(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2020年1月30日(30.01.2020)

WIPO | PCT

(10) 国際公開番号
WO 2020/021845 A1

(51) 国際特許分類:
G06F 16/35 (2019.01)

(21) 国際出願番号： PCT/JP2019/021289

(22) 国際出願日： 2019年5月29日(29.05.2019)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(30) 優先権データ：
特願 2018-138453 2018年7月24日(24.07.2018) JP

(71) 出願人:株式会社ＮＴＴドコモ(**NTT DOCOMO, INC.**) [JP/JP]; 〒1006150 東京都千代田区永田町二丁目１１番１号 Tokyo (JP).

(72) 発明者: 池田 大志 (**IKEDA Taishi**); 〒1006150 東京都千代田区永田町二丁目１１番１号 山王パークタワー 株式会社ＮＴＴドコモ 知的財産部内 Tokyo (JP).

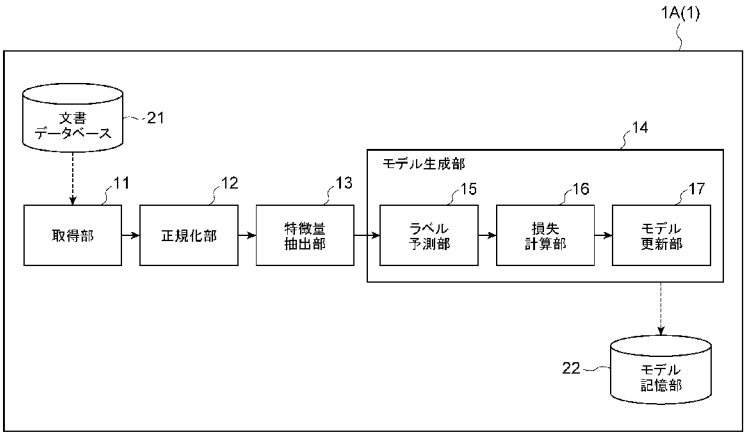
(74) 代理人: 長谷川 芳樹, 外 (**HASEGAWA Yoshiki** et al.); 〒1000005 東京都千代田区丸の内二丁目１番１号丸の内 ＭＹ ＰＬＡＺＡ（明治安田生命ビル） ９階 創英国際特許法律事務所 Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,

(54) **Title:** DOCUMENT CLASSIFICATION DEVICE AND TRAINED MODEL

(54) 発明の名称： 文書分類装置及び学習済みモデル

[図1]



11 Acquisition unit
12 Normalization unit
13 Feature quantity extraction unit
14 Model generation unit
15 Label prediction unit
16 Loss calculation unit
17 Model update unit
21 Document database
22 Model storage unit

(57) **Abstract:** A document classification device for generating, through machine learning, a document classification model, which is a model for classifying documents and which, on the basis of an input document, outputs identification information identifying a classification result, said document classification device being provided with: an acquisition unit which acquires learning data including a document and identification information associated with the document; a feature quantity extraction unit which extracts, as feature quantities, words included in the document and character information

WO 2020/021845 A1

［続葉有］

HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告 (条約第21条(3))

comprising one or a plurality of items of information extractable from the words, the character information comprising one of the characters constituting a word or a plurality of successive characters in the word; and a model generation unit which performs machine learning on the basis of the feature quantities extracted from the document and the identification information associated with the document, and generates a document classification model.

(57) 要約:文書分類装置は、文書を分類するためのモデルであって、入力された文書に基づいて、分類の結果を識別する識別情報を出力する文書分類モデルを機械学習により生成する装置であって、文書と該文書に関連付けられた識別情報とを含む学習データを取得する取得部と、文書に含まれる単語と、単語を構成する文字のうちの一の文字または単語において連続する複数の文字からなる文字列であって、前記単語から1または複数抽出可能な情報である文字情報と、を特徴量として抽出する特徴量抽出部と、文書から抽出された特徴量及び当該文書に関連付けられた識別情報に基づいて機械学習を行い、文書分類モデルを生成するモデル生成部と、を備える。

明　細　書

発明の名称 ： 文書分類装置及び学習済みモデル

技術分野

[0001] 本発明は、文書分類装置及び学習済みモデルに関する。

背景技術

[0002] 機械学習により生成されたモデルを用いて文書を分類する技術が知られている。このような技術においては、学習用の文書に対して形態素解析を行うことにより単語を抽出し、抽出した単語を特徴量として機械学習に供することが一般的である。

[0003] 非特許文献１には、機械学習手法の一種であるフィードフォワードニューラルネットワークを用いて単語単位の特徴量に基づいて文書分類モデルを生成し、文書分類を行うことが記載されている。

[0004] 特許文献１には、所定の照合条件に該当する文字列を特徴トークンとして文書から抽出し、特徴トークンとして抽出されなかった文字列を分割した文字を非特徴トークンとして抽出して、特徴トークン及び非特徴トークンを文書分類のための学習に用いる技術が記載されている。

[0005] 特許文献２には、語の表記揺れに対応するための辞書を作成する技術が記載されている。

先行技術文献

特許文献

[0006] 特許文献1：特開２００９－９８９５２号公報

特許文献2：特開２０１６－１３９１６４号公報

非特許文献

[0007] 非特許文献1：Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber and Hal Daume III. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joi

nt Conference on Natural Language Processing, pages 1681-1691.

発明の概要

発明が解決しようとする課題

[0008] 分類対象として入力された文書において、例えば、チャットボットを利用するユーザーによるタイプミス及び誤認識等に起因して、単語に誤字及び脱字等が含まれる場合がある。入力された文書に誤字及び脱字等が含まれていると、分類精度の低下を招くこととなる。単語の表記揺れに対しては、特許文献２に記載されたような、一の単語に対して複数の表記を関連付けた辞書を用いることにより対応することができる。しかし、このような辞書を予め作成することにはコストがかかる。

[0009] 特許文献１に記載された技術では、ある単語に関して、表記が揺れた複数のバリエーションを学習に用いるわけではないので、誤字及び脱字等を含む文書が入力された場合の分類精度の低下を防止できない。

[0010] そこで、本発明は、上記問題点に鑑みてなされたものであり、機械学習により生成されるモデルを用いた文書の分類において、誤字及び脱字等を含む文書の分類の精度の低下を防止可能な文書分類装置及び学習済モデル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

[0011] 上記課題を解決するために、本発明の一形態に係る文書分類装置は、文書を分類するためのモデルであって、入力された文書に基づいて、分類の結果を識別する識別情報を出力する文書分類モデルを機械学習により生成する文書分類装置であって、文書と該文書に関連付けられた識別情報とを含む学習データを取得する取得部と、文書に含まれる単語と、当該単語を構成する文字のうちの一の文字または当該単語において連続する複数の文字からなる文字列であって、当該単語から１または複数抽出可能な情報である文字情報と、を特徴量として抽出する特徴量抽出部と、文書から抽出された特徴量及び当該文書に関連付けられた識別情報に基づいて機械学習を行い、文書分類モデルを生成するモデル生成部と、を備える。

[0012] 上記の形態によれば、文書分類モデルの生成のための学習用の文書から抽出された単語と共に、当該単語に含まれる文字からなる文字情報が特徴量として用いられて文書分類モデルが生成されるので、単語に含まれる文字単位の情報も文書分類モデルに反映される。これにより、分類対象の文書に含まれる単語に誤字及び脱字等が含まれている場合であっても、その単語に含まれる文字も特徴量として文書分類モデルに入力することにより、文書に含まれる単語を構成する文字に関する特徴が分類に加味される。従って、文書の分類の精度の低下を防止できる。

[0013] 上記課題を解決するために、本発明の一形態に係る学習済モデルは、入力された文書を分類し、分類の結果を識別する識別情報を出力するよう、コンピュータを機能させるための学習済みモデルであって、文書に含まれる単語と、当該単語を構成する文字のうちの一の文字または当該単語において連続する複数の文字からなる文字列であって、当該単語から１または複数抽出可能な情報である文字情報と、を含む特徴量を入力値とし、文書に関連付けられた識別情報を出力値とする学習データに基づく機械学習により構成される。

[0014] 上記の形態によれば、分類対象の文章から抽出された単語だけではなく、当該単語に含まれる文字からなる文字情報を含む特徴量が学習データとして用いられた機械学習により、文書を分類するための学習済みモデルが構成される。これにより、分類対象の文書に含まれる単語に誤字及び脱字等が含まれている場合であっても、その単語に含まれる文字も特徴量として学習済みモデルに入力することにより、文書に含まれる単語を構成する文字に関する特徴が分類に加味される。従って、文書の分類の精度の低下を防止できる。

発明の効果

[0015] 機械学習により生成されるモデルを用いた文書の分類において、誤字及び脱字等を含む文書の分類の精度の低下を防止可能な文書分類装置及び学習済モデルを提供することが可能となる。

図面の簡単な説明

[0016] [図1]文書分類モデルの機械学習による生成の局面における本実施形態に係る文書分類装置の機能的構成を示す機能ブロック図である。

[図2]文書分類モデルを用いた文書の分類（予測）の局面における本実施形態に係る文書分類装置の機能的構成を示す機能ブロック図である。

[図3]需要予測システムのハードブロック図である。

[図4]文書データベースに記憶されているラベル付き文書の例を示す図である。

[図5]正規化部による文書の正規化から特徴量抽出部による特徴量の抽出に至る処理の具体例を示す図である。

[図6]本実施形態の生成方法の処理内容を示すフローチャートである。

[図7]本実施形態の分類方法の処理内容を示すフローチャートである。

[図8]文書分類プログラムの構成を示す図である。

発明を実施するための形態

[0017] 本発明に係る文書分類装置及び学習済モデルの実施形態について図面を参照して説明する。なお、可能な場合には、同一の部分には同一の符号を付して、重複する説明を省略する。

[0018] 図１及び図２は、本実施形態に係る文書分類装置１の機能的構成を示す図である。文書分類装置１は、入力された文書に基づいて、文書の分類の結果を識別する識別情報を出力する文書分類モデルを機械学習により生成し、生成された文書分類モデルを用いて文書を分類し、分類の結果を示す識別情報を出力する。

[0019] 図１は、文書を分類するためのモデルである文書分類モデルの機械学習による生成の局面における文書分類装置１Ａ（１）の機能的構成を示す機能ブロック図である。図２は、文書分類モデルを用いた文書の分類（予測）の局面における文書分類装置１Ｂ（１）の機能的構成を示すブロック図である。

[0020] 図１に示すように、文書分類装置１Ａは、取得部１１、正規化部１２、特徴量抽出部１３及びモデル生成部１４を含む。モデル生成部１４は、ラベル予測部１５、損失計算部１６及びモデル更新部１７を含む。また、文書分類

装置１Ａは、文書データベース２１及びモデル記憶部２２を含む。

[0021] 図２に示すように、文書分類装置１Ｂは、取得部１１Ｂ（１１）、正規化部１２Ｂ（１２）、特徴量抽出部１３Ｂ（１３）及びラベル予測部１５Ｂ（１５）を含む。また、文書分類装置１Ｂは、モデル記憶部２２及び文書データベース２３を含む。

[0022] 文書分類装置１Ａ及び文書分類装置１Ｂは、各々が別の装置として構成されてもよいし、一体に構成されてもよい。また、文書分類装置１Ａ及び文書分類装置１Ｂに含まれる各機能部は、複数の装置に分散されて構成されてもよい。

[0023] なお、図１及び図２に示したブロック図は、機能単位のブロックを示している。これらの機能ブロック（構成部）は、ハードウェア及び／又はソフトウェアの任意の組み合わせによって実現される。また、各機能ブロックの実現手段は特に限定されない。すなわち、各機能ブロックは、物理的及び／又は論理的に結合した１つの装置により実現されてもよいし、物理的及び／又は論理的に分離した２つ以上の装置を直接的及び／又は間接的に(例えば、有線及び／又は無線)で接続し、これら複数の装置により実現されてもよい。

[0024] 例えば、本発明の一実施の形態における文書分類装置１は、コンピュータとして機能してもよい。図３は、本実施形態に係る文書分類装置１のハードウェア構成の一例を示す図である。文書分類装置１は、物理的には、プロセッサ１００１、メモリ１００２、ストレージ１００３、通信装置１００４、入力装置１００５、出力装置１００６、バス１００７などを含むコンピュータ装置として構成されてもよい。

[0025] なお、以下の説明では、「装置」という文言は、回路、デバイス、ユニットなどに読み替えることができる。文書分類装置１のハードウェア構成は、図３に示した各装置を１つ又は複数含むように構成されてもよいし、一部の装置を含まずに構成されてもよい。

[0026] 文書分類装置１における各機能は、プロセッサ１００１、メモリ１００２などのハードウェア上に所定のソフトウェア（プログラム）を読み込ませる

ことで、プロセッサ１００１が演算を行い、通信装置１００４による通信や、メモリ１００２及びストレージ１００３におけるデータの読み出し及び／又は書き込みを制御することで実現される。

[0027] プロセッサ１００１は、例えば、オペレーティングシステムを動作させてコンピュータ全体を制御する。プロセッサ１００１は、周辺装置とのインターフェース、制御装置、演算装置、レジスタなどを含む中央処理装置（ＣＰＵ：Central Processing Unit）で構成されてもよい。また、プロセッサ１００１は、ＧＰＵ（Graphics Processing Unit）を含んで構成されてもよい。例えば、図１及び図２に示した各機能部１１～１７などは、プロセッサ１００１で実現されてもよい。

[0028] また、プロセッサ１００１は、プログラム（プログラムコード）、ソフトウェアモジュールやデータを、ストレージ１００３及び／又は通信装置１００４からメモリ１００２に読み出し、これらに従って各種の処理を実行する。プログラムとしては、上述の実施の形態で説明した動作の少なくとも一部をコンピュータに実行させるプログラムが用いられる。例えば、文書分類装置１の各機能部１１～１７は、メモリ１００２に格納され、プロセッサ１００１で動作する制御プログラムによって実現されてもよい。上述の各種処理は、１つのプロセッサ１００１で実行される旨を説明してきたが、２以上のプロセッサ１００１により同時又は逐次に実行されてもよい。プロセッサ１００１は、１以上のチップで実装されてもよい。なお、プログラムは、電気通信回線を介してネットワークから送信されても良い。

[0029] メモリ１００２は、コンピュータ読み取り可能な記録媒体であり、例えば、ＲＯＭ（Read Only Memory）、ＥＰＲＯＭ（Erasable Programmable ＲＯＭ）、ＥＥＰＲＯＭ（Electrically Erasable Programmable ＲＯＭ）、ＲＡＭ（Random Access Memory）などの少なくとも１つで構成されてもよい。メモリ１００２は、レジスタ、キャッシュ、メインメモリ（主記憶装置）などと呼ばれてもよい。メモリ１００２は、本発明の一実施の形態に係る文書分類方法を実施するために実行可能なプログラム（プログラムコード）

、ソフトウェアモジュールなどを保存することができる。

[0030] ストレージ１００３は、コンピュータ読み取り可能な記録媒体であり、例えば、ＣＤ－ＲＯＭ（Compact Disc ＲＯＭ）などの光ディスク、ハードディスクドライブ、フレキシブルディスク、光磁気ディスク(例えば、コンパクトディスク、デジタル多用途ディスク、Ｂｌｕ－ｒａｙ（登録商標）ディスク)、スマートカード、フラッシュメモリ(例えば、カード、スティック、キードライブ)、フロッピー（登録商標）ディスク、磁気ストリップなどの少なくとも１つで構成されてもよい。ストレージ１００３は、補助記憶装置と呼ばれてもよい。上述の記憶媒体は、例えば、メモリ１００２及び／又はストレージ１００３を含むデータベース、サーバその他の適切な媒体であってもよい。

[0031] 通信装置１００４は、有線及び／又は無線ネットワークを介してコンピュータ間の通信を行うためのハードウェア（送受信デバイス）であり、例えばネットワークデバイス、ネットワークコントローラ、ネットワークカード、通信モジュールなどともいう。

[0032] 入力装置１００５は、外部からの入力を受け付ける入力デバイス（例えば、キーボード、マウス、マイクロフォン、スイッチ、ボタン、センサなど）である。出力装置１００６は、外部への出力を実施する出力デバイス（例えば、ディスプレイ、スピーカー、LEDランプなど）である。なお、入力装置１００５及び出力装置１００６は、一体となった構成（例えば、タッチパネル）であってもよい。

[0033] また、プロセッサ１００１やメモリ１００２などの各装置は、情報を通信するためのバス１００７で接続される。バス１００７は、単一のバスで構成されてもよいし、装置間で異なるバスで構成されてもよい。

[0034] また、文書分類装置１は、マイクロプロセッサ、デジタル信号プロセッサ（ＤＳＰ：Digital Signal Processor）、ＡＳＩＣ（Application Specific Integrated Circuit）、ＰＬＤ（Programmable Logic Device）、ＦＰＧＡ（Field Programmable Gate Array）などのハードウェアを含んで

構成されてもよく、当該ハードウェアにより、各機能ブロックの一部又は全てが実現されてもよい。例えば、プロセッサ１００１は、これらのハードウェアの少なくとも１つで実装されてもよい。

[0035] 再び図１を参照して、文書分類装置１Ａの各機能部について説明する。取得部１１は、文書と当該文書に関連付けられた識別情報とを含む学習データを取得する。本実施形態では、取得部１１は、識別情報を構成するラベルが付されたラベル付き文書を文書データベース２１から取得する。

[0036] 文書データベース２１は、ラベル付き文書を記憶している記憶手段である。図４は、文書データベース２１に記憶されているラベル付き文書の例を示す図である。図４に示すように、文書データベース２１は、文書を識別する文書インデックスに関連付けて文書及びラベルを記憶している。例えば、文書インデックス「１」の文書は、「Ｂｌｕｅｔｏｏｔｈがつながらない」という文書であり、この文書には、ラベル「Ｃ００１」が付されている。

[0037] 正規化部１２は、取得部１１により取得された文書を正規化する。具体的には、正規化部１２は、例えば、文書の表記を統一するために、文書に対してＵｎｉｃｏｄｅ正規化を実施する。正規化形式は、例えば、ＮＦＫＣ（Normalization Form Compatibility Composition）であってもよい。また、正規化部１２は、さらに、正規化処理として、文書に含まれるアルファベットを小文字に変換してもよい。

[0038] 特徴量抽出部１３は、文書に含まれる単語と、当該単語を構成する文字に関する文字情報とを特徴量として抽出する。文字情報は、文書に含まれる単語を構成する文字のうちの一の文字または単語において連続する複数の文字からなる文字列である。文字情報は、一の単語から１または複数抽出可能な情報である。

[0039] 具体的には、特徴量抽出部１３は、周知の形態素解析の手法により、正規化部１２により正規化された文書に含まれる単語及び当該単語の品詞を抽出する。さらに、特徴量抽出部１３は、抽出した単語から、当該単語を構成する文字のうちの一の文字または当該単語において連続する複数の文字からな

る文字列を、文字情報として抽出する。このように文書から抽出された単語及び文字情報が、文書の特徴量を構成する。

[0040] ここで、図５を参照して、正規化部１２による文書の正規化から特徴量抽出部１３による特徴量の抽出に至る処理の具体例を説明する。まず、正規化部１２は、取得部１１により取得された入力文ｆ１「Ｂｌｕｅｔｏｏｈがつながらない」に対して文字列正規化処理を実施して、正規化結果ｆ２「ｂｌｕｅｔｏｏｈがつながらない」を得る。なお、この入力文ｆ１では、「Ｂｌｕｅｔｏｏｔｈ」という単語が誤入力されて、「Ｂｌｕｅｔｏｏｈ」となっている。

[0041] 次に、特徴量抽出部１３は、正規化結果ｆ２「ｂｌｕｅｔｏｏｈがつながらない」に対して形態素解析処理を実施して、正規化結果ｆ２から単語及びその単語の品詞を抽出する。図５の例では、特徴量抽出部１３は、形態素解析結果ｆ３として、「ｂｌｕｅｔｏｏｈ／名詞」、「が／助詞」、「つながら／動詞」、「ない／助動詞」を得る。これらの単語はそれぞれ、単語ＩＤ＝１～４により識別される。

[0042] 特徴量抽出部１３は、形態素解析結果ｆ３として抽出した各単語から部分文字列を抽出する部分文字列抽出処理を実施して、部分文字列抽出結果ｆ４を得る。図５に示す例では、特徴量抽出部１３は、抽出された一の単語からなる単語ユニグラムを抽出すると共に、抽出された単語を構成する一の文字からなる文字ユニグラム、及び、抽出された単語において連続する２文字からなる文字バイグラムを、各単語から抽出する。

[0043] 具体的には、特徴量抽出部１３は、符号ｆ４１に示すように、単語ＩＤが１である単語「ｂｌｕｅｔｏｏｈ／名詞」から、「単語ユニグラム：ｂｌｕｅｔｏｏｈ」、「文字ユニグラム：ｂ、ｌ、ｕ、ｅ、ｔ、ｏ、ｈ」及び「文字バイグラム：ｂｌ、ｌｕ、ｕｅ、ｅｔ、ｔｏ、ｏｏ、ｏｈ、ｈが」を抽出する。ここで、抽出される文字バイグラムの最後の要素には、入力文において当該単語の次に位置する単語の一文字目が付加されることとしてもよい。

[0044] 同様に、特徴量抽出部１３は、符号ｆ４２に示すように、単語ＩＤが２で

ある単語「が／助詞」から、「単語ユニグラム：が」、「文字ユニグラム：が」及び「文字バイグラム：がつ」を抽出する。

[0045] 同様に、特徴量抽出部１３は、符号ｆ４３に示すように、単語ＩＤが３である単語「つながら／動詞」から、「単語ユニグラム：つながら」、「文字ユニグラム：つ、な、が、ら」及び「文字バイグラム：つな、なが、がら、らな」を抽出する。

[0046] 同様に、特徴量抽出部１３は、符号ｆ４４に示すように、単語ＩＤが４である単語「ない／助動詞」から、「単語ユニグラム：ない」、「文字ユニグラム：な、い」及び「文字バイグラム：ない、い＜Ｅ＞」を抽出する。ここで、入力文において文末に位置する単語から文字バイグラムを抽出する場合において、抽出される文字バイグラムの最後の要素には、例えば「＜Ｅ＞」のような、文末を示す特殊文字が付加されることとしてもよい。

[0047] 特徴量抽出部１３は、部分文字列抽出結果ｆ４に対して特徴量選択処理を実施して、特徴量ｆ５を抽出する。特徴量抽出部１３は、図５に示す例では、入力文から抽出された単語である単語ユニグラム「ｂｌｕｅｔｏｏｈ、が、つながら、ない」を特徴量として抽出する。また、特徴量抽出部１３は、入力文から抽出された単語に含まれる文字情報である文字ユニグラム及び文字バイグラムを特徴量として抽出する。

[0048] 特徴量抽出部１３は、入力文に含まれる全ての単語から文字情報を抽出してもよいが、図５に示す例のように、入力文に含まれる単語のうち所定条件に該当する単語から文字情報を抽出することとしてもよい。入力される文書に含まれる全ての単語から文字情報を抽出して抽出した全ての文字情報を特徴量とすると、文書分類モデルに入力する特徴量（ベクトル）の次元数が膨大になってしまい、モデルの生成及び文書の分類に係る処理負荷が大きくなる。所定条件に該当する単語に含まれる文字のみを文字情報として抽出することにより、処理負荷を軽減できる。

[0049] 特徴量抽出部１３は、入力文に含まれる単語のうち、所定の品詞に該当する単語から文字情報を抽出する。誤字及び脱字等が発生しやすい品詞を所定

の品詞として、所定の品詞に該当する単語から文字情報を抽出することにより、そのような単語が誤字等を含んで、分類対象の文書として入力された場合であっても、分類の精度の低下を防止できる。

[0050] 図５に示す例では、特徴量抽出部１３は、具体的には名詞に含まれる文字情報を特徴量として抽出することを文字情報利用条件として、名詞である単語「ｂｌｕｅｔｏｏｈ」に含まれる文字情報群（「文字ユニグラム：ｂ、ｌ、ｕ、ｅ、ｔ、ｏ、ｈ」及び「文字バイグラム：ｂｌ、ｌｕ、ｕｅ、ｅｔ、ｔｏ、ｏｏ、ｏｈ、ｈが」）を特徴量として抽出する。

[0051] そして、特徴量抽出部１３は、入力文に含まれる単語「ｂｌｕｅｔｏｏｈ、が、つながら、ない」及び単語「ｂｌｕｅｔｏｏｈ」に含まれる文字情報群「ｂ、ｌ、ｕ、ｅ、ｔ、ｏ、ｈ、ｂｌ、ｌｕ、ｕｅ、ｅｔ、ｔｏ、ｏｏ、ｏｈ、ｈが」を特徴量ｆ５として抽出する。

[0052] また、特徴量抽出部１３は、入力文に含まれる単語のうち所定の品詞に該当する単語を、文字情報を抽出するための単語から除外することとしてもよい。例えば助詞等のような文書の分類に大きな影響を与えない品詞を所定の品詞として、そのような所定の品詞からの文字情報の抽出を行わないことにより、分類精度の低下を防止しつつ、文書分類モデルに入力する情報量を適切に削減できる。

[0053] また、特徴量抽出部１３は、取得部１１により取得された学習データにおいて、所定頻度以下で含まれる文字情報を特徴量として抽出することとしてもよい。機械学習に用いる複数の学習データの文書において、所定頻度を超えて含まれる文字は、助詞等のように文書の分類の精度に大きく寄与しない文字である可能性が高い。適切に設定された所定頻度以下で学習データに含まれる文字を文字情報として抽出することにより、分類精度の低下の防止に効果的に寄与する特徴量を構成できる。

[0054] また、特徴量抽出部１３は、文書に含まれる単語から抽出した文字情報から、重複する文字情報を除去することとしてもよい。例えば、図５に示す例において、文字「な」は、単語「つながら」及び単語「ない」に重複して含

まれているので、それぞれの単語から取得された文字「な」のうちの一方を除去することとしてもよい。このように重複する文字を特徴量から除去することにより、文書分類モデルに入力する情報量を削減できる。

[0055] 再び図１を参照して、文書分類装置１Ａの機能部を説明する。モデル生成部１４は、文書から抽出された特徴量及び当該文書に関連付けられた識別情報に基づいて機械学習を行い、文書分類モデルを生成する。本実施形態では、モデル生成部１４は、入力文から抽出された特徴量及び当該入力文に関連付けられているラベルに基づいて機械学習を行う。前述のとおり、モデル生成部１４は、ラベル予測部１５、損失計算部１６及びモデル更新部１７を含む。

[0056] 本実施形態の文書分類モデルは、ニューラルネットワーク（フィードフォワードニューラルネットワーク）を含んで構成される。学習済みのニューラルネットワークを含むモデルである文書分類モデルは、コンピュータにより読み込まれ又は参照され、コンピュータに所定の処理を実行させ及びコンピュータに所定の機能を実現させるプログラムとして捉えることができる。

[0057] 即ち、本実施形態の学習済みモデルは、ＣＰＵ及びメモリを備えるコンピュータにおいて用いられる。具体的には、コンピュータのＣＰＵが、メモリに記憶された学習済みモデルからの指令に従って、ニューラルネットワークの入力層に入力された入力データ（例えば、単語及び文字情報を含む特徴量）に対し、各層に対応する学習済みの重み付け係数と応答関数等に基づく演算を行い、出力層から結果（ラベル）を出力するよう動作する。

[0058] ラベル予測部１５は、フィードフォワードニューラルネットワークを用いて、入力文より抽出した特徴量を特徴ベクトルに変換し、さらにアフィン変換を行う。そして、ラベル予測部１５は、特徴ベクトルに活性化関数を適応させ、入力文に対応するラベルの確率値を求め、確率値が最大となる予測ラベルを取得する。ラベル予測部１５は、入力文に関連付けられているラベルを正解ラベルとして取得する。ここで、「正解ラベル」は、入力文書の分類結果を識別する識別情報である。

[0059] 損失計算部１６は、ラベル予測部１５により求められた各ラベルの確率値を用いて、正解ラベルと予測ラベルのとの間の損失を算出する。モデル更新部１７は、損失をフィードフォワードニューラルネットワークに逆伝搬させ、ニューラルネットワークの重みを更新する。

[0060] モデル生成部１４は、ラベル予測部１５、損失計算部１６及びモデル更新部１７による機械学習を実施し、更新されたニューラルネットワークが予め設定されたモデルの生成に関する所定条件を満たした場合に、ニューラルネットワークを含むモデルを、学習済みの文書分類モデルとして、モデル記憶部２２に保存する。

[0061] 以下にモデル生成部１４によるモデル生成の例をより具体的に説明する。ラベル予測部１５は、特徴量抽出部１３により抽出された特徴量に基づいてラベルを予測する。ここで、入力文より得られた特徴量は、Ｘ＝｛ｘ＿１，ｘ＿２，・・・、ｘ＿ｎ｝のように表される。Ｘの各要素はそれぞれ、各特徴量を示す。nは入力文に含まれる特徴量の数を示す。一般的なフィードフォワードニューラルネットワークは、入力層、中間層、出力層により構成される。まず、入力層では、入力文より抽出した特徴量を特徴ベクトルに変換する。ここで特徴ベクトルは、ｄ次元のベクトルであり、文書データベース２１から取得された入力文（文書）の特徴量と一対一に対応している。また、特徴ベクトルを格納するｄ１×Ｖ次元の行列をＥとする。入力文から抽出された特徴量Ｘの各要素は、特徴ベクトルに変換され、特徴ベクトル列Ｃ＝｛ｃ＿１，ｃ＿２，・・・、ｃ＿ｎ｝となる。

[0062] 次に、中間層の入力とするため、特徴ベクトル列の平均をとり、ｄ１次元のベクトルとなるＳを求める。次に、Ｓに対してアフェン変換を行い、ｄ２次元の中間層ｈを得る。さらに、以下の式の計算により、ｈを得る。

ｈ＝Ｗ＿ｉｎ・Ｓ＋ｂ＿ｉｎ

ただし、Ｗ＿ｉｎはｄ２×ｄ１次元の重み行列であり、ｂ＿ｉｎは次元ｄ２のバイアスベクトルである。

[0063] 次に、活性化関数ｔａｎｈを用いて、ｈ’＝ｔａｎｈ（ｈ）を得る。そし

て、ｈ’に対してアフェン変換を行い、以下のようにＮ次元の出力層Ｏを求める。ここで、Ｎはラベルが関連付けられた文書データベースから抽出されたラベルのラベル数である。

Ｏ＝Ｗ＿ｏｕｔ・ｈ’＋ｂ＿ｏｕｔ

ただし、Ｗ＿ｏｕｔはＮ×ｄ２次元の重み行列であり、ｂ＿ｏｕｔは次元Ｎのバイアスベクトルである。

[0064] 求められたｈ’に、ソフトマックス関数を適応することで、各ラベルに対応する確率値を求めることができるので、各ラベルの確率値を用いて、正解ラベルと予測ラベルとの間の損失Ｌを算出する。損失Ｌは、クロスエントロピーにより以下の式により求められる。

[数1]

$$L = -\sum_{i=1}^{N} y\_i \cdot \log(o\_i)$$

ただし、上記式において、ｙ＿ｉ∈｛０，１｝であり、ラベルが正解ラベルである場合には、ｙ＿ｉ＝１となり、正解ラベルでない場合には、ｙ＿ｉ＝０となる。

[0065] モデル更新部１７は、誤差伝搬法を用いて、損失をフィードフォワードニューラルネットワークに逆伝搬させ、ネットワークの重みを更新する。ここで更新すべき重みは、Ｗ＝｛Ｅ，Ｗ＿ｉｎ、ｂ＿ｉｎ、Ｗ＿ｏｕｔ、ｂ＿ｏｕｔ｝である。

[0066] モデル生成部１４は、ラベルが関連付けられた文書のＮ回の読み込み及びモデルの更新の後、生成されたモデルを学習済みの文書分類モデルとしてモデル記憶部２２に記憶させる。

[0067] 次に、図２を参照して、文書分類装置１Ｂの各機能部について説明する。文書分類装置１Ｂの取得部１１Ｂ、正規化部１２Ｂ、特徴量抽出部１３Ｂ及びラベル予測部１５Ｂの機能は、文書分類装置１Ａの取得部１１、正規化部１２、特徴量抽出部１３及びラベル予測部１５の機能と同様であるので、簡単に説明する。

[0068] 取得部１１Ｂは、文書データベース２３から分類対象の文書を取得する。文書データベース２３は、分類対象の文書を記憶している記憶手段である。分類対象の文書は、ラベルが関連付けられていない文書である。

[0069] 正規化部１２Ｂは、取得部１１Ｂにより取得された文書を正規化する。具体的には、正規化部１２Ｂは、例えば、文書の表記を統一するために、文書に対してＵｎｉｃｏｄｅ正規化及び文書に含まれるアルファベットを小文字に変換する処理等を実施する。

[0070] 特徴量抽出部１３Ｂは、文書に含まれる単語と、当該単語を構成する文字に関する文字情報とを特徴量として抽出する。

[0071] ラベル予測部１５Ｂは、特徴量抽出部１３Ｂにより抽出された特徴量に基づいて、モデル記憶部２２から読み出した学習済みの文書分類モデルを用いて、文書を分類する。具体的には、ラベル予測部１５Ｂは、文書分類モデルに特徴量を入力し、文書分類モデルから出力されるラベルを、文書の分類の結果を識別する識別情報として出力する。

[0072] 次に、図６を参照して、文書分類装置１における、文書分類モデルを機械学習により生成する生成方法について説明する。図６は、本実施形態の生成方法の処理内容を示すフローチャートである。

[0073] ステップＳ１において、取得部１１は、入力文書と当該入力文書に関連付けられたラベルとを含む学習データを取得する。

[0074] ステップＳ２において、特徴量抽出部１３は、ステップＳ１において取得された入力文書から、入力文書に含まれる単語と、当該単語を構成する文字に関する文字情報とを特徴量として抽出する。なお、特徴量の抽出に先立って、正規化部１２による入力文書の正規化が実施されてもよい。

[0075] ステップＳ３及びステップ４において、モデル生成部１４は、文書から抽出された特徴量及び当該文書に関連付けられた識別情報に基づいて機械学習を行い、文書分類モデルを生成する。具体的には、ステップＳ３において、ラベル予測部１５は、ラベル予測部１５は、フィードフォワードニューラルネットワークを用いて、入力文書より抽出した特徴量に基づいて、確率値が

最大となる予測ラベルを取得する。

[0076] ステップＳ４において、損失計算部１６は、ステップＳ３において求められたラベルの確率値を用いて、正解ラベルと予測ラベルとの間の損失を算出する。正解ラベルはステップＳ１において取得された入力文書に関連付けられているラベルである。そして、モデル更新部１７は、損失をフィードフォワードニューラルネットワークに逆伝搬させ、ニューラルネットワークの重みを更新する。

[0077] ステップＳ５において、モデル生成部１４は、ステップＳ１～Ｓ４のモデルの生成のための学習処理を終了するか否かを判定する。この判定は、例えば、所定条件が充足されたか否かにより行われる。所定条件は、例えば、学習に用いた学習データの数が閾値を超えること等とすることができるか、この条件に限定されない。学習処理を終了すると判定された場合には、処理はステップＳ１に戻る。一方、学習処理を終了すると判定されなかった場合には、処理はステップＳ６に進む。

[0078] ステップＳ６において、モデル生成部１４は、生成された文書分類モデルを出力する。

[0079] 次に、図７を参照して、文書分類装置１Ｂ（１）における、文書分類モデルを用いた文書の分類（予測）方法について説明する。図７は、本実施形態の分類方法の処理内容を示すフローチャートである。

[0080] ステップＳ１１において、取得部１１Ｂは、分類対象の文書である入力文書を取得する。

[0081] ステップＳ１２において、特徴量抽出部１３Ｂは、ステップＳ１において取得された入力文書から、入力文書に含まれる単語と、当該単語を構成する文字に関する文字情報とを特徴量として抽出する。なお、特徴量の抽出に先立って、正規化部１２Ｂによる入力文書の正規化が実施されてもよい。

[0082] ステップＳ１３において、ラベル予測部１５Ｂは、モデル記憶部２２に記憶されている学習済みの文書分類モデルを読み込む。

[0083] ステップＳ１４において、ラベル予測部１５Ｂは、フィードフォワードニ

ューラルネットワークを含む文書分類モデルを用いて、入力文書を分類する。具体的には、ラベル予測部１５Ｂは、文書分類モデルに特徴量を入力し、文書分類モデルから出力されるラベルを、文書の分類の結果を識別する識別情報として出力する。

[0084] 次に、コンピュータを、本実施形態の文書分類装置１として機能させるための文書分類プログラムについて説明する。図８は、文書分類プログラムＰ１の構成を示す図である。

[0085] 文書分類プログラムＰ１は、文書分類装置１におけるモデルの生成処理を統括的に制御するメインモジュールｍ１０、取得モジュールｍ１１、正規化モジュールｍ１２、特徴量抽出モジュールｍ１３及びモデル生成モジュールｍ１４を備えて構成される。モデル生成モジュールｍ１４は、ラベル予測モジュールｍ１５、損失計算モジュールｍ１６及びモデル更新モジュールｍ１７を含む。そして、各モジュールｍ１１～ｍ１７により、文書分類装置１における取得部１１、正規化部１２、特徴量抽出部１３、モデル生成部１４、ラベル予測部１５、損失計算部１６及びモデル更新部１７のための各機能が実現される。なお、文書分類プログラムＰ１は、通信回線等の伝送媒体を介して伝送される態様であってもよいし、図８に示されるように、記録媒体Ｍ１に記憶される態様であってもよい。

[0086] 以上説明した本実施形態の文書分類装置１では、文書分類モデルの生成のための学習用の文書から抽出された単語と共に、当該単語に含まれる文字からなる文字情報が特徴量として用いられて文書分類モデルが生成されるので、単語に含まれる文字単位の情報も文書分類モデルに反映される。これにより、分類対象の文書に含まれる単語に誤字及び脱字等が含まれている場合であっても、その単語に含まれる文字も特徴量として文書分類モデルに入力することにより、文書に含まれる単語を構成する文字に関する特徴が分類に加味される。従って、文書の分類の精度の低下を防止できる。

[0087] また、本実施形態の学習済みの文書分類モデルでは、分類対象の文章から抽出された単語だけではなく、当該単語に含まれる文字からなる文字情報を

含む特徴量が学習データとして用いられた機械学習により、文書を分類するための学習済みモデルが構成される。これにより、分類対象の文書に含まれる単語に誤字及び脱字等が含まれている場合であっても、その単語に含まれる文字も特徴量として学習済みモデルに入力することにより、文書に含まれる単語を構成する文字に関する特徴が分類に加味される。従って、文書の分類の精度の低下を防止できる。

[0088] また、別の形態に係る文書分類装置では、特徴量抽出部は、文書に含まれる単語のうち所定条件に該当する単語から、文字情報を抽出することとしてもよい。

[0089] 入力される文書に含まれる全ての文字を特徴量とすると、文書分類モデルに入力する特徴量（ベクトル）の次元数が膨大になってしまい、モデルの生成及び文書の分類に係る処理負荷が大きくなる。上記形態によれば、所定条件に該当する単語に含まれる文字のみを文字情報として抽出することにより、処理負荷を軽減できる。

[0090] また、別の形態に係る文書分類装置は、特徴量抽出部は、文書に含まれる単語のうち第１の品詞に該当する単語から、文字情報を抽出することとしてもよい。

[0091] 上記形態によれば、誤字及び脱字等が発生しやすい品詞を所定の品詞として、所定の品詞に該当する単語から文字情報を抽出することにより、そのような単語が誤字等を含んで、分類対象の文書として入力された場合であっても、分類の精度の低下を防止できる。

[0092] また、別の形態に係る文書分類装置では、特徴量抽出部は、文書に含まれる単語のうち第２の品詞に該当する単語を、文字情報を抽出する単語から除外することとしてもよい。

[0093] 上記形態によれば、例えば助詞等のような文書の分類に大きな影響を与えない品詞を所定の品詞として、そのような所定の品詞からの文字情報の抽出を行わないことにより、分類精度の低下を防止しつつ、文書分類モデルに入力する情報量を適切に削減できる。

[0094] また、別の形態に係る文書分類装置では、特徴量抽出部は、取得部により取得された学習データにおいて、所定頻度以下で含まれる文字情報を、特徴量として抽出することとしてもよい。

[0095] 機械学習に用いる複数の学習データの文書において、所定頻度を超えて含まれる文字は、助詞等のように文書の分類の精度に大きく寄与しない文字である可能性が高い。上記形態によれば、適切に設定された所定頻度以下で学習データに含まれる文字を文字情報として抽出することにより、分類精度の低下の防止に効果的に寄与する特徴量を構成できる。

[0096] また、別の形態に係る文書分類装置では、特徴量抽出部は、文書に含まれる単語から抽出した文字情報から、重複する文字情報を除去することとしてもよい。

[0097] 上記形態によれば、分類精度の低下を招くことなく、文書分類モデルに入力する情報量を削減できる。

[0098] 以上、本実施形態について詳細に説明したが、当業者にとっては、本実施形態が本明細書中に説明した実施形態に限定されるものではないということは明らかである。本実施形態は、特許請求の範囲の記載により定まる本発明の趣旨及び範囲を逸脱することなく修正及び変更態様として実施することができる。したがって、本明細書の記載は、例示説明を目的とするものであり、本実施形態に対して何ら制限的な意味を有するものではない。

[0099] 本明細書で説明した各態様／実施形態の処理手順、シーケンス、フローチャートなどは、矛盾の無い限り、順序を入れ替えてもよい。例えば、本明細書で説明した方法については、例示的な順序で様々なステップの要素を提示しており、提示した特定の順序に限定されない。

[0100] 情報等は、上位レイヤ(または下位レイヤ)から下位レイヤ(または上位レイヤ)へ出力され得る。複数のネットワークノードを介して入出力されてもよい。

[0101] 入出力された情報等は特定の場所(例えば、メモリ)に保存されてもよいし、管理テーブルで管理してもよい。入出力される情報等は、上書き、更新、

または追記され得る。出力された情報等は削除されてもよい。入力された情報等は他の装置へ送信されてもよい。

[0102] 判定は、１ビットで表される値（０か１か）によって行われてもよいし、真偽値（Boolean：trueまたはfalse）によって行われてもよいし、数値の比較（例えば、所定の値との比較）によって行われてもよい。

[0103] 本明細書で説明した各態様／実施形態は単独で用いてもよいし、組み合わせて用いてもよいし、実行に伴って切り替えて用いてもよい。また、所定の情報の通知（例えば、「Ｘであること」の通知）は、明示的に行うものに限られず、暗黙的（例えば、当該所定の情報の通知を行わない）ことによって行われてもよい。

[0104] ソフトウェアは、ソフトウェア、ファームウェア、ミドルウェア、マイクロコード、ハードウェア記述言語と呼ばれるか、他の名称で呼ばれるかを問わず、命令、命令セット、コード、コードセグメント、プログラムコード、プログラム、サブプログラム、ソフトウェアモジュール、アプリケーション、ソフトウェアアプリケーション、ソフトウェアパッケージ、ルーチン、サブルーチン、オブジェクト、実行可能ファイル、実行スレッド、手順、機能などを意味するよう広く解釈されるべきである。

[0105] また、ソフトウェア、命令などは、伝送媒体を介して送受信されてもよい。例えば、ソフトウェアが、同軸ケーブル、光ファイバケーブル、ツイストペア及びデジタル加入者回線（ＤＳＬ）などの有線技術及び／又は赤外線、無線及びマイクロ波などの無線技術を使用してウェブサイト、サーバ、又は他のリモートソースから送信される場合、これらの有線技術及び／又は無線技術は、伝送媒体の定義内に含まれる。

[0106] 本明細書で説明した情報、信号などは、様々な異なる技術のいずれかを使用して表されてもよい。例えば、上記の説明全体に渡って言及され得るデータ、命令、コマンド、情報、信号、ビット、シンボル、チップなどは、電圧、電流、電磁波、磁界若しくは磁性粒子、光場若しくは光子、又はこれらの任意の組み合わせによって表されてもよい。

[0107] なお、本明細書で説明した用語及び／又は本明細書の理解に必要な用語については、同一の又は類似する意味を有する用語と置き換えてもよい。

[0108] 本明細書で使用する「システム」および「ネットワーク」という用語は、互換的に使用される。

[0109] また、本明細書で説明した情報、パラメータなどは、絶対値で表されてもよいし、所定の値からの相対値で表されてもよいし、対応する別の情報で表されてもよい。

[0110] 本明細書で使用する「に基づいて」という記載は、別段に明記されていない限り、「のみに基づいて」を意味しない。言い換えれば、「に基づいて」という記載は、「のみに基づいて」と「に少なくとも基づいて」の両方を意味する。

[0111] 本明細書で「第１の」、「第２の」などの呼称を使用した場合においては、その要素へのいかなる参照も、それらの要素の量または順序を全般的に限定するものではない。これらの呼称は、２つ以上の要素間を区別する便利な方法として本明細書で使用され得る。したがって、第１および第２の要素への参照は、２つの要素のみがそこで採用され得ること、または何らかの形で第１の要素が第２の要素に先行しなければならないことを意味しない。

[0112] 「含む（include）」、「含んでいる（including）」、およびそれらの変形が、本明細書あるいは特許請求の範囲で使用されている限り、これら用語は、用語「備える(comprising)」と同様に、包括的であることが意図される。さらに、本明細書あるいは特許請求の範囲において使用されている用語「または（or）」は、排他的論理和ではないことが意図される。

[0113] 本明細書において、文脈または技術的に明らかに1つのみしか存在しない装置である場合以外は、複数の装置をも含むものとする。

[0114] 本開示の全体において、文脈から明らかに単数を示したものではなければ、複数のものを含むものとする。

符号の説明

[0115] １，１Ａ，１Ｂ…文書分類装置、１１，１１Ｂ…取得部、１２，１２Ｂ…

正規化部、１３，１３Ｂ…特徴量抽出部、１４…モデル生成部、１５，１５Ｂ…ラベル予測部、１６…損失計算部、１７…モデル更新部、２１…文書データベース、２２…モデル記憶部、２３…文書データベース、Ｍ１…記録媒体、ｍ１０…メインモジュール、ｍ１１…取得モジュール、ｍ１２…正規化モジュール、ｍ１３…特徴量抽出モジュール、ｍ１４…モデル生成モジュール、ｍ１５…ラベル予測モジュール、ｍ１６…損失計算モジュール、ｍ１７…モデル更新モジュール、Ｐ１…文書分類プログラム。

請求の範囲

[請求項1] 文書を分類するためのモデルであって、入力された文書に基づいて、分類の結果を識別する識別情報を出力する文書分類モデルを機械学習により生成する文書分類装置であって、

文書と該文書に関連付けられた前記識別情報とを含む学習データを取得する取得部と、

前記文書に含まれる単語と、前記単語を構成する文字のうちの一の文字または前記単語において連続する複数の文字からなる文字列であって、前記単語から１または複数抽出可能な情報である文字情報と、を特徴量として抽出する特徴量抽出部と、

前記文書から抽出された特徴量及び当該文書に関連付けられた前記識別情報に基づいて機械学習を行い、前記文書分類モデルを生成するモデル生成部と、

を備える文書分類装置。

[請求項2] 前記特徴量抽出部は、前記文書に含まれる単語のうち所定条件に該当する単語から、前記文字情報を抽出する、

請求項１に記載の文書分類装置。

[請求項3] 前記特徴量抽出部は、前記文書に含まれる単語のうち第１の品詞に該当する単語から、前記文字情報を抽出する、

請求項２に記載の文書分類装置。

[請求項4] 前記特徴量抽出部は、前記文書に含まれる単語のうち第２の品詞に該当する単語を、前記文字情報を抽出する単語から除外する、

請求項２または３に記載の文書分類装置。

[請求項5] 前記特徴量抽出部は、前記取得部により取得された前記学習データにおいて、所定頻度以下で含まれる文字情報を、特徴量として抽出する、

請求項１～４のいずれか一項に記載の文書分類装置。

[請求項6] 前記特徴量抽出部は、前記文書に含まれる単語から抽出した文字情

報から、重複する文字情報を除去する、

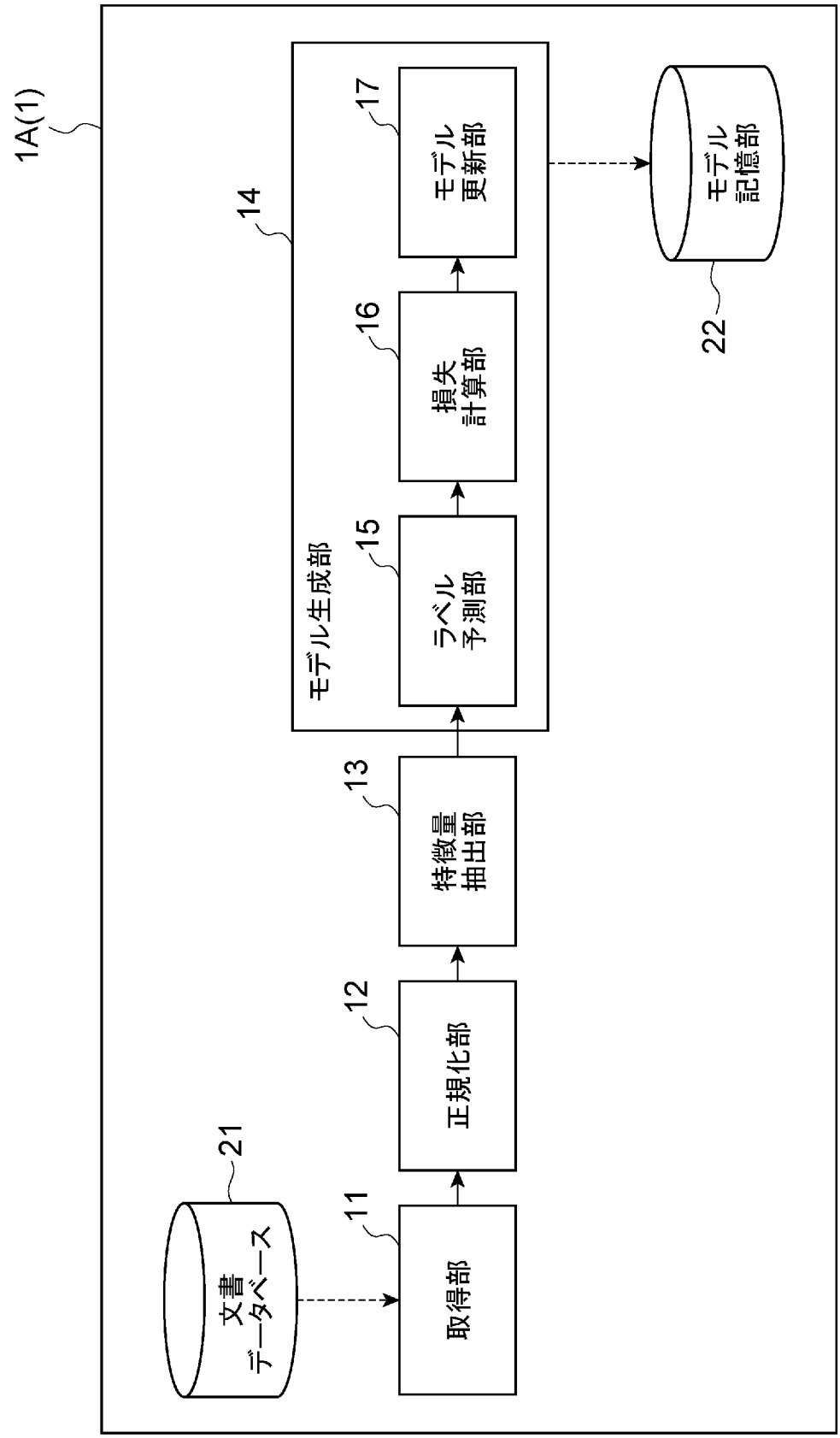
請求項１～５のいずれか一項に記載の文書分類装置。

[請求項7] 入力された文書を分類し、分類の結果を識別する識別情報を出力するよう、コンピュータを機能させるための学習済みモデルであって、

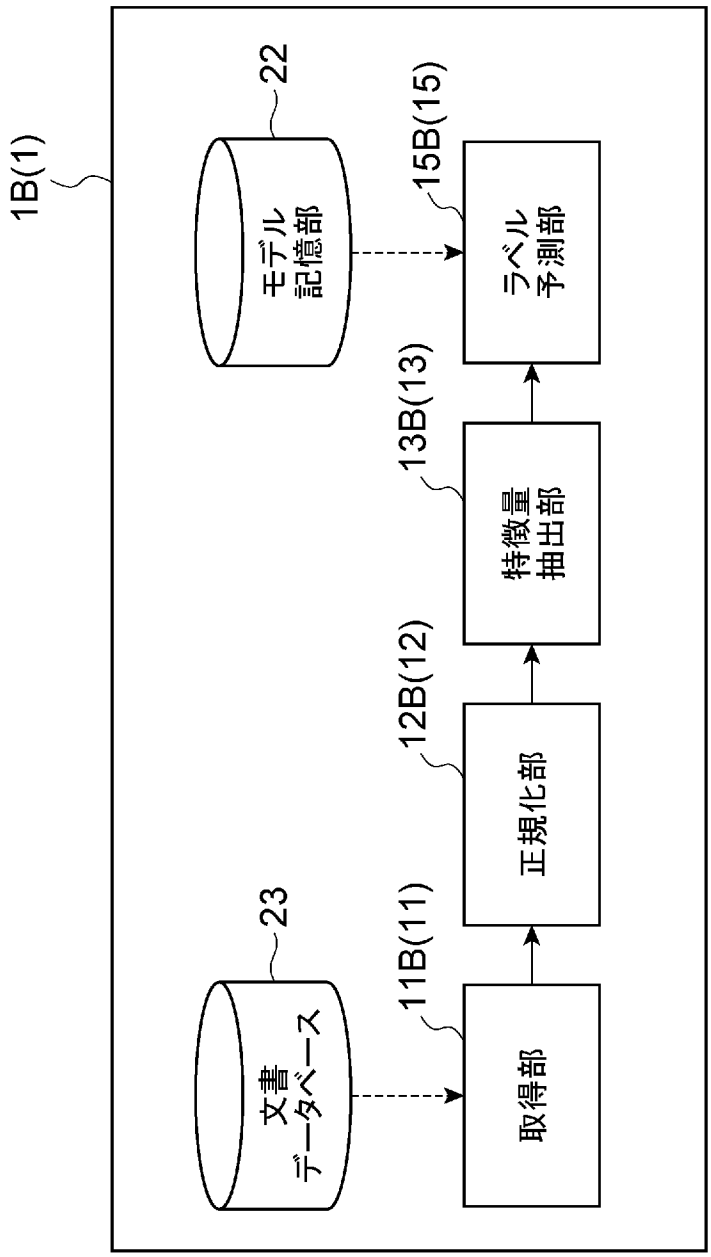
前記文書に含まれる単語と、前記単語を構成する文字のうちの一の文字または前記単語において連続する複数の文字からなる文字列であって、前記単語から１または複数抽出可能な情報である文字情報と、を含む特徴量を入力値とし、前記文書に関連付けられた前記識別情報を出力値とする学習データに基づく機械学習により構成された、

学習済みモデル。

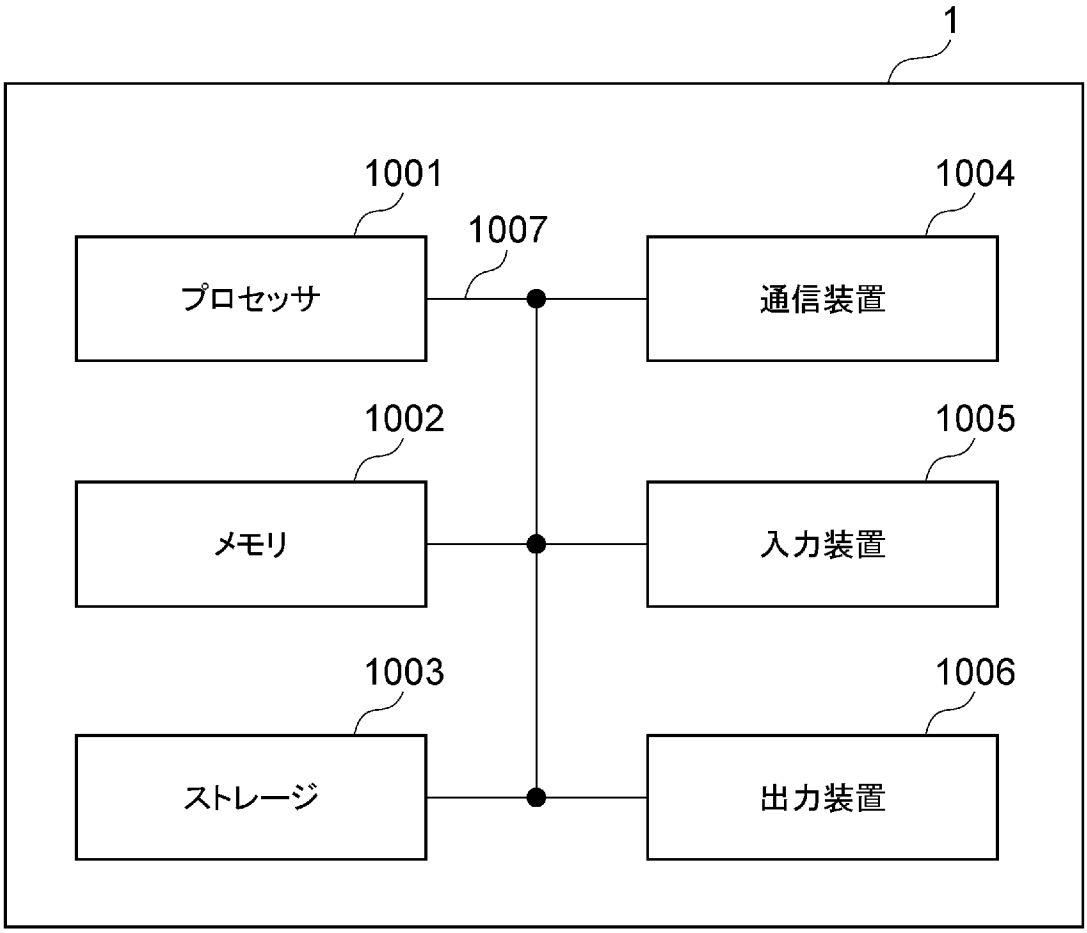
[図1]


1A(1)
14
モデル生成部
17
モデル
更新部
16
損失
計算部
15
ラベル
予測部
13
特徴量
抽出部
12
正規化部
11
取得部
21
文書
データベース
22
モデル
記憶部

[図2]


1B(1)
22
モデル記憶部
15B(15)
ラベル予測部
13B(13)
特徴量抽出部
12B(12)
正規化部
23
文書データベース
11B(11)
取得部

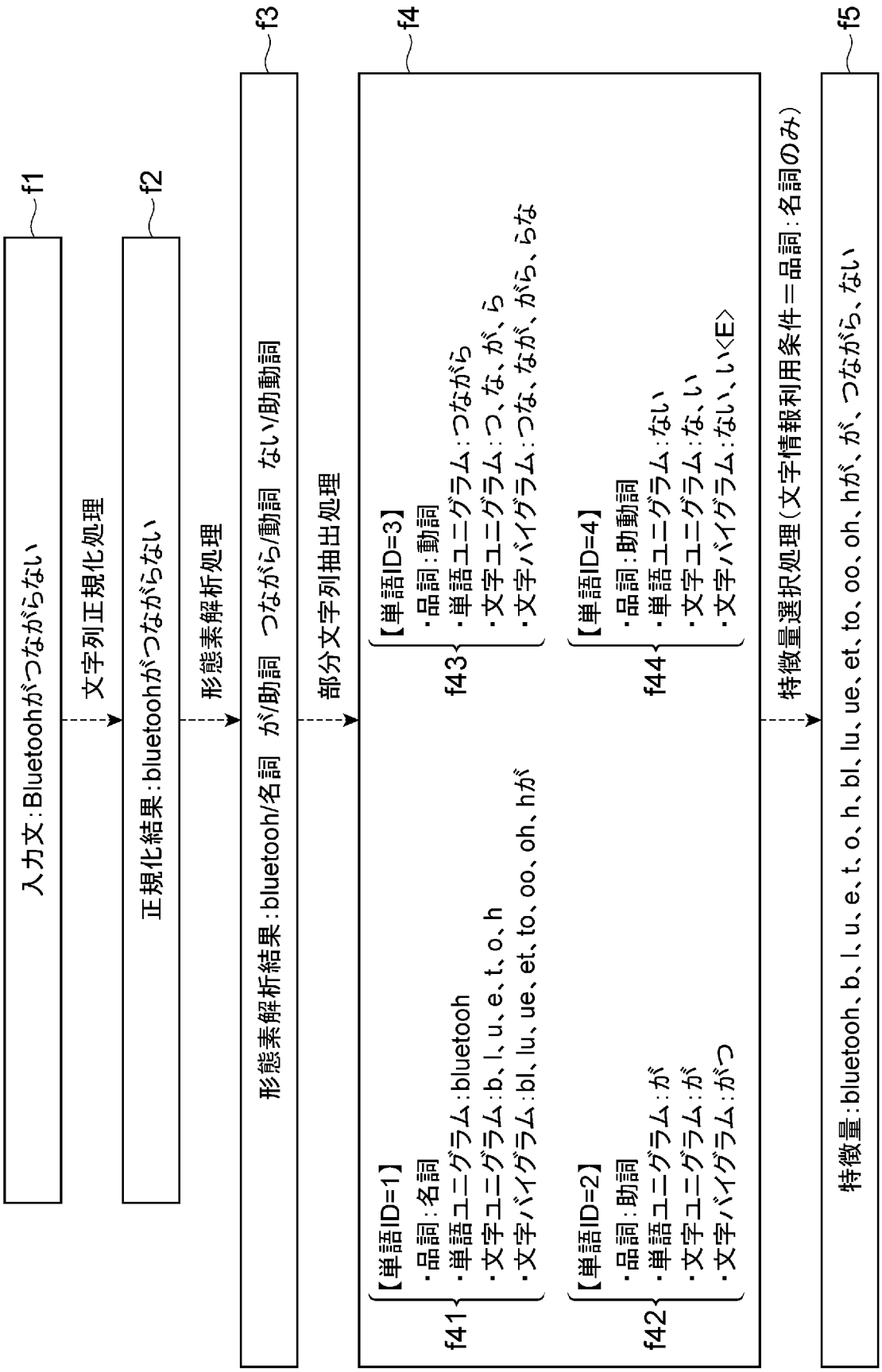
[図3]


1
1001
プロセッサ
1007
1004
通信装置
1002
メモリ
1005
入力装置
1003
ストレージ
1006
出力装置

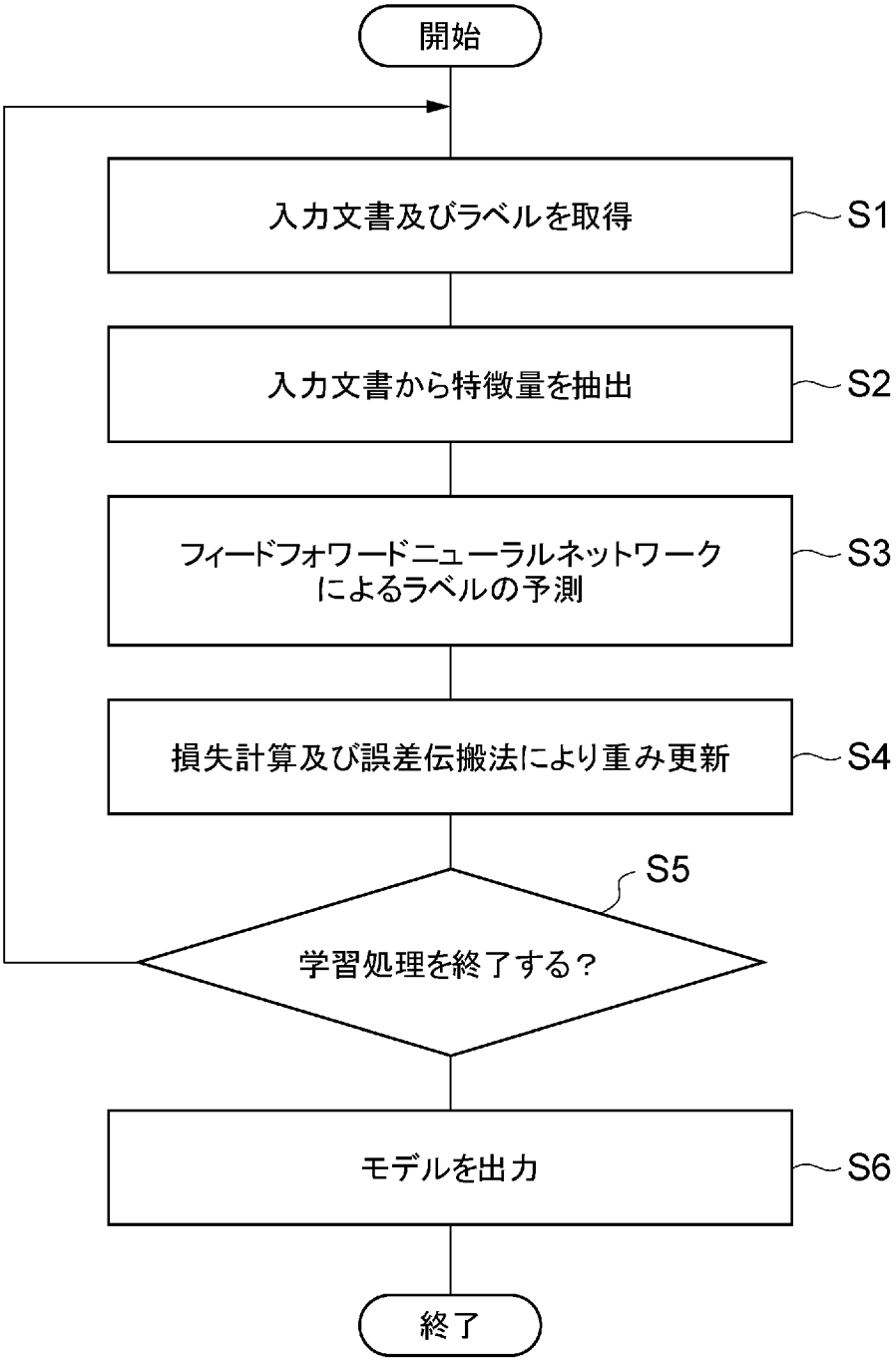
[図4]

文書インデックス	ラベル	文書
1	C001	Bluetoothがつながらない
2	C002	迷惑メールの設定方法
3	C003	写真を削除する
・・・	・・・	・・・

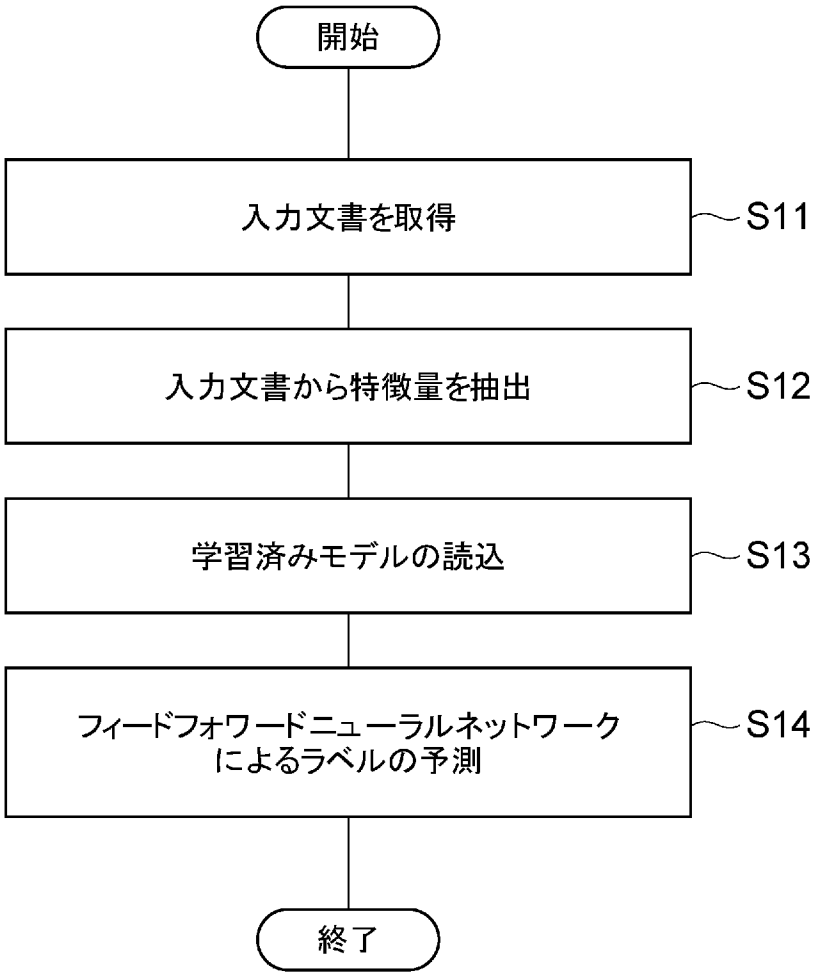
[図5]


f1
入力文:Bluetoohがつながらない
文字列正規化処理
f2
正規化結果:bluetoohがつながらない
形態素解析処理
f3
形態素解析結果:bluetooh/名詞 が/助詞 つながら/動詞 ない/助動詞
部分文字列抽出処理
f4
f41
【単語ID=1】
・品詞:名詞
・単語ユニグラム:bluetooh
・文字ユニグラム:b、l、u、e、t、o、h
・文字バイグラム:bl、lu、ue、et、to、oo、oh、hが
f42
【単語ID=2】
・品詞:助詞
・単語ユニグラム:が
・文字ユニグラム:が
・文字バイグラム:がつ
f43
【単語ID=3】
・品詞:動詞
・単語ユニグラム:つながら
・文字ユニグラム:つ、な、が、ら
・文字バイグラム:つな、ながが、がら、らな
f44
【単語ID=4】
・品詞:助動詞
・単語ユニグラム:ない
・文字ユニグラム:な、い
・文字バイグラム:ない、い<E>
特徴量選択処理(文字情報利用条件=品詞:名詞のみ)
f5
特徴量:bluetooh、b、l、u、e、t、o、h、bl、lu、ue、et、to、oo、oh、hが、が、つながら、ない

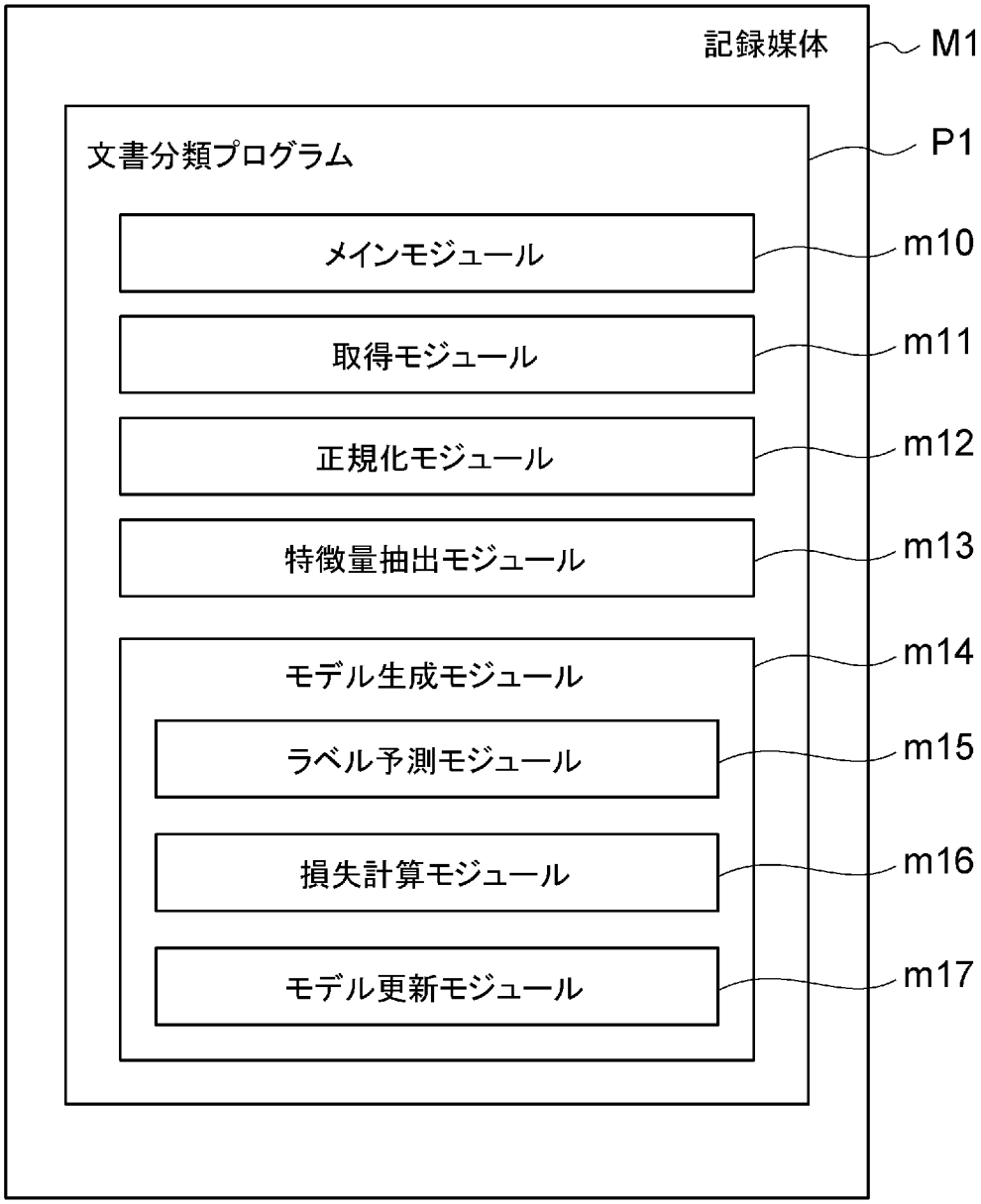
[図6]


開始
入力文書及びラベルを取得
S1
入力文書から特徴量を抽出
S2
フィードフォワードニューラルネットワーク
によるラベルの予測
S3
損失計算及び誤差伝搬法により重み更新
S4
S5
学習処理を終了する？
モデルを出力
S6
終了

[図7]


開始
入力文書を取得
S11
入力文書から特徴量を抽出
S12
学習済みモデルの読込
S13
フィードフォワードニューラルネットワーク
によるラベルの予測
S14
終了

[図8]


記録媒体
M1
文書分類プログラム
P1
メインモジュール
m10
取得モジュール
m11
正規化モジュール
m12
特徴量抽出モジュール
m13
モデル生成モジュール
m14
ラベル予測モジュール
m15
損失計算モジュール
m16
モデル更新モジュール
m17

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/021289

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl. G06F16/35(2019.01)i

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)
Int. Cl. G06F16/35

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2019
Registered utility model specifications of Japan 1996-2019
Published registered utility model applications of Japan 1994-2019

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	US 2006/0089924 A1 (RASKUTTI, Bhavani et al.) 27 April 2006, Title of the Invention, paragraphs [0059], [0060], [0073] & WO 2002/025479 A1 & EP 1323078 A1 & CA 2423033 A1 & NZ 524988 A	1, 2, 5-7 3, 4
X	CN 108108351 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 01 June 2018, Title of the Invention, abstract, claim 5 (Family: none)	1, 7
Y	JP 8-166965 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 25 June 1996, abstract (Family: none)	3, 4
Y	JP 5-54037 A (FUJITSU LTD.) 05 March 1993, abstract (Family: none)	3, 4

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

☐ See patent family annex.

* 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
23.07.2019	30.07.2019
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Authorized officer Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１９／０２１２８９

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F16/35(2019.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F16/35

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１９年
日本国実用新案登録公報	１９９６－２０１９年
日本国登録実用新案公報	１９９４－２０１９年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
X Y	US 2006/0089924 A1 (RASKUTTI, Bhavani et al.) 2006.04.27, 発明の名称、段落００５９，００６０，００７３ & WO 2002/025479 A1 & EP 1323078 A1 & CA 2423033 A1 & NZ 524988 A	1,2,5-7 3,4
X	CN 108108351 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 2018.06.01, 発明の名称、要約、請求項５（ファミリーなし）	1,7
Y	JP 8-166965 A (日本電信電話株式会社) 1996.06.25, 要約 (ファミリーなし)	3,4

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
２３．０７．２０１９	３０．０７．２０１９

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｎ　９６５２
早川　学
電話番号　０３－３５８１－１１０１　内線　３５８６

C（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Y	JP 5-54037 A（富士通株式会社）1993.03.05, 要約 (ファミリーなし)	3,4