(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4496347号
(P4496347)
(45)発行日　平成22年7月7日(2010.7.7)　(24)登録日　平成22年4月23日(2010.4.23)

(51)Int.Cl.　　F I
G06F 17/28 (2006.01)　　G06F 17/28　U
G06F 17/30 (2006.01)　　G06F 17/30　２１０Ｄ

請求項の数 7　(全 27 頁)

(21)出願番号　特願2003-418871(P2003-418871)
(22)出願日　平成15年12月17日(2003.12.17)
(65)公開番号　特開2005-182218(P2005-182218A)
(43)公開日　平成17年7月7日(2005.7.7)
審査請求日　平成15年12月17日(2003.12.17)
審判番号　不服2007-4688(P2007-4688/J1)
審判請求日　平成19年2月15日(2007.2.15)

特許法第３０条第１項適用　平成１５年９月２１日平成１５年度電気関係学会北陸支部連合大会実行委員会事務局発行の「平成１５年度電気関係学会北陸支部連合大会講演論文集」に発表

(73)特許権者　591040236
石川県
石川県金沢市鞍月１丁目１番地
(74)代理人　100111822
弁理士　渡部　章彦
(72)発明者　上田　芳弘
石川県金沢市鞍月２丁目１番地　石川県工業試験場内
(72)発明者　加藤　直孝
石川県金沢市鞍月２丁目１番地　石川県工業試験場内
(72)発明者　林　克明
石川県金沢市鞍月２丁目１番地　石川県工業試験場内

最終頁に続く

(54)【発明の名称】文書分類装置及びそのプログラム

(57)【特許請求の範囲】
【請求項１】
入力された文書に出現する単語を用いて、複数のカテゴリの各々について設けられる複数のカテゴリ辞書からなり、各々のカテゴリ辞書が、当該カテゴリにおいて出現した単語毎に、文書中に現れた単語の頻度を表すパラメータである単語の重要性を示すtf値と、単語がどの程度当該カテゴリに現れるかを表すパラメータである単語の特定性を示すidf 値とに基づいて、単語の重要性と特定性とを加味した重みを表すパラメータである単語単独の重要度を示すtf・idf 値を算出することにより作成された、当該カテゴリにおいて出現した単語毎に単語単独の重要度を表すtf・idf 値を格納する重要性辞書と、当該カテゴリにおいて出現した単語についての複数の単語の組み合わせ毎に、各々の単語の前記idf 値と、単語の共起性を示す確信度を表すパラメータであるconf値とに基づいて、特定性の高い単語間の共起性を示すidf/conf値を算出することにより作成された、当該カテゴリにおいて出現した単語についての第１単語と第２単語の組み合わせ毎に２単語間の共起性を表すidf/conf値を格納する同時出現性辞書とからなる辞書で、当該単語を照合して、前記辞書の単語毎のtf・idf 値とidf/conf値とを求める手段と、
前記辞書の単語毎のtf・idf 値とidf/conf値と入力された文書に出現する単語との一致率との積を算出して前記単語毎のスコアを算出する手段と、
前記単語毎のスコアに基づいて前記カテゴリ毎のスコアを算出する手段と、
これに基づいて前記入力された文書を前記複数のカテゴリのいずれかに分類する手段とを備える

ことを特徴とする文書分類装置。
【請求項２】
前記文書は電子メールであり、前記カテゴリは文書を分配すべき担当者であり、前記文書分類装置は前記入力された文書を当該分類された担当者に配信する
ことを特徴とする請求項１記載の文書分類装置。
【請求項３】
前記分類する手段は、前記カテゴリ毎のスコアに基づいて、前記入力された文書を、当該スコアが上位から所定の数のカテゴリに分類する
ことを特徴とする請求項１記載の文書分類装置。
【請求項４】
当該文書分類装置が、更に、
所定の文書を前記文書分類装置により前記複数のカテゴリのいずれかに分類した結果に基づいて、前記重要性辞書におけるtf・idf 値を更新し、前記同時出現性辞書におけるidf/conf値を更新する学習装置とからなる
ことを特徴とする請求項１記載の文書分類装置。
【請求項５】
前記学習装置は、
前記所定の文書が真のカテゴリに分類された場合、前記所定の文書に出現する単語であって当該真のカテゴリの辞書で照合された単語を重要語として、その重みを大きくする手段と、
前記所定の文書が真のカテゴリに分類されなかった場合、前記所定の文書に出現する単語の中で当該真のカテゴリについて特定性の高い単語を再評価することにより、当該真のカテゴリの特定語の重みを大きくする手段と、
前記所定の文書が真のカテゴリ以外のカテゴリに分類された場合、前記所定の文書に出現する単語であって当該誤って分類されたカテゴリの辞書で照合した単語を不要語として、当該誤って分類されたカテゴリの辞書における重みを小さくし、真のカテゴリ以外のカテゴリの辞書における重みを小さくする手段とを備える
ことを特徴とする請求項４記載の文書分類装置。
【請求項６】
前記重要性辞書は、属するカテゴリの判っている文書に基づいて辞書を作成する辞書編集装置に設けられた手段であって、当該カテゴリにおいて出現した単語毎に、文書中に現れた単語の頻度を表すパラメータである単語の重要性を示すtf値を算出し、単語がどの程度当該カテゴリに現れるかを表すパラメータである単語の特定性を示すidf 値を算出し、これらに基づいて、単語の重要性と特定性とを加味した重みを表すパラメータである単語単独の重要度を示すtf・idf 値を算出することにより、当該カテゴリにおいて出現した単語毎にtf・idf 値を格納する重要性辞書を作成する手段により作成され、
前記同時出現性辞書は、前記辞書編集装置に設けられた手段であって、当該カテゴリにおいて出現した単語についての複数の単語の組み合わせ毎に、各々の単語の前記idf 値を算出し、単語の共起性を示す確信度を表すパラメータであるconf値を算出し、これらに基づいて、特定性の高い単語間の共起性を示すidf/conf値を算出することにより、当該カテゴリにおいて出現した単語についての複数の単語の組み合わせ毎に当該単語間の共起性を表すidf/conf値を格納する同時出現性辞書を作成する手段により作成される
ことを特徴とする請求項１記載の文書分類装置
【請求項７】
文書分類装置を実現するプログラムであって、
前記プログラムは、
入力された文書に出現する単語を用いて、複数のカテゴリの各々について設けられる複数のカテゴリ辞書からなり、各々のカテゴリ辞書が、当該カテゴリにおいて出現した単語毎に、文書中に現れた単語の頻度を表すパラメータである単語の重要性を示すtf値と、単語がどの程度当該カテゴリに現れるかを表すパラメータである単語の特定性を示すidf 値

とに基づいて、単語の重要性と特定性とを加味した重みを表すパラメータである単語単独の重要度を示すtf・idf 値を算出することにより作成された、当該カテゴリにおいて出現した単語毎に単語単独の重要度を表すtf・idf 値を格納する重要性辞書と、当該カテゴリにおいて出現した単語についての複数の単語の組み合わせ毎に、各々の単語の前記idf 値と、単語の共起性を示す確信度を表すパラメータであるconf値とに基づいて、特定性の高い単語間の共起性を示すidf/conf値を算出することにより作成された、当該カテゴリにおいて出現した単語についての第１単語と第２単語の組み合わせ毎に２単語間の共起性を表すidf/conf値を格納する同時出現性辞書とからなる辞書で、当該単語を照合して、前記辞書の単語毎のtf・idf 値とidf/conf値とを求める処理手段と、

前記辞書の単語毎のtf・idf 値とidf/conf値と入力された文書に出現する単語との一致率との積を算出して前記単語毎のスコアを算出する処理手段と、

前記単語毎のスコアに基づいて前記カテゴリ毎のスコアを算出する処理手段と、

前記カテゴリ毎のスコアに基づいて前記入力された文書を前記複数のカテゴリのいずれかに分類する処理手段、

の各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である

ことを特徴とする文書分類プログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、辞書編集装置、文書分類装置及びそのプログラムに関し、特に、電子メールやＷｅｂ等で入力した電子文書を、適切な担当者に自動的に正確に分類するための辞書編集装置、文書分類装置及びそのプログラムに関する。

【背景技術】

【０００２】

最近、企業のコールセンター等では、電話やＦＡＸに依らず、電子メールやＷｅｂで顧客からの問い合わせに対応することへのニーズが高まっている。しかし、電子メール等による問い合わせ（問い合わせメール）に少人数で対応するためには、適切な担当者へ問い合わせメールを分類し、回答しなければならない場合が多い。電子メールやＷｅｂにより充実したサービスを提供するためには、このような分類業務の効率化が重要であり、これを自動分類するシステムが強く望まれている。

【０００３】

このような電子メールの自動分類の手段としては、電子メールを分類するための類推ルールをサンプルデータから学習して生成し、これに基づいて電子メールを自動分類できるようにした装置が知られている（例えば、特許文献１参照）。

【０００４】

一方、文書分類に関する多くの研究において、単語単独の重要度であるtf・idf 値(term frequency times inverse document frequency)が応用されている。更に、本発明者等により、tf・idf 値の他に、２単語間の共起性を表すidf/conf値を用いて、電子メール等の文書分類を行うことが提案されている（非特許文献１参照）。

【特許文献１】特開２００１－２５６２５１号公報

【非特許文献１】成田他、（Ｆ－６７）テキストマイニングと強化学習による電子メール自動分配、平成１４年度電気関係学会北陸支部連合大会

【発明の開示】

【発明が解決しようとする課題】

【０００５】

前述のように、本発明者等は、先にtf・idf 値及びidf/conf値を用いて電子メール等の文書分類を行うことを提案した。しかし、その後の本発明者の検討によれば、tf・idf 値とidf/conf値を用いただけの文書分類では、電子メールの分類等に用いた場合、実用的な分類精度が得られないことが判った。即ち、tf・idf 値及びidf/conf値を併せて１個のパラメータとして用いて１個の辞書を作成及び使用したのでは実用に耐え得る分類結果が得

られず、一方、tf・idf 値とidf/conf値とを別々の独立した２個のパラメータとして用いて２個の辞書を作成及び使用すると、実用的な分類の実現に有効であることが判った。更に、辞書の作成後も単語の重み（ウェイト）を学習することが有効であるが、別々の独立した２個のパラメータであるtf・idf 値とidf/conf値とについて個々に学習すると、実用的な分類の実現に有効であることが判った。そして、この学習の課程において、３種類の学習（重要語、特定語、不要語）を行うことが有効であることが判った。

【０００６】

本発明は、文書を適切な担当者に自動的に正確に分類するための辞書を作成する辞書編集装置を提供することを目的とする。

【０００７】

また、本発明は、文書を適切な担当者に自動的に正確に分類する文書分類装置を提供することを目的とする。

【０００８】

また、本発明は、文書を適切な担当者に自動的に正確に分類する文書分類プログラムを提供することを目的とする。

【課題を解決するための手段】

【００１０】

本発明の文書分類装置は、入力された文書に出現する単語を用いて、複数のカテゴリの各々について設けられる複数のカテゴリ辞書からなり、各々のカテゴリ辞書が、当該カテゴリにおいて出現した単語毎に、文書中に現れた単語の頻度を表すパラメータである単語の重要性を示すtf値と、単語がどの程度当該カテゴリに現れるかを表すパラメータである単語の特定性を示すidf 値とに基づいて、単語の重要性と特定性とを加味した重みを表すパラメータである単語単独の重要度を示すtf・idf 値を算出することにより作成された、当該カテゴリにおいて出現した単語毎に単語単独の重要度を表すtf・idf 値を格納する重要性辞書と、当該カテゴリにおいて出現した単語についての複数の単語の組み合わせ毎に、各々の単語の前記idf 値と、単語の共起性を示す確信度を表すパラメータであるconf値とに基づいて、特定性の高い単語間の共起性を示すidf/conf値を算出することにより作成された、当該カテゴリにおいて出現した単語についての第１単語と第２単語の組み合わせ毎に２単語間の共起性を表すidf/conf値を格納する同時出現性辞書とからなる辞書で、当該単語を照合して、前記辞書の単語毎のtf・idf 値とidf/conf値とを求める手段と、前記辞書の単語毎のtf・idf 値とidf/conf値と入力された文書に出現する単語との一致率との積を算出して前記単語毎のスコアを算出する手段と、前記単語毎のスコアに基づいて前記カテゴリ毎のスコアを算出する手段と、これに基づいて前記入力された文書を前記複数のカテゴリのいずれかに分類する手段とを備える。

【００１１】

本発明のプログラムは、文書分類装置を実現するプログラムであって、前記プログラムは、入力された文書に出現する単語を用いて、複数のカテゴリの各々について設けられる複数のカテゴリ辞書からなり、各々のカテゴリ辞書が、当該カテゴリにおいて出現した単語毎に、文書中に現れた単語の頻度を表すパラメータである単語の重要性を示すtf値と、単語がどの程度当該カテゴリに現れるかを表すパラメータである単語の特定性を示すidf 値とに基づいて、単語の重要性と特定性とを加味した重みを表すパラメータである単語単独の重要度を示すtf・idf 値を算出することにより作成された、当該カテゴリにおいて出現した単語毎に単語単独の重要度を表すtf・idf 値を格納する重要性辞書と、当該カテゴリにおいて出現した単語についての複数の単語の組み合わせ毎に、各々の単語の前記idf 値と、単語の共起性を示す確信度を表すパラメータであるconf値とに基づいて、特定性の高い単語間の共起性を示すidf/conf値を算出することにより作成された、当該カテゴリにおいて出現した単語についての第１単語と第２単語の組み合わせ毎に２単語間の共起性を表すidf/conf値を格納する同時出現性辞書とからなる辞書で、当該単語を照合して、前記辞書の単語毎のtf・idf 値とidf/conf値とを求める処理手段と、前記辞書の単語毎のtf・idf 値とidf/conf値と入力された文書に出現する単語との一致率との積を算出して前記単

語毎のスコアを算出する処理手段と、前記単語毎のスコアに基づいて前記カテゴリ毎のスコアを算出する処理手段と、前記カテゴリ毎のスコアに基づいて前記入力された文書を前記複数のカテゴリのいずれかに分類する処理手段、の各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である。

【発明の効果】

【００１２】

本発明の辞書編集装置によれば、単語単独の重要度を表すtf・idf 値と複数の単語間の共起性を表すidf/conf値とを別々の独立した２個のパラメータとして用いて２個の辞書（tf・idf辞書、idf/conf辞書）を作成することができるので、これを文書分類装置の辞書として用いることにより、電子メール等の文書の分類に用いた場合、実用的な分類精度を得ることができる。従って、文書を適切な担当者に自動的に正確に分類するための辞書を容易に作成することができる。

【００１３】

本発明の文書分類装置によれば、前述の２個の辞書tf・idf辞書、idf/conf辞書を用いることにより、入力した文書を基本的には単語単独の重要度を表すtf・idf 値と複数（２つ）の単語間の共起性を表すidf/conf値とに基づいてカテゴリに分類できるので、電子メール等の文書の分類に用いた場合、実用的な分類精度を得ることができる。従って、文書を適切な担当者に自動的に正確に分類することができる。

【００１４】

本発明の文書分類プログラムによれば、これをフレキシブルディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ／Ｗ、ＤＶＤ等の媒体に格納すること、又は、インターネット等のネットワークを介してダウンロードすることにより供給することができ、これにより前述の文書分類システムを容易に実現することができ、正確な文書分類を可能とすることができる。

【発明を実施するための最良の形態】

【００１５】

図１は、文書分類システム構成図であり、本発明の文書分類システムの構成を示す。文書分類システムは、３個のサブシステム、即ち、辞書編集装置１、文書分類装置２、強化学習装置３からなる。これらの間は、例えばＬＡＮ（Local Area Network）により相互に接続される。この例では、文書分類装置２は、例えばメール解析装置２からなる。従って、以下の例においては、分類（解析）対象である前記入力された文書は例えば電子メール８であり、前記カテゴリは文書を分配すべき担当者（回答者）であり、前記文書分類装置２は入力された文書を当該分類された担当者に配信（分配）する。

【００１６】

辞書編集装置１は、文書ファイル６に基づいて辞書を作成する辞書編集装置であって、当該担当者において出現した単語毎に、単語単独の重要度を表すtf・idf 値を格納する重要性辞書（tf・idf辞書）４１を作成し、また、当該担当者において出現した単語についての複数の単語の組み合わせ毎に、当該単語間の共起性を表すidf/conf値を格納する同時出現性辞書（idf/conf辞書）４２を作成する。文書ファイル６は、各担当者が日常業務で作成した文書である（文書を格納している）ので、属するカテゴリ即ちその担当者の判っている文書である。従って、当該担当者において出現した単語とは、当該担当者の作成した文書ファイル６に出現した単語である。この例では、同時出現性辞書４２は、当該担当者において出現した単語についての第１単語と第２単語の組み合わせ毎に２単語間の共起性を表すidf/conf値を格納する。従って、辞書４は、複数の担当者の各々について設けられる複数の担当者（の）辞書４０からなり、各々の担当者辞書４０が重要性辞書４１と同時出現性辞書４２とからなる。

【００１７】

辞書編集装置１は、実際には、文書ファイル（担当者毎のフォルダ）６に保存された各種の文書を変換して得たテキストファイルを読み込み、読み込んだテキストデータから改行とスペースを取り除いた後に、周知の形態素解析処理により単語に分割することにより抽出した単語を、担当者毎の辞書４０に登録する。抽出した単語において、品詞が助詞、

助動詞、接続詞、接頭詞、副詞、連体詞、感動詞、記号は不要語と考えられるので、辞書４には登録しない。

【００１８】

即ち、本発明では、辞書４の作成及び編集において、単語単独の重要度であるtf・idf値を用い、更に、単語間の共起性をidf/conf値（idf divided by confidence）をも用いる。ここで、tf値は文書中に現れた単語の頻度（即ち、単語の重要性）を示し、idf 値は単語がどの程度特定の分類に現れるか（即ち、単語の特定性）を示し、tf・idf 値は文書分類の研究で用いられているものであり単語の重要性と特定性とを加味した重み（即ち、単語単独の重要度）を示し、conf値は単語の共起性を示す確信度を示し、idf/conf値は特定性の高い単語間の共起性を示す。

【００１９】

図２（Ａ）はtf・idf 値の辞書（重要性辞書）４１の一例を示す。重要性辞書４１は、その担当者が使用した単語ａ１、ａ２、ａ３、・・毎に、そのtf値、idf 値、tf・idf 値を格納する。tf値、idf 値、tf・idf 値の算出については後述する。図２（Ｂ）はidf/conf値の辞書（同時出現性辞書）４２の一例を示す。同時出現性辞書４２は、その担当者が使用した単語ｂ１、ｂ２、ｂ３、・・における第１単語と第２単語の組み合わせ毎に、当該第１単語のidf 値、第１単語と第２単語間のconf値、第１単語と第２単語間のidf/conf値を格納する。前述のように、この例におけるidf/conf値は、当該２単語間の特定性を考慮した共起性を示す。conf値、idf/conf値の算出については後述する。

【００２０】

メール解析装置２は、入力された文書である問い合わせの電子メール（問い合わせメール）８に出現する単語を用いて辞書４で当該単語を照合して、辞書４の単語毎のtf・idf値とidf/conf値とを求め、これらに基づいて所定の演算を行って単語毎のスコアを算出し、単語毎のスコアに基づいて担当者毎のスコアを算出し、これに基づいて入力された文書を複数の担当者のいずれかに分類する。この例では、メール解析装置２は、前記所定の演算として、辞書４の単語毎のtf・idf 値とidf/conf値と入力された文書に出現する単語との一致率との積を算出して、単語毎のスコアを算出する。また、メール解析装置２は、担当者毎のスコアに基づいて、入力された文書である問い合わせメール（以下、単にメール）８を、当該スコアが上位から所定の数の担当者に分類する。

【００２１】

メール解析装置２は、辞書編集装置１と同様に、入力されたメール８の本文について、改行とスペースを取り除いた後に周知の形態素解析処理を行って、単語に分割することにより得られた単語と辞書４とを照合する。この時、メール解析装置２は、マッチした（照合できた）単語又は単語の組み合わせに関するtf・idf 値とidf/conf値と単語の一致率とから、当該メール８に対する全ての担当者のスコアを算出して、最後にスコアの高い一人又は複数の担当者を当該メール８への回答者（当該メール８に回答すべき担当者）として推定し、図３に示すように、その結果を例えばメール解析装置２に表示する。即ち、質問者からのメール８を入力として、これに回答すべき担当者の候補（即ち、図３）を出力として得る。また、この例では、メール解析装置２は、この分類の結果（スコア）に従って、図３において「○」の付された担当者の担当者（回答者）端末５に、当該入力されたメール８を分配する（配信する）。

【００２２】

図３は、メール解析結果の一例を示す。当該結果は、担当者ｄ毎に、回答者とするか否かの結果、スコアScore(d)、メール8と辞書とで照合できた単語t、当該単語毎のスコア(d, t)からなる。担当者ｄ及び照合単語ｔはスコアの高い順に表示される。最もスコアの高いもの（第１(i)位）を第１(i)位照合単語という。図３において、「○」の付された担当者が回答者とされる。この例では上位２人が回答者として推定される。入力された（受信された）メール８は、回答者とされた担当者にメール８として分類（送信）される。一方、「×」の付された担当者は回答者と推定されない。人間の分類者を介在させる場合や分類を受けた担当者が他の担当者に再分類する場合は、図３に示す推定結果を参照して分類

を行うようにすれば良い。

【００２３】

なお、図示はしないが、メール解析装置２は、例えばＬＡＮを介して回答者端末５に接続され、また、メールサーバに接続される。メールサーバは、インターネットに接続され、外部からのメール８を受信し、これを内部の各端末５に配信するために、メール解析装置２に送る。メール８は、例えば質問者が周知のＷｅｂブラウザを用いて、問い合わせ内容、氏名、返信用電子メールアドレス等を入力することにより作成（入力）する。

【００２４】

本発明では、２個の辞書４１、４２に基づいて得た２個の重みをそのまま用いるのではなく、マッチした（照合できた）単語の一致率をも考慮してスコアを算出する。即ち、処理対象であるメール８に出現する単語と辞書の単語との照合を、完全一致ではなく、部分一致で行うと共に、当該部分一致の割合（単語の一致率）をスコアの算出に用いる。これにより、同一の用語についての僅少な個人差の影響等を排除することができる。

【００２５】

強化学習装置３は、所定の文書（即ち、別文書ファイル７）をメール解析装置２により複数の担当者のいずれかに分類した結果に基づいて、一旦作成した重要性辞書４１におけるtf・idf 値を更新し、同時出現性辞書４２におけるidf/conf値を更新する（強化する）。即ち、強化学習装置３は、重要語学習、特定語学習、不要語学習を行う。重要語学習処理は、所定の文書が真の担当者に分類された場合、前記所定の文書に出現する単語であって当該真の担当者の辞書４０で照合された単語を重要語として、その重みを大きくする。特定語学習の処理は、所定の文書が真の担当者に分類されなかった場合、前記所定の文書に出現する単語の中で当該真の担当者について特定性の高い単語を再評価することにより、当該真の担当者の特定語の重みを大きくする。不要語学習は、所定の文書が真の担当者以外の担当者に分類された場合、前記所定の文書に出現する単語であって当該誤って分類された担当者（真の担当者以外の担当者（カテゴリ））の辞書４０で照合した単語を不要語として、当該誤って分類された担当者の辞書４０における重みを小さくし、真の担当者以外の担当者の辞書４０における重みを小さくする。これにより、メール８の分類精度を更に向上することができる。

【００２６】

本発明では、一旦作成した辞書４におけるtf・idf 値とidf/conf値を強化するために、別文書ファイル７、即ち、分類を受ける各担当者自身が様々な業務の中で作成した文書を読み込んだファイルを用いる。別文書ファイル７は、文書ファイル６及びメール８とは別の文書（文書ファイル）であって、各担当者が作成した別文書である（別文書を格納する）。即ち、別文書ファイル７は、担当者が随時作成する文書を随時取り込んで格納したものである。別文書ファイル７は、文書ファイル６と同様にして、メール解析装置２により複数の担当者のいずれかに分類される。別文書ファイル７の担当者も明確であるのでこの結果は正確に評価することができ、また、担当者の業務内容の変化に応じて随時tf・idf 値とidf/conf値を強化することができ、メール８の正確な分類に有効である。

【００２７】

また、本発明では、重要語学習、特定語学習、不要語学習即ち、tf・idf 値とidf/conf値の強化を、プロフィットシェアリング（以下ＰＳと言う）により行う。即ち、別文書ファイル７についてその担当者（分類先）を推定し、その推定結果の適切さに応じた報酬（プロフィット）によりtf・idf 値とidf/conf値を補正する。ＰＳは強化学習において注目されている。これにより、分類の専門家と同等レベルの分類精度を得ることができる。ＰＳは、周知のように、報酬を得たときに複数ルール（本発明では複数単語の重み）を一括して強化するので、効率的にメール８の分類精度を向上することができる。なお、このように強化学習を文書分類に取り入れたシステムは、これまでに開発されていない。

【００２８】

以上のように、本発明によれば、２個の辞書４１、４２に基づいて得た２個の重みを用いてメール８等の文書を分類することにより、実用的な分類を実現することができる。即

ち、日常業務としてメール８を分類している専門家の分類結果（分類精度）が、実用上必要な精度と考えることができる。本発明のメール解析装置２による分類精度として、当該専門家の分類精度とほぼ同等の精度を得ることができる。従って、本発明のメール解析装置２（による分類）は十分に実用に耐え得るものである。なお、本発明者の検討によれば、実際は、２個の辞書４１、４２のみを用いた分類精度は当該専門家の精度を少し下回るが、tf・idf 値とidf/conf値を別文書ファイル７を用いて強化学習することにより、分類の専門家と同等な精度でメール８を分類することができる。

【００２９】

以下、処理フローを参照して、本発明の文書分類システムにおける処理について、詳細に説明する。図４は、文書分類処理フローであり、本発明の図１に示す文書分類システムにおける文書分類処理を示す。辞書編集装置１が、各担当者が作成した文書ファイル６を収集して、この中の出現単語の重みを、単語単独の重要度を表すtf・idf 値と２単語間の共起性を表すidf/conf値として算出し、この２種類の辞書４１、４２を担当者毎に作成する（ステップS101）。強化学習装置３が、ＰＳを応用して、これらの重み又はウェイト（を示すtf・idf 値及びidf/conf値）を強化学習する（ステップS102）。即ち、メール解析装置２が、メール８の受信の有無を判断し、当該受信がない場合、別文書ファイル７の文書と２種類の辞書４１、４２を照合して、単語の重みと単語の一致率とから、担当者毎にスコアを算出し、このスコアが高い担当者を当該メール８への回答者として推定し、当該推定の結果を強化学習装置３に入力してtf・idf 値及びidf/conf値の重みを更新する。一方、メール８を受信すると、メール解析装置２は当該メール８を解析する（ステップS103）。即ち、メール８と２種類の辞書４１、４２を照合して、単語の重みと単語の一致率とから、担当者毎にスコアを算出する。そして、メール解析装置２は、このスコアが高い担当者を当該メール８への回答者として推定し、これをメール解析装置２の端末（図示せず）に表示する（ステップS104）。

【００３０】

図５は、辞書編集処理フローであり、図１の辞書編集装置１がステップS101において実行する辞書編集処理を示す。辞書編集装置１が、全担当者の全文書ファイル６を読み込み（ステップS111）、全文書について前処理を行う（ステップS112）。即ち、全文書から、その改行とスペースとを除去し、残りの部分について周知の形態素解析を行い、分かち書きした単語を得る。そして、当該得られた全単語から不要な品詞の単語を削除する。次に、辞書編集装置１が、図２（Ｃ）に示す担当者・単語テーブルを生成し（ステップS113）、これに基づいてtf・idf 値を算出して（ステップS114）、当該単語及びtf・idf 値をtf・idf 辞書（重要性辞書）４１に書き込む（ステップS115）。これにより、tf・idf 辞書４１が作成される。担当者・単語テーブル及びその生成については、図６を参照して後述する。次に、辞書編集装置１が、図８に示す単語の組み合わせテーブルを生成し（ステップS116）、これに基づいてidf/conf値を算出して（ステップS117）、当該単語とその組み合わせ及びidf/conf値をidf/conf辞書（同時出現性辞書）４２に書き込む（ステップS118）。これにより、idf/conf辞書４２が作成される。単語の組み合わせテーブル及びその生成については、図７を参照して後述する。

【００３１】

図６は、担当者・単語テーブル生成処理フローであり、辞書編集装置１がステップS113において実行する担当者・単語テーブル生成処理を示す。辞書編集装置１が、ｄに最初の担当者を設定し（ステップS121）、図２（Ｃ）の担当者・単語テーブルの当該列にｄを追加しその要素tf(t, d) を全て「０」に初期化し（ステップS122）、ｔに最初の単語を設定し（ステップS123）、担当者・単語テーブルの行t1、t2、t3、・・にｔが存在するか否かを調べる（ステップS124）。存在しない場合、辞書編集装置１は、担当者・単語テーブルの行にｔを追加し、その要素tf(t, d) を全て「０」に初期化する（ステップS125）。存在する場合、辞書編集装置１は、ステップS125を省略し、当該担当者ｄ及び単語ｔに対応する要素に「１」を加算する（ステップS126）。即ち、tf(t, d) ＝tf(t, d) ＋１を実行する。この後、辞書編集装置１は、担当者ｄの全単語について処理が終了したか否かを

調べ（ステップS127）、終了していない場合、ｔに次の単語を設定し（ステップS128）、ステップS124以下を繰り返す。終了している場合、辞書編集装置１は、全担当者について処理が終了したか否かを調べ（ステップS129）、終了していない場合、ｄに次の担当者を設定し（ステップS1210 ）、ステップS122以下を繰り返す。終了している場合、当該処理を終了する。これにより、担当者・単語テーブルが生成される。

【００３２】

今、tf・idf 値は、式（１）のように定義される。即ち、

【００３３】

【数１】

$$tf_t^d \cdot idf_t^d = \frac{tf(t,d)}{\sum_{t' \in T^d} tf(t',d)} \cdot \log \frac{N}{df(t)} \qquad (1)$$

【００３４】

ここで、右辺第１項のtf(t, d) は、担当者ｄが作成した全ての文書（文書ファイル６）中における単語ｔの出現回数を表す。従って、右辺第１項は同一の担当者が何度も繰り返して使用する単語に大きい重みを与える。例えば、図２（Ｃ）において、担当者d1における単語t1の出現回数tf(t1, d1)＝１である。右辺第１項の分母は担当者ｄが使用した単語の頻度の総和を表している。例えば、図２（Ｃ）において、担当者d1についての当該値Σtf(t', d1)＝８５３８である。右辺第２項のＮは分類先である全ての担当者数を、df(t)は単語ｔを使用した担当者数を表す。従って、右辺第２項は特定少数の担当者が使用する単語に大きな重みを与え、各担当者を特徴付ける特定性の指標となる。以上から、担当者・単語テーブルに基づいて、tf値（tf_t^d）、idf 値（idf_t^d ）、tf・idf 値を算出することができる。

【００３５】

図７は、単語の組み合わせテーブル生成処理フローであり、辞書編集装置１がステップS116において実行する単語の組み合わせテーブル生成処理を示す。辞書編集装置１が、最小支持度及び最小確信度を読み込み（ステップS131）、ｄに最初の担当者を設定し（ステップS132）、句点の検索をしてセンテンス集合S^d を生成し（ステップS133）、最小支持度とセンテンス集合S^d の全センテンス数とからＬＳＣを算出する（ステップS134）。次に、辞書編集装置１が、単語の組み合わせ数を「１」とし、即ち、単語組み合わせテーブル（以下、同じ）Ｃ＝Ｃ１とし（ステップS135）、（Ｃの単語の組み合わせ）＝（センテンス集合S^d から抽出した各センテンスにおける重複のない単語）とし（ステップS136）、Ｃのカウント値（Count ）をＣの単語の組み合わせが出現するセンテンス集合S^d のセンテンス数とする（ステップS137）。次に、辞書編集装置１が、Ｃにおける単語の組み合わせの全てについて、当該組み合わせについてのＣのカウント値がＬＳＣ以上であるか否かを調べ、そうでない場合にはＣから当該単語の組み合わせを除き、そうである場合にはＣから当該単語の組み合わせを除かないようにする（ステップS138）。この後、辞書編集装置１が、単語の組み合わせ数が「２」か否かを調べる（ステップS139）。「２」でない場合、即ち、単語の組み合わせ数が「1」の場合、辞書編集装置１が、単語の組み合わせ数を「２」とし、即ち、Ｃ＝Ｃ２とし（ステップS1310 ）、（Ｃの単語の組み合わせ）＝（Ｃ１から生成した２単語の組み合わせ）とし（ステップS1311 ）、ステップS137以下を繰り返す。

【００３６】

ステップS139において単語の組み合わせ数が「２」である場合、辞書編集装置１が、t1をＣ２の単語の組み合わせの第１単語とし（ステップS1312）、t1のカウント値（t1__Count）をＣ１のt1のカウント値（Count）とし（ステップS1313 ）、最小確信度とt1のＣ１のカウント値とからＬＣＣを算出する（ステップS1314 ）。次に、辞書編集装置１が、Ｃ

２における単語の組み合わせの全てについて、当該組み合わせについてのＣ２のカウント値がＬＣＣ以上か否かを調べ、そうでない場合にはＣ２から当該単語の組み合わせを除き、そうである場合にはＣ２から当該単語の組み合わせを除かないようにする（ステップS1315 ）。この後、辞書編集装置１が、全担当者について処理済みか否かを調べ（ステップS1316 ）、そうでない場合、ｄに次の担当者を設定し（ステップS1317 ）、ステップS133以下を繰り返す。全担当者について処理済みである場合、処理を終了する。

【００３７】

ここで、idf/conf値は、相関ルールの抽出に用いられる周知の確信度に基づいている。相関ルールの手法を本発明に応用するために、担当者ｄが使用した単語を要素とする集合をT^d ＝（t1，t2，t3，・・，tm）とする。また、担当者ｄが作成した文書から抽出したセンテンス(句点で区切られる1文)の集合をS^d＝（s1，s2，s3，・・sn）（si⊆T^d ）とする。ここで、単語ｔの支持度support(t)はS^d 全体に対しｔを含むセンテンスの割合を表す。また、相関ルールはｔ⇒ｔ’で表現され、単語ｔが出現したセンテンスには単語ｔ’が出現する確率が高いこと、即ちｔとｔ’の共起性が高いことを表す。相関ルールは支持度(support)及び確信度(confidence)の２つのパラメータを有し、これらの値により相関ルールの有意性を示す。ここで、support(t⇒t')はS^d 全体に対しｔとｔ’を共に含むセンテンスの割合、confidence(t⇒t')はｔを含むセンテンスの中でｔ’を含むセンテンスの割合と定義されている。

【００３８】

本発明では、図８に示すように、最小支持度と最小確信度を設定して、共起性の高い単語の組み合わせを求め、この共起性を表す重みとして確信度を用いる。この例では、最小支持度及び最小確信度を、各々、０．３及び０．７とした。これらの値は経験的に定めることができる。なお、索引語抽出の研究で良く知られているように、頻度の低い単語は不要語であるが、頻度が上位の単語も特徴語ではなく一般語であることが多く、不要語となる。ここでも同様に、最小支持度と最小確信度を満足しない単語の組み合わせは不要であり、辞書４には登録しない。これと同時に、支持度もしくは確信度が上位の組み合わせは一般語の組み合わせとなり不要である。そこで、上述のように最小支持度と最小確信度を満足した単語の組み合わせの共起性を表す具体的な重みとして、確信度の逆数に単語の特定性を考慮して第１単語のidf 値を積算した値を用いる。なお、更に、第１単語に加えて第２単語のidf 値も積算した値を重みとしても良いし、確信度の代わりに支持度を用いても良い。

【００３９】

例えば、図８に示すように、担当者ｄのセンテンス集合S^dが求まるとする。単語の組み合わせを１個とすると、各々の出現回数Count が求まる（最初のＣ１）。これから、（最小支持度）×（S^dのセンテンス数）＝０．３×４よりもCount 値の小さい単語｛単語４｝を除く（２番目のＣ１）。これにより、各々の単語についての最小支持度を満足する値が定まる。次に、残りの単語について出現する２個の単語の組み合わせの全てについて、各々の出現回数Count を求め（最初のＣ２）、これから、０．３×４よりもCount 値の小さい単語の組み合わせ｛単語１，単語２｝及び｛単語１，単語５｝を除く（２番目のＣ２）。次に、残りの単語の組み合わせについて、（最小確信度）×（各々の第１単語についての最小支持度を満足する値）よりもCount 値の小さい単語の組み合わせ｛単語２，単語３｝及び｛単語３，単語５｝を除く（３番目のＣ２）。即ち、｛単語２，単語３｝については２（Ｃ２におけるCount 値、以下同じ）＜０．７×３（２番目のＣ１の単語２のCount 値、以下同じ）であり、｛単語３，単語５｝については２＜０．７×３であり、除かれる。一方、｛単語１，単語３｝については２＞０．７×２であり、｛単語２，単語５｝については３＞０．７×３であり、残される。

【００４０】

なお、本発明者の検討によれば、単語の組み合わせ数を３以上にしても、メール８（文書）の分類精度は向上しないことが判った。従って、この例においては、計算量の低減のために、単語の組み合わせは「２」に制限される。従って、この例では、同時出現性辞書

４２は、第１単語と第２単語の組み合わせ毎にidf/conf値を格納する。

【００４１】

今、単語ｔが出現したとき単語ｔ’が共起する指標であるidf/conf値を式（２）で定義する。

【００４２】

【数２】

$$idf_t^d / conf_{t \Rightarrow t'}^d = idf_t^d \cdot \frac{confidence_{max}^d}{confidence^d(t \Rightarrow t')} \quad (2)$$

【００４３】

ここで、右辺第2項の分子はある担当者ｄにおけるconfidence値（即ち、確信度）の最大値（max ）であり、担当者毎に大きさの異なるconfidence値を標準化している。以上から、図８の単語の組み合わせテーブルに基づいて、conf値、idf/conf値を算出することができる。conf値の定義は、式（２）の両辺よりidf_t^d を除くことにより明らかであろう。

【００４４】

図９は、メール（問い合わせメール）解析処理フローであり、図１のメール解析装置２がステップS103及びS104において実行するメール解析処理を示す。メール解析装置２が、処理対象である受信したメール（問い合わせメール）８を読み込み（ステップS141）、これについて、その改行とスペースとを除去し、残りの部分について周知の形態素解析を行い（ステップS142）、分かち書きした単語を得る。次に、メール解析装置２が、ｄに最初の担当者を設定し（ステップS143）、t1にメール８の最初の単語を設定し（ステップS144）、tf・idf 値重み付き加算を行い（ステップS145）、idf/conf値重み付き加算を行い（ステップS146）、メール８の全単語について処理を終了した否かを調べる（ステップS147）。tf・idf 値重み付き加算については図１０を参照して後述し、idf/conf値重み付き加算については図１１を参照して後述する。

【００４５】

全単語について処理を終了していない場合、メール解析装置２は、t1にメール８の次の単語を設定し（ステップS148）、ステップS145以下を繰り返す。全単語について処理を終了している場合、メール解析装置２は、担当者ｄのスコアを当該メール８の全単語についてのスコアの総計、即ち、Score(d)＝ΣScore(d, t1)として算出し（ステップS149）、全担当者について処理を終了した否かを調べる（ステップS1410 ）。全担当者について処理を終了していない場合、メール解析装置２は、ｄに次の担当者を設定し（ステップS1411）、ステップS144以下を繰り返す。全担当者について処理を終了している場合、メール解析装置２は、全担当者についてのScore(d)の平均及び標準偏差を算出して、これに基づいて、当該メール８に回答すべき担当者（回答者）の候補を決定し（ステップS1412 ）、これを表示する（ステップS1413 ）。

【００４６】

図１０は、tf・idf 値重み付き加算処理フローであり、メール解析装置２がステップS145において実行するtf・idf 値重み付き加算処理を示す。メール解析装置２が、Max__tf・idf __Ratio に「０」を設定（代入）し、Score(d, t1)に「０」を設定し、ｍを読み込み（ステップS151）、t__dic に担当者ｄのtf・idf 辞書４１の最初の単語を設定し（ステップS152）、t1とt__dic とが部分一致するか否かを調べる（ステップS153）。部分一致する場合、メール解析装置２が、tf・idf にt__dic のtf・idf 値を設定し、Matched__Ratio に、t1とt__dic との一致率のｍ乗を設定し、tf・idf とMatched__Ratio とからtf・idf __Ratio を算出し（ステップS154）、tf・idf __Ratio がMax __tf・idf __Ratio よりも大きいか否かを調べる（ステップS155）。大きい場合、メール解析装置２が、Max __tf・idf __Ratio にtf・idf __Ratio を設定し（ステップS156）、担当者ｄのtf・idf

辞書４１の全単語について処理が終了したか否かを調べる（ステップS157）。

【００４７】

全単語について処理が終了していない場合、t__dic に担当者ｄのtf・idf 辞書４１の次の単語を設定し（ステップS158）、ステップS153以下を繰り返す。ステップS153においてt1とt__dic とが部分一致しない場合、ステップS154～ステップS156を省略して、ステップS157を実行する。ステップS155においてtf・idf __Ratio がMax __tf・idf __Ratio よりも大きくない場合、ステップS156を省略して、ステップS157を実行する。ステップS157において全単語について処理が終了している場合、Score(d, t1)にMax __tf・idf __Ratio を設定して処理を終了する（ステップS159）。

【００４８】

図１１は、idf/conf値重み付き加算処理フローであり、メール解析装置２がステップS145において実行するidf/conf値重み付き加算処理を示す。メール解析装置２が、Max __idf/conf __Ratioに「０」を設定し、ｍを読み込み（ステップS161）、t__dic1に担当者ｄのidf/conf辞書４２の最初の第１単語を設定し、t__dic2に担当者ｄのidf/conf辞書４２の最初の第２単語を設定し（ステップS162）、t1とt__dic1とが部分一致するか否かを調べる（ステップS163）。部分一致する場合、メール解析装置２が、t2にメール８のt1の次の単語を設定し（ステップS164）、t2とt__dic2とが部分一致するか否かを調べる（ステップS165）。部分一致する場合、メール解析装置２が、idf/confにt__dic1とt__dic2のidf/conf値を設定し、Matched __Ratio1にt1とt__dic1の一致率のｍ乗を設定し、Matched __Ratio2にt2とt__dic2の一致率のｍ乗を設定し、idf/confとMatched __Ratio1とMatched __Ratio2とに基づいてidf/conf__Ratio を算出する（ステップS166）。

【００４９】

ステップS165においてt2とt__dic2とが部分一致しない場合、メール解析装置２が、メール８の最後の単語まで処理済か否かを調べ（ステップS1611 ）、最後の単語まで処理を終了していない場合、t2にメール８のt2の次の単語を設定し（ステップS1612 ）、ステップS165以下を繰り返す。

【００５０】

ステップS166の後、メール解析装置２が、idf/conf__Ratio がMax __idf/conf__Ratio よりも大きいか否かを調べ（ステップS167）、大きい場合、Max __idf/conf__Ratio にidf/conf __Ratio を設定して（ステップS168）、担当者ｄのidf/conf辞書４２の全単語について処理済か否かを調べる（ステップS169）。全単語について処理を終了していない場合、メール解析装置２が、t__dic1に担当者ｄのidf/conf辞書４２の次の第１単語を設定し、t__dic2に担当者ｄのidf/conf辞書４２の次の第２単語を設定し（ステップS1610 ）、ステップS163以下を繰り返す。全単語について処理を終了している場合、メール解析装置２が、定数αを読み込み（ステップS1613 ）、Score(d, t1)にαとMax __idf/conf__Ratio とにより求まる値を加算することによりScore(d, t1)を算出する（ステップS1614 ）。

【００５１】

ステップS163においてt1とt__dic1とが部分一致しない場合、メール解析装置２が、ステップS164～S168（S1611 及びS1612 を含む）を省略して、ステップS169を実行する。ステップS167においてidf/conf__Ratio がMax __idf/conf__Ratio よりも大きくない場合、メール解析装置２が、ステップS168を省略して、ステップS169を実行する。ステップS1611 においてメール８の最後の単語まで処理を終了している場合、メール解析装置２が、ステップS1612 （S166～S168を含む）を省略して、ステップS169を実行する。

【００５２】

このように、本発明では、メール８中の全ての出現単語を辞書４と照合して、メール８に対する各担当者のスコアを算出する。このとき、形態素解析により分割された単語の単語長が変化することがある。また、質問者と担当者とが同一の単語を使用する保証はない。このため、メール８に含まれる単語と辞書４中の単語を完全一致を条件に照合するとマッチする単語が少なくなり、結局は分類の精度が向上しない。そこで、本発明では、単語

の照合は部分一致によって行い、この一致率を単語の重みに乗ずることによる重み付き加算を行って、メール８に対する担当者ｄのスコアを算出する。

【００５３】

まず、メール８中のある単語ｔのtf・idf 値とidf/conf値を加味したスコアを式（３）で定義する。なお、単語ｔ及びこれと共起性を示す単語ｔ’は、ともに上記の部分一致のために辞書４中の複数の単語又は単語の組み合わせに照合する可能性があるので、このうち、各々tf・idf 値とidf/conf値に一致率を乗じた値が最大値となるものを加算する。

【００５４】

【数３】

$$Score_t^d = tf_t^d \cdot idf_t^d \cdot Match\_Ratio(t) + \alpha \cdot \max_{\substack{t' \in Mail, \\ t' \neq t}} [conf_{t \Rightarrow t'}^d \cdot idf_t^d \cdot Match\_Ratio(t) \cdot Match\_Ratio(t')] \quad (3)$$

【００５５】

ここで、Match _Ratio (t)は単語の一致率、即ちメール８中の単語ｔと辞書４中の単語の単語長に対する一致した文字数の比率を表し、式（４）で定義される。

【００５６】

【数４】

$$Match\_Ratio(t) = \left(\frac{Match\_Length(t, t_{dic})^2}{Length(t) \cdot Length(t_{dic})} \right)^m \quad (4)$$

【００５７】

ここで、t_{dic}は辞書４中の単語を表し、Length(t)は単語ｔの文字数、Match _Length(t, t_{dic})はｔとt_{dic}の一致した文字数を表している。また、式（３）のαは、第１項と第２項の重みを決める係数である。更に、式（４）のｍは単語の一致率をどのオーダでスコアに反映するかを表す。この一致率は1以下であるため、ｍを大きくするほどこの率によってスコアに差が生じる。なお、このαとｍのいずれも適当な範囲で変動させて、経験則により、最大の分類精度を得るαとｍを採用することができる。

【００５８】

最後に、式（３）で与えられる単語ｔのスコア$Score_t^d$について、メール８本文に出現する単語で総和を取ることにより、入力されたメール８に対する担当者ｄのスコアを式（５）で定義する。

【００５９】

【数５】

$$Score^d = \sum_{t \in Mail} Score_t^d \quad (5)$$

【００６０】
ここで、入力したメール８に対する回答者は、全ての担当者のScore^{d} の分布を正規分布と仮定して、その平均スコアより２σ（σは標準偏差）以上大きいスコアを持つ担当者と推定する。これは、実際のメール８において、特に問い合わせ内容を限定しない場合は、複数の事柄について複合して質問していることや複数の専門分野にまたがって質問していることが多く、複数の担当者が連携して回答する必要があるためである。
【００６１】
図１２は、強化学習処理フローであり、主として図１の強化学習装置３がステップS102において実行する単語の重みの強化学習処理を示す。強化学習装置３が、別文書ファイル７を読み込み（ステップS171）、真の回答者集合を抽出する（ステップS172）。なお、別文書ファイル７には、例えば学習用メール８が含まれる（以下同じ）。メール解析装置２が、別文書ファイル７について前述のようにメール解析を実行し、図３の結果を得る。強化学習装置３が、当該結果をメール解析結果テーブルとして取得し（ステップS173）、メール解析結果テーブル中の全ての回答者について、当該回答者が真の回答者集合に含まれるか否かを調べ（ステップS174）、含まれる場合には当該回答者について重要語学習処理を実行し（ステップS175）、含まれない場合には当該回答者について不要語学習処理を実行する（ステップS176）。なお、実際には、点線で示すように、一人の回答者についてステップS174とS175又はS176とを実行することを、各回答者について繰り返す。この後、強化学習装置３が、メール解析結果テーブル中の全ての非回答者について、当該非回答者が真の回答者集合に含まれるか否かを調べ（ステップS177）、含まれる場合には当該非回答者について特定語学習処理を実行し（ステップS178）、含まれない場合にはステップS178を省略する。なお、実際には、点線で示すように、一人の非回答者についてステップS177又はS177及びS178を実行することを、各非回答者について繰り返す。この後、強化学習装置３が、全ての別文書ファイル７について処理済か否かを調べ（ステップS179）、処理済でない場合にはステップS171以下を繰り返し、処理済である場合には処理を終了する。
【００６２】
本発明では、別文書ファイル7の解析結果を受けて、ＰＳにより、推定した回答者の正しさから担当者毎の辞書４のtf・idf 値とidf/conf値を更新する。この例では、作成者即ち担当者が判明している別文書ファイル７をメール解析装置２に入力する。この場合、真の文書作成者が判っているので、当該システムの分類結果の正否を判断できる。なお、人間の分類者が参照して別文書ファイル7を分類した結果を正しいとして学習しても良い。
【００６３】
本発明では、以下の３種類の強化学習を行う。即ち、重要語の学習は、真の回答者をシステムが回答者であると推定できた場合に、別文書ファイル7とこの回答者の辞書４０で照合した単語を重要語として、その重みを大きくする。また、特定語の学習は、真の回答者をシステムが回答者であると推定できなかった場合に、別文書ファイル7に出現した単語の中で真の回答者について特定性の高い単語を再評価することにより、この回答者の特定語の重みを大きくする。更に、不要語の学習は、真の回答者以外をシステムが回答者であると誤推定した場合に、別文書ファイル7とこの担当者の辞書４０で照合した単語を不要語として、この担当者の辞書４０における不要語の重みを小さくする。これと同時に、この単語を真の回答者以外の他の担当者の辞書４０でも、同様に不要語として学習する。
【００６４】
図１３は、重要語学習処理フローであり、強化学習装置３がステップS175において実行する重要語の学習処理を示す。強化学習装置３が、定数c1、Ｌ、Ｓを定義ファイルから読み込み（ステップS181）、単語ｔにメール解析結果テーブルにおいて対象回答者の第１位照合単語を設定する（ステップS182）。次に、強化学習装置３が、対象回答者のtf・idf辞書４１でｔを検索してtf・idf にｔと照合した単語のtf・idf 値を設定し、tf・idf とc1とからf__tf・idf を算出し（ステップS183）、対象回答者のidf/conf辞書４２の第１単語でｔを検索してidf/confにｔと照合した単語のidf/conf値を設定し、idf/confとc1とからf__idf/confを算出し（ステップS184）、tf・idf 辞書４１及びidf/conf辞書４２に

おけるｔのtf・idf 値及びidf/conf値を更新する（ステップS185）。即ち、新たなtf・idf 値を、それまでのtf・idf 値にf__tf・idf を加算した値とする。新たなidf/conf値を、それまでのidf/conf値にf__idf/confを加算した値とする。

【００６５】

次に、強化学習装置３が、ｉ＝２として（ステップS186）、単語ｔにメール解析結果テーブルにおいて対象回答者の第ｉ位照合単語を設定する（ステップS187）。次に、強化学習装置３が、対象回答者のtf・idf 辞書４１でｔを検索してtf・idf にｔと照合した単語のtf・idf 値を設定し、f__tf・idf とＳとから新たなf__tf・idf を算出し（ステップS188）、対象回答者のidf/conf辞書４２の第１単語でｔを検索してidf/confにｔと照合した単語のidf/conf値を設定し、f__idf/confとＳとから新たなf__idf/confを算出し（ステップS189）、tf・idf 辞書４１及びidf/conf辞書４２におけるｔのtf・idf 値及びidf/conf値を更新する（ステップS1810 ）。即ち、新たなtf・idf 値を、それまでのtf・idf 値にf__tf・idf を加算した値とし、新たなidf/conf値を、それまでのidf/conf値にf__idf/confを加算した値とする。この後、強化学習装置３が、ｉがＬに等しいか否かを調べ（ステップS1811 ）、等しくない場合、ｉをｉ＋１とし（ステップS1812 ）、ステップS187以下を繰り返す。等しい場合、ステップS1812 を省略して、処理を終了する。なお、メール8の文章が極めて短いとｉの最大値がＬより小さくなることがあるが、この場合、読み込んだＬに代えて当該ｉの最大値が用いられる。即ち、ステップS1811においてｉが当該ｉの最大値に等しい場合、処理を終了する（以下、ステップS1911及びS2120において同じ）。

【００６６】

ＰＳでは、エピソード単位にルールに付加された重みを強化する。エピソードとは、初期状態あるいは報酬を得た直後から、次の報酬までのルール系列を表す。この強化には報酬からどれだけ過去かを引数とする強化関数が用いられる。長さlのエピソード（rl, ・・ri, ・・r2, r1）に対して、ルールｒ$_i$ の重みｗ$_i$ は式（６）のように強化関数ｆ$_i$ で強化される。

【００６７】

【数６】

$$w_{r_i} = w_{r_i} + f_i \qquad (6)$$

【００６８】

本発明では、ルールは辞書４中の単語又は単語の組み合わせに相当し、エピソードのルール系列は式（５）でスコアScore^d を決定した単語（tl, ・・ti, ・・t2, t1）である。即ち別文書ファイル7と担当者ｄの辞書４０中で照合した単語である。更に、系列中の順序ｉは式（5）のスコアScore^d への寄与の順、即ち式（３）のScore$_t{}^d$の大きさの順とし、このScore$_t{}^d$が最大となるスコア１位の単語ｔ（第１位照合単語）をt1とする。また、本発明では、単語の重みをtf・idf 値とidf/conf値の２種類で構成しているので、式（６）はそれぞれ式（７）及び式（８）のようになり、当該システムが推定した回答者と真の回答者が一致した場合に担当者ｄの単語tiに、ｉが上位であるほど大きな正の報酬を与える。

【００６９】

【数７】

$$tf_{t_i}^d \cdot idf_{t_i}^d = tf_{t_i}^d \cdot idf_{t_i}^d + f_{tf,i}^d \quad (7)$$

$$idf_{t_i}^d / conf_{t_i \Rightarrow t_i'}^d = idf_{t_i}^d / conf_{t_i \Rightarrow t_i'}^d + f_{conf,i}^d \quad (8)$$

【００７０】

但し、本発明の学習方法においては、tf・idf 値とidf/conf値について、それぞれ別に強化するが、その方法は全く同様である（図１３中の処理を参照）。

【００７１】

ＰＳでは、有効な単語の重みが強化され、無効な単語の重みが抑制されることを保証しなければならない。この条件は合理性定理により式（９）を満足することであることが周知のように証明されている。

【００７２】

【数８】

$$\forall k = 2, \cdots, W+1, \qquad L \sum_{i=k}^{W+1} f_i^d < f_{i-1}^d \quad (9)$$

【００７３】

ここで、Ｗはエピソードの最大長、Ｌは同一感覚入力下に存在する有効ルールの最大数であり、本発明では学習を有効にする単語数を限定し、別文書ファイル7と辞書４で一致した単語でスコア$Score^d$の上位何単語を学習するかを設定することとし、その単語数をＬとした（例えば、Ｌ＝１０なら上位１０位の単語を学習する、以下同じ）。また、この式（９）の定理を満足する最も簡単な強化関数には、式（１０）で表される等比減少関数を用いる。

【００７４】

【数９】

$$f_n^d = \frac{1}{S} f_{n-1}^d, \qquad n = 2, \cdots, W \quad (10)$$

【００７５】

ここで、周知のように、Ｓ≧Ｌ＋１を満足しなければならない。また、式（１０）の初期値f_1^d、即ちスコア$Score_t^d$が最大となる単語の強化値は以下の式（１１）とする。ここで、c1は定数である。

【００７６】

【数１０】

$$f_1^d = c_1 w_1^d \quad (11)$$

【００７７】
このように、重要語の学習は、推定した回答者が真の回答者であった場合に行い、スコア$Score_t^d$が上位である単語の重みに正の報酬を与え、その重みを大きくする。
【００７８】
図１４は、不要語学習処理フローであり、強化学習装置３がステップS176において実行する不要語の学習処理を示す。強化学習装置３が、定数c2、Ｌ、Ｓを定義ファイルから読み込み（ステップS191）、単語ｔにメール解析結果テーブルにおいて対象回答者の第１位照合単語を設定する（ステップS192）。次に、強化学習装置３が、対象回答者のtf・idf辞書４１でｔを検索してtf・idf にｔと照合した単語のtf・idf 値を設定し、tf・idf とc2とからf__tf・idf を算出し（ステップS193）、対象回答者のidf/conf辞書４２の第１単語でｔを検索してidf/confにｔと照合した単語のidf/conf値を設定し、idf/confとc2とからf__idf/confを算出し（ステップS194）、不要語強化処理を実行する（ステップS195）。不要語強化処理については図１５を参照して後述する。
【００７９】
次に、強化学習装置３が、ｉ＝２として（ステップS196）、単語ｔにメール解析結果テーブルにおいて対象回答者の第ｉ位照合単語を設定する（ステップS197）。次に、強化学習装置３が、対象回答者のtf・idf 辞書４１でｔを検索してtf・idf にｔと照合した単語のtf・idf 値を設定し、f__tf・idf とＳとから新たなf__tf・idf を算出し（ステップS198）、対象回答者のidf/conf辞書４２の第１単語でｔを検索してidf/confにｔと照合した単語のidf/conf値を設定し、f__idf/confとＳとから新たなf__idf/confを算出し（ステップS199）、ステップS195と同様の不要語強化処理を実行する（ステップS1910 ）。この後、強化学習装置３が、ｉがＬに等しいか否かを調べ（ステップS1911 ）、等しくない場合、ｉをｉ＋１とし（ステップS1912 ）、ステップS197以下を繰り返す。等しい場合、ステップS1912 を省略して、処理を終了する。
【００８０】
図１５は、不要語強化処理フローであり、強化学習装置３がステップS195及びS1910 において実行する不要語の強化処理を示す。強化学習装置３が、担当者ｄに対象回答者を設定し（ステップS201）、tf・idf －f__tf・idf が「０」より大きいか否かを調べ（ステップS202）、大きい場合、tf・idf 辞書４１のｔのtf・idf 値をtf・idf －f__tf・idfに更新し（ステップS203）、大きくない場合、tf・idf 辞書４１のｔのtf・idf 値を「０」に更新する（ステップS204）。次に、強化学習装置３が、idf/conf－f__idf/confが「０」より大きいか否かを調べ（ステップS205）、大きい場合、idf/conf辞書４２のｔのidf/conf値をidf/conf－f__idf/confに更新し（ステップS206）、大きくない場合、idf/conf辞書４２のｔのidf/conf値を「０」に更新する（ステップS207）。次に、強化学習装置３が、真の回答者集合以外の全担当者について処理済か否かを調べ（ステップS208）、処理済でない場合、担当者ｄに次の担当者（対象回答者）を設定し（ステップS209）、担当者ｄのtf・idf 辞書４１でｔを検索してtf・idf にｔと照合した単語のtf・idf 値を設定し（ステップS2010 ）、担当者ｄのidf/conf辞書４２の第１単語でｔを検索してidf/confにｔと照合した単語のidf/conf値を設定し（ステップS2011 ）、ステップS202以下を繰り返す。ステップS208において全担当者について処理済である場合、処理を終了する。
【００８１】
不要語の学習は、真の回答者以外をシステムが回答者であると誤推定した場合に行う。この場合は、推定した回答者の上位単語は本来不要語であるが、大きな重みが付与されているものと考えられる。そこで、このような単語の重みには負の報酬を与え、式（１２）によりその重みを小さくする。
【００８２】

【数１１】

$$w_i^d = \begin{cases} w_i^d - f_i^d & (\text{if} \quad w_i^d \geqq f_i^d) \\ 0 & (\text{if} \quad w_i^d < f_i^d) \end{cases} \qquad (12)$$

【００８３】
なお、式（１２）では重みが負になることを防いでいる。また、強化値の等比関数は式（１０）と同様で、その初期値は式（１３）で与える。
【００８４】
【数１２】

$$f_1^d = c_2 w_1^d \qquad (13)$$

【００８５】
ここで、c2は定数である。更に、経験的にこのような不要語は一般用語であることが大多数であり、システムが回答者として推定した者だけではなく、真の回答者以外の辞書４０でも負の報酬を与えることにより、不要語の学習効率が大きく向上する。そこで、この場合の学習は、システムが誤推定した回答者の上位単語について、真の回答者以外の他の担当者の辞書４０にその単語が登録されている場合、同様に負の報酬を与え、その重みを小さくする。
【００８６】
図１６は、特定語学習処理フローであり、強化学習装置３がステップS178において実行する特定語の学習処理を示す。強化学習装置３が、定数c3、Ｌ、Ｓ、min __idf 、max __ave __ratio を定義ファイルから読み込み（ステップS211）、idf 値再評価処理を実行する（ステップS212）。idf 値再評価処理については図１７を参照して後述する。次に、強化学習装置３が、単語ｔにメール解析結果テーブルにおいて対象非回答者の第１位照合単語を設定し（ステップS213）、対象非回答者のtf・idf 辞書４１でｔを検索してtf・idf にｔと照合した単語のtf・idf 値を設定し（ステップS214）、max __tf・idf に対象非回答者のtf・idf 辞書４１での最大tf・idf 値を設定し（ステップS215）、tf・idf がmax __tf・idf とmax __ave __ratio との積以上であるか否かを調べる（ステップS216）。なお、max __ave __ratio は全担当者のtf・idf 値及びidf/conf値の最大値で標準化した平均的な重みを表し、このmax __ave __ratio とmax __tf・idf との積は後述する式（１４）のw_p^dとなり、対象非回答者のスコアに有効となる重みを表す。max __ave __ratio の算出方法は（図１８で）後述する。f__tf・idf がmax __tf・idf とmax __ave __ratio との積以上でない場合、f__tf・idf にmax __tf・idf とmax __ave __ratio との積を設定し（ステップS217）、積以上である場合、f__tf・idf にtf・idf とc3とから求まる値を設定する（ステップS218）。
【００８７】
この後、強化学習装置３が、対象非回答者のidf/conf辞書４２の第１単語でｔを検索してidf/confにｔと照合した単語のidf/conf値を設定し（ステップS219）、max __idf/conf に対象非回答者のidf/conf辞書４２での最大idf/conf値を設定し（ステップS2110 ）、idf/confがmax __idf/confとmax __ave __ratio との積以上であるか否かを調べる（ステップS2111 ）。なお、max __ave __ratio については、図１６において前述した通りである。f__idf/confがmax __idf/confとmax __ave __ratio との積以上でない場合、f__idf/confにmax __idf/confとmax __ave __ratio との積を設定し（ステップS2112 ）、積以上である場合、f__idf/confにidf/confとc3とから求まる値を設定する（ステップS2113 ）。

【００８８】
この後、強化学習装置３が、tf・idf 辞書４１及びidf/conf辞書４２におけるｔのtf・idf 値及びidf/conf値を更新する（ステップS2114 ）。即ち、新たなtf・idf 値を、それまでのtf・idf 値にf__tf・idf を加算した値とし、新たなidf/conf値を、それまでのidf/conf値にf__idf/confを加算した値とする。この後、強化学習装置３が、ｉ＝２として（ステップS2115 ）、単語ｔにメール解析結果テーブルにおいて対象回答者の第ｉ位照合単語を設定する（ステップS2116 ）。次に、強化学習装置３が、対象回答者のtf・idf 辞書４１でｔを検索してtf・idf にｔと照合した単語のtf・idf 値を設定し、f__tf・idf とＳとから新たなf__tf・idf を算出し（ステップS2117 ）、対象回答者のidf/conf辞書４２の第１単語でｔを検索してidf/confにｔと照合した単語のidf/conf値を設定し、f__idf/confとＳとから新たなf__idf/confを算出し（ステップS2118 ）、tf・idf 辞書４１及びidf/conf辞書４２におけるｔのtf・idf 値及びidf/conf値を更新する（ステップS2119 ）。即ち、新たなtf・idf 値を、それまでのtf・idf 値にf__tf・idf を加算した値とし、新たなidf/conf値を、それまでのidf/conf値にf__idf/confを加算した値とする。この後、強化学習装置３が、ｉがＬに等しいか否かを調べ（ステップS2120 ）、等しくない場合、ｉをｉ＋１とし（ステップS2121 ）、ステップS2116 以下を繰り返す。等しい場合、ステップS2121 を省略して、処理を終了する。

【００８９】
図１７は、idf 値再評価処理フローであり、強化学習装置３がステップS212において実行するidf 値再評価処理を示す。強化学習装置３が、メール解析結果テーブルにおいて全非回答者の上位Ｌ位までの単語集合でidf 値を算出し（ステップS221）、メール解析結果テーブルにおいて対象非回答者の上位Ｌ位までの単語のScore(d, t) をidf 値に書き換え（ステップS222）、メール解析結果テーブルから対象非回答者のmin __idf 未満のidf 値の単語と、Ｌ＋１位以下の単語とを削除し（ステップS223）、メール解析結果テーブルにおいて対象非回答者の単語をidf 値で降順に整列させる（ステップS224）。なお、ステップS223において、min __idf 未満のidf 値の単語、又は、Ｌ＋１位以下の単語が存在しない場合には、削除すべき単語が存在しない場合がある。また、ステップS223における単語の削除によりｉの最大値がＬより小さくなることがあるが、この場合、読み込んだＬに代えて当該ｉの最大値が用いられる。即ち、ステップS2120においてｉが当該ｉの最大値に等しい場合、処理を終了する。

【００９０】
特定語の学習は、真の回答者をシステムが回答者であると推定できなかった場合に行う。この場合は、真の回答者の辞書４０内で重要語の重みが本来の値より小さくなっている恐れがある。これは、真の回答者がある専門用語を固定的に多く使用し、その同意語や類似語を滅多に使用しない場合等では同意語や類似語のtf値が極端に小さくなるために起こり得る。そこで、真の回答者のスコア上位単語と他の担当者の上位単語を比較して、単語の特定性をidf 値の算出方法と同様にして再評価する。即ち、回答者としては推定されなかった真の回答者の上位単語に特定性が認められれば特定語として正の報酬を与え、その重みを大きくする。ここでの強化方法は、idf 値で単語系列を決定する以外は式（１０）の真の回答者を推定できた場合と同様であるが、idf 値が最大となる単語の強化値は式（１１）の代わりに式（１４）を用いる。

【００９１】
【数１３】

$$f_1^d = \begin{cases} c_3 w_1^d & (\text{if} \quad w_1^d \geqq w_p^d) \\ w_p^d & (\text{if} \quad w_1^d < w_p^d) \end{cases} \qquad (14)$$

【００９２】

ここで、式（１１）と同様にc3は定数である。なお、辞書４中の全ての単語について、標準化した順位と重みの関係は図１８に示したような曲線となる。図１８の横軸は、各担当者の単語の最大順位（即ち、登録単語数）で各単語の順位を割った値を表し、縦軸は、同様に重みの最大値で各単語の重みを割った値を表す。図１８は横軸の各順位に該当する全担当者の平均重み（tf・idf 値とidf/conf値の両方の平均）をプロットしたものである。この曲線の平均傾きを「－１」と仮定し、この曲線と、傾きが「－１」の直線との接点における重みの値を図１６におけるmax _ave _ratio とした。式（１４）のw_p^dはこのmax _ave _ratio と最大重みとの積を表す。真の回答者を当該システムが回答者であると推定（分類）できなかった場合は、真の回答者のスコア$Score_t^d$の上位単語は当該担当者の辞書４０内で低順位であることが多いと考えられる。このような低順位のものの重みは、高順位のものの重みと比較して何桁も小さな値となっている。このため、強化値を元の重みの定数倍として与えても全体のスコアに影響する大きさにならない。そこで、学習の効果が現れないことを受け、スコア1位のw_1^dがこのw_p^dより小さければ、一度にw_p^dまで引き上げ、強化することとした。なお、本発明者の検討によれば、単純な重みの平均値をw_p^dとすると、重みの小さい単語が非常に多いために、w_p^dは小さな値となり、学習効果が上がらないことが判った。

【００９３】

学習時の単語の照合は、この例では完全一致としている（従って、図１３～図１７においては単語は必ず照合されることになる）が、これは本発明者の検討により、別文書ファイル７を学習に用いる場合、メール解析のように部分一致としなくても学習効果が上がることが確かめられたためであり、学習にメール８を用いる場合には、部分一致として強化値に単語の一致率を掛けて、強化しても良い。

【００９４】

以上、本発明をその実施の形態に従って説明したが、本発明は、その主旨の範囲内で種々の変形が可能である。例えば、本発明の文書分類システムは、メール８に限られず、広く文書の分類に用いることができる。更に、分類の対象である文書に、例えばホームページを含むことができる。更には、分類の対象である文書に、例えば音声入力されたデータを音声認識して得た電子データを含むことができる。本発明の文書分類システムは、自然言語（記号を含む）により構成された電子データであれば、これを分類することができる。また、文書の分類先は、個々の担当者に限られず、種々の組織内における担当部署であっても良い。本発明の文書分類システムは、３個のサブシステムの全てを備えなくとも良い。即ち、辞書編集装置１、文書分類装置２、強化学習装置３の各々を独立に設けても良く、文書分類装置２のみを設けても良く、文書分類装置２に辞書編集装置１又は強化学習装置３を併設しても良い。

【産業上の利用可能性】

【００９５】

以上説明したように、本発明によれば、辞書編集装置において、tf・idf 値とidf/conf値とを別々の独立した２個のパラメータとして用いてtf・idf辞書、idf/conf辞書を作成することができるので、これを文書分類装置の辞書として用いることにより、電子メール等の文書の分類に用いた場合、実用的な分類精度を得ることができ、文書を適切な担当者に自動的に正確に分類するための辞書を容易に作成することができる。

【００９６】

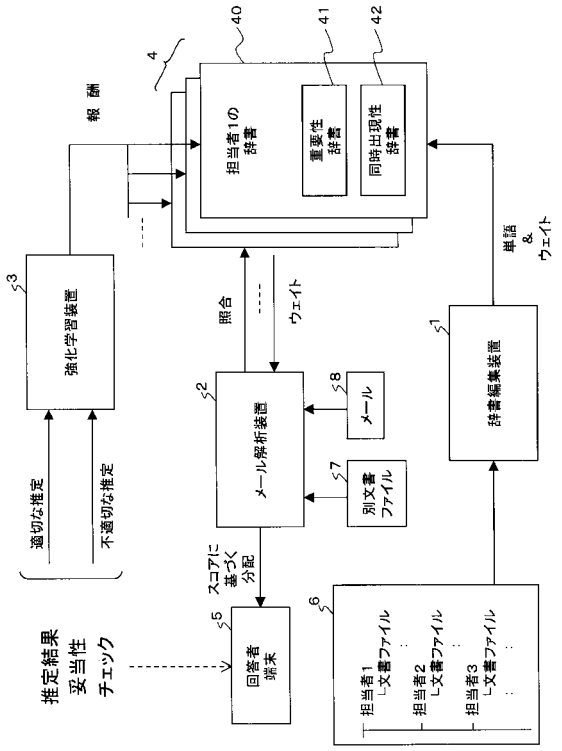
また、本発明によれば、文書分類装置において、前述の２個の辞書tf・idf辞書、idf/conf辞書を用いることにより、電子メール等の文書の分類に用いた場合、実用的な分類精度を得ることができ、文書を適切な担当者に自動的に正確に分類することができる。

【００９７】

また、本発明によれば、文書分類システムプログラムを、フレキシブルディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ／Ｗ、ＤＶＤ等の媒体に格納すること、又は、インターネット等のネットワークを介してダウンロードすることにより供給することができ、前述の文書分類装置を容易に実現することができ、正確な文書分類を可能とすることができる。

【図面の簡単な説明】
【００９８】
【図１】文書分類システム構成図である。
【図２】辞書説明図である。
【図３】文書分類結果（メール解析結果）説明図である。
【図４】文書分類処理フローである。
【図５】辞書編集処理フローである。
【図６】担当者・単語テーブル生成処理フローである。
【図７】単語の組み合わせテーブル生成処理フローである。
【図８】単語の組み合わせテーブル説明図である。
【図９】問い合わせメール解析処理フローである。
【図１０】tf・idf 値重み付き加算処理フローである。
【図１１】idf/conf値重み付き加算処理フローである。
【図１２】強化学習処理フローである。
【図１３】重要語学習処理フローである。
【図１４】不要語学習処理フローである。
【図１５】不要語強化処理フローである。
【図１６】特定語学習処理フローである。
【図１７】idf 値再評価処理フローである。
【図１８】順位と重みの関係説明図である。
【符号の説明】
【００９９】
１　辞書編集装置
２　文書分類装置（メール解析装置）
３　強化学習装置
４　辞書
４０　カテゴリ辞書
４１　重要性辞書（tf・idf 辞書）
４２　同時出現性辞書（idf/conf辞書）

【図１】



【図２】

(A)

単語	tf値	idf値	tf･idf値
単語a1	0.00424	1.838	0.00779
単語a2	0.00354	1.1451	0.00405
単語a3	0.0114	0.3502	0.00399
・	・	・	・
・	・	・	・
・	・	・	・

(B)

第1単語	第2単語	第1単語のidf値	conf値	idf/conf値
単語b1	単語b2	1.838	0.01207	152.3
単語b3	単語b4	1.1451	0.01263	90.67
単語b5	単語b6	1.8382	0.02241	82.03
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・

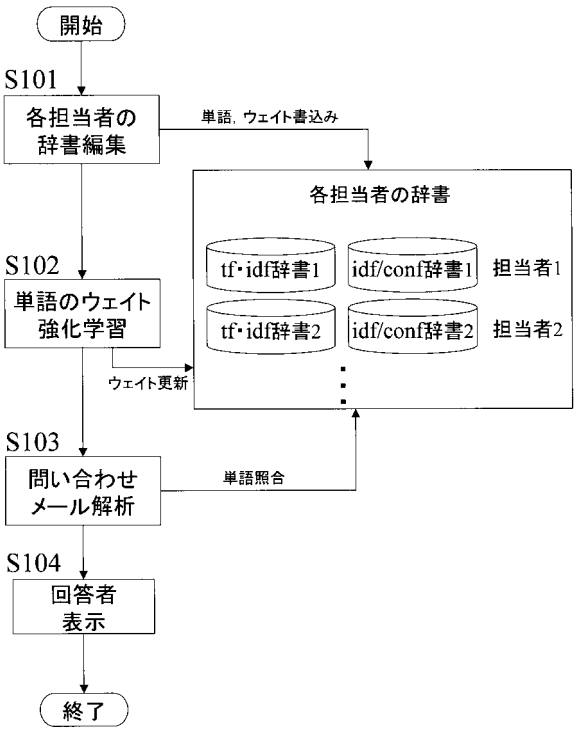
(C)

単語＼担当者	d1	d2	d3･･･	df(t)
t1	1	11	5･･･	32
t2	0	25	2･･･	14
t3	33	0	0･･･	2
・	・	・	・	
・	・	・	・	
Σtf(t',d)	8538	12945	35792	

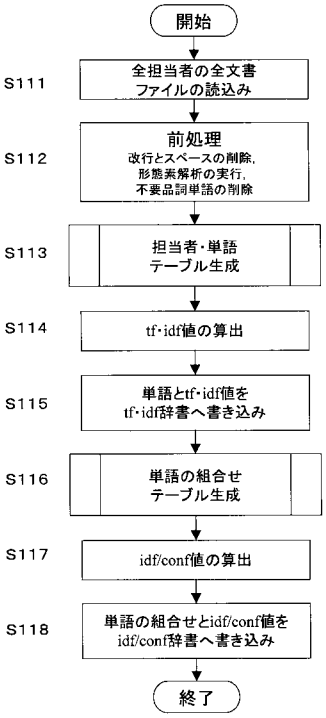
【図３】

回答者	担当者名d	Score(d)	照合単語t	Score(d,t)
○	担当者1	0.008297	単語11	0.003221
			単語12	0.002776
			単語13	0.000904
			単語14	0.000586
			単語15	0.000277
			・	・
			・	・
			・	・
○	担当者2	0.003147	単語21	0.000917
			単語22	0.000742
			単語23	0.000447
			単語24	0.000421
			単語25	0.000183
			・	・
			・	・
			・	・
×	担当者3	0.003051	単語31	0.001869
			単語32	0.000291
			単語33	0.000241
			単語34	0.000122
			単語35	0.000088
			・	・
			・	・
			・	・
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・

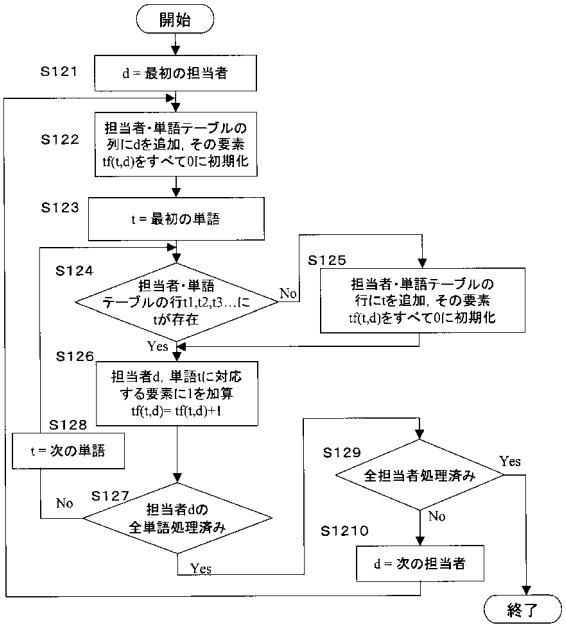
【図４】



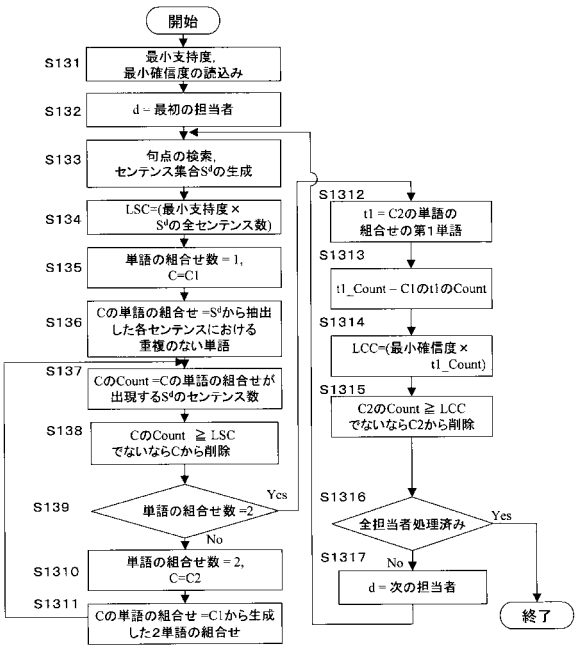
【図５】



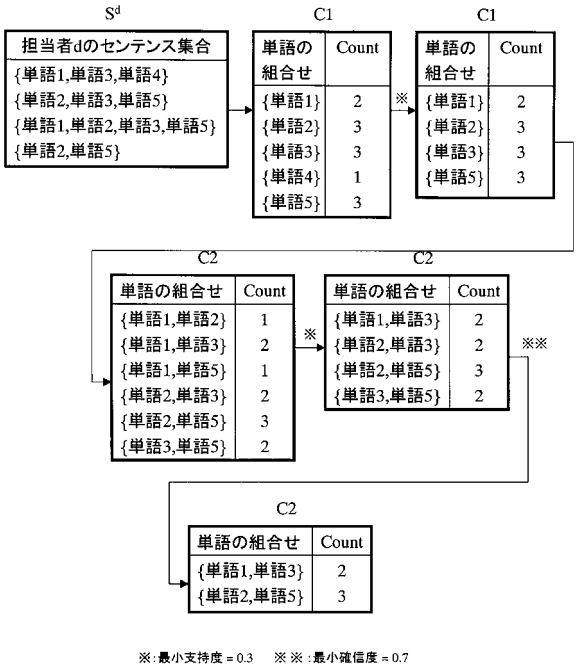
【図６】



【図７】



【図８】



S^d

担当者dのセンテンス集合
{単語1,単語3,単語4}
{単語2,単語3,単語5}
{単語1,単語2,単語3,単語5}
{単語2,単語5}

C1

単語の組合せ	Count
{単語1}	2
{単語2}	3
{単語3}	3
{単語4}	1
{単語5}	3

※

C1

単語の組合せ	Count
{単語1}	2
{単語2}	3
{単語3}	3
{単語5}	3

C2

単語の組合せ	Count
{単語1,単語2}	1
{単語1,単語3}	2
{単語1,単語5}	1
{単語2,単語3}	2
{単語2,単語5}	3
{単語3,単語5}	2

※

C2

単語の組合せ	Count
{単語1,単語3}	2
{単語2,単語3}	2
{単語2,単語5}	3
{単語3,単語5}	2

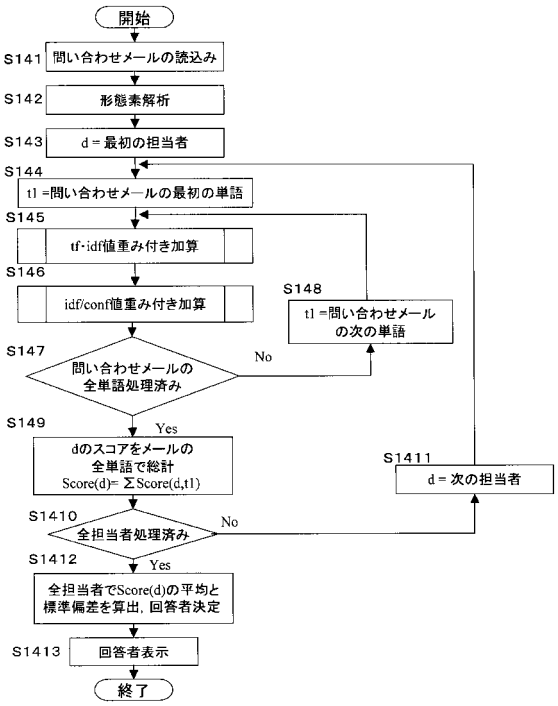
※※

C2

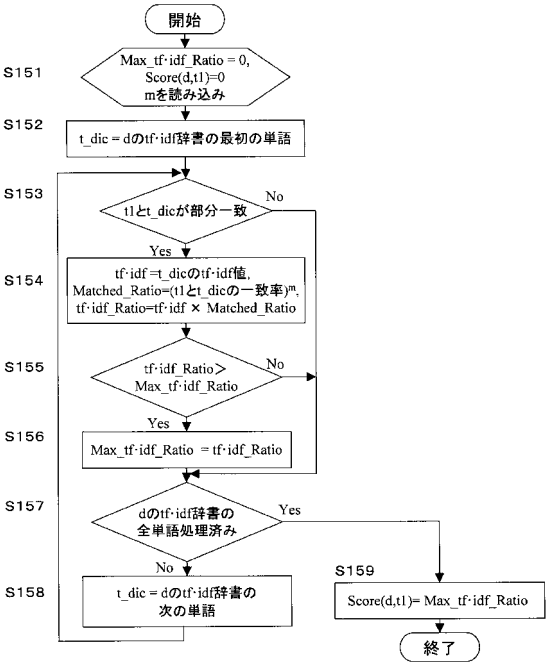
単語の組合せ	Count
{単語1,単語3}	2
{単語2,単語5}	3

※：最小支持度 = 0.3　※※：最小確信度 = 0.7

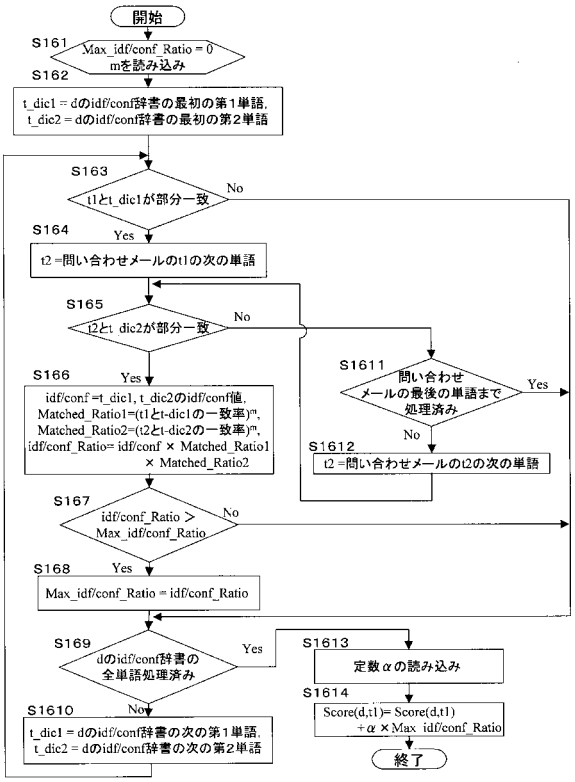
【図９】



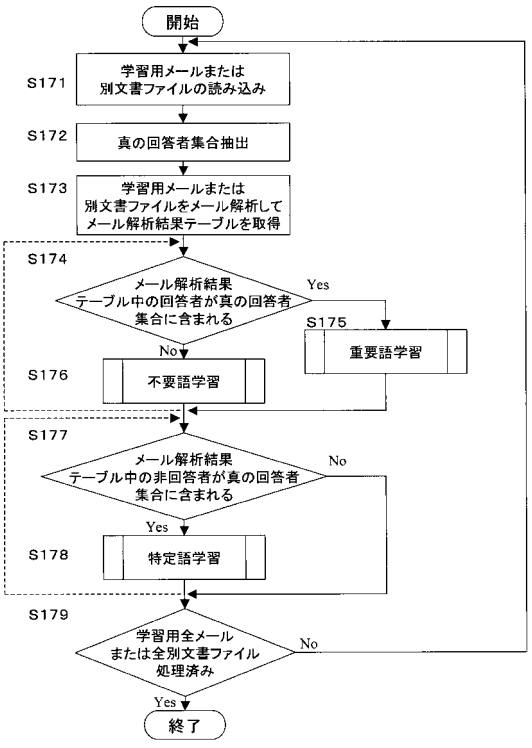
【図１０】



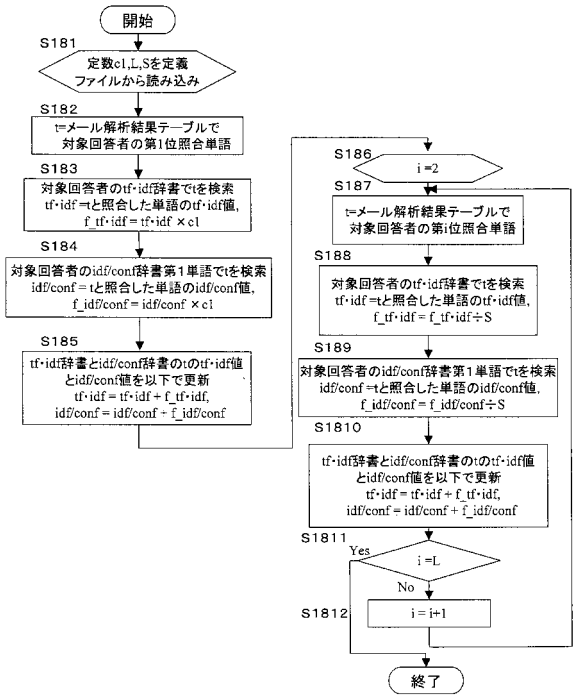
【図１１】



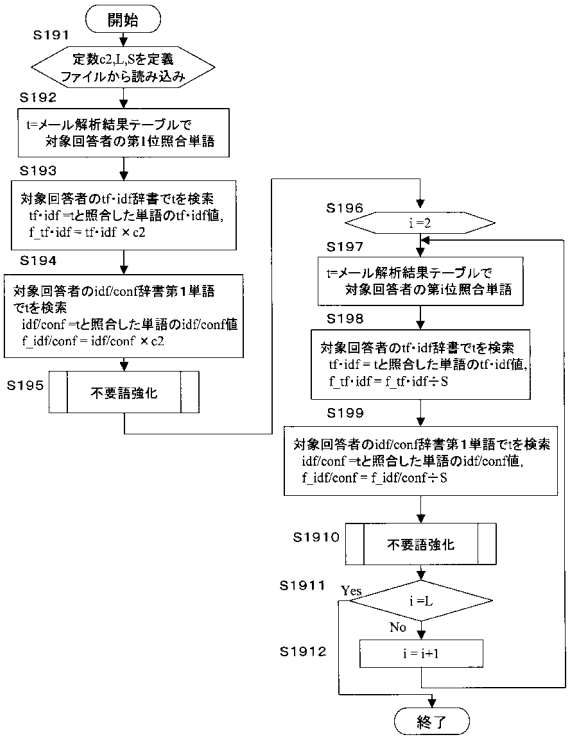
【図１２】



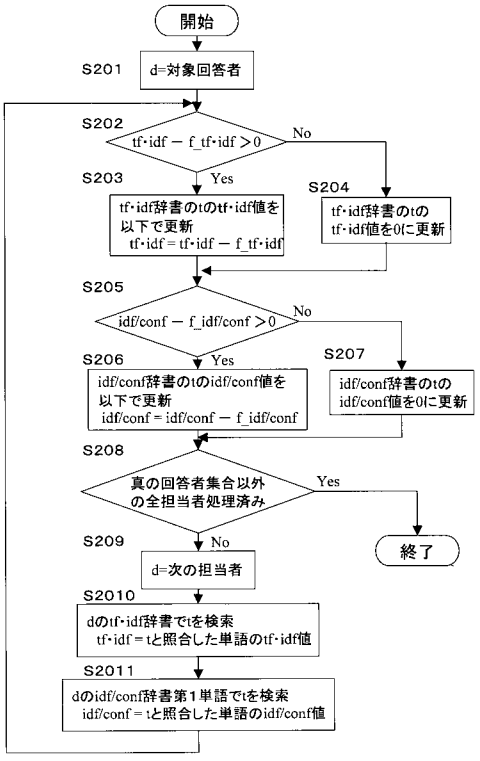
【図１３】



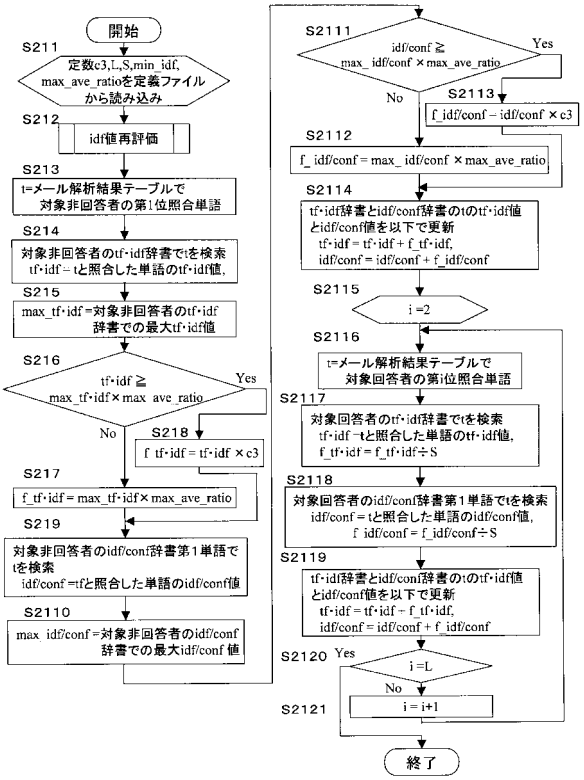
【図１４】



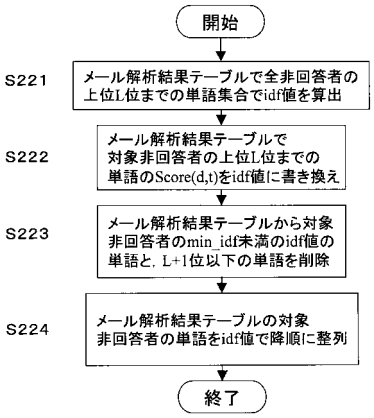
【図１５】



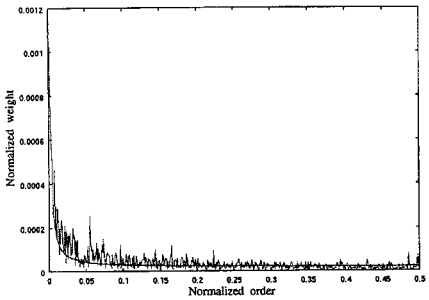
【図１６】



【図１７】


開始
S221
メール解析結果テーブルで全非回答者の上位L位までの単語集合でidf値を算出
S222
メール解析結果テーブルで対象非回答者の上位L位までの単語のScore(d,t)をidf値に書き換え
S223
メール解析結果テーブルから対象非回答者のmin_idf未満のidf値の単語と，L+1位以下の単語を削除
S224
メール解析結果テーブルの対象非回答者の単語をidf値で降順に整列
終了

【図１８】


Normalized weight
0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Normalized order

フロントページの続き

合議体
審判長　田口　英雄
審判官　飯田　清司
審判官　池田　聡史

(56)参考文献　特開平１１－１６７５８１（ＪＰ，Ａ）
特開２００１－２５６２５１（ＪＰ，Ａ）
特開２００２－１８３１７５（ＪＰ，Ａ）
上田他、テキストマイニングと強化学習を用いた電子メール自動分配、電子情報通信学会論文誌、Ｄ－Ｉ　Ｖｏｌ．Ｊ８７－Ｄ－Ｉ　Ｎｏ．１０、２００４年10月、ｐ８８７－８９８
「特別企画　日本語完全対応　ｅメール自動分類・配信ソリューション－ＭａｔｃｈＭａｉｌ－ＣａｌｌＣｅｎｔｅｒ－」、ビジネスコミュニケーション、Ｖｏｌ．３７　Ｎｏ．５、２０００年5月、ｐ．３６－４１