(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4925293号
(P4925293)
(45)発行日　平成24年4月25日(2012.4.25)　(24)登録日　平成24年2月17日(2012.2.17)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 370Z

請求項の数 15　(全 50 頁)

(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)
(65)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)
審査請求日　平成21年9月30日(2009.9.30)

(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　野崎　大進

最終頁に続く

(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)【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。
【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。

【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、
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。
【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。
【請求項１２】
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１３】
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１４】
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１５】
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に、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するときに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、前記従来の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。

【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。
【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。
【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。
【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。

【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。
【００２２】
（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。
【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。
【００２７】
（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。
【００２８】
（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。
【００２９】
（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。
【００３０】
（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。
【００３１】
（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。

【００３２】
（１２）：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。
【００３３】
（１３）：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】
【００３４】
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。
【００３５】
§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。
【００３６】
（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。
【００３７】
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（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。
【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。

【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。
【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。
【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】
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。
【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。
【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。
【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 $_{M1}$(x) ）を計算する。

【００５３】
【数１】

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

【００５４】
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。

【００５５】
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。

【００５６】
{ x | Score $_{M1}$(x) ≧ k_p × 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。

【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）
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。

【００５８】
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。

【００５９】
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。

【００６０】
{ x | F_{KDOC}(x) ≧ k_u ×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。

【００６１】

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。

【００６２】

（５）：対応表の説明

上記方法１～３でkp、k_u 、k_f を変化すると、取り出すF-termの数が変化するこになる。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kp、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。

【００６３】

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

kp=0.9の時に選ばれたF-termの精度→95%
kp=0.8の時に選ばれたF-termの精度→85%
kp=0.7の時に選ばれたF-termの精度→80%
kp=0.6の時に選ばれたF-termの精度→75%
kp=0.5の時に選ばれたF-termの精度→65%
kp=0.4の時に選ばれたF-termの精度→50%
kp=0.3の時に選ばれたF-termの精度→45%
kp=0.2の時に選ばれたF-termの精度→20%
kp=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、kp=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。

【００６４】

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。

【００６５】

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。

【００６６】

（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。

【００６７】

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。

【００６８】

【数２】

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

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

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

【００６９】

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。

【００７０】

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）。

【００７１】

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。

【００７２】

【数３】

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

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

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

【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。

【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。

【００７５】

【数４】

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

【００７６】

この式で、T 、tf、N 、 nは、 SMARTのものと同一である。

【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 $Sim_{Overlap}$ 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。

【００７８】

【数５】

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

【００７９】

この式で、T は、 SMARTのものと同一である。

【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になる。

【００８１】

【数６】

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

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

【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筌 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。

【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
ＥＯＳ

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。

【００８６】

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。

【００８７】

（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。

【００８８】

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。

【００８９】

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。

【００９０】

各F-termのkpが求まれば、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。

【００９１】

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。

【００９２】

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。

【００９３】

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。

【００９４】

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。

【００９５】

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）。

【００９６】

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。

【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。

【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。
【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。
【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）。
【０１１０】
各Fterm のkjが求まれば、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）。
【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。
【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）。

【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。

【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。

【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Fterm のklが求まれば、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。（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。

【０１３９】

§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。

【０１４０】

（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。

【０１４１】

(1) 基本的な方法 (tf・idf 法) の説明

$$score(D) = \sum_{w \in W} (tf(w,D) * \log(N/df(w)))$$

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。

【０１４２】

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

【０１４３】

Okapi の式なら

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

で加算

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。

【０１４４】

さらに、以下の情報検索を行うこともできる。

【０１４５】

（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。

【０１４６】

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。

【０１４７】

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。

【０１４８】

（文献）

"J. J. Rocchio", "Relevance feedback in information retrieval","The SMART retrieval System", "Edited by G. Salton", "Prentice Hall, Inc.","page 313-323", 1971

この方法は、log(N/df(w))のかわりに、

$$\{E(t) + k\_af * (RatioC(t) - RatioD(t))\} * \log(N/df(w))$$

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)
RatioC(t) は記事群Ｂでのt の出現率
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。

【０１５０】

Score(D)のΣ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Σ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。

【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0kapi 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。

【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、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。

【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が求まれば、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 __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）で文書を出力する。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が求まれば、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。

【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）。

【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。

【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

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が求まれば、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（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の式のΣ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、Σで加算するということをやめて、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。

【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Σの内部の部分の式を利用して求めて、入力の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Σ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、Σで加算するということをやめて、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、入力の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、kpと kl の二つを使うことを考える。kp = 0, 0.1, 0.2, ,, 1.0と kj = 1, 2, 3, ,, 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】
kp = 0, kj = 1 の場合の確信度 ...
kp = 0.1, kj = 1 の場合の確信度 ...
...
kp = 0, kj = 2 の場合の確信度 ...
kp = 0.1, kj = 2 の場合の確信度 ...
...
...
kp = 0, kj = 1000 の場合の確信度 ...
kp = 0.1, kj = 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 kp, kj を求める。このkp, kjは、１つのときと同じ方法で求めれる。kp, kj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 kp, kj 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 kp が kp1で kj が kj1であったとする。そして、kp1 , kj1 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、kp1 と最も近い値の kp と、kj1 と最も近い値の kj 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、kp1 をはさむ二つの kp 、kj1 をはさむ二つの kj を用い、二つの kp と二つの kj から kp, kj 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kp1 をはさむ二つの kp の２つのデータ kp2, kp3 (kp2>kp>kp3) をkj1 をはさむ二つの kj の２つのデータ kj2, kj3 (kj2>kj>kj3)を用い、kp2, kj2のときの確信度を p(2,2) 、kp3, kj2のときの確信度を p(3,2) 、kp2, kj3のときの確信度を p(2,3) 、kp3, kj3のときの確信度を p(3,3) とし、
r(2,2) = sqrt((kp-kp2) ^2 + a(kj-kj2) ^2)
r(3,2) = sqrt((kp-kp3) ^2 + a(kj-kj2) ^2)
r(2,3) = sqrt((kp-kp2) ^2 + a(kj-kj3) ^2)
r(3,3) = sqrt((kp-kp3) ^2 + a(kj-kj3) ^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。
【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。^はべき乗を意味し、sqrt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。kp kj kl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）kpの値を利用する場合の説明

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kp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。

【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（特許文書）のF-term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。

【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。

【０２１４】

各F-termのkpが求まれば、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。

【０２１６】

Ｓ１４２：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）。

【０２１７】

Ｓ１４３：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。

【０２１８】

Ｓ１４４：対応表作成部１１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。

【０２１９】

Ｓ１４５：対応表作成部１１は、前記Ｓ１４１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、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１３に格納される）。

【０２２０】

図２１は確信度付与処理フローチャートである。以下、図２１の処理Ｓ１５１～Ｓ１５５にしたがって説明する。

【０２２１】

Ｓ１５１：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。

【０２２２】
Ｓ１５２：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Ｓ１４１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）。
【０２２３】
Ｓ１５３：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。
【０２２４】
Ｓ１５４：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。
【０２２５】
Ｓ１５５：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。
【０２２６】
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。
【０２２７】
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）。
【０２２８】
各Fterm のkjが求まれば、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。
【０２２９】
図２２は対応表作成処理フローチャートである。以下、図２２の処理Ｓ１６１～Ｓ１６５にしたがって説明する（確信度付与装置は図７参照）。
【０２３０】
Ｓ１６１：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。
【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Fterm のklが求まれば、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。

【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。

【０２４５】

Ｓ１８２：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）。

【０２４６】

Ｓ１８３：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。

【０２４７】

Ｓ１８４：対応表作成部１１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。

【０２４８】

Ｓ１８５：対応表作成部１１は、これを前記Ｓ１８１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１３に格納される）。

【０２４９】

図２５は確信度付与処理フローチャートである。以下、図２５の処理Ｓ１９１～Ｓ１９５にしたがって説明する（確信度付与装置は図７参照）。

【０２５０】

Ｓ１９１：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。

【０２５１】

Ｓ１９２：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Ｓ１８１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）。

【０２５２】

Ｓ１９３：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。

【０２５３】

Ｓ１９４：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。

【０２５４】

Ｓ１９５：確信度付与部１２は、格納部１３の対応表から前記求めたklに対応する正解率を各F-termに付与して出力部より出力する。

【０２５５】

（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。

【０２５６】

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。

【０２５７】

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。

【０２５８】

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。

【０２５９】

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。

【０２６２】

各記事のkpが求まれば、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）。

【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、各記事がきりぎり出力される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。
【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。
【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）。
【０２７５】
各Fterm のkjが求まれば、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。

【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）。
【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。
【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
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 kl のものについて、その記事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。
【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klを求める。すると各記事のスコアが 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）
各記事のklが求まれば、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。
【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。

【０２９２】

Ｓ２４３：情報検索システムは、スコアがkl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 kl のものについて、その記事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。

【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klに対応する正解率の平均値を求め、kl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）。

【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。

【０２９８】

Ｓ２５３：確信度付与部１２は、各記事がぎりぎり出力されるkl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kl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 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

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 の部分を 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)〕

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

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-siryou.pdf

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。

【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。

【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。

【０３１２】

【数７】

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

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

【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≦ｊ≦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、スームージングを行う。ここでは、以下の式（１３）を用いてスームージングを行ったものを用いる。

【０３１４】

【数８】

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

【０３１５】

ただし、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、ｆｒｅｑ（ａ）は、分類がａである事例の個数を意味する。

【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が与えられたときに、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≦ｊ≦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｜ｆmax）　　（１４）

ただし、ｆmax は以下の式によって与えられる。

【０３１９】

【数９】

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

【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（$1 \leqq j \leqq k$）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ま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。
【０３２２】
【数１０】

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

$$for\ \forall f_j(1 \leq j \leq k)$$

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

【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。
【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】
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。
【０３２６】
図３３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３３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３３（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３３（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。

【０３２７】

このとき、二つの分類が正例と負例からなるものとすると、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３３（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。

【０３２８】

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。

【０３２９】

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。

【０３３０】

【数１１】

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

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

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

【０３３１】

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，…，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、

ｓｇｎ（ｘ）＝１（ｘ≧０）

－１（otherwise ）

であり、また、各α_iは式（２０）と式（２１）の制約のもと式（１９）を最大にする場合のものである。

【０３３２】

【数１２】

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

$$0 \le \alpha_i \le C \quad (i = 1,...,l) \quad (20)$$

$$\sum_{j=1}^{l}\alpha_i y_i = 0 \quad (21)$$

【０３３３】

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。

【０３３４】

Ｋ（ｘ，ｙ）＝（ｘ・ｙ＋１）ｄ　　（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。

【０３３５】

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。

【０３３６】

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)

参考文献（７）：Taku Kudoh, Tinysvm:Support Vector machines,(http://cl.aist-nara.ac.jp/taku-ku//software/Tiny SVM/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。

【０３３７】

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。

【０３３８】

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。推定するべき候補が、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

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。

【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「抽出するべき」という分類が獲得した票数とする。

【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「抽出するべき」という分類になる確率とする。

【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「抽出するべき」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。

【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

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「抽出するべき」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（抽出するべき表現対）の空間への距離の大きさとする。より詳しくは、抽出するべき表現対を正例、抽出する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「抽出するべき事例」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。

【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。

【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。

【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。

【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

（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kp、kj、kl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良いし、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。

【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いなかという素性や、その文字列が該当特許に含まれるかいな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kpを求める。この求め方は、以下のようにする。

【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kpとなる。また、その記事の順位kjも求める。また、その記事のスコア＝ kl 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、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。

【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。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。

【０３６３】

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。

【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kpを求める。この求め方は、以下の通りである。

【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kpとなる。また、その記事の順位kjも求める。また、その記事のスコア＝kl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kp、kj、kl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】

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kp、kj、kl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良いし、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。

【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）

重回帰分析では、x1，x2，x3,,, yの組のデータがあるときに、x1，x2，x3,,, から y を求める。

【０３７２】

y = a0 + a1 * x1 + a2 * x2 + a3 * x3 + ...

の式の係数 a0，a1，...を、データから適切にもとめることができる。

【０３７３】

(1) kp - yの組からの予測

y は確信度

x1 = kp として、

y = a0 + a1 * kpとして、

回帰分析により kp - y の組のデータから a0，a1 を求める。kpから yを求める式が求まる。

【０３７４】

(2) kp - yの組からの予測 (2 次の利用)

y は確信度

x1 = kp 、 x2 = kp＾2 として、

y = a0 + a1 * kp + a2 * kp＾2 として、

重回帰分析により kp - y の組のデータから a0，a1，a2 を求める。kpから yを求める式が求まる。

【０３７５】

(3) kp、kj - yの組からの予測

y は確信度

x1 = kp 、x2 = kj として、

y = a0 + a1 * kp + a2 * kjとして、

重回帰分析により kp 、 kj - y の組のデータから a0，a1，a2 を求める。kp，kjから yを求める式が求まる。

【０３７６】

(4) kp 、 kj - y の組からの予測 (2 次の利用)

y は確信度

x1 = kp，x2 = kjとして、

y = a0 + a1 * kp + a2 * kj + a3 * kp＾2 + a4 * kp * kj + a5 * kj＾2 として、
重回帰分析により kp 、 kj - y の組のデータから a0, a1, a2,,, a5 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。

【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。

【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もよい。

【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。

【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３４で示している。

【０３８６】

図３４は実験結果の説明図であり、図３４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３４では、４つの方法を試してある。ここの確信度（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

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】

〔(kp -kp3) pr2 + (kp2 - kp) pr3〕／〔(kp2 - kp) + (kp -kp3)〕

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前後と少し大きいがそれを除くと、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

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。

【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。

【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。

【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。

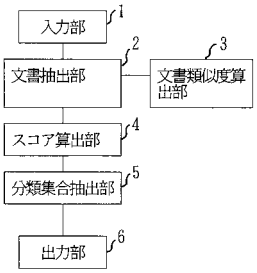
【符号の説明】

【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）

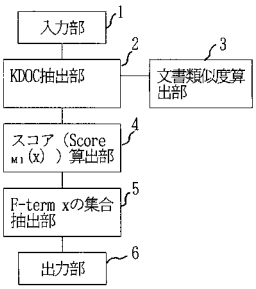
【図１】

文書分類装置の説明図



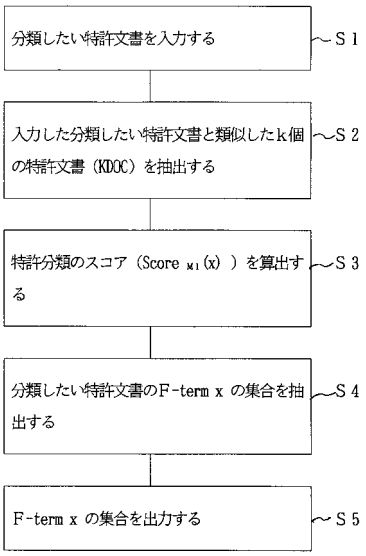
【図２】

特許文書分類装置の説明図



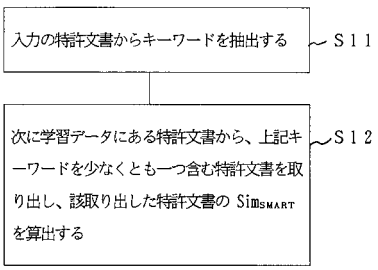
【図３】

特許文書の分類処理フローチャート



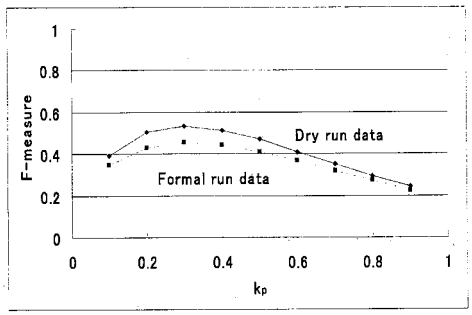
【図４】

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



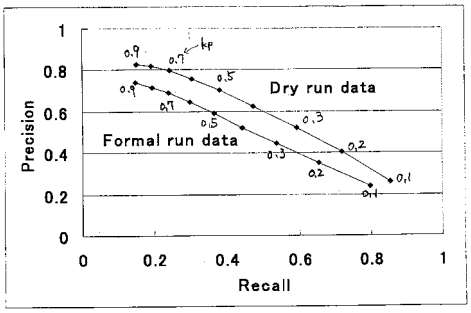
【図５】

kpとF値の対応の説明図



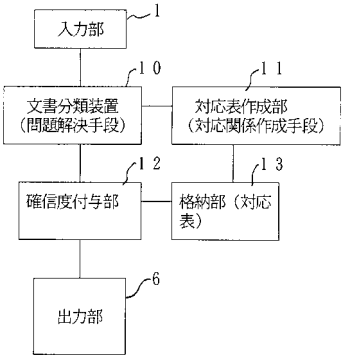
【図６】

kpと再現率と精度の対応の説明図

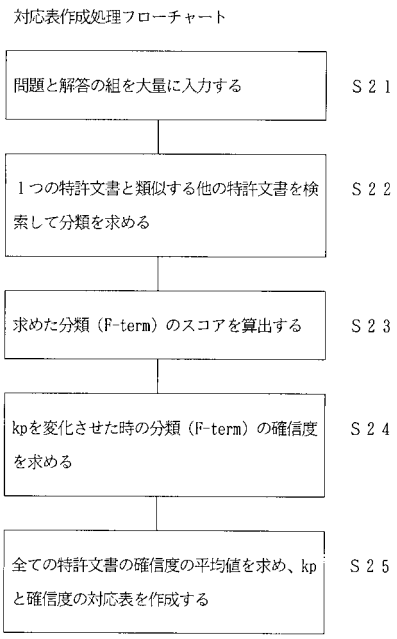


【図７】

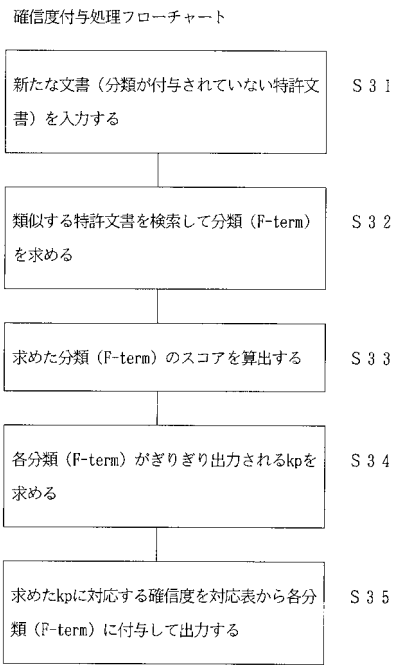
確信度付与装置の説明図



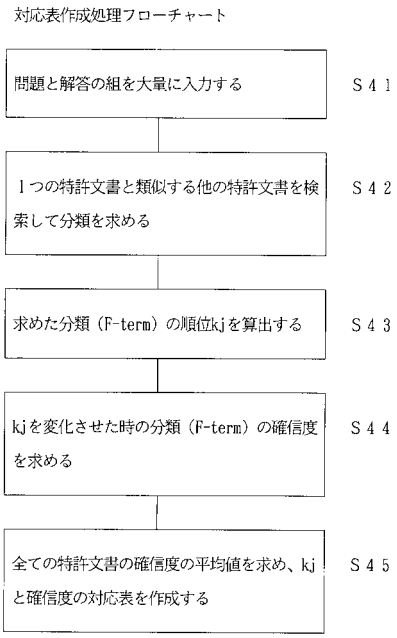
【図８】


対応表作成処理フローチャート
問題と解答の組を大量に入力する
Ｓ２１
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
Ｓ２２
求めた分類（F-term）のスコアを算出する
Ｓ２３
kpを変化させた時の分類（F-term）の確信度を求める
Ｓ２４
全ての特許文書の確信度の平均値を求め、kpと確信度の対応表を作成する
Ｓ２５

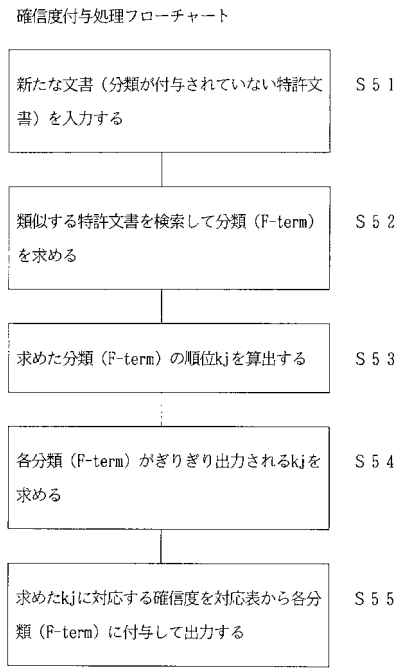
【図９】


確信度付与処理フローチャート
新たな文書（分類が付与されていない特許文書）を入力する
Ｓ３１
類似する特許文書を検索して分類（F-term）を求める
Ｓ３２
求めた分類（F-term）のスコアを算出する
Ｓ３３
各分類（F-term）がぎりぎり出力されるkpを求める
Ｓ３４
求めたkpに対応する確信度を対応表から各分類（F-term）に付与して出力する
Ｓ３５

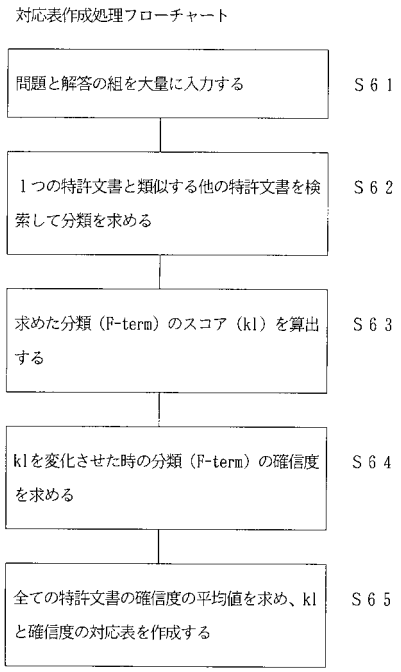
【図１０】


対応表作成処理フローチャート
問題と解答の組を大量に入力する
Ｓ４１
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
Ｓ４２
求めた分類（F-term）の順位kjを算出する
Ｓ４３
kjを変化させた時の分類（F-term）の確信度を求める
Ｓ４４
全ての特許文書の確信度の平均値を求め、kjと確信度の対応表を作成する
Ｓ４５

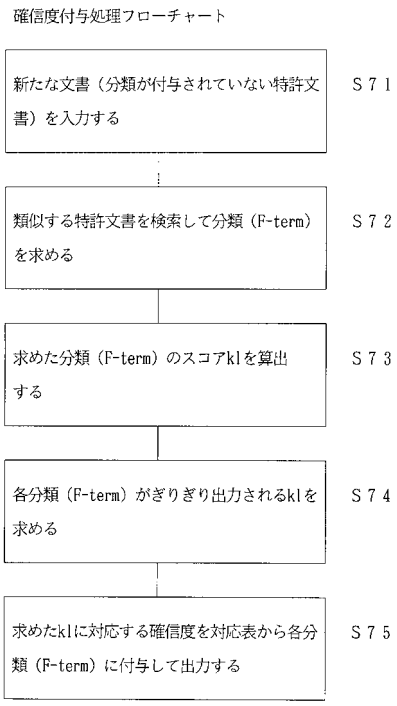
【図１１】


確信度付与処理フローチャート
新たな文書（分類が付与されていない特許文書）を入力する
Ｓ５１
類似する特許文書を検索して分類（F-term）を求める
Ｓ５２
求めた分類（F-term）の順位kjを算出する
Ｓ５３
各分類（F-term）がぎりぎり出力されるkjを求める
Ｓ５４
求めたkjに対応する確信度を対応表から各分類（F-term）に付与して出力する
Ｓ５５

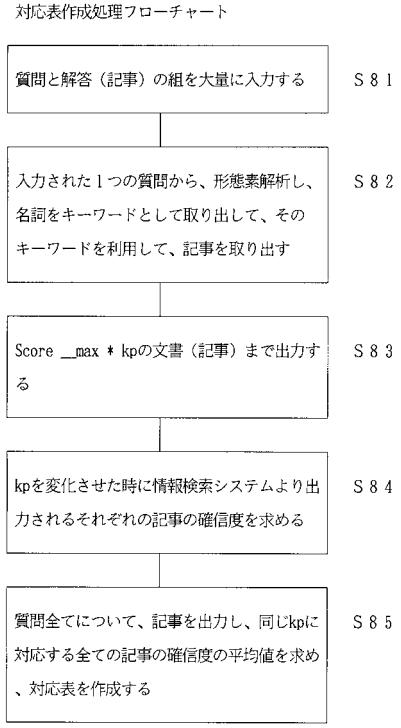
【図１２】



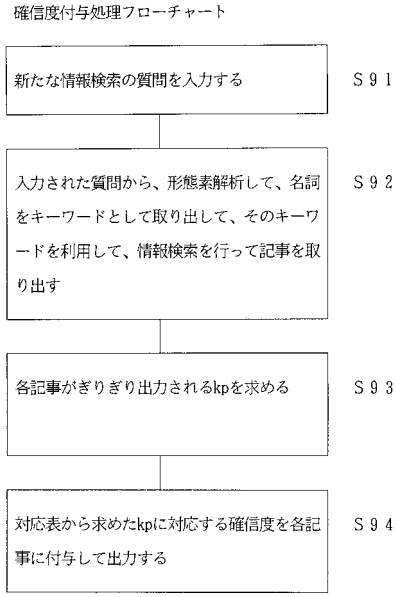
【図１３】



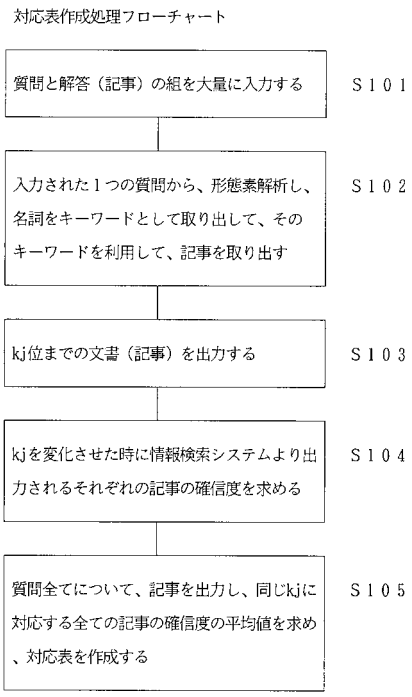
【図１４】



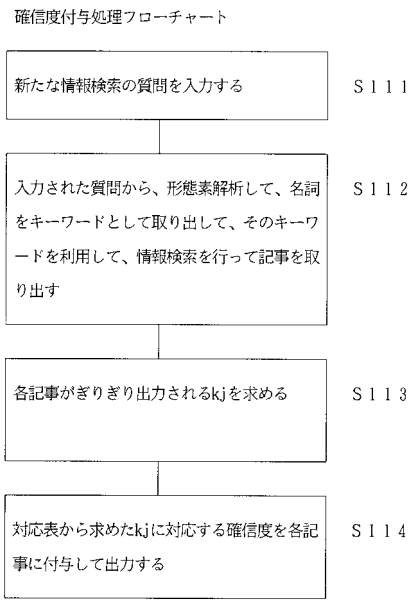
【図１５】



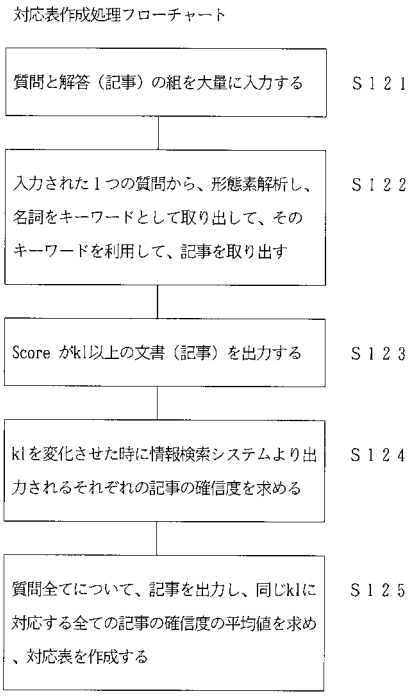
【図１６】



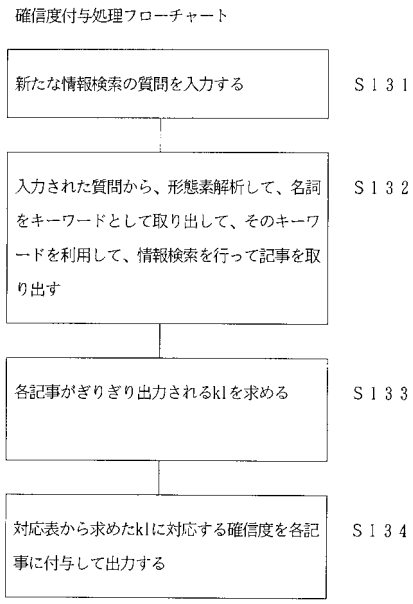
【図１７】



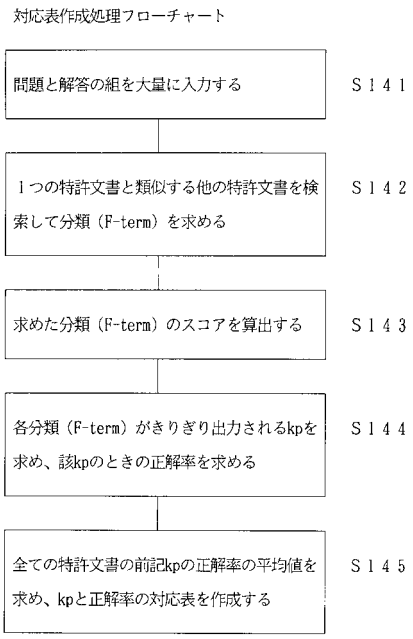
【図１８】



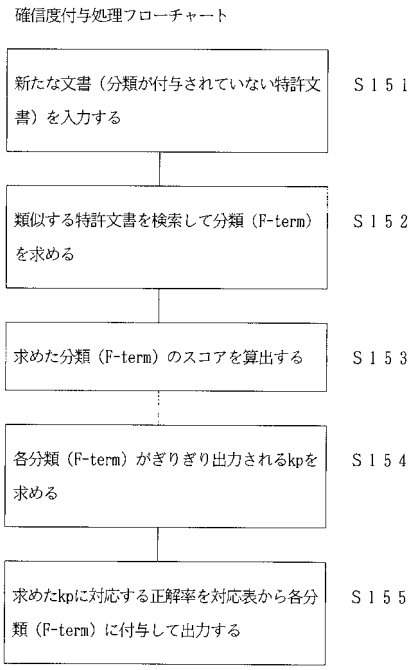
【図１９】



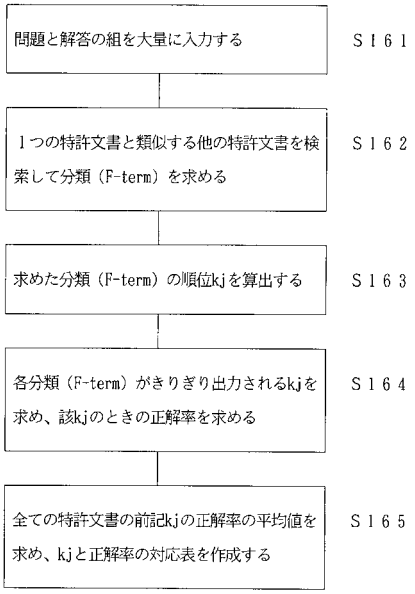
【図２０】



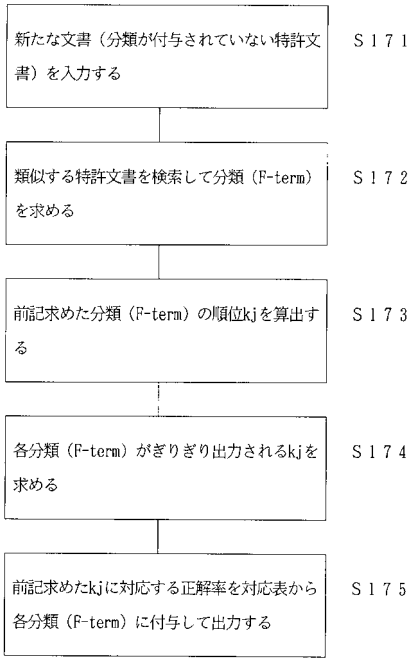
【図２１】



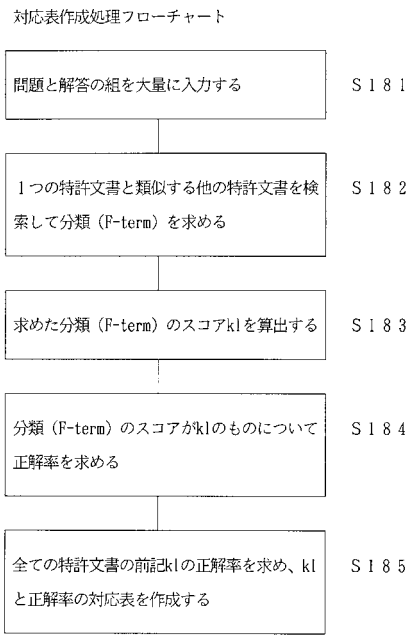
【図２２】



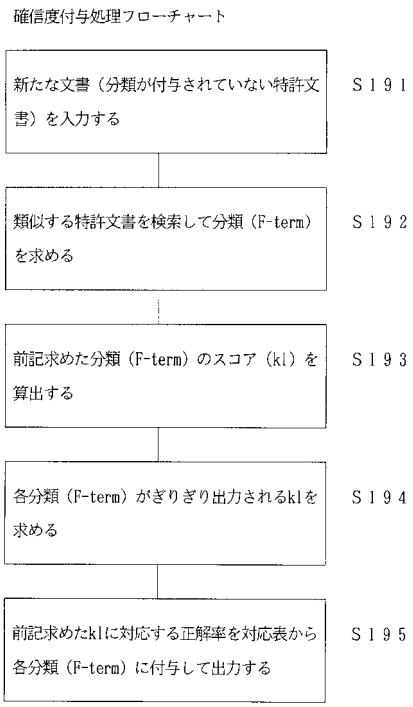
【図２３】



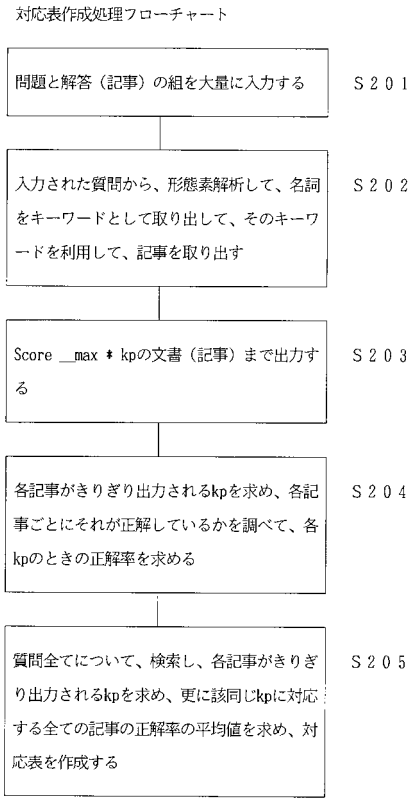
【図２４】



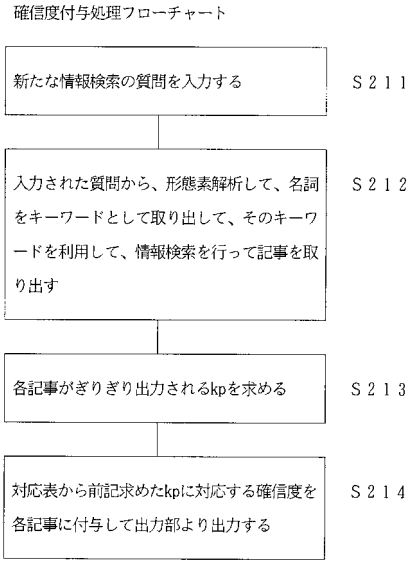
【図２５】



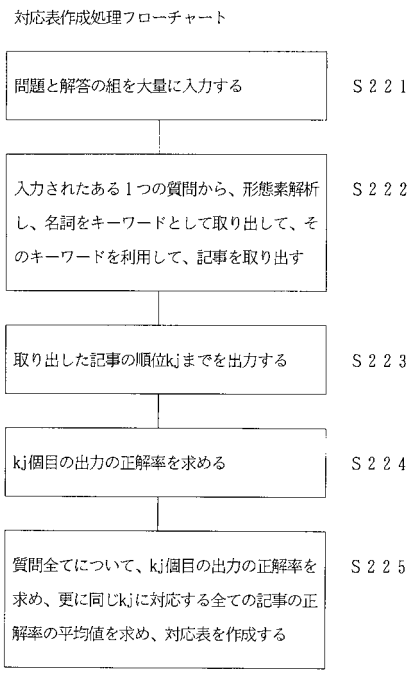
【図２６】



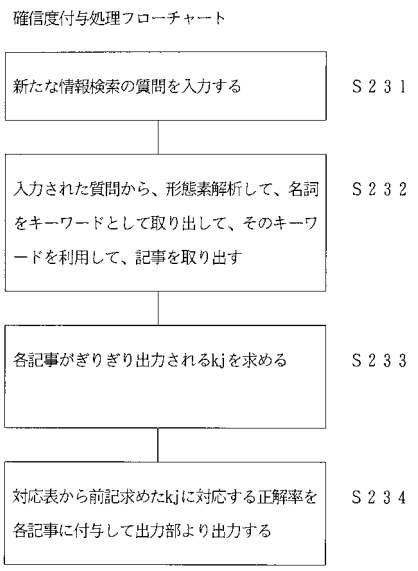
【図２７】



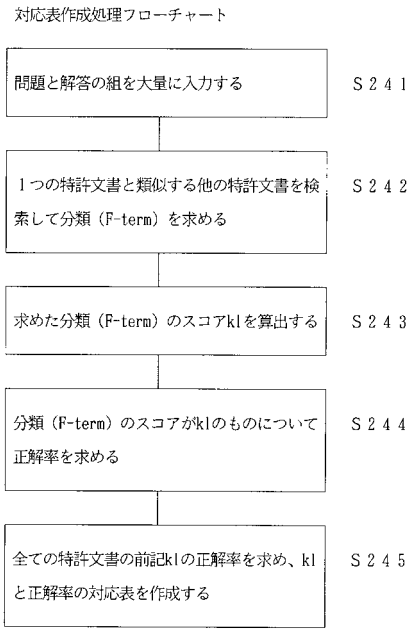
【図２８】



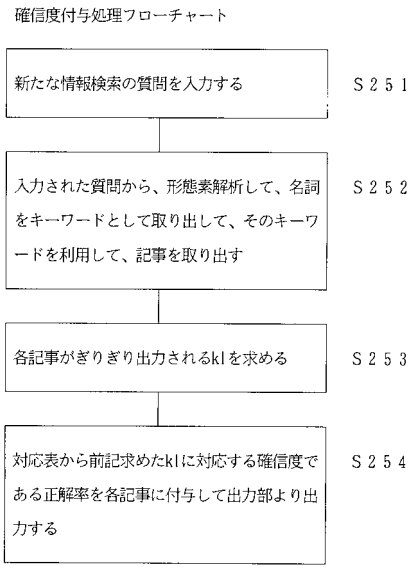
【図２９】



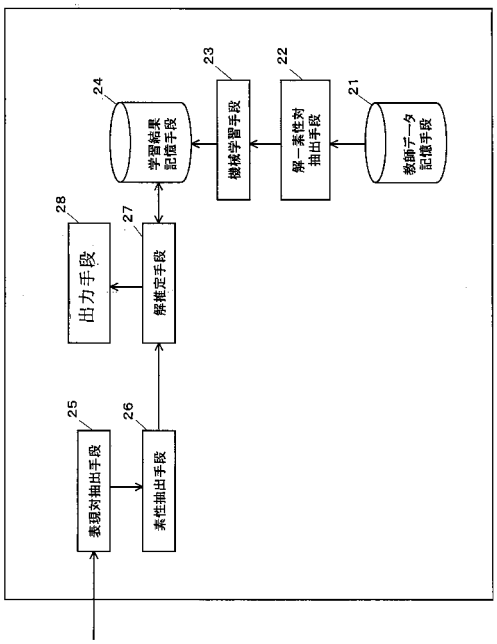
【図３０】



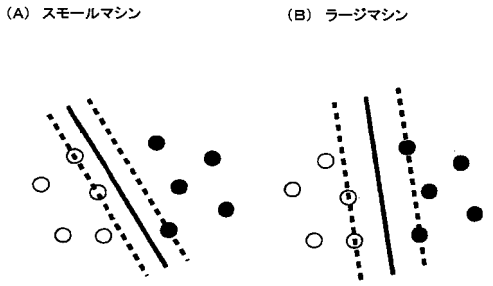
【図３１】



【図３２】



【図３３】



【図３４】

実験結果の説明図

	base0.5			kp			kj			kp, kj		
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

フロントページの続き

(56)参考文献 特開２００３－０２２２７５（ＪＰ，Ａ）
乾 裕子 他，表層表現に着目した自由回答アンケートの意図に基づく自動分類，自然言語処理，日本，言語処理学会，２００３年 4月１０日，Vol.10,No.2，PP.19-42.
金丸 敏幸 他，話者の意図に関わる副詞辞書の構築，言語処理学会第１２回年次大会ワークショップ「感情・評価・態度と言語」論文集，日本，言語処理学会，２００６年 3月１３日，PP.41-44.

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