(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4067603号
(P4067603)
(45)発行日　平成20年3月26日(2008.3.26)　(24)登録日　平成20年1月18日(2008.1.18)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 210D
G06F 17/30 350C
G06F 17/30 170A

請求項の数 3　(全 15 頁)

(21)出願番号　特願平9-217131
(22)出願日　平成9年7月27日(1997.7.27)
(65)公開番号　特開平11-45247
(43)公開日　平成11年2月16日(1999.2.16)
審査請求日　平成16年7月2日(2004.7.2)

前置審査

(73)特許権者　390024350
株式会社ジャストシステム
徳島県徳島市川内町平石若松１０８番地４
(74)代理人　100096655
弁理士　川井　隆
(74)代理人　100091225
弁理士　仲野　均
(72)発明者　野村　直之
徳島県徳島市沖浜東３丁目４６番地　株式会社ジャストシステム内

審査官　辻本　泰隆

最終頁に続く

(54)【発明の名称】文書分類装置、文書分類プログラムが記憶された記憶媒体、及び文書分類方法

(57)【特許請求の範囲】
【請求項１】
予め決められた複数分類のセットの範囲内で、人手によって対象文書を分類した手動分類結果を取得する手動分類結果取得手段と、
前記対象文書を取得する対象文書取得手段と、
前記対象文書取得手段で取得された対象文書を、前記複数分類のセットの範囲内で、自動的に分類して自動分類結果を得る自動分類結果取得手段と、
前記手動分類結果と前記自動分類結果とから前記対象文書に対する分類を最終決定する分類決定手段と、
各分類に対する評価値を担当者ごとに格納する評価関数データベースと、
対象文書を手動で分類した担当者の情報を取得する分類担当取得手段と、を具備し、
前記自動分類結果取得手段は、
前記対象文書を特徴づける対象文書ベクトルを取得する文書ベクトル取得手段と、
前記各分類を特徴づける典型文書の典型文書ベクトルを取得する典型文書ベクトル取得手段と、
前記対象文書ベクトルと前記各典型文書ベクトルとの類似度を算出して各分類に対する類似度を得る類似度算出手段と
を有し、前記類似度算出手段によって得られた各分類に対する類似度を分類結果とし、
前記分類決定手段は、前記手動分類結果と前記取得した分類担当者に対応する評価値とに基づいて手動分類の点数を算出し、また、前記類似度算出手段によって得られた各分類

に対する類似度に基づいて自動分類の点数を算出し、そして、前記算出された手動分類の点数と自動分類の点数との合計値を分類ごとに算出し、この算出結果に基づいて、前記対象文書に対する分類を最終決定することを特徴とする文書分類装置。

【請求項２】

各分類に対する評価値を担当者ごとに格納する評価関数データベースを備えたコンピュータに、

予め決められた複数分類のセットの範囲内で、人手によって対象文書を分類した手動分類結果を取得する手動分類結果取得機能と、

前記対象文書を取得する対象文書取得機能と、

前記対象文書取得機能で取得された対象文書を、前記複数分類のセットの範囲内で、自動的に分類して自動分類結果を得る自動分類結果取得機能と、

前記手動分類結果と前記自動分類結果とから前記対象文書に対する分類を最終決定する分類決定機能と、

対象文書を手動で分類した担当者の情報を取得する分類担当取得機能と、を実現させるためのコンピュータ読取り可能な文書分類プログラムが記憶された記憶媒体であって、

前記自動分類結果取得機能は、

前記対象文書を特徴づける対象文書ベクトルを取得する文書ベクトル取得機能と、

前記各分類を特徴づける典型文書の典型文書ベクトルを取得する典型文書ベクトル取得機能と、

前記対象文書ベクトルと前記各典型文書ベクトルとの類似度を算出して各分類に対する類似度を得る類似度算出機能と

を有し、前記類似度算出機能によって得られた各分類に対する類似度を分類結果とし、

前記分類決定機能は、前記手動分類結果と前記取得した分類担当者に対応する評価値とに基づいて手動分類の点数を算出し、また、前記類似度算出機能によって得られた各分類に対する類似度に基づいて自動分類の点数を算出し、そして、前記算出された手動分類の点数と自動分類の点数との合計値を分類ごとに算出し、この算出結果に基づいて、前記対象文書に対する分類を最終決定することを特徴とする文書分類プログラムが記憶された記憶媒体。

【請求項３】

各分類に対する評価値を担当者ごとに格納する評価関数データベース、手動分類結果取得手段、対象文書取得手段、自動分類結果取得手段、分類決定手段、分類担当取得手段、文書ベクトル取得手段、典型文書ベクトル取得手段、類似度算出手段を有する文書分類装置において用いられる文書分類方法であって、

前記手動分類結果取得手段が、予め決められた複数分類のセットの範囲内で、人手によって対象文書を分類した手動分類結果を取得する第１ステップと、

前記対象文書取得手段が、前記対象文書を取得する第２ステップと、

前記自動分類結果取得手段が、前記第２ステップで取得された対象文書を、前記複数分類のセットの範囲内で、自動的に分類して自動分類結果を得る第３ステップと、

前記分類決定手段が、前記手動分類結果と前記自動分類結果とから前記対象文書に対する分類を最終決定する第４ステップと、

前記分類担当取得手段が、対象文書を手動で分類した担当者の情報を取得する第５ステップと、を有し、

前記第３ステップは、

前記文書ベクトル取得手段が、前記対象文書を特徴づける対象文書ベクトルを取得する第６ステップと、

前記典型文書ベクトル取得手段が、前記各分類を特徴づける典型文書の典型文書ベクトルを取得する第７ステップと、

前記類似度算出手段が、前記対象文書ベクトルと前記各典型文書ベクトルとの類似度を算出して各分類に対する類似度を得る第８ステップと

を有し、前記第８ステップによって得られた各分類に対する類似度を分類結果とし、

前記第４ステップは、前記手動分類結果と前記取得した分類担当者に対応する評価値とに基づいて手動分類の点数を算出し、また、前記第８ステップによって得られた各分類に対する類似度に基づいて自動分類の点数を算出し、そして、前記算出された手動分類の点数と自動分類の点数との合計値を分類ごとに算出し、この算出結果に基づいて、前記対象文書に対する分類を最終決定することを特徴とする文書分類方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書分類装置、文書分類プログラムが記憶された記憶媒体、及び文書分類方法に係り、詳細には、取得した対象文書に対する分類精度の向上に関する。

【０００２】

【従来の技術】

文書をファイルしたり、電子的に配信したり、記憶媒体に記憶させたりする場合、その対象文書を予め決められたカテゴリに分類する場合がある。

このように対象文書の分類を行う場合、従来では分類者担当ものがその対象文書を読んだ後に手動分類をしたり、コンピュータシステムを使用して文書内容を解析することで対象文書を自動的に分類したりしている。

【０００３】

【発明が解決しようとする課題】

しかし、従来の人手による文書の手動分類では、必ずしも正確に分類付けがされない場合があった。

一方、コンピュータシステムによる判断は高速に大量の文書を分類することが可能であるが、この分類も必ず下正確であるとは限らなかった。

また、従来の手動分類と自動分類とでは、分類形態が全く異なるため両者を融合したシームレスな使い勝手が実現しないかった。

【０００４】

本発明は、このような従来技術の課題を解決するために成されたもので、手動分類と自動分類の両分類結果を使用して、対象文書に対してより精度の高い分類を行うことが可能な文書分類装置を提供することを第１の目的とする。

また、本発明は、手動分類と自動分類の両分類結果を使用して、対象文書に対してより精度の高い分類を行うことが可能な文書分類プログラムが記録された記憶媒体を提供することを第１の目的とする。

また、本発明は、手動分類と自動分類の両分類結果を使用して、対象文書に対してより精度の高い分類を行うことが可能な文書分類方法を提供することを第３の目的とする。

【０００５】

【課題を解決するための手段】

前記第１の目的を達成するために、請求項１に記載した発明では、予め決められた複数分類のセットの範囲内で、人手によって対象文書を分類した手動分類結果を取得する手動分類結果取得手段と、前記対象文書を取得する対象文書取得手段と、前記対象文書取得手段で取得された対象文書を、前記複数分類のセットの範囲内で、自動的に分類して自動分類結果を得る自動分類結果取得手段と、前記手動分類結果と前記自動分類結果とから前記対象文書に対する分類を最終決定する分類決定手段と、各分類に対する評価値を担当者ごとに格納する評価関数データベースと、対象文書を手動で分類した担当者の情報を取得する分類担当取得手段と、を具備し、前記自動分類結果取得手段は、前記対象文書を特徴づける対象文書ベクトルを取得する文書ベクトル取得手段と、前記各分類を特徴づける典型文書の典型文書ベクトルを取得する典型文書ベクトル取得手段と、前記対象文書ベクトルと前記各典型文書ベクトルとの類似度を算出して各分類に対する類似度を得る類似度算出手段とを有し、前記類似度算出手段によって得られた各分類に対する類似度を分類結果とし、前記分類決定手段は、前記手動分類結果と前記取得した分類担当者に対応する評価値とに基づいて手動分類の点数を算出し、また、前記類似度算出手段によって得られた各分

類に対する類似度に基づいて自動分類の点数を算出し、そして、前記算出された手動分類の点数と自動分類の点数との合計値を分類ごとに算出し、この算出結果に基づいて、前記対象文書に対する分類を最終決定することを特徴とする文書分類装置を提供する。

前記第２の目的を達成するために、請求項２に記載した発明では、各分類に対する評価値を担当者ごとに格納する評価関数データベースを備えたコンピュータに、予め決められた複数分類のセットの範囲内で、人手によって対象文書を分類した手動分類結果を取得する手動分類結果取得機能と、前記対象文書を取得する対象文書取得機能と、前記対象文書取得機能で取得された対象文書を、前記複数分類のセットの範囲内で、自動的に分類して自動分類結果を得る自動分類結果取得機能と、前記手動分類結果と前記自動分類結果とから前記対象文書に対する分類を最終決定する分類決定機能と、対象文書を手動で分類した担当者の情報を取得する分類担当取得機能と、を実現させるためのコンピュータ読取り可能な文書分類プログラムが記憶された記憶媒体であって、前記自動分類結果取得機能は、前記対象文書を特徴づける対象文書ベクトルを取得する文書ベクトル取得機能と、前記各分類を特徴づける典型文書の典型文書ベクトルを取得する典型文書ベクトル取得機能と、前記対象文書ベクトルと前記各典型文書ベクトルとの類似度を算出して各分類に対する類似度を得る類似度算出機能とを有し、前記類似度算出機能によって得られた各分類に対する類似度を分類結果とし、前記分類決定機能は、前記手動分類結果と前記取得した分類担当者に対応する評価値とに基づいて手動分類の点数を算出し、また、前記類似度算出機能によって得られた各分類に対する類似度に基づいて自動分類の点数を算出し、そして、前記算出された手動分類の点数と自動分類の点数との合計値を分類ごとに算出し、この算出結果に基づいて、前記対象文書に対する分類を最終決定することを特徴とする文書分類プログラムが記憶された記憶媒体を提供する。

前記第３の目的を達成するために、請求項３に記載した発明では、各分類に対する評価値を担当者ごとに格納する評価関数データベース、手動分類結果取得手段、対象文書取得手段、自動分類結果取得手段、分類決定手段、分類担当取得手段、文書ベクトル取得手段、典型文書ベクトル取得手段、類似度算出手段を有する文書分類装置において用いられる文書分類方法であって、前記手動分類結果取得手段が、予め決められた複数分類のセットの範囲内で、人手によって対象文書を分類した手動分類結果を取得する第１ステップと、前記対象文書取得手段が、前記対象文書を取得する第２ステップと、前記自動分類結果取得手段が、前記第２ステップで取得された対象文書を、前記複数分類のセットの範囲内で、自動的に分類して自動分類結果を得る第３ステップと、前記分類決定手段が、前記手動分類結果と前記自動分類結果とから前記対象文書に対する分類を最終決定する第４ステップと、前記分類担当取得手段が、対象文書を手動で分類した担当者の情報を取得する第５ステップと、を有し、前記第３ステップは、前記文書ベクトル取得手段が、前記対象文書を特徴づける対象文書ベクトルを取得する第６ステップと、前記典型文書ベクトル取得手段が、前記各分類を特徴づける典型文書の典型文書ベクトルを取得する第７ステップと、前記類似度算出手段が、前記対象文書ベクトルと前記各典型文書ベクトルとの類似度を算出して各分類に対する類似度を得る第８ステップとを有し、前記第８ステップによって得られた各分類に対する類似度を分類結果とし、前記第４ステップは、前記手動分類結果と前記取得した分類担当者に対応する評価値とに基づいて手動分類の点数を算出し、また、前記第８ステップによって得られた各分類に対する類似度に基づいて自動分類の点数を算出し、そして、前記算出された手動分類の点数と自動分類の点数との合計値を分類ごとに算出し、この算出結果に基づいて、前記対象文書に対する分類を最終決定することを特徴とする文書分類方法を提供する。

【０００６】

【発明の実施の形態】

以下、本発明の文書分類装置、文書分類プログラムが記憶された記憶媒体、及び文書分類方法における好適な実施の形態について、図１から図７を参照して説明する。

（１）実施形態の概要

本実施形態による文書分類処理では、過去に行った分類に対する正解率から求めた重み付

け等による評価関数を各分類担当者毎にデータベース化しておくと共に、各分類毎にその分類を特徴づける典型文書を予め用意しておく。
そして、分類担当者（人手）による対象文書の分類結果と評価関数とから、各分類に対する手動分類の点数化を行う。また、対象文書と典型文書との類似度を算出し、この類似度を用いて各分類に対する自動分類の点数化を行う。この両点数を各分類毎に合計した値が最も高い分類を最終分類結果とする。
このように、手動分類と自動分類とを融合化することで、より正確な分類結果を得ることができる。
【０００７】
（２）実施の形態の詳細
本実施形態の文書分類装置は、パーソナルコンピュータやワードプロセッサ等を含むコンピュータシステムで構成するだけでなく、ＬＡＮ（ローカル・エリア・ネットワーク）のサーバ、コンピュータ（パソコン）通信のホスト、インターネット上に接続されたコンピュータシステム等によって構成することも可能である。また、ネットワーク上の各機器に機能分散させ、ネットワーク全体で文書分類装置を構成することも可能である。
【０００８】
図１は、文書分類装置の構成を表したブロック図である。
文書分類装置は、図１に示すようにシステム全体を制御するための制御部１１を備えている。この制御部１１には、データバス等のバスライン２１を介して、入力装置としてのキーボード１２やマウス１３、表示装置１４、印刷装置１５、記憶装置１６、記憶媒体駆動装置１７、通信制御装置１８、入出力Ｉ／Ｆ１９、及び文字認識装置２０が接続されている。
制御部１１は、ＣＰＵ１１１、ＲＯＭ１１２、ＲＡＭ１１３を備えている。
ＲＯＭ１１２は、ＣＰＵ１１１が各種制御や演算を行うための各種プログラムやデータが予め格納されたリードオンリーメモリである。
【０００９】
ＲＡＭ１１３は、ＣＰＵ１１１にワーキングメモリとして使用されるランダムアクセスメモリである。このＲＡＭ１１３には、本実施形態による文書分類処理を行うためのエリアとして、自動分類と手動分類の分類結果を点数化して正規化等の処理を行う分類処理表が格納される分類処理表格納エリア１１３１、分類の対象となる対象文書が格納される対象文書格納エリア１１３２、抽出したキーワードの重要度等を要素値として対象文書を特徴づける対象文書ベクトルが格納される対象文書ベクトル格納エリア、典型文書を特徴づける典型文書ベクトルが格納される典型文書ベクトル格納エリア１１３４、対象文書と各典型文書との類似度が格納される類似度格納エリア１１３５、…、その他の各種エリアが確保されるようになっている。
【００１０】
キーボード１２は、自装置内で対象文書を作成する場合の対象文書取得手段や群類担当者による分類結果を入力する場合の手動分類結果取得手段の一部を構成し、かな文字を入力するためのかなキーやテンキー、各種機能を実行するための機能キー、カーソルキー、等の各種キーが配置されている。
マウス１３は、ポインティングデバイスであり、表示装置１４に表示されたキーやアイコン等を左クリックすることで対応する機能の指定を行う入力装置である。
表示装置１４は、例えばＣＲＴや液晶ディスプレイ等が使用される。この表示装置には、キーボード１２やマウス１３による入力結果が表示されたり、最終分類結果が表示されたりするようになっている。
印刷装置１５は、表示装置１４に表示された文書や、記憶装置１６の文書格納部１６４に格納された文書等の印刷を行うためのものである。この印刷装置としては、レーザプリンタ、ドットプリンタ、インクジェットプリンタ、ページプリンタ、感熱式プリンタ、熱転写式プリンタ、等の各種印刷装置が使用される。
【００１１】

記憶装置１６は、読み書き可能な記憶媒体と、その記憶媒体に対してプログラムやデータ等の各種情報を読み書きするための駆動装置で構成されている。この記憶装置１６に使用される記憶媒体としては、主としてハードディスクが使用されるが、後述の記憶媒体駆動装置１７で使用される各種記憶媒体のうちの読み書き可能な記憶媒体を使用するようにしてもよい。
記憶装置１６は、仮名漢字変換辞書１６１、プログラム格納部１６２、データ格納部１６３、文書データベース１６４、評価関数データベース１６５、文書ベクトルデータベース１６６、図示しないその他の格納部（例えば、この記憶装置１６内に格納されているプログラムやデータ等をバックアップするための格納部）等を有している。
プログラム格納部１６２には、本実施形態における文書分類処理プログラム、文書ベクトル作成処理プログラム等の各種プログラムの他、仮名漢字変換辞書１６１を使用して入力された仮名文字列を漢字混り文に変換する仮名漢字変換プログラム等の各種プログラムが格納されている。
データ格納部１６３には、ユーザに関するデータ等の、システムが必要とする各種データが格納されている。

【００１２】

文書データベース１６４には、各の分類を特徴づける典型文書や、典型文書以外の通常の文書等が格納されている。この文書データベース１６４に格納される各文書の形式は特に限定されるものではなく、テキスト形式の文書、ＨＴＭＬ（Hyper Text Markup Language）形式の文書、ＪＩＳ形式の文書等の各種形式の文書の格納が可能である。
この典型文書により特徴づけられる分類としては、技術動向報告、主張報告、新プロジェクト等の社内用の分類や、政治、経済、健康等の一般的な分類、図書館等弟子用される一般図書や科学技術文献に関する分類、その他各種分類が使用目的によって適宜選択可能になっている。

【００１３】

図２は、評価関数データベース１６５の内容を概念的に表したものである。
この図２に示すように、評価関数は各分類担当者花子、太郎、四郎、…毎に、各分類甲、乙、丙、…に対する、「重み」が評価値として格納されている。
「重み」は各分類に対する分類担当者の正解率（または誤り率）等に基づいて決定される。この「重み」は、各担当者が対象文書に対する分類を決定する毎に、最終分類結果と比較して、変更される。
この図２に示すように、分類担当者花子さんは、分類甲に対しての正解率が低く、分類丙に対する正解率が高いことが理解される。

【００１４】

図３は、文書ベクトルデータベース１６６の内容を概念的に表したものである。
この図３に示されるように、文書Ａｊｋの中から自動抽出されたキーワードｘに対して求められた重要度ｆ（ｘ）が文書ベクトルの要素値ｆ（ｘ）として格納されている。この文書ベクトルは各文書ｊｋ（ｊ＝１～、ｋ＝１～）毎に格納され、文書データベース１６４に格納されている各文書と対応づけられている。
各文書ベクトルの次元は採用するキーワードｘ（重要語句）の数であるが、２文書間の類似度を両文書ベクトルから求める場合には、両文書のキーワードの和集合の数が両文書ベクトルの次元となる。この場合、一方の文書ベクトルにのみ含まれるキーワードに対する他方の文書ベクトルの要素値は、”０”に定義される。

【００１５】

例えば図３おいて、文書Ｂのキーワードは「重要、重要語、重要度、…」、文書Ｃのキーワードは「重要、…、政治、…」であり、両文書の文書ベクトルは次の通りである。
文書Ｂの文書ベクトル＝（　１，１８，１９，…）
文書Ｃの文書ベクトル＝（１８，…，２１，…）
これに対して文書Ｂと文書Ｃとの類似度を算出する場合には、両文書のキーワードを「重要、重要語、重要度、…、政治、…」とし、両文書の文書ベクトルはつぎの通り定義され

る。
文書Ａの文書ベクトル＝（　１，１８，１９，…，　０，…）、
文書Ｃの文書ベクトル＝（１８，　０，　０，…，２１，…）
【００１６】
記憶媒体駆動装置１７（図１）は、ＣＰＵ１１１が外部の記憶媒体からコンピュータプログラムや文書を含むデータ等を読み込むための駆動装置である。記憶媒体に記憶されているコンピュータプログラム等には、本実施形態の文書分類装置により実行される文書分類処理等の各種処理プログラム、および、そこで使用される辞書、データ等も含まれる。
ここで、記憶媒体とは、コンピュータプログラムやデータ等が記憶される記憶媒体をいい、具体的には、フロッピーディスク、ハードディスク、磁気テープ等の磁気記憶媒体、メモリチップやＩＣカード等の半導体記憶媒体、ＣＤ－ＲＯＭやＭＯ、ＰＤ（相変化書換型光ディスク）等の光学的に情報が読み取られる記憶媒体、紙カードや紙テープ等の用紙（および、用紙に相当する機能を持った媒体）を用いた記憶媒体、その他各種方法でコンピュータプログラム等が記憶される記憶媒体が含まれる。
本実施形態の文書分類装置において使用される記憶媒体としては、主として、ＣＤ－ＲＯＭやフロッピーディスク等の記憶媒体が使用される。
記憶媒体駆動装置１７は、これらの各種記憶媒体からコンピュータプログラムを読み込む他に、フロッピーディスクのような書き込み可能な記憶媒体に対してＲＡＭ１１３や記憶装置１６に格納されているデータ等を書き込むことが可能である。
【００１７】
本実施形態の文書分類装置では、制御部１１のＣＰＵ１１１が、記憶媒体駆動装置１７にセットされた外部の記憶媒体からコンピュータプログラムを読み込んで、記憶装置１６の各部に格納（インストール）する。そして、本実施形態による文書分類処理等の各種処理を実行する場合、記憶装置１６から該当プログラムをＲＡＭ１１３に読み込み、実行するようになっている。
但し、記憶装置１６からではなく、記憶媒体駆動装置１７により外部の記憶媒体から直接ＲＡＭ１１３にプログラムを読み込んで実行することも可能である。また、文書分類装置によっては、本実施形態の文書分類処理プログラム等を予めＲＯＭ１１２に記憶させておき、これをＣＰＵ１１１が実行するようにしてもよい。
さらに、本実施形態の文書分類処理プログラム等の各種プログラムやデータを、通信制御装置１８を介して他の記憶媒体からダウンロードし、実行するようにしてもよい。
【００１８】
通信制御装置１８は、文書分類装置と他のパーソナルコンピュータやワードプロセッサ等の各種電子機器との間をネットワーク接続するための制御装置である。
通信制御装置１８は、これら各種電子機器が有している対象文書と同一の言語の文書、入力された他言語の文書、および同一言語や他言語の文書のデータベースを検索対象としてアクセスすることが可能になっている。対象となる文書には、テキスト形式やＨＴＭＬ形式等の各種形式の文書の他、ビットマップデータ等の各種データも含まれる。
入出力Ｉ／Ｆ１９は、音声や音楽等の出力を行うスピーカ等の各種機器を接続するためのインターフェースである。
文字認識装置２０は、用紙等に記載された文字をテキスト形式やＨＴＭＬ等の各種形式で認識する装置であり、イメージスキャナや文字認識プログラム等で構成されている。
【００１９】
本実施形態では、キーボード１２の入力操作により作成した文書（ＲＡＭ１１３の所定格納エリアに格納）の他、外部で作成して所定の記憶媒体に格納した文書で記憶媒体駆動装置１７から読み込んだ文書、予め文書データベースに格納されている文書、通信制御装置１８からダウンロードした文書、及び文字認識装置２０で文字認識した文書、等の各種文書を検索の元になる対象文書として取得する（文書取得手段）ことが可能である。
【００２０】
以上のように構成された本実施形態の文書分類装置による文書分類処理の動作について、

図４を使用して説明する。
図４は文書分類処理のメイン動作を表したフローチャートである。
ＣＰＵ１１１は、まず分類を希望する対象文書Ｔを取得しＲＡＭ１１３の対象文書格納エリア１１３２に格納する（ステップ１１）。
【００２１】
そして、ＣＰＵ１１１は、分類担当者と、その分類担当者によって分類された手動分類結果を取得し、ＲＡＭ１１３の分類処理表格納エリア１１３１の分類処理表に格納する（ステップ１２）。
図６は、ＲＡＭ１１３１の作業領域としてエリアが確保されている自動分類表の内容を概念的に表したものである。
分類担当者花子が対象文書を読んで決定した分類が分類甲であった場合、図６に示すように、花子の分類結果として花子ａ欄６１における、分類甲の点数が１点で他の分類が０点となる。
【００２２】
次にＣＰＵ１１１は、取得した対象文書Ｔの文書ベクトルＢｔが既に作成されていて文書ベクトルデータベース１６６中に格納されているか否かを確認し（ステップ１４）、格納されていれば（；Ｙ）、その文書ベクトルＢｔを読み込んでＲＡＭ１１３の対象文書ベクトル格納エリア１１３３に格納する（ステップ１５）。
対象文書の文書ベクトルＢｔが文書ベクトルデータベース１６６に格納されていない場合（ステップ１４；Ｎ）、ＣＰＵ１１１は、対象文書に対する文書ベクトルＢｔを作成する（ステップ１６）。
【００２３】
図５は、文書ベクトル作成処理の動作を表したフローチャートである。
ＣＰＵ１１１は、形態素解析を行うことで対象文書Ｔから自立語を抽出する（ステップ１３１）と共に、名詞句、複合名詞句等を含めた候補語（句）を対象文書Ｔから抽出しＲＡＭ１１３の所定作業領域に格納する（ステップ１３２）。
そして抽出した候補語（句）の対象文書Ｔでの出現頻度、評価関数から、各候補語（句）重要度ｆ（ｘ）を決定する（ステップ１３３）。ここで、評価関数としては、例えば、所定の重要語が予め指定されている場合にはその重要語に対する重み付け、単語、名詞句、複合名詞句等の候補語（句）の種類による重み付け等が使用される。
さらにＣＰＵ１１１は、決定した重要度ｆ（ｘ）の値から対象文書Ｔのキーワードａ，ｂ，…を決定する（ステップ１３４）。そして、各キーワードの重要度ｆ（ｘ）を要素として、文書ベクトルＢ＝（ｆ（ａ），ｆ（ｂ），…）をＲＡＭ１１３の対象文書ベクトル格納エリア１１３３に格納して（ステップ１３５）、図４の文書分類処理ルーチンにリターンする。
【００２４】
次にＣＰＵ１１１は、対象文書Ｔと分類甲、乙、丙、…の各典型文書との類似度Ｓを算出する（ステップ１７）。
すなわち、ＣＰＵ１１１は、図７に示すように、対象文書の文書ベクトルＢｔと典型文書の文書ベクトルＢｊｋとを比較し、両者ベクトルの角度に依存するコサインにより両文書間の類似度Ｓを算出する。
一般に、文書Ａｘの文書ベクトルＢｘと文書Ａｙの文書ベクトルＢｙとの間の角度をθとし、両文書ベクトルの内積をＢｘ・Ｂｙとし、両文書ベクトルの大きさをそれぞれ｜Ｂｘ｜、｜Ｂｙ｜とした場合、両文書ベクトルの類似度Ｓは次の数式１により求まる。
【００２５】
【数１】
類似度Ｓ＝ＣＯＳ（θ）＝（Ｂｘ・Ｂｙ）／（｜Ｂｘ｜×｜Ｂｙ｜）
【００２６】
この類似度Ｓの値は－１≦Ｓ≦１の値をとり、１に近いほど２つの文書ベクトルが互いに平行に近く、２つの文書Ａｘと文書Ａｙは互いに類似していると考えることができる。

【００２７】
次に、ＣＰＵ１１１は、各分類の典型文書に対して算出した類似度Ｓの合計値が１になるように正規化し、正規化後の類似度を自動分類の点数として分類処理表エリア１１３１の自動ｂ欄６２（図６）に格納する（ステップ１８）。
【００２８】
そして、ＣＰＵ１１１は、手動分類と自動分類による点数に対して評価関数の処理を行う（ステップ１９）。
すなわち、分類担当花子の評価関数のうち、分類甲に対象文書を分類した場合の評価関数（重みｗ＝０．５）を評価関数データベース１６５から読み出し、図６の分類処理表における、花子ａ欄６１の各分類の点数に、乗じて花子ｃ欄６３に格納する。また、自動ｂ欄６２における各分類の点数に（１－ｗ＝０．５）を乗じて、自動ｄ欄６４に格納する。
【００２９】
さらにＣＰＵ１１１は、評価関数処理を行った後の手動分類の各点数（花子ｃ欄６３）と、に評価関数処理後の自動分類の各点数（自動ｄ欄６４）との合計値（ｃ＋ｄ）を各分類毎に求め、合計値が最大となる分類を対象文書Ｔに対する分類として最終決定する（ステップ２０；分類決定手段）。
ＣＰＵ１１１は、最終決定した分類により、分類目的に応じて対象文書を処理し（ステップ２１）、処理を終了する。対象文書の処理の例としては、分類目的が配信であればその分類に属するユーザに対象文書を配信する。
【００３０】
以上説明したように本実施形態によれば、各分類担当者による手動分類の結果にから各分類に対する手動分類の点数化を行い、各分類を特徴づける典型文書の文書ベクトルと対象文書の対象文書ベクトルとの類似度から各分類に対する自動分類の点数化を行うことで、手動分類と自動分類とを融合させることができ、より正確な分類結果を得ることができる。
【００３１】
以上、本実施形態の構成および他言語文書検索の処理について説明したが、本発明では、これらの各形態に限定されるものではなく、請求項に記載された発明の範囲内で種々の変形をすることが可能である。
例えば、典型文書は、必ずしも予め選ばれている必要がなく、文書データベース１６４に格納されてる通常の文書を典型文書として使用してもよい。
また、文書データベース１６３に格納されている文書の中から、クラスタリング処理により自動抽出した文書を典型文書として使用するようにしてもよい。
【００３２】
説明した実施形態では、典型文書とその典型文書ベクトルとがそれぞれ文書データベース１６４、文書ベクトルデータベース１６６に格納されていることを前提に説明したが、必ずしも両者が存在する必要はない。
すなわち、典型文書に対する典型文書ベクトルが存在すれば（文書ベクトルデータベース１６６に格納されていれば）、対象文書Ｔとの類似度Ｓを算出することができるので、典型文書自体は必ずしも必要ではない。
逆に、各分類毎にその分類を特徴づける典型文書が存在すれば（文書データベース１６４に格納されて入れば）、図５に示した文書ベクトル作成処理により、典型文書ベクトルを作成することができるので、典型文書ベクトル自体は必ずしも必要ではない。
【００３３】
また、説明した実施形態では、１分類に対する典型文書の数については特に限定しなかったが、典型文書は必ずしも１分類に１典型文書である必要はなく、１分類に複数の典型文書を用意するようにしてもよい。この場合、各分類に対する対象文書の類似度としては合計値または平均値（正規化処理を行うのでどちらを使用することも可能である。）を使用する。このように１分類複数典型文書とすることで、各をより的確に特徴づけることができ、自動分類側の精度を上げることができる。

【００３４】
また、最終分類結果と分類対象者による分類結果が異なる場合には、評価関数の重み付けを変えることで、学習を行うようにしても良い。で文書分類装置を構成することも可能である。
また、自動分類による分類結果（例えば、ステッ１８による正規化後の類似度の値）に対して、手動分類の場合と同様に、重み付け（自動分類に対する評価関数）を規定するようにしてもよい。そして、この場合の重み付けに対しても、学習により変更するようにしてもよい。
【００３５】
さらに、説明した実施形態では、対象文書の言語については特に言及しなかったが、本発明では日本語に限定されるものではなく、あらゆる言語の対象文書に適用することが可能である。この場合、対象文書の言語用の形態素解析アルゴリズム等を使用するといった、本発明の構成には影響のない部分を変更するだけでよい。
但し、典型文書の言語は対象文書の言語と同一である必要がある。
【００３６】
以上の実施形態において説明した、各装置、各部、各動作、各処理等に対しては、それらを含む上位概念としての各手段（～手段）により、実施形態を構成することが可能である。
例えば、「ＣＰＵ１１１は、図７に示すように、対象文書の文書ベクトルＢｔと典型文書の文書ベクトルＢｊｋとを比較し、両者ベクトルの角度に依存するコサインにより両文書間の類似度Ｓを算出する。」との記載に対して「類似度算出手段」を構成するようにしてもよい。
同様に、その他各種動作に対して「～（動作）手段」等の上位概念で実施形態を構成するようにしてもよい。
例えば、以下のように構成するようにしてもよい。
（１）図８に示すように、予め決められた複数分類のセットの範囲内で、人手によって対象文書を分類した手動分類結果を取得する手動分類結果取得手段と、前記対象文書を取得する対象文書取得手段と、前記対象文書取得手段で取得された対象文書を、前記複数分類のセットの範囲内で、自動的に分類して自動分類結果を得る自動分類結果取得手段と、前記手動分類結果と前記自動分類結果とから前記対象文書に対する分類を最終決定する分類決定手段と、を文書分類装置に具備させる。
（２）図９に示すように、上記（１）に記載した文書分類装置において、前記自動分類結果取得手段は、前記対象文書を特徴づける対象文書ベクトルを取得する文書ベクトル取得手段と、前記各分類を特徴づける典型文書の典型文書ベクトルを取得する典型文書ベクトル取得手段と、前記対象文書ベクトルと前記各典型文書ベクトルとの類似度を算出して各分類に対する類似度を得る類似度算出手段とを有し、前記類似度算出手段によって得られた各分類に対する類似度を分類結果とする。
（３）図１０に示すように、予め決められた複数分類のセットの範囲内で、人手によって対象文書を分類した手動分類結果を取得する手動分類結果取得機能と、前記対象文書を取得する対象文書取得機能と、前記対象文書取得機能で取得された対象文書を、前記複数分類のセットの範囲内で、自動的に分類して自動分類結果を得る自動分類結果取得機能と、前記手動分類結果と前記自動分類結果とから前記対象文書に対する分類を最終決定する分類決定機能と、をコンピュータに実現させるためのコンピュータ読取り可能な文書分類プログラムを記憶媒体に記憶させる。
（４）図１１に示すように、前記自動分類結果取得機能は、前記対象文書を特徴づける対象文書ベクトルを取得する文書ベクトル取得機能と、前記各分類を特徴づける典型文書の典型文書ベクトルを取得する典型文書ベクトル取得機能と、前記対象文書ベクトルと前記各典型文書ベクトルとの類似度を算出して各分類に対する類似度を得る類似度算出機能とを有し、前記類似度算出機能によって得られた各分類に対する類似度を分類結果とする。
（５）図１２に示すように、予め決められた複数分類のセットの範囲内で対象文書を自動

的に分類し、この自動分類結果と、前記複数分類のセットの範囲内で、人手によって前記対象文書を分類した手動分類結果とから前記対象文書に対する分類を最終決定する。

【００３７】

【発明の効果】

本発明によれば、同一の複数分類のセットの範囲内で、手動分類と自動分類を行うと共に、両分類結果を使用して対象文書に対する最終分類を決定するようにしたので、手動分類と自動分類の両分類結果を使用して、対象文書に足してより精度の高い分類を行うことができる。

【図面の簡単な説明】

【図１】本発明の１実施形態における文書分類装置の構成を表したブロック図である。

【図２】同上、実施形態における評価関数データベースの内容を概念的に表した説明図である。

【図３】同上、実施形態における文書ベクトルデータベースの内容を概念的に表した説明図である。

【図４】同上、実施形態における文書分類処理のメイン動作を表したフローチャートである。

【図５】同上、実施形態の文書分類処理における文書ベクトル作成処理の動作を表したフローチャートである。

【図６】同上、実施形態において分類の最終決定までの分類処理表での処理を表した説明図である。

【図７】同上、実施形態においける対象文書と典型文書との類似関係を文書ベクトルを用いて表した説明図である。

【図８】請求項１に記載した発明のクレーム対応図である。

【図９】請求項２に記載した発明のクレーム対応図である。

【図１０】請求項３に記載した発明のクレーム対応図である。

【図１１】請求項４に記載した発明のクレーム対応図である。

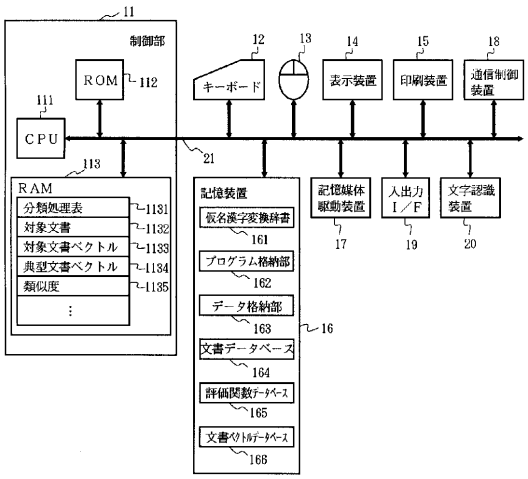
【図１２】請求項５に記載した発明のクレーム対応図である。

【符号の説明】

１１　制御部
１１２　ＲＯＭ
１１３　ＲＡＭ
１１３１　分類処理表
１１３２　対象文書格納エリア
１１３３　対象文書ベクトル格納エリア
１１３４　典型文書ベクトル格納エリア
１１３５　類似度格納エリア
１２　キーボード
１３　マウス
１４　表示装置
１５　印刷装置
１６　記憶装置
１６１　仮名漢字変換辞書
１６２　プログラム格納部
１６３　データ格納部
１６４　文書データベース
１６５　評価関数データベース
１６６　文書ベクトルデータベース
１７　記憶媒体駆動装置
１８　通信制御装置
１９　入出力Ｉ／Ｆ

２０　文字認識装置

【図１】



【図２】

評価関数データベース；165

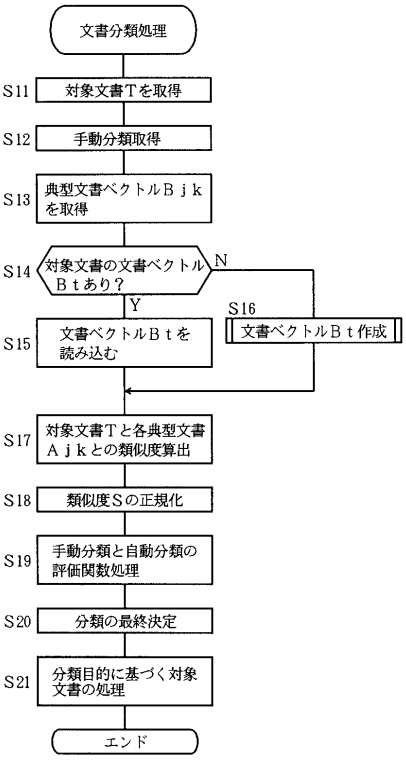
分　類	花　子	太　郎	四　郎	・・・
甲	0.5	0.8	0.9	・・・
乙	0.9	0.7	0.6	・・・
丙	0.9	0.9	0.8	・・・
・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	

【図３】

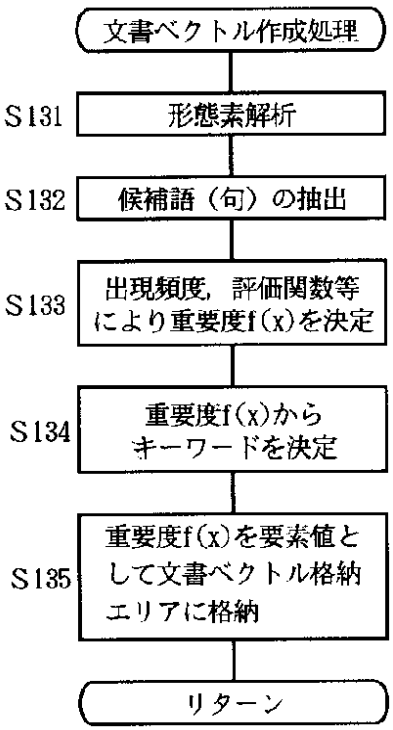
文書ベクトルデータベース；166

文書	キーワードの要素値ｆ（ｘ）						
	……	重要	重要語	重要度	……	政治	……
A	……	2	20	21	……	—	……
B	……	1	18	19	……	—	……
C	……	18	—	—	……	21	……
⋮							

【図４】



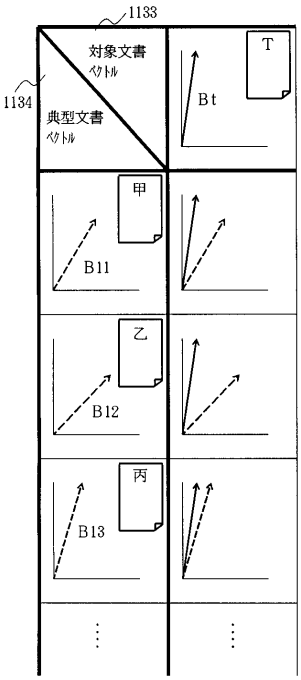
【図５】



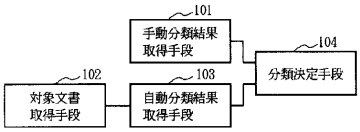
【図６】

分類	花子a (61)	自動b (62)	花子c a×w (63)	自動d b×(1-w) (64)	最終評価 c+d (65)
甲	1	0.2	0.5	0.1	0.6
乙	0	0.1	0	0.05	0.05
丙	0	0.7	0	0.35	0.35
⋮	⋮	⋮	⋮	⋮	⋮

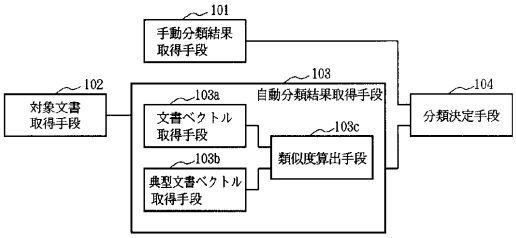
【図７】



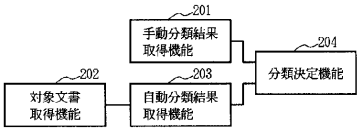
【図８】



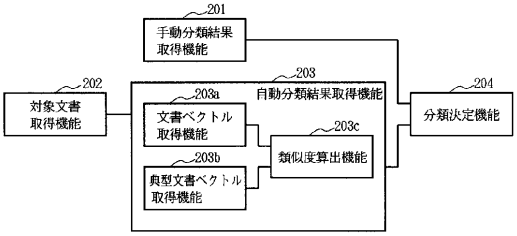
【図９】



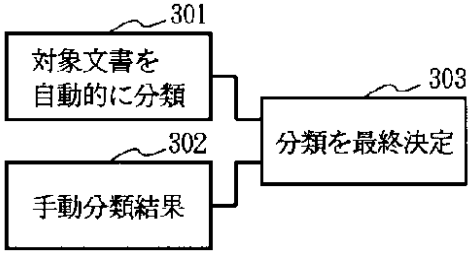
【図１０】



【図１１】



【図１２】



フロントページの続き

(56)参考文献 特開平０９－０２２４１４（ＪＰ，Ａ）
特開平７－２８２０７８（ＪＰ，Ａ）
特開平９－２６９６３（ＪＰ，Ａ）
田中 栄治，情報探索支援システムの構築（１），電子情報通信学会技術研究報告，日本，社団法人 電子情報通信学会，１９９６年１２月１４日，第96巻 第431号，55～62

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/30