(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2007-323454
(P2007-323454A)
(43)公開日　平成19年12月13日(2007.12.13)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	210D	5B075
G06F 19/00	(2006.01)	G06F	17/30	350C	
		G06F	19/00	130	

審査請求 未請求　請求項の数 7　OL　（全 15 頁）

(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)

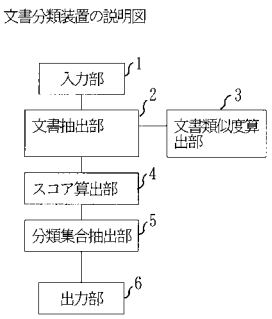
特許法第３０条第１項適用申請有り　２００５年１２月６日～９日　国立情報学研究所主催の「ＮＴＣＩＲ Ｗｏｒｋｓｈｏｐ ５ Ｍｅｅｔｉｎｇ（Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ Ｆｉｆｔｈ ＮＴＣＩＲ Ｗｏｒｋｓｈｏｐ Ｍｅｅｔｉｎｇ ｏｎ Ｅｖａｌｕａｔｉｏｎ ｏｆ Ｉｎｆｏｒｍａｔｉｏｎ Ａｃｃｅｓｓ Ｔｅｃｈｎｏｌｏｇｉｅｓ：Ｉｎｆｏｒｍａｔｉｏｎ Ｒｅｔｒｉｅｖａｌ，Ｑｕｅｓｔｉｏｎ Ａｎｓｗｅｒｉｎｇ ａｎｄ Ｃｒｏｓｓ－Ｌｉｎｇｕａｌ Ｉｎｆｏｒｍａｔｉｏｎ Ａｃｃｅｓｓ）」で発表

(71)出願人　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

Ｆターム(参考)　5B075 ND03　ND40　NR05　NR12　PR04　QM08　UU06

(54)【発明の名称】文書分類装置及びプログラム

(57)【要約】
【課題】入力した文書に、自動で分類を付与すること。
【解決手段】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。
【選択図】図１



【特許請求の範囲】
【請求項１】
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類似する指定数の文書を抽出する文書抽出手段と、
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とを備えることを特徴とした文書分類装置。
【請求項２】
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類似する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】
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類似する指定数の文書を抽出する文書抽出手段と、
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】
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類似する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】
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。
【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】
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】
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】
（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。

【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。

【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。

【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。

【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】

（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。

【００２６】

（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Ｓ１～Ｓ５に従って説明する。

【００２７】

Ｓ１：入力部１に、分類したい特許文書を入力する。

【００２８】

Ｓ２：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。

【００２９】

Ｓ３：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。

【００３０】

Ｓ４：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。

【００３１】

Ｓ５：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。

【００３２】

図４は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。以下、図４の処理Ｓ１１～Ｓ１２に従って説明する。

【００３３】

Ｓ１１：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。

【００３４】

Ｓ１２：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。

【００３５】

（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。

【００３６】

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。

【００３７】

（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/mutiyama/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。

【００３８】

【数１】

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

【００３９】

ここで、

role(x,i) = 1 （もしi 番目の文書が F-term x の分類を持つ場合）

= 0（その他の場合）

ただし、$score_{doc}(i)$は、入力文書と選択された文書の間の類似度が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】

$\{ x \mid Score_{M1}(x) \geqq k_p \times \max_y Score_{M1}(y) \}$ ・・・・（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、kr = 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】

kp = 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】

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

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 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。

【００４７】

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。

【００４８】

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。

【００４９】

(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。

【００５０】

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。

【００５１】

(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。

【００５２】

(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。

【００５３】

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。

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

【００５４】

（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。

【００５５】

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】

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

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

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

【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。

【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。

【００６０】

【数３】

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

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

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

【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】

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

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。

【００６６】

【数５】

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

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。

【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、NTCI

R-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、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。

【００６９】

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のようになる。

【００７０】

【数６】

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

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

【００７１】

（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。

【００７２】

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。

【００７３】

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。

【００７４】

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。

【００７５】
（技術的観点の並べ替えの説明）
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。
【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。そのリストのならびの順に並べ替える。方法も同様にリストを作成し並べ替える。
【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は、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。
【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。
【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。新たな特許の可能性を探すために、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、フロッピィ、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。

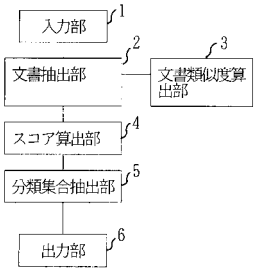
【符号の説明】

【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）

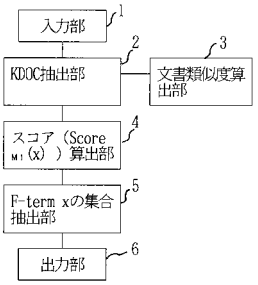
【図１】

文書分類装置の説明図



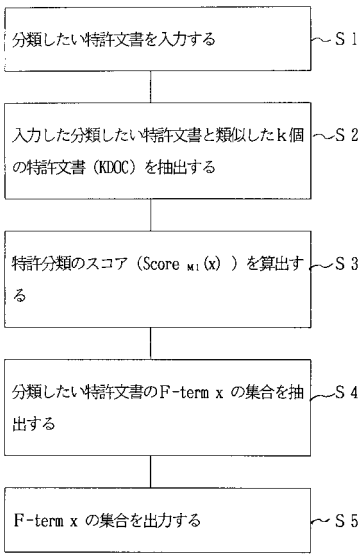
【図２】

特許文書分類装置の説明図



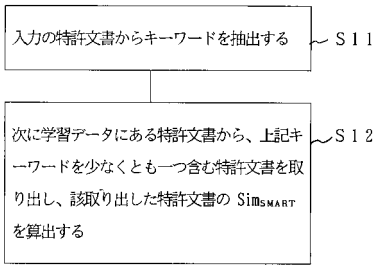
【図３】

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^{--}

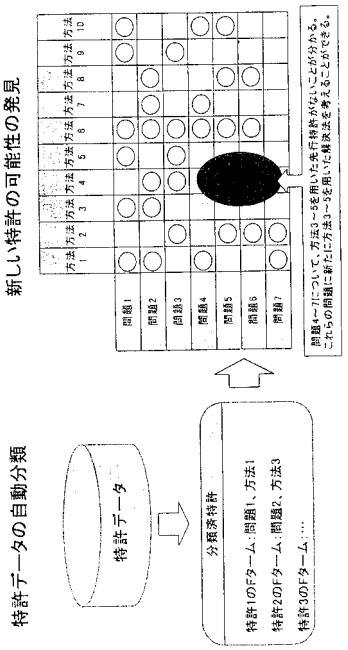
【図６】

NTCIR-5 Patent Workshop でのFormal runの説明図

システム	F-measure
我々のシステム１	0.4379
我々のシステム２	0.4190
我々のシステム３	0.4168
我々のシステム４	0.4258
我々のシステム５	0.4393
チーム１のシステム１	0.1604
チーム１のシステム２	0.1648
チーム１のシステム３	0.1646
チーム１のシステム４	0.1652
チーム１のシステム５	0.1447
チーム１のシステム６	0.1614
チーム１のシステム７	0.1606
チーム１のシステム８	0.1581
チーム１のシステム９	0.1696
チーム１のシステム１０	0.1579
チーム２のシステム１	0.2830
チーム２のシステム２	0.1435
チーム２のシステム３	0.1110

【図７】

新しい特許の可能性の発見の説明図



【図８】

技術的観点の並べ替えの説明図