

(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2008-165480

(P2008-165480A)

(43) 公開日 平成20年7月17日(2008.7.17)

(51) Int.Cl. F I テーマコード (参考)
G06F 17/30 (2006.01) G06F 17/30 370Z 5B075

審査請求 未請求 請求項の数 15 O L (全 49 頁)

<p>(21) 出願番号 特願2006-354123 (P2006-354123) (22) 出願日 平成18年12月28日 (2006.12.28)</p>	<p>(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 NR00 PQ36</p>
--	---

(54) 【発明の名称】 確信度付与装置及び方法及びプログラム

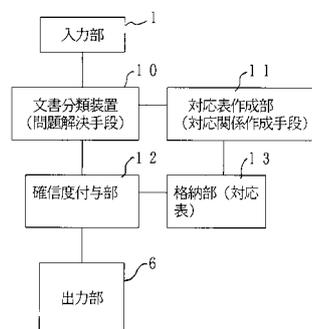
(57) 【要約】

【課題】 出力される解答の確信度を付与すること。

【解決手段】 問題を入力する入力手段1と、前記入力された問題を解いてその解答を複数抽出し、該抽出した前記解答と所定値とを出力する問題解決手段10と、予め解答が付与された問題を複数個用意し、該問題をそれぞれ前記問題解決手段10に入力してそれぞれの解答を出力するときに、前記所定値と前記解答を出力し、前記所定値と前記解答と前記解答の確信度を求め、前記所定値と確信度の対応関係を作成する対応関係作成手段11と、前記入力手段1より新しい問題を入力して前記問題解決手段10で解答を順序化して出力するとき、ある解答が出力される前記所定値を求め、前記対応関係からある解答の確信度を付与して出力する確信度付与手段12とを備える。

【選択図】 図7

確信度付与装置の説明図



【特許請求の範囲】**【請求項 1】**

問題を入力する入力手段と、

前記入力された問題を解いてその解答を複数抽出し、該抽出した前記解答と所定値とを出力する問題解決手段と、

予め解答が付与された問題を複数個用意し、該問題をそれぞれ前記問題解決手段に入力してそれぞれの解答を出力するときに、前記所定値と前記解答と前記解答の確信度を求め、前記所定値と確信度の対応関係を作成する対応関係作成手段と、

前記入力手段より新しい問題を入力して前記問題解決手段で解答を順序化して出力するとき、ある解答が出力される前記所定値を求め、前記対応関係からある解答の確信度を付与して出力する確信度付与手段とを備えることを特徴とした確信度付与装置。

10

【請求項 2】

前記確信度として、全ての出力のうちの正解出力の割合である適合率を用いることを特徴とした請求項 1 記載の確信度付与装置。

【請求項 3】

前記確信度として、正解数のうち、正解出力の割合である再現率を用いることを特徴とした請求項 1 記載の確信度付与装置。

【請求項 4】

前記確信度として、再現率の逆数と適合率の逆数の平均の逆数である F 値を用いることを特徴とした請求項 1 記載の確信度付与装置。

20

【請求項 5】

前記確信度付与手段により確信度を付与して出力する数を、F 値を最大にする数とすることを特徴とした請求項 1 ~ 4 のいずれかに記載の確信度付与装置。

【請求項 6】

前記確信度として、個々の解答の正解率を用いることを特徴とした請求項 1 記載の確信度付与装置。

【請求項 7】

予め解答が付与された問題を複数個用意し、該問題をそれぞれ前記問題解決手段に入力してそれぞれの解答を出力するときに、該解答がぎりぎり出力される前記所定値を求め、該ぎりぎり出力される解答が正解しているかを調べて前記所定値の時の正解率を求め、どうゆう所定値なら正解か不正解かの事例を機械学習して学習結果を蓄える機械学習手段を備え、

30

前記確信度付与手段は、前記対応関係として前記学習結果を用いることを特徴とした請求項 6 記載の確信度付与装置。

【請求項 8】

前記所定値として、複数観点の所定値を用い、前記対応関係作成手段で前記複数観点の所定値と確信度の対応関係を作成することを特徴とした請求項 1 ~ 7 記載のいずれかに記載の確信度付与装置。

【請求項 9】

前記問題解決手段が文書分類装置であり、前記問題が分類を付与する文書であり、前記解答が前記文書の分類であることを特徴とした請求項 1 ~ 8 のいずれかに記載の確信度付与装置。

40

【請求項 10】

前記問題解決手段が情報検索装置であり、前記問題が質問の文書であり、前記解答が前記質問の文書より検索された文書であることを特徴とした請求項 1 ~ 8 のいずれかに記載の確信度付与装置。

【請求項 11】

前記問題解決手段で、スコアを求めて解答を出力する場合、前記所定値として、ある解答のスコアを最初の解答のスコアで割った値 (kp) を用いることを特徴とした請求項 1 ~ 10 のいずれかに記載の確信度付与装置。

50

【請求項 1 2】

前記問題解決手段で、解答を出力するときの前記所定値として出力順位 (kj) を用いることを特徴とした請求項 1 ~ 10 のいずれかに記載の確信度付与装置。

【請求項 1 3】

前記問題解決手段で、スコアを求めて解答を出力する場合、前記所定値として、スコア (kl) を用いることを特徴とした請求項 1 ~ 10 のいずれかに記載の確信度付与装置。

【請求項 1 4】

入力手段より問題を入力し、

問題解決手段で前記入力された問題を解いてその解答を複数抽出し、該抽出した前記解答とその解答を順序化する所定値とを出力し、

対応関係作成手段で予め解答が付与された問題を複数個用意し、該問題をそれぞれ前記問題解決手段に入力してそれぞれの解答を出力するときに、前記所定値と前記解答を出力し、同じ前記所定値と前記出力したそれぞれの解答の確信度の平均を求め、前記所定値と確信度の対応関係を作成する対応関係作成手段と、

確信度付与手段で前記入力手段より新しい問題を入力して前記問題解決手段で解答を順序化して出力するとき、ある解答が出力される前記所定値を求め、前記対応関係からある解答の確信度を付与して出力することを特徴とした確信度付与方法。

【請求項 1 5】

問題を入力する入力手段と、

前記入力された問題を解いてその解答を複数抽出し、該抽出した前記解答とその解答を順序化する所定値とを出力する問題解決手段と、

予め解答が付与された問題を複数個用意し、該問題をそれぞれ前記問題解決手段に入力してそれぞれの解答を出力するときに、前記所定値と前記解答を出力し、同じ前記所定値と前記出力したそれぞれの解答の確信度の平均を求め、前記所定値と確信度の対応関係を作成する対応関係作成手段と、

前記入力手段より新しい問題を入力して前記問題解決手段で解答を順序化して出力するとき、ある解答が出力される前記所定値を求め、前記対応関係からある解答の確信度を付与して出力する確信度付与手段として、

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、例えば google (登録商標) などの検索結果で、検索の上位から検索結果の文書を提示するとき、その各文書に確信度 (精度、再現率、F 値、正解率等) を付与するものである。確信度 (正解率) は、その文書が検索結果として正しいかどうかを意味する値である。50% なら、半分の確率であっているもので、100% なら、ほぼ 100% あっているものである。これを自動で付与する検索結果への確信度付与装置及び方法及びプログラムに関する。

【背景技術】

【0002】

従来、キーワードにより文書を検索して、キーワードの出現確率等により検索結果を順序つけて出力するシステムはあった (特許文献 1 参照)。

【特許文献 1】特許 3799447 号公報

【発明の開示】

【発明が解決しようとする課題】

【0003】

上記従来の検索結果を順序つけて出力するシステムは、効果的な方法で各文書に確信度 (正解率等) を付与する技術はなかった。

【0004】

本発明は上記問題点の解決を図り、検索結果で、検索の上位から検索結果の文書を提示

10

20

30

40

50

するときその各文書に確信度（正解率等）を自動で付与することを目的とする。

【課題を解決するための手段】

【0005】

図7は確信度付与装置の説明図である。図7中、1は入力部（入力手段）、6は出力部（出力手段）、10は文書分類装置（問題解決手段）、11は対応表作成部（対応関係作成手段）、12は確信度付与部（確信度付与手段）、13は格納部（対応表）である。

【0006】

本発明は、前記従来課題を解決するため次のような手段を有する。

【0007】

(1)：問題を入力する入力手段1と、前記入力された問題を解いてその解答を複数抽出し、該抽出した前記解答と所定値とを出力する問題解決手段10と、予め解答が付与された問題を複数個用意し、該問題をそれぞれ前記問題解決手段10に入力してそれぞれの解答を出力するときに、前記所定値と前記解答と前記解答の確信度を求め（即ち、この確信度は、前記解答と、予め解答が付与された問題を照らし合わせて、出力した解答がどのくらいあっているかを調べて、確信度の定義にしたがって求める）、前記所定値と確信度の対応関係を作成する対応関係作成手段11と、前記入力手段1より新しい問題を入力して前記問題解決手段10で解答を順序化して出力するとき、ある解答が出力される前記所定値を求め、前記対応関係からある解答の確信度を付与して出力する確信度付与手段12とを備える。このため、出力される解答の確信度を付与することができ、どの解答が信頼できるかを容易に判断できる。

10

20

【0008】

(2)：前記(1)の確信度付与装置において、前記確信度として、全ての出力のうちの正解出力の割合である適合率を用いる。このため、出力された分類又は文書までの適合率（精度）を容易に判断することができる。

【0009】

(3)：前記(1)の確信度付与装置において、前記確信度として、正解数のうち、正解出力の割合である再現率を用いる。このため、再現率により正解のもれ量を容易に判断することができる。

【0010】

(4)：前記(1)の確信度付与装置において、前記確信度として、再現率の逆数と適合率の逆数の平均の逆数であるF値を用いる。このため、F値を用いて、適合率（精度）ともれ量を考慮した確信度を付与することができる。

30

【0011】

(5)：前記(1)～(4)の確信度付与装置において、前記確信度付与手段12により確信度を付与して出力する数を、F値を最大にする数とする。このため、確信度の高い分類又は文書のみを出力することができる。

【0012】

(6)：前記(1)の確信度付与装置において、前記確信度として、個々の解答の正解率を用いる。このため、出力される個々の解答の正解率（確信度）を付与することができ、どの解答が信頼できるかを容易に判断できる。

40

【0013】

(7)：前記(6)の確信度付与装置において、予め解答が付与された問題を複数個用意し、該問題をそれぞれ前記問題解決手段10に入力してそれぞれの解答を出力するときに、該解答がぎりぎり出力される前記所定値を求め、該ぎりぎり出力される解答が正解しているかを調べて前記所定値の時の正解率を求め、どうゆう所定値なら正解か不正解かの事例を機械学習して学習結果を蓄える機械学習手段を備え、前記確信度付与手段12は、前記対応関係として前記学習結果を用いる。このため、機械学習により、出力される正解の正解率を容易に付与することができる。

【0014】

(8)：前記(7)の確信度付与装置において、前記所定値として、複数観点の所定値

50

を用い、前記機械学習手段に前記複数観点の所定値のときの正解か不正解かの事例を機械学習させる。このため、複数観点の所定値により、より正確な機械学習を行うことができる。

【0015】

(9)：前記(1)～(8)の確信度付与装置において、前記問題解決手段10が文書分類装置であり、前記問題が分類を付与する文書であり、前記解答が前記文書の分類である。このため、出力される分類に確信度を付与することができる。

【0016】

(10)：前記(1)～(8)の確信度付与装置において、前記問題解決手段10が情報検索装置であり、前記問題が質問の文書であり、前記解答が前記質問の文書より検索された文書である。このため、出力される解答の文書に確信度を付与することができる。

10

【0017】

(11)：前記(1)～(10)の確信度付与装置において、前記問題解決手段10で、スコアを求めて解答を出力する場合、前記所定値として、ある解答のスコアを最初の解答のスコアで割った値(kp)を用いる。このため、あるkpにより解答に確信度を付与することができる。

【0018】

(12)：前記(1)～(10)の確信度付与装置において、前記問題解決手段10で、解答を出力するときの前記所定値として出力順位(kj)を用いる。このため、kjにより解答に確信度を付与することができる。

20

【0019】

(13)：前記(1)～(10)の確信度付与装置において、前記問題解決手段10で、スコアを求めて解答を出力する場合、前記所定値として、スコア(kl)を用いる。このため、klにより解答に確信度を付与することができる。

【発明の効果】

【0020】

本発明によれば次のような効果がある。

【0021】

(1)：予め解答が付与された問題から対応関係作成手段で、所定値と確信度の対応関係を作成しておき、新しい問題を入力して問題解決手段で解答を順序化して出力するとき、ある解答が出力される所定値を求め、前記対応関係からある解答の確信度を付与して出力するため、出力される解答の確信度を付与することができ、どの解答まで信頼できるかを容易に判断できる。

30

【0022】

(2)：前記確信度として、全ての出力のうち、正解出力の割合である適合率を用いるため、出力された分類又は文書までの適合率(精度)を容易に判断することができる。

【0023】

(3)：前記確信度として、正解数のうち、正解出力の割合である再現率を用いるため、再現率により正解のもれ量を容易に判断することができる。

【0024】

(4)：前記確信度として、再現率の逆数と適合率の逆数の平均の逆数であるF値を用いるため、F値を用いて、適合率(精度)ともれ量を考慮した確信度を付与することができる。

40

【0025】

(5)：前記確信度付与手段により確信度を付与して出力する数を、F値を最大にする数とするため、確信度の高い分類又は文書のみを出力することができる。

【0026】

(6)：前記確信度として、個々の解答の正解率を用いるため、出力される個々の解答の正解率(確信度)を付与することができ、どの解答が信頼できるかを容易に判断できる。

50

【 0 0 2 7 】

(7) : 機械学習手段を備え前記確信度付与手段で、前記対応関係として前記学習結果を用いるため、機械学習により、出力される正解の正解率を容易に付与することができる。

【 0 0 2 8 】

(8) : 前記所定値として、複数観点の所定値を用い、機械学習手段に前記複数観点の所定値のときの正解か不正解かの事例を機械学習させるため、複数観点の所定値により、より正確な機械学習を行うことができる。

【 0 0 2 9 】

(9) : 前記問題解決手段が文書分類装置であり、前記問題が分類を付与する文書であり、前記解答が前記文書の分類であるため、出力される分類に確信度を付与することができる。

10

【 0 0 3 0 】

(1 0) : 前記問題解決手段が情報検索装置であり、前記問題が質問の文書であり、前記解答が前記質問の文書より検索された文書であるため、出力される解答の文書に確信度を付与することができる。

【 0 0 3 1 】

(1 1) : 前記問題解決手段で、スコアを求めて解答を出力する場合、前記所定値として、ある解答のスコアを最初の解答のスコアで割った値 (kp) を用いるため、あるkpにより解答に確信度を付与することができる。

20

【 0 0 3 2 】

(1 2) : 問題解決手段で、解答を出力するときの前記所定値として出力順位 (kj) を用いるため、kjにより解答に確信度を付与することができる。

【 0 0 3 3 】

(1 3) : 前記問題解決手段で、スコアを求めて解答を出力する場合、前記所定値として、スコア (kl) を用いるため、klにより解答に確信度を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 】

【 0 0 3 4 】

本発明は、情報検索結果で、検索の上位から検索結果の文書を提示するときにその各文書に確信度 (精度、再現率、F 値、正解率等) を自動で付与するものである。付与の方法は、あらかじめ正解のセットを用意しておき、その正解セットでどういう場合に、どのくらいの精度かの対応表を求めておく。そして新しい文書がきたとき、その文書がどういう場合か調べて、先に求めた表から確信度を求める。なお、表以外に他の同様の方法でも可能である。また、文書検索以外の、出力がリスト化されているものならばどのようなものも扱える。

30

【 0 0 3 5 】

§ 1 : 表に基づく確信度付与の説明

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とき確信度の付与を行う。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確信度を参考とすることができる。

40

【 0 0 3 6 】

(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。

【 0 0 3 7 】

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

50

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 (画面表示、印刷) する出力手段である。この出力部 6 の出力は、画面表示せず、プログラム内部で、他のプログラムに出力したり、プログラム内部で変数の値として、算出したりすることも含むものである。

【0038】

(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。

10

【0039】

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。

20

【0040】

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。

【0041】

(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。

30

【0042】

S1 : 入力部 1 に、分類したい特許文書を入力する。

【0043】

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。

40

【0044】

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。

【0045】

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。

【0046】

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。

【0047】

50

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S11～S12に従って説明する。

【0048】

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。

【0049】

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。

10

【0050】

(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。

【0051】

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。

20

【0052】

(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($Score_{M_1}(x)$)を計算する。

【0053】

【数1】

$$Score_{M_1}(x) = \sum_{i=1}^k ((k_r)^i \times score_{doc}(i) \times role(x, i)), \quad (1)$$

30

【0054】

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。

【0055】

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

40

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。

【0056】

$$\{ x \mid Score_{M_1}(x) \geq k_p \times \max_y Score_{M_1}(y) \} \cdots (2)$$

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。

【0057】

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)

50

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。

【 0 0 5 8 】

$k_p = 0.9$ 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。

【 0 0 5 9 】

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、最終的に以下の式を満足する F-term x の集合を取り出すことになる。

【 0 0 6 0 】

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

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。

【 0 0 6 1 】

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。

【 0 0 6 2 】

(5) : 対応表の説明

上記方法 1 ~ 3 で k_p 、 k_u 、 k_f を変化すると、取り出す F-term の数が変化することになる。ここで入力文書に正解データ (正しい F-term が付与されている) がある場合、変化させた各 k_p 、 k_u 、 k_f と確信度 (精度、再現率、F 値) の対応表を作成することができる。

【 0 0 6 3 】

例えば、方法 1 を利用した場合の k_p と特許文書の F-term の精度 (適合率) の対応の場合、

$k_p=0.9$ の時に選ばれた F-term の精度	95%
$k_p=0.8$ の時に選ばれた F-term の精度	85%
$k_p=0.7$ の時に選ばれた F-term の精度	80%
$k_p=0.6$ の時に選ばれた F-term の精度	75%
$k_p=0.5$ の時に選ばれた F-term の精度	65%
$k_p=0.4$ の時に選ばれた F-term の精度	50%
$k_p=0.3$ の時に選ばれた F-term の精度	45%
$k_p=0.2$ の時に選ばれた F-term の精度	20%
$k_p=0.1$ の時に選ばれた F-term の精度	10%

上記の対応が各入力文書 (正しい F-term が付与されている特許文書) ごとに出力される。したがって、精度 (適合率) は、特許文書分類装置に入力された特許文書ごとに出力され、特許文書ごとに異なる精度となることがあるので、各特許文書の精度の平均をとる。例えば、 $k_p=0.9$ の時の各特許文書の精度の平均を取るものである。なお、再現率、F 値の

10

20

30

40

50

場合も精度と同様に各特許文書の平均を取って対応表を作成する。

【 0 0 6 4 】

図 5 はkpとF値の対応の説明図である。図 5 において、kpとF値 (F-measure) の対応は、kpが 0.1から 0.3まではF値が上昇し、0.4から 0.9までF値が低下している。kpが 0.3の時F値が最大となっている。なお、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ことになる。

【 0 0 6 5 】

図 6 はkpと再現率と精度の対応の説明図である。図 6 において、横軸が再現率(Recall)、縦軸が精度(Precision)であり、グラフの黒点の数字がkpの値である。この図では、再現率が大きくなるほど精度は低下している。すなわち、kpが小さくなる(選ばれるF-termの数が増える)ほど精度が低下し、再現率が上がっていることがわかる。

10

【 0 0 6 6 】

(6) : 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、入力の特許文書との間の類似度を計算するために以下の四つの方法を利用できる。

【 0 0 6 7 】

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、名詞を取り出す。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。

20

【 0 0 6 8 】

【数 2】

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

30

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

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

【 0 0 6 9 】

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。

40

【 0 0 7 0 】

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)。

【 0 0 7 1 】

b) BM25の説明

50

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。

【0072】

【数3】

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

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

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

10

【0073】

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。

20

【0074】

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）。

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。

30

【0075】

【数4】

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

40

【0076】

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

【0077】

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_{Ov}

50

$erlap$ 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。

【0078】

【数5】

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

【0079】

この式で、 T は、SMARTのものと同一である。

10

【0080】

(7) : 文書検索結果の評価の説明

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。この評価には、図5のようにF-measure (F値) を使かうことができる。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(再現率が大きいと正解の漏れが少なくなる)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なになる。

【0081】

20

【数6】

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

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

30

【0082】

(8) : 単語の認識の説明

a) 形態素解析システムの説明

日本語を単語に分割するために、単語抽出部が行う形態素解析システムが必要になる。ここではChaSenについて説明する(奈良先端大で開発されている形態素解析システム茶釜 <http://chasen.aist-nara.ac.jp/index.html.jp> で公開されている)。

【0083】

これは、日本語文を分割し、さらに、各単語の品詞も推定してくれる。例えば、「学校へ行く」を入力すると以下の結果を得ることができる。

40

【0084】

学校	ガッコウ	学校	名詞 - 一般		
へ	へ	へ	助詞 - 格助詞 - 一般		
行く	イク	行く	動詞 - 自立	五段・カ行促音便	基本型

E O S

このように各行に一個の単語が入るように分割され、各単語に読みや品詞の情報が付与される。

【0085】

b) 英語の品詞タグつけの説明

英語の品詞タグつけシステムとしては、次の Brillのものが有名である。

50

【 0 0 8 6 】

Eric Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, Computational Linguistics, Vol. 21, No. 4, p.543-565, 1995.

これは、英語文の各単語の品詞を推定してくれるものである。

【 0 0 8 7 】

(9) : 表に基づく確信度付与の説明

a) kpを利用する場合の説明

予め問題と解答の組を大量に集める。問題は、F-termをふるべき特許であり、解答は、その特許のF-termである。これを評価データと呼ぶ。前記文書分類装置でいくつかのkpごとに、上記評価データでF-termを出力し評価し、そのときの精度（適合率）、再現率、F値等の確信度を求める。更に同じkpに対応する全ての特許のF-termの精度（適合率）、再現率、F値の平均値を求める。そうすると、kpと精度（適合率）、再現率、F値の対応表が完成する。

【 0 0 8 8 】

次に、新しい特許が文書分類装置に入ってくると、F-termが出力される。各F-termがぎりぎり出力されるkpを求める。この求め方は、以下のようになる。

【 0 0 8 9 】

あるF-termのスコア（Score）を最初のF-term（最もスコアの大きいF-term）のスコアで割った値がそのF-termがぎりぎり出力されるkpとなる。（kpの定義によりこうなる、式（2）を参照こと）。スコアは式（1）等を利用して求める。

【 0 0 9 0 】

各F-termのkpが求めれば、先の対応表に基づいて、各F-termに対応する精度（適合率）、再現率、F値をくっつけて表示する。そのF-termまでのF-term群に対する精度（適合率）、再現率、F値である（個々のF-termの精度（適合率）、再現率、F値ではない）。個々のF-termのものについては、後に説明する。

【 0 0 9 1 】

以下、図面に基づいて説明する。図7は確信度付与装置の説明図である。図7において、確信度付与装置には、入力部1、出力部6、文書分類装置（問題解決手段）10、対応表作成部（対応関係作成手段）11、確信度付与部12、格納部（対応表）13が設けてある。

【 0 0 9 2 】

入力部1は、情報を入力する入力手段である。出力部6は、情報を出力する出力手段である。文書分類装置10は、前に説明した文書の分類を行う文書分類手段（問題解決手段）である（図1、図2参照）。対応表作成部（対応関係作成手段）11は、kpと精度（適合率）、再現率、F値の対応関係（表）を作成する対応関係（表）作成手段である。確信度付与部12は、文書分類装置10で付与した分類に精度（適合率）、再現率、F値、正解率等の確信度を付与する確信度付与手段である。格納部（対応表）13は、対応表作成部11が作成した対応表を格納する格納手段である。

【 0 0 9 3 】

図8は対応表作成処理フローチャートである。以下、図8の処理S21～S25にしたがって説明する。

【 0 0 9 4 】

S21：入力部1より、予め問題と解答の組（ここでは特許文書とそのF-term）を大量に入力し、文書分類装置10の格納手段に格納する。

【 0 0 9 5 】

S22：文書分類装置10は、前記入力された1つの特許文書と類似する他の特許文書を検索して分類を求める（F-termを求める）。

【 0 0 9 6 】

S23：文書分類装置10は、前記類似する他の特許文書の分類（F-term）が何個の特

許文書に現れたか等により、前記求めた分類 (F-term) のスコアを算出する。

【0097】

S24: 対応表作成部11は、kpを変化させた時に文書分類装置10より出力されるそれぞれの分類 (F-term) の確信度を求める。

【0098】

S25: 対応表作成部11は、前記S21で入力した特許文書全てについて、文書分類装置10で分類を付与 (F-termを求め) し、kpを変化させて確信度を求め、更に同じkpに対応する全ての特許文書の確信度の平均値を求め、対応表を作成する。

【0099】

図9は確信度付与処理フローチャートである。以下、図9の処理S31~S35にしたがって説明する。

10

【0100】

S31: 入力部1より、新たな文書 (F-termが付与されていない特許文書) を入力する。

【0101】

S32: 文書分類装置10は、前記入力された特許文書と類似する特許文書 (前記処理S21で入力された特許文書) を検索して分類を求める (F-termを求める)。

【0102】

S33: 文書分類装置10は、前記類似する特許文書の分類 (F-term) が何個の特許文書に現れたか等により、前記付与した分類 (F-term) のスコアを算出する。

20

【0103】

S34: 確信度付与部12は、各分類 (F-term) がぎりぎり出力されるkpを求める。

【0104】

S35: 確信度付与部12は、格納部13の対応表から前記求めたkpに対応する確信度を各F-termに付与して出力部より出力する。

【0105】

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確信度により容易に決定することができる。

30

【0106】

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。

【0107】

なお、確信度付与部12で、確信度を付与して出力する分類 (F-term) の数は、F値の最大のところまで、精度 (適合率) がある値 (規定値) 以上のところまで、再現率がある値 (規定値) 以下のところまで出力する等を行うことにより、不要な出力を少なくすることができる。

40

【0108】

b) 出力順位を利用する場合の説明

出力順位を利用する方法の場合、文書分類装置で出力する分類 (F-term) をkj位までを出力システムとする。いくつかkjの値を変えて、このシステムで評価データの問題を解き、精度 (適合率)、再現率、F値の値を求める。そうすると、kj (順位) と精度 (適合率)、再現率、F値の対応表が完成する。

【0109】

新しい特許が入ってくると、文書分類装置で先の方法でF-termを出力する。各F-termがぎりぎり出力されるkjを求める。すると、出力される順位がkjとなる (kjの定義によりこうなる、他の方法ではこの部分は異なった方法になる)。

50

【0110】

各Ftermのkjが求めれば、先の対応表に基づいて、各Ftermに対応する精度（適合率）、再現率、F値をくっつけて表示する。これは、そのF-termまでの文書群に対する精度（適合率）、再現率、F値であることに注意。（これは個々のF-termの精度（適合率）、再現率、F値でない。個々のF-termのものについて、以下の個々の値の算出の場合を参照のこと）。

【0111】

図10は対応表作成処理フローチャートである。以下、図10の処理S41～S45にしたがって説明する（確信度付与装置は図7参照、但し、ここでは図7で説明したkpの代わりにkjを用いるものである）。

10

【0112】

S41：入力部1より、予め問題と解答の組（ここでは特許文書とそのF-term）を大量に入力し、文書分類装置10の格納手段に格納する。

【0113】

S42：文書分類装置10は、前記入力された1つの特許文書と類似する他の特許文書を検索して分類を求める（F-termを求める）。

【0114】

S43：文書分類装置10は、前記類似する他の特許文書の分類（F-term）が何個の特許文書に現れたか等により、前記求めた分類（F-term）の順位kjを算出する。

20

【0115】

S44：対応表作成部11は、kjを変化させた時に文書分類装置10より出力されるそれぞれの分類（F-term）の確信度を求める。

【0116】

S45：対応表作成部11は、前記S41で入力した特許文書全てについて、文書分類装置10で分類を付与（F-termを求め）し、kjを変化させて確信度を求め、更に同じkjに対応する全ての特許文書の確信度の平均値を求め、対応表を作成する。

【0117】

図11は確信度付与処理フローチャートである。以下、図11の処理S51～S55にしたがって説明する（確信度付与装置は図7参照、但し、ここでは図7で説明したkpの代わりにkjを用いるものである）。

30

【0118】

S51：入力部1より、新たな文書（F-termが付与されていない特許文書）を入力する。

【0119】

S52：文書分類装置10は、前記入力された特許文書と類似する特許文書（前記処理S41で入力された特許文書）を検索して分類を求める（F-termを求める）。

【0120】

S53：文書分類装置10は、前記類似する特許文書の分類（F-term）が何個の特許文書に現れたか等により、前記求めた分類（F-term）の順位kjを算出する。

【0121】

S54：確信度付与部12は、各F-termがぎりぎり出力されるkjを求める。

40

【0122】

S55：確信度付与部12は、格納部13の対応表から前記求めたkjに対応する確信度を各F-termに付与して出力部より出力する。

【0123】

c) スコア (Score) を利用する場合の説明

スコアを利用する方法の場合、文書分類装置で出力する分類（F-term）をスコアがkl以上のものまでを出力システムとする。いくつかklの値を変えて、このシステムで評価データの問題を解き、精度（適合率）、再現率、F値の値を求める。そうすると、kl（スコア）と精度（適合率）、再現率、F値の対応表が完成する。

50

【0124】

新しい特許が入ってくると、文書分類装置は先の方法でF-termを出力する。各F-termがぎりぎり出力されるklを求める。ここで各F-termのスコアがklとなる。(klの定義によりこうなる。他の方法ではこの部分は異なった方法になる)。

【0125】

各F-termのklが求めれば、先の対応表に基づいて、各F-termに対応する精度(適合率)、再現率、F値をくっつけて表示する。(そのF-termまでの文書群に対する精度(適合率)、再現率、F値であることに注意。個々のF-termの精度(適合率)、再現率、F値でない。個々のF-termのものについては、以下の個々の値の算出の場合を参照)。

【0126】

図12は対応表作成処理フローチャートである。以下、図12の処理S61~S65にしたがって説明する(確信度付与装置は図7参照、但し、ここでは図7で説明したkpの代わりにklを用いるものである)。

【0127】

S61:入力部1より、予め問題と解答の組(ここでは特許文書とそのF-term)を大量に入力し、文書分類装置10の格納手段に格納する。

【0128】

S62:文書分類装置10は、前記入力された1つの特許文書と類似する他の特許文書を検索して分類を求める(F-termを求める)。

【0129】

S63:文書分類装置10は、前記類似する他の特許文書の分類(F-term)が何個の特許文書に現れたか等により、前記求めた分類(F-term)のスコア(kl)を算出する。

【0130】

S64:対応表作成部11は、klを変化させた時に文書分類装置10より出力されるそれぞれの分類(F-term)の確信度を求める。

【0131】

S65:対応表作成部11は、前記S61で入力した特許文書全てについて、文書分類装置10で分類を付与(F-termを求め)し、klを変化させて確信度を求め、更に同じklに対応する全ての特許文書の確信度の平均値を求め、対応表を作成する。

【0132】

図13は確信度付与処理フローチャートである。以下、図13の処理S71~S75にしたがって説明する(確信度付与装置は図7参照、但し、ここでは図7で説明したkpの代わりにklを用いるものである)。

【0133】

S71:入力部1より、新たな文書(F-termが付与されていない特許文書)を入力する。

【0134】

S72:文書分類装置10は、前記入力された特許文書と類似する特許文書(前記処理S61で入力された特許文書)を検索して分類を求める(F-termを求める)。

【0135】

S73:文書分類装置10は、前記類似する特許文書の分類(F-term)が何個の特許文書に現れたか等により、前記求めた分類(F-term)のスコアklを算出する。

【0136】

S74:確信度付与部12は、各F-termがぎりぎり出力されるklを求める。

【0137】

S75:確信度付与部12は、格納部13の対応表から前記求めたklに対応する確信度を各F-termに付与して出力部より出力する。

【0138】

以上kp、順位、スコアを利用する方法を示したが、順序化して出力するシステムであれば、他のものを利用することもできる。

10

20

30

40

50

【 0 1 3 9 】

§ 2 : 情報検索の場合の説明

(1) : 情報検索システム (情報検索装置) の説明

キーワードから文書を検索する技術 (文書検索の技術) は、例えば、次のものがある。

【 0 1 4 0 】

(単語群 A をより多く含む記事の抽出方法の説明)

情報検索の基礎知識として以下の式がある。ここで、Score(D) が大きいものを取る。

【 0 1 4 1 】

(1) 基本的な方法 (tf · idf 法) の説明

$$\text{score}(D) = \sum_w \left(\frac{\text{tf}(w,D)}{W} * \log(N/\text{df}(w)) \right)$$

10

W はユーザーが入力するキーワードの集合

tf(w,D) は文書 D での w の出現回数

df(w) は全文書で W が出現した文書の数

N は文書の総数

score(D) が高い文書を検索結果として出力する。

【 0 1 4 2 】

(2) Robertson らの Okapi weighting の説明

(文献)

村田真樹, 馬青, 内元清貴, 小作浩美, 内山将夫, 井佐原均 “ 位置情報と分野情報を用いた情報検索 ” 自然言語処理 (言語処理学会誌) 2000年 4月, 7 巻, 2 号, p.141 ~ p.160

20

の (1) 式、が性能がよいことが知られている。これの式(1) の で積を取る前の tf 項と idf 項の積が Okapi のウェイト法になって、この値を単語の重みに使う。

【 0 1 4 3 】

Okapi の式なら

$$\text{Score}(D) = \sum_w \left(\frac{\text{tf}(w,D)}{W} / (\text{tf}(w,D) + \text{length}/\text{delta}) * \log(N/\text{df}(w)) \right)$$

length は記事 D の長さ、delta は記事の長さの平均、

記事の長さは、記事のバイト数、また、記事に含まれる単語数などを使う。

30

【 0 1 4 4 】

さらに、以下の情報検索を行うこともできる。

【 0 1 4 5 】

(Okapi の参考文献)

S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford Okapi at TREC-3, TREC-3, 1994

(SMART の参考文献)

Amit Singhal AT&T at TREC-6, TREC-6, 1997

より高度な情報検索の方法として、tf · idf を使うだけの式でなく、これらの Okapi や SMART の式を用いてもよい。

40

【 0 1 4 6 】

これらの方法では、tf · idf だけでなく、記事の長さなども利用して、より高精度な情報検索を行うことができる。

【 0 1 4 7 】

今回の、単語群 A をより多く含む記事の抽出方法では、さらに、Rocchio's formula を使うことができる。

【 0 1 4 8 】

(文献)

"J. J. Rocchio", "Relevance feedback in information retrieval", "The SMART retrieval System", "Edited by G. Salton", "Prentice Hall, Inc.", "page 313-323", 19

50

71

この方法は、 $\log(N/df(w))$ のかわりに、

$$\{E(t) + k_{af} * (\text{RatioC}(t) - \text{RatioD}(t))\} * \log(N/df(w))$$

を使う。

【0149】

$E(t) = 1$ (元の検索にあったキーワード)

$= 0$ (それ以外)

$\text{RatioC}(t)$ は記事群 B での t の出現率

$\text{RatioD}(t)$ は記事群 C での t の出現率

$\log(N/df(w))$ を上式でおきかえた式でScore(D)を求めて、その値が大きいものほど、単語群 A をより多く含む記事として取り出すものである。 10

【0150】

Score(D)の の加算の際に足す単語 w の集合 W は、元のキーワードと、単語群 A の両方とする。ただし、元のキーワードと、単語群 A は重ならないようにする。

【0151】

また、他の方法として、Score(D)の の加算の際に足す。単語 w の集合 W は、単語群 A のみとする。ただし、元のキーワードと、単語群 A は重ならないようにする。

【0152】

ここでは Roccio の式で複雑な方法をとったが、単純に、単語群 A の単語の出現回数の和が大きいものほど、単語群 A をより多く含む記事として取り出すようにしてもよいし、また、単語群 A の出現の異なりの大きいものほど、単語群 A をより多く含む記事として取り出すようにしてもよい。 20

【0153】

(2) : 確信度付与の説明

予め問題と解答の組を大量に集める。問題は、情報検索の質問 (例えば、企業合併に関する記事を取り出すこと) であり、解答は、その質問に対応する記事群である。これを評価データと呼ぶ、ここで上記 (1) で説明したような情報検索システム (情報検索装置) を一つ用意する。

【0154】

質問から、形態素解析して、名詞をキーワードと取り出して、そのキーワードを利用して上記情報検索システムで記事を取り出す。そうすると、各記事はOkapi の式ならScore(D)の値を持ち、この値の大きいものが出力される。 30

【0155】

a) kp の値を利用する場合の説明

kp の値を利用する方法の場合は、ある質問の場合のScore(D)の最大値を Score_max とする。そして、 $\text{Score_max} * kp$ の文書まで出力する。いくつか kp の値を変えて、このシステムで評価データの問題を解き、精度 (適合率)、再現率、F 値等の (確信度) の値を求める。そうすると、 kp と精度 (適合率)、再現率、F 値の対応表が完成する。

【0156】

次に、新しい情報検索の質問が入ってくる。先の方法 (情報検索システム) で文書を出力する。各文書がぎりぎり出力される kp を求める。 40

【0157】

この求め方は、以下のようにする。

【0158】

ある文書のScore を最初の文書 (最もScore の大きい文書) のScore で割った値がその文書がぎりぎり出力される kp となる。(kp の定義によりこうなる。順位による方法や他の方法ではこの部分は異なった方法になる) 。

【0159】

各文書の kp が求めれば、先の対応表に基づいて、各文書に対応する精度 (適合率)、再現率、F 値をくっつけて表示する。(これは、その文書までの文書群に対する精度 (適合 50

率)、再現率、F値であることに注意。個々の文書の精度(適合率)、再現率、F値でない。個々の文書のものについては、個々の値の算出の場合を参照)。

【0160】

図14は対応表作成処理フローチャートである。以下、図14の処理S81~S85にしたがって説明する(確信度付与装置は図7参照、但し、ここでは図7の文書分類装置の代わりに情報検索システム(情報検索装置)を用いるものである)。

【0161】

S81:入力部1より、予め質問(問題)と記事(解答)の組を大量に入力し、情報検索システムの格納手段に格納する。

【0162】

S82:情報検索システムは、前記入力されたある1つの質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、前記入力された記事(解答)の情報検索を行って記事を取り出す。

【0163】

S83:情報検索システムは、Score \leq max * kpの文書(記事)まで出力する。

【0164】

S84:対応表作成部11は、kpを変化させた時に情報検索システムより出力されるそれぞれの記事の確信度を求める。

【0165】

S85:対応表作成部11は、前記S81で入力した質問全てについて、情報検索システムで記事を出力し、kpを変化させて確信度を求め、更に同じkpに対応する全ての記事の確信度の平均値を求め、対応表を作成する(対応表は格納部13に格納する)。

【0166】

図15は確信度付与処理フローチャートである。以下、図15の処理S91~S94にしたがって説明する。

【0167】

S91:入力部1より、新たな情報検索の質問を入力する。

【0168】

S92:情報検索システムは、前記入力された質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、情報検索を行って記事を取り出す。

【0169】

S93:確信度付与部12は、情報検索システムにより、各記事がぎりぎり出力されるkpを求める。

【0170】

S94:確信度付与部12は、格納部13の対応表から前記求めたkpに対応する確信度を記事に付与して出力部より出力する。

【0171】

b)出力順位を利用する場合の説明

出力順位を利用する方法の場合は、kj位までの文書(記事)を出力システムとする。これは、いくつかkjの値を変えて、この情報検索システムで、評価データの問題を解き、精度(適合率)、再現率、F値の値を求める。そうすると、kjと精度(適合率)、再現率、F値の対応表が完成する。

【0172】

次に、新しい情報検索の質問が入ってくると、先の方法(対応表作成時の)で文書出力する。そして、各文書がぎりぎり出力されるkjを求める。この出力される順位がkjとなる。(これはkjの定義によりこうなる。他の方法ではこの部分は異なった方法になる)。各文書のkjが求めれば、先の対応表に基づいて、各文書に対応する精度(適合率)、再現率、F値をくっつけて表示する。(その文書までの文書群に対する精度(適合率)、再現率、F値であることに注意、個々の文書の精度(適合率)、再現率、F値でない。個々の

10

20

30

40

50

文書のものについては、以下の個々の値の算出の場合を参照)。

【0173】

図16は対応表作成処理フローチャートである。以下、図16の処理S101～S105にしたがって説明する(確信度付与装置は図7参照、但し、ここでは図7の文書分類装置の代わりに情報検索システムを用い、kpの代わりにkjを用いるものである)。

【0174】

S101:入力部1より、予め質問(問題)と記事(解答)の組を大量に入力し、情報検索システムの格納手段に格納する。

【0175】

S102:情報検索システムは、前記入力されたある1つの質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、記事(解答)の情報検索を行って記事を取り出す。

10

【0176】

S103:情報検索システムは、kj位までの文書(記事)を出力する。

【0177】

S104:対応表作成部11は、kjを変化させた時に情報検索システムより出力されるそれぞれの文書(記事)の確信度を求める。

【0178】

S105:対応表作成部11は、前記S101で入力した質問全てについて、情報検索システムで文書(記事)を出力し、kjを変化させて確信度を求め、更に同じkjに対応する全ての文書(記事)の確信度の平均値を求め、対応表を作成する(対応表は格納部13に格納する)。

20

【0179】

図17は確信度付与処理フローチャートである。以下、図17の処理S111～S114にしたがって説明する(確信度付与装置は図7参照、但し、ここでは図7の文書分類装置の代わりに情報検索システムを用い、kpの代わりにkjを用いるものである)。

【0180】

S111:入力部1より、新たな情報検索の質問を入力する。

【0181】

S112:情報検索システムは、前記入力された質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、情報検索を行って記事を取り出す。

30

【0182】

S113:確信度付与部12は、情報検索システムにより、各記事がぎりぎり出力されるkjを求める。

【0183】

S114:確信度付与部12は、格納部13の対応表から前記求めたkjに対応する確信度を記事に付与して出力部より出力する。

【0184】

c)スコア(Score)を利用する場合の説明

40

Scoreを利用する方法の場合は、Scoreがkl以上の文書までを出力システムとする。いくつか、klの値を変えて、この情報検索システムで、評価データの問題を解き、精度(適合率)、再現率、F値の値を求める。そうすると、Scoreであるklと精度(適合率)、再現率、F値の対応表が完成する。

【0185】

次に、新しい情報検索の質問が入ってくる。先の方法(対応表の作成方法)で文書を出力する。ここで、各文書がぎりぎり出力されるklを求める。すると各文書のScoreがklとなる。(klの定義によりこうなる。他の方法ではこの部分は異なった方法になる)。

【0186】

各文書のklが求めれば、先の対応表に基づいて、各文書に対応する精度(適合率)、再

50

現率、F値をくっつけて表示する。(その文書までの文書(記事)群に対する精度(適合率)、再現率、F値であることに注意。個々の文書の精度(適合率)、再現率、F値でない。個々の文書のものについては、以下の個々の値の算出の場合を参照)

図18は対応表作成処理フローチャートである。以下、図18の処理S121~S125にしたがって説明する(確信度付与装置は図7参照、但し、ここでは図7の文書分類装置の代わりに情報検索システムを用い、kpの代わりにklを用いるものである)。

【0187】

S121:入力部1より、予め質問(問題)と記事(解答)の組を大量に入力し、情報検索システムの格納手段に格納する。

【0188】

S122:情報検索システムは、前記入力されたある1つの質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、記事(解答)の情報検索を行って記事を取り出す。

【0189】

S123:情報検索システムは、Scoreがkl以上の文書(記事)までを出力する。

【0190】

S124:対応表作成部11は、klを変化させた時に情報検索システムより出力されるそれぞれの文書(記事)の確信度を求める。

【0191】

S125:対応表作成部11は、前記S121で入力した質問全てについて、情報検索システムで文書(記事)を出力し、klを変化させて確信度を求め、更に同じklに対応する全ての文書(記事)の確信度の平均値を求め、対応表を作成する(対応表は格納部13に格納する)。

【0192】

図19は確信度付与処理フローチャートである。以下、図19の処理S131~S134にしたがって説明する(確信度付与装置は図7参照、但し、ここでは図7の文書分類装置の代わりに情報検索システムを用い、kpの代わりにklを用いるものである)。

【0193】

S131:入力部1より、新たな情報検索の質問を入力する。

【0194】

S132:情報検索システムは、前記入力された質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、情報検索を行って記事を取り出す。

【0195】

S133:確信度付与部12は、情報検索システムにより、各記事がぎりぎり出力されるklを求める。

【0196】

S134:確信度付与部12は、格納部13の対応表から前記求めたklに対応する確信度を記事に付与して出力部より出力する。

【0197】

d)klのスコアの正規化の説明

klとしては、スコアの正規化を行ったものを用いてもよい。スコアの正規化としてはいくつか方法がある。前記(1):情報検索システムの説明で説明したOkapiの式で単語分だけ加算するが、その単語の数で、元のスコアを割るという方法が最も単純な正規化としてありえる。

【0198】

次に、Wの単語が1回ずつ出現する記事を想定して、その記事のスコアで、元のスコアを割るという方法がある。

【0199】

また、方法自体を変更して、で加算するというをやめて、ベクトルにしてから、

10

20

30

40

50

スコアを求めることで正規化と同じ効果をもたせてもよい。

【0200】

例えば、あらゆる種類の単語分だけ、要素とするベクトルを作成して、各ベクトルの要素の値は、前記 Okapi の式の 内部の部分の式を利用して求めて、入力キーワードでもベクトルを作成し、検索対象の文書でもベクトルを作成する。これらベクトルの角度をスコアとする。角度を利用することで、正規化と同じ効果をもつ。

【0201】

F タームの話だと、BM25 と Okapi はほぼ同じ式だが、BM25 の式 (6) の で単語分だけ加算するが、その単語の数で、元のスコアを割るという方法が最も単純な正規化としてありえる。

【0202】

次に、W の単語が 1 回ずつ出現する記事を想定してその記事のスコアで、元のスコアを割るという方法がある。

【0203】

また、方法自体を変更して、 で加算するというをやめて、ベクトルにしてから、スコアを求めることで正規化と同じ効果をもたせてもよい。例えば、あらゆる種類の単語分だけ、要素とするベクトルを作成して、各ベクトルの要素の値は、入力の単語については、式 (8) を使い文書の単語については、式 (7) を使い、求めて、入力キーワードでもベクトルを作成し、検索対象の文書でもベクトルを作成する。これらベクトルの角度をスコアとする。角度を利用することで、正規化と同じ効果をもつ。

【0204】

e) 観点の異なる 2 つの所定値を使う場合の説明

例えば、 k_p と k_l の二つを使うことを考える。 $k_p = 0, 0.1, 0.2, \dots, 1.0$ と $k_j = 1, 2, 3, \dots, 1000$ の二つを使うことを考える。これらのあらゆる組み合わせの場合の、確信度の平均を求めて、対応表を作る。

【0205】

$k_p = 0, k_j = 1$ の場合の確信度 ...

$k_p = 0.1, k_j = 1$ の場合の確信度 ...

...

$k_p = 0, k_j = 2$ の場合の確信度 ...

$k_p = 0.1, k_j = 2$ の場合の確信度 ...

...

...

$k_p = 0, k_j = 1000$ の場合の確信度 ...

$k_p = 0.1, k_j = 1000$ の場合の確信度 ...

...

上のように対応表が求まる。

ここで、あたらしい問題が入力される。そして、解答を出力させる。解答を出力させる時点の k_p, k_j を求める。この k_p, k_j は、1 つのときと同じ方法で求められる。 k_p, k_j がわかれば上記の対応表を調べて、その場合の確信度を求めて出力する。解答を出力させる時点の k_p, k_j とぴったり同じときのデータが対応表にない場合は補間処理を行う。

【0206】

例えば、このあと、新しい入力で k_p が k_{p1} で k_j が k_{j1} であったとする。そして、 k_{p1}, k_{j1} の場合の値が表にのっていないとする。そうすると、ある種の補間処理が必要になる。その場合は、表にのっている、 k_{p1} と最も近い値の k_p と、 k_{j1} と最も近い値の k_j との組み合わせの時点の値を使ってもいいし、表にのっている、 k_{p1} をはさむ二つの k_p 、 k_{j1} をはさむ二つの k_j を用い、二つの k_p と二つの k_j から k_p, k_j をひとつずつ選ぶ全ての組み合わせの 4 つのデータの平均を使ってもよい。

【0207】

また、表にのっている k_{p1} をはさむ二つの k_p の 2 つのデータ k_{p2}, k_{p3} ($k_{p2} > k_{p3}$)

10

20

30

40

50

を k_{j1} をはさむ二つの k_j の2つのデータ k_{j2} , k_{j3} ($k_{j2} > k_{j3}$)を用い、 k_{p2} , k_{j2} のときの確信度を $p(2,2)$ 、 k_{p3} , k_{j2} のときの確信度を $p(3,2)$ 、 k_{p2} , k_{j3} のときの確信度を $p(2,3)$ 、 k_{p3} , k_{j3} のときの確信度を $p(3,3)$ とし、

$$r(2,2) = \sqrt{(k_p - k_{p2})^2 + a(k_j - k_{j2})^2}$$

$$r(3,2) = \sqrt{(k_p - k_{p3})^2 + a(k_j - k_{j2})^2}$$

$$r(2,3) = \sqrt{(k_p - k_{p2})^2 + a(k_j - k_{j3})^2}$$

$$r(3,3) = \sqrt{(k_p - k_{p3})^2 + a(k_j - k_{j3})^2}$$

として、

$$p(2,2)/r(2,2) + p(3,2)/r(3,2) + p(2,3)/r(2,3) + p(3,3)/r(3,3)$$

を

$$1/r(2,2) + 1/r(3,2) + 1/r(2,3) + 1/r(3,3)$$

で割ったものを確信度に用いてよい。

10

【0208】

ここで、 a は定数であり、あらかじめ実験で定めるか、システム利用者が予め値を与える。 \wedge はべき乗を意味し、 $\sqrt{\quad}$ は平方根を意味する。これに類する方法でもよい。他の補間方法でもよい。 k_p k_j k_l など3つ以上使う場合も同様である。

【0209】

§3：個々の値の算出の説明

(1)：文書分類装置を用いる場合の説明

a) k_p の値を利用する場合の説明

20

個々の値の算出の場合は、予め問題と解答の組を大量に集める。問題は、F-termをふるべき特許、解答は、その特許のF-termである。これを評価データと呼ぶ。

【0210】

前記文書分類装置（特許文書分類装置）で上記評価データでF-termを出力する。ここで各F-termがぎりぎり出力される k_p を求める。この求め方は、以下のようにする。

【0211】

あるF-termのスコア（Score）を最初のF-term（最もスコアの大きいF-term）のスコアで割った値がそのF-termがぎりぎり出力される k_p となる。（ k_p の定義によりこうなる、式（2）を参照のこと、順位による方法や他の方法ではこの部分は異なった方法になることに注意）スコアは式（1）等を利用して求める。

30

【0212】

前記出力された上記評価データの各F-termごとにそれが正解しているかを調べて、各 k_p の時の正解率を求める。更に同じ k_p に対応する全ての上記評価データ（特許文書）のF-termの正解率の平均値を求める。そうすると、 k_p と正解率の対応表が完成する。

【0213】

新しい特許（分類が付与されていない）が入ってくると、前記文書分類装置でF-termを出力する。各F-termがぎりぎり出力される k_p を求める。この求め方は、上記対応表作成の場合と同様であり、あるF-termのスコアを最初のF-term（最もScoreの大きいF-term）のスコアで割った値がそのF-termがぎりぎり出力される k_p となる。（ k_p の定義によりこうなる、式（2）を参照のこと、順位による方法や他の方法ではこの部分は異なった方法になることに注意）スコアは式（1）等を利用して求める。

40

【0214】

各F-termの k_p が求めれば、先の対応表に基づいて、各F-termに対応する正解率をくっつけて表示する。（この正解率は、個々のF-termの正解率であることに注意。そのF-termまでのF-term群に対する精度（適合率）、再現率、F値などとは異なるものである。）

図20は対応表作成処理フローチャートである。以下、図20の処理S141～S145にしたがって説明する（確信度付与装置は図7参照）。

【0215】

S141：入力部1より、予め問題と解答の組（ここでは特許文書とそのF-term）を大量に入力し、文書分類装置10の格納手段に格納する。

50

【 0 2 1 6 】

S 1 4 2 : 文書分類装置 1 0 は、前記入力された 1 つの特許文書と類似する他の特許文書を検索して分類を求める (F-term を求める)。

【 0 2 1 7 】

S 1 4 3 : 文書分類装置 1 0 は、前記類似する他の特許文書の分類 (F-term) が何個の特許文書に現れたか等により、前記求めた分類 (F-term) のスコアを算出する。

【 0 2 1 8 】

S 1 4 4 : 対応表作成部 1 1 は、各分類 (F-term) がきりぎり出力されるkpを求め、各分類 (F-term) ごとにそれが正解しているかを調べて、各kpのときの正解率を求める。

【 0 2 1 9 】

S 1 4 5 : 対応表作成部 1 1 は、前記 S 1 4 1 で入力した特許文書全てについて、文書分類装置 1 0 で分類を付与 (F-term を求め) し、各分類 (F-term) がきりぎり出力されるkpを求め、更に該同じkpに対応する全ての特許文書の正解率の平均値を求め、対応表を作成する (対応表は格納手段 1 3 に格納される)。

【 0 2 2 0 】

図 2 1 は確信度付与処理フローチャートである。以下、図 2 1 の処理 S 1 5 1 ~ S 1 5 5 にしたがって説明する。

【 0 2 2 1 】

S 1 5 1 : 入力部 1 より、新たな文書 (F-term が付与されていない特許文書) を入力する。

【 0 2 2 2 】

S 1 5 2 : 文書分類装置 1 0 は、前記入力された特許文書と類似する特許文書 (前記処理 S 1 4 1 で入力されたの特許文書) を検索して分類を求める (F-term を求める)。

【 0 2 2 3 】

S 1 5 3 : 文書分類装置 1 0 は、前記類似する特許文書の分類 (F-term) が何個の特許文書に現れたか等により、前記付与した分類 (F-term) のスコアを算出する。

【 0 2 2 4 】

S 1 5 4 : 確信度付与部 1 2 は、各分類 (F-term) がきりぎり出力されるkpを求める。

【 0 2 2 5 】

S 1 5 5 : 確信度付与部 1 2 は、格納部 1 3 の対応表から前記求めたkpに対応する確信度を各F-termに付与して出力部より出力する。

【 0 2 2 6 】

b) 出力順位を利用する場合の説明

出力順位を利用する方法の場合は、F-termを出力システム (前記特許文書分類装置) とする。このシステムで評価データの問題を解き、kj個目の出力のF-termがあっているかまちがっているかを調べて、kj個目の出力の正解率を求める。そうすると、kjと正解率の対応表が完成する。

【 0 2 2 7 】

新しい特許が入ってくると、先の方法 (特許文書分類装置) でF-termを出力する。そして、各F-termがきりぎり出力されるkjを求める。そのF-termが出力される順位がkjとなる。(kjの定義によりこうなる。他の方法ではこの部分は異なった方法になる)。

【 0 2 2 8 】

各F-termのkjが求めれば、先の対応表に基づいて、各F-termに対応する正解率をくっつけて表示する。

【 0 2 2 9 】

図 2 2 は対応表作成処理フローチャートである。以下、図 2 2 の処理 S 1 6 1 ~ S 1 6 5 にしたがって説明する (確信度付与装置は図 7 参照)。

【 0 2 3 0 】

S 1 6 1 : 入力部 1 より、予め問題と解答の組 (ここでは特許文書とそのF-term) を大量に入力し、文書分類装置 1 0 の格納手段に格納する。

10

20

30

40

50

【0231】

S162：文書分類装置10は、前記入力された1つの特許文書と類似する他の特許文書を検索して分類を求める（F-termを求める）。

【0232】

S163：文書分類装置10は、前記類似する他の特許文書の分類（F-term）が何個の特許文書に現れたか等により、前記求めた分類（F-term）の順位kjを算出する。

【0233】

S164：対応表作成部11は、kj個目の分類（F-term）の出力があっているか間違っているかを調べて、kj個目の出力の正解率を求める。

【0234】

S165：対応表作成部11は、前記S161で入力した特許文書全てについて、kj個目の出力の正解率を求め、更に同じkjに対応する全ての特許文書の正解率の平均値を求め、対応表を作成する（対応表は格納手段13に格納される）。

【0235】

図23は確信度付与処理フローチャートである。以下、図23の処理S171～S175にしたがって説明する（確信度付与装置は図7参照）。

【0236】

S171：入力部1より、新たな文書（F-termが付与されていない特許文書）を入力する。

【0237】

S172：文書分類装置10は、前記入力された特許文書と類似する特許文書（前記処理S161で入力されたの特許文書）を検索して分類を求める（F-termを求める）。

【0238】

S173：文書分類装置10は、前記類似する特許文書の分類（F-term）が何個の特許文書に現れたか等により、前記求めた分類（F-term）の順位kjを算出する。

【0239】

S174：確信度付与部12は、各F-termがぎりぎり出力されるkjを求める。

【0240】

S175：確信度付与部12は、格納部13の対応表から前記求めたkjに対応する確信度を各F-termに付与して出力部より出力する。

【0241】

c)スコア(score)を利用する場合の説明

スコアを利用する方法の場合は、F-termを出力システム（前記特許文書分類装置）を使用する。このシステムで評価データの問題を解き、出力される各F-termを評価する。F-termのスコアがklのものについて、そのF-termがあっているかどうかを調べて、klの場合の正解率を求める。これをあらゆるklについて求める。そうすると、klと正解率の対応表が完成する。

【0242】

新しい特許（F-termが付与されていない）が入ってくると、特許文書分類装置でF-termを出力する。ここで各F-termがぎりぎり出力されるklを求める。すると各F-termのスコアがklとなる。（klの定義によりこうなる。他の方法ではこの部分は異なった方法になる）

各F-termのklが求めれば、先の対応表に基づいて、各F-termに対応する正解率をくっつけて表示する。

【0243】

図24は対応表作成処理フローチャートである。以下、図24の処理S181～S185にしたがって説明する（確信度付与装置は図7参照）。

【0244】

S181：入力部1より、予め問題と解答の組（ここでは特許文書とそのF-term）を大量に入力し、文書分類装置10の格納手段に格納する。

10

20

30

40

50

【 0 2 4 5 】

S 1 8 2 : 文書分類装置 1 0 は、前記入力された 1 つの特許文書と類似する他の特許文書を検索して分類を求める (F-term を求める) 。

【 0 2 4 6 】

S 1 8 3 : 文書分類装置 1 0 は、前記類似する他の特許文書の分類 (F-term) が何個の特許文書に現れたか等により、前記求めた分類 (F-term) のスコア (kl) を算出する。

【 0 2 4 7 】

S 1 8 4 : 対応表作成部 1 1 は、F-term のスコアが kl のものについて、その F-term があっているかどうかを調べて、kl の場合の正解率を求める。

【 0 2 4 8 】

S 1 8 5 : 対応表作成部 1 1 は、これを前記 S 1 8 1 で入力した特許文書の分類 (F-term) のあらゆる kl についてその正解率を求める。そうすると、kl と正解率の対応表が完成する (対応表は格納手段 1 3 に格納される) 。

【 0 2 4 9 】

図 2 5 は確信度付与処理フローチャートである。以下、図 2 5 の処理 S 1 9 1 ~ S 1 9 5 にしたがって説明する (確信度付与装置は図 7 参照) 。

【 0 2 5 0 】

S 1 9 1 : 入力部 1 より、新たな文書 (F-term が付与されていない特許文書) を入力する。

【 0 2 5 1 】

S 1 9 2 : 文書分類装置 1 0 は、前記入力された特許文書と類似する特許文書 (前記処理 S 1 8 1 で入力されたの特許文書) を検索して分類を求める (F-term を求める) 。

【 0 2 5 2 】

S 1 9 3 : 文書分類装置 1 0 は、前記類似する特許文書の分類 (F-term) が何個の特許文書に現れたか等により、前記求めた分類 (F-term) のスコア (kl) を算出する。

【 0 2 5 3 】

S 1 9 4 : 確信度付与部 1 2 は、各 F-term がぎりぎり出力される kl を求める。

【 0 2 5 4 】

S 1 9 5 : 確信度付与部 1 2 は、格納部 1 3 の対応表から前記求めた kl に対応する正解率を各 F-term に付与して出力部より出力する。

【 0 2 5 5 】

(2) : 情報検索装置を用いる場合の説明

予め問題と解答の組を大量に集める。問題は、情報検索の質問 (例えば、企業合併に関する記事を取り出すこと) であり、解答は、その質問に対応する記事 (文書) 群である。これを評価データと呼ぶ、ここで前に説明したような情報検索システム (情報検索装置) を一つ用意する。

【 0 2 5 6 】

質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して上記情報検索システムで記事を取り出す。そうすると、各記事は Okapi の式なら Score(D) の値を持ち、この値の大きいものが出力される。

【 0 2 5 7 】

a) kp の値を利用する場合の説明

kp の値を利用する方法の場合は、ある質問の場合の Score(D) の最大値を Score_max とする。そして、Score_max * kp の記事 (文書) まで出力する。

【 0 2 5 8 】

前記情報検索システムで上記評価データで記事 (文書) 群を出力する。ここで各記事 (文書) がぎりぎり出力される kp を求める。この求め方は、以下のようにする。

【 0 2 5 9 】

ある記事のスコア (Score) を最初の記事 (最もスコアの大きい記事) のスコアで割った値がその記事がぎりぎり出力される kp となる。 (kp の定義によりこうなる、式 (2) を

10

20

30

40

50

参照のこと、順位による方法や他の方法ではこの部分は異なった方法になることに注意) スコアは式(1)等を利用して求める。

【0260】

前記出力された上記評価データの各記事ごとにそれが正解しているかを調べて、各kpの時の正解率を求める。更に同じkpに対応する全ての上記評価データ(質問)の記事の正解率の平均値を求める。そうすると、kpと正解率の対応表が完成する。

【0261】

新しい情報検索の質問が入ってくると、前記情報検索システムで記事を出力する。各記事がぎりぎり出力されるkpを求める。この求め方は、上記対応表作成の場合と同様であり、ある記事のスコアを最初の記事(最もScoreの大きい記事)のスコアで割った値がその記事がぎりぎり出力されるkpとなる。(kpの定義によりこうなる、式(2)を参照のこと、順位による方法や他の方法ではこの部分は異なった方法になることに注意) スコアは式(1)等を利用して求める。

10

【0262】

各記事のkpが求めれば、先の対応表に基づいて、各記事に対応する正解率をくっつけて表示する。(この正解率は、個々のF-termの正解率であることに注意。そのF-termまでのF-term群に対する精度(適合率)、再現率、F値などとは異なるものである。)

図26は対応表作成処理フローチャートである。以下、図26の処理S201~S205にしたがって説明する(確信度付与装置は図7参照、但し、図7の文書分類装置の代わりに情報検索システム(装置)を用いる)。

20

【0263】

S201:入力部1より、予め問題と解答の組(ここでは情報検索の質問とその質問に対応する記事群)を大量に入力し、情報検索システムの格納手段に格納する。

【0264】

S202:情報検索システムは、前記入力されたある1つの質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、前記入力された記事群の情報検索を行って記事を取り出す。

【0265】

S203:情報検索システムは、Score_{max} * kpの文書(記事)まで出力する。

【0266】

S204:対応表作成部11は、各記事がぎりぎり出力されるkpを求め、各記事ごとにそれが正解しているかを調べて、各kpのときの正解率を求める。

30

【0267】

S205:対応表作成部11は、前記S201で入力した質問全てについて、情報検索システムで記事を検索し、各記事がぎりぎり出力されるkpを求め、更に該同じkpに対応する全ての特許文書の正解率の平均値を求め、対応表を作成する(対応表は格納手段13に格納される)。

【0268】

図27は確信度付与処理フローチャートである。以下、図27の処理S211~S214にしたがって説明する。

40

【0269】

S211:入力部1より、新たな情報検索の質問を入力する。

【0270】

S212:情報検索システムは、前記入力された質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、情報検索を行って記事を取り出す。

【0271】

S213:確信度付与部12は、各記事がぎりぎり出力されるkpを求める。

【0272】

S214:確信度付与部12は、格納部13の対応表から前記求めたkpに対応する確信

50

度である正解率を各記事に付与して出力部より出力する。

【0273】

b) 出力順位を利用する場合の説明

出力順位を利用する方法の場合は、情報検索システムを用いる。このシステムで評価データの問題を解き、kj個目の出力の記事があっているかまちがっているかを調べて、kj個目の出力の正解率を求める。そうすると、kjと正解率の対応表が完成する。

【0274】

新しい特許が入ってくると、先の方法(特許情報検索システム)で記事を出力する。そして、各記事がぎりぎり出力されるkjを求める。そのF-termが出力される順位がkjとなる。(kjの定義によりこうなる。他の方法ではこの部分は異なった方法になる)。

10

【0275】

各Ftermのkjが求めれば、先の対応表に基づいて、各Ftermに対応する正解率をくっつけて表示する。

【0276】

図28は対応表作成処理フローチャートである。以下、図28の処理S221~S225にしたがって説明する(確信度付与装置は図7参照、但し、図7の文書分類装置の代わりに情報検索システム(装置)を用いる)。

【0277】

S221: 入力部1より、予め問題と解答の組(ここでは情報検索の質問とその質問に対応する記事群)を大量に入力し、情報検索システムの格納手段に格納する。

20

【0278】

S222: 情報検索システムは、前記入力されたある1つの質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、前記入力された記事群の情報検索を行って記事を取り出す。

【0279】

S223: 情報検索システムは、前記取り出した記事の順位kjをまで出力する。

【0280】

S224: 対応表作成部11は、kj個目の記事の出力があっているか間違っているかを調べて、kj個目の出力の正解率を求める。

【0281】

S225: 対応表作成部11は、前記S221で入力した質問全てについて、kj個目の出力の正解率を求め、更に同じkjに対応する全ての記事の正解率の平均値を求め、対応表を作成する(対応表は格納手段13に格納される)。

30

【0282】

図29は確信度付与処理フローチャートである。以下、図29の処理S231~S234にしたがって説明する(確信度付与装置は図7参照、但し、図7の文書分類装置の代わりに情報検索システム(装置)を用いる)。

【0283】

S231: 入力部1より、新たな情報検索の質問を入力する。

【0284】

S232: 情報検索システムは、前記入力された質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、情報検索を行って記事を取り出す。

40

【0285】

S233: 確信度付与部12は、各記事がぎりぎり出力されるkjを求める。

【0286】

S234: 確信度付与部12は、格納部13の対応表から前記求めたkjに対応する確信度である正解率を各記事に付与して出力部より出力する。

【0287】

c) スコア(Score)を利用する場合の説明

50

スコアを利用する方法の場合は、前記情報検索システムを使用する。このシステムで評価データの問題を解き、出力される各記事を評価する。記事のスコアが k_I のものについて、その記事が持っているかどうかを調べて、 k_I の場合の正解率を求める。これをあらゆる k_I について求める。そうすると、 k_I と正解率の対応表が完成する。

【0288】

新しい情報検索の質問が入ってくると、情報検索システムで記事を出力する。ここで各記事がぎりぎり出力される k_I を求める。すると各記事のスコアが k_I となる。(k_I の定義によりこうなる。他の方法ではこの部分は異なった方法になる)

各記事の k_I が求めれば、先の対応表に基づいて、各記事に対応する正解率をくっつけて表示する。

10

【0289】

図30は対応表作成処理フローチャートである。以下、図30の処理S241~S245にしたがって説明する(確信度付与装置は図7参照、但し、図7の文書分類装置の代わりに情報検索システム(装置)を用いる)。

【0290】

S241: 入力部1より、予め問題と解答の組(ここでは情報検索の質問とその質問に対応する記事群)を大量に入力し、情報検索システムの格納手段に格納する。

【0291】

S242: 情報検索システムは、前記入力されたある1つの質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、前記入力された記事群の情報検索を行って記事を取り出す。

20

【0292】

S243: 情報検索システムは、スコアが k_I 以上の記事を出力する。

【0293】

S244: 対応表作成部11は、記事のスコアが k_I のものについて、その記事が持っているかどうかを調べて、 k_I の場合の正解率を求める。

【0294】

S245: 対応表作成部11は、これを前記S241で入力した質問全てについて、記事を出力し、同じ k_I に対応する正解率の平均値を求め、 k_I と正解率の対応表が完成する(対応表は格納手段13に格納される)。

30

【0295】

図31は確信度付与処理フローチャートである。以下、図31の処理S251~S254にしたがって説明する(確信度付与装置は図7参照、但し、図7の文書分類装置の代わりに情報検索システム(装置)を用いる)。

【0296】

S251: 入力部1より、新たな情報検索の質問を入力する。

【0297】

S252: 情報検索システムは、前記入力された質問から、形態素解析して、名詞をキーワードとして取り出して、そのキーワードを利用して、情報検索を行って記事を取り出す。

40

【0298】

S253: 確信度付与部12は、各記事がぎりぎり出力される k_I を求める。

【0299】

S254: 確信度付与部12は、格納部13の対応表から前記求めた k_I に対応する確信度である正解率を各記事に付与して出力部より出力する。

【0300】

以上 k_p 、順位、スコアを利用する方法を示したが、順序化して出力する他のものを利用することができる。

【0301】

(3): データの補間、補正の説明

50

表（対応表）に基づく方法で、例えば、kpと正解率の対応表が作成できたとする。このあと、新しい入力でkpがkp1の場合の正解率が表から必要になったが、kp1の値が表にのっていないとする。そうすると、ある種の補間処理が必要になる。その場合は、表にのっている、kp1と最も近い値のkpの部分でkp1の代りにつかってもいいし、表にのっている、kp1をはさむ二つのkpの2行のデータを用い、その2行のデータの正解率の平均をkp1の正解率としてもよい。

【0302】

また、表にのっている、kp1をはさむ二つのkpの2行のデータkp2、kp3 (kp2>kp>kp3)を用い、その2行のデータの正解率pr2、pr3を利用して

$$\{(kp - kp3) pr2 + (kp2 - kp) pr3\} / \{(kp2 - kp) + (kp - kp3)\}$$

10

を正解率としてもよい。その他の補完処理によりkpに対応する正解率を求めてもよい。

【0303】

また、kpと正解率の対に対して、単回帰式近似、又は、多項式近似、又は、対数近似、又は、指数近似などをして求めた近似式によりkpに対応する正解率を求めるようにしてもよい（例えば、「Excelで学ぶ時系列分析と予測」（オーム社）2章の“単回帰分析”3章の“重回帰分析”参照）。また、上記回帰分析的な近似以外の補正処理を行ってもよい。なお、データの補間、補正は、kl等の他のデータについても同様である。

【0304】

§4：機械学習を用いる場合の説明

a) 機械学習法の詳細な説明

20

図32は機械学習法の説明図である。図32において、機械学習法には、教師データ記憶手段21、解-素性対抽出手段22、機械学習手段23、学習結果記憶手段24、表現対抽出手段25、素性抽出手段26、解推定手段27、出力手段28を備える。

【0305】

ここで、機械学習手段23による機械学習の手法について説明する。機械学習の手法は、問題-解の組のセットを多く用意し、それで学習を行ない、どういう問題のときにどういう解になるかを学習し、その学習結果を利用して、新しい問題のときも解を推測できるようにする方法である（例えば、下記の参考文献(1)～参考文献(3)参照）。

【0306】

参考文献(1)：村田真樹，機械学習に基づく言語処理，龍谷大学理工学部，招待講演，2004．<http://www2.nict.go.jp/jt/a132/members/murata/ps/rk1-siryoku.pdf>

30

参考文献(2)：サポートベクトルマシンを用いたテンス・アスペクト・モダリティの日英翻訳，村田真樹，馬青，内元清貴，井佐原均，電子情報通信学会言語理解とコミュニケーション研究会 NLC2000-78，2001年．

参考文献(3)：SENSEVAL2J辞書タスクでのCRLの取り組み，村田真樹，内山将夫，内元清貴，馬青，井佐原均，電子情報通信学会言語理解とコミュニケーション研究会 NLC2001-40，2001年．

どういう問題のときに、という、問題の状況を機械に伝える際に、素性（解析に用いる情報で問題を構成する各要素）というものが必要になる。問題を素性によって表現するのである。例えば、日本語文末表現の時制の推定の問題において、問題：「彼が話す。」 - - 解「現在」が与えられた場合に、素性の一例は、「彼が話す。」「が話す。」「話す。」「す」「。」となる。

40

【0307】

すなわち、機械学習の手法は、素性の集合-解の組のセットを多く用意し、それで学習を行ない、どういう素性の集合のときにどういう解になるかを学習し、その学習結果を利用して、新しい問題のときもその問題から素性の集合を取り出し、その素性の場合の解を推測する方法である。

【0308】

機械学習手段23は、機械学習の手法として、例えば、k近傍法、シンプルベイズ法、決定リスト法、最大エントロピー法、サポートベクトルマシン法などの手法を用いる。

50

【0309】

k近傍法は、最も類似する一つの事例のかわりに、最も類似するk個の事例を用いて、このk個の事例での多数決によって分類先(解)を求める手法である。kは、あらかじめ定める整数の数字であって、一般的に、1から9の間の奇数を用いる。

【0310】

シンプルベイズ法は、ベイズの定理にもとづいて各分類になる確率を推定し、その確率値が最も大きい分類を求める分類先とする方法である。

【0311】

シンプルベイズ法において、文脈bで分類aを出力する確率は、以下の式(11)で与えられる。

10

【0312】

【数7】

$$p(a|b) = \frac{p(a)}{p(b)} p(b|a) \quad (11)$$

$$\cong \frac{\tilde{p}(a)}{p(b)} \prod_i \tilde{p}(f_i|a) \quad (12)$$

20

【0313】

ただし、ここで文脈bは、あらかじめ設定しておいた素性 f_j ($F, 1 \leq j \leq k$)の集合である。 $p(b)$ は、文脈bの出現確率である。ここで、分類aに非依存であって定数のために計算しない。 $P(a)$ (ここでPはpの上部にチルダ)と $P(f_i|a)$ は、それぞれ教師データから推定された確率であって、分類aの出現確率、分類aのときに素性 f_i を持つ確率を意味する。 $P(f_i|a)$ として最尤推定を行って求めた値を用いると、しばしば値がゼロとなり、式(12)の値がゼロで分類先を決定することが困難な場合が生じる。そのため、スムージングを行う。ここでは、以下の式(13)を用いてスムージングを行ったものを用いる。

30

【0314】

【数8】

$$p(f_i|a) = \frac{\text{freq}(f_i, a) + 0.01 * \text{freq}(a)}{\text{freq}(a) + 0.01 * \text{freq}(a)} \quad (13)$$

【0315】

ただし、 $\text{freq}(f_i, a)$ は、素性 f_i を持ちかつ分類がaである事例の個数、 $\text{freq}(a)$ は、分類がaである事例の個数を意味する。

40

【0316】

決定リスト法は、素性と分類先の組とを規則とし、それらをあらかじめ定めた優先順序でリストに蓄えおき、検出する対象となる入力を与えられたときに、リストで優先順位の高いところから入力のデータと規則の素性とを比較し、素性が一致した規則の分類先をその入力の分類先とする方法である。

【0317】

決定リスト方法では、あらかじめ設定しておいた素性 f_j ($F, 1 \leq j \leq k$)のうち、いずれか一つの素性のみを文脈として各分類の確率値を求める。ある文脈bで分類aを

50

出力する確率は以下の式によって与えられる。

【0318】

$$p(a|b) = p(a|f_{\max}) \quad (14)$$

ただし、 f_{\max} は以下の式によって与えられる。

【0319】

【数9】

$$f_{\max} = \arg \max_{f_j \in F} \max_{a_i \in A} \tilde{p}(a_i|f_j) \quad (15)$$

10

【0320】

また、 $P(a_i|f_j)$ (ここで P は p の上部にチルダ) は、素性 f_j を文脈に持つ場合の分類 a_i の出現の割合である。

【0321】

最大エントロピー法は、あらかじめ設定しておいた素性 f_j ($1 \leq j \leq k$) の集合を F とするとき、以下所定の条件式 (式 (16)) を満足しながらエントロピーを意味する式 (17) を最大にするときの確率分布 $p(a, b)$ を求め、その確率分布にしたがって求める各分類の確率のうち、最も大きい確率値を持つ分類を求める分類先とする方法である。

20

【0322】

【数10】

$$\sum_{a \in A, b \in B} p(a, b) g_j(a, b) = \sum_{a \in A, b \in B} \tilde{p}(a, b) g_j(a, b) \quad (16)$$

$$\text{for } \forall f_j (1 \leq j \leq k)$$

30

$$H(p) = - \sum_{a \in A, b \in B} p(a, b) \log(p(a, b)) \quad (17)$$

【0323】

ただし、 A 、 B は分類と文脈の集合を意味し、 $g_j(a, b)$ は文脈 b に素性 f_j があって、なおかつ分類が a の場合 1 となり、それ以外で 0 となる関数を意味する。また、 $P(a_i|f_j)$ (ここで P は p の上部にチルダ) は、既知データでの (a, b) の出現の割合を意味する。

40

【0324】

式 (16) は、確率 p と出力と素性の組の出現を意味する関数 g をかけることで出力と素性の組の頻度の期待値を求めることになっており、右辺の既知データにおける期待値と、左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として、エントロピー最大化 (確率分布の平滑化) を行なって、出力と文脈の確率分布を求めるものとなっている。最大エントロピー法の詳細については、以下の参考文献 (4) および参考文献 (5) に記載されている。

【0325】

50

参考文献 (4) : Eric Sven Ristad, Maximum Entropy Modeling for Natural Language, (ACL/EACL Tutorial Program, Madrid, 1997)

参考文献 (5) : Eric Sven Ristad, Maximum Entropy Modeling Toolkit, Release 1.6beta, (<http://www.mnemonic.com/software/memt>, 1998))

サポートベクトルマシン法は、空間を超平面で分割することにより、二つの分類からなるデータを分類する手法である。

【 0 3 2 6 】

図 3 3 はサポートベクトルマシン法のマージン最大化の概念図である。図 3 3 において、白丸は正例、黒丸は負例を意味し、実線は空間を分割する超平面を意味し、破線はマージン領域の境界を表す面を意味する。図 3 3 (A) は、正例と負例の間隔が狭い場合 (スモールマージン) の概念図、図 3 3 (B) は、正例と負例の間隔が広い場合 (ラージマージン) の概念図である。

10

【 0 3 2 7 】

このとき、二つの分類が正例と負例からなるものとする、学習データにおける正例と負例の間隔 (マージン) が大きいものほどオープンデータで誤った分類をする可能性が低いと考えられ、図 3 3 (B) に示すように、このマージンを最大にする超平面を求めそれを用いて分類を行なう。

【 0 3 2 8 】

基本的には上記のとおりであるが、通常、学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や、超平面の線形の部分を非線型にする拡張 (カーネル関数の導入) がなされたものが用いられる。

20

【 0 3 2 9 】

この拡張された方法は、以下の識別関数を用いて分類することと等価であり、その識別関数の出力値が正か負かによって二つの分類を判別することができる。

【 0 3 3 0 】

【 数 1 1 】

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^l \alpha_i y_i K(x_i, x) + b\right) \quad (18)$$

30

$$b = -\frac{\max_{i, y_i=-1} b_i + \min_{i, y_i=1} b_i}{2}$$

$$b_i = \sum_{j=1}^l \alpha_j y_j K(x_j, x_i)$$

40

【 0 3 3 1 】

ただし、 x は識別したい事例の文脈 (素性の集合) を、 x_i と y_j ($i = 1, \dots, l$, $y_j \in \{1, -1\}$) は学習データの文脈と分類先を意味し、関数 sgn は、

$$\operatorname{sgn}(x) = \begin{cases} 1 & (x > 0) \\ -1 & (\text{otherwise}) \end{cases}$$

であり、また、各 α_i は式 (2 0) と式 (2 1) の制約のもと式 (1 9) を最大にする場合のものである。

【 0 3 3 2 】

【数 1 2】

$$L(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (19)$$

$$0 \leq \alpha_i \leq C \quad (i=1, \dots, l) \quad (20)$$

$$\sum_{j=1}^l \alpha_j y_j = 0 \quad (21)$$

10

【0 3 3 3】

また、関数 K はカーネル関数と呼ばれ、様々なものが用いられるが、本形態では以下の多項式のものを用いる。

【0 3 3 4】

$$K(x, y) = (x \cdot y + 1)^d \quad (22)$$

C 、 d は実験的に設定される定数である。例えば、 C はすべての処理を通して 1 に固定した。また、 d は、1 と 2 の二種類を試している。ここで、 $\alpha_i > 0$ となる x_i は、サポートベクトルと呼ばれ、通常、式 (18) の和をとっている部分は、この事例のみを用いて計算される。つまり、実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない。

20

【0 3 3 5】

なお、拡張されたサポートベクトルマシン法の詳細については、以下の参考文献 (6) および参考文献 (7) に記載されている。

【0 3 3 6】

参考文献 (6) : Nello Cristianini and John Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods, (Cambridge University Press, 2000)

30

参考文献 (7) : Taku Kudoh, Tinysvm: Support Vector machines, (<http://cl.aist-nara.ac.jp/taku-ku//software/TinySVM/index.html>, 2000)

サポートベクトルマシン法は、分類の数が 2 個のデータを扱うものである。したがって、分類の数が 3 個以上の事例を扱う場合には、通常、これにペアワイズ法またはワン VS レスト法などの手法を組み合わせて用いることになる。

【0 3 3 7】

ペアワイズ法は、 n 個の分類を持つデータの場合に、異なる二つの分類先のあらゆるペア ($n(n-1)/2$ 個) を生成し、各ペアごとにどちらがよいかを二値分類器、すなわちサポートベクトルマシン法処理モジュールで求めて、最終的に、 $n(n-1)/2$ 個の二値分類による分類先の多数決によって、分類先を求める方法である。

40

【0 3 3 8】

ワン VS レスト法は、例えば、 a 、 b 、 c という三つの分類先があるときは、分類先 a とその他、分類先 b とその他、分類先 c とその他、という三つの組を生成し、それぞれの組についてサポートベクトルマシン法で学習処理する。そして、学習結果による推定処理において、その三つの組のサポートベクトルマシンの学習結果を利用する。推定すべき候補が、その三つのサポートベクトルマシンではどのように推定されるかを見て、その三つのサポートベクトルマシンのうち、その他でないほうの分類先であって、かつサポートベクトルマシンの分離平面から最も離れた場合のものの分類先を求める解とする方法である。例えば、ある候補が、「分類先 a とその他」の組の学習処理で作成したサポートベク

50

トルマシンにおいて分離平面から最も離れた場合には、その候補の分類先は、a と推定する。

【0339】

解推定手段27が推定する、各表現対についての、どのような解（分類先）になりやすいかの度合いの求め方は、機械学習手段23が機械学習の手法として用いる様々な方法によって異なる。

【0340】

例えば、本発明の実施の形態において、機械学習手段23が、機械学習の手法としてk近傍法を用いる場合、機械学習手段23は、教師データの事例同士で、その事例から抽出された素性の集合のうち重複する素性の割合（同じ素性をいくつ持っているかの割合）にもとづく事例同士の類似度を定義して、前記定義した類似度と事例とを学習結果情報として学習結果記憶手段24に記憶しておく。

10

【0341】

そして、解推定手段27は、表現対抽出手段25によって新しい表現対（の候補）が抽出されたときに、学習結果記憶手段24において定義された類似度と事例を参照して、表現対抽出手段25によって抽出された表現対の候補について、その候補の類似度が高い順にk個の事例を学習結果記憶手段24の事例から選択し、選択したk個の事例での多数決によって決まった分類先を、表現対の候補の分類先（解）として推定する。すなわち、解推定手段27では、各表現対についての、どのような解（分類先）になりやすいかの度合いを、選択したk個の事例での多数決の票数、ここでは「抽出すべき」という分類が獲得した票数とする。

20

【0342】

また、機械学習手法として、シンプルベイズ法を用いる場合には、機械学習手段23は、教師データの事例について、前記事例の解と素性の集合との組を学習結果情報として学習結果記憶手段24に記憶する。そして、解推定手段27は、表現対抽出手段25によって新しい表現対（の候補）が抽出されたときに、学習結果記憶手段24の学習結果情報の解と素性の集合との組をもとに、ベイズの定理にもとづいて素性抽出手段26で取得した表現対の候補の素性の集合の場合の各分類になる確率を算出して、その確率の値が最も大きい分類を、その表現対の候補の素性の分類（解）と推定する。すなわち、解推定手段27では、表現対の候補の素性の集合の場合にある解となりやすさの度合いを、各分類になる確率、ここでは「抽出すべき」という分類になる確率とする。

30

【0343】

また、機械学習手法として決定リスト法を用いる場合には、機械学習手段23は、教師データの事例について、素性と分類先との規則を所定の優先順序で並べたリストを学習結果記憶手段24に記憶する。そして、表現対抽出手段15によって新しい表現対（の候補）が抽出されたときに、解推定手段27は、学習結果記憶手段24のリストの優先順位の高い順に、抽出された表現対の候補の素性と規則の素性とを比較し、素性が一致した規則の分類先をその候補の分類先（解）として推定する。すなわち、解推定手段27では、表現対の候補の素性の集合の場合にある解となりやすさの度合いを、所定の優先順位またはそれに相当する数値、尺度、ここでは「抽出すべき」という分類になる確率のリストにおける優先順位とする。

40

【0344】

また、機械学習手法として最大エントロピー法を使用する場合には、機械学習手段23は、教師データの事例から解となりうる分類を特定し、所定の条件式を満足しかつエントロピーを示す式を最大にするときの素性の集合と解となりうる分類の二項からなる確率分布を求めて学習結果記憶手段24に記憶する。そして、表現対抽出手段25によって新しい表現対（の候補）が抽出されたときに、解推定手段27は、学習結果記憶手段24の確率分布を利用して、抽出された表現対の候補の素性の集合についてその解となりうる分類の確率を求めて、最も大きい確率値を持つ解となりうる分類を特定し、その特定した分類をその候補の解と推定する。すなわち、解推定手段27では、表現対の候補の素性の集合

50

の場合にある解となりやすさの度合いを、各分類になる確率、ここでは「抽出すべき」という分類になる確率とする。

【0345】

また、機械学習手法としてサポートベクトルマシン法を使用する場合には、機械学習手段23は、教師データの事例から解となりうる分類を特定し、分類を正例と負例に分割して、カーネル関数を用いた所定の実行関数にしたがって事例の素性の集合を次元とする空間上で、その事例の正例と負例の間隔を最大にし、かつ正例と負例を超平面で分割する超平面を求めて学習結果記憶手段24に記憶する。そして表現対抽出手段25によって新しい表現対(の候補)が抽出されたときに、解推定手段27は、学習結果記憶手段24の超平面を利用して、抽出された表現対の候補の素性の集合が超平面で分割された空間において正例側か負例側のどちらにあるかを特定し、その特定された結果にもとづいて定まる分類を、その候補の解と推定する。すなわち、解推定手段27では、表現対の候補の素性の集合の場合にある解となりやすさの度合いを、分離平面からの正例(抽出すべき表現対)の空間への距離の大きさとする。より詳しくは、抽出すべき表現対を正例、抽出すべきではない表現対を負例とする場合に、分離平面に対して正例側の空間に位置する事例が「抽出すべき事例」と判断され、その事例の分離平面からの距離をその事例の度合いとする。

10

【0346】

b) 機械学習を用いる場合の説明(文書分類装置を使用する場合)

確信度付与装置で機械学習の方法の場合は、予め問題と解答の組を大量に集める。問題は、F-termをふるべき特許、解答は、その特許のF-termとする。これを評価データと呼ぶ。

20

【0347】

ここで前記文書分類装置を用いて上記評価データでF-termを出力する。そして、各F-termがぎりぎり出力されるkpを求める。この求め方は、以下のようにする。

【0348】

あるF-termのスコア(Score)を最初のF-term(最もScoreの大きいF-term)のスコアで割った値がそのF-termがぎりぎり出力されるkpとなる。また、そのF-termの順位kjも求める。また、そのF-termのスコア=klも求める。スコアは前記式(1)等で求める。

【0349】

各F-termごとにそれが正解しているかどうかを調べる。正解していれば、kp、kj、klのときに正解とし、正解していなければkp、kj、klのときに不正解という事例になる。

30

【0350】

出力した各F-termについて上記事例を作成する。次に、機械学習(機械学習手段23)を利用する。kp、kj、klのときに正解、kp、kj、klのときに不正解、といった事例を学習データ(解-素性対抽出手段22)として、機械学習を行う。ここで、kp、kj、klがそれぞれ素性となる。正解、不正解は求める分類先となる。

【0351】

機械学習により、どういうkp、kj、klなら、正解に、どういうkp、kj、klなら、不正解になりやすいかを学習し、それを学習結果(学習結果記憶手段24)に蓄える。

40

【0352】

ここで、新しい特許(F-termが付与されていない)が入ってくる。前記文書分類装置を用いて、F-termを出力する。そして、各F-termがぎりぎり出力されるkpを求める。この求め方は、以下の通りである。

【0353】

あるF-termのスコアを最初のF-term(最もスコアの大きいF-term)のスコアで割った値がそのF-termがぎりぎり出力されるkpとなる。また、そのF-termの順位kjも求める。また、そのF-termのスコア=klも求める。スコアは前記式(1)等で求める。

【0354】

先の学習結果により、このときのkp、kj、klの場合に正解になりやすい確信度を求める

50

(解推定手段27)。ここでは、確信度も出力できる機械学習(機械学習手段)を用いる。

【0355】

この確信度を各F-termに対応する正解率としてくっつけて表示する。(この正解率は、個々のF-termの正解率であることに注意。そのF-termまでのF-term群に対する精度(適合率)、再現率、F値などとは異なるものである)。

【0356】

ここで、機械学習の素性を k_p 、 k_j 、 k_l としたが、これの一部のみを素性としてもよいし、逆に他のものもこの素性に加えてもよいし、これらの一部と他のものの組み合わせを素性としてもよい。

10

【0357】

例えば、特許文書群に含まれる単語や文字列を利用して、その単語が該当特許に含まれるか否かという素性や、その文字列が該当特許に含まれるか否かという素性を利用してもよい。

【0358】

c) 機械学習を用いる場合の説明(情報検索システムを使用する場合)

確信度付与装置で機械学習の方法の場合は、予め問題と解答の組を大量に集める。問題は、情報検索の質問、解答はその質問に対応する記事群である。これを評価データと呼ぶ。

【0359】

ここで前記情報検索システムを用いて上記評価データで記事を出力する。そして、各記事がぎりぎり出力される k_p を求める。この求め方は、以下のようにする。

20

【0360】

ある記事のスコア(Score)を最初の記事(最もScoreの大きい記事)のスコアで割った値がその記事がぎりぎり出力される k_p となる。また、その記事の順位 k_j も求める。また、その記事のスコア = k_l も求める。スコアは前記式(1)等で求める。

【0361】

各記事ごとにそれが正解しているかどうかを調べる。正解していれば、 k_p 、 k_j 、 k_l のときに正解とし、正解していなければ k_p 、 k_j 、 k_l のときに不正解という事例になる。

【0362】

出力した各記事について上記事例を作成する。次に、機械学習(機械学習手段23)を利用する。 k_p 、 k_j 、 k_l のときに正解、 k_p 、 k_j 、 k_l のときに不正解、といった事例を学習データ(解-素性対抽出手段22)として、機械学習を行う。ここで、 k_p 、 k_j 、 k_l がそれぞれ素性となる。正解、不正解は求める分類先となる。

30

【0363】

機械学習により、どういう k_p 、 k_j 、 k_l なら、正解に、どういう k_p 、 k_j 、 k_l なら、不正解になりやすいかを学習し、それを学習結果(学習結果記憶手段24)に蓄える。

【0364】

ここで、新しい情報検索に質問が入ってくる。前記情報検索システムを用いて、記事を出力する。そして、各記事がぎりぎり出力される k_p を求める。この求め方は、以下の通りである。

40

【0365】

ある記事のスコアを最初の記事(最もスコアの大きい記事)のスコアで割った値がその記事がぎりぎり出力される k_p となる。また、その記事の順位 k_j も求める。また、その記事のスコア = k_l も求める。スコアは前記式(1)等で求める。

【0366】

先の学習結果により、このときの k_p 、 k_j 、 k_l の場合に正解になりやすい確信度を求める(解推定手段27)。ここでは、確信度も出力できる機械学習(機械学習手段)を用いる。

【0367】

50

この確信度を各記事に対応する正解率としてくっつけて表示する。(この正解率は、個々の記事の正解率であることに注意。その記事までの記事群に対する精度(適合率)、再現率、F値などとは異なるものである)。

【0368】

ここで、機械学習の素性を k_p 、 k_j 、 k_l としたが、これの一部のみを素性としてもよいし、逆に他のものもこの素性に加えてもよいし、これらの一部と他のものの組み合わせを素性としてもよい。

【0369】

以上、分類を付与する場合と情報検索の場合に機械学習により確信度(正解率)を出力する説明をしたが、この機械学習法としては、ニューラルネットワークや重回帰分析を用いてもよい。重回帰分析の説明は、「Excelで学ぶ時系列分析と予測」(オーム社)3章の“重回帰分析”で求めてもよい。重回帰分析の場合は、「正解」を値1「不正解」を値0として求めればよい。

10

【0370】

すなわち、求める分類が2種類ならば、重回帰分析が利用できる。重回帰分析の場合は、素性の数だけ説明変数 x を用意し、素性のありなしを、その説明変数 x の値を1、0で表現する。目的変数(被説明変数)は、ある分類の場合を値1、他の分類の場合を値0として求めればよい。

【0371】

(重回帰分析の利用の説明)

20

重回帰分析では、 x_1, x_2, x_3, \dots, y の組のデータがあるときに、 x_1, x_2, x_3, \dots から y を求める。

【0372】

$$y = a_0 + a_1 * x_1 + a_2 * x_2 + a_3 * x_3 + \dots$$

の式の係数 a_0, a_1, \dots を、データから適切にもとめることができる。

【0373】

(1) $k_p - y$ の組からの予測

y は確信度

$x_1 = k_p$ として、

$y = a_0 + a_1 * k_p$ として、

30

重回帰分析により $k_p - y$ の組のデータから a_0, a_1 を求める。 k_p から y を求める式が求まる。

【0374】

(2) $k_p - y$ の組からの予測(2次の利用)

y は確信度

$x_1 = k_p, x_2 = k_p^2$ として、

$y = a_0 + a_1 * k_p + a_2 * k_p^2$ として、

重回帰分析により $k_p - y$ の組のデータから a_0, a_1, a_2 を求める。 k_p から y を求める式が求まる。

40

【0375】

(3) $k_p, k_j - y$ の組からの予測

y は確信度

$x_1 = k_p, x_2 = k_j$ として、

$y = a_0 + a_1 * k_p + a_2 * k_j$ として、

重回帰分析により $k_p, k_j - y$ の組のデータから a_0, a_1, a_2 を求める。 k_p, k_j から y を求める式が求まる。

【0376】

(4) $k_p, k_j - y$ の組からの予測(2次の利用)

y は確信度

$x_1 = k_p, x_2 = k_j$ として、

50

$y = a_0 + a_1 * kp + a_2 * kj + a_3 * kp^2 + a_4 * kp * kj + a_5 * kj^2$ として、重回帰分析により kp 、 $kj - y$ の組のデータから $a_0, a_1, a_2, \dots, a_5$ を求める。 kp 、 kj から y を求める式が求まる。

【0377】

これらの処理は、重回帰分析を機械学習手法として用いている方法ともとらえられるし、また、重回帰分析を補間手法として利用しているともとらえられる。

【0378】

(機械学習、重回帰分析の利用の説明)

機械学習、重回帰分析を利用するときには、 $kp - y$ の組のデータや、 $kp, kj - y$ の組のデータなどを利用する。

10

【0379】

このとき、同じ kp について、 y の平均を求めて、各 kp ごとに y の値が一つだけあるデータを作って、それを $kp - y$ の組のデータとしてもよい。この場合、 $kp - y$ の平均の組のデータになっている。

【0380】

また、これとは別の方法として、このとき、同じ kp について、データをまとめることをせずに、元のすべてのデータ自体を使って、それを $kp - y$ の組のデータとしてもよい。すなわち、確信度の平均をとるという操作をせずに、元の、問題の個数分だけ、 $kp - y$ の組のデータの個数があるようにしてもよい。

20

【0381】

(確信度についての説明)

前記の説明において使用する確信度としては、適合率の偏差値、再現率の偏差値、F値の偏差値、正解率の偏差値を用いてもよい。また、これらに類するものでもよい。数値的に求められるものなら、これら以外のものでよい。

【0382】

なお、値が大きいものを取り出すなどについては、「値が閾値以上のものを取り出す」「値が大きいものを所定の値の個数以上のものを大きい順に取り出す」「取り出されたものの値の最大値に対して所定の割合をかけた値を求め、その求めた値以上の値を持つものを取り出す」等の表現とすることができる。また、これら閾値、所定の値を、あらかじめ定めることも、適宜ユーザが値を変更、設定できることも可能である。

30

【0383】

また、入力された問題を解いてその解答を複数順序化して抽出し、該抽出した解答と所定値を出力するとき、この所定値は、前に説明した kp, kj, kl のように解答の順序化と同じ順(又は逆の順)となる(複数観点の所定値を用いる場合は除く)。

【0384】

§5: 実験結果の説明

次に実際に実験を行なった結果の説明をする。NTCIR-5 Patent 分類タスクのデータを使用した。ここで分類対象の特許文書は 1201 件あった。そして、この特許文書を次のように分割した。

【0385】

close ... 600

open ... 601

ここで close のデータを使って、対応表を求めて、確信度を予測する。そして、open のデータを使って、予測した確信度の妥当性を確認する。実験結果の表を図 3 4 で示している。

40

【0386】

図 3 4 は実験結果の説明図であり、図 3 4 において、表の値は、真の値と、本発明により予測した値の差の絶対値(絶対誤差)を示している。図 3 4 では、4 つの方法を試してある。この確信度(再現率、適合率、F 値)は、 kj 個目までの F タームを出力させた場合の確信度(再現率、適合率、F 値)である。すなわち、 kj 個目の F タームの確信度でな

50

く、kj個目までのFタームの確信度となっている。

【0387】

(1) base0.5 --- すべて確信度を0.5 とする方法。

【0388】

(2) kp --- kp と確信度の対応表を求めて予測する方法（ここでは、kp = 0, 0.1, 0.2, ..., 1.0の値の場合の対応表を求めた）。

【0389】

(3) kj --- kj と確信度の対応表を求めて予測する方法（ここでは、kj = 1, 2, 3, ..., 200の値の場合の対応表を求めた）。

【0390】

(4) kp, kj --- kp, kjと確信度の対応表を求めて予測する方法（ここでは、kp = 0, 0.1, 0.2, ..., 1.0の値とkj = 1, 2, 3, ..., 200の値のすべての組み合わせの場合の対応表を求めた）。

【0391】

kjについては補間処理は必要ない。kpについては補間処理を行った。この補間処理はすでに説明した次の式でおこなった。

【0392】

$$\left[(kp - kp3) pr2 + (kp2 - kp) pr3 \right] / \left[(kp2 - kp) + (kp - kp3) \right]$$

図34の表のkjは、システムの出力の何個目のFタームのときの結果を示すかをあらわしている。例えば、kj = 1だと、システムの出力の1個目のFタームのときの結果を示している。図34の表の値は、真の値と、本発明により予測した値の差の絶対値（絶対誤差）と書いたが、正確には、記事ごとに、kj個目のFタームのときの真の値（確信度）と、本発明により予測した値（確信度）の差の絶対値（絶対誤差）を求めて、それを加えて、記事の総数で割った。つまり、表の値は、絶対誤差の平均である。

【0393】

図34の表では、全般的にbase0.5に比べて他の方法の誤差はかなり小さい。このため、本発明の有効性がわかる。また、kp、kj単独のものを利用したのに比べてkp、kj両方を利用したものは、すこしではあるが誤差が小さくなっている。また、kpまたkjまたkp、kjのそれぞれの手法とも、kjが小さい、上位の出力において、適合率の誤差が0.2前後と少し大きいをそれを除くと、誤差は0.1前後であり、かなりよい予測が実現できていることがわかる。

【0394】

§6：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（スコア_{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、文書分類装置（文書分類手段）10、対応表作成部（対応関係作成手段）11、確信度付与部（確信度付与手段）12、格納部（対応表）13、教師データ記憶手段21、解-素性対抽出手段22、機械学習手段23、学習結果記憶手段24、表現対抽出手段25、素性抽出手段26、解推定手段27、出力手段28、問題解決手段、情報検索システム（装置）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。

【0395】

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

10

20

30

40

50

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】

【0396】

【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kpとF値の対応の説明図である。

10

【図6】本発明のkpと再現率と精度の対応の説明図である。

【図7】本発明の確信度付与装置の説明図である。

【図8】本発明の対応表作成処理フローチャートである。

【図9】本発明の確信度付与処理フローチャートである。

【図10】本発明の対応表作成処理フローチャートである。

【図11】本発明の確信度付与処理フローチャートである。

【図12】本発明の対応表作成処理フローチャートである。

【図13】本発明の確信度付与処理フローチャートである。

【図14】本発明の対応表作成処理フローチャートである。

【図15】本発明の確信度付与処理フローチャートである。

20

【図16】本発明の対応表作成処理フローチャートである。

【図17】本発明の確信度付与処理フローチャートである。

【図18】本発明の対応表作成処理フローチャートである。

【図19】本発明の確信度付与処理フローチャートである。

【図20】本発明の対応表作成処理フローチャートである。

【図21】本発明の確信度付与処理フローチャートである。

【図22】本発明の対応表作成処理フローチャートである。

【図23】本発明の確信度付与処理フローチャートである。

【図24】本発明の対応表作成処理フローチャートである。

【図25】本発明の確信度付与処理フローチャートである。

30

【図26】本発明の対応表作成処理フローチャートである。

【図27】本発明の確信度付与処理フローチャートである。

【図28】本発明の対応表作成処理フローチャートである。

【図29】本発明の確信度付与処理フローチャートである。

【図30】本発明の対応表作成処理フローチャートである。

【図31】本発明の確信度付与処理フローチャートである。

【図32】本発明の機械学習法の説明図である。

【図33】本発明のサポートベクトルマシン法のマージン最大化の概念図である。

【図34】本発明の実験結果の説明図である。

【符号の説明】

40

【0397】

1 入力部（入力手段）

6 出力部（出力手段）

10 文書分類装置（問題解決手段）

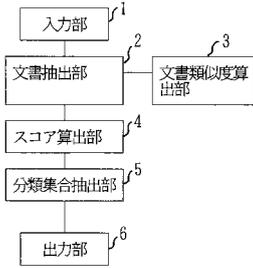
11 対応表作成部（対応関係作成手段）

12 確信度付与部（確信度付与手段）

13 格納部（対応表）

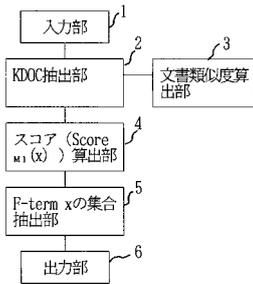
【 図 1 】

文書分類装置の説明図



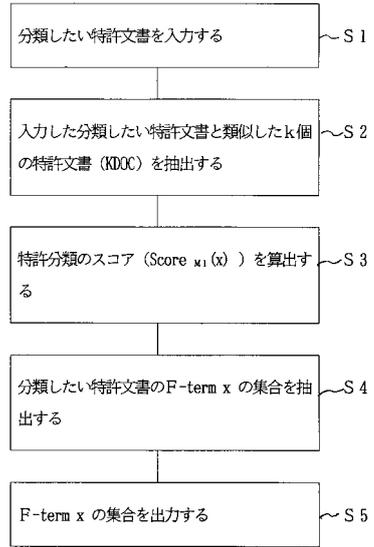
【 図 2 】

特許文書分類装置の説明図



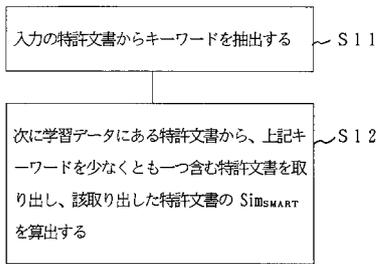
【 図 3 】

特許文書の分類処理フローチャート



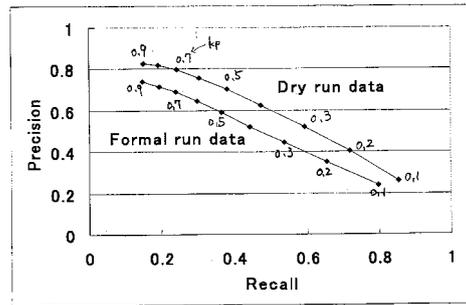
【 図 4 】

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



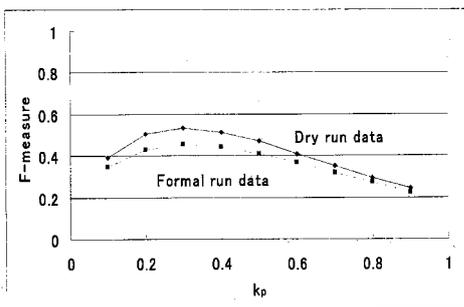
【 図 6 】

kpと再現率と精度の対応の説明図



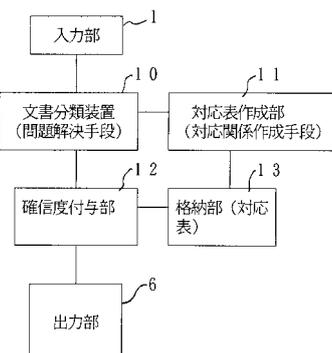
【 図 5 】

kpとF値の対応の説明図

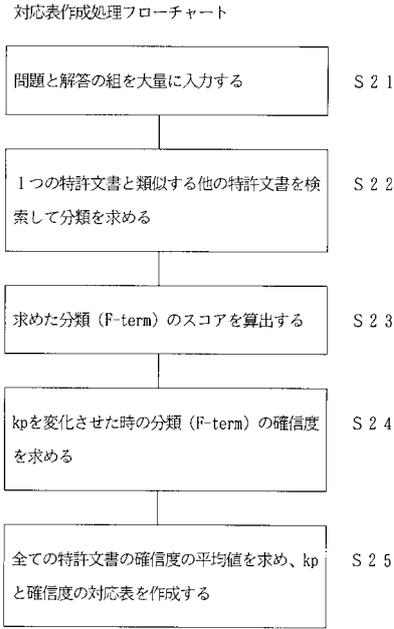


【 図 7 】

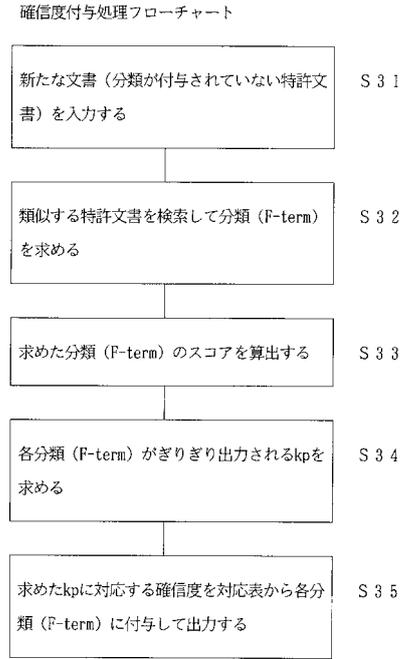
確信度付与装置の説明図



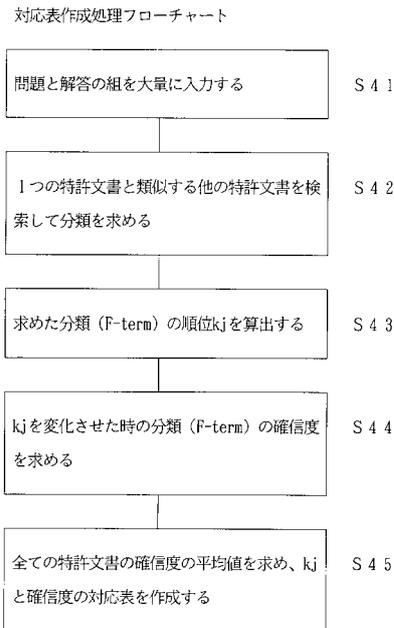
【 図 8 】



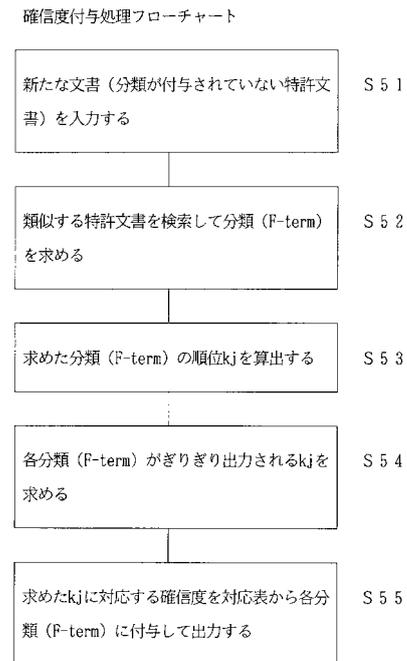
【 図 9 】



【 図 1 0 】

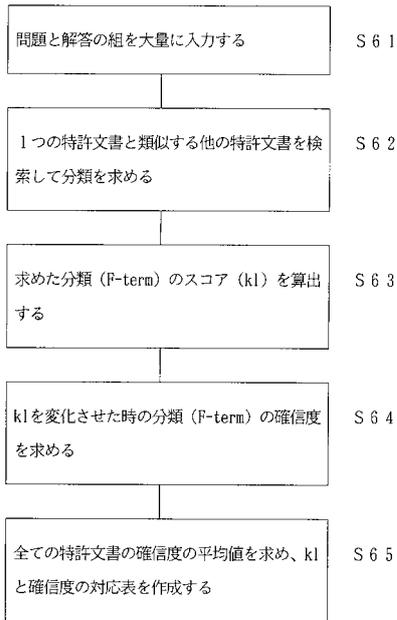


【 図 1 1 】



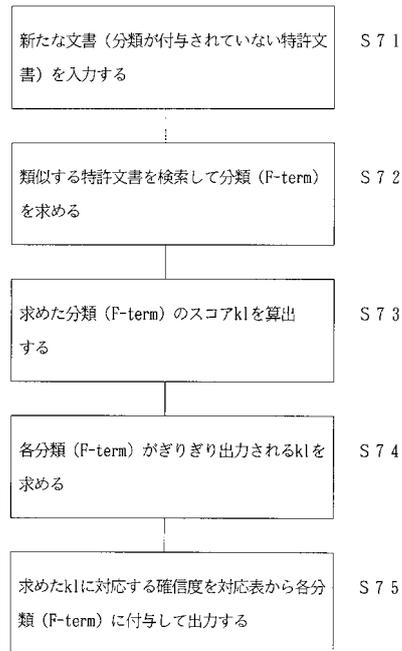
【 図 1 2 】

対応表作成処理フローチャート



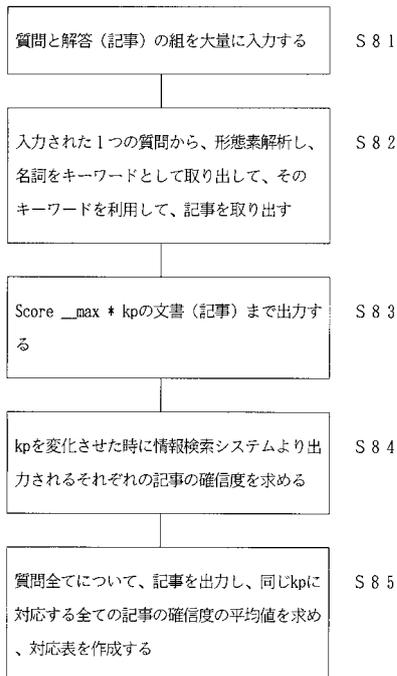
【 図 1 3 】

確信度付与処理フローチャート



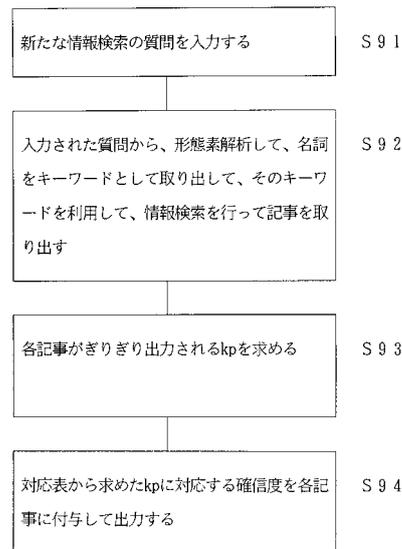
【 図 1 4 】

対応表作成処理フローチャート



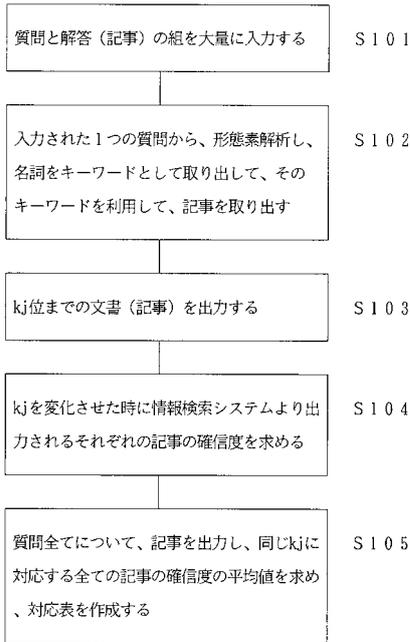
【 図 1 5 】

確信度付与処理フローチャート



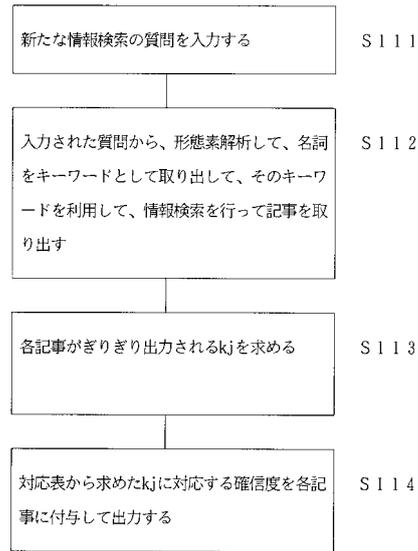
【 図 1 6 】

対応表作成処理フローチャート



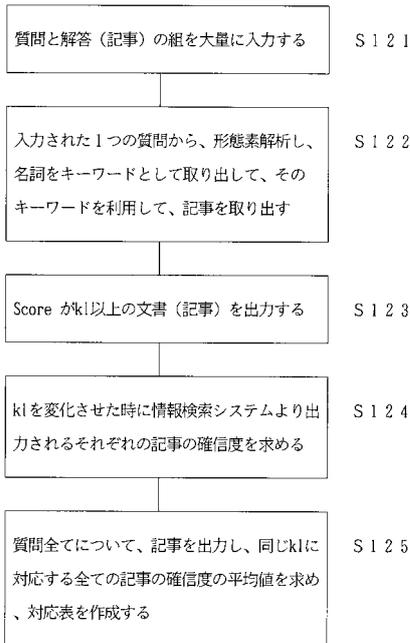
【 図 1 7 】

確信度付与処理フローチャート



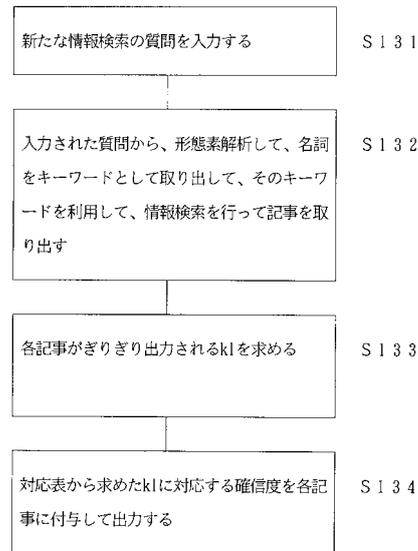
【 図 1 8 】

対応表作成処理フローチャート



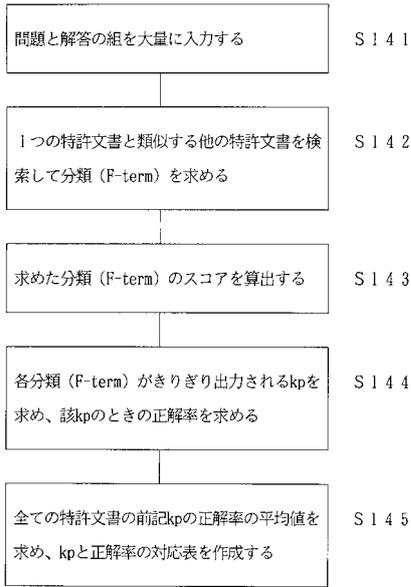
【 図 1 9 】

確信度付与処理フローチャート



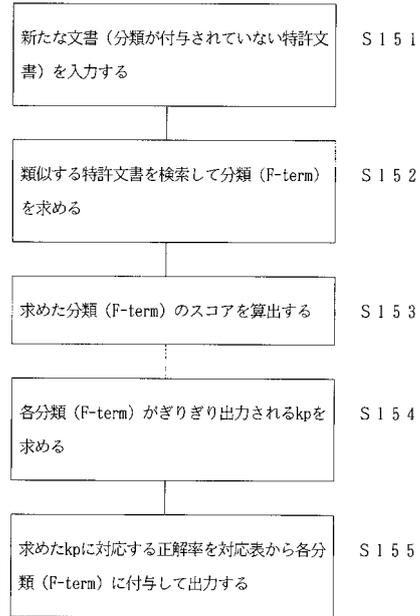
【 図 2 0 】

対応表作成処理フローチャート



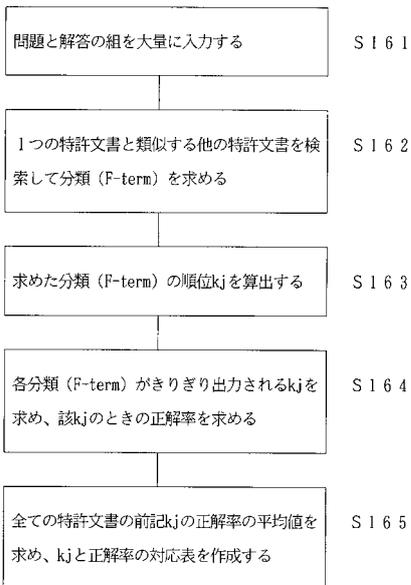
【 図 2 1 】

確信度付与処理フローチャート



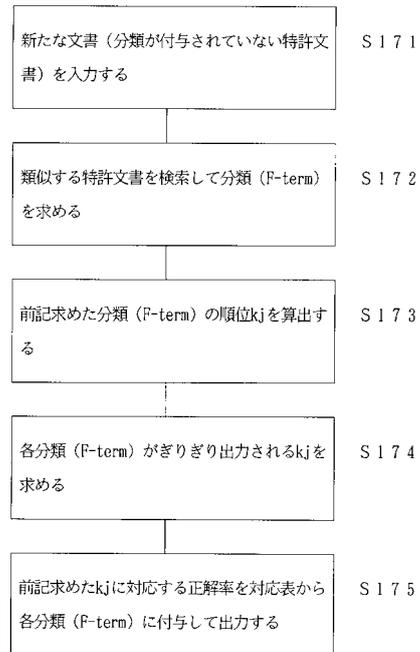
【 図 2 2 】

対応表作成処理フローチャート



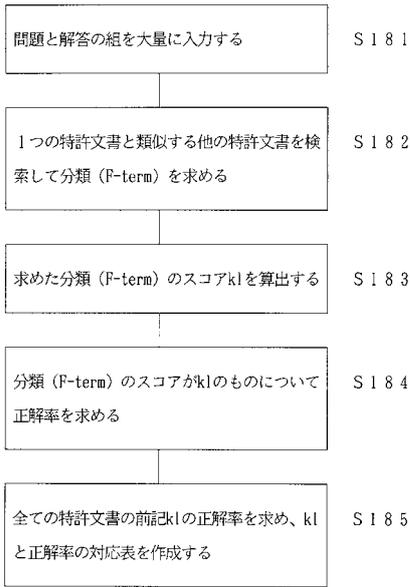
【 図 2 3 】

確信度付与処理フローチャート



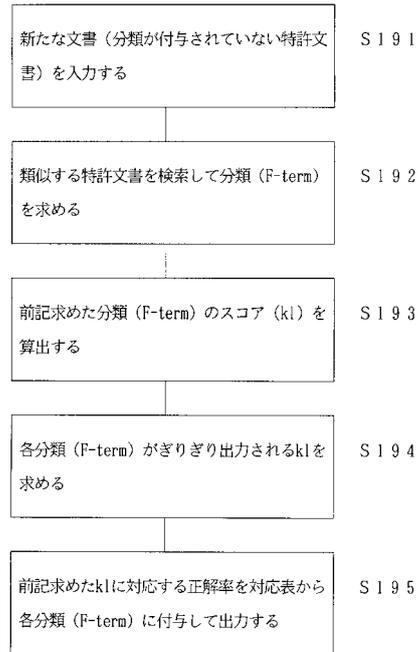
【 図 2 4 】

対応表作成処理フローチャート



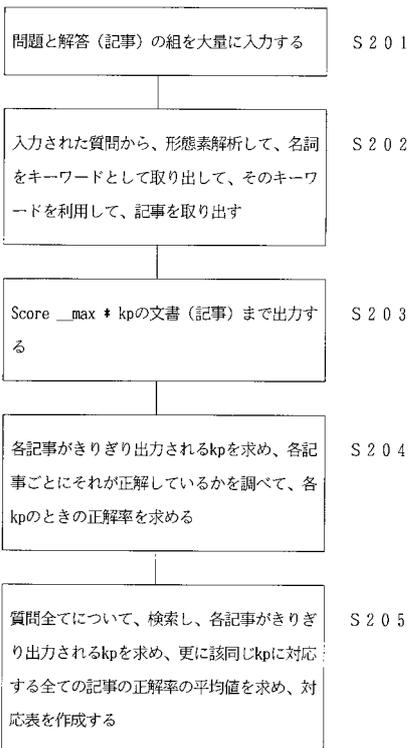
【 図 2 5 】

確信度付与処理フローチャート



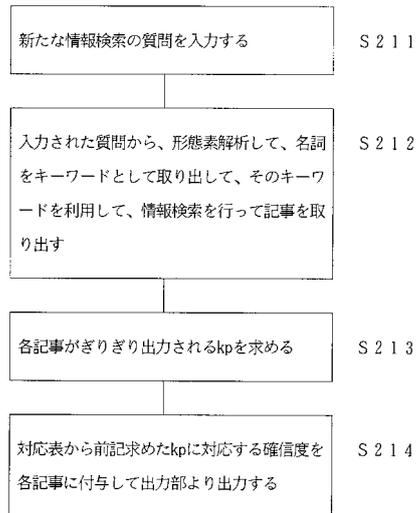
【 図 2 6 】

対応表作成処理フローチャート



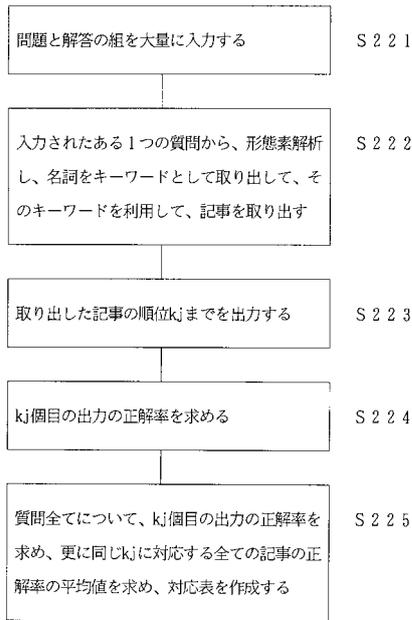
【 図 2 7 】

確信度付与処理フローチャート



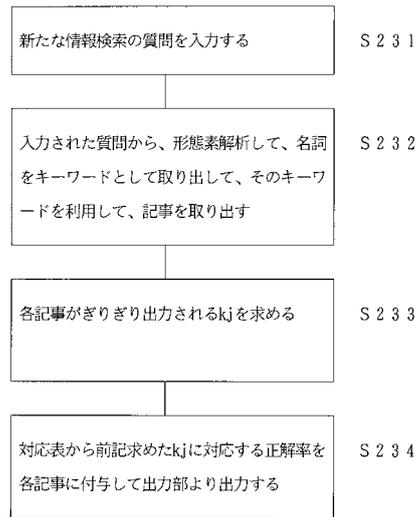
【 図 2 8 】

対応表作成処理フローチャート



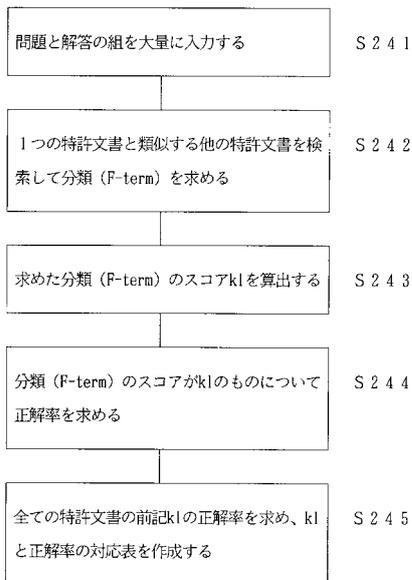
【 図 2 9 】

確信度付与処理フローチャート



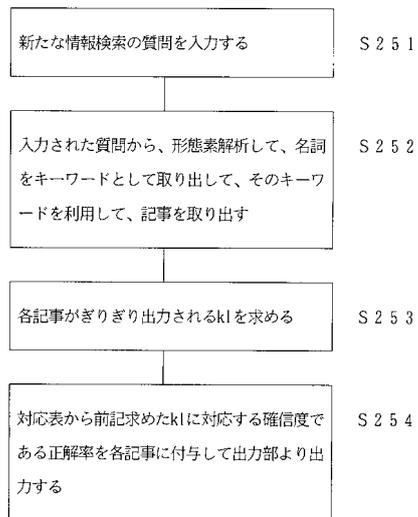
【 図 3 0 】

対応表作成処理フローチャート

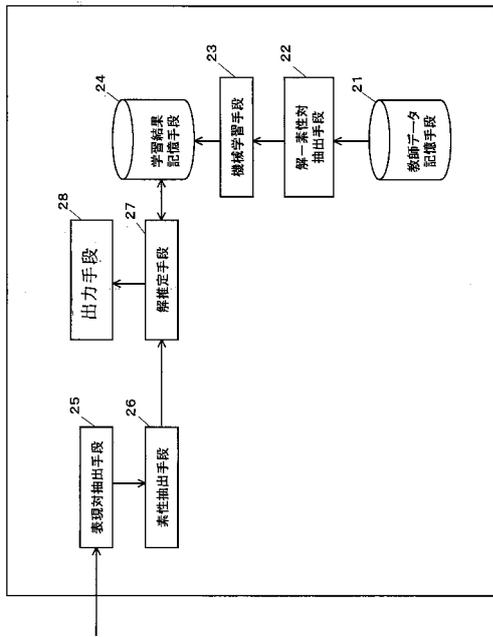


【 図 3 1 】

確信度付与処理フローチャート



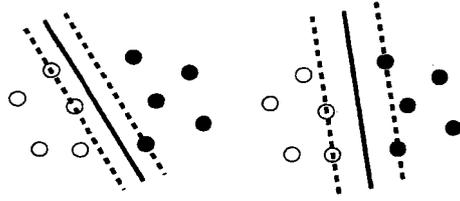
【 図 3 2 】



【 図 3 3 】

(A) スモールマシン

(B) ラージマシン



【 図 3 4 】

実験結果の説明図

kj	base0.5			kp			kj			kp, kj		
	再現率	適合率	F値	再現率	適合率	F値	再現率	適合率	F値	再現率	適合率	F値
1	0.40	0.50	0.32	0.05	0.29	0.08	0.05	0.29	0.08	0.05	0.29	0.08
2	0.31	0.34	0.22	0.10	0.26	0.13	0.08	0.27	0.11	0.09	0.26	0.12
3	0.25	0.32	0.17	0.13	0.22	0.14	0.10	0.22	0.12	0.11	0.22	0.13
4	0.20	0.26	0.14	0.14	0.19	0.14	0.12	0.19	0.13	0.13	0.18	0.13
5	0.17	0.23	0.13	0.15	0.17	0.14	0.13	0.18	0.13	0.14	0.18	0.13
10	0.15	0.13	0.11	0.15	0.14	0.12	0.15	0.13	0.11	0.15	0.13	0.11
20	0.28	0.17	0.11	0.14	0.09	0.10	0.14	0.10	0.10	0.14	0.09	0.10
30	0.35	0.25	0.14	0.11	0.07	0.09	0.12	0.07	0.09	0.12	0.07	0.09
40	0.39	0.30	0.18	0.09	0.06	0.08	0.10	0.06	0.08	0.10	0.06	0.08
50	0.42	0.33	0.22	0.08	0.05	0.07	0.09	0.05	0.07	0.08	0.05	0.07
100	0.47	0.41	0.33	0.04	0.03	0.05	0.04	0.03	0.04	0.04	0.03	0.04