

(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2010-9254

(P2010-9254A)

(43) 公開日 平成22年1月14日(2010.1.14)

(51) Int.Cl.	F I	テーマコード (参考)
G 0 6 F 17/30 (2006.01)	G O 6 F 17/30 2 1 O B	5 B O 7 5
	G O 6 F 17/30 1 7 O A	
	G O 6 F 17/30 3 4 O B	

審査請求 有 請求項の数 5 O L (全 14 頁)

(21) 出願番号	特願2008-166566 (P2008-166566)	(71) 出願人	500257300
(22) 出願日	平成20年6月25日 (2008. 6. 25)		ヤフー株式会社
			東京都港区赤坂9丁目7番1号
		(74) 代理人	100106002
			弁理士 正林 真之
		(72) 発明者	新田 清
			東京都港区六本木六丁目10番1号 ヤフー株式会社内
		Fターム(参考)	5B075 ND03 NK32 PR04 UU40

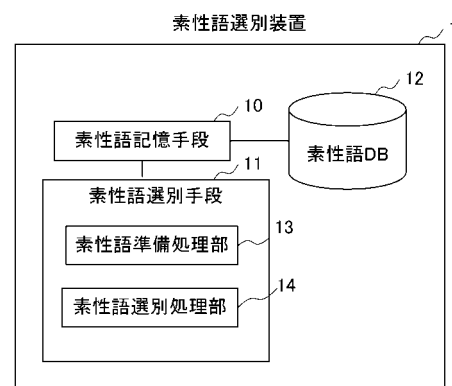
(54) 【発明の名称】 素性語選別装置、方法及びプログラム

(57) 【要約】 (修正有)

【課題】素性語を所定のルールに基づいて選別することにより、ウェブページにカテゴリを割り付ける作業負担を軽減できる素性語選別装置を提供する。

【解決手段】通信ネットワークN上に配信されるウェブページを特徴付ける素性語を、ウェブページに関連付けて記憶する素性語記憶手段10と、素性語記憶手段10に記憶されている素性語毎に、素性語を検索キーワードとした検索エンジンによる検索によって、素性語に関連付けて記憶されているウェブページが合致した頻度を示すヒット回数をカウントし、ヒット回数が、所定の閾値以上又は以下となった際に、素性語を素性語記憶手段10から削除することにより、素性語記憶手段10に記憶されている素性語を選別する素性語選別手段11と、を備える。

【選択図】 図2



【特許請求の範囲】

【請求項 1】

通信ネットワーク上に配信されるウェブページを特徴付ける素性語を、前記ウェブページに関連付けて記憶する素性語記憶手段と、

前記素性語記憶手段に記憶されている前記素性語毎に、前記素性語を検索キーワードとした検索エンジンによる検索によって、前記素性語に関連付けて前記素性語記憶手段に記憶されている前記ウェブページが合致した頻度を示すヒット回数をカウントし、前記ヒット回数が、所定の閾値以上又は以下となった際に、前記素性語を前記素性語記憶手段から削除することにより、前記素性語記憶手段に記憶されている前記素性語を選別する素性語選別手段と、

10

を備えることを特徴とする素性語選別装置。

【請求項 2】

前記素性語選別手段は、

前記素性語が、前記素性語記憶手段においてその素性語に関連付けて記憶されている前記ウェブページ中に出現する出現頻度をカウントし、前記出現頻度が所定の閾値以下となった際に、前記素性語を前記素性語記憶手段から削除することにより、前記素性語記憶手段に記憶されている前記素性語を選別することを特徴とする、請求項 1 に記載の素性語選別装置。

【請求項 3】

前記素性語選別手段は、

前記素性語が、前記素性語記憶手段においてその素性語に関連付けて記憶されている前記ウェブページのタイトル又は本文又は概要中に出現する出現頻度をカウントし、その出現頻度が所定の閾値以下となった際に、前記素性語を前記素性語記憶手段から削除することにより、前記素性語記憶手段に記憶されている前記素性語を選別することを特徴とする、請求項 1 に記載の素性語選別装置。

20

【請求項 4】

コンピュータが、

予め前記コンピュータに記憶され、通信ネットワーク上に配信されるウェブページを特徴付ける素性語毎に、前記素性語を検索キーワードとした検索エンジンによる検索によって、前記素性語に関連付けて記憶されている前記ウェブページが合致した頻度を示すヒット回数をカウントし、前記ヒット回数が、所定の閾値以上又は以下となった際に、前記素性語を削除することにより、前記素性語を選別する素性語選別ステップを、少なくとも実行することを特徴とする方法。

30

【請求項 5】

請求項 4 に記載の方法をコンピュータに実行させることを特徴とするプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、通信ネットワーク上に配信されるウェブページを、予め定められたカテゴリ別に分類するための特徴語を選別する素性語選別装置、方法及びプログラムに関する。

40

【背景技術】

【0002】

近年、インターネット上に無数のウェブページが存在しており、目的にあったウェブページを検索することは容易ではない。そのため、検索エンジンが提供されているが、このような検索エンジンを使用したとしても、目的とするウェブページが検索できるとは限らない。

したがって、YAHOO! JAPAN（登録商標）等、いくつかのサービス業者は、検索エンジンの他に、下記非特許文献 1 に示すように、ウェブページをカテゴリ別に分類する分類サービスを行っており、これにより、ユーザは、目的とするウェブページを見つけやすくなっている。

50

【非特許文献1】[online]、2008年4月21月号、YAHOO! JAPAN
カテゴリ、[2008年4月23日検索]、インターネット<URL: http://dir.yahoo.co.jp/>

【発明の開示】

【発明が解決しようとする課題】

【0003】

このようなカテゴリ別に分類する分類サービスでは、カテゴリの規模が大きい（数万ノードの）場合は手作業でウェブページを分類している。カテゴリ規模が小さい（数百ノードの）場合は自動的な分類処理が多数試みられており、インターネット上に存在するウェブページに含まれる語句と、予め記憶された素性語とを対比して、その対比結果に基づいて、予め記憶された素性語に対応するカテゴリを付与している。

10

カテゴリの規模を大きくすると、ユーザはより意図に近い分類カテゴリを指定することができる。しかしながら、カテゴリの規模を大きくすると所属する文書の数が増え、分類処理時に取り扱う素性語の数も増加し、ウェブページにマッチするカテゴリを割り付ける作業が複雑となり時間を要することになる。また、そのような作業に耐えうる分類装置を導入するにはコストも高くなる。

【0004】

本発明は、このような事情を考慮して提案されるものであり、予め記憶された素性語を所定のルールに基づいて選別することにより、ウェブページにカテゴリを割り付ける作業負担を軽減することができる素性語選別装置、方法及びプログラムを提供することを目的とする。

20

【課題を解決するための手段】

【0005】

上記目的を達成するために、本発明者は、予め記憶された素性語を所定のルールに基づいて選別できる仕組みを見出し、本発明を想到するに至った。

本発明に係る素性語選別装置は、予め定められた一定のルールに従って素性語を素性語データベースから削除することにより、予め記憶されている素性語に基づいたカテゴリ分類処理に要する負担を軽減するものである。

【0006】

(1) 通信ネットワーク上に配信されるウェブページを特徴付ける素性語を、前記ウェブページに関連付けて記憶する素性語記憶手段と、

30

前記素性語記憶手段に記憶されている前記素性語毎に、前記素性語を検索キーワードとした検索エンジンによる検索によって、前記素性語に関連付けて前記素性語記憶手段に記憶されている前記ウェブページが合致した頻度を示すヒット回数をカウントし、前記ヒット回数が、所定の閾値以上又は以下となった際に、前記素性語を前記素性語記憶手段から削除することにより、前記素性語記憶手段に記憶されている前記素性語を選別する素性語選別手段と、

を備えることを特徴とする素性語選別装置。

【0007】

(1)の発明によれば、検索エンジンによる検索によって、素性語に関連付けて素性語記憶手段に記憶されているウェブページにヒットしたヒット回数が、所定の閾値以上又は以下となった際に、その素性語を素性語記憶手段から削除するので、以下の効果を奏する。

40

【0008】

すなわち、前記ヒット回数が所定の閾値以上となった際に、素性語を素性語記憶手段から削除するので、その素性語はありふれた用語、つまり、ウェブページを特徴づけることができない用語であると判断して、その素性語を基にしたカテゴリ分類を不可能にすることができる。

【0009】

また、前記ヒット回数が所定の閾値以下となった際に、素性語を素性語記憶手段から削

50

除するので、その素性語は一般ユーザに連想されない用語、つまり、ウェブページを特徴づけることができない用語であると判断して、その素性語を基にしたカテゴリ分類を不可能にすることができる。

【0010】

そのため、真にウェブページを特徴づけるような素性語のみが素性語記憶手段に残留するため、素性語に基づいたカテゴリ分類を効率的に行うことができ、カテゴリ分類装置の導入コストを低く抑えることができる。

【0011】

(2) 前記素性語選別手段は、前記素性語が、前記素性語記憶手段においてその素性語に関連付けて記憶されている前記ウェブページ中に出現する出現頻度をカウントし、前記出現頻度が所定の閾値以下となった際に、前記素性語を前記素性語記憶手段から削除することにより、前記素性語記憶手段に記憶されている前記素性語を選別することを特徴とする、(1)に記載の素性語選別装置。

10

【0012】

(2)の発明によれば、素性語に関連付けて記憶されているウェブページ中に出現する出現頻度が所定の閾値以下となった際に、その素性語を素性語記憶手段から削除するので、出現頻度が閾値以下である素性語は、ウェブユーザの作成者にとって連想されない用語、つまり、ウェブページを特徴づけることができない用語であると判断して、その素性語を基にしたカテゴリ分類を不可能にすることができる。

【0013】

20

そのため、真にウェブページを特徴づけるような素性語のみが素性語記憶手段に残留するため、素性語に基づいたカテゴリ分類を効率的に行うことができ、カテゴリ分類装置の導入コストを低く抑えることができる。

【0014】

(3) 前記素性語選別手段は、前記素性語が、前記素性語記憶手段においてその素性語に関連付けて記憶されている前記ウェブページのタイトル又は本文又は概要中に出現する出現頻度をカウントし、その出現頻度が所定の閾値以下となった際に、前記素性語を前記素性語記憶手段から削除することにより、前記素性語記憶手段に記憶されている前記素性語を選別することを特徴とする、(1)に記載の素性語選別装置。

【0015】

30

(3)の発明によれば、素性語に関連付けて記憶されているウェブページのタイトル中に出現する出現頻度が所定の閾値以下となった際に、素性語を素性語記憶手段から削除する。一般に、タイトル中の出現頻度が閾値以下である素性語は、ウェブページ中のその他の項目よりも更にウェブページ作成者の印象に依存するため、そのようなタイトル中の出現頻度が閾値以下である素性語を削除することにより、ウェブページを特徴づけるような素性語のみを素性語記憶手段に残すことができる。

【0016】

(4) コンピュータが、

予め前記コンピュータに記憶され、通信ネットワーク上に配信されるウェブページを特徴付ける素性語毎に、前記素性語を検索キーワードとした検索エンジンによる検索によって、前記素性語に関連付けて記憶されている前記ウェブページが合致した頻度を示すヒット回数をカウントし、前記ヒット回数が、所定の閾値以上又は以下となった際に、前記素性語を削除することにより、前記素性語を選別する素性語選別ステップを、少なくとも実行することを特徴とする方法。

40

【0017】

(5) (4)に記載の方法をコンピュータに実行させることを特徴とするプログラム。

【発明の効果】

【0018】

この発明によれば、真にウェブページを特徴づけるような素性語のみが素性語記憶手段に残留するため、素性語に基づいたカテゴリ分類を効率的に行うことができ、カテゴリ分

50

類装置の導入コストを低く抑えることができる。

【発明を実施するための最良の形態】

【0019】

以下、本発明を実施するための最良の形態について図を参照しながら説明する。なお、これはあくまでも一例であって、本発明の技術的範囲はこれに限られるものではない。

【0020】

[素性語選別装置と関連要素の全体構成]

図1は、本実施形態に係る素性語選別装置と関連要素の全体構成を示す図である。

【0021】

本実施形態に係る素性語選別装置1は、インターネット等の通信ネットワークNと、分類装置2とに接続されている。検索装置3は、通信ネットワークNと、素性語選別装置1とに接続されている。

【0022】

分類装置2は、通信ネットワーク上のウェブページをカテゴリに分類するものである。その分類手法は種々考えられるが、例えば、以下の分類手法が考えられる。

【0023】

すなわち、分類装置2は、ウェブページに含まれる文字データについて、形態素解析を行い、文章（複数の単語から構成される文字データ）を単語毎に分けて、分けた単語を品詞毎に分類して、特定の品詞（例えば名詞）のみを抽出する。そして、後述する素性語データベース12（図2参照）に記憶された素性語の中に、一致する素性語があるか否かを判断し、一致するものがあれば、一致した素性語に関係付けられたカテゴリとの平均相互情報量により、平均相互情報量が所定値以上である場合には、当該カテゴリを、このウェブページのカテゴリとする。

【0024】

なお、平均相互情報量の算出には、例えば、以下の式が使用される。

【数1】

$$MI(w,c) = \sum_{W \in \{w,w\}} \sum_{C \in \{c,c\}} P(W,C) \log \frac{P(W,C)}{P(W)P(C)}$$

30

MI (w, c) : 形態素となる単語wとカテゴリcの間の平均相互情報量

Pを確率として、P (X, Y) は、「X」と「Y」とが共起する確率であり、P (X)、P (Y) は、それぞれ個別に生起する確率である。

【0025】

検索装置3は、通信ネットワークN上に接続されたユーザ端末（図示せず）から、検索キーワードを含む検索要求（以下、クエリと呼ぶ）を受け付けて、検索キーワードにマッチするウェブページの通信アドレス、タイトル、要約を少なくとも含む検索結果を返信する。

【0026】

検索装置3は、通信ネットワークN上に接続されたユーザ端末からクエリを受け付けるクエリ受付部30と、検索処理部32にクエリを送信するクエリ送信部31と、クエリ送信部31から送信されたクエリに基づいて、検索キーワードにマッチするウェブページを検索する検索処理部32と、検索処理部32により検索されたウェブページの通信アドレス、タイトル、要約を少なくとも含む検索結果を生成する検索結果生成部33と、ユーザ端末の他に、素性語選別装置1に検索結果を送信する検索結果送信部34と、検索履歴を記録する検索ログ記録部35と、を備える。

【0027】

[素性語選別装置の機能構成]

図2は、本実施形態に係る素性語選別装置1の機能構成の概要を示す図である。本実施形態に係る素性語選別装置1は、素性語記憶手段10と、素性語選別手段11と、を備え

50

る。素性語選別手段 11 は、更に素性語準備処理部 13 と素性語選別処理部 14 を備える。

【0028】

素性語記憶手段 10 は、ウェブページを特徴づける単語となる素性語を予め記憶している。尚、素性語記憶手段 10 の詳細については後述する。素性語選別手段 11 は、分類装置 2 と協働した処理により、後述する所定のルールに基づいて素性語を選別する。素性語データベース 12（素性語 DB）は素性語記憶手段 10 の一部を構成しており、予め前記素性語を記憶している。

【0029】

〔素性語データベースと関連要素の構成〕

図 3 は、素性語データベース 12 と関連要素の構成の概要を示す図である。

素性語データベース 12 は、構築済みのカテゴリ構造データを用いて、予め素性語 ID と素性語とを対応付けて記憶した素性語 ID テーブル 121 と、予め文書 ID と文書（ウェブページ）の通信アドレス（例えば URL）とを対応付けて記憶した文書 ID テーブル 122 と、予め素性語 ID と文書 ID とを対応付けて記憶した ID 対応テーブル 123 と、後述する頻度テーブル 124 と、を備える。

【0030】

素性語 ID テーブル 121 には、素性語に素性語 ID が割り付けられて記憶されている。文書 ID テーブル 122 には、ウェブページの通信アドレス（URL 等）に文書 ID が割り付けられて記憶されている。ID 対応テーブル 123 には、素性語に割り付けられた素性語 ID と、ウェブページに割り付けられた文書 ID と、該素性語が該ウェブページに出現する出現位置を表す出現位置タイプと、が対応して記憶されている。

【0031】

頻度テーブル 124 には、素性語 ID 毎に、検索キーワードを基にした検索により、文書 ID テーブル 122 に記憶されているウェブページにヒットしたヒット回数が「頻度」として記憶される。図中では、検索キーワードを基にしたウェブページのヒット回数を頻度として記憶するという意味で、「頻度」がタイプ「検索」に対応して記憶される。

【0032】

また、頻度テーブル 124 には、素性語 ID 毎に、その素性語 ID に対応して素性語 ID テーブル 121 に記憶されている素性語が、文書 ID テーブル 122 に記憶されているウェブページにおける「タイトル」及び「本文」中における出現頻度が記憶される。図中では、ウェブページのタイトル及び本文における素性語の出現頻度を記憶するという意味で、「頻度」が、タイプ「タイトル」及び「本文」に対応して記憶される。

【0033】

素性語選別装置 1 は、ヒット回数及び出現頻度をカウントする対象となる素性語を、素性語 ID テーブル 121 を参照して抽出する。

【0034】

〔素性語選別装置のハードウェア構成図〕

図 4 は、本実施形態に係る素性語選別装置のハードウェア構成を示す図である。

素性語選別装置 1 は、制御部 300 を構成する CPU（Central Processing Unit）310（マルチプロセッサ構成では CPU 320 等複数の CPU が追加されてもよい）、バスライン 200、通信 I/F（I/F：インタフェース）330、メインメモリ 340、BIOS（Basic Input Output System）350、I/O コントローラ 360、ハードディスク 370、光ディスクドライブ 380、並びに半導体メモリ 390 を備える。尚、ハードディスク 370、光ディスクドライブ 380、並びに、半導体メモリ 390 はまとめて記憶装置 410 と呼ばれる。

【0035】

制御部 300 は、素性語選別装置 1 を統括的に制御する部分であり、ハードディスク 370（後述）に記憶された各種プログラムを適宜読み出して実行することにより、上述したハードウェアと協働し、本発明に係る各種機能を実現している。

10

20

30

40

50

【0036】

通信 I / F 330 は、素性語選別装置 1 が、インターネット N (図 1) を介してユーザ端末等と情報を送受信する場合のネットワーク・アダプタである。通信 I / F 330 は、モデム、ケーブル・モデム及びイーサネット (登録商標) ・アダプタを含んでよい。

【0037】

B I O S 350 は、素性語選別装置 1 の起動時に C P U 310 が実行するブートプログラムや、素性語選別装置 1 がハードウェアに依存するプログラム等を記録する。

【0038】

I / O コントローラ 360 には、ハードディスク 370、光ディスクドライブ 380、及び半導体メモリ 390 等の記憶装置 410 を接続することができる。

10

【0039】

ハードディスク 370 は、本ハードウェアを素性語選別装置 1 として機能させるための各種プログラム、本発明の機能を実行するプログラム及び後述するテーブル及びレコードを記憶する。なお、素性語選別装置 1 は、外部に別途設けたハードディスク (図示せず) を外部記憶装置として利用することもできる。

【0040】

光ディスクドライブ 380 としては、例えば、DVD-ROM ドライブ、CD-ROM ドライブ、DVD-RAM ドライブ、CD-RAM ドライブを使用することができる。この場合は各ドライブに対応した光ディスク 400 を使用する。光ディスク 400 から光ディスクドライブ 380 によりプログラム又はデータを読み取り、I / O コントローラ 360 を介してメインメモリ 340 又はハードディスク 370 に提供することもできる。

20

【0041】

なお、本発明でいうコンピュータとは、記憶装置、制御部等を備えた情報処理装置をいい、素性語選別装置 1 は、記憶装置 410、制御部 300 等を備えた情報処理装置により構成され、この情報処理装置は、本発明のコンピュータの概念に含まれる。

【0042】

この素性語選別装置 1 では、制御部 300 が主として素性語選別手段 11 に対応し、記憶装置 410 が主として素性語記憶手段 10 に対応する。

【0043】

[検索装置 3 のハードウェア構成]

30

分類装置 2、検索装置 3 も、上述の素性語選別装置 1 と同様な構成を持つ。

【0044】

[本発明の実施形態に係るフローチャート]

[素性語選別のための準備処理]

図 5 は、本発明の実施形態に係る素性語選別のための準備処理のフローチャートを示している。

S 1 : 分類装置 2 は、既存のカテゴリデータを集合 D とする。

S 2、S 3 : 分類装置 2 は、集合 D が空でなければ、集合 D からカテゴリ (クラス) c 毎に、属する文書 d を取り出す。

S 4、S 5 : 分類装置 2 は、文書 d の素性 f (素性語ベクトル f) を生成し、素性語データベース 12 を更新する。

40

【0045】

[分類装置の訓練処理]

図 6 は、本発明の実施形態に係る訓練処理のフローチャートを示している。

S 6 ~ S 9 : 分類装置 2 は、既存のカテゴリデータを集合 D とする。集合 D が空でなければ、集合 D からカテゴリ (クラス) c 毎に、属する文書 d を取り出す。分類装置 2 は、文書 d の素性 f を生成する。

S 10 : 素性語選別装置 1 は、後述する素性語選別処理を行う。

S 11 : 分類装置 2 は、選別された素性語を用いて訓練用データを書き出す。

S 12 : ステップ S 7 において集合 D が空の場合には、ステップ S 11 で書き出された

50

訓練用データを用いて訓練を行いモデルMを生成する。

【0046】

[分類装置の予測処理]

図7は、本発明の実施形態に係る予測処理のフローチャートを示している。

S13、S14：分類装置2は、入力文書をdとし、文書dの素性fを生成する。

S15：素性語選別装置1は、後述する素性語選別処理を行う。

S16、S17：分類装置2は、選別された素性f'をモデルMで予測することにより、属するカテゴリcを予測する。

【0047】

[素性語選別処理]

図8は、本実施形態に係る素性語選別処理を表すフローチャートを示している。

S18：素性fの要素を集合Eとする。

S19、S20：集合Eが空ではなければ、集合Eから素性要素を取り出す。尚、図中、場所1とは、後述するように、素性要素が位置する文書中の場所を示す。語tとは、素性語となりうる文書中の語を示す。

S21：素性語DBを参照して(1, t)を選別する。

S22：削除対象であるかどうかを判定し、削除対象であれば、ステップS19に戻る。そうでなければステップS23に進む。

S23：(1, t)を出力し、ステップS19に戻る。

【0048】

[素性語選別処理の詳細]

以下に素性語選別処理の詳細について説明する。

【0049】

(場所1、語t)の選別条件には次の種類がある。

(条件a) 語tのウェブ検索頻度が一定範囲である。

(条件b) 既存カテゴリデータにおいて、場所1に語tが出現した頻度が一定範囲である。

(条件c) 既存カテゴリデータにおいて、任意の場所1に語tが出現した頻度が一定範囲である。

(条件d) 語tのウェブ検索頻度、既存カテゴリデータにおける特定場所での出現頻度が一定範囲の特定の組み合わせ条件を満たす。

ここで、場所1には、本文、タイトル、メタ記述、メタキーワードの種類が存在する。

【0050】

以下、図9～図13を参照し、準備処理と選別処理に関する具体的な処理フローを条件a、b、cについて示す。

【0051】

[素性語選別準備処理(条件a)]

図9は、素性語準備処理部13による条件aにおける素性語選別の準備処理を示したものである。

S30：素性fの要素を集合Eとする。

S31、S32：集合Eが空でない場合は、Eから素性要素(場所1、語t)を取り出す。集合Eが空の場合は、処理を終了する。

S33、S34：語tの検索頻度がDBに登録済みの場合は、ステップS31に戻り、そうでない場合は、検索装置3によって、語tをインターネット検索させる。

S35、S36：検索装置3からのヒット件数が取得できた場合は、素性語記憶手段10は、語tの検索頻度nをDBに登録する。そうでない場合は、ステップS31の処理に戻る。

【0052】

[素性語選別処理(条件a)]

図10は、素性語選別処理部14による条件aにおける素性語選別処理を示したもので

10

20

30

40

50

ある。

S 4 0 : 素性語 D B から語 t の検索頻度を取得する。

S 4 1 : 頻度 n が取得できた場合は、S 4 2 に処理を移し、そうでない場合は処理を終了する。

S 4 2、S 4 4 : 頻度 n が検索頻度の下限值に等しいか小さい場合は、削除判定を出力する。そうでない場合は、S 4 3 に処理を移す。

S 4 3、S 4 4 : 頻度 n が検索上限値を超える場合は、削除判定を出力する。そうでない場合は、処理を終了する。

【0053】

[素性語選別準備処理 (条件 b)]

図 1 1 は、素性語準備処理部 1 3 による条件 b における素性語選別の準備処理を示したものである。

S 5 0 : 素性 f の要素を集合 E とする。

S 5 1、S 5 2 : 集合 E が空でない場合は、E から素性要素 (場所 l、語 t) を取り出す。集合 E が空の場合は、処理を終了する。

S 5 3 : 素性語 D B から語 t の場所 l の頻度を取得する。

S 5 4 : 頻度 n が取得できた場合は、ステップ S 5 6 の処理に移り、取得できなかった場合は、ステップ S 5 5 の処理に戻る。

S 5 5 : 素性語記憶手段 1 0 により、語 t の場所 l の頻度 1 で D B に登録し、ステップ S 5 1 の処理に再び戻る。

S 5 6 : D B の語 t の場所 l の頻度を更新する。そしてステップ S 5 1 へ処理を移す。

【0054】

[素性語選別処理 (条件 b)]

図 1 2 は、素性語選別処理部 1 4 による条件 b における素性語選別処理を示したものである。

S 6 0 : 素性語 D B から語 t の場所 l の頻度 n を取得する。

S 6 1 : 頻度 n が取得できた場合は、ステップ S 6 2 に移り、そうでない場合は処理を終了する。

S 6 2、S 6 4 : 頻度 n が場所頻度の下限值に等しいか大きい場合は、ステップ S 6 3 に移り、そうでない場合は、ステップ S 6 4 に移り、削除判定を出力し、その後処理を終了する。

S 6 3、S 6 4 : 頻度 n が場所頻度の上限値を超えない場合は、処理を終了し、そうでない場合は、ステップ S 6 4 に移り、削除判定を出力し、その後処理を終了する。

【0055】

[素性語選別準備処理 (条件 c)]

素性語準備処理部 1 3 による条件 c における素性選別準備処理は、図 1 1 の条件 b の場合と同様である。素性語準備処理部 1 3 は、条件 b の準備処理が行われていない場合のみ同一の処理を実行する。

【0056】

[素性語選別処理 (条件 c)]

図 1 3 は、素性語選別処理部 1 4 による条件 c における素性語選別処理を示したものである。

S 7 0 : 素性語 D B から語 t の全ての場所における頻度を取得し、合計する。

S 7 1 : 頻度合計 n a が取得できた場合は、ステップ S 7 2 に移り、そうでない場合は処理を終了する。

S 7 2 : 頻度合計 n a が場所頻度合計の下限值に等しいか大きい場合は、ステップ S 7 3 に移り、そうでない場合は、ステップ S 7 4 に移る。

S 7 3 : 頻度合計 n a が場所頻度合計の上限値を超える場合は、ステップ S 7 4 に移り、そうでない場合は、処理を終了する。

S 7 4 : 削除判定を出力し、処理を終了する。

【0057】

以上、本発明を実施形態に即して説明したが、本発明の実施形態に記載された効果は、本発明から生じる最も好適な効果を列挙したに過ぎず、本発明による効果は、本発明の実施例に記載されたものに限定されるものではない。

【図面の簡単な説明】

【0058】

【図1】本実施形態に係る素性語選別装置と関連要素の全体構成を示す図である。

【図2】本実施形態に係る素性語選別装置の機能構成の概要を示す図である。

【図3】素性語データベースと関連要素の構成の概要を示す図である。

【図4】本実施形態に係る素性語選別装置のハードウェア構成を示す図である。

10

【図5】本発明の実施形態に係る素性語選別のための準備処理のフローチャートを示している。

【図6】本発明の実施形態に係る訓練処理のフローチャートを示している。

【図7】本発明の実施形態に係る予測処理のフローチャートを示している。

【図8】本実施形態に係る素性語選別処理を表すフローチャートを示している。

【図9】本発明の実施形態に係る素性語選別準備処理（条件a）のフローチャートを示している。

【図10】本発明の実施形態に係る素性語選別処理（条件a）のフローチャートを示している。

【図11】本発明の実施形態に係る素性語選別準備処理（条件b）のフローチャートを示している。

20

【図12】本発明の実施形態に係る素性語選別処理（条件b）のフローチャートを示している。

【図13】本発明の実施形態に係る素性語選別処理（条件c）のフローチャートを示している。

【符号の説明】

【0059】

1 素性語選別装置

3 検索装置

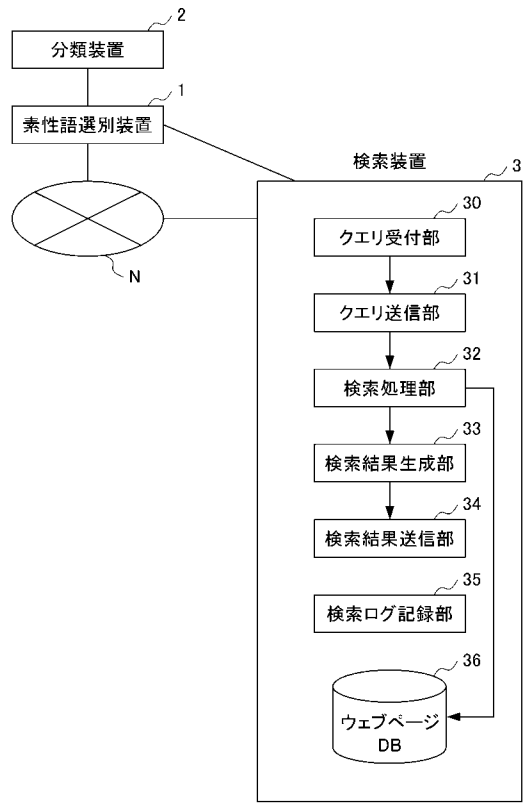
10 素性語記憶手段

11 素性語選別手段

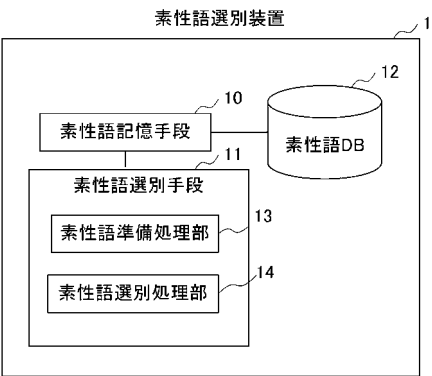
12 素性語データベース

30

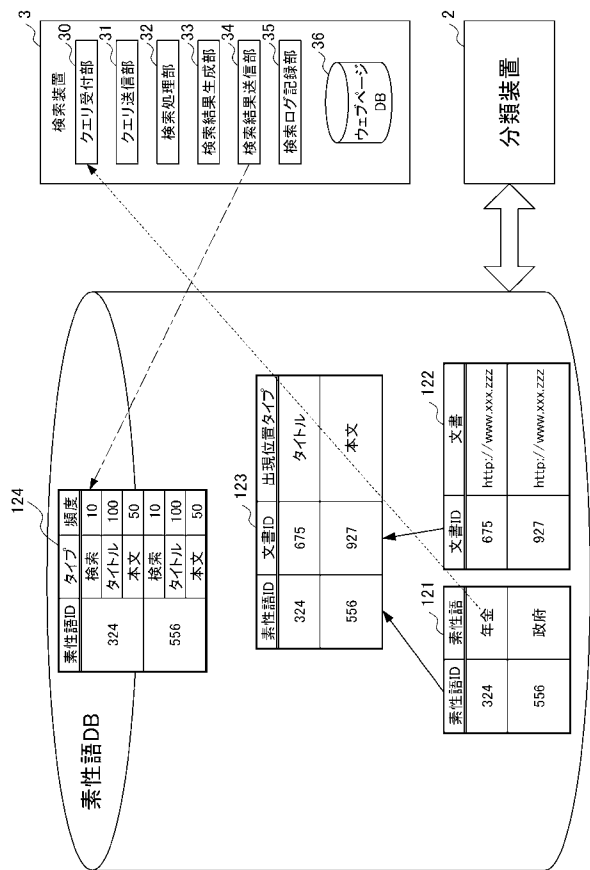
【図 1】



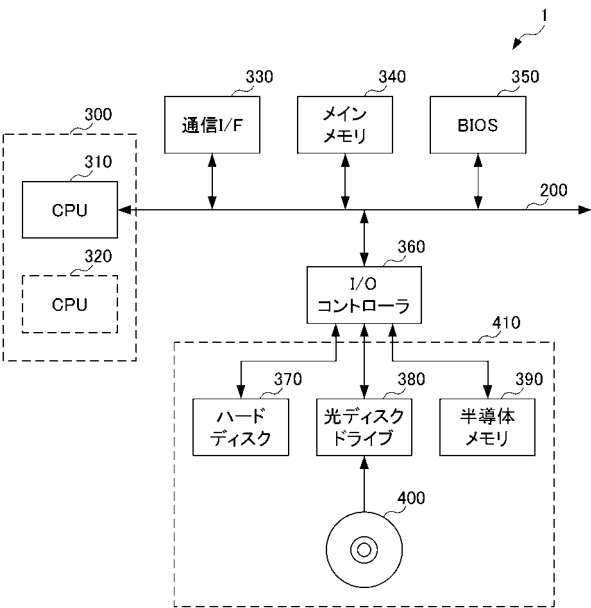
【図 2】



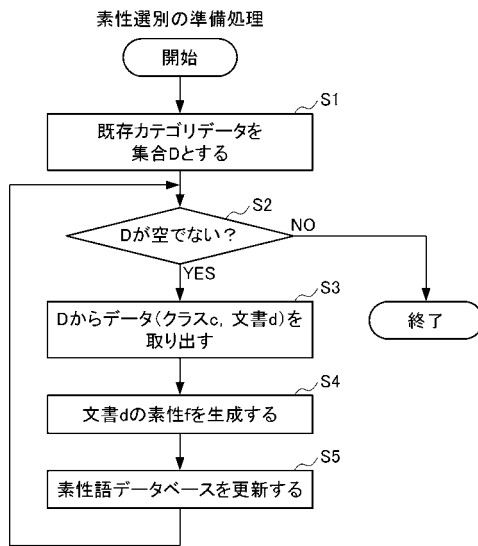
【図 3】



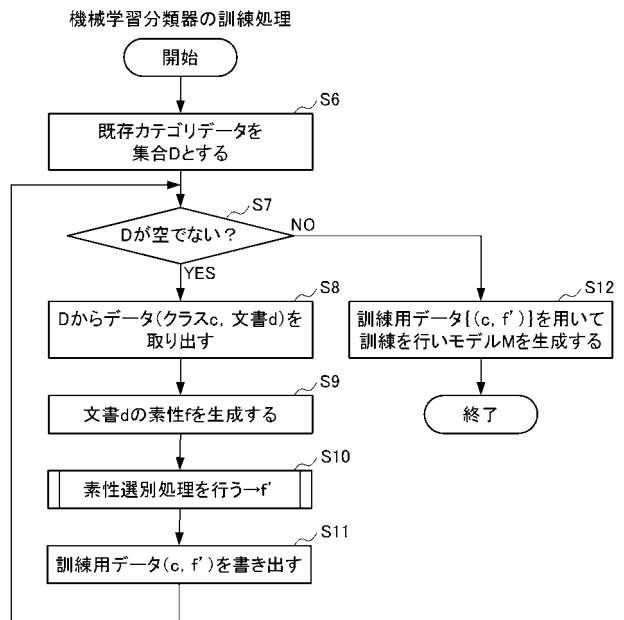
【図 4】



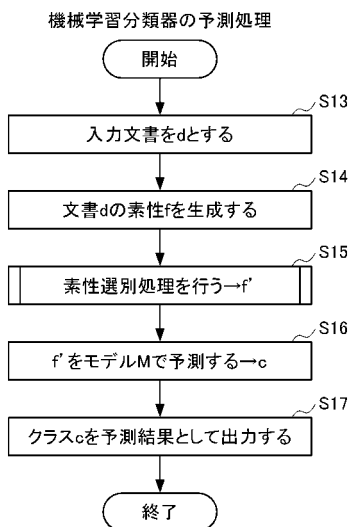
【図 5】



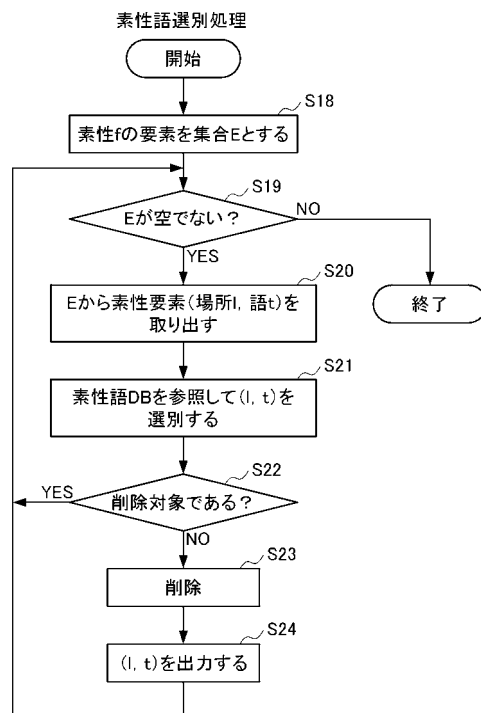
【図 6】



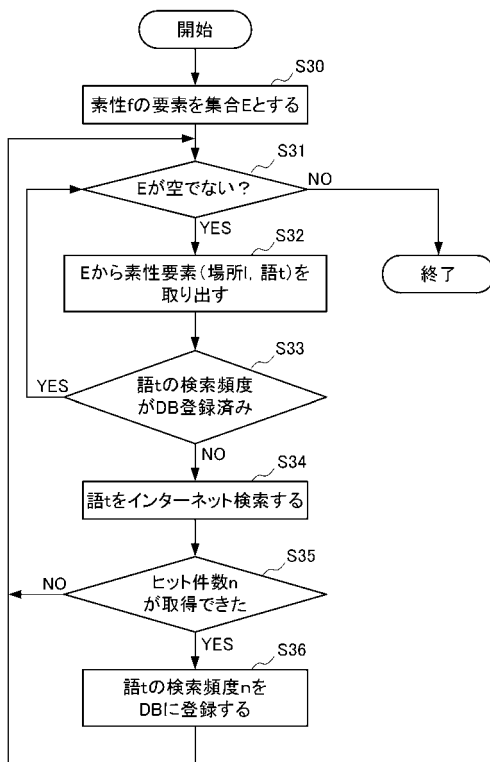
【図 7】



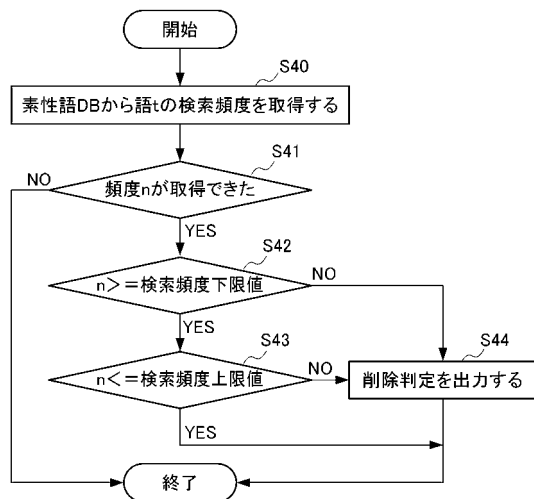
【図 8】



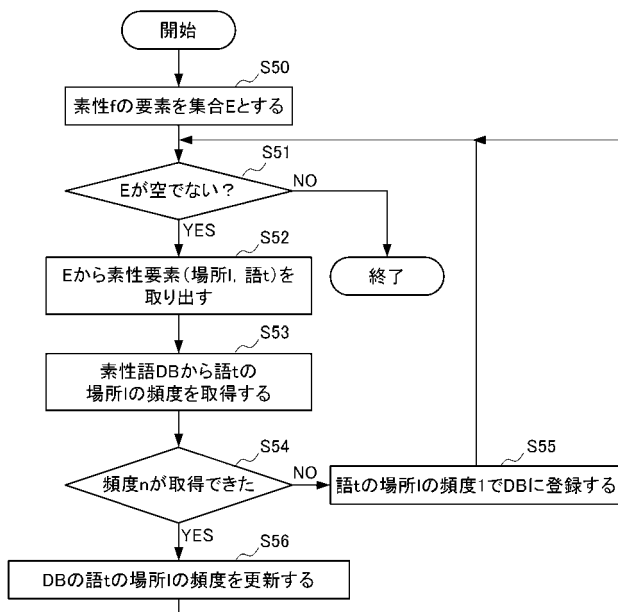
【図 9】



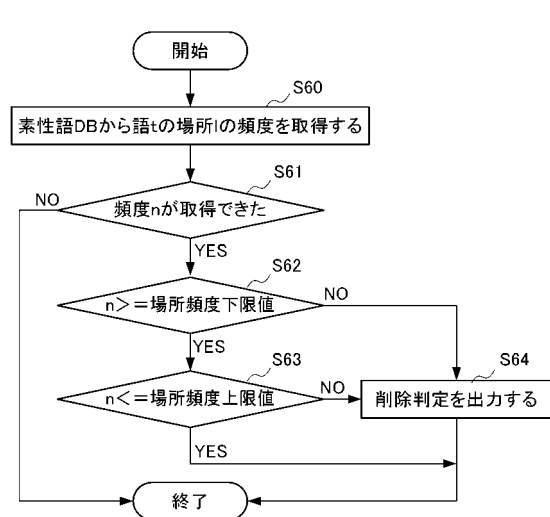
【図 10】



【図 11】



【図 12】



【図 13】

