

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2021-179747

(P2021-179747A)

(43) 公開日 令和3年11月18日(2021.11.18)

(51) Int.Cl.	F I	テーマコード (参考)
G 1 6 Z 99/00 (2019.01)	G 1 6 Z 99/00	5 B 0 2 9
G 0 6 K 9/20 (2006.01)	G 0 6 K 9/20 3 4 0 C	5 B 0 6 4
G 0 6 K 9/66 (2006.01)	G 0 6 K 9/66	5 L 0 4 9
G 0 6 T 7/00 (2017.01)	G 0 6 T 7/00 3 5 0 C	5 L 0 9 6

審査請求 未請求 請求項の数 6 O L (全 20 頁)

(21) 出願番号	特願2020-84142 (P2020-84142)	(71) 出願人	000006150
(22) 出願日	令和2年5月12日 (2020.5.12)		京セラドキュメントソリューションズ株式会社
		(74) 代理人	100140796
			弁理士 原口 貴志
		(72) 発明者	樫本 陽介
			大阪府大阪市中央区玉造 1 丁目 2 番 2 8 号
			京セラドキュメントソリューションズ株式会社内
		F ターム (参考)	5B029 AA01 BB02 CC26
			5B064 AA01 AB02 BA01 DA27 DA32
			EA08 EA21 EA27
			5L049 EE01
			5L096 BA17 DA02 HA11 KA04

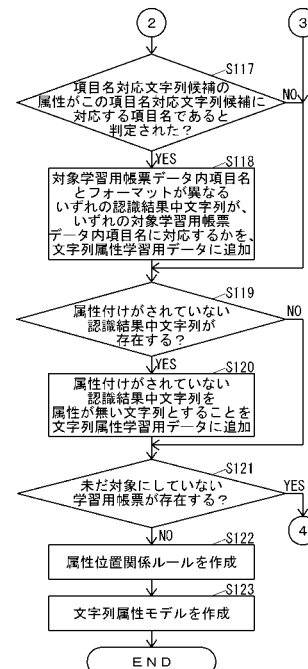
(54) 【発明の名称】 帳票データ取得システムおよび帳票データ取得プログラム

(57) 【要約】

【課題】 帳票から自動で取得した帳票データの適切性を向上することができる帳票データ取得システムおよび帳票データ取得プログラムを提供する。

【解決手段】 情報処理装置は、学習用帳票における属性の位置関係を学習することによって、帳票における文字列の属性の位置関係のルールを示す属性位置関係ルールを作成し (S 1 2 2)、学習用帳票の画像に対する文字認識の結果における文字列と、学習用帳票に基づいた学習用帳票データにおける属性との対応付けを実行することによって、帳票における文字列の属性の確率を取得するための文字列属性モデルを作成し (S 1 2 3)、帳票の画像に対する文字認識の結果における文字列に対して、文字列属性モデルを適用して、属性の確率を取得し、帳票の画像に対する文字認識の結果における文字列の、帳票における位置と、属性位置関係ルールとに基づいて、確率を補正することを特徴とする。

【選択図】 図 9



【特許請求の範囲】**【請求項 1】**

帳票に基づいたデータとしての帳票データを、この帳票から自動で取得する帳票データ取得システムであって、

帳票における文字列の属性の確率を取得するための文字列属性モデルを作成する文字列属性学習部と、

帳票における文字列の属性の位置関係のルールを示す属性位置関係ルールを作成する属性位置関係学習部と、

帳票の画像に対する文字認識の結果における文字列に対して、前記文字列属性モデルを適用して、属性の確率を取得する属性確率取得部と、

帳票の画像に対する文字認識の結果における文字列の、帳票における位置と、前記属性位置関係ルールとに基づいて、前記確率を補正する属性確率補正部と

を備え、

前記文字列属性学習部は、

帳票から事前に作成された前記帳票データとしての学習用帳票データの作成の基になった帳票としての学習用帳票の画像に対する文字認識の結果における文字列と、

前記学習用帳票データにおける属性と

の対応付けを実行することによって、前記文字列属性モデルを作成し、

前記属性位置関係学習部は、前記学習用帳票における属性の位置関係を学習することによって、前記属性位置関係ルールを作成することを特徴とする帳票データ取得システム。

【請求項 2】

前記文字列属性学習部は、前記対応付けの結果において、前記学習用帳票データにおける属性のうちの、項目の値とフォーマットが異なる文字列が存在する場合に、この文字列がいずれの値に対応するかということにも基づいて、前記文字列属性モデルを作成することを特徴とする請求項 1 に記載の帳票データ取得システム。

【請求項 3】

前記対応付けの結果において、前記学習用帳票データにおける属性のうちの、項目の値とフォーマットが異なる文字列が存在する場合に、前記帳票データにおける値のフォーマットのルールを取得するフォーマット取得部と、

前記フォーマット取得部によって取得されたルールに基づいて前記帳票データにおける値のフォーマットを変更するフォーマット変更部と

を備えることを特徴とする請求項 2 に記載の帳票データ取得システム。

【請求項 4】

前記文字列属性学習部は、前記学習用帳票データにおける属性のうちの、項目の値に対応する文字列が前記学習用帳票の画像に対する文字認識の結果に存在するが、この項目の項目名に対応する文字列がこの文字認識の結果に存在しない項目としての項目名対応文字列無項目が存在する場合に、前記項目名対応文字列無項目の前記学習用帳票データにおける値に、前記学習用帳票の画像に対する文字認識の結果において対応する文字列との、前記学習用帳票の画像における位置関係に基づいて、前記項目名対応文字列無項目の項目名に対応する文字列の候補を探索し、探索によって抽出された前記候補と、この候補に対応する項目名との類似度に基づいて、この候補の属性が、この項目名であると判定したとき、この候補がいずれの項目名に対応するかということにも基づいて、前記文字列属性モデルを作成することを特徴とする請求項 1 から請求項 3 までのいずれかに記載の帳票データ取得システム。

【請求項 5】

帳票に存在しない値を算出するための算出式を導き出す算出式導出部と、

帳票に存在しない値を前記算出式に基づいて算出して前記帳票データに追加する値算出部と

を備え、

前記算出式導出部は、前記学習用帳票の画像に対する文字認識の結果に存在する文字列

10

20

30

40

50

に対応するものが存在しない、前記学習用帳票データにおける属性のうちの、項目の値について、前記学習用帳票データにおける値のそれぞれの間の関係性を分析して、前記学習用帳票データにおける他の値からの前記算出式を導き出すことを特徴とする請求項 1 から請求項 4 までのいずれかに記載の帳票データ取得システム。

【請求項 6】

帳票に基づいたデータとしての帳票データを、この帳票から自動で取得するための帳票データ取得プログラムであって、

帳票における文字列の属性の確率を取得するための文字列属性モデルを作成する文字列属性学習部と、

帳票における文字列の属性の位置関係のルールを示す属性位置関係ルールを作成する属性位置関係学習部と、

帳票の画像に対する文字認識の結果における文字列に対して、前記文字列属性モデルを適用して、属性の確率を取得する属性確率取得部と、

帳票の画像に対する文字認識の結果における文字列の、帳票における位置と、前記属性位置関係ルールとに基づいて、前記確率を補正する属性確率補正部と

をコンピュータに実現させ、

前記文字列属性学習部は、

帳票から事前に作成された前記帳票データとしての学習用帳票データの作成の基になった帳票としての学習用帳票の画像に対する文字認識の結果における文字列と、

前記学習用帳票データにおける属性と

の対応付けを実行することによって、前記文字列属性モデルを作成し、

前記属性位置関係学習部は、前記学習用帳票における属性の位置関係を学習することによって、前記属性位置関係ルールを作成することを特徴とする帳票データ取得プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、帳票に基づいたデータとしての帳票データを、この帳票から自動で取得する帳票データ取得システムおよび帳票データ取得プログラムに関する。

【背景技術】

【0002】

従来の帳票データ取得システムとして、単語と単語との位置関係と、各単語の属性とに基づいて単語と単語との関連性を決定するためのルールである関連性決定ルールと、非定型の帳票に含まれる各単語の、辞書を用いて特定した属性と、帳票における各単語の位置とに基づいて、各単語の属性を特定するものが知られている（例えば、特許文献 1 参照。）。

【先行技術文献】

【特許文献】

【0003】

【特許文献 1】特開 2019 - 079488 号公報

【発明の概要】

【発明が解決しようとする課題】

【0004】

しかしながら、従来の帳票データ取得システムにおいては、帳票データを適切に取得することができる帳票が関連性決定ルールおよび辞書によって定まっているので、利用者が希望する帳票に対して帳票データを適切に取得することができない場合があるという問題がある。

【0005】

そこで、本発明は、帳票から自動で取得した帳票データの適切性を向上することができる帳票データ取得システムおよび帳票データ取得プログラムを提供することを目的とする

。

【課題を解決するための手段】

【0006】

本発明の帳票データ取得システムは、帳票に基づいたデータとしての帳票データを、この帳票から自動で取得する帳票データ取得システムであって、帳票における文字列の属性の確率を取得するための文字列属性モデルを作成する文字列属性学習部と、帳票における文字列の属性の位置関係のルールを示す属性位置関係ルールを作成する属性位置関係学習部と、帳票の画像に対する文字認識の結果における文字列に対して、前記文字列属性モデルを適用して、属性の確率を取得する属性確率取得部と、帳票の画像に対する文字認識の結果における文字列の、帳票における位置と、前記属性位置関係ルールとに基づいて、前記確率を補正する属性確率補正部とを備え、前記文字列属性学習部は、帳票から事前に作成された前記帳票データとしての学習用帳票データの作成の基になった帳票としての学習用帳票の画像に対する文字認識の結果における文字列と、前記学習用帳票データにおける属性との対応付けを実行することによって、前記文字列属性モデルを作成し、前記属性位置関係学習部は、前記学習用帳票における属性の位置関係を学習することによって、前記属性位置関係ルールを作成することを特徴とする。

10

【0007】

この構成により、本発明の帳票データ取得システムは、帳票における文字列の属性の確率を取得するための文字列属性モデルと、帳票における文字列の属性の位置関係のルールを示す属性位置関係ルールとを、帳票から事前に作成された帳票データとしての学習用帳票データと、学習用帳票データの作成の基になった帳票としての学習用帳票の画像に対する文字認識の結果における文字列とを用いて学習するので、学習用帳票に適した文字列属性モデルおよび属性位置関係ルールを得ることができ、その結果、帳票から自動で取得した帳票データの適切性を向上することができる。

20

【0008】

本発明の帳票データ取得システムにおいて、前記文字列属性学習部は、前記対応付けの結果において、前記学習用帳票データにおける属性のうちの、項目の値とフォーマットが異なる文字列が存在する場合に、この文字列がいずれの値に対応するかということにも基づいて、前記文字列属性モデルを作成しても良い。

30

【0009】

この構成により、本発明の帳票データ取得システムは、学習用帳票データにおける項目の値とフォーマットが異なるが、この値に対応する文字列を、項目の値として取得することができるように文字列属性モデルを作成するので、帳票から自動で取得した帳票データの適切性を更に向上することができる。

【0010】

本発明の帳票データ取得システムは、前記対応付けの結果において、前記学習用帳票データにおける属性のうちの、項目の値とフォーマットが異なる文字列が存在する場合に、前記帳票データにおける値のフォーマットのルールを取得するフォーマット取得部と、前記フォーマット取得部によって取得されたルールに基づいて前記帳票データにおける値のフォーマットを変更するフォーマット変更部とを備えても良い。

40

【0011】

この構成により、本発明の帳票データ取得システムは、帳票データにおける項目の値とフォーマットが異なるが、この値に対応する文字列のフォーマットを、帳票データにおける項目の値のフォーマットに変更するので、帳票から自動で取得した帳票データの適切性を更に向上することができる。

【0012】

本発明の帳票データ取得システムにおいて、前記文字列属性学習部は、前記学習用帳票データにおける属性のうちの、項目の値に対応する文字列が前記学習用帳票の画像に対する文字認識の結果に存在するが、この項目の項目名に対応する文字列がこの文字認識の結果に存在しない項目としての項目名対応文字列無項目が存在する場合に、前記項目名対応

50

文字列無項目の前記学習用帳票データにおける値に、前記学習用帳票の画像に対する文字認識の結果において対応する文字列との、前記学習用帳票の画像における位置関係に基づいて、前記項目名対応文字列無項目の項目名に対応する文字列の候補を探索し、探索によって抽出された前記候補と、この候補に対応する項目名との類似度に基づいて、この候補の属性が、この項目名であると判定したとき、この候補がいずれの項目名に対応するかということにも基づいて、前記文字列属性モデルを作成しても良い。

【0013】

この構成により、本発明の帳票データ取得システムは、学習用帳票データにおける項目の値に対応する文字列が学習用帳票の画像に対する文字認識の結果に存在するが、この項目の項目名に対応する文字列がこの文字認識の結果に存在しない項目としての項目名対応文字列無項目が存在する場合に、項目名対応文字列無項目の学習用帳票データにおける値に、学習用帳票の画像に対する文字認識の結果において対応する文字列との、学習用帳票の画像における位置関係に基づいて、項目名対応文字列無項目の項目名に対応する文字列の候補を探索し、探索によって抽出された候補と、この候補に対応する項目名との類似度に基づいて、この候補の属性が、この項目名であると判定したとき、この候補がいずれの項目名に対応するかということにも基づいて、文字列属性モデルを作成するので、帳票から自動で取得した帳票データの適切性を更に向上することができる。

10

【0014】

本発明の帳票データ取得システムは、帳票に存在しない値を算出するための算出式を導き出す算出式導出部と、帳票に存在しない値を前記算出式に基づいて算出して前記帳票データに追加する値算出部とを備え、前記算出式導出部は、前記学習用帳票の画像に対する文字認識の結果に存在する文字列に対応するものが存在しない、前記学習用帳票データにおける属性のうちの、項目の値について、前記学習用帳票データにおける値のそれぞれの間の関係性を分析して、前記学習用帳票データにおける他の値からの前記算出式を導き出しても良い。

20

【0015】

この構成により、本発明の帳票データ取得システムは、学習用帳票の画像に対する文字認識の結果に存在する文字列に対応するものが存在しない、学習用帳票データにおける値について、学習用帳票データにおける値のそれぞれの間の関係性を分析して、学習用帳票データにおける他の値からの算出式を導き出し、導き出した算出式に基づいて、帳票に存在しない値を帳票データに追加するので、帳票から自動で取得した帳票データの適切性を更に向上することができる。

30

【0016】

本発明の帳票データ取得プログラムは、帳票に基づいたデータとしての帳票データを、この帳票から自動で取得するための帳票データ取得プログラムであって、帳票における文字列の属性の確率を取得するための文字列属性モデルを作成する文字列属性学習部と、帳票における文字列の属性の位置関係のルールを示す属性位置関係ルールを作成する属性位置関係学習部と、帳票の画像に対する文字認識の結果における文字列に対して、前記文字列属性モデルを適用して、属性の確率を取得する属性確率取得部と、帳票の画像に対する文字認識の結果における文字列の、帳票における位置と、前記属性位置関係ルールとに基づいて、前記確率を補正する属性確率補正部とをコンピューターに実現させ、前記文字列属性学習部は、帳票から事前に作成された前記帳票データとしての学習用帳票データの作成の基になった帳票としての学習用帳票の画像に対する文字認識の結果における文字列と、前記学習用帳票データにおける属性との対応付けを実行することによって、前記文字列属性モデルを作成し、前記属性位置関係学習部は、前記学習用帳票における属性の位置関係を学習することによって、前記属性位置関係ルールを作成することを特徴とする。

40

【0017】

この構成により、本発明の帳票データ取得プログラムを実行するコンピューターは、帳票における文字列の属性の確率を取得するための文字列属性モデルと、帳票における文字列の属性の位置関係のルールを示す属性位置関係ルールとを、帳票から事前に作成された

50

帳票データとしての学習用帳票データと、学習用帳票データの作成の基になった帳票としての学習用帳票の画像に対する文字認識の結果における文字列とを用いて学習するので、学習用帳票に適した文字列属性モデルおよび属性位置関係ルールを得ることができ、その結果、帳票から自動で取得した帳票データの適切性を向上することができる。

【発明の効果】

【0018】

本発明の帳票データ取得システムおよび帳票データ取得プログラムは、帳票から自動で取得した帳票データの適切性を向上することができる。

【図面の簡単な説明】

【0019】

10

【図1】本発明の一実施の形態に係る情報処理システムのブロック図である。

【図2】図1に示す情報処理システムによって扱われる帳票の一例を示す図である。

【図3】図1に示す情報処理システムによって扱われる帳票の、図2に示す例とは異なる一例を示す図である。

【図4】MFPによって構成される場合の図1に示す画像形成装置のブロック図である。

【図5】図1に示す情報処理装置のブロック図である。

【図6】(a)図5に示す学習用データの伝票情報テーブルの一例を示す図である。(b)図5に示す学習用データの品目テーブルの一例を示す図である。

【図7】文字列属性モデルおよび属性位置関係ルールを作成する場合の図5に示す情報処理装置の動作の一部のフローチャートである。

20

【図8】図7に示すフローチャートの続きのフローチャートである。

【図9】図8に示すフローチャートの続きのフローチャートである。

【図10】帳票から帳票データを自動で作成する場合の図5に示す情報処理装置の動作のフローチャートである。

【図11】文字列属性モデルを適用して判定された、図2に示す帳票に含まれる文字列の属性の一例を示す図である。

【図12】属性位置関係ルールを適用して補正された、図2に示す帳票に含まれる文字列の属性の一例を示す図である。

【図13】(a)図12に示す属性に基づいて作成された帳票データの伝票情報テーブルの一例を示す図である。(b)図12に示す属性に基づいて作成された帳票データの品目テーブルの一例を示す図である。

30

【図14】(a)図2に示す帳票から自動で作成された帳票データの伝票情報テーブルの一例を示す図である。(b)図2に示す帳票から自動で作成された帳票データの品目テーブルの一例を示す図である。

【図15】図1に示す情報処理システムによって扱われる帳票の、図2および図3に示す例とは異なる一例を示す図である。

【発明を実施するための形態】

【0020】

以下、本発明の実施の形態について、図面を用いて説明する。

【0021】

40

まず、本発明の一実施の形態に係る情報処理システムの構成について説明する。

【0022】

図1は、本実施の形態に係る情報処理システム10のブロック図である。

【0023】

図1に示すように、情報処理システム10は、例えばMFP(Multifunction Peripheral)などの画像形成装置20と、例えばPC(Personal Computer)などのコンピューターとしての情報処理装置30とを備えている。

【0024】

画像形成装置20と、情報処理装置30とは、LAN(Local Area Net

50

work)などのネットワーク11を介して通信可能に接続可能である。

【0025】

情報処理装置30は、帳票に基づいたデータとしての帳票データを、この帳票から自動で取得する帳票データ取得システムである。

【0026】

図2および図3は、それぞれ、情報処理システム10によって扱われる帳票の一例を示す図である。

【0027】

図2および図3のそれぞれに示す帳票は、請求書である。情報処理システム10によって扱われる請求書は、例えば図2および図3に示すように、例えば請求元によってフォーマットが異なる。

10

【0028】

図4は、MFPによって構成される場合の画像形成装置20のブロック図である。

【0029】

図4に示すように、画像形成装置20は、種々の操作が入力される例えばボタンなどの入力デバイスである操作部21と、種々の情報を表示する例えばLCD(Liquid Crystal Display)などの表示デバイスである表示部22と、用紙などの記録媒体に画像を印刷する印刷デバイスであるプリンター23と、原稿から画像を読み取る読取デバイスであるスキャナー24と、図示していない外部のファクシミリ装置と公衆電話回線などの通信回線経由でファックス通信を行うファックスデバイスであるファックス通信部25と、LAN、インターネットなどのネットワーク経由で、または、ネットワークを介さずに有線または無線によって直接に、外部の装置と通信を行う通信デバイスである通信部26と、各種の情報を記憶する例えば半導体メモリー、HDD(Hard Disk Drive)などの不揮発性の記憶デバイスである記憶部27と、画像形成装置20全体を制御する制御部28とを備えている。

20

【0030】

制御部28は、例えば、CPU(Central Processing Unit)と、プログラムおよび各種のデータを記憶しているROM(Read Only Memory)と、制御部28のCPUの作業領域として用いられる揮発性の記憶デバイスとしてのメモリーであるRAM(Random Access Memory)とを備えている。制御部28のCPUは、記憶部27または制御部28のROMに記憶されているプログラムを実行する。

30

【0031】

図5は、情報処理装置30のブロック図である。

【0032】

図5に示すように、情報処理装置30は、種々の操作が入力される例えばキーボード、マウスなどの操作デバイスである操作部31と、種々の情報を表示する例えばLCDなどの表示デバイスである表示部32と、LAN、インターネットなどのネットワーク経由で、または、ネットワークを介さずに有線または無線によって直接に、外部の装置と通信を行う通信デバイスである通信部33と、各種の情報を記憶する例えば半導体メモリー、HDDなどの不揮発性の記憶デバイスである記憶部34と、情報処理装置30全体を制御する制御部35とを備えている。

40

【0033】

記憶部34は、帳票に基づいた帳票データを、この帳票から自動で取得するための帳票データ取得プログラム34aを記憶している。帳票データ取得プログラム34aは、例えば、情報処理装置30の製造段階で情報処理装置30にインストールされていても良いし、CD(Compact Disk)、DVD(Digital Versatile Disk)、USB(Universal Serial Bus)メモリーなどの外部の記憶媒体から情報処理装置30に追加でインストールされても良いし、ネットワーク上から情報処理装置30に追加でインストールされても良い。

50

【 0 0 3 4 】

記憶部 3 4 は、情報処理システム 1 0 によって扱われる帳票から例えば人間が認識して抜き出すなどして、情報処理システム 1 0 によって扱われる帳票から事前に作成された帳票データ（以下「学習用帳票データ」という。）を含む学習用データ 3 4 b を記憶可能である。学習用データ 3 4 b は、複数の帳票の学習用帳票データを含むことが可能である。

【 0 0 3 5 】

図 6 は、学習用データ 3 4 b の一例を示す図である。

【 0 0 3 6 】

図 6 に示す学習用データ 3 4 b は、請求書から抜き出されたデータであり、請求書における基本的な情報を示す図 6 (a) に示す伝票情報テーブルと、請求書における品目の詳細な情報を示す図 6 (b) に示す品目テーブルとを備えている。

10

【 0 0 3 7 】

図 6 に示す学習用データ 3 4 b において具体的な値が示されているデータは、図 2 に示す請求書と、図 3 に示す請求書とから抜き出されたデータである。図 6 に示す学習用データ 3 4 b において項目「請求書番号」の値が「1 2 3 4 - 5 6 7 8」であるデータは、図 2 に示す請求書から抜き出されたデータであり、図 6 に示す学習用データ 3 4 b において項目「請求書番号」の値が「1 9 - 9 8 7 6」であるデータは、図 3 に示す請求書から抜き出されたデータである。

【 0 0 3 8 】

伝票情報テーブルにおける項目「請求書番号」、「支払い期限」、「総消費税」に対応する項目名は、それぞれ、図 2 に示す請求書において「No」、「お支払い期限」、「消費税」と表されている。伝票情報テーブルにおける項目「請求元名」、「請求元郵便番号」、「請求元住所」、「請求先名」、「請求先郵便番号」、「請求先住所」に対応する項目名は、図 2 に示す請求書に表れていない。伝票情報テーブルにおける項目「請求元名」、「請求元郵便番号」、「請求元住所」の値は、それぞれ、図 2 に示す請求書における右上側の特定の位置に表れている。伝票情報テーブルにおける項目「請求先名」、「請求先郵便番号」、「請求先住所」の値は、それぞれ、図 2 に示す請求書における左上側の特定の位置に表れている。伝票情報テーブルにおける項目「請求日」、「支払い期限」の値の日付のフォーマットは、図 2 に示す請求書における項目「請求日」、「お支払い期限」の値の日付のフォーマットとは異なる。

20

30

【 0 0 3 9 】

伝票情報テーブルにおける項目「請求書番号」、「請求日」、「支払い期限」、「小計」、「総消費税」、「合計」に対応する項目名は、それぞれ、図 3 に示す請求書において「請求番号」、「発行日」、「支払予定日」、「買上げ額」、「消費税」、「請求額」と表されている。伝票情報テーブルにおける項目「請求元名」、「請求元郵便番号」、「請求元住所」、「請求先名」、「請求先郵便番号」、「請求先住所」に対応する項目名は、図 3 に示す請求書に表れていない。伝票情報テーブルにおける項目「請求元名」、「請求元郵便番号」、「請求元住所」の値は、それぞれ、図 3 に示す請求書における右上側の特定の位置に表れている。伝票情報テーブルにおける項目「請求先名」、「請求先郵便番号」、「請求先住所」の値は、それぞれ、図 3 に示す請求書における左上側の特定の位置に表れている。伝票情報テーブルにおける項目「請求日」、「支払い期限」の値の日付のフォーマットは、図 3 に示す請求書における項目「発行日」、「支払予定日」の値の日付のフォーマットとは異なる。伝票情報テーブルにおける項目「請求元名」、「請求先名」の値の「株式会社」のフォーマットは、図 3 に示す請求書におけるフォーマットとは異なる。

40

【 0 0 4 0 】

品目テーブルにおける項目「請求書番号」、「商品名」、「単価」、「価格」に対応する項目名は、それぞれ、図 2 に示す請求書において「No」、「商品」、「単価（円）」、「価格（円）」と表されている。品目テーブルにおける項目「商品コード」、「消費税」に対応する項目名は、図 2 に示す請求書に表れていない。品目テーブルにおける項目「

50

商品名」の値は、図 2 に示す請求書において表の項目「商品」欄のセルにおける右側の値に対応している。品目テーブルにおける項目「商品コード」の値は、図 2 に示す請求書において表の項目「商品」欄のセルにおける左側の値に対応している。品目テーブルにおける項目「消費税」の値は、図 2 に示す請求書において表れておらず、図 2 に示す請求書における項目「単価（円）」の値から算出されたものである。

【 0 0 4 1 】

品目テーブルにおける項目「請求書番号」、「商品名」、「価格」に対応する項目名は、それぞれ、図 3 に示す請求書において「請求番号」、「商品名称」、「金額」と表されている。

【 0 0 4 2 】

図 5 に示すように、記憶部 3 4 は、帳票における文字列の属性の確率を取得するための文字列属性モデル 3 4 c を記憶することができる。

【 0 0 4 3 】

記憶部 3 4 は、文字列が帳票データにおけるいずれの属性に対応するかを学習するための文字列属性学習用データ 3 4 d を記憶している。ここで、帳票データにおける属性としては、いずれかの項目の項目名であることと、いずれかの項目の値であることが存在する。

【 0 0 4 4 】

記憶部 3 4 は、単語の分散表現辞書 3 4 e を記憶している。

【 0 0 4 5 】

記憶部 3 4 は、帳票における文字列の属性の位置関係のルールを示す属性位置関係ルール 3 4 f を記憶することができる。

【 0 0 4 6 】

記憶部 3 4 は、帳票データに存在するが、帳票に存在しない値を算出するための算出式を示す算出式情報 3 4 g を記憶している。

【 0 0 4 7 】

記憶部 3 4 は、帳票データにおける値のフォーマットのルールを示すフォーマット情報 3 4 h を記憶している。

【 0 0 4 8 】

制御部 3 5 は、例えば、CPU と、プログラムおよび各種のデータを記憶している ROM と、制御部 3 5 の CPU の作業領域として用いられる揮発性の記憶デバイスとしてのメモリーである RAM とを備えている。制御部 3 5 の CPU は、記憶部 3 4 または制御部 3 5 の ROM に記憶されているプログラムを実行する。

【 0 0 4 9 】

制御部 3 5 は、帳票データ取得プログラム 3 4 a を実行することによって、文字列属性モデル 3 4 c を作成する文字列属性学習部 3 5 a と、属性位置関係ルール 3 4 f を作成する属性位置関係学習部 3 5 b と、帳票データにおける値のフォーマットのルールを取得するフォーマット取得部 3 5 c と、帳票に存在しない値を算出するための算出式を導き出す算出式導出部 3 5 d と、帳票の画像に対する文字認識の結果における文字列に対して、文字列属性モデル 3 4 c を適用して、属性の確率を取得する属性確率取得部 3 5 e と、属性位置関係ルール 3 4 f に基づいて文字列に対する属性の確率を補正する属性確率補正部 3 5 f と、帳票に存在しない値を算出して帳票データに追加する値算出部 3 5 g と、帳票データにおける値のフォーマットを変更するフォーマット変更部 3 5 h とを実現する。

【 0 0 5 0 】

次に、情報処理システム 1 0 の動作について説明する。

【 0 0 5 1 】

まず、文字列属性モデル 3 4 c および属性位置関係ルール 3 4 f を作成する場合の情報処理装置 3 0 の動作について説明する。

【 0 0 5 2 】

図 7 は、文字列属性モデル 3 4 c および属性位置関係ルール 3 4 f を作成する場合の情

10

20

30

40

50

報処理装置30の動作の一部のフローチャートである。図8は、図7に示すフローチャートの続きのフローチャートである。図9は、図8に示すフローチャートの続きのフローチャートである。

【0053】

図7～図9に示すように、文字列属性学習部35aは、学習用データ34bに学習用帳票データが含まれる帳票（以下「学習用帳票」という。）のうち未だ対象にしていない1つの学習用帳票のみを対象にする（S101）。

【0054】

文字列属性学習部35aは、現在の対象の学習用帳票の画像を取得する（S102）。例えば、文字列属性学習部35aは、現在の対象の学習用帳票の画像を、この学習用帳票からスキャナ24によって取得しても良い。

10

【0055】

文字列属性学習部35aは、S102の処理の後、S102において取得された、学習用帳票の画像（以下「学習用帳票画像」という。）において文字認識の対象になる部分の位置情報をレイアウト認識などの技術を使用して取得する（S103）。ここで、文字列属性学習部35aは、学習用帳票画像において、文字列の部分や、枠部分の位置情報を取得することができる。文字列属性学習部35aは、例えば文書の塊、表、表のセルなどのように、順次、文字列になるまで、位置情報を取得する。例えば、文字列属性学習部35aは、いずれの表におけるいずれのセルの文字列であるかや、いずれのセルにおけるいずれの位置の文字列であるかなど、文字列の位置情報を取得する。

20

【0056】

文字列属性学習部35aは、S103の処理の後、学習用帳票画像のうち、S103において位置情報が取得された文字列の部分に対して、OCR（Optical Character Recognition）技術を使用して文字認識を実行する（S104）。

【0057】

文字列属性学習部35aは、S104の処理の後、現在の対象の学習用帳票に対応する学習用帳票データ（以下「対象学習用帳票データ」という。）を学習用データ34bから取得する（S105）。ここで、学習用データ34bにおける各テーブルには、それぞれ、学習用データ34bの作成者の意図が存在するので、各テーブル内のデータ同士の関係性は高い。そのため、文字列属性学習部35aは、S105において対象学習用帳票データを取得する際に、現在の対象の学習用帳票データにおける個々のデータがいずれのテーブルのデータであるかについても取得する。

30

【0058】

文字列属性学習部35aは、S105の処理の後、S104における文字認識の結果（以下「文字認識結果」という。）における文字列（以下「認識結果中文字列」という。）と、対象学習用帳票データにおける項目名（以下「対象学習用帳票データ内項目名」という。）および値（以下「対象学習用帳票データ内値」という。）との対応付けを実行する（S106）。

【0059】

文字列属性学習部35aは、文字認識結果に改行が含まれている可能性も考慮して、S106における対応付けを実行する。例えば、図2に示す請求書および図3に示す請求書における請求先住所と、図2に示す請求書における請求元住所とには、改行が含まれているが、文字列属性学習部35aは、改行の前後の文字列が一連の文字列であると判断して、S106における対応付けを実行する。

40

【0060】

文字列属性学習部35aは、認識結果中文字列と、対象学習用帳票データ内項目名および対象学習用帳票データ内値との間に、文字列のフォーマットの差異が存在することを考慮して、S106における対応付けを実行する。例えば、対象学習用帳票データにおける項目「請求日」、「支払い期限」の値の日付のフォーマットは、図2に示す請求書におけ

50

るフォーマットと差異があるが、文字列属性学習部 3 5 a は、同一の日付を意味していると判断して、S 1 0 6 における対応付けを実行する。また、対象学習用帳票データにおける項目「請求元名」、「請求先名」の値の「株式会社」のフォーマットは、図 3 に示す請求書におけるフォーマットと差異があるが、文字列属性学習部 3 5 a は、同一の株式会社を意味していると判断して、S 1 0 6 における対応付けを実行する。なお、文字列のフォーマットの差異の種類には、文字列の一部の省略も含まれる。

【 0 0 6 1 】

文字列属性学習部 3 5 a は、S 1 0 6 の処理の後、S 1 0 6 における対応付けの結果（以下「対応付け結果」という。）において、対象学習用帳票データ内値とフォーマットが異なる認識結果中文字列が存在するか否かを判断する（S 1 0 7 ）。

10

【 0 0 6 2 】

フォーマット取得部 3 5 c は、対象学習用帳票データ内値とフォーマットが異なる認識結果中文字列が対応付け結果において存在すると S 1 0 7 において判断されると、対応付け結果においてフォーマットが異なる対象学習用帳票データ内値および認識結果中文字列の組み合わせに基づいたフォーマットのルールをフォーマット情報 3 4 h に追加する（S 1 0 8 ）。

【 0 0 6 3 】

文字列属性学習部 3 5 a は、S 1 0 8 の処理の後、対象学習用帳票データ内値とフォーマットが異なる、いずれの認識結果中文字列がいずれの対象学習用帳票データ内項目名の項目の値に対応するか、すなわち、いずれの対象学習用帳票データ内値に対応するかを、文字列属性学習用データ 3 4 d に追加する（S 1 0 9 ）。

20

【 0 0 6 4 】

値算出部 3 5 g は、対象学習用帳票データ内値とフォーマットが異なる認識結果中文字列が対応付け結果において存在しないと S 1 0 7 において判断されるか、S 1 0 9 の処理が終了すると、認識結果中文字列に対応するものが存在しない対象学習用帳票データ内値が存在するか否かを判断する（S 1 1 0 ）。例えば、小計や、消費税などは、S 1 0 2 において画像が取得された帳票にそもそも含まれておらず対象学習用帳票データの作成時に追加される場合がある。すなわち、認識結果中文字列に対応するものが存在しない対象学習用帳票データ内値が存在する場合がある。

【 0 0 6 5 】

30

値算出部 3 5 g は、認識結果中文字列に対応するものが存在しない対象学習用帳票データ内値が存在すると S 1 1 0 において判断すると、認識結果中文字列に対応するものが存在しない対象学習用帳票データ内値について、学習用帳票データにおける値のそれぞれの間の関係性を分析して、学習用帳票データにおける他の値からの算出式を導き出す（S 1 1 1 ）。ここで、値算出部 3 5 g は、1 つの帳票の学習用帳票データにおける値のそれぞれの間の関係性を分析して算出式を導き出すよりも、多数の帳票の学習用帳票データにおける値のそれぞれの間の関係性を分析して算出式を導き出す方が好ましい。

【 0 0 6 6 】

値算出部 3 5 g は、S 1 1 1 の処理の後、S 1 1 1 において導き出した算出式を算出式情報 3 4 g に追加する（S 1 1 2 ）。

40

【 0 0 6 7 】

文字列属性学習部 3 5 a は、認識結果中文字列に対応するものが存在しない対象学習用帳票データ内値が存在しないと S 1 1 0 において判断されるか、S 1 1 2 の処理が終了すると、対象学習用帳票データにおける項目のうち、対象学習用帳票データ内値に対応する文字列が文字認識結果に存在するが、対象学習用帳票データ内項目名に対応する文字列が文字認識結果に存在しないもの（以下「項目名対応文字列無項目」という。）が存在するか否かを判断する（S 1 1 3 ）。

【 0 0 6 8 】

文字列属性学習部 3 5 a は、項目名対応文字列無項目が存在すると S 1 1 3 において判断すると、項目名対応文字列無項目の対象学習用帳票データ内値に文字認識結果において

50

対応する文字列との、学習用帳票画像における位置関係に基づいて、項目名対応文字列無項目の項目名に対応する文字列の候補（以下「項目名対応文字列候補」という。）を探索する（S 1 1 4）。例えば、文字列属性学習部 3 5 a は、項目名対応文字列無項目の対象学習用帳票データ内値に文字認識結果において対応する文字列に対して、学習用帳票画像において水平方向における左側に存在する文字列を文字認識結果において探索するとともに、学習用帳票画像において垂直方向における上側に存在する文字列を文字認識結果において探索することによって、探索された文字列を項目名対応文字列候補として抽出する。

【0069】

文字列属性学習部 3 5 a は、S 1 1 4 の処理の後、S 1 1 4 における探索によって項目名対応文字列候補が抽出されたか否かを判断する（S 1 1 5）。10

【0070】

文字列属性学習部 3 5 a は、S 1 1 4 における探索によって項目名対応文字列候補が抽出されたと S 1 1 5 において判断すると、S 1 1 4 における探索によって抽出された項目名対応文字列候補と、この項目名対応文字列候補に対応する項目名との類似度を分散表現辞書 3 4 a を使用して判定することによって、S 1 1 4 における探索によって抽出された項目名対応文字列候補の属性が、この項目名対応文字列候補に対応する項目名であるか否かを判定する（S 1 1 6）。

【0071】

文字列属性学習部 3 5 a は、S 1 1 6 の処理の後、S 1 1 4 における探索によって抽出された項目名対応文字列候補の属性が、この項目名対応文字列候補に対応する項目名であると S 1 1 6 において判定されたか否かを判断する（S 1 1 7）。20

【0072】

文字列属性学習部 3 5 a は、S 1 1 4 における探索によって抽出された項目名対応文字列候補の属性が、この項目名対応文字列候補に対応する項目名であると S 1 1 6 において判定されたと S 1 1 7 において判断すると、対象学習用帳票データ内項目名とフォーマットが異なる、いずれの認識結果中文字列がいずれの対象学習用帳票データ内項目名に対応するかを、文字列属性学習用データ 3 4 d に追加する（S 1 1 8）。

【0073】

文字列属性学習部 3 5 a は、項目名対応文字列無項目が存在しないと S 1 1 3 において判断するか、S 1 1 4 における探索によって項目名対応文字列候補が抽出されなかったと S 1 1 5 において判断するか、S 1 1 4 における探索によって抽出された項目名対応文字列候補が、この項目名対応文字列候補に対応する項目名であると S 1 1 6 において判定されなかったと S 1 1 7 において判断するか、S 1 1 8 の処理を実行すると、対象学習用帳票データ内項目名および対象学習用帳票データ内値のいずれにも対応しない認識結果中文字列、すなわち、属性付けがされていない認識結果中文字列が存在するか否かを判断する（S 1 1 9）。30

【0074】

文字列属性学習部 3 5 a は、属性付けがされていない認識結果中文字列が存在すると S 1 1 9 において判断すると、属性付けがされていない認識結果中文字列を属性が無い文字列とすることを、文字列属性学習用データ 3 4 d に追加する（S 1 2 0）。40

【0075】

文字列属性学習部 3 5 a は、属性付けがされていない認識結果中文字列が存在しないと S 1 1 9 において判断するか、S 1 2 0 の処理が終了すると、未だ対象にしていない学習用帳票が存在するか否かを判断する（S 1 2 1）。

【0076】

文字列属性学習部 3 5 a は、未だ対象にしていない学習用帳票が存在すると S 1 2 1 において判断すると、S 1 0 1 の処理を実行する。

【0077】

属性位置関係学習部 3 5 b は、未だ対象にしていない学習用帳票が存在しないと S 1 2 1 において判断されると、全ての学習用帳票における属性の位置関係を学習することによ50

って、属性位置関係ルール 3 4 f を作成する (S 1 2 2)。属性位置関係ルール 3 4 f には、例えば、属性同士のグループ (以下「属性グループ」という。) 内における属性の位置関係のルールと、属性グループ同士の位置関係のルールとを含んでいる。属性グループとしては、例えば、帳票において必ず近傍に存在する属性同士のグループと、帳票において必ず水平線上に存在する属性同士のグループと、帳票において必ず垂直線上に存在する属性同士のグループとが存在する。例えば、属性グループには、請求元名、請求元郵便番号、請求元住所の 3 つの属性が存在する、請求元に関する属性グループと、請求先名、請求先郵便番号、請求先住所の 3 つの属性が存在する、請求先に関する属性グループとが存在する。

【 0 0 7 8 】

属性グループ内における属性の位置関係のルールは、帳票において左もしくは上に出現する確率を属性毎に求めることによって作成されるルールであり、帳票における属性同士の上下左右の位置関係のルールである。例えば、請求元に関する属性グループ内において、請求元名は、帳票において上から 1 番目に出現する確率が高く、請求元郵便番号は、帳票において上から 2 番目に出現する確率が高く、請求元住所は、帳票において上から 3 番目に出現する確率が高い。

【 0 0 7 9 】

属性グループ同士の位置関係のルールは、帳票において左もしくは上に出現する確率を属性グループ毎に求めることによって作成されるルールであり、帳票における属性グループ同士の上下左右の位置関係のルールである。例えば、請求先に関する属性グループは、請求元に関する属性グループに対して、帳票において左側に出現する確率が高い。

【 0 0 8 0 】

文字列属性学習部 3 5 a は、S 1 2 2 の処理の後、学習用データ 3 4 b と、文字列属性学習用データ 3 4 d とに基づいて、RNN (R e c u r r e n t N e u r a l N e t w o r k) などを利用して、文字列属性モデル 3 4 c を作成して (S 1 2 3)、図 7 ~ 図 9 に示す動作を終了する。

【 0 0 8 1 】

次に、帳票から帳票データを自動で作成する場合の情報処理装置 3 0 の動作について説明する。

【 0 0 8 2 】

図 1 0 は、帳票から帳票データを自動で作成する場合の情報処理装置 3 0 の動作のフローチャートである。

【 0 0 8 3 】

図 1 0 に示すように、属性確率取得部 3 5 e は、S 1 0 2 の処理と同様に、対象の帳票の画像を取得する (S 1 3 1)。例えば、属性確率取得部 3 5 e は、対象の帳票の画像を、この帳票からスキャナ 2 4 によって取得しても良い。

【 0 0 8 4 】

属性確率取得部 3 5 e は、S 1 3 1 の処理の後、S 1 0 3 の処理と同様に、S 1 3 1 において取得された、帳票の画像 (以下「帳票画像」という。) において文字認識の対象になる部分の位置情報をレイアウト認識などの技術を使用して取得する (S 1 3 2)。

【 0 0 8 5 】

属性確率取得部 3 5 e は、S 1 3 2 の処理の後、S 1 0 4 の処理と同様に、帳票画像のうち、S 1 3 2 において位置情報が取得された文字列の部分に対して、OCR 技術を使用して文字認識を実行する (S 1 3 3)。

【 0 0 8 6 】

属性確率取得部 3 5 e は、S 1 3 3 の処理の後、S 1 3 3 において文字認識した文字列に対して、文字列属性モデル 3 4 c を適用して、属性の確率を取得する (S 1 3 4)。

【 0 0 8 7 】

属性確率補正部 3 5 f は、S 1 3 4 の処理の後、S 1 3 3 において文字認識した文字列の、帳票における位置と、属性位置関係ルール 3 4 f とに基づいて、S 1 3 4 において取

10

20

30

40

50

得した確率を補正する（Ｓ１３５）。例えば、属性確率補正部３５ｆは、Ｓ１３３において文字認識した文字列に対して、帳票における位置と、属性位置関係ルール３４ｆとに基づいて取得した、属性の確率と、Ｓ１３４において取得した、属性の確率とを掛け合わせることで、最終的な属性の確率を求める。

【００８８】

属性確率補正部３５ｆは、Ｓ１３５の処理の後、Ｓ１３３において文字認識した文字列の属性を、Ｓ１３５において補正した確率に基づいて判定することによって、帳票データを作成する（Ｓ１３６）。

【００８９】

値算出部３５ｇは、Ｓ１３６の処理の後、算出式情報３４ｇに算出式が示されているかどうかを判断する（Ｓ１３７）。

【００９０】

値算出部３５ｇは、算出式情報３４ｇに算出式が示されているとＳ１３７において判断すると、対象の帳票に存在しない値を、算出式情報３４ｇに示される算出式に基づいて算出する（Ｓ１３８）。

【００９１】

値算出部３５ｇは、Ｓ１３８の処理の後、Ｓ１３８において算出した値を、Ｓ１３６において作成された帳票データに追加する（Ｓ１３９）。

【００９２】

フォーマット変更部３５ｈは、算出式情報３４ｇに算出式が示されていないとＳ１３７において判断されるか、Ｓ１３９の処理が終了すると、フォーマット情報３４ｈにフォーマットのルールが示されているかどうかを判断する（Ｓ１４０）。

【００９３】

フォーマット変更部３５ｈは、フォーマット情報３４ｈにフォーマットのルールが示されているとＳ１４０において判断すると、フォーマット情報３４ｈに示されるルールに基づいて帳票データにおける値のフォーマットを変更する（Ｓ１４１）。

【００９４】

フォーマット変更部３５ｈは、フォーマット情報３４ｈにフォーマットのルールが示されていないとＳ１４０において判断するか、Ｓ１４１の処理が終了すると、図１０に示す動作を終了する。

【００９５】

図１０に示す動作の具体例として、図２に示す帳票から帳票データを自動で作成する場合について説明する。

【００９６】

図１１は、文字列属性モデル３４ｃを適用して判定された、図２に示す帳票に含まれる文字列の属性の一例を示す図である。図１２は、属性位置関係ルール３４ｆを適用して補正された、図２に示す帳票に含まれる文字列の属性の一例を示す図である。

【００９７】

図２に示す帳票から帳票データを自動で作成する場合、Ｓ１３４において取得した確率に基づいて判定される属性は、例えば、図１１に示すようになる。

【００９８】

Ｓ１３４の処理の後、Ｓ１３５において補正した確率に基づいて判定される属性は、例えば、図１２に示すようになる。図１２に示す例では、図１１に示す例と比較して、帳票における右上側に配置されている、項目「請求先名」の値と判定されていた文字列「株式会社」、項目「請求先郵便番号」の値と判定されていた文字列「〒９８７－６５４３」、および、項目「請求先住所」の値と判定されていた文字列「神奈川県 市 区

９－８－７ ６Ｆ」が、それぞれ、項目「請求元名」の値、項目「請求元郵便番号」の値、および、項目「請求元住所」の値と判定されている。また、帳票における左上側に配置されている、項目「合計」の項目名と判定されていた文字列「合計（税込）」、および、項目「合計」の値と判定されていた文字列「¥２６４０」が、それぞれ「属性

10

20

30

40

50

無し」と判定されている。また、帳票における左下側に配置されている、項目「請求書番号」の値と判定されていた文字列「０００００００」が、「属性無し」と判定されている。

【００９９】

図１３（ａ）は、図１２に示す属性に基づいて作成された帳票データの伝票情報テーブルの一例を示す図である。図１３（ｂ）は、図１２に示す属性に基づいて作成された帳票データの品目テーブルの一例を示す図である。

【０１００】

図１２に示す属性に基づいてＳ１３６において作成された帳票データは、図１３に示すようになる。図１３（ｂ）においては、項目「消費税」の値が入力されていない。

10

【０１０１】

図１４（ａ）は、図２に示す帳票から自動で作成された帳票データの伝票情報テーブルの一例を示す図である。図１４（ｂ）は、図２に示す帳票から自動で作成された帳票データの品目テーブルの一例を示す図である。

【０１０２】

図１４（ａ）に示す伝票情報テーブルは、図１３（ａ）に示す伝票情報テーブルと比較して、項目「請求日」、「支払い期限」の値のフォーマットがＳ１４１において変更されている。図１４（ｂ）に示す品目テーブルは、図１３（ｂ）に示す品目テーブルと比較して、項目「消費税」の値がＳ１３９において追加されている。図１４（ｂ）に示す品目テーブルは、図１３（ｂ）に示す品目テーブルと比較して、項目「価格」の値「１，０００」のフォーマットがＳ１４１において「１０００」に変更されている。

20

【０１０３】

なお、図１０に示す動作によって作成された帳票データは、利用者による確認を経て、新たな学習用帳票データになることができる。制御部３５は、新たな学習用帳票データに基づいて図７～図９に示す動作を追加で実行することによって、帳票から自動で作成する帳票データの精度を向上することができる。

【０１０４】

以上に説明したように、情報処理装置３０は、学習用帳票データと、学習用帳票の画像に対する文字認識の結果における文字列とを用いて文字列属性モデル３４ｃおよび属性位置関係ルール３４ｆを学習する（Ｓ１０１～Ｓ１２３）ので、学習用帳票に適した文字列属性モデル３４ｃおよび属性位置関係ルール３４ｆを得ることができ、その結果、帳票から自動で取得した帳票データの適切性を向上することができる。

30

【０１０５】

情報処理装置３０は、学習用帳票データにおける項目の値とフォーマットが異なるが、この値に対応する文字列を、項目の値として取得できるように文字列属性モデル３４ｃを作成する（Ｓ１０９およびＳ１２３）ので、帳票から自動で取得した帳票データの適切性を更に向上することができる。

【０１０６】

情報処理装置３０は、帳票データにおける項目の値とフォーマットが異なるが、この値に対応する文字列のフォーマットを、帳票データにおける項目の値のフォーマットに変更する（Ｓ１４１）ので、帳票から自動で取得した帳票データの適切性を更に向上することができる。

40

【０１０７】

情報処理装置３０は、項目名対応文字列無項目が存在する場合に（Ｓ１１３でＹＥＳ）、項目名対応文字列無項目の学習用帳票データにおける値に、学習用帳票の画像に対する文字認識の結果において対応する文字列との、学習用帳票の画像における位置関係に基づいて、項目名対応文字列無項目の項目名に対応する文字列の項目名対応文字列候補を探索し（Ｓ１１４）、探索によって抽出された項目名対応文字列候補と、この項目名対応文字列候補に対応する項目名との類似度に基づいて、この項目名対応文字列候補の属性が、この項目名であると判定したとき（Ｓ１１７でＹＥＳ）、この項目名対応文字列候補がいず

50

れの項目名に対応するかということにも基づいて、文字列属性モデルを作成する（S 1 1 8 および S 1 2 3）ので、帳票から自動で取得した帳票データの適切性を更に向上することができる。

【0 1 0 8】

情報処理装置 3 0 は、学習用帳票の画像に対する文字認識の結果に存在する文字列に対応するものが存在しない、学習用帳票データにおける値について、学習用帳票データにおける値のそれぞれの間の関係性を分析して、学習用帳票データにおける他の値からの算出式を導き出し（S 1 1 1）、導き出した算出式に基づいて、帳票に存在しない値を帳票データに追加する（S 1 3 8 ~ S 1 3 9）ので、帳票から自動で取得した帳票データの適切性を更に向上することができる。

10

【0 1 0 9】

情報処理装置 3 0 は、学習用帳票および学習用帳票データを用いて文字列属性モデル 3 4 c および属性位置関係ルール 3 4 f を学習するので、利用者の希望する帳票について学習用帳票および学習用帳票データが用意されることによって、利用者の希望する帳票から適切に帳票データを取得することができる。例えば、請求書ではなく図 1 5 に示す健康診断書から情報処理装置 3 0 に帳票データを取得させる場合、利用者は、学習用帳票として図 1 5 に示す健康診断書などの健康診断書を用いて、健康診断書用の文字列属性モデルと、健康診断書用の属性位置関係ルールとを情報処理装置 3 0 に新たに学習させる必要がある。

20

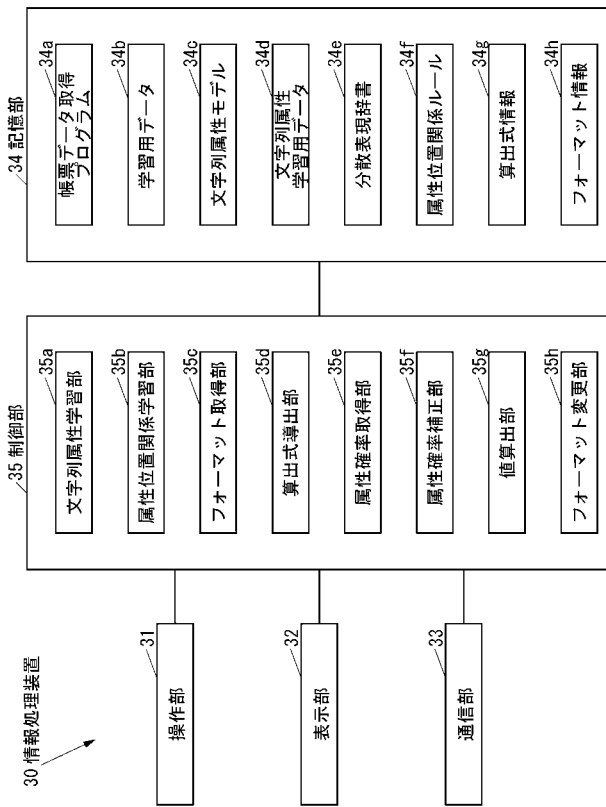
【符号の説明】

【0 1 1 0】

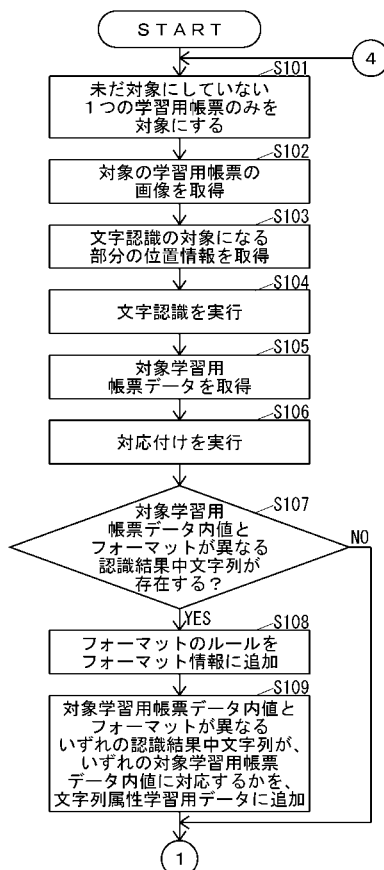
- 3 0 情報処理装置（帳票データ取得システム、コンピューター）
- 3 4 a 帳票データ取得プログラム
- 3 4 b 学習用データ（学習用帳票データ）
- 3 4 c 文字列属性モデル
- 3 4 f 属性位置関係ルール
- 3 5 a 文字列属性学習部
- 3 5 b 属性位置関係学習部
- 3 5 c フォーマット取得部
- 3 5 d 算出式導出部
- 3 5 e 属性確率取得部
- 3 5 f 属性確率補正部
- 3 5 g 値算出部
- 3 5 h フォーマット変更部

30

【 図 5 】



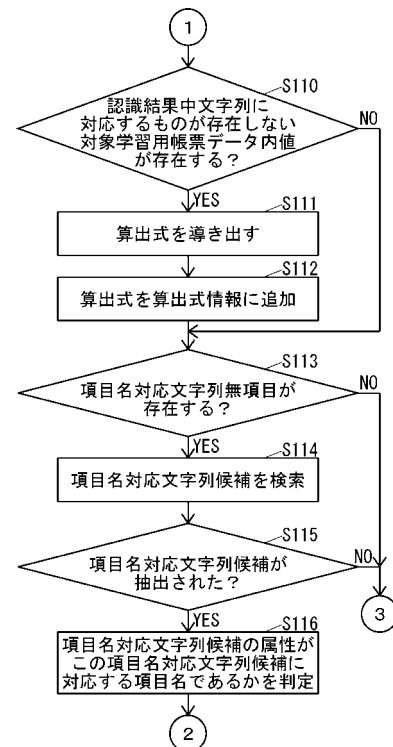
【 図 7 】



【 図 6 】

[illegible]

【 図 8 】

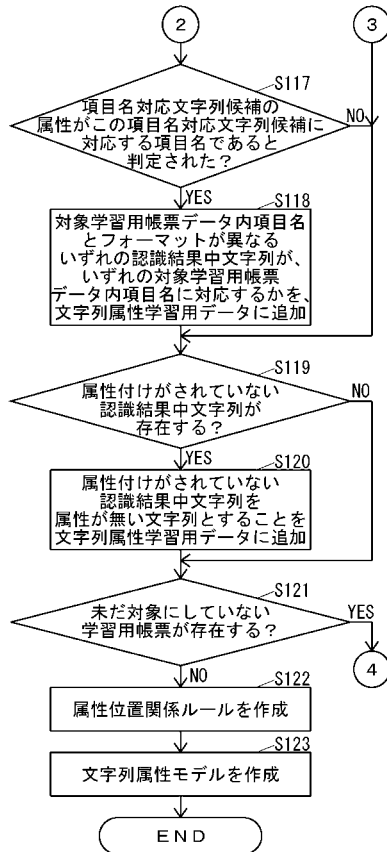


品目コード	商品名	商品コード	単価	数量	価格	消費税
1234-5678	ボールペン	3456-78	500	1	500	50
1234-5678	シャープペンシル	1234-56	200	2	400	40
1234-5678	用紙	9876	100	5	500	50
1234-5678	名刺	5432	100	10	1000	100
19-9876	書籍	2222	800	100	80000	8000
19-9876	名札	3333	100	120	12000	1200
19-9876	花束	1-0001	2000	20	40000	4000
・	・	・	・	・	・	・
・	・	・	・	・	・	・
・	・	・	・	・	・	・

(9)

(9)

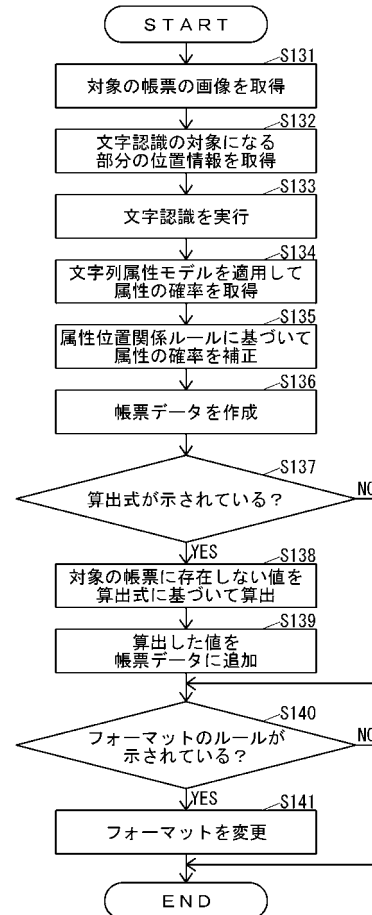
【図 9】



【図 1 1】

請求先名		請求書番号項目		請求書番号	
請求先郵便番号		請求日項目		請求日	
請求先住所		請求先名		請求先郵便番号	
請求先住所		請求先住所		請求先住所	
合計項目		合計			
商品名項目	単価項目	数量項目	価格項目		
商品コード	商品名	単価	数量	価格	
商品コード	商品名	単価	数量	価格	
商品コード	商品名	単価	数量	価格	
商品コード	商品名	単価	数量	価格	
小計項目		小計			
総消費税項目		総消費税			
合計項目		合計			
支払い期限項目		支払い期限			

【図 1 0】



【図 1 2】

請求先名		請求書番号項目		請求書番号	
請求先郵便番号		請求日項目		請求日	
請求先住所		請求元名		請求元郵便番号	
請求先住所		請求元住所		請求元住所	
合計項目		合計			
商品名項目	単価項目	数量項目	価格項目		
商品コード	商品名	単価	数量	価格	
商品コード	商品名	単価	数量	価格	
商品コード	商品名	単価	数量	価格	
商品コード	商品名	単価	数量	価格	
小計項目		小計			
総消費税項目		総消費税			
合計項目		合計			
支払い期限項目		支払い期限			

【図 1 3】

伝票情報テーブル									
請求書番号	請求日	支払い期限	請求元名	請求元郵便番号	請求元住所	請求先名	請求先郵便番号	請求先住所	合計
1234-5678	2019年11月22日	2019年12月25日	株式会社△△	〒887-6543	神奈川県△△市△△△△△△F	株式会社△△	〒123-4567	東京都△△区△△町1-2-3 △△ビル4F	2640
									消費税
									240
									小計
									2400

(a)

品目テーブル					消費税	
請求書番号	商品名	商品コード	単価	数量	価格	
1234-5678	ボールペン	3456-78	500	1	500	unknown
1234-5678	シャープペンシル	1234-56	200	2	400	unknown
1234-5678	用紙	9876	100	5	500	unknown
1234-5678	名札	5432	100	10	1000	unknown

(b)

【図 1 4】

伝票情報テーブル									
請求書番号	請求日	支払い期限	請求元名	請求元郵便番号	請求元住所	請求先名	請求先郵便番号	請求先住所	合計
1234-5678	2019/11/22	2019/12/25	株式会社△△	〒987-6543	神奈川県△△市△△△△△△F	株式会社△△	〒123-4567	東京都△△区△△町1-2-3 △△ビル4F	2640
									消費税
									240
									小計
									2400

(a)

品目テーブル					消費税	
請求書番号	商品名	商品コード	単価	数量	価格	
1234-5678	ボールペン	3456-78	500	1	500	50
1234-5678	シャープペンシル	1234-56	200	2	400	40
1234-5678	用紙	9876	100	5	500	50
1234-5678	名札	5432	100	10	1000	100

(b)

【図 1 5】

健康診断報告書					検査日: 2019/11/26
氏名		京セラ 太郎			
個人番号		12345678			
生年月日		1990/11/11			身長/体重: 175cm 70kg
検査項目	基準値	前々回	前回	今回	
栄養					
総蛋白					
アルブミン					
肝・胆					
総ビリルビン					
GOT					
GPT					
ALP					
脂質					
HDLコレステロール					
LDLコレステロール					
中性脂肪					
尿検査					
血糖					
酸化あるブトン					
血液検査					
赤血球数					
ヘモグロビン					
ヘマトクリット					
MCV					
MCH					
MCHC					
血小板					
内分泌					
T4					
FT4					
感染					
白血球数					
フィラリア抗原					