【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】平成23年10月6日(2011.10.6)

【公開番号】特開2010-55512(P2010-55512A)
【公開日】平成22年3月11日(2010.3.11)
【年通号数】公開・登録公報2010-010
【出願番号】特願2008-221912(P2008-221912)
【国際特許分類】
Ｇ０６Ｆ　17/21　　(2006.01)
Ｇ０６Ｔ　11/60　　(2006.01)
【ＦＩ】
Ｇ０６Ｆ　17/21　　５３０Ａ
Ｇ０６Ｔ　11/60　　１００Ａ

【手続補正書】
【提出日】平成23年8月24日(2011.8.24)
【手続補正１】
【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。
【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】
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バロ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】
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、
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、
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、
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。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】００２１
【補正方法】変更
【補正の内容】
【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。
【手続補正３】
【補正対象書類名】明細書
【補正対象項目名】００２２
【補正方法】変更
【補正の内容】
【００２２】

また、本発明は、電子文書処理方法であって、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、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、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、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。