
803
804
805

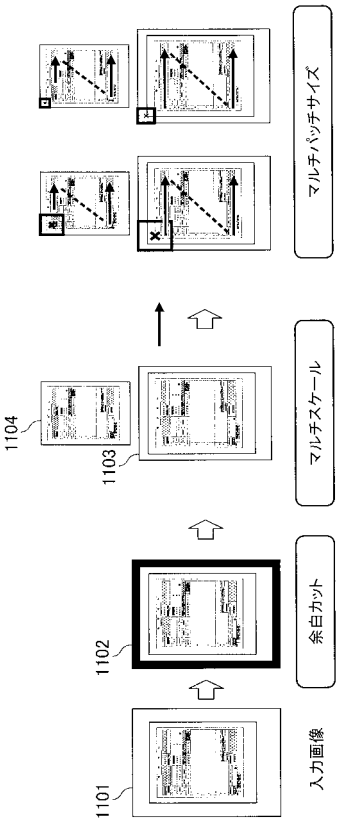
【図９】


学習データ
正例データ候補群
負例データ候補群
文書A
文書B
文書C
その他
正例
負例
VS
文書A 分類器
文書B 分類器
文書C 分類器

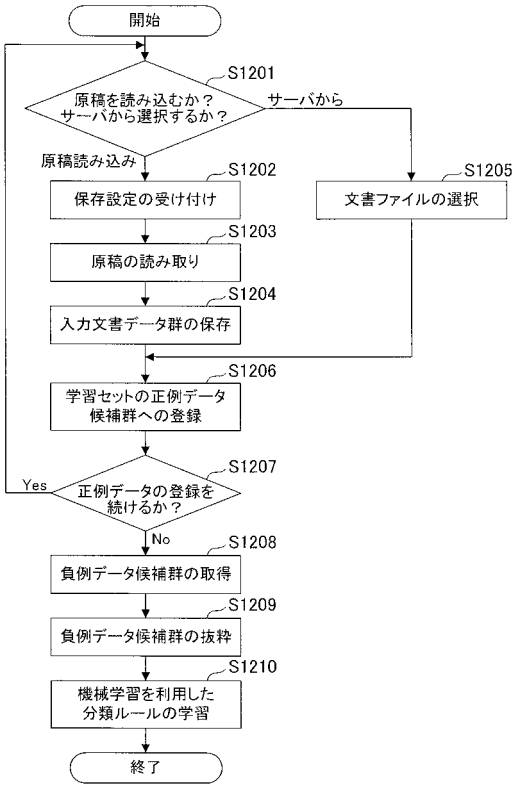
【図１０】


入力画像
パッチ内の
画像毎のエッジ判定結果
エッジ画素について
非エッジ画素について
パッチ内特徴量次元数:9
0°
180°
1001
1002
1003
1004

【図１１】


マルチバッチサイズ
マルチスケール
余白カット
入力画像
1101
1102
1103
1104

【図１２】


開始
S1201
原稿を読み込むか？
サーバから選択するか？
サーバから
原稿読み込み
S1202
保存設定の受け付け
S1205
文書ファイルの選択
S1203
原稿の読み取り
S1204
入力文書データ群の保存
S1206
学習セットの正例データ
候補群への登録
S1207
Yes
正例データの登録を
続けるか？
No
S1208
負例データ候補群の取得
S1209
負例データ候補群の抜粋
S1210
機械学習を利用した
分類ルールの学習
終了