(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)公開番号
特開2024-172025
(P2024-172025A)
(43)公開日　令和6年12月12日(2024.12.12)

(51)国際特許分類	FI	テーマコード（参考）
G06F 16/35 (2019.01)	G06F 16/35	5B175

審査請求　未請求　請求項の数　11　ＯＬ　（全13頁）

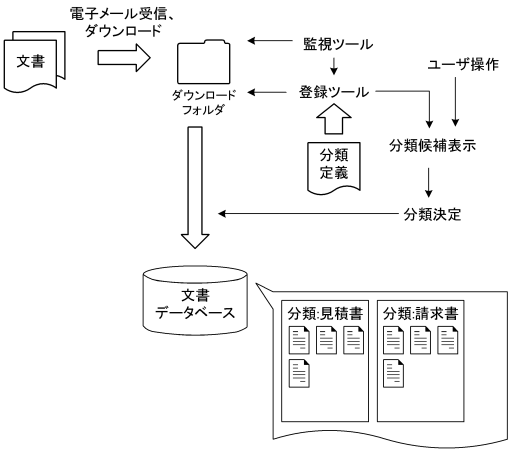
(21)出願番号　特願2023-89438(P2023-89438)
(22)出願日　令和5年5月31日(2023.5.31)

(71)出願人　000001443
カシオ計算機株式会社
東京都渋谷区本町１丁目６番２号
(74)代理人　110001254
弁理士法人光陽国際特許事務所
(72)発明者　上坂　重樹
東京都八王子市石川町２９５１番地の５
カシオ計算機株式会社　八王子技術センター内

Ｆターム（参考）　5B175　DA01 FA03

(54)【発明の名称】　情報処理装置、情報処理方法及びプログラム

(57)【要約】
【課題】ユーザによる電子文書データの分類に要する手間を低減することのできる情報処理装置、情報処理方法及びプログラムを提供する。
【解決手段】情報処理装置は、電子文書から１行ごとに抜き出した行に基づいて対象文字列を生成し、生成された対象文字列と所定の文字列とを比較した比較の結果に基づいて電子文書の分類先候補を出力する処理を行う処理部を備える。処理部は、抜き出した行に空白が含まれる場合には行から空白を削除して対象文字列を生成する。
【選択図】図２



【特許請求の範囲】
【請求項１】
電子文書から１行ごとに抜き出した行に基づいて対象文字列を生成し、生成された前記対象文字列と所定の文字列とを比較した比較の結果に基づいて前記電子文書の分類先候補を出力する処理を行う処理部を備え、
前記処理部は、抜き出した前記行に空白が含まれる場合には当該行から空白を削除して前記対象文字列を生成する情報処理装置。
【請求項２】
前記処理部は、電子文書の先頭の行から予め定められた行数だけ、前記処理を行う、請求項１記載の情報処理装置。
【請求項３】
前記処理部は、除外対象が前記対象文字列に含まれる場合に、前記対象文字列の前記出力の優先度を下げる、請求項１記載の情報処理装置。
【請求項４】
前記処理部は、前記電子文書が予め定められた位置に記憶された場合に、当該電子文書に対する前記処理を行う、請求項１記載の情報処理装置。
【請求項５】
前記位置は、電子メールに添付された文書が格納される設定位置である、請求項４記載の情報処理装置。
【請求項６】
前記電子文書は、電子帳簿に係る書類である、請求項１記載の情報処理装置。
【請求項７】
前記処理部は、前記電子文書から全文を抜き出して、前記全文中の改行の指定位置に基づいて前記行を各々決定する、請求項１記載の情報処理装置。
【請求項８】
前記処理部は、前記対象文字列を生成するときに、文字サイズが最大である文字が含まれる行を抜き出す、請求項１記載の情報処理装置。
【請求項９】
表示部と、操作受付部と、を備え、
前記処理部は、
前記分類先候補を前記表示部により表示させ、
前記操作受付部が受け付けた入力操作に応じた分類先に前記電子文書を分類する
請求項１記載の情報処理装置。
【請求項１０】
電子文書から１行ごとに抜き出した行に基づいて対象文字列を生成し、生成された前記対象文字列と所定の文字列とを比較した比較の結果に基づいて前記電子文書の分類先候補を出力する処理を行う情報処理方法であって、
抜き出した前記行に空白が含まれる場合には当該行から空白を削除して前記対象文字列を生成する
情報処理方法。
【請求項１１】
コンピュータに、
電子文書から１行ごとに抜き出した行に基づいて対象文字列を生成し、生成された前記対象文字列と所定の文字列とを比較した比較の結果に基づいて前記電子文書の分類先候補をかする処理を実行させ、
前記処理では、抜き出した前記行に空白が含まれる場合には当該行から空白を削除して前記対象文字列を生成する
プログラム。
【発明の詳細な説明】
【技術分野】

【０００１】
この発明は、情報処理装置、情報処理方法及びプログラムに関する。
【背景技術】
【０００２】
従来、画像から文字を認識して、分類項目を取得する技術が知られている（特許文献１）。また、特許文献２には、文書の画像データから認識された文字列から元の文書でなされていた強調表示を適切に読み取る技術が開示されている。
【０００３】
一方、多くの取引が電子文書により行われるようになってきている。電子文書は、ネットワークを介して迅速にやり取りされる。電子文書としては、ＰＤＦ（Portable Document Format）が幅広く利用されている。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】特開２００８－１７６６２５号公報
【特許文献２】特開２０１７－１２６２７０号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
ネットワークを介して受信された電子文書のデータは、担当者が分類して処理、保管する。しかしながら、電子文書の分量が増えるのに従って、分類作業に要する担当者の手間も大きくなっているという課題がある。
【０００６】
この発明の目的は、担当者による電子文書データの分類に要する手間を低減することのできる情報処理装置、情報処理方法及びプログラムを提供することにある。
【課題を解決するための手段】
【０００７】
上記目的を達成するため、本発明は、
電子文書から検出する内容を記憶する記憶部と、
電子文書に含まれる文字を抜き出して順番に並べたテキストから前記内容に応じた文字列を検出し、前記文字列に対応する分類を出力する処理を行う処理部と、
を備える情報処理装置である。
【発明の効果】
【０００８】
本発明に従うと、担当者による電子文書データの分類に要する手間を低減することができるという効果がある。
【図面の簡単な説明】
【０００９】
【図１】本実施形態の情報処理装置の機能構成を示すブロック図である。
【図２】情報処理装置における文書分類の流れについて説明する図である。
【図３】分類定義データを説明する図である。
【図４】電子帳簿書類の先頭付近の例を示す図である。
【図５】監視制御処理の制御手順を示すフローチャートである。
【図６】登録ツールにより実行される文書分類制御処理の制御手順を示すフローチャートである。
【発明を実施するための形態】
【００１０】
以下、本発明の実施の形態を図面に基づいて説明する。
図１は、本実施形態の情報処理装置１の機能構成を示すブロック図である。
【００１１】

情報処理装置１は、通常のＰＣ（Personal Computer）であってもよい。情報処理装置１は、ＣＰＵ１１（Central Processing Unit）（制御部）と、ＲＡＭ１２（Random Access Memory）と、記憶部１３と、表示部１４と、操作受付部１５と、通信部１６などを備える。

【００１２】

ＣＰＵ１１は、演算処理を行うプロセッサである。ＣＰＵ１１は、単一であってもよいし、複数のものが並列に動作、又は用途などに応じて各々独立に動作するのであってもよい。ＣＰＵ１１は、汎用プロセッサだけではなく、マイコン又はＡＳＩＣ（Application Specific Integrated Circuit）などであってもよい。ＣＰＵ１１は、処理部として情報処理装置１が取得した電子文書を分類して保管する。

【００１３】

ＲＡＭ１２は、ＣＰＵ１１に作業用のメモリ空間を提供し、一時データを記憶する。

【００１４】

記憶部１３は、不揮発性メモリを含む。不揮発性メモリは、例えば、フラッシュメモリやＨＤＤ（Hard Disk Drive）などである。不揮発性メモリには、各種プログラム及び設定データなどが記憶される。また、記憶部１３は、文書データベース１３４を記憶している。各種プログラムには、メールソフト１３１と、後述の監視ツール及び登録ツールを含む文書分類のためのプログラム１３２とが含まれる。設定データには、電子文書から検出して当該電子文書の分類に利用する内容を定義した分類定義データ１３５が含まれる。ダウンロードデータ１３３は、外部から取得されてダウンロードフォルダ（又はディレクトリ）に記憶されたデータである。このデータには、例えば、受信した電子メールに添付されたデータ、及びＨＴＴＰにより外部のＷｅｂサイトなどからダウンロードされたデータなどが含まれる。なお、ダウンロードフォルダの名称は、他のものであってもよい。

【００１５】

表示部１４は、デジタル表示画面を有し、ＣＰＵ１１の制御に基づいて種々の情報をデジタル表示画面に表示する。デジタル表示画面は、例えば、液晶ディスプレイ（ＬＣＤ）又は有機ＥＬ（Electro-Luminescent）ディスプレイなどである。

【００１６】

操作受付部１５は、ユーザなどの外部からの入力操作を受け付ける。操作受付部１５は、受け付けられた入力操作に応じた操作信号をＣＰＵ１１へ出力する。操作受付部１５は、例えば、キーボード及びポインティングデバイスなどを含み得る。ポインティングデバイスには、マウスが含まれていてもよい。また、操作受付部１５は、デジタル表示画面に重なって位置するタッチパネルを有していてもよい。

なお、情報処理装置１は、表示部１４及び操作受付部１５を有していなくてもよい。これらは、周辺機器として、ＵＳＢ（Universal Serial Bus）端子又はＰＳ／２端子などの接続端子を介して情報処理装置１に外付けされていてもよい。あるいは、これらは外部機器であって、通信部１６を介して情報処理装置１と通信接続されてもよい。

【００１７】

通信部１６は、外部機器との通信を所定の規約（プロトコル）に従って制御する。所定の規約には、例えば、ＬＡＮ（Local Area Network）におけるＴＣＰ／ＩＰなどが含まれ得る。通信部１６は、ブルートゥース（登録商標）やＷｉＦｉなどの無線通信を制御するためのネットワークカードを有していてもよい。通信部１６は、各々の通信規約に従って、外部機器と通信が可能であってもよい。外部機器には、上記のように表示部１４及び操作受付部１５の構成が含まれていてもよい。

本実施形態のコンピュータは、少なくともＣＰＵ１１とＲＡＭ１２とを含み、記憶部１３及び通信部１６などを含み得る。また、本実施形態の情報処理装置１の全体がコンピュータに対応してもよい。

【００１８】

次に、本実施形態の情報処理装置１による文書分類処理について説明する。

図２は、情報処理装置１における文書分類の流れについて説明する図である。

【００１９】
情報処理装置１では、電子帳簿に係る電子文書データが分類されて文書データベース１３４に登録される。分類対象とされる電子文書のフォーマットは、ＰＤＦファイルである。電子文書が特定のフォルダ又はディレクトリ（予め定められた位置）、例えば、ダウンロードフォルダに格納、記憶されると、監視ツールが電子文書の追加を検出する。すなわち、監視ツールは、常駐プログラムであってもよい。監視ツールは、電子文書の追加を検出すると、登録ツールを起動させる。
【００２０】
登録ツールは、追加されたＰＤＦファイルを解析する。登録ツールは、ＰＤＦファイル（電子文書）からテキスト（表示される文字；数字、記号及び標識などを含む）を１行分ずつ順番に抜き出したテキストデータを対象文字列として取得する。ＰＤＦファイルがタグ付きデータの場合には、登録ツールは、このＰＤＦファイルを構造解析して、表示内容のテキストを全文抜き出す。登録ツールは、テキスト内で改行を指示する位置を特定して、当該位置を区切りとして、１行分のテキスト（行）ごとに分割する。登録ツールは、行に分割する際に、改行を示す制御コードやタグを削除してもよい。また、登録ツールは、全文を抜き出す際、又は行に分割する際に、対象文字列内で表示上の改行とは関係のないデータ上の改行を削除又は無視してもよい。さらに、登録ツールは、このときに抜き出したテキストに対応するタグデータからフォントサイズ、フォント種別及びフォントカラーなどの表示設定を特定してもよい。登録ツールは、抜き出された各行のテキスト（対象文字列）をそれぞれ分類定義データ１３５と比較して、分類定義データ１３５により定義されている検出内容であるキーワード（所定の文字列）を検索する。登録ツールは、検出された内容（文字列）に応じて分類先候補を抽出して、表示部１４により候補を表示（出力）させる。この候補に対して操作受付部１５が選択に係る入力操作を受け付けると、入力操作に応じて分類先が決定されて、ＰＤＦファイルが分類情報とともに文書データベース１３４に登録される。
【００２１】
なお、ダウンロードフォルダには、登録すべき電子文書データ以外のファイルが記憶され得る。ＰＤＦ形式以外のファイルが追加された場合には、監視ツールは、登録ツールを起動させない。ＰＤＦ形式のファイルが追加されて登録ツールが起動された場合でも、ユーザは、このファイルを登録しない選択操作を行うことができる。この場合には、ＰＤＦファイルは、そのままダウンロードフォルダに残され、他の任意の用途などに用いられ得る。
【００２２】
また、監視ツールは、ダウンロードフォルダ以外の設定されたフォルダ（設定位置）に追加される新規ファイルを監視するのであってもよい。電子帳簿に係るＰＤＦファイルが他のファイルとは異なる専用フォルダに一時記憶されることで、分類が必要なファイルが記憶された場合にのみ登録ツールが起動される。
【００２３】
また、ＰＤＦファイルは、テキスト部分を含めて表示内容が全て画像データである場合がある。この場合には、登録ツールは、電子帳簿データではないと判断してもよい。あるいは、登録ツールは、周知の文字認識技術を利用して、画像からテキストを読み取ってもよい。この場合、登録ツールは、テキストの内容とともに、各文字のフォントサイズ、フォント種別及びフォントカラーなどを読み取ってもよい。
【００２４】
図３は、分類定義データ１３５を説明する図である。
図３（ａ）に示すように、分類項目（キー）には、電子文書の表題に応じた文書種別が含まれ得る。「見積書」、「請求書」、「注文書」などは、表題（タイトル）がそのまま電子文書の分書種別（キー）に係る分類種別を表す文字列（キーワード）であり得る。表題は、文書の先頭にあることが多いので、文書の先頭付近で優先的にこれらの文字列が検索されてもよい。あるいは、文書の先頭の行から予め定められた行数、例えば、１行目か

ら３行目までの３行だけで、これらの文字列が検索されてもよい。また、上記のように構造解析により表示設定が取得されている場合には、特定の表示設定、例えば、フォントサイズが他の部分よりも大きい行で優先的に又は選択的に文字列を検索してもよい。

【００２５】

また、このような表題では、しばしば各文字の間にスペース（空白。全角半角、数を問わない。また、タブなどによるもの、タグによって空白両脇の文字の位置が別個に指定されたものなども含まれる）が挿入されている。登録ツールは、抽出したテキスト（全文まとめて又は行ごと）からこのスペースを削除して、検出内容が検索される対象の文字列（対象文字列）を特定（生成）する。テキスト内にスペースがない場合には、行のテキストがそのまま対象文字列とされればよい。テキストが一行ごとに区分されることで、対象文字列では、複数の行の文字が不必要につながらない。

【００２６】

「注文書」及び「発注書」は、異なる表題であるが、文書としては同種のものである。したがって、ここでは、分類定義データ１３５においてカンマにより区切られて同一行に記載されることで、同一分類とされる。

【００２７】

図３（ｂ）に示すように、分類は、また、書類の自社（団体）からの宛先又は自社（団体）への発送元など取引先種別に応じてなされ得る。すなわち、取引先種別を分類項目（キー）として、宛先又は発送元を表す法人の名称が分類のキーワードであってもよい。宛先は、文書の上部にあることが多いが、先頭には限られない。情報処理装置１（登録ツール）は、抽出したテキストから宛先に含まれることの多い法人の種別を含む文字列（キーワード）を検出する。法人の種別は、例えば、株式会社、有限会社、合名会社、合資会社、相互会社、合同会社などであり、しばしば「（株）」などのように括弧付きで省略表記され得る。登録ツールは、これらの文字列の候補を正規表現により予め設定しておくことで、宛先の候補を検索する。また、宛先候補の文字列と同一行内の「御中」、「様」、「送付先」、「送付元」などは、通常、取引先の名称ではない。したがって、これらは、法人の種別の候補を表す文字列から除外する用語（除外ワード）、又は候補の先頭又は末尾を示す用語として予め設定され得る。ユーザは、予め設定されていない法人の種別を分類定義データ１３５に追加設定することができてもよい。

【００２８】

取引先の法人が営利会社などではないことが多い場合には、図３（ｃ）に示すように、「法人」が含まれる名称を検出対象の法人の種別とする設定がなされてもよい。この場合の法人の種別には、例えば、社団法人、財団法人、ＮＰＯ法人（非営利活動法人）、学校法人、医療法人、独立行政法人、社会福祉法人などが含まれ得る。

【００２９】

更に、図３（ｄ）に示すように、法律又は経理などと関係が強い場合には、例えば、弁護士法人、税理士法人、弁理士法人、司法書士法人、行政書士法人、及び法律事務所、法務事務所、会計事務所、税理士事務所、司法書士事務所、行政書士事務所などを検索可能な正規表現を分類定義データ１３５に設定することができる。また、情報処理装置１は、図３（ｂ）～図３（ｄ）の設定を全て有し、ユーザが必要なもののみが選択的に利用されてもよい。その他、日本国外との取引が多い場合などには、例えば、ＬＬＣ，Ｃｏ．Ｌｔｄ．，Ｉｎｃ．,などが分類定義データ１３５に設定されてもよい。

【００３０】

このような正規表現を用いた文字列の検索では、取引先だけではなく、自社（団体）の法人名などが併せて検出されやすい。登録ツールは、選択から除外する自社、仲介業者や金融機関などを除外対象として、除外対象をまとめた除外リストを保持していてもよい検出された法人のうち、除外リストに含まれる除外対象は、分類種別の候補としての優先順位が下げられる。あるいは、除外対象は、完全に分類種別の候補から除外されてもよい。反対に、一度分類先候補から分類として選択された法人名は、分類リストに登録されて、次回以降に優先的に分類先候補として表示部１４により表示され得る。なお、対象文字列

内に除外対象と優先的な分類先候補とが同時に含まれる場合には、この対象文字列が分類先候補を記載する行のテキストではないと判断されてもよい。この場合には、キーワードの有無にかかわらず、除外対象が含まれる対象文字列全体が分類先候補の検出対象から除外されてもよい。あるいは、対象文字列内で検出された分類先候補の用語のみが出力されてもよいし、上記のように優先度が低下された対象文字列が分類先文字列とされてもよい。

【００３１】

特に長い法人名では、複数行に跨って宛先名が記載される場合があり得る。例えば、上記検索された語のみがある行に記載され残りの固有名称が別の行にある場合には、登録ツールは、検索された語を含む行の前後の行を統合することができる。また、例えば、分類リストに登録済の法人名の一部との合致が検出された場合には、登録ツールは、当該合致部分を含む行を前後の行と統合して再度分類リストなどと比較してもよい。上記のように改行を示す制御コードやタグが予め除去されていない場合には、削除される複数行を統合する場合にこれらが除去されてもよい。

【００３２】

同一の文字列が複数回検出された場合には、当該文字列が一回のみ候補として出力されればよい。検出されたある文字列を内包する文字列が別個に検出された場合には、いずれか一方のみが候補として出力されてもよい。一方が登録リストに含まれている場合には、文字列の長短にかかわらず登録されている文字列が優先的に出力されてもよい。検出された文字列のいずれも登録リストに含まれていない場合には、長い方又は短い方のいずれが優先的に出力されるかが予め設定されていてもよい。

【００３３】

図４は、電子帳簿書類の先頭付近の例を示す図である。

上記のように、スペースを含む「請　求　書」との記載から文書種別の候補として「請求書」が検出される。また、取引先種別として、「株式会社ＡＡＡ御中」及び「ＢＢＢ株式会社」が検出される。このうち、「御中」は、上記除外ワードとして削除され得る。株式会社ＡＡＡ及びＢＢＢ株式会社のうちいずれかが自社である場合には、除外リストに従って自社名が除外されて、他方が分類の候補とされる。反対に、いずれか一方が分類リストに登録済の場合には、登録済の法人名が上位の分類先候補とされる。複数の分類先候補がある場合には、当該複数の分類先候補が並列に表示部１４により表示されて、これらがいずれもユーザにより選択可能とされればよい。分類先候補が１つの場合には、ユーザは、単純に候補を承認することができる。

【００３４】

データベースでは、複数のキーについてそれぞれ分類種別が定められ得る。上記のように、文書種別と取引先種別のいずれについても選択及び登録操作が可能である。選択のための表示及び入力操作は、複数のキーについて並列に行われてもよいし、順番に一つずつ行われてもよい。

【００３５】

図５は、本実施形態の情報処理装置１で監視ツールにより実行される監視制御処理のＣＰＵ１１による制御手順を示すフローチャートである。この監視制御処理は、例えば、情報処理装置１の起動時に自動で起動され、又はユーザの入力操作などにより任意のタイミングで起動され得る。一度起動された監視制御処理は、別途割込み処理などにより終了命令がなされるまで繰り返し継続的に実行される。

【００３６】

ＣＰＵ１１は、監視対象のフォルダのファイルリストを取得する（Ｓ１）。上記のように、監視対象のフォルダは、「ダウンロードフォルダ」であってもよい。ＣＰＵ１１は、ファイルリストを前回の処理Ｓ１で取得したファイルリストと比較する（Ｓ２）。

【００３７】

ＣＰＵ１１は、監視の結果、前回のファイルリストに対して追加されたファイルがあるか否かを判別する（Ｓ３）。なお、ＣＰＵ１１は、追加ファイルだけではなく、同名で更

新されたファイルを併せて検出してもよい。追加ファイルがないと判別された場合には（Ｓ３；ＮＯ）、ＣＰＵ１１の処理は、処理Ｓ１に戻る。

【００３８】

追加ファイルがあると判別された場合には（Ｓ３；ＹＥＳ）、ＣＰＵ１１は、追加ファイルはＰＤＦ形式であるか否かを判別する（Ｓ４）。追加ファイルがＰＤＦ形式ではないと判別された場合には（Ｓ４；ＮＯ）、ＣＰＵ１１の処理は、処理Ｓ１に戻る。追加ファイルがＰＤＦ形式であると判別された場合には（Ｓ４；ＹＥＳ）、ＣＰＵ１１は、登録ツールによる文書分類制御処理を呼び出して実行する（Ｓ５）。それから、ＣＰＵ１１の処理は、処理Ｓ１に戻る。

【００３９】

なお、処理Ｓ１に戻る前に、所定の待機時間が設定されてもよい。あるいは、ＣＰＵ１１は、対象フォルダに対する操作が検出されるまで、処理Ｓ１を実行せずに待機してもよい。

【００４０】

図６は、情報処理装置１で登録ツールにより実行される文書分類制御処理のＣＰＵ１１による制御手順を示すフローチャートである。

【００４１】

ＣＰＵ１１は、対象フォルダの文書データを取得する（Ｓ５１）。ＣＰＵ１１は、文書データからテキストデータを全文抽出する。ＣＰＵ１１は、全文テキストデータにおける改行位置を特定し、上から順に一行分ずつ抜き出した行データを得る（Ｓ５２）。ＣＰＵ１１は、抽出した各行のデータにおけるスペース（インデント、タブ、タグ指定なども含む）を削除して対象文字列を生成する（Ｓ５３）。

【００４２】

ＣＰＵ１１は、分類定義データ１３５を参照して、一行分のテキストデータに含まれるキーワードをそれぞれ検索する（Ｓ５４）。上記のように、ＣＰＵ１１は、キーワードの検索対象とする行を先頭の予め定められた行に限定してもよい。あるいは、ＣＰＵ１１は、キーワードを検索する対象とする行を、各行の文字サイズに基づいて（例えば、最大の文字サイズの行を）選択してもよい。ＣＰＵ１１は、キーワードに対応する分類先候補を設定する（Ｓ５５）。分類先候補は、例えば、単純にキーワードを含む一行のテキストから除外ワードを削除したものであってもよい。また、分類定義データ１３５において、キーワードと異なる分類が設定されている場合には、設定されている分類名が分類先候補とされる。

【００４３】

ＣＰＵ１１は、設定した分類先候補を表示部１４により一覧表示させる（Ｓ５６）。ＣＰＵ１１は、一覧表示の際に、分類先候補が複数ある場合に、自社名のような除外リストに含まれる除外対象を含む分類先候補を除外リストに含まれる除外対象を含まない分類先候補よりも下に表示させるなどして、表示の優先度を下げてもよい。ＣＰＵ１１は、操作受付部１５への入力操作を待ち受け、分類の選択操作を受け付ける。ＣＰＵ１１は、選択操作に応じて分類を確定する（Ｓ５７）。なお、表示された候補内に適切な分類が含まれていない場合や、候補名が不正確な場合などには、ユーザは、適切な分類の名称を操作受付部１５により直接入力することができる。ＣＰＵ１１は、入力された名称を新たな分類として設定し、分類リストに追加登録する。

【００４４】

ＣＰＵ１１は、文書データが分類されてデータベースに登録される対象のファイル、すなわち電子帳簿データであるか否かを判別する（Ｓ５８）。分類、登録対象のファイルではないと判別された場合には（Ｓ５８；ＮＯ）、ＣＰＵ１１は、文書分類制御処理を終了して、処理を監視制御処理に戻す。

【００４５】

文書データが分類、登録対象のファイルであると判別された場合には（Ｓ５８；ＹＥＳ）、ＣＰＵ１１は、文書データに分類情報を付加する（Ｓ５９）。ＣＰＵ１１は、文書デ

ータをデータベースに登録する（Ｓ６０）。ＣＰＵ１１は、登録済の文書ファイルのデータを対象フォルダ（ダウンロードフォルダ）から削除する（Ｓ６１）。ＣＰＵ１１は、文書分類制御処理を終了して、処理を監視制御処理に戻す。

この文書分類制御処理のうち少なくとも処理Ｓ５２、Ｓ５４、Ｓ５５は、本実施形態の情報処理方法を構成し、本実施形態のプログラム１３２における処理手段をなす。

【００４６】

以上のように、本実施形態の情報処理装置１は、ＣＰＵ１１を備える。ＣＰＵ１１は、電子文書から１行ごとに抜き出した行に基づいて対象文字列を生成し、生成された対象文字列と所定の文字列（キーワード）とを比較した比較の結果に基づいて電子文書の分類先候補を出力する処理を行う。ＣＰＵ１１は、抜き出した行に空白が含まれる場合には当該行から空白を削除して対象文字列を生成する。

このように、情報処理装置１は、同種の電子文書が複数、特に多数ある場合に、容易に分類先候補を特定してユーザに示すことができる。特に、分類に使われる文書タイトルなどは、しばしばスペースやタブなどの空白を挟む。これを除外して文字列の検索を行うことで、情報処理装置１は、容易に検出漏れを低減して分類先候補を検出、出力することができる。したがって、ユーザは、容易に電子文書を分類して管理し、以後により容易に当該電子文書を呼び出すことが可能になる。

【００４７】

また、ＣＰＵ１１は、電子文書の先頭の行から予め定められた行数だけ、前記処理を行ってもよい。

分類種別に用いられ得る文書種別を表す文書名や、取引先種別を表す宛先又は書類の作成元などは、文書の先頭付近に記載されていることが多い。したがって、情報処理装置１は、処理を行う行数を先頭から特定の行数に絞ることで、処理が簡易化される。また、他の行から不要な検索結果を得ないので、不要な分類先候補が増えず、ユーザの選択が容易になる。

【００４８】

また、ＣＰＵ１１は、除外対象が前記対象文字列に含まれる場合に、対象文字列の出力の優先度を下げてもよい。自社名など、分類対象ではないが分類先候補として検出されやすい文字列がある。このような文字列を除外対象として予め除外リストなどに登録しておくことで、情報処理装置１は、不要な分類先候補を選択しづらくすることができる。

【００４９】

また、ＣＰＵ１１は、電子文書が予め定められた位置（フォルダなど）に記憶された場合に、当該電子文書に対する処理を行う。すなわち、情報処理装置１は、ダウンロードフォルダなど特定のフォルダに追加された電子文書に対して自動的に分類及びデータベースへの登録に係る処理を開始する。したがって、いちいちユーザが登録処理を起動する必要がなく、ユーザの処理の手間が軽減される。

【００５０】

また、上記特定のフォルダは、電子メールに添付された文書がデフォルトで格納される設定位置であってもよい。これにより、情報処理装置１は、電子メールに添付されて送られた電子文書も容易に分類してデータベースに登録することができる。したがって、ユーザの手間がより軽減される。

【００５１】

また、電子文書は、電子帳簿に係る書類であってもよい。近年、会計処理が電子処理に移行して、電子帳簿に係る処理が増大している。これに伴い、注文書、見積書、請求書などの決まった電子文書が多数電子的にやり取りされる。情報処理装置１によれば、このような電子文書の分類及び管理の手間が大いに低減される。

【００５２】

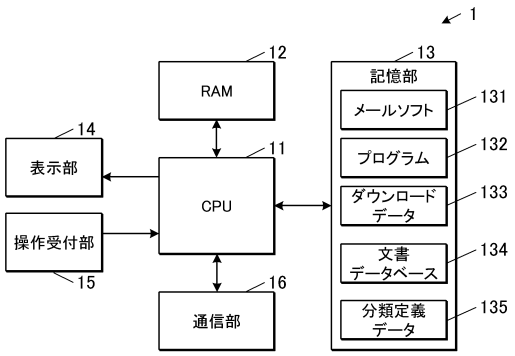
ＣＰＵ１１は、先に電子文書からテキストの内容全文を抜き出した後、全文中の改行の指定位置に基づいて行を各々決定してもよい。文書データによって、改行位置がテキストの逐次抽出では分かりづらい場合もあるので、全文データから改行位置を特定していくこ

とで、改行位置の誤認定などをより確実に避けることができる。
【００５３】
　ＣＰＵ１１は、１行ごとにテキストデータを抜き出して対象文字列を生成するときに、文字サイズが最大である文字が含まれる行を抜き出してもよい。上記のように文書の構造解析を行う場合には、各テキストの文字サイズを特定することができる。文書名などは、タイトルとして最も大きいフォントサイズで記載されていることが多い。したがって、情報処理装置１は、このような行を選択的に抜き出して対象文字列を生成することで、容易に適切な分類先候補を得ることができる。
【００５４】
　また、情報処理装置１は、表示部１４と、操作受付部１５と、を備える。ＣＰＵ１１は、分類先候補を表示部１４により表示させる。ＣＰＵ１１は、操作受付部１５が受け付けた入力操作に応じた分類先に電子文書を分類する。
　したがって、ユーザは容易に分類先を適切に決定することができる。
【００５５】
　また、本実施形態の情報処理方法は、電子文書から１行ごとに抜き出した行に基づいて対象文字列を生成し、生成された対象文字列と所定の文字列（キーワード）とを比較した比較の結果に基づいて電子文書の分類先候補を出力する処理を行う。この処理では、抜き出した行に空白が含まれる場合には当該行から空白を削除して対象文字列を生成する。
　この情報処理方法によれば、同種の電子文書が複数、特に多数ある場合に、ＣＰＵ１１が電子文書から余分な空白を削除して適切に検索の対象文字列を設定し、機械的かつ容易に分類先候補を特定してユーザに示すことができる。したがって、ユーザは、容易に電子文書を分類して管理し、以後により容易に当該電子文書を呼び出すことが可能になる。
【００５６】
　また、本実施形態のプログラム１３２をコンピュータにインストールして実行可能とすることで、ユーザは容易かつ、より正確に多くの同種の電子文書を仕分けして管理することができる。よって、ユーザの手間が大いに低減される。
【００５７】
　なお、本発明は、上記実施の形態に限られるものではなく、様々な変更が可能である。
　例えば、上記実施の形態では、同一分類に含まれる複数の文字列がカンマで区切られて登録されていたが、これに限られない。例えば、スペース又はタブなどにより区切られてもよい。あるいは、検出対象の文字列が全て別個に登録されてもよい。この場合に、文字列と分類とが異なる場合には、当該文字列と分類とが対応付けられて記憶されてもよい。
【００５８】
　また、上記実施の形態では、電子文書の分類項目（キー）として文書種別及び取引先種別を考慮したが、分類項目は、これらに限られない。例えば、取引日時、取引金額、商品（サービス）などが分類項目とされてもよい。
【００５９】
　また、上記では、正規表現を用いて検出する文字列を表したが、正規表現を用いなくてもよい。検出対象の全パターンが網羅されてもよい。また、上記のように、正規表現は、取引先の偏りなどに応じて「会社」、「法人」、「事務所」及び英語表現などのうち一部が選択可能であってもよい。あるいは、初めから全ての正規表現に基づく文字列が検索されてもよい。ただし、選択対象の候補の数が多くなると、自身で直接入力する手間に比して、候補から選択する手間が大きくなり得る。したがって、あまり余計な候補が多く選択されないように正規表現が選択されるのが好ましい。
【００６０】
　また、上記では、電子メールの添付ファイル及びネットワークを介したダウンロードデータを例に挙げて説明したが、これらに限られない。例えば、電子文書ファイルは、ＵＳＢメモリなどの可搬型記録媒体などにより取得されてもよい。また、外部から取得した書類に加えて又は代えて、自身で作成して外部へ送付する電子文書ファイルも分類の対象とされ得る。

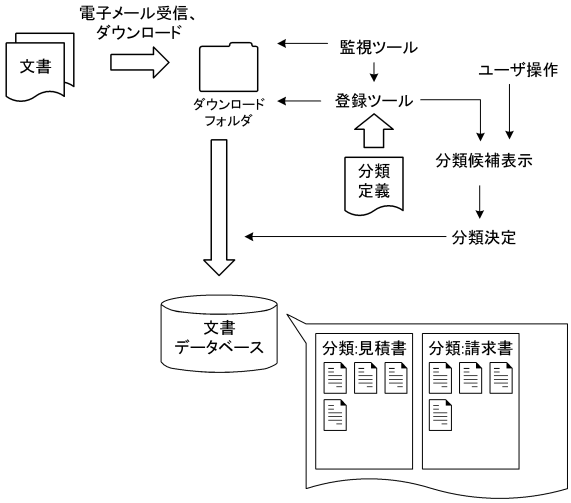
【００６１】
　また、上記では、電子帳簿に係る電子文書データが分類対象とされたが、これに限られない。定型的であって、文書に含まれるテキストの内容から分類が可能なものであれば、分類の対象とされてよい。また、電子文書がＰＤＦであるものとして説明されたが、電子文書はこれに限られない。定型的な取引文書などとして用いられるフォーマットのものであれば、分類対象は、他の形式の電子文書であってもよい。また、電子文書から各行のテキストを抽出する処理は、全文抽出後に各文に分割されるものに限られない。逐次改行が検出されて、１行ずつ行のテキストが抽出されてもよい。
【００６２】
　また、上記では、ＰＣなどの情報処理装置１が単独で文書の分類及び格納を行ったが、これに限られない。情報処理装置１は、分類に係る動作を他の装置に要求して、分類結果のみを取得してもよい。あるいは、情報処理装置１は、分類情報を含む電子文書データを外部のデータベースサーバなどに送信して、当該データベースサーバにより電子文書データを記憶させてもよい。また、データベース装置は、外付けの補助記憶装置、ネットワーク上の記憶装置、あるいはクラウドサーバなどであってもよい。
【００６３】
　また、以上の説明では、本発明の文書分類制御に係るプログラム１３２を記憶するコンピュータ読み取り可能な媒体としてＨＤＤ、フラッシュメモリなどの不揮発性メモリなどからなる記憶部１３を例に挙げて説明したが、これらに限定されない。その他のコンピュータ読み取り可能な媒体として、ＭＲＡＭなどの他の不揮発性メモリや、ＣＤ－ＲＯＭ、ＤＶＤディスクなどの可搬型記録媒体を適用することが可能である。また、本発明に係るプログラムのデータを、通信回線を介して提供する媒体として、キャリアウェーブ（搬送波）も本発明に適用される。
　その他、上記実施の形態で示した具体的な構成、処理動作の内容及び手順などは、本発明の趣旨を逸脱しない範囲において適宜変更可能である。本発明の範囲は、特許請求の範囲に記載した発明の範囲とその均等の範囲を含む。
【符号の説明】
【００６４】
１　　情報処理装置
１１　ＣＰＵ
１２　ＲＡＭ
１３　記憶部
１３１ メールソフト
１３２ プログラム
１３３ ダウンロードデータ
１３４ 文書データベース
１３５ 分類定義データ
１４　表示部
１５　操作受付部
１６　通信部

【図面】

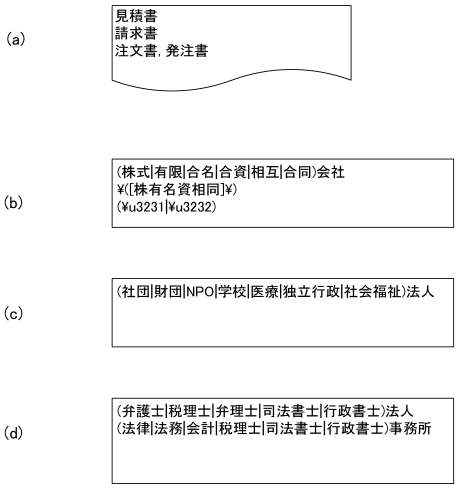
【図１】


1
12
RAM
13
記憶部
メールソフト
131
プログラム
132
ダウンロードデータ
133
文書データベース
134
分類定義データ
135
14
表示部
11
CPU
操作受付部
15
16
通信部

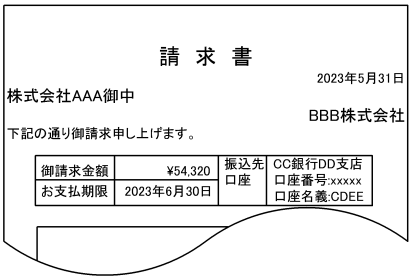
【図２】


電子メール受信、ダウンロード
文書
ダウンロードフォルダ
監視ツール
登録ツール
ユーザ操作
分類定義
分類候補表示
分類決定
文書データベース
分類:見積書
分類:請求書

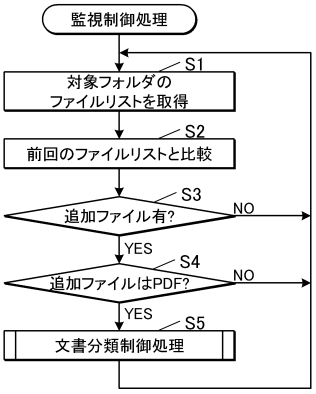
【図３】


(a)
見積書
請求書
注文書, 発注書
(b)
(株式|有限|合名|合資|相互|合同)会社
¥([株有名資相同]¥)
(¥u3231|¥u3232)
(c)
(社団|財団|NPO|学校|医療|独立行政|社会福祉)法人
(d)
(弁護士|税理士|弁理士|司法書士|行政書士)法人
(法律|法務|会計|税理士|司法書士|行政書士)事務所

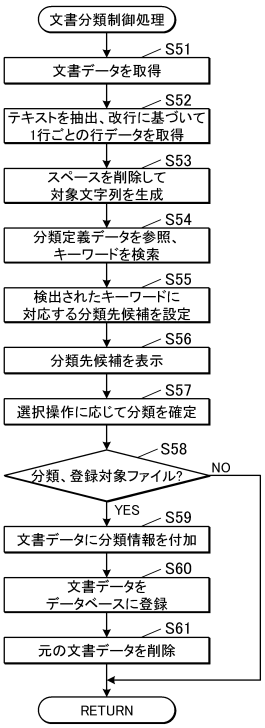
【図４】


請 求 書
2023年5月31日
株式会社AAA御中
BBB株式会社
下記の通り御請求申し上げます。
御請求金額
¥54,320
お支払期限
2023年6月30日
振込先口座
CC銀行DD支店
口座番号:xxxxx
口座名義:CDEE

【図５】


監視制御処理
S1
対象フォルダの
ファイルリストを取得
S2
前回のファイルリストと比較
S3
追加ファイル有?
NO
YES
S4
追加ファイルはPDF?
NO
YES
S5
文書分類制御処理

【図６】


文書分類制御処理
S51
文書データを取得
S52
テキストを抽出、改行に基づいて
1行ごとの行データを取得
S53
スペースを削除して
対象文字列を生成
S54
分類定義データを参照、
キーワードを検索
S55
検出されたキーワードに
対応する分類先候補を設定
S56
分類先候補を表示
S57
選択操作に応じて分類を確定
S58
分類、登録対象ファイル?
NO
YES
S59
文書データに分類情報を付加
S60
文書データを
データベースに登録
S61
元の文書データを削除
RETURN