

(19) 日本国特許庁(JP)

(12) 特許公報(B2)

(11) 特許番号

特許第4931111号
(P4931111)

(45) 発行日 平成24年5月16日(2012.5.16)

(24) 登録日 平成24年2月24日(2012.2.24)

(51) Int.Cl. F I
G06F 17/30 (2006.01) G06F 17/30 210D
 G06F 17/30 350C

請求項の数 7 (全 16 頁)

<p>(21) 出願番号 特願2006-154126 (P2006-154126) (22) 出願日 平成18年6月2日(2006.6.2) (65) 公開番号 特開2007-323454 (P2007-323454A) (43) 公開日 平成19年12月13日(2007.12.13) 審査請求日 平成21年3月26日(2009.3.26)</p> <p>特許法第30条第1項適用 2005年12月6日~9日 国立情報学研究所主催の「NTCIR Workshop 5 Meeting (Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access)」で発表</p>	<p>(73) 特許権者 301022471 独立行政法人情報通信研究機構 東京都小金井市貫井北町4-2-1 (74) 代理人 100103827 弁理士 平岡 憲一 (74) 代理人 100119161 弁理士 重久 啓子 (72) 発明者 村田 真樹 東京都小金井市貫井北町4-2-1 独立行政法人情報通信研究機構内 審査官 岩田 淳</p>
---	--

最終頁に続く

(54) 【発明の名称】 文書分類装置及びプログラム

(57) 【特許請求の範囲】

【請求項1】

適切な個数の複数の分類を求める対象として分類する文書を入力する入力手段と、
 前記入力された文書と予め1つの文書に複数の分類の付与された文書集合との類似度をキーワードを抽出して算出する文書類似度算出手段と、

前記予め分類の付与された文書集合から前記入力された文書と最も高い類似度から順に指定数の文書を抽出する文書抽出手段と、

前記抽出した指定数の文書について、前記類似度の高いものほど順位を上位とする順位を前記文書について求め、前記抽出した指定数の文書において出現した複数の分類のそれぞれについて、その分類が出現した文書の順位が上位のものほど大とする値と、前記類似度をかけあわせた値とを、前記抽出した指定数の文書のうちその分類が出現した文書分だけ足し合わせて値を求め、その値をその分類のスコアとすることで、その分類のスコアを算出するスコア算出手段と、

前記算出した分類のスコアのうち最も高い分類を特定し、該特定した分類のスコアに指定値をかけた値を求め、該求めた値より大きい複数の分類を前記入力された文書の分類として抽出する分類集合抽出手段とを備えることを特徴とした文書分類装置。

【請求項2】

適切な個数の複数の分類を求める対象として分類する文書を入力する入力手段と、

前記入力された文書と予め1つの文書に複数の分類の付与された文書集合との類似度をキーワードを抽出して算出する文書類似度算出手段と、

10

20

前記予め分類の付与された文書集合から前記入力された文書と最も高い類似度から順に指定数の文書を抽出する文書抽出手段と、

前記抽出した指定数の文書において出現していた複数の分類それぞれについて、その分類が前記抽出した指定数の文書のうち出現した文書の個数を求め、その個数をその分類のスコアとすることで、その分類のスコアを算出するスコア算出手段と、

前記算出したスコアが大きい分類順に、前記抽出した指定数の文書に付与されている複数の分類の数の平均の分類数分を前記入力された文書の分類として抽出する分類集合抽出手段とを備えることを特徴とした文書分類装置。

【請求項 3】

前記抽出した種々の技術的観点から多観点で分類されている複数の文書の該分類の技術的観点を二次元の表にし、該表に分類された技術がどこにあるかの印を設けることを特徴とした請求項 1 又は 2 記載の文書分類装置。

【請求項 4】

前記技術的観点を並べ替え、前記印が付いていない箇所を集め直すことを特徴とした請求項 3 記載の文書分類装置。

【請求項 5】

前記文書は、特許文書であることを特徴とした請求項 1 ~ 4 のいずれかに記載の文書分類装置。

【請求項 6】

適切な個数の複数の分類を求める対象として分類する文書を入力する入力手段と、
前記入力された文書と予め 1 つの文書に複数の分類の付与された文書集合との類似度をキーワードを抽出して算出する文書類似度算出手段と、

前記予め分類の付与された文書集合から前記入力された文書と最も高い類似度から順に指定数の文書を抽出する文書抽出手段と、

前記抽出した指定数の文書について、前記類似度の高いものほど順位を上位とする順位を前記文書について求め、前記抽出した指定数の文書において出現した複数の分類のそれぞれについて、その分類が出現した文書の順位が上位のものほど大とする値と、前記類似度をかけあわせた値とを、前記抽出した指定数の文書のうちその分類が出現した文書だけ足し合わせて値を求め、その値をその分類のスコアとすることで、その分類のスコアを算出するスコア算出手段と、

前記算出した分類のスコアのうち最も高い分類を特定し、該特定した分類のスコアに指定値をかけた値を求め、該求めた値より大きい複数の分類を前記入力された文書の分類として抽出する分類集合抽出手段として

コンピュータを機能させるためのプログラム。

【請求項 7】

適切な個数の複数の分類を求める対象として分類する文書を入力する入力手段と、
前記入力された文書と予め 1 つの文書に複数の分類の付与された文書集合との類似度をキーワードを抽出して算出する文書類似度算出手段と、

前記予め分類の付与された文書集合から前記入力された文書と最も高い類似度から順に指定数の文書を抽出する文書抽出手段と、

前記抽出した指定数の文書において出現していた複数の分類それぞれについて、その分類が前記抽出した指定数の文書のうち出現した文書の個数を求め、その個数をその分類のスコアとすることで、その分類のスコアを算出するスコア算出手段と、

前記算出したスコアが大きい分類順に、前記抽出した指定数の文書に付与されている複数の分類の数の平均の分類数分を前記入力された文書の分類として抽出する分類集合抽出手段として

コンピュータを機能させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

10

20

30

40

50

本発明は、分類したい文書と類似した文書を、検索において高精度で知られるBM25やSMARTの方式で収集し、その収集した文書群で出現頻度の大きい分類にその文書を分類する文書分類装置及びプログラムに関する発明である。本発明は特に、一つの文書に複数の分類が付与される、Multi-classの分類問題を扱い、出現頻度の大きい分類のうち、どの分類までを、その文書の分類とするかを自動で決定する枠組みとなっている。

【背景技術】

【0002】

従来、サポートベクトルマシン法や最大エントロピー法などの機械学習法を利用した、Multi-classの分類問題に関する研究（非特許文献1参照）では、効果的な方法があった。しかし、類似文書を収集し、それら文書を利用して、Multi-classの分類問題を扱う方法では、効果的な方法がなかった。特に特許分類では、文書数が多くサポートベクトルマシン法や最大エントロピー法などの機械学習法は利用しにくい問題もあった。

10

【非特許文献1】平博順、春野雅彦、Support Vector Machineによるテキスト分類における属性選択、情報処理学会論文誌、Vol.41, No.4, 2000, p.1113-1123。

【発明の開示】

【発明が解決しようとする課題】

【0003】

上記従来の機械学習法を利用して分類する方法では、文書数が多く、しかも、一つの文書に複数の分類が付与されるものは、正確に分類を付与できるものではなかった。

【0004】

20

本発明は上記問題点の解決を図り、一つの文書に複数の分類が付与される、Multi-classの分類問題を扱い、出現頻度の大きい分類のうち、どの分類までを、その文書の分類とするかを自動で決定することを目的とする。

【課題を解決するための手段】

【0005】

図1は本発明の文書分類装置の説明図である。図1中、1は入力部（入力手段）、2は文書抽出部（文書抽出手段）、3は文書類似度算出部（文書類似度算出手段）、4はスコア算出部（スコア算出手段）、5は分類集合抽出部（分類集合抽出手段）、6は出力部（出力手段）である。

【0006】

30

本発明は、前記従来の課題を解決するため次のような手段を有する。

【0007】

(1)：分類する文書を入力する入力手段1と、前記入力された文書と予め分類の付与された文書集合との類似度をキーワードを抽出して算出する文書類似度算出手段3と、前記予め分類の付与された文書集合から前記入力された文書と最も類似する指定数の文書を抽出する文書抽出手段2と、前記類似度を加味した同じ分類の数により前記抽出した指定数の文書の分類のスコアを算出するスコア算出手段4と、前記算出したスコアが指定値より大きい分類を抽出する分類集合抽出手段5とを備える。このため、入力した文書に、自動で分類を付与することができる。

【0008】

40

(2)：分類する文書を入力する入力手段1と、前記入力された文書と予め分類の付与された文書集合との類似度をキーワードを抽出して算出する文書類似度算出手段3と、前記予め分類の付与された文書集合から前記入力された文書と最も類似する指定数の文書を抽出する文書抽出手段2と、前記抽出した指定数の文書の分類が何個の文書に現れたかにより、前記分類のスコアを算出するスコア算出手段4と、前記算出したスコアが大きい分類順に前記抽出した指定数の文書に付けられた平均の分類数分抽出する分類集合抽出手段5とを備える。このため、予め分類の付与された文書集合から抽出する文書の指定数を設定するだけで、入力した文書に、自動で分類を付与することができる。

【0009】

(3)：前記(1)又は(2)の文書分類装置において、前記抽出した複数の文書の分

50

類の技術的観点を二次元の表にし、該表に分類された技術がどこにあるかの印を設ける。このため、分類の付与されていない技術的観点（開発されていない技術）が何であるかを容易に見つけることができる。

【0010】

(4)：前記(3)の文書分類装置において、前記技術的観点を並べ替え、前記印が付いていない箇所を集め直す。このため、分類のない穴をより容易に見つけることができる。

【0011】

(5)：前記(1)～(4)の文書分類装置において、前記文書は、特許文書とする。このため、特許文書にFターム等の分類を自動で付与することができる。

10

【0012】

(6)：分類する文書を入力する入力手段1と、前記入力された文書と予め分類の付与された文書集合との類似度をキーワードを抽出して算出する文書類似度算出手段3と、前記予め分類の付与された文書集合から前記入力された文書と最も類似する指定数の文書を抽出する文書抽出手段2と、前記類似度を加味した同じ分類の数により前記抽出した指定数の文書の分類のスコアを算出するスコア算出手段4と、前記算出したスコアが指定値より大きい分類を抽出する分類集合抽出手段5として、コンピュータを機能させるためのプログラムとする。このため、このプログラムをコンピュータにインストールすることで、入力した文書に、自動で分類を付与することができる文書分類装置を容易に提供することができる。

20

【0013】

(7)：分類する文書を入力する入力手段1と、前記入力された文書と予め分類の付与された文書集合との類似度をキーワードを抽出して算出する文書類似度算出手段3と、前記予め分類の付与された文書集合から前記入力された文書と最も類似する指定数の文書を抽出する文書抽出手段2と、前記抽出した指定数の文書の分類が何個の文書に現れたかにより、前記分類のスコアを算出するスコア算出手段4と、前記算出したスコアが大きい分類順に前記抽出した指定数の文書に付けられた平均の分類数分抽出する分類集合抽出手段5として、コンピュータを機能させるためのプログラムとする。このため、このプログラムをコンピュータにインストールすることで、予め分類の付与された文書集合から抽出する文書の指定数を設定するだけで、入力した文書に、自動で分類を付与することができる文書分類装置を容易に提供することができる。

30

【発明の効果】

【0014】

本発明によれば次のような効果がある。

【0015】

(1)：文書類似度算出手段で入力された文書と予め分類の付与された文書集合との類似度をキーワードを抽出して算出し、文書抽出手段で前記予め分類の付与された文書集合から前記入力された文書と最も類似する指定数の文書を抽出し、スコア算出手段で前記類似度を加味した同じ分類の数により前記抽出した指定数の文書の分類のスコアを算出し、分類集合抽出手段で前記算出したスコアが指定値より大きい分類を抽出するため、入力した文書に、自動で分類を付与することができる。

40

【0016】

(2)：文書類似度算出手段で入力された文書と予め分類の付与された文書集合との類似度をキーワードを抽出して算出し、文書抽出手段で予め分類の付与された文書集合から前記入力された文書と最も類似する指定数の文書を抽出し、スコア算出手段で抽出した指定数の文書の分類が何個の文書に現れたかにより、前記分類のスコアを算出し、分類集合抽出手段で算出したスコアが大きい分類順に前記抽出した指定数の文書に付けられた平均の分類数分抽出するため、予め分類の付与された文書集合から抽出する文書の指定数を設定するだけで、入力した文書に、自動で分類を付与することができる。

【0017】

50

(3) : 抽出した複数の文書の分類の技術的観点を二次元の表にし、該表に分類された技術がどこにあるかの印を設けるため、分類の付与されていない技術的観点(開発されていない技術)が何であるかを容易に見つけることができる。

【0018】

(4) : 技術的観点を並べ替え、印が付いていない箇所を集め直すため、分類のない穴をより容易に見つけることができる。

【0019】

(5) : 文書は、特許文書とするため、特許文書にFターム等の分類を自動で付与することができる。

【発明を実施するための最良の形態】

【0020】

本発明は、分類したい文書と類似した文書を、検索において高精度で知られるBM25やSMARTの方式で収集し、その文書群で出現頻度の大きい分類にその文書を分類する。特に、一つの文書に複数の分類が付与される、Multi-classの分類問題を扱い、出現頻度の大きい分類のうち、どの分類までを、その文書の分類とするかを自動で決定する枠組みとなっている。

【0021】

(1) : 文書分類装置の説明

図1は文書分類装置の説明図である。図1において、文書分類装置には、入力部(入力手段)1、文書抽出部(文書抽出手段)2、文書類似度算出部(文書類似度算出手段)3、スコア算出部(スコア算出手段)4、分類集合抽出部(分類集合抽出手段)5、出力部(出力手段)6が設けてある。

【0022】

入力部1は、特許文書等の文書を入力する入力手段である。文書抽出部2は、分類したい文書と類似した文書(k個)を抽出する文書抽出手段である。文書類似度算出部3は、文書間の類似度を算出する文書類似度算出手段である。スコア算出部4は、分類のスコアを算出するスコア算出手段である。分類集合抽出部5は、分類のスコアにより、分類したい文書の分類集合(スコアが指定値以上のもの)を抽出する分類集合抽出手段である。出力部6は、分類したい文書の分類を出力する出力手段である。

【0023】

(2) : 特許の文書分類装置の説明

特許文書(特許文献)は、IPC、FI、Fターム(F-term)等で分類されている。特に、F-termは、一定の技術範囲(テーマ)を種々の技術的観点から多観点で区別したものであり、例えば、目的、用途、構造、材料、製法、処理操作方法、制御手段など多数の技術的観点から技術を区別したタームリストに基づいている。このため、一つの特許文書には、通常、複数のF-term(特許分類)が付与されている。以下、文書として特許文書を用いる場合の説明をする。

【0024】

図2は特許文書分類装置の説明図である。図2において、特許文書分類装置には、入力部(入力手段)1、KDOC抽出部(KDOC抽出手段)2、文書類似度算出部(文書類似度算出手段)3、スコア($Score_{M1}(x)$)算出部(スコア算出手段)4、F-term xの集合抽出部(F-term xの集合抽出手段)5、出力部(出力手段)6が設けてある。

【0025】

入力部1は、特許文書を入力する入力手段である。KDOC抽出部2は、分類したい特許文書と類似した特許文書(k個)を抽出するKDOC抽出手段である。なお、ここでKDOCは、抽出したk個の特許文書である。文書類似度算出部3は、特許文書間の類似度を算出する文書類似度算出手段である。スコア($Score_{M1}(x)$)算出部4は、特許分類のスコア($Score_{M1}(x)$)を算出するスコア算出手段である。F-term xの集合抽出部5は、特許分類のスコアにより、分類したい特許文書のF-term xの集合を抽出する分類集合抽出手段である。出力部6は、分類したい特許文書のF-term xの集合を出力する出力手段である。

10

20

30

40

50

【 0 0 2 6 】

(3) : 特許文書の分類処理の説明

図 3 は特許文書の分類処理フローチャートである。以下、図 3 の処理 S 1 ~ S 5 に従って説明する。

【 0 0 2 7 】

S 1 : 入力部 1 に、分類したい特許文書を入力する。

【 0 0 2 8 】

S 2 : KDOC抽出部 2 は、入力した分類したい特許文書と類似した k 個の特許文書 (KDOC) を抽出する。ここで、文書類似度算出部 3 で、入力した分類したい特許文書と学習データとして与えられた特許文書集合 (データベース等の格納手段内の) との類似度を求める。学習データとして与えられた特許文書集合は、正しい F-term の分類の付与された文書集合である。k 個の特許文書の取り出しには、ruby-ir toolkit を利用した。k は実験で定める値である。

10

【 0 0 2 9 】

S 3 : スコア (Score_{M1}(x)) 算出部 4 は、特許分類のスコア (Score_{M1}(x)) を算出する。

【 0 0 3 0 】

S 4 : F-term x の集合抽出部 5 は、特許分類のスコアにより、分類したい特許文書の F-term x の集合 (スコアが指定値以上のもの) を抽出する。

【 0 0 3 1 】

S 5 : 出力部 6 は、分類したい特許文書の F-term x の集合を出力する。

20

【 0 0 3 2 】

図 4 は入力特許文書と選択された特許文書との類似度を求める処理フローチャートである。以下、図 4 の処理 S 1 1 ~ S 1 2 に従って説明する。

【 0 0 3 3 】

S 1 1 : 文書類似度算出部 3 は、入力の特許文書からキーワードを抽出する。このキーワードとしては、形態素解析技術を利用して、名詞を取り出した。

【 0 0 3 4 】

S 1 2 : 文書類似度算出部 3 は、次に学習データにある与えられた入力のテーマ (テーマは特に与えなくてもよい) を持つすべての特許文書から、上記キーワードを少なくとも一つ含む特許文書を取り出し、該取り出した特許文書の Sim_{SMART} を算出する。この Sim_{SMART} を学習データにあるそれぞれの特許文書との間の類似度として用いる。

30

【 0 0 3 5 】

(4) : F-term x の集合の取り出しの説明

F-term x の集合の取り出しには、以下のように四つの方法がある。

【 0 0 3 6 】

a) 方法 1 の説明

特許分類装置 (KDOC抽出部 2) は、まず、入力と最も類似した k 個の特許文書を、学習データとして与えられた特許文書集合 (正しい F-term の分類の付与された文書集合) から取り出す。この k 個の特許文書を KDOC と呼ぶことにする。文書の取り出しには、ruby-ir toolkit を利用した。k は、実験で定める値である。

40

【 0 0 3 7 】

(ruby-ir toolkit の参考文献)

ruby-ir-eng, "Masao Utiyama", "Information Retrieval Module for Ruby", 2005, ("www2.nict.go.jp/jt/a132/members/mutyama/software")

特許分類装置 (スコア算出部 4) は、次に、KDOC を以下の式 (1) にしたがってソートすることで、F-term x のスコア (Score_{M1}(x)) を計算する。

【 0 0 3 8 】

【数 1】

$$Score_{M1}(x) = \sum_{i=1}^k ((k_r)^i \times score_{doc}(i) \times role(x, i)), \quad (1)$$

【0039】

ここで、

$$\begin{aligned} role(x, i) &= 1 \quad (\text{もし } i \text{ 番目の文書が F-term } x \text{ の分類を持つ場合}) \\ &= 0 \quad (\text{その他の場合}) \end{aligned}$$

ただし、 $score_{doc}(i)$ は、入力文書と選択された文書との類似度が i 番目に大きいとされた文書の類似度の値であり、 k_r は実験により定められる定数である。なお、 $score_{doc}(i)$ を、次のように簡単にすることもできる。

【0040】

$$score_{doc}(i) = 1001 - i$$

特許分類装置（分類集合抽出部 5）は、最終的に、以下の式（2）を満足する F-term x の集合を取り出す。

【0041】

$$\{ x \mid Score_{M1}(x) \geq k_p \times \max_y Score_{M1}(y) \} \dots (2)$$

ただし、 k_p は、実験により定められる定数である。この取り出された F-term x の集合が求める分類である。

【0042】

方法 1 の利用例の説明

（下の F-term1、F-term2 などは、各文書にふられている F-term である）

文書 A 入力文書との類似度 100 F-term1

文書 B 入力文書との類似度 90 F-term1 F-term2

文書 C 入力文書との類似度 80 F-term1

文書 D 入力文書との類似度 70 F-term3

だったとし、 $k_r = 0.99$ とすると、

F-term1 のスコアは、 $100+90*0.99+80*0.99^2=267.5$

F-term2 のスコアは、 $90*0.99=89.1$

F-term3 のスコアは、 $70*0.99^3=67.9$

となる。

【0043】

$k_p = 0.9$ とすると、トップのスコアの 267.5 の 0.9 倍の 240.8 以上のスコアの分類を取り出す。この場合、F-term1 だけがそれを満足するので、F-term1 だけが答えとして取り出されることになる。

【0044】

b) 方法 2 の説明

文書分類装置は、まず、方法 1 と同様に KDOC を取り出す。文書分類装置は、次に、F-term x が KDOC において、何個の文書に現れたかを数える。この数を $F_{KDOC}(x)$ で記すと、文書分類装置は、最終的に以下の式（3）を満足する F-term x の集合を取り出すことになる。

【0045】

$$\{ x \mid F_{KDOC}(x) \geq k_u \times k \},$$

ただし、 k_u は、実験により定められる定数である。ただし、 $k_u = 0.5$ のとき、この方法は、オリジナルの k 近傍法と同一になる。

【0046】

c) 方法 3 の説明

文書分類装置は、まず、方法 1 と同様に KDOC を取り出す。文書分類装置は、次に、 $F_{KDOC}(x)$ を計算する。文書分類装置は、最終的に、 $F_{KDOC}(x)$ の値の大きい順に k_f 個の

10

20

30

40

50

F-term を取り出し、これを求める分類とする。ここで、 k_f は、実験により定める定数である。

【 0 0 4 7 】

d) 方法 4 の説明

文書分類装置は、まず、方法 1 と同様に KDOC を取り出す。文書分類装置は、次に、 $F_{KDOC}(x)$ を計算する。文書分類装置は、最終的に、 $F_{KDOC}(x)$ の値の大きい順に k_a 個の F-term を取り出し、これを求める分類とする。ただし、 k_a は、KDOC にあるそれぞれの文書にふられた F-term の分類の個数の平均である。

【 0 0 4 8 】

上記それぞれの方法の有効性を確認するために、以下のベースラインとなる方法を実験で利用した。

【 0 0 4 9 】

(1) ベースライン 1

文書分類装置は、まず、学習データにある、与えられたテーマ分類を持つすべての特許文書から全てのF-term分類を取り出す。

【 0 0 5 0 】

文書分類装置は、ランダムに k_b 個のF-termを取り出し、これを求める分類とする。

ただし、 k_b は、与えられたテーマ分類を持つ特許文書にふられたF-term分類の個数の平均である。

【 0 0 5 1 】

(2) ベースライン 2

文書分類装置は、まず、学習データにある、与えられたテーマ分類を持つすべての特許文書から全てのF-term分類を取り出し、それをその分類が出現した文書数の大きい順に並べかえる。文書分類装置は、分類が出現した文書数の大きい順に k_b 個のF-termを取り出しそれを求める分類とする。ただし、 k_b は、与えられたテーマ分類を持つ特許文書にふられたF-term分類の個数の平均である。

【 0 0 5 2 】

(3) オリジナルの k 近傍法

(引用文献)

Fukunaga,1972;Okamoto and Yugami,1997;Yang and Liu,1999;Duda et al.,2001;Guo et al.,2004

オリジナルの k 近傍法をそれぞれのF-term分類に用いる方法である。文書分類装置は、まず、方法 1 と同様に KDOC を取り出す。それぞれの F-term 分類ごとに文書分類装置は、KDOCの中でその分類を持った記事数 (NUM_+) と、その分類を持たない記事数 (NUM_-) を求める。文書分類装置は、 NUM_+ の値が NUM_- 以上のF-term分類を取り出し、これを求める分類とする。この方法は、次の説明とも等価である。

【 0 0 5 3 】

文書分類装置は、まず、方法 1 と同様に KDOC を取り出す。文書分類装置は、次に、 $F_{KDOC}(x)$ を計算する。文書分類装置は、最終的に、以下の式を満足するF-term xの集合を取り出す。

$$\{ x \mid F_{KDOC}(x) \geq 0.5 \times k \}$$

【 0 0 5 4 】

(5) : 文書間の類似度の計算の説明

学習データにおけるそれぞれの特許文書と、入力の特許文書の間の類似を計算するために以下の四つの方法を利用できる。

【 0 0 5 5 】

a) SMART の説明

文書分類装置は、まず、入力の特許文書からキーワードを取り出す。キーワードとしては、形態素解析技術を利用して、名詞を取り出した。次に、学習データにある与えられた入力のテーマを持つすべての特許文書から、上記キーワードを少なくとも一つ含む文書を

10

20

30

40

50

取り出す。文書分類装置（文書類似度算出部3）は、それぞれの取り出した文書の Sim_{SMART} を算出するために以下の式（3）を使う。 Sim_{SMART} を入力文書と学習データにあるそれぞれの特許文書との間の類似度として用いる。

【0056】

【数2】

$$Sim_{SMART} = \sum_{t \in T} (W_d \times W_q), \quad (3)$$

$$W_d = \frac{1 + \log(tf)}{1 + \log(avtf)} \times \frac{1}{0.8 + 0.2 \frac{utf}{pivot}}, \quad (4)$$

10

$$W_q = (1 + \log(qtf)) \times \log \frac{N+1}{n} \quad (5)$$

【0057】

この式において、T は入力の特許文書と取り出された特許文書の両方に現れたキーワードの集合を意味し、tf はキーワード t が取り出された文書において出現した回数を意味し、avtf は取り出された文書において取り出されたキーワードそれぞれの出現の平均を意味し、qtf は入力の文書におけるキーワード t の出現した回数を意味し、utf は取り出された文書におけるキーワードの異なりの数を意味し、pivot は学習データの全文書における文書ごとのキーワードの異なりの数の平均を意味し、N は学習データにおける与えられた入力のテーマ分類をもつ特許文書の総数を意味し、n はキーワード t が現れた文書の数を意味する。

20

【0058】

SMART は、情報検索のキーワードの重み付け法のひとつである（引用文献；Singhal et al., 1996; Singhal, 1997）。

【0059】

b) BM25の説明

文書分類装置は、まず、入力の特許文書からキーワードを取り出す。キーワードとしては、形態素解析技術を利用して、名詞を取り出した。次に、学習データにある与えられた入力のテーマ分類を持つすべての特許文書から、上記キーワードを少なくとも一つ含む文書を取り出す。文書分類装置（文書類似度算出部3）は、それぞれの取り出した文書の Sim_{BM25} を算出するために以下の式（6）を使う。 Sim_{BM25} を入力文書と学習データにあるそれぞれの特許文書との間の類似度として用いる。

30

【0060】

【数3】

$$Sim_{BM25} = \sum_{t \in T} (W_d \times W_q), \quad (6)$$

40

$$W_d = \frac{(k_1 + 1)tf}{k_1((1 - b) + b \frac{dl}{avdl}) + tf}, \quad (7)$$

$$W_q = \frac{(k_3 + 1)qtf}{k_3 + qtf} \log \frac{N}{n} \quad (8)$$

【0061】

この式に置いて T、tf、qtf、N、n は、SMART のものと同じである。dl は取り出した記事の長さであり、avdl は全文書での記事の長さの平均であり、 k_1 、 k_3 それと b は実験で

50

定める定数である。ruby-ir toolkitのデフォルト値として、 $k_1 = 1$ 、 $k_3 = 1000$ 、 $b = 1$ の値を利用した。BM25のオリジナルの式の $\log \{ (N-n+0.5)/(n+0.5) \}$ の代りに $\log(N/n)$ を利用した。これは、オリジナルの式だとマイナスのスコアを出力するためである。実験において修正した式の方が高い精度を出すことを確認した。

【0062】

BM25は、情報検索のキーワードの重み付け手法の一つである（引用文献；Robertson et al., 1994）。

【0063】

c) Tfidfの説明

文書分類装置は、まず、入力の特許文書からキーワードを取り出す。キーワードとしては、形態素解析技術を利用して、名詞を取り出した。次に、学習データにある与えられた入力のテーマ分類を持つすべての文書から、上記キーワードを少なくとも一つ含む文書を取り出す。文書分類装置（文書類似度算出部3）は、それぞれの取り出した文書の Sim_{Tfidf} を算出するために以下の式（9）を使う。 Sim_{Tfidf} を入力文書と学習データにあるそれぞれの文書との間の類似度として用いる。

【0064】

【数4】

$$Sim_{Tfidf} = \sum_{t \in T} tf \times \log \frac{N}{n}, \quad (9)$$

10

20

この式で、 T 、 tf 、 N 、 n は、SMARTのものと同一である。

【0065】

d) Overlapの説明

文書分類装置は、まず、入力の特許文書からキーワードを取り出す。キーワードとしては、形態素解析技術を利用して、名詞を取り出した。次に、学習データにある与えられた入力のテーマ分類を持つすべての文書から、上記キーワードを少なくとも一つ含む文書を取り出す。文書分類装置（文書類似度算出部3）は、それぞれの取り出した文書の $Sim_{Overlap}$ を算出するために以下の式（10）を使う。 $Sim_{Overlap}$ を入力文書と学習データにあるそれぞれの文書との間の類似度として用いる。

30

【0066】

【数5】

$$Sim_{Overlap} = \sum_{t \in T} 1, \quad (10)$$

この式で、 T は、SMARTのものと同一である。

【0067】

（6）：実験結果の説明

図5は実験結果の説明図である。図5において、キーワードは、特許文書の要約の部分と請求項の部分から取り出した。Dry runのデータは、各手法のパラメータを決めるのに利用した。Formal runのデータでの実験結果が、手法の性能を示している。図5の表で最も性能の高い方法に*を付与し、--は0.01の有意差を持って*の方法より劣っていることを意味する。この有意差検定には、両側検定のt検定を利用している。実験結果からSMARTと方法1を利用する方法が最もよいことがわかる。

40

【0068】

（7）：文書分類コンテストの説明

図6はNTCIR-5 Patent WorkshopでのFormal runの説明図である。図6において、NTCIR-5 Patent Workshopは、文書分類のコンテストであり、我々のチームも含めて、3チームが参加した。我々のシステム（文書分類装置）は他のチームと圧倒的な精度差があり、

50

システムの優秀性がうかがえる。我々のシステム 1 は、BM25と方法 1 を用いる方法で細かい実装は上述の手法の比較実験のときとは異なっている。

【 0 0 6 9 】

なお、上記のコンテストは、特許文書のテーマ分類が与えられたときに、入力の日本語特許文書の F-term の分類を求めるもので、評価には、F-measure を使っている。F-measure は、再現率(recall)の逆数と適合率(precision)の逆数の平均の逆数である。再現率は、正解の分類のうち、正解の出力の割合であり、適合率は、すべての出力のうち、正解の出力の割合である。式で表現すると以下ようになる。

【 0 0 7 0 】

【数 6】

$$F\text{-measure} = \frac{2}{\frac{1}{\text{再現率}} + \frac{1}{\text{適合率}}}$$

$$\text{再現率} = \frac{\text{正解出力数}}{\text{正解分類数}}, \quad \text{適合率} = \frac{\text{正解出力数}}{\text{すべての出力数}}$$

10

20

【 0 0 7 1 】

(8) : 新しい特許の可能性の発見の説明

このように、本発明は、文書分類に関する発明である。分類したい文書と類似した文書を、検索において高精度で知られる BM25 や SMART の方式で収集し、その文書群で出現頻度の大きい分類にその文書を分類する。特に、一つの文書に複数の分類が付与される、Multi-class の分類問題を扱い、出現頻度の大きい分類のうち、どの分類までを、その文書の分類とするかを自動で決定する枠組みとなっている。

【 0 0 7 2 】

特許文書には、複数の特許を分類するためのコードがふられている。そのコードは一般には人手で付与されているが、本発明を利用すれば、ある程度自動でもコードを付与することができるようになり、人手の作業を軽減する効果がある。また、特許データを自動分類できると、以下の効果もある。特許文書には、Fタームという種々の観点から特許を分類するための分類コードがあり、これを使うと、各特許がどの問題を、どういう方法で扱っているかがわかる。各特許ごとにこれらの情報を整理し、図 7 の表のデータを(自動で)作成すれば、どの問題を、どの方法で扱った特許はあって、どの問題を、どの方法で扱った特許はないかがわかる。

30

【 0 0 7 3 】

図 7 は新しい特許の可能性の発見の説明図である。例えば、図 7 の新しい特許の可能性の発見の表で、左から右にある技術的観点である方法 1 ~ 方法 10、上から下に他の技術的観点である問題 1 ~ 問題 7 が設けてある。丸は Fタームが付けられたものを示している。例えば、左上の丸は特許 1 の Fターム(方法 1、問題 1)が付けられたものである。

40

【 0 0 7 4 】

楕円で示したところは、特許がなく、問題 3 ~ 6 を、方法 3 ~ 5 で扱った新しい特許を考えることができ、新しい特許を発見することができる可能性がある。本発明は、そのような特許の可能性を発見することを支援するシステムにおいても利用できる。なお、図 7 では、Fタームが付けられたものを単に丸で示したが、頻度情報を付加することもできる。例えば、同じ Fタームが付けられた特許の数の大小により、丸、二重丸、三重丸のように異なる表示をすることもできる。

【 0 0 7 5 】

(技術的観点の並べ替えの説明)

50

図7の表を出したあと、さらに、問題1、2、3・・・方法1、2、3・・・を並べ替えて、空白の箇所集め直すということが考えられる。

【0076】

・方法A

問題のベクトルの次元を、方法の個数とし、方法のベクトルの次元を、問題の個数とし、それぞれのベクトルの要素には、その問題の方法またはその方法の問題にあたる特許があれば「1」なければ「0」として作成し、そのベクトルの近さに基づいて各問題のベクトル、各方法のベクトルを並べる。

【0077】

近い問題同士はなるべく隣同士に、遠い問題同士はなるべく離れた場所に、近い方法同士はなるべく隣同士に、遠い方法同士はなるべく離れた場所になるように、それぞれ並べ替える。

10

【0078】

・方法B

方法Aをより詳細にした方法であり、まず、最も近い問題同士をくっつける。そのくっつけたリストの端の問題のいずれかと最も近い問題をくっつける。さらに、そのくっつけたリストの端の問題のいずれかと最も近い問題をくっつける。

【0079】

これを繰り返す。そのリストのならばの順に並べ替える。方法も同様にリストを作成し並べ替える。

20

【0080】

・方法C

双対尺度法を利用するものである。

【0081】

(文献)

上田太一郎、刈田正雄、本田和恵",実践ワークショップExcel 徹底活用多変量解析", "秀和システム", 2003,

双対尺度法では、固有値計算により、問題と方法の両方を一つの二次元の図に似たもの同士を配置する能力がある。双対尺度法への入力、問題と方法の二次元の表であり、それぞれの問題1、2、3・・・と方法1、2、3・・・が交わる欄にはその問題をその方法で扱う特許があれば「1」なければ「0」を記入する。そのような表を入力すれば双対尺度法では、固有値計算により、問題と方法の両方を一つの二次元の図に似たもの同士を配置できる。

30

【0082】

この二次元の図は、第1固有値に基づく軸と第2固有値に基づく軸の二つを利用するが、ここで第1固有値に基づく軸のみを利用する。問題1、2、3・・・と方法1、2、3・・・、それぞれで、第1固有値に基づく軸でのそれぞれの値を利用して、もとの問題と方法の表において、問題1、2、3・・・と方法1、2、3・・・、それぞれをその値の順に並べ替える。そうすると、表の対角線付近に「1」をより多く持つように並べ替えることができる。

40

【0083】

上記方法A、B、Cの説明では、ベクトルでの要素の値、双対尺度法への入力の表の各要素の値を、その問題をその方法で扱う特許があれば「1」なければ「0」としていたが、逆にその問題をその方法で扱う特許があれば「0」なければ「1」としてもよい。

【0084】

例えば方法Cで、その問題をその方法で扱う特許があれば「0」なければ「1」として表を並べ替えれば、図8のようになる。

【0085】

図8は技術的観点の並べ替えの説明図である。図8において、表の対角線付近には空欄(先行特許がないことを示している。)が集中していることがわかる。新たな特許の可能

50

性を探するために、大きな先行特許がないところを探すのにこの方法は役に立つ。ここでは、上記方法 A、B、C を示したが、他の方法で並べ替える方法でもよい。例えば、ランダムに並べ替える表を複数作り、そして空欄が連続して出現し表において大きな長方形を形成し、その長方形の大きさで判断して複数作った表のうち、その長方形の大きさが最も大きいものがよいとして、それに並べ替えるという方法を利用してもよい。

【0086】

なお、実際に本発明の手法を利用することで、2005年度に国立情報学研究所において開催された評価型ワークショップNTCIR5特許文書分類タスクのFターム分類のサブタスクにおいて参加3団体のうち、最もよい精度を出すことができた。

【0087】

(9)：プログラムインストールの説明

入力部(入力手段)1、文書抽出部(文書抽出手段)2、KDOC抽出部(KDOC抽出手段)2、文書類似度算出部(文書類似度算出手段)3、スコア算出部(スコア算出手段)4、スコア(Score_{M1}(x))算出部4、分類集合抽出部(分類集合抽出手段)5、F-term xの集合抽出部(F-term xの集合抽出手段)5、出力部(出力手段)6等は、プログラムで構成でき、主制御部(CPU)が実行するものであり、主記憶に格納されているものである。このプログラムは、一般的な、コンピュータ(情報処理装置)で処理されるものである。このコンピュータは、主制御部、主記憶、ファイル装置、表示装置、キーボード等の入力手段である入力装置などのハードウェアで構成されている。

【0088】

このコンピュータに、本発明のプログラムをインストールする。このインストールは、フロッピー、光磁気ディスク等の可搬型の記録(記憶)媒体に、これらのプログラムを記憶させておき、コンピュータが備えている記録媒体に対して、アクセスするためのドライブ装置を介して、或いは、LAN等のネットワークを介して、コンピュータに設けられたファイル装置にインストールされる。そして、このファイル装置から処理に必要なプログラムステップを主記憶に読み出し、主制御部が実行するものである。

【図面の簡単な説明】

【0089】

【図1】本発明の文書分類装置の説明図である。

【図2】本発明の特許文書分類装置の説明図である。

【図3】本発明の特許文書の分類処理フローチャートである。

【図4】本発明の入力特許文書と選択された特許文書の間の類似度を求める処理フローチャートである。

【図5】本発明の実験結果の説明図である。

【図6】本発明のNTCIR-5 Patent WorkshopでのFormal runの説明図である。

【図7】本発明の新しい特許の可能性の発見の説明図である。

【図8】本発明の技術的観点の並べ替えの説明図である。

【符号の説明】

【0090】

- 1 入力部(入力手段)
- 2 文書抽出部(文書抽出手段)
- 3 文書類似度算出部(文書類似度算出手段)
- 4 スコア算出部(スコア算出手段)
- 5 分類集合抽出部(分類集合抽出手段)
- 6 出力部(出力手段)

10

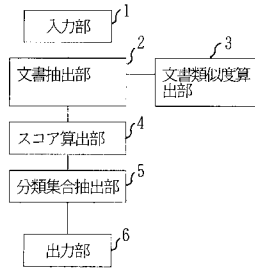
20

30

40

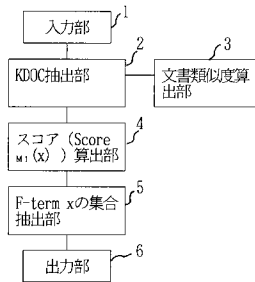
【図 1】

文書分類装置の説明図



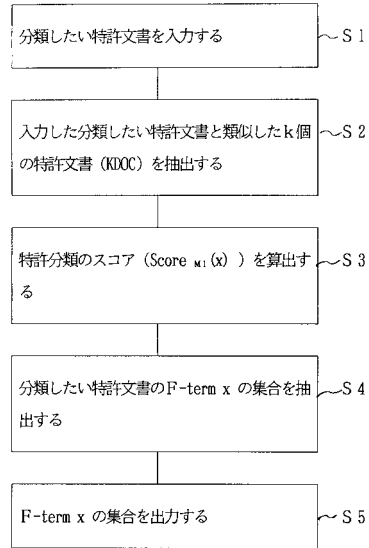
【図 2】

特許文書分類装置の説明図



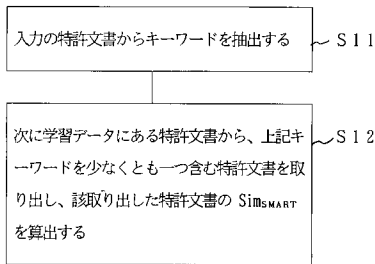
【図 3】

特許文書の分類処理フローチャート



【図 4】

入力特許文書と選択された特許文書の間の類似度を求める
処理フローチャート



【図 5】

実験結果の説明図

類似度算出手法	パラメータ	F-measure	
		Dry run	Formal run
ベースライン 1	—	0.0324	0.0396
ベースライン 2	—	0.3991	0.2962
オリジナルの k 近傍法			
SMART	k = 31	0.4186	0.2941
BM25	k = 21	0.4131	0.3009
Overlap	k = 21	0.3823	0.2689
Tfidf	k = 101	0.3196	0.1998
Method 1			
SMART	k = 101, k _r = 0.99, k _p = 0.3	0.5350*	0.4525*
BM25	k = 101, k _r = 0.99, k _p = 0.3	0.5237	0.4403
Overlap	k = 101, k _r = 0.99, k _p = 0.3	0.4764	0.4040
Tfidf	k = 301, k _r = 1, k _p = 0.3	0.4323	0.3766
Method 2			
SMART	k = 51, k _u = 0.3	0.5232	0.4048
BM25	k = 101, k _u = 0.2	0.5161	0.3979
Overlap	k = 51, k _u = 0.3	0.4721	0.3610
Tfidf	k = 101, k _u = 0.2	0.4271	0.3621
Method 3			
SMART	k = 51, k _f = 8	0.5148	0.4067
BM25	k = 101, k _f = 8	0.5054	0.3941
Overlap	k = 101, k _f = 9	0.4640	0.3644
Tfidf	k = 501, k _f = 10	0.4319	0.3431
Method 4			
SMART	k = 101	0.5254	0.4346
BM25	k = 51	0.5132	0.4289
Overlap	k = 101	0.4650	0.3937
Tfidf	k = 301	0.4293	0.3634

フロントページの続き

(56)参考文献 特開2006-099478(JP,A)
特開2005-070989(JP,A)
エクスメディア,Excel 2000 for Windows SUPER MASTER
 ,株式会社エクスメディア,2000年 8月 2日,第1版,P.400-403

(58)調査した分野(Int.Cl.,DB名)
G06F 17/30
JSTPlus(JDreamII)