(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4686724号
(P4686724)
(45)発行日　平成23年5月25日(2011.5.25)　(24)登録日　平成23年2月25日(2011.2.25)

(51)Int.Cl.　　F I
H04L 12/58 (2006.01)　　H04L 12/58 100F
G06F 13/00 (2006.01)　　G06F 13/00 610Q

請求項の数 8　(全 10 頁)

(21)出願番号　特願2006-320004(P2006-320004)
(22)出願日　平成18年11月28日(2006.11.28)
(65)公開番号　特開2008-135926(P2008-135926A)
(43)公開日　平成20年6月12日(2008.6.12)
審査請求日　平成21年6月8日(2009.6.8)

(73)特許権者　304020177
国立大学法人山口大学
山口県山口市吉田１６７７－１
(72)発明者　杉井　学
山口県山口市吉田１６７７－１
(72)発明者　松野　浩嗣
山口県山口市吉田１６７７－１

審査官　安藤　一道

最終頁に続く

(54)【発明の名称】迷惑メールのフィルタ機能を有する電子メールシステム

(57)【特許請求の範囲】
【請求項１】
外部からの電子メールを受信する電子メール受信部と、
前記電子メール受信部によって受信された電子メールが迷惑メールか否かを判定する迷惑メール判定部と、
前記迷惑メール判定部の判定結果に応じて前記電子メールをフィルタリングする迷惑メールフィルタ部と、
前記迷惑メールフィルタ部によってフィルタリングされた前記電子メールをローカルメールボックス又は外部に送信する電子メール送信部と、
を有する電子メールシステムであって、
前記迷惑メール判定部は、
決定木学習部によって予め生成された単語出現頻度データベースにより、前記電子メール中の全ての単語を出現頻度に応じた符号に変換する単語符号化部と、
前記単語符号化部により符号化された電子メール符号化データに、決定木学習部によって予め生成された決定木を適用することにより迷惑メールか否かを判定する判定部と、
を有し、
前記決定木学習部は、
前記迷惑メール判定部と同一サーバ内又は異なるサーバ内にあり、
迷惑メールを保存した迷惑メールデータベースと、通常メールを保存した通常メールデータベースと、

前記迷惑メールデータベース及び前記通常メールデータベース内の電子メール中の単語の出願頻度を求めて前記単語出現頻度データベースを生成する単語出現頻度データベース生成部と、
前記単語出現頻度データベースにより、前記迷惑メールデータベース及び前記通常メールデータベース内の電子メール中の全ての単語を出現頻度に応じた符号に変換する単語符号化部と、
前記単語符号化部により符号化された電子メール符号化データのパターンに基づいて、迷惑メールと通常メールとを振り分ける最適な決定木を生成する学習部と、
を有し、
前記迷惑メール判定部及び前記決定木学習部における電子メールはヘッダ部分及び本文の両方を含むものであり、前記迷惑メール判定部及び前記決定木学習部は、前記電子メールにおけるヘッダ部分及び本文を分けずに同一アルゴリズムにより処理する
ことを特徴とする電子メールシステム。
【請求項２】
前記単語出現頻度データベース生成部は、単語の出現頻度とともに、前記単語が迷惑メールと通常メールのどちらに多く含まれるかを示す出現偏りも求めて前記単語出現頻度データベースを生成し、
前記単語符号化部は、前記電子メール中の全ての単語を前記出現頻度及び前記出現偏りに応じた符号に変換する
ことを特徴とする請求項１記載の電子メールシステム。
【請求項３】
前記学習部は、前記電子メール符号化データ内の符号を、最適な決定木を求められるグループに分け、前記グループ分けの結果により前記符号をさらに第２の符号に変換する機能を有することを特徴とする請求項１又は２記載の電子メールシステム。
【請求項４】
前記学習部に、ＢＯＮＳＡＩプログラムを用いることを特徴とする請求項３記載の電子メールシステム。
【請求項５】
外部からの電子メールを受信する電子メール受信部と、
前記電子メール受信部によって受信された電子メールが迷惑メールか否かを判定する迷惑メール判定部と、
前記迷惑メール判定部の判定結果に応じて前記電子メールをフィルタリングする迷惑メールフィルタ部と、
前記迷惑メールフィルタ部によってフィルタリングされた前記電子メールをローカルメールボックス又は外部に送信する電子メール送信部と、
を有する電子メールプログラムであって、
前記迷惑メール判定部は、
決定木学習部によって予め生成された単語出現頻度データベースにより、前記電子メール中の全ての単語を出現頻度に応じた符号に変換する単語符号化部と、
前記単語符号化部により符号化された電子メール符号化データに、決定木学習部によって予め生成された決定木を適用することにより迷惑メールか否かを判定する判定部と、
を有し、
前記決定木学習部は、
前記迷惑メール判定部と同一サーバ内又は異なるサーバ内にあり、
迷惑メールを保存した迷惑メールデータベースと、通常メールを保存した通常メールデータベースと、
前記迷惑メールデータベース及び前記通常メールデータベース内の電子メール中の単語の出願頻度を求めて前記単語出現頻度データベースを生成する単語出現頻度データベース生成部と、
前記単語出現頻度データベースにより、前記迷惑メールデータベース及び前記通常メー

ルデータベース内の電子メール中の全ての単語を出現頻度に応じた符号に変換する単語符号化部と、
前記単語符号化部により符号化された電子メール符号化データのパターンに基づいて、迷惑メールと通常メールとを振り分ける最適な決定木を生成する学習部と、
を有し、
前記迷惑メール判定部及び前記決定木学習部における電子メールはヘッダ部分及び本文の両方を含むものであり、前記迷惑メール判定部及び前記決定木学習部は、前記電子メールにおけるヘッダ部分及び本文を分けずに同一アルゴリズムにより処理する
ことを特徴とする電子メールプログラム。
【請求項６】
前記単語出現頻度データベース生成部は、単語の出現頻度とともに、前記単語が迷惑メールと通常メールのどちらに多く含まれるかを示す出現偏りも求めて前記単語出現頻度データベースを生成し、
前記単語符号化部は、前記電子メール中の全ての単語を前記出現頻度及び前記出現偏りに応じた符号に変換する
ことを特徴とする請求項５記載の電子メールプログラム。
【請求項７】
前記学習部は、前記電子メール符号化データ内の符号を、最適な決定木を求められるグループに分け、前記グループ分けの結果により前記符号をさらに第２の符号に変換する機能を有することを特徴とする請求項５又は６記載の電子メールプログラム。
【請求項８】
前記学習部に、ＢＯＮＳＡＩプログラムを用いることを特徴とする請求項７記載の電子メールプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、学習型の決定木アルゴリズムを用いた、迷惑メールのフィルタ機能を有する電子メールシステム及びプログラムに関する。
【背景技術】
【０００２】
インターネット上を流れる迷惑メールの割合は、全電子メール中の６０％を越えると言われており、現在では、迷惑メール対策として、さまざまな自動分類方法が用いられている。開発初期のもっとも単純な方法に、メールヘッダに記述される特定の迷惑メール送信サーバやメールのＦｒｏｍ行記載のメールアドレスを、管理者やユーザがひとつひとつ登録し、合致するメールを排除する方法がある。しかし迷惑メール送信者は、このような対策をかいくぐる新しい方法で次々に迷惑メールを送信してくるため、手作業で分類やアドレスの登録などを行うには作業コストが大きすぎ、現実的ではなくなってきている。また、これまでの方法では、通常の電子メールを迷惑メールと間違えて判断するケースおよびその逆のケースも増えている。近年、メール本文などの単語の出現頻度による特徴を分類に役立てるベイズ理論を応用した方法が注目されているが、未だ利用者および管理者の作業コストは大きく、分類精度もそれほど高くない。
【０００３】
従来技術として、特許文献１乃至３が挙げられる。
特許文献１には、文字列の一部をわざと間違えたり文字間に無意味な記号を挿入した電子メールであっても、迷惑メール等の電子メールを効果的に分類できる電子メール処理装置が記載されている。電子メールに含まれる単語について単語情報データベース内の迷惑メール対象文字列と相同性検索をすることによって迷惑メールの判定を行っている。
特許文献２には、電子メールのヘッダ情報に含まれるメール中継装置によって、該当電子メールが迷惑メールか否かをベイズ確率モデルを用いて判定する電子メールフィルタリングシステムが記載されている。

特許文献３には、ユーザが通常メールと迷惑メールとを分類し、その分類された内容を分析してフィルタルールを追加する電子メールフィルタリングシステムが記載されている。

特許文献１乃至３のいずれにも、迷惑メールの判定に学習型の決定木アルゴリズムを用いることついて記載されていない。

【特許文献１】特開２００６－２９３５７３号公報
【特許文献２】特開２００６－２６０５１５号公報
【特許文献３】特開２００６－２４５８１３号公報

【発明の開示】

【発明が解決しようとする課題】

【０００４】

本発明は、学習型の決定木アルゴリズムを用いて、迷惑メールを正確に効率良くフィルタリングする電子メールシステ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】

【０００５】

前記目的を達成するため、本発明は以下の構成を有する。

外部からの電子メールを受信する電子メール受信部と、前記電子メール受信部によって受信された電子メールが迷惑メールか否かを判定する迷惑メール判定部と、前記迷惑メール判定部の判定結果に応じて前記電子メールをフィルタリングする迷惑メールフィルタ部と、前記迷惑メールフィルタ部によってフィルタリングされた前記電子メールをローカルメールボックス又は外部に送信する電子メール送信部と、を有する電子メールシステム及びプログラムであって、前記迷惑メール判定部は、決定木学習部によって予め生成された単語出現頻度データベースにより、前記電子メール中の全ての単語を出現頻度に応じた符号に変換する単語符号化部と、前記単語符号化部により符号化された電子メール符号化データに、決定木学習部によって予め生成された決定木を適用することにより迷惑メールか否かを判定する判定部と、を有し、前記決定木学習部は、前記迷惑メール判定部と同一サーバ内又は異なるサーバ内にあり、迷惑メールを保存した迷惑メールデータベースと、通常メールを保存した通常メールデータベースと、前記迷惑メールデータベース及び前記通常メールデータベース内の電子メール中の単語の出願頻度を求めて前記単語出現頻度データベースを生成する単語出現頻度データベース生成部と、前記単語出現頻度データベースにより、前記迷惑メールデータベース及び前記通常メールデータベース内の電子メール中の全ての単語を出現頻度に応じた符号に変換する単語符号化部と、前記単語符号化部により符号化された電子メール符号化データのパターンに基づいて、迷惑メールと通常メールとを振り分ける最適な決定木を生成する学習部と、を有し、前記迷惑メール判定部及び前記決定木学習部における電子メールはヘッダ部分及び本文の両方を含むものであり、前記迷惑メール判定部及び前記決定木学習部は、前記電子メールにおけるヘッダ部分及び本文を分けずに同一アルゴリズムにより処理することを特徴とする電子メールシステム及びプログラム。

【０００６】

また、以下の実施態様を有する。

前記単語出現頻度データベース生成部は、単語の出現頻度とともに、前記単語が迷惑メールと通常メールのどちらに多く含まれるかを示す出現偏りも求めて前記単語出現頻度データベースを生成し、前記単語符号化部は、前記電子メール中の全ての単語を前記出現頻度及び前記出現偏りに応じた符号に変換する。

前記学習部は、前記電子メール符号化データ内の符号を、最適な決定木を求められるグループに分け、前記グループ分けの結果により前記符号をさらに第２の符号に変換する機能を有する。

前記学習部に、ＢＯＮＳＡＩプログラムを用いる。

【発明の効果】

【０００７】

学習型の決定木アルゴリズムを用いることで、従来のシステムに比べて、迷惑メールを正確に効率よくフィルタリングできる。また、決定木の学習及び適用の前に、電子メールを単語の出現頻度及び出現偏りに応じて符号化しておくことで、効果的に決定木の学習及び適用ができる。本発明のアルゴリズムは電子メールのヘッダ情報及び本文の両方に分け隔てなく適用でき、両方の情報を用いることでより簡単で正確に電子メールのフィルタリングが可能である。

決定木の学習には時間が掛かるが、予め生成された決定木に基づいて電子メールを分類するのは短時間でできる。本発明の決定木学習部と迷惑メール判定部とは独立して実行可能であるので、決定木を事前に学習しておいたり、決定木の学習を別サーバで実行することが可能である。迷惑メール判定部は、既に生成された決定木に基づいて電子メールを分類すればよいので、リアルタイムで電子メールのフィルタリングが可能である。

【発明を実施するための最良の形態】

【０００８】

図面を用いて本発明の実施形態について説明する。図１は、本電子メールシステムのブロック図である。電子メールシステム１は、インターネットから電子メールを受信する電子メール受信部２と、電子メール受信部２で受信された電子メールが迷惑メールか否かを判定する迷惑メール判定部３と、迷惑メール判定部３の判定結果に応じて電子メールをフィルタリングする迷惑メールフィルタ部４と、迷惑メールフィルタ部４によってフィルタリングされた電子メールをローカルメールボックス又は外部に送信する電子メール送信部５とからなる。迷惑メールフィルタ部４は、迷惑メールの削除、迷惑メールにフラグを付与、迷惑メールを別フォルダに移動などの動作を行う。電子メール送信部５は、本電子メールシステムの使用形態に応じて、フィルタリングされた電子メールを同一サーバ内のローカルメールボックスに振り分けて送信しても良いし、外部のメールサーバに転送しても良い。この電子メールシステムは、インターネットに接続されたサーバ上で動作させても良いし、電子メールを受信する端末上で動作させても良い。

【０００９】

図２は、迷惑メール判定部３のブロック図である。迷惑メール判定部３は、決定木学習部６によって予め生成された単語出現頻度データベース３３及び決定木３４と、単語出現頻度データベース３３により電子メール中の全ての単語を出現頻度に応じた符号に変換する単語符号化部３１と、単語符号化部３１により符号化された電子メール符号化データに決定木３４を適用することにより迷惑メールか否かを判定する判定部３２とからなる。単語出現頻度データベース３３及び決定木３４は決定木学習部６により予め生成しておき、必要に応じて転送しておくなどして迷惑メール判定部３で利用可能にしておく。迷惑メール判定部３と決定木学習部６とは同一サーバ上で実行させても良いし、それぞれ異なるサーバ上で実行させても良い。決定木学習部６で生成された単語出現頻度データベース３３及び決定木３４を、複数のサーバ上の迷惑メール判定部３で利用しても良い。

【００１０】

図３は、決定木学習部６のブロック図である。決定木学習部６は、迷惑メールを保存した迷惑メールデータベース６１と、通常メールを保存した通常メールデータベース６２と、迷惑メールデータベース６１及び通常メールデータベース６２内の電子メール中の単語の出願頻度を求めて単語出現頻度データベース３３を生成する単語出現頻度データベース生成部６３と、単語出現頻度データベース３３により迷惑メールデータベース６１及び前記通常メールデータベース６２内の電子メール中の全ての単語を出現頻度に応じた符号に変換する単語符号化部６４と、単語符号化部６４により符号化された電子メール符号化データのパターンに基づいて迷惑メールと通常メールとを分類する最適な決定木３４を生成する学習部６５とからなる。迷惑メールデータベース６１及び通常メールデータベース６２には、予めユーザによって分類された迷惑メール及び通常メールが蓄積されている。単語出現頻度データベース生成部６３は、単語の出現頻度及び出現偏り（単語が迷惑メールと通常メールのどちらに多く含まれるか）を求めて単語出現頻度データベース３３を生成する。単語符号化部６４は、単語出現頻度データベース３３に含まれる単語の出現頻度及

び偏りの情報から、各単語をＡ、Ｂ、Ｃなどの符号に変換する。迷惑メール判定部３内の単語符号化部３１も、単語符号化部６４と同様な動作を行う。学習部６５は、後述のＢＯＮＳＡＩプログラムを用いて決定木３４の生成を行う。迷惑メール判定部３及び決定木学習部６は、電子メールにおけるヘッダ部分及び本文を分けずに同一アルゴリズムにより処理する。

【００１１】

以下、ＢＯＮＳＡＩについて簡単に説明する（ＢＯＮＳＡＩの詳細については、Ｓｈｉｍｏｚｏｎｏ，Ｓ．、Ｓｈｉｎｏｈａｒａ、Ａ．，Ｍｉｙａｎｏ，Ｓ．，Ｋｕｈａｒａ，Ｓ．，Ａｒｉｋａｗａ，Ｓ．“Ｋｎｏｗｌｅｄｇｅ Ａｃｑｕｉｓｉｔｉｏｎ ｆｒｏｍ Ａｍｉｎｏ Ａｃｉｄ Ｓｅｑｕｅｎｃｅ ｂｙ Ｍａｃｈｉｎｅ Ｌｅａｒｎｉｎｇ Ｓｙｓｔｅｍ ＢＯＮＳＡＩ”，Ｔｒａｎｓ．Ｉｎｆｏｒｍ．Ｐｒｏｃｅｓｓ．Ｓｏｃ．Ｊａｐａｎ，３５（１０）：２００９－２０１８，１９９４参照）。ＢＯＮＳＡＩは、確率的近似学習と呼ばれる学習パラダイムに基づいて開発された機会学習プログラムで、正の学習グループと負の学習グループを与えると決定木を作成する。決定木の作成については、Ｊ．Ｒ．Ｑｕｉｎｌａｎの決定木学習アルゴリズムＩＤ３の枝狩り規準を改良した“Ｃ４．５”というアルゴリズムに基づいている。さらに、ＢＯＮＳＡＩはｉｎｄｅｘｉｎｇというグルーピングの機能を持っている。もともとＢＯＮＳＡＩは生物ゲノム情報から重要な遺伝子配列などを抽出する目的で開発された機械学習システムであるが、本発明者の工夫によって迷惑メールの分類に利用可能であることが見出された。ＢＯＮＳＡＩは、正の例と負の例として二つのデータ集団を入力すると、正の例には存在するが負の例には存在しないパターンを見つけ出すことができるので、この機能を利用して迷惑メールの分類を行う。

【００１２】

図４は、本システムの決定木学習の流れ図である。図４に示すように正の学習グループとして迷惑メール群、負の学習グループとして通常メール群を作成し、迷惑メール群に存在する特徴的なパターンの抽出を試みる。まず、両群の電子メールの文字列を単語に分解し、両群に存在するすべての単語について正の学習グループでの出現頻度を算出し、出現頻度の高いものからＡ～Ｅまでのグルーピングを行う。出現頻度を表すA～Eの文字で電子メール内のすべての文字列を置換してから、機械学習システムＢＯＮＳＡＩに投入する。ＢＯＮＳＡＩ（東京大学医科学研究所ヒトゲノム解析センター宮野研究室開発）は、正の学習グループと負の学習グループとして二つのデータ集団を入力すると、正の学習グループには存在するが、負の学習グループには存在しないといったパターンを見つけ出し、二つの学習例を正しく分けることができる決定木（Ｄｅｃｉｓｉｏｎ Ｔｒｅｅ）を作成する。また同時に、正の学習グループと負の学習グループを最も効率よく分類できる条件で、それぞれのグループ例を構成する要素もグループ分けする機能を持っている（Ｉｎｄｅｘｉｎｇ）。単語の出現頻度を反映させた学習グループ例をＢＯＮＳＡＩに投入することで、単語の出現頻度とその語順を考慮したパターン抽出が可能になる。つまりＢＯＮＳＡＩは、出現頻度を反映したＡ～Ｅの文字で置換された電子メール内の文字列を、図４のようにｉｎｄｅｘｉｎｇによってさらにグルーピングし、例えば０～２のような文字で置き換えながら、電子メール内の文字列中に存在するパターンを抽出する。また同時に正および負の学習グループ例を最も正しく分ける規則を提示する。

【００１３】

図５は、決定木の例である。例えば図４のケースでは、電子メールを単語分解及び２段階グルーピング（単語出現頻度、Ｉｎｄｅｘｉｎｇ）して０～２の符号に変換されたデータが、パターン「２０」を含んでいたら「迷惑メール」（正の学習グループ）と判定する。パターン「２０」を含んでいない場合は、さらにパターン「０２１」の検索を行い、パターン「０２１」を含んでいたら「迷惑メール」、含んでいなければ「通常メール」（負の学習グループ）と判定する。図４及び５は説明のための簡単な事例であるが、実際に利用する場合はパターン長はもっと長く、枝分岐ももっと複雑である。決定木学習及び決定木の適用には単語を符号化したものを利用するので、単語分解できるデータであれば何で

も利用可能であり、電子メールのヘッダ部分及び本文について同じアルゴリズムを適用できる。

【実施例】

【００１４】

以下、実施例について説明する。

決定木の学習手順は、以下の通りである。

１．サンプル電子メール（迷惑メール[正の例]：５００通、通常メール[負の例]：５００通）の準備。

２．サンプル電子メール（ヘッダ及び本文）を単語に分解。

３．単語の出現率と出現偏りの計算

・出現率＝ｌｏｇ（出現数の総和）

（出現率が小さいものは除外）

・出現偏り＝正の例での出現数／正の例及び負の例での出現数の総和

４．出現頻度に応じた符号化。

Ｘ：　０．８＜（出現偏り）

Ｙ：　０．６≦（出現偏り）≦０．８

Ｚ：　　（出現偏り）＜０．６

Ｏ：　その他[出現数少]

５．ＢＯＮＳＡＩにより最適な決定木の生成。

図６に、ＢＯＮＳＡＩにより生成された決定木の例を示す。この例では、ＢＯＮＳＡＩのグルーピング機能（ｉｎｄｅｘｉｎｇ）により、Ｘ→０、Ｙ→０、Ｚ→１、Ｏ→１のさらなる符号化が行われている。

【００１５】

生成された決定木に基づいて、７１２通の一般の受信メールを振り分けてみた結果は以下の通りである。

通常メール分類の正解率：９４．４％（２３８／２５２通）

迷惑メール分類の正解率：９７．８％（４５０／４６０通）

この結果から、高い正解率で迷惑メールと通常メールの振り分けが可能であることがわかる。

【００１６】

別の実施例について説明する。前述の実施例では、単語出現頻度による符号化の符号数は４個（Ｘ，Ｙ，Ｚ，Ｏ）、ＢＯＮＳＡＩのグルーピング機能（ｉｎｄｅｘｉｎｇ）による符号化の符号数は２個（０，１）であったが、単語出現頻度による符号化の符号数を６個（Ｘ，Ｙ，Ｚ，Ｏ，Ａ，Ｂ）、ＢＯＮＳＡＩのグルーピング機能（ｉｎｄｅｘｉｎｇ）による符号化の符号数を３個（０，１，２）にした場合の決定木の例を図７に示す。

この決定木に基づいて、８０６通の一般の受信メールを振り分けてみた結果は以下の通りである。

通常メール分類の正解率：９６．１％（２７３／２８４通）

迷惑メール分類の正解率：９８．６％（５１５／５２２通）

前述の実施例よりもさらに高い正解率であることがわかる。

「単語出現頻度による符号化の符号数」及び「ＢＯＮＳＡＩのグルーピング機能（indexing）による符号化の符号数」はこの他の組み合わせも可能であり、演算速度、サンプル電子メール数、学習に掛けられる時間等に応じて任意に設定できる。

【００１７】

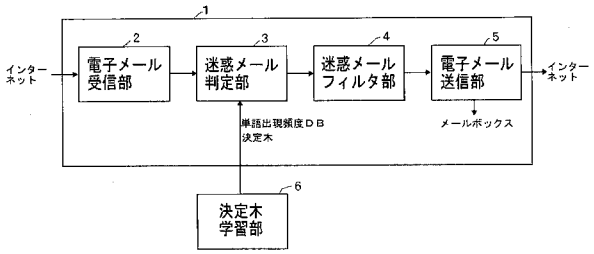
以上、本発明の実施形態の一例を説明したが、本発明はこれに限定されるものではなく、特許請求の範囲に記載された技術的思想の範疇において各種の変更が可能であることは言うまでもない。

【図面の簡単な説明】

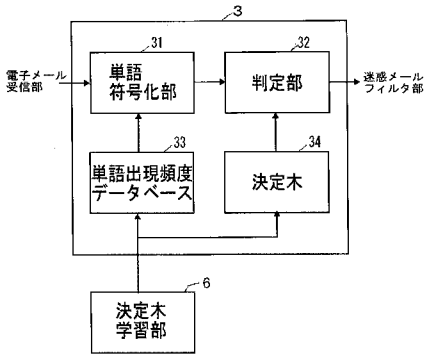
【００１８】

【図１】本システムのブロック図
【図２】迷惑メール判定部のブロック図
【図３】決定木学習部のブロック図
【図４】決定木学習の流れ図
【図５】決定木の例
【図６】実施例における決定木
【図７】別の実施例における決定木
【符号の説明】
【００１９】
１：電子メールシステム、　２：電子メール受信部、　３：迷惑メール判定部、　４：迷惑メールフィルタ部、　５：電子メール送信部、　６：決定木学習部、
３１：単語符号化部、　３２：判定部、　３３：単語出現頻度データベース、　３４：決定木、
６１：迷惑メールデータベース、　６２：通常メールデータベース、　６３：単語出現頻度データベース生成部、　６４：単語符号化部、６５：学習部（ＢＯＮＳＡＩ）

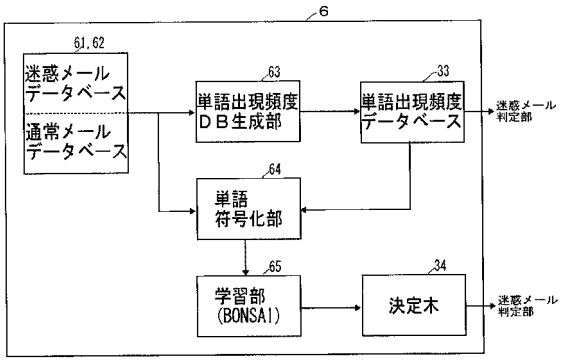
【図１】



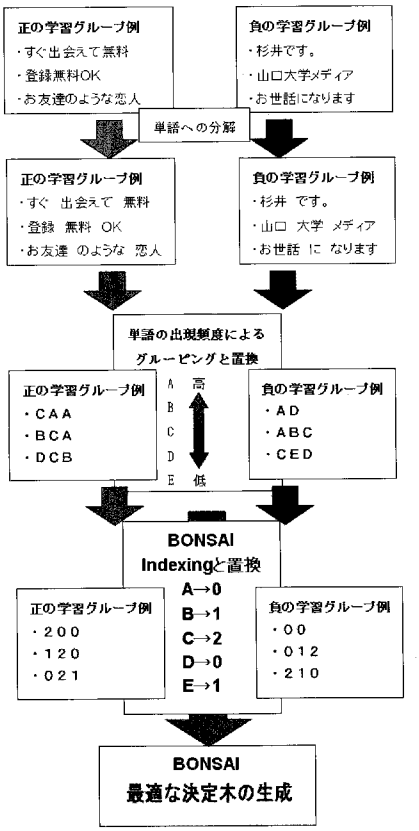
【図２】



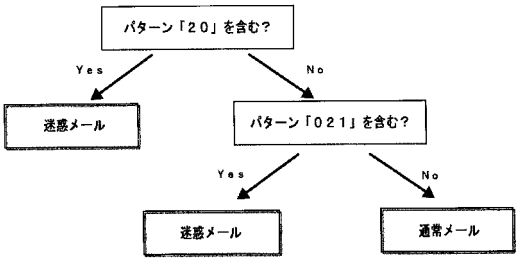
【図３】



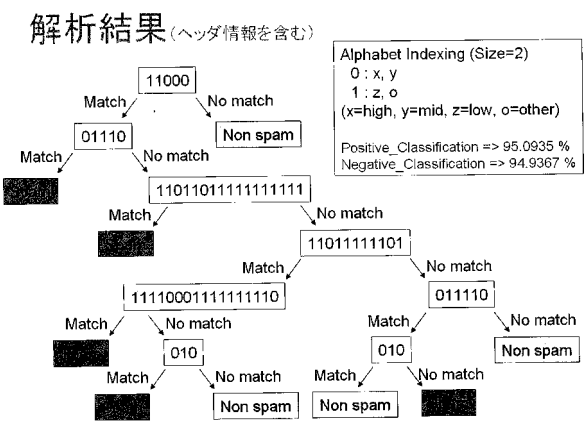
【図４】


正の学習グループ例
・すぐ出会えて無料
・登録無料OK
・お友達のような恋人
負の学習グループ例
・杉井です。
・山口大学メディア
・お世話になります
単語への分解
正の学習グループ例
・すぐ 出会えて 無料
・登録 無料 OK
・お友達 のような 恋人
負の学習グループ例
・杉井 です。
・山口 大学 メディア
・お世話 に なります
単語の出現頻度による
グルーピングと置換
正の学習グループ例
・CAA
・BCA
・DCB
A 高
B
C
D
E 低
負の学習グループ例
・AD
・ABC
・CED
BONSAI
Indexingと置換
A→0
B→1
C→2
D→0
E→1
正の学習グループ例
・200
・120
・021
負の学習グループ例
・00
・012
・210
BONSAI
最適な決定木の生成

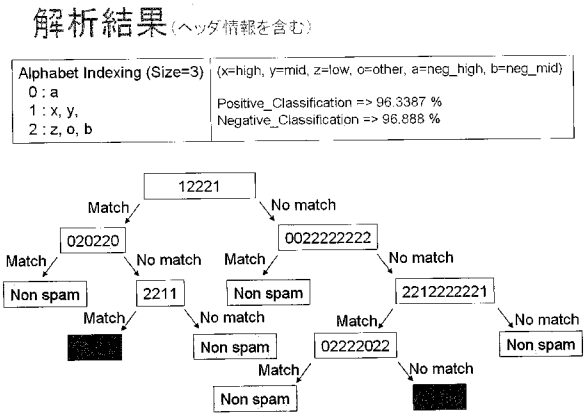
【図５】


パターン「20」を含む?
Yes
No
迷惑メール
パターン「021」を含む?
Yes
No
迷惑メール
通常メール

【図６】


解析結果(ヘッダ情報を含む)
Alphabet Indexing (Size=2)
0 : x, y
1 : z, o
(x=high, y=mid, z=low, o=other)
Positive_Classification => 95.0935 %
Negative_Classification => 94.9367 %
11000
Match
No match
01110
Non spam
Match
No match
11011011111111111
Match
No match
11011111101
Match
No match
11110001111111110
011110
Match
No match
010
Match
No match
Non spam
Match
No match
Non spam
Non spam

【図７】


解析結果(ヘッダ情報を含む)
Alphabet Indexing (Size=3)
0 : a
1 : x, y,
2 : z, o, b
(x=high, y=mid, z=low, o=other, a=neg_high, b=neg_mid)
Positive_Classification => 96.3387 %
Negative_Classification => 96.888 %
12221
Match
No match
020220
0022222222
Match
No match
Match
No match
Non spam
2211
Non spam
2212222221
Match
No match
Match
No match
Non spam
02222022
Non spam
Match
No match
Non spam

フロントページの続き

(56)参考文献 特開２００６－２９３５７３（ＪＰ，Ａ）
特開２００６－２６０５１５（ＪＰ，Ａ）
特開２００６－２４５８１３（ＪＰ，Ａ）
特開２００４－３４８５２３（ＪＰ，Ａ）
Shinichi Shimozono et al, Knowledge Acquisition from Amino Acid Sequences by Machine Learning System BONSAI, pp.1-18
大福、松浦，ベイジアンフィルタと社会ネットワーク手法を統合した迷惑メールフィルタリングとその最適統合法，情報処理学会論文誌，社団法人情報処理学会，２００６年 ８月，Vol.47,No.8, pp.2548-2555

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０４Ｌ １２／５８
Ｇ０６Ｆ １３／００