

(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2020-4156

(P2020-4156A)

(43) 公開日 令和2年1月9日 (2020.1.9)

(51) Int.Cl.	F I	テーマコード (参考)
G 0 6 F 16/00 (2019.01)	G 0 6 F 17/30 2 1 0 D	5 B 1 0 9
G 0 6 F 16/90 (2019.01)	G 0 6 F 17/30 1 8 0 A	
G 0 6 F 40/131 (2020.01)	G 0 6 F 17/22 6 2 9	
G 0 6 F 16/30 (2019.01)	G 0 6 F 17/30 1 7 0 A	

審査請求 未請求 請求項の数 9 O L (全 18 頁)

(21) 出願番号	特願2018-123996 (P2018-123996)	(71) 出願人	000005223
(22) 出願日	平成30年6月29日 (2018.6.29)		富士通株式会社
			神奈川県川崎市中原区上小田中4丁目1番1号
		(74) 代理人	100079049
			弁理士 中島 淳
		(74) 代理人	100084995
			弁理士 加藤 和詳
		(74) 代理人	100099025
			弁理士 福田 浩志
		(72) 発明者	工藤 淳真
			神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内

最終頁に続く

(54) 【発明の名称】 分類方法、装置、及びプログラム

(57) 【要約】

【課題】 テキストの分類精度を向上させる。

【解決手段】 受付解析部12が、複数のテキストを受け付け、分割部14が、受け付けた複数のテキストのうち、いずれかのテキストが特定の条件を満たす区切れ箇所を含むことを検出すると、いずれかのテキストを区切れ箇所分割して、新たな複数のテキストを生成し、分類部16が、受け付けた複数のテキストのうち、特定の条件を満たす区切れ箇所を含まないテキストと、生成した新たな複数のテキストと、を複数のクラスタに分類する。

【選択図】 図9



【特許請求の範囲】**【請求項 1】**

複数のテキストを受け付け、

受け付けた前記複数のテキストのうち、いずれかのテキストが特定の条件を満たす区切れ箇所を含むことを検出すると、前記いずれかのテキストを前記区切れ箇所分割して、新たな複数のテキストを生成し、

受け付けた前記複数のテキストのうち、前記特定の条件を満たす区切れ箇所を含まないテキストと、生成した前記新たな複数のテキストと、を複数のクラスタに分類する、

処理をコンピュータが実行することを特徴とする分類方法。

【請求項 2】

前記複数のクラスタそれぞれに分類されたテキストに基づき、前記複数のクラスタそれぞれの特徴情報を生成し、

生成した前記特徴情報を、前記複数のクラスタそれぞれに対応付けて表示部に表示する、

ことを特徴とする請求項 1 に記載の分類方法。

【請求項 3】

前記複数のクラスタについて、前記複数のクラスタそれぞれに分類されたテキストと、該テキストに含まれる複数の形態素の、受け付けた前記複数のテキストでの出現状況とに基づき、前記複数のクラスタをそれぞれ代表する複数の代表形態素を決定し、

決定した前記複数の代表形態素それぞれを、前記複数の代表形態素それぞれが代表する複数のクラスタそれぞれに対応付けて前記表示部に表示する、

ことを特徴とする請求項 2 に記載の分類方法。

【請求項 4】

決定した前記複数の代表形態素それぞれを、前記複数のクラスタそれぞれに分類された前記テキストの数に応じた順序で並べて前記表示部に表示する、

ことを特徴とする請求項 3 に記載の分類方法。

【請求項 5】

前記複数のテキストにおける各テキストの出現状況に関する指標に基づいて、前記複数のクラスタの各々に含まれるテキストについての前記指標が、出現頻度が高いことを示す順に前記クラスタを並べて前記表示部に表示する

ことを特徴とする請求項 2 ～請求項 4 のいずれか 1 項に記載の分類方法。

【請求項 6】

前記特定の条件を満たす区切れ箇所を読点とする請求項 1 ～請求項 5 のいずれか 1 項に記載の分類方法。

【請求項 7】

前記テキストを区切れ箇所分割した前半部分及び後半部分のいずれか一方の、前記複数のテキストにおける出現状況が予め定めた条件を満たす場合に、前記区切れ箇所を前記特定の条件を満たす区切れ箇所とする請求項 1 ～請求項 5 のいずれか 1 項に記載の分類方法。

【請求項 8】

複数のテキストを受け付ける受付部と、

受け付けた前記複数のテキストのうち、いずれかのテキストが特定の条件を満たす区切れ箇所を含むことを検出すると、前記いずれかのテキストを前記区切れ箇所分割して、新たな複数のテキストを生成する生成部と、

受け付けた前記複数のテキストのうち、前記特定の条件を満たす区切れ箇所を含まないテキストと、生成した前記新たな複数のテキストと、を複数のクラスタに分類する分類部と、

を含むことを特徴とする分類装置。

【請求項 9】

複数のテキストを受け付け、

受け付けた前記複数のテキストのうち、いずれかのテキストが特定の条件を満たす区切れ箇所を含むことを検出すると、前記いずれかのテキストを前記区切れ箇所分割して、新たな複数のテキストを生成し、

受け付けた前記複数のテキストのうち、前記特定の条件を満たす区切れ箇所を含まないテキストと、生成した前記新たな複数のテキストと、を複数のクラスタに分類する、

処理をコンピュータに実行させることを特徴とする分類プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

開示の技術は、分類方法、分類装置、及び分類プログラムに関する。

10

【背景技術】

【0002】

従来、自然言語で記述された文書（テキストデータ）を、記述されている内容に基づいて分類することが行われている。

【0003】

例えば、互いに対応付けられた質問と回答とを文書記憶部に記憶しておき、文書記憶部中の各回答の特徴ベクトルに基づいて回答をクラスタ分類しておく情報検索システムが提案されている。

【0004】

また、談話データ及び談話セマンティクスを入力とし、談話データからFAQ候補となる質問文を抽出して出力するFAQ候補抽出システムが提案されている。このシステムでは、談話セマンティクスは各ステートメントのフロー情報を含み、談話データから、顧客によって発話され、質問文若しくは要求文であることを示すフローが設定された質問・要求ステートメントを抽出する。そして、質問・要求ステートメントから指定されたキーワードを含むものを抽出し、質問・要求ステートメントについてクラスタリングし、各クラスタの代表となる質問・要求ステートメントをFAQ候補として出力する。

20

【0005】

また、ツリー状の観点及び属性単語を含む観点リストを記憶した観点リスト記憶手段と、各属性単語に関連する多数の学習文章情報を記憶した学習文章情報記憶手段とを有する装置が提案されている。この装置は、共有コンテンツから複数のキーワードを抽出し、キーワードを要素とし且つその出現頻度を値とする第1のベクトルを導出する。そして、キーワード毎に、当該キーワードと一致する属性単語における学習文章情報について、当該学習文章情報に含まれる単語を要素とし且つその出現頻度を値とする第2のベクトルを導出する。さらに、両ベクトルの類似度を算出し、類似度を対応付けた類似度付き観点リストを生成し、観点リストのレイヤ毎に、類似度の分散が最も大きい観点及び属性単語を導出する。

30

【先行技術文献】

【特許文献】

【0006】

【特許文献1】特開2002-41573号公報

40

【特許文献2】特開2012-3704号公報

【特許文献3】特開2012-70036号公報

【発明の概要】

【発明が解決しようとする課題】

【0007】

しかしながら、例えば、各テキストに定型的な表現が含まれている場合には、その定型的な表現が影響して、各文書から適切な特徴を抽出することができず、文書の分類を適切に行うことができない場合がある。

【0008】

一つの側面として、開示の技術は、テキストの分類精度を向上させることを目的とする

50

。

【課題を解決するための手段】

【0009】

一つの態様として、開示の技術は、複数のテキストを受け付ける。また、開示の技術は、受け付けた前記複数のテキストのうち、いずれかのテキストが特定の条件を満たす区切れ箇所を含むことを検出すると、前記いずれかのテキストを前記区切れ箇所分割して、新たな複数のテキストを生成する。そして、開示の技術は、受け付けた前記複数のテキストのうち、前記特定の条件を満たす区切れ箇所を含まないテキストと、生成した前記新たな複数のテキストと、を複数のクラスタに分類する。

【発明の効果】

10

【0010】

一つの側面として、テキストの分類精度を向上させることができる、という効果を有する。

【図面の簡単な説明】

【0011】

【図1】文書の分類を説明するための図である。

【図2】定型表現が文書の分類に与える影響を説明するための図である。

【図3】定型表現が文書の分類に与える影響を説明するための図である。

【図4】定型表現を抽出するためのテキストの分類を説明するための図である。

【図5】定型表現を抽出するためのテキストの分類の問題点を説明するための図である。

20

【図6】本実施形態に係る分類装置の機能ブロック図である。

【図7】受付解析部による解析結果の一例を示す図である。

【図8】単語モデルの一例を示す図である。

【図9】テキストの分割を説明するための図である。

【図10】分類結果画面の一例を示す図である。

【図11】本実施形態に係る分類装置として機能するコンピュータの概略構成を示すブロック図である。

【図12】本実施形態における分類処理の一例を示すフローチャートである。

【図13】分割処理の一例を示すフローチャートである。

【図14】クラスタリング処理の一例を示すフローチャートである。

30

【図15】表示制御処理の一例を示すフローチャートである。

【発明を実施するための形態】

【0012】

以下、図面を参照して、開示の技術に係る実施形態の一例を説明する。

【0013】

本実施形態に係る分類装置では、定型表現を抽出するために、テキスト集合に含まれる各テキストを複数のクラスタに分類する。

【0014】

ここで、実施形態の詳細を説明する前に、定型表現を抽出するために、テキストを分類する理由について説明する。例えば、システムに関するインシデント対応時のメールなどの文書を分類し、各文書が表すインシデントが何の事案に関するインシデントかを特定する場合を考える。

40

【0015】

例えば、図1に示すように、インシデント1件における文章の集合を1つの文書とし、複数のインシデントに関する文書集合に含まれる各文書を、B o W (Bag of Words) などの手法により、文書に含まれる単語の出現頻度でベクトル化する。そして、ベクトルのコサイン類似度(図1中の「0.7」、「0.0」、「0.4」)が高い文書同士をクラスタにまとめることにより、文書を分類する。

【0016】

そして、T F (Term Frequency) - I D F (Inverse Document Frequency) などにより

50

、各クラスタに属する文書に含まれる特徴語を抽出し（図 1 中の下線で示す単語）、各クラスタに対応付けることで、各クラスタに含まれる文書が何の事案のインシデントかを把握可能にする。なお、TF 値及び IDF 値は、以下のように定義される。

【0017】

単語 w の TF 値

= 文書中の単語 w の出現数 / 文書中の全ての単語の出現数

単語 w の IDF 値

= $\log$ (文書の総数 / 単語 w を含む文書の数)

【0018】

しかし、上記のベクトル化の際には、各文書に含まれる定型表現もベクトル化されてしまい、定型表現がノイズとなってコサイン類似度に影響を及ぼしてしまう。例えば、図 2 に示すように、質問文 1 は、サーバに関する問い合わせであり、質問文 2 は、ネットワークに関する問い合わせであり、それぞれ質問内容が異なる。しかし、両質問文とも、「お疲れさまです、サービス第一開発部の工藤です。」のような定型表現を含むため、コサイン類似度が高くなってしまう。

【0019】

また、図 3 に示すように、文書が定型表現を含む場合には、文書に出現する単語数が多くなるため、特徴語となるべき単語の TF 値が下がってしまう。

【0020】

特に、システムに関するインシデント対応の分野では、顧客からのメールなどでの問合せの文章に、挨拶や結びの言葉などの定型表現が現れ易い。そこで、定型表現が文書の分類に与える影響を低減するために、各文書から定型表現を削除することが考えられる。しかし、定型表現の中には、「サービス第一開発部の工藤です。」などの固有名詞や、顧客毎の固有の表現等が含まれるため、定型表現を事前に定義しておくことは困難である。

【0021】

そこで、例えば、図 4 に示すように、文書集合に含まれる各文書を、文書中に含まれる「。（句点）」や「¥n（改行コード）」といった一文の区切りを表す箇所で分割することによって一文のテキスト集合を作成する。そして、各文をベクトル化してクラスタリングし、各クラスタに含まれるテキストを確認することで、定型表現が分類されたクラスタを特定し、定型表現を抽出することが考えられる。そして、抽出した定型表現を各文書から削除する。

【0022】

このように、テキストを分類して定型表現を抽出することで、固有の表現等を含む定型表現も抽出することができる。したがって、本実施形態では、定型表現を抽出するために、テキストを分類する。

【0023】

しかし、各文書から作成された 1 文が複文の場合、定型表現を一部に含むにも関わらず、全体としては定型表現とは見做されず、定型表現が分類されるクラスタに分類されない場合がある。例えば、図 5 に示すように、クラスタ 2 に分類されている一文は、クラスタ 1 に分類されている定型表現と同様の「対処方法を教えて下さい」という定型表現を含むが、「予想外のメッセージが出力されている為、」の部分の影響で、クラスタ 1 に分類されない。

【0024】

そこで、本実施形態では、複文に含まれる定型表現も抽出可能に、テキストを分類する。以下、本実施形態の詳細について説明する。

【0025】

図 6 に示すように、本実施形態に係る分類装置 10 は、機能的には、受付解析部 12 と、分割部 14 と、分類部 16 と、表示制御部 18 とを含む。また、分類装置 10 の所定の記憶領域には、単語モデル 24 が記憶される。なお、受付解析部 12 は、開示の技術の受付部の一例であり、分割部 14 は、開示の技術の生成部の一例である。

10

20

30

40

50

【0026】

受付解析部12は、分類装置10に入力されるテキスト集合を受け付ける。例えば、受付解析部12は、システムに関するインシデント対応時のメールなどの文章を含み、インシデント1件における文章の集合を1つの文書とし、文書集合に含まれる各文書を一文のテキストに整形したテキスト集合を受け付ける。一文のテキストの整形は、例えば、文書中に含まれる「。（句点）」や「¥n（改行コード）」といった一文の区切りを表す箇所

【0027】

受付解析部12は、受け付けたテキスト集合に含まれる各テキストに対して形態素解析を行い、テキストを形態素に分割すると共に、各形態素に品詞、形態素情報等の属性情報を付与する。また、受付解析部12は、形態素解析の結果を用いて、各テキストに対して係り受け解析を行い、文節毎の係り受け関係を解析する。

【0028】

図7に、「予想外のメッセージが出力されている為、対処方法を教えて下さい」というテキストに対する、受付解析部12による形態素解析結果及び係り受け解析結果の一例を示す。図7中のAが形態素解析結果である。図7の例では、テキストに含まれる形態素毎に、テキストの先頭から順に番号を付すと共に、形態素解析の結果得られた属性情報が各形態素に対応付けられている。また、図7中のBが係り受け解析結果である。図7の例では、形態素解析結果に基づいて、テキストを各文節に区切り、各文節間の係り受け関係を解析した例であり、各文節をボックスで、文節間の係り受け関係を矢印で表している。

【0029】

分割部14は、受付解析部12により受け付けられたテキスト集合に含まれるテキストのうち、いずれかのテキストが特定の条件を満たす区切れ箇所を含むことを検出すると、そのテキストを区切れ箇所

【0030】

具体的には、分割部14は、各テキストについての受付解析部12による解析結果に基づいて、各テキストを所定の区切れ箇所

【0031】

分割部14は、テキストを所定の区切れ箇所

【0032】

より具体的には、分割部14は、単語モデル24に含まれるIDF値テーブル24Aを参照して、テキストに含まれる各単語のIDF値を取得する。

【0033】

ここで、図8に、単語モデル24の一例を示す。単語モデル24は、IDF値テーブル24Aと、単語ベクトルテーブル24Bとを含む。IDF値テーブル24Aは、単語とその単語のIDF値とを対応付けて記憶したテーブルである。IDF値は、予め任意の文書集合から生成しておいてもよいし、入力されたテキスト集合の元になった文書集合から生成してもよい。単語ベクトルテーブル24Bについては後述する。

【0034】

分割部14は、IDF値テーブル24Aから取得した単語毎のIDF値を用いて、各テキストを区切れ箇所

10

20

30

40

50

それぞれ単文として、単文集合に入れる。

【0035】

分類部16は、テキスト集合に含まれるテキストのうち、特定の条件を満たす区切れ箇所を含まないテキスト、及び生成した新たな複数のテキスト、すなわち、単文集合に含まれる単文の各々を、複数のクラスタに分類する。

【0036】

具体的には、分類部16は、各単文を、単語モデル24の単語ベクトルテーブル24Bを用いてベクトル化する。単語ベクトルテーブル24Bは、単語と、その単語をTF-IDFやword2vec等によりベクトル表現した単語ベクトルとを対応付けて記憶したテーブルである。

【0037】

分類部16は、各単文の単語ベクトルのコサイン類似度などを用いて、k-meansやsimple linkageなどの従来既知のクラスタリング手法により、各単文をクラスタリングする。

【0038】

また、分類部16は、複数のクラスタの各々に分類された単文に含まれる単語の出現状況に基づいて、複数のクラスタの各々から特徴語を抽出し、抽出した特徴語を、各クラスタに対応付ける。なお、単語の出現状況としては、TF-IDFなどを用いることができる。また、特徴語は、開示の技術の特徴情報及び代表形態素の一例である。

【0039】

表示制御部18は、テキスト集合における各単文の出現状況に関する指標に基づいて、複数のクラスタの各々に含まれる単文についての指標が、出現頻度が高いことを示す順にクラスタを並べて、表示装置（図示省略）に表示する。

【0040】

例えば、表示制御部18は、単語モデル24に含まれるIDF値テーブル24Aを参照して、各単文に含まれる各単語のIDF値を取得し、各単文のIDF値ベクトルのノルムを算出する。また、表示制御部18は、クラスタ毎に、そのクラスタに含まれる単文の各々のIDF値ベクトルのノルムの平均を算出する。そして、表示制御部18は、IDF値ベクトルのノルムの平均が小さい順にクラスタをソートし、表示装置に表示する。IDF値ベクトルのノルムの平均が小さいクラスタは、そのクラスタに含まれる単文が、テキスト集合において横断的に出現していることを表しているため、定型表現が分類されたクラスタであると思われるものである。

【0041】

図10に、表示装置に表示される分類結果画面30の一例を示す。図10の例では、各クラスタを1つの枠で囲み、枠内に、そのクラスタに含まれる単文を表示している。また、各クラスタには、分類部16により、そのクラスタに対応付けられた特徴語を対応付けて表示している。図10では、サーバやネットワークなどのインシデントに関する具体的な内容を示す単文が分類されたクラスタよりも上位に、定型表現が分類されたクラスタが表示されている例を示している。

【0042】

なお、分類結果画面30は、図10の例に限定されない。例えば、クラスタに対応付けられた特徴語のみをソート順に表示装置に表示し、画面からその特徴語を選択することで、その特徴語が表すクラスタに含まれる単文を表示するような表示形態としてもよい。

【0043】

分類装置10は、例えば図11に示すコンピュータ40で実現することができる。コンピュータ40は、CPU（Central Processing Unit）41と、一時記憶領域としてのメモリ42と、不揮発性の記憶部43とを備える。また、コンピュータ40は、入力装置、表示装置等の入出力装置44と、記憶媒体49に対するデータの読み込み及び書き込みを制御するR/W（Read/Write）部45とを備える。また、コンピュータ40は、インターネット等のネットワークに接続される通信I/F46を備える。CPU41、メモリ42

10

20

30

40

50

、記憶部 43、入出力装置 44、R/W部 45、及び通信 I/F 46は、バス 47を介して互いに接続される。

【0044】

記憶部 43は、HDD (Hard Disk Drive)、SSD (Solid State Drive)、フラッシュメモリ等によって実現できる。記憶媒体としての記憶部 43には、コンピュータ 40を、分類装置 10として機能させるための分類プログラム 50が記憶される。分類プログラム 50は、受付解析プロセス 52と、分割プロセス 54と、分類プロセス 56と、表示制御プロセス 58とを有する。また、記憶部 43は、単語モデル 24を構成する情報が記憶される情報記憶領域 60を有する。

【0045】

CPU 41は、分類プログラム 50を記憶部 43から読み出してメモリ 42に展開し、分類プログラム 50が有するプロセスを順次実行する。CPU 41は、受付解析プロセス 52を実行することで、図 6に示す受付解析部 12として動作する。また、CPU 41は、分割プロセス 54を実行することで、図 6に示す分割部 14として動作する。また、CPU 41は、分類プロセス 56を実行することで、図 6に示す分類部 16として動作する。また、CPU 41は、表示制御プロセス 58を実行することで、図 6に示す表示制御部 18として動作する。また、CPU 41は、情報記憶領域 60から情報を読み出して、単語モデル 24をメモリ 42に展開する。これにより、分類プログラム 50を実行したコンピュータ 40が、分類装置 10として機能することになる。なお、プログラムを実行する CPU 41はハードウェアである。

【0046】

なお、分類プログラム 50により実現される機能は、例えば半導体集積回路、より詳しくはASIC (Application Specific Integrated Circuit) 等で実現することも可能である。

【0047】

次に、本実施形態に係る分類装置 10の作用について説明する。

【0048】

例えば、システムに関するインシデント対応時のメールなどの文章を含み、インシデント 1件における文章の集合を 1つの文書とし、文書集合に含まれる各文書を一文のテキストに整形したテキスト集合が、分類装置 10へ入力される。そして、分類装置 10において、図 12に示す分類処理が実行される。なお、分類処理は、開示の技術の分類方法の一例である。

【0049】

図 12に示す分類処理のステップ S10で、受付解析部 12が、分類装置 10に入力されたテキスト集合 Sを受け付ける。テキスト集合 Sは、テキスト 1、テキスト 2、・・・、テキスト N (Nはテキスト集合 Sに含まれるテキストの数) を含む。

【0050】

次に、ステップ S20で、図 13に示す分割処理が実行される。

【0051】

図 13に示す分割処理のステップ S22で、単文集合 Pとして空集合を用意し、次のステップ S24で、テキストを識別するための変数 s に 1を設定する。

【0052】

次に、ステップ S26で、受付解析部 12が、テキスト s に対して形態素解析を行い、テキストを形態素に分割すると共に、各形態素に属性情報を付与する。また、受付解析部 12が、形態素解析の結果を用いて、テキスト s に対して係り受け解析を行い、文節毎の係り受け関係を解析する。

【0053】

次に、ステップ S28で、分割部 14が、上記ステップ S26での解析結果に基づいて、テキスト s を「、(読点)」などの所定の区切れ箇所で、後半部分 s __ 1 と前半部分 s __ 2 とに区切る。また、分割部 14は、単語モデル 24に含まれる IDF 値テーブル 24

10

20

30

40

50

Aを参照して、テキストに含まれる各単語のIDF値を取得する。そして、分割部14は、IDF値テーブル24Aから取得した単語毎のIDF値を用いて、後半部分s__1及び前半部分s__2の各々について、IDF値のベクトルのノルムを算出する。

【0054】

次に、ステップS30で、分割部14が、後半部分s__1のIDF値のノルムv__1、及び前半部分s__2のIDF値のノルムv__2のいずれか一方が予め定めた閾値TH以下か否かを判定する。v__1及びv__2のいずれか一方が閾値TH以下の場合、処理はステップS32へ移行し、v__1及びv__2共に閾値TH以下、又はv__1及びv__2共に閾値THより大きい場合には、処理はステップS34へ移行する。

【0055】

ステップS32では、分割部14が、テキストsを単文s__1とs__2とに分割し、単文集合Pに追加する。一方、ステップS34では、テキストsをそのまま単文集合Pに追加する。

【0056】

次に、ステップS38で、受付解析部12が、sがNか否かを判定することにより、受け付けたテキスト集合Sに含まれるテキストの全てについて、上記ステップS26～S32又はS34の処理が終了したか否かを判定する。sがまだNに達していない場合には、処理はステップS36へ移行し、受付解析部12がsを1インクリメントして、処理はステップS26に戻る。s=Nの場合には、分割処理は終了して、分類処理に戻る。

【0057】

次に、図12に示す分類処理のステップS50で、図14に示すクラスタリング処理が実行される。

【0058】

図14に示すクラスタリング処理のステップS52で、分類部16が、単文集合Pに含まれる各単文を、単語モデル24の単語ベクトルテーブル24Bを用いてベクトル化する。

【0059】

次に、ステップS54で、分類部16が、各単文の単語ベクトルのコサイン類似度などを用いて、k-meansやsimple linkageなどの従来既知のクラスタリング手法により、各単文をクラスタリングする。

【0060】

次に、ステップS56で、分類部16が、複数のクラスタの各々に分類された単文に含まれる単語のTF-IDFなどの出現状況を示す指標に基づいて、複数のクラスタの各々から特徴語を抽出し、抽出した特徴語を各クラスタに対応付ける。そして、クラスタリング処理は終了して、分類処理に戻る。

【0061】

次に、図12に示す分類処理のステップS60で、図15に示す表示制御処理が実行される。

【0062】

図15に示す表示制御処理のステップS62で、表示制御部18が、単語モデル24に含まれるIDF値テーブル24Aを参照して、各単文に含まれる各単語のIDF値を取得し、各単文のIDF値ベクトルのノルムを算出する。

【0063】

次に、ステップS64で、表示制御部18が、クラスタ毎に、そのクラスタに含まれる単文の各々のIDF値ベクトルのノルムの平均を算出する。

【0064】

次に、ステップS66で、表示制御部18が、IDF値ベクトルのノルムの平均が小さい順にクラスタをソートし、例えば、図10に示すような分類結果画面30を表示装置に表示する。そして、表示制御処理は終了し、分類処理も終了する。

【0065】

10

20

30

40

50

以上説明したように、本実施形態に係る分類装置によれば、テキスト集合に含まれる各テキストを特定の区切り箇所分割した上で、単語ベクトルのコサイン類似度などに基づいてクラスタリングして分類する。これにより、テキストが複文で、その一部に定型表現を含む場合でも、定型表現を抽出するためのテキストの分類精度を向上させることができる。

【0066】

また、テキストを読点等の区切れ箇所区切った前半部分及び後半部分のいずれか一方のIDF値ベクトルのノルムが、予め定めた閾値以下の場合に、その区切れ箇所を特定の区切り箇所とする。これにより、定型表現を抽出するためのテキストの分類精度をより向上させることができる。

10

【0067】

なお、上記実施形態では、クラスタに含まれる単文の各々のIDF値ベクトルのノルムの平均が小さい順にクラスタをソートする場合について説明したが、これに限定されない。例えば、各クラスタに分類された単文の数が多い順にクラスタをソートしてもよい。定型表現は、テキスト集合での出現頻度が高いことが想定されるため、多くの単文が含まれるクラスタは、定型表現が分類されたクラスタであると見做すものである。

【0068】

また、上記実施形態では、テキストの区切れ箇所を、テキストの形態素解析結果及び係り受け解析結果に基づいて特定する場合について説明したが、これに限定されない。例えば、予め定めた文字列の前又は後を区切れ箇所にするなど、形態素解析結果及び係り受け解析結果に基づいて特定可能な区切れ箇所を用いるようにしてもよい。この場合、受付解析部における形態素解析及び係り受け解析の処理は省略することができる。また、解析済みのテキスト集合を受け付けることにより、受付解析部における形態素解析及び係り受け解析の処理を省略するようにしてもよい。

20

【0069】

また、上記実施形態では、システムのインシデント対応に関する文書を整形したテキスト集合を入力する場合について説明したが、これに限定されず、開示の技術は、様々な文書に適用可能である。特に、文書内に定型的な表現が多く含まれる文書に対して、開示の技術は有効である。

【0070】

また、上記実施形態では、分類プログラムが記憶部に予め記憶（インストール）されている態様を説明したが、これに限定されない。開示の技術に係るプログラムは、CD-ROM、DVD-ROM、USBメモリ等の記憶媒体に記憶された形態で提供することも可能である。

30

【0071】

以上の実施形態に関し、更に以下の付記を開示する。

【0072】

（付記1）

複数のテキストを受け付け、

受け付けた前記複数のテキストのうち、いずれかのテキストが特定の条件を満たす区切れ箇所を含むことを検出すると、前記いずれかのテキストを前記区切れ箇所分割して、新たな複数のテキストを生成し、

40

受け付けた前記複数のテキストのうち、前記特定の条件を満たす区切れ箇所を含まないテキストと、生成した前記新たな複数のテキストと、を複数のクラスタに分類する、

処理をコンピュータが実行することを特徴とする分類方法。

【0073】

（付記2）

前記複数のクラスタそれぞれに分類されたテキストに基づき、前記複数のクラスタそれぞれの特徴情報を生成し、

生成した前記特徴情報を、前記複数のクラスタそれぞれに対応付けて表示部に表示する

50

、
ことを特徴とする付記 1 に記載の分類方法。

【0074】

(付記 3)

前記複数のクラスタについて、前記複数のクラスタそれぞれに分類されたテキストと、
該テキストに含まれる複数の形態素の、受け付けた前記複数のテキストでの出現状況とに
基づき、前記複数のクラスタをそれぞれ代表する複数の代表形態素を決定し、

決定した前記複数の代表形態素それぞれを、前記複数の代表形態素それぞれが代表する
複数のクラスタそれぞれに対応付けて前記表示部に表示する、

ことを特徴とする付記 2 に記載の分類方法。

10

【0075】

(付記 4)

決定した前記複数の代表形態素それぞれを、前記複数のクラスタそれぞれに分類された
前記テキストの数に応じた順序で並べて前記表示部に表示する、

ことを特徴とする付記 3 に記載の分類方法。

【0076】

(付記 5)

前記複数のテキストにおける各テキストの出現状況に関する指標に基づいて、前記複数
のクラスタの各々に含まれるテキストについての前記指標が、出現頻度が高いことを示す
順に前記クラスタを並べて前記表示部に表示する

20

ことを特徴とする付記 2 ～付記 4 のいずれか 1 項に記載の分類方法。

【0077】

(付記 6)

前記特定の条件を満たす区切れ箇所を読点とする付記 1 ～付記 5 のいずれか 1 項に記載
の分類方法。

【0078】

(付記 7)

前記テキストを区切れ箇所で区切った前半部分及び後半部分のいずれか一方の、前記複
数のテキストにおける出現状況が予め定めた条件を満たす場合に、前記区切れ箇所を前記
特定の条件を満たす区切れ箇所とする付記 1 ～付記 5 のいずれか 1 項に記載の分類方法。

30

【0079】

(付記 8)

複数のテキストを受け付ける受付部と、

前記受付部により受け付けられた前記複数のテキストのうち、いずれかのテキストが特
定の条件を満たす区切れ箇所を含むことを検出すると、前記いずれかのテキストを前記区
切れ箇所で分割して、新たな複数のテキストを生成する生成部と、

前記受付部により受け付けられた前記複数のテキストのうち、前記特定の条件を満たす
区切れ箇所を含まないテキストと、前記生成部により生成された前記新たな複数のテキス
トと、を複数のクラスタに分類する分類部と、

を含むことを特徴とする分類装置。

40

【0080】

(付記 9)

前記分類部は、前記複数のクラスタそれぞれに分類されたテキストに基づき、前記複数
のクラスタそれぞれの特徴情報を生成し、

前記分類部により生成された前記特徴情報を、前記複数のクラスタそれぞれに対応付け
て表示部に表示する表示制御部を更に含む

ことを特徴とする付記 8 に記載の分類装置。

【0081】

(付記 10)

前記分類部は、前記複数のクラスタについて、前記複数のクラスタそれぞれに分類され

50

たテキストと、該テキストに含まれる複数の形態素の、受け付けた前記複数のテキストでの出現状況とに基づき、前記複数のクラスタをそれぞれ代表する複数の代表形態素を決定し、

前記表示制御部は、前記分類部により決定された前記複数の代表形態素それぞれを、前記複数の代表形態素それぞれが代表する複数のクラスタそれぞれに対応付けて前記表示部に表示する、

ことを特徴とする付記 9 に記載の分類装置。

【0082】

(付記 11)

前記表示制御部は、前記分類部により決定された前記複数の代表形態素それぞれを、前記複数のクラスタそれぞれに分類された前記テキストの数に応じた順序で並べて前記表示部に表示する、

ことを特徴とする付記 10 に記載の分類装置。

【0083】

(付記 12)

前記表示制御部は、前記複数のテキストにおける各テキストの出現状況に関する指標に基づいて、前記複数のクラスタの各々に含まれるテキストについての前記指標が、出現頻度が高いことを示す順に前記クラスタを並べて前記表示部に表示する

ことを特徴とする付記 9 ～付記 11 のいずれか 1 項に記載の分類装置。

【0084】

(付記 13)

前記生成部は、前記特定の条件を満たす区切れ箇所を読点とする付記 8 ～付記 12 のいずれか 1 項に記載の分類装置。

【0085】

(付記 14)

前記生成部は、前記テキストを区切れ箇所で区切った前半部分及び後半部分のいずれか一方の、前記複数のテキストにおける出現状況が予め定めた条件を満たす場合に、前記区切れ箇所を前記特定の条件を満たす区切れ箇所とする付記 8 ～付記 12 のいずれか 1 項に記載の分類装置。

【0086】

(付記 15)

複数のテキストを受け付け、

受け付けた前記複数のテキストのうち、いずれかのテキストが特定の条件を満たす区切れ箇所を含むことを検出すると、前記いずれかのテキストを前記区切れ箇所で分割して、新たな複数のテキストを生成し、

受け付けた前記複数のテキストのうち、前記特定の条件を満たす区切れ箇所を含まないテキストと、生成した前記新たな複数のテキストと、を複数のクラスタに分類する、

処理をコンピュータに実行させることを特徴とする分類プログラム。

【0087】

(付記 16)

前記複数のクラスタそれぞれに分類されたテキストに基づき、前記複数のクラスタそれぞれの特徴情報を生成し、

生成した前記特徴情報を、前記複数のクラスタそれぞれに対応付けて表示部に表示する、

ことを特徴とする付記 15 に記載の分類プログラム。

【0088】

(付記 17)

前記複数のクラスタについて、前記複数のクラスタそれぞれに分類されたテキストと、該テキストに含まれる複数の形態素の、受け付けた前記複数のテキストでの出現状況とに基づき、前記複数のクラスタをそれぞれ代表する複数の代表形態素を決定し、

10

20

30

40

50

決定した前記複数の代表形態素それぞれを、前記複数の代表形態素それぞれが代表する複数のクラスタそれぞれに対応付けて前記表示部に表示する、

ことを特徴とする付記 16 に記載の分類プログラム。

【0089】

(付記 18)

決定した前記複数の代表形態素それぞれを、前記複数のクラスタそれぞれに分類された前記テキストの数に応じた順序で並べて前記表示部に表示する、

ことを特徴とする付記 17 に記載の分類プログラム。

【0090】

(付記 19)

前記複数のテキストにおける各テキストの出現状況に関する指標に基づいて、前記複数のクラスタの各々に含まれるテキストについての前記指標が、出現頻度が高いことを示す順に前記クラスタを並べて前記表示部に表示する

ことを特徴とする付記 16 ～付記 18 のいずれか 1 項に記載の分類プログラム。

【0091】

(付記 20)

複数のテキストを受け付け、

受け付けた前記複数のテキストのうち、いずれかのテキストが特定の条件を満たす区切れ箇所を含むことを検出すると、前記いずれかのテキストを前記区切れ箇所分割して、新たな複数のテキストを生成し、

受け付けた前記複数のテキストのうち、前記特定の条件を満たす区切れ箇所を含まないテキストと、生成した前記新たな複数のテキストと、を複数のクラスタに分類する、

処理をコンピュータに実行させることを特徴とする分類プログラムを記憶した記憶媒体。

【符号の説明】

【0092】

- 10 分類装置
- 12 受付解析部
- 14 分割部
- 16 分類部
- 18 表示制御部
- 24 単語モデル
- 24A IDF 値テーブル
- 24B 単語ベクトルテーブル
- 30 分類結果画面
- 40 コンピュータ
- 41 CPU
- 42 メモリ
- 43 記憶部
- 49 記憶媒体
- 50 分類プログラム

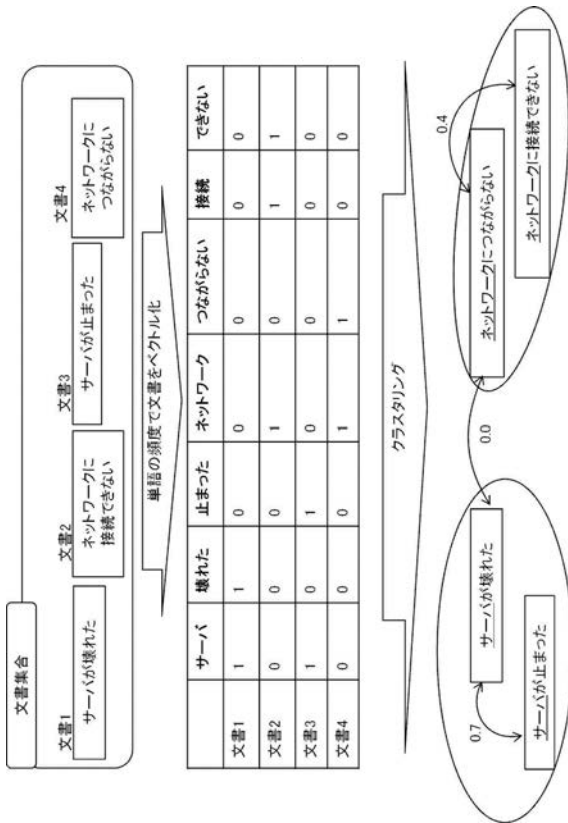
10

20

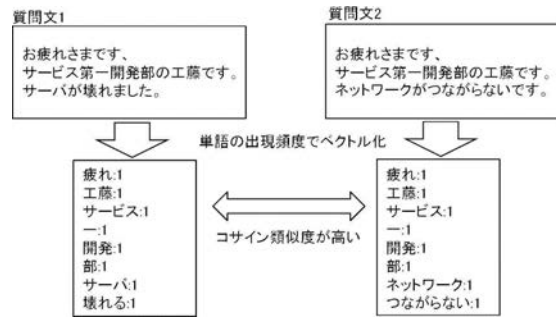
30

40

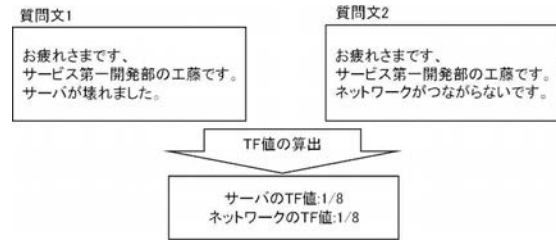
【図 1】



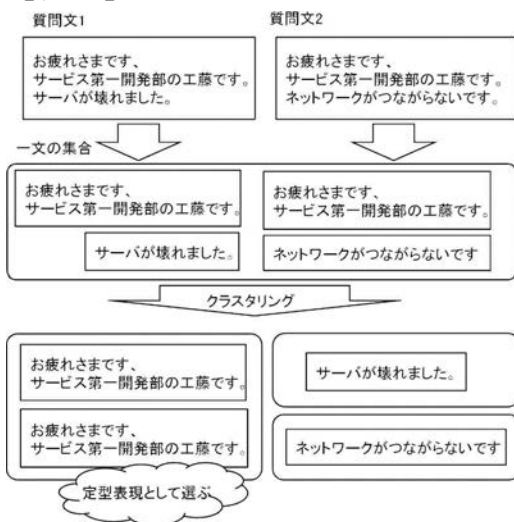
【図 2】



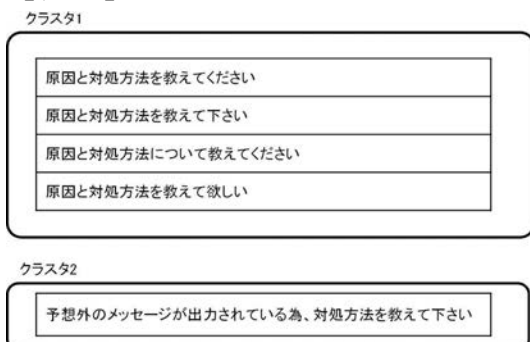
【図 3】



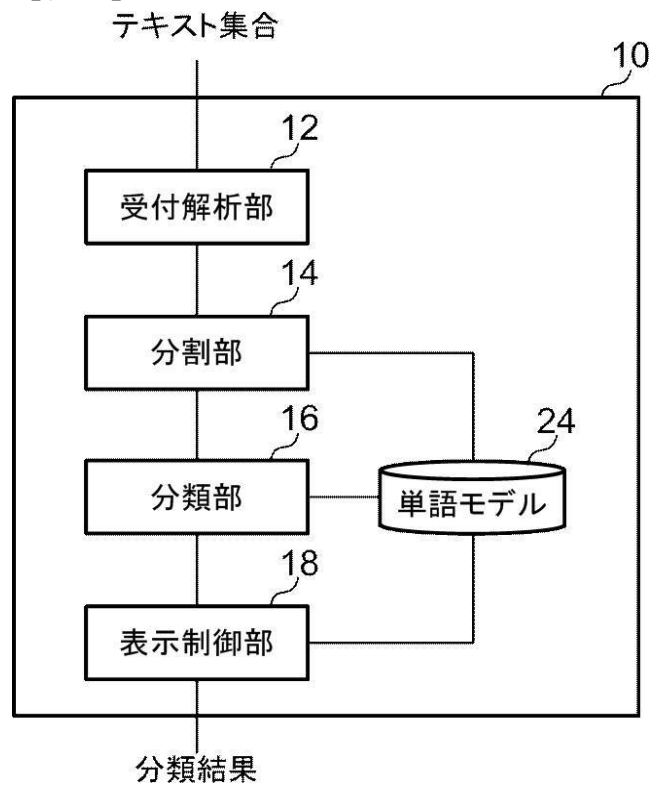
【図 4】



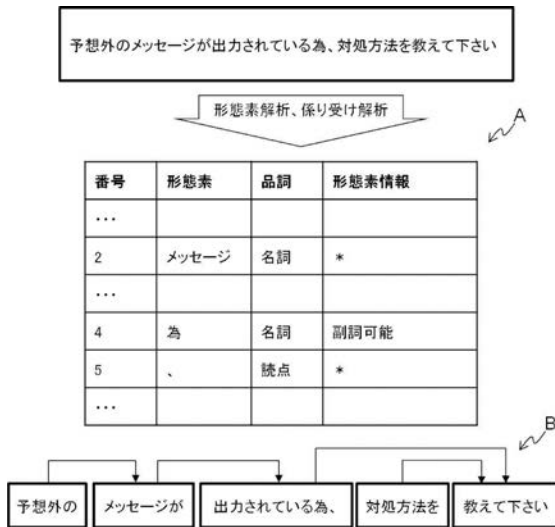
【図 5】



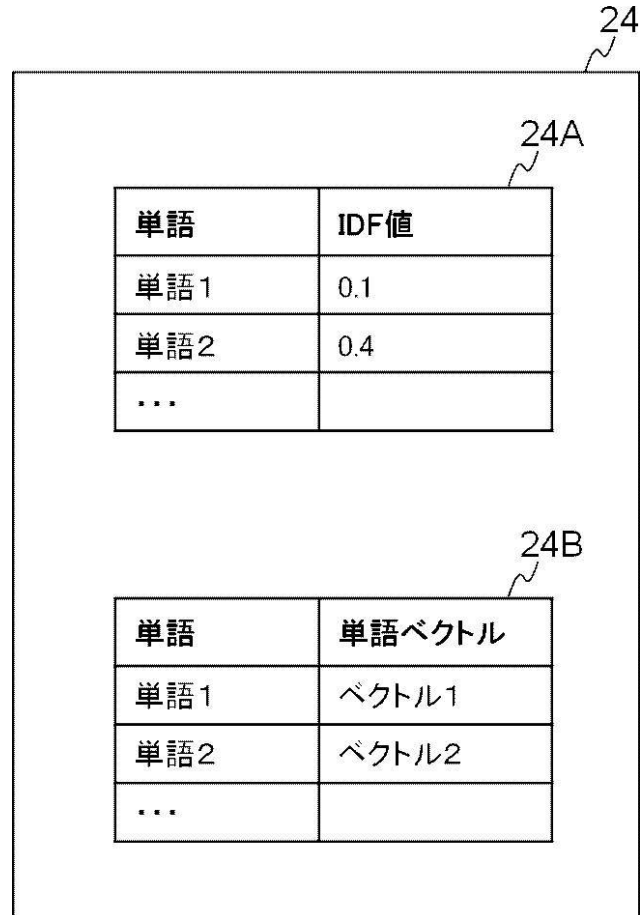
【図 6】



【図 7】



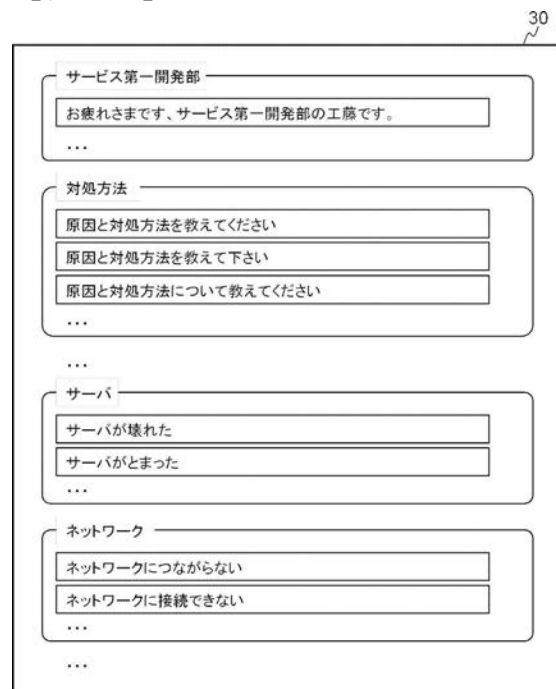
【図 8】



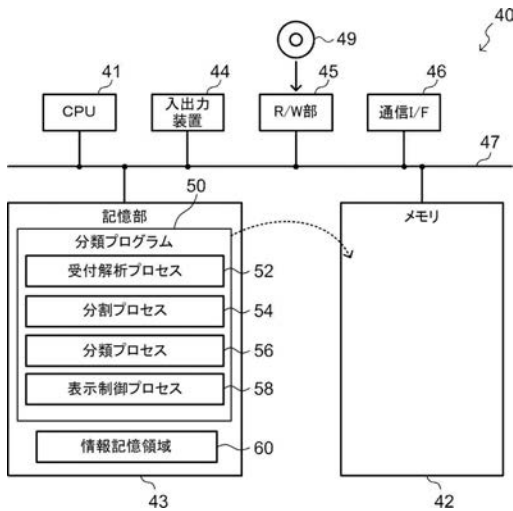
【図 9】



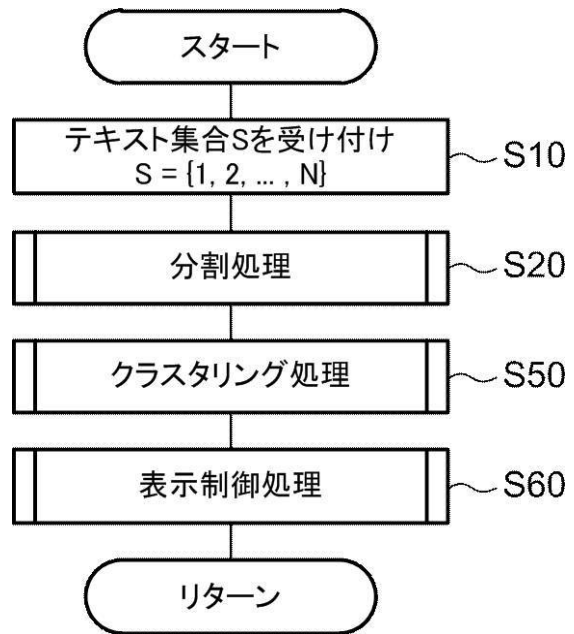
【図 10】



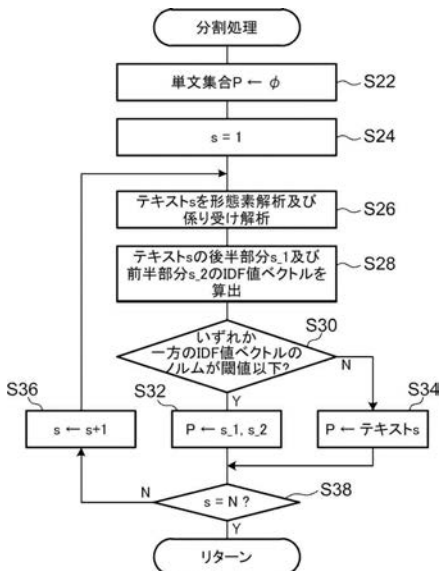
【図 1 1】



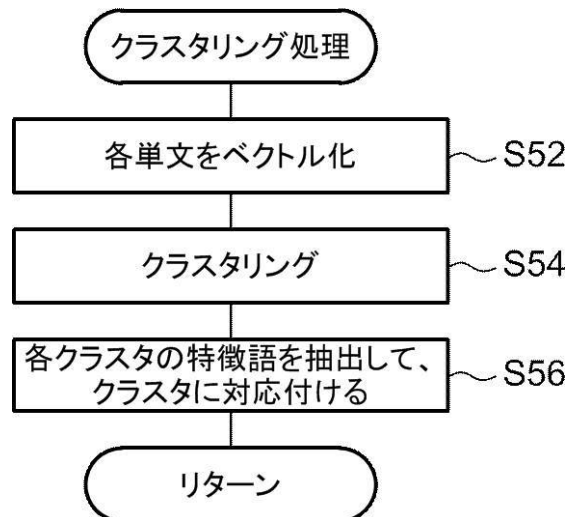
【図 1 2】



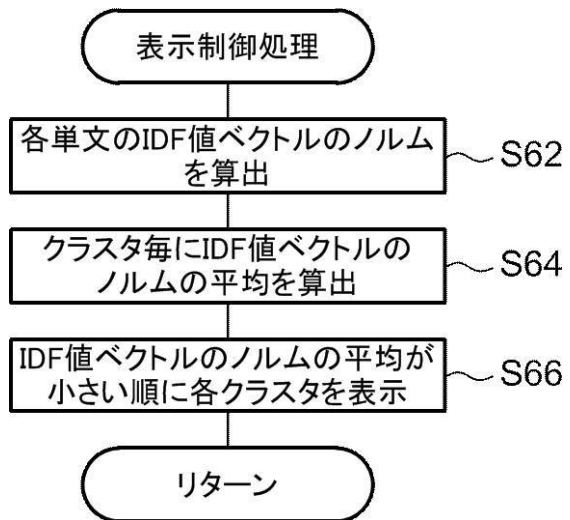
【図 1 3】



【図 1 4】



【図 15】



フロントページの続き

- (72)発明者 埴 大紀
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
- (72)発明者 宮城 俊秀
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
- (72)発明者 山越 幸太
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
- (72)発明者 廣田 佳祐
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
- Fターム(参考) 5B109 QA02 QA03 VC01