(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2010-509656
(P2010-509656A)
(43)公表日　平成22年3月25日(2010.3.25)

(51)Int.Cl.		F I			テーマコード（参考）
G06K 9/20	(2006.01)	G06K	9/20	340C	5B029
G06F 17/30	(2006.01)	G06F	17/30	220Z	5B075
G06F 17/21	(2006.01)	G06F	17/21	530A	5B109

審査請求 有　予備審査請求 未請求　（全 39 頁）

(21)出願番号　特願2009-535346 (P2009-535346)
(86)(22)出願日　平成19年11月5日 (2007.11.5)
(85)翻訳文提出日　平成21年6月29日 (2009.6.29)
(86)国際出願番号　PCT/US2007/023233
(87)国際公開番号　WO2008/057473
(87)国際公開日　平成20年5月15日 (2008.5.15)
(31)優先権主張番号　11/592,268
(32)優先日　平成18年11月3日 (2006.11.3)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　11/644,009
(32)優先日　平成18年12月22日 (2006.12.22)
(33)優先権主張国　米国 (US)

(71)出願人　502208397
グーグル　インコーポレイテッド
アメリカ合衆国　カリフォルニア州　９４０４３　マウンテン　ビュー　アンフィシアター　パークウェイ　１６００
(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
(72)発明者　ファーマニアーク，　ラルフ
カナダ国　エヌ５エックス－３ティー９
オンタリオ，　ロンドン，　ウィンダミア　ロード　４３－６８３

最終頁に続く

(54)【発明の名称】連続する記事部分の媒体資料解析

(57)【要約】
本発明は、複数のページにわたって連続する記事を有する媒体資料を解析するシステムおよび方法に関する。媒体資料アナライザは、セグメンタと記事コンポーザとを含む。セグメンタは、媒体資料内のコラム状の本文テキストに関連するブロックセグメントを識別する。記事コンポーザは、言語統計情報および連続移行情報に基づいて、識別されたブロックセグメントのうちのいずれが、媒体資料内の複数のページにわたって広がる連続する記事に属するかを決定する。

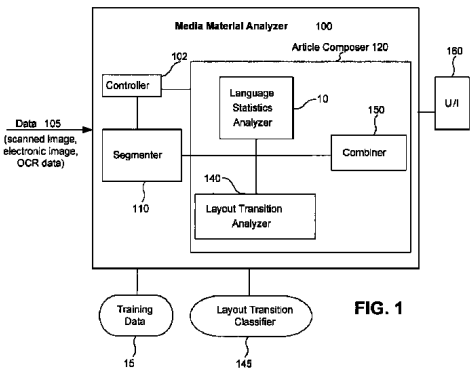


FIG. 1

【特許請求の範囲】
【請求項１】
レイアウトを有し、かつ複数のページにわたって広がる１つ以上の連続する記事を含む媒体資料を表すデータを解析する媒体資料アナライザであって、
（ａ）該媒体資料のページ内のコラム状の本文テキストと関連するブロックセグメントを識別するセグメンタと、
（ｂ）言語統計情報および連続移行情報に基づいて、該識別されたブロックセグメントのいずれが該媒体資料内の複数のページにわたって広がる連続する記事に属するかを決定する記事コンポーザと
を備えている、媒体資料アナライザ。
【請求項２】
前記記事コンポーザは、連続レイアウト移行アナライザを含み、該連続レイアウト移行アナライザは、第１のページ内の候補となる連続する記事部分に関連する最後のブロックセグメントを識別し、該最後のブロックセグメントの下の１つ以上の項目を識別し、該識別された１つ以上の項目の少なくとも１つの特徴に基づいて、該最後のブロックセグメントを分類し、そして、決定ツリーを適用して、該最後のブロックセグメントが連続する記事内にある確率を示す１つ以上の連続移行特徴を選び出す、請求項１に記載の媒体資料アナライザ。
【請求項３】
前記記事コンポーザは、連続言語統計アナライザを含み、該連続言語統計アナライザは、前記連続する記事の前記第１のページから連続するページにおける記事部分に対して、言語統計情報を計算し、そして該計算された連続する言語統計情報に基づいて、前記候補となる連続する記事部分内の最後のブロックセグメントが、連続する記事部分を有する確率を決定する、請求項２に記載の媒体資料アナライザ。
【請求項４】
前記連続レイアウト移行アナライザは、さらに、連続するページ内の候補となる連続する記事と関連する第１のブロックセグメントを識別し、該第１のブロックセグメントの上の１つ以上の項目を識別し、該１つ以上の識別された項目の少なくとも１つの特徴に基づいて、該第１のブロックセグメントを分類し、そして決定ツリーを適用して、該第１のブロックセグメントが連続する記事内にある確率を示す１つ以上の連続移行特徴を選び出し、該適用された決定ツリーに基づいて、該第１のブロックセグメントが連続する記事である確率を決定する、請求項３に記載の媒体資料アナライザ。
【請求項５】
前記連続言語統計アナライザは、さらに、前記第１のブロックセグメントを有するページよりも前のページ内の記事部分に対して、言語統計情報を計算し、該計算された、より前のページの言語統計情報に基づいて、前記候補となる連続する記事部分内の該第１のブロックセグメントが連続する記事部分を有する確率を決定する、請求項４に記載の媒体資料アナライザ。
【請求項６】
前記連続レイアウト移行アナライザは、さらに、候補となる一対の最後および第１のブロックセグメントに対する連続移行特徴を識別し、該一対の最後および第１のブロックセグメントに対して一組の連続移行特徴を決定し、そして決定ツリーを適用して、該一組の決定された連続移行特徴に基づいて、該候補となる一対の最後および第１のブロックセグメントが、前記媒体資料内の複数のページにわたる同一の連続する記事に属する確率を決定する、請求項５に記載の媒体資料アナライザ。
【請求項７】
前記言語統計情報は、単語頻度情報を備え、前記連続言語統計アナライザは、前記最後のブロックセグメント内のテキストおよび連続するページ上の前記記事部分内のテキストにおける単語頻度に基づいて、マッチスコアを計算する、請求項３に記載の媒体資料アナライザ。

【請求項８】
前記言語統計情報は、単語頻度情報を備え、前記連続言語統計アナライザは、前記第１のブロックセグメント内のテキストおよびより前のページ上の前記記事部分内のテキストにおける単語頻度に基づいて、マッチスコアを計算する、請求項５に記載の媒体資料アナライザ。
【請求項９】
レイアウトを有し、かつ複数のページにわたって広がる１つ以上の連続する記事を含む媒体資料を表すデータを解析するコンピュータ実装された方法であって、
（ａ）該媒体資料のページ内のコラム状の本文テキストに関連するブロックセグメントを識別することと、
（ｂ）言語統計情報および連続移行情報に基づいて、該識別されたブロックセグメントのいずれが該媒体資料内の複数のページにわたって広がる連続する記事に属するかを決定することと
を包含する、方法。
【請求項１０】
前記記事決定ステップは、
候補となる連続する記事部分内のブロックセグメントに対する連続レイアウト移行情報を解析することと、
該候補となる連続する記事部分内のテキストに対する言語統計を解析することと
を含む、請求項９に記載の方法。
【請求項１１】
前記連続レイアウト移行情報を解析するステップは、
第１のページ内の候補となる連続記事部分に関連する最後のブロックセグメントを識別することと、
該最後のブロックセグメントより下の１つ以上の項目を識別することと、
該識別された１つ以上の項目の少なくとも１つの特徴に基づいて、該最後のブロックセグメントを分類することと、
決定ツリーを適用して、該最後のブロックセグメントが連続する記事内に存在する確率を示す１つ以上の連続移行特徴を選ぶことと
を包含する、請求項１０に記載の方法。
【請求項１２】
前記言語統計解析ステップは、
前記連続する記事の第１のページから連続するページ上の記事部分に対する、言語統計情報を計算することと、
該計算された連続する言語統計情報に基づいて、前記候補となる連続する記事部分内の前記最後のブロックセグメントが連続する記事部分を有する確率を決定することと
を包含する、請求項１１に記載の方法。
【請求項１３】
前記連続レイアウト移行情報を解析するステップは、
連続するページ内の候補となる連続する記事に関連する第１のブロックセグメントを識別することと、
該第１のブロックセグメントより上の１つ以上の項目を識別することと、
該１つ以上の識別された項目の少なくとも１つの特徴に基づいて、該第１のブロックセグメントを分類することと、
決定ツリーを適用して、該第１のブロックセグメントが連続する記事内に存在する確率を示す１つ以上の連続移行特徴を選び出し、そして該適用された決定ツリーに基づいて、該第１のブロックセグメントが連続する記事内に存在する確率を決定することと
をさらに包含する、請求項１２に記載の方法。
【請求項１４】
前記言語統計解析ステップは、

前記第１のブロックセグメントを有するページよりも前のページ内の記事部分に対して、言語統計情報を計算することと、
該計算されたより前のページの言語統計情報に基づいて、前記候補となる連続する記事部分内の該第１のブロックセグメントが連続する記事部分を有する確率を決定することと
をさらに包含する、請求項１３に記載の方法。
【請求項１５】
前記連続レイアウト移行解析ステップは、
候補となる一対の最後および第１のブロックセグメントを識別することと、
該一対の最後および第１のブロックセグメントに対する一組の連続移行特徴を決定することと、
決定ツリーを適用して、該一組の決定された連続移行特徴に基づいて、該候補となる一対の最後および第１のブロックセグメントが、前記媒体資料内の複数のページにわたる同一の連続する記事に属する確率を決定することと
をさらに含む、請求項１２に記載の方法。
【請求項１６】
前記言語統計情報は、単語頻度情報を備え、前記連続言語統計解析ステップは、前記最後のブロックセグメント内のテキストおよび連続するページの前記記事部分内のテキストにおける単語頻度に基づいて、マッチスコアを計算することを含む、請求項１２に記載の方法。
【請求項１７】
前記言語統計情報は、単語頻度情報を備え、前記連続言語統計解析ステップは、前記第１のブロックセグメント内のテキストおよび以前のページの前記記事部分内のテキストにおける単語頻度に基づいて、マッチスコアを計算することを含む、請求項１４に記載の方法。
【請求項１８】
レイアウトを有する媒体資料内の複数のページにわたって広がる連続する記事を構成する記事コンポーザであって、
連続レイアウト移行アナライザと、
連続言語統計アナライザと
を備え、該連続レイアウト移行アナライザは、異なるページ上の候補となる記事の最後のブロックセグメントおよび第１のブロックセグメントが同じ連続する記事内に存在する確率を示す１つ以上の連続移行特徴を選び出すために、決定ツリーを適用し、
該連続言語統計アナライザは、異なるページ上の異なる記事部分に対する言語統計情報を計算し、該計算された言語統計情報に基づいて、候補となる記事部分の第１および最後のブロックセグメントが連続する記事部分を有する確率を決定し、それにより、該記事コンポーザは、解析された連続レイアウト移行特徴および該計算された言語統計に従って、該第１および最後のブロックセグメントが同じ連続する記事に属する確率に基づいて、複数のページにわたる連続する記事を構成することが可能である、記事コンポーザ。
【発明の詳細な説明】
【技術分野】
【０００１】
本願は、２００６年１１月３日出願の米国出願第１１／５９２，２６８号（代理人整理番号第２５２５．００１００００）の一部継続出願であり、該米国出願の全内容は本明細書において参照により援用される。
【０００２】
（発明の分野）
本発明は、媒体資料のコンピュータ補助による解析に関する。
【背景技術】
【０００３】
（発明の背景）

（関連技術）
文書および印刷された資料の解析を行うか、または補助するために、コンピュータがますます使用されている。レイアウト解析技術およびシステムは、文書中のテキストおよび画像の位置および相対的な配列を解析するために使用されてきた。このような文書レイアウト解析は、多くの文書画像化用途において重要であり得る。例えば、文書レイアウト解析は、レイアウトベースの文書検索、光学文字認識を用いたテキスト抽出、および文書画像の電子形式への変換の一部として、使用され得る。文書レイアウト解析は、概して、単純な文書（例えば、ビジネスレターまたは一列の報告書）において最良に機能し、かつレイアウトが複雑であるか、または可変であるときには、困難であり得るかまたは機能不可能でさえあり得る。例えば、自動の文書レイアウト解析または半自動の文書レイアウト解析は、しばしば、複雑なレイアウトに分類され、かつ再ソートがレイアウトの手動解析に対してなされなければならない。

【０００４】
レイアウトに配列された本文のテキストのコラムを有する媒体資料は、文書レイアウト解析に対する特別な挑戦を生み出す。例えば、新聞のレイアウトは、概して、非常に複雑であり、多くの記事と論理的な要素とがページ上で一緒に接近して組み合わせられる。新聞の構造を理解することは、記事の文脈、パターンマッチングおよび可能性としては新聞のスタイル、すなわちコンピュータよりも人間に対して自然な要素によって、人間によって自然に行われる。自動的な方法は、概して、図形的な特徴または幾何学的な特徴のみにほとんど依存しており、その結果、全ての新聞にわたって機能する一貫した一組の単純な罫線がないので、多くの間違いを生じる。このような限定された自動的な方法は、媒体資料の２つ以上のページにわたって連続する記事を解析するさらなる困難性を有する。

【発明の概要】
【発明が解決しようとする課題】
【０００５】
レイアウトを有する媒体資料を解析する向上したシステムおよび方法が必要である。
【課題を解決するための手段】
【０００６】
（発明の概要）
本発明は、レイアウトを有する資料媒体を解析するシステムおよび方法に関する。

【０００７】
一実施形態において、媒体資料アナライザは、セグメンタと記事コンポーザとを含み得る。セグメンタは、媒体資料内のコラム状の本文のテキストと関連するブロックセグメントを識別する。一例において、セグメンタは、画像データ内の画素データを解析して、類似の画素値変化の複雑性（ｐｉｘｅｌ　ｖａｌｕｅ　ｃｈａｎｇｅ　ｃｏｍｐｌｅｘｉｔｙ）を有する領域を識別する。画素値変化は、画素から水平方向および垂直方向に沿って識別される。記事コンポーザは、識別されたブロックセグメントのいずれが媒体資料中の１つ以上の記事に属するかを決定する。記事コンポーザは、言語統計情報、レイアウト移行情報、または言語統計情報およびレイアウト移行情報の両方に基づいて、候補となるブロックセグメントが同一の記事に属するかどうかを決定し得る。

【０００８】
別の実施形態において、記事コンポーザは、言語統計アナライザを含み得る。言語統計アナライザは、言語統計情報に基づいて、セグメンタから出力されたブロックセグメントのいずれが、媒体資料中の１つ以上の記事に属するかを決定する。特に、言語統計アナライザは、セグメンタによって出力された候補となるブロックセグメントに対して、言語統計を計算し、そして、言語統計情報におけるオーバーラップに基づいて、候補となるブロックセグメントが同一の記事に属する確率を決定する。

【０００９】
さらなる実施形態において、記事コンポーザは、レイアウト移行アナライザを含み得る。レイアウト移行アナライザは、セグメンタによって出力された候補となるブロックセグ

メントにおけるレイアウト移行特徴を解析し、そしてレイアウト移行解析に基づいて、候補となるブロックセグメントが媒体資料内の同一の記事に属するかどうかを決定する。一例において、レイアウト移行特徴は、垂直方向および水平方向の移行特徴を含む。

【００１０】

本発明の一局面に従って、コンピュータ実装された方法は、レイアウトを有する媒体資料を表すデータを解析する。この方法は、媒体資料内のコラム状の本体テキストと関連するブロックセグメントを識別することと、言語統計情報およびレイアウト情報に基づいて、識別されたブロックセグメントのいずれが、媒体資料内の１つ以上の記事に属するかを決定することとを含み得る。

【００１１】

さらなる実施形態において、ネットワークを介して（例えば、ウェブを介して）、ブラウザを通して、レイアウトを有する媒体資料を探索するシステムが提供される。ブラウザは、サーチ要求を満たすことにおいて識別された同一の記事内の１つ以上のブロックセグメントからテキストを受信し得る。

【００１２】

さらなる実施形態において、レイアウトを有する媒体資料を表し、かつ複数のページにわたって広がる１つ以上の連続する記事を含むデータを解析する媒体資料アナライザが提供される。媒体資料アナライザは、媒体資料ページ内のコラム状本体テキストに関連するブロックセグメントを識別するセグメンタと、言語統計情報および連続移行情報に基づいて、識別されたブロックセグメントのいずれが、媒体資料内の複数のページにわたって広がる連続する記事に属するかを決定する記事コンポーザとを含む。

【００１３】

なおさらなる実施形態において、レイアウトを有する媒体資料を表し、かつ複数のページにわたって広がる１つ以上の連続する記事を含むデータを解析するコンピュータ実装された方法が提供される。この方法は、媒体資料ページ内のコラム状本体テキストに関連するブロックセグメントを識別することと、言語統計情報および連続移行情報に基づいて、識別されたブロックセグメントのいずれが、媒体資料内の複数のページにわたって広がる連続する記事に属するかを決定することとを含む。

【００１４】

さらに、一実施形態において、レイアウトを有する媒体資料内の複数のページにわたって広がる連続する記事を構成する記事コンポーザは、連続レイアウト移行アナライザと、連続統計アナライザとを含む。連続レイアウト移行アナライザは、１つ以上の連続移行特徴を選び出すために決定ツリーを適用し、この１つ以上の連続移行特徴は、異なるページ上の候補となる記事部分の最後および第１のブロックセグメントが、同一の連続記事に存在する確率を示す。連続言語統計アナライザは、異なるページ上の異なる記事部分に対する言語統計情報を計算し、そして、該計算された統計情報に基づいて、候補となる記事部分の第１および最後のブロックセグメントが連続する記事部分を有する確率を決定する。このようにして、記事コンポーザは、解析された連続レイアウト移行特徴および計算された言語統計に従って、第１および最後のブロックセグメントが同一の連続する記事に属する確率に基づき、複数のページにわたる連続する記事を構成し得る。

【００１５】

本発明のさらなる実施形態、特徴および利点と、本発明の様々な実施形態の構造および動作とが、添付の図面を参照して以下に詳細に記載される。

【００１６】

特許または出願のファイルは、カラーで制作された少なくとも１つの図面を含む。カラーの図面を有する、この特許または特許出願刊行物のコピーは、要求し、必要な料金を支払うと、官庁より提供される。

【００１７】

本発明の実施形態は、添付の図面を参照して記載される。図面において、同様の参照番号は同一の要素または機能的に類似の要素を示し得る。ある要素が第１に現れる図面は、

対応する参照番号のもっとも左の桁によって概して示される。

【図面の簡単な説明】

【００１８】

【図１】図１は、本発明の実施形態に従う、媒体資料アナライザの図である。

【図２】図２は、本発明の実施形態に従う、媒体資料を解析する方法の図である。

【図３】図３は、図２の方法における、ブロックセグメント識別ステップを実行する例示的なルーチンを示す図である。

【図４】図４は、図３のルーチンに従って識別されたブロックセグメントを有する媒体資料の画像を示し、カラーを含む、図である。

【図５】図５は、本発明のさらなる実施形態に従う、ＯＣＲデータからテキストを抽出し、そしてブロックセグメント領域を調整するステップを示すフローチャート図である。

【図６】図６は、本発明の一実施形態に従う、言語統計を解析して、記事内のブロックセグメントを識別する方法を示す図である。

【図７Ａ】図７Ａおよび図７Ｂは、本発明の一実施形態に従う、訓練モードにおいてレイアウト移行アナライザの動作を示すフローチャート図である。図７Ａは、垂直方向の移行特徴を決定するために訓練モードにおいて動作する方法を示す。図７Ｂは、水平方向の移行特徴に基づいて、訓練モードにおいて動作する方法を示す。

【図７Ｂ】図７Ａおよび図７Ｂは、本発明の一実施形態に従う、訓練モードにおけるレイアウト移行アナライザの動作を示すフローチャート図である。図７Ａは、垂直方向の移行特徴を決定するために訓練モードにおいて動作する方法を示す。図７Ｂは、水平方向の移行特徴に基づいて、訓練モードにおいて動作する方法を示す。

【図８】図８は、本発明の一実施形態に従う、レイアウト移行分類子を訓練し、かつ構築するために用いられ得る、候補となるブロックセグメントを示す例示的な媒体資料である。

【図９Ａ】図９Ａおよび図９Ｂは、本発明の一実施形態に従う、実行モードにおけるレイアウト移行アナライザの動作を示すフローチャート図である。図９Ａは、媒体資料レイアウト内の垂直方向移行特徴に基づいた実行モード動作を示す。図９Ｂは、レイアウト内の水平方向移行特徴に基づいた実行モード動作を示す。

【図９Ｂ】図９Ａおよび図９Ｂは、本発明の一実施形態に従う、実行モードにおけるレイアウト移行アナライザの動作を示すフローチャート図である。図９Ａは、媒体資料レイアウト内の垂直方向移行特徴に基づいた実行モード動作を示す。図９Ｂは、レイアウト内の水平方向移行特徴に基づいた実行モード動作を示す。

【図１０Ａ】図１０Ａ～図１０Ｄは、カラーを含み、本発明の例示的な実施形態に従って解析された新聞のページを含む例示的な媒体資料を示す。

【図１０Ｂ】図１０Ａ～図１０Ｄは、カラーを含み、本発明の例示的な実施形態に従って解析された新聞のページを含む例示的な媒体資料を示す。

【図１０Ｃ】図１０Ａ～図１０Ｄは、カラーを含み、本発明の例示的な実施形態に従って解析された新聞のページを含む例示的な媒体資料を示す。

【図１０Ｄ】図１０Ａ～図１０Ｄは、カラーを含み、本発明の例示的な実施形態に従って解析された新聞のページを含む例示的な媒体資料を示す。

【図１１】図１１は、本発明のさらなる実施形態に従う、Ｗｏｒｌｄ Ｗｉｄｅ Ｗｅｂを介して、レイアウトを有する媒体資料をサーチするシステムを示す図である。

【図１２】図１２は、本発明の一実施形態に従う、媒体資料アナライザによって解析されたデータのサーチにおけるサーチ結果の例示的な表示を示す図である。

【図１３】図１３は、本発明のさらなる実施形態に従う、連続する記事部分を解析し得る媒体資料アナライザの図である。

【図１４Ａ】図１４Ａ～図１４Ｅは、本発明の一実施形態に従う、図１３の媒体資料アナライザにおける記事コンポーザの動作を示すフローチャート図である。

【図１４Ｂ】図１４Ａ～図１４Ｅは、本発明の一実施形態に従う、図１３の媒体資料アナライザにおける記事コンポーザの動作を示すフローチャート図である。

【図１４Ｃ】図１４Ａ～図１４Ｅは、本発明の一実施形態に従う、図１３の媒体資料アナライザにおける記事コンポーザの動作を示すフローチャート図である。
【図１４Ｄ】図１４Ａ～図１４Ｅは、本発明の一実施形態に従う、図１３の媒体資料アナライザにおける記事コンポーザの動作を示すフローチャート図である。
【図１４Ｅ】図１４Ａ～図１４Ｅは、本発明の一実施形態に従う、図１３の媒体資料アナライザにおける記事コンポーザの動作を示すフローチャート図である。
【図１５】図１５は、本発明の実施形態を実装するために使用され得る例示的なコンピュータシステムの図である。
【発明を実施するための形態】
【００１９】
【表１】

目次

（実施形態の詳細な説明）
本発明は、特定の用途に対する例示的な実施形態を参照して本明細書に記載されるが、本発明がそれらの実施形態に限定されないことが理解されるべきである。本明細書において提供される教示を利用する機会を有する当業者は、本発明の範囲内のさらなる修正、用途および実施形態、ならびに本発明が大いに有効であるさらなる分野を認識するだろう。
【００２０】
（概観）
本発明は、レイアウトを有する媒体資料を解析するシステムおよび方法に関する。例として、本文テキストのコラムを伴うレイアウトを有する媒体資料を含むが、それに限定はされない。このような例は、新聞、雑誌、カタログ、小冊子、パンフレットおよび他のタイプの印刷資料を含むがこれらに限定はされない。
【００２１】
（媒体資料アナライザ）

図１は、本発明の実施形態に従う媒体資料アナライザ１００を示す。媒体資料アナライザ１００は、コントローラ１０２と、セグメンタ１１０と、記事コンポーザ１２０とを含む。記事コンポーザ１２０は、純粋な言語統計モード、純粋なレイアウト移行モードまたは二つの組み合わせにおいて動作し得る。

【００２２】

図１に示される実施形態において、記事コンポーザ１２０は、言語統計アナライザ１３０と、レイアウト移行アナライザ１４０と、コンバイナ１５０とを含む。媒体資料アナライザ１００は、データ１０５、訓練データ１３５およびレイアウト移行分類子（ｃｌａｓｓｉｆｉｅｒ）１４５を受信し得るか、またはそれらにアクセスし得る。媒体資料アナライザ１００はまた、ユーザインターフェース１６０に連結され得る。

【００２３】

データ１０５は、媒体資料の画像データを含み得る。このような画像データは、電子的またはスキャンされた画像データと、画像データから抽出された光学文字認識（ＯＣＲ）データとを含み得る。データ１０５は、任意のタイプのファイルフォーマットで提供され得る。

【００２４】

訓練データ１３５は、媒体資料内の記事に属するブロックセグメントのポジティブな例およびネガティブな例を含み得る。レイアウト移行分類子１４５は、候補となるブロックセグメントが媒体資料内の記事に属するように分類されることを可能にする移行特徴情報を含むデータ構造を含むがこれらに限定されない。このようなデータ構造は、決定ツリーを含み得るが、これに限定はされない。

【００２５】

セグメンタ１１０は、データ１０５内の媒体資料内のコラム状の本文テキストに関連するブロックセグメントを識別する。記事コンポーザ１２０は、言語統計情報および／またはレイアウト移行情報に基づいて、識別されたブロックセグメントのうちのいずれが媒体資料のうちの１つ以上の記事に属することを決定する。

【００２６】

一実施形態において、言語統計アナライザ１３０は、セグメンタ１１０によって出力された候補となるブロックセグメントに対する言語統計を計算する。次いで、言語統計アナライザ１３０は、言語統計情報のオーバーラップに基づいて、候補となるブロックセグメントが同一の記事に属する確率を決定する。

【００２７】

レイアウト移行アナライザ１４０は、レイアウト移行特徴およびセグメンタ１１０によって出力された候補となるブロックセグメントをさらに解析する。次いで、レイアウト移行アナライザ１４０は、レイアウト移行特徴に基づいて、候補となるブロックセグメントが媒体内の同一の記事に属するかどうかを決定する。

【００２８】

コンバイナ１５０は、言語統計アナライザ１３０およびレイアウト移行アナライザ１４０によって解析された候補となるブロックセグメントが同一の記事に属するかどうかを識別する。一例において、コンバイナ１５０は、言語統計アナライザ１３０によって決定された確率と、レイアウト移行アナライザ１４０から出力されたレイアウト移行特徴に基づく、ブロックが同一の記事に属するかどうかという決定との両方の出力に基づいて、候補となるブロックセグメントが同一の記事に属するかどうかを識別する。

【００２９】

あるいは、コンバイナ１５０は、言語統計アナライザ１３０のみによって決定された確率に基づいて、同一の記事に属する候補となるブロックセグメントを識別し得る。コンバイナ１５０は、また、レイアウト移行アナライザ１４０のみによって解析されたレイアウト移行特徴に基づいて、候補となるブロックセグメントが同一の記事に属するかどうかを決定し得る。

【００３０】

コントローラ１０２は、セグメンタ１１０と記事コンポーザ１２０とを制御および管理する。ユーザからのさらなる制御は、ユーザインターフェース１６０を介して提供され得る。例えば、ユーザは、動作を開始し得るか、またはデータ１０５、訓練データ１３５もしくはレイアウト移行分類子１４５の入力を開始し得る。ユーザは、媒体資料アナライザ１００と相互作用して、訓練データ１３５の生成またはレビューを助け得る。例えば、ユーザは、訓練データ１３５の質を向上させるために、所与の媒体資料内の記事に属するブロックセグメントのポジティブな例とネガティブな例とを選択し得る。ユーザはまた、レイアウト移行分類子１４５を構築または修正するために、媒体資料アナライザ１００と相互作用し得る。

【００３１】

媒体資料アナライザ１００はまた、スキャンされたデータ１０５または媒体資料アナライザ１００から出力されたデータの画像を表示し得る。表示のための出力データは、媒体資料アナライザ１００の解析に従って構成されたハイライトされたブロックセグメントを示すために解析された媒体資料の表示を含み得る。特定のレイアウトに対して、ユーザはフィードバックを提供し得るか、またはハイライトされたブロックセグメントを選択し得る。他のタイプの情報は、この記載を提供された当業者に対して明らかであるように、表示され得る。

【００３２】

媒体資料アナライザ１００（その構成要素モジュールを含む）は、ソフトウェア、ファームウェア、ハードウェアまたはそれらの任意の組み合わせにおいて実装され得る。媒体資料アナライザ１００は、コンピュータ、ワークステーション、分散コンピューティングシステム、埋め込みシステム、スタンドアロン電子デバイス、ネットワーク化されたデバイス、モバイルデバイス、セットトップボックス、テレビ、あるいは他のタイプのプロセッサまたはコンピュータシステムを含むが、これらに限定されない任意のタイプの処理デバイス上で実行するように実装され得る。

【００３３】

媒体資料アナライザ１００は、また、種々の用途において使用され得る。データ１０５において自動的に、または半自動的に動作させることによって、媒体資料アナライザ１００は、格納された画像データ（例えば、アーカイブされた媒体資料）を解析し得る。マイクロフィッシュ、フィルムおよび他のストレージ媒体は、入力用に画像データを取得するためにスキャンされ得る。任意のファイルフォーマットの電子ファイルがまた入力され得る。解析は、ユーザからの最小の入力で、またはユーザからの入力なしで、自動的に、または半自動的に実行され得る。このようにして、媒体資料アナライザ１００は、種々の媒体に対するブロックセグメントから成り立つ記事を構成するために使用され得る。次いで、媒体資料アナライザ１００は、レイアウトを有する媒体資料内の記事から成り立つテキストデータのブロックセグメントを出力し得る。このような出力は、媒体資料のコンテンツをレビューまたはサーチすることを望むローカルユーザおよびリモートユーザに、配信され得るか、または格納され得る。

【００３４】

自動的に、または半自動的に動作することによって、複数の媒体資料アナライザ１００は、大量の媒体資料を解析するために、使用され得、かつ縮尺が合わせられ得る。このようにして、媒体資料レイアウト内のコンテンツは、広範囲のユーザに対して、ローカルに、そしてネットワークを介してリモートに利用可能にし得る。媒体資料アナライザ１００は、ユーザが、図書館、大学、政府機関、会社および他の場所において、ローカルにまたはリモートにアクセスされる媒体資料内のテキストデータをレビューすることを可能にし得る。媒体資料アナライザ１００は、サーチエンジン、ウェブポータルまたは他のウェブサイトと共に使用され得、リモートユーザが、レイアウトを有する媒体資料をレビューおよびサーチすることを可能にする。

【００３５】

媒体資料アナライザ１００、ならびにセグメンタ１１０および記事コンポーザ１２０を

含む構成要素の動作は、図２に示される媒体資料データを解析する方法に関して、以下にさらに詳細に記載される。

【００３６】

（媒体資料アナライザの動作）

さらなる実施形態に従って、媒体資料からのデータを解析する方法２００が提供される（図２）。簡潔さのために、方法２００は、媒体資料アナライザ１００を参照して記載されるが、必ずしも媒体資料アナライザ１００の構造に限定されることを意図されない。

【００３７】

（ブロックセグメント化）

ステップ２１０において、特徴に従って、媒体資料内のコラム状の本文テキストと関連するブロックセグメントが識別される。図３は、ステップ２１０のブロックセグメント化を実行する例示的なルーチンをさらに詳細に示す（ステップ３０５～ステップ３２０）。明確さのために、ブロックセグメント化ルーチンは、図４に示されるように、新聞のページの例示的な画像４１０に関して記載される。この例において、データ１０５は、画像４１０を表す画素データを含む。画素データは、特定のピクチャ要素（画素）位置における画像の強度を表す画素値からなる。画素値は、グレイスケール、カラー、バイナリまたは他のタイプの画素データを含むが、これらに限定されない任意のタイプの画素値であり得る。

【００３８】

ループ３０５において、セグメンタ１１０は、類似の画素値変化複雑性を有する領域を識別するために、画像データ内の画素を解析する。画素の全てまたは画素のサンプルが解析され得る。セグメンタ１１０は、解析されている各画素から、水平方向および垂直方向に沿った画素値変化を解析する。類似の画素値変化複雑性のこれらの領域は、ブロックセグメントを含み得る。特に、のど空き（ｇｕｔｔｅｒ）および他の境界を有するレイアウトにおいて配列されたコラム状の本文テキストを覆う領域に対して、本文テキストのブロックセグメントが得られる。

【００３９】

一例において、セグメンタ１１０は、バイナリ（純粋な黒および白の）画像における一致する複雑性の領域を見出すテクスチャ方法を実行する。画像内の各画素位置に対して、セグメンタ１１０は、カラーが両方の側でｎ回変化するまで、水平方向（左および右）を見る必要がある距離のログを計算する。テキストの領域は、相対的に一致する小さい値を有するが、のど空きおよび他の単純な領域はより大きい値を有する。ｎ＝１に設定することは、例示的な実行長のアプローチを与える。図４の例において、ｎ＝２が用いられ、ページにわたって水平方向に広がるブロックセグメントという結果をもたらす（画像４２０を参照）。ｎに対するより大きな値は、よりむらのない領域をもたらすが、同様にカットオフされる境界をもたらす。この計算は、垂直方向において繰り返され、ｎ＝２のとき、ページにわたって垂直方向に広がるブロックセグメントをもたらす（画像４３０を参照）。記事内の支配的な本文テキスト（例えば、ページの見出しではなく記事内の本文テキスト）を探すとき、水平方向および垂直方向の計算は、最終的なセグメント化された画像を得るために、共に加算され得、次いで、閾値化され得る（カラー画像４４０を参照）。さらなる例において、異なるサイズのテキストの領域を区別するために見る場合には、セグメンタ１１０は、最小値から開始し得、許容値でｆｌｏｏｄ－ｆｉｌｌを行い得る。

【００４０】

データ１０５は、また、媒体資料レイアウトの画像データに関連するテキストデータを含み得る。例えば、媒体資料内のテキストを表す光学文字認識（ＯＣＲ）データが提供され得る。あるいは、媒体資料アナライザ１００はＯＣＲモジュール（図示せず）を含み得、このＯＣＲモジュールは、スキャンされた画像データまたは電子画像データに関連するＯＣＲデータを生成する。

【００４１】

ループ３１５において、画像データから抽出されたＯＣＲテキストデータに対して、ス

テップ３１０において、セグメンタ１１０は、テキストデータを、類似の画素値変化複雑性を有するように識別された対応する画像領域と関連付ける。セグメンタ１１０はまた、テキストデータ内のテキストサイズを識別し得、特に、コラム状の本文テキストに関連する本文テキストサイズを識別し得る。このようにして、本文テキストサイズを有する記事に属する候補であるテキストデータのブロックセグメントが識別され得る。

【００４２】

図５に示されるさらなる実施形態において、ループ３１５’（ステップ５１０～ステップ５５０）は、ループ３１５の代わりに用いられる。ステップ５１０において、セグメンタ１１０は、単語を見出すために、画像データから抽出されたＯＣＲテキストデータを解析する。セグメンタ１１０は、テキストデータ内で見出された単語を、類似の画素値変化複雑性を有するとして、ステップ３１０において識別された初期の組の領域にマッピングする。セグメンタ１１０は、領域のテキストデータにおけるテキストサイズを決定し、特に、コラム状の本文テキストと関連する本文テキストサイズを識別し得る（ステップ５３０）。セグメンタ１１０は、どの領域がコラム状の本文テキストと関連するかを決定する（ステップ５４０）。

【００４３】

次いで、セグメンタ１１０は、コラム状の本文テキスト（本文セグメントとも呼ばれる）を有する初期の組の領域を調整し、マッピングされた単語の分布に基づいて、最終の組の画像領域を得る。ＯＣＲセグメント化が良好であるとき、この調整は、画素値変化および閾値解析をただ用いることではなく、ＯＣＲによって見出された領域における値の分布を見ることによって、問題のあるレイアウトエリアを固定することに役立ち得る。

【００４４】

特徴に従って、セグメンタ１１０は、コラム状の本文テキストを対応する最終の組の画像領域とさらに関連付ける。例えば、どのテキストが本文テキストであるかを決定するために、テキストサイズを、所与の許容値を有する支配的なテキストサイズと比較する比較がなされ得る。

【００４５】

画素値変化に基づくセグメント化は、例示的に上記され、本発明を必ずしも限定することを意図されない。現在公知の、または将来的に開発される他のセグメント化技術がデータ１０５をセグメント化するために使用され得る。

【００４６】

（記事構成）

ステップ２２０において、特徴に従って、記事コンポーザ１２０は、言語統計情報およびレイアウト移行情報に基づいて、いずれの候補となるブロックセグメントが同一の記事に属するかを決定する。この言語統計およびレイアウト移行の組み合わせは、片方のみで達成し得る精度を超えて、精度を向上させ得る。

【００４７】

（言語統計）

一実施形態において、ステップ２２０はルーチン６００（ステップ６１０～ステップ６３０）を含む。言語統計アナライザ１３０は、ルーチン６００を実行する。ルーチン６００は、セグメンタ１１０によって出力された複数の対の候補となるブロックセグメントに対するマッチスコアを計算する。マッチスコアは、スコア関数に従って計算される。特定のブロックのテキスト（好ましくは３０を超える単語を有する）に対して、各単語が、全体の言語資料（ｃｏｒｐｕｓ）に対してブロック内にどの程度あるかを計算する。単語が記事内のＸパーセントの単語と、言語資料内のＹパーセントの単語とを形成する場合、用いられる正しい式はｌｏｇ（（Ｘ／Ｙ）＋１）である。各ブロックに対して、値のベクトルが得られる。複数の対のブロック間のコサイン距離を得るために、これらのベクトルを用いることは、１（正しいマッチ）～０（単語にオーバーラップしない）の範囲のスコアを与える。

【００４８】

ステップ６２０において、言語統計アナライザ１３０は、計算されたマッチスコアおよび訓練データ１３５に基づいて、複数の対の候補となるブロックセグメントが同一の記事に属する一組の言語統計確率を計算する。この訓練データ１３５は、訓練の組および／またはユーザ入力から得られた確率データを含む。このような訓練は、利用可能な場合には同一の媒体資料によって実行され得る（例えば、図４の例に対して、他の発行者の新聞の１０００ページに対する画像データ）。他の場合には、言語統計目的のための訓練が異なる媒体資料によってなされ得る。

【００４９】

例えば、ステップ６２０において、複数の対のブロック間のスコア関数（ステップ６１０において計算される）を仮定すると、言語統計アナライザ１３０は、２つの任意のブロックが同一の記事からである確率を計算する。言語のために、大きな集合の記事へのアクセスがある場合には、様々なサイズのブロックに記事を分割する。同一の記事からの複数の対を、ポジティブな例として使用し、別の記事からの複数の対を、ネガティブな例として使用する。このような例がない場合には、明確に同一の記事であるブロック（同一のセグメント化された領域）およびほぼ明確に異なるブロック（異なるページ／発行者または離れている）を選び出すためにＯＣＲ化された文書自体をその代りに使用し得る。特有の単語長およびコサイン距離を有する一対のブロックを仮定する場合には、類似の例を見て、どの程度の割合がポジティブな例であるかを見出す。データ点の数に依存して、これは、ｋｅｒｎｅｌ　ｓｍｏｏｔｈｅｒまたはｌｏｃａｌ　ｒｅｇｒｅｓｓｉｏｎによって向上する。

【００５０】

最終的に、ステップ６３０において、言語統計アナライザ１３０は、決定された確率に基づいて同一の記事に属するブロックを識別する。例えば、確率が５０％を超える場合に、ブロックセグメントは、同一の記事に属するように識別され得る。精度が相対的に重要である一例においては、確率が９０％を超える場合に、ブロックセグメントが同一の記事に属するように識別される。これらは例示である。他の確率閾値が使用され得る。

【００５１】

これは、自動的または半自動的（半管理された習得タスク）であり得、このことはテキストの一部のブロックを仮定すると、一対のブロックが同一の記事に由来する確率を出力する。

【００５２】

このマッチスコア関数ならびにコサイン距離および単語頻度の使用は、例示的であり、本発明を限定することを意図していない。現在公知の、または将来開発される他の関連のある技術が、２つのテキストのブロックの関連性を決定するか、またはスコア付けするために使用され得る。

【００５３】

（レイアウト移行）

特徴に従って、レイアウト移行解析が、訓練モードにおいて、または実行モードにおいて実行され得る。一実施形態において、ステップ２２０は、訓練モードルーチン７００および実行モードルーチン９００を含む。レイアウト移行アナライザ１４０は、訓練モードまたは実行モードにおいて動作する。訓練モードにおいて、レイアウト移行アナライザ１４０は、媒体資料の複数のサンプルからの収集されたデータに対して動作し、レイアウト移行分類子１４５を構築する。実行モードにおいて、レイアウト移行アナライザ１４０は、レイアウト移行分類子１４５を、解析される媒体資料レイアウト内のデータに適用する。

【００５４】

（訓練モード）

レイアウト移行アナライザ１４０は、訓練モードルーチン７００を実行する。レイアウト移行アナライザ１４０は、垂直方向の移行（図７Ａ、ステップ７１０～７６０）および水平方向の移行（図７Ｂ、ステップ７７０～７９６）について媒体資料のレイアウトを解

析する。

【００５５】

（垂直方向の移行）

ステップ７１０において、レイアウト移行アナライザ１４０は、１つのブロックがもう１つのブロックよりも上にあり、垂直方向に整列されたブロックの間に本文テキストのブロックがないように垂直方向に整列された、複数の対の本文テキストのブロックセグメントを発見する（図７Ａ）。これらの垂直方向に整列された複数の対のブロックセグメントは、セグメンタ１１０から出力されたブロックセグメント内に発見され得る。例えば、図８に示される新聞のページにおいて、一対のブロックセグメント８１０、８３０は、垂直方向に整列されているものとして識別され得る。

【００５６】

複数の対の垂直方向に整列されたブロックセグメントの間におけるレイアウトに配置された項目が識別される（ステップ７２０）。そして、ブロックセグメントの対は、（少なくとも１つの）介在項目（ｉｎｔｅｒｖｅｎｉｎｇ　ｉｔｅｍ）のうちのいずれか１つ以上に基づいて分類される（ステップ７３０）。例えば、垂直方向に整列されたブロックセグメント８１０、８３０の場合、介在項目８２０は、水平方向の罫線、テキストのライン、下線、水平方向の罫線、テキストのライン、水平方向の罫線である。そして、１つの分類は、これらの項目の特定の移行特徴、例えば、水平方向の罫線と、１６ｐｔのテキストのラインと、下線と、水平方向の罫線と、２４ｐｔのテキストのラインと、水平方向の罫線とによって分離されたブロックであり得る。

【００５７】

次に、レイアウト移行アナライザ１４０は、分類されたブロックセグメントに対して、一組の移行特徴を計算する（ステップ７４０）。例えば、介在項目８２０を有するブロックセグメント８１０、８３０の場合、計算された移行特徴の組は、ブロックの垂直方向の分離の全て、どの程度良好にブロックが整列するか、ブロックの幅に対する罫線の幅、テキストのフォントサイズ、ブロックの幅に対するテキストのラインの幅、等であり得る。新聞のレイアウトのデータの一例においては、ブロックセグメントの対に対して用いられ得る垂直方向の移行特徴（例えば、ほぼ同一の平均テキストサイズの本文テキストの領域であって、それぞれの上に配置され、頂部ブロックおよび底部ブロックと称される）のリストは、（１）頂部ブロックおよび底部ブロックの平均の幅、（２）頂部ブロックと底部ブロックとの間の垂直方向の距離、（３）頂部ブロックの幅と底部ブロックの幅との間に本文でないテキストのブロックが存在するときの、頂部ブロックの幅と底部ブロックの幅との間の平均の幅の小部分としての差、（４）頂部ブロックおよび底部ブロックの左拡張部（ｌｅｆｔ　ｅｘｔｅｎｔ）、（５）頂部ブロックおよび底部ブロックの右拡張部（ｒｉｇｈｔ　ｅｘｔｅｎｔ）、（６）頂部ブロックおよび底部ブロックの頂部、（７）頂部ブロックおよび底部ブロックの底部、（８）頂部ブロックおよび底部ブロックの頂部と底部との間の距離、（９）頂部ブロックおよび底部ブロックの左と右との間の距離、（１０）これらの頂部ブロックおよび底部ブロックにおける平均フォントサイズ、ならびに（１１）これらの頂部ブロックおよび底部ブロックにおける最大フォントサイズ、を含む。

【００５８】

これらの例の垂直方向の移行特徴は、例示的なものに過ぎず、本発明を限定することを意図されていない。計算されるべき正確な組の移行特徴は、訓練される分類子１４５に必要とされる所望の精度と、媒体資料の複雑性とに依存する。より単純なレイアウトは、より少ない計算されるべき移行特徴を必要とし得る。精度が比較的重要であるより複雑なレイアウトまたはアプリケーションは、計算されるべきより多くの組の移行特徴を見込み得る。計算されるべき正確な組の特徴は、手動または自動で変更され得る。手動の変更は、ユーザインターフェース１６０を介するユーザからの入力に基づいて実行され得る。

【００５９】

ステップ７５０において、レイアウト移行アナライザ１４０は、ブロックセグメントが同一の記事にある確率を決定する。ステップ７３０における各分類と、一組の垂直方向の

特徴とに対し、一連の数字によって概略される多数の垂直方向の特徴が存在する。また、言語統計アナライザ１３０の出力から、移行が記事の一部分である確率が知られる。

【００６０】

ここで、セグメントが同一の記事内にないときに、ブロックセグメントの区分の尤度を最大化するために、レイアウト移行分類子１４５、例えば決定ツリーが、自動的に形成され得る（ステップ７６０）。この分類子１４５は、各垂直方向の移行に対して、２つの垂直方向に整列されたブロックが融合されるかどうかを決定するために用いられ得る。このようにして、垂直方向に整列されたブロックセグメントは、ここで、最大のコラムの集合であり、水平方向の移行の解析のための準備が出来ている。垂直方向の移行特徴に基づいてレイアウト移行分類子１４５を形成するこの訓練は、利用可能なデータの集合体（例えば、１つ以上の画像からの複数のブロックセグメント）にわたって実行され得る。一例では、本発明を限定することを意図しないが、訓練は、画像の大きな集合、例えば新聞の異なる記事からの１００ページ以上にわたって実行され得、レイアウト移行分類子１４５を形成し得る。

【００６１】

（水平方向の移行）

ステップ７７０において、レイアウト移行アナライザ１４０は、複数の対の水平方向に整列された本文テキストのブロックセグメントを発見し、１つのブロックはその他のブロックの近くにあり、本文テキストのブロックは、水平方向に整列されたブロックの間には存在しない（図７Ｂ）。これらの複数の対の水平方向に整列されたブロックセグメントは、セグメンタ１１０によって出力されたブロックセグメント内に発見される。例えば、図８に示されている新聞のページにおいて、一対のブロックセグメント８４０、８５０は、水平方向に整列されているとして識別され得る。

【００６２】

複数の対の水平方向に整列されたブロックセグメントの間のレイアウトに配置された介在項目が識別される（ステップ７８０）。そして、複数の対のブロックセグメントは、（少なくとも１つの）介在項目のうちの任意の１つ以上に基づいて、分類される（ステップ７９０）。例えば、水平方向に整列されたブロックセグメント８４０、８５０の場合、介在項目はのど空きである。そして、１つの分類は、これらの（少なくとも１つの）項目の特定の移行特徴（例えば、のど空きおよびその幅）によって分離されたブロックであり得る。

【００６３】

次に、レイアウト移行アナライザ１４０は、分類されたブロックセグメントに対して、一組の移行特徴を計算する（ステップ７９２）。例えば、ブロックセグメント８４０、８５０ならびにそれらの介在項目の場合、計算される一組の移行特徴は、のど空きおよびその幅、ブロックの水平方向の水平方向の分離の全体、ブロックがどの程度良好に整列されているか、等であり得る。新聞のレイアウトデータの一例において、ほぼ同一の平均テキストサイズであって、互いに近くに配置される複数の対のブロックセグメントまたは本文テキストの領域（左ブロックおよび右ブロックとも称される）に対して用いられ得る水平方向の移行特徴のリストは、（１）右ブロックの右縁と左ブロックの左縁との間の距離、（２）左ブロックおよび右ブロックの頂部の垂直方向の整列、（３）左ブロックと右ブロックとの間の水平方向の距離、（４）２つの左ブロックと右ブロックとの幅の間の差、（５）本文ではないテキストの近傍に対する関係、を含む。本文ではないテキストの近傍に対するそのような関係は、例えば、左ブロックの頂部の最も近くの本文ではないテキストのブロックと、右ブロックの頂部の最も近くのブロックとを発見し、近傍の本文ではないテキストのブロックの各々に対して、２つの左ブロックおよび右ブロックの最も遠くからの本文ではないテキストのブロックの垂直方向の距離、２つの左ブロックおよび右ブロックの最も近くからの本文ではないテキストのブロックの垂直方向の距離、左ブロックを越えた本文ではないテキストのブロックの左の範囲、右ブロックを越えた本文ではないテキストのブロックの右の範囲、２つの左ブロックおよび右ブロックの頂部の平均からの本文

ではないテキストのブロックの距離、２つの左ブロックおよび右ブロックの底部の平均からの本文ではないテキストのブロックの距離、本文ではないテキストのブロックの幅、本文ではないテキストのブロックの高さ、本文ではないテキスト内のフォントサイズおよび本文ではないテキストのブロックの単語数、を含む。

【００６４】

これらの例の水平方向の移行特徴は、例示的なものであり、本発明を限定することを意図されていない。計算される完全な組の移行特徴は、訓練される分類子１４５に必要とされる所望の精度と、媒体資料の複雑性とに依存する。より単純なレイアウトは、より少ない計算されるべき移行特徴を必要とし得る。精度が比較的重要であるより複雑なレイアウトまたはアプリケーションは、計算されるべきより大きな組の移行特徴を見込み得る。計算されるべき正確な組の特徴は、手動または自動で変更され得る。手動の変更は、ユーザインターフェース１６０を介するユーザからの入力に基づいて実行され得る。

【００６５】

ステップ７９４において、レイアウト移行アナライザ１４０は、ブロックセグメントが同一の記事にある確率を決定する。ステップ７９０における各分類と、一組の水平方向の特徴とに対し、一連の数字によって概略される多数の移行特徴が存在する。また、言語統計アナライザ１３０の出力から、移行が記事の一部分である確率が知られる。ここで、セグメントが同一の記事内にないときに、ブロックセグメントの区分の尤度を最大化するために、レイアウト移行アナライザ１４５、例えば決定ツリーが、自動的に形成され得る（ステップ７９６）。この分類子１４５は、水平方向の移行の各々を決定し、２つの水平方向に整列されたブロックのグループが同一の記事内に存在するかどうかを決定するために用いられ得る。このようにして、ブロックセグメントの最大のコラムは、それらが同一の記事に属するときに、水平方向にさらにグルーピングされる。水平方向の移行特徴に基づいてレイアウト移行分類子１４５を形成するこの訓練は、利用可能なデータの集合体（例えば、１つ以上の画像からの複数のブロックセグメント）にわたって実行され得る。一例では、本発明を限定することを意図しないが、訓練は、画像の大きな集合、例えば新聞の異なる記事からの１００ページ以上にわたって実行され得、レイアウト移行分類子１４５を形成し得る。

【００６６】

１つの利点は、このレイアウト移行解析が、構成される記事の精度を向上させるための言語統計解析を補足し得るということである。レイアウト手段において垂直方向および水平方向の移行特徴に基づいてブロックセグメントを分類する分類子１４５を用いると、一対のブロックセグメントが言語統計解析のみに基づいたときには無関係に見えるがいくつかの強く関連する対のパターンにフィットするという場合でさえも、コンバイナ１５０は、その対を同一の記事の一部分として配置し得る。このようにして、言語統計とレイアウト移行とのこの組み合わせは、言語統計とレイアウト移行との一方のみで達成し得る精度を超えて、精度を向上させ得る。

【００６７】

上述のように、発明者がテキストの２つのブロックがどのように関連しているかを評価するために単語の頻度を用いて実行した一部の例において、媒体資料解析ルーチン２００は、２つのブロックが同一の記事からのものだったかどうかを示すことに関して、約９０％の精度であった。媒体資料解析ルーチン２００は、一般に正しい規則を発見するために、ページの大きな集合にわたってこれらの予測を組み合わせる。例えば、２つのブロックが１６～２０ｐｔのヘルベティカ（Ｈｅｌｖｅｔｉｃａ）テキストによって分離されるときに、これら２つのブロックが通常関連のないテキストを有している場合に、これは、記事を分離するものである可能性がある。

【００６８】

この方法は、言語統計確率の評価を形成し、レイアウト統計特徴に基づいて、承認される一組の確率を形成し、このプロセスは、新しいデータが解析されるときに繰り返され、各実行可能性が、幾分か多くの情報を追加し、分類子の精度を向上させる。

【００６９】
訓練分類子１４５がデータの集合体にわたって訓練され、構築されると、訓練分類子１４５は、実行モードにおいて動作するレイアウト移行アナライザ１４０によって用いられ得る。
【００７０】
（実行モード）
実行モードは、図７に関連して上述された訓練モードに類似している。レイアウト移行アナライザ１４０は、実行モードルーチン９００を実行する。レイアウト移行アナライザ１４０は、垂直方向の移行（図９Ａ、ステップ７１０～７４０および９１０）と水平方向の移行（図９Ｂ、ステップ７７０～７９２および９２０）とについて、媒体資料のレイアウトを解析する。
【００７１】
実行モードにおいて、レイアウト移行アナライザ１４０は、上述のように垂直方向に整列された複数の対のブロックセグメントに対して、ステップ７１０～７４０を実行する。分類子１４５を構築する代わりに、レイアウト移行アナライザ１４０は、分類子１４５（例えば、決定ツリー）を適用して、垂直方向に整列されたブロックセグメントが同一の記事に属するかどうかを決定する（ステップ９１０）。
【００７２】
同様に、実行モードにおいて、レイアウト移行アナライザ１４０は、上述のように水平方向に整列された複数の対のブロックセグメントに対して、ステップ７７０～７９２を実行する。レイアウト移行アナライザ１４０は、分類子１４５（例えば、決定ツリー）を適用して、水平方向に整列されたブロックセグメントが同一の記事に属するかどうかを決定する（ステップ９２０）。
【００７３】
訓練モードと実行モードとの間のこの分割は、例示的なものであり、本発明の実施形態を限定することを意図されていない。別の実施形態において、実行モード中に媒体資料アナライザ１００の実行の間に出力された結果は、新しいデータ１０５が解析されるときに分類子が定期的に更新され得るように、分類子１４５を改変するために用いられ得る。
【００７４】
プロセスの流れを垂直方向の流れと垂直方向の流れとの２つのタイプに分離することによってレイアウトを解析し、その後、垂直方向の移行と水平方向の移行とに何が似ているかと、それらの間に何が存在しているかとに基づいて、垂直方向の移行と水平方向の移行とを合算する方法は、固有の特徴であるが、本発明を限定することは意図されていない。代替的に、レイアウト移行解析は、プロセスの流れを２つの部分に分離することなしに、レイアウト移行特徴に基づいてなされ得る。また、本文テキストのブロックを見る代わりに、ページ上の全ての要素に対してツリー構造を形成し、任意の２つのタイプの要素の間の移行の規則を形成するように試みることができる。
【００７５】
分類子１４５を構築する際に用いられ得る複数の機械学習アプローチが存在する。本記載を与えられた当業者には明白なように、決定ツリーの実装の他に、ｌｉｎｅａｒ ｓｅｐａｒａｔｏｒ ａｆｔｅｒ ａ ｂａｓｉｓ ｅｘｐａｎｓｉｏｎ、ｋ－ｍｅａｎｓ ｃｌｕｓｔｅｒｉｎｇ、ｋｅｒｎｅｌ ｓｍｏｏｔｈｉｎｇ ｍｅｔｈｏｄｓ等を用いることができる。別のアプローチは、単純に、特徴を離散化し、これらのバケットに分類し、見てきたケースが十分な例であったことを望むことである。
【００７６】
（表示例）
図１０Ａ～１０Ｄは、本発明の実施形態にしたがって解析される新聞のページを含む、例示的な媒体資料を示す。
【００７７】
図１０Ａは、言語統計解析とレイアウト移行解析とに基づいて、媒体資料アナライザ１

００によって解析された新聞の第一面（ｆｒｏｎｔ　ｐａｇｅ）の表示である。本文テキストを含むブロックセグメントは、媒体資料アナライザ１００によって解析されるときに、本文テキストが属する対応する記事内でハイライトされる。同一の記事内のコラム状の本文テキストに対応するブロックセグメントは、アナライザ１００がどのようにしてデータを細分化し、適切なセグメントを用いてどのように記事を構成したのかを示すために、同一のカラーによってカラー表示されたり、または陰影付けされたりする。図１０Ｂは、レイアウト移行分類子とレイアウト移行アナライザとを有する媒体資料アナライザ１００によって解析される、比較的トリッキーなレイアウトの新聞の内部ページの例の表示である。

【００７８】

図１０Ｃおよび図１０Ｄは、本発明の一実施形態にしたがって言語統計解析（純粋な言語統計モード）に基づいて解析される媒体材料のハイライトされた例を示している。図１０Ｃは、新聞の第一面の例を示しており、ブロックセグメントがハイライトされ、第一面に記事がある。図１０Ｄは、新聞の内部のページを示しており、記事内でブロックセグメントがハイライトされている。この例では、同一の記事内のブロックセグメントが同一の色でハイライトされているが、本発明は、そのように限定されるわけではない。その他のタイプのハイライト表示（例えば、グレイスケールの陰影付け、境界、テクスチャあるいはその他のマークまたは印）が、カラーの代わりに、またはカラーに加えて用いられ得る。また、ハイライト表示は、必要に応じて用いられなくてもよく、記事のセグメントまたはその一部分のみが表示されることがあり得る。

【００７９】

（ワールドワイドウェブへの応用）

本発明のさらなる実施形態にしたがうと、ワールドワイドウェブを介して、レイアウトを有する媒体資料をサーチするシステムが提供される（図１１）。図１１に示されているように、媒体資料をサーチするシステム１１００は、クライアント１１１０と、ウェブサーバ１１３０と、サーバ１１４０と、データベース１１４５とを含む。クライアント１１１０は、ネットワーク１１２０を介して、ウェブサーバ１３０に結合されている。ネットワーク１１２０は、ローカルエリアネットワーク、中規模エリアネットワーク、またはワイドエリアネットワークを含むがそれらには限定されない任意のタイプの１つ以上の任意のネットワーク、例えばインターネットであり得る。一例において、クライアント１１１０は、ネットワーク１１２０を介して通信するブラウザを含み得る。任意のタイプのブラウザが用いられ得る。ウェブサーバ１１３０は、サーバ１１４０に結合されている。

【００８０】

サーバ１４０は、上述のように、媒体資料アナライザ１００を含むか、それに結合されている。サーバ１４０はまた、データベース１１４５に結合されている。データベース１１４５は、媒体資料アナライザ１００をサポートするためにデータを格納する任意のタイプのデータベースまたはメモリである。データベース１１４５は、上述のように、例えば、訓練データ１３５と、レイアウト移行分類子１４５と、データ１０５とを格納し得る。データベース１１４５はまた、画像データそのものを表すデータを含む媒体資料アナライザ１００からの任意の出力を、そして媒体資料アナライザ１００によって識別された記事に属するブロックセグメントと共に、格納し得る。勿論、特定の用途に依存して、インデックス付けおよびその他の操作が実行され得、それにより、出力されたデータは、サーチ要求またはその他のタイプのデータ要求を満足するように、容易に検索される。

【００８１】

操作中、ユーザは、クライアント１１１０においてサーチクエリを入力し得る。そして、クライアント１１１０におけるブラウザは、ネットワーク１１２０を介して、ウェブサーバ１１３０にサーチクエリを転送する。ウェブサーバ１１３０は、サーバ１１４０と通信し、オプションとして、媒体資料アナライザ１００と直接的に通信する。一実施形態において、媒体資料アナライザ１００は、画像データそのものを表すデータと共に、記事に属するブロックセグメントを識別するメタデータを出力する。この出力は、データベース

１１４５に格納される。サーバ１１４０は、サーチ要求を満たすように、キーワードまたはサーチ用語について、データベース１１４５をサーチする。そして、サーバ１１４０は、ウェブサーバ１１３０に、サーチ要求を満足する結果を返す。そして、ウェブサーバ１１３０は、表示のために、満足されたサーチ結果を、クライアント１１１０におけるブラウザに転送する。このようにして、ウェブサーバ１１３０およびサーバ１１４０は、連携して動作し、任意のサーチエンジン、ポータルまたはウェブサイトの一部分となり得る。

【００８２】

図１２は、新聞の実施形態と共に用いられえる例示的なディスプレイ１２００を示している。ディスプレイ１２００に示されているように、フィールド１２１０は、サーチ結果を入力するために用いられ得る。そして、ボタン１２１５は、サーチを開始するためにユーザによって選択され得る。フィールド１２２０は、サーチからの出力結果を表示するように用いられ得る。一実施形態においては、サーチ結果を示すために多数のウィンドウが表示される。例えば、記事内でサーチ結果がヒットした場合、同一の記事からの２つの断片が、２つのウィンドウ１２２２、１２２４内に表示され得る。これらの断片は、サーチ用語と、サーチ用語の周辺の情報とを含み得る。これは、単なる例であり、本発明を限定するように意図されていない。１つ以上の記事からの１つ以上の断片が表示され得る。さらに、記事のテキスト全体、または記事内のサーチ用語のみ、または断片、例えばサーチ用語を包囲する領域が、表示され得る。任意の数のヒット、断片またはサーチを満足する所望のテキストが、表示され得る。図１０に示されているようなハイライト表示されたブロックセグメントを有する新聞のページの全体の画像（またはその一部分)もまた、表示され得る。

【００８３】

さらなる例にしたがうと、その他のタイプの情報が、ディスプレイ１２００内に表示され得る。図１２に示されているように、フィールド１２３０は、関連情報を示すために表示され得る。フィールド１２４０は、新聞に関する書誌情報、例えば、発行元、新聞が発行された日時、リポーターの署名欄および他の情報等の情報を表示するように用いられ得る。記事、タイトル、および新聞の名前をエリア１２５４に表示するために、別のフィールド１２５０が提供される。サーチを実行するユーザが記事をオーダーすること（ボタン１２６２)および新聞に契約すること（ボタン１２６４）を可能にするように、追加的な制御フィールド、例えばフィールド１２６０が提供され得る。ナビゲーション制御もまた提供され得る。例えば、ナビゲーションエリア１２７０は、ページ番号を表示するジャンプフィールド１２７２を含み得、該ジャンプフィールドは、ユーザが新聞の異なるページにジャンプすること、または、異なるサーチ結果にジャンプすることを可能にする。ユーザが、ユーザに提示された媒体情報の表示を、スクロールすること、ズームインすること、ズームアウトすること、または、変更することを可能にする、その他のナビゲーション制御（図示されず)が提供され得る。

【００８４】

（さらなる特徴および利点）

特に、スキャンされた（または電子的に生成された）新聞ならびに関連する資料（例えば、雑誌、カタログ等）のレイアウトをセグメント化および解析するための新規なアプローチが提供される。高度な形態学関連アルゴリズム（ｍｏｒｐｈｏｌｏｇｙ－ｒｅｌａｔｅｄ　ａｌｇｏｒｉｔｈｍ）は、ページを物理的なブロックに分解する。テキスト情報（ページ内に存在する、または、ＯＣＲから抽出される）は、テキストブロックをどのようにして記事に構成すべきか、および、どのようにしてテキストが流れるかを、決定するために用いられ得る。加えて、多数のページのテキスト解析を通して収集された情報が、レイアウト解析のために、集合に特有（ｃｏｌｌｅｃｔｉｏｎ－ｓｐｅｃｉｆｉｃ）の幾何学的規則を推察するために用いられ得る。

【００８５】

レイアウトのセグメント化は、１つの単位として記事または結合した実体（ｃｏｈｅｓｉｖｅ　ｅｎｔｉｔｙ）が何であるかを理解し、そしてインデックス付けすることを可能

にする。レイアウトのセグメント化はまた、便利にも記事にズームインすること、文脈内の記事を抽出すること、実際にテキストを提示することなしに、テキスト情報をリフローすることを可能にする。言い換えると、レイアウトのセグメント化は、スキャンされた新聞および雑誌をナビゲートするための強力かつ便利なユーザ経験を可能にし、実際、レイアウトのセグメント化はまた、電子ソース（例えば、ＰＤＦ)に適用される。

【００８６】

さらなる特徴にしたがうと、媒体資料アナライザ１００の実施形態は、言語統計を用いることにより、様々な幾何学的要素の規則を学習し、例えば新聞のような媒体資料に特有の規則を計算し得る。そのようなアナライザおよび方法は、訓練データ内にどのような例示的なセグメント化も必要とせず、画像とＯＣＲ出力とから決定された言語統計から純粋に機能する。

【００８７】

さらに、その他の上述の制限された幾何学ベースのレイアウト解析とは異なり、本発明の発明者によって、本明細書中に実施形態が提供され、この実施形態は、記事レベルのセグメント化を推察するために用いられるべきテキストデータから言語統計が引き出されることと、新聞／雑誌のページの特定の集合をセグメント化するために用いられ得る幾何学的規則を推察することとを可能にする実施形態が提供される。

【００８８】

単純に最終的な推測を提供する代わりに、媒体資料アナライザは、記事に対する最良の推測のリストを保持し、ユーザがユーザインターフェースにおいて、不都合なものがある場合に、一部分の代替物を見ることを可能にする。ユーザは、概して、最も容易に読み取ることが可能なオプションを発見するまで、これを行うことを望み得る。ユーザが選択するものを観察することによって、記事コンポーザは、確率を変更することにより、リアルタイムで選択を更新し、ユーザ選択からトリッキーな領域を学習することを可能にする。

【００８９】

記事のセグメント化の計算後、記事を表示するための多くの方法が存在する。１つの方法は、ユーザが記事を選択することを可能にし、その時点で、ユーザは、ページのズームイン画面を入手し、ユーザがスクロールホイールを用いて記事の複数の部分にわたってナビゲートすることを可能にする。バウンディングボックスと共に機能して、個々の単語の画像が抽出され、リフローされ得る。これは、新聞に似ているがより読みやすい別個のページを介して行われ得るか、あるいは、ユーザがテキストまたはコラムのサイズを変更し、その結果、新聞がスタイルおよびパラメータにフィットするように「再生成される」ことを可能にすることにより、行われ得る。

【００９０】

（さらなる応用－連続する記事）

コラムの規則を発見するために多くのページにわたる集合体内で言語統計を用いる技術が、いくつかのその他の問題を解決するために用いられ得る。一実施形態にしたがうと、新聞のページ間で連続する記事内のブロックセグメントを決定するために、さらなる解析が実行され得る。異なるページ上の記事部分が同一の連続する記事内に属するかどうかの解析は、言語統計と、連続レイアウト移行情報とを用いて行われる。

【００９１】

図１３に示されているように、媒体資料アナライザ１３００は、上述の媒体資料アナライザ１００を含み、連続言語統計アナライザ１３３０と、連続レイアウトアナライザ１３４０とを有する記事コンポーザ１３２０をさらに含む。連続言語統計アナライザ１３３０と連続レイアウトアナライザ１３４０とは、図１３に示されているように、記事コンポーザ１３２０内に含まれえる。これは例示であり、本発明を限定することは意図されていない。例えば、連続言語統計アナライザ１３３０と連続レイアウトアナライザ１３４０とは、別個または組み合わせで提供されるか、または、言語統計アナライザ１３０とレイアウト移行アナライザ１４０との一部分として、それぞれ追加され得る。連続言語統計アナライザ１３３０と連続レイアウトアナライザ１３４０とは、ソフトウェア、ファームウェア

、ハードウェアまたはそれらの組み合わせで実行され得る。連続言語統計アナライザ１３３０と連続レイアウトアナライザ１３４０との機能は、明確化のために別個に記載されるが、１つのモジュールまたはデバイス内で組み合わされ得るか、または、より多くのモジュールまたはデバイスにわたって分散され得る。

【００９２】

媒体資料アナライザ１３００は、複数のページにわたって広がっている１つ以上の連続する記事を含むレイアウトを有する媒体資料を表すデータを解析する。媒体資料アナライザ１３００は、コントローラ１０５と、セグメンタ１１０と、記事コンポーザ１３２０とを含む。セグメンタ１１０は、上述のように、媒体資料のページ内のコラム状の本文テキストに関連するブロックセグメントを識別する。記事コンポーザ１３２０は、言語統計情報と連続移行情報とに基づいて、識別されたブロックセグメントのどれが、媒体資料内の複数のページにわたって広がっている連続する記事に属しているかを決定する。

【００９３】

コントローラ１０２は、セグメンタ１１０と記事コンポーザ１３２０とを制御および管理する。ユーザからのさらなる制御が、ユーザインターフェース１６０を介して提供される。例えば、ユーザは、動作を開始すること、すなわち、データ１０５、訓練データ１３３５またはレイアウト移行分類子１３４５の入力を開始することができる。ユーザは、媒体資料アナライザ１３００と相互作用することにより、訓練データ１３３５を形成またはレビューすることを助け得る。例えば、ユーザは、所与の媒体資料内の複数のページにわたって広がっている連続する記事に属するブロックセグメントのポジティブおよびネガティブな例を用いることにより、訓練データ１３３５の品質を向上させ得る。ユーザはまた、媒体資料アナライザ１３００と相互作用することにより、レイアウト移行分類子１３４５を構築または修正し得る。

【００９４】

訓練データ１３３５は、媒体資料内の連続する記事に属するブロックセグメントのポジティブおよびネガティブな例を含み得る。レイアウト移行分類子１３４５は、候補となるブロックセグメントが媒体資料内の記事に属するとして分類されることを可能にする連続移行特徴を含むデータ構造を含み得るが、それには限定されない。そのようなデータ構造は、決定ツリーを含み得るが、それには限定されない。図１において上述された訓練データ１３５と分類子１４５とはまた、媒体資料アナライザ１３００と共に用いられ得、特に、別個に、または、訓練データ１３３５と分類子１３４５のそれぞれの一部分として用いられ得る。

【００９５】

一実施形態において、記事コンポーザ１３２０は、上述のような言語統計アナライザ１３０、レイアウト移行アナライザ１４０、コンバイナ１５０を含み、連続レイアウト移行アナライザ１３４０と、連続言語統計アナライザ１３３０とをさらに含む。連続レイアウト移行アナライザ１３４０は、決定ツリー１３４５を適用し、異なるページ上の候補となる記事部分の最後のブロックセグメントと第１のブロックセグメントとが同一の連続する記事内にある確率を示している１つ以上の連続移行特徴をピックアップする。連続言語統計アナライザ１３３０は、計算された言語統計情報に基づいて、異なるページ上の異なる記事部分に対する言語統計情報を計算し、候補となる記事部分内の第１および最後のブロックセグメントが、連続する記事部分を有する確率を決定する。このようにして、記事コンポーザ１３２０は、解析された連続レイアウト移行特徴と、計算された言語統計とにしたがって、第１および最後のブロックセグメントが同一の連続する記事に属する確率に基づいて、複数のページにわたって連続する記事を構成し得る。本明細書中の計算される確率は、ｙｅｓ／ｎｏまたはブール値表示、確率または信頼データを表す数値、あるいは、確率または信頼データを表す値の数値範囲を含み得るが、それらには限定されない。

【００９６】

連続レイアウト移行アナライザ１３４０と連続言語統計アナライザ１３３０との動作は、図１４Ａ～Ｅに示されているルーチン１４００に関連して（ステップ１４０２～１４３

６）、以下でさらに詳細に記載される。ルーチン１４００は、セグメンタ１１０がデータ１０５内のブロックセグメントを識別した後に開始し、言語統計アナライザ１３０、レイアウト移行アナライザ１４０およびコンバイナ１５０は、図１～１０に関連して上述されたように、媒体材料の複数のページ上の記事内のブロックセグメントを組み合わせている。

【００９７】

ステップ１４０２において、連続レイアウト移行アナライザ１３４０は、候補となる連続する記事部分に関連する最後のブロックセグメントを識別する。例えば、連続レイアウト移行アナライザ１３４０は、全てのコラム状のブロックを見て、記事内の最後のブロックセグメントである最後のブロックセグメントを識別する。そして、連続レイアウト移行アナライザ１３４０は、ブロックセグメントの下にある１つ以上の項目を識別する（ステップ１４０４）。そのような項目は、最後のブロックの終わりに現れる、単語（例えば、「連続している（ｃｏｎｔｉｎｕｅｄ）」、「連続（ｃｏｎｔ．）」、「～を参照（ｓｅｅ）」または「～に行く（ｇｏ　ｔｏ）」）あるいはレイアウト項目（例えば、矢印または直線、ドローイングボックス）であり得る。

【００９８】

連続レイアウト移行アナライザ１３４０は、識別された１つ以上の項目の少なくとも1つの特徴に基づいて、最後のブロックセグメントを分類する（ステップ１４０６）。項目の特徴は、項目そのもの、または、項目およびレイアウトに関する特性を含み得る。例えば、単語「連続している」の特徴は、用語「連続している」のフォントサイズ、フォントスタイル、間隔または配置、および／または、単純に用語「連続している」そのものの存在であり得る。これらの特徴を分類することは、共通する特徴を有するブロックをそれぞれのグループに配置することを助け得る。例えば、ドローイングボックスを下に有するブロックは、１つのグループに分類され得るが、用語「連続している」を下に有するブロックは、別のグループに配置され得る。

【００９９】

そして、連続レイアウト移行アナライザ１３４０は、決定ツリー１３４５を適用することにより、最後のブロックセグメントが連続する記事内にある確率を示す１つ以上の連続移行特徴をピックアップする（ステップ１４０８）。連続移行特徴は、特定のレイアウトに依存して、最後のブロックが別のページ上に連続する記事部分を有するより高い確率に関連する特徴であり得る。例えば、例示的なレイアウトにおける連続移行特徴は、１２ｐｔのイタリック体の用語「連続する（ｃｏｎｔｉｎｕｅｄ　ｏｎ）」と１２ｐｔの矢印とに関連する特徴であり得る。これらの特徴および例は、例示的なものであり、本発明を限定することは意図されていない。当業者には明白であり得るように、様々な組み合わせにおけるその他の特徴が、解析されるレイアウトと訓練データとに依存して用いられ得る。

【０１００】

連続レイアウト移行アナライザ１３４０が最後のブロックを学習または識別すると（ステップ１４０２～１４０８）、連続言語統計アナライザ１３３０は、以後のページ上の記事部分を、最後のブロックがあるページからサーチする（図１４Ｂ、ステップ１４１０）。連続言語統計アナライザ１３３０は、連続する記事の第１のページから、一連のページ上の記事部分に対する言語統計情報を計算し（ステップ１４１２）、計算された一連の言語統計情報に基づいて、候補となる連続記事部分内の最後のブロックセグメントが連続する記事部分を有している確率を決定する（ステップ１４１４）。例えば、言語統計情報は、単語の頻度情報であり得、連続言語統計アナライザ１３３０は、最後のブロックセグメント内のテキストおよび一連のページ上の記事部分内のテキストにおける単語の頻度に基づいて、マッチスコアを計算し得る。そして、最後のブロックセグメントが連続する記事部分を有する確率が、マッチスコアに基づいて決定され得る。訓練データ１３３５もまた、マッチスコアから確率を決定するために用いられ得る。

【０１０１】

ブロックが第１のブロックであるかどうかと、連続する記事部分内に存在する可能性が

あるかどうかとを決定するために、同様の解析が実行される。図１４Ｃに示されているように、連続レイアウト移行アナライザ１３４０は、一連のページ内の候補となる連続する記事に関連する第１のブロックセグメントを識別する（ステップ１４１６）。例えば、連続レイアウト移行アナライザ１３４０は、全てのコラム状のブロックを見て、記事内の第１のブロックセグメントであるブロックセグメントを識別する。連続レイアウト移行アナライザ１３４０は、第１のブロックセグメントの上の１つ以上の項目を識別する（ステップ１４１８）。そのような項目は、第１のブロックの上に現れる単語（例えば、「から連続している（ｃｏｎｔｉｎｕｅｄ　ｆｒｏｍ）」または「からの連続（ｃｏｎｔ．ｆｒｏｍ）」）であり得る。

【０１０２】

そして、連続レイアウト移行アナライザ１３４０は、１つ以上の識別された項目の少なくとも１つの特徴に基づいて、第１のブロックセグメントを分類する（ステップ１４２０）。項目の特徴は、項目そのものまたは項目およびレイアウトに関する特性を含み得る。例えば、単語「連続している（ｃｏｎｔｉｎｕｅｄ）」に対する特徴は、用語「連続している（ｃｏｎｔｉｎｕｅｄ）」のフォントサイズ、フォントスタイル、間隔または配置、および／または、単純に用語「連続している（ｃｏｎｔｉｎｕｅｄ）」そのものの存在であり得る。これらの特徴を分類することは、共通する特徴を有するブロックをそれぞれのグループに配置することを助け得る。例えば、ドローイングボックスを下に有するブロックは、１つのグループに分類され得るが、用語「連続している（ｃｏｎｔｉｎｕｅｄ）」を下に有するブロックは、別のグループに配置され得る。

【０１０３】

連続レイアウト移行アナライザ１３４０は、決定ツリー１３４５を適用することにより、第１のブロックセグメントが連続する記事内にある確率を示す１つ以上の連続移行特徴をピックアップする（ステップ１４２２）。そして、連続レイアウト移行アナライザ１３４０は、適用される決定ツリー１３４５に基づいて、第１のブロックセグメントが連続する記事内にある確率を決定し得る（ステップ１４２４）。そのような連続移行特徴は、特定のレイアウトに依存して、第１のブロックが以前のページ上に連続する記事部分を有するより高い確率に関連する特徴であり得る。例えば、例示的なレイアウトにおける連続移行特徴は、１２ｐｔのイタリック体の用語「～から連続する（ｃｏｎｔｉｎｕｅｄ　ｆｒｏｍ）」に関連する特徴であり得る。これらの特徴および例は、例示的なものであり、本発明を限定することは意図されていない。当業者には明白であり得るように、様々な組み合わせにおけるその他の特徴が、解析されるレイアウトと訓練データとに依存して用いられ得る。

【０１０４】

連続レイアウト移行アナライザ１３４０が第１のブロックを学習または識別すると（ステップ１４１６～１４２４）、連続言語統計アナライザ１３３０は、以前のページ上の記事部分を、第１のブロックがあるページからサーチする（図１４Ｄ、ステップ１４２６）。連続言語統計アナライザ１３３０は、第１のブロックのページから、以前のページ上の記事部分に対する言語統計情報を計算し（ステップ１４２８）、計算された以前のページの言語統計情報に基づいて、候補となる連続記事部分内の第１のブロックセグメントが連続する記事部分である確率を決定する（ステップ１４３０）。例えば、言語統計情報は、単語の頻度情報であり得、連続言語統計アナライザ１３３０は、第１のブロックセグメント内のテキストおよび一連のページ上の記事部分内のテキストにおける単語の頻度に基づいて、マッチスコアを計算し得る。そして、第１のブロックセグメントが連続する記事部分である確率が、マッチスコアに基づいて決定され得る。訓練データもまた、マッチスコアから確率を決定するために用いられ得る。

【０１０５】

最後に、最後および第１のブロックが候補となる連続する記事に対して識別されると、別個のページ上の記事部分の間で連続しているかどうかについて学習され得る。図１４Ｅに示されているように、連続レイアウト移行アナライザ１３４０は、一対の候補となる最

後および第１のブロックセグメントを識別し（ステップ１４３２）、一対の候補となる最後および第１のブロックセグメントに対する一組の連続移行特徴を決定し（ステップ１４３４）、決定された連続移行特徴の組に基づいて、決定ツリー１３４５を適用し、一対の候補となる最後および第１のブロックセグメントが媒体資料内の複数のページにわたって同一の連続する記事に属する確率を決定する（ステップ１４３６）。一組の連続移行特徴は、１つの記事部分から別の記事部分への進行（ｐｒｏｇｒｅｓｓｉｏｎ）に関係している特性によって、拡張され得る。例えば、最後のブロックセグメント内の最後の単語、最後のブロックの下の単語または図面、第１のブロック内の第１の単語、第１のブロックの上の単語または図面を考え、様々なセクションの間の単語の重複を見る。決定ツリー１３４５が形成されたときに、「連続している」を含み、第２のセクション内の第１の３つの単語のうちの１つにマッチする単語を有している第１の区分の下に太字が存在するかどうかが分かり、そして、これが特定のレイアウトに対して右の連続であり得る可能性が分かる。

【０１０６】

本記載を与えられた当業者には明白であり得るように、上述の実施形態の媒体資料アナライザ１３００は、訓練データ１３３５の使用と共に、または、訓練データ１３３５の使用を伴わずに、動作され得る。訓練データ１３３５の使用を伴わないと、連続の決定は、見られている特定のブロックのみに基づき得る。例えば、媒体資料アナライザ１３００は、言語スコアが最大になるように、一部の記事内の最後のブロックを取り得、異なるページ上にある一部の記事内の最初のブロックを発見し得る。これは、時々機能するが、多くの可能性のある答えが存在し得、正しいものは、これらの段落内で最良のテキストマッチを有していない可能性があるというだけの理由で、ミスをする比較的高い可能性がある。

【０１０７】

精度を向上させるために、媒体資料アナライザ１３００は、所定の訓練データ１３３５の利点を用いて動作し得るか、または、訓練モードで動作して、訓練データ１３３５を入手または補足し得る。本質的に、訓練は、図１４における実行モードに関連して記載された上述のアプローチを用い、より正確な決定ツリーまたは分類子１３４５を形成するために、多くのページにわたって結果を組み合わせる。そのためものとして、上述では、異なるページ上の複数の領域の対、１つの領域からその他の領域への移行の記載（述べられている特徴)、および、該対の言語スコアの例があった。これらのスコアを（非常に弱いスコアであったとしても)組み合わせることにより、傾向が抽出され、決定ツリーまたは分類子１３４５を構築するために用いられ得る。例えば、第１のブロックの下に小さい三角形があったり、または、第１のブロック内に単語「ｃｏｎｔ’ｄ」があったりし、ブロック間で特定の量の単語が重複している場合には、これらは有効な連続であり得る。

【０１０８】

効率化のために、１つの例では、これらは３つのタスクに分割される。なぜならば、対応のために全ての対のブロックをテストすることは、過度に複雑であり得るからである。

【０１０９】

代わりに、３つの部分またはテストが存在する：

（１）所定のブロックは連続している（ｃｏｎｔｉｎｕｅｄ）か？

（２）所定のブロックは連続（ｃｏｎｔｉｎｕａｔｉｏｎ）か？

（３）これらの２つのタイプのブロックの対が与えられたとき、それらは同一の記事か？

これを同時に行うことを試みることは、分類子１３４５が、（３）を理解する必要があるのみならず、（１）および（２）を２つの部分に組み込む必要があり得ることを意味する。単一の分類子のアプローチが可能であるが、上記のものは、より安全であり、テストがより容易である。

【０１１０】

（例示的なコンピュータシステムの実装)

本発明の様々な局面は、ソフトウェア、ファームウェア、またはそれらの組み合わせに

よって実装され得る。図１５は、例示的なコンピュータシステム１５００を示しており、このコンピュータシステムにおいては、本発明またはその一部分が、コンピュータ読み取り可能なコードとして実行され得る。例えば、図２の方法２００を実行する媒体資料アナライザ１００、図１４の方法１４００を実行する媒体資料アナライザ１３００が、システム１３００内に実装され得る。本発明の様々な実施形態が、この例のコンピュータシステム１５００の観点で記載される。この記載を読んだ後、当業者は、その他のコンピュータシステムおよび／またはコンピュータアーキテクチャを用いて、本発明をどのように実装するかを明白に理解し得る。

【０１１１】

コンピュータシステム１５００は、１つ以上のプロセッサ、例えばプロセッサ１５０４を含む。プロセッサ１５０４は、特定用途向けプロセッサまたは汎用のプロセッサであり得る。プロセッサ１５０４は、通信インフラストラクチャ１５０６（例えば、バスまたはネットワーク）に連結され得る。

【０１１２】

コンピュータシステム１５００はまた、メインメモリ１５０８（好適には、ランダムアクセスメモリ（ＲＡＭ））を含み得、さらには、２次メモリ１５１０を含み得る。２次メモリ１５１０は、例えば、ハードディスクドライブ１５１２および／またはリムーバブル可能格納ドライブ１５１４を含み得る。リムーバブル可能格納ドライブ１５１４は、フロッピー（登録商標）ディスクドライブ、磁気テープドライブ、光ディスクドライブ、フラッシュメモリ等を含み得る。リムーバブル格納ドライブ１５１４は、周知の方法で、リムーバブル格納ユニット１５１８からの読み取り、および／または、リムーバブル格納ユニット１５１８への書き込みを行い得る。リムーバブル格納ユニット１５１８は、フロッピー（登録商標）ディスク、磁気テープ、光ディスク等を含み得、リムーバブル格納ユニット１５１８は、リムーバブル格納ドライブ１５１４によって読み取りおよび書き込みがなされる。当業者は、リムーバブル格納ユニット１５１８が、コンピュータソフトウェアおよび／またはデータが格納されたコンピュータ使用可能な媒体を含むことを理解し得る。

【０１１３】

代替的な実装において、２次メモリ１５１０は、コンピュータプログラムまたはその他の命令がコンピュータシステム１５００にロードされることを可能にするその他の類似の手段を含み得る。そのような手段は、例えば、リムーバブル格納ユニット１５２２およびインターフェース１５２０を含み得る。そのような手段の例は、プログラムカートリッジおよびカートリッジインターフェース（例えば、ビデオゲームデバイスに見られるようなもの）、リムーバブルメモリチップ（例えば、ＥＰＲＯＭまたはＰＲＯＭ）および関連ソケット、ならびに、ソフトウェアおよびデータがリムーバブル格納ユニット１５２２からコンピュータシステム１５００に転送されることを可能にするその他のリムーバブル格納ユニット１５２２およびコンピュータシステム１５００を含み得る。

【０１１４】

コンピュータシステム１５００はまた、通信インターフェース１５２４を含み得る。通信インターフェース１５２４は、ソフトウェアおよびデータが、コンピュータシステムと外部デバイスとの間で転送されることを可能にする。通信インターフェース１５２４は、モデム、ネットワークインターフェース（例えば、イーサネット（登録商標）カード）、通信ポート、ＰＣＭＣＩＡスロットおよびカード等を含み得る。通信インターフェース１５２４を介して転送されるソフトウェアおよびデータは、信号の形態であり得、この信号は、通信インターフェース１５２４によって受信されることが可能な電子的、電磁的、光学的、またはその他の信号であり得る。これらの信号は、通信パス１５２６を介して通信インターフェース１５２４に提供される。通信パス１５２６は、信号を搬送し、ワイヤまたはケーブル、光ファイバー、電話回線、携帯電話リンク、ＲＦリンクまたはその他の通信チャネルを用いて実装され得る。

【０１１５】

本書面において、用語「コンピュータプログラム媒体」および「コンピュータ使用可能

媒体」は、例えばリムーバブル格納ユニット１５１８、リムーバブル格納ユニット１５２２、ハードディスクドライブ１５１２にインストールされたハードディスク、ならびに、通信パス１５２６を介して搬送される信号等の媒体を広く意味するように用いられる。また、コンピュータプログラム媒体およびコンピュータ使用可能媒体は、例えばメインメモリ１５０８および２次メモリ１５１０等のメモリを意味し得、このメモリは、メモリ半導体（例えば、ＤＲＡＭ等）であり得る。これらのコンピュータプログラム製品は、コンピュータシステム１５００にソフトウェアを提供する手段である。

【０１１６】

コンピュータプログラム（コンピュータ制御論理とも称される）は、メインメモリ１５０８および／または２次メモリ１５１０に格納される。コンピュータプログラムはまた、通信インターフェース１５２４を介して受信され得る。そのようなコンピュータプログラムは、実行されたときに、コンピュータシステム１５００が、本明細書中で議論された本発明を実行することを可能にする。特に、これらのコンピュータプログラムは、実行されたときに、プロセッサ１５０４が、例えば上記で議論された図２のフローチャート２００によって示された方法のステップのような、本発明のプロセスを実行することを可能にする。したがって、そのようなコンピュータプログラムは、コンピュータシステム１５００のコントローラを代表する。ソフトウェアを用いて本発明が実行される場合、そのソフトウェアは、リムーバブル格納ドライブ１５１４、インターフェース１５２０、ハードドライブ１５１２または通信インターフェース１５２４を用いることにより、コンピュータプログラム製品に格納されたり、コンピュータシステム１５００にロードされたりし得る。

【０１１７】

本発明の実施形態はまた、任意のコンピュータ使用可能媒体上に格納されたソフトウェアを含むコンピュータ製品に関係し得る。そのようなソフトウェアは、１つ以上のデータ処理デバイス内で実行されたときに、（少なくとも１つの）データ処理デバイスに、本明細書中で議論されたような動作を行わせる。任意のコンピュータ使用可能または読み取り可能な媒体を利用する本発明の実施形態は、現在公知であるか、将来公知になる。コンピュータ使用可能媒体の例は、プライマリ格納デバイス（例えば、任意のタイプのランダムアクセスメモリ）、２次格納デバイス（例えば、ハードドライブ、フロッピー（登録商標）ディスク、ＣＤ　ＲＯＭ、ＺＩＰディスク、テープ、磁気格納デバイス、光学格納デバイス、ＭＥＭＳ、ナノテクノロジー格納デバイス等）、通信媒体（例えば、有線および無線の通信ネットワーク、ローカルエリアネットワーク、ワイドエリアイントラネット等）を含むが、これらには限定されない。

【０１１８】

（結論）

本発明の例示的な実施形態が提示されてきた。本発明は、これらの例に限定されるものではない。これらの例は、本明細書中では、例示を目的として示されており、限定目的として示されてはいない。当業者は、本明細書中に含まれる教示に基づくことにより、代替案（本明細書中に記載されているものの均等、拡張、変形、逸脱（ｄｅｖｉａｔｉｏｎ）等）を明白に理解し得る。そのような代替案は、本発明の範囲および精神に含まれる。

【図１】

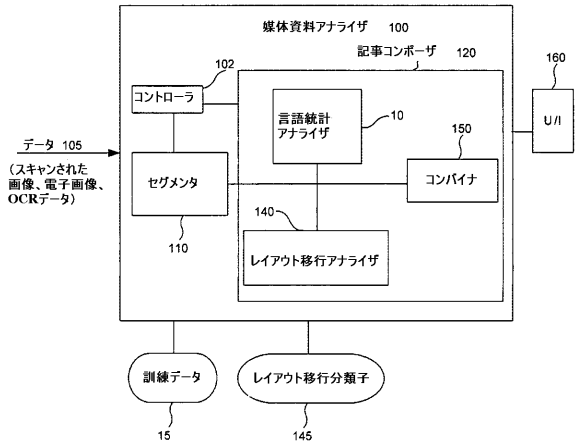


FIG. 1

【図２】

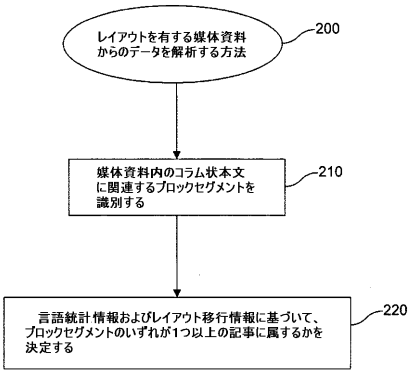


FIG. 2

【図３】

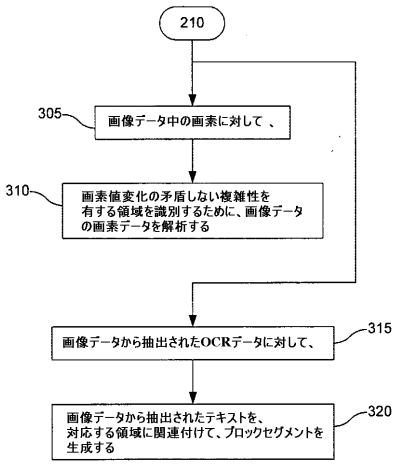


FIG. 3

【図４】

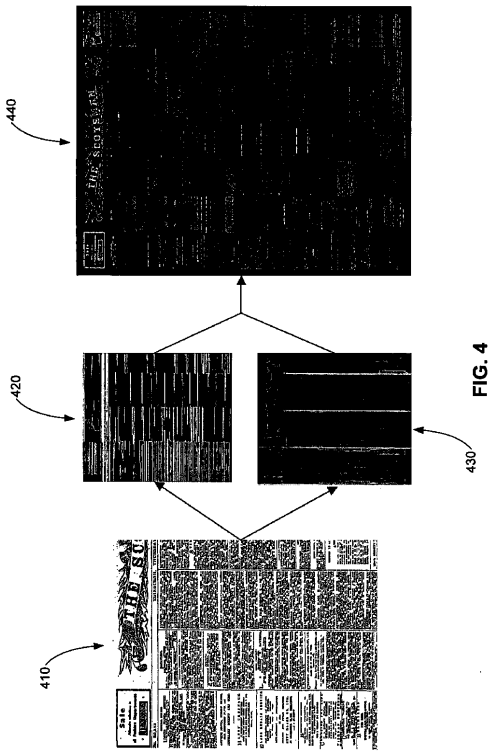


FIG. 4

【図５】

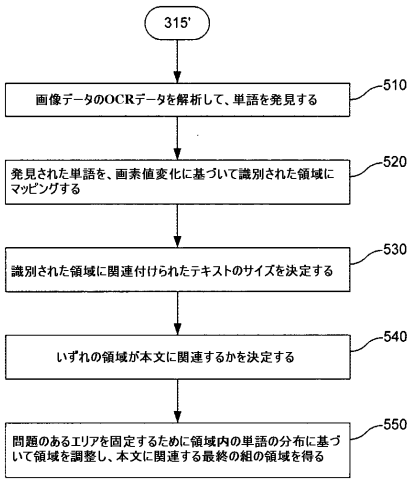


FIG. 5

【図６】

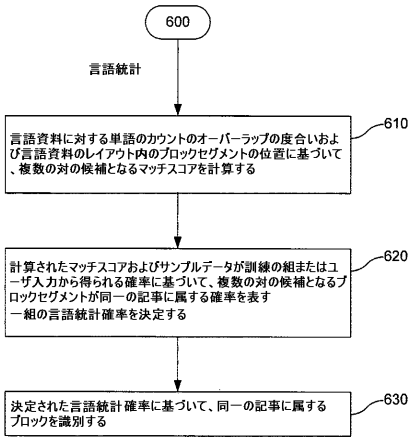


FIG. 6

【図７Ａ】

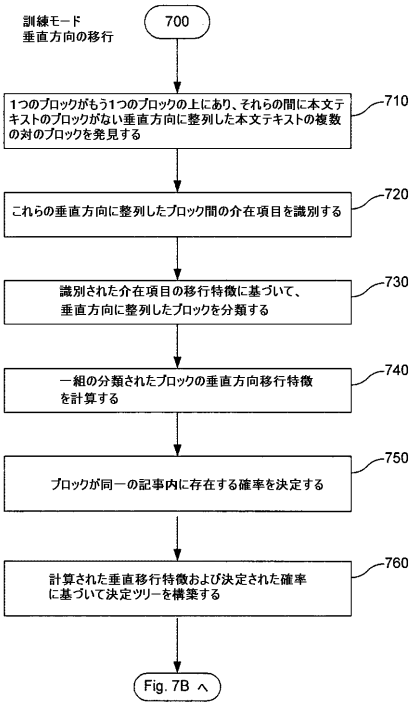


FIG. 7A

【図７Ｂ】

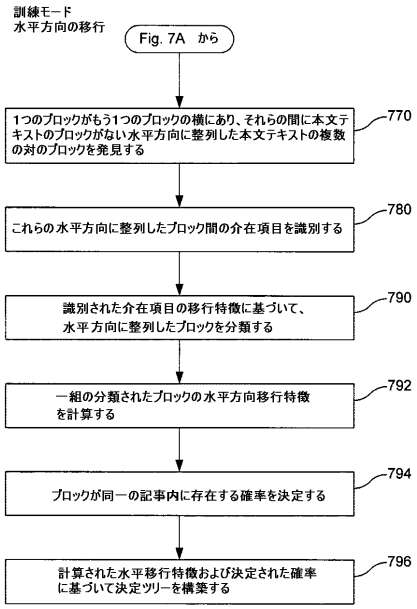


FIG. 7B

【図８】

FIG. 8

【図９Ａ】

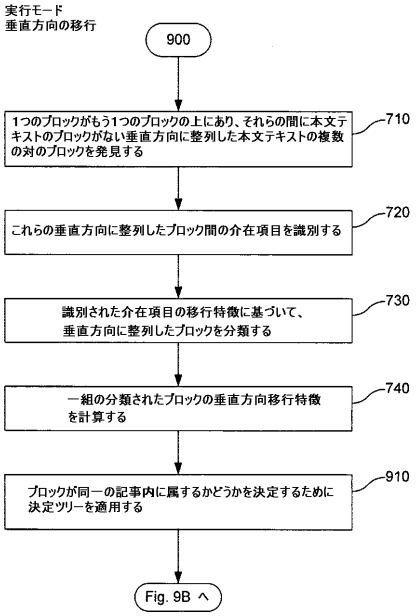


FIG. 9A

【図９Ｂ】

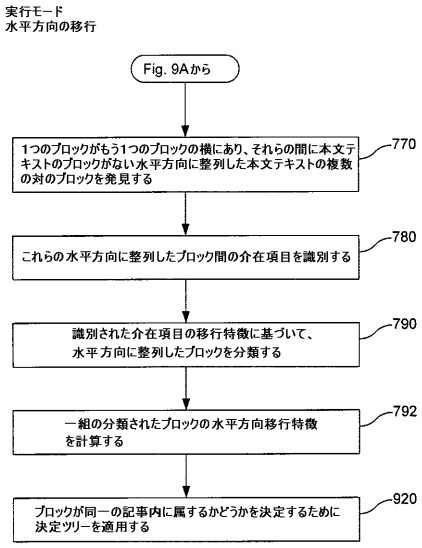


FIG. 9B

【図１０Ａ】

ページ1－最終的な決定ツリー例

FIG. 10A

【図１０Ｂ】

ページ14　－最終的な決定ツリー例

FIG. 10B

【図１０Ｃ】

ページ1　－純粋な言語統計例

FIG. 10C

【図１０Ｄ】

ページ14　－純粋な言語統計例

FIG. 10D

【図１１】

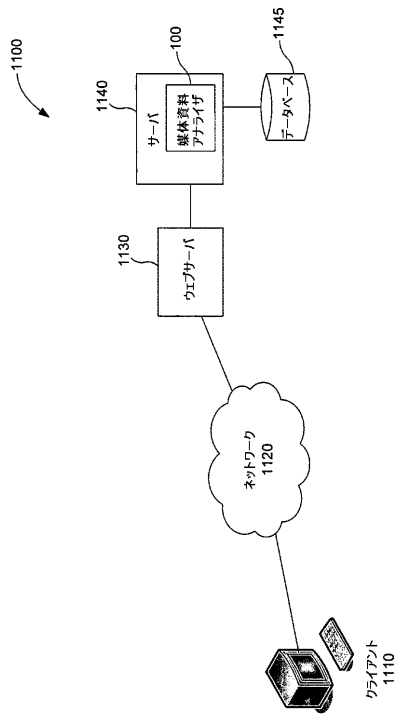


FIG. 11

【図１２】

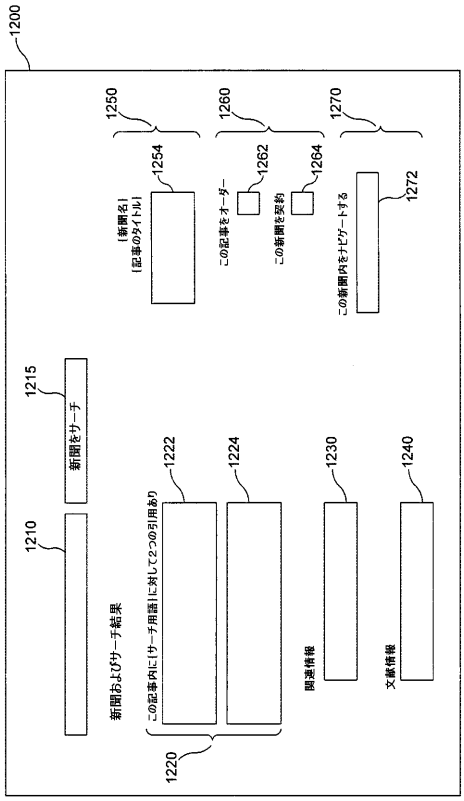


FIG. 12

【図１３】

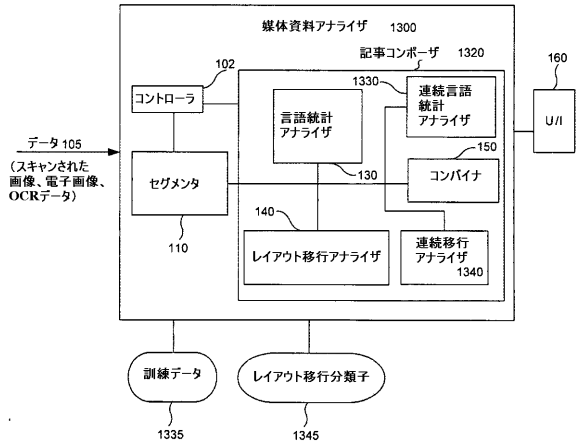
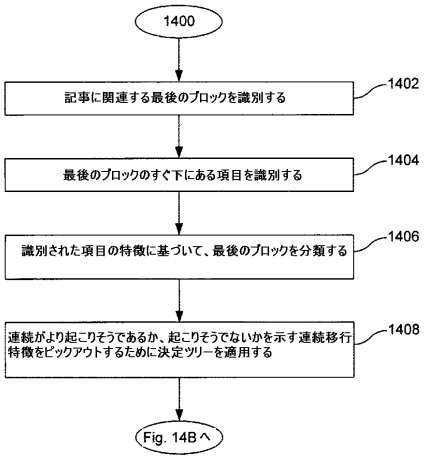


FIG. 13

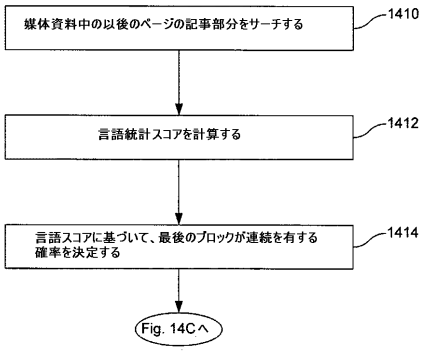
【図１４Ａ】



連続移行実行モード

FIG. 14A

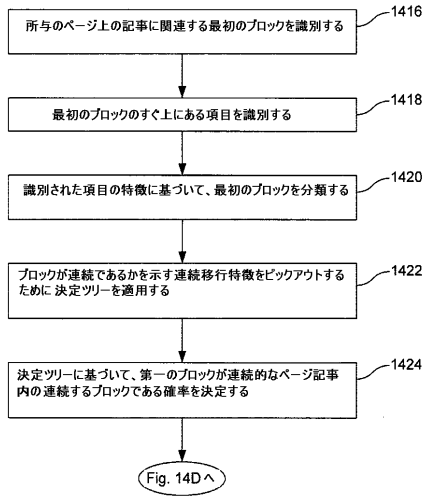
【図１４Ｂ】



連続移行実行モード

FIG. 14B

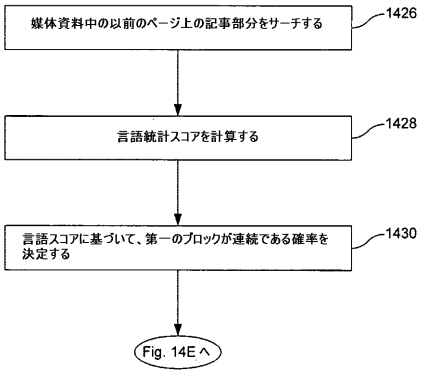
【図１４Ｃ】



連続移行実行モード

FIG. 14C

【図１４Ｄ】



連続移行実行モード

FIG. 14D

【図１４Ｅ】

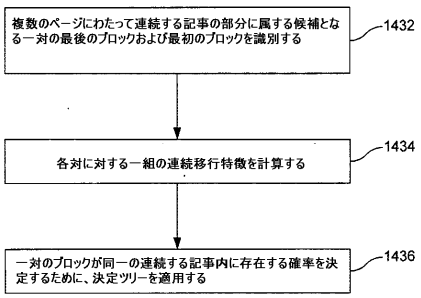


FIG. 14E

【図１５】

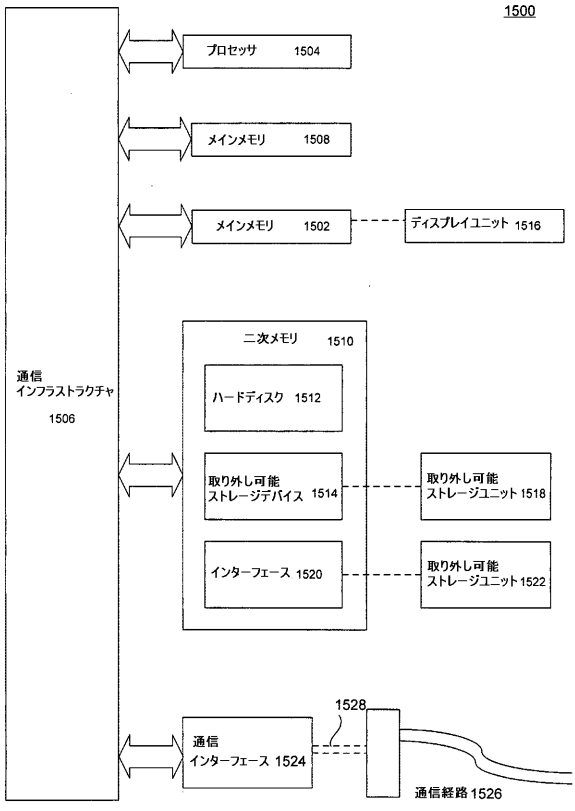


FIG. 15

【手続補正書】
【提出日】平成21年6月29日(2009.6.29)
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】００１７
【補正方法】変更
【補正の内容】
【００１７】
本発明の実施形態は、添付の図面を参照して記載される。図面において、同様の参照番号は同一の要素または機能的に類似の要素を示し得る。ある要素が第１に現れる図面は、対応する参照番号のもっとも左の桁によって概して示される。
例えば、本発明は以下の項目を提供する。
（項目１）
レイアウトを有し、かつ複数のページにわたって広がる１つ以上の連続する記事を含む媒体資料を表すデータを解析する媒体資料アナライザであって、
（ａ）該媒体資料のページ内のコラム状の本文テキストと関連するブロックセグメントを識別するセグメンタと、
（ｂ）言語統計情報および連続移行情報に基づいて、該識別されたブロックセグメントのいずれが該媒体資料内の複数のページにわたって広がる連続する記事に属するかを決定する記事コンポーザと
を備えている、媒体資料アナライザ。
（項目２）
上記記事コンポーザは、連続レイアウト移行アナライザを含み、該連続レイアウト移行アナライザは、第１のページ内の候補となる連続する記事部分に関連する最後のブロックセグメントを識別し、該最後のブロックセグメントの下の１つ以上の項目を識別し、該識別された１つ以上の項目の少なくとも１つの特徴に基づいて、該最後のブロックセグメントを分類し、そして、決定ツリーを適用して、該最後のブロックセグメントが連続する記事内にある確率を示す１つ以上の連続移行特徴を選び出す、項目１に記載の媒体資料アナライザ。
（項目３）
上記記事コンポーザは、連続言語統計アナライザを含み、該連続言語統計アナライザは、上記連続する記事の上記第１のページから連続するページにおける記事部分に対して、言語統計情報を計算し、そして該計算された連続する言語統計情報に基づいて、上記候補となる連続する記事部分内の最後のブロックセグメントが、連続する記事部分を有する確率を決定する、項目２に記載の媒体資料アナライザ。
（項目４）
上記連続レイアウト移行アナライザは、さらに、連続するページ内の候補となる連続する記事と関連する第１のブロックセグメントを識別し、該第１のブロックセグメントの上の１つ以上の項目を識別し、該１つ以上の識別された項目の少なくとも１つの特徴に基づいて、該第１のブロックセグメントを分類し、そして決定ツリーを適用して、該第１のブロックセグメントが連続する記事内にある確率を示す１つ以上の連続移行特徴を選び出し、該適用された決定ツリーに基づいて、該第１のブロックセグメントが連続する記事である確率を決定する、項目３に記載の媒体資料アナライザ。
（項目５）
上記連続言語統計アナライザは、さらに、上記第１のブロックセグメントを有するページよりも前のページ内の記事部分に対して、言語統計情報を計算し、該計算された、より前のページの言語統計情報に基づいて、上記候補となる連続する記事部分内の該第１のブロックセグメントが連続する記事部分を有する確率を決定する、項目４に記載の媒体資料アナライザ。
（項目６）

上記連続レイアウト移行アナライザは、さらに、候補となる一対の最後および第１のブロックセグメントに対する連続移行特徴を識別し、該一対の最後および第１のブロックセグメントに対して一組の連続移行特徴を決定し、そして決定ツリーを適用して、該一組の決定された連続移行特徴に基づいて、該候補となる一対の最後および第１のブロックセグメントが、上記媒体資料内の複数のページにわたる同一の連続する記事に属する確率を決定する、項目５に記載の媒体資料アナライザ。

（項目７）

上記言語統計情報は、単語頻度情報を備え、上記連続言語統計アナライザは、上記最後のブロックセグメント内のテキストおよび連続するページ上の上記記事部分内のテキストにおける単語頻度に基づいて、マッチスコアを計算する、項目３に記載の媒体資料アナライザ。

（項目８）

上記言語統計情報は、単語頻度情報を備え、上記連続言語統計アナライザは、上記第１のブロックセグメント内のテキストおよびより前のページ上の上記記事部分内のテキストにおける単語頻度に基づいて、マッチスコアを計算する、項目５に記載の媒体資料アナライザ。

（項目９）

レイアウトを有し、かつ複数のページにわたって広がる１つ以上の連続する記事を含む媒体資料を表すデータを解析するコンピュータ実装された方法であって、

（ａ）該媒体資料のページ内のコラム状の本文テキストに関連するブロックセグメントを識別することと、

（ｂ）言語統計情報および連続移行情報に基づいて、該識別されたブロックセグメントのいずれが該媒体資料内の複数のページにわたって広がる連続する記事に属するかを決定することと

を包含する、方法。

（項目１０）

上記記事決定ステップは、

候補となる連続する記事部分内のブロックセグメントに対する連続レイアウト移行情報を解析することと、

該候補となる連続する記事部分内のテキストに対する言語統計を解析することと

を含む、項目９に記載の方法。

（項目１１）

上記連続レイアウト移行情報を解析するステップは、

第１のページ内の候補となる連続記事部分に関連する最後のブロックセグメントを識別することと、

該最後のブロックセグメントより下の１つ以上の項目を識別することと、

該識別された１つ以上の項目の少なくとも１つの特徴に基づいて、該最後のブロックセグメントを分類することと、

決定ツリーを適用して、該最後のブロックセグメントが連続する記事内に存在する確率を示す１つ以上の連続移行特徴を選ぶことと

を包含する、項目１０に記載の方法。

（項目１２）

上記言語統計解析ステップは、

上記連続する記事の第１のページから連続するページ上の記事部分に対する、言語統計情報を計算することと、

該計算された連続する言語統計情報に基づいて、上記候補となる連続する記事部分内の上記最後のブロックセグメントが連続する記事部分を有する確率を決定することと

を包含する、項目１１に記載の方法。

（項目１３）

上記連続レイアウト移行情報を解析するステップは、

連続するページ内の候補となる連続する記事に関連する第１のブロックセグメントを識別することと、
該第１のブロックセグメントより上の１つ以上の項目を識別することと、
該１つ以上の識別された項目の少なくとも１つの特徴に基づいて、該第１のブロックセグメントを分類することと、
決定ツリーを適用して、該第１のブロックセグメントが連続する記事内に存在する確率を示す１つ以上の連続移行特徴を選び出し、そして該適用された決定ツリーに基づいて、該第１のブロックセグメントが連続する記事内に存在する確率を決定することと
をさらに包含する、項目１２に記載の方法。
（項目１４）
上記言語統計解析ステップは、
上記第１のブロックセグメントを有するページよりも前のページ内の記事部分に対して、言語統計情報を計算することと、
該計算されたより前のページの言語統計情報に基づいて、上記候補となる連続する記事部分内の該第１のブロックセグメントが連続する記事部分を有する確率を決定することと
をさらに包含する、項目１３に記載の方法。
（項目１５）
上記連続レイアウト移行解析ステップは、
候補となる一対の最後および第１のブロックセグメントを識別することと、
該一対の最後および第１のブロックセグメントに対する一組の連続移行特徴を決定することと、
決定ツリーを適用して、該一組の決定された連続移行特徴に基づいて、該候補となる一対の最後および第１のブロックセグメントが、上記媒体資料内の複数のページにわたる同一の連続する記事に属する確率を決定することと
をさらに含む、項目１２に記載の方法。
（項目１６）
上記言語統計情報は、単語頻度情報を備え、上記連続言語統計解析ステップは、上記最後のブロックセグメント内のテキストおよび連続するページの上記記事部分内のテキストにおける単語頻度に基づいて、マッチスコアを計算することを含む、項目１２に記載の方法。
（項目１７）
上記言語統計情報は、単語頻度情報を備え、上記連続言語統計解析ステップは、上記第１のブロックセグメント内のテキストおよび以前のページの上記記事部分内のテキストにおける単語頻度に基づいて、マッチスコアを計算することを含む、項目１４に記載の方法。
（項目１８）
レイアウトを有する媒体資料内の複数のページにわたって広がる連続する記事を構成する記事コンポーザであって、
連続レイアウト移行アナライザと、
連続言語統計アナライザと
を備え、該連続レイアウト移行アナライザは、異なるページ上の候補となる記事の最後のブロックセグメントおよび第１のブロックセグメントが同じ連続する記事内に存在する確率を示す１つ以上の連続移行特徴を選び出すために、決定ツリーを適用し、
該連続言語統計アナライザは、異なるページ上の異なる記事部分に対する言語統計情報を計算し、該計算された言語統計情報に基づいて、候補となる記事部分の第１および最後のブロックセグメントが連続する記事部分を有する確率を決定し、それにより、該記事コンポーザは、解析された連続レイアウト移行特徴および該計算された言語統計に従って、該第１および最後のブロックセグメントが同じ連続する記事に属する確率に基づいて、複数のページにわたる連続する記事を構成することが可能である、記事コンポーザ。

【手続補正２】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
レイアウトを有する媒体資料を表すデータを解析する媒体資料アナライザであって、
該媒体資料内のコラム状の本文テキストと関連するブロックセグメントを識別するセグメンタと、
該セグメンタによって出力された候補となるブロックセグメントに対する言語統計を計算し、言語統計情報内のオーバーラップに基づいて、候補となるブロックセグメントが同一の記事に属する確率を決定する言語統計アナライザと
を備えている、媒体資料アナライザ。
【請求項２】
言語統計情報およびレイアウト移行情報に基づいて、前記識別されたブロックセグメントのうちのいずれが前記媒体資料内の１つ以上の記事に属するかを決定する記事コンポーザをさらに備えている、請求項１に記載の媒体資料アナライザ。
【請求項３】
前記記事コンポーザは、連続レイアウト移行アナライザを含み、該連続レイアウト移行アナライザは、第１のページ内の候補となる連続する記事部分に関連する前記識別されたブロックセグメントのうちの最後のブロックセグメントを識別し、該最後のブロックセグメントの下の１つ以上の項目を識別し、該識別された１つ以上の項目の少なくとも１つの特徴に基づいて、該最後のブロックセグメントを分類し、そして、決定ツリーを適用して、該最後のブロックセグメントが連続する記事にある確率を示す１つ以上のレイアウト移行特徴を選び出し、該識別されたブロックセグメントは、１つ以上の記事に属し、該１つ以上の記事は、該媒体資料内の複数のページにわたって連続し、かつ広がる、請求項２に記載の媒体資料アナライザ。
【請求項４】
前記言語統計情報は、単語頻度情報を備え、前記言語統計アナライザは、言語資料全体に対する各ブロックセグメントにおける単語頻度と、一対の候補となるブロックセグメント間のコサイン距離類似性とに基づいて、該一対の候補となるブロックセグメントに対するマッチスコアを計算する、請求項２に記載の媒体資料アナライザ。
【請求項５】
前記言語統計アナライザは、計算されたマッチスコアと、同一の記事に属するブロックセグメントの所定のポジティブな例と同一の記事に属さないブロックセグメントの所定のネガティブな例とを有するサンプルデータと、に基づいて、一対の候補となるブロックセグメントが前記資料媒体内の同一の記事に属する確率を決定する、請求項２に記載の媒体資料アナライザ。
【請求項６】
前記言語統計アナライザは、訓練データセット内の記事の集合から、前記ポジティブなデータ例とネガティブなデータ例とを自動的に選択する、請求項５に記載の媒体資料アナライザ。
【請求項７】
前記所定のポジティブなデータ例とネガティブなデータ例とは、前記媒体資料の画像からの光学文字認識を介して抽出されたテキストデータの表示から、ユーザによって、ユーザインタフェースにおいて選択される、請求項５に記載の媒体資料アナライザ。
【請求項８】
レイアウトを有する媒体資料を表すデータを解析するコンピュータ実装された方法であって、

該媒体資料内のコラム状の本文テキストに関連するブロックセグメントを識別することと、
言語統計情報およびレイアウト情報に基づいて、該識別されたブロックセグメントのいずれが該媒体資料内の１つ以上の記事に属するかを決定することと
を包含し、
該決定することは、
候補となるブロックセグメントに対する言語統計を計算することと、
言語統計情報におけるオーバーラップに基づいて、比較されるブロックセグメントが同一の記事に属することを決定することと
を含む、方法。
【請求項９】
前記識別されたブロックセグメントは、１つ以上の記事に属し、該１つ以上の記事は前記媒体資料内の複数のページにわたって連続し、かつ広がる、請求項８に記載の方法。
【請求項１０】
前記同一の記事内に存在することが決定された１つ以上のブロックセグメントからテキストを表示することをさらに含む、請求項８に記載の方法。
【請求項１１】
前記言語統計情報は、単語頻度情報を備え、前記計算することは、言語資料全体に対する各ブロックセグメントにおける単語頻度と、一対の候補となるブロックセグメント間のコサイン距離類似性とに基づいて、該一対の候補となるブロックセグメントに対するマッチスコアを計算することを含む、請求項８に記載の方法。
【請求項１２】
前記確率を決定するステップは、前記計算されたマッチスコアと、同一の記事に属するブロックセグメントの所定のポジティブな例と同一の記事に属さないブロックセグメントの所定のネガティブな例とを有するサンプルデータと、に基づいて、前記一対の候補となるブロックセグメントが前記資料媒体内の同一の記事に属する確率を決定することを含む、請求項１１に記載の方法。
【請求項１３】
ユーザが、前記ポジティブなデータ例とネガティブなデータ例とを、前記媒体資料の画像からの光学文字認識を介して抽出されたテキストデータの表示から選択することを可能にすることをさらに含む、請求項１２に記載の方法。
【請求項１４】
前記決定することは、言語統計情報における前記オーバーラップに基づいて決定された確率に基づいて、前記候補となるブロックセグメントが、前記媒体資料内の同一の記事に属するかどうかを識別することを含む、請求項８に記載の方法。
【請求項１５】
レイアウトを有する媒体資料を表すデータを解析する媒体資料アナライザであって、
該媒体資料内のコラム状の本文テキストと関連するブロックセグメントを識別するセグメンタと、
言語統計情報およびレイアウト移行情報に基づいて、該識別されたブロックセグメントのいずれが該媒体資料内の１つ以上の記事に属するかを決定する記事コンポーザと
を備え、
該記事コンポーザは、レイアウト移行アナライザを含み、該レイアウト移行アナライザは、該セグメンタによって出力された候補となるブロックセグメント内のレイアウト移行特徴を解析し、該候補となるブロックセグメントが該媒体資料内の同一の記事に属するか動かを決定し、
該レイアウト移行アナライザは、該レイアウト移行特徴を該候補となるブロックセグメントから計算し、該計算されたレイアウト移行特徴に基づいて、該候補となるブロックセグメントが該媒体資料内の同一の記事内に属するかどうかを決定するために、所定のレイアウト移行分類子を適用する、媒体資料アナライザ。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 07/23233

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/00 (2008.04)
USPC - 715/255
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 715/255

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 715/200, 204, 243, 255, 264, 272

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Electronic Databases Searched: USPTO WEST (PGPUB,USPAT,USOCR,EPAB,JPAB); GOOGLE SCHOLAR; DIALOG PRO
Search Terms Used: text/printed/media/digital articles/document extraction/metadata/segmentation comparison, multiple/successive column/page spanning, textual keyword/metadata/language/syntax/linguistic analysis, etc.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2006/0080309 A1 (YACOUB et al.) 13 April 2006 (13.04.2006) Abstract, Para [0017]-[0141]	1-18
A	US 2006/0184525 A1 (JONES et al.) 17 August 2006 (17.08.2006) Entire Document	1-18
A	US 2004/0122811 A1 (PAGE) 24 June 2004 (24.06.2004) Entire Document	1-18
A	US 2003/0229854 A1 (LEMAY) 11 December 2003 (11.12.2003) Entire Document	1-18

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21 April 2008 (21.04.2008)	14 MAY 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LS,MW,MZ,NA,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IS,IT,LT,LU,LV,MC,MT,NL,PL,PT,RO,SE,SI,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BH,BR,BW,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IS,JP,KE,KG,KM,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LT,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PG,PH,PL,PT,RO,RS,RU,SC,SD,SE,SG,SK,SL,SM,SV,SY,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,ZA,ZM,ZW

(72)発明者 スミス, レイ
アメリカ合衆国 カリフォルニア ９４０４３, マウンテン ビュー, アンフィシアター パークウェイ １６００, グーグル インコーポレイテッド 気付
(72)発明者 ビンセント, ルーク
アメリカ合衆国 カリフォルニア ９４０４３, マウンテン ビュー, アンフィシアター パークウェイ １６００, グーグル インコーポレイテッド 気付
(72)発明者 ブルームバーグ, ダン
アメリカ合衆国 カリフォルニア ９４０４３, マウンテン ビュー, アンフィシアター パークウェイ １６００, グーグル インコーポレイテッド 気付

Ｆターム(参考) 5B029 AA01 AA02 BB02 CC26 CC27 CC29
5B075 NR03 NR12
5B109 NA03