(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)公開番号
特開2022-144738
(P2022-144738A)
(43)公開日　令和4年10月3日(2022.10.3)

(51)国際特許分類			FI			テーマコード（参考）
G06F	*16/35*	*(2019.01)*	G06F	16/35		5B175
G06N	*20/00*	*(2019.01)*	G06N	20/00	160	

審査請求　未請求　請求項の数　7　OL　（全17頁）

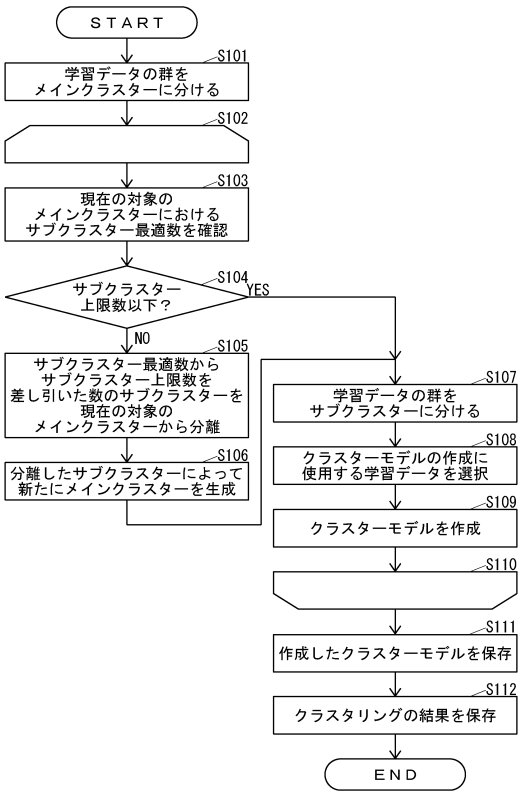
(21)出願番号　特願2021-45884(P2021-45884)
(22)出願日　令和3年3月19日(2021.3.19)

(71)出願人　000006150
京セラドキュメントソリューションズ株式会社
大阪府大阪市中央区玉造１丁目２番２８号
(74)代理人　100140796
弁理士　原口 貴志
(72)発明者　庄司 秀典
大阪府大阪市中央区玉造１丁目２番２８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）　5B175　DA01 FA03

(54)【発明の名称】　情報抽出システムおよび情報抽出プログラム

(57)【要約】
【課題】 情報抽出モデルの作成のための計算量を低減することができる情報抽出システムおよび情報抽出プログラムを提供する。
【解決手段】 情報抽出システムは、請求書データから情報を抽出するための情報抽出モデルであるクラスターモデルの作成のための学習データの群をクラスタリングすることによって、学習データのそれぞれをいずれかのメインクラスターに分け（Ｓ１０１）、メインクラスター毎に学習データを使用して学習を実行することによって、メインクラスター毎のクラスターモデルを作成する（Ｓ１０９）ことを特徴とする。
【選択図】 図３



【特許請求の範囲】
【請求項１】
文書のデータから情報を抽出するための情報抽出モデルの作成のための学習データの群をクラスタリングすることによって、前記学習データのそれぞれをいずれかのメインクラスターに分ける文書クラスタリング部と、
前記メインクラスター毎に前記学習データを使用して学習を実行することによって、前記メインクラスター毎の前記情報抽出モデルを作成するモデル学習部と
を備えることを特徴とする情報抽出システム。
【請求項２】
前記文書クラスタリング部は、前記メインクラスター内の前記学習データの群をクラスタリングすることによって、前記メインクラスター内の前記学習データのそれぞれをいずれかのサブクラスターに分け、
前記モデル学習部は、前記情報抽出モデルの作成に使用する前記学習データを前記サブクラスター毎に選択し、選択した前記学習データを使用して学習を実行することによって、前記メインクラスター毎の前記情報抽出モデルを作成することを特徴とする請求項１に記載の情報抽出システム。
【請求項３】
前記モデル学習部は、重心が前記メインクラスターの重心に最も近い前記サブクラスターにおいて、重心が前記メインクラスターの重心に最も近い前記学習データを、前記情報抽出モデルの作成に使用する前記学習データとして選択することを特徴とする請求項２に記載の情報抽出システム。
【請求項４】
前記モデル学習部は、重心が前記メインクラスターの重心に最も近い前記サブクラスター以外の前記サブクラスターのそれぞれにおいて、重心が前記メインクラスターの重心から最も遠い前記学習データを、前記情報抽出モデルの作成に使用する前記学習データとして選択することを特徴とする請求項３に記載の情報抽出システム。
【請求項５】
前記文書クラスタリング部は、前記メインクラスターにおける前記サブクラスターの最適数をクラスター数自動推定法によって確認し、確認した前記最適数が特定の上限数を超える場合に、前記最適数から前記上限数を差し引いた数の前記サブクラスターを、このメインクラスターから分離することを特徴とする請求項２から請求項４までのいずれかに記載の情報抽出システム。
【請求項６】
前記文書クラスタリング部は、前記最適数から前記上限数を差し引いた数の前記サブクラスターを前記メインクラスターから分離する場合に、重心がこのメインクラスターの重心から遠い前記サブクラスターを優先して、このメインクラスターから分離することを特徴とする請求項５に記載の情報抽出システム。
【請求項７】
文書のデータから情報を抽出するための情報抽出モデルの作成のための学習データの群をクラスタリングすることによって、前記学習データのそれぞれをいずれかのメインクラスターに分ける文書クラスタリング部と、
前記メインクラスター毎に前記学習データを使用して学習を実行することによって、前記メインクラスター毎の前記情報抽出モデルを作成するモデル学習部と
をコンピューターに実現させることを特徴とする情報抽出プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、文書のデータから特定の項目に対する値を抽出する情報抽出システムおよび情報抽出プログラムに関する。
【背景技術】

【０００２】
従来、文書のデータから情報を抽出するための情報抽出モデルを使用して文書のデータから情報を抽出する情報抽出システムが知られている（例えば、特許文献１、２参照。）。
【先行技術文献】
【特許文献】
【０００３】
【特許文献１】米国特許出願公開第２０１４／０１７７９５１号明細書
【特許文献２】特許第６６２９９４２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、従来の情報抽出システムにおいては、情報抽出モデルの作成のための計算量が多いという問題がある。
【０００５】
そこで、本発明は、情報抽出モデルの作成のための計算量を低減することができる情報抽出システムおよび情報抽出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】
本発明の情報抽出システムは、文書のデータから情報を抽出するための情報抽出モデルの作成のための学習データの群をクラスタリングすることによって、前記学習データのそれぞれをいずれかのメインクラスターに分ける文書クラスタリング部と、前記メインクラスター毎に前記学習データを使用して学習を実行することによって、前記メインクラスター毎の前記情報抽出モデルを作成するモデル学習部とを備えることを特徴とする。
【０００７】
この構成により、本発明の情報抽出システムは、メインクラスター毎に情報抽出モデルを作成するので、情報抽出モデル毎の特徴を単純化することができ、その結果、情報抽出モデル毎に必要な学習データの数を低減することができる。したがって、本発明の情報抽出システムは、情報抽出モデルの作成のための計算量を低減することができる。
【０００８】
本発明の情報抽出システムにおいて、前記文書クラスタリング部は、前記メインクラスター内の前記学習データの群をクラスタリングすることによって、前記メインクラスター内の前記学習データのそれぞれをいずれかのサブクラスターに分け、前記モデル学習部は、前記情報抽出モデルの作成に使用する前記学習データを前記サブクラスター毎に選択し、選択した前記学習データを使用して学習を実行することによって、前記メインクラスター毎の前記情報抽出モデルを作成しても良い。
【０００９】
この構成により、本発明の情報抽出システムは、情報抽出モデルの作成に使用する学習データをサブクラスター毎に選択し、選択した学習データを使用して学習を実行することによって、メインクラスター毎の情報抽出モデルを作成するので、情報抽出モデル毎に必要な学習データの数を低減することができ、その結果、情報抽出モデルの作成のための計算量を低減することができる。
【００１０】
本発明の情報抽出システムにおいて、前記モデル学習部は、重心が前記メインクラスターの重心に最も近い前記サブクラスターにおいて、重心が前記メインクラスターの重心に最も近い前記学習データを、前記情報抽出モデルの作成に使用する前記学習データとして選択しても良い。
【００１１】
この構成により、本発明の情報抽出システムは、重心がメインクラスターの重心に最も近いサブクラスターにおいて、重心がメインクラスターの重心に最も近い学習データを、

情報抽出モデルの作成に使用する学習データとして選択するので、メインクラスターの特徴を最も強く表す学習データを使用して情報抽出モデルを作成することができ、その結果、メインクラスターの特徴が適切に反映された情報抽出モデルを作成することができる。

【００１２】

本発明の情報抽出システムにおいて、前記モデル学習部は、重心が前記メインクラスターの重心に最も近い前記サブクラスター以外の前記サブクラスターのそれぞれにおいて、重心が前記メインクラスターの重心から最も遠い前記学習データを、前記情報抽出モデルの作成に使用する前記学習データとして選択しても良い。

【００１３】

この構成により、本発明の情報抽出システムは、重心がメインクラスターの重心に最も近いサブクラスター以外のサブクラスターのそれぞれにおいて、重心がメインクラスターの重心から最も遠い学習データを、情報抽出モデルの作成に使用する学習データとして選択するので、メインクラスターにおいて広範囲に散らばった学習データを使用して情報抽出モデルを作成することができ、その結果、メインクラスターの特徴が適切に反映された情報抽出モデルを作成することができる。

【００１４】

本発明の情報抽出システムにおいて、前記文書クラスタリング部は、前記メインクラスターにおける前記サブクラスターの最適数をクラスター数自動推定法によって確認し、確認した前記最適数が特定の上限数を超える場合に、前記最適数から前記上限数を差し引いた数の前記サブクラスターを、このメインクラスターから分離しても良い。

【００１５】

この構成により、本発明の情報抽出システムは、メインクラスターにおけるサブクラスターの最適数が特定の上限数を超える場合に、最適数から上限数を差し引いた数のサブクラスターを、このメインクラスターから分離するので、情報抽出モデル毎に必要な学習データの数を低減することができ、その結果、情報抽出モデルの作成のための計算量を低減することができる。

【００１６】

本発明の情報抽出システムにおいて、前記文書クラスタリング部は、前記最適数から前記上限数を差し引いた数の前記サブクラスターを前記メインクラスターから分離する場合に、重心がこのメインクラスターの重心から遠い前記サブクラスターを優先して、このメインクラスターから分離しても良い。

【００１７】

この構成により、本発明の情報抽出システムは、最適数から上限数を差し引いた数のサブクラスターをメインクラスターから分離する場合に、重心がこのメインクラスターの重心から遠いサブクラスターを優先して、このメインクラスターから分離するので、メインクラスターの特徴を強く表す学習データを使用して情報抽出モデルを作成することができ、その結果、メインクラスターの特徴が適切に反映された情報抽出モデルを作成することができる。

【００１８】

本発明の情報抽出プログラムは、文書のデータから情報を抽出するための情報抽出モデルの作成のための学習データの群をクラスタリングすることによって、前記学習データのそれぞれをいずれかのメインクラスターに分ける文書クラスタリング部と、前記メインクラスター毎に前記学習データを使用して学習を実行することによって、前記メインクラスター毎の前記情報抽出モデルを作成するモデル学習部とをコンピューターに実現させることを特徴とする。

【００１９】

この構成により、本発明の情報抽出プログラムを実行するコンピューターは、メインクラスター毎に情報抽出モデルを作成するので、情報抽出モデル毎の特徴を単純化することができ、その結果、情報抽出モデル毎に必要な学習データの数を低減することができる。したがって、本発明の情報抽出プログラムを実行するコンピューターは、情報抽出モデル

の作成のための計算量を低減することができる。
【発明の効果】
【００２０】
本発明の情報抽出システムおよび情報抽出プログラムは、情報抽出モデルの作成のための計算量を低減することができる。
【図面の簡単な説明】
【００２１】
【図１】本発明の一実施の形態に係る情報抽出システムのブロック図である。
【図２】図１に示す記憶部に記憶される情報抽出モデルの一例を示す図である。
【図３】クラスターモデルを作成する場合の図１に示す情報抽出システムの動作のフローチャートである。
【図４】図３に示す動作において学習データの群をメインクラスターに分ける処理のイメージを示す図である。
【図５】図３に示す動作においてメインクラスターからサブクラスターを分離する処理のイメージを示す図である。
【図６】図３に示す動作においてクラスターモデルの作成に使用する学習データを選択する処理のイメージを示す図である。
【図７】請求書データから特定の項目に対する値を抽出する場合の図１に示す情報抽出システムの動作のフローチャートである。
【図８】クラスターモデルを更新する場合の図１に示す情報抽出システムの動作の一部のフローチャートである。
【図９】図８に示す動作の続きの動作のフローチャートである。
【発明を実施するための形態】
【００２２】
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。
【００２３】
まず、本発明の一実施の形態に係る情報抽出システムの構成について説明する。
【００２４】
図１は、本実施の形態に係る情報抽出システム１０のブロック図である。
【００２５】
図１に示すように、情報抽出システム１０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１１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１２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１３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１４と、情報抽出システム１０全体を制御する制御部１５とを備えている。情報抽出システム１０は、例えば、ＰＣ（Ｐｅｒｓｏｎａｌ　Ｃｏｍｐｕｔｅｒ）またはサーバーによって構成されても良いし、プリンター専用機などの画像形成装置によって構成されても良い。
【００２６】
記憶部１４は、文書としての請求書のデータ（以下「請求書データ」という。）から情報を抽出するための情報抽出モデルを使用して請求書データから情報を抽出するための情報抽出プログラム１４ａを記憶している。情報抽出プログラム１４ａは、例えば、情報抽出システム１０の製造段階で情報抽出システム１０にインストールされていても良いし、ＵＳＢ（Ｕｎｉｖｅｒｓａｌ　Ｓｅｒｉａｌ　Ｂｕｓ）メモリーなどの外部の記憶媒体から情報抽出システム１０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抽出システム１０に追加でインストールされても良い。
【００２７】
記憶部１４は、複数のフォーマットの請求書を学習済みの情報抽出モデル（以下「ベー

スモデル」という。）１４ｂを記憶している。ベースモデル１４ｂは、情報抽出システム１０の利用者に情報抽出システム１０を提供する者が用意しても良い。

【００２８】

記憶部１４は、後述のメインクラスター毎の情報抽出モデル（以下「クラスターモデル」という。）１４ｃを記憶可能である。クラスターモデルによる値の抽出の対象の請求書データ（以下「抽出対象データ」という。）は、請求書内の文字と、請求書内の文字以外の素性とを含む請求書データである。請求書内の文字以外の素性には、請求書における各文字の座標が含まれる。また、請求書内の文字以外の素性には、例えば、請求書内の画像と、請求書における各画像の座標とが含まれても良い。請求書内の文字と、請求書における各文字の座標とは、例えば、請求書の画像に対するＯＣＲ（Ｏｐｔｉｃａｌ Ｃｈａｒａｃｔｅｒ Ｒｅｃｏｇｎｉｔｉｏｎ）処理によって取得されることが可能である。請求書内の画像と、請求書における各画像の座標とは、これらを請求書の画像から取得することが可能なシステムによって取得されることが可能である。

【００２９】

記憶部１４は、メインクラスターのクラスタリングの結果（以下「クラスタリング結果」という。）１４ｄを記憶可能である。

【００３０】

制御部１５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１５のＣＰＵの作業領域として用いられるメモリーとしてのＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）とを備えている。制御部１５のＣＰＵは、記憶部１４または制御部１５のＲＯＭに記憶されているプログラムを実行する。

【００３１】

制御部１５は、情報抽出プログラム１４ａを実行することによって、請求書データをクラスタリングする文書クラスタリング部１５ａと、クラスターモデルを作成するモデル学習部１５ｂと、クラスターモデルを使用して請求書データから特定の項目に対する値を抽出するデータ抽出実行部１５ｃとを実現する。

【００３２】

文書クラスタリング部１５ａにおいてクラスタリングに使用されるアルゴリズムとしては、例えば、ＤＢＳＣＡＮ、ｇ－ｍｅａｎｓ、エルボー法など、クラスターの数を自動で決定することが可能なアルゴリズムが採用される。文書クラスタリング部１５ａにおいてクラスタリングに使用される素性としては、例えば、単語ベクトル、単語の座標が採用される。単語ベクトルとしては、例えば、ｏｎｅ－ｈｏｔベクトル、ｔｆ－ｉｄｆ、ｗｏｒｄ２ｖｅｃなどのベクトル表現が採用される。

【００３３】

モデル学習部１５ｂにおいてクラスターモデルの作成に使用されるアルゴリズムとしては、例えば、ＬＳＴＭ、Ｔｒａｎｓｆｏｒｍｅｒなどの自然言語処理を使用したアルゴリズムをベースにしたものが採用される。モデル学習部１５ｂにおいてクラスターモデルの作成に使用される素性としては、例えば、テキスト情報、文字の座標が採用される。

【００３４】

データ抽出実行部１５ｃによって値を抽出する対象の文書には、文書毎に値の記載の位置が異なる場合がない定型文書と、文書毎に値の記載の位置が異なる場合がある準定型文書とが含まれるが、非定型文書は含まれない。

【００３５】

文書クラスタリング部１５ａ、モデル学習部１５ｂおよびデータ抽出実行部１５ｃにおいてデータの距離の計算に使用されるアルゴリズムとしては、例えば、コサイン距離、マンハッタン距離、ユークリッド距離が採用される。

【００３６】

図２は、記憶部１４に記憶される情報抽出モデル２０の一例を示す図である。

【００３７】

図２に示す情報抽出モデル２０は、抽出対象データ４０における「請求書内の文字」に基づいて各文字を取得し（Ｓ２１）、Ｓ２１において取得した各文字に対して、各文字に基づいたベクトル情報を付与し（Ｓ２２）、Ｓ２２の出力をＢｉ－ＬＳＴＭに入力する（Ｓ２３）。

【００３８】

また、情報抽出モデル２０は、抽出対象データ４０における「請求書内の文字」に基づいて各単語を取得し（Ｓ２４）、Ｓ２４において取得した各単語に対して、各単語に基づいたベクトル情報を付与する（Ｓ２５）。

【００３９】

また、情報抽出モデル２０は、抽出対象データ４０における「請求書における各文字の座標」に基づいて各単語の座標を取得し（Ｓ２６）、Ｓ２６において取得した各単語の座標を全結合層に入力する（Ｓ２７）。

【００４０】

そして、情報抽出モデル２０は、Ｓ２３の出力と、Ｓ２５の出力と、Ｓ２７の出力とを連結する（Ｓ２８）。

【００４１】

次いで、情報抽出モデル２０は、Ｓ２８の出力をＢｉ－ＬＳＴＭに入力し（Ｓ２９）、Ｓ２９の出力を全結合層に入力し（Ｓ３０）、Ｓ３０の出力を全結合層に入力し（Ｓ３１）、Ｓ３１の出力をＣＲＦに入力する（Ｓ３２）。

【００４２】

次に、情報抽出システム１０の動作について説明する。

【００４３】

まず、クラスターモデルを作成する場合の情報抽出システム１０の動作について説明する。

【００４４】

図３は、クラスターモデルを作成する場合の情報抽出システム１０の動作のフローチャートである。

【００４５】

利用者は、クラスターモデルの作成のための学習データの群を用意し、用意した学習データの群を使用した学習を、操作部１１から、または、図示していないコンピューターから通信部１３を介して、情報抽出システム１０に指示することができる。ここで、学習データは、請求書内の文字と、請求書内の文字以外の素性と、請求書から抽出されることを利用者が希望する項目に対する正解ラベルとを含む、請求書毎の請求書データである。請求書内の文字以外の素性には、請求書における各文字の座標が含まれる。また、請求書内の文字以外の素性には、例えば、請求書内の画像と、請求書における各画像の座標とが含まれても良い。請求書から抽出されることを利用者が希望する項目とは、例えば、文書が請求書である場合には、請求先、請求日、締切日、請求金額などである。文書から抽出されることを利用者が希望する項目に対する正解ラベルは、請求書内の文字と、請求書内の文字以外の素性とから、利用者によって選択された値である。請求書内の文字と、請求書における各文字の座標とは、例えば、請求書の画像に対するＯＣＲ処理によって取得されることが可能である。請求書内の画像と、請求書における各画像の座標とは、これらを請求書の画像から取得することが可能なシステムによって取得されることが可能である。

【００４６】

情報抽出システム１０の制御部１５は、学習データの群を使用した学習が指示されると、図３に示す動作を実行する。

【００４７】

図３に示すように、文書クラスタリング部１５ａは、学習データの群をクラスタリングすることによって、学習データのそれぞれをいずれかのメインクラスターに分ける（Ｓ１０１）。

【００４８】

図４は、図３に示す動作において学習データの群をメインクラスターに分ける処理のイメージを示す図である。なお、図４（ｂ）において、学習データは、学習データ自身が所属するメインクラスター毎のマークで表示されている。

【００４９】

図４に示すように、文書クラスタリング部１５ａは、学習データの群をクラスタリングするために、学習データの対象の請求書内の文字を学習データ同士で比較することができるように学習データを図４（ａ）に示すようにベクトル化する。

【００５０】

次いで、文書クラスタリング部１５ａは、ベクトル化した学習データの群をクラスタリングすることによって、学習データのそれぞれを図４（ｂ）に示すようにメインクラスターＡ～Ｅのいずれかに分ける（Ｓ１０１）。

【００５１】

図３に示すように、制御部１５は、Ｓ１０１の処理の後、図３に示す動作の今回の実行において未だＳ１０３の処理の対象にしていないメインクラスターの１つを対象にする（Ｓ１０２）。

【００５２】

次いで、文書クラスタリング部１５ａは、現在の対象のメインクラスターにおけるサブクラスターの最適数（以下「サブクラスター最適数」という。）をクラスター数自動推定法によって確認する（Ｓ１０３）。

【００５３】

次いで、文書クラスタリング部１５ａは、Ｓ１０３において確認したサブクラスター最適数が、サブクラスターの上限数（以下「サブクラスター上限数」という。）以下であるか否かを判断する（Ｓ１０４）。ここで、サブクラスター上限数は、本実施の形態において例えば５である。

【００５４】

文書クラスタリング部１５ａは、Ｓ１０３において確認したサブクラスター最適数がサブクラスター上限数以下ではないとＳ１０４において判断すると、Ｓ１０３において確認したサブクラスター最適数からサブクラスター上限数を差し引いた数のサブクラスターを現在の対象のメインクラスターから分離する（Ｓ１０５）。ここで、文書クラスタリング部１５ａは、重心が現在の対象のメインクラスターの重心から遠いサブクラスターを優先して現在の対象のメインクラスターから分離する。なお、メインクラスターの重心は、例えば、このメインクラスターに所属する学習データの文書ベクトルの平均値である。同様に、サブクラスターの重心は、例えば、このサブクラスターに所属する学習データの文書ベクトルの平均値である。

【００５５】

文書クラスタリング部１５ａは、Ｓ１０５の処理の後、Ｓ１０５において現在の対象のメインクラスターから分離したサブクラスターによって新たにメインクラスターを生成する（Ｓ１０６）。すなわち、文書クラスタリング部１５ａは、Ｓ１０５において現在の対象のメインクラスターから分離したサブクラスターを新たなメインクラスターにする。

【００５６】

図５は、図３に示す動作においてメインクラスターからサブクラスターを分離する処理のイメージを示す図である。なお、図５は、図４（ｂ）に示すメインクラスターＢの例である。図５（ａ）、（ｂ）において、学習データは、学習データ自身が所属するサブクラスター毎のマークで表示されている。図５（ｃ）において、学習データは、学習データ自身が所属するメインクラスター毎のマークで表示されている。

【００５７】

図５（ａ）に示すように、文書クラスタリング部１５ａは、メインクラスターＢにおけるサブクラスター最適数を確認する（Ｓ１０３）。図５（ａ）に示す例では、文書クラスタリング部１５ａは、メインクラスターＢにおけるサブクラスター最適数をクラスター数自動推定法によって７と確認している。

【００５８】
次いで、文書クラスタリング部１５ａは、Ｓ１０３において確認したサブクラスター最適数がサブクラスター上限数以下ではない場合に（Ｓ１０４でＮＯ）、Ｓ１０３において確認したサブクラスター最適数からサブクラスター上限数を差し引いた数のサブクラスターを図５（ｂ）に示すようにメインクラスターＢから分離する（Ｓ１０５）。すなわち、文書クラスタリング部１５ａは、サブクラスターＦ、ＧをメインクラスターＢから分離する。図５（ｂ）に示す例は、サブクラスター上限数が５の場合の例である。
【００５９】
文書クラスタリング部１５ａは、Ｓ１０５の処理の後、Ｓ１０５においてメインクラスターＢから分離したサブクラスターＦ、Ｇを図５（ｃ）に示すようにそれぞれ新たなメインクラスターＦ、Ｇにする（Ｓ１０６）。
【００６０】
図３に示すように、文書クラスタリング部１５ａは、Ｓ１０３において確認した最適数がサブクラスター上限数以下であるとＳ１０４において判断するか、Ｓ１０６の処理が終了すると、現在の対象のメインクラスター内の学習データの群をサブクラスター最適数でクラスタリングすることによって、現在の対象のメインクラスター内の学習データのそれぞれをいずれかのサブクラスターに分ける（Ｓ１０７）。
【００６１】
次いで、モデル学習部１５ｂは、現在の対象のメインクラスター内のサブクラスターから、クラスターモデルの作成に使用する学習データを選択する（Ｓ１０８）。ここで、モデル学習部１５ｂは、現在の対象のメインクラスター内のサブクラスターのうち、重心が現在の対象のメインクラスターの重心に最も近いサブクラスターにおいて、重心が現在の対象のメインクラスターの重心に最も近い学習データを、クラスターモデルの作成に使用する学習データとして選択する。また、モデル学習部１５ｂは、現在の対象のメインクラスター内のサブクラスターのうち、重心が現在の対象のメインクラスターの重心に最も近いサブクラスター以外のサブクラスターのそれぞれにおいて、重心が現在の対象のメインクラスターの重心から最も遠い学習データを、クラスターモデルの作成に使用する学習データとして選択する。なお、学習データの重心は、例えば、この学習データの文書ベクトルである。
【００６２】
図６は、図３に示す動作においてクラスターモデルの作成に使用する学習データを選択する処理のイメージを示す図である。なお、図６は、図５（ｃ）に示すメインクラスターＢの例である。なお、図６において、学習データは、学習データ自身が所属するサブクラスター毎のマークで表示されている。
【００６３】
図６に示すように、モデル学習部１５ｂは、メインクラスターＢ内のサブクラスターのうち、重心がメインクラスターＢの重心に最も近いサブクラスターＤにおいて、重心がメインクラスターＢの重心に最も近い学習データをクラスターモデルの作成に使用する学習データとして選択するとともに、メインクラスターＢ内のサブクラスターのうち、サブクラスターＤ以外のサブクラスターのそれぞれにおいて、重心がメインクラスターＢの重心から最も遠い学習データをクラスターモデルの作成に使用する学習データとして選択する（Ｓ１０８）。なお、図６において、右上にチェックマークが付されている学習データが、クラスターモデルの作成に使用する学習データとして選択されたものである。
【００６４】
図３に示すように、モデル学習部１５ｂは、Ｓ１０８の処理の後、Ｓ１０８において選択した学習データを使用して学習を実行することによって、現在の対象のメインクラスター用のクラスターモデルを作成する（Ｓ１０９）。ここで、モデル学習部１５ｂは、ベースモデル１４ｂを基にしてクラスターモデルを作成する。
【００６５】
文書クラスタリング部１５ａは、Ｓ１０９の処理の後、図３に示す動作の今回の実行に

おいて未だＳ１０３の処理の対象にしていないメインクラスターが存在する場合には、図３に示す動作の今回の実行において未だＳ１０３の処理の対象にしていないメインクラスターの１つを対象にして（Ｓ１１０）、Ｓ１０３の処理を実行する。

【００６６】

モデル学習部１５ｂは、Ｓ１０９の処理の後、図３に示す動作の今回の実行において未だＳ１０３の処理の対象にしていないメインクラスターが存在しない場合には、図３に示す動作の今回の実行において新たに作成した全てのクラスターモデルを記憶部１４に保存する（Ｓ１１１）。

【００６７】

次いで、文書クラスタリング部１５ａは、図３に示す動作におけるメインクラスターのクラスタリングの結果をクラスタリング結果１４ｄに保存して（Ｓ１１２）、図３に示す動作を終了する。

【００６８】

次に、請求書データから特定の項目に対する値を抽出する場合の情報抽出システム１０の動作について説明する。

【００６９】

図７は、請求書データから特定の項目に対する値を抽出する場合の情報抽出システム１０の動作のフローチャートである。

【００７０】

利用者は、抽出対象データを用意し、用意した抽出対象データからの特定の項目に対する値の抽出を、操作部１１から、または、図示していないコンピューターから通信部１３を介して、情報抽出システム１０に指示することができる。ここで、特定の項目とは、クラスターモデルの作成時に使用された学習データにおいて正解ラベルに対する項目、すなわち、請求書から抽出されることを利用者が希望する項目である。

【００７１】

情報抽出システム１０の制御部１５は、抽出対象データからの特定の項目に対する値の抽出が指示されると、図７に示す動作を実行する。

【００７２】

図７に示すように、文書クラスタリング部１５ａは、クラスタリング結果１４ｄを使用して、抽出対象データが所属するメインクラスターを判定する（Ｓ１２１）。

【００７３】

データ抽出実行部１５ｃは、Ｓ１２１の処理の後、抽出対象データが所属するメインクラスターがＳ１２１において特定されたか否かを判断する（Ｓ１２２）。

【００７４】

データ抽出実行部１５ｃは、抽出対象データが所属するメインクラスターがＳ１２１において特定されたとＳ１２２において判断すると、抽出対象データが所属するとＳ１２１において特定されたメインクラスター用のクラスターモデルを使用して請求書データから特定の項目に対する値を抽出して（Ｓ１２３）、図７に示す動作を終了する。

【００７５】

データ抽出実行部１５ｃは、抽出対象データが所属するメインクラスターがＳ１２１において特定されなかった、すなわち、抽出対象データがいずれのメインクラスターにも所属しない外れ値であるとＳ１２２において判断すると、抽出対象データに適合するクラスターモデルが存在しないことを利用者に通知する（Ｓ１２４）。ここで、利用者への通知の方法としては、例えば、抽出対象データからの特定の項目に対する値の抽出が操作部１１から指示された場合には、表示部１２における表示でも良いし、抽出対象データからの特定の項目に対する値の抽出が図示していないコンピューターから通信部１３を介して指示された場合には、通信部１３を介した、このコンピューターへの出力でも良い。

【００７６】

データ抽出実行部１５ｃは、Ｓ１２４の処理の後、抽出対象データに最も近いメインクラスター用のクラスターモデルを使用して抽出対象データから特定の項目に対する値を抽

出して（Ｓ１２５）、図７に示す動作を終了する。

【００７７】

なお、Ｓ１２３またはＳ１２５において抽出された値は、様々な用途に活用されることが可能である。例えば、Ｓ１２３またはＳ１２５において抽出された値は、抽出対象データの基になった請求書の画像ファイルのファイル名に使用されても良い。

【００７８】

次に、クラスターモデルを更新する場合の情報抽出システム１０の動作について説明する。

【００７９】

図８は、クラスターモデルを更新する場合の情報抽出システム１０の動作の一部のフローチャートである。図９は、図８に示す動作の続きの動作のフローチャートである。

【００８０】

利用者は、クラスターモデルの更新のための学習データ（以下「追加データ」という。）を用意し、用意した追加データを使用した学習を、操作部１１から、または、図示していないコンピューターから通信部１３を介して、情報抽出システム１０に指示することができる。ここで、利用者は、例えば、クラスターモデルを使用して抽出された値が適切ではなかった請求書データに、正解ラベルを付与することによって、追加データとしても良い。

【００８１】

情報抽出システム１０の制御部１５は、追加データを使用した学習が指示されると、図８および図９に示す動作を実行する。

【００８２】

図８および図９に示すように、文書クラスタリング部１５ａは、クラスタリング結果１４ｄを使用して、追加データが所属するメインクラスターを判定する（Ｓ１４１）。

【００８３】

文書クラスタリング部１５ａは、Ｓ１４１の処理の後、追加データが所属するメインクラスターがＳ１４１において特定されたか否かを判断する（Ｓ１４２）。

【００８４】

文書クラスタリング部１５ａは、追加データが所属するメインクラスターがＳ１４１において特定されたとＳ１４２において判断すると、追加データが所属するとＳ１４１において特定されたメインクラスターに追加データを追加する（Ｓ１４３）。

【００８５】

次いで、文書クラスタリング部１５ａは、追加データが所属するとＳ１４１において特定されたメインクラスターを対象にする（Ｓ１４４）。

【００８６】

次いで、文書クラスタリング部１５ａは、現在の対象のメインクラスターにおけるサブクラスター最適数をクラスター数自動推定法によって確認する（Ｓ１４５）。

【００８７】

次いで、文書クラスタリング部１５ａは、Ｓ１４５において確認したサブクラスター最適数がサブクラスター上限数以下であるか否かを判断する（Ｓ１４６）。

【００８８】

文書クラスタリング部１５ａは、Ｓ１４６の処理の後、Ｓ１４５において確認したサブクラスター最適数がサブクラスター上限数以下ではないとＳ１４６において判断すると、Ｓ１４５において確認したサブクラスター最適数からサブクラスター上限数を差し引いた数のサブクラスターを現在の対象のメインクラスターから分離する（Ｓ１４７）。ここで、文書クラスタリング部１５ａは、重心が現在の対象のメインクラスターの重心から遠いサブクラスターを優先して現在の対象のメインクラスターから分離する。

【００８９】

文書クラスタリング部１５ａは、Ｓ１４７の処理の後、Ｓ１４７において現在の対象のメインクラスターから分離したサブクラスターによって新たにメインクラスターを生成す

る（Ｓ１４８）。すなわち、文書クラスタリング部１５ａは、Ｓ１４７において現在の対象のメインクラスターから分離したサブクラスターを新たなメインクラスターにする。

【００９０】

文書クラスタリング部１５ａは、Ｓ１４５において確認した最適数がサブクラスター上限数以下であるとＳ１４６において判断するか、Ｓ１４８の処理が終了すると、現在の対象のメインクラスター内の学習データの群をサブクラスター最適数でクラスタリングすることによって、現在の対象のメインクラスター内の学習データのそれぞれをいずれかのサブクラスターに分ける（Ｓ１４９）。

【００９１】

次いで、モデル学習部１５ｂは、現在の対象のメインクラスター内のサブクラスターから、クラスターモデルの作成に使用する学習データを選択する（Ｓ１５０）。ここで、モデル学習部１５ｂは、現在の対象のメインクラスター内のサブクラスターのうち、重心が現在の対象のメインクラスターの重心に最も近いサブクラスターにおいて、重心が現在の対象のメインクラスターの重心に最も近い学習データを、クラスターモデルの作成に使用する学習データとして選択する。また、モデル学習部１５ｂは、現在の対象のメインクラスター内のサブクラスターのうち、重心が現在の対象のメインクラスターの重心に最も近いサブクラスター以外のサブクラスターのそれぞれにおいて、重心が現在の対象のメインクラスターの重心から最も遠い学習データを、クラスターモデルの作成に使用する学習データとして選択する。

【００９２】

モデル学習部１５ｂは、Ｓ１５０の処理の後、Ｓ１５０において選択された学習データを使用して学習を実行することによって、現在の対象のメインクラスター用のクラスターモデルを作成する（Ｓ１５１）。ここで、モデル学習部１５ｂは、ベースモデル１４ｂを基にしてクラスターモデルを作成する。

【００９３】

文書クラスタリング部１５ａは、Ｓ１５１の処理の後、図８および図９に示す動作の今回の実行において新たに生成したメインクラスターに、図８および図９に示す動作の今回の実行において未だＳ１４５の処理の対象にしていないメインクラスターが存在する場合には、図８および図９に示す動作の今回の実行において新たに生成したメインクラスターのうち、図８および図９に示す動作の今回の実行において未だＳ１４５の処理の対象にしていないメインクラスターの１つを対象にして（Ｓ１５２）、Ｓ１４５の処理を実行する。

【００９４】

データ抽出実行部１５ｃは、Ｓ１５１の処理の後、図８および図９に示す動作の今回の実行において新たに生成したメインクラスターに、図８および図９に示す動作の今回の実行において未だＳ１４５の処理の対象にしていないメインクラスターが存在しない場合には、図８および図９に示す動作の今回の実行において新たに作成した全てのクラスターモデルが、クラスターモデル自身の対象のメインクラスターに含まれる全ての学習データに対して特定の程度以上に高い精度で特定の項目に対する値を抽出することができるか否かを判断する（Ｓ１５３）。ここで、データ抽出実行部１５ｃは、高い精度で特定の項目に対する値を抽出することができるか否かを、利用者によって判定されても良いし、データ抽出実行部１５ｃ自身が精度の閾値に基づいて自動で判定しても良い。

【００９５】

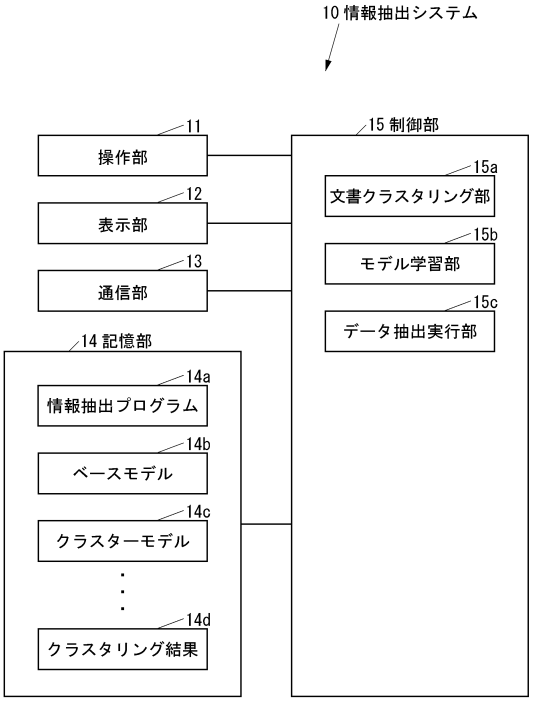
モデル学習部１５ｂは、図８および図９に示す動作の今回の実行において新たに作成した全てのクラスターモデルが、クラスターモデル自身の対象のメインクラスターに含まれる全ての学習データに対して特定の程度以上に高い精度で特定の項目に対する値を抽出することができるとＳ１５３において判断すると、追加データが所属するとＳ１４１において特定されたメインクラスター用のクラスターモデルを記憶部１４から削除し（Ｓ１５４）、図８および図９に示す動作の今回の実行において新たに作成した全てのクラスターモデルを記憶部１４に保存する（Ｓ１５５）。

【００９６】
文書クラスタリング部１５ａは、図８および図９に示す動作の今回の実行において新たに作成したいずれかのクラスターモデルが、クラスターモデル自身の対象のメインクラスターに含まれるいずれかの学習データに対して特定の程度以上に高い精度で特定の項目に対する値を抽出することができないとＳ１５３において判断すると、図８および図９に示す動作の今回の実行におけるこれまでのクラスタリングの結果を全て廃棄する（Ｓ１５６）。したがって、文書クラスタリング部１５ａは、追加データが現在所属するメインクラスターから追加データを分離する。
【００９７】
文書クラスタリング部１５ａは、追加データが所属するメインクラスターがＳ１４１において特定されなかった、すなわち、追加データがいずれのメインクラスターにも所属しない外れ値であるとＳ１４２において判断するか、Ｓ１５６の処理が終了すると、追加データによって新たにメインクラスターを生成する（Ｓ１５７）。
【００９８】
モデル学習部１５ｂは、Ｓ１５７の処理の後、追加データを使用して学習を実行することによって、追加データが所属するメインクラスター用のクラスターモデルを作成する（Ｓ１５８）。ここで、モデル学習部１５ｂは、ベースモデル１４ｂを基にしてクラスターモデルを作成する。
【００９９】
モデル学習部１５ｂは、Ｓ１５８の処理の後、Ｓ１５８において新たに作成したクラスターモデルを記憶部１４に保存する（Ｓ１５９）。
【０１００】
文書クラスタリング部１５ａは、Ｓ１５５またはＳ１５９の処理の後、図８および図９に示す動作におけるメインクラスターのクラスタリングの結果をクラスタリング結果１４ｄに保存して（Ｓ１６０）、図８および図９に示す動作を終了する。
【０１０１】
以上に説明したように、情報抽出システム１０は、メインクラスター毎に情報抽出モデルとしてのクラスターモデルを作成する（Ｓ１０９、Ｓ１５１およびＳ１５８）ので、クラスターモデル毎の特徴を単純化することができ、その結果、クラスターモデル毎に必要な学習データの数を低減することができる。したがって、情報抽出システム１０は、クラスターモデルの作成のための計算量を低減することができる。
【０１０２】
情報抽出システム１０は、クラスターモデルの作成に使用する学習データをサブクラスター毎に選択し（Ｓ１０８およびＳ１５０）、選択した学習データを使用して学習を実行することによって、メインクラスター毎のクラスターモデルを作成する（Ｓ１０９およびＳ１５１）ので、クラスターモデル毎に必要な学習データの数を低減することができ、その結果、クラスターモデルの作成のための計算量を低減することができる。
【０１０３】
情報抽出システム１０は、重心がメインクラスターの重心に最も近いサブクラスターにおいて、重心がメインクラスターの重心に最も近い学習データを、クラスターモデルの作成に使用する学習データとして選択する（Ｓ１０８およびＳ１５０）ので、メインクラスターの特徴を最も強く表す学習データを使用してクラスターモデルを作成することができ、その結果、メインクラスターの特徴が適切に反映されたクラスターモデルを作成することができる。
【０１０４】
情報抽出システム１０は、重心がメインクラスターの重心に最も近いサブクラスター以外のサブクラスターのそれぞれにおいて、重心がメインクラスターの重心から最も遠い学習データを、クラスターモデルの作成に使用する学習データとして選択する（Ｓ１０８およびＳ１５０）ので、メインクラスターにおいて広範囲に散らばった学習データを使用してクラスターモデルを作成することができ、その結果、メインクラスターの特徴が適切に

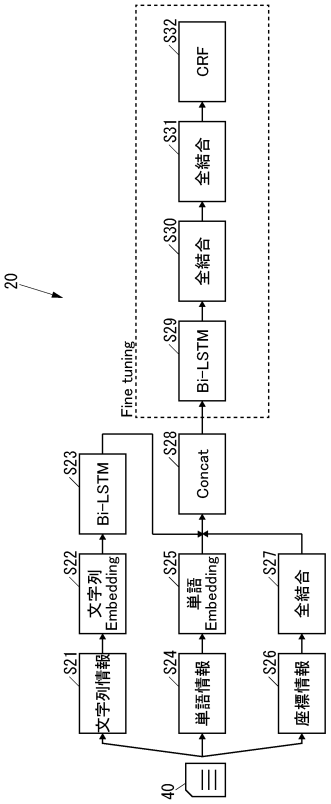
反映されたクラスターモデルを作成することができる。
【０１０５】
情報抽出システム１０は、メインクラスターにおけるサブクラスター最適数がサブクラスター上限数を超える場合に、サブクラスター最適数からサブクラスター上限数を差し引いた数のサブクラスターを、このメインクラスターから分離する（Ｓ１０５およびＳ１４７）ので、クラスターモデル毎に必要な学習データの数を低減することができ、その結果、クラスターモデルの作成のための計算量を低減することができる。
【０１０６】
情報抽出システム１０は、クラスター最適数からクラスター上限数を差し引いた数のサブクラスターをメインクラスターから分離する場合に、重心がこのメインクラスターの重心から遠いサブクラスターを優先して、このメインクラスターから分離する（Ｓ１０５およびＳ１４７）ので、メインクラスターの特徴を強く表す学習データを使用して情報抽出モデルを作成することができ、その結果、メインクラスターの特徴が適切に反映された情報抽出モデルを作成することができる。
【０１０７】
情報抽出システム１０は、クラスターモデルの作成のための計算量を低減することができるので、例えば、一般的なＰＣの計算リソースでも深層学習の学習処理を実行することができる。したがって、情報抽出システム１０は、情報を抽出する対象の文書が、例えば個人情報や取引情報など、保護すべき情報が含まれる、例えば請求書などの文書である場合に、文書のデータをローカル環境外にアップロードすることなく、ローカル環境内の一般的なＰＣでクラスターモデルを作成することができる。
【０１０８】
以上においては、モデル学習部１５ｂは、クラスターモデルを更新する場合に、ベースモデル１４ｂを基にしてクラスターモデルを作成する。しかしながら、モデル学習部１５ｂは、クラスターモデルを更新する場合に、更新の対象のクラスターモデルが記憶部１４に既に記憶されている場合には、更新の対象のクラスターモデルを基にして新たなクラスターモデルを作成しても良い。
【０１０９】
以上においては、情報抽出システム１０は、請求書のデータから情報を抽出する。しかしながら、情報抽出システム１０は、請求書の場合と同様にして、例えば答案用紙など、請求書以外の種類の文書のデータから情報を抽出することが可能である。なお、情報抽出システム１０は、文書の種類毎のベースモデルを使用しても良いし、複数の種類の文書に共通のベースモデルを使用しても良い。ここで、情報抽出システム１０は、文書の種類毎のベースモデルを使用する方が、複数の種類の文書に共通のベースモデルを使用するよりも、情報の抽出の精度を向上することができる。しかしながら、情報抽出システム１０は、複数の種類の文書に共通のベースモデルを使用する方が、文書の種類毎のベースモデルを使用するよりも、ベースモデルの用意の労力を低減することができる。
【符号の説明】
【０１１０】
１０　情報抽出システム
１４ａ　情報抽出プログラム
１４ｃ　クラスターモデル（情報抽出モデル）
１５ａ　文書クラスタリング部
１５ｂ　モデル学習部
２０　情報抽出モデル

【図面】

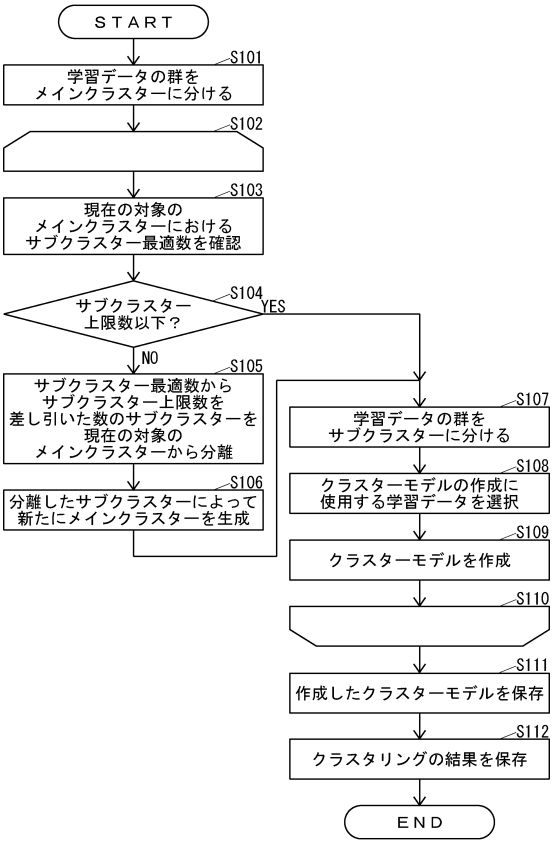
【図１】


10 情報抽出システム
11
操作部
12
表示部
13
通信部
14 記憶部
14a
情報抽出プログラム
14b
ベースモデル
14c
クラスターモデル
14d
クラスタリング結果
15 制御部
15a
文書クラスタリング部
15b
モデル学習部
15c
データ抽出実行部

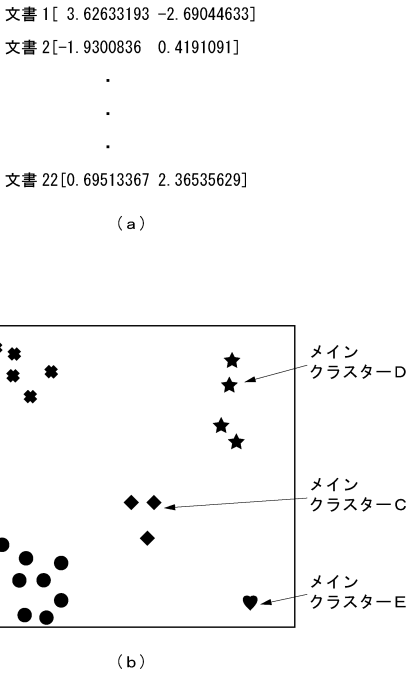
【図２】


20
Fine tuning
S32
CRF
S31
全結合
S30
全結合
S29
Bi-LSTM
S23
Bi-LSTM
S28
Concat
S22
文字列 Embedding
S25
単語 Embedding
S27
全結合
S21
文字列情報
S24
単語情報
S26
座標情報
40

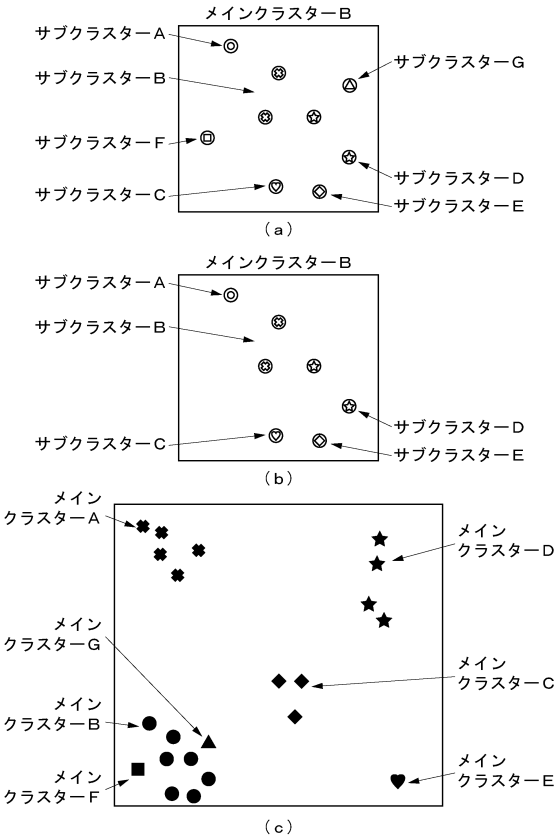
【図３】


START
S101
学習データの群を
メインクラスターに分ける
S102
S103
現在の対象の
メインクラスターにおける
サブクラスター最適数を確認
S104
サブクラスター
上限数以下？
YES
NO
S105
サブクラスター最適数から
サブクラスター上限数を
差し引いた数のサブクラスターを
現在の対象の
メインクラスターから分離
S106
分離したサブクラスターによって
新たにメインクラスターを生成
S107
学習データの群を
サブクラスターに分ける
S108
クラスターモデルの作成に
使用する学習データを選択
S109
クラスターモデルを作成
S110
S111
作成したクラスターモデルを保存
S112
クラスタリングの結果を保存
END

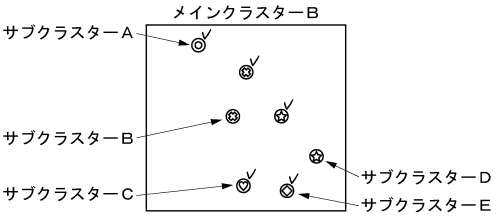
【図４】


文書 1[3.62633193 -2.69044633]
文書 2[-1.9300836 0.4191091]
文書 22[0.69513367 2.36535629]
（a）
メイン
クラスターA
メイン
クラスターD
メイン
クラスターC
メイン
クラスターB
メイン
クラスターE
（b）

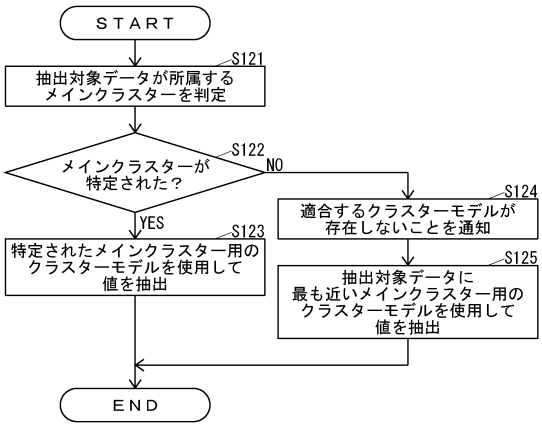
【図５】


メインクラスターB
サブクラスターA
サブクラスターB
サブクラスターF
サブクラスターC
サブクラスターG
サブクラスターD
サブクラスターE
（a）
メインクラスターB
サブクラスターA
サブクラスターB
サブクラスターC
サブクラスターD
サブクラスターE
（b）
メイン
クラスターA
メイン
クラスターD
メイン
クラスターG
メイン
クラスターC
メイン
クラスターB
メイン
クラスターF
メイン
クラスターE
（c）

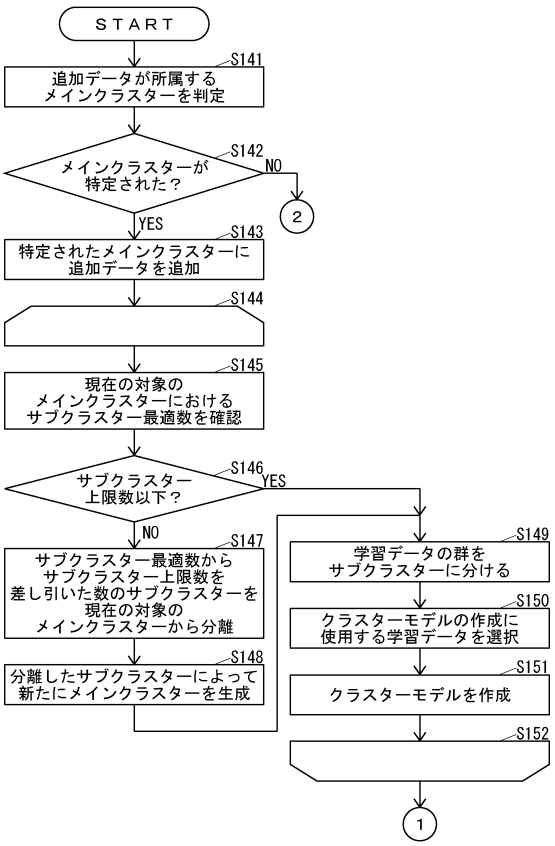
【図６】


メインクラスターB
サブクラスターA
サブクラスターB
サブクラスターC
サブクラスターD
サブクラスターE

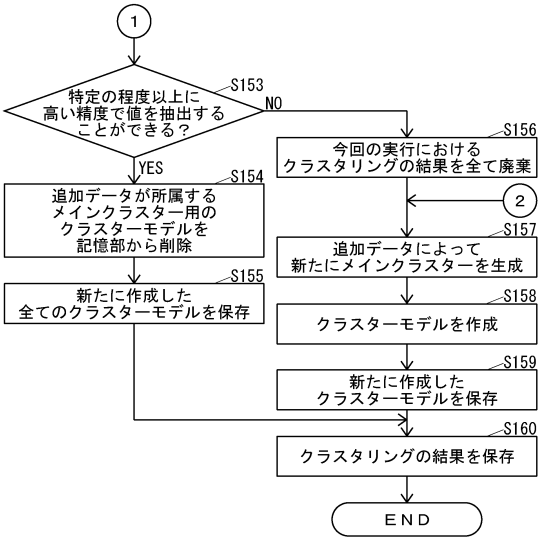
【図７】


START
抽出対象データが所属する
メインクラスターを判定
S121
メインクラスターが
特定された？
S122
NO
YES
適合するクラスターモデルが
存在しないことを通知
S124
特定されたメインクラスター用の
クラスターモデルを使用して
値を抽出
S123
抽出対象データに
最も近いメインクラスター用の
クラスターモデルを使用して
値を抽出
S125
END

【図８】


START
追加データが所属する
メインクラスターを判定
S141
メインクラスターが
特定された？
S142
NO
2
YES
特定されたメインクラスターに
追加データを追加
S143
S144
現在の対象の
メインクラスターにおける
サブクラスター最適数を確認
S145
サブクラスター
上限数以下？
S146
YES
NO
サブクラスター最適数から
サブクラスター上限数を
差し引いた数のサブクラスターを
現在の対象の
メインクラスターから分離
S147
分離したサブクラスターによって
新たにメインクラスターを生成
S148
学習データの群を
サブクラスターに分ける
S149
クラスターモデルの作成に
使用する学習データを選択
S150
クラスターモデルを作成
S151
S152
1

【図９】


1
S153
特定の程度以上に
高い精度で値を抽出する
ことができる？
NO
YES
S154
追加データが所属する
メインクラスター用の
クラスターモデルを
記憶部から削除
S155
新たに作成した
全てのクラスターモデルを保存
S156
今回の実行における
クラスタリングの結果を全て廃棄
2
S157
追加データによって
新たにメインクラスターを生成
S158
クラスターモデルを作成
S159
新たに作成した
クラスターモデルを保存
S160
クラスタリングの結果を保存
END