

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5247311号
(P5247311)

(45) 発行日 平成25年7月24日 (2013. 7. 24)

(24) 登録日 平成25年4月19日 (2013. 4. 19)

(51) Int. Cl.

F I

G 0 6 F 17/21 (2006. 01)

G 0 6 F 17/21 5 3 0 A

G 0 6 T 11/60 (2006. 01)

G 0 6 T 11/60 1 0 0 A

請求項の数 12 (全 20 頁)

(21) 出願番号 特願2008-221912 (P2008-221912)
 (22) 出願日 平成20年8月29日 (2008. 8. 29)
 (65) 公開番号 特開2010-55512 (P2010-55512A)
 (43) 公開日 平成22年3月11日 (2010. 3. 11)
 審査請求日 平成23年8月24日 (2011. 8. 24)

(73) 特許権者 000001007
 キヤノン株式会社
 東京都大田区下丸子3丁目30番2号
 (74) 代理人 110001243
 特許業務法人 谷・阿部特許事務所
 (74) 代理人 100077481
 弁理士 谷 義一
 (74) 代理人 100088915
 弁理士 阿部 和夫
 (72) 発明者 伊丹 剛
 東京都大田区下丸子3丁目30番2号 キ
 ヤノン株式会社内

審査官 成瀬 博之

最終頁に続く

(54) 【発明の名称】 電子文書処理装置および電子文書処理方法

(57) 【特許請求の範囲】

【請求項 1】

レイアウト情報を含む電子文書から複数のテキスト列を抽出する手段と、
 前記抽出されたテキスト列それぞれのベースラインを検出する手段と、
 前記抽出されたテキスト列それぞれに対して、前記ベースライン上の、前方に第1の線分を設け、かつ後方に前記第1の線分とは異なる種類の第2の線分を設ける手段と、
 異なる複数のテキスト列について、前記異なるテキスト列に設けられる前記第1の線分と前記第2の線分とが重なる場合に、当該異なる複数のテキスト列は連鎖していると判断する手段と

を備えることを特徴とする電子文書処理装置。

10

【請求項 2】

前記設ける手段は、前記電子文書の座標上に配置される前記第1の線分および第2の線分を設け、前記判断する手段は、前記異なるテキスト列に設けられる前記第1の線分と前記第2の線分とが前記電子文書上の前記座標上で重なる場合に、前記異なる複数のテキスト列は連鎖していると判断することを特徴とする請求項1に記載の電子文書処理装置。

【請求項 3】

前記判断する手段は、前記第1の線分と前記第2の線分とが、条件とする角度で交差している場合に、前記異なるテキスト列は連鎖していると判断することを特徴とする請求項1または2に記載の電子文書処理装置。

【請求項 4】

20

前記条件とする角度の許容範囲を切り替える手段をさらに備えることを特徴とする請求項 3 に記載の電子文書処理装置。

【請求項 5】

前記抽出されたテキスト列に該テキスト列の識別用の識別子を割り当てる手段をさらに備え、

前記判断する手段は、前記連鎖していると判断されたテキスト列のペア毎に前記識別子のペアを作成し、前記識別子のペアの組み合わせに基づいて、前記識別子のペアを 1 次元にソートすることを特徴とする請求項 1 乃至 4 のいずれかに記載の電子文書処理装置。

【請求項 6】

前記連鎖を判定する上でのパラメータとなるベースライン上の前方および後方に設ける第 1 および第 2 の線分の長さを切り替える手段をさらに備えることを特徴とする請求項 1 乃至 5 のいずれかに記載の電子文書処理装置。

【請求項 7】

前記判断する手段は、前記連鎖の判断において、1 つのテキスト列の後段に複数のテキスト列の連鎖が認められる場合に、テキスト列の連鎖が認められる中で最も、前段のテキスト列と後段のテキスト列との成す角度が小さい方のテキスト列を優先して、テキスト列を連鎖させることを特徴とする請求項 1 乃至 6 のいずれかに記載の電子文書処理装置。

【請求項 8】

連鎖されたと判断されたテキスト列を 1 つのグループにグループ化する手段をさらに備えることを特徴とする請求項 1 乃至 7 のいずれかに記載の電子文書処理装置。

【請求項 9】

前記グループに該グループの識別用の識別を与える手段をさらに備えることを特徴とする請求項 8 に記載の電子文書処理装置。

【請求項 10】

前記グループ化されたテキスト列のオフセット座標を検出し、該オフセット座標の座標値によって整列させることで、テキスト列の抽出順位を決める手段をさらに備えることを特徴とする請求項 8 に記載の電子文書処理装置。

【請求項 11】

電子文書処理装置の抽出手段が、レイアウト情報を含む電子文書から複数のテキスト列を抽出する工程と、

電子文書処理装置の検出手段が、前記抽出されたテキスト列それぞれのベースラインを検出する工程と、

電子文書処理装置の解析手段が、前記抽出されたテキスト列それぞれに対して、前記ベースライン上の、前方に第 1 の線分を設け、かつ後方に前記第 1 の線分とは異なる種類の第 2 の線分を設ける工程と、

電子文書処理装置の判定手段が、異なる複数のテキスト列について、前記異なるテキスト列に設けられる前記第 1 の線分と前記第 2 の線分とが重なった場合に、当該異なる複数のテキスト列は連鎖していると判断する工程と

を有することを特徴とする電子文書処理方法。

【請求項 12】

コンピュータを、請求項 1 乃至 10 のいずれかに記載の電子文書処理装置として機能させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は電子文書処理装置及び電子文書処理方法に関し、特に、電子文書レイアウト情報を保持した電子文書からテキストを抽出するための電子文書処理装置及び電子文書処理方法に関するものである。

【背景技術】

【0002】

従来、電子文書におけるテキストの検索は、電子文書内に含まれるテキストを抽出し、その中に検索キーが含まれるかどうかを判定することによって行っている。検索手法としては、抽出された文字の中に検索キーが一部でも含まれるかどうかによって判定する方法が一般的である。

【 0 0 0 3 】

より限定的な検索手法としては、単語が完全に含まれるかどうかで判定する完全一致単語検索、空白を含めた複数の単語などからなるフレーズを検索キーとするフレーズ検索がある。また、電子文書を総なめにして検索語がないかを探し出し、仮に検索語と一致するテキストを含んだ電子文書が見つければ、そのテキストがどのような個所に書かれたものかを検索結果として持ってくる全文検索がある。

10

さらには、検索したい内容を文章として入力し、その文章に近い内容の情報を探すことができる概念検索といった高度な検索手法も存在している。

【 0 0 0 4 】

たとえば、電子文書から抽出されるテキストが、“He is a good boy. But, she is a bad girl.” であるとする。概念検索では“good boy”は勿論のこと、“nice boy”でも検索がヒットする。また、“bad boy”では検索にヒットしない。これは、抽出されたテキストが意味する概念に則しているかについて、検索処理を提供できているからである。

したがって概念検索を行う場合、抽出されたテキストは、日本語であれば日本語としての文章性であったり、英語であれば英語としての文章性であったりが保たれている必要がある。

20

【 0 0 0 5 】

一方、文字のレイアウト情報を保持する電子文書において、テキスト描画を表すコマンド（以降、テキスト描画コマンドと呼ぶ）の順番と、テキスト描画コマンドがページ上で指定される描画開始位置が、それぞれ独立している場合がある。例えば、ページ上の中心あたりから、1つ目のテキスト描画コマンドがはじまり、次にページ上の下方あたりから2つ目のテキスト描画コマンドがはじまり、最後にページ上の上方あたりから最後のテキスト描画コマンドがはじまるといったケースである。

【 0 0 0 6 】

実際の電子文書の例で、PDF (Portable Document Format ; 登録商標) や PDL (Page Description Language) などでのそのような表現がなされる場合がある。例えば、PDFを作成するソフトウェアの1つに、プリンタドライバ型のソフトウェアがある。これは、元の下稿を作成したワープロアプリケーションやドローアプリケーションにおいて、印刷指示を行う際、通常のプリンタドライバの代わりに、このPDF作成用ドライバを選択すると、印刷コマンドからPDFファイルを作成するソフトウェアである。

30

【 0 0 0 7 】

このとき、元の下稿を作成したアプリケーションが、どのようなテキスト描画の順番で、印刷コマンドをPDF作成ドライバに渡すかは、そのアプリケーションに依存する。例えば、そのアプリケーションがレイアウトフリーな電子原稿作成アプリケーション（例えば、Microsoft Office PowerPoint、Microsoft Office Visio ; 共に登録商標）である場合等である。このアプリケーションでは、ページ上の文章の構成を考慮しないようにテキスト描画をしている場合、PDF作成ドライバで作成されるPDFファイルには、テキスト描画コマンドが並べられる順番として文章性を大きく欠いた結果で入力される場合がある。これは、テキスト描画コマンドが表示ページ上の座標位置は正しいが、そのテキスト描画コマンドのPDFファイル内での順番はばらばらであるということである。

40

【 0 0 0 8 】

レイアウトフリーな電子原稿作成アプリケーションの場合、操作者がテキストオブジェクトを生成した順に、そのオブジェクトをシーケンシャルに番号を振って管理している。しかし、操作者はレイアウトフリーに操作できる特性を活かして原稿を作成するので、テ

50

キストオブジェクトを生成した順番に則して文章性を保つようにテキストオブジェクトを配置していくことは必ずしも行われぬ。このような電子原稿から、PDFファイルを作成した場合、例えば図1に示すようなPDFファイルが生成される。

【0009】

図1では、レイアウトフリーな電子原稿作成アプリケーションから作成されたPDFファイルのプレビュー101と、そのPDFファイル内のテキスト描画コマンドの配列102の例を示している。配列102のようなテキスト描画コマンドの配列になった原因は、レイアウトフリーな電子原稿作成アプリケーションにて、テキストオブジェクトを生成したことに起因する。このときのテキストオブジェクトを生成した順番は、「Michael」「Confidential」「sushi」「Michael」「Possibly」「appreciates」の順である。ただし、その後並べ変えて符号101のプレビューで示すような配置にし、操作者が意図する通りの文章に構成している。この電子原稿から、PDFファイルに変換した場合、配列102に示す順番でテキスト描画コマンドが入力されてしまう。

10

【0010】

このようなPDFファイルからテキストを抽出しても文章性は保たれていないので、このような結果を受け取った検索エンジンは単語検索程度しかまなず、概念検索等の高度な検索において精度を落としてしまうという問題がある。

【0011】

こういった基本的な問題に対応するため、ページ内のテキストを取得する際にテキスト描画コマンドを座標に応じてソートする先行技術の例として、特許文献1では「ページ記述言語ファイル内に記載されている単語を識別する装置及び方法」が開示されている。特許文献1では、電子文書内のテキスト描画コマンドが記載されている順にテキストを抽出するのではなく、一旦すべてのテキスト描画コマンドと、それに紐付いた座標等のリソース情報を取り出す。さらに、テキスト描画コマンドのオフセット座標(テキスト描画の開始位置)をソートして、そのソート結果の順にテキストを抽出し、テキストの配置位置に則したテキスト抽出結果を得る技術がある。

20

【0012】

【特許文献1】特開平08-194697号公報

【発明の開示】

【発明が解決しようとする課題】

30

【0013】

しかし、レイアウトフリーな電子原稿作成ソフト(アプリケーション)では、例えばアーチ状に並ぶような、装飾されたテキスト列を作成されることもあり、テキスト描画コマンドのオフセット座標の順にソートしてもうまくテキストが取得できないケースがある。

【0014】

なお、本明細書において、「装飾されたテキスト列」とは、例えば、アーチ状、ウェーブ状、サークル(ループ)状、スクエア状、スター状というように、所定の一方方向にテキストを整然と並べて形成するテキスト列ではないテキスト列を指す。

【0015】

図2は、図1に示したPDFファイルを例に、テキスト描画コマンドのオフセット座標の順にソートしてもうまくテキストが取得できない例を示している。

40

【0016】

図2において、テキスト群(テキスト列)が、アーチ状に並ぶように装飾されている。テキスト202は、「Possibly」を示しており、その座標は(4、20)である。テキスト203は、「Michael」を示しており、その座標は(8、25)である。テキスト204は、「appreciates」を示しており、その座標は(12、25)である。テキスト205は、「sushi」を示しており、その座標は(20、17)である。テキスト206は、「Michael」を示しており、その座標は(5、10)である。テキスト207は、「Confidential」を示しており、その座標は(10、10)である。

【0017】

50

これをテキスト描画コマンドのオフセット座標順にソートすると、上記各テキストの座標に基づいて、図2に示す各テキストは次の順番で取得される。すなわち、テキスト203、テキスト204、テキスト202、テキスト205、テキスト206、テキスト207の順で取得されてしまう。しかしながら、図2のようなPDFファイルの場合、このページのデザインが意図するテキストの順番は、テキスト202、テキスト203、テキスト204、テキスト205、テキスト206、テキスト207が自然である。

【0018】

装飾されたテキスト列の装飾によっては、上述の、テキスト描画コマンドのオフセット座標順のソートにより、取得されるテキスト列がバラバラになることもある。

【0019】

このように、従来では、装飾されたテキスト列を含む画像において概念検索といったページ全体の文章性が必要な高度な検索を行う場合、テキスト列の抽出で得られたテキスト列が本来の文章とは異なる場合があった。

【0020】

本発明は、このような問題に鑑みてなされたもので、その目的とするところは、装飾されたテキスト列を含む画像であっても、概念検索といった高度な検索を高精度に実行可能な電子文書処理装置および電子文書処理方法を提供することにある。

【課題を解決するための手段】

【0021】

このような目的を達するために、本発明は、レイアウト情報を含む電子文書から複数のテキスト列を抽出する手段と、前記抽出されたテキスト列それぞれのベースラインを検出する手段と、前記抽出されたテキスト列それぞれに対して、前記ベースライン上の、前方に第1の線分を設け、かつ後方に前記第1の線分とは異なる種類の第2の線分を設ける手段と、異なる複数のテキスト列について、前記異なるテキスト列に設けられる前記第1の線分と前記第2の線分とが重なる場合に、当該異なる複数のテキスト列は連鎖していると判断する手段とを備えることを特徴とする。

【0022】

また、本発明は、電子文書処理方法であって、電子文書処理装置の抽出手段が、レイアウト情報を含む電子文書から複数のテキスト列を抽出する工程と、電子文書処理装置の検出手段が、前記抽出されたテキスト列それぞれのベースラインを検出する工程と、電子文書処理装置の解析手段が、前記抽出されたテキスト列それぞれに対して、前記ベースライン上の、前方に第1の線分を設け、かつ後方に前記第1の線分とは異なる種類の第2の線分を設ける工程と、電子文書処理装置の判定手段が、異なる複数のテキスト列について、前記異なるテキスト列に設けられる前記第1の線分と前記第2の線分とが重なった場合に、当該異なる複数のテキスト列は連鎖していると判断する工程とを有することを特徴とする。

【発明の効果】

【0023】

本発明によれば、テキスト列の連鎖を検知することができるので、装飾されたテキスト列を含む画像であっても、概念検索といった高度な検索を高精度に実行することができる。

【発明を実施するための最良の形態】

【0024】

以下、図面を参照して本発明の実施形態を詳細に説明する。なお、以下で説明する図面で、同一機能を有するものは同一符号を付け、その繰り返しの説明は省略する。

本発明の一実施形態は、装飾されたテキスト列を含む画像に対しても、画像に含まれる文章を再現しつつテキスト列を抽出可能とするために、上記装飾されたテキスト列においてテキスト列の連鎖を検知する。

【0025】

すなわち、本発明の一実施形態に係る電子文書処理装置（情報処理装置とも言う）は、

テキスト抽出する電子文書が先述したような装飾により並べられた場合、テキストの連鎖を検知するために、以下の構成をシーケンシャルに実行する機能を備える。

【 0 0 2 6 】

工程 1 ; 電子文書処理装置は、各テキスト描画コマンドが指すテキスト列のベースラインを元に、2 種類の線分を作成する（例えば、テキスト列の前方に線分 A（第 1 の線分）、テキスト列の後方に線分 B（第 2 の線分）を作成する）。

工程 2 ; 電子文書処理装置は、工程 1 で作成した 2 種類の線分である、線分 A と線分 B を座標上にマッピングする。

工程 3 ; 電子文書処理装置は、該電子文書処理装置にて定義した条件で、第 1 のテキスト列の線分 A と第 2 のテキスト列の線分 B とが重なる場合、その線分 A 側のテキスト列（第 1 のテキスト列）と線分 B 側のテキスト列とは連鎖していると判断する。そして、該連鎖していると判断されたテキスト列をつなげる（グループ化する）。

工程 4 ; 電子文書処理装置は、工程 3 にてグループ化したテキスト列を座標順にソートする。

工程 5 ; 電子文書処理装置は、工程 4 でソートした結果のテキストを抽出する。

【 0 0 2 7 】

このように、本発明の一実施形態では、テキスト列の連鎖を検知してテキスト列をつなげる仕組みを有している。よって、レイアウトフリーな電子原稿作成機能を有するアプリケーションが起源の、テキスト列に装飾が施されているような電子文書であっても、ページ全体の文章性を損なうことなくテキストを取得することができる。

【 0 0 2 8 】

よって、例えば、検索エンジンに本発明の一実施形態にかかる電子文書処理装置を適用することで、概念検索などページ全体の文章性が必要とされるような高度な検索において、その精度を向上させることができる。

【 0 0 2 9 】

（第 1 の実施形態）

図 3 は、本実施形態の電子文書処理装置（情報処理装置）の構成例を示すブロック図である。

【 0 0 3 0 】

符号 3 0 1 は演算制御用の CPU である。符号 3 0 2 はデータ及び指示入力のためのキーボードである。符号 3 0 3 は電子文書画像を表示するディスプレイである。符号 3 0 4 は電子文書を格納するハードディスクである。3 0 5 は装置を制御するプログラムや必要な情報をあらかじめ記憶する ROM である。符号 3 0 6 は様々なワークエリアとして利用される RAM である。

【 0 0 3 1 】

符号 3 0 7 は、電子文書の構造を解析する解析手段に相当し、電子文書内のテキスト描画コマンドが指すテキスト列の配置特性を解析するレイアウト解析処理部である。符号 3 0 8 は、異なるテキスト列が連鎖しているか否かを判断し、連鎖している場合は、該異なるテキスト列をつなげてグループ化するテキスト列結合処理部である。すなわち、テキスト列結合処理部 3 0 8 は、本電子文書処理装置に予め定義された条件で、異なるテキスト列同士の線分 A と線分 B とが重なる場合、その線分 A 側のテキスト列と線分 B 側のテキスト列とは連鎖していると判断してテキスト列をつなげる。次いで、該つながったテキスト列をグループ化する。符号 3 0 9 はグループ化されたテキスト列群を座標順にソートする整列処理部である。

【 0 0 3 2 】

尚、レイアウト解析処理部 3 0 7 や、テキスト列結合部 3 0 8 や、整列処理部 3 0 9 は、ROM 3 0 5 に格納されたコンピュータプログラムを CPU 3 0 1 が実行することで実現される。すなわち、コンピュータプログラムは、本実施形態に係る電子文書処理装置といったコンピュータを、レイアウト解析処理部 3 0 7 や、テキスト列結合部 3 0 8 や、整列処理部 3 0 9 として機能させることができる。なお、コンピュータプログラムが記憶さ

れるコンピュータ読み取り可能な記憶媒体は、ROM 305に限るものではなく、例えばハードディスク 304であってもよい。

【0033】

(本実施形態の電子文書処理装置(情報処理装置)の動作例)

図4は、本実施形態に係る文書処理を行うフローチャートである。

【0034】

ステップS401では、レイアウト解析処理部307は、図5にて後述するレイアウト解析処理を行う。ステップS402では、テキスト列結合処理部308は、図8、9、11にて後述するテキスト列結合処理を行う。ステップS403では、整列処理部309は、図12にて後述する整列処理を行う。ステップS404では、電子文書処理装置は、ステップS401 - S403の処理結果をもって、整列された順番にテキスト列のテキストを抽出する。

10

【0035】

図5は、S401にて行うレイアウト解析処理を表すフローチャートと、その補足図である。

ステップS501において、レイアウト解析処理部307は、電子文書が保持するページ内のテキスト描画コマンドのすべてを取得する。なお、電子文書としてPDFファイルを例とする場合、テキスト描画コマンドはTjで表され、図1の配列102で示すような記載が電子文書中でなされている。

【0036】

20

ステップS502では、レイアウト解析処理部307は、符号5Aで示すように、取得したテキスト描画コマンドに対応するフォント情報から、テキスト描画コマンドが表すテキスト列のベースラインを取得する。ベースラインは図6の符号601で示すように、書体の構成要素であって、フォント情報の構成要素のひとつである。このベースラインは、テキスト列の並びの視線を誘導する役割として必要なラインで、単にラインと呼ばれることもある。テキスト列を並べる際には、図6に示すようにベースライン601の一直線上にそれぞれの文字が並ぶことになる。

【0037】

ステップS503では、レイアウト解析処理部307は、符号5Bで示すように、ステップS502にて取得されたベースラインのオフセット座標を取得する。ベースラインのオフセット座標は、先頭文字のベースラインの開始位置となる。

30

【0038】

ステップS504では、レイアウト解析処理部307は、符号5Cで示すように、ベースライン上において、テキスト列の先頭文字から前方に第1の線分としての線分Aを作成する。続いてステップS505にて、レイアウト解析処理部307は、符号5Dで示すように、ベースライン上において、テキスト列の最終文字から後方に、第1の線分とは種類が異なる第2の線分としての線分Bを作成する。ステップS505の後にはレイアウト解析処理を終了する。

【0039】

このように、本実施形態では、レイアウト解析処理部307は、テキスト列から検出されたベースラインの前方および後方のそれぞれに、互いに異なる種類の線分を設け、該線分を文書画像の座標上に配置する。

40

本実施形態では、後述するが、ある2つのテキスト列(第1のテキスト列および第2のテキスト列とする)に着目した場合、第1および第2のテキスト列の種類の異なる線分同士が所定の条件で重なるか否かで連鎖しているか否かを判断する。そして、互いに種類の異なる線分のうち、線分Aはテキスト列の前方に配置される線分であり、線分Bはテキスト列の後方に配置される線分である。従って、線分Aと線分Bとが重なるか否かで連鎖性を判断することで、前段のテキスト列の最後尾の後に後段のテキスト列の最前を配置することができる。

【0040】

50

なお、レイアウト解析処理部 307 は、線分 A、B を作成する際に、ベースラインの座標上での方向や予め定められた線分の長さ等に基づいて、各テキスト列に対する線分 A および B のそれぞれの関数を求める。この線分 A (線分 B) の関数情報には、線分 A の関数 (軸に平行な関数か、傾きを持った関数か)、線分の長さ、線分の両端の座標を含めることができる。従って、線分 A および線分 B の関数情報を用いれば、線分 A および線分 B を座標上に配置することができる。すなわち、線分 A および線分 B の関数情報は、各線分を座標上に配置するための情報といえる。

【0041】

レイアウト解析処理部 307 は、該求められた線分 A の関数情報および線分 B の関数情報を RAM 306 に格納する。

10

【0042】

線分 A と線分 B の長さは、テキスト列の連鎖を検知する上で、最も重要なバロメータのひとつとなる。本実施形態における電子文書処理装置には、テキスト列の連鎖検知のための線分 A と線分 B の長さを操作者に入力させることができるユーザーインターフェースを提供することができる。なお図 7 は、バロメータを入力するためのユーザーインターフェースの好適な例を示している。

【0043】

図 7 に示すユーザーインターフェースがディスプレイ 303 に表示されると、操作者は、キーボード 302 やポインティングデバイスを用いて、連鎖検知用に設けられる線分の長さを所望に応じて入力することができる。該入力が行われると、CPU 301 は、線分の長さに関するユーザ入力を受付け、該ユーザ入力に応じて、ベースライン上の前方および後方に設ける線分の長さを切り替える。レイアウト解析処理部 307 は、この切り替えられた線分の長さにて線分を作成することができる。

20

なお、本実施形態では、上記線分の長さをユーザ入力により変更可能な構成としても良いし、予め定められた長さを用いるようにしても良い。

【0044】

図 8 は、本実施形態に係る、テキスト列結合処理を表すフローチャートと、その補足図である。

ステップ S801 において、テキスト列結合処理部 308 は、ステップ 501 にて取得されたテキスト描画コマンドの数と同じ数だけバッファを用意する。それぞれのバッファには、テキスト列を構成するテキストと、そのテキスト列が表す線分 A と線分 B の関数情報と、識別用の ID (識別子) を収める。識別用の ID は処理順にシーケンシャルに割り当てて。

30

【0045】

ステップ S802 では、テキスト列結合処理部 308 は、テキスト列の連鎖検知として、バッファの全数 (n1 とする) から、2 個を選ぶ組み合わせに対し、線分 A および線分 B の関数情報に基づいて線分 A と線分 B が平行に重なるケースがあるか確認する。これは、符号 8B で示すように、線分 A と線分 B の関数が同じで、さらに線分 A と線分 B の X 座標領域が共有する箇所があるかの確認処理を行えばよい。すなわち、符号 8B において、テキスト列 81 における、線分 A を線分 a とし、線分 B を線分 b とし、テキスト列 82 における、線分 A を線分 c とし、線分 B を線分 d とする。このとき、テキスト列結合処理部 308 は、線分 A および線分 B の関数情報に基づいて、テキスト列 81 とテキスト列 82 の互いに種類が異なる線分である線分 b と線分 c とを比較して、線分 b と線分 c とが重なるか否かを判断する。なお、傾きが同じ線分 A と線分 B の位置が、誤差により、上下 (または左右) に少しズレている場合が考えられるので、数ドット (例えば 1 ~ 3 ドット) のズレがあったとしても重なっていると判断するように、誤差の許容範囲を予め設けておくのが望ましい。例えば、傾きが同じ線分 A と線分 B それぞれに数ドットの線幅を定義したとき、その線幅で各線分が描画される領域が重なれば、それらのテキスト列はほぼ平行に重なっていると判断するようにしてもよい。また、各線分の傾きが同じである場合に線分 A と線分 B の距離を求め、その距離が所定の閾値以内であれば、それらの線分に関連する

40

50

テキスト列は平行に重なっていると判断するようにしてもよい。

【 0 0 4 6 】

なお、上記 2 個を選ぶ組み合わせは、コンビネーションの数式を用いると、符号 8 A に示すように ${}_{n_1}C_2$ と表せる。

【 0 0 4 7 】

ステップ S 8 0 3 では、テキスト列結合処理部 3 0 8 は、ステップ S 8 0 2 の処理結果を踏まえて、線分 A と線分 B が平行に重なるケースがあるのかを判断し、Y e s の場合にはステップ S 8 0 4 に進み、N o の場合にはステップ S 8 0 5 に進む。ステップ S 8 0 4 では、テキスト列結合処理部 3 0 8 は、線分 A と線分 B が平行に重なるバッファ同士を、符号 8 C に示すようにグループ化する。

10

【 0 0 4 8 】

図 9 は、ステップ S 8 0 4 で行う、線分 A と線分 B が平行に重なるバッファ同士をグループ化する処理の詳細である。本実施形態では、テキスト列の終端から後方に延びる線分 B と、テキスト列の先端から前方に延びる線分 A の重なりはテキスト列が連鎖している判断する。

ここで、ステップ S 9 0 1 では、テキスト列結合処理部 3 0 8 は、符号 9 A に示すように、重なり合った線分 B → 線分 A のバッファの I D がペアになったものを作成する。すなわち、テキスト列結合処理部 3 0 8 は、線分 B と線分 A とが重なったテキスト列のペア毎に、これら重なっているテキスト列に対する I D のペアを作成する。

上述したように、各テキスト列において線分 B は、テキスト列の後方に延びているので、ある 2 つのテキスト列に着目した場合、線分 B の後に該線分 B と重なる線分 A が配置されると、一方のテキスト列の後に他方のテキスト列が配置されるようになる。

20

【 0 0 4 9 】

なお、例えば、符号 9 A に示すように 4 つのテキスト列が一直線に連鎖するのであれば、バッファの I D のペアは 3 つ作成される。

符号 9 A において、テキスト列 9 1 の、線分 A を線分 i とし、線分 B を線分 j とする。また、テキスト列 9 2 の、線分 A を線分 k とし、線分 B を線分 l とする。また、テキスト列 9 3 の、線分 A を線分 m とし、線分 B を線分 n とする。さらに、テキスト列 9 4 の、線分 A を線分 o とし、線分 B を線分 p とする。また、テキスト列 9 1 の I D を「 7 」とし、テキスト列 9 2 の I D を「 2 」とし、テキスト列 9 3 の I D を「 9 」とし、テキスト列 9 4 の I D を「 3 」とする。

30

なお、本明細書において、「 X (X ; 自然数) 」は、テキスト列の識別用の I D を指す。

【 0 0 5 0 】

符号 9 A に示すように、テキスト列 9 1 の線分 B である線分 j とテキスト列 9 2 の線分 A である線分 k とが重なっているので、「 7 」と「 2 」とがペアになる。同様に、各テキスト列の線分 A、B の重なりにより、「 2 」と「 9 」とがペアになり、「 9 」と「 3 」とがペアになる。

【 0 0 5 1 】

ステップ S 9 0 2 では、テキスト列結合処理部 3 0 8 は、ステップ S 9 0 1 にて取得されたバッファの I D のペアの組み合わせから、I D を 1 次元に並ぶようにソートする。例えば、3 つの I D のペアを 1 次元にソートすると、符号 9 B に示すような順でソートされる。

40

【 0 0 5 2 】

ステップ S 9 0 3 では、テキスト列結合処理部 3 0 8 は、バッファの I D のペアを 1 次元にソートできないケースがあるか否かを判断する。例えば符号 9 C に示すように、テキスト列が重ねられて配置されるケース (符号 9 C では、テキスト列 9 3 とテキスト列 9 5 とが重なっている) があり、この場合、バッファの I D のペアをソートしても分岐ができてしまう。

なお、符号 9 C において、テキスト列 9 5 は、線分 A として線分 q を有し、線分 B とし

50

て線分 r を有する。また、テキスト列 95 の ID は、「10」である。

ステップ S903 の判断において、Yes の場合にはステップ S904 に進み、No の場合にはステップ S905 に進む。なお、ステップ S903 において、Yes となるケースは少なく、例外的な対応の処理としてステップ S904 を用意している。

【0053】

ステップ S904 では、テキスト列結合処理部 308 は、バッファの ID のペアをソートしても分岐ができてしまっているケースについて、バッファの ID の数字が小さい方を優先してソートし、バッファの ID を 1 次元にソートする処理を行う。

例えば、符号 9C に示すように、テキスト列 93 とテキスト列 95 とが重なっている場合、バッファの ID の並びは「7」「2」の後に「9」「3」と「10」「3」の系ができるので、1 次元にソートできていない。これを、符号 9D に示すように、「7」「2」の後の、「9」「3」と「10」「3」に対しバッファの ID の数字が小さい方の「9」を優先して、「7」「2」「9」「3」「10」「3」の順で一旦並べる。「3」は重複しているので、1 つになるように削除する。ここで、連結後のテキスト列の最後尾に線分 B が出やすいように、重複する場合は最後方の重複しているバッファの ID（本例の場合「10」の後ろの「3」）が残るように削除処理を行う。その結果、ID が「7」「2」「9」「10」「3」の 1 次元の並びとなる。

【0054】

なお、本実施形態では、分岐ができる場合にバッファの ID の数字が小さい方を優先してソートしているが、これに限定されず、バッファの ID の数字が大きい方を優先してソートするようにしても良い。

【0055】

ステップ S905 では、テキスト列結合処理部 308 は、符号 9E に示すように、グループ化の処理として、新たにバッファを用意し、ソートの結果を結合したものをそのバッファに収める。該新たなバッファの ID にも、以前の処理で使用されていない新しい ID（符号 9E では、「100」）を付与する。すなわち、テキスト列結合処理部 308 は、連鎖したテキスト列を 1 つのグループとし、該グループ毎にグループを識別するための ID（識別子）を付与する。

ステップ S906 では、テキスト列結合処理部 308 は、ステップ S905 で新たにグループ化されて不要になったバッファの削除を行う。

【0056】

ステップ S805 において、レイアウト解析処理部 307 は、テキスト列の連鎖検知として、電子文書処理装置に設定された条件として定義されている角度（条件とする角度）で交差（クロス）しているか否かを確認する。すなわち、ステップ S803 とステップ S804 の処理を通過した後の、残りのバッファの全数（ n_2 とする）から、2 個を選ぶ組み合わせに対し、線分 A と線分 B の関数が互いに交じり合い、かつ、条件として定義されている角度にてクロスしているか確認する。条件とする角度は符号 8E に示すように、例えば、線分 A と線分 B がクロスする角度が 90 度以内の場合に連鎖検知するように定義する。

なお、上記 2 個を選ぶ組み合わせは、コンビネーションの数式を用いると符号 8D に示すように ${}_n C_2$ と表せる。

【0057】

線分 A と線分 B のクロスする角度の条件は、テキスト列の連鎖を検知する上で、最も重要なパラメータのひとつとなる。本実施形態における電子文書処理装置には、テキスト列の連鎖検知のための線分 A と線分 B のクロスする角度の条件を操作者に入力させることができるユーザーインターフェースを提供する。なお図 10 は、パラメータを入力するためのユーザーインターフェースの好適な例を示している。

【0058】

図 10 に示すユーザーインターフェースがディスプレイ 303 に表示されると、操作者は、キーボード 302 やポインティングデバイスを用いて、連鎖検知用に設けられる線分

10

20

30

40

50

が交差される角度 θ を所望に応じて入力することができる。該入力が行われると、CPU 301 は、線分が交差される角度 θ に関するユーザ入力を受け、該ユーザ入力に応じて、線分が交差する角度の許容範囲を切り替える。

なお、本実施形態では、上記角度 θ をユーザ入力により変更可能な構成としても良いし、予め定められた角度を用いるようにしても良い。

【0059】

ステップ S806 では、レイアウト解析処理部 307 は、ステップ S805 の処理結果を踏まえて、線分 A と線分 B が定義された連鎖条件でクロスしているケースがあるのかを判断する。Yes の場合にはステップ S807 に進み、No の場合にはテキスト列結合処理を終了する。

10

【0060】

ステップ S807 では、テキスト列結合処理部 308 は、線分 A と線分 B が定義された連鎖条件でクロスしているバッファ同士を、符号 8F に示すようにグループ化する。すなわち、符号 8F において、テキスト列 83 における、線分 A を線分 e とし、線分 B を線分 f とし、テキスト列 84 における、線分 A を線分 g とし、線分 B を線分 h とする。このとき、テキスト列結合処理部 308 は、線分 A および線分 B の関数情報に基づいて、テキスト列 81 とテキスト列 82 の互いに種類が異なる線分である線分 f と線分 g とを比較して、線分 f と線分 g とがクロスするか否かを判断する。

【0061】

図 11 は、ステップ S807 で行う、線分 A と線分 B が定義された連鎖条件でクロスしているバッファ同士をグループ化する処理の詳細である。本実施形態では、テキスト列の終端から後方に延びる線分 B と、テキスト列の先端から前方に延びる線分 A の条件下でのクロスはテキスト列の連鎖であると判断する。

20

【0062】

ここで、ステップ S1101 では、テキスト列結合処理部 308 は、符号 11A に示すように、クロスした線分 B → 線分 A のバッファの ID がペアになったものを作成する。なお、例えば、符号 11A に示すように、4 つのテキスト列が条件下のクロスにより連鎖するのであれば、バッファの ID のペアは 3 つ作成される。

符号 11A において、テキスト列 111 の、線分 A を線分 i とし、線分 B を線分 j とする。また、テキスト列 112 の、線分 A を線分 k とし、線分 B を線分 l とする。また、テキスト列 113 の、線分 A を線分 m とし、線分 B を線分 n とする。さらに、テキスト列 114 の、線分 A を線分 o とし、線分 B を線分 p とする。また、テキスト列 111 の ID を「7」とし、テキスト列 112 の ID を「2」とし、テキスト列 113 の ID を「9」とし、テキスト列 114 の ID を「3」とする。

30

【0063】

ステップ S1102 では、テキスト列結合処理部 308 は、ステップ S1101 にて取得されたバッファの ID のペアの組み合わせから、ID を 1 次元に並ぶようにソートする。例えば、4 つの ID のペアを 1 次元にソートすると、符号 11B に示すような順でソートされる。

【0064】

ステップ S1103 では、テキスト列結合処理部 308 は、バッファの ID のペアが、連鎖のループにより 1 次元にソートできないケースがあるか否かを判断する。

40

例えば符号 11C に示すように、テキスト列がループして配置されるケースがある。この場合、バッファの ID のペアをソートしてもループになっていることにより、バッファ ID の始点と終点が定まらないことが確認できる。

【0065】

符号 11A において、テキスト列 115 の、線分 A を線分 C とし、線分 B を線分 D とする。また、テキスト列 116 の、線分 A を線分 E とし、線分 B を線分 F とする。また、テキスト列 117 の、線分 A を線分 G とし、線分 B を線分 H とする。また、テキスト列 118 の、線分 A を線分 I とし、線分 B を線分 J とする。また、テキスト列 119 の、線分 A

50

を線分Kとし、線分Bを線分Lとする。また、テキスト列120の、線分Aを線分Mとし、線分Bを線分Nとする。また、テキスト列121の、線分Aを線分Oとし、線分Bを線分Pとする。さらに、テキスト列122の、線分Aを線分Qとし、線分Bを線分Rとする。

【0066】

また、テキスト列115のIDを「1」とし、テキスト列116のIDを「9」とし、テキスト列117のIDを「3」とし、テキスト列118のIDを「6」とし、テキスト列119のIDを「7」とする。さらに、テキスト列120のIDを「4」とし、テキスト列121のIDを「5」とし、テキスト列122のIDを「8」とする。

【0067】

ステップS1103の判断において、Yesの場合にはステップS1104に進み、Noの場合にはステップS1105に進む。

ステップS1104では、テキスト列結合処理部308は、バッファのIDのペアをソートしてもループ状態となってしまうケースについて、線分B側に最も小さい数のバッファのIDを持つバッファをテキスト列連鎖の開始位置とし、1次元にソートする。例えば、符号11Cに示すようなテキスト列の並びの場合、線分B側に最も小さい数のバッファのIDを持つバッファは、テキスト列115(「1」)であるので、テキスト列連鎖の開始は、テキスト列115となる。従って、バッファIDの並びは符号11Dに示すように「1」「9」「3」「6」「7」「4」「5」「8」の結果となる。

【0068】

なお、テキスト列連鎖の開始位置となるテキスト列は、線分B側に最も大きい数のバッファIDを持つものとしても良い。また、線分B側に最も小さい数(または最も大きい数)から所定番目に大きい(小さい)数のバッファIDを持つものとしても良い。

【0069】

さらに、ステップS1105では、テキスト列結合処理部308は、バッファのIDのペアが分岐により1次元にソートできないケースがあるか判断する。

例えば符号11Eに示すように、テキスト列の連鎖において、分岐が発生するように配置されるケースがあり、この場合、バッファのIDのペアをソートしても分岐ができてしまう。ステップS1105の判断において、Yesの場合にはステップS1106に進み、Noの場合にはステップS1107に進む。なお、ステップS1105の判定において、Yesとなるケースは少なく、例外的な対応の処理としてステップS1106を用意している。

【0070】

符号11Eにおいて、テキスト列123の、線分Aを線分qとし、線分Bを線分rとする。また、テキスト列124の、線分Aを線分sとし、線分Bを線分tとする。さらに、テキスト列125の、線分Aを線分uとし、線分Bを線分vとする。また、テキスト列123のIDを「10」とし、テキスト列124のIDを「15」とし、テキスト列125のIDを「16」とする。

【0071】

ステップS1106では、テキスト列結合処理部308は、バッファのIDのペアをソートしても分岐ができてしまっているケースについて、1次元のソートする処理を行う。

例えば、符号11Eに示すようなテキスト列の並びの場合、バッファIDの並びは「7」「2」の後に「9」「3」「10」と「15」「16」「10」の系ができる。符号11Fに示すように、分岐の場合は、「7」「2」の後の、「9」「3」「10」と「15」「16」「10」に対しバッファIDの数字が小さい方の「9」を優先して、「7」「2」「9」「3」「10」「15」「16」「10」の順で一旦並べる。

【0072】

上記一旦並べられた配列において、「10」は重複しているので、1つになるように削除する。重複IDの削除の条件として、分岐の時点で「9」「3」「10」と「15」「16」「10」の系に対し、「9」「3」「10」の系を優先したので「9」「3」「1

10

20

30

40

50

0」の並びが維持されるように「10」を残す。よって、「15」「16」「10」の系に属する「10」を削除する条件にすると、「7」「2」「9」「3」「10」「15」「16」「10」の並びとなる。

【0073】

ステップS1106のソート方法にはテキスト列の分岐ポイントの交わる角度によって、合理的に判断する方法もあるので、後述する他の実施形態で示す。

【0074】

なお、本実施形態では、分岐ができる場合にバッファのIDの数字が小さい方を優先してソートしているが、これに限定されず、バッファのIDの数字が大きい方を優先してソートするようにしても良い。

10

【0075】

ステップS1107では、テキスト列結合処理部308は、符号9Eに示したものと同様に、グループ化の処理として、新たにバッファを用意し、ソートの結果を結合したものをそのバッファに収める。バッファのIDにも、以前の処理で使用されていない新しいIDを付与する。すなわち、テキスト列結合処理部308は、連鎖したテキスト列を1つのグループとし、該グループ毎にグループを識別するためのID（識別子）を付与する。

【0076】

ステップS1108では、テキスト列結合処理部308は、ステップS1107で新たにグループ化されて不要になったバッファの削除を行う。ステップS1108の後には線分Aと線分Bが定義された連鎖条件でクロスしているバッファ同士をグループ化する処理を終了し、さらにテキスト列結合処理も終了する。

20

【0077】

ステップS403では、ステップS402でグループ化したバッファ群の整列処理を行う。ステップS403で行う整列処理の詳細を図12に示す。

ステップS1201では、整列処理部309は、ステップS402にてグループ化できなかったテキスト列について、そのテキスト列単体で1グループとみなし、ページ内での整列処理をページ内のグループ間で行えるようにする。すなわち、本ステップにより、対象ページ内の全てのテキスト列が何らかのグループにグループ化されることになる。

ステップS1202では、整列処理部309は、符号12Aに示すように、グループ内で持つ先頭テキスト列のオフセット座標をそのグループの基準座標とする。

30

【0078】

ステップS1203では、整列処理部309は、グループの基準座標の縦方向（Y座標方向）の座標の数字が大きいほどソート順位が優先されるルールでグループ群をソートする。つまり、ページ上において、上方に位置するテキスト列を持つグループが、ソート後の1次元の並びとして前方にくるようにソートする。

【0079】

ステップS1204では、整列処理部309は、縦方向の座標が同じになるグループがあるか否かの判断を行う。ステップS1204の判断において、Yesの場合にはステップS1205に進み、Noの場合には整列処理を終了する。

【0080】

40

ステップS1205では、整列処理部309は、横方向（X座標方向）の座標が小さいほどソート順位が優先されるルールでグループ群をソートする。つまり、ページ上において、左方に位置するテキスト列を持つグループが、ソート後の1次元の並びとして前方にくるようにソートする。ステップS1205の処理の後には整列処理を終了する。

【0081】

このように整列処理部309は、グループ化されたテキスト列のオフセット座標を基準座標として検出し、該オフセット座標の座標値によって各テキスト列を整列させる。このような整列により、整列処理部309は、テキスト列の抽出準位を決定することができる。

【0082】

50

ステップS404では、CPU301は、ステップS403の整列処理によってソートされた順にグループのテキストを抽出する。このように、文書レイアウトに則したテキスト抽出を行うことで、ページ全体の文章性を損なうことなくテキストを取得することができる。

【0083】

このように、本実施形態では、電子文書処理装置は、テキスト列の前方を表すものとして機能する線分A（第1の線分）と、テキスト列の後方を表すものとして機能する線分B（第2の線分）とをベースラインに基づいて各テキスト列に設けている。そして、各線分を座標上で展開するための情報を線分の関数情報として保持し、該関数情報に基づいて、各テキスト列の線分A、Bを座標上にマッピングする。そして、該マッピングされた線分のうち、異なる種類の線分で重なった場合に、それら線分を有するテキスト列が連鎖していると判断し、グループ化している。よって、隣り合うテキスト列がある角度を持って配列されていても、それらテキスト列が連鎖していると判断することができる。すなわち、装飾されたテキスト列についても、高精度にテキスト列の連鎖の再現を行うことができる。

10

【0084】

よって、例えば、検索エンジン等に本実施形態の電子文書処理装置を適用すると、概念検索などページ全体の文章性が必要とされるような高度な検索において、その精度を向上させることができる。

【0085】

20

（第2の実施形態）

上述した実施形態のステップS1106では、テキスト列の分岐ポイントにおいて、IDの数字が小さい方のテキスト列を優先してソートされるルールとしていた。しかし、ステップS1106において、よりテキストレイアウトに則してテキスト列の連鎖の優先順位を付ける方法がある。

【0086】

図13は、本実施形態に係る、テキスト列が連鎖し、かつ、分岐が出来ている状態を示している。

図13において、テキスト列131の、線分Aを線分aとし、線分Bを線分bとする。また、テキスト列132の、線分Aを線分cとし、線分Bを線分dとする。さらに、テキスト列133の、線分Aを線分eとし、線分Bを線分fとする。また、テキスト列1131のIDを「5」とし、テキスト列132のIDを「8」とし、テキスト列133のIDを「7」とする。

30

【0087】

図13では、IDが「5」のテキスト列131に対し、IDが「8」のテキスト列132とIDが「5」のテキスト列133とが連鎖し、かつ、分岐状態となっている。テキスト列131の線分Bである線分bの関数に対し、テキスト列132の線分Aである線分cの関数は $\theta 1$ の角度を成している。また、テキスト列131の線分bの関数に対し、テキスト列133の線分Aである線分eの関数は $\theta 2$ の角度を成している。

なお、互いに交わる2つの1次関数が成す角度 θ は、平面三角法により容易に算出できることは言うまでも無い。

40

【0088】

ここで、 $\theta 1 < \theta 2$ の場合に、角度が小さい方の $\theta 1$ を形成する、「5」のテキスト列131から「8」のテキスト列132をつなぐ方が連鎖として自然である。よって、テキスト列の連鎖時に分岐が出来る場合には、お互いの交わる角度が小さい方を優先してテキスト列をソートするのが良い。すなわち、テキスト列結合処理部308は、1つのテキスト列の後に複数のテキスト列の連鎖が認められる場合には、前方のテキスト列と後方のテキスト列との角度に応じて抽出順位を決定することができる。すなわち、テキスト列の連鎖が認められる中で、後段にある複数のテキスト列において、前段のテキスト列と後段のテキスト列との成す角度が小さい方のテキスト列を優先して、テキスト列を連鎖させるよう

50

にしても良い。

【 0 0 8 9 】

ただ、テキストレイアウトに則した厳密なソート処理はシステムにとって高負荷であり、パフォーマンスに影響を与えかねない。よって、この方法による処理を行うかどうかは選択できたほうが良い。本実施形態における電子文書処理装置には、テキストレイアウトに則した厳密な処理を行うか、またはパフォーマンスを優先して処理を行うかについて、操作者に選択させることができるユーザーインターフェースを提供する。なお図 1 4 は、ユーザに処理レベルを選択させるためのユーザーインターフェースの好適な例を示している。

【 0 0 9 0 】

10

(その他の実施形態)

本発明は、複数の機器(例えばコンピュータ、インターフェース機器、リーダー、プリンタなど)から構成されるシステムに適用することも、1つの機器からなる装置(複合機、プリンタ、ファクシミリ装置など)に適用することも可能である。

【 0 0 9 1 】

前述した実施形態の機能を実現するように前述した実施形態の構成を動作させるプログラムを記憶媒体に記憶させ、該記憶媒体に記憶されたプログラムをコードとして読み出し、コンピュータにおいて実行する処理方法も上述の実施形態の範疇に含まれる。即ちコンピュータ読み取り可能な記憶媒体も実施例の範囲に含まれる。また、前述のコンピュータプログラムが記憶された記憶媒体はもちろんそのコンピュータプログラム自体も上述の実

20

【 0 0 9 2 】

かかる記憶媒体としてはたとえばフロッピー(登録商標)ディスク、ハードディスク、光ディスク、光磁気ディスク、CD-ROM、磁気テープ、不揮発性メモリカード、ROMを用いることができる。

【 0 0 9 3 】

また前述の記憶媒体に記憶されたプログラム単体で処理を実行しているものに限らず、他のソフトウェア、拡張ボードの機能と共同して、OS上で動作し前述の実施形態の動作を実行するものも前述した実施形態の範疇に含まれる。

【図面の簡単な説明】

30

【 0 0 9 4 】

【図 1】従来の、PDFファイルのプレビュー例と、そのPDFファイル内のテキスト描画コマンドの配列の例を示す図である。

【図 2】図 1 で示したPDFファイルを例に、テキスト描画コマンドのオフセット座標の順にソートしてもうまくテキストが取得できない例を示す図である。

【図 3】本発明の一実施形態の、電子文書処理装置を実現するハードウェア構成例を示すブロック図である。

【図 4】本発明の一実施形態の、文書処理の手順例を示すフローチャートである。

【図 5】本発明の一実施形態の、レイアウト解析処理を示すフローチャートと、その補足図である。

40

【図 6】本発明の一実施形態の、書体の構成要素となるベースラインの例を示す図である。

【図 7】本発明の一実施形態の、条件入力のためのユーザーインターフェースの例を示す図である。

【図 8】本発明の一実施形態の、テキスト列結合処理を表すフローチャートと、その補足図である。

【図 9】本発明の一実施形態の、テキスト列のグループ化処理を示すフローチャートである。

【図 1 0】本発明の一実施形態の、条件入力のためのユーザーインターフェースの例を示す図である。

50

【図 1 1】本発明の一実施形態の、テキスト列のグループ化処理を示すフローチャートである。

【図 1 2】本発明の一実施形態の、グループ化されたテキスト列の整列処理を示すフローチャートである。

【図 1 3】本発明の一実施形態の、テキスト列が連鎖し、かつ、分岐が出来ている状態の例を示す図である。

【図 1 4】本発明の一実施形態の、条件入力のためのユーザーインターフェースの例を示す図である。

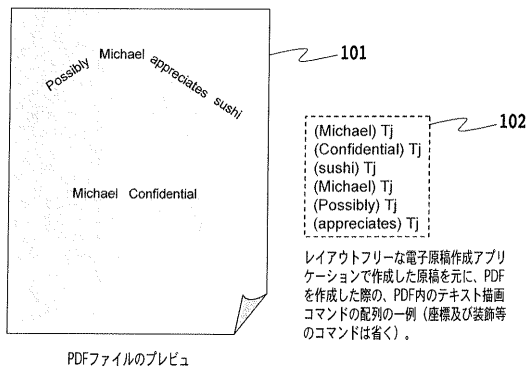
【符号の説明】

【 0 0 9 5 】

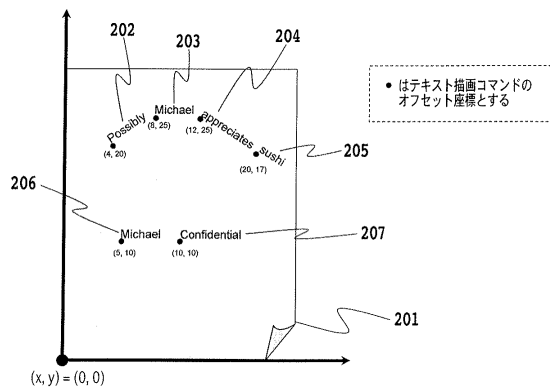
- 3 0 1 C P U
- 3 0 2 キーボード
- 3 0 3 ディスプレイ
- 3 0 4 ハードディスク
- 3 0 5 R O M
- 3 0 6 R A M
- 3 0 7 レイアウト解析処理部
- 3 0 8 テキスト列結合処理部
- 3 0 9 整列処理部

10

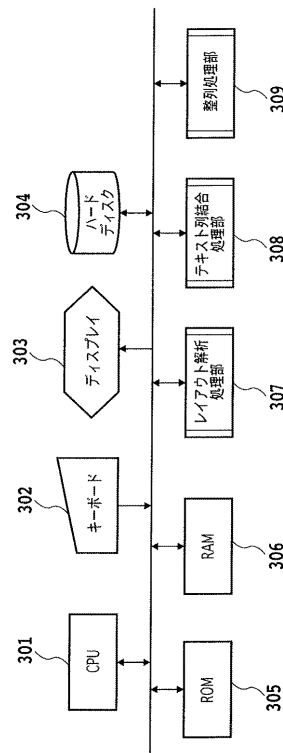
【図 1】



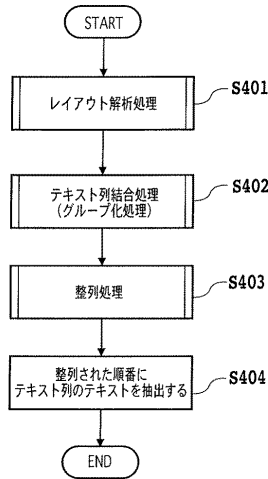
【図 2】



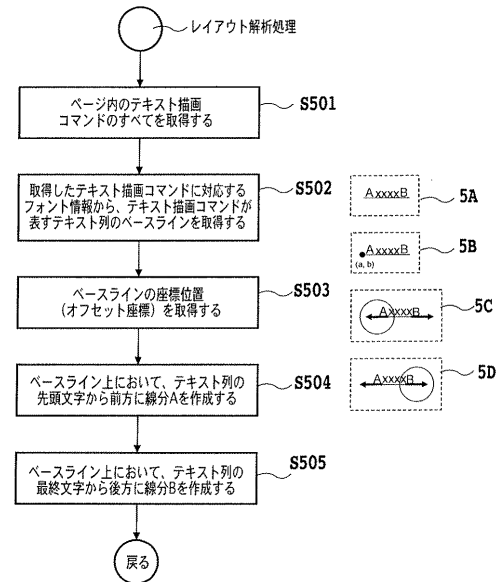
【図 3】



【図 4】



【図 5】



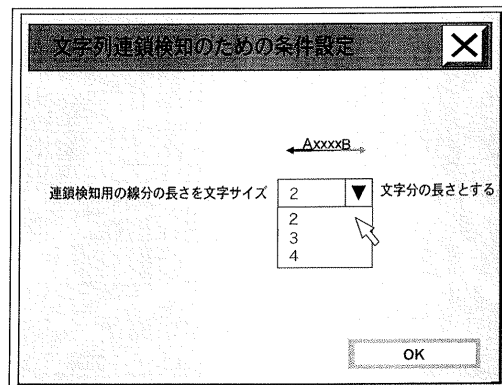
【図 6】

good 良い 601

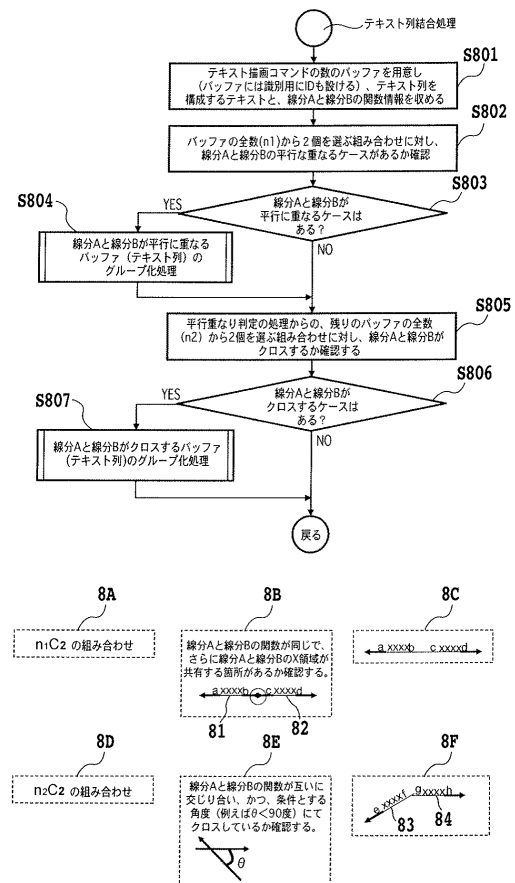
ベースライン

書体の構成要素であり、単にラインとも呼ばれる。
テキスト列の並びの視線を誘導する上で重要な要素のひとつである。

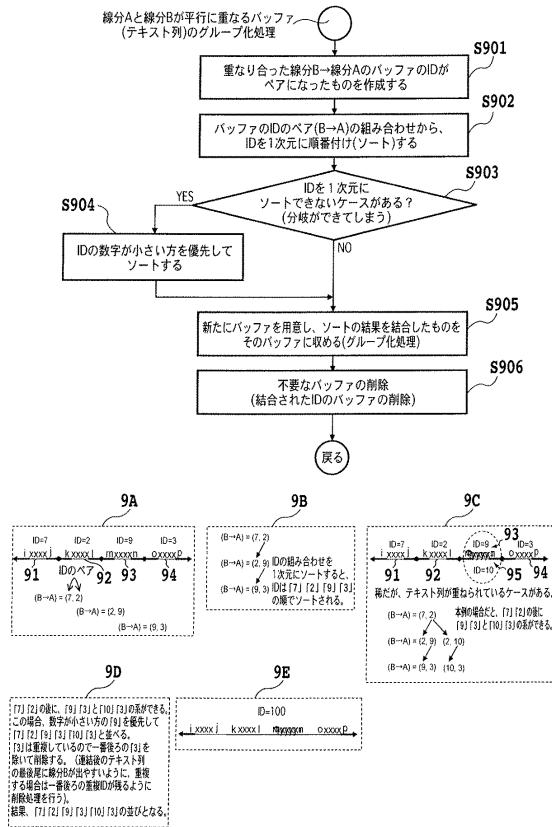
【図 7】



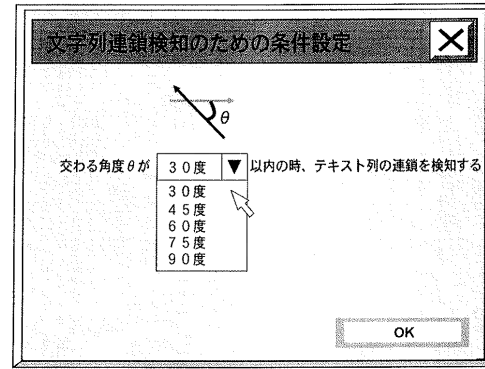
【図 8】



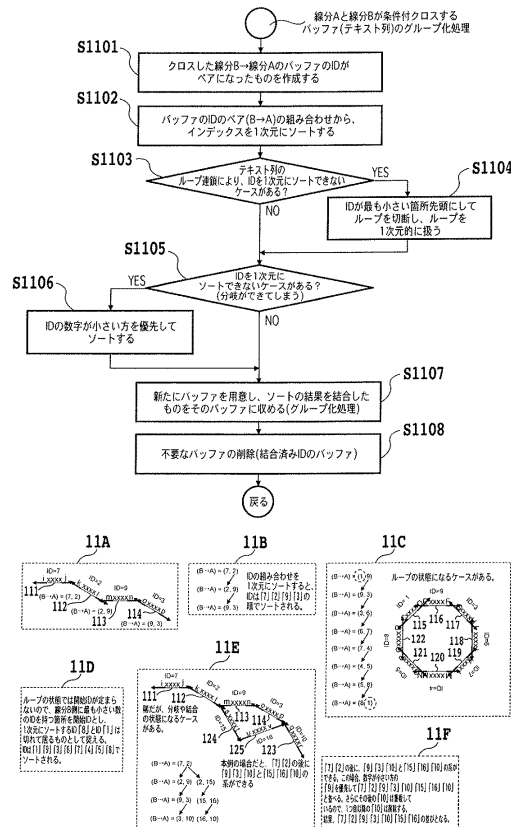
【図 9】



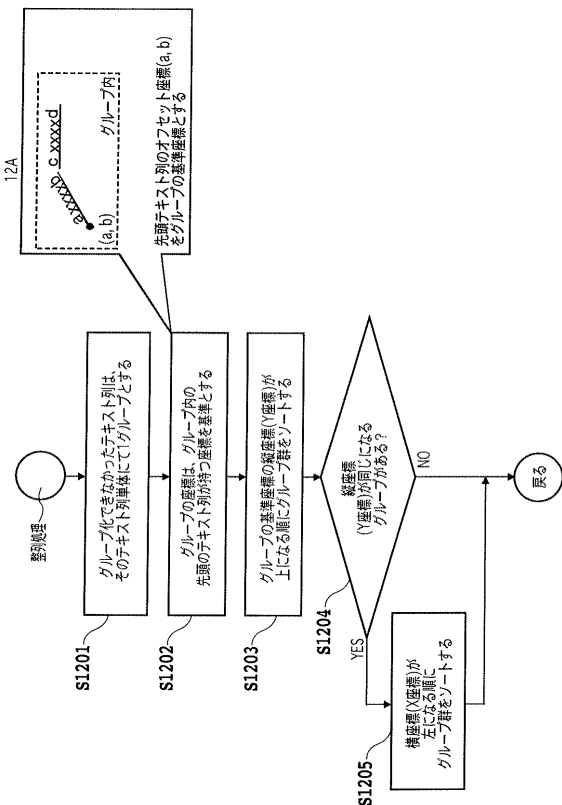
【図 10】



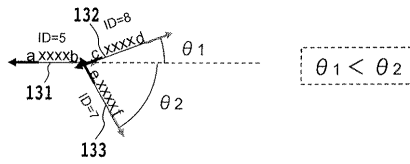
【図 11】



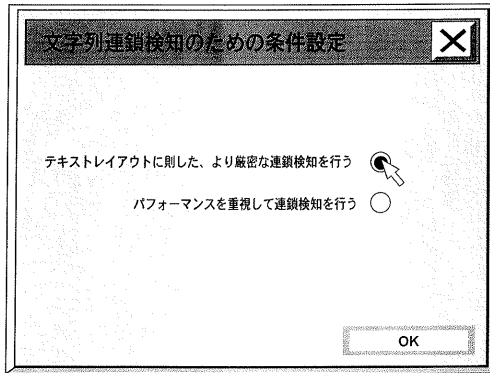
【図 12】



【図 13】



【図 14】



フロントページの続き

(56)参考文献 特開平 1 1 - 2 0 3 4 9 4 (J P , A)
特開 2 0 0 3 - 2 8 8 6 0 7 (J P , A)
特開 2 0 0 7 - 0 8 6 9 5 5 (J P , A)
特開平 0 5 - 0 1 2 3 9 8 (J P , A)

(58)調査した分野(Int.Cl. , D B 名)
G 0 6 F 1 7 / 2 0 - 1 7 / 2 6
G 0 6 T 1 1 / 6 0