(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5173721号
(P5173721)
(45)発行日　平成25年4月3日(2013.4.3)　(24)登録日　平成25年1月11日(2013.1.11)

(51)Int.Cl.　F I
G06T 1/00 (2006.01)　G06T 1/00 200C
G06F 17/30 (2006.01)　G06F 17/30 220C
G06F 17/30 350C

請求項の数 8　(全 48 頁)

(21)出願番号　特願2008-256639(P2008-256639)
(22)出願日　平成20年10月1日(2008.10.1)
(65)公開番号　特開2010-86412(P2010-86412A)
(43)公開日　平成22年4月15日(2010.4.15)
審査請求日　平成23年9月27日(2011.9.27)

(73)特許権者　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　100076428
弁理士　大塚　康徳
(74)代理人　100112508
弁理士　高柳　司郎
(74)代理人　100115071
弁理士　大塚　康弘
(74)代理人　100116894
弁理士　木村　秀二
(74)代理人　100130409
弁理士　下山　治
(74)代理人　100134175
弁理士　永川　行光

最終頁に続く

(54)【発明の名称】文書処理システム及びその制御方法、プログラム、記憶媒体

(57)【特許請求の範囲】
【請求項１】
複数の文書データと、各文書データのメタデータと、各文書データの関連を示す関連情報とを格納する格納手段と、
原稿上の画像を読み取って当該画像に基づく文書データを入力する入力手段と、
前記入力手段により入力された文書データの特徴を抽出する特徴抽出手段と、
前記特徴抽出手段により抽出された前記文書データの特徴と、前記格納手段に格納された文書データのメタデータとに基づいて、前記入力手段により入力された文書データに関連する関連文書データを、前記格納手段に格納されている複数の文書データから特定する特定手段と、
前記特定手段により特定された関連文書データのメタデータを、前記入力手段により入力された文書データと前記関連文書データとの関連情報に基づく前記関連文書データのメタデータの確実性を示す確信度とともに、前記入力された文書データのメタデータとして前記格納手段に格納する制御手段と、
を有することを特徴とする文書処理システム。
【請求項２】
前記特定手段は更に、前記入力手段により入力された文書データと前記格納手段に格納されている各文書データに含まれる画像との類似度を判定し、当該判定した類似度に基づいて、前記関連文書データを特定することを特徴とする請求項１に記載の文書処理システム。

【請求項３】
前記メタデータの検索キーに基づいて前記格納手段に格納されている文書データを検索する検索手段と、
前記検索手段により検索された文書データが複数存在する場合、前記メタデータの確信度に基づいて前記検索された文書データを提示する順番を決定して提示する提示手段と、を更に備えることを特徴とする請求項１又は２に記載の文書処理システム。
【請求項４】
前記関連情報に基づいて、前記格納手段に格納されている複数の文書データの文書ランクを決定する決定手段を更に備えることを特徴とする請求項１乃至３のいずれか１項に記載の文書処理システム。
【請求項５】
前記入力手段により文書データが入力されたことに応じて、前記特定手段により特定された当該文書データに関連する関連文書データの存在を通知する通知手段を更に備えることを特徴とする請求項１乃至４のいずれか１項に記載の文書処理システム。
【請求項６】
複数の文書データと、各文書データのメタデータと、各文書データの関連を示す関連情報とを格納する格納手段を備えた文書処理システムの制御方法であって、
原稿上の画像を読み取って当該画像に基づく文書データを入力する入力工程と、
前記入力工程で入力された文書データの特徴を抽出する特徴抽出工程と、
前記特徴抽出工程で抽出された前記文書データの特徴と、前記格納手段に格納された文書データのメタデータとに基づいて、前記入力工程で入力された文書データに関連する関連文書データを、前記格納手段に格納されている複数の文書データから特定する特定工程と、
前記特定工程で特定された関連文書データのメタデータを、前記入力工程で入力された文書データと前記関連文書データとの関連情報に基づく前記関連文書データのメタデータの確実性を示す確信度とともに、前記入力された文書データのメタデータとして前記格納手段に格納する制御工程と、
を備えることを特徴とする文書処理システムの制御方法。
【請求項７】
請求項６に記載の文書処理システムの制御方法をコンピュータに実行させるためのプログラム。
【請求項８】
請求項６に記載の文書処理システムの制御方法をコンピュータに実行させるためのプログラムを記憶したコンピュータが読み取り可能な記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、複数の文書データと、該複数の文書データ間の関連を示す関連情報とを格納する文書処理システム及びその制御方法、プログラム、記憶媒体に関するものである。
【背景技術】
【０００２】
ストレージ技術の進化及び低コスト化が進んで、従来では考えられない程の大量の文書データを蓄積管理することが可能となり、このような機能を実現するファイルサーバ、文書管理システム及びグループウェア等が普及している。またＰＣ等の情報処理装置が進化する一方で、コピー機、プリンタ、イメージスキャナ、ファクス、デジタルカメラ、文書ストレージや画像の送受信機能を備えた複合機（ＭＦＰ）等の各種の機器がネットワークと通信可能に構成されている。また顧客のネットワーク環境では、情報処理装置や各種事務機器との間で大量の文書データ等の交換が行われており、オフィスのネットワークを流通する文書トラフィックを積極的にストアするストレージ・インフラストラクチャが実用化されつつある。

【０００３】
ストレージ・インフラストラクチャの一例として、特許文献１には、操作者の手を煩わすことなく、確実に必要な画像の控えを残す複合画像処理装置を提供するために、少なくとも２つの画像出力装置が接続できる複合画像処理装置が開示されている。この装置は、画像処理ジョブの処理パラメータを監視し、起動されたジョブが所定の条件を満たしているかどうかを判定している。そして、その条件を満足していると判定したジョブの実行に際して、本来の画像データの出力先に加えて、更にもう１つの画像出力装置（画像ファイルなど）にも画像データを送ることが記載されている。このストレージ・インフラストラクチャは、機密漏えいの抑止などセキュリティを目的とする監査のため、或いは以前に作成した文書や以前に実施した処理に類似した無駄な処理をできるだけ省いて、既存の資産をうまく再利用するため等の理由が挙げられる。
【０００４】
このようなオフィスのネットワークを流通する文書トラフィックを積極的に格納するストレージインフラストラクチャでは、文書内容データをストアするだけでなく、その文書に関連する各種の付加情報、即ちメタデータも格納する。例えば、文書と他の文書の関連情報や、文書のライフサイクルに関連した履歴情報が、メタデータとしてその文書と関連付けて格納される。関連文書としては、例えば同一カテゴリに属する文書のグルーピング、旧版と改訂版、アプリケーションデータと印刷時に収集されたスナップショット文書、類似文書、同一ページ含む文書、類似画像を含む文書等がある。また文書のライフサイクルに関連するメタデータには、例えば文書に対して施された処理の内容、パラメータ、時刻、用いた装置、場所、及び処理の操作者の情報などが含まれる。
【０００５】
特許文献２は、文書を扱う装置（プリンタ、スキャナ、コピー機、ＦＡＸ、プロジェクタ、デジタルカメラ等）に文書管理機能の一部を実装したファイリングシステムを開示している。これによれば、文書を扱うごとに、その文書情報と、その文書を扱った関係者に関する付加情報とを文書管理サーバへ送信している。
【０００６】
電子的な文書データファイルの分野では、文書内容データに付随するメタデータを文書データに関連付けて表現するファイル形式が使われている。OpenDocument Format (ISO/IEC 26300) やOffice Open XML (Ecma-376) では、文書ファイル形式の中にＸＭＬ文書によるメタデータの表現を含んでいる。
【０００７】
特許文献３は、コンピュータ等のデジタルの世界と紙の文書の間に情報の連続性や関連性を構築した書情報管理システムを開示している。これによれば、紙の文書をデジタルの世界の文書情報管理システム内に組み込むとともに、紙の文書を媒体としてデジタルの世界に直接アクセス可能とし、更に紙の文書を用いたハイパーテキストを実現している。このシステムでは、媒体用紙上の任意の位置に記録した記載情報に選択情報を付与することにより、所望の関連情報ファイル（電子化した文書）を検索して出力している。関連情報ファイルを検索するための連結情報も媒体用紙に記録されている。
【０００８】
また特許文献５及び非特許文献１は、ＰａｇｅＲａｎｋ（登録商標）としてよく知られている技術のアイデアを開示している。これによれば、Ｗｅｂの膨大なリンク構造を用いて、ページから別のページへのリンクを支持投票とみなし、その投票数によりそのページの重要性を判断している。この際、単に票数、つまりリンク数を見るだけではなく、票を投じたページについても分析し、「重要度」の高いページによって投じられた票は、より高く評価されて、それを受け取ったページを「重要なもの」にしている。
【特許文献１】特許３４８６４５２号公報
【特許文献３】特開２００４－７８７３５号公報
【特許文献４】特開平０９－９１３０１号公報
【特許文献５】米国特許第６，２８５，９９９号公報

【非特許文献１】Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, 'The PageRank Citation Ranking: Bringing Order to the Web', 1998, http://www-db.stanford.edu/~backrub/pageranksub.ps

【発明の開示】

【発明が解決しようとする課題】

【０００９】

オフィス等における最重要資源の一つである文書を格納する量は、今後、益々多くなり膨大な量になると予想される。また文書の生成と処理はオフィスの基本活動であり、その蓄積容量は増え続け、膨大でダイナミックに文書が蓄積される空間をカテゴリ等の木構造の分類で整理することは難しい。従って、膨大で未整理の文書ストレージから効率良く検索する手法を充実させる必要がある。この検索には、インターネットにおける検索サービスだけでなく、エンタープライズサーチと呼ばれる企業ネットワーク内での全文検索やコンテンツ検索の活用が普及しつつある。

【００１０】

ストアされた膨大な文書の中から所望の文書を効率的に検索するためには、文書データだけでなく、その文書に付随する各種メタデータや、他の文書との関連を活用することが重要となる。例えば、ユーザが文書に対して行った処理のような、ユーザのオフィスにおけるアクティビティを反映するメタデータをキーとして検索できるようになれば、より高度で、実用的な検索機能が提供できる。

【００１１】

また複数の文書とメタデータをノードとし、文書間、メタデータ間の関連から構成される意味的なネットワークを一種の知識表現として活用することにより、種々の応用の可能性が広がる。文書とメタデータのネットワークを分類、分析、加工することによって、いわゆるデータマイニングやビジネスインテリジェンスに用いることができる。また、このネットワークは、文書や文書に関連したオフィスワーカの行動を表現しているので、統計処理等による統合を施すことにより、いわゆる「群集の叡智」或は「集合知」を引き出して活用できる。尚、「群集の叡智」は、インターネットにおいて「Web 2.0」の潮流を特徴付ける一つの要素として注目を集めている。今後はイントラネットにおいても活用することで、オフィス全体の生産性を著しく高めることが期待できる。

【００１２】

ところが、この意味的なネットワークと電子的にリンクされているオンライン文書やメタデータを内包するファイル形式の電子文書は、一度紙媒体へ印刷、又はファクス送信されると、そのメタデータや他の文書との関連データを失ってしまう。即ち紙媒体文書やファクス文書のような、ネットワーク的にオフラインとなる文書は、メタデータや意味的な関連のネットワークから切り離されたものとなる。

【００１３】

前述の特許文献４の先行技術では、紙媒体上にデジタルの世界の関連情報ファイルを検索するための連結情報を記録している。しかし、紙のスキャンやファクス受信といった処理に際して、そのオフラインの文書及びその処理に関するメタデータを、オンラインの意味的なネットワーク中に再結合できない。即ち、ストレージ・インフラストラクチャ内に存在するオンライン文書と、それらの処理の対象となった紙媒体の文書（及びその処理に関するメタデータ）を関連付けて蓄積、管理することができなかった。

【００１４】

本発明の目的は、このような従来の問題点を解決することにある。

【００１５】

本発明の特徴は、原稿上の画像を読み取って入力された文書データを、格納手段に格納された複数の文書データのうち該入力された文書データに関連する文書データと関連付ける技術を提供することにある。

【課題を解決するための手段】

【００１６】

上記目的を達成するために本発明の一態様に係る文書処理システムは、　複数の文書データと、各文書データのメタデータと、各文書データの関連を示す関連情報とを格納する格納手段と、

原稿上の画像を読み取って当該画像に基づく文書データを入力する入力手段と、

前記入力手段により入力された文書データの特徴を抽出する特徴抽出手段と、

前記特徴抽出手段により抽出された前記文書データの特徴と、前記格納手段に格納された文書データのメタデータとに基づいて、前記入力手段により入力された文書データに関連する関連文書データを、前記格納手段に格納されている複数の文書データから特定する特定手段と、

前記特定手段により特定された関連文書データのメタデータを、前記入力手段により入力された文書データと前記関連文書データとの関連情報に基づく前記関連文書データのメタデータの確実性を示す確信度とともに、前記入力された文書データのメタデータとして前記格納手段に格納する制御手段と、を有することを特徴とする。

【発明の効果】

【００１７】

本発明によれば、原稿上の画像を読み取って入力された文書データを、格納手段に格納された複数の文書データのうち該入力された文書データに関連する文書データと関連付けることができる。

【発明を実施するための最良の形態】

【００１８】

以下、添付図面を参照して本発明の好適な実施形態を詳しく説明する。尚、以下の実施形態は特許請求の範囲に係る本発明を限定するものでなく、また本実施形態で説明されている特徴の組み合わせの全てが本発明の解決手段に必須のものとは限らない。

【００１９】

図１は、本発明の一実施形態に係る文書処理システムの全体構成を示すブロック図である。

【００２０】

この文書処理システムは、互いにネットワークを介して接続された画像処理装置１１０，１２０，１３０とパーソナルコンピュータ（情報処理装置）１０１，１０２とサーバシステム１４０とを有している。ネットワークは、例えばＬＡＮ（Local Area Network）１００で構成される。尚、これら画像処理装置１１０，１２０，１３０は、文書処理装置としても適用可能であるが、以下では画像処理装置と呼ぶこととする。

【００２１】

画像処理装置１１０は、画像入力デバイスであるスキャナ１１３、画像出力デバイスであるプリンタ１１４、コントローラ１１１、ユーザインタフェースである操作部１１２を備えている。スキャナ１１３、プリンタ１１４、操作部１１２はそれぞれ、コントローラ１１１に接続されて、コントローラ１１１からの命令によって制御される。またコントローラ１１１は、ＬＡＮ１００に接続されている。画像処理装置１２０，１３０は画像処理装置１１０と同様の構成であるため、その説明を省略する。

【００２２】

パーソナルコンピュータ１０１，１０２は、複数のユーザのそれぞれが主に個人的に使用する情報処理装置であり、ユーザが利用するアプリケーションプログラムやユーザのデータ等を格納している。サーバシステム１４０は、サーバコンピュータ１４１と大規模ストレージ装置１４２を具備している。サーバコンピュータ１４１は、複数のユーザやクライアントシステムに対してサービスを提供するサーバアプリケーションや共有データ等を格納している。また大規模ストレージ装置１４２は、高性能で信頼性が高い大規模な二次記憶装置であり、主にサーバコンピュータ１４１上で稼動するデータベース管理システム（ＤＢＭＳ）のデータ等を格納している。また、以下の説明では、画像処理装置１１０、パーソナルコンピュータ１０１を参照して、このシステムの動作を説明するが、他の画像処理装置やパーソナルコンピュータでも同様の処理が実施できることは言うまでもない。

【００２３】
このサーバシステム１４０によってサービスされるサーバアプリケーションの一つは、ネットワーク全域に亘って流通するジョブ文書をアーカイブ（即ち、蓄積管理）するデータベース（ＤＢ）アプリケーションである。これを以下、ジョブアーカイブ・アプリケーションと呼ぶ。このジョブアーカイブ・アプリケーションは、ネットワーク１００に接続された他の装置群にそれぞれ組み込まれたソフトウェアと連携して、ジョブアーカイブシステムと呼ばれる分散アプリケーションを構成する。
【００２４】
パーソナルコンピュータ１０１は、画像処理装置１１０，１２０，１３０やサーバシステム１４０等とＬＡＮ１００を介して連携する。例えばパーソナルコンピュータ１０１は、画像処理装置１１０に対して文書データを送信、又は画像処理装置１１０から文書データ受信して、印刷、スキャン、ファクス送受信を行う。またボックス（画像処理装置１１０に組み込みの文書管理システム）へ文書データを蓄積したり、そこから取り出す等のジョブを実行する。このネットワーク上で文書データを処理するジョブを実行するとき、サーバシステム１４０で稼動するジョブアーカイブ・アプリケーションが、ジョブ情報とジョブの処理対象文書データの控えをアーカイブする。例えば、印刷ジョブの場合、パーソナルコンピュータ１０１のプリンタドライバが画像処理装置１１０へジョブを投入するとともに、サーバシステム１４０へもそのジョブに関連する情報と処理対象文書のデータを送信することでアーカイブが達成される。
【００２５】
また画像処理装置１１０は、他の画像処理装置１２０，１３０やパーソナルコンピュータ１０１，１０２や、サーバシステム１４０等とＬＡＮ１００を介して連携する。例えば、画像処理装置１１０は、原稿の画像をスキャンしてデジタルデータ化して他の装置へ送信したり、他の装置が保有しているデータをリトリーブして印刷、或はボックスへ蓄積したり、更に他の装置へ転送したりするジョブを実行する。これら文書データを処理するジョブを実行する際にも、サーバシステム１４０上で稼動するジョブアーカイブ・アプリケーションが、ジョブ情報とジョブの処理対象文書データの控えをアーカイブする。例えば、プッシュスキャンジョブの場合、画像処理装置１１０の「送信」アプリケーションが、原稿文書をスキャナ１１３で読み取ったデジタル文書データを本来の送信宛先に送信する。これと同時に、サーバシステム１４０へ、そのジョブに関連する情報と処理対象文書のデータを送信することによりアーカイブが達成される。
【００２６】
このようにして、ネットワーク全域に亘って流通する文書データは、ジョブアーカイブ・アプリケーションによりアーカイブされている。
【００２７】
図２は、本実施形態に係るサーバシステム１４０で稼動するジョブアーカイブ・アプリケーションのソフトウェア構成を示すブロック図である。
【００２８】
ＤＢ管理システム２０１はデータベース管理システムであり、大量のレコードを含む大容量のデータを、レコード間の関連とともに構造化したデータベースとして格納する。このＤＢ管理システム２０１のデータは、上述したように大規模ストレージ装置１４２に格納されている。また、ＤＢ管理システム２０１は、ＳＱＬ等の問い合わせ言語による問い合わせに応じて、条件に合致するレコードをデータベースから高速にリトリーブする。ＤＢ管理システム２０１は、文書ＤＢ２０２、ジョブＤＢ２０３、インデクスＤＢ２０４を含み、このＤＢ管理システム２０１は、よく知られたリレーショナルデータベースやオブジェクト指向データベース等の実装によって実現できる。
【００２９】
文書ＤＢ２０２は、ジョブアーカイブシステムが蓄積管理する文書データを格納するデータベースである。文書内容データと、その文書に関連するメタデータとを文書レコードとして格納している。文書ＤＢ２０２とジョブＤＢ２０３とは、格納されるレコード間で

相互に関連している。ジョブＤＢ２０３は、ジョブアーカイブシステムが蓄積管理するジョブデータをジョブレコードとして格納するデータベースである。ジョブＤＢ２０３と文書ＤＢ２０２とは、格納されるレコード間で相互に関連している。インデクスＤＢ２０４は、ジョブアーカイブシステムが蓄積管理する文書データやジョブデータから、所望のデータを高速に検索するためのインデクスレコードを格納するデータベースである。インデクスＤＢ２０４に格納されるインデクスレコードは、文書ＤＢ２０２及びジョブＤＢ２０３内のレコードを参照している。

【００３０】

ストア部２０５は、画像処理装置１１０やパーソナルコンピュータ１０１等のクライアント装置から文書データ及びジョブデータを受信して、ＤＢ管理システム２０１に格納する格納要求受け付けモジュールである。このストア部２０５は、受信した文書データとジョブデータをＤＢ管理システム２０１に格納する。またストア部２０５は、受信した文書データのデータ形式に応じてメタデータを生成するための処理を切り替える。即ち、受信した文書データが、スキャナで読み取った、或はデジタルカメラで撮影した、或はファクスで受信した画像データである場合、その画像データをラスタ画像ページ処理部２０６に送る。一方、受信した文書データがコード化された文書データの場合、即ち、ページ記述言語やベクタ表現された各種文書フォーマットや、ＤＴＰやワードプロセッサや表計算等の各種アプリケーションの文書フォーマットの場合は展開部２１０に送る。展開部２１０は、そのコード文書データをラスタ画像データに展開してラスタ画像ページ処理部２０６に出力する。

【００３１】

ラスタ画像ページ処理部２０６は、ラスタ画像データから、そのデータを構成するページを切り分けて、各ページごとに処理するモジュールである。ラスタ画像ページ処理部２０６は、その切り分けた各ページ画像を画像特徴抽出部２０７及び画像構造解析部２０８に送る。ここで、ラスタ画像とは、スキャナで読み取った、或いはファクシミリ受信したような画像データを言う。従ってラスタ画像とは、画像中の各文字等がコード化されていないデータである。一方、ラスタ画像ではない文書データとは、これとは逆に、そのデータに含まれる各文字や記号などがコード化されており、その文書のレイアウトや内容等の編集・変更可能なデータである。

【００３２】

画像特徴抽出部２０７は、１ページのラスタ画像データを解析して画像間の類似性判定の基準として用いる特徴を抽出するモジュールである。ここで抽出された特徴は、ＤＢ管理システム２０１に送られて、そこに格納される。類似画像検索に有効な特徴抽出の手法は数多く知られているが、本実施形態では、特定のアルゴリズムには依存せず有効な手法を複数併用する。ここで採用可能な手法には、例えば以下のものを含む。画像中のエッジなどからオブジェクトを抽出して形状を判定し、その形状やその配置や配色や複数のオブジェクト間の位置関係等を用いるもの、また画像全体を構成する支配的な色の組み合わせや配色パターンをヒストグラムなどで抽出して用いるものがある。更には、認知的な類似性判定に近い特性を持つ特徴量を導き出す各種の数学処理（例えばフーリエ・メリン変換）を用いるものがある。

【００３３】

画像構造解析部２０８は、１ページのラスタ画像データからその構造を解析するモジュールである。ここではブロックセレクション或は像域分離等の手法を用いて、ひとかたまりの画像領域（ページ）から、それを構成する特性の異なる複数の領域（文字領域、画像領域、写真領域、グラフィクス領域、白黒領域、カラー領域など）に分解する。そして、各領域の領域構造に関する解析と分類を行う。また背景等の下地パターンとその上に配置された文字や形状等のオブジェクトとの、レイヤ構造に関する解析と分類も行う。この解析の結果得られた画像領域（或は画像レイヤ）のラスタ画像データは、画像特徴抽出部２０７に送られる。またこの解析の結果得られたテキスト領域（又はテキストレイヤ）のラスタ画像データは、ＯＣＲ２０９に送られる。また解析の結果得られた構造情報は、ＤＢ

管理システム２０１に送られて、そこに格納される。ＯＣＲ２０９は、文字が描画されたラスタ画像データを入力し、それを解析して文字認識するモジュールである。文字認識したテキストデータ（即ち、Ｕｎｉｃｏｄｅ等によってコード化されたデータ）をＤＢ管理システム２０１に送って格納する。

【００３４】

インデクス生成部２１１は、文書ＤＢ２０２やジョブＤＢ２０３から高速にデータを検索するためのインデクス情報を生成するモジュールである。インデクス情報は、検索キーとして与えられる画像に類似した画像を含む文書レコードを高速に検索したり、検索キーとして与えられるテキストを文書内容データやページ内容データの中に含む文書レコードを高速に全文検索するのに使用される。また、検索キーとして与えられる条件に合致するメタデータを持つ文書レコードやジョブレコードを高速に検索するのに使用される。このインデクス情報の生成もまた、周知の複数の手法を併用できる。全文検索のためのインデクス情報の生成には、例えばＮ－グラム（N-gram）の手法を用いる。また類似画像検索のためのインデクス情報の生成には、画像の特徴を表現する特徴ベクトルを予め分類（クラスタリング）したりハッシュ関数等によって順序付けたりしておく。このインデクス生成部２１１によるインデクス情報の生成は、文書データやジョブデータの追加登録や編集等によって文書ＤＢ２０２やジョブＤＢ２０３が更新されたときに行われる。また、各ＤＢの更新とは非同期に、バッチ処理として生成することもできる。その生成したインデクス情報は、ＤＢ管理システム２０１のインデクスＤＢ２０４に格納される。

【００３５】

リトリーブ部２１２は、画像処理装置１１０やパーソナルコンピュータ１０１等のクライアント装置から検索キー画像又は検索キーテキストとその検索要求を受け付けて、これに応じてＤＢ管理システム２０１から文書データを検索するモジュールである。そして、ヒットした文書データや、その文書に関連するサムネール画像やジョブデータ等のメタデータをクライアント装置に返信する。文書検索部２１３は、文書検索要求に合致する文書を検索するモジュールである。リトリーブ部２１２からの検索要求と与えられた検索キーの型に応じて、文書内容データに基づく検索、文書に含まれるページデータに基づく検索、文書のメタデータに基づく検索、文書に関連するジョブに基づく検索を組み合わせて文書を検索する。そして、その検索要求に合致する文書レコードの候補を複数探し出す。ページ検索部２１４は、文書データに含まれるページデータに基づく検索の要求に応じて、文書ＤＢ２０２から、検索要求の条件に合致するページレコードの候補（及びそのページを含む文書）を複数探し出す。類似画像検索部２１５は、検索キーとして与えられた画像に基づく類似画像検索の要求に応じて、検索キーである画像に類似する画像を含むページ内容データを持つページレコード（及びそのページを含む文書）を複数探し出す。尚、この類似画像検索は、画像特徴抽出部２０７と同様の画像特徴抽出を検索キーである画像に対して行い、画像の特徴の類似性を基に類似画像を検索する。この実施形態では、周知である、画像を検索キーとして類似画像を検索する類似画像検索の手法を組み合わせて適用する。これには、画像のエッジ等からオブジェクトを抽出して形状を判定し、その形状や配置や配色や複数のオブジェクト間の位置関係等を用いるもの、また画像全体を構成する支配的な色の組み合わせや配色パターンをヒストグラム等で抽出して用いるもの等がある。

【００３６】

ＤＢ操作部２１６は、サーバコンピュータ１４１の管理コンソール又は画像処理装置１１０やパーソナルコンピュータ１０１等のクライアント装置から、ＤＢ管理システム２０１に対する操作要求を受け付けて処理するデータベース操作モジュールである。尚、データベースのレコードに対する操作は、例えば、メタデータ（タグなど）の追加や編集といった操作を含む。

【００３７】

図３は、本実施形態に係る画像処理装置のハードウェア構成を示すブロック図である。尚、画像処理装置１１０，１２０，１３０は同じ構成であるため、ここでは画像処理装置

１１０を例にして説明する。

【００３８】

コントローラ１１１は、スキャナ１１３やプリンタ１１４と接続され、一方ではＬＡＮ１００や公衆回線（ＷＡＮ）と接続することで、画像情報やデバイス情報の入出力を行なっている。ＣＰＵ３０１は、コントローラ１１１全体を制御するコントローラである。ＲＡＭ３０２は、ＣＰＵ３０１が動作するために使用するシステムワークエリアを提供している。またＲＡＭ３０２は、画像データを一時記憶するための画像メモリとしても使用される。ＲＯＭ３０３はブートＲＯＭであり、システムのブートプログラムが格納されている。ＨＤＤ３０４はハードディスクドライブで、システムソフトウェア、画像データ等を格納する。操作部Ｉ／Ｆ３０６は、操作部（ＵＩ）１１２との間のインタフェースを司り、操作部１１２に表示すべき画像データを操作部１１２に対して出力する。また使用者が操作部１１２を介して入力した情報を、ＣＰＵ３０１に伝える役割を果たす。ネットワークインタフェース（Network）３０８はＬＡＮ１００との接続を司り、ＬＡＮ１００に対して情報の入出力を行なう。モデム（ＭＯＤＥＭ）３０９は公衆回線との接続を司り、公衆回線に対して情報の入出力を行なう。以上のデバイスがシステムバス３０７上に配置される。

【００３９】

イメージバスインターフェース（Image Bus I/F）３０５は、システムバス３０７と画像データを高速で転送する画像バス３１０とを接続し、データ構造を変換するバスブリッジである。画像バス３１０は、ＰＣＩバス又はＩＥＥＥ１３９４で構成される。この画像バス３１０には以下のデバイスが配置される。ラスタイメージプロセッサ（ＲＩＰ）３１１は、ネットワーク１００から送信されたＰＤＬコードをビットマップイメージに展開する。デバイスＩ／Ｆ部３１２は、スキャナ１１３やプリンタ１１４とコントローラ１１１とを接続し、画像データの同期系／非同期系の変換を行なう。スキャナ画像処理部３１３は、スキャナ１１３で入力した画像データに対して補正、加工、編集を行なう。プリンタ画像処理部３１４は、プリンタ１１４に出力する画像データに対して、プリンタ１１４の性能に応じた補正、解像度変換等を行なう。画像回転部３１５は画像データの回転を行なう。画像圧縮部３１６は、多値画像データに対してはＪＰＥＧ圧縮伸長処理を行い、２値画像データに対してはＪＢＩＧ，ＭＭＲ，ＭＨの圧縮伸長処理を行なう。

【００４０】

図４は、本実施形態に係る画像処理装置１１０の外観を示す斜視図である。尚、画像処理装置１２０，１３０も同等の外観を備える。

【００４１】

スキャナ１１３は、原稿となる紙上の画像を照明し、ＣＣＤラインセンサ（図示せず）を走査することによって、ラスタイメージデータを生成する。使用者が原稿を原稿フィーダ４０５のトレイ４０６にセットして、操作部１１２で読み取りの起動を指示する。これによりコントローラ１１１のＣＰＵ３０１がスキャナ１１３に指示を与え、トレイ４０６にセットされた原稿を１枚ずつフィードしてスキャナ１１３が原稿上の画像の読取動作を行なう。

【００４２】

プリンタ１１４は、ラスタイメージデータをシートに印刷する。その印刷方式は、感光体ドラムや感光体ベルトを用いた電子写真方式、微少ノズルアレイからインクを吐出してシート上に直接画像を印刷するインクジェット方式等のいずれでもよい。尚、プリンタ１１４の印刷動作は、ＣＰＵ３０１からの指示によって起動される。プリンタ１１４は、異なる用紙サイズ又は異なる用紙向きを選択できるように複数の給紙段を持ち、それに対応した用紙カセット４０１，４０２，４０３を有している。また排紙トレイ４０４は、印刷が終了して排紙されたシートを積載して載置する。

【００４３】

図５は、本実施形態に係る画像処理装置の操作部の構成を示す平面図である。

【００４４】

ＬＣＤ表示部５０１は、ＬＣＤ（液晶表示装置）上にタッチパネル５０２が貼られており、画像処理装置１１０の操作画面及びソフトキーを表示する。そして使用者により表示されているキーが押されると、その押された位置を示す位置情報がコントローラ１１１のＣＰＵ３０１に伝えられる。スタートキー５０５は、原稿の読み取り動作を指示する場合等に操作されるキーである。このスタートキー５０５の中央部には、緑と赤の２色ＬＥＤ５０６があり、その色によってスタートキー５０５を操作できる状態であるか否かを判別できる。ストップキー５０３は、稼働中の画像処理装置１１０の動作を停止させる場合に操作されるキーである。ＩＤキー５０７は、使用者のユーザＩＤを入力するときに操作されるキーである。またリセットキー５０４は、操作部１１２からの設定を初期化するときに操作されるキーである。

【００４５】

図６は、本実施形態に係る画像処理装置の操作部及び操作部Ｉ／Ｆの構成をコントローラの構成と対応させて示すブロック図である。

【００４６】

上述したように、操作部１１２は、操作部Ｉ／Ｆ３０６を介してシステムバス３０７に接続される。システムバス３０７には、ＣＰＵ３０１，ＲＡＭ３０２，ＲＯＭ３０３，ＨＤＤ３０４が接続されている。ＣＰＵ３０１は、ＲＯＭ３０３とＨＤＤ３０４に記憶された制御プログラム等に基づいて、システムバス３０７に接続される各種デバイスとのアクセスを総括的に制御する。

【００４７】

タッチパネル５０２や各種ハードキー５０３，５０４，５０５，５０７からのユーザ入力情報は、入力ポート６０１を介してＣＰＵ３０１に渡される。ＣＰＵ３０１は、ユーザによる入力情報の内容と制御プログラムとに基づいて表示データを生成し、出力ポート６０２を介してＬＣＤ表示部５０１に、その表示データを出力する。また必要に応じて２色ＬＥＤ５０６の表示を制御する。

【００４８】

図７は、本実施形態に係る画像処理装置の操作部に表示される標準的な操作画面の一例を示す図である。

【００４９】

図７の最上部の表示領域７０１に並んでいるボタン群は、この画像処理装置１１０が提供する各種機能から１つを選択するためのボタン群である。「コピー」は、スキャナ１１３でスキャンし読み取った原稿の画像データをプリンタ１１４で印刷して原稿の複写物を得るための機能である。「送信」は、スキャナ１１３で読み取った原稿データやＨＤＤ３０４に蓄積されている画像データを各種出力先に送信するための機能である。この場合の出力先としては、ネットワークインタフェース３０８経由で各種のプロトコルによって送信可能な各種の出力先、及び、モデム３０９経由でファクシミリ等のプロトコルによって送信可能な各種の出力先がある。そして、それらの中から複数の出力先を選択して送信することができる。「ボックス」は、ＨＤＤ３０４に蓄積されている画像データやコードデータ等の文書ファイルを閲覧、編集、印刷、及び送信する機能である。ＨＤＤ３０４に蓄積される文書ファイルは、スキャナ１１３によって読み取った原稿の画像データ、ネットワークインタフェース３０８経由で受信したデータを含む。更には、ネットワークインタフェース３０８経由で他の装置から受信した印刷データを蓄積したデータ、モデム３０９経由で他の装置から受信したファクシミリデータ等をも含む。このボックス機能は、ユーザのオフィス環境において電子的なメールボックスとして利用できる。またパスワードを入力して初めてシートへの印刷を許可することによって、ＰＤＬ印刷ジョブの守秘性を高めるセキュアド印刷として利用することもできる。また、このボックス機能は、画像処理装置１１０のＨＤＤ３０４だけでなく、他の画像処理装置１２０，１３０のＨＤＤや、情報処理装置１０１，１０２が公開する共有ファイルシステムにも適用できる。更には、サーバシステム１４０がサービスする共有ファイルシステムやデータベースシステム等に蓄積されている画像データやコードデータ等の文書ファイルにネットワーク１００を介して

アクセスし、閲覧、編集、印刷及び送信する場合にも適用できる。「拡張」は、スキャナ１１３を外部装置から利用するためにロックするなど、各種の拡張機能を呼び出すための機能である。「検索」は、画像処理装置１１０や他の画像処理装置のボックス機能、情報処理装置が公開する共有ファイルシステム、サーバシステム１４０がサービスする共有ファイルシステムやデータベースシステム等から、所望の文書を検索する機能である。

【００５０】

図７の７０２は、コピー機能が選択された場合の操作画面の一例を示している。７０３はステータス表示領域であり、表示領域７０１で選択された機能の如何に関わらず、この画像処理装置１１０の各機能や装置自体の情報等の各種のメッセージをユーザに対して表示するのに使用される。

【００５１】

図８は、本実施形態に係るＤＢ管理システム２０１に格納される各データベースの抽象的なデータ構造を示す模式図である。

【００５２】

文書ＤＢ２０２は、複数の文書レコード８０１、複数の関連レコード８１１を含む。文書レコード８０１は、ユーザが取り扱う紙文書や電子的な文書ファイルに対応するレコードである。この文書レコード８０１は、文書メタデータ８０２、文書内容データ８０３、及びその文書のページ数分のページレコード８０４を含む。

【００５３】

文書メタデータ８０２は、文書レコード８０１に対応する文書に関連する各種のメタデータを格納するレコードである。文書メタデータ８０２は、対応する文書に関して、文書名、作者、作成日付、データ形式、データサイズ、ページ数、タグ、関連文書（関連メタデータ）、ジョブ履歴、検索履歴（操作履歴メタデータ）等の情報を含む。ジョブ履歴や検索履歴は、その文書データを入力した画像処理装置１１０から取得しても良い。ここでタグとは、文書にユーザが付けた任意の文字列からなるキーワードのようなもので、ユーザは一つの文書に対して複数のタグを自由に付すことができるので、文書を種々の基準で分類したり検索し易くするのに役立つ。また共有の文書に対して、その文書を後で参照したり利用する複数のユーザが、タグを追加していくこともできる。これによって文書を分類や検索するための意味的なメタデータを飛躍的に充実させることが期待できる。このアプローチをフォークソノミー（folksonomy）と呼ぶ場合がある。このフォークソノミーは、「folks」（人々・民衆）と「taxonomy」（分類学）を組み合わせた用語である。ジョブ履歴は、この文書を処理対象として実行された一連のジョブを特定する参照情報のリストである。１つの文書レコードは複数のジョブレコードへの参照を保持する場合がある。例えば、明らかに同一と特定できる文書を複数のジョブが処理対象とした場合、その文書と複数のジョブレコードとが関連付けられる。

【００５４】

文書内容データ８０３は、文書そのものの内容に対応するデータである。コード化された文書データが格納された場合は、テキストやアプリケーションプログラムのデータなどが文書内容データとなる。紙の原稿に対応し画像スキャナで読み取られたラスタ画像データのように、文書を構成するページが明確に分離している場合は、ページレコード８０４内部に内容データを含める。

【００５５】

ページレコード８０４は、文書を構成するページのそれぞれに対応するレコードである。スキャナ１１３で原稿の表面と裏面をそれぞれ読み取ったラスタ画像データや、アプリケーションプログラムのデータを展開部２１０で展開してページ単位に分割した画像データ及び構造情報やテキストやメタデータ等が、それぞれのページレコードに対応する。ページレコード８０４は、ページメタデータ８０５とページ内容データ８０６等を含む。

【００５６】

ページメタデータ８０５は、ページレコード８０４に対応するページに関連する各種のメタデータを格納するレコードである。このページメタデータ８０５は、構造情報、特徴

、サムネール、検索履歴、媒体ＩＤ（媒体特徴データ）等を含む。構造情報は、画像構造解析部２０８や展開部２１０が解析して格納したページの構造に関する情報である。特徴は、画像特徴抽出部２０７が抽出して格納したページを構成する画像の特徴を表現する情報である。サムネールは、ページ全体の画像やページに含まれる画像要素を、解像度変換（又は縮小変倍）して、比較的小さくて扱い易いサイズにした画像である。このサムネール画像は、ページメタデータ８０５の生成時に生成しても良く、或は外部からのリトリーブに応えるために必要となったときオンデマンドに生成してもよい。また、スケジューリングされたバッチ処理によって、まだ生成されていないサムネール画像群をまとめて生成するタスクを非同期に実行してもよい。検索履歴は、対応するページに関する検索が行われた履歴情報を表現するデータである。媒体ＩＤは、対応するページに関連する紙等の記録媒体を識別する情報である。例えば、媒体ＩＤは、紙に埋め込まれた超小型無線ＩＣチップの識別情報を用いて構成する。又は、ペーパーフィンガー印刷（紙指紋）技術等に基づき、シート毎に固有な紙の繊維パターンを識別情報として用いて構成する。又は、シートに印刷される可視又は不可視の画像パターンを識別情報として用いて構成する。画像パターンによって媒体識別情報を符号化する技術として、１次元ならびに２次元バーコード技術や、透明インクや透明トナー技術、磁性インクや磁性トナー技術、等の技術を用いることが好適である。

【００５７】

印刷ジョブに伴って文書レコード８０１を生成する場合、印刷に用いる媒体が超小型無線ＩＣチップが埋め込まれたシートであれば、図４の用紙カセット４０１，４０２，４０３又は出力用紙の搬送経路に配備された受信機（不図示）が識別情報を読み取る。そしてその識別情報をページレコード８０４のページメタデータ８０５中の媒体ＩＤに格納する。またスキャンジョブに伴って文書レコードを生成する場合、スキャンした媒体が超小型無線ＩＣチップが埋め込まれたシートであれば、原稿フィーダ４０５の用紙搬送経路に配備された受信機（不図示）によって識別情報を読み取る。そして、その識別情報をページレコード８０４のページメタデータ８０５中の媒体ＩＤに格納する。また印刷ジョブで、シートごとに固有な紙の繊維パターンを識別情報として用いる場合は、用紙カセット４０１，４４０２，４０３又は出力用紙の搬送経路に配備された受信機（不図示）によって出力用紙の繊維パターンを読み取って符号化する。そして、ページレコード８０４のページメタデータ８０５中の媒体ＩＤに格納する。またスキャンジョブに伴って文書レコードを生成する場合は、スキャナ１１３、又は原稿フィーダ４０５の用紙搬送経路に配備された繊維パターン読み取り専用スキャナ（不図示）によって、入力シートの繊維パターンを読み取って符号化する。そしてページレコード８０４のページメタデータ８０５の媒体ＩＤにストアする。

【００５８】

またシートに印刷される可視又は不可視の画像パターンを識別情報として用いる場合は、印刷ジョブに際して、まずページごとに、又は、文書ごとにユニークな値をＵＵＩＤ等の技術を用いて生成する。そして、文書ごとにユニークな値を符号化して画像パターンを生成する。更に、その画像パターンと印刷ジョブの画像データ（ページ内容データ）とをオーバレイした画像データをプリンタ１１４によって印刷する。こうして印刷されたシートが正常に排紙されると、文書ごとにユニークな値をページレコード８０４のページメタデータ８０５の媒体ＩＤに格納する。一方、スキャンジョブに伴って文書レコード８０１を生成する場合は、スキャナ１１３によって原稿に埋め込まれた画像パターンを読み取って復号化する。次に、得られた文書ごとにユニークな値をページレコード８０４のページメタデータ８０５の媒体ＩＤに格納する。

【００５９】

ページ内容データ８０６は、ページそのものの内容に対応するデータである。ここには紙原稿のページをスキャナ１１３で読み取ったラスタ文書データや、ファクスで受信した各ページのラスタ文書データが格納される。またコード文書を展開部２１０でレンダリングした画像データ等のページ単位の画像データも格納される。また、ページ画像をＯＣＲ

２０９で文字認識して得たテキストデータや、コード文書を展開部２１０が展開して得たページ単位のテキスト情報なども、このページ内容データ８０６に格納される。

【００６０】

関連レコード８１１は、複数の文書レコード８０１の組に関連付けられ、文書とその関連文書との間の関連を表現するためのレコードである。この関連レコード８１１は、文書レコード８０１からみると付随するメタデータの一種とみなすことができる。関連レコード８１１は、関連文書リスト及び関連情報等を含む。関連文書リストは、関連レコード８１１によって関連を記述する複数の文書レコードを表現するデータである。関連情報は、関連文書リストによって結合される複数の文書データ間の関連を表現するデータである。

【００６１】

ジョブＤＢ２０３は、複数のジョブレコード８０８を含む。ジョブレコード８０８は、ユーザが実行した文書処理ジョブの各々に対応するレコードである。ジョブレコード８０８は、文書レコード８０１からみると付随するメタデータの一種とみなすことができる。ジョブレコード８０８は、日時、操作者、要求した装置、処理した装置、処理内容、及び、処理文書等を含む。日時は、ジョブを実行した日時を表現するデータである。操作者は、ジョブを実行したユーザを特定するデータである。要求した装置は、ジョブ実行の要求元になった装置である（例えば、パーソナルコンピュータ１０１から画像処理装置１１０に印刷した場合、要求した装置はパーソナルコンピュータ１０１となる）。処理した装置は、ジョブを実質的に処理した装置である（例えば、パーソナルコンピュータ１０１から画像処理装置１１０に印刷した場合、処理した装置は画像処理装置１１０となる）。処理内容は、ジョブの処理内容を特定する情報である。この処理内容は、ジョブの種別、及びそれぞれのジョブ種別において選択可能な各種オプションと設定可能な各種パラメータをどのように選択・設定して処理したか特定する情報を含む。処理文書は、このジョブが処理対象とした文書を特定する参照情報のリストである。１つのジョブレコードが複数の文書レコードを参照する場合がある。これは例えば、１つのジョブが複数の文書を処理対象として実行された場合である。

【００６２】

インデクスＤＢ２０４は、複数のインデクスレコード８０９を含む。インデクスレコード８０９は、文書ＤＢ２０２やジョブＤＢ２０３から高速にデータを検索するためのインデクス情報であり、複数の文書レコード８０１及び複数のジョブレコード８０８を参照している。インデクス情報は、検索キーとして与えられる画像に類似した画像を含む文書レコードを高速に検索するのに使用される。また、検索キーとして与えられるテキストで、文書内容データ８０３やページ内容データ８０６に含んでいる文書レコード８０１を高速に全文検索するのにも使用される。また、検索キーとして与えられる条件に合致するメタデータを持つ文書レコード８０１やジョブレコード８０８を高速に検索したりするために使用され、このインデックス情報は、インデクス生成部２１１によって生成される。

【００６３】

［実施形態１］

図９は、本実施形態１において、ある時点でＤＢ管理システム２０１に格納された各データベースの具体的なデータ構造例を示すインスタンス関係図である。

【００６４】

ＤＢ管理システムデータ構造９０１は、図８に示す抽象的なデータ構造に則った、ＤＢ管理システム２０１に構築された、文書レコード、関連レコード、ジョブレコードの各インスタンス群とその関連を例示している。ＤＢ管理システムデータ構造９０２は、ある時点で存在するインスタンス群とその関連を例示している。文書レコードインスタンスｄ１は、具体的な一つの文書に対応する文書レコード８０１のインスタンスを示し、文書レコードインスタンスｄ２，ｄ３，ｄ４，ｄ５，ｄ６，ｄ７，ｄ８，ｄ９も同様である。関連レコードインスタンスｒ１は、具体的な一つの関連に対応する関連レコード８１１のインスタンスを示し、図示しない文書レコードインスタンスと文書レコードインスタンスｄ１とを関連付けている。関連レコードインスタンスｒ２，ｒ３，ｒ４，ｒ５，ｒ６，ｒ７，

ｒ８も、関連レコードインスタンスｒ１と同様である。ジョブレコードｊ１は具体的な一つのジョブに対応するジョブレコード８０８のインスタンスを示し、文書レコードインスタンスｄ１を対象として実施されたジョブの情報を保持し、文書レコードインスタンスｄ１と関連付けられている。ジョブレコードｊ２，ｊ３，ｊ４，ｊ５，ｊ６，ｊ７，ｊ８，ｊ９，ｊ１０，ｊ１１も同様である。

【００６５】

図１０は、本実施形態１に係る文書処理システムの画像処理装置における文書入力処理の手順を説明するフローチャートである。このフローチャートで示す手順は画像処理装置１１０のＣＰＵ３０１により実行される組み込みアプリケーションプログラムによって達成される。

【００６６】

このフローチャートの一連の手順は、画像処理装置１１０の印刷機能、文書転送機能、文書蓄積機能等に対して、パーソナルコンピュータ１０１から送られた文書データを受信することにより開始される。また或は、このフローチャートの一連の手順は、画像処理装置１１０のファクス受信機能によって、モデム３０９が公衆回線から画像データを受信することにより開始されても良い。この場合、文書入力処理とはファクス受信処理に相当している。また、このフローチャートで示す手順は、ユーザが操作部１１２の表示領域７０１のコピー、送信、ボックス機能等で、スキャナ１１３によって原稿の画像データを読み取る処理を選択し、スタートキー５０５で読み取り動作を起動したときに開始されても良い。この場合、文書入力処理とは、原稿をスキャンして文書データを読み取る処理に相当する。

【００６７】

まずステップＳ１で、画像処理装置１１０は各種の文書入力処理を行う。この文書入力処理は、印刷、画像処理装置１１０ストレージへの蓄積、ファクスやＩＦＡＸ、電子メール等への転送等のためにパーソナルコンピュータ１０１から送られた文書データを入力する処理を含む。またファクス受信やＩＦＡＸ受信、電子メールの受信等の受信処理の結果として遠隔の装置から送られた文書データを入力する処理でも良い。またコピー、画像処理装置１１０のストレージ（ボックス機能）への蓄積、ファクスやＩＦＡＸ、電子メール等への送信等のために、スキャナ１１３で読み取った紙媒体上の画像データを文書データとして入力する処理でも良い。このように画像処理装置１１０が行う文書入力処理は、ネットワークやシリアルインタフェース等を介してオンライン文書データを入力するオンライン文書入力と、紙媒体のスキャン等によりオフライン文書を入力するオフライン文書入力とに大別される。オンライン文書データとは、内容データを計算処理によって一意に解析可能であり、また文書管理システムが文書データを管理するために使用するメタデータを含むものである。文書管理システムは、このメタデータを用いて、文書データの検索を行ったり、複数の文書データを関連付けて管理したりする。一方、紙媒体から読み取ったりファクスで受信したラスタ画像データが含まれるオフライン文書データは、文書管理システムに対してオフラインの状態となっている。つまり、オフライン文書データには文書管理システムが文書データを管理するために使用するメタデータが含まれていない。なお、ラスタ画像データには、画像自体の属性を示す画像作成日時や解像度などの簡易的な属性情報が付加されていてもよい。また、ラスタ画像データとは、例えば、ビットマップ形式の画像データや、ビットマップ形式の画像データを圧縮した圧縮画像データなどのことを指す。

【００６８】

次にステップＳ２に進み、ステップＳ１で行ったジョブ処理に対応するジョブレコード８０８を生成してジョブＤＢ２０３に格納する。次にステップＳ３に進み、ステップＳ１で入力したジョブ処理で入力した文書データに対応する文書レコード８０１を生成して文書ＤＢ２０２に格納する。またステップＳ２で生成したジョブレコード８０８を、ステップＳ３で生成した文書レコード８０１に対するメタデータの一つとして関連付ける。また文書データに付随する他のメタデータも同様に、文書メタデータ８０２として文書ＤＢ２

０２に格納する。

【００６９】

次にステップＳ４に進み、文書入力処理がラスタ文書データのオフライン入力処理か否かを判定する。ここでラスタ文書データのオフライン入力処理であればステップＳ６へ進むが、ラスタ文書データのオフライン入力処理でなければステップＳ５へ進む。ステップＳ５では、入力文書のメタデータと内容データとに基づいて、入力文書と関連する文書をジョブアーカイブ・アプリケーションから検索する。即ち、入力文書と関連する文書レコードを、既にＤＢ管理システム２０１に既に格納されている文書レコード中から検索する。この文書入力処理は、オンライン入力であるか、或はコード文書の入力処理であるため、リレーショナルデータベース管理システム（ＲＤＢＭＳ）等の分野で公知のデータ検索技術を駆使して関連文書レコードを検索できる。こうしてステップＳ５の処理を終了するとステップＳ９に進む。

【００７０】

ステップＳ６～Ｓ８では、ラスタ文書データに関連する文書データを特定する関連文書特定処理を実行する。即ち、媒体に基づく関連文書検索処理を行う。紙文書のスキャンによる文書入力処理の場合、前述したように紙媒体の媒体ＩＤを識別し、それがページメタデータ８０５の媒体ＩＤデータと同一又は類似しているページレコード８０４を検索する。こうしてページレコード８０４が見つかると、そのページレコード８０４を含む文書レコード８０１は、その入力文書の物理的なページ媒体（紙）を過去に扱った際に格納した文書レコードであると識別できる。即ち、その紙に印刷したときに生成した文書レコード８０１として、入力文書との関連を見出すことができる。或はまた、過去にその紙をスキャンして、コピー、送信、或はボックスに蓄積したり、紙をキーとした画像検索をした場合等に生成した文書レコード８０１として、入力文書との関連を見出すことができる。

【００７１】

次にステップＳ７に進み、画像として埋め込まれたコードデータに基づく関連文書検索処理を行う。ラスタ文書データの入力処理の場合、前述したようにラスタ画像（文書）に含まれる二次元バーコード等の解析、復号によって、画像として埋め込まれたメタデータや内容データを抽出できる。その抽出したコードデータに基づいて、入力文書と関連する文書をジョブアーカイブ・アプリケーションから検索する。即ち、入力文書と関連する文書レコードを、ＤＢ管理システム２０１に既に格納されている文書レコードから検索する。検索キーは、画像から復号したコードデータであるため、リレーショナルデータベース管理システム（ＲＤＢＭＳ）等の分野で公知のデータ検索技術を駆使して関連文書レコードを検索できる。

【００７２】

次にステップＳ８に進み、ラスタ文書データと類似する文書データをジョブアーカイブ・アプリケーションから検索する。ここで関連する文書とは、文書レコード８０１の類似度が高い文書、即ち、文書内容データ８０３の類似度が高い文書、文書メタデータの類似度が高い文書等を関連文書として検索する。またページレコード８０４の類似度が高いページ（類似ページ）を含む文書、即ち、ページ内容データ８０６の類似度が高いページ、ページメタデータ８０５の類似度が高いページを含む文書も関連文書として検索する。特に、ページメタデータ８０５の構造情報データと特徴データを用いて、画像を構成する複数の領域の構造と特徴が類似しているページや、類似の領域要素を含むページを、類似度が高いページであると判定する。そしてステップＳ８からステップＳ９へ進む。

【００７３】

次にステップＳ９で、関連文書の検索結果を判定し、少なくとも１つの関連文書の検索に成功した場合はステップＳ１０に進み、失敗した場合は終了する。ステップＳ１０では、ステップＳ１で生成した文書レコード８０１と、ステップＳ５乃至ステップＳ８で検索した関連文書の文書レコード８０１とを、相互に関連付ける関連レコード８１１を関連文書の数だけ生成して文書ＤＢ２０２に格納する。それぞれの関連レコード８１１の関連文書リストデータには、入力文書及び関連文書に対応する２つの文書レコード８０１への参

照を記録する。また関連情報データには、ステップＳ３で説明した各種の関連を識別する情報を記録する。類似度に基づく関連については、その類似度の程度を表現する値もここに記録する。

【００７４】

図１１は、本実施形態１において、印刷、受信、蓄積等に伴うコード文書やメタデータつき文書の文書入力処理を完了した時点でＤＢ管理システム２０１に格納された各データベースの具体的なデータ構造例を示すインスタンス関係図である。尚、ＤＢ管理システムデータ構造９０２は、図９のＤＢ管理システムデータ構造９０２と同じである。図１１では、図９に示すデータ構造例に対してデータ構造１１０１が追加されている。

【００７５】

データ構造１１０１は、文書レコードインスタンスｄ１０、ジョブレコードインスタンスｊ１２、及び関連レコードインスタンスｒ９、ｒ１０を含む。文書レコードインスタンスｄ１０は、印刷、受信、蓄積等によって文書入力されたコード文書やメタデータ付文書に対応する文書レコード８０１のインスタンスである。ジョブレコードインスタンスｊ１２は、この文書入力処理に関する情報を記録したジョブレコード８０８のインスタンスである。関連レコードインスタンスｒ９は、ステップＳ５の検索によってヒットした、ＤＢ２０２に既に存在した関連文書レコードｄ２と、文書入力された文書に対応する文書レコードｄ１０とを関連付けるために生成され蓄積されたインスタンスである。関連レコードインスタンスｒ１０は、同様にステップＳ５の検索によってヒットした、ＤＢ２０２に存在した関連文書レコードｄ５と、文書入力された文書に対応する文書レコードｄ１０とを関連付けるために生成され蓄積されたインスタンスである。

【００７６】

図１２は、本実施形態１において、紙媒体として与えられた文書のスキャンやラスタ画像のファクス受信等による文書入力処理を完了した時点でＤＢ管理システム２０１に格納された各データベースの具体的なデータ構造例を示すインスタンス関係図である。ここでは、図１１に示したＤＢ管理システム２０１に格納された各データベースの具体的なデータ構造例を示すインスタンス関係図にデータ構造１２０１が追加されている。それ以外は前述の図１１と同じであるため、それらの説明を省略する。

【００７７】

データ構造１２０１は、文書レコードインスタンスｄ１１、ジョブレコードインスタンスｊ１３、及び推定関連レコードインスタンスｒ１１，ｒ１２を含む。

【００７８】

文書レコードインスタンスｄ１１は、スキャンやファクス受信等によって入力されたラスタ文書データに対応する文書レコード８０１のインスタンスである。この文書レコードインスタンスｄ１１は、オフライン入力によって得られた文書であるため、文書メタデータや文書内容データをまったく持たないか、又は、比較的貧弱なデータしか持たない（図では×印によってこれを示している）。ジョブレコードインスタンスｊ１３は、この文書入力処理に関する情報を記録したジョブレコード８０８のインスタンスである。推定関連レコードインスタンスｒ１１は、ステップＳ８の類似画像検索によってヒットした、ＤＢ２０２に存在する関連文書レコードｄ５と、入力された文書に対応する文書レコードｄ１１とを関連付けるために生成され蓄積されたインスタンスである。推定関連レコードインスタンスｒ１２もまた、ステップＳ６の媒体ＩＤ検索によってヒットした、ＤＢ２０２に存在する関連文書レコードｄ９と、入力された文書に対応する文書レコードｄ１１とを関連付けるために生成され蓄積されたインスタンスである。

【００７９】

図１３は、本実施形態１に係る関連レコード８１１のインスタンス群に記録される関連情報をテーブル構造によって表現したデータ表現の一例を示す図である。このデータ表現は、図８のデータ構造における文書ＤＢ２０２を表現するためにＤＢ管理システム２０１によって管理される。図１３は、図１２に例示したインスタンス群とそれらの関連に対応している。図において、各行は、関連の参照元文書から参照先文書への有向グラフの情報

に対応し、各列は、関連を構成する関連ＩＤ、参照元文書ＩＤ、参照先文書ＩＤ、関連種別、関連度の情報を示している。

【００８０】

関連ＩＤは、図９～図１２で関連レコードインスタンスｒとして示された、関連レコード８１１の各インスタンスを識別するＩＤである。参照元文書ＩＤと参照先文書ＩＤは、それぞれ文書レコード８０１のインスタンスを識別するＩＤであり、この参照元文書から参照先文書への関連を記述している。関連種別は、参照元から参照先への関連の種別を示す。関連度は、関連の程度を示す数値である。この関連度は、「０」よりも大きく「１」以下の値をとり、値が大きいほど関連の度合が大きいことを示している。

【００８１】

以下、関連種別について説明する。

【００８２】

「文書一致（旧版）」は、文書を識別する情報により同一文書の異なる版であることが特定された場合に付与される関連情報であり、参照元文書ＩＤの文書が参照先文書ＩＤの旧版であることを表現する。ここで同一文書の異なる版であることは、以下に挙げるような各種の文書識別情報の比較によって特定できる。例えば、参照元と参照先とで文書メタデータ８０２の所在情報のＵＲＬが等しい、或は最新版を示す関連文書の所在を示すＵＲＬが等しい、或は文書名等の文書ＩＤが等しい場合には、これら文書は同一文書であると判定できる。また例えば、印刷された紙文書の場合には、その媒体ＩＤが印刷ジョブレコードに記録されており、その紙文書と印刷ジョブのソースデータとなった文書が等しい場合も同一文書と判定できる。また例えば、文書内容データ８０３やページ内容データ８０６が等しい場合も同一文書と判定できる。「文書一致（新版）」は、「文書一致（旧版）」と逆方向の関連を表現する。

【００８３】

「手動関連付け（参照先）」は、ユーザによって手動で付与された関連を表現する。ユーザは、ジョブアーカイブ・アプリケーションやボックス等の文書管理システムを介して文書ＤＢ２０２の文書間に、手動で関連を付与できる。いまユーザがある文書Ａを別の文書Ｂに関連付けた場合は、「手動関連付け（参照先）」の参照元文書ＩＤは文書ＡのＩＤとなり、参照先文書ＩＤは文書ＢのＩＤとなる。「手動関連付け（参照元）」は、「手動関連付け（参照先）」と逆方向の関連を表現する。

【００８４】

「作者一致」は、両文書の文書メタデータ８０２の作者情報が等しい場合に付与される関連情報である。「作者一致」は、一般に双方向の関連である。複数の著者からなる共著の文書の場合、作者ごとに対応する複数の関連を他の文書との間に持つ場合もある。

【００８５】

「包含（含まれる）」は、関連する両文書の間に内容の包含関係が特定される場合に付与される関連情報である。文書の内容の包含関係は、文書内容データ８０３又はページレコード８０４の比較によって判定できる。「包含（含まれる）」は、参照元文書ＩＤの文書の内容が、参照先文書ＩＤの文書の内容に含まれることを意味する。また「包含（含む）」は、「包含（含まれる）」の逆である、参照元文書ＩＤの文書の内容が、参照先文書ＩＤの文書の内容を含むことを意味する。

【００８６】

「作成日一致」は、文書メタデータ８０２の作成日付が等しい場合に付与される関連情報である。「作成日一致」は一般に双方向の関連である。

【００８７】

「タグ一致」は、文書メタデータ８０２のタグ情報に等しいタグを持つ場合に付与される関連情報である。「タグ一致」は一般に双方向の関連である。複数のタグが付けられた文書の場合、タグごとに対応する複数の関連を他の文書との間に持つ場合もある。

【００８８】

「文書内容データ類似」は、文書内容データ８０３やページレコード８０４の類似性を

判定し、その類似度が閾値を超えていると判定された場合に付与される関連情報である。「文書内容データ類似」は、一般に双方向の関連である。

【００８９】

「同一ジョブ処理対象」は、同一のジョブの処理対象となった文書群に付与される関連情報である。ジョブレコード８０８の処理文書リストに含まれる文書群の各組み合わせに対して付与される。「同一ジョブ処理対象」は一般に双方向の関連情報である。

【００９０】

「画像類似（再オンライン化）」は、紙媒体のスキャンやラスタ画像（文書）のファクス受信等による文書入力処理によって文書ＤＢ２０２に追加された文書レコードと、既にＤＢ２０２に存在した文書レコードとの間に付与される関連情報である。この関連情報は、文書入力時に図１０の手順によって生成され格納される。また文書入力と同時でなく、後に図１０のステップＳ６乃至ステップＳ１０と同等のバッチ処理によって関連レコードを生成して格納しても良い。このバッチ処理によって関連レコードを生成する場合は、文書入力処理を高速化できるという効果や、文書入力時に実行可能な関連文書の検索処理よりも、より高度な検索を実現できる効果などがある。

【００９１】

「画像類似（再オンライン化）」の参照元文書ＩＤは、ＤＢ２０２に存在した関連文書レコードであり、参照先文書ＩＤは、追加された文書レコードを表わしている。「画像類似（オンライン）」は、「画像類似（再オンライン化）」と逆方向の関連である。

【００９２】

「媒体ＩＤ一致（再オンライン化）」は、紙媒体のスキャンやラスタ画像（文書）のファクス受信等による文書入力処理によって文書ＤＢ２０２に追加された文書レコードと、ＤＢ２０２に存在する文書レコードとの間に付与される関連情報である。この関連情報は、文書入力時に図１０の手順によって生成され格納される。また文書入力と同時でなく、後に図１０のステップＳ６乃至ステップＳ１０と同等のバッチ処理によって関連レコードを生成して格納しても良い。このバッチ処理によって関連レコードを生成する場合は、文書入力処理を高速化できるとともに、文書入力時に実行可能な関連文書の検索処理よりも、より高度な検索を実現できる効果などがある。「画像類似（再オンライン化）」の参照元文書ＩＤは、ＤＢ２０２に存在した関連文書レコードであり、参照先文書ＩＤは、追加された文書レコードを表現する。「媒体ＩＤ一致（オンライン）」は、「媒体ＩＤ一致（再オンライン化）」と逆方向の関連である。

【００９３】

図１４（Ａ）（Ｂ）は、本実施形態１に係る文書検索アプリケーションの基本画面である文書検索画面の一例を示す図である。尚、以下の図面において、下線が付してある文字列は、その表示領域を押すと対応する詳細情報表示ウィンドウが開き、それぞれの情報のより詳細な情報を確認できることを表している。

【００９４】

文書検索画面１４００は、文書検索アプリケーションの基本画面である。本実施形態に係る文書検索アプリケーションは、文書検索画面を操作部１１２の表示領域７０２（図７）に、この検索画面１４００を表示する。文書検索画面１４００は、検索条件設定領域１４０１、検索キー入力領域１４０２、及び検索スタート指示領域１４０３を有している。

【００９５】

検索条件設定領域１４０１は、検索条件を設定したり確認したりするための領域である。検索条件ラジオボタン１４０４は、基本的な検索条件を選択し、また選択されている設定を確認するためのラジオボタンである。選択肢の「全てのキーを含む」は、セットした全ての検索キーにヒットする文書を検索することを示す。「いくつかのキーを含む」は、セットした検索キーのうちのいずれかにヒットする文書を検索することを示す。「高度な検索」は、検索オプションボタン１４０５によって設定した、より詳細な検索条件の設定に基づいて、ヒットする文書を検索することを示す。検索オプションボタン１４０５は、詳細な検索条件を設定するウィンドウを開くためのボタンである。この詳細な検索条件の

設定は、高度な検索モードで検索が実行されたときヒットする文書を判定する基準として用いる高度な検索条件の設定を含む。この詳細な検索のオプションとして、メタデータ検索や全文検索を併用する条件を、類似画像検索と併用して設定できる。

【００９６】

メタデータ検索は、文書に対応する文書レコード８０１に関して、その文書メタデータやページメタデータ８０５群や対応するジョブレコード８０８にそれぞれ格納されているデータ項目毎に検索条件を指定する検索方法である。このメタデータ検索は以下の検索条件を設定できる。即ち、文書名、所有者、作成日付、データ形式、ページ数、タグ、関連文書、ジョブ履歴（日時、操作者、要求した装置、処理した装置、処理内容、このジョブにおいて処理した他の処理対象文書）、ページの構造情報等に基づく検索条件を指定できる。従って、文書名や所有者や作成日時やタグ等に基づく一般的な検索に加えて、関連文書や過去にその文書が検索された履歴に基づいて検索することもできる。また文書を構成するページに関して、方向がポートレート（縦長）かランドスケープ（横長）か、用紙のサイズ、ページ数がｎページ以上ｍページ未満、カラーかモノクロか、画像とテキストの割合はどの程度か等に基づいて検索できる。また、いつ、どこで、誰が、どのように処理した文書であるかという、ジョブの履歴に基づいて検索することもできる。

【００９７】

全文検索は、検索キーとしてテキスト（文字列）を設定し、文書の全テキスト中に設定された文字列を含む文書を検索する。文書のテキストは、文書レコード８０１に含まれる文書内容データ８０３、ページレコード８０４のいずれかに含まれるページ内容データに含まれているテキストである。また文書メタデータ８０２やページメタデータ８０５に含まれているテキスト形式のデータを全文検索の対象に加えることもできる。また、文書と関連するジョブレコード８０８に含まれているテキスト形式のデータを全文検索の対象に加え、ジョブレコード８０８がヒットした場合は、対応する文書レコード８０１がヒットするように設定することもできる。

【００９８】

図１４（Ａ）の検索キー入力領域１４０２は、検索キーを入力するための領域であり、類似画像検索の検索キーとする画像を設定したり確認するための情報が表示されている状態を示している。

【００９９】

原稿スキャンボタン１４０６は、画像処理装置１１０のスキャナ１１３を用いて原稿を読み取り、その画像データを類似画像検索の検索キーとするためのボタンである。この原稿スキャンボタン１４０６が押されると画像スキャンウィンドウを開く。この画像スキャンウィンドウでは、コピー機能や送信機能における原稿読み取り設定や、ＴＷＡＩＮ等のよく知られたインタフェースに基づく一般的なスキャナデバイスドライバの原稿読み取り設定等と同様に、原稿読み取りのパラメータを設定できる。そして操作部１１２のスタートキー５０５が押されると、設定されている原稿読み取りパラメータに従って原稿をスキャンし、その読み取った画像データを検索キー画像として入力する。このとき原稿のスキャンが完了したとき画像スキャンウィンドウが開かれていれば閉じる。原稿スキャンボタン１４０６を押さずにスタートキー５０５が押された場合は、デフォルトの原稿読み取りパラメータ、又は、その時点までに設定されている原稿読み取りパラメータに従って原稿をスキャンする。

【０１００】

ボックス画像選択ボタン１４０７は、画像処理装置１１０のボックス機能を利用して、予め格納されている文書群の中から検索キー画像を選択するためのボタンである。ボックス機能によってＨＤＤ３０４を閲覧して、検索キー画像として利用したい画像を含む文書を選択できる。また他の画像処理装置１２０，１３０のＨＤＤや、情報処理装置１０１，１０２が公開する共有ファイルシステム等に記憶されている画像データやコードデータ等も同様に、検索キー画像として選択できる。更には、サーバシステム１４０がサービスする共有ファイルシステムやデータベースシステム等に蓄積されている画像データやコード

データ等も同様に、検索キー画像として選択できる。

【０１０１】

検索キー画像設定領域１４０８は、セットされている検索キー画像の組を確認し操作するための領域である。検索キー画像設定状況メッセージ１４０９は、検索キー画像のセット状況を示すメッセージであり、セットされている検索キー画像の個数等を表示する。検索キー画像表示領域１４１０は、セットされている検索キー画像群をブラウズする領域である。この領域１４１０に、検索キーとしてセットされた画像に対応する検索キーアイコンの組が並べて表示される。原稿スキャンボタン１４０６やボックス画像選択ボタン１４０７を用いて検索キー画像を入力すると、対応する検索キーアイコンがこの領域に追加される。原稿スキャンボタン１４０６を用いて原稿の表面と裏面や、複数の原稿をまとめてスキャンした場合、或は、ボックス画像選択ボタン１４０７を用いて複数ページから構成される文書を選択することができる。この場合、それぞれのページを読み取った画像データに対応する複数の検索キーアイコンを追加することを選択できる。また、複数ページ画像を含む文書に対応する１つの検索キーアイコンを追加するようにも選択できる。検索キーアイコン１４１１は、１つの検索キー画像に対応するアイコンである。このアイコン１４１１を介して、検索キーに対する各種の操作を指示できる。検索キーＩＤ１４１２は、この検索キーを特定するための識別子である。検索キーサムネール１４１３は、この検索キーのサムネール画像である。検索キーサムネール１４１３が押されると、画像ビューアウィンドウを開いて、そのサムネールよりも大きなサイズで検索キー画像を表示する。この画像ビューアウィンドウによって、ユーザは検索キー画像の詳細を確認できる。検索キー概要１４１４は、この検索キーに関する簡単な説明の表示である。検索キー詳細ボタン１４１５は、この検索キーに関する詳細情報を確認するためのボタンである。このボタン１４１５により、検索キー概要１４１４よりも詳細に検索キーに関する情報を表示する検索キー詳細ウィンドウを開くことができる。この検索キー詳細ウィンドウでは、この検索キーに固有の検索条件を設定することもできる。また今後の検索するときこの検索キーを再利用するために、検索キーをボックスに保存することもできる。検索キー編集ボタン１４１６は、この検索キーを編集するためのボタンで、このボタン１４１６が押下されると、検索キーを編集するための検索キー編集ウィンドウが開かれる。この検索キー編集ウィンドウでは、検索キー画像に対してトリミング、マスキング、ノイズ除去等の各種画像処理を施して、所望の検索キー画像へと編集できる。また、検索キー画像を切り分けて、複数の検索キー画像に分割できる。また、複数ページ画像を含む文書に対応する１つの検索キーをページ画像単位に切り分けて、それぞれのページ画像に対応する検索キー画像に分割できる。検索キー削除ボタン１４１７は、この検索キを検索キーの組から取り除くためのボタンである。検索キーＩＤ１４１２が「キー＃２」であるボックスから選択した画像の検索キーアイコンも同様であるが、図面を簡略化するために各キーの参照記号は省略している。

【０１０２】

検索スタート指示領域１４０３は、検索処理を起動するための領域である。検索開始ボタン１４１８は、検索処理を開始させるためのボタンである。この検索開始ボタン１４１８が押されると、サーバシステム１４０がサービスするジョブアーカイブ・アプリケーションに対して検索処理要求を発行する。この際、検索条件設定領域１４０１で設定した検索条件と、検索キー入力領域１４０２でセットした検索キーとを用いた検索処理を要求する。

【０１０３】

一方、図１４（Ｂ）の検索キー入力領域１４０２は、検索キーを入力するための領域であり、キーワード検索の検索キーとするキーワードを設定したり確認したりするための情報が表示されている状態を示している。検索キーワードフィールド１４１９は、キーワード検索に用いるキーワード群を表示する領域である。入力リセットボタン１４２０は、設定中の検索キーワードをクリアするためのボタンである。スクリーンキーボード１４２１は、検索キーワードを設定するために用いる画面上の仮想キーボードである。

【０１０４】
図１５は、本実施形態１に係る文書検索アプリケーションにおける文書検索結果リスト画面の一例を示す図である。図において、斜体の文字列は、実際の画面表示では、その文書が持つ、対応するメタデータの実際の値が表示されることを示している。
【０１０５】
この文書検索結果リスト画面１５００は、文書検索アプリケーションがジョブアーカイブ・アプリケーションから検索処理要求の応答を受信したときその検索結果を表示する画面の一例を示す。本実施形態に係る文書検索アプリケーションは、この文書検索結果リスト画面を操作部１１２の表示領域７０２に表示する。この文書検索結果リスト画面１５００は、検索リスト操作領域１５０１、検索リスト表示領域１５０２、スクロールバー１５０３を有している。
【０１０６】
検索リスト操作領域１５０１は、検索結果リストの表示制御等を操作するための領域である。表示フィルタリング状態１５０４は、検索リスト表示領域１５０２に表示されている文書が、検索によりヒットした複数の文書のうち、どのような表示フィルタを施した結果として得られた文書であるかを表示している。ここではサーバシステム１４０のリトリーブ部２１２から受信したヒット文書を全て表示することもできるし（即ち、「全文書」、フィルタ無し）、またヒットした文書の中から表示フィルタ設定した条件に従って選別した結果を表示することもできる。
【０１０７】
表示フィルタ設定ボタン１５０５は、表示フィルタ条件を設定するためのボタンである。表示フィルタ設定ボタン１５０５が押されると、表示フィルタ設定ウィンドウを開き、ユーザに所望のフィルタ条件を設定させる。ヒットした文書群の文書レコード８０１に含まれる各種の情報に基づく条件をフィルタ条件に設定できる。即ち、文書メタデータ８０２、ヒットしたページのページレコード８０４のページメタデータ８０５、文書に関連付けられたジョブレコード８０８等に格納された各情報に対するパターンマッチング条件等を設定できる。言い換えると、検索オプションボタン１４０５で設定できる詳細な検索のオプションと同様のフィルタ条件を設定できる。例えば、文書名や作成日時やタグ等に基づく一般的なフィルタリングに加えて、関連文書や過去にその文書が検索された履歴に基づいてフィルタリングすることもできる。また文書を構成するページに関して、方向がポートレート（縦長）かランドスケープ（横長）か、用紙のサイズ、ページ数がｎページ以上ｍページ未満に基づいてフィルタリングすることもできる。更には、カラーかグレースケール（連続階調画像）か白黒二値画像か、画像とテキストの割合はどの程度か等の基準に基づいてフィルタリングすることもできる。また、いつ、どこで、誰が、どのように処理した文書であるかという、ジョブに関連する基準に基づいてフィルタリングすることもできる。
【０１０８】
表示項目設定領域１５０６は、検索でヒットした文書を検索リスト表示領域１５０２に表示する際に、文書ごとに表示する項目を制御する領域である。チェックボックスの矩形又はチェックボックスにつけられたラベル文字列を押すたびに、チェックボックスの選択状態と非選択状態とが交互に切り替わる。「属性情報を表示」が選択されている場合、文書名、データ形式、ページ数、文書の所在情報、等の文書に関するメタデータを検索リスト表示領域１５０２に表示する。また「サムネールを表示」が選択されている場合、検索条件にヒットしたページのサムネール画像を検索リスト表示領域１５０２に表示する。
【０１０９】
文書サマリーサムネール設定領域１５０７は、検索でヒットした文書を検索リスト表示領域１５０２に表示する際に、各文書の文書サマリーサムネールの表示形式を制御する領域である。表示項目設定領域１５０６の「サムネールを表示」が選択されており、かつ、「文書サマリーサムネールを表示」チェックボックスが選択されている場合は、文書サマリーサムネールを表示する。この文書サマリーサムネールとは、その文書の概要を視覚的

に把握しやすくするために、文書を構成するページに対応する一組のサムネールを並べたものである。

【０１１０】

文書サマリーサムネール構成設定領域１５０８は、文書サマリーサムネールを構成するサムネールの構成を設定する領域である。文書サマリーサムネール構成設定領域１５０８には、４つの数値入力用のテキスト入力フィールドが設けられており、それぞれに「先頭」、「前」、「後」、「末尾」のラベル文字列をつけてある。「先頭」の数値によって、文書の先頭ページから何ページ分のサムネールを表示するかを設定する。「前」の数値によって、検索でヒットしたページに先行するページのサムネールを何ページ分表示するか設定する。「後」の数値によって、検索でヒットしたページに後続するページのサムネールを何ページ分表示するか設定する。更に「末尾」の数値によって、文書の末尾ページから何ページ分のサムネールを表示するか設定する。文書サマリーサムネールアニメーション表示チェックボックス１５０９は、文書サマリーサムネールをアニメーション表示するか否かを設定するためのチェックボックスである。再検索ボタン１５１０は、図１４に示す文書検索画面１４００に戻るためのボタンである。絞り込み検索ボタン１５１１は、文書検索画面１４００に戻って絞り込み再検索を行うためのボタンである。検索リスト表示領域１５０２に表示された文書の中から検索キーとして追加したい文書（検索キーとして追加したい画像を含む文書）をマークしてから絞り込み検索ボタン１５１１を押す。これにより、マークをつけられた文書が検索キーとして検索キー画像表示領域１４１０に追加された状態で文書検索画面１４００に戻り、絞込み再検索を実行できる。

【０１１１】

的確な検索キー画像をできるだけ多く、かつ簡便に追加できることにより、所望の文書の検索ヒット率を向上し、見つけ出しやすくできる。また追加された検索キー画像の特徴量を分析し、類似度の判定における各種特徴量の配点を調整することによって、よりユーザの意図に即した類似画像の検索を行うことが可能となる。即ち、ユーザが絞り込み検索によって追加した検索キー画像は、検索を行うユーザの観点からみても主観的に類似度が高いサンプル画像であると判断できる。従って、この検索キー画像の類似度が、より高く評価されるように、複数の特徴量と類似度判定アルゴリズムとを組み合わせる配点を調整する。例えば、元の検索キー画像と追加された検索キー画像の間で、形状に基づく類似度が高く色合いに基づく類似度が低かった場合は、絞り込み再検索では形状ベースの類似度を色合いよりも優先する。同様にして、色合い優先、配色パターン優先、オブジェクト構造木の類似度優先など、適切な調整を行うことができる。

【０１１２】

検索リスト表示領域１５０２は、検索した結果、検索条件に合致した文書の一覧を表示する領域である。検索ヒット文書表示１５１２，１５１３，１５１４，１５１５は、それぞれ検索条件に合致した文書に対応する情報を表示している。デフォルトの設定では、ヒット率が高い文書ほどリストの上位に表示するようにしている。同等のヒット率の場合、文書の価値を数値化した文書ランクが高い文書ほど上位に表示する。このときフィルタ設定ボタン１５０５を押して、デフォルト以外の順序で並べ替えて文書リストを表示し直すこともできる。例えば、文書の作成日、最終参照日、文書名、データ形式、ページ数、文書の所在情報、その文書を対象として行われたジョブの日時や操作者や装置や処理内容など、文書に関連付けられた各種メタデータに基づいて、昇順又は降順に表示できる。尚、文書リストの表示順序を設定し直すと、即時にリスト表示が更新される。

【０１１３】

次にデフォルトの表示順序の拠り所となる文書のヒット率について簡単に説明する。類似画像検索は、アルゴリズムごとに固有の類似度に基づくが、一般に類似度は「似ている程度」を表現する連続量であり、「似ているか、又は、似ていない」の二値ではない。但し、本実施形態の実装上、類似度が所定の閾値よりも低い画像は似ていないものとして切り捨てる。また類似度が所定の閾値より高い画像は、相対的に類似度の高い画像と低い画像とを区別する。与えられた検索キー画像との類似度が高い画像を含む文書の方が、比較

的低い画像を含む文書よりも、ヒット率を高く算出する。また、検索キーは複数指定できるので、より多くの検索条件に合致する文書のヒット率は、より少ない検索条件だけに合致する文書よりもヒット率を高くする。また類似画像検索の検索キー画像が複数指定される場合、類似度の高い画像を多く含む画像のヒット率を高くする。尚、「全てのキーを含む」ラジオボタンが選択されて検索された場合は、与えられた検索キーの全てに合致しなければヒットしない。尚、検索リスト表示領域１５０２に表示される文書の内、リストの下位に表示される文書は、上位に表示される文書よりも、文書表示をより簡略化したり縮小したりすることによって、一画面の中に表示可能な文書の総件数を増やすようにしてもよい。

【０１１４】

スクロールバー１５０３は、文書検索結果リスト画面１５００をスクロールするためのスクロールバーである。多くの場合、検索リスト表示領域１５０２には大量の文書が表示されるので、操作部１１２の表示部５０１の表示領域に納まらない場合が多い。そこでユーザは、画面をスクロールしながら文書を一覧してその中から所望の文書を見つけ出す。尚、検索リスト表示領域１５０２の最下部等にページ送りのためのボタンなど（不図示）を配置して、検索結果文書のリストを複数のページに分割して表示してもよい。尚、検索リスト表示領域１５０２の最下部等に配置したリスト印刷ボタン（不図示）を押すと、文書検索結果リストを印刷するように構成してもよい。

【０１１５】

図１６は、本実施形態１に係る検索ヒット文書表示の一例を示す図である。尚、ここで検索ヒット文書表示１５１２～１５１５はそれぞれ同様に構成されているので、検索ヒット文書表示１５１２を例にして説明する。

【０１１６】

データ形式アイコン１６０１は、対応する文書のデータ形式を表現するためのアイコンである。文書名１６０２は、対応する文書の文書名を表示する。データ形式１６０３は、対応する文書のデータ形式を表示する。ページ数１６０４は、対応する文書のページ数を表示する。文書の所在情報１６０５は、対応する文書が保存されているファイルサーバ等の格納位置を特定する情報を表示する。この文書の所在情報は、ＵＲＩや、又はファイルサーバとそのファイルシステム中のファイルパス文字列等によって識別される。ジョブアーカイブシステムがアーカイブした文書の場合、そのジョブにおいて収集された処理対象文書の控えデータが保存されている位置を表示しても良い。また或は、処理対象文書のオリジナルデータが保存されている位置が特定できる場合はその位置を表示してもよい。履歴情報１６０６は、対応する文書を処理対象として過去に施されたジョブ処理や検索等の履歴を表示する。これにより、いつ、誰が、どんな処理を、どの装置において、この文書に対して施したかといった履歴情報を確認できる。ページ１６０７は、対応する文書を構成するページの内、検索キーの条件にヒットしたページのページ番号を表示する。ヒットページサムネール１６０８は、対応する文書を構成するページの内、検索キーの条件にヒットしたページの概観を表現するためのサムネール画像を表示する。先頭ページサムネール１６０９は、対応する文書の先頭のページの概観を表現するサムネール画像を表示する。ここでは図１５の文書サマリーサムネール構成設定領域１５０８で設定されたページ数分のサムネール画像を並べて表示する。前ページサムネール１６１０は、検索キーにヒットしたページに先行するページの概観を表現するサムネール画像を表示する。ここでは、文書サマリーサムネール構成設定領域１５０８で設定されたページ数分のサムネール画像を並べて表示する。後ページサムネール１６１１は、検索キーにヒットしたページに後続するページの概観を表現するサムネール画像を表示する。ここでは、文書サマリーサムネール構成設定領域１５０８において設定されたページ数分のサムネール画像を並べて表示する。末尾ページサムネール１６１２は、対応する文書の末尾のページの概観を表現するサムネール画像を表示する。ここでは、文書サマリーサムネール構成設定領域１５０８において設定されたページ数分のサムネール画像を並べて表示する。

【０１１７】

尚、非常に多くのページを文書サマリーサムネールに表示しようとした場合、より縮小率の高い小さなサムネールを表示して、限られた表示領域の中に、多くのサムネール画像を表示するように調整する。或は、比較的優先度の低いページのサムネールをより小さく縮小して表示したり、先行するページの裏側に重ね合わせページの一部が隠れるように配置して表示しても良い。また或は、表示を省略したりすることによって、限られた表示領域の中に収まるように調整するのが望ましい。尚、表示領域が不十分な場合は、文書サマリーサムネール中に優先的に表示する優先度の高いページは、次のようなアルゴリズムに従って選択する。例えば、文書の前の方のページをより優先したり、先に指定された検索キーに対応してヒットしたページをより優先させて表示する。また或は、類似画像検索の条件にヒットした場合は、類似度の高いページを優先して表示するようにしても良い。

【０１１８】

印刷ボタン１６１３は、対応する文書を印刷するためのボタンである。保存ボタン１６１４は、対応する文書をボックス機能に保存するためのボタンである。送信ボタン１６１５は、対応する文書を送信機能によって送信するためのボタンである。タグ付けボタン１６１６は、対応する文書のタグを操作するためのボタンである。タグ付けボタン１６１６を押すと、文書タグウィンドウが開き、既に、その文書に設定されているタグを閲覧及び編集するとともに、任意のタグを新たに追加登録できる。関連文書ボタン１６１７は、対応する文書の関連文書を操作するためのボタンである。この関連文書ボタン１６１７を押すと、関連文書ウィンドウが開き、その文書に関連付けられている文書を閲覧及び編集したり、当該文書と他の文書の関連を追加登録したりできる。マーク付けチェックボックス１６１８は、対応する文書をマークするためのチェックボックスである。リストに表示された文書群の内、幾つかの文書に選択的に働く操作を行うと、このチェックボックスが選択状態にある文書が対象となる。例えば、マーク付けチェックボックス１６１８を選択状態にしてから、絞り込み検索ボタン１５１１を押すと、そのマークされた文書群が検索キーに追加された状態で再検索を続けられる。オンライン属性１６１９は、対応する文書がオフライン入力処理によって入力された文書であるか否かの区別を表示する。その文書がオフライン入力で入力されたものであれば「再オンライン化」と表示し、そうでなければ「オンライン」と表示する。

【０１１９】

図１７は、本実施形態１に係る文書検索アプリケーションにより検索された文書データに関連する関連文書データを表示する処理の手順を示すフローチャートである。この手順は、文書検索アプリケーションを構成する処理の一部であり、画像処理装置１１０のＣＰＵ３０１等によって実行される。この手順は、ユーザが注目している注目文書に対応する文書に関して、例えば検索ヒット文書表示１５１２の関連文書ボタン１６１７（図１６）が押されたとき等に起動される。

【０１２０】

まずステップＳ２１で、検索する関連文書の関連距離ｎ（ＲＡＭ３０２の変数エリア）に「１」をセットする。次にステップＳ２２に進み、注目文書から関連距離ｎにある文書レコードを検索して選択する。この関連距離とは、ＤＢ管理システムデータ構造９０１において、注目文書レコードと、それに結び付けられた関連レコードを経由して到達できる関連文書との間に存在する関連レコード数の最小値を指す。いまｎが「１」の場合は、注目文書レコードから見て１つの関連レコードを経由して到達できる文書レコードが検索され、その１つが選択される。次にステップＳ２３に進み、その選択された関連文書がオフライン入力された文書であるか否かを判定する。そうであればステップＳ２４に進み、そうでないときはステップＳ２５に進む。ステップＳ２４では、その選択された関連文書を再オンライン化文書としてマークしてステップＳ２６に進む。一方、ステップＳ２５では、その選択された関連文書をオンライン文書としてマークしてステップＳ２６に進む。ステップＳ２６では、関連距離ｎにある全ての文書レコードを選択したか否か判定する。全て選択した時はステップＳ２７に進むが、そうでないときはステップＳ２２に戻って前述の処理を繰り返す。

【０１２１】
ステップＳ２７では、選択された関連距離ｎの文書レコード群において、再オンライン化文書がオンライン文書よりも下位に表示されるように並べ替えを行う。即ち、推定関連レコードに基づく関連文書レコードが、より明確な関連レコードに基づく関連文書レコードよりも下位に表示されるように並べ替える。次にステップＳ２８に進み、関連距離ｎに１を加える。次にステップＳ２９に進み、関連距離ｎがシステム既定値、又はユーザによって指定された関連距離を越えたか否かを判定する。超えていないときはステップＳ２２に戻って前述の処理を実行するが、超えたときはステップＳ３０に進み、検索された関連文書レコード群を表示して、この処理を終了する。この際、ステップＳ２７で、再オンライン化文書を下位に並べ替えているため、同一の関連距離の文書群中ではオンライン文書の方が再オンライン化文書よりも上位に表示される。

【０１２２】
尚、図の手順では、同一関連距離にある関連文書レコードの中で再オンライン化文書を下位に表示するための並べ替えを行ったが、検索された全ての関連文書レコードの中で再オンライン化文書を下位に表するように構成してもよい。

【０１２３】
図１８は、本実施形態１に係る文書検索アプリケーションにおける注目文書に対する関連文書検索結果リストの表示結果の画面例を示す図である。この画面では、図１４に示す画面で指定された検索条件に従って検索された文書と、この文書に関連する文書として図１７に示すフローチャートに沿って検索された関連文書とを、対応付けてユーザに提示する。ここで図１５と共通する部分は同じ記号で示し、それらの説明を省略する。この関連文書検索結果リスト画面は、文書検索アプリケーションが図１７のフローチャートに示した手順等によって表示する画面である。

【０１２４】
関連距離が１であるオンライン文書ラベル１８０１は、以下に表示される文書レコード群が、注目文書と関連距離１の関連で結び付けられたオンライン文書であることを示すラベルである。オンライン文書表示１８０２は、オンライン属性１６１９（図１６）の表示例であり、ここでは、検索ヒット文書表示１５１２に対応する文書レコードがオンライン文書であることを示している。関連距離が１である再オンライン化文書ラベル１８０３は、以下に表示される文書レコード群が、注目文書と関連距離１の関連で結び付けられた再オンライン化文書であることを示すラベルである。再オンライン化文書表示１８０４は、オンライン属性１６１９の表示例を示し、ここでは検索結果表示１５１４に対応する文書レコードが再オンライン化文書であることを示している。

【０１２５】
このようにして、関連文書リストの表示において、再オンライン化文書はオンライン文書よりも下位に表示される。

【０１２６】
尚、入力文書に対する関連文書の検索と関連付けは、入力処理の直後に全て完了する必要はなく、後で十分な時間をかけて行うバッチ処理をスケジューリングするように構成してもよい。

【０１２７】
また、ジョブアーカイブシステムのＤＢ管理システム２０１は、大規模ストレージ装置１４２に集中して配備しなくてもよい。ストレージ及びデータベース管理システムが複数の装置に分散した分散データベースとして配備し、分散検索できるように構成してもよい。例えば、パーソナルコンピュータ１０１，１０２が備えるストレージや画像処理装置１１０，１２０，１３０が備えるＨＤＤ３０４に基づく分散データベースシステムとして構成することもできる。

【０１２８】
以上説明したように本実施形態１によれば、スキャンやファクス受信による文書のオフライン入力処理において、ＤＢ上の既存文書レコード群の中から関連文書を検索し、その

入力する文書を、検索された関連文書レコードと関連付けて記憶できる。このため、ＤＢ中の文書レコードの関連文書レコードを検索するときに、オフライン入力処理で格納された文書も関連文書として検索できる。また文書のオフライン入力処理で生成される文書レコードには、スキャンジョブやファクス受信ジョブに固有のジョブ履歴など固有のメタデータが付与される。これにより注目文書に対する関連文書検索の結果得られる情報量と価値を増大できる。例えば、その文書を誰がいつどこで処理したかといったジョブ処理の履歴情報も保持できるため、この情報を基に、紙の形態で存在するはずの文書を探すことも容易となった。

【０１２９】

更に、従来は、オフライン入力された文書を、キーワード検索等のコードデータに基づく検索によって見つけ出すことは困難であった。これに対して本実施形態の構成では、検索にヒットしたオンライン文書から関連文書として辿ることによって、再オンライン化された所望の文書を容易に見つけ出すことができる。

【０１３０】

また逆に、展開部２１０を持たないシステムでは、コード文書を類似画像検索やページ構造情報に基づく検索によって見つけ出すのが困難だった。本実施形態によれば、これらの検索にヒットした再オンライン化文書から関連文書として辿ることによって、コード文書として入力された所望の文書も容易に見つけ出すことができる。

【０１３１】

また本実施形態１では、関連文書検索等の検索結果表示において、文書のオフライン入力処理によって格納された文書レコードを識別するための表示を行うことができる。このため、ユーザは検索結果リストの文書群の中から、どれがオンライン文書データで、どれがオンライン化文書データであるかを一目で識別できる。

【０１３２】

また本実施形態１では、関連文書検索等の検索結果表示において、文書のオフライン入力処理によって格納された文書レコードを、それ以外の文書レコード群よりも下位に表示している。このため、ユーザは推定関連による関連文書の情報よりも、コード情報やオンライン情報等に基づく関連文書の情報の方を先に見ることが容易となった。

【０１３３】

［実施形態２］

次に本発明の実施形態２について説明する。この実施形態２では、関連付けられた再オンライン化文書をオンライン文書から参照できるようにする。

【０１３４】

図１９は、本発明の実施形態２に係る文書検索アプリケーションで、再オンライン化された文書レコードに対して既存文書レコードからメタデータや内容データを伝播する処理の手順を示すフローチャートである。尚、この実施形態２に係るシステム及び画像処理装置等のハードウェア構成は前述の実施形態１と同様であるため、その説明を省略する。

【０１３５】

この手順は、例えば図１２に示したＤＢ管理システム２０１のデータ構造を操作する処理として、例えば画像処理装置１１０のＣＰＵ３０１により実行される。この手順は、推定によって追加された関連レコードｒ１１，ｒ１２等に関して実行される。従ってこの手順は、前述の図１０に示したような文書入力処理手順におけるラスタ文書データのオフライン入力処理の後処理のための追加ステップとして起動される。また或は、文書入力手順とは独立したバッチ処理として起動されても良い。また或は、後述する図２３に示すような検索処理の前処理として起動されても良い。

【０１３６】

先ずステップＳ３１で、推定関連レコードの１つに注目する。次にステップＳ３２に進み、その注目している推定関連レコードに付与されている関連度が所定の閾値以上であるか否かを判定する。閾値以上であればステップＳ３３に進むが、そうでないときはステップＳ３５に進む。ステップＳ３３では、推定関連レコードが関連付ける文書レコードの組

について、ＤＢ２０２に従来から存在したオンライン文書に付与されたメタデータ群を、ＤＢ２０２に追加された再オンライン化文書へ伝播する。即ち、前者の文書のメタデータ群に付与されているものと同等のメタデータを後者の文書にも付与する。このメタデータの伝播は、文書メタデータ８０２をコピーすることで行ってもよい。また或は、後者の文書レコードの文書メタデータ８０２が、前者の文書レコード中の文書メタデータ８０２を参照するようにリンクを張ることで行ってもよい。

【０１３７】

次にステップＳ３４に進み、推定関連レコードが関連付ける文書レコードの組について、ＤＢ２０２に従来から存在したオンライン文書に付与された文書内容データを、ＤＢ２０２に追加された再オンライン化文書へ伝播する。即ち、前者の文書レコードが持つ文書内容データと同等の文書内容データを後者の文書にも持たせる。この文書内容データの伝播は、文書内容データ８０３をコピーすることで行ってもよい。また或は、後者の文書レコードの文書内容データ８０３が、前者の文書レコードの文書内容データ８０３を参照するようにリンクを張ることで行ってもよい。そしてステップＳ３５に進む。ステップＳ３５では、全ての推定関連レコードが注目済みであるか否かを判定する。注目済みのときは一連の手順を終了するが、そうでないときはステップＳ３１へ戻り、新たな推定関連レコードに注目して一連の手順を繰り返す。

【０１３８】

図２０は、本実施形態２において、再オンライン化文書の文書レコードにメタデータや内容データを伝播した結果としてＤＢ管理システム２０１に構築されるデータ構造の一例を示す図である。

【０１３９】

ラスタ文書データのオフライン入力処理よりも前の時点で、ＤＢ管理システム２０１に存在したデータ構造９０２は、既存の文書レコードｄ５，ｄ９を含んでいる。文書レコードｄ５は、それに対応する文書メタデータｄ５ｍと、文書内容データｄ５ｃとを保持している。文書メタデータｄ５ｍのタグには、例えば「プロダクトＸ」、「性能」「機能」という３つの文字列が割り当てられている。文書レコードｄ９は、それに対応する文書メタデータｄ９ｍと、文書内容データｄ９ｃとを保持している。文書メタデータｄ９ｍのタグには、例えば「プロジェクトＡ」、「日程」、「要員」という３つの文字列が割り当てられている。

【０１４０】

ラスタ文書データのオフライン入力処理によって追加されたデータ構造１２０１は、入力処理によって生成された文書レコードｄ１１と、その文書と関連を推定された既存文書レコードとの間を関連付ける推定関連レコードｒ１１，ｒ１２を含む。推定関連レコードｒ１１は、既存のオンライン側文書レコードｄ５と、再オンライン化された文書レコードｄ１１とを結び付けている。推定関連レコードｒ１２は、既存のオンライン文書レコードｄ９と再オンライン化された文書レコードｄ１１とを結び付けている。文書レコードｄ１１は、ラスタ文書データのオフライン入力処理によって生成された文書レコードであるため、文書入力処理そのものから得られる文書メタデータ及びコード化された内容データは非常に貧弱であるか又は空である。そこで図１９のフローチャートで示したメタデータ等の伝播処理手順によって、関連文書からメタデータと内容データの伝播を受ける。

【０１４１】

文書メタデータｄ１１ｍは、文書メタデータｄ５ｍから伝播された文書メタデータｄ５ｍ－ｐの情報と、文書メタデータｄ９ｍから伝播された文書メタデータｄ９ｍ－ｐの情報を含む。即ち、文書メタデータｄ１１ｍのタグには、例えば「プロダクトＸ」、「性能」「機能」、「プロジェクトＡ」、「日程」、「要員」の６つの文字列が割り当てられているものと同等に扱われる。文書内容データｄ１１ｃは、文書内容データｄ５ｃから伝播された文書内容データｄ５ｃ－ｐの内容と、文書内容データｄ９ｃから伝播された文書内容データｄ９ｃ－ｐの内容を含む。

【０１４２】

図２１は、実施形態２に係る文書検索アプリケーションで、再オンライン化された文書レコードに対して既存文書レコードからメタデータや内容データを確信度に基づき伝播する処理の手順を示すフローチャートである。このフローチャートは、図１９に示した手順の変形例を示している。この手順は、例えば図１２に示したＤＢ管理システム２０１のデータ構造を操作する処理として、例えば画像処理装置１１０のＣＰＵ３０１において実行される。この手順は、推定によって追加された関連レコードｒ１１，ｒ１２等に関して実行される。従ってこの手順は図１０に示したような文書入力処理手順におけるラスタ文書データのオフライン入力処理の後処理のための追加ステップとして起動される。また或は、文書入力手順とは独立したバッチ処理として起動されても良い。また或は、図２３に示すような検索処理の前処理として起動するようにしてもよい。

【０１４３】

先ずステップＳ４１で、推定関連レコードの１つに注目する。次にステップＳ４２に進み、注目している推定関連レコードに付与されている関連度が所定の閾値以上であるか否かを判定する。ここで閾値以上であればステップＳ４３に進み、そうでないときはステップＳ４６に進む。ステップＳ４３では、注目している推定関連レコードに割り当てられた関連度に基づき関連の確信度を算出する。再オンライン化文書レコードと既存の文書レコードとの間の関連推定の根拠には、文書ＩＤや媒体ＩＤの一致に基づくような確実なものもあれば、ページ画像の類似判定に基づく、ある程度不確実な推定もある。例えば画像類似度によって判定される推定関連レコードには、類似度の大小等に応じてある範囲を持つ関連度が割り当てられ、関連の確実性が表現されている。この関連の種別と、種別ごとの関連度の大小に応じて、定められたアルゴリズムに従って、推定された関連の確信度を算出する。

【０１４４】

次にステップＳ４４に進み、推定関連レコードが関連付ける文書レコードの組について、ＤＢ２０２に従来から存在したオンライン文書に付与されたメタデータ群を、ＤＢ２０２に追加された再オンライン化文書へ確信度付きで伝播する。即ち、前者の文書のメタデータ群に付与されているものと同等のメタデータを後者の文書にも付与する。このメタデータの伝播は、文書メタデータ８０２をコピーすることで行ってもよい。また後者の文書レコードの文書メタデータ８０２が、前者の文書レコード中の文書メタデータ８０２を参照するようにリンクを張ることで行ってもよい。

【０１４５】

次にステップＳ４５に進み、推定関連レコードが関連付ける文書レコードの組について、ＤＢ２０２に従来から存在したオンライン文書に付与された文書内容データを、ＤＢ２０２に追加された再オンライン化文書へ確信度付きで伝播する。即ち、前者の文書レコードが持つ文書内容データと同等の文書内容データを後者の文書にも持たせる。この文書内容データの伝播は、文書内容データ８０３をコピーすることで行ってもよい。また後者の文書レコードの文書内容データ８０３が、前者の文書レコード中の文書内容データ８０３を参照するようにリンクを張ることで行ってもよい。そしてステップＳ４６に進む。

【０１４６】

ステップＳ４６では、全ての推定関連レコードに注目済みであるか否かを判定する。全てに注目済みでないときはステップＳ４１に戻るが、全てに注目済みのときは、この処理を終了する。

【０１４７】

本実施形態２によれば、再オンライン化文書レコードに関連付けられた文書レコードが持つメタデータや内容データを伝播するように構成したため、再オフライン化文書をメタデータや内容データに基づく検索の対象とすることが可能となった。

【０１４８】

図２２は、本実施形態２に係る文書検索アプリケーションにおいて、再オンライン化文書の文書レコードにメタデータや内容データを確信度付きで伝播した結果としてＤＢ管理システム２０１に構築されるデータ構造の一例を示す図である。このデータ構造は、図２

０に示したデータ構造の変形例の１つであり、図２０と共通する部分は同じ記号で示している。

【０１４９】

ラスタ文書データのオフライン入力処理よりも前の時点で、ＤＢ管理システム２０１に存在したデータ構造９０２は、既存の文書レコードｄ５，ｄ９を含んでいる。文書レコードｄ５は、それに対応する文書メタデータｄ５ｍと文書内容データｄ５ｃを保持している。この文書レコードｄ５は、例えばコードデータ文書の印刷処理等によって生成された文書レコードであるため、付与されたメタデータと内容データは全て確信度１の確実性を持っている。文書メタデータｄ５ｍのタグには、例えば「プロダクトＸ」、「性能」「機能」という３つの文字列が割り当てられている。

【０１５０】

文書レコードｄ９は、それに対応する文書メタデータｄ９ｍと文書内容データｄ９ｃを保持している。この文書レコードｄ９は、例えばコードデータ文書の蓄積処理等によって生成された文書レコードであるため、付与されたメタデータと内容データは全て確信度１の確実性を持っている。文書メタデータｄ９ｍのタグには、例えば「プロジェクトＡ」、「日程」、「要員」という３つの文字列が割り当てられている。

【０１５１】

ラスタ文書データのオフライン入力処理によって追加されたデータ構造１２０１は、入力処理によって生成された文書レコードｄ１１と、その文書と関連を推定された既存文書レコードとの間を関連付ける推定関連レコードｒ１１，ｒ１２を含む。推定関連レコードｒ１１は、既存のオンライン側文書レコードｄ５と再オンライン化された文書レコードｄ１１とを結び付けている。推定関連レコードｒ１１は、例えば画像類似判定によって推定された関連であるため、推定の関連度として「０．６」が割り当てられている。推定関連レコードｒ１２は、既存のオンライン側文書レコードｄ９と再オンライン化された文書レコードｄ１１とを結び付けている。推定関連レコードｒ１２は、例えば紙媒体の繊維パターン（紙指紋）の類似性判定によって推定された関連であるため、推定の関連度として「０．９」が割り当てられている。

【０１５２】

文書レコードｄ１１は、ラスタ文書データのオフライン入力処理によって生成された文書レコードであるため、文書入力処理そのものから得られる文書メタデータ及びコード化された内容データは非常に貧弱であるか又は空である。そこで図１９に示したメタデータ等の伝播処理手順によって、関連文書からメタデータと内容データの関連度に基づく伝播を受ける。文書メタデータｄ１１ｍは、文書メタデータｄ５ｍから伝播された文書メタデータｄ５ｍ－ｐの情報と、文書メタデータｄ９ｍから伝播された文書メタデータｄ９ｍ－ｐの情報とを含む。即ち、文書メタデータｄ１１ｍのタグには、例えば「プロダクトＸ」、「性能」「機能」の３つの文字列がそれぞれ確信度０．６で割り当てられているものと同等に扱われる。更に、文書メタデータｄ１１ｍのタグには、「プロジェクトＡ」、「日程」、「要員」の３つの文字列がそれぞれ確信度０．９で割り当てられているものと同等に扱われる。また文書内容データｄ１１ｃは、文書内容データｄ５ｃから伝播された文書内容データｄ５ｃ－ｐの内容と、文書内容データｄ９ｃから伝播された文書内容データｄ９ｃ－ｐの内容とを含む。伝播に用いられた関連の関連度に応じて、前者の内容データｄ５ｃ－ｐには確信度０．６が、後者の内容データｄ９ｃ－ｐには確信度０．９が割り当てられる。

【０１５３】

尚、２以上の関連距離を持つ文書レコード間でメタデータや内容データを伝播する場合、関連の距離が大きくなるほど関連度を減少させて伝播を行う。即ち、距離が離れた文書レコードから伝播されたデータは、より小さな確信度を持つように構成する。

【０１５４】

図２３は、本実施形態２に係る文書検索アプリケーションにおけるキーワード検索と結果表示処理の手順を示すフローチャートである。この手順は、例えば図１２に示したＤＢ

管理システム２０１のデータ構造を操作する処理として、例えば画像処理装置１１０のＣＰＵ３０１において実行される。

【０１５５】

先ずステップＳ５１で、与えられた検索キーリスト中の注目キーを指し示すキー番号ｉを「１」に、検索ヒット文書リストＲを空集合に初期化する。次にステップＳ５２に進み、検索キーにメタデータ又は内容データがヒットする文書群を選択してヒット文書リストＲｉを作成する。尚、ここでキー番号ｉ、検索ヒット文書リストＲ、ヒット文書リストＲｉはＲＡＭ３０２に設定される。次にステップＳ５３に進み、検索条件がＡＮＤ検索であるか否かを判定する。そうであればステップＳ５４へ進むが、そうでないときはステップＳ５５へ進む。ステップＳ５４では、検索ヒット文書リストＲに含まれる文書集合とヒット文書リストＲｉに含まれる文書集合の積集合を、新しい検索ヒット文書リストＲとしてステップＳ５６へ進む。一方ステップＳ５５では、検索ヒット文書リストＲに含まれる文書集合とヒット文書リストＲｉに含まれる文書集合の和集合を、新しい検索ヒット文書リストＲとしてステップＳ５６へ進む。ステップＳ５６では、与えられた検索キーの全てについて検索済みであるか否かを判定し、検索済みでないときはステップＳ５７に進んで、与えられた検索キーリスト中の注目キーを指し示すキー番号ｉに１を加えてステップＳ５２に進む。

【０１５６】

一方、ステップＳ５６で、全てについて検索済みであると判断するとステップＳ５８に進み、検索ヒット文書リストＲ中の文書群について、より多くの検索キーにヒットした文書を上位に並び替える。次にステップＳ５９に進み、同じ数のキーにヒットした文書群ごとに、より確信度の高いメタデータ、又は内容データにヒットした文書を上位に並び替える。即ち、オンライン文書を再オンライン化文書よりも上位に並び替える。また、再オンライン化文書同士では、より関連度の高い推定関連によって伝播したデータにヒットした文書がより上位にリストされるように並び替える。そしてステップＳ６０に進み、適切な並び替えを終えた検索ヒット文書レコード群を表示して、この処理を終了する。

【０１５７】

図２４は、本実施形態２において、複数の推定関連によって伝播したメタデータを持つ再オンライン化文書が検索結果の上位にヒットする例を示す図である。

【０１５８】

図２４（Ａ）は、図１４（Ｂ）に示した文書検索画面に対して検索キーワードを入力した例を示している。図１４（Ｂ）と対応する構成要素には同一の符号をつけて説明を省略する。

【０１５９】

検索条件ラジオボタン１４０４では、「いくつかのキーを含む」が選択されており、セットした検索キーのうちのいずれかにヒットする文書を検索することが指定されている。検索キーワードフィールド１４１９は、キーワード検索に用いるキーワード群を表示する領域であり、図の検索では「プロジェクトＡ」と「プロダクトＸ」の２つのキーワードが指定されている。

【０１６０】

図２４（Ｂ）は、図２４（Ａ）に示した検索の結果として表示される検索結果リストの画面例を示しており、前述の図１５に示した検索結果リスト表示の一例である。ここでも図１５と対応する構成要素には同一の符号をつけて説明を省略する。

【０１６１】

キーワード検索結果ラベル２４０１は、このラベル２４０１以下に表示される文書が検索にヒットしたことを示すラベルである。検索ヒット文書表示１５１２，１５１３，１５１４，１５１５は、それぞれ検索条件に合致した文書に対応する情報を表示している。キーワード検索の場合、通常はオンライン文書の方がキーワードや内容データの確実度が高いためより上位に表示される傾向にある。しかしながら、複数の検索キーが指定されたキーワード検索においては、複数の推定関連によってメタデータや内容データを伝播された

再オフライン化文書がより上位にヒットする場合もある。図２４（Ｂ）はこの例を示しており、図２３の手順に従って、図２２のデータ構造を処理した場合（図２０でも同様）、再オンライン化された文書レコードｄ１１が検索ヒット文書表示１５１２として、文書レコードｄ４，ｄ１０よりも上位に表示されている。

【０１６２】

尚、この検索手順の一連の処理は、情報処理装置１０１で実行してもよい。或は、一連の処理を部分に分割してそれぞれの処理を担当するソフトウェアを複数の装置上に配備して実行する分散アプリケーションとして構成することもできる。例えば、検索画面や検索結果リストの表示とユーザからの指示入力を画像処理装置１１０で実行し、それ以外の処理を情報処理装置１０１やサーバシステム１４０や他の画像処理装置１２０，１３０等で実行してもよい。逆に、検索画面や検索結果リストの表示とユーザからの指示入力を情報処理装置１０１で実行し、それ以外の処理を画像処理装置１１０やサーバシステム１４０で実行するように構成してもよい。尚、分散アプリケーションを構成する方法の１つとして、ＷｅｂブラウザとＷｅｂサーバの組み合わせによって実現するＷｅｂアプリケーションの形態がよく知られている。

【０１６３】

以上説明したように本実施形態２によれば、再オンライン化文書レコードに関連付けられた文書レコードが持つメタデータや内容データを伝播するため、再オフライン化文書をメタデータや内容データに基づく検索の対象とすることが可能となった。

【０１６４】

更に、複数の検索キーが与えられるキーワード検索等においては、複数の関連によって伝播されたデータを持つ文書レコードを、伝播元の文書レコードよりもむしろ上位にヒットさせることも可能となった。例えば、複数のオンライン文書の印刷を合わせて一連の紙文書が構成されており、更にこの紙文書のコピーに伴って再オンライン化文書レコードが生成された場合を考えると、この性質が有益であることがわかる。

【０１６５】

更に、本実施形態２では、メタデータや内容データの伝播において、関連の種別とその種別ごとの関連度とに基づく「確信度」も伝播している。また、より確からしい推定関連に基づいて伝播されたデータが、以降の検索等の処理結果に対してより強い影響を与えるようにしている。このため、人が行う知的な判断基準に、より良く合致するように、オフライン入力されたラスタ文書データをＤＢに再オンライン化することが可能となった。従って、例えば同じように検索にヒットした場合、推定関連に基づく再オンライン化文書よりも、オンライン文書の方が検索結果リストの上位に表示されるようにできる。また例えば、より確からしい推定関連に基づき伝播されたデータに基づいて検索ヒットした文書を、検索結果リストの、より上位に表示することもできる。

【０１６６】

［実施形態３］

次に、再オンライン化文書による文書ランク伝播を行う実施形態３について説明する。尚、この実施形態３に係るシステム及び画像処理装置等のハードウェア構成は前述の実施形態１と同様であるため、その説明を省略する。

【０１６７】

前述の非特許文献１に開示された手法は、米国ＧｏｏｇｌｅのＰａｇｅＲａｎｋ（登録商標）技術に採用されていることでよく知られている。

【０１６８】

Ｗｅｂのように他の文書を参照するリンクを含む文書からなるデータベースにおいて、他からの参照をその文書への人気投票と考えて、文書の重要度を判定する。ある文書はその文書が持つＰａｇｅＲａｎｋをその文書から参照する文書群へ分配することで、ＰａｇｅＲａｎｋの高い多くの文書から参照されている文書のＰａｇｅＲａｎｋは高くなる。ＰａｇｅＲａｎｋは、その文書の重要度を示す値として、検索エンジンにおける検索ヒット文書の表示順制御等に活用されている。

【０１６９】
図２５は、本実施形態３に係る関連文書の相互参照ネットワークに基づき文書ランクを決定する処理を概念的に説明するフローチャートである。この手順は、例えば図１２に示したＤＢ管理システム２０１のデータ構造を操作する処理として、例えば画像処理装置１１０のＣＰＵ３０１において実行される。
【０１７０】
本実施形態３における文書ランクとは、ＤＢ管理システム２０１の文書レコードが持つ値である。文書ランクの基本概念は、従来技術であるＰａｇｅＲａｎｋと同等の概念であり、文書間の参照のネットワーク関係に応じて決定される。即ち、文書ランクが高い、多くの文書から参照されている文書ほど高い文書ランクを持つように構成されている。
【０１７１】
先ずステップＳ７１で、ある文書に注目する。次にステップＳ７２に進み、その文書が持つ文書ランクを、その文書から参照している文書数で割る。次にステップＳ７３に進み、その割った文書ランクを参照先の文書にそれぞれ配分し加算する。そしてステップＳ７４に進み、全ての文書の文書ランクが決定したかを判定し、全ての文書ランクが決定していないときはステップＳ７１に戻って前述の処理を実行する。こうして全ての文書ランクが決定すると、この処理を終了する。
【０１７２】
本実施形態３の応用において、文書ランクは文書の意味的な重要度を表す指標として、文書間の参照のネットワーク関係を含む各種の情報から総合的に算出される。この文書ランクは、文書のメタデータとして明示的に割り付けられた重要度にも基づく。また機密度、所有者、作者、保管場所、ページ数、等の文書の属性に基づいて文書ランクを算出することもできる。更に、その文書に、後から付けられたタグの数や種類、参照された回数、関連文書の参照関係のネットワーク等に基づいて文書ランクを算出しても良い。関連文書の相互参照関係のネットワークに基づく文書ランクに関して、上述のアルゴリズムのように、文書ランクの高い文書から多く参照されている文書のランクが高くなる。また文書ランクの高い文書と同時に処理（即ち、同時に印刷、送信、保存、リトリーブ、ジョブ結合など）された履歴を持つ文書のランクは高くなる。このような基準に基づいて文書ランクが算出される。
【０１７３】
図２６は、本実施形態３に係る文書インスタンス間の関連種別に対応する、参照関係に基づく文書ランクの伝播を説明する図である。
【０１７４】
ＤＢ管理システム２０１の文書レコード群は、関連レコードによって表現される参照関係の有向グラフに応じて文書ランクを伝播し、文書ランクの配分を行う。
【０１７５】
関連種別（ａ）は「手動関連付け」であり、「参照元」文書レコードから「参照先」文書レコードへの方向に文書ランクを伝播する。
【０１７６】
関連種別（ｂ）は「包含」であり、「含む」文書レコードから「含まれる」文書レコードへの方向に文書ランクを伝播する。
【０１７７】
関連種別（ｃ）は「同一部分共有（引用）」であり、「引用する」文書レコードから「引用される」文書レコードへの方向に文書ランクを伝播する。
【０１７８】
関連種別（ｄ）は「同一部分共有（密度）」であり、「同一部分低密度」文書レコードから「同一部分高密度」文書レコードへの方向に文書ランクを伝播する。
【０１７９】
関連種別（ｅ）は「同一文書新旧版」であり、「同一文書旧版」文書レコードから「同一文書新版」文書レコードへの方向に文書ランクを伝播する。

【０１８０】
関連種別（ｆ）は「同一ジョブ処理対象」であり、「同一ジョブ処理対象」文書レコード間で双方向に文書ランクを伝播する。
【０１８１】
関連種別（ｇ）は「タグ一致」であり、「タグ一致」文書レコード間で双方向に文書ランクを伝播する。
【０１８２】
関連種別（ｈ）は「画像類似」であり、「画像類似」文書レコード間で双方向に文書ランクを伝播する。
【０１８３】
関連種別（ｉ）は「媒体ＩＤ一致」であり、「媒体ＩＤ一致」文書レコード間で双方向に文書ランクを伝播する。
【０１８４】
図２７は、本実施形態３に係るＤＢ管理システム２０１に格納された各データベースの具体的なデータ構造例において文書ランクの伝播と決定例を示すインスタンス関係図である。図において、関連レコードに付された矢印は図２６で説明した文書ランク伝播の方向を示している。また前述の実施形態１の図１２と共通する部分は同じ記号で示し、それらの説明を省略する。
【０１８５】
文書レコードｄ１には、４つの関連レコードから文書ランク（ＤｏｃＲａｎｋ）の配分が流入している。即ち、関連レコードｒ１，ｒ２，ｒ３，ｒ８をそれぞれ経由して文書ランクの配分１５，１００，３５，５０を受け取っている。従って、文書レコードｄ１の文書ランクの値は、流入した配分の和である「２００」となる。また文書レコードｄ１は、１つの関連レコードを経由して文書ランクの配分を流出している。即ち、関連レコードｒ８を経由して文書ランク「２００」を文書レコードｄ３へ渡している。この文書レコードｄ３には、１つの関連レコードから文書ランクの配分が流入している。即ち、関連レコードｒ８を経由して文書ランク「２００」を受け取っている。従って、文書レコードｄ３の文書ランクの値は、流入した配分の和である「２００」となる。また文書レコードｄ３は、４つの関連レコードを経由して文書ランクの配分を流出している。従って、文書レコードｄ３から関連文書の各々へ伝播する文書ランクの配分は「５０」となる。即ち、関連レコードｒ８を経由して文書レコードｄ１へ、関連レコードｒ９とｒ１０を経由して不図示の文書レコードへ、関連レコードｒ１１を経由して文書レコードｄ４へ、それぞれ文書ランク「５０」を渡している。
【０１８６】
以下同様に、関連のネットワークにおける文書ランク伝播の相互関係によって、各文書レコードに固有の文書ランクが決定されている。
【０１８７】
データ構造１２０１は、スキャンやファクス受信等によるラスタ文書データの入力によって追加されたデータ構造である。推定関連レコードｒ１０２，ｒ１０３は、オンライン文書やコード文書同士の関連と同様に、文書ランクを伝播する関連のネットワーク中に組み込まれている。この結果、再オンライン化された文書レコードｄ１１に対して適切な文書ランクが決定されている。またスキャンやファクス受信等によるラスタ文書データの入力処理に伴いデータ構造１２０１が追加される以前には、既にＤＢ２０２に存在したオンライン文書レコードｄ５とｄ９の間の関連は見出されていなかった。しかしデータ構造１２０１が追加されたため、文書レコードｄ１１を介して、文書レコードｄ５と文書レコードｄ９の文書ランクが伝播されている。これにより、オンライン文書レコード群にも、より適切な文書ランクが割り当てられるようになった。この例の場合、文書レコードｄ１１との関連の成立によって、文書レコードｄ５から文書レコードｄ９へ文書ランクの配分が伝播して、文書レコードｄ９の文書ランクが高くなっている。即ち、文書レコードｄ１１に対応するラスタ文書データの入力処理によって、文書レコードｄ９の価値が再発見され

て、その文書の評価が上昇したことになる。

【０１８８】

図２８は、本実施形態３に係る文書検索アプリケーションにおける関連レコード８１１のインスタンス群に記録される文書ランク伝播を伴う関連情報をテーブル構造によって表現したデータ表現の一例を示す図である。このデータ表現は、図８のデータ構造における文書ＤＢ２０２を表現するためにＤＢ管理システム２０１によって管理される。この図２８は、図２７に例示したインスタンス群とそれらの関連に対応している。尚、図２８において、各行は、関連の参照元文書から参照先文書への有向グラフの情報に対応している。また各列は、関連を構成する関連ＩＤ、参照元文書ＩＤ、参照先文書ＩＤ、関連種別、ランク伝播の情報を示している。

【０１８９】

関連ＩＤは、関連レコード８１１（図８）の各インスタンスを識別するＩＤである。参照元文書ＩＤと参照先文書ＩＤは、それぞれ文書レコード８０１のインスタンスを識別するＩＤであり、この行が前者から後者への関連を記述していることを示す。関連種別は、関連方向に対応した関連種別を示す。この関連種別の内容は、図２６で説明したものである。ランク伝播は、関連する方向への文書ランク伝播の有無を示し「１」は参照元文書から参照先文書へ文書ランクを配分することを示している。「０」は配分しないことを示す。

【０１９０】

以上説明したように本実施形態３によれば、スキャンやファクス受信により得られるラスタ文書データのオフライン入力処理において、ＤＢの既存文書レコード群の中から検索された関連文書と、入力する文書との関連に従って文書ランクを伝播している。このため、オフライン入力された文書の文書レコードに対して、その文書の価値を示す適切な文書ランクを決定できるようになった。

【０１９１】

また再オンライン化文書との関連が新たに格納されることによって、従来関連していなかったＤＢの既存のオンライン文書レコード間に新たな関連が生じる。この結果、既存のオンライン文書の文書ランクも、より適切に再計算できるようになった。即ち、ラスタ文書データのオフライン入力処理によって、ＤＢの文書レコードの個々に固有の重要度を示す文書レコードの計算精度を高めることが可能となった。

【０１９２】

即ち、本実施形態３によれば、更に、ある文書を対象として行われた処理に基づいて、また関連文書の相互参照関係のネットワークに基づいて、文書の文書ランクが高まるように構成できる。このため、群集の叡智をより活用できるようになった。即ち、紙のスキャンやファクス受信で得られるラスタ文書データに対するユーザの行動によって、文書の文書ランクも自動的に高まるようになった。従って、電子的な形態ばかりでなく紙等の形態においても、頻繁に処理されている文書（及び関連するオンライン文書）はユーザにとって重要な文書であるという、現実世界の傾向をより反映した重要度判定が可能となった。この文書ランクに基づいて、例えば検索結果リストの表示順序等を制御することによって、ユーザが求める文書を、より迅速に見つけ出し易いシステムを提供できる。

【０１９３】

［実施形態４］

次に本発明の実施形態４について説明する。この実施形態４では、ジョブの処理内容を加味して文書ランクを改善している。

【０１９４】

図２９は、本実施形態３に係る文書検索アプリケーションにおけるＤＢ管理システム２０１に格納された各データベースの具体的なデータ構造例においてジョブレコードを加味した文書ランクの伝播と決定例を示すインスタンス関係図である。この図は、図２７のインスタンス関係図の一部に対応する。対応する構成要素には同一の符号をつけて説明を省略する。

【０１９５】
ジョブレコードｊ１３は、ジョブレコード８０８（図８）のインスタンスの一つである。ジョブレコード８０８は、ユーザが実行した文書処理ジョブの各々に対応するレコードである。本実施形態４に係るジョブレコード８０８では、図８に示した日時、操作者、要求した装置、処理した装置、処理内容、及び、処理文書等の属性に加えて「擬似ＤｏｃＲａｎｋ」属性データが記録されている。また、このジョブレコードｊ１３には擬似文書ランクの「４」が割り当てられている。
【０１９６】
擬似ＤｏｃＲａｎｋデータは、ユーザが実行した文書処理ジョブの属性に応じて決定される擬似的な文書ランクである。ジョブ処理が示唆する対象文書の重要性を反映するように設計された所定のアルゴリズム（後述）に従って、擬似的な文書ランクの値が決定される。擬似文書ランクは、上述した実施形態３と同様の構成に従ってＤＢ管理システム２０１内に構築された文書レコードインスタンスのネットワークを伝播する。
【０１９７】
ジョブレコードｊ１３に割り当てられた擬似文書ランク「４」は、ジョブレコードｊ１３の処理対象文書として参照されている文書レコードｄ１１へ伝播する。もし複数の文書を処理対象とするジョブレコードであれば、この擬似文書ランクの値は、対象文書の文書レコード群に対して分配される。ジョブレコードｊ１３から流入した擬似文書ランク「４」は、文書レコードｄ１１の文書ランク決定において他の文書レコードから分配された文書ランクと同等に扱われ、文書レコードｄ１１の文書ランク決定に寄与する。図の例では、文書レコードｄ１１の文書ランクの値は、他の文書レコードから分配された文書ランク「５０」と「２５」と加算されて「７９」と決定されている。このジョブレコードｊ１３によって寄与された文書レコードｄ１１の文書ランクは、この文書から他の文書への参照関係を表す関連レコードによって他の文書へ伝播される。図の例では、ジョブレコードｊ１３による影響を受けた文書ランク「７９」が、推定関連レコードｒ１０３を介して他の文書へと伝播している。
【０１９８】
図３０は、本実施形態３に係る文書検索アプリケーションにおけるジョブレコードインスタンスに対して擬似的な文書ランクの値を決定する手順を説明するフローチャートである。この手順は、例えば図１２に示したＤＢ管理システム２０１のデータ構造を操作する処理として、例えば画像処理装置１１０のＣＰＵ３０１において実行される。
【０１９９】
ジョブレコードインスタンスに固有の擬似的な文書ランクは、ジョブ処理が示唆する対象文書の重要性を反映するように決定される。ジョブレコードインスタンスの擬似文書ランクは、ジョブレコード８０８の処理属性、日時、操作者、要求した装置、処理した装置、処理内容、及び、処理文書等に応じて、例えば以下のように決定される。
【０２００】
重要な文書を扱う役割を担った操作者によって操作されたジョブには高い擬似的な文書ランクを割り当てる（ステップＳ８１，Ｓ８２）。また重要な文書を扱うように管理された端末装置からジョブ投入要求されたジョブには高い擬似的な文書ランクを割り当てる（ステップＳ８３，Ｓ８４）。また品位が高い仕上げ用装置で処理されたジョブには、品位が低いドラフト確認用装置で処理されたジョブよりも、より高い擬似文書ランクを割り当てる（ステップＳ８５，Ｓ８６）。また、大量部数の処理や長時間をかけた処理にはより高い擬似文書ランクを割り当てる（ススステップＳ８７，Ｓ８８）。例えば大量のコピージョブや、大量の送信ジョブには、より高い擬似文書ランクを割り当てる。
【０２０１】
またスキャン、印刷、送信、受信、蓄積等のジョブにおいて、高品位の処理パラメータが設定されて処理されたジョブには高い擬似文書ランクを割り当てる。例えば、カラー処理にはモノクロ処理よりも高い擬似文書ランクを割り当てる（ステップＳ８９，Ｓ９０）。階調処理は、ビット深度の高い処理により高い擬似文書ランクを割り当てる（ステップ

Ｓ９１，Ｓ９２）。無圧縮又は可逆圧縮が指定された設定の処理には、非可逆圧縮が指定された処理よりも高い擬似文書ランクを割り当てる。また非可逆圧縮同士では低い圧縮率が指定された処理には高い圧縮率が指定された処理よりも高い擬似文書ランクを割り当てる（ステップＳ９３，Ｓ９４）。大きなサイズが設定されたジョブには高い擬似文書ランクを割り当てる（ステップＳ９５，Ｓ９６）。縮小レイアウト指定されず、原稿ページを出力ページの１ページに割り当てるように設定されたジョブには、縮小レイアウトによって２ｕｐ設定されたジョブよりもより高い擬似文書ランクを割り当てる。また、２ｕｐ設定されたジョブには４ｕｐ設定されたジョブよりも高い擬似文書ランクを割り当てる（ステップＳ９７，Ｓ９８）。また用紙への印刷を伴うジョブにおいては、製本設定されたジョブには製本設定されていないジョブよりも高い擬似文書ランクを割り当てる（ステップＳ９９，Ｓ１００）。用紙コスト削減のために両面印刷が奨励されているユーザ環境においては、片面印刷のジョブには両面印刷のジョブよりも高い擬似文書ランクを割り当てる。また用紙コスト削減のために、印刷済み用紙の裏面の再利用が奨励されているユーザ環境においては、両面印刷のジョブには片面印刷のジョブよりも高い擬似文書ランクを割り当てる（ステップＳ１０１，Ｓ１０２）。より高品位の用紙を給紙するために、給紙カセットや用紙銘柄の指定が行われたジョブには、より高い擬似文書ランクを割り当てる（ステップＳ１０３，Ｓ１０４）。更に、ジョブレコード８０８が保持するジョブ処理の日時属性に従って、最近実行されたジョブは以前に実行されたジョブよりも高い擬似文書ランクを持つように、経過時間に応じて擬似文書ランクを割り当てる（ステップＳ１０５，Ｓ１０６）。このようなアルゴリズムに基づいて決定される擬似文書ランクは、算出のタイミングによって可変となる。

【０２０２】

ジョブレコードインスタンスに割り当てる擬似文書ランクの算出は、それが対応するジョブの実行とは独立のタイミングで算出することもできる。文書レコードや他のジョブレコードの追加に伴って、文書ランクや擬似文書ランクを決定する必要が生じたときに算出することもできるし、定期的、不定期的なバッチ処理によって算出することもできる。

【０２０３】

図３１は、本実施形態４に係る画像処理装置１１０の操作部１１２に表示される入力文書の関連文書に関する情報を表示し操作するための画面の一例を示す図である。この画面例は、図７のコピー操作画面の上にダイアログウィンドウが表示された様子を示している。図７と同様の構成は同一の符号をつけて説明を省略する。

【０２０４】

スキャン完了ダイアログウィンドウ３１０１は、コピーのためのスキャン処理が完了したことを示すダイアログウィンドウである。関連文書情報３１０２は、スキャンし終わった入力文書の関連文書に関する情報を表示し、関連文書を操作するためのユーザインタフェース領域である。関連文書サマリ情報３１０３は、入力文書に関連付けられた文書レコード８０１群の自動的な解析と統計処理によって導かれる各種のサマリ情報を示すメッセージ文字列である。例えば、入力文書に関連付けられた文書レコード８０１群の解析によって、入力文書のオリジナルに相当する文書の、より新しいバージョンのオリジナル文書が検索された場合を考える。この場合は、スキャンした文書の改訂版が存在することを示唆するメッセージを表示する。また入力文書に関連付けられた文書レコード８０１群の解析により、多くの関連文書から参照されていたり、多くのジョブ処理（スキャン、印刷、送信、蓄積、検索など）の対象となっている場合がある。また、多くのメタデータ（タグなど）が付与されている文書レコード８０１が検索された場合等には、スキャンした文書よりも重要度が高い可能性がある文書の存在が示唆されていると考えられる。また入力文書に関連付けられた文書レコード８０１群の解析によって、関連文書を対象とするジョブが最近いつ行われていたかを示す情報を表示する。また、入力文書に関連付けられた文書レコード８０１群の解析によって、関連文書を対象とするジョブが最近の一定期間の間にどの程度頻繁に行われているかを示す情報を表示する。

【０２０５】

関連文書表示ボタン３１０４は、入力文書に関連付けられた文書レコード８０１群の情報を表示するための関連文書表示ウィンドウを開くためのボタンである。関連文書表示ウィンドウは、図１８に示した画面と同様に構成され、関連文書のリストを表示する。また、関連文書の関連の意味的なネットワークを、文書をノードとし関連をアークとしてグラフ表現したネットワーク図としてグラフィカルに表示することによって、ユーザによるブラウズの利便性を高めることもできる。「閉じる」ボタン３１０５は、スキャン完了ダイアログウィンドウ３２０１を閉じて元の画面表示に復帰するためのボタンである。

【０２０６】

以上説明したように本実施形態４によれば、ジョブ処理の情報に応じて擬似文書ランクを算出し、擬似文書ランクがジョブ処理対象の文書レコードの文書ランクへ配分されるようにした。このため、その文書を対象として実行されたジョブの実行のされ方に応じて、その文書の重要度をより適切に算出できる。特に、紙文書のスキャンやファクス受信のようなラスタ文書データのオフライン入力処理においても、コード文書やオンライン処理のジョブと同様に、ジョブ情報を加味した文書ランクの算出が可能となった。

【０２０７】

従って、例えば大量にコピーされている文書や、高品位に丁寧にコピーされている文書（及びその関連文書）はユーザにとって重要な文書であるという、現実世界の傾向をより反映した重要度判定が可能となった。この文書ランクに基づいて、例えば検索結果リストの表示順序等を制御することによって、ユーザが求める文書を、より迅速に見つけ出し易いシステムを提供できる。

【０２０８】

また本実施形態４によれば、ジョブ処理が実行された日時からの経過時間に応じて擬似文書ランクを繰り返し算出し、擬似文書ランクがジョブ処理対象の文書レコードの文書ランクへ配分されるように構成した。このため、その文書を対象として実行されたジョブに応じて、その文書の重要度を、より適切に算出できる。従って、例えば最近頻繁にコピー、ファクス受信、ボックス蓄積、検索ヒット等している関連文書群はユーザにとって重要な文書であるという、現実世界の傾向をより反映した重要度判定が可能となった。

【０２０９】

また本実施形態４によれば、画像処理装置においてユーザが各種文書処理を操作する際に、その文書に関連付けられたストレージ上の関連文書の存在をユーザに通知することができる。このため群集の叡智を活用しやすくなった。即ち、その文書に関する他のユーザの行動を簡単に把握できるようになった。例えば入力文書に対応するより新しいバージョンや、より注目を集めている文書があること、入力文書に対する他のユーザからの注目の度合い等を容易に把握できるようになった。

【０２１０】

（他の実施形態）

以上、本発明の実施形態について詳述したが、本発明は、複数の機器から構成されるシステムに適用しても良いし、また一つの機器からなる装置に適用しても良い。

【０２１１】

なお本発明は、前述した実施形態の機能を実現するソフトウェアのプログラムを、システム或いは装置に直接或いは遠隔から供給し、そのシステム或いは装置のコンピュータが該供給されたプログラムを読み出して実行することによっても達成され得る。その場合、プログラムの機能を有していれば、形態は、プログラムである必要はない。

【０２１２】

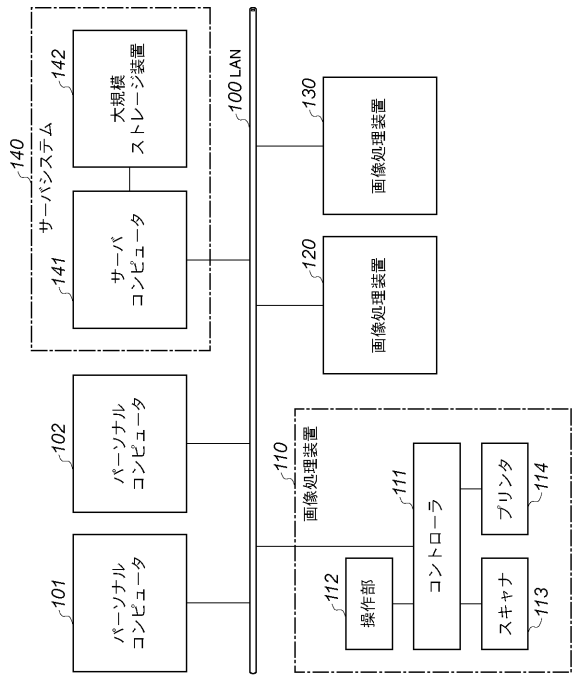
従って、本発明の機能処理をコンピュータで実現するために、該コンピュータにインストールされるプログラムコード自体も本発明を実現するものである。つまり、本発明のクレームでは、本発明の機能処理を実現するためのコンピュータプログラム自体も含まれる。その場合、プログラムの機能を有していれば、オブジェクトコード、インタプリタにより実行されるプログラム、ＯＳに供給するスクリプトデータ等、プログラムの形態を問わない。

【０２１３】
プログラムを供給するための記録媒体としては、様々なものが使用できる。例えば、フロッピー（登録商標）ディスク、ハードディスク、光ディスク、光磁気ディスク、ＭＯ、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＣＤ－ＲＷ、磁気テープ、不揮発性のメモリカード、ＲＯＭ、ＤＶＤ（ＤＶＤ－ＲＯＭ，ＤＶＤ－Ｒ）などである。
【０２１４】
その他、プログラムの供給方法としては、クライアントコンピュータのブラウザを用いてインターネットのホームページに接続し、該ホームページからハードディスク等の記録媒体にダウンロードすることによっても供給できる。その場合、ダウンロードされるのは、本発明のコンピュータプログラムそのもの、もしくは圧縮され自動インストール機能を含むファイルであってもよい。また、本発明のプログラムを構成するプログラムコードを複数のファイルに分割し、それぞれのファイルを異なるホームページからダウンロードすることによっても実現可能である。つまり、本発明の機能処理をコンピュータで実現するためのプログラムファイルを複数のユーザに対してダウンロードさせるＷＷＷサーバも、本発明のクレームに含まれるものである。
【０２１５】
また、本発明のプログラムを暗号化してＣＤ－ＲＯＭ等の記憶媒体に格納してユーザに配布する形態としても良い。その場合、所定の条件をクリアしたユーザに対し、インターネットを介してホームページから暗号化を解く鍵情報をダウンロードさせ、その鍵情報を使用することにより暗号化されたプログラムが実行可能な形式でコンピュータにインストールされるようにする。
【０２１６】
また、コンピュータが、読み出したプログラムを実行することによって、前述した実施形態の機能が実現される形態以外の形態でも実現可能である。例えば、そのプログラムの指示に基づき、コンピュータ上で稼動しているＯＳなどが、実際の処理の一部または全部を行ない、その処理によっても前述した実施形態の機能が実現され得る。
【０２１７】
更に、記録媒体から読み出されたプログラムが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書き込まれるようにしてもよい。この場合、その後で、そのプログラムの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部または全部を行ない、その処理によって前述した実施形態の機能が実現される。
【０２１８】
以上説明したように本実施形態によれば、オフラインで入力されたラスタ文書データと、その文書データに対して施された処理のメタデータを、データベースに既に登録されている文書データとメタデータとに関連付けることができる。
【０２１９】
これにより、ラスタ文書データの検索に際して、その文書データに関連するストレージ上の文書データのメタデータも活用でき、より高度な文書データの検索ができる文書データベースシステムならびに画像入力装置を提供できる。
【０２２０】
またこれにより、文書データとメタデータとそれらの関連から構成される意味的ネットワークから「群集の叡智」を導き出す際に、ラスタ文書データに対してオフラインで実施したユーザの行動も活用できるという効果がある。
【図面の簡単な説明】
【０２２１】
【図１】本発明の一実施形態に係る文書処理システムの全体構成を示すブロック図である。
【図２】本実施形態に係るサーバシステムで稼動するジョブアーカイブ・アプリケーションのソフトウェア構成を示すブロック図である。

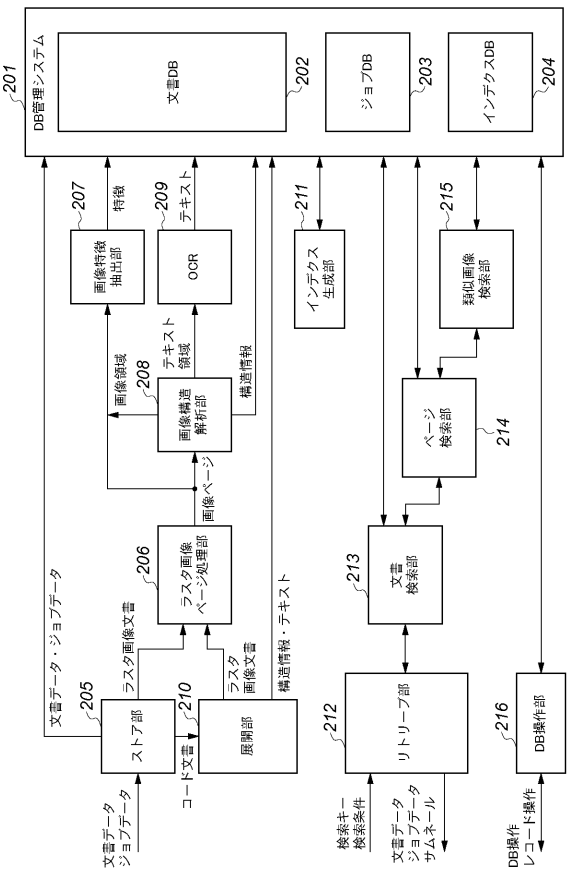
【図３】本実施形態に係る画像処理装置のハードウェア構成を示すブロック図である。
【図４】本実施形態に係る画像処理装置の外観を示す斜視図である。
【図５】本実施形態に係る画像処理装置の操作部の構成を示す平面図である。
【図６】本実施形態に係る画像処理装置の操作部及び操作部Ｉ／Ｆの構成をコントローラの構成と対応させて示すブロック図である。
【図７】本実施形態に係る画像処理装置の操作部に表示される標準的な操作画面の一例を示す図である。
【図８】本実施形態に係るＤＢ管理システムに格納される各データベースの抽象的なデータ構造を示す模式図である。
【図９】本実施形態１において、ある時点でＤＢ管理システムに格納された各データベースの具体的なデータ構造例を示すインスタンス関係図である。
【図１０】本実施形態１に係る文書処理システムの画像処理装置における文書入力処理の手順を説明するフローチャートである。
【図１１】本実施形態１において、印刷、受信、蓄積等に伴うコード文書やメタデータつき文書の文書入力処理を完了した時点でＤＢ管理システムに格納された各データベースの具体的なデータ構造例を示すインスタンス関係図である。
【図１２】本実施形態１において、紙媒体として与えられた文書のスキャンやラスタ文書データのファクス受信等による文書入力処理を完了した時点でＤＢ管理システムに格納された各データベースの具体的なデータ構造例を示すインスタンス関係図である。
【図１３】本実施形態１に係る関連レコードのインスタンス群に記録される関連情報をテーブル構造によって表現したデータ表現の一例を示す図である。
【図１４】本実施形態１に係る文書検索アプリケーションの基本画面である文書検索画面の一例を示す図である。
【図１５】本実施形態１に係る文書検索アプリケーションにおける文書検索結果リスト画面の一例を示す図である。
【図１６】本実施形態１に係る検索ヒット文書表示の一例を示す図である。
【図１７】本実施形態１に係る文書検索アプリケーションにおける注目文書の関連文書を表示する処理の手順を示すフローチャートである。
【図１８】本実施形態１に係る文書検索アプリケーションにおける注目文書に対する関連文書検索結果リストの表示結果の画面例を示す図である。
【図１９】本発明の実施形態２に係る文書検索アプリケーションで、再オンライン化された文書レコードに対して既存文書レコードからメタデータや内容データを伝播する処理の手順を示すフローチャートである。
【図２０】本実施形態２において、再オンライン化文書の文書レコードにメタデータや内容データを伝播した結果としてＤＢ管理システムに構築されるデータ構造の一例を示す図である。
【図２１】実施形態２に係る文書検索アプリケーションで、再オンライン化された文書レコードに対して既存文書レコードからメタデータや内容データを確信度に基づき伝播する処理の手順を示すフローチャートである。
【図２２】本実施形態２に係る文書検索アプリケーションにおいて、再オンライン化文書の文書レコードにメタデータや内容データを確信度付きで伝播した結果としてＤＢ管理システムに構築されるデータ構造の一例を示す図である。
【図２３】本実施形態２に係る文書検索アプリケーションにおけるキーワード検索と結果表示処理の手順を示すフローチャートである。
【図２４】本実施形態２において、複数の推定関連によって伝播したメタデータを持つ再オンライン化文書が検索結果の上位にヒットする例を示す図である。
【図２５】本実施形態３に係る関連文書の相互参照ネットワークに基づき文書ランクを決定する処理を概念的に説明するフローチャートである。
【図２６】本実施形態３に係る文書インスタンス間の関連種別に対応する、参照関係に基づく文書ランクの伝播を説明する図である。

【図２７】本実施形態３に係るＤＢ管理システムに格納された各データベースの具体的なデータ構造例において文書ランクの伝播と決定例を示すインスタンス関係図である。
【図２８】本実施形態３に係る文書検索アプリケーションにおける関連レコードのインスタンス群に記録される文書ランク伝播を伴う関連情報をテーブル構造によって表現したデータ表現の一例を示す図である。
【図２９】本実施形態３に係る文書検索アプリケーションにおけるＤＢ管理システムに格納された各データベースの具体的なデータ構造例においてジョブレコードを加味した文書ランクの伝播と決定例を示すインスタンス関係図である。
【図３０】本実施形態３に係る文書検索アプリケーションにおけるジョブレコードインスタンスに対して擬似的な文書ランクの値を決定する手順を説明するフローチャートである。
【図３１】本実施形態４に係る画像処理装置の操作部に表示される入力文書の関連文書に関する情報を表示し操作するための画面の一例を示す図である。

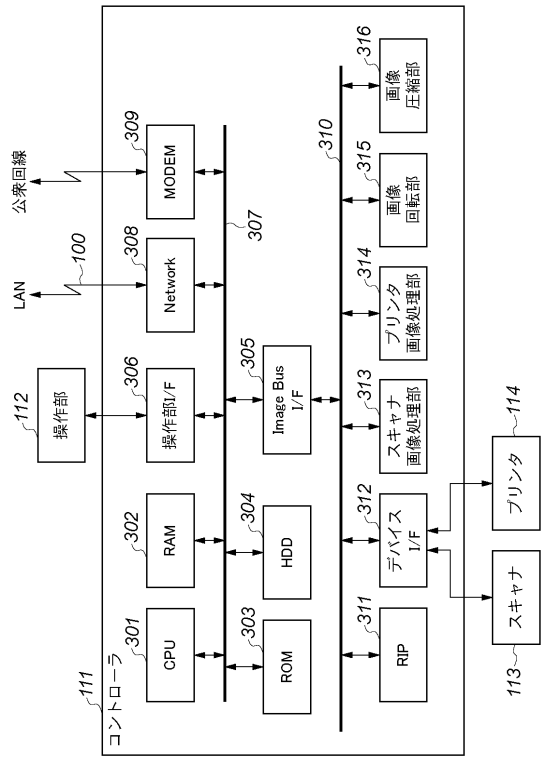
【図１】



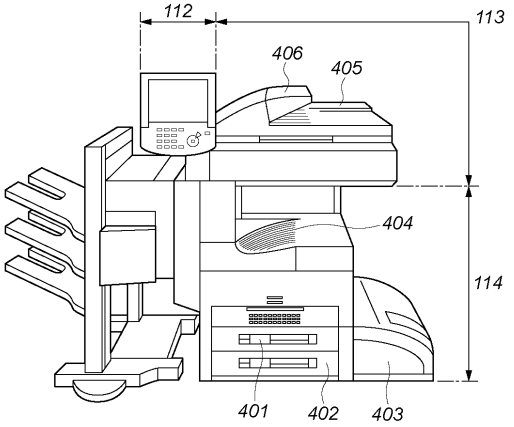
【図２】



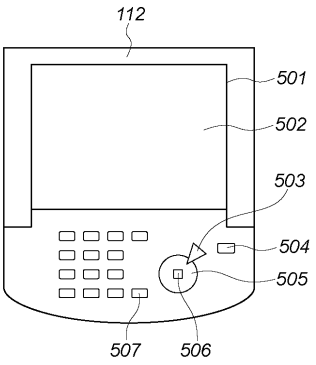
【図３】


公衆回線
LAN
100
MODEM
Network
操作部I/F
Image Bus I/F
RAM
HDD
CPU
ROM
RIP
デバイス I/F
スキャナ画像処理部
プリンタ画像処理部
画像回転部
画像圧縮部
操作部
コントローラ
プリンタ
スキャナ
307
309
308
306
305
302
304
301
303
311
312
313
314
315
316
112
111
114
113
310

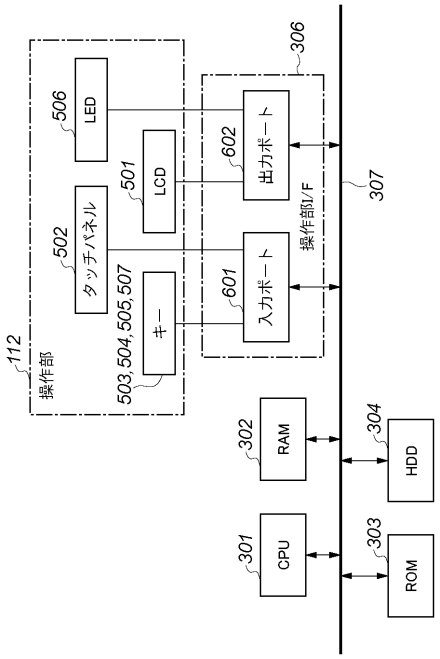
【図４】


112
113
406
405
404
114
401
402
403

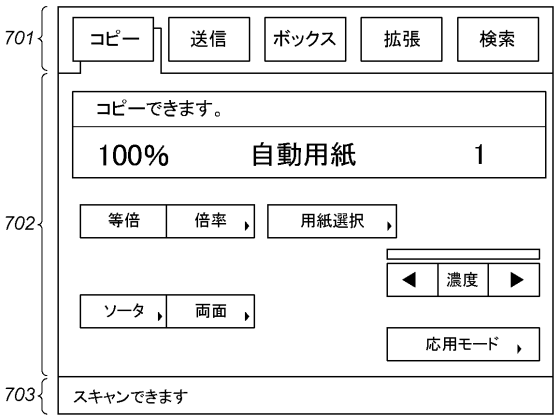
【図５】


112
501
502
503
504
505
507
506

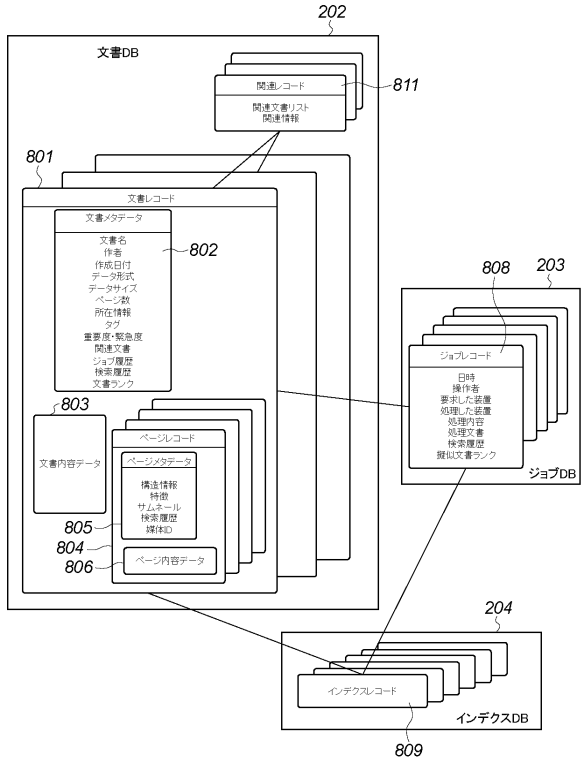
【図６】


306
LED
506
LCD
501
タッチパネル
502
キー
503,504,505,507
出力ポート
602
入力ポート
601
操作部I/F
307
112
操作部
RAM
302
304
HDD
CPU
301
303
ROM

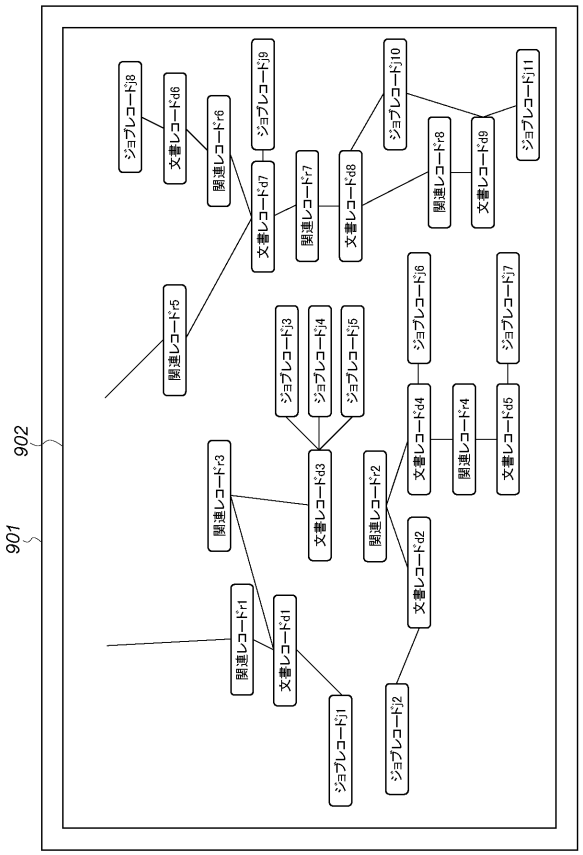
【図７】


701
コピー
送信
ボックス
拡張
検索
コピーできます。
100%
自動用紙
1
702
等倍
倍率
用紙選択
濃度
ソータ
両面
応用モード
703
スキャンできます

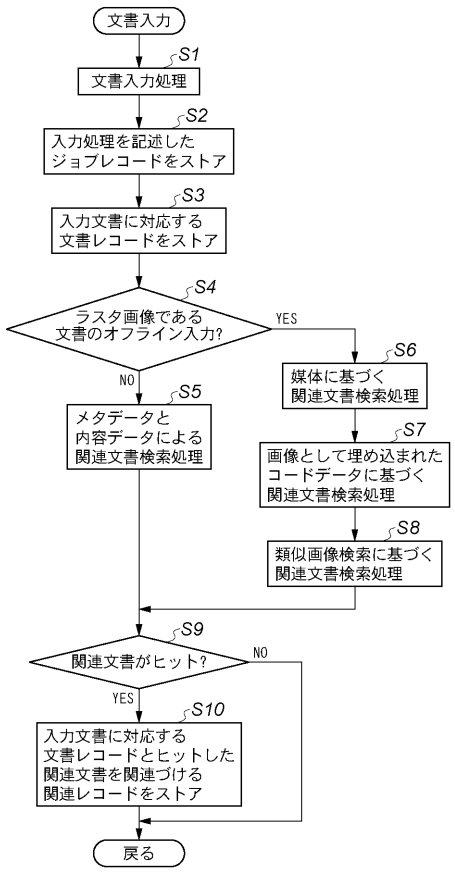
【図８】


202
文書DB
関連レコード
関連文書リスト
関連情報
811
801
文書レコード
文書メタデータ
文書名
作者
作成日付
データ形式
データサイズ
ページ数
所在情報
タグ
重要度・緊急度
関連文書
ジョブ履歴
検索履歴
文書ランク
802
808
203
ジョブレコード
日時
操作者
要求した装置
処理した装置
処理内容
処理文書
検索履歴
擬似文書ランク
ジョブDB
803
文書内容データ
ページレコード
ページメタデータ
構造情報
特徴
サムネール
検索履歴
媒体ID
805
804
806
ページ内容データ
204
インデクスレコード
インデクスDB
809

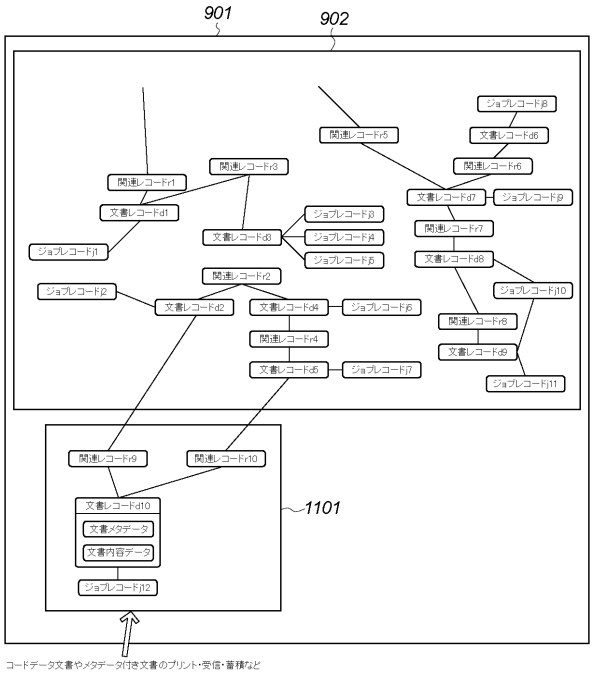
【図９】


902
901
ジョブレコードj8
文書レコードd6
関連レコードr6
ジョブレコードj9
文書レコードd7
関連レコードr7
文書レコードd8
ジョブレコードj10
関連レコードr8
文書レコードd9
ジョブレコードj11
関連レコードr5
ジョブレコードj3
ジョブレコードj4
ジョブレコードj5
ジョブレコードj6
文書レコードd4
関連レコードr4
文書レコードd5
ジョブレコードj7
関連レコードr3
文書レコードd3
関連レコードr2
文書レコードd2
関連レコードr1
文書レコードd1
ジョブレコードj1
ジョブレコードj2

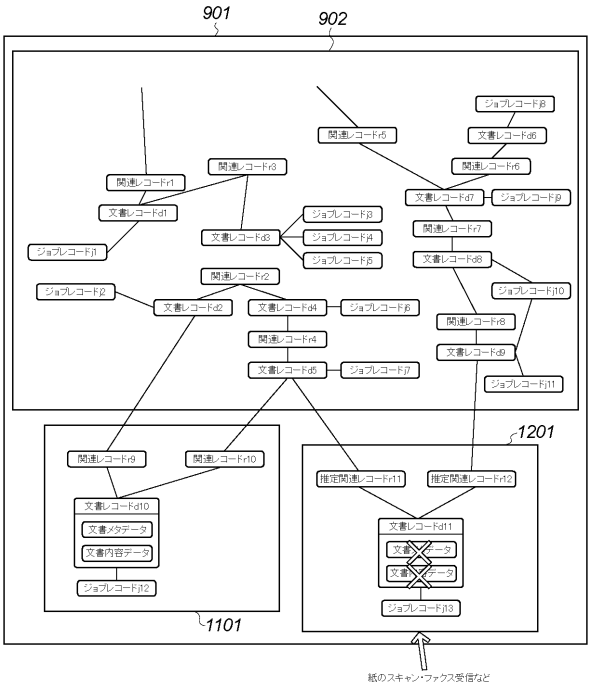
【図１０】


文書入力
S1
文書入力処理
S2
入力処理を記述した
ジョブレコードをストア
S3
入力文書に対応する
文書レコードをストア
S4
ラスタ画像である
文書のオフライン入力?
YES
NO
S5
メタデータと
内容データによる
関連文書検索処理
S6
媒体に基づく
関連文書検索処理
S7
画像として埋め込まれた
コードデータに基づく
関連文書検索処理
S8
類似画像検索に基づく
関連文書検索処理
S9
関連文書がヒット?
NO
YES
S10
入力文書に対応する
文書レコードとヒットした
関連文書を関連づける
関連レコードをストア
戻る

【図１１】


901
902
関連レコードr5
ジョブレコードj8
文書レコードd6
関連レコードr6
関連レコードr3
関連レコードr1
文書レコードd1
ジョブレコードj1
文書レコードd7
ジョブレコードj9
関連レコードr7
文書レコードd8
ジョブレコードj10
関連レコードr8
文書レコードd9
ジョブレコードj11
文書レコードd3
ジョブレコードj3
文書レコードd4
ジョブレコードj4
ジョブレコードj6
関連レコードr2
文書レコードd2
ジョブレコードj2
関連レコードr4
文書レコードd5
ジョブレコードj5
ジョブレコードj7
関連レコードr9
関連レコードr10
文書レコードd10
文書メタデータ
文書内容データ
ジョブレコードj12
1101
コードデータ文書やメタデータ付き文書のプリント・受信・蓄積など

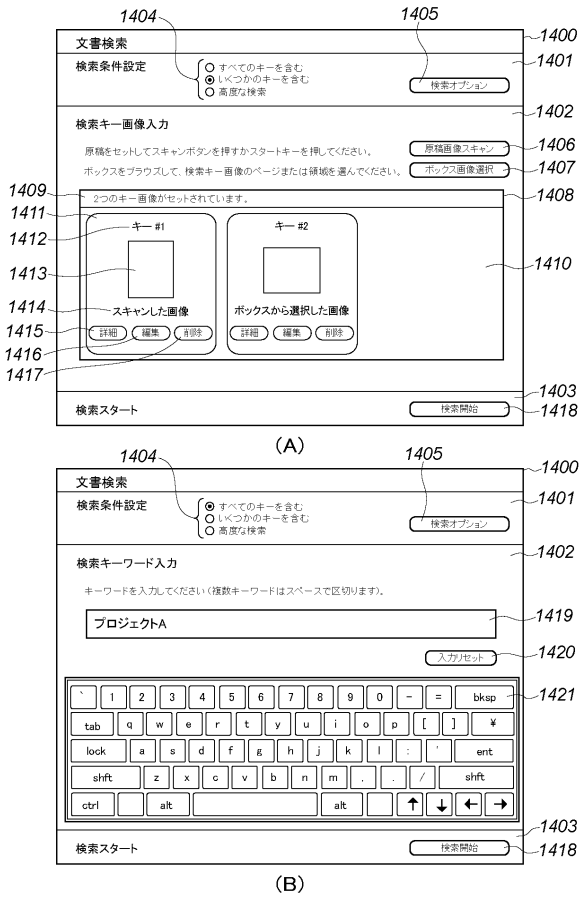
【図１２】



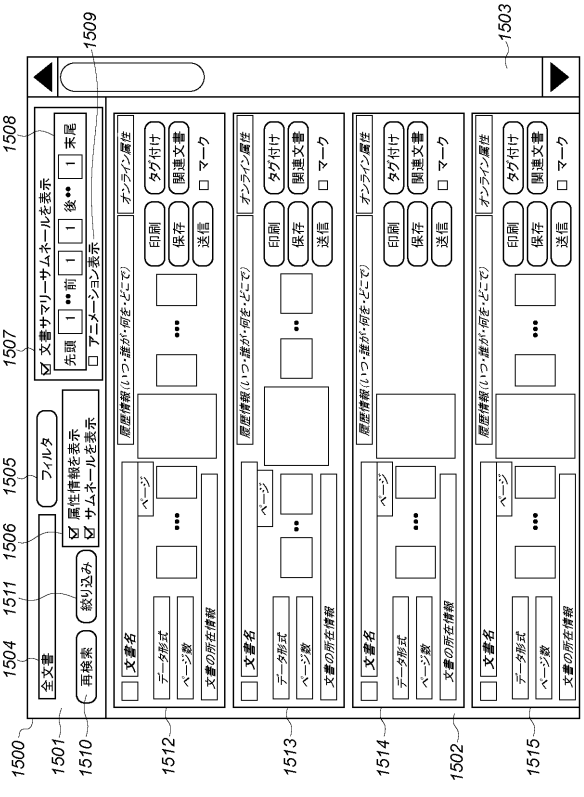
【図１３】

関連ID	参照元文書ID	参照先文書ID	関連種別	関連度
r1	d1	…	文書一致(旧版)	1
r1	…	d1	文書一致(新版)	1
r2	d2	d4	手動関連づけ(参照先)	1
r2	d4	d2	手動関連づけ(参照元)	1
r3	d1	d3	作者一致	1
r3	d3	d1	作者一致	1
r4	d4	d5	包含(含まれる)	1
r4	d5	d4	包含(含む)	1
r5	d7	…	作成日一致	1
r5	…	d7	作成日一致	1
r6	d6	d7	タグ一致	1
r6	d7	d6	タグ一致	1
r7	d7	d8	文書内容データ類似	1
r7	d8	d7	文書内容データ類似	1
r8	d8	d9	同一ジョブ処理対象	1
r8	d9	d8	同一ジョブ処理対象	1
r9	d2	d10	タグ一致	1
r9	d10	d2	タグ一致	1
r10	d5	d10	文書一致(新版)	1
r10	d10	d5	文書一致(旧版)	1
r11	d5	d11	画像類似(再オンライン化)	0.6
r11	d11	d5	画像類似(オンライン)	0.6
r12	d9	d11	媒体ID一致(再オンライン化)	0.9
r12	d11	d9	媒体ID一致(オンライン)	0.9

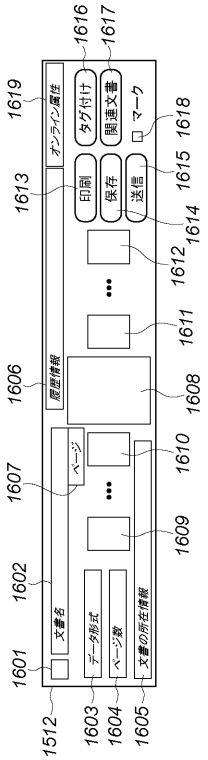
【図１４】



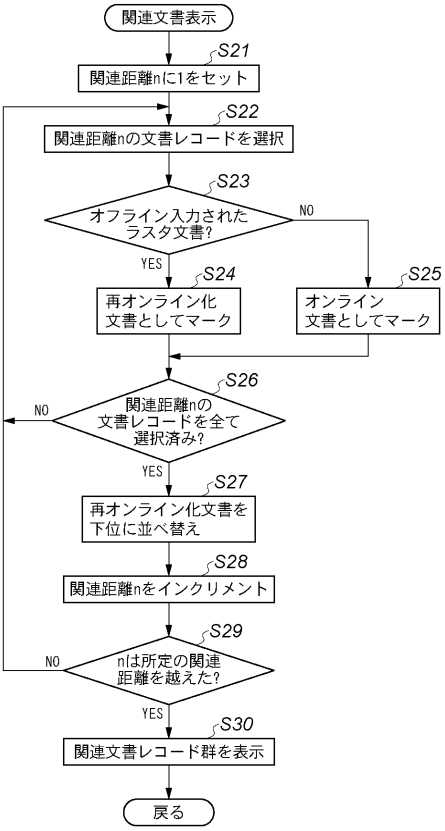
【図１５】



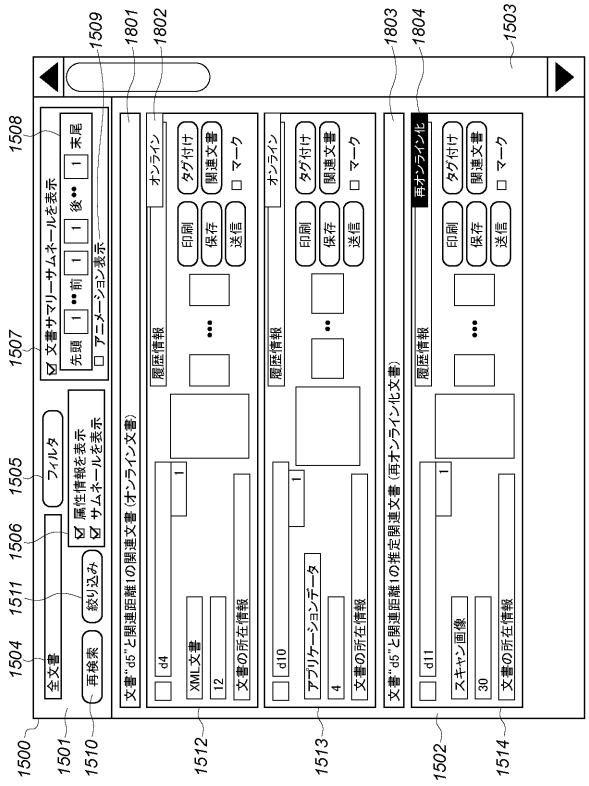
【図１６】



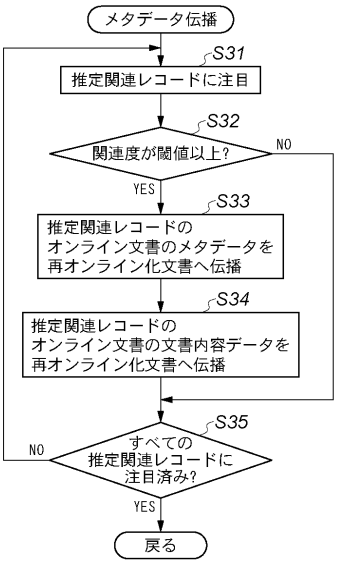
【図１７】



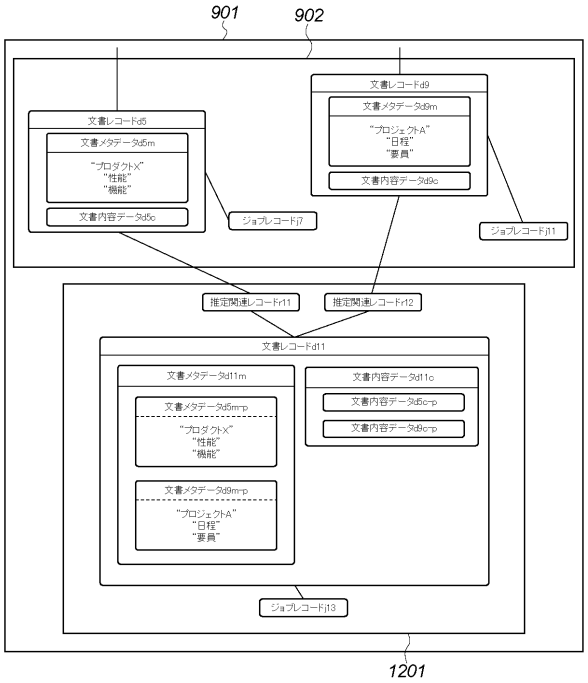
【図１８】



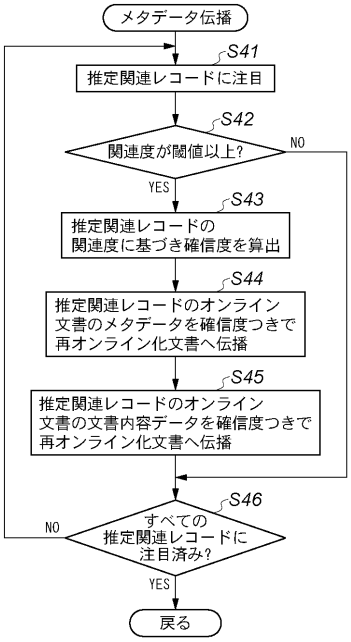
【図１９】



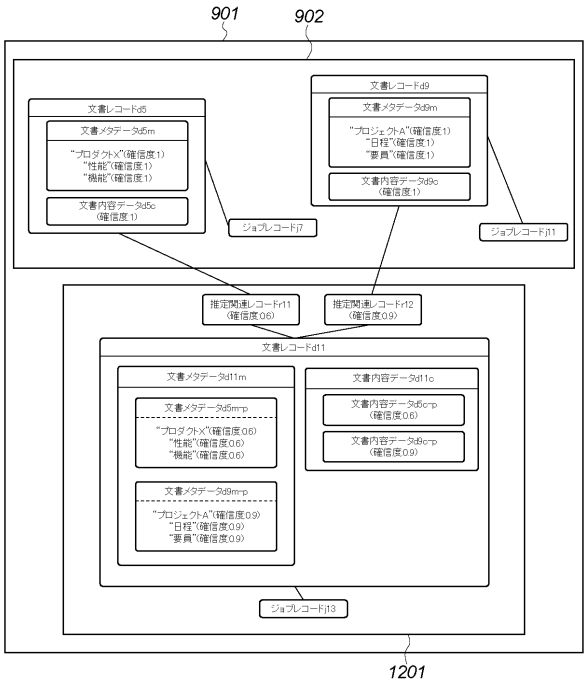
【図２０】



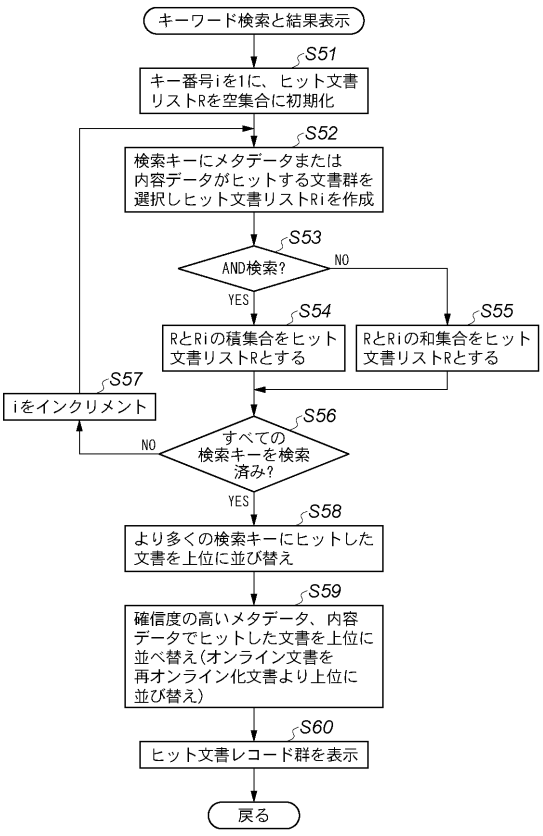
【図２１】



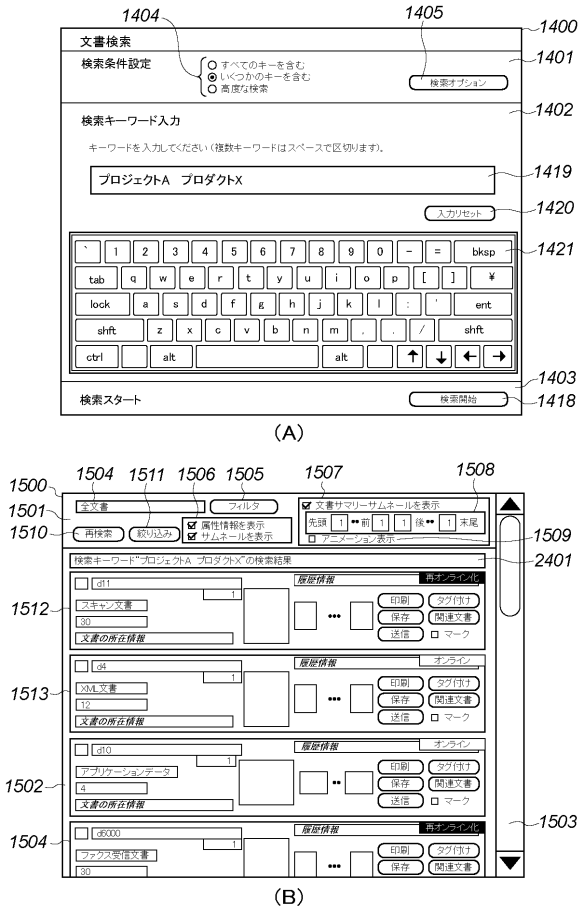
【図２２】



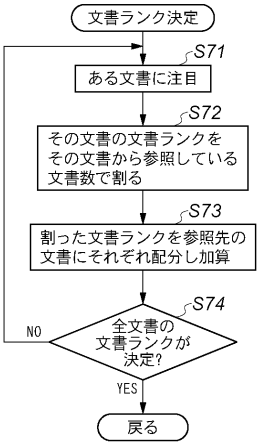
【図２３】



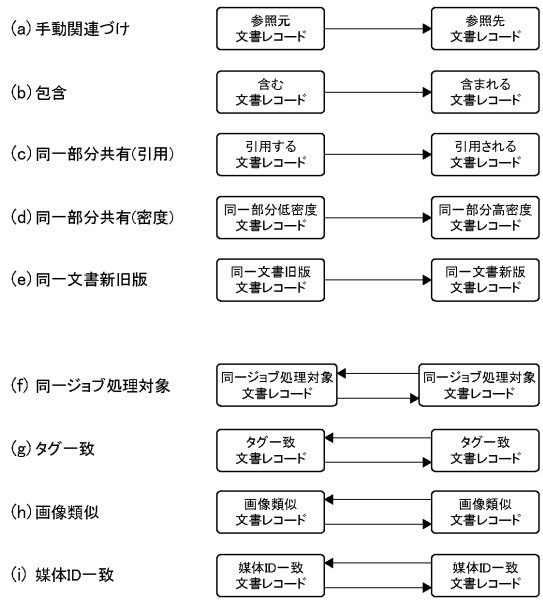
【図２４】



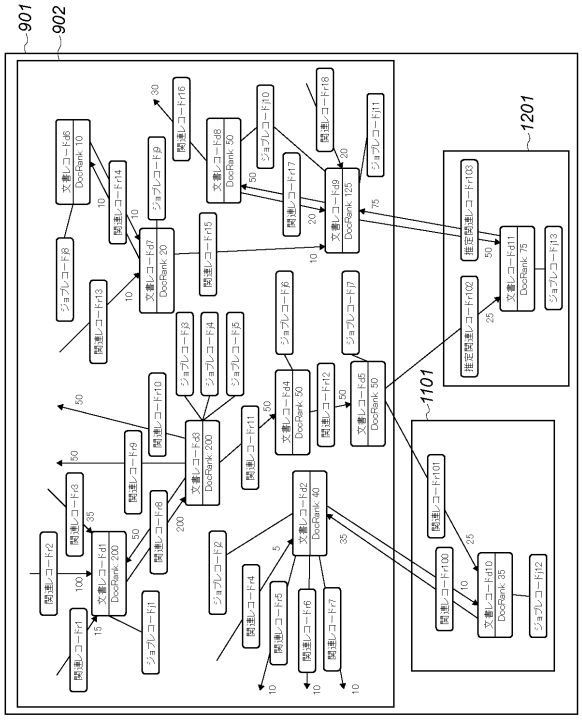
【図２５】



【図２６】



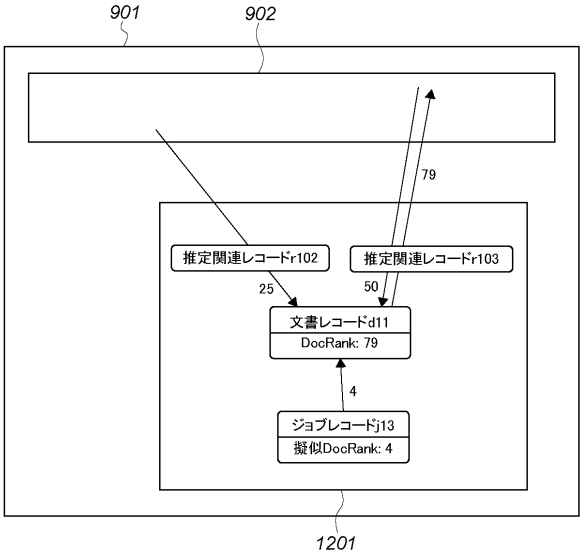
【図２７】



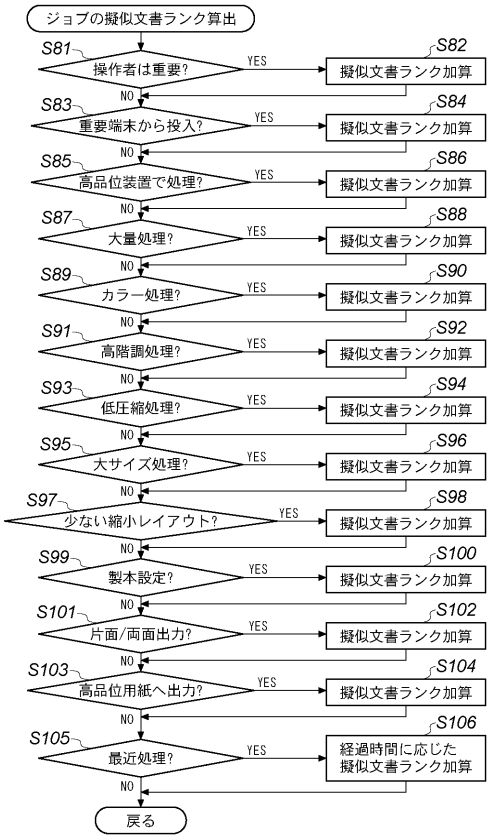
【図２８】

関連ID	参照元文書ID	参照先文書ID	関連種別	ランク伝播
r1	d1	…	手動関連づけ（参照元）	0
r1	…	d1	手動関連づけ（参照先）	1
r2	d1	…	包含（含む）	0
r2	…	d1	包含（含まれる）	1
r3	d1	…	同一部分共有（引用する）	0
r3	…	d1	同一部分共有（引用される）	1
r4	d2	…	同一部分共有（低密度）	0
r4	…	d2	同一部分共有（高密度）	1
r5	d2	…	同一文書新旧版（新版）	1
r5	…	d2	同一文書新旧版（旧版）	0
r6	d2	…	手動関連づけ（参照先）	1
r6	…	d2	手動関連づけ（参照元）	0
r7	d2	…	包含（含まれる）	1
r7	…	d2	包含（含む）	0
r8	d1	d3	同一ジョブ処理対象	1
r8	d3	d1	同一ジョブ処理対象	1
r9	d3	…	同一部分共有（引用される）	1
r9	…	d3	同一部分共有（引用する）	0
r10	d3	…	同一部分共有（高密度）	1
r10	…	d3	同一部分共有（低密度）	0
r11	d3	d4	同一文書新旧版（新版）	1
r11	d4	d3	同一文書新旧版（旧版）	0
r12	d4	d5	手動関連づけ（参照先）	1
r12	d5	d4	手動関連づけ（参照元）	0
…	…	…	…	…
r100	d2	d10	画像類似	1
r100	d10	d2	画像類似	1
r101	d5	d10	同一内容共有（高密度）	1
r101	d10	d5	同一内容共有（低密度）	0
r102	d5	d11	包含（含まれる）	1
r102	d11	d5	包含（含む）	0
r103	d9	d11	媒体ID一致	1
r103	d11	d9	媒体ID一致	1

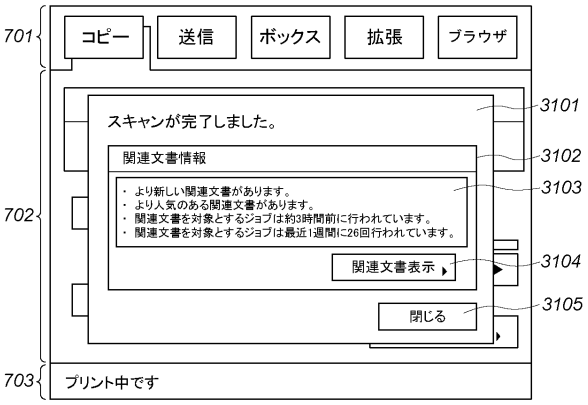
【図２９】



【図３０】



【図３１】



フロントページの続き

(72)発明者　山本　雅仁
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

審査官　村松　貴士

(56)参考文献　特開平１１－１４３９０８（ＪＰ，Ａ）
特開２００５－２７５８４７（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　　１／００
Ｇ０６Ｆ　１７／３０