(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4311552号
(P4311552)

(45)発行日　平成21年8月12日(2009.8.12)　(24)登録日　平成21年5月22日(2009.5.22)

(51)Int.Cl.　F I
G06T 1/00 (2006.01)　G06T 1/00 200C

請求項の数 17　(全 23 頁)

(21)出願番号	特願2004-47112(P2004-47112)
(22)出願日	平成16年2月23日(2004.2.23)
(65)公開番号	特開2005-182730(P2005-182730A)
(43)公開日	平成17年7月7日(2005.7.7)
審査請求日	平成18年10月2日(2006.10.2)
(31)優先権主張番号	10/742,131
(32)優先日	平成15年12月19日(2003.12.19)
(33)優先権主張国	米国(US)

(73)特許権者　505149170
コファックス,　インコーポレイテッド
アメリカ合衆国　カリフォルニア　９２６１８－３６０３,アーバイン,ラグーナ　キャニオン　ロード　１６２４５
(74)代理人　100078282
弁理士　山本　秀策
(74)代理人　100062409
弁理士　安村　高明
(74)代理人　100113413
弁理士　森下　夏樹

最終頁に続く

(54)【発明の名称】ドキュメントの自動分離

(57)【特許請求の範囲】
【請求項１】
コンピュータベースの方法であって、
該方法は、
カテゴリの分類ルールに従って複数のドキュメントイメージを複数の所定のカテゴリに自動的にカテゴリ化することを包含し、
該カテゴリ化することは、
該複数のドキュメントイメージの複数の可能なカテゴリ化シーケンスの各々について、該複数のドキュメントイメージの各ドキュメントイメージに対してそれぞれの出力スコアを生成することであって、該それぞれの出力スコアは、該ドキュメントイメージがそれぞれのカテゴリ化シーケンスのそれぞれのカテゴリに属することに関する情報をエンコードし、該それぞれの出力スコアは、各ドキュメントイメージの条件付き確率に基づいて計算され、該条件付き確率は、それぞれのカテゴリ化シーケンスにおける少なくとも１つの先行するドキュメントイメージまたは後続するドキュメントイメージの少なくとも１つのカテゴリと、各ドキュメントイメージのグラフィック情報およびテキスト情報のうちの少なくとも１つとに依存する、ことと、
探索アルゴリズムを用いて、それぞれのカテゴリ化シーケンスのドキュメントイメージの出力スコアに基づいて、該複数の可能なカテゴリ化シーケンスから最高の確率を有するカテゴリ化シーケンスを決定することと
により実行され、

該方法は、
該最高の確率を有すると決定されたカテゴリ化シーケンスの各ドキュメントイメージのカテゴリ化に基づいて、該複数のドキュメントイメージのうちのどれがどのカテゴリに属するのかを識別する少なくとも１つの識別子を自動的に生成することをさらに包含する、方法。
【請求項２】
前記自動的にカテゴリ化するステップは、機械学習法を用いる、請求項１に記載の方法。
【請求項３】
前記自動的にカテゴリ化するステップは、手動で入力されたユーザ特有の分類ルールを用いる、請求項１または２に記載の方法。
【請求項４】
前記少なくとも１つの識別子は、コンピュータにより生成された分離ページであって、前記複数のカテゴリのうちの異なったカテゴリに属する連続したイメージを分離するようにドキュメントイメージ間に挿入された分離ページを含む、請求項１に記載の方法。
【請求項５】
前記少なくとも１つの識別子は、ＸＭＬメッセージのフォーマットである、請求項１に記載の方法。
【請求項６】
前記少なくとも１つの識別子は、少なくとも１つのコンピュータにより生成されたラベルであって、前記複数のドキュメントイメージのうちの少なくとも１つに挿入されたラベルを含む、請求項１に記載の方法。
【請求項７】
前記複数のカテゴリは、金融取引において用いられる少なくとも２つの異なったフォームタイプを含む、請求項１～６のいずれか一項に記載の方法。
【請求項８】
前記複数のカテゴリは、前記少なくとも２つの異なったフォームタイプの各々について、フォームタイプの最初のページを示す最初のページカテゴリ、フォームタイプの中間のページを示す中間のページカテゴリ、およびフォームタイプの最終のページを示す最後のページカテゴリをさらに含む、請求項７に記載の方法。
【請求項９】
前記探索アルゴリズムは、全ての可能なカテゴリ化シーケンスおよび対応する確率を表すグラフ構造に適用されるグラフ探索アルゴリズムである、請求項１に記載の方法。
【請求項１０】
前記グラフ探索アルゴリズムを用いるステップは、
前記グラフ構造を用いることにより、前記全ての可能なカテゴリ化シーケンスの各々について、前記複数のドキュメントイメージの各々に対する前記出力スコアに基づいて、合計出力スコアを計算することと、
該グラフ探索アルゴリズムを適用することにより、どのカテゴリ化シーケンスが最高の合計出力スコアをもたらすかを決定することと
を包含する、請求項９に記載の方法。
【請求項１１】
前記グラフ構造は、有限状態変換器を用いてインプリメントされ、前記複数のドキュメントイメージは、該有限状態変換器の入力であり、前記複数のカテゴリ化シーケンスは、該有限状態変換器の出力である、請求項１０に記載の方法。
【請求項１２】
前記グラフ構造は、重みつき有限状態変換器を用いてインプリメントされ、該重みつき有限状態変換器は、前記複数のドキュメントイメージを入力として有し、前記複数のカテゴリ化シーケンスを出力として有し、該入力および該出力のうちの少なくとも１つは、重みつき値と関連する、請求項１１に記載の方法。

【請求項１３】
前記自動的にカテゴリ化するステップは、手動で生成されたルールであって、少なくとも１つの可能なカテゴリ化シーケンスを除去するルールを適用することをさらに包含する、請求項１０～１２のいずれか一項に記載の方法。
【請求項１４】
前記少なくとも１つの可能なカテゴリ化シーケンスは、第１のドキュメントタイプの最後のページが識別される前に第２のドキュメントタイプの最初のページが該第１のドキュメントタイプの最初のページの後に続くというシーケンスを含む、請求項１３に記載の方法。
【請求項１５】
前記少なくとも１つの可能なカテゴリ化シーケンスは、所定の数のページを有することが知られる第１のドキュメントタイプの２つの連続的ページのシーケンスを含む、請求項１３または１４に記載の方法。
【請求項１６】
前記複数のドキュメントイメージは、
銀行ローンドキュメントと、
保険フォームと、
納税フォームと、
雇用フォームと、
健康管理フォームと、
請求書フォームと
のうちの複数個を表す、請求項１～１５のいずれか一項に記載の方法。
【請求項１７】
プログラムが記録されたコンピュータ読み取り可能な格納媒体であって、該プログラムは、該プログラムがコンピュータ上で実行されたときに該コンピュータに請求項１～１６のいずれか一項に記載の方法を実行させる、コンピュータ読み取り可能な格納媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
（発明の分野）
本発明は、デジタルスキャナによって生成されたドキュメントまたはサブドキュメント等のイメージのシーケンスにおける境界を効果的に見つけるシステムおよび方法に関する。
【背景技術】
【０００２】
（関連技術の説明）
ドキュメントまたはサブドキュメントの境界を見つけることは、大量のドキュメントおよび／またはサブドキュメントをそれらのドキュメントまたはドキュメントタイプに従って処理するという意味合いで有用である。本明細書中で用いられるように、「ドキュメント」という用語は、通常、開始部分の境界（例えば、第１ページ、第１パラグラフ等）および終わり部分の境界（例えば、最後のページ、最後のパラグラフ等）を有する媒体に含まれる情報に関し、「サブドキュメント」は、「ドキュメント」（例えば、ページ（単数または複数）、セクション（単数または複数）、パラグラフ（単数または複数）等）に含まれる任意の定義可能な情報のサブセットであり得る。以下において、「ドキュメント」および「サブドキュメント」は、集合的に「ドキュメント」と呼ばれる。
【０００３】
大量デジタルドキュメントの走査、および、これに続くドキュメントの処理のために通常用いられる現在の方法は、例えば、Ｒａｏによる特許文献１に記載されるように、ドキュメントを分類するために、物理的セパレータシートを用いるステップを包含する。大量の走査作業において、走査の前に物理的セパレータページを挿入するという手動の労力は

、著しく費用がかかり、かつ時間を要し得る。例えば、米国の大規模なローン処理会社は、現在、１ヶ月に２０００万件のローンイメージを処理するために、セパレータページの印刷に、１年間に１００万ドルを費やすと推定する。さらに、これらのローン処理会社は、ローンドキュメントごとに、少なくとも２０秒の手動の労力を推定する。従って、セパレータページを用いることが、全ドキュメント作成費の実質的な部分を占め得、この労力のレベルは、処理されるフォームの量とともに直線的に拡大縮小する。

【０００４】

同様の量が与えられた場合、人間によって構築されたルールに基づいたシステム（ここで、分類および／または分離ルールは、人間のオペレータによって定められる）は、特定の種類のタスクにとっては首尾よく働く。しかしながら、このようなルールに基づくシステムのコストは、処理されるドキュメントの数とともに直線的に増減せず、ドキュメントのタイプと業務ルールとの組み合わせの数が増加すると、より不十分な増減になりさえし得る。これは、時間とともに、システムが新しい制約に適応することを強いられ、かつ、新旧のルール間での対応の相互関係が正確であることを保証することが厄介であり、時間を要し、かつ高度に熟練した（従って、高価な）労働力を必要とし得る。

【０００５】

ごく最近、ルールの生成の処理を自動化する研究が行われた。非特許文献１（以後、「Ｃｏｌｌｉｎｓ－Ｔｈｏｍｐｓｏｎ」」）に記載される研究は、特定の順序でないページを有する一括のドキュメントをとり入れ、かつ、同じドキュメントからのページを共に、自動的にグループ分けする。この研究は、３ステップの方法を用いる。第１に、ページの各ペアに、ドキュメント構造情報、テキストレイアウト情報、テキストの類似性、および一般的イメージコンテンツフィーチャに基づいた４つの類似性スコアが割り当てられる。これらのスコアは、その後、２つのページ間の全体的類似性を計算するために用いられる。最後に、ページのペアは、全部が互いに類似であるページのより大きいグループを得るために、類似性スコアによってクラスタ形成される。その結果、複数のドキュメントから、大きいページのセットのドキュメントが分離される。

【０００６】

Ｃｏｌｌｉｎｓ－Ｔｈｏｍｐｓｏｎによって提示される方法は、ページをドキュメントに対応するグループに区分し、これは、どのタイプのドキュメントが集合の中に存在するかを識別することを試みない。しかしながら、このアプローチは、ビジネス全体の問題に対処するには及ばない。あるドキュメントが開始し、別のドキュメントが終了する場所をコンピュータに教示し、かつ、セパレータページの後に続くドキュメントのタイプを識別するために、かなり頻繁に、ドキュメント間にセパレータページが挿入される。情報の両方の情報部分が、特定のビジネスプロセスを強化するために重要である。このドキュメントのタイプの識別は、特定のドキュメントに対してどのさらなる処理が行われることが必要であるかを決定するために用いられる。以下の実施例は、両方のステップを実行することの価値を示す。

【０００７】

抵当リファイナンス会社は、ローンリファイナンス申し込みのドキュメントの作成を自動化することを所望する。作成プロセスは、今日、各ドキュメント間にバーコードセパレータを挿入するステップを包含する。セパレータは、１つのドキュメントが開始および終了する場所をコンピュータに教示する。バーコードは、どのドキュメントタイプがセパレータの後にあるかをコンピュータに教示する。ドキュメントタイプに基づいて、自動化された抽出およびルーティング技術が、各ドキュメントからの正確な情報を抜き取り得る。以前は、このすべての作業が手動で行われなければならなかった。ドキュメントタイプを識別しない場合、技術による節約が大幅に低減される。ドキュメントは、分離されるが、識別されない。人間のオペレータは、同定するために、各ドキュメントに目を通す必要がある。このプロセスは、各ドキュメントに目を通し、バーコードセパレータページを挿入するくらい長い。

【０００８】

さらに、Ｃｏｌｌｉｎｓ－Ｔｈｏｍｐｓｏｎによって記載されたシステムは、特定の基準（ページは、同じドキュメントからのものである）に従ってドキュメントを互いに分離するように構築された。しかしながら、ビジネスプロセスのグループ分け基準を再定義することは有用であり得る。例えば、納税申告用紙から証書を分けることは、１つの分離タスクであり得る。別のビジネスプロセスにおいて、１人の個人に属するすべてのフォームを識別することは、所望される分離タスクであり得る。Ｃｏｌｌｉｎｓ－Ｔｈｏｍｐｓｏｎにおいて用いられる方法は、システムのユーザが、類似であるとは何を意味するかを容易に定義すること、従って、分離タスクを再定義することを可能にしない。その代わりに、ユーザは、分類を再プログラムし、システムでクラスタ形成すること、および、システムが入力として用いるドキュメントから用いられるフィーチャを設計し直すことを必要とする。

【特許文献１】米国特許第６，１１８，５４４号明細書

【非特許文献１】Ｃｏｌｌｉｎｓ－Ｔｈｏｍｐｓｏｎら、“Ａ Ｃｌｕｓｔｅｒｉｎｇ－Ｂａｓｅｄ Ａｌｇｏｒｉｔｈｍ ｆｏｒ Ａｕｔｏｍａｔｉｃ Ｄｏｃｕｍｅｎｔ Ｓｅｐａｒａｔｉｏｎ”、ＡＣＭ Ｓｐｅｃｉａｌ Ｉｎｔｅｒｅｓｔ Ｇｒｏｕｐ ｉｎ Ｉｎｆｏｒｍａｔｉｏｎ Ｒｅｔｒｉｅｖａｌ、２００２年

【発明の開示】

【発明が解決しようとする課題】

【０００９】

本発明は、ドキュメントの境界の線引き、およびドキュメントタイプの識別を、コンピュータベースのシステムにおいて達成することを課題とする。

【課題を解決するための手段】

【００１０】

（発明の簡単な要旨）

本発明により、コンピュータベースのシステムにおいて、ドキュメントの境界に線引きし、かつドキュメントタイプを識別する方法であって、カテゴリの分類ルールにより、複数のドキュメントイメージを複数の所定のカテゴリに自動的にカテゴリ化するステップと、該複数のドキュメントイメージのどれが、少なくとも２つのカテゴリのどれに属するかを識別する少なくとも１つの識別子を自動的に生成するステップとを包含する、方法が提供され、それにより、上記目的を達成する。

【００１１】

前記少なくとも１つの識別子は、前記複数のカテゴリのうちの異なったカテゴリに属するイメージに線引きするために、ドキュメントイメージ間に挿入されるコンピュータによって生成された分離ページを備えてもよい。

【００１２】

前記少なくとも１つの識別子は、前記複数のデジタルイメージのカテゴリ化により、該複数のデジタルイメージのカテゴリ化シーケンスを識別するコンピュータ可読記述を含んでもよい。

【００１３】

前記コンピュータ可読記述は、ＸＭＬメッセージを含んでもよい。

【００１４】

前記少なくとも１つの識別子は、前記複数のドキュメントイメージの少なくとも１つと電子的に関連付けられた、少なくとも１つのコンピュータによって生成されたラベルを含んでもよい。

【００１５】

前記複数のカテゴリは、金融取引において用いられる少なくとも２つの異なったフォームタイプを含んでもよい。

【００１６】

前記複数のカテゴリは、前記少なくとも２つの異なったフォームタイプごとに、最初、中間、および最後のページカテゴリをさらに含んでもよい。

【００１７】
前記自動的にカテゴリ化するステップは、ドキュメントイメージごとに出力スコアを生成するステップと、前記複数のドキュメントイメージの複数の可能なカテゴリ化シーケンスから、該出力スコアに基づいて最適カテゴリ化シーケンスを決定するためにグラフ探索アルゴリズムを用いるステップとを包含してもよい。
【００１８】
前記出力スコアは、各ドキュメントイメージが、前記複数のカテゴリからの少なくとも１つのそれぞれのカテゴリに属する確率を表してもよい。
【００１９】
前記グラフ探索アルゴリズムを用いるステップは、前記可能なカテゴリ化シーケンスごとに、前記複数のドキュメントイメージごとの前記出力スコアに基づいて、合計出力スコアを計算するためにグラフ構造を用いるステップと、どのカテゴリ化シーケンスが最高の合計出力スコアをもたらすかを決定するステップとを包含してもよい。
【００２０】
前記グラフ構造は、有限状態変換器を用いて実現され、前記複数のドキュメントイメージは、入力を含み、前記複数のカテゴリ化シーケンスは、出力を含んでもよい。
【００２１】
前記グラフ構造は、前記複数のドキュメントイメージを入力として、および前記複数のカテゴリ化シーケンスを出力として有する重みつき有限状態変換器を用いて実現され、各カテゴリ化シーケンスは、前記出力スコアに基づいて、重みつき値をそれらと関連付けてもよい。
【００２２】
前記出力スコアは、各ドキュメントイメージが、前記複数のカテゴリからの少なくとも１つのそれぞれのカテゴリに属する条件付確率を表し、ドキュメントイメージごとの該条件付確率は、少なくとも１つの先行するか、または後続のドキュメントイメージについて選択される少なくとも１つのカテゴリに依存してもよい。
【００２３】
前記自動的にカテゴリ化するステップは、少なくとも１つの可能なカテゴリ化シーケンスを除去する、手動で生成された分類ルールを適用するステップをさらに包含してもよい。
【００２４】
前記少なくとも１つの可能なカテゴリ化シーケンスは、第１のドキュメントタイプの最後のページが識別される前に、第２のドキュメントタイプの最初のページによって追従される第１のドキュメントタイプの最初のページを含んでもよい。
【００２５】
前記少なくとも１つの可能なカテゴリ化シーケンスは、所定の数のページを有することが知られる第１のドキュメントタイプの２つの連続的ページを含んでもよい。
【００２６】
本発明により、コンピュータによって実行される場合、ドキュメントの境界に線引きし、かつ、ドキュメントタイプを識別する方法を実行する、コンピュータによって実行可能な命令を格納するコンピュータ可読媒体であって、該方法は、カテゴリの分類ルールにより、複数のドキュメントイメージを複数の所定のカテゴリに自動的にカテゴリ化するステップと、該複数のドッキュメントイメージのどれが、該少なくとも２つのカテゴリのどれに属するかを識別する少なくとも１つの識別子を自動的に生成するステップとを包含してもよい。
【００２７】
前記複数のカテゴリのうちの異なったカテゴリに属するイメージに線引きするために、前記少なくとも１つの識別子は、ドキュメントイメージ間に挿入されるコンピュータ可読分離ページを含んでもよい。
【００２８】

前記少なくとも１つの識別子は、前記複数のデジタルイメージのカテゴリ化により、該複数のデジタルイメージのカテゴリ化シーケンスを識別するコンピュータ可読記述を含んでもよい。

【００２９】

前記コンピュータ可読記述は、ＸＭＬメッセージを含んでもよい。

【００３０】

前記少なくとも１つの識別子は、前記複数のドキュメントイメージの少なくとも１つと電子的に関連した少なくとも１つのコンピュータによって生成されたラベルを備えてもよい。

【００３１】

前記複数のカテゴリは、金融取引において用いられる少なくとも２つの異なったフォームタイプを含んでもよい。

【００３２】

前記複数のカテゴリは、前記少なくとも２つの異なったフォームタイプごとに最初、中間、および最後のページカテゴリをさらに含んでもよい。

【００３３】

前記自動的にカテゴリ化するステップは、ドキュメントイメージごとに出力スコアを生成するステップと、前記出力スコアに基づいて、前記複数のドキュメントイメージごとの複数の可能なカテゴリ化シーケンスからの最適なカテゴリ化シーケンスを決定するためにグラフ探索アルゴリズムを用いるステップとを包含してもよい。

【００３４】

前記出力スコアは、各ドキュメントイメージが前記複数のカテゴリからの少なくとも１つのそれぞれのカテゴリに属する確率を表してもよい。

【００３５】

前記グラフ探索アルゴリズムを用いるステップは、前記可能なカテゴリ化シーケンスごとに、前記複数のドキュメントイメージごとの前記出力スコアに基づいて、合計出力スコアを計算するためにグラフ構造を用いるステップと、どのカテゴリ化シーケンスが最高の合計出力スコアをもたらすかを決定するステップとを包含してもよい。

【００３６】

前記グラフ構造は、有限状態変換器を用いて実現され、前記複数のドキュメントイメージは、入力を含み、前記複数のカテゴリ化シーケンスは、出力を含んでもよい。

【００３７】

前記グラフ構造は、前記複数のドキュメントイメージを入力として、および前記複数のカテゴリ化シーケンスを出力として有する重みつき有限状態変換器を用いて実現され、各カテゴリ化シーケンスは、該出力スコアに基づいて、重みつき値とこれらを関連付けてもよい。

【００３８】

前記出力スコアは、各ドキュメントイメージが、前記複数のカテゴリからの少なくとも１つのそれぞれのカテゴリに属する条件付確率を表し、各ドキュメントイメージごとの該条件付確率は、少なくとも１つの先行するか、または後続のドキュメントイメージについて選択された少なくとも１つのカテゴリに依存してもよい。

【００３９】

前記自動的にカテゴリ化するステップは、少なくとも１つの可能なカテゴリ化シーケンスを除去する手動で生成された分類ルールを適用するステップをさらに包含してもよい。

【００４０】

前記少なくとも１つの可能なカテゴリ化シーケンスは、第１のドキュメントタイプの最後のページが識別される前に、第２のドキュメントタイプの最初のページによって追従される該第１のドキュメントタイプの最初のページを含んでもよい。

【００４１】

前記少なくとも１つの可能なカテゴリ化シーケンスは、所定の数のページを有すること

が知られた第１のドキュメントタイプの２つの連続的ページを含んでもよい。

【００４２】

前記少なくとも１つのドキュメントは、複数の銀行ローンドキュメント（ｂａｎｋ　ｌｏａｎ　ｄｏｃｕｍｅｎｔ）を含んでもよい。

【００４３】

前記少なくとも１つのドキュメントは、複数の保険フォーム（ｉｎｓｕｒａｎｃｅ　ｆｏｒｍ）を含んでもよい。

【００４４】

前記少なくとも１つのドキュメントは、複数の納税フォーム（ｔａｘ　ｆｏｒｍ）を含んでもよい。

【００４５】

前記少なくとも１つのドキュメントは、複数の雇用フォーム（ｅｍｐｌｏｙｍｅｎｔ　ｆｏｒｍ）を含んでもよい。

【００４６】

前記少なくとも１つのドキュメントは、複数の健康管理フォーム（ｈｅａｌｔｈｃａｒｅ　ｆｏｒｍ）を含んでもよい。

【００４７】

前記少なくとも１つのドキュメントは、複数の請求書フォーム（ｉｎｖｏｉｃｅ　ｆｏｒｍ）を含んでもよい。

【００４８】

本発明は、デジタル走査した後で、ドキュメントまたはサブドキュメントの分離および識別に伴う手動の労力を低減する方法およびシステムを提供する。特に、本方法およびシステムは、手動で構築されたルールに基づくシステムのように、ほとんどの入来するドキュメントを自動的に処理するが、さらに、システムのセットアップ、メンテナンスおよび拡張に伴う構成時間を著しく低減するという利益をももたらす。ある実施形態において、これは、ドキュメントおよび／またはサブドキュメントを分離するために用いられるルールを自動的に構築する監視付き（ｓｕｐｅｒｖｉｓｅｄ）機械学習法を用いることによって達成される。

【００４９】

さらなる実施形態において、本発明は、テキストおよびイメージのイメージ分類を適用し、これらの結果を、ルールベースのフレームワークで組み合わせ、これにより、分離の最も見込みのある構成が、容易に構成可能な制約のセットのもとで見出され得るシステムおよび方法を提供する。

【００５０】

別の実施形態において、本発明は、高品質の分離を自動的に生成するために、確率的ネットワークを用いる。確率的ネットワークは、原則に基づいて、情報の複数のソースを組み合わせ得、当業者は、すべての利用可能な情報から最も見込みのある分離を推論するために、公知の推論アルゴリズムを用い得る。情報の例示的ソースは、サブドキュメントシーケンス情報、各イメージにおけるグラフィカル情報、各イメージにおけるテキスト情報、ドキュメントおよび／またはサブドキュメントの頻度の分布、ドキュメントおよび／またはサブドキュメントの長さの分布、ならびにビジネスプロセスルールを含むが、これらに限定されない。単一の確率フレームワークに組み込まれるべき情報の種々のソースは、確率推定、およびなされた独立性の仮定を明確にするネットワーク構造の構築を必要とする。これらの仮定は、ネットワークにおける各確率の推定および推論の間に利用可能な情報を定義する。

【００５１】

ある実施形態において、各イメージのグラフィカル情報は、イメージのドキュメントまたはサブドキュメントタイプを予測する分類ルールを学習するために、機械学習アルゴリズムによって用いられる。別の実施形態において、機械学習アルゴリズムは、光学式文字認識（ＯＣＲ）によって取得されたイメージにおけるテキスト情報に基づいて、イメージ

ごとに分類ルールを学習する。さらに、別の実施形態は、２つのそのような分類器の出力を組み合わせ得、かつ、これらから単一の出力スコアを生成し得る。別の実施形態において、これらのフィーチャの２つのセットは、１つのフィーチャ空間、およびドキュメントまたはサブドキュメント分類ルールを構築するために、すべてのフィーチャが同時に用いられる１つの機械学習アルゴリズムに組み合わせられる。

【００５２】

さらなる実施形態において、分類ルールからの出力スコアは、推定されるクラスメンバーシップの確率として解釈され得る。これは、スコアが、推定するように構築された真のクラス確率分布関数と良好に相関することを意味する。これらのスコアは、確率に対して較正されるので、誤分類コストおよび事前分類（ｃａｔｅｇｏｒｙ　ｐｒｉｏｒｓ）を考慮して決定を下す（例えば、Ｂａｙｅｓ最適決定）ために用いられ得る。確率を厳密に推定するための出力スコアの較正は、異なった情報ソースの組み合わせがより容易に達成されることを可能にする。なぜなら、種々のソースのからの情報の組み合わせは、通常、どのように進行するか、またはどれほどの改善が可能であるかを決定するために、原則に基づいた方法を用いずに、発見的に実行されるからである。

【００５３】

ある実施形態において、本発明は、１．０または０．０の「ハード」確率を生成する方法およびシステムに基づいたルールを含み得る。他の実施形態において、本発明は、より高度のレベルの分解能で、より平滑な確率密度関数を推定する能力を有する方法およびシステムを含む。

【００５４】

別の実施形態において、本発明は、ユーザが分類ルールまたはさらなる問題の制約を手動で明確にすることを可能にする。これは、関係／制約がネットワークにおいて容易にエンコードされ、かつユーザにすでに知られている場合、機械学習技術を用いるよりも、制約および関係を取得するために、より効率的な方法であり得る。

【００５５】

さらなる実施形態において、本発明は、ドキュメントの境界に線引きし、かつ、ドキュメントタイプを識別する方法を含む。この方法は、各カテゴリの、サンプルドキュメントイメージに基づいて自動的に生成された分類ルールに従って、複数のドキュメントイメージを複数の所定のカテゴリに自動的に分類するステップと、複数のドキュメントイメージのどれが、少なくとも２つのカテゴリのいずれに属するかを識別するための少なくとも１つの識別子を自動的に生成するステップとを包含する。

【００５６】

ある実施形態において、少なくとも１つの識別子は、複数のカテゴリのうちの異なったカテゴリに属するイメージに線引きするために、ドキュメントイメージ間に挿入された、コンピュータによって生成された分離ページを含む。別の実施形態において、少なくとも１つの識別子は、それぞれのカテゴリ分類に従って、複数のデジタルイメージのカテゴリ分類シーケンスを識別するコンピュータ可読記述（例えば、ＸＭＬメッセージ）を含む。さらに別の実施形態において、少なくとも１つの識別子は、複数のドキュメントイメージの少なくとも１つと電子的に関連付けられた、コンピュータによって生成された少なくとも１つのラベルを含む。

【００５７】

本発明のある実施形態によると、ネットワーク構造の構成可能性、およびルール構築の種々の方法を活用してネットワークの確率を推定する能力のために、本発明は、他の形態の情報、あるいは他の種類のドキュメントまたはサブドキュメントタイプを含むように、容易に保守および拡張される。

【発明の効果】

【００５８】

本発明により、ドキュメントの境界の線引き、およびドキュメントタイプの識別が、コンピュータベースのシステムにおいて達成される。

【発明を実施するための最良の形態】
【００５９】
（好適な実施形態の詳細な説明）
本発明は、以下において図を参照して詳細に説明され、ここで、同じ要素は、一貫して同じ符号で示される。
【００６０】
本発明は、コンピュータシステムまたは他の処理システム上でソフトウェアを用いて実行され得る。図１は、本明細書中に記載された本発明の機能性を実行することができる例示的コンピュータシステム１００のブロック図である。各コンピュータシステム１００は、Ｉｎｔｅｌ Ｃｏｒｐｏｒａｔｉｏｎ（米国カリフォルニア州Ｓａｎｔａ Ｃｌａｒａ）から販売される「Ｐｅｎｔｉｕｍ（Ｒ）」マイクロプロセッサおよび関連した集積回路チップ等の１つ以上の中央演算処理ユニット（ＣＰＵ）の制御下で動作する。コンピュータシステム１００は、キーボードおよびマウス１０４からコマンドおよびデータを入力し得、ユーザは、ディスプレイ１０６で入力およびコンピュータ出力を閲覧し得る。このディスプレイは、通常、ビデオモニタまたはフラットパネルディスプレイデバイスであり、コンピュータ１００は、さらに、予め組み込まれたハードディスクドライブ等の、ダイレクトアクセス格納デバイス（ＤＡＳＤ）もまた含む。メモリ１０８は、通常、揮発性半導体ランダムアクセスメモリ（ＲＡＭ）を備える。各コンピュータは、好適には、プログラム製品リーダがデータを読み出し得（および、データを適宜書き込み得）るプログラム製品格納デバイス１１２を収容するプログラム製品リーダ１１０を含む。プログラム製品リーダは、例えば、ディスクドライブを備え得、プログラム製品格納デバイスは、フロッピー（Ｒ）ディスク、光学式ＣＤ－ＲＯＭディスク、ＣＤ－Ｒディスク、ＣＤ－ＲＷディスク、ＤＶＤディスク等の取り外し可能な格納媒体を備え得る。各コンピュータ１００は、コンピュータネットワーク１１３を介して接続された他のコンピュータと、ネットワーク１１３とコンピュータ１００との間の接続１１６を介して通信を可能にするネットワークインターフェース１１４を通じて通信し得る。これらのデバイスは、通信バス１１７を通じて他のデバイスと通信する能力を有する。
【００６１】
ＣＰＵ１０２は、ＤＡＳＤ１０７に格納され、および／またはコンピュータ１００のメモリ１０８に一時的に格納されるソフトウェアプログラムのプログラミングステップの制御下で動作する。プログラミングステップが実行された場合、関連システムコンポーネントは、機能を実行する。従って、ある実施形態において、プログラミングステップは、本明細書中に記載されたシステムの機能性を実現する。プログラミングステップは、プログラム製品１１２、またはネットワーク接続１１６を通じて、ＤＡＳＤ１０７から受信され得る。格納ドライブ１１０は、ＣＰＵ１０２による実行のために、プログラム製品を受信し、その上に記録されたプログラミングステップを読み出し、かつ、プログラミングステップをメモリ１０８に転送する。上述のように、プログラム製品格納デバイスは、磁気フロッピー（Ｒ）ディスク、ＣＤ－Ｒｏｍ、およびＤＶＤ格納ディスクを含む、記録されたコンピュータ可読命令を有する複数の取り外し可能媒体の任意の１つを備え得る。他の適切なプログラム製品格納デバイスは、磁気テープおよび半導体メモリチップを含み得る。このようにして、本発明による動作のために必要な処理ステップは、プログラム製品上で具体化され得る。
【００６２】
あるいは、プログラムステップは、ネットワーク１１３を介してオペレーティングメモリ１０８に収容され得る。ネットワーク方法において（さらなる説明がなくても当業者に理解される周知の方法により）、コンピュータは、ネットワーク通信がネットワーク接続１１６にわたって確立された後、プログラムステップを含むデータを、ネットワークインターフェース１１４を通じてメモリ１０８に収容する。その後、システムの処理を実行するために、プログラムステップがＣＰＵ１０２によって実行される。当業者に公知のように、本明細書中に記載される本発明の種々の機能をサポートするために、代替的アーキテ

クチャおよび構成を有する他のコンピューティングマシンおよびシステムが実現され得る。

【００６３】

１実施形態において、デジタルスキャナ１２０は、任意の公知の周辺バスインターフェースまたはアーキテクチャを用いてコンピュータシステム１００に接続される。スキャナ１２０は、アナログイメージ（例えば、グラフィックおよび／またはテキスト情報）を走査して、これらをＣＰＵ１０２によって処理するために、デジタルイメージに変換するか、またはファイルする。スキャナ１２０は、市販される任意の適切なスキャナであり得る。１実施形態において、スキャナ１２０は、イリノイ州Ｌｉｎｃｏｌｎｗｏｏｄに位置するＢｏｅｗｅ　Ｂｅｌｌ＆Ｈｏｗｅｌｌによって製造されるＢｏｅｗｅ　Ｂｅｌｌ＆Ｈｏｗｅｌｌ８１２５である。

【００６４】

１実施形態において、本発明は、ドキュメントに線引きするためにセパレータページを用いる従来技術のプロセスを改善するように設計される。例示的従来技術のプロセスは、図２に示される。２０１で開始して、ドキュメントページの集合は、人が異なったドキュメントタイプまたは目的のセクションに対応するページ間に物理セパレータシートを手動で挿入することによって処理される。目的のセクションは、各ドキュメントに必要とされるアプリケーションおよびさらなる処理に依存する。ステップ２０２にて、ドキュメントページおよびセパレータページの集合が、その後、デジタルスキャナに供給され、セパレータページを含む、各ページを表すデジタルイメージのシーケンスが生成される。このイメージのシーケンスは、その後、セパレータページによって識別および区別されるドキュメントまたはサブドキュメントのタイプに基づいて、システム１００内に常駐するさらなるソフトウェアコンポーネントによってさらに処理され得る。ドキュメントまたはサブドキュメント専用の処理がここで可能である。なぜなら、セパレータページのイメージは、ドキュメントまたはサブドキュメントを線引きし、かつ、システム１００によって容易に検出され得るからである。

【００６５】

本発明は、ページのドキュメントグループまたはサブドキュメントグループを線引きするプロセスを自動化する。１実施形態は、図３に示される。ステップ３０１で開始して、ドキュメントページは、デジタルスキャナ１２０に挿入され、かつ、デジタルイメージのシーケンスに変換される。このデジタルイメージのシーケンスは、その後、本発明により処理される（ステップ３０２）。ステップ３０２の出力は、ステップ２０２の出力、すなわち、自動的に生成されたセパレータシートのイメージがインターリーブされたデジタル化されたページのシーケンスと同じである。差異は、ステップ３０２において、本発明は、セパレータシートイメージをイメージシーケンスに自動的に挿入していることである。１実施形態において、ソフトウェアによって生成されたセパレータページは、さらに、セパレータページの直後に追従するか、またはこれに先行するドキュメントのタイプを示す。本発明がセパレータページをどのように決定するか、および、セパレータページをどこで挿入するかの方法は、本発明の種々の実施形態によりさらに詳細に後述される。

【００６６】

作業の流れのルーティングシステムが、ドキュメントシーケンス情報を直接的に解釈するように構成された場合、将来のサブシステムをセパレータイメージの処理または格納から解放して、さらなる経済性が得られ得る。この代替的実施形態は、図４に示される。ステップ４０１で開始して、ページは、デジタルスキャナに挿入され、かつ、デジタルイメージのシーケンスに変換される。このデジタルイメージのシーケンスは、その後、本発明により処理される（ステップ４０２）。このステップにおいて、セパレータシートイメージをデジタルイメージのシーケンスに挿入する代わりに、ステップ４０２は、変更されない元のデジタル化されたイメージシーケンスを出力し、かつ、さらに、イメージのシーケンスの記述を出力する。この記述は、ドキュメントまたはサブドキュメントの境界がコンピュータシステム１００によって解釈されることを可能にする。１実施形態において、こ

の記述は、ドキュメントの境界およびタイプを決定するために、システム１００によって読み出されかつ処理されるＸＭＬメッセージである。ドキュメント分離に対応する例示的ＸＭＬメッセージは、

【００６７】

【数１】

```
<SeparationDescription>
     <Section type="FormA">
          <Image SeqOffset="1"/>
     </Section>
     <Section type="FormB">
          <Image SeqOffset="2"/>
          <Image SeqOffset="3"/>
     </Section>
     <Section type="FormC">
          <Image SeqOffset="4"/>
     </Section>
</SeparationDescription>
```

のように提供される。

【００６８】

しかしながら、当業者は、シーケンスを行う情報を生成および提供するために代替的方法が存在することを理解する。例えば、１実施形態において、コンピュータシステム１００は、電子ラベルまたは他の識別子を各スキャナによって生成されたデジタルイメージ上に挿入または貼り付けて、一連のフォームにおける各フォームの最初、最後、および任意の中間のページを識別し得る。次に続く、これらのページの処理は、その後、各ページのラベルまたは識別子により実行される。

【００６９】

１実施形態において、本発明は、ページのシーケンスを自動的に分離するために分類ルールを構築および組み合わせる。ルールのセットは、確率ネットワークによって定義される。１実施形態において、このネットワークは、Ｍｏｈｒｉ，Ｍによる「Ｆｉｎｉｔｅ－Ｓｔａｔｅ Ｔｒａｎｓｄｕｃｅｒｓ ｉｎ Ｌａｎｇｕａｇｅ ａｎｄ Ｓｐｅｅｃｈ Ｐｒｏｅｓｓｉｎｇ」（以後、「Ｍｏｈｒｉ」）Ａｓｓｏｃｉａｔｉｏｎ ｆｏｒ Ｃｏｍｐｕｔａｔｉｏｎａｌ Ｌｉｎｇｕｉｓｔｉｃｓ（１９９７年）に記載の有限状態機械（ＦＳＭ）の公知の形態である、有限状態変換器として実現され得る。１実施形態によると、本明細書中で記載されるＦＳＭのタイプは、入力値、状態またはアイテム（例えば、ページのデジタルイメージ）を表す入力アーク（ｉｎｐｕｔ ａｒｃｓ）、および、可能な次の値を表す出力アーク（ｏｕｔｐｕｔ ａｒｃｓ）を有する１つ以上の状態遷移または決定ポイントとして表され得る。当該分野で公知のように、各状態遷移または決定ポイントは、入力アーク上の入力、出力アーク上の出力を受取り、１実施形態において、入力アークおよび／または出力アークと関連した確率重み値を有する。入力アークおよび出力アークは、さらに、ε（イプシロン）とよく呼ばれる空値またはシンボルを表し得る。１実施形態において、この確率重み値は、確率の負の対数として解釈され、ここで、Ｐは、アークによって表される確率である。

【００７０】

図５は、単一のドキュメント内の３つのフォームまたはサブドキュメントを分離するように設計された単純なＦＳＭまたは確率ネットワークのグラフィック表現を示す。ＦＳＭは、３つのアークを有する単一の遷移状態またはポイントを含み、各アークは、入力および出力状態の両方を表す。コロンの前の各アークに関する情報は、入力アイテムである。図５の場合、これは入力イメージである。この入力イメージは、イメージのシーケンスにおいて、各イメージがそのアークの入力として考えられることを示すために、下付き数字ｔがインデクス付けされる（例えば、イメージ０は、最初の入力、イメージ１は次の入力等）。コロンの後であるが“／”の前の情報は、出力である。この場合、これは「Ａ」「Ｂ」または「Ｃ」であり、３つのフォーム、すなわちフォームＡ、フォームＢ、またはフォームＣのうちの１つにページを割り当てることに対応する。“／”の右側の情報は、モデル化されたイベントの確率である。当業者に公知のように、変換器は、ある「通常言語」をもう一方の「通常言語」にマップする。この場合、図５における変換器は、イメージのシーケンスをＡ、ＢおよびＣシンボル、および、これらと関連した確率にマップする。実際、図５における変換器が、入来するイメージのシーケンスに適用された場合、フォームのサブシーケンスのすべての可能な組み合わせが、それらが生じる確率と共に列挙される。グラフ探索アルゴリズムは、その後、最高の確率でイメージのシーケンスが与えられた、フォームのシーケンスを見つけるために用いられ得る。例示的グラフ探索アルゴリズムは、深さ優先探索および幅優先探索アルゴリズムであり、これらは、当業者に周知であり、かつ、例えば、Ｒｕｓｓｅｌｌ，Ｓ．、Ｎｏｒｖｉｇ，Ｐ．による「Ａｒｔｉｆｉｃｉａｌ Ｉｎｔｅｌｌｉｇｅｎｃｅ：Ａ Ｍｏｄｅｒｎ Ａｐｐｒｏａｃｈ」Ｐｒｅｎｔｉｃｅ－Ｈａｌｌ，Ｉｎｃ．（１９９５年）７０～８４ページ、５３１～５４４ページに記載される。図５における変換器について、これは、各イメージに与えられた、最も見込みのあるフォームを欲張り（ｇｒｅｅｄｙ）な態様で選択することと同じである。これは、各イメージが、他のイメージと別個であると考えられ、かつ、他のイメージがどのフォームに割り当てられたかを考慮に入れないからである。しかしながら、複数のフォームタイプが存在し得る任意のネットワーク構造について、高い確率を有するシーケンスは、必ずしも、イメージごとに最高のイメージ対フォームの確率アークを順番に選択することによって構築されるシーケンスではない。これは、フォームの特定のシーケンスが、例えば、フォームの（ページの）長さ、または先行するか、または次に続くイメージと関連した確率といった他のファクタに基づいて、多少見込みがあることが可能だからである。

【００７１】

図５において、「$image_t$：Ａ／ｐ（ＦｏｒｍＡ｜$image_t$）」とラベル付けされた第１の最大のアークは、ＦＳＭの可能な経路または出力を表し、かつ、走査されたイメージは「ＦｏｒｍＡ」イメージである確率を提供する。「$image_t$：Ｂ／ｐ（ＦｏｒｍＢ｜$image_t$）」とラベル付けされた第２の中間アークは、ＦＳＭの別の可能な経路または出力を表し、かつ、走査されたイメージが「ＦｏｒｍＢ」イメージである確率を提供する。同様に、「$image_t$：Ｃ／ｐ（ＦｏｒｍＣ｜$image_t$）」とラベル付けされた最小のアークは、そのイメージが「ＦｏｒｍＣ」である確率を提供する。１実施形態において、各経路と関連した確率は、各イメージのテキストおよび／またはグラフィカルコンテンツを解析し、その後、このコンテンツを既知のモデルと比較するか、または各カテゴリまたはフォームタイプと関連したセットをトレーニングすることによって生成される。このタイプの解析および確率分類を実行する例示的方法およびシステムは、「Ｅｆｆｅｃｔｉｖｅ Ｍｕｌｔｉ－Ｃｌａｓｓ Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅ Ｃｌａｓｓｉｆｉｃａｔｉｏｎ」と称される米国特許出願第６０／３４１，２９１号（２００３年3月１０日出願）、アトーニードケット番号第５２９２３－２００８００号に記載され、この出願は参考のため、本明細書中にその全体が援用される（以後、「Ｈａｒｒｉｓ」）。

【００７２】

図６は、同じ問題を解決するために、本発明の別の実施形態を表す、より複雑な変換器を示す。この変換器のアーク確率は、入来するイメージ（$image_t$と示される）に依

存し、このフォームに前のイメージが割り当てられる（$image_{t-1}$と示される）。例えば、「$image_t$：A／p（FormA｜$image_t$，$image_{t-1}$＝FormA）」とラベル付けされたアークは、ＦＳＭの１つの可能な経路または結果を表し、走査されたイメージが、現在のイメージ、$image_t$のプロパティが与えられた「FormA」イメージである確率、および、前のイメージが「FormA」イメージであったという情報を提供する。あるいは、「$image_t$：A／p（FormA｜$image_t$，$image_{t-1}$＝FormB）」とラベル付けされたアークは、同じ入力イメージの異なった確率を生成する。なぜなら、このアークは、前のイメージ、すなわち$image_{t-1}$が「FormB」イメージであり、「FormA」イメージではなかったという情報を用いるからである。前のイメージの分類に関する情報を利用することによって、このＦＳＭは、現在のイメージをどのように分類するかについて、より識別力がある。ＦＳＭをこのように構築することによって、当該の問題についてのより複雑な確率モデルが表現される。

【００７３】

さらなる実施形態において、ＦＳＭの構築および最適化は、例えば、Mohriに記載されるような関係代数の方法を用いて行われ得る。当業者に公知のように、変換器の入力（または、同様に、出力）側が、通常言語を表す。１実施形態において、通常言語は、セット、場合によっては、無限のイメージ（フォーム）の入力（出力）シーケンスである。従って、結合、クロス乗積、否定、減算および交差等のセットを演算は、他の変換器を生成するために、変換器の群上で実行され得る。さらに、変換器は、有理関数であり、従って、例えば、Mohriに記載されるように、このような投影および生成等の演算もまた可能である。これらの演算は、Mohriに示されるように、変換器を構築、操作および最適化する際に有用であることが証明される。

【００７４】

例えば、図６がほぼ正確であったが、FormAに割り当てられる２つの連続したイメージのシーケンスを否認することを所望していたことを前提とする。おそらく、これは、FormAがあるページフォームであり、かつ、別のFormAの次に現れ得ないというビジネスルールを強化する。図７における変換器は、ＦＳＭであり、これは、図６におけるＦＳＭで生成された場合、正確に所望の結果をもたらす。図７において、アークは、フォームタイプシンボル（「A」、「B」、または「C」）である入力および出力アークの両方でラベル付けされ、かつ、確率を有しない。入力シンボルは、フォームタイプである。なぜなら、このＦＳＭは、図５～図６に記載されるようなＦＳＭの出力を、入力としてとるように設計されるからである。さらに、アーク上には確率はない。なぜなら、このＦＳＭは、特定の経路が他よりもより見込みがあるか、または好適であると判定するように設計されないからであり、このＦＳＭは、フォームタイプのシーケンスを単に認めるか、または否認するように設計される。これは、２つのFormAイメージのシーケンスを有するすべての経路に０確率を、および、すべての他の経路に１．０確率を均等に割り当てると考えられ得る。例えば、「A：A」がラベル付けされたアークは、ＦＳＭがFormAシンボルをFormAシンボルに無条件にマッピングすることを意味する。同様に、「B：B」および「C：C」とラベル付けされたアークは、FormBシンボル対FormBシンボル、およびFormCシンボル対FormCシンボルにそれぞれマッピングする。これらのアークは、図５～図６におけるもの等のＦＳＭを用いて決定された任意のイメージについて、フォームタイプを変更せず、２つのFormAタイプイメージを有するシーケンスのみを除去することに留意されたい。これは、一旦「A」シンボルが読み込まれると、唯一の許容され得る出て行くアークは「B」および「C」だからである。従って、２つの連続する「A」出力シンボルを含む任意の経路は、最良の解決策を見つけるためにグラフ探索アルゴリズムが用いられる間、廃棄されるからである。

【００７５】

図８は、図５における変換器が与えられた６つの入力イメージのフォームシーケンスのすべての組み合わせの表現を示す。図５における変換器が与えられた６つの入力イメージ

に対して７２０の可能なフォームシーケンスがある。図７におけるフィルタが提供された後、Ｍｏｈｒｉにおいて記載されるような重み付きＦＳＭの生成を用いて、隣り合う２つのＦｏｒｍＡイメージを有するすべてのシーケンスが除去される（図９に示される）。状態およびアークの数が図８よりも図９において、より多い一方で、一意的シーケンスまたは経路の数は、図９において、より小さいことに留意されたい。図９における６つの入力イメージに対して、４４８のフォームのシーケンスのみがある。他の７２０－４４８＝２７２は、すべて、２つの連続的ＦｏｒｍＡを中に有し、従って、可能なシーケンスとして除去された。

【００７６】

このフレームワークにおいて、イメージごとの情報、イメージ情報のシーケンス、フォームごとのシーケンス情報、イメージ情報のシーケンス、およびフォーム情報のシーケンスを利用する確率およびカスタムアプリケーションルール（例えば、２つの連続的ＦｏｒｍＡイメージは、許されない）は、すべて、許容可能なシーケンスのセットを制約するように原則に基づいて組み合わされ、かつ、次に、最高の確率を有する許容可能なシーケンスを見つけるために最適化され得る。

【００７７】

本発明は、各アークに対して分類ルールを確立する周知のマシン学習技術を用いる。例示的技術は、例えば、Ｂｉｓｈｏｐ，Ｃ．による「Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋｓ ｆｏｒ Ｐａｔｔｅｒｎ Ｒｅｃｏｇｎｉｔｉｏｎ」Ｏｘｆｏｒｄ Ｕｎｉｖｅｒｓｉｔｙ Ｐｒｅｓｓ、Ｉｎｃ．（２００２年）、２７、７７～８５、２３０～２４７、２９５～３００、および３４３～３４５ページに記載されるようなニューラルネットワーク（以後、「Ｂｉｓｈｏｐ」）、Ｖａｐｎｉｋ，Ｖ．による「Ｔｈｅ Ｎａｔｕｒｅ ｏｆ Ｓｔａｔｉｓｔｉｃａｌ Ｌｅａｒｎｉｎｇ Ｔｈｅｏｒｙ：Ｓｅｃｏｎｄ Ｅｄｉｔｉｏｎ」Ｓｐｒｉｎｇｅｒ－Ｖｅｒｌａｇ Ｎｅｗ Ｙｏｒｋ，Ｉｎｃ．（２０００年）１３８～１４２ページに記載されるようなサポートベクトルマシーンである。他の技術は、例えば、Ｒｕｓｓｅｌｌ，Ｓ．およびＮｏｖｉｇ，Ｐ．による「Ａｒｔｉｆｉｃｉａｌ Ｉｎｔｅｌｌｉｇｅｎｃｅ：Ａ Ｍｏｄｅｒｎ Ａｐｐｒｏａｃｈ」Ｐｒｅｎｔｉｃｅ－Ｈａｌｌ，Ｉｎｃ．（１９９５年）、５３１～５４４ページに記載されるような学習された決定ツリーの利用を含む。別の実施形態において、これらの方法は、例えば、Ｂｉｓｈｏｐ、Ｈａｒｒｉｓ、およびＺａｄｒｏｚｎｙ，Ｂ．らによる「Ｔｒａｎｓｆｏｒｍｉｎｇ ｃｌａｓｓｉｆｉｅｒ ｓｃｏｒｅｓ ｉｎｔｏ ａｃｃｕｒａｔｅ ｍｕｌｔｉｃｌａｓｓ ｐｒｏｂａｂｉｌｉｔｙ ｅｓｔｉｍａｔｅｓ」Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ Ｅｇｙｐｔ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｋｎｏｗｌｅｄｇｅ Ｄｉｓｃｏｖｅｒｙ ａｎｄ Ｄａｔａ Ｍｉｎｉｎｇ、（２００２年）、ならびにＰｌａｔｔ，Ｊ．Ｃ．による「Ｐｒｏｂａｂｉｌｉｓｔｉｃ ｏｕｔｐｕｔｓ ｆｏｒ Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅｓ ａｎｄ Ｃｏｍｐａｒｉｓｏｎｓ ｔｏ Ｒｅｇｕｌａｒｉｚｅｄ Ｌｉｋｅｌｉｈｏｏｄ Ｍｅｔｈｏｄｓ」Ａｄｖａｎｃｅｓ ｉｎ Ｌａｒｇｅ Ｍａｒｇｉｎ Ｃｌａｓｓｉｆｉｅｒｓ、ＭＩＴ Ｐｒｅｓｓ（１９９９年）に記載されるような較正された出力確率を出力し、従って、上述のネットワークの最適化が原則に基づいて行われる。

【００７８】

分類および確率ルールの適用と共に、本発明は、カスタムアプリケーションまたは「フィルタ」ルールをさらに含み、これらは、各ドキュメント、サブドキュメント、フォームまたは他のアイテムの公知の特性に基づいてアプリケーションごとに合わせられる。例えば、上述のように、特定のアプリケーションにおいて、ＦｏｒｍＡが単一のページフォームにすぎないことが知られ得る。従って、２つの連続的ＦｒｏｍＡに至るＦＳＭのすべての可能な経路は、ＦＳＭから除去される。別の例のように、ＦｏｒｍＣは、常に３つのページの長さであることが公知であり得る。従って、ＦｏｒｍＣの開始ページが中間および終わりのページによって追従されなければならないカスタムロールは、このカスタムルールを満たさない任意の経路を削除するように実現され得る。これらは、次に続くハンドリ

ング／プロセシングのために、ドキュメントまたは他のアイテムを分類および分離する際に支援し得る多くの可能なカスタムルールのほんのわずかな例である。

【００７９】

本発明の１実施形態による、ドキュメントセパレーションプログラムは、ページのグループを異なったローン申し込みフォームに分離するために、各イメージからテキストフィーチャのみを用いて構築された。２０の異なった可能なフォームがあるが、これらのフォームの１３のみが２５よりも多い例示的ページを有し、従って、分類子は、これらの１３のフォームのためにだけ構成される。この実施例は、２５，５２６ローンアプリケーションページを含む。これらのページは、Ｂｏｅｗｅ　Ｂｅｌｌ＆Ｈｏｗｅｌｌ（イリノイ州Ｌｉｎｃｏｌｎｗｏｏｄに位置する、によって製造されたＢｏｅｗｅ　Ｂｅｌｌ＆Ｈｏｗｅ　８１２５デジタルスキャナを用いてデジタル的に走査され、ドキュメントページごとに単一の．ｔｉｆ　ｆｉｌｅが生成された。これらのイメージは、ロシアのＭｏｓｃｏｗに位置するＡｂｂｙｙ　Ｓｏｆｔｗａｒｅ　Ｈｏｕｓｅによって製造されたＡｂｂｙＹＯＣＲと呼ばれる第３パーティ光学式文字認識（ＯＣＲ）パッケージによって処理された。ＯＣＲ処理は、ページごとに単一のＡＳＣＩＩテキストファイルをもたらす。

【００８０】

単一のローンの申し込みに対応するすべてのテキストファイルは、プログラムに送信される。ＡＳＣＩＩエンコードテキストファイルのシーケンスは、ローン申し込み＃を含んだファイル名を有するディスクに保存される。これらの個々のテキストファイルは、その後、２０のフォームタイプのうちの１つにより手動で分類され、順序どおりにローン申し込みに現れる。

【００８１】

分類子を構築するために用いられる１３のフォームの各々について、Ｆｏｒｍ＿ｓｔａｒｔ、Ｆｏｒｍ＿ｍｉｄ、およびＦｏｒｍ＿ｅｎｄの３つのカテゴリが構成される。これらの３つのクラスは、フォームに最初に現れるページ、フォームの中間ページ、およびフォームの最後のページをそれぞれ表すように構成される。３つ以上のページを含むフォームについては、ページ１は、Ｆｏｒｍ＿ｓｔａｒｔに割り当てられ、最後のページは、Ｆｏｒｍ＿ｅｎｄに割り当てられ、および、すべての他のページは、Ｆｏｒｍ＿ｍｉｄに割り当てられる。２つのページのみを含むフォームについては、最初および最後のフォームがＦｏｒｍ＿ｓｔａｒｔ、およびＦｒｏｍ＿ｅｎｄのそれぞれに割り当てられ、さらに、最初および最後のページの両方が、Ｆｏｒｍ＿ｍｉｄに割り当てられる。最後に、長さにおけるただ１つのページであるフォームについて、このページは、すべての３つのカテゴリに割り当てられる。従って、１３のフォームタイプ×フォームタイプごとの３つのカテゴリが、３９のバイナリ分類子の構成をもたらす。それぞれがＨａｒｒｉｓに記載されたように、クラスメンバーシップの確率を出力するように構成される（例えば、ｐ（ＬｏａｎＡｐｐｌｉｃａｔｉｏｎ＿ｓｔａｒｔ｜ｉｍａｇｅ）、ｐ（Ａｐｐｒａｉｓａｌ＿ｅｎｄ｜Ｉｍａｇｅ）等）。各場合における正のクラスは、クラス（例えば、Ａｐｐｒａｉｓａｌ＿ｅｎｄ）における例によって定義され、負のクラスは、すべて、他のページである（分類子が構成されない７つのフォームの一部分であったものを含む）。ドキュメントの分離の有効性を試験するために、試験セットは、トレーニングセットを構成するために用いられたものと同じ方法で構成される。このセットは、２０個の異なったフォームからの５，３５７ページを有する。

【００８２】

３９個のページごとのテキスト分類子のみを用いた結果、図５と類似の構成になる。無効のシーケンスの１つの原因は、ページが、Ｆｏｒｍ＿ｓｔａｒｔに割り当てられる前に現れるページの前に、Ｆｏｒｍ＿ｅｎｄに不正確に分類されることである。エラーの別の形態は、フォームが開始した場合に、別のフォームが開始し、その後、最初のフォームが終了し、その後、第２のフォームが終了することである。これらのシーケンスは、無意味であるので、以下のフィルタルールを強化したフィルタＦＳＭが構成される。これらのルールは、一旦Ｆｏｒｍ＿ｓｔａｒｔがページに割り当てられると、次のページは、すべて

、対応するＦｏｒｍ＿ｍｉｄまたはＦｏｒｍ＿ｅｎｄに割り当てられなければならず、一旦Ｆｏｒｍ＿ｅｎｄがページに割り当てられると、パージは、Ｆｏｒｍ＿ｓｔａｒｔにのみ割り当てられ得る。これにより、有効なシーケンスのみが製造される。上述の例示的ＦＳＭの単純な構造を仮定して、類似の態様で残りの１０個のフォームをこのＦＳＭに追加することは、当業者にとって通常の手順である。次に、ページのシーケンスが与えられたフォームの最も見込みのあるシーケンスを見つけるために、生じたＦＳＭに深さ優先探索アルゴリズムが適用される。このシーケンスは、テキストベースの分類子によって自動的に構築されたルールを用いてフォームにページが割り当てられることによって定義される。見出されたシーケンスにおけるフォームに対する、ページの、ページごとの割り当ての最高確率からの唯一の逸脱は、見出されるシーケンスが上述の制約を満たすという意味で「許容可能」でなければならず、すなわち、（ａ）は、フォームはオーバーラップし得ず、かつ（ｂ）終了する前に開始しなければならないことである。試験セットに対してこのプログラムを用いることによって、自動フォーム分離の実行が続く。

【００８３】

【表１】

NAME	TP	FP	FN	精度	リコール	F- 計測
AppraisalOrigination	22	1	4	95.65%	84.62%	89.80%
FinalLoanApplication	101	9	9	91.82%	91.82%	91.82%
FloodCertificate	34	0	0	100.00%	100.00%	100.00%
HUD1Settlement	100	9	15	91.74%	86.96%	89.29%
Note	431	8	18	98.18%	95.99%	92.62%
OriginationFundingTransmittal	69	1	10	98.57%	87.34%	92.62%
OriginationHUDReqCopyTaxForm	40	117	57	25.48%	98.97%	99.48%
OriginationInitialEscrow	81	0	1	100.00%	98.78%	99.39%
OriginationLimitedPowerOfAttorney	105	2	6	98.13%	94.59%	96.33%
OriginationMiscRiders	24	24	18	50.00%	57.14%	53.33%
OriginationTitleCommitment	24	8	15	75.00%	61.54%	67.61%
OriginationTruthInLending	103	3	5	97.17%	95.37%	96.26%
OriginationUnrecordedMortgage	85	7	17	92.39%	83.33%	87.63%
Summaries (Known forms)	1219	189	175	86.58%	87.45%	87.01%

カラム「ＮＡＭＥ」は、試験されるフォームの名称に対応する。カラム「ＴＰ」、「ＦＰ」および「ＦＮ」は、システムによってなされる真正、偽正および偽負フォーム分離をそれぞれ示す。精度は、ＴＰ／（ＴＰ＋ＦＰ）として定義され、リコールは、ＴＰ／（ＴＰ＋ＦＮ）と定義される。Ｆ測定値は、精度とリコールとの間の調和平均と定義される。上述の表において、各ＴＰ、ＦＰ、およびＦＮは、完成したフォームである（例えば、イメージのシーケンス）。従って、３ページＡｐｐｒａｉｓａｌＯｒｉｇｉｎａｔｉｏｎフォームが、１ページＡｐｐｒａｉｓａｌＯｒｉｇｉｎａｔｉｏｎフォームによって追従される２ページＡｐｐｒａｉｓａｌＯｒｉｇｉｎａｔｉｏｎフォームに不正確に分割された場合、これは、ＡｐｐｒａｉｓａｌＯｒｉｇｉｎａｔｉｏｎについては２ＦＰおよび１ＦＮになる。同じ３ページフォームが３ページＮｏｔｅとして不正確に識別された場合、Ｎｏｔｅについては、これは１ＦＰと記録され、ＡｐｐｒａｉｓａｌＯｒｉｇｉｎａｔｉｏｎについては１ＦＮとして記録される。

【００８４】

この手順によってなされる別のタイプの間違いは、１つのフォームのページの長いシーケンスが、まさに隣り合うフォームの２つのシーケンスに分割されることである。例えば、単一の４ページフォームは、２つの隣接する２ページフォームに分割され得る。特定のローン処理アプリケーションがなされた場合、任意のタイプのフォームの２つの発生（ｏｃｃｕｒｒｅｎｃｅ）が同じローンアプリケーションにおいて現れることは不可能である。従って、別の実施形態において、図７におけるものと同じフィルタが、まさに隣り合って現れる反復形態を除去するように構成される。従って、テキスト分類子が、すべての４

つのページを同じフォームタイプに割り当てることが所望された場合、これらは、２ページの２つのシーケンスの代わりに、４ページの１つのシーケンスに押し込まれる。これは、システムの精度を劇的に改善する。フィルタは、約１時間のうちに構成され、これは、カスタムビジネスルールを、特に、このローンアプリケーション問題に対して実施する。さらなるフィルタルールを有するこのシステムの性能は、以下のテーブルで提供される。

【００８５】

【表２】

NAME	TP	FP	FN	精度	リコール	F-計測
AppraisalOrigination	22	0	4	100.00%	84.62%	91.67%
FinalLoanApplication	103	3	7	97.17%	93.64%	95.37%
FloodCertificate	34	0	0	100.00%	100.00%	100.00%
HUD1Settlement	101	5	14	95.28%	87.83%	91.40%
Note	431	0	18	100.00%	95.99%	97.95%
OriginationFundingTransmittal	69	1	10	98.57%	87.34%	92.62%
OriginationHUDReqCopyTaxForm	96	0	1	100.00%	98.97%	99.48%
OriginationInitialEscrow	81	0	1	100.00%	98.78%	99.39%
OriginationLimitedPowerOfAttorney	105	2	6	98.13%	94.59%	96.33%
OriginationMiscRiders	31	0	11	100.00%	73.81%	84.93%
OriginationTitleCommitment	24	1	15	96.00%	61.54%	75.00%
OriginationTruthInLending	103	3	5	97.17%	95.37%	96.26%
OriginationUnrecordedMortgage	88	0	14	100.00%	86.27%	92.63%
Summaries (Known forms)	1288	15	106	98.85%	92.40%	95.51%

上記の結果は、カスタムフィルタを組み込んで、ドキュメントおよび／またはアイテムの自動分類および分離を改善するために、カスタムフィルタルールを取り入れることができるという点で、本発明の有用性を示す。１実施形態において、フィルタルールは、処理されるドキュメントまたはアイテムの公知の特徴、フィーチャ等を手動で用いて構成され得る。別の実施形態において、フィルタルールは、上述のように、例示的ドキュメントまたはアイテムのトレーニングセットを利用する公知の機械学習技術を用いて自動的に構築され得、システムを構成または適合することが必要とされる非常にわずかな時間で非常に正確なシステムを達成する。

【００８６】

別の実施形態において、前の発明は、ローンアプリケーションをデジタルで走査および処理するための大規模プロセスに統合される。このプロセスは、作業の流れ、および１９個のＢｅｌｌ＆Ｈｏｗｅｌｌ８１２５デジタルスキャナおよび２２個の人間が見るリビュアの統合化を管理するためのＫｏｆａｘ　Ａｓｃｅｎｔ　Ｃａｐｔｕｒｅ５．５１ソフトウェアを用いる。この統合は、上述のようなＸＭＬメッセージを戻す方法を用いて行われる。この統合は、各フォームに割り当てられるページごとのすべての確率の平均にすぎないフォームごとの「信頼スコア」を戻すことを含む。このプロセスは、偽正に非常に敏感であるので、９５％未満の信頼スコアを有するフォームがＡｓｃｅｎｔによって検閲のために人間にルーティングされる。この検閲は、コンピュータ端末にて手動で実行され、ページのシーケンスに割り当てられた適切なフォーム（単数または複数）が人間の検閲者によって決定され、その後、ページは、割り当てられたフォームタイプにより処理される。手動のフォーム分離ステップをコンピュータ端末に移動することによって、物理的セパレータページを印刷する必要を除去する。このプロセスについて、年間１，０００，０００ドル以上が節約され得ることが推定される。単一のローンアプリケーションにおけるフォーム間にセパレータシートを物理的に挿入するために約２０秒を要し、かつ、単一のローン処理企業が月間２０００万を越えるローンアプリケーションを受取り得ると想定して、フォームの大多数を自動的に分離することによって節約される人間の時間の量が、より重要ですらあるさらなる節約である。上述の自動フォーム分離システムは、２週間のうちにこのプロジェクトのために実施される。これは、通常、何ヶ月間も測定される、システム

に基づく伝統的ルールを構成するためにかかる時間に対する著しい改善であり、本発明は、任意の自動システムのこのタスクに関してこれまで報告されたよりも著しく正確な結果を示す。

【００８７】

一旦人間の検閲者が、例えば、２５個の十分なページを処理すると、ページごとの確率の推定をより良好に構築するために、分類子が保持される。このことの特に有用な表れは、十分なページが手動で検閲されて、新しいフォームタイプモデルの追加を可能にすることである。これは、自動分類子が将来においてフォームタイプのさらなるカテゴリを処理することを可能にする。ドキュメント、サブドキュメント、またはフォームが識別および分離されると、電子セパレータシートまたはラベルが各フォームタイプを識別するために「挿入され」得る。例えば、これらのセパレータシートは、デジタル化されたドキュメントイメージシーケンスまたはＸＭＬ記述、またはドキュメントまたはサブドキュメントまたは他のアイテムのシーケンスにおける各ページと電子的に関連付けられえる他の電子ラベルの一部分になる実際のコンピュータによって生成されたイメージのフォームであり得る。

【００８８】

本発明は、これまで、銀行のローンドキュメントを線引きおよび識別するという意味合いで説明されてきたが、当業者は、例えば、保険フォーム、納税フォーム、雇用フォーム、健康管理フォーム、請求書等の他のタイプのドキュメントを所望の分類ルールに基づいて処理するという意味合いで方法およびシステムを線引きおよび識別する新規のドキュメントを提供するために、通常の実験以外は用いずに本発明を実現し得る。

【００８９】

本発明により提供されるのは、１つ以上のドキュメントのデジタルイメージを解析し、１つ以上のドキュメント内の１つ以上のページまたはサブドキュメントを自動的に分類し、かつ、異なったカテゴリに属するデジタルイメージ間に挿入された分離ページのコンピュータによって生成されたイメージ、デジタルイメージのカテゴリ化シーケンスの記述、またはデジタルイメージに貼り付けられたか、これに関連したコンピュータによって生成された電子ラベル等の線引き識別子を自動的に生成することによって、ドキュメントの境界に線引きし、かつドキュメントのタイプを識別するための方法およびシステムである。

【００９０】

上述のように、本発明は、分類および／または確率ルールと、カスタムメイドフィルタルールとの組み合わせを用いて、目的のドキュメント、サブドキュメント、または他のアイテムの自動的分離を確実かつ効率的に実行するための改善された方法およびシステムを提供する。本発明の好ましい実施形態を用いて本発明を例示してきたが、本発明は、この実施形態に限定して解釈されるべきものではない。本発明は、特許請求の範囲によってのみその範囲が解釈されるべきであることが理解される。当業者は、本発明の具体的な好ましい実施形態の記載から、本発明の記載および技術常識に基づいて等価な範囲を実施することができることが理解される。本明細書において引用した特許、特許出願および文献は、その内容自体が具体的に本明細書に記載されているのと同様にその内容が本明細書に対する参考として援用されるべきであることが理解される。

【図面の簡単な説明】

【００９１】

【図１】図１は、本発明により用いられ得る例示的コンピュータシステムを表すブロック図を示す。

【図２】図２は、従来技術のドキュメント分離手順のプロセスフローチャートを示す。

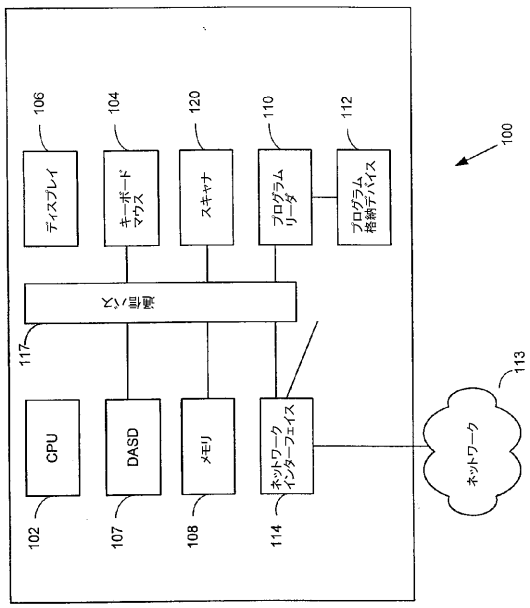
【図３】図３は、本発明の１実施形態による、ドキュメント分離手順のプロセスフローチャートを示す。

【図４】図４は、本発明の別の実施形態による、ドキュメント分離手順のプロセスフローチャートを示す。

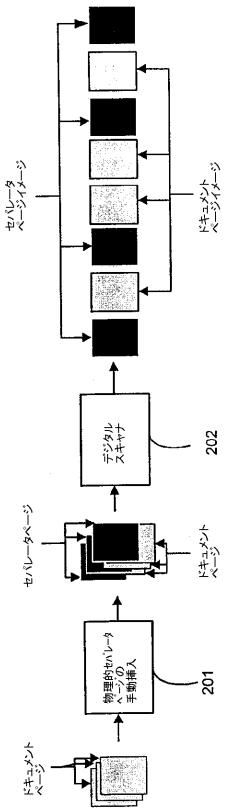
【図５】図５は、本発明の１実施形態による、３つの異なった形態またはドキュメントタ

イプを分離するための例示的有限状態機械図を示す。
【図６】図６は、本発明のさらなる実施形態による、例示的有限状態機械図を示す。
【図７】図７は、本発明の別の実施形態による、例示的有限状態機械図を示す。
【図８】図８は、図５の有限状態機械が与えられた、６つの入力イメージの形態のシーケンスのすべての可能な組み合わせを表す図を提供する。
【図９】図９は、本発明の１実施形態による、図７のフィルタ変換器を図５の変換器に適用した後の、形態のシーケンスの可能な組み合わせを表す図を提供する。
【符号の説明】
【００９２】
１００　コンピュータシステム
１０２　ＣＰＵ
１０４　キーボードおよびマウス
１０６　ディスプレイ
１０７　ＤＡＳＤ
１０８　メモリ
１１０　プログラムリーダ
１１２　プログラム格納デバイス
１１３　ネットワーク
１１４　ネットワークインターフェース
１１７　通信バス

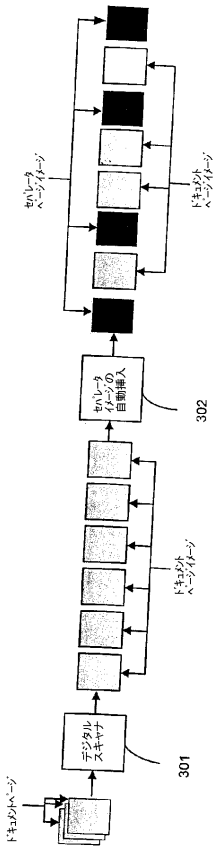
【図１】



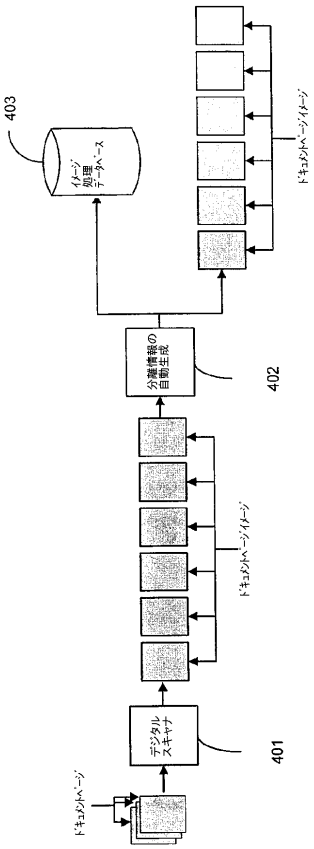
【図２】



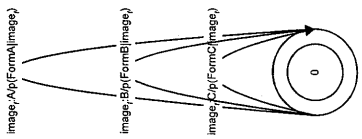
【図３】


ドキュメントページ
デジタルスキャナ
301
ドキュメントページイメージ
セパレータページの自動挿入
302
セパレータページイメージ
ドキュメントページイメージ

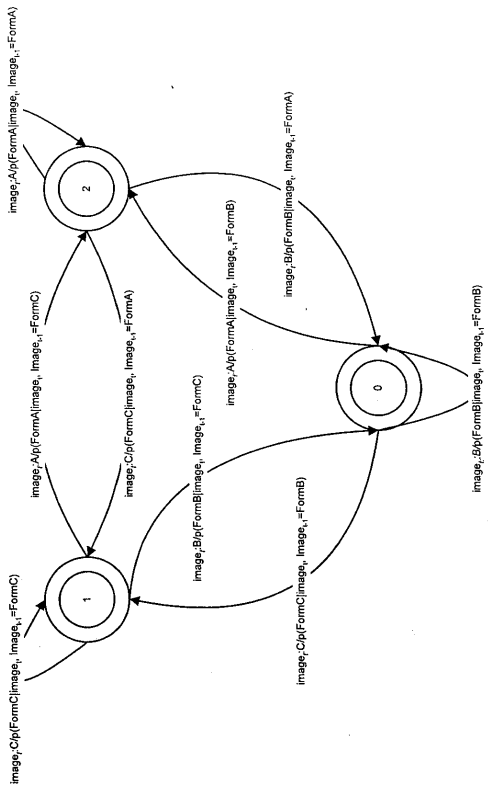
【図４】


403
イメージ設置データベース
ドキュメントページイメージ
分離情報の自動生成
402
ドキュメントページイメージ
デジタルスキャナ
401
ドキュメントページ

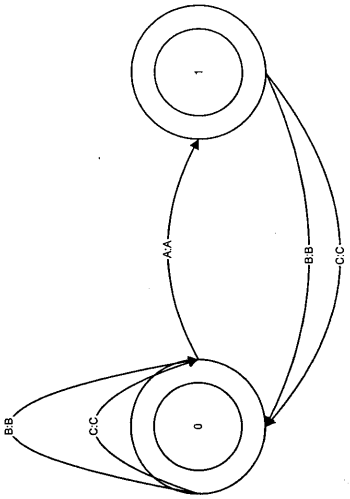
【図５】


image_i:A/p(FormA|image_i)
image_i:B/p(FormB|image_i)
image_i:C/p(FormC|image_i)
0

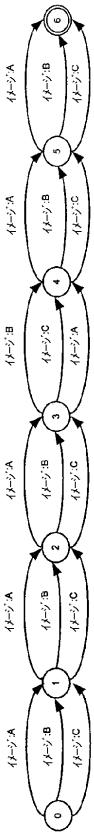
【図６】


image_i:A/p(FormA|image_i, Image_i+1=FormA)
image_i:A/p(FormA|image_i, Image_i+1=FormC)
image_i:C/p(FormC|image_i, Image_i+1=FormA)
image_i:B/p(FormB|image_i, Image_i+1=FormA)
image_i:A/p(FormA|image_i, Image_i+1=FormB)
image_i:B/p(FormB|image_i, Image_i+1=FormC)
image_i:C/p(FormC|image_i, Image_i+1=FormB)
image_i:C/p(FormC|image_i, Image_i+1=FormC)
image_i:B/p(FormB|image_i, Image_i+1=FormB)
2
1
0

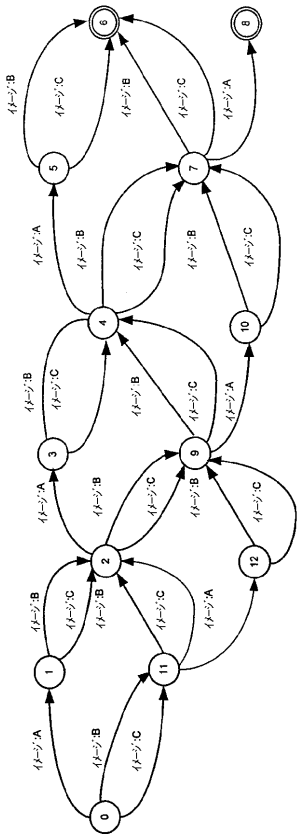
【図７】



【図８】



【図９】



フロントページの続き

(72)発明者 モウリティアス エイ. アール. シュミットラー
アメリカ合衆国 カリフォルニア ９２０２９, エスコンディド, クラレンス レーン ５５８
(72)発明者 スコット スチュワート テキセイラ
アメリカ合衆国 カリフォルニア ９２１２９, サン ディエゴ, ビア クレスタ ロード ７３３７
(72)発明者 クリストファー ケイ. ハリス
アメリカ合衆国 カリフォルニア ９２１１６, サン ディエゴ, ノース アベニュー ４４３６
(72)発明者 サミーヤ サマット
アメリカ合衆国 カリフォルニア ９２１２９, サン ディエゴ, ピピット プレイス ７７９０
(72)発明者 ローランド ボレイ
アメリカ合衆国 カリフォルニア ９２８６１, ビラ パーク, ロマ ストリート ９１３１
(72)発明者 アンソニー マッキオラ
アメリカ合衆国 カリフォルニア ９１７０９, チノ ヒルズ, グレン リッジ ドライブ １５２８３

審査官 相澤 祐介

(56)参考文献 特開２０００－３５４１４４（ＪＰ, Ａ）
特開２００２－０２４２５８（ＪＰ, Ａ）
特開２０００－０６７０６５（ＪＰ, Ａ）
特開２００３－０９１５２１（ＪＰ, Ａ）
特開２００２－３１２３８５（ＪＰ, Ａ）

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｔ １／００、７／００
Ｈ０４Ｎ １／３８７