(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4598774号
(P4598774)

(45)発行日　平成22年12月15日(2010.12.15)　(24)登録日　平成22年10月1日(2010.10.1)

(51)Int.Cl.　F I
G06F 13/00 (2006.01)　G06F 13/00 610Q

請求項の数 24 (全 21 頁)

(21)出願番号	特願2006-533134 (P2006-533134)
(86)(22)出願日	平成16年5月14日 (2004.5.14)
(65)公表番号	特表2007-503660 (P2007-503660A)
(43)公表日	平成19年2月22日 (2007.2.22)
(86)国際出願番号	PCT/US2004/015383
(87)国際公開番号	WO2004/105332
(87)国際公開日	平成16年12月2日 (2004.12.2)
審査請求日	平成19年5月2日 (2007.5.2)
(31)優先権主張番号	60/471,242
(32)優先日	平成15年5月15日 (2003.5.15)
(33)優先権主張国	米国(US)

(73)特許権者　501113353
シマンテック　コーポレイション
Ｓｙｍａｎｔｅｃ　Ｃｏｒｐｏｒａｔｉｏｎ
アメリカ合衆国　カリフォルニア州　９５０１４　クーパーティノ、スティーブンス　クリーク　ブルバード、２０３３０
２０３３０　Ｓｔｅｖｅｎｓ　Ｃｒｅｅｋ　Ｂｏｕｌｅｖａｒｄ，　Ｃｕｐｅｒｔｉｎｏ，　Ｃａｌｉｆｏｒｎｉａ　９５０１４，　ＵＳＡ
(74)代理人　100064621
弁理士　山川　政樹
(74)代理人　100098394
弁理士　山川　茂樹

最終頁に続く

(54)【発明の名称】類似性測度に基づいて電子メール・スパムをフィルタ処理するための方法および装置

(57)【特許請求の範囲】

【請求項１】

電子メール・メッセージを受信するための受信手段と、

前記電子メール・メッセージに、第１の文字セット内の文字の位置を示す文字参照が一つ以上含まれているか否かを検出する検出手段と、

前記検出した一つ以上の文字参照の内の少なくとも一つを前記第１の文字セット内の位置に対応する文字に変更することによって、前記電子メール・メッセージのコンテンツを修正する修正手段と、

前記電子メール・メッセージの修正されたコンテンツに基づいて、前記電子メール・メッセージを特徴付けるデータを生成するデータ生成手段と、

前記電子メール・メッセージを特徴付ける前記データを、複数のスパム・メッセージを特徴付けるデータのセットと比較する比較手段と、

前記電子メール・メッセージを特徴付ける前記データと、前記複数のスパム・メッセージを特徴付けるデータの前記セット内のいずれかのデータ項目の間の類似が、閾値を超えるかどうかを判定する判定手段と

を含む装置。

【請求項２】

前記一つ以上の文字参照は、「＆＃＜数値＞」の構文を持つＨＴＭＬ文字参照（ただし、「＜数値＞」は１０進文字番号または１６進文字番号）である請求項１に記載の装置。

【請求項３】

スパム・メッセージを受信するための受信手段と、
前記スパム・メッセージに、第１の文字セット内の文字の位置を示す文字参照が一つ以上含まれているか否かを検出するための検出手段と、
前記検出した一つ以上の文字参照の内の少なくとも一つを前記第１の文字セット内の位置に対応する文字に変更することによって、前記スパム・メッセージのコンテンツを修正する修正手段と、
前記スパム・メッセージの修正されたコンテンツに基づいて、前記スパム・メッセージを特徴付けるデータを生成する生成手段と、
前記スパム・メッセージに類似する着信メッセージを探し出すのに後に使用される、前記スパム・メッセージを特徴付ける前記データを、サーバに転送する転送手段と
を含む装置。
【請求項４】
前記第１の文字セットがＡＳＣＩＩ文字セットである請求項３に記載の装置。
【請求項５】
前記一つ以上の文字参照の内の少なくとも一つは、スパム・メッセージ内のＵＲＬ内にある請求項３に記載の装置。
【請求項６】
処理システム上で実行されると、
前記処理システムの受信手段が、電子メール・メッセージを受信すること、
前記電子メール・メッセージに、第１の文字セット内の文字の位置を示す文字参照が一つ以上含まれているか否かを、前記処理システムの検出手段が検出すること、
前記検出した一つ以上の文字参照の内の少なくとも一つを前記第１の文字セット内の位置に対応する文字に変更することによって、前記電子メール・メッセージのコンテンツを、前記処理システムの修正手段が修正すること、
前記電子メール・メッセージの修正されたコンテンツに基づいて、前記電子メール・メッセージを特徴付けるデータを、前記処理システムの生成手段が生成すること、
前記処理システムの比較手段が、前記電子メール・メッセージを特徴付ける前記データを、複数のスパム・メッセージを特徴付けるデータのセットと比較すること、
前記処理システムの判定手段が、前記電子メール・メッセージを特徴付ける前記データと、前記複数のスパム・メッセージを特徴付けるデータの前記セット内のいずれかのデータ項目の間の類似が、閾値を超えるかどうかを判定すること
を含む方法を実行するようにさせる実行可能命令を含むコンピュータ可読媒体。
【請求項７】
前記第１の文字セットがＡＳＣＩＩ文字セットである請求項６に記載のコンピュータ可読媒体。
【請求項８】
前記一つ以上の文字参照は数値文字参照または文字エンティティ参照を含む請求項６に記載のコンピュータ可読媒体。
【請求項９】
処理システム上で実行されると、
前記処理システムの受信手段が、スパム・メッセージを受信すること、
前記スパム・メッセージに、第１の文字セット内の文字の位置を示す文字参照が一つ以上含まれているか否かを、前記処理システムの検出手段が検出すること、
前記処理システムの修正手段が、前記検出した一つ以上の文字参照の内の少なくとも一つを前記第１の文字セット内の位置に対応する文字に変更することによって、前記スパム・メッセージのコンテンツを修正すること、
前記処理システムの生成手段が、前記スパム・メッセージの修正されたコンテンツに基づいて、前記スパム・メッセージを特徴付けるデータを生成すること、
前記処理システムの転送手段が、前記スパム・メッセージに類似する着信メッセージを探し出すのに後に使用される、前記スパム・メッセージを特徴付ける前記データを、サー

バに転送すること
を含む方法を実行するようにさせる実行可能命令を含むコンピュータ可読媒体。
【請求項１０】
前記第１の文字セットがＡＳＣＩＩ文字セットである請求項９に記載のコンピュータ可読媒体。
【請求項１１】
前記スパム・メッセージのコンテンツを修正することが、
前記スパム・メッセージ内の第１のグループの文字参照を前記第１の文字セット内の対応する文字に、前記修正手段が変更すること、
前記変更された第１のグループの文字参照が新たなグループの文字参照を形成するとの決定を、前記修正手段がなすこと、
前記決定に応答して、前記新たなグループの文字参照を前記第１の文字セット内の対応する文字に、前記修正手段が変更すること
を含む請求項９に記載のコンピュータ可読媒体。
【請求項１２】
電子メール・メッセージの中で、スパム・フィルタ処理を回避するように前記電子メール・メッセージに追加された雑音を示すデータを検出し、前記電子メール・メッセージのコンテンツを変更して、前記雑音を低減させるメッセージ・クリーニング手段であって、
前記電子メール・メッセージ内の一つ以上の文字参照を検出するための文字参照検出手段と、
前記一つ以上の文字参照の少なくとも一つを、第１の文字セット内の対応する文字に変更する文字参照変更手段と
を有したメッセージ・クリーニング手段と、
前記電子メール・メッセージの前記変更されたコンテンツをスパム・メッセージのコンテンツと比較する類似アイデンティファイア手段と
を含むシステム。
【請求項１３】
前記一つ以上の文字参照は数値文字参照を含む請求項１２に記載のシステム。
【請求項１４】
前記メッセージ・クリーニング手段は、例外として適格でないフォーマット・データを前記電子メール・メッセージから抜き取ることによって、前記電子メール・メッセージの前記コンテンツを変更することを含む請求項１２に記載のシステム。
【請求項１５】
前記一つ以上の文字参照は文字エンティティ参照を含む請求項１２に記載のシステム。
【請求項１６】
前記文字参照検出手段が、変更された第１のグループの文字参照が新たなグループの文字参照を形成するとの決定をなす手段をさらに含み、
前記文字参照変更手段が、前記決定に応答して、前記新たなグループの文字参照を前記第１の文字セット内の対応する文字に変更する手段をさらに含む請求項１２に記載のシステム。
【請求項１７】
前記第１の文字セットがＡＳＣＩＩ文字セットである請求項１２に記載のシステム。
【請求項１８】
前記メッセージ・クリーニング手段は、ＵＲＬから、事前定義されたカテゴリの一意識別子データを除去することによって、前記電子メール・メッセージの前記コンテンツを変更する手段を含む請求項１２に記載のシステム。
【請求項１９】
前記メッセージ・クリーニング手段は、ＵＲＬから、事前定義されたカテゴリのクエリ・データを除去することによって、前記電子メール・メッセージの前記コンテンツを変更する手段を含む請求項１２に記載のシステム。

【請求項２０】
前記類似アイデンティファイア手段は、前記電子メール・メッセージの前記変更されたコンテンツが、前記スパム・メッセージの前記コンテンツに類似しているかどうかを判定することによって、前記電子メール・メッセージの前記変更されたコンテンツを、前記スパム・メッセージの前記コンテンツと比較する手段を含む請求項１２に記載のシステム。
【請求項２１】
電子メール・メッセージの中で、スパム・フィルタ処理を回避するために前記電子メール・メッセージに追加された一つ以上の文字参照であって数値文字参照または文字エンティティ参照を含む文字参照を検出する検出手段と、
前記一つ以上の文字参照の少なくとも一つを、第１の文字セット内の対応する文字に変更することによって、前記電子メール・メッセージのコンテンツを変更する変更手段と、
前記電子メール・メッセージの前記変更されたコンテンツを、スパム・メッセージのコンテンツと比較する比較手段と
を含む装置。
【請求項２２】
前記第１の文字セットがＡＳＣＩＩ文字セットである請求項２１に記載の装置。
【請求項２３】
処理システム上で実行されると、
前記処理システムの検出手段が、電子メール・メッセージの中で、スパム・フィルタ処理を回避するために前記電子メール・メッセージに追加された一つ以上の文字参照であって数値文字参照または文字エンティティ参照を含む文字参照を検出すること、
前記処理システムの変更手段が、前記一つ以上の文字参照の少なくとも一つを、第１の文字セット内の対応する文字に変更することによって、前記電子メール・メッセージのコンテンツを変更すること、
前記処理システムの比較手段が、前記電子メール・メッセージの前記変更されたコンテンツを、スパム・メッセージのコンテンツと比較すること
を含む方法を実行するようにさせる実行可能命令を含むコンピュータ可読媒体。
【請求項２４】
前記一つ以上の文字参照の内の少なくとも一つは、前記電子メール・メッセージ内のＵＲＬ内にある請求項２３に記載のコンピュータ可読媒体。
【発明の詳細な説明】
【関連出願】
【０００１】
本出願は、参照により全体が本明細書に組み込まれている、２００３年５月１５日に出願した米国仮出願第６０／４７１２４２号の優先権を主張する。
【技術分野】
【０００２】
本発明は、電子メール（ｅｍａｉｌ）をフィルタ処理することに関し、より詳細には、本発明は、類似性測度に基づいて電子メール・スパムをフィルタ処理することに関する。
【背景技術】
【０００３】
インターネットの普及が進んでおり、ますます多くの人々が、電子的大量メールを生成し、送信することによって製品やサービスを宣伝して、インターネットを介してビジネスを行っている。それらの電子メッセージ（電子メール）は、普通、求められておらず、受信者によって迷惑と見られている。というのは、それらのメッセージが、必要で重要なデータ処理のために必要とされる記憶スペースの多くを占有するからである。例えば、メール・サーバは、そのサーバの記憶容量が、宣伝を含む不要な電子メールで最大限度までいっぱいになった場合、重要で、かつ／または所望される電子メールを受け入れることを拒否しなければならなくなる。さらに、セットトップ・ボックス、ＰＤＡ、ネットワーク・コンピュータ、ポケットベルなどのシン・クライアント・システムはすべて、限られた記

憶容量を有する。そのようなシステムのいずれにおける不要な電子メールも、ユーザのための有限なリソースを縛りつけることになる。さらに、通常のユーザは、大量であるが、無用な宣伝情報をダウンロードすることによって、時間を浪費する。それらの不要な電子メールは、一般にスパムと呼ばれる。

【０００４】

現在、不要なメッセージをフィルタにかけて除くことができる製品が存在する。例えば、すべてのスパム・エージェント（すなわち、大量の求められていない電子メールを生成する企業）の索引リストを保持し、そのリスト上の企業から送信された電子メールをブロックする手段を提供するスパム・ブロック方法が存在する。

【０００５】

現在、入手可能な別の「ジャンク・メール」フィルタは、事前定義された語や、前述したパターンに基づくフィルタ群を使用する。着信メールは、件名が既知のスパム・パターンを含む場合、不要なメールと指定される。

【発明の開示】

【発明が解決しようとする課題】

【０００６】

しかし、スパム・フィルタ処理が高度になるにつれ、フィルタを回避するスパム発信者の技術も高度になる。最近の世代のスパム発信者によって採り入れられている戦術の例には、ランダム化、発信元隠蔽、ＨＴＭＬを使用したフィルタ回避が含まれる。

【課題を解決するための手段】

【０００７】

類似性測度に基づいて電子メール・スパムをフィルタ処理するための方法とシステムを説明する。一態様によれば、方法は、着信電子メール・メッセージを受信し、着信電子メール・メッセージのコンテンツに基づいて、着信電子メール・メッセージを特徴付けるデータを生成し、その生成されたデータを、スパム・メッセージを特徴付けるデータのセットと比較することを含む。方法は、着信電子メール・メッセージを特徴付けるデータと、スパム・メッセージを特徴付けるデータのセットからのいずれかのデータ項目の間の類似が、閾値を超えるかどうかを判定することをさらに含む。

【発明を実施するための最良の形態】

【０００８】

本発明のその他の特徴は、添付の図面、および以下の詳細な説明から明白となろう。

【０００９】

本発明は、以下に提供する詳細な説明や、本発明の様々な実施形態の添付の図面からより完全に理解されるが、説明と図面は、本発明を特定の実施形態に限定するものと解釈されるべきではなく、単に説明および理解を目的とする。

【００１０】

類似性測度に基づいて電子メール・スパムをフィルタ処理するための方法とシステムを説明する。以下の説明では、多数の詳細を提示する。しかし、本発明は、それらの特定の詳細なしに実施することもできることが、当業者には理解されよう。その他、周知の構造やデバイスは、本発明を不明瞭にするのを避けるため、詳細にではなく、ブロック図の形態で示す。

【００１１】

以下の詳細な説明のいくつかの部分は、コンピュータ・メモリ内のデータ・ビットに対するオペレーションのアルゴリズムと記号表現として提示する。それらのアルゴリズム記述やアルゴリズム表現は、データ処理分野の業者が、自らの作業の内容を他の同業者に最も効果的に伝えるのに使用する手段である。アルゴリズムは、本明細書では、また、一般に、所望の結果をもたらす、自己矛盾のない一連のステップであると考えられる。ステップは、物理的量の物理的な操作を要するステップである。通常、ただし必然的にではなく、それらの量は、格納されること、転送されること、結合されること、比較されること、それ以外の形で操作されることが可能な、電気信号または磁気信号の形態をとる。ときと

して、主に、一般的な用法の理由で、それらの信号をビット、値、要素、シンボル、文字、項目、数などと呼ぶのが好都合であることが分かっている。

【００１２】

しかし、それらの用語や類似する用語のすべては、適切な物理的量に関連付けられるべきであり、それらの量に適用された便利なラベルに過ぎないことに留意されたい。特に明記しない限り、以下の説明から明白なとおり、説明全体で、「処理する」または「演算する」または「計算する」または「算出する」または「表示する」などの用語を利用する説明は、コンピュータ・システムのレジスタ内やメモリ内の物理的（電子的）量として表されるデータを操作し、変換して、コンピュータ・システムのメモリまたはレジスタ、あるいは他のそのような情報記憶デバイス、情報伝送デバイス、または情報表示デバイスの内部の物理的量として同様に表される他のデータにする、コンピュータ・システム、または類似の電子コンピューティング・デバイスのアクションやプロセスを指すものと理解されるべきである。

【００１３】

また、本発明は、本明細書のオペレーションを実行するための装置にも関する。この装置は、要求される目的のために特別に構築されることも、コンピュータの中に格納されたコンピュータ・プログラムによって選択的に起動される、または再構成される汎用コンピュータを含むことも可能である。そのようなコンピュータ・プログラムは、フロッピー・ディスク、光ディスク、ＣＤ－ＲＯＭ、光磁気ディスクを含む任意のタイプのディスク、読み取り専用メモリ（ＲＯＭ）、ランダム・アクセス・メモリ（ＲＡＭ）、ＥＰＲＯＭ、ＥＥＰＲＯＭ、磁気カードまたは光カード、あるいは電子命令を格納するのに適した任意のタイプの媒体などの、ただし、それらには限定されない、コンピュータ可読記憶媒体の中に格納されることが可能であり、各媒体は、コンピュータ・システム・バスに結合される。

【００１４】

本明細書で提示するアルゴリズムと表示は、いずれの特定のコンピュータ、またはその他の装置にも本質的に関連していない。様々な汎用システムを、本明細書の教示によるプログラムとともに使用することができ、あるいは、要求される方法ステップを実行する、より特殊化された装置を構築することが好都合であると判明する。様々なそれらのシステムの要求される構造は、以下の説明から明らかとなる。さらに、本発明は、いずれの特定のプログラミング言語に関連しても説明しない。様々なプログラミング言語を使用して、本明細書で説明する本発明の教示を実施することができることが理解されよう。

【００１５】

マシン可読媒体には、マシン（例えば、コンピュータ）による読み取りが可能な形態で情報を格納する、または伝送するための任意の機構が含まれる。例えば、マシン可読媒体には、読み取り専用メモリ（「ＲＯＭ」）、ランダム・アクセス・メモリ（「ＲＡＭ」）、磁気ディスク記憶媒体、光記憶媒体、フラッシュメモリ・デバイス、電気、光、音響、またはその他の形態の伝搬される信号（例えば、搬送波、赤外線信号、デジタル信号など）、その他が含まれる。

【００１６】

類似性測度に基づいて電子メール・スパムをフィルタ処理すること

図１は、スパム電子メール（ｅｍａｉｌ）の配信を制御するためのシステムの一実施形態のブロック図である。システムは、公共ネットワーク（例えば、インターネット、無線ネットワークなど）、または私設ネットワーク（例えば、ＬＡＮ、イントラネットなど）などの通信ネットワーク１００に結合された制御センタ１０２を含む。制御センタ１０２は、ネットワーク１００を介して複数のネットワーク・サーバ１０４と通信する。各サーバ１０４は、私設ネットワークまたは公共ネットワークを使用して、ユーザ端末装置１０６と通信する。

【００１７】

制御センタ１０２は、スパムと識別されたメッセージを分析し、スパムを検出するため

のフィルタ処理規則を開発し、フィルタ処理規則をサーバ群１０４に配信することを担うスパム対策設備である。メッセージは、既知のスパム源（例えば、「スパム・プローブ」、すなわち、可能な限り多くのスパム発信者メーリング・リストに入り込むように特別に選択された電子メール・アドレスを使用して特定された）によって送信されれば、スパムとして識別される。

【００１８】

サーバ１０４は、送信された対応するユーザ端末装置のユーザにアドレス指定されたメッセージを受信し、格納するメール・サーバである。代替として、サーバ１０４は、メール・サーバ１０４に結合された、異なるサーバであってもよい。サーバ群１０４は、制御センタ１０２から受け取られたフィルタ処理規則に基づいて、着信メッセージをフィルタ処理することを担う。

【００１９】

一実施形態では、制御センタ１０２は、スパム攻撃に関連するコンテンツを特徴付けるデータを生成し、そのデータをサーバ群１０４に送信することを担うスパム・コンテンツ準備モジュール１０８を含む。各サーバ１０４は、制御センタ１０２から受信されたスパム・データを格納し、格納されたデータを使用して、スパム・コンテンツに類似する着信電子メール・メッセージを識別することを担う、類似度算出モジュール１１０を含む。

【００２０】

代替の実施形態では、各サーバ１０４は、スパム攻撃に関連するコンテンツを特徴付けるデータを生成するスパム・コンテンツ準備モジュール１０８と、生成されたデータを使用して、スパム・コンテンツに類似する電子メール・メッセージを識別する類似度算出モジュール１１０をともにホストする。

【００２１】

図２は、スパム・コンテンツ準備モジュール２００の一実施形態のブロック図である。スパム・コンテンツ準備モジュール２００は、スパム・コンテンツ・パーサ２０２、スパム・データ・ジェネレータ２０６、スパム・データ・トランスミッタ２０８を含む。

【００２２】

スパム・コンテンツ・パーサ２０２は、スパム攻撃による電子メール・メッセージの本文（スパム・メッセージと呼ぶ）を解析することを担う。

【００２３】

スパム・データ・ジェネレータ２０６は、スパム・メッセージを特徴付けるデータを生成することを担う。一実施形態では、スパム・メッセージを特徴付けるデータには、スパム・メッセージを構成するトークン（例えば、文字、語、行など）のセットに関して計算されたハッシュ値のリストが含まれる。スパム・メッセージ、または他の任意の電子メール・メッセージを特徴付けるデータを本明細書では、メッセージ・シグネチャと呼ぶ。スパム・メッセージ、または他の任意の電子メール・メッセージのシグネチャは、メッセージ・コンテンツを識別する様々なデータを含むことが可能であり、異なる電子メール・メッセージのシグネチャを比較する際に類似性測度の使用を可能にする、様々なアルゴリズムを使用して作成される。

【００２４】

一実施形態では、スパム・コンテンツ準備モジュール２００は、雑音を示すデータを検出すること、さらに、スパム・メッセージのシグネチャを生成するのに先立って、スパム・メッセージからその雑音を除去することを担う雑音低減アルゴリズム２０４も含む。雑音は、スパムの性質を隠すようにスパム・メッセージに追加されている、受信者には見えないデータを表す。

【００２５】

一実施形態では、スパム・コンテンツ準備モジュール２００は、単一のスパム攻撃を元とするメッセージをグループ化することを担うメッセージ・グループ化アルゴリズム（図示せず）も含む。グループ化は、スパム・メッセージの指定された特性（例えば、含まれるＵＲＬ、メッセージ部分など）に基づいて実行される。グループ化が使用される場合、

スパム・データ・ジェネレータ２０６は、それぞれの個別メッセージに関してではなく、スパム・メッセージのグループに関するシグネチャを生成することができる。

【００２６】

スパム・データ・トランスミッタ２０８は、スパム・メッセージのシグネチャを、図１のサーバ群１０４のような、参加するサーバ群に配信することを担う。一実施形態では、各サーバ１０４は、コール・センタ１０２に対する接続（例えば、セキュリティで保護されたＨＴＴＰＳ接続）を定期的に（例えば、５分毎に）開始する。このプル・ベースの接続を使用して、シグネチャは、コール・センタ１０２から妥当なサーバ１０６に伝送される。

【００２７】

図３は、類似度算出モジュール３００の一実施形態のブロック図である。類似度算出モジュール３００は、着信メッセージ・パーサ３０２と、スパム・データ・レシーバ３０６と、メッセージ・データ・ジェネレータ３１０と、類似アイデンティファイア３１２と、スパム・データベース３０４とを含む。

【００２８】

着信メッセージ・パーサ３０２は、着信電子メール・メッセージの本文を解析することを担う。

【００２９】

スパム・データ・レシーバ３０６は、スパム・メッセージのシグネチャを受信して、そのシグネチャをスパム・データベース３０４の中に格納することを担う。

【００３０】

メッセージ・データ・ジェネレータ３１０は、着信電子メール・メッセージのシグネチャを生成することを担う。一部の実施形態では、着信電子メール・メッセージのシグネチャは、その着信電子メール・メッセージを構成するトークン（例えば、文字、語、行など）のセットに関して計算されたハッシュ値のリストを含む。他の諸実施形態では、着信電子メール・メッセージのシグネチャは、電子メール・メッセージのコンテンツ（例えば、着信電子メール・メッセージを構成するトークン・セットのサブセット）を特徴付ける、他の様々なデータを含む。前述したとおり、電子メール・メッセージのシグネチャは、異なる電子メール・メッセージのシグネチャを比較する際に類似性測度の使用を可能にする、様々なアルゴリズムを使用して作成される。

【００３１】

一実施形態では、類似度算出モジュール３００は、以下により詳細に説明するとおり、雑音を示すデータを検出することと、着信電子メール・メッセージから、そのメッセージのシグネチャを生成するのに先立って、その雑音を除去することを担う、着信メッセージ・クリーニング・アルゴリズム３０８も含む。

【００３２】

類似アイデンティファイア３１２は、各着信電子メール・メッセージのシグネチャを、スパム・データベース３０４の中に格納されたスパム・メッセージのシグネチャと比較することと、その比較に基づき、着信電子メール・メッセージが、何らかのスパム・メッセージに類似しているかどうかを判定することを担う。

【００３３】

一実施形態では、スパム・データベース３０４は、スパム・メッセージが雑音低減プロセスを受ける前のスパム・メッセージ（すなわち、雑音のあるスパム・メッセージ）に対して生成されたシグネチャと、スパム・メッセージが雑音低減プロセスを受けた後のスパム・メッセージ（すなわち、低減された雑音を有するスパム・メッセージ）に対して生成されたシグネチャを格納する。この実施形態では、メッセージ・データ・ジェネレータ３１０がまず、雑音低減に先立って、着信電子メール・メッセージのシグネチャを生成し、類似アイデンティファイア３１２が、そのシグネチャを、雑音のあるスパム・メッセージのシグネチャと比較する。この比較により、着信電子メール・メッセージが、それらのスパム・メッセージの１つに類似していることが示された場合、類似アイデンティファイア

３１２が、その着信電子メール・メッセージにスパムとしてマークを付ける。代替として、類似アイデンティファイア３１２は、着信メッセージ・クリーニング・アルゴリズム３０８を呼び出して、着信電子メール・メッセージから雑音を除去する。次に、メッセージ・データ・ジェネレータ３１０が、その変更された着信メッセージに関するシグネチャを生成し、このシグネチャが、次に、類似アイデンティファイア３１２によって、低減された雑音を有するスパム・メッセージのシグネチャと比較される。

【００３４】

図４は、スパム・メッセージを扱うためのプロセス４００の一実施形態を示す流れ図である。プロセスは、ハードウェア（例えば、専用論理、プログラマブル論理、マイクロコードなど）、ソフトウェア（汎用コンピュータ・システム上、または専用マシン上で実行されるような）、またはそれらの組み合わせを含む処理論理によって実行される。一実施形態では、処理論理は、図１の制御センタ１０２に存在する。

【００３５】

図４を参照すると、プロセス４００は、処理論理が、スパム・メッセージを受け取ることから始まる（処理ブロック４０２）。

【００３６】

処理ブロック４０４で、処理論理は、スパム・メッセージを変更して、雑音を低減させる。雑音低減アルゴリズムの一実施形態は、図９と図１０に関連して、以下により詳細に説明する。

【００３７】

処理ブロック４０６で、処理論理はスパム・メッセージのシグネチャを生成する。一実施形態では、スパム・メッセージのシグネチャは、図６Ａに関連して以下により詳細に説明するとおり、着信電子メール・メッセージを構成するトークン（例えば、文字、語、行など）のセットに関して計算されたハッシュ値のリストを含む。他の諸実施形態では、着信電子メール・メッセージのシグネチャは、電子メール・メッセージのコンテンツを特徴付ける、他の様々なデータを含む。

【００３８】

処理ブロック４０８で、処理論理は、スパム・メッセージのシグネチャをサーバ（例えば、図１のサーバ１０４）に転送し、サーバは、スパム・メッセージのそのシグネチャを使用して、そのスパム・メッセージに類似する着信電子メール・メッセージを探し出す（ブロック４１０）。

【００３９】

図５は、類似性測度に基づいて電子メール・スパムをフィルタ処理するためのプロセス５００の一実施形態の流れ図である。プロセスは、ハードウェア（例えば、専用論理、プログラマブル論理、マイクロコードなど）、ソフトウェア（汎用コンピュータ・システム上、または専用マシン上で実行されるような）、またはそれらの組み合わせを含む処理論理によって実行される。一実施形態では、処理論理は、図１のサーバ１０４に存在する。

【００４０】

図５を参照すると、プロセス５００は、処理論理が着信電子メール・メッセージを受け取ることから始まる（処理ブロック５０２）。

【００４１】

処理ブロック５０４で、処理論理は、着信メッセージを変更して、雑音を低減させる。雑音低減アルゴリズムの一実施形態は、図９と図１０に関連して、以下により詳細に説明する。

【００４２】

処理ブロック５０６で、処理論理は、着信メッセージのコンテンツに基づき、着信メッセージのシグネチャを生成する。一実施形態では、着信電子メール・メッセージのシグネチャは、図６Ａに関連して以下により詳細に説明するとおり、着信電子メール・メッセージを構成するトークン（例えば、文字、語、行など）のセットに関して計算されたハッシュ値のリストを含む。他の諸実施形態では、着信電子メール・メッセージのシグネチャは

、電子メール・メッセージのコンテンツを特徴付ける、他の様々なデータを含む。

【００４３】

処理ブロック５０８で、処理は、着信メッセージのシグネチャをスパム・メッセージのシグネチャと比較する。

【００４４】

処理ブロック５１０で、着信メッセージのシグネチャと何らかのスパム・メッセージのシグネチャの間の類似が、閾値類似性測度を超えている、と処理論理が判定する。２つのメッセージ間の類似を判定するためのプロセスの一実施形態は、図６Ｂに関連して以下により詳細に説明する。

【００４５】

処理ブロック５１２で、処理論理は、着信電子メール・メッセージにスパムとしてマークを付ける。

【００４６】

図６Ａは、電子メール・メッセージのシグネチャを作成するためのプロセス６００の一実施形態の流れ図である。プロセスは、ハードウェア（例えば、専用論理、プログラマブル論理、マイクロコードなど）、ソフトウェア（汎用コンピュータ・システム上、または専用マシン上で実行されるような）、またはそれらの組み合わせを含む処理論理によって実行される。一実施形態では、処理論理は、図１のサーバ１０４に存在する。

【００４７】

図６Ａを参照すると、プロセス６００は、処理論理が、電子メール・メッセージをトークン・セットに分割することから始まる（処理ブロック６０２）。各トークン・セットは、電子メール・メッセージからの事前定義された数の順次ユニットを含む。事前定義された数は、１以上である。ユニットは、電子メール・メッセージの中の文字、語、または行を表わす。一実施形態では、各トークン・セットは、電子メール・メッセージの中のそのトークン・セットの出現の回数と組み合わせられる。

【００４８】

処理ブロック６０４で、処理論理は、それらのトークン・セットに関するハッシュ値を計算する。一実施形態では、ハッシュ値は、トークン・セットと対応するトークン出現回数との各組み合わせにハッシュ関数を適用することによって計算される。

【００４９】

処理ブロック６０６で、処理論理は、計算されたハッシュ値を使用して、電子メール・メッセージに関するシグネチャを作成する。一実施形態では、シグネチャは、計算されたハッシュ値のサブセットを選択し、電子メール・メッセージを特徴付けるパラメータを、計算されたハッシュ値の選択されたサブセットに加えることによって作成される。パラメータは、例えば、電子メール・メッセージのサイズ、計算されるハッシュ値の数、電子メール・メッセージに関連するキーワード、添付ファイルの名前などを指定することが可能である。

【００５０】

一実施形態では、電子メール・メッセージに関するシグネチャは、図７と図８に関連して以下により詳細に説明する、文字ベースのドキュメント比較機構を使用して作成される。

【００５１】

図６Ｂは、電子メール・メッセージのシグネチャを使用してスパムを検出するためのプロセス６５０の一実施形態の流れ図である。プロセスは、ハードウェア（例えば、専用論理、プログラマブル論理、マイクロコードなど）、ソフトウェア（汎用コンピュータ・システム上、または専用マシン上で実行されるような）、またはそれらの組み合わせを含む処理論理によって実行される。一実施形態では、処理論理は、図１のサーバ１０４に存在する。

【００５２】

図６Ｂを参照すると、プロセス６５０は、着信電子メール・メッセージのシグネチャの

中のデータを、各スパム・メッセージのシグネチャの中のデータと比較する。シグネチャ・データは、電子メール・メッセージのコンテンツを特徴付けるパラメータと、電子メール・メッセージの中に含まれるトークンに関して生成されたハッシュ値のサブセットとを含む。パラメータは、例えば、電子メール・メッセージのサイズ、電子メール・メッセージの中のトークンの数、電子メール・メッセージに関連するキーワード、添付ファイルの名前などを指定する。

【００５３】

処理論理は、着信電子メール・メッセージのシグネチャの中のパラメータを、各スパム・メッセージのシグネチャの中の対応するパラメータと比較することから始まる（処理ブロック６５２）。

【００５４】

判定ボックス６５４で、処理論理は、いずれかのスパム・メッセージ・シグネチャが、着信メッセージ・シグネチャのパラメータに類似するパラメータを含むかどうかを判定する。類似性は、例えば、２つのパラメータ間の許容される差、または２つのパラメータの許容される比に基づいて判定される。

【００５５】

スパム・メッセージ・シグネチャのいずれも、着信メッセージ・シグネチャのパラメータに類似するパラメータを含まない場合、処理論理は、その着信電子メール・メッセージが正当である（すなわち、スパムではない）と判定する（処理ブロック６６２）。

【００５６】

代替として、１つまたは複数のスパム・メッセージ・シグネチャが、類似するパラメータを有する場合、処理論理は、最初のスパム・メッセージのシグネチャが、着信電子メールのシグネチャの中のハッシュ値に類似するハッシュ値を有するかどうかを判定する（判定ボックス６５６）。類似性閾値に基づき、ハッシュ値は、例えば、ある数のハッシュ値が一致する場合、または一致するハッシュ値と一致しないハッシュ値の比が、指定された閾値を超えた場合、類似していると見る。

【００５７】

最初のスパム・メッセージ・シグネチャが、着信電子メール・シグネチャのハッシュ値に類似するハッシュ値を有する場合、処理論理は、着信電子メールはスパムであると判定する（処理ブロック６７０）。それ以外の場合、処理論理は、類似するパラメータを有する、さらなるスパム・メッセージ・シグネチャが存在するかどうかをさらに判定する（判定ブロック６５８）。存在する場合、処理論理は、次のスパム・メッセージ・シグネチャが、着信電子メール・シグネチャのハッシュ値に類似するハッシュ値を有するかどうかを判定する（判定ボックス６５６）。類似するハッシュ値を有する場合、処理論理は、その着信電子メール・メッセージはスパムであると判定する（処理ブロック６７０）。類似するハッシュ値を有さない場合、処理論理は、処理ブロック６５８に戻る。

【００５８】

処理論理は、類似するパラメータを有する他のスパム・メッセージ・シグネチャが全く存在しないと判定した場合、その着信メール・メッセージはスパムではないと判定する（処理ブロック６６２）。

【００５９】

文字ベースのドキュメント比較機構

図７は、ドキュメントの文字ベースの比較のためのプロセス７００の一実施形態の流れ図である。プロセスは、ハードウェア（例えば、専用論理、プログラマブル論理、マイクロコードなど）、ソフトウェア（汎用コンピュータ・システム上、または専用マシン上で実行されるような）、またはそれらの組み合わせを含む処理論理によって実行される。

【００６０】

図７を参照すると、プロセス７００は、処理論理がドキュメントを前処理することから始まる（処理ブロック７０２）。一実施形態では、ドキュメントは、ドキュメント内のそれぞれの大文字の英字を小文字の英字に変更することによって前処理される。例えば、「

Ｉ　ａｍ　Ｓａｍ，Ｓａｍ　Ｉ　ａｍ」というメッセージが前処理されて、「ｉ．ａｍ．ｓａｍ．ｓａｍ．ｉ．ａｍ」という表現になる。

【００６１】

処理ブロック７０４で、処理論理は、ドキュメントをトークンに分割し、各トークンが、ドキュメントからの所定の数の順次の文字を含む。一実施形態では、各トークンは、そのトークンの出現回数と組み合わせられる。この組み合わせは、ラベル付きシングル（ｓｈｉｎｇｌｅ）と呼ばれる。例えば、トークン内の順次の文字の所定の数が３に等しい場合、上記に規定した表現は、ラベル付きシングルの以下のセットを含む。

【００６２】

【表１】

i.al
.aml
am.l
m.sl
.sal
saml
sm.2
m.sl
.sm2
sam2
am.3
m.il
.i.l
i.a2
.am4

【００６３】

一実施形態では、シングルは、ヒストグラムとして表される。

【００６４】

処理ブロック７０６で、処理論理は、トークンに関するハッシュ値を計算する。一実施形態では、ハッシュ値は、ラベル付きシングルに関して計算される。例えば、上記の各ラベル付きシングルにハッシュ関数Ｈ（ｘ）が適用された場合、以下の結果がもたらされる。すなわち、

【００６５】

【表２】

H(i.a1) -> 458348732
H(.am1) -> 200404023
H(am.1) -> 692939349
H(m.s1) -> 220443033
H(.sa1) -> 554034022
H(8am1) -> 542929292
H(am.2) -> 629292229
H(m.s1) -> 702202232
H(.sa2) -> 322243349
H(8am2) -> 993923828
H(am.3) -> 163393269
H(m.i1) -> 595437753
H(.i.1) -> 843438583
H(i.a2) -> 244485639
H(.am4) -> 493869359

【００６６】
一実施形態では、処理論理は、ハッシュ値を以下のとおり並べ替える。すなわち、
【００６７】
【表３】

163393269
200604023
220643033
246685639
322263369
458368732
493869359
542929292
554034022
595637753
629292229
692939349
702202232
843438583
993923828

【００６８】
処理ブロック７０８で、処理論理は、計算されたハッシュ値からハッシュ値のサブセッ

トを選択する。一実施形態では、処理論理は、並べ替えられたハッシュ値から、小さい方からＸ個の値を選択し、それらの値から、ドキュメントの「スケッチ」を作成する。例えば、Ｘ＝４の場合、スケッチは、以下のとおり表現される。すなわち、

[163393269 200404023 220443033 244485639]

【００６９】

処理ブロック７１０で、処理論理は、スケッチに、ドキュメントのトークンに関するパラメータを加えることによって、ドキュメントのシグネチャを作成する。一実施形態では、パラメータは、ドキュメントの中における最初のトークンの数を指定する。前述の例では、最初のトークンの数は、１５である。したがって、ドキュメントのシグネチャは、以下のとおり表現される。すなわち、

[15 163393269 200404023 220443033 244485639]

代替として、パラメータは、ドキュメントのコンテンツの他の任意の特性（例えば、ドキュメントのサイズ、ドキュメントに関連するキーワードなど）を指定してもよい。

【００７０】

図８は、２つのドキュメントが類似しているかどうかを判定するためのプロセス８００の一実施形態の流れ図である。プロセスは、ハードウェア（例えば、専用論理、プログラマブル論理、マイクロコードなど）、ソフトウェア（汎用コンピュータ・システム上、または専用マシン上で実行されるような）、またはそれらの組み合わせを含む処理論理によって実行される。

【００７１】

図８を参照すると、プロセス８００は、処理論理が、ドキュメント１のシグネチャの中で指定されたトークン数と、ドキュメント２のシグネチャの中で指定されたトークン数を比較して、第１のシグネチャの中のトークン数が、第２のシグネチャからのトークン数に対して、許容される範囲内にあるかどうかを判定することから始まる（判定ブロック８０２）。例えば、許容される範囲は、１つ以内の違い、または９０パーセント以上の比である。

【００７２】

第１のシグネチャの中のトークン数が、第２のシグネチャからのトークン数に対して、許容される範囲外である場合、処理論理は、ドキュメント１とドキュメント２は異なると判定する（処理ブロック８０８）。そうではなく、第１のシグネチャの中のトークン数が、第２のシグネチャからのトークン数に対して、許容される範囲内にある場合、処理論理は、シグネチャ１の中のハッシュ値と、シグネチャ２の中のハッシュ値の間の類似が閾値を超えている（例えば、９５パーセントを超えるハッシュ値が同一である）かどうかを判定する（判定ボックス８０４）。超えている場合、処理論理は、２つのドキュメントは類似していると判定する（処理ブロック８０６）。超えていない場合、処理論理は、ドキュメント１とドキュメント２は異なると判定する（処理ブロック８０８）。

【００７３】

雑音低減を使用する電子メールスパムフィルタ処理

図９は、電子メール・メッセージの中の雑音を低減するためのプロセス９００の一実施形態の流れ図である。プロセスは、ハードウェア（例えば、専用論理、プログラマブル論理、マイクロコードなど）、ソフトウェア（汎用コンピュータ・システム上、または専用マシン上で実行されるような）、またはそれらの組み合わせを含む処理論理によって実行される。

【００７４】

図９を参照すると、プロセス９００は、処理論理が、電子メール・メッセージの中で、雑音を示すデータを検出することから始まる（処理ブロック９０２）。前述したとおり、雑音は、メール・メッセージの受信者に見えず、スパム・フィルタ処理を回避するために電子メール・メッセージに追加されているデータを表す。そのようなデータには、例えば、フォーマット・データ（例えば、ＨＴＭＬタグ）、数値文字参照、文字エンティティ参照、事前定義されたカテゴリのＵＲＬデータなどが含まれる。数値文字参照は、ドキュメ

ント文字セットのなかの文字のコード位置を指定する。文字エンティティ参照は、シンボリック名を使用して、作成者が、コード位置を憶えている必要がないようにする。例えば、＆ａｒｉｎｇという文字エンティティ参照は、ｒｉｎｇリングの上に置かれた小文字「ａ」を指す。

【００７５】

処理ブロック９０４で、処理論理は、電子メール・メッセージのコンテンツを変更して、雑音を低減する。一実施形態では、コンテンツ変更には、フォーマット・データを除去することや、数値文字参照や文字エンティティ参照をＡＳＣＩＩ等価物に変換すること、ＵＲＬデータを変更することが含まれる。

【００７６】

処理ブロック９０６で、処理論理は、電子メール・メッセージの変更されたコンテンツをスパム・メッセージのコンテンツと比較する。一実施形態では、比較は、厳密な一致を識別するように実行される。代替として、比較は、２つのドキュメントが類似しているかどうかを判定するように実行される。

【００７７】

図１０は、電子メール・メッセージを変更して、雑音を低減させるためのプロセス１００００の一実施形態の流れ図である。プロセスは、ハードウェア（例えば、専用論理、プログラマブル論理、マイクロコードなど）、ソフトウェア（汎用コンピュータ・システム上、または専用マシン上で実行されるような）、またはそれらの組み合わせを含む処理論理によって実行される。

【００７８】

図１０を参照すると、プロセス１０００は、処理論理が、フォーマット・データ（例えば、ＨＴＭＬタグ）を求めて、電子メール・メッセージを調べることから始まる（処理ブロック１００２）。

【００７９】

判定ボックス１００４で、処理論理は、見つかったフォーマット・データが、例外として適格であるかどうかを判定する。通常、ＨＴＭＬフォーマットは、メッセージの情報コンテンツに何も追加しない。しかし、いくつかの例外が存在する。それらの例外は、メッセージのさらなる処理のための有用な情報を含むタグ（例えば、＜ＢＯＤＹ＞、＜Ａ＞、＜ＩＭＧ＞、＜ＦＯＮＴ＞などのタグ）である。例えば、＜ＦＯＮＴ＞タグと＜ＢＯＤＹ＞タグは、「白地に白(white on white)」のテキストを無くすために必要とされ、＜Ａ＞タグと＜ＩＭＧ＞タグは、通常、データをシステムの他のコンポーネントに渡すために使用されるリンク情報を含む。

【００８０】

フォーマット・データが、例外として適格でない場合、そのフォーマット・データは、電子メール・メッセージから抜き取られる（処理ブロック１００６）。

【００８１】

次に、処理論理は、各数値文字参照と各文字エンティティ参照を、対応するＡＳＣＩＩ文字に変換する（処理ブロック１００８）。

【００８２】

ＨＴＭＬでは、数値文字参照は、以下の２つの形態をとる。すなわち、

１．Ｄが１０進数である、「＆＃Ｄ；」という構文が、ＩＳＯ１０６４６の１０進文字番号Ｄを指し、

２．Ｈが１６進数である「＆＃ｘＨ；」または「＆＃ＸＨ；」という構文が、ＩＳＯ１０６４６の１６進文字番号Ｈを指す。数値文字参照における１６進数字は、大文字と小文字の区別がない。

【００８３】

例えば、本文中のランダム化された文字は、以下の表現のとおり現れる。すなわち、

【００８４】

【表４】

Th&#101&#32&#83a&#118&#105n&#103&#115R&#101&#103is
&#116e&#114&#119&#97&#110&#116&#115&#32yo&#117.

【００８５】
この表現は、「Ｔｈｅ　ＳａｖｉｎｇｓＲｅｇｉｓｔｅｒ　ｗａｎｔｓ　ｙｏｕ」という句の意味を有する。
【００８６】
ときとして、処理ブロック１００８において実行される変換は、繰り返される必要がある。例えば、「＆＃３８；」という文字列は、ＡＳＣＩＩにおいて「＆」という文字列に対応し、「＆＃３５；」という文字列は、ＡＳＣＩＩにおいて「＃」という文字列に対応し、「＆＃５１；」という文字列は、ＡＳＣＩＩにおいて３に対応し、「＃５６；」という文字列は、ＡＳＣＩＩにおいて８に対応し、「＃５９；」は、ＡＳＣＩＩにおいて「；」という文字列に対応する。このため、結合された文字列、「＆＃３８；＆＃３５；＆＃５１；＃５６；＃５９；」は、変換されると、変換される必要がある「＆＃３８；」という文字列をもたらす。
【００８７】
したがって、処理ブロック１００８における最初の変換オペレーションの後、処理論理は、変換済みのデータが、数値文字参照または文字エンティティ参照を依然として含むかどうかを調べる（判定ボックス１０１０）。検査が肯定的であった場合、処理論理は、処理ブロック１００８において変換オペレーションを繰り返す。肯定的ではなかった場合、処理論理は、処理ブロック１０１２に進む。
【００８８】
処理ブロック１０１２で、処理論理は、事前定義されたカテゴリのＵＲＬデータを変更する。それらのカテゴリには、例えば、処理論理によって、対応するＡＳＣＩＩ文字に変換される、ＵＲＬの中に含まれる数値文字参照が含まれる。さらに、ＵＲＬ「パスワード」構文を使用して、ＵＲＬホスト名の中で「@」の前に文字を追加することができる。それらの文字は、目標のＷｅｂサーバによって無視されるが、相当な量の雑音情報を各ＵＲＬに追加する。処理論理は、それらの追加の文字を除去することによってＵＲＬデータを変更する。最後に、処理論理は、ＵＲＬの終端における「？」という文字列の後に続く、ＵＲＬの「クエリ」部分を除去する。
【００８９】
ＵＲＬの例は、以下のとおりである。すなわち、
ｈｔｔｐ％３ａ％２ｆ％２ｆｌｏｔｓｏｆｊｕｎｋ@ｗｗｗ．ｆｏｏ．ｃｏｍ％２ｆｂａｒ．ｈｔｍｌ？ｍｕｃｈｍｏｒｅｊｕｎｋ
処理論理は、上記のＵＲＬデータをｈｔｔｐ：／／ｗｗｗ．ｆｏｏ．ｃｏｍ／ｂａｒ．ｈｔｍｌに変更する。
【００９０】
例示的なコンピュータ・システム
図１１は、本明細書で説明するオペレーションの１つまたは複数を実行するのに使用することができる例示的なコンピュータ・システム１１００のブロック図である。代替の実施形態では、マシンは、ネットワーク・ルータ、ネットワーク・スイッチ、ネットワーク・ブリッジ、パーソナル・デジタル・アシスタント（ＰＤＡ）、セルラー電話、Ｗｅｂ機器、あるいはそのマシンによって行われるアクションを指定する一連の命令を実行することができる任意のマシンを含む。
【００９１】
コンピュータ・システム１１００は、バス１１０８を介して互いに通信する、プロセッサ１１０２、メイン・メモリ１１０４、スタティック・メモリ１１０６を含む。コンピュータ・システム１１００は、ビデオ・ディスプレイ・ユニット１１１０（例えば、液晶デ

ィスプレイ（ＬＣＤ）または陰極線管（ＣＲＴ））をさらに含むことが可能である。また、コンピュータ・システム１１００は、英数字入力デバイス１１１２（例えば、キーボード）、カーソル制御デバイス１１１４（例えば、マウス）、ディスク・ドライブ・ユニット１１１６、信号生成デバイス１１２０（例えば、スピーカ）、ネットワーク・インタフェース・デバイス１１２２も含む。

【００９２】

ディスク・ドライブ・ユニット１１１６は、前述した方法のいずれか１つ、またはすべてを実施する命令セット（すなわち、ソフトウェア）１１２６が格納されているコンピュータ可読媒体１１２４を含む。ソフトウェア１１２６は、メイン・メモリ１１０４内、および／またはプロセッサ１１０２内にも、完全に、または部分的に存在するように図示されている。ソフトウェア１１２６は、ネットワーク・インタフェース・デバイス１１２２を介して、さらに送信、または受信される。本明細書では、「コンピュータ可読媒体」という用語は、コンピュータによる実行のための、コンピュータが、本発明の方法のいずれか１つを実行するようにさせる一連の命令を格納する、または符号化することができる任意の媒体を含むと解釈される。したがって、「コンピュータ可読媒体」には、ソリッドステート・メモリ、光ディスクや磁気ディスク、搬送波信号が含まれるが、これらに限定されない。

【００９３】

以上の説明を読んだ後、本発明の多くの代替形態および変更形態が当業者には明白となるに違いないが、例示として図示し、説明したいずれの特定の実施形態も、限定するものと見なされることは全く意図していないことを理解されたい。したがって、様々な実施形態の詳細についての言及は、本発明に不可欠であると見なされる特徴だけを記載する特許請求の範囲を限定することを意図していない。

【図面の簡単な説明】

【００９４】

【図１】スパム電子メールの配信を制御するためのシステムの一実施形態を示すブロック図である。

【図２】スパム・コンテンツ準備モジュールの一実施形態を示すブロック図である。

【図３】類似度算出モジュールの一実施形態を示すブロック図である。

【図４】スパム・メッセージを扱うためのプロセスの一実施形態を示す流れ図である。

【図５】類似点測度に基づいて電子メール・スパムをフィルタ処理するためのプロセスの一実施形態を示す流れ図である。

【図６Ａ】電子メール・メッセージのシグネチャを作成するためのプロセスの一実施形態を示す流れ図である。

【図６Ｂ】電子メール・メッセージのシグネチャを使用してスパムを検出するためのプロセスの一実施形態を示す流れ図である。

【図７】ドキュメントの文字ベースの比較のためのプロセスの一実施形態を示す流れ図である。

【図８】２つのドキュメントが類似しているかどうかを判定するためのプロセスの一実施形態を示す流れ図である。

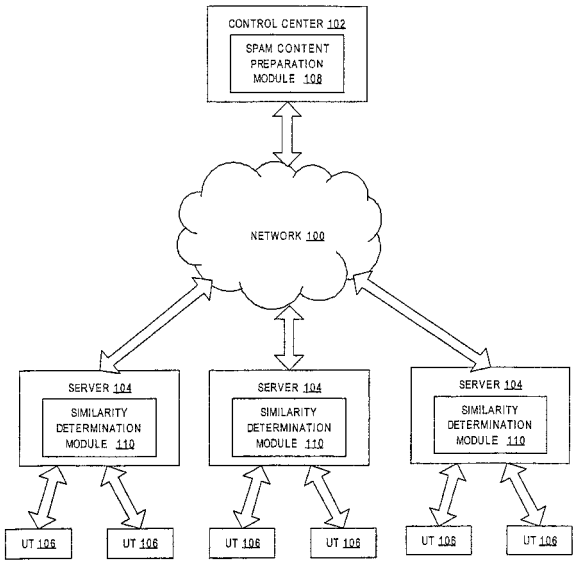
【図９】電子メール・メッセージの中の雑音を低減するためのプロセスの一実施形態を示す流れ図である。

【図１０】電子メール・メッセージを変更して雑音を低減するためのプロセスの一実施形態を示す流れ図である。

【図１１】例示的なコンピュータ・システムを示すブロック図である。

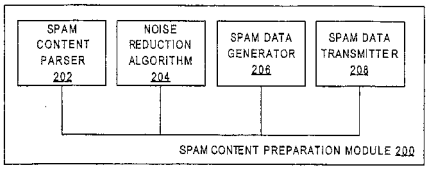
【図１】

１００　ネットワーク
１０２　制御センタ
１０４　サーバ
１０８　スパム・コンテンツ準備モジュール
１１０　類似度算出モジュール



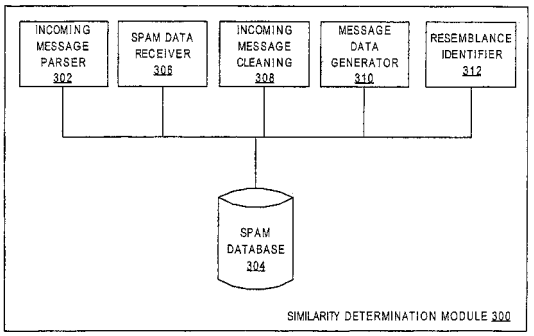
【図２】

２００　スパム・コンテンツ準備モジュール
２０２　スパム・コンテンツ・パーサ
２０４　雑音低減アルゴリズム
２０６　スパム・データ・ジェネレータ
２０８　スパム・データ・トランスミッタ

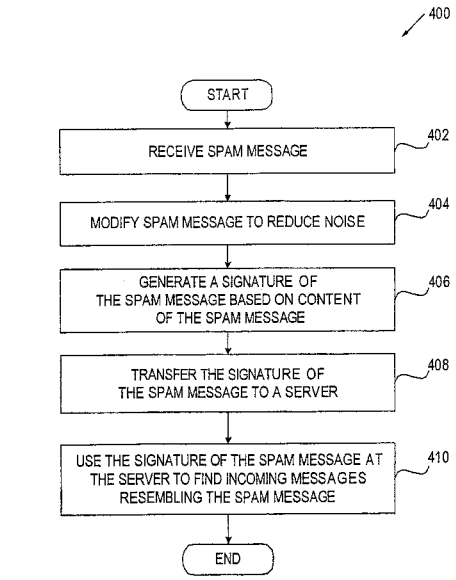


【図３】

３００　類似度算出モジュール
３０２　着信メッセージ・パーサ
３０４　スパム・データベース
３０６　スパム・データ・レシーバ
３０８　着信メッセージ・クリーニング
３１０　メッセージ・データ・ジェネレータ
３１２　類似アイデンティファイア

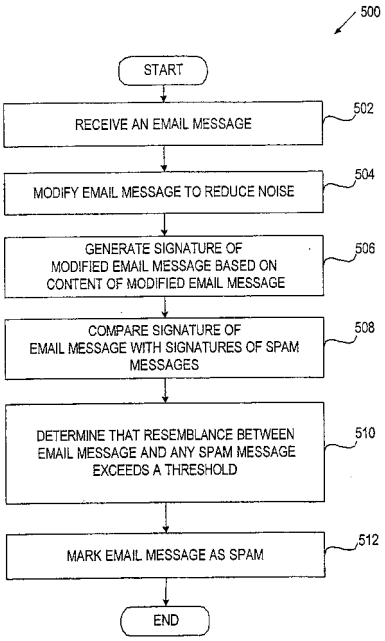


【図４】



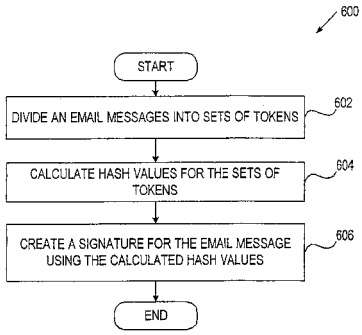
４０２　スパム・メッセージを受信する
４０４　スパム・メッセージを変更して、雑音を低減する
４０６　スパム・メッセージのコンテンツに基づいて、スパム・メッセージのシグネチャを生成する
４０８　スパム・メッセージのシグネチャをサーバに転送する
４１０　サーバにおいてスパム・メッセージのシグネチャを使用して、スパム・メッセージに類似する着信メッセージを探し出す

【図５】



５０２　電子メール・メッセージを受信する
５０４　電子メール・メッセージを変更して、雑音を低減する
５０６　変更された電子メール・メッセージのコンテンツに基づいて、変更された電子メール・メッセージのシグネチャを生成する
５０８　電子メール・メッセージのシグネチャをスパム・メッセージのシグネチャと比較する
５１０　電子メール・メッセージといずれかのスパム・メッセージの間の類似が、閾値を超えているかどうかを判定する
５１２　電子メール・メッセージにスパムとしてマークを付ける

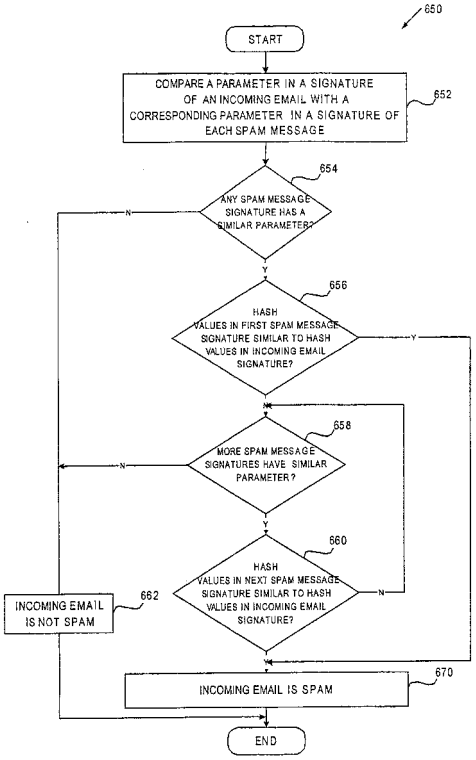
【図６Ａ】



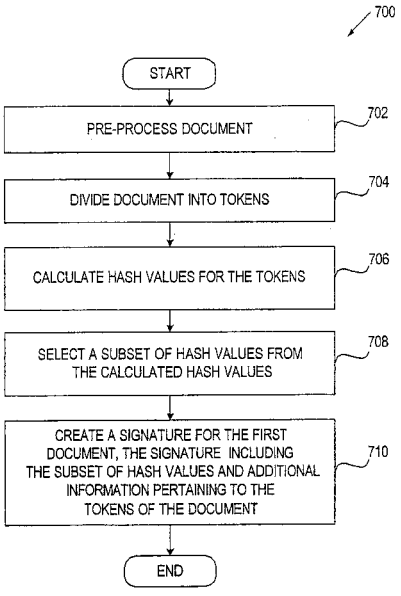
６０２　電子メール・メッセージをトークン・セットに分割する
６０４　トークン・セットに関するハッシュ値を計算する
６０６　計算されたハッシュ値を使用して、電子メール・メッセージに関するシグネチャを作成する

【図６Ｂ】

６５２　着信電子メールのシグネチャの中のパラメータを、各スパム・メッセージのシグネチャの中の対応するパラメータと比較する
６５４　いずれかのスパム・メッセージ・シグネチャが、類似するパラメータを有するか
６５６　最初のスパム・メッセージ・シグネチャの中のハッシュ値は、着信電子メール・シグネチャの中のハッシュ値に類似しているか
６５８　さらなるスパム・メッセージ・シグネチャが、類似したパラメータを有するか
６６０　次のスパム・メッセージ・シグネチャの中のハッシュ値は、着信電子メール・シグネチャの中のハッシュ値に類似しているか
６６２　着信電子メールは、スパムではない
６７０　着信電子メールは、スパムである

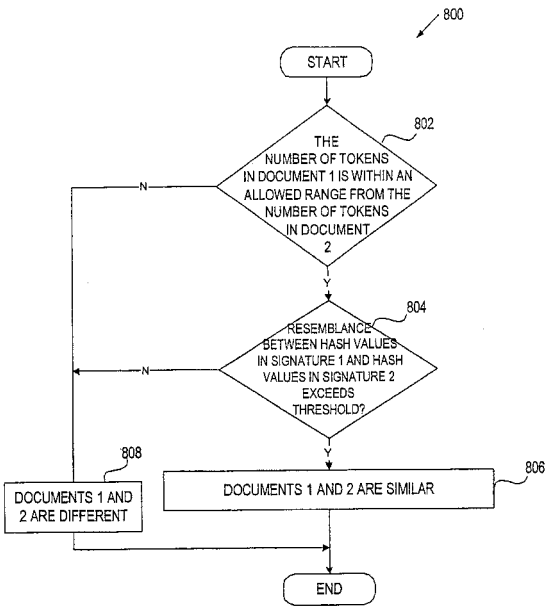


【図７】



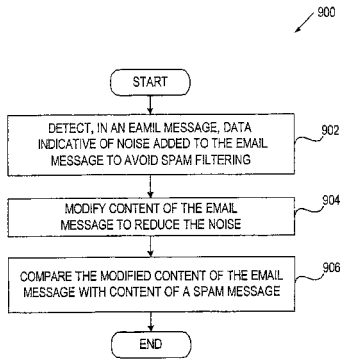
７０２　ドキュメントを前処理する
７０４　ドキュメントをトークンに分割する
７０６　トークンに関するハッシュ値を計算する
７０８　計算されたハッシュ値からハッシュ値のサブセットを選択する
７１０　ハッシュ値のサブセットと、ドキュメントのトークンに関する追加の情報とを含む、第１のドキュメントに関するシグネチャを作成する

【図８】



８０２　ドキュメント１の中のトークンの数は、ドキュメント２の中のトークンの数からの許容される範囲内にあるか
８０４　シグネチャ１の中のハッシュ値とシグネチャ２の中のハッシュ値の間の類似は、閾値を超えているか
８０６　ドキュメント１とドキュメント２は、類似している
８０８　ドキュメント１とドキュメント２は、異なる

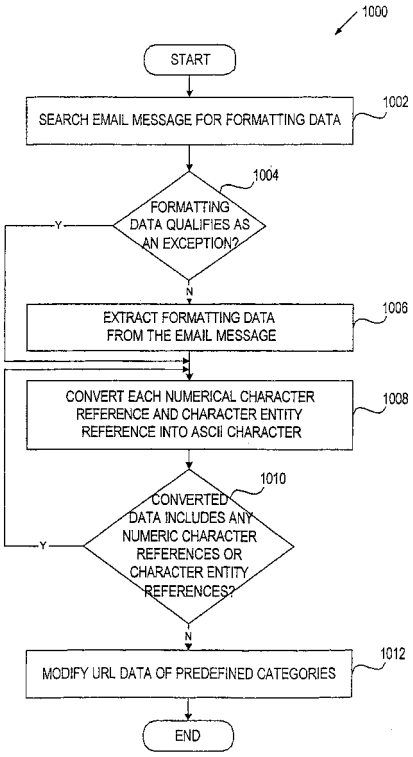
【図９】



９０２　電子メール・メッセージの中で、スパム・フィルタ処理を回避するために電子メール・メッセージに追加された雑音をパラメータを検出する
９０４　電子メール・メッセージのコンテンツを変更して、雑音を低減する
９０６　電子メール・メッセージの変更されたコンテンツをスパム・メッセージのコンテンツと比較する

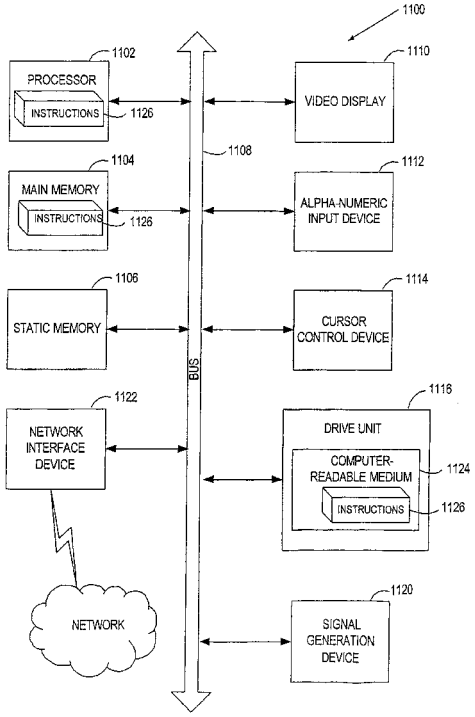
【図１０】

１００２　フォーマット・データを求めて、電子メール・メッセージを調べる
１００４　フォーマット・データは、例外として適格であるか
１００６　電子メール・メッセージからフォーマット・データを抜き取る
１００８　各数値文字参照および各文字エンティティ参照をＡＳＣＩＩ文字に変換する
１０１０　変換されたデータは、何らかの数値文字参照または文字エンティティ参照を含むか
１０１２　事前定義されたカテゴリのＵＲＬデータを変更する



【図１１】

NETWORK:　ネットワーク
１１０２　プロセッサ
１１０４　メイン・メモリ
１１０６　スタティック・メモリ
１１０８　バス
１１１０　ビデオ・ディスプレイ
１１１２　英数字入力デバイス
１１１４　カーソル制御デバイス
１１１６　ドライブ・ユニット
１１２０　信号生成デバイス
１１２２　ネットワーク・インタフェース・デバイス
１１２４　コンピュータ可読媒体
１１２６　命令



フロントページの続き

(72)発明者　グレッソン，マット
アメリカ合衆国
(72)発明者　ホーグストレイト，デイビッド
アメリカ合衆国
(72)発明者　ジェンセン，サンディ
アメリカ合衆国
(72)発明者　マンテル，エリ
アメリカ合衆国
(72)発明者　メドラー，アート
アメリカ合衆国
(72)発明者　シュナイダー，ケン
アメリカ合衆国

審査官　千本　潤介

(56)参考文献　特開平１１－０１５７５６（ＪＰ，Ａ）
特開２０００－３５３１３３（ＪＰ，Ａ）
特開２００６－２９３５７３（ＪＰ，Ａ）
国際公開第２００４／１１４６１４（WO，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 13/00