(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4297345号
(P4297345)
(45)発行日　平成21年7月15日(2009.7.15)　(24)登録日　平成21年4月24日(2009.4.24)

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

請求項の数 4　(全 12 頁)

(21)出願番号　特願2004-6918(P2004-6918)
(22)出願日　平成16年1月14日(2004.1.14)
(65)公開番号　特開2005-202590(P2005-202590A)
(43)公開日　平成17年7月28日(2005.7.28)
審査請求日　平成18年4月27日(2006.4.27)

(73)特許権者　000208891
ＫＤＤＩ株式会社
東京都新宿区西新宿二丁目３番２号
(73)特許権者　504016525
吉田　健一
埼玉県北本市高尾２－２３２
(74)代理人　100084870
弁理士　田中　香樹
(74)代理人　100079289
弁理士　平木　道人
(74)代理人　100119688
弁理士　田邊　壽二
(72)発明者　山崎　克之
埼玉県上福岡市大原二丁目１番１５号　株式会社　ＫＤＤＩ研究所内

最終頁に続く

(54)【発明の名称】マスメイル検出方式およびメイルサーバ

(57)【特許請求の範囲】
【請求項１】
配送対象の電子メイルを収集する電子メイル収集手段と、
該収集した電子メイルを特徴量に変換する特徴量変換手段と、
該変換した特徴量を使ってマスメイルを検出するマスメイル検出手段とを具備し、
前記特徴量変換手段は電子メイルの本文から部分文字列を抽出し、その部分文字列から計算したハッシュ値の集まりを特徴量として用い、
前記マスメイル検出手段は、特徴量データベースと特徴量データベースへのポインタを具備し、新規電子メイルの特徴量のハッシュ値が該ポインタにエントリされているか否かを判断し、エントリされている場合には、該ポインタを用いて特徴量データベースをアクセスし、該特徴量データベースに既に登録されている電子メイルのハッシュ値と比較することにより、新規電子メイルと既登録の電子メイルとの類似度を判定し、一定数以上のハッシュ値が一致した電子メイルを類似電子メイルと判定し、該類似電子メイルが所定数検出された時に該類似電子メイルをマスメイルと判定することを特徴とするマスメイル検出方式。
【請求項２】
前記特徴量データベースには、類似メイル数と、前記ポインタにより参照される被参照数と、電子メイルのハッシュ値とが登録され、
類似メイルが受信された時には、既登録の類似メイルにおける類似メイル数が１増加し、被参照数が前記類似メイルと既登録の類似メイルとのハッシュ値の一致した数に応じて

増加することを特徴とする請求項１に記載のマスメイル検出方式。
【請求項３】
前記マスメイル検出手段が、directed map cache方式（管理マップキャッシュ方式）またはLRU方式を利用する事を特徴とする請求項２に記載のマスメイル検出方式。
【請求項４】
前記請求項１ないし３のいずれかのマスメイル検出方式を備えたメイルサーバ。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は電子メイルのスパム処理に係わり，特に携帯電話やＩＳＰなど、大規模な電子メイルサーバを運用する事業者が電子メイルサーバを経由して配送される電子メイルに含まれている未承諾広告などの迷惑メイルを検出するのに好適なマスメイル検出方式および該マスメイル検出方式を備えたメイルサーバに関する。
【背景技術】
【０００２】
電子メイルの普及に従い電子メイルを搬送手段とした迷惑メイルが増加し、社会問題となっている。従来、このような迷惑メイルを防止する手段としては、電子メイルの受信者が受信に用いる端末に迷惑メイルを検知する仕組を用意し、その仕組によって迷惑メイルを自動的に削除するなどの方法が一般的であった。
【０００３】
例えば、SpamAssassinはルールベース方式を用いたソフトであり、bogofilterは機械学習方式を用いたソフトであり、どちらも主としてPCユーザの間で有効な仕組として使われている。なお、これらのソフトは、それぞれ、非特許文献１，２に示されている。
【非特許文献１】http://www.au.spamassassin.org/presentations/SAGE_IE_2002/
【非特許文献２】http://bogofilter.sourceforge.net/bogofilter_man.html
【発明の開示】
【発明が解決しようとする課題】
【０００４】
上記した従来技術は、電子メイルの受信者がPCのように一定の水準以上の情報処理能力を持つ受信端末を使用することが前提であり、携帯電話などのような比較的低い能力の受信端末には不向きである。携帯電話などのような比較的低い能力の受信端末を支援するには、事業者側のメイルサーバにマスメイルを検出する手段を備える事が望ましかった。
【０００５】
しかし、上記の従来技術は、サーバで用いるには処理速度が遅く大規模な設備を必要とするという問題点があった。また、大人数のユーザに共通したマスメイルの検知ルールや機械学習結果を作成するのが困難であり、かつ、新種のスパムに対応するための維持管理のコストが膨大であるといった問題点もあった。
【０００６】
本発明の目的は、前記した問題点を解決するために、事前のルール作成や学習が不要で、かつ、高速に動作するマスメイル検出方式および該マスメイル検出方式を備えたメイルサーバを提供することにある。
【課題を解決するための手段】
【０００７】
上記した目的を達成するために、本発明は、配送対象の電子メイルを収集する電子メイル収集手段と、該収集した電子メイルを特徴量に変換する特徴量変換手段と、該変換した特徴量を使ってマスメイルを検出するマスメイル検出手段とを具備し、前記特徴量変換手段は電子メイルの本文から部分文字列を抽出し、その部分文字列から計算したハッシュ値の集まりを特徴量として用い、前記マスメイル検出手段は、特徴量データベースと特徴量データベースへのポインタを具備し、新規電子メイルの特徴量のハッシュ値が該ポインタにエントリされているか否かを判断し、エントリされている場合には、該ポインタを用い

て特徴量データベースをアクセスし、該特徴量データベースに既に登録されている電子メイルのハッシュ値と比較することにより、新規電子メイルと既登録の電子メイルとの類似度を判定し、一定数以上のハッシュ値が一致した電子メイルを類似電子メイルと判定し、該類似電子メイルが所定数検出された時に該類似電子メイルをマスメイルと判定するようにした点に第１の特徴がある。

【０００８】

また、本発明は、前記マスメイル検出手段が、記憶領域に記憶しておく電子メイルとして頻繁に配送される電子メイルを優先的に記憶するための手段をもち、該手段としてdirected map cache方式またはLRU方式を用いるようにした点に第２の特徴がある。

【発明の効果】

【０００９】

本発明によれば，携帯電話やISPなど、大規模な電子メイルサーバを運用する事業者が電子メイルサーバを経由して配送される電子メイルに含まれている未承諾広告などの迷惑メイルを検出するのに好適なマスメイル検出方式を提供できる。

【発明を実施するための最良の形態】

【００１０】

以下に、図面を参照して本発明を詳細に説明する。図１は、本発明が適用されるシステム構成の一例を示すブロック図である。

【００１１】

図において、１は例えば通信事業者（プロバイダ）のメイルサーバ群、２はインターネット、３はこれらの間のデータ配送に用いられるスイッチングハブであり、メイルサーバ群１はユーザ端末４と接続されている。また、スイッチングハブ３には、本発明のマスメイル検出装置５が接続されている。

【００１２】

本実施形態は、メイルサーバ群１とインターネット２の間でSMTPプロトコルを用いて配送される電子メイルの中から、前記マスメイル検出装置５を用いてマスメイルを検出するものである。

【００１３】

該マスメイル検出装置５は、電子メイル収集手段５１，特徴量変換手段５２およびマスメイル検出手段５３から構成されている。電子メイル収集手段５１は配送対象の電子メイルを収集するものであり適切な計算機上のプログラムで良い。特徴量変換手段５２は電子メイル収集手段５１で収集した電子メイルを特徴量に変換するものであり適切な計算機上のプログラムで良い。マスメイル検出手段５３は変換した特徴量を使ってマスメイルを検出するものであり適切な計算機上のプログラムで良い。５５は検出結果のマスメイルである。

【００１４】

次に、本実施形態の動作を説明する。電子メイル収集手段５１は、ネットワーク上に流れる電子メイル配送プロトコルを解析し、ネットワーク上に流れる電子メイルトラフィックから電子メイル本文を抽出する。次に、特徴量変換手段５２が電子メイル本文から、例えば幾つかのハッシュ値を計算し該メイルの特徴量とする。最後に、マスメイル検出手段５３が記憶しておいた過去の電子メイルと新たに受信した電子メイルとを、前記特徴量を用いて比較し、特定の基準に従って類似度を判定し、類似している場合はマスメイルの候補（類似メイル）として判定し、一定数以上の類似メイルが検出されるとこれをマスメイルと判定する。

【００１５】

図２は、前記電子メイル収集手段５１の処理手続きの一例を示すフローチャートである。メイルサーバ群１とインターネット２の間では複数のメイルが並行して配送されている。そこで、ステップＳ１０では、電子メイル収集手段５１は、スイッチングハブ３でタッピングすることでＴＣＰパケットを受信する。該受信したＴＣＰパケットは、複数の電子メイルの情報が混ったものである。ステップＳ１５では、電子メイル収集手段５１は、パ

ケットの種類を判断する。すなわち、パケットがメイルであるか否か、メイルであれば、新規メイルのパケットであるか、処理中メイルのパケットであるか、あるいは処理中メイルの終了パケットであるかの判断をする。

【００１６】

そして、受信したパケットが新規メイルのものであれば、ステップＳ１１に進んで、新規メイル用記憶領域を初期設定する。一方、受信したパケットが処理中のメイルの終了を意味するパケットであれば、ステップＳ１３に進み、処理中のメイル本文を特徴量変換手段５２に送信し、次にステップＳ１４に進んで、処理中のメイル用記憶領域を廃棄／解放する。また、受信したパケットが終了以外の処理中のメイルパケットであれば、ステップＳ１２に進んで、処理中のメイル用記憶領域にＴＣＰパケットに含まれるメイルの内容を記録する。前記ステップＳ１５で、メイル以外のパケットであると判断された場合には、何も処理を行わず終了する。図２では便宜的にエンド（終了）と記したが、実際の処理は終了することなく、ステップＳ１０～Ｓ１４の処理が継続的に繰り返されることは明らかである。

【００１７】

図３は前記特徴量変換手段５２の処理の一例を示す説明図であり、図４は該特徴量変換手段５２の処理手続きの一例を示すフローチャートである。

【００１８】

本実施形態においては、メイル本文の特徴量として、事前に定めた長さＬの文字列（例えば、４文字）のハッシュ値の集合を用いる。具体的には、図４の手順に従い、始にステップＳ２１にて、メイル本文１００の先頭から順番にＬ文字づつ取り出し、そのハッシュ値を計算する。次に、ステップＳ２２に進み、計算したハッシュ値をソートし、始めのＮ個（例えば１００個）を特徴量としてマスメイル検出手段５３に送信する。

【００１９】

例えば、図３に示されているように、長さＬが４で、メイル本文１００が「メイル本文の文章」であったとすれば、「メイル本」（図３の１０１）、「イル本文」、「ル本文の」、「本文の文」（図３の１０２）などのハッシュ値２０１～２０２を計算する（図４のステップＳ２１）。次いで、該ハッシュ値２０１～２０２をソートした後、始めのＮ個を特徴量２００としてマスメイル検出手段５３に送信する（図４のステップＳ２２）。該ハッシュ値としては、例えば６４ビットの整数で表すことができる。

【００２０】

図５は前記マスメイル検出手段５３が利用するデータ構造の例である。３００は、頻繁に配送される電子メイルを特徴量データベース３１０の中に優先的に記憶するためのデータ構造Directed Map Cache（以下、DMC）、すなわち管理マップキャッシュ方式を示す。該ＤＭＣ３００は、特徴量データベース３１０と、該特徴量データベース３１０へのポインタ３１１を有する。特徴量データベース３１０は、電子メイル毎に、特徴量（ハッシュ値１～Ｎ）、該当電子メイルの類似メイル数、およびポインタ３１１のエントリで該当メイルを参照しているポインタの数(DMC被参照数）を記憶した計算機上のデータ構造である。該ポインタ３１１の各々は、例えば６４ビットで表現することができる。

【００２１】

電子メイル収集手段５１が電子メイルを抽出すると、特徴量変換手段５２がその電子メイル本文から特徴量２００（図３参照）を計算し、最後にマスメイル検出手段５３が図６に例示した手順に従い、多量に配送されている類似したメイルをマスメイルとして検知する。具体的には、１通の電子メイルに対して特徴量変換手段５２が計算する特徴量２００はＮ個（Ｎは、正の整数）のハッシュ値を持つが、マスメイル検出手段５３は各メイルに対して図６の手順に従いステップＳ３１からステップＳ４１までの処理を最大Ｎ回繰り返す。

【００２２】

ステップＳ３０では、前記特徴量２００を基に、前記電子メイル収集手段５１で収集された電子メイルに類似するメイルが既にあるか否かの判定が行われる。この処理の一具体

例を、図７のフローチャートを参照して説明する。

【００２３】

ステップＳ３０１では、前記特徴量２００の番号を示す数ｍを１と置き、ステップＳ３０２では新規のメイルの特徴量２００の中のｍ番目のハッシュ値を抽出する。ステップＳ３０３では、該ハッシュ値がポインタ３１１にエントリされているか否かの判断がなされる。この判断が肯定の場合にはステップＳ３０４に進んで、現在のポインタ３１１から参照されている特徴量データベース３１０中のエントリと類似度の判定を行う。そして、例えば８０％の類似があれば類似メイル、８０％より小さければ非類似メイルと判定する。ステップＳ３０５では、ｍ＝Ｎが成立したか否かの判断がなされ、否定の場合には、ステップＳ３０６に進んでｍに１が加算される。次に、ステップＳ３０２に戻って、２番目のハッシュ値が抽出される。以下、同様にして、前記した処理が繰り返し行われ、ステップＳ３０５の判断が肯定になると、前記ステップＳ３０の処理は終了する。

【００２４】

前記ステップＳ３０４の類似度判定は、例えば新規の電子メイルのハッシュ値２００（図３参照）と特徴量データベース３１０内のハッシュ値が一致した数を利用する。例えば、特徴量の数Ｎが１００個、類似度の閾値が８０％の時、８０個のハッシュ値が一致すると類似のメイルと判断する。一致した数の計測処理を速めるために、予め前記ハッシュ値はソートしておくと好適である。なお、図７ではＮ個のハッシュ値について類似度を判定したが、必ずしもＮ個のハッシュ値について類似度を判定する必要はなく、Ｎ個より少ないハッシュ値で類似度を判定しても良い。

【００２５】

図６に戻って説明を続けると、ステップＳ３１において前記ｍを再度ｍ＝１とし、ステップＳ３２において、特徴量２００中のｍ番目のハッシュ値を抽出する。次いで、ステップＳ３３に進み、該ｍ番目のハッシュ値が類似メイルのハッシュ値であるか否かの判断がなされる。この判断が否定、すなわち新規の電子メイル（非類似メイル）の場合には、類似メイルは特徴量データベース３１０に記憶されていないので、ステップＳ３４に進んで、該当メイルの特徴量が特徴量データベース３１０の新規エントリとして登録される。具体的には、新規電子メイルの特徴量２００（図３参照）を特徴量データベース３１０のハッシュ値１～ハッシュ値Ｎとして記憶する。次に、ステップＳ３５に進み、特徴量のベクトル値でDMC３００の内容を更新する。

【００２６】

前記ステップＳ３３の判断が肯定の場合、すなわち類似メイルがある場合には、ステップＳ３７に進む。該ステップＳ３７では、特徴量データベース３１０に既に記憶されているメイルの類似メイル数（図５参照）を１加算する。次に、ステップＳ３８に進んで、該特徴量データベース３１０に記憶されたベクトル値でDMC３００の内容を更新する。なお、該ステップＳ３８は前記ステップＳ３５と同一の処理であり、その具体例を図８を参照して後述する。

【００２７】

ステップＳ３９では、前記類似メイル数が予め定められた値Ｓ以上になったか否かの判断がなされ、Ｓ以上になった場合にはステップＳ４０に進んで該当メイルをスパムと判定する。一方、ステップＳ３９の判断が否定の時にはステップＳ３６に進む。ステップＳ３６では、ｍ＝Ｎになったか否かの判断がなされ、この判断が否定の時には、ステップＳ４１に進んで、ｍに１が加算される。そして、再度ステップＳ３２からの動作が繰り返される。

【００２８】

図８は図６のステップＳ３５、Ｓ３８のDMC更新処理の手順を例示したものである。マスメイル検出手段５３は各電子メイルの処理にあたり、まずステップＳ３５１の判断をする。すなわち、当該ハッシュ値は、現在のポインタ３１１から特徴量データベース３１０の古いエントリを参照しているか否かの判断をする。この判断が否定の時には、ステップＳ３５２に進んで、ポインタ３１１の対応エントリが新しい特徴量データベース３１０の

エントリを指すように設定し、該特徴量データベースの被参照数に１を加算する。

【００２９】

一方、前記ステップＳ３５１の判断が肯定の時、すなわち当該ハッシュ値が現在のポインタ３１１から特徴量データベース３１０の古いエントリを参照している時には、ステップＳ３５３に進んで、該ハッシュ値が自分自身のエントリを参照しているか否かの判断をする。すなわち、当該ハッシュ値が前記類似メイルの中に含まれているか否かの判断をする。この判断が肯定の時には、何の処理も行わずに図６の処理に抜ける。

【００３０】

ステップＳ３５３の判断が否定の時、すなわち当該ハッシュ値が前記類似メイルの中に含まれていない時には、ステップＳ３５４に進み、現在のポインタ３１１から参照されている特徴量データベース３１０中の古いエントリのDMC被参照数を１減算する。次いで、ステップＳ３５５に進み、DMC被参照数が０であるか否かの判断がなされる。この判断が肯定の時には、ステップＳ３５６に進んで、DMC被参照数が０になった過去のメイルのエントリを、特徴量データベース３１０から削除する。前記ステップＳ３５５の判断が否定の時には前記ステップＳ３５２に進み、ポインタ３１１の対応エントリが新しい特徴量データベース３１０のエントリを指すように設定すると共に、該特徴量データベースの被参照数を１加算する。

【００３１】

以上の処理によると、類似したメイルが多いメイルは頻繁に図６のステップＳ３８から起動されて図８の更新処理（より具体的には、ステップＳ３５２）が動くのでDMC被参照数は０になりにくいが、類似メイルがないものはハッシュ値がぶつかったデータを上書きする事で時間の経過とともにDMC被参照数が減少し、前記ステップＳ３５６で最終的には削除される。

【００３２】

次に、前記した図６～図８の動作を具体例を、図９～図１３を参照して参照して説明する。今、インターネットを介して新規のメイルが図９に示されているように、メイル１、２、３、４の順に収集されたものとし、該メイルの特徴量（前記図３の特徴量２００）が、メイル１に関してはハッシュ値ｈ１，ｈ２，ｈ３，ｈ４、メイル２に関してはハッシュ値ｈ２，ｈ３，ｈ６，ｈ７、メイル３に関してはハッシュ値ｈ４，ｈ８，ｈ９，ｈ０、メイル４に関してはハッシュ値ｈ１，ｈ２，ｈ３，ｈ０であるとする。ここで、類似メイルと判定する基準を７５％以上の一致とすると、メイル４はメイル１と類似になる。この判定は図７の処理により行われる。なお、ここでは、説明を簡単にするために、各メイルの特徴量が４個であるとした。

【００３３】

さて、まずインターネットを介してメイル１が抽出されると、図６のステップＳ３３の判断は否定になるので、ステップＳ３４，Ｓ３５の処理が行われる。ステップＳ３４の処理により特徴量データベース３１０は図１０（ｂ）のハッシュ値１～４にｈ１～ｈ４が登録され、ステップＳ３５の処理によりDMC３００のポインタ３１１は同図（ａ）のようになると共に、DMC被参照数が４となる。

【００３４】

次に、メイル２が抽出されると、前記ステップＳ３３の判断は否定になるので、ステップＳ３４，Ｓ３５の処理に進む。ステップＳ３４の処理により特徴量データベース３１０は図１１（ｂ）のメイル２のハッシュ値１～４にｈ２、ｈ３、ｈ６、ｈ７が登録され、ステップＳ３５の処理によりDMC３００のポインタ３１１は同図（ａ）のようになると共に、メイル１のDMC被参照数が２となり、メイル２のDMC被参照数が４となる。

【００３５】

続いて、メイル３が抽出されると、前記ステップＳ３３の判断は否定になるので、ステップＳ３４，Ｓ３５の処理が行われる。ステップＳ３４の処理により特徴量データベース３１０は図１２（ｂ）のように、メイル３のハッシュ値１～４にｈ４、ｈ８、ｈ９、ｈ０が登録され、ステップＳ３５の処理によりDMC３００のポインタは同図（ａ）のようにな

ると共に、メイル１，２，３のDMC被参照数がそれぞれ、１，４，４となる。

【００３６】

さらに、メイル４が抽出されると、このメイル４は既登録のメイル１と類似するものであるので、前記ステップＳ３３の判断は肯定になり、ステップＳ３７，Ｓ３８の処理が行われる。ステップＳ３７の処理により特徴量データベース３１０のメイル１の類似メイル数に１が加算され、図１３（ｂ）のようになる。また、ステップＳ３８の処理により、DMC３００のポインタは同図（ａ）のようになると共に、メイル１，２，３のDMC被参照数はそれぞれ、４，２，３となる。

【００３７】

つまり、類似メイルが到着すると、前記ステップＳ３７でメイル１の類似メイル数に１が加算される。次に、ステップＳ３８の処理、つまり図８の処理において、ハッシュ値ｈ１はポインタ３１が自分自身のメイル１を指しているのでステップＳ３５３の判断は肯定になり図８の処理を抜ける。次のハッシュ値ｈ２，ｈ３は、ポインタが共にメイル２を指しているので、ステップＳ３５３の判断は否定となり、ステップＳ３５４以下の処理に移る。そして、ステップＳ３５４でメイル２のDMC被参照数が１減算され、ステップＳ３５２でメイル１のDMC被参照数が１加算される。次のハッシュ値ｈ０についても、同様に処理される。

【００３８】

以上のようにして、類似したメイルが多いメイルは頻繁に図６のステップＳ３８から起動されて図８の更新処理が起動され、DMC被参照数が増加する。一方では、メイル２を見れば明らかなように、類似メイルがないものはハッシュ値がぶつかったデータを上書きする事で時間の経過とともにDMC被参照数が減少する。

【００３９】

なお、本発明は前記実施形態に限定されることなく、次のように変形することも可能である。上記の実施形態においては、電子メイル収集手段５１はネットワーク上に流れる電子メイルをスイッチングハブ３でタッピングする事で収集していたが、メイルサーバのソフトを変更し、メイルサーバが配送対象のメイルを直接特徴量変換手段５２に送信するようにしてもよい。またメイルの配送プロトコルはSMTPを想定していたが、HTTPを用いたWWWメイルのような別の配送形態であってもかまわない。

【００４０】

また、メイルサーバが配送対象のメイルを特徴量変換手段５２に送信する時に既にスパムと判定したメイルについてはスパムであるとのマークをつけて送り、その情報を使ってマスメイル検出手段５３が、マークのついたメイルと類似したメイルは即座にスパムと判定してもかまわない。また、メイルサーバは前記特徴量変換手段５２までを含むように構成し、該特徴量変換手段５２にて変換された特徴量がネットワーク経由でマスメイル検出手段５３に送信される構成にしてもよい。

【００４１】

また、上記の実施形態においては、マスメイル検出手段５３が、記憶領域に頻繁に配送される電子メイルを優先的に記憶するための仕組みとしてDMC３００（図５参照）を利用していたが、LRU方式のような別の仕組みを利用するのでもかまわない。LRUを使う場合、具体的には特徴量データベース３１０のエントリを管理するLRUリストを作成し、前記ステップＳ３７（図６参照）で処理対象とした特徴量データベース３１０のエントリをLRUの先頭に移動する処理までステップＳ３７に含める。また、ステップＳ３４で特徴量データベース３１０に新規のエントリを作る時に必要な記憶領域はLRUの最後のエントリを廃棄して確保し、新規のエントリをLRUの先頭に加える。

【００４２】

また、上記の実施形態においては、特徴量変換手段５２の前処理については説明しなかったが、図１の電子メイル収集手段５１と特徴量変換手段５２との間に前処理手段を設けても良い。この前処理手段は、文字列を抽出する手段であってもよく、この前処理手段により、メールアドレス、電話番号などを抽出するようにしてもよい。また、その他の何ら

かの前処理を行う手段でもかまわない。この前処理は、電子メイルが受信者に表示されるときの仕様に従って行うものであってもかまわない。この前処理は、例えば受信者の端末の初めの部分（例えば、始めの２頁分）に表示される文字を処理の対象として選択するのでもかまわない。この時に想定される仕様としては、表示に影響を与えるHTMLやMIMEの処理などが上げられるが、これ以外でもかまわない。また、表示の時に大文字と小文字、全角文字と半角文字などを似た文字として扱い、同じ特徴量が計算される仕組みを持つ（例えば、事前に全ての全角大文字を半角小文字に変換する仕組みを持つ）ものでもかまわない。

【００４３】

また、上記の実施形態においては、特徴量として電子メイル本文に含まれる文字列のハッシュ値を用いたが、バイグラムや単語の出現頻度など、その他の特徴量を用いてもかまわない。

【００４４】

図１４は、本発明の他の実施形態の構成を示すブロック図であり、本発明をメイルサーバに組み込んだ構成例である。なお、図１４において図１と同一または同等物には同じ符号が付されており、図１と重複する説明は省略する。

【００４５】

図１４(a)のメイルサーバ群１が複数のメイルサーバ１ａ、１ｂ、１ｃなどから構成されているとすると、本実施形態は、同図(b)に示されているように、電子メイル収集手段５１、特徴量変換手段５２、マスメイル検出手段５３およびメイル処理手段５７を、各メイルサーバ１ａ、１ｂ、１ｃに組み込んだ点に特徴がある。

【００４６】

本実施形態では、マスメイルであるか否かの検出結果は、メイル処理手段５７へ送られる。メイル処理手段５７の行う処理は、マスメイル検出結果に基づいて、当該マスメイルの削除、メイル表題部へのマスメイルの表示などを行う。また、メイルサーバ運用者へのマスメイルの通知であってもかまわない。

【００４７】

以上の説明から明らかなように、本発明によれば、事前のルール作成や学習が不要である。また、単に電子メイルの特徴量を比較することにより類似メイルを検出し、該類似メイルが一定数に達するとマスメイルと判定するので、高速にマスメイル検出動作を行うことができる。

【図面の簡単な説明】

【００４８】

【図１】本発明を含むシステム構成を示すブロック図である。

【図２】電子メイル収集手段の処理手続きの一例を示すフローチャートである。

【図３】特徴量変換手段の処理の一例を示す説明図である。

【図４】特徴量変換手段の処理手続きの一例を示すフローチャートである。

【図５】マスメイル検出手段が利用するデータ構造の一例であるDirected Mapped Cacheを示す。

【図６】マスメイル検出手段の処理の一例を示すフローチャートである。

【図７】図６のステップＳ３０の一具体例を示すフローチャートである。

【図８】図６のステップＳ３５およびＳ３８の一具体例を示すフローチャートである。

【図９】順次抽出されるメイル１、２、３、４の特徴量の説明図である。

【図１０】メイル１に対するDirected Mapped Cacheのデータ例を示す説明図である。

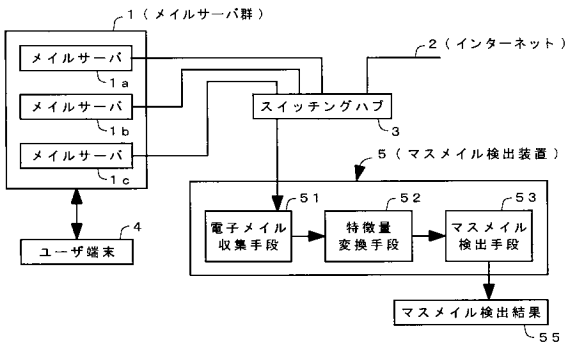
【図１１】メイル１、２に対するDirected Mapped Cacheのデータ例を示す説明図である。

【図１２】メイル１、２、３に対するDirected Mapped Cacheのデータ例を示す説明図である。

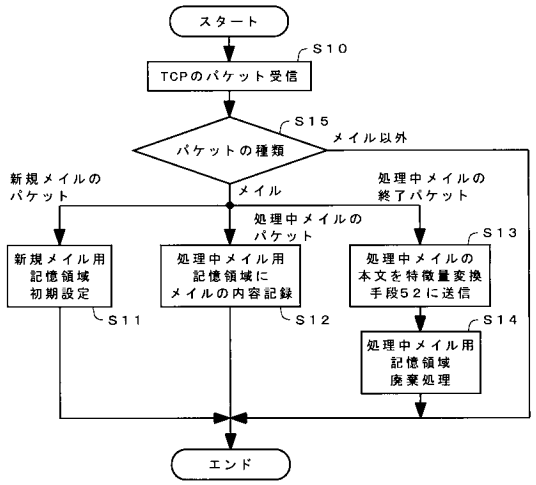
【図１３】メイル１、２、３、４に対するDirected Mapped Cacheのデータ例を示す説明図である。

【図１４】本発明の他の実施形態の要部を示すブロック図である。
【符号の説明】
【００４９】
１・・・メイルサーバ群、２・・・インターネット、３・・・スイッチングハブ、５・・・マスメイル検出装置、５１・・・電子メイル収集手段、５２・・・特徴量変換手段、５３・・・マスメイル検出手段、５５・・・マスメイル検出結果、３００・・・Directed Mapped Cache、３１０・・・特徴量データベース、３１１・・・ポインタ。

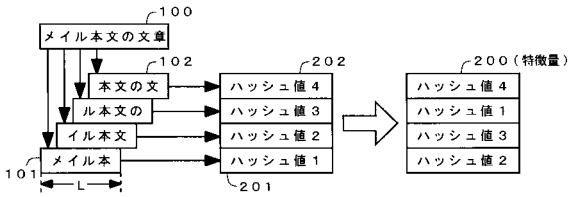
【図１】



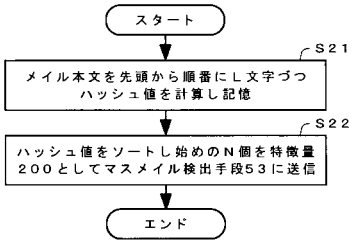
【図２】



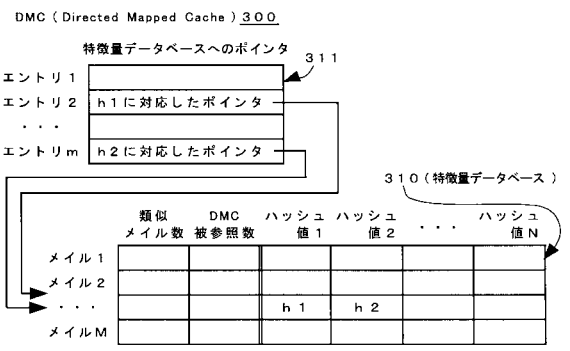
【図３】



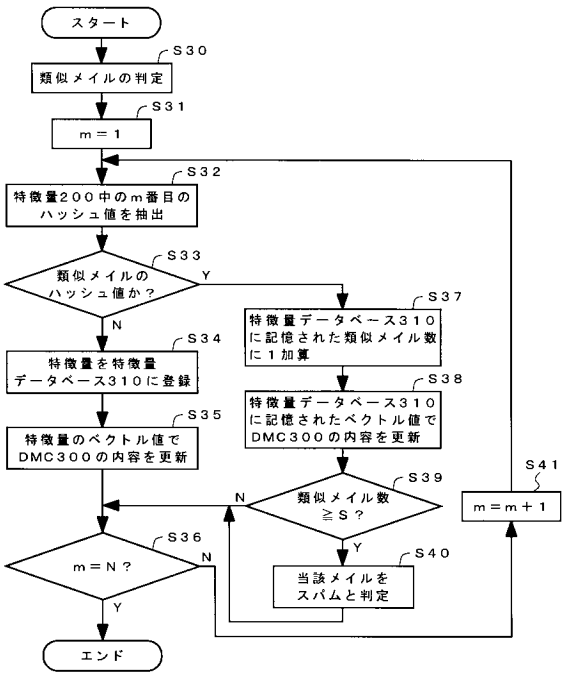
【図４】



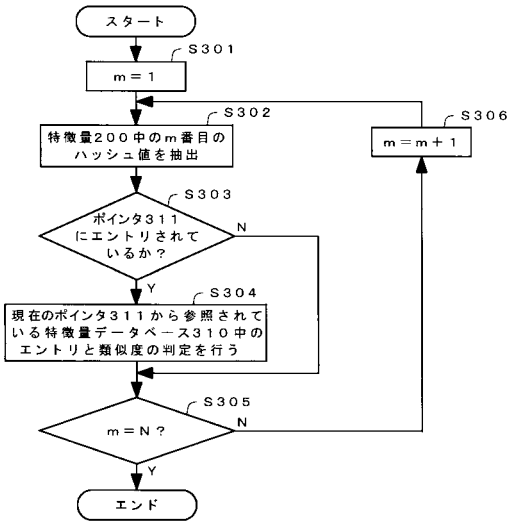
【図５】



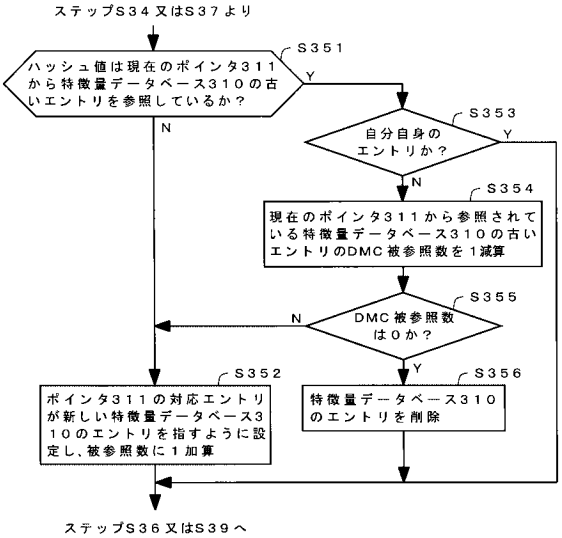
【図６】



【図７】



【図８】



【図９】

メイル１	メイル２	メイル３	メイル４
ｈ１	ｈ２	ｈ４	ｈ１
ｈ２	ｈ３	ｈ８	ｈ２
ｈ３	ｈ６	ｈ９	ｈ３
ｈ４	ｈ７	ｈ０	ｈ０

（類似は、3／4＝75%とする。よって、メイル４はメイル１と類似）

【図１０】

（ａ）

ｈ０	
ｈ１	メイル１
ｈ２	メイル１
ｈ３	メイル１
ｈ４	メイル１
ｈ５	
ｈ６	
ｈ７	
ｈ８	
ｈ９	

（ｂ）

	類似メイル数	DMC被参照数	ハッシュ値１	ハッシュ値２	ハッシュ値３	ハッシュ値４
メイル１	０	４	ｈ１	ｈ２	ｈ３	ｈ４

【図１１】

（ａ）

ｈ０	
ｈ１	メイル１
ｈ２	メイル２
ｈ３	メイル２
ｈ４	メイル１
ｈ５	
ｈ６	メイル２
ｈ７	メイル２
ｈ８	
ｈ９	

（ｂ）

	類似メイル数	DMC被参照数	ハッシュ値１	ハッシュ値２	ハッシュ値３	ハッシュ値４
メイル１	０	２	ｈ１	ｈ２	ｈ３	ｈ４
メイル２	０	４	ｈ２	ｈ３	ｈ６	ｈ７

【図１２】

（ａ）

ｈ０	メイル３
ｈ１	メイル１
ｈ２	メイル２
ｈ３	メイル２
ｈ４	メイル３
ｈ５	
ｈ６	メイル２
ｈ７	メイル２
ｈ８	メイル３
ｈ９	メイル３

（ｂ）

	類似メイル数	DMC被参照数	ハッシュ値１	ハッシュ値２	ハッシュ値３	ハッシュ値４
メイル１	０	１	ｈ１	ｈ２	ｈ３	ｈ４
メイル２	０	４	ｈ２	ｈ３	ｈ６	ｈ７
メイル３	０	４	ｈ４	ｈ８	ｈ９	ｈ０

【図１３】

（ａ）

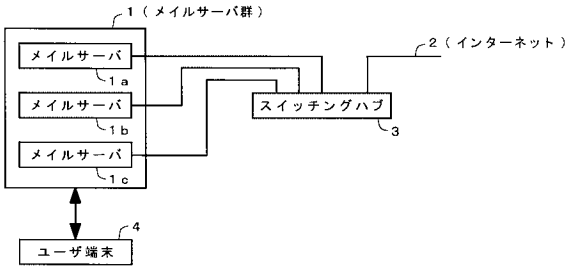
ｈ０	メイル３
ｈ１	メイル１
ｈ２	メイル１
ｈ３	メイル１
ｈ４	メイル１
ｈ５	
ｈ６	メイル２
ｈ７	メイル２
ｈ８	メイル３
ｈ９	メイル３

（ｂ）

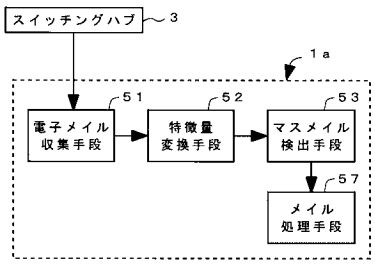
	類似メイル数	DMC被参照数	ハッシュ値１	ハッシュ値２	ハッシュ値３	ハッシュ値４
メイル１	１	４	ｈ１	ｈ２	ｈ３	ｈ４
メイル２	０	２	ｈ２	ｈ３	ｈ６	ｈ７
メイル３	０	３	ｈ４	ｈ８	ｈ９	ｈ０

【図１４】

（ａ）



（ｂ）



フロントページの続き

(72)発明者　藤川　裕充
埼玉県上福岡市大原二丁目１番１５号　株式会社　ＫＤＤＩ研究所内
(72)発明者　中島　昭浩
東京都新宿区西新宿二丁目３番２号　ＫＤＤＩ株式会社内
(72)発明者　本間　輝彰
東京都新宿区西新宿二丁目３番２号　ＫＤＤＩ株式会社内
(72)発明者　吉田　健一
埼玉県北本市高尾２－２３２

審査官　須藤　竜也

(56)参考文献　特表２００４－５００７６１（ＪＰ，Ａ）

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