

(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2010-39671

(P2010-39671A)

(43) 公開日 平成22年2月18日(2010.2.18)

(51) Int.Cl.

G06F 17/30 (2006.01)

F I

G06F 17/30 220Z
G06F 17/30 170A

テーマコード (参考)

5B075

審査請求 未請求 請求項の数 6 O L (全 13 頁)

(21) 出願番号 特願2008-200574 (P2008-200574)
(22) 出願日 平成20年8月4日 (2008.8.4)(71) 出願人 000004226
日本電信電話株式会社
東京都千代田区大手町二丁目3番1号
(74) 代理人 100121706
弁理士 中尾 直樹
(74) 代理人 100066153
弁理士 草野 卓
(74) 代理人 100128705
弁理士 中村 幸雄
(72) 発明者 野本 済央
東京都千代田区大手町二丁目3番1号 日
本電信電話株式会社内
(72) 発明者 野田 喜昭
東京都千代田区大手町二丁目3番1号 日
本電信電話株式会社内

最終頁に続く

(54) 【発明の名称】 テキストマイニング装置、方法、プログラム及びその記録媒体

(57) 【要約】

【課題】テキストマイニングにおいて、単語ランキングから定型部の影響をより適切に除去する。

【解決手段】各複数の分析対象文書を複数の単語に分割する。複数の分割された単語のそれぞれが分析対象文書に出現する頻度（以下、単語出現頻度とする。）を求める。ある分析対象文書の定型部をその分析対象文書の主題とは関係がない定型的な部分とし、ある単語の定型部平均単語出現頻度をその単語が各上記複数の分析対象文書の定型部に出現する推定平均頻度として、分割された単語について求めた単語出現頻度からその単語の定型部平均単語出現頻度を減算して、その単語についての定型部影響除去後単語出現頻度を求める。ある単語の文書頻度をその単語を含む上記複数の分析対象文書の数として、定型部影響除去後単語出現頻度が予め定められた頻度よりも高い又は以上である単語のそれぞれについての文書頻度を求める。

【選択図】図1

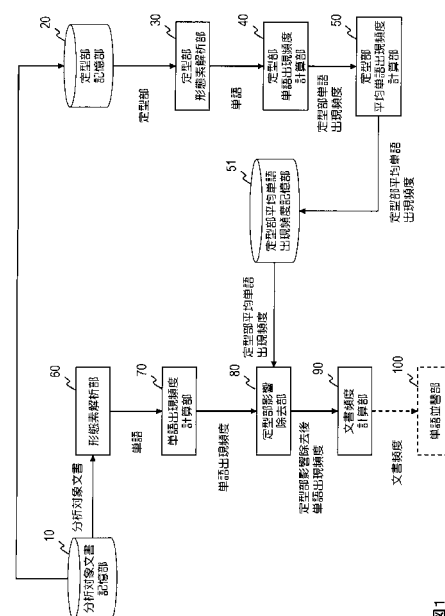


図1

【特許請求の範囲】**【請求項 1】**

複数の分析対象文書を格納する分析対象文書記憶部と、

上記分析対象文書記憶部から読み込んだ各上記複数の分析対象文書を複数の単語に分割する形態素解析部と、

上記複数の分割された単語の全部又は一部のそれぞれが上記読み込んだ分析対象文書に出現する頻度（以下、単語出現頻度とする。）を求める単語出現頻度計算部と、

ある分析対象文書の定型部をその分析対象文書の主題とは関係がない定型的な部分とし、ある単語の定型部平均単語出現頻度をその単語が各上記複数の分析対象文書の定型部に出現する推定平均頻度として、上記分割された単語について求めた上記単語出現頻度からその単語の定型部平均単語出現頻度を減算して、その単語についての定型部影響除去後単語出現頻度を求める定型部影響除去部と、

ある単語の文書頻度をその単語を含む上記複数の分析対象文書の数として、上記複数の分割された単語のうち、定型部影響除去後単語出現頻度が予め定められた頻度よりも高い又は以上である単語のそれぞれについての文書頻度を求める文書頻度計算部と、

を含むテキストマイニング装置。

【請求項 2】

請求項 1 に記載のテキストマイニング装置において、

上記複数の分析対象文書から選択された一部の分析対象文書のそれぞれの定型部を複数の単語に分割する定型部形態素解析部と、

各上記複数の分割された単語の全部又は一部のそれぞれが、上記定型部に出現する頻度を求める定型部単語出現頻度計算部と、

上記求めた頻度を単語ごとに加算した後に、上記一部の分析対象文書の数で割ることにより、その単語の定型部平均単語出現頻度を求める定型部平均単語出現頻度計算部と、

を更に含むことを特徴とするテキストマイニング装置。

【請求項 3】

請求項 1 又は 2 に記載のテキストマイニング装置において、

上記文書頻度が高い順に単語を出力する単語並替部を更に含む、

ことを特徴とするテキストマイニング装置。

【請求項 4】

分析対象文書記憶部には複数の分析対象文書を格納され、

上記分析対象文書記憶部から読み込んだ各上記複数の分析対象文書を複数の単語に分割する形態素解析ステップと、

上記複数の分割された単語の全部又は一部のそれぞれが上記読み込んだ分析対象文書に出現する頻度（以下、単語出現頻度とする。）を求める単語出現頻度計算ステップと、

ある分析対象文書の定型部をその分析対象文書の主題とは関係がない定型的な部分とし、ある単語の定型部平均単語出現頻度をその単語が各上記複数の分析対象文書の定型部に出現する推定平均頻度として、上記分割された単語について求めた上記単語出現頻度からその単語の定型部平均単語出現頻度を減算して、その単語についての定型部影響除去後単語出現頻度を求める定型部影響除去ステップと、

ある単語の文書頻度をその単語を含む上記複数の分析対象文書の数として、上記複数の分割された単語のうち、定型部影響除去後単語出現頻度が予め定められた頻度よりも高い又は以上である単語のそれぞれについての文書頻度を求める文書頻度計算ステップと、

を含むテキストマイニング方法。

【請求項 5】

請求項 1 から 3 の何れかに記載のテキストマイニング装置の各部としてコンピュータを機能させるためのプログラム。

【請求項 6】

請求項 5 に記載のプログラムを記憶したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

10

20

30

40

50

【技術分野】

【0001】

この発明は、形式化されていないテキストデータを単語等に分割し、その出現頻度や相関関係などをデータマイニングの手法を使って解析することで、一定の知見や発想を得るテキストマイニング技術に関する。

【背景技術】

【0002】

商品等に対する自由記述アンケート及びBlog記事等のテキストで書かれた文書を複数集めてきて、それらの主題傾向を調べようとした際、その中でどのような主題がどのくらいあるかを調べるために文書頻度(Document Frequency, 略してDFとも呼ぶ。)の単語ランキングが用いられる。

10

【0003】

ある単語についての文書頻度とは、その単語を含む文書の数のことである。例えば、図5においては、3つの文書1, 2, 3のそれぞれが「電話」という単語を含んでいる。よって、「電話」の文書頻度は3となる。文書頻度は、文書に含まれる単語の数を考慮しない概念である。例えば、文書1, 2, 3のそれぞれは「電話」を2つ含んでいるが、文書頻度はそれについては考慮しない。ある文書がある単語を含めば、その文書に含まれるその単語の数に関わらず、文書頻度として1を計上する。

【0004】

文書頻度の単語ランキングとは、文書頻度が大きい順に単語を並び替えて必要に応じて順位付けしたものである。図6に、図5の右側の欄に記載された単語及びその文書頻度についての単語ランキングを例示する。この単語ランキングにより、「電話」「横須賀」「市役所」という単語が出現した文書の数は3であり最も多いということがわかる。

20

【0005】

ところで、各文書にほぼ共通して現れる定型的であり、主題とは関係のない部分(以下、定型部という。)が存在した場合、その定型部に含まれる単語が単語ランキングの上位を占めてしまい、正確な主題傾向を掴むことができないという問題が生じ得る。図7に、下線で定型部を例示する。文書1の下線部分である「はい、お電話ありがとうございます。横須賀市役所市民窓口です。」「お電話ありがとうございました。それでは失礼致します。」等の挨拶はどの文書にも共通して現れる内容であり定型部と言える。そして、この定型部に含まれる「電話」「横須賀」「市役所」が、図6の単語ランキングの上位を占めている。このように、定型部に含まれる単語が単語ランキングの上位を占めると、どのような主題を持つ文書が多いのかを理解するのが難しくなる。したがって、単語ランキングから定型部の影響を取り除く必要がある。

30

【0006】

以下、定型部の影響を取り除くための2つの方法を説明する。

第1の方法は、定型部に多く現れる単語をストップワードとして登録しておき、このストップワードとして登録された単語を除いて単語ランキングを生成する方法である(例えば、非特許文献1参照。)。ストップワードの例を図8の右上に示す。この例では、文書1, 2, 3の定型部に多く現れる「電話」「横須賀」「市役所」「市民」「窓口」「ありがとう」「失礼」が、ストップワードとして登録される。そして、これらのストップワードを除いて生成された単語ランキングの例を図8の右下に示す。この単語ランキングは、図6の単語ランキングから、これらのストップワードを除いたものである。

40

【0007】

第2の方法は、テキストタイリング法等のテキストセグメンテーション技術を用いた方法である(例えば、非特許文献2参照。)。テキストタイリング法を用いて、図9の左側に例示するように、文書を定型部と主題部とに分離して、主題部を用いて文書頻度を計算して、単語ランキングの生成を行う。すなわち、単語が文書の主題部に出現した場合には文書頻度として1を計上する。文書の定型部のみに出現する単語は文書頻度を計算する際に考慮しない。

50

【 0 0 0 8 】

図 10 を用いて、テキストタイリング法により文書を区切る方法を簡単に説明する。文書中に基準点を定めて、その基準点の左右に所定の数の文からなる窓を設定する。図 10 の例では、窓のサイズは 3 であり、それぞれ 3 つの文からなる 2 つの窓をそれぞれ基準点の左側と右側とに設けている。基準点の左側の窓を左窓といい、基準点の右側の窓を右窓という。左窓に含まれる各単語が左窓に出現する頻度である左窓出現頻度、及び、右窓に含まれる各単語が右窓に出現する頻度である右窓出現頻度を計算する。そして、左窓出現頻度と右窓出現頻度との類似度を計算する。基準点を一定間隔でスライドさせながら、その類似度の変化を見て行き、類似度が低くなる位置を文書の切れ目として見つけ出して行く。話題が変化する位置では、左窓と右窓とは語彙的関連性が低く、類似度が小さくなる

10

【非特許文献 1】北研二，津田和彦，獅々堀正幹，「情報検索アルゴリズム」，共立出版，p. 29 - 30

【非特許文献 2】Maria A.Hearst. Multi-Paragraph Segmentation of Expository Text. 32nd Annual Meeting of the Association for Computation Linguistics. P.9-16. 1944

【発明の開示】

【発明が解決しようとする課題】

【 0 0 0 9 】

上記第 1 の方法においては、ストップワードとして登録された単語が主題部にも登場してその文書の主題を構成する場合には、その主題を見つけることはできない。例えば、図 8 の例では、「市役所」は文書 2 の主題部において「市役所の開庁時間帯」という主題の一部を構成しており、「窓口」は文書 3 の主題部において「窓口の受付時間帯」という主題の一部を構成している。しかし、「市役所」「窓口」は定型部に多く出現する単語であるため、ストップワードとして登録されている。したがって、この例では、単語ランキングから「市役所」「窓口」は除外されてしまい、文書 2 の主題及び文書 3 の主題の把握に失敗している。このように、上記第 1 の方法においては、主題を構成する単語を必要以上に除外しており、必ずしも定型部の影響を適切に取り除いているとは言えないという問題がある。

20

【 0 0 1 0 】

上記第 2 の方法においては、定型部が窓サイズよりも少ない数の文からなる場合には、その定型部を区切ることは難しい。これに対して、窓サイズを小さくすることによりさらに小さい定型部を区切ろうとする方法もあり得る。しかし、窓サイズを小さくするとそこに含まれる単語の数が少なくなり、左窓出現頻度と右窓出現頻度との類似度が著しく低くなり、統計的に信頼性のある類似度を計算することができなくなる。また、文書の最初又は最後においては左窓の窓サイズと右窓の窓サイズとを同じにすることができないため、定型部が文書の最初付近又は最後付近にある場合にも、その定型部を区切ることは難しい。このように上記第 2 の方法においては、定型部を適切に区切ることが難しいことに起因して、必ずしも定型部の影響を適切に取り除いているとは言えないという問題がある。

30

【 0 0 1 1 】

この発明は、上記問題に鑑みて、より適切に定型部の影響を取り除くことができるテキストマイニング装置、方法、プログラム及びその記録媒体を提供することを目的とする。

40

【課題を解決するための手段】

【 0 0 1 2 】

請求項 1 に記載されたテキストマイニング装置によれば、複数の分析対象文書を格納する分析対象文書記憶部と、分析対象文書記憶部から読み込んだ各複数の分析対象文書を複数の単語に分割する形態素解析部と、複数の分割された単語の全部又は一部のそれぞれが読み込んだ分析対象文書に出現する頻度（以下、単語出現頻度とする。）を求める単語出現頻度計算部と、ある分析対象文書の定型部をその分析対象文書の主題とは関係がない定型的な部分とし、ある単語の定型部平均単語出現頻度をその単語が各複数の分析対象文書

50

の定型部に出現する推定平均頻度として、分割された単語について求めた単語出現頻度からその単語の定型部平均単語出現頻度を減算して、その単語についての定型部影響除去後単語出現頻度を求める定型部影響除去部と、ある単語の文書頻度をその単語を含む複数の分析対象文書の数として、複数の分割された単語のうち、定型部影響除去後単語出現頻度が予め定められた頻度よりも高い又は以上である単語のそれぞれについての文書頻度を求める文書頻度計算部と、を含む。

【発明の効果】

【0013】

単語出現頻度から定型部平均単語出現頻度を減算することにより、より適切に定型部の影響を取り除くことができる。

定型部平均単語出現頻度は単語が定型部に出現する推定平均頻度であるから、単語出現頻度からこの定型部平均単語出現頻度を減算しても、上記第1の方法とは異なり、主題を構成する単語を必要以上に除外することにはならない。また、上記第2の方法とは異なり、上記テキストタイリング法を用いないため、定型部を適切に区切ることができないという問題は生じない。

【発明を実施するための最良の形態】

【0014】

以下、この発明の一実施形態を説明する。図1に、この発明によるテキストマイニング装置の機能構成の例を示す。図2に、この発明によるテキストマイニング方法の例を示す。

【0015】

分析対象文書記憶部10には、分析対象となる文書が複数記憶されている。分析対象となる文書のことを、分析対象文書と呼ぶことにする。分析対象文書は、例えば、電話の応対をテキストに書き起こしたもの、商品等に対する自由記述アンケート又はBlog記事等のテキストデータである。

分析対象文書記憶部10に記憶された複数の分析対象文書から一部の分析対象文書が選択される(ステップS1)。そして、これらの選択された分析対象文書のそれぞれから定型部が取り出されて、定型部記憶部20に記憶される。

【0016】

この発明では、ある分析対象文書の主題とは関係がない定型的な部分をその分析対象文書の定型部とする。ひとつの分析対象文書に複数の定型部の部分が含まれている場合には、これら複数の定型部の部分のすべてを、この分析対象文書の定型部と呼ぶ。例えば、図7の文書1では、第一の定型部の部分「はい、お電話ありがとうございます。横須賀市役所市民窓口です。」、第二の定型部の部分「お電話ありがとうございました。それでは失礼致します。」がある。したがって、この文書1の定型部といった場合には、「はい、お電話ありがとうございます。横須賀市役所市民窓口です。」と「お電話ありがとうございました。それでは失礼致します。」との両方を意味し、何れか一方を意味しない。

【0017】

分析対象文書の選択は、選択された分析対象文書の定型部における各単語の単語頻出頻度と、分析対象文書記憶部10に記憶された複数の分析対象文書の定型部における各単語の単語頻出頻度とに大きな差がないように行うことが望ましい。後述する定型部平均単語出現頻度の推定精度を上げるためである。そのように行えば、分析対象文書の選択は、コンピュータ等を用いてランダムに行っても良いし、人が任意に行ってもよい。例えば、50個程度の分析対象文書が選択される。

【0018】

分析対象文書からの定型部の取り出しは、例えば手作業等の、定型部を精度良く取り出すことができる方法で行う。この発明では、定型部の取り出しを、すべての分析対象文書に対して行う必要はなく、選択された分析対象文書、すなわち一部の分析対象文書に対して行えば足りる。したがって、定型部を取り出すために、手作業等の比較的時間が経っても精度が高い方法を用いることができるのである。もちろん、コンピュータを用いて精

10

20

30

40

50

度良く定型部を取り出すことができる方法があれば、コンピュータを用いてその方法により機械的に定型部の取り出しを行ってもよい。

【 0 0 1 9 】

定型部形態素解析部 3 0 は、定型部記憶部 2 0 から各定型部を読み出して、各定型部を複数の単語に分割する（ステップ S 2）。分割された単語は、定型部単語出現頻度計算部 4 0 に送られる。

【 0 0 2 0 】

単語への分割は、周知の形態素解析方法を用いることができる。例えば、形態素解析結果候補の中から最も長い文字数を含む候補を選択する最長一致法、文節数が最小になる候補を選択する方法、自立語の長い候補を選択する方法、予め定めた規則をもとにして形態素解析を行う方法、隠れマルコフモデルを用いて統計的にもっともらしい候補を選択する方法を例えば用いる（例えば、参考文献 1 参照。）。

〔参考文献 1〕石崎俊，「自然言語処理」，昭晃堂，P．2 7 - 2 9

【 0 0 2 1 】

定型部単語出現頻度計算部 4 0 は、分割された単語が、定型部に出現する頻度を求める（ステップ S 3）。つまり、定型部ごとに、定型部に出現した各単語がその定型部に出現する数をカウントする。求まった頻度は、定型部平均単語出現頻度計算部 5 0 に送られる。

【 0 0 2 2 】

例えば、図 3 の文書 1 の定型部においては、「電話」は 3 回出現するので「電話」が出現する頻度は 3 となる。同様に、文書 1 の定型部に出現する「ありがとう」「横須賀」「市役所」等の各単語が出現する頻度を求める。文書 2 についても同様に文書 2 に出現する各単語が文書 2 の定型部に出現する頻度を求め、文書 3 についても同様に文書 3 に出現する各単語が文書 2 の定型部に出現する頻度を求める。なお、図 3 は、選択された分析対象文書が、文書 1，2，3 の 3 つである場合の例である。

【 0 0 2 3 】

定型部平均単語出現頻度計算部 5 0 は、求まった頻度を単語ごとに加算した後に、選択された分析対象文書の数で割ることにより、定型部平均単語出現頻度を求める（ステップ S 4）。ある単語についての定型部平均単語出現頻度とは、その単語が分析対象文書の定型部に出現する推定平均頻度のことである。「求まった頻度を単語ごとに加算」とは、言い換えると、「単語が各定型部に出現する頻度を加算すること」を意味する。選択された分析対象文書の数は定型部の数と等しいため、選択された分析対象文書の数ではなく定型部の数で割ると考えてもよい。求まった定型部平均単語出現頻度は、定型部平均単語出現頻度記憶部 5 1 に記憶される。

【 0 0 2 4 】

例えば、図 3 の例では、「電話」が、文書 1 の定型部に出現する回数は 3 回であり、文書 2 の定型部に出現する回数は 1 回であり、文書 3 の定型部に出現する回数は 2 回である。定型部平均単語出現頻度計算部 5 0 は、これらの回数を加算（3 回 + 1 回 + 2 回 = 6 回）して、「電話」が、選択された分析対象文書の定型部の全体に出現する回数（6 回）を求める。そして、この加算値（6 回）を、選択された分析対象文書の数である 3 で割ることにより、「電話」の定型部平均単語出現頻度（2 回）を求める。「ありがとう」「横須賀」等の他の単語についても同様に定型部平均単語出現頻度を求める。

【 0 0 2 5 】

このように、この例では、全部の分析対象文書からではなく、選択された一部の分析対象文書から、定型部平均単語出現頻度を推定する。この定型部平均単語出現頻度の計算は、事前に行っておいてもよいし、後述するステップ S 5 からステップ S 8 の処理と並行して行ってもよい。

【 0 0 2 6 】

形態素解析部 6 0 は、分析対象文書記憶部 1 0 から読み込んだ各複数の分析対象文書を複数の単語に分割する（ステップ S 5）。分割された単語は、単語出現頻度計算部 7 0 に

10

20

30

40

50

送られる。単語への分割は、定型部形態素解析部 30 と同様に周知の形態素解析方法を用いて行うことができる。

【0027】

単語出現頻度計算部 70 は、形態素解析部 60 が分割した各単語が上記読み込んだ分析対象文書に出現する頻度を求める（ステップ S6）。つまり、分析対象文書ごとに、分析対象文書に含まれる各単語がその分析対象文書に出現する回数をカウントする。この求めた頻度のことを、単語出現頻度（Term Frequency, 略して TF と呼ぶ。）と表現する。すなわち、ある単語とある分析対象文書とについて単語出現頻度は、その単語がその分析対象文書に出現する回数を意味する。求めた頻度、すなわち単語出現頻度は、定型部影響除去部 80 に送られる。

10

【0028】

例えば、図 4 においては、「電話」は文書に 2 回出現するので「電話」の単語出現頻度は 2 回となり、「横須賀」は文書に 1 回出現するので「横須賀」の単語出現頻度は 1 回となる。

【0029】

定型部影響除去部 80 は、分割された単語について求めた上記頻度、すなわち分割された単語の単語出現頻度から、その単語の定型部平均単語出現頻度を減算して、その単語についての定型部影響除去後単語出現頻度を求める（ステップ S7）。すなわち、分割された単語ごとに、分割された単語の単語出現頻度から、その単語の定型部平均単語出現頻度を減算して、その減算結果をその単語の定型部影響除去後単語出現頻度とする。求めた定型部影響除去後単語出現頻度は、文書頻度計算部 90 に送られる。

20

【0030】

なお、定型部平均単語出現頻度が事前計算されている場合には、定型部影響除去部 80 は、定型部平均単語出現頻度記憶部 51 から対応する定型部平均単語出現頻度を適宜読み込む。定型部平均単語出現頻度の計算がステップ S5 からステップ S8 の処理と並行して行われる場合には、定型部平均単語出現頻度計算部 50 が計算した対応する定型部平均単語出現頻度が定型部影響除去部 80 に直接送られてもよい。

【0031】

文書頻度計算部 90 は、ある単語の文書頻度をその単語を含む分析対象文書の数として、複数の分割された単語のうち、定型部影響除去後単語出現頻度が予め定められた頻度よりも高い単語のそれぞれについての文書頻度を求める（ステップ S8）。なお、複数の分割された単語のうち、定型部影響除去後単語出現頻度が予め定められた頻度以上の単語のそれぞれについての文章頻度を求めてもよい。

30

【0032】

このようにして、単語出現頻度から定型部平均単語出現頻度を減算することにより、より適切に定型部の影響を取り除くことができる。すなわち、定型部平均単語出現頻度は単語が定型部に出現する推定平均頻度であるから、単語出現頻度からこの定型部平均単語出現頻度を減算しても、背景技術の欄に記載された第 1 の方法とは異なり、主題を構成する単語を必要以上に除外することにはならない。また、背景技術の欄に記載された第 2 の方法とは異なり、上記テキストタイリング法を用いないため、定型部を適切に区切ることができないという問題は生じない。

40

【0033】

〔変形例等〕

定型部形態素解析部 30 が定型部を分割して単語を出力する際、その分割の方法によっては、助詞、接続詞等の単体で主題を構成しない単語が出力される場合がある。この場合、定型部単語出現頻度計算部 40 は、名詞、動詞等の単体で主題を構成する単語についての出現頻度を求め、助詞、接続詞等の単体で主題を構成しない単語についての出現頻度を求めなくてもよい。つまり、定型部単語出現頻度計算部 40 は、定型部形態素解析部 30 が分割した単語の全部ではなく、一部の単語についての出現頻度を求めてもよい。

【0034】

50

同様に、形態素解析部 60 が分析対象文書を分割して単語を出力する際、その分割の方法によっては、助詞、接続詞等の単体で主題を構成しない単語が出力される場合がある。この場合、単語出現頻度計算部 70 は、名詞、動詞等の単体で主題を構成する単語についての単語出現頻度を求め、助詞、接続詞等の単体で主題を構成しない単語についての単語出現頻度を求めなくてもよい。つまり、単語出現頻度計算部 70 は、形態素解析部 60 が分割した単語の全部ではなく、一部の単語についての単語出現頻度を求めてもよい。

【0035】

なお、図 1 に点線で示すように、文書頻度が高い順に単語を出力する単語並替部 100 を設けてもよい。これにより、単語ランキングを生成することができ、主題傾向を把握し易くなる。単語並替部 100 は、単語の文書頻度に応じて順位付けをしてもよい。また、単語並替部 100 は、並び替えた単語の全部を出力する必要はなく、並び替えた単語の一部のみを出力してもよい。例えば、予め定められた文書頻度以上の文書頻度を有する単語のみや、予め定められた順位以上の順位を有する単語のみを出力してもよい。これにより、さらに主題傾向を把握し易くなる。

10

【0036】

上記の例では、分析対象文書の一部から定型部平均単語出現頻度を計算したが、分析対象文書ではない文書から上記と同様にして定型部平均単語出現頻度を計算してもよい。例えば、過去に分析対象であったが、今回は分析対象ではない文書であり、定型部における各単語の単語出現頻度に大きな変化がないような場合には、その過去に分析対象であった文書から定型部平均単語出現頻度を計算してもよい。同様に、分析対象文書と分析対象ではない文書とを含む文書から上記と同様にして定型部平均単語出現頻度を計算してもよい。

20

【0037】

上述の構成をコンピュータによって実現する場合、テキストマイニング装置の各部が有する機能の処理内容はプログラムによって記述される。そして、このプログラムをコンピュータで実行することにより、上記各部の機能がコンピュータ上で実現される。

【0038】

すなわち、CPU がプログラムを逐次読み込んで実行することにより、定型部形態素解析部 30、定型部単語出現頻度計算部 40、定型部平均単語出現頻度計算部 50、形態素解析部 60、単語出現頻度計算部 70、定型部影響除去部 80、文書頻度計算部 90 及び単語並替部 100 の機能がそれぞれ実現される。また、補助記憶装置又はメモリが、分析対象文書記憶部 10、定型部記憶部 20 及び定型部平均単語出現頻度記憶部 51 として機能する。

30

【0039】

テキストマイニング装置の各部として機能する CPU は、メモリ又は補助記憶装置から読み込み込んだデータに対して処理を行い、処理を行った後のデータをメモリ又は補助記憶装置に格納する。すなわち、メモリ又は補助記憶装置を介して、音響再生装置の各部間でデータがやり取りされる。

【0040】

この処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体としては、例えば、磁気記録装置、光ディスク、光磁気記録媒体、半導体メモリ等のようなものでもよいが、具体的には、例えば、磁気記録装置として、ハードディスク装置、フレキシブルディスク、磁気テープ等を、光ディスクとして、DVD (Digital Versatile Disc)、DVD-RAM (Random Access Memory)、CD-ROM (Compact Disc Read Only Memory)、CD-R (Recordable) / RW (ReWritable) 等を、光磁気記録媒体として、MO (Magnetooptical disc) 等を、半導体メモリとしてEEPROM (Electrically Erasable and Programmable-Read Only Memory) 等を用いることができる。

40

【0041】

また、このプログラムの流通は、例えば、そのプログラムを記録したDVD、CD-R

50

ＯＭ等の可搬型記録媒体を販売、譲渡、貸与等することによって行う。さらに、このプログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することにより、このプログラムを流通させる構成としてもよい。

【００４２】

また、上述した実施形態とは別の実行形態として、コンピュータが可搬型記録媒体から直接このプログラムを読み取り、そのプログラムに従った処理を実行することとしてもよく、さらに、このコンピュータにサーバコンピュータからプログラムが転送されるたびに、逐次、受け取ったプログラムに従った処理を実行することとしてもよい。また、サーバコンピュータから、このコンピュータへのプログラムの転送は行わず、その実行指示と結果取得のみによって処理機能を実現する、いわゆるＡＳＰ（Application Service Provider）型のサービスによって、上述の処理を実行する構成としてもよい。なお、本形態におけるプログラムには、電子計算機による処理の用に供する情報であってプログラムに準ずるもの（コンピュータに対する直接の指令ではないがコンピュータの処理を基底する性質を有するデータ等）を含むものとする。

10

【００４３】

また、この形態では、コンピュータ上で所定のプログラムを実行させることにより、本装置を構成することとしたが、これらの処理内容の少なくとも一部をハードウェア的に実現することとしてもよい。

また、上述の各種の処理は、記載に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的にあるいは個別に実行されてもよい。

20

その他、本発明の趣旨を逸脱しない範囲で適宜変更が可能であることはいうまでもない。

【図面の簡単な説明】

【００４４】

【図１】テキストマイニング装置の機能構成の例を示す図。

【図２】テキストマイニング方法の例を示す図。

【図３】定型部平均単語出現頻度を説明するための図。

【図４】単語出現頻度を説明するための図。

【図５】文書頻度を説明するための図。

30

【図６】単語ランキングを説明するための図。

【図７】定型部及び主題部を説明するための図。

【図８】定型部の影響を除去する従来の第１の方法を説明するための図。

【図９】定型部の影響を除去する従来の第１の方法を説明するための図。

【図１０】テキストタイリング法を説明するための図。

【符号の説明】

【００４５】

- １０ 分析対象文書記憶部
- ２０ 定型部記憶部
- ３０ 定型部形態素解析部
- ４０ 定型部単語出現頻度計算部
- ５０ 定型部平均単語出現頻度計算部
- ５１ 定型部平均単語出現頻度記憶部
- ６０ 形態素解析部
- ７０ 単語出現頻度計算部
- ８０ 定型部影響除去部
- ９０ 文書頻度計算部
- １００ 単語並替部

40

【図 1】

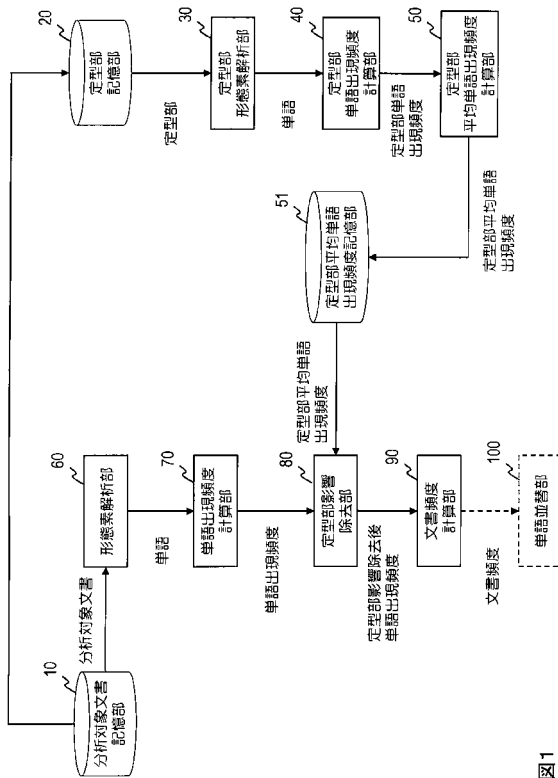


図 1

【図 2】

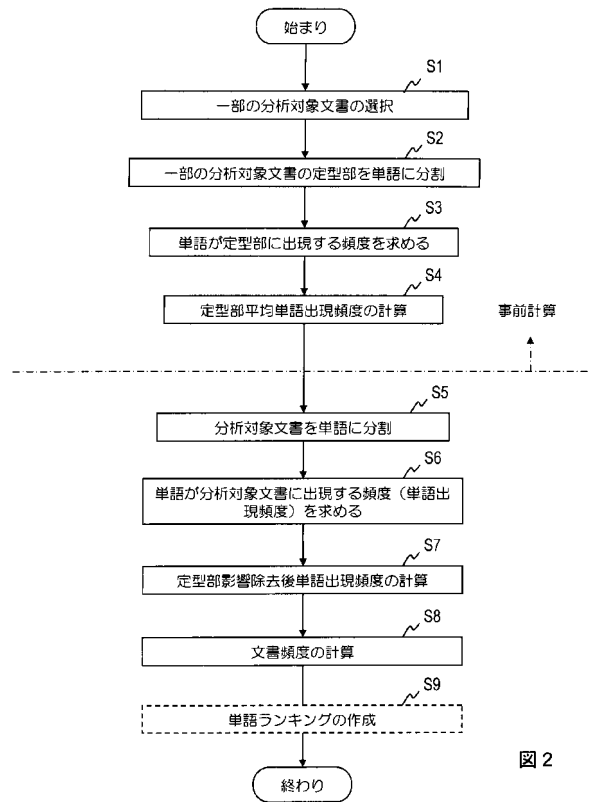


図 2

【図 3】

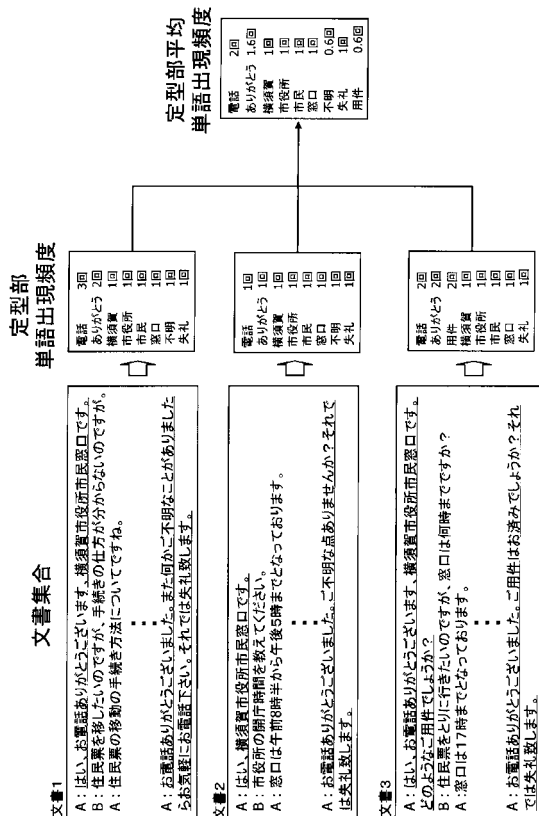


図 3

【図 4】

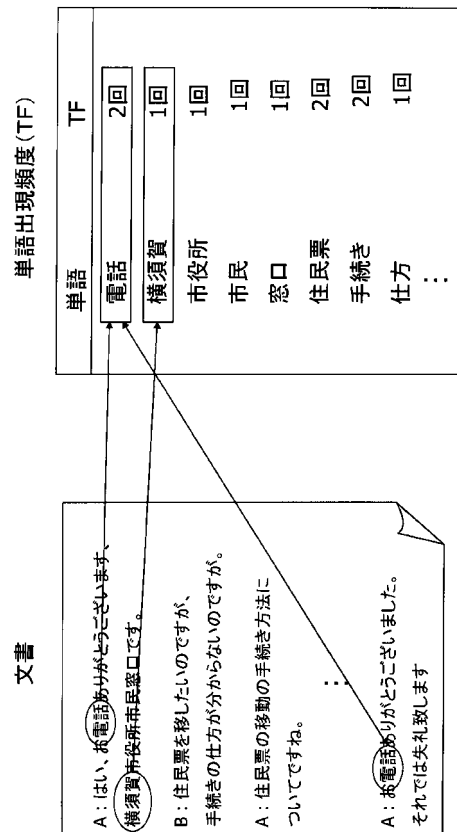


図 4

【 図 5 】

文書集台		文書頻度 (DF)	
文書 1	A: はい、お電話ありがとうございます。	単語	DF
	B: 住民票を授けたいのですが、手続きの仕方が分からないのですが。	電話	3
	A: 住民票の移動の手続き方法についてですね。	横須賀 市役所 住民票 手続き 仕方	3 3 2 1 1
文書 2	A: お電話ありがとうございます。それでは失礼致します。	市民 窓口	2 2
	A: はい、お電話ありがとうございます。横須賀市役所市民窓口です。	市民 窓口	2 2
	B: 市役所の制作時間を教えてください。	市民 窓口	2 2
文書 3	A: 窓口は午前8時半から午後5時までとなっております。	市民 窓口	2 2
	A: お電話ありがとうございます。それでは失礼致します。	市民 窓口	2 2
	A: はい、お電話ありがとうございます。横須賀市役所市民窓口です。	市民 窓口	2 2
文書 4	B: 住民票をとりに行きたいのですが、窓口は何時までですか？	市民 窓口	2 2
	A: 窓口は17時までとなっております。	市民 窓口	2 2
	A: お電話ありがとうございます。それでは失礼致します。	市民 窓口	2 2

5 ☒

【 図 6 】

文書集令		文書 3	
文書 1 A: はい、お電話ありがとうございます。横須賀市役所市民窓口です。 B: 生民票を想たいのですが、手続きの仕方が分らないのですが。 A: 生民票の移動の手続き方法についてですね。 ... A: お電話ありがとうございます。それでは失礼致します。		文書 3 A: はい、お電話ありがとうございます。横須賀市役所市民窓口です。 B: 市役所の開庁時間を教えて下さい。 A: 窓口は午前8時半から午後5時までとなっております。 ... A: お電話ありがとうございます。それでは失礼致します。	
文書 2 A: はい、お電話ありがとうございます。横須賀市役所市民窓口です。 B: 市役所の開庁時間を教えて下さい。 A: 窓口は午前8時半から午後5時までとなっております。 ... A: お電話ありがとうございます。それでは失礼致します。		文書 3 A: はい、お電話ありがとうございます。横須賀市役所市民窓口です。 B: 生民票をとりに行きたいのですが、窓口は何時までですか？ A: 窓口は17時までとなっております。 ... A: お電話ありがとうございます。それでは失礼致します。	

9
☒

【圖 7】

文書1	A: はい、お電話ありがとうございます、機須賀市役所市民窓口です。 B: 住民票を移したいのですが、手続きの仕方が分からないのですが。 A: 住民票の移動の手続き方法についてですね。 A: お電話ありがとうございます。それでは失礼致します。	【定型部】 【主題部】 【主題部】 【定型部】
文書2	A: はい、お電話ありがとうございます、機須賀市役所市民窓口です。 B: 市役所の開庁時間を教えてください。 A: 窓口は午前8時半から午後5時までとなっております。 A: お電話ありがとうございます。それでは失礼致します。	【定型部】 【主題部】 【主題部】 【定型部】
文書3	A: はい、お電話ありがとうございます、機須賀市役所市民窓口です。 B: 住民票をとりに行きたいのですが、窓口は何時までですか？ A: 窓口は17時までとなっております。 A: お電話ありがとうございます。それでは失礼致します。	【定型部】 【主題部】 【主題部】 【定型部】

7 ☒

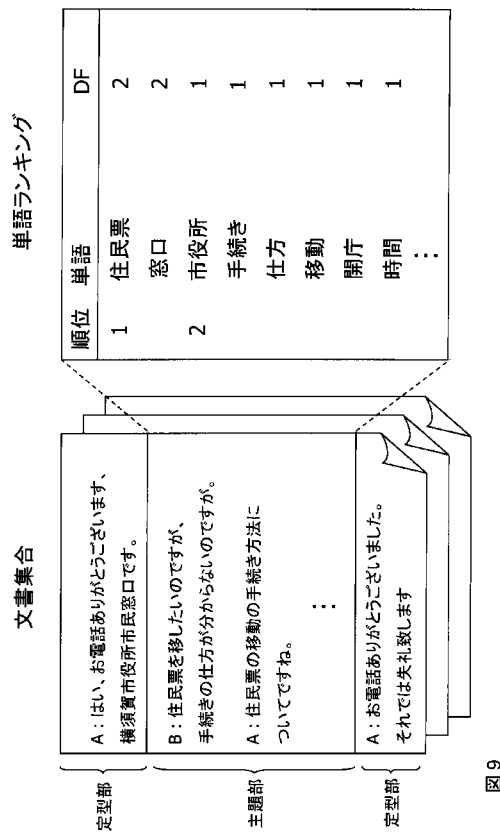
【圖 8】

文書 1	
A: はい、お電話ありがとうございます。機須賀市役所市民窓口です。	
B: 住民票を移したいのですが、手続きの仕方が分からないのです。	
A: 住民票の移動の手続き方法についてですね。	
A: お電話ありがとうございました。それでは失礼致します。	
文書 2	
A: はい、お電話ありがとうございます。機須賀市役所市民窓口です。	
B: 市役所の開庁時間を教えてください。	
A: 窓口は午前8時半から午後5時までとなっております。	
A: お電話ありがとうございました。それでは失礼致します。	
文書 3	
A: はい、お電話ありがとうございます。機須賀市役所市民窓口です。	
B: 住民票をとりに行きたいのですが、窓口は何時までですか？	
A: 窓口は17時までとなっております。	
A: お電話ありがとうございました。それでは失礼致します。	

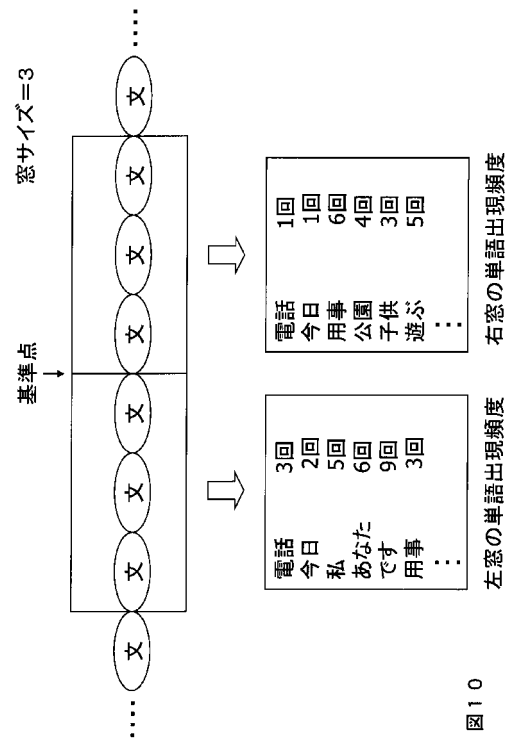
98

8
✗

【図 9】



【図 10】



フロントページの続き

(72)発明者 甘粕 哲郎

東京都千代田区大手町二丁目3番1号 日本電信電話株式会社内

Fターム(参考) 5B075 ND03 NS10 PQ02 UU06