(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4906273号
(P4906273)
(45)発行日　平成24年3月28日(2012.3.28)　(24)登録日　平成24年1月20日(2012.1.20)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 220Z
G06F 17/30 180Z

請求項の数 25　外国語出願　(全 29 頁)

(21)出願番号　特願2005-147965 (P2005-147965)
(22)出願日　平成17年5月20日 (2005.5.20)
(65)公開番号　特開2005-339545 (P2005-339545A)
(43)公開日　平成17年12月8日 (2005.12.8)
審査請求日　平成20年5月20日 (2008.5.20)
(31)優先権主張番号　10/850,623
(32)優先日　平成16年5月21日 (2004.5.21)
(33)優先権主張国　米国(US)

(73)特許権者　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100077481
弁理士　谷　義一
(74)代理人　100088915
弁理士　阿部　和夫
(72)発明者　バマ　ラマラスナム
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション内

最終頁に続く

(54)【発明の名称】外部データを使用した検索エンジンスパムの検出

(57)【特許請求の範囲】
【請求項１】
少なくともプロセッサ、メモリ、入力手段を有し、ネットワークを経由して電子ドキュメントにアクセス可能なコンピュータによって実行され、検索と関連して電子ドキュメントを評価する方法において、
前記プロセッサが、
電子ドキュメントを解析し、ならびに、前記電子ドキュメントの第１の属性および第２の属性を識別するステップであって、前記電子ドキュメントは、前記入力手段を経由してユーザから受信された検索要求および前記電子ドキュメントが前記要求された検索に関連しているという検索エンジンによる決定に応答して、前記検索エンジンによって検索可能であり、前記第１の属性は電子メールメッセージ属性に対応しており、前記第２の属性は、前記検索要求に対する前記電子ドキュメントの関連性決定を操作するパターンを特性付けているステップと、
ネットワークを経由して前記検索エンジンの外部のソースから、好ましくない電子メールメッセージに関連している電子メールメッセージ属性を含んだ情報を受信するステップと、
前記電子ドキュメントの前記第１の属性に基づいて、前記電子ドキュメントの第１の信頼レベルを決定するステップであって、前記第１の信頼レベルは前記電子ドキュメントが前記好ましくない電子メールメッセージに関連付けられている可能性を示しているステップと、

前記電子ドキュメントの前記第２の属性に基づいて、前記電子ドキュメントの第２の信頼レベルを決定するステップであって、前記第２の信頼レベルは前記電子ドキュメントが前記検索要求に対して不満足なものであるという可能性を示しているステップと、

前記決定された第１の信頼レベルおよび前記決定された第２の信頼レベルの関数として、前記電子ドキュメントに対する格付けを生成するステップと、

前記電子ドキュメントの前記生成された格付けに基づいて、前記検索要求に関して、前記電子ドキュメントが不満足なものであると指定するステップと

を備えることを特徴とする方法。

【請求項２】

前記外部ソースは、電子メールスパム検出システムを含むことを特徴とする請求項１に記載の方法。

【請求項３】

前記電子ドキュメントは、ウェブページおよびマルチメディアファイルの内の１つ以上を含むことを特徴とする請求項１に記載の方法。

【請求項４】

前記電子ドキュメントからリンクされた１つ以上の他の電子ドキュメントについての前記第１の信頼レベルを指定するステップをさらに備えることを特徴とする請求項１に記載の方法。

【請求項５】

前記電子ドキュメントを解析するステップは、前記入力手段を経由して、検索結果の中で前記電子ドキュメントを好ましくないものとして指定するユーザ提供情報を受信するステップに応答して行われることを特徴とする請求項１に記載の方法。

【請求項６】

前記受信された検索要求に応答して、検索結果を前記ユーザに提供するステップと、

前記提供された検索結果中において不満足なものであるとして指定された前記電子ドキュメントを格下げするステップ、

前記提供された検索結果から不満足なものであるとして指定された前記電子ドキュメントを除外するステップ、および

前記電子ドキュメントのランキングが前記提供された検索結果中において所定のランクを超えるときに、前記提供された検索結果中の前記電子ドキュメントの前記ランキングを保持するステップのうちの１つ以上のステップを実施するステップと

をさらに備えることを特徴とする請求項１に記載の方法。

【請求項７】

前記受信された情報は、前記電子メールメッセージ属性が、前記好ましくない電子メールメッセージと関連付けられている予め定められた可能性を含んでおり、前記第１の信頼レベルは、前記予め定められた可能性に基づいていることを特徴とする請求項１に記載の方法。

【請求項８】

前記電子メールメッセージ属性は、ホスト名であって、前記電子ドキュメントの前記第１の属性は、前記電子ドキュメントが前記ホスト名によって提供されたことを示している前記ホスト名に対応していることを特徴とする請求項７に記載の方法。

【請求項９】

前記電子メールメッセージ属性は、ネットワークアドレスであって、前記電子ドキュメントの前記第１の属性は、前記電子ドキュメントが前記ネットワークアドレスに位置していることを示している前記ネットワークアドレスに対応していることを特徴とする請求項７に記載の方法。

【請求項１０】

前記電子メールメッセージ属性は、１つ以上の語であり、前記電子ドキュメントの前記第１の属性は前記１つ以上の語に対応していることを特徴とする請求項７に記載の方法。

【請求項１１】

前記電子メールメッセージ属性は、前記電子メールメッセージ属性に関連した前記電子メールメッセージを望ましくないものと指定するユーザが提供する情報を受信するのに応答して、前記外部ソースによって識別されることを特徴とする請求項１に記載の方法。

【請求項１２】

請求項１乃至１１いずれかの方法を実行するためのコンピュータ実行可能な命令を含むコンピュータ読み取り可能記憶媒体。

【請求項１３】

検索と関連して、ネットワークを経由してアクセス可能な電子ドキュメントを評価するシステムにおいて、

入力手段を経由して、ユーザからの検索要求を受信し、および、電子ドキュメントが前記受信された検索要求に関連しているという決定に基づいて、該電子ドキュメントを識別するプロセッサと、

前記プロセッサの外部のソースによって提供され、好ましくない電子メールメッセージに関連している電子メールメッセージ属性を示しているデータをストアする、前記プロセッサからネットワークを経由してアクセス可能なメモリ領域と

を備え、

前記プロセッサは、前記電子ドキュメントを解析して、前記電子ドキュメントの第１の属性および第２の属性を識別するよう構成され、前記第１の属性は前記電子メールメッセージ属性に対応し、前記第２の属性は前記検索要求に対する前記電子ドキュメントの関連性決定を操作するパターンを特性付けており、

前記プロセッサは、前記電子ドキュメントの前記第１の属性に基づいて、前記電子ドキュメントの第１の信頼レベルを決定するようさらに構成され、前記第１の信頼レベルは前記電子ドキュメントが好ましくない電子メールメッセージに関連付けられている可能性を示しており、

前記プロセッサは、前記電子ドキュメントの前記第２の属性に基づいて、前記電子ドキュメントの第２の信頼レベルを確立するようさらに構成され、前記第２の信頼レベルは前記電子ドキュメントが前記電子ドキュメントの１つ以上の属性に基づいて、検索に対して不満足なものであるという可能性を示しており、

前記プロセッサは、前記決定された第１の信頼レベルおよび前記確立された第２の信頼レベルの関数として、前記電子ドキュメントに対する格付けを生成し、前記電子ドキュメントの前記生成された格付けに基づいて、前記受信された検索要求に関して、前記電子ドキュメントが不満足なものであると分類するようさらに構成されていること

を特徴とするシステム。

【請求項１４】

前記外部ソースは、電子メールスパム検出システムを含むことを特徴とする請求項１３に記載のシステム。

【請求項１５】

前記プロセッサは、前記受信された検索要求に応答して検索結果を前記ユーザに提供し、

前記提供された検索結果中において不満足なものであるとして分類された前記電子ドキュメントを格下げすること、

前記提供された検索結果から不満足なものであるとして分類された前記電子ドキュメントを除外すること、および

前記電子ドキュメントのランキングが前記提供された検索結果中において所定のランクを超えるときに、前記提供された検索結果中の前記電子ドキュメントの前記ランキングを保持すること

のうちの１つ以上を実行するようにさらに構成されていることを特徴とする請求項１３に記載のシステム。

【請求項１６】

前記外部ソースによって提供されるデータは、前記電子メールメッセージ属性が、好ま

しくない電子メールメッセージと関連付けられている予め定められた可能性を含んでおり、前記第１の信頼レベルは、前記予め定められた可能性に基づいていることを特徴とする請求項１３に記載のシステム。

【請求項１７】

前記電子メールメッセージ属性は、ホスト名であって、前記電子ドキュメントの前記第１の属性は、前記電子ドキュメントが前記ホスト名によって提供されたことを示している前記ホスト名に対応していることを特徴とする請求項１６に記載のシステム。

【請求項１８】

前記電子メールメッセージ属性は、ネットワークアドレスであって、前記電子ドキュメントの前記第１の属性は、前記電子ドキュメントが前記ネットワークアドレスに位置していることを示している前記ネットワークアドレスに対応していることを特徴とする請求項１６に記載のシステム。

【請求項１９】

前記電子メールメッセージ属性は、１つ以上の語であり、前記電子ドキュメントの前記第１の属性は前記１つ以上の語に対応していることを特徴とする請求項１６に記載のシステム。

【請求項２０】

検索と関連してネットワークを経由してアクセス可能な電子ドキュメントを評価する機能を、少なくともプロセッサ、メモリ、入力手段を有し、ネットワークを経由して電子ドキュメントにアクセス可能なコンピュータにおいて実現するコンピュータプログラムをストアしたコンピュータ読み取り可能記憶媒体において、前記プログラムは、前記コンピュータに、

ユーザからの検索要求を受信し、および、電子ドキュメントが前記受信された検索要求に関連しているという決定に基づいて、該電子ドキュメントを識別するクエリコンポーネントと、

前記電子ドキュメントが好ましくない電子メールメッセージに関連しているかどうかを評価するのに使用する電子メールメッセージ属性を示しているデータを提供する外部コンポーネントと通信するように構成され、

電子ドキュメントを解析して、前記電子ドキュメントの第１の属性および第２の属性を識別するよう構成され、前記第１の属性は電子メールメッセージ属性に対応しており、前記第２の属性は、前記検索要求に対する前記電子ドキュメントの関連性決定を操作するパターンを特性付けられており、

前記電子ドキュメントの前記第１の属性に基づいて、前記電子ドキュメントが前記好ましくない電子メールメッセージに関連付けられている可能性を示している、電子ドキュメントの第１の信頼レベルを決定するよう構成され、ならびに

前記電子ドキュメントの前記第２の属性に基づいて、前記電子ドキュメントが前記電子ドキュメントの１つ以上の属性に基づいて検索に対して不満足なものであるという可能性を示している、前記電子ドキュメントの第２の信頼レベルを確立するように構成されている内部コンポーネントと、

前記決定された第１の信頼レベルおよび前記確立された第２の信頼レベルの関数として、前記電子ドキュメントの格付けを生成するよう構成された解析コンポーネントと

の各機能を実行させ、

前記クエリコンポーネントは、前記電子ドキュメントの前記生成された格付けに基づいて、前記受信された検索要求に関して、前記電子ドキュメントが不満足なものであると分類するようさらに構成されていることを特徴とするコンピュータ読み取り可能記憶媒体。

【請求項２１】

前記クエリコンポーネントは、前記受信された検索要求に応答して検索結果を前記ユーザに提供し、

前記提供された検索結果中において不満足なものであるとして分類された前記電子ドキュメントを格下げすること、

前記提供された検索結果から不満足なものであるとして分類された前記電子ドキュメントを除外すること、および
前記電子ドキュメントのランキングが前記提供された検索結果中において所定のランクを超えるときに、前記提供された検索結果中の前記電子ドキュメントの前記ランキングを保持すること
のうちの１つ以上を実行するようにさらに構成されていることを特徴とする請求項２０に記載のコンピュータ読み取り可能記憶媒体。
【請求項２２】
前記外部コンポーネントによって提供されるデータは、前記電子メールメッセージ属性が、好ましくない電子メールメッセージと関連付けられている予め定められた可能性を含んでおり、前記第１の信頼レベルは、前記予め定められた可能性に基づいていることを特徴とする請求項２０に記載のコンピュータ読み取り可能記憶媒体。
【請求項２３】
前記電子メールメッセージ属性は、ホスト名であって、前記電子ドキュメントの前記第１の属性は、前記電子ドキュメントが前記ホスト名によって提供されたことを示している前記ホスト名に対応していることを特徴とする請求項２２に記載のコンピュータ読み取り可能記憶媒体。
【請求項２４】
前記電子メールメッセージ属性は、ネットワークアドレスであって、前記電子ドキュメントの前記第１の属性は、前記電子ドキュメントが前記ネットワークアドレスに位置していることを示している前記ネットワークアドレスに対応していることを特徴とする請求項２２に記載のコンピュータ読み取り可能記憶媒体。
【請求項２５】
前記電子メールメッセージ属性は、１つ以上の語であり、前記電子ドキュメントの前記第１の属性は前記１つ以上の語に対応していることを特徴とする請求項２２に記載のコンピュータ読み取り可能記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の実施形態は、通信ネットワークを使用した関連性のあるデータエンティティについての検索の分野に関する。詳細には、本発明の実施形態は、電子ドキュメントの作成者が誤解をさせることにより、検索エンジンにその電子ドキュメントに対して不当に高いランクを付与させる作為的な操作を、外部データを使用して防止することに関する。
【背景技術】
【０００２】
インターネットは、多数のコンピュータ上に分散された膨大な量の情報を有している。したがって、様々な話題に関する大量の情報をユーザに提供している。これは、イントラネットやエクストラネット（ｅｘｔｒａｎｅｔ）など、他のいくつかの通信ネットワークについても言えることである。大量の情報がネットワーク上で利用可能ではあるが、所望の情報を見出すことは、通常簡単でもなく、また迅速にも行うこともできない。
【０００３】
ネットワーク上で所望の情報を見出すという課題に対処するために、検索エンジンが、開発されてきている。一般的に、所望の情報のタイプについてのアイデアを持つユーザは、検索エンジンに対して１つまたは複数の検索用語を入力する。次いで、ユーザが指定した検索用語に関連した電子ドキュメントを含んでいるとこの検索エンジンが判断したネットワークロケーション（例えば、ＵＲＬ）のリストを、この検索エンジンは返す。多数の検索エンジンはまた、関連性ランキング（ｒｅｌｅｖａｎｃｅ　ｒａｎｋｉｎｇ）を提供する。典型的な関連性ランキングは、所与のネットワークロケーションにある電子ドキュメントが、他の電子ドキュメントに対して、相対的にこのユーザが指定した検索用語に関連している可能性についての相対的評価のことである。例えば、従来の検索エンジンは、

特定の検索用語が電子ドキュメント中で出現する回数、その電子ドキュメント中における配置（例えば、しばしば、その電子ドキュメントの末尾に出現する場合に比べて、タイトル中に出現する用語は、より重要であると見なすことができる）に基づいて関連性ランキングを提供することもある。さらに、リンク解析が、ウェブページおよび他のハイパーリンクされたドキュメントをランク付けする際における強力な技法となってきてもいる。アンカーテキスト解析（ａｎｃｈｏｒ－ｔｅｘｔ　ａｎａｌｙｓｉｓ）、ウェブページ構造解析、キー用語リストの使用、およびＵＲＬテキストが、関連性ランキングを提供するのに使用される他の技法である。

【０００４】

電子ドキュメントの作成者は、しばしば、彼らの電子ドキュメントをユーザに表示する作為的な努力を介して、関連性ランキングの問題を複雑にしてしまう。例えば、一部の作成者は、彼らの電子ドキュメントについて、別の方法で保証され得るよりも高いランクの統計を生成するように、検索エンジンを誘導しようと試みる。検索エンジンからの不当に高いランクを達成しようという試みにおける、作成者による電子ドキュメントの作為的な操作は、一般に検索エンジンスパミング（ｓｅａｒｃｈ　ｅｎｇｉｎｅ　ｓｐａｍｍｉｎｇ）と呼ばれる。検索エンジンスパム（ｓｅａｒｃｈ　ｅｎｇｉｎｅ　ｓｐａｍ）の目的は、ユーザを詐欺的に誘導して操作された電子ドキュメントを訪問させることである。操作の一形態は、電子ドキュメント中に（例えば、電子ドキュメントのメタタグ中に）数百個のキー用語を挿入すること、または、１つもしくは複数の検索用語に関するその電子ドキュメントの関連性について検索エンジンを混乱させて過大評価させる（または誤って識別させることさえする）他の技法を利用することを含んでいる。例えば、自動車についてのクラシファイド広告（ｃｌａｓｓｉｆｉｅｄ　ａｄｖｅｒｔｉｓｉｎｇ）ウェブページの作成者は、用語「ｃａｒ（車）」の反復を用いてこの「キー用語」セクションを満たすこともある。この作成者がこれを行う結果、検索エンジンは、ユーザが用語「ｃａｒ」を検索するときはいつでも、そのウェブページをより関連性の高いウェブページとして識別するようになる。しかし、このウェブページのトピックをより正確に表す「キー用語」セクションは、「ａｕｔｏｍｏｂｉｌｅ（自動車）」、「ｃａｒ」、「ｃｌａｓｓｉｆｉｅｄ（クラシファイド）」、および「ｆｏｒ　ｓａｌｅ（売り物）」という用語を含むこともあるだろう。

【０００５】

検索エンジンスパムを作成する他の一部の技法には、実際のユーザに対するものと異なる電子ドキュメントを検索エンジンに対して返すこと（すなわち、クローキング（ｃｌｏａｋｉｎｇ）技法）、電子ドキュメントと無関係のキー用語を対象とすること、キー用語カウントを増大させるためにユーザが見ることのない区域にキー用語を挿入すること、リンクの人気を増大させるためにユーザが見ることのない区域にリンクを挿入すること、低品質のドアウェイウェブページ（ｄｏｏｒｗａｙ　ｗｅｂ　ｐａｇｅ）を生成すること、高ランク付け電子ドキュメントから無関連の電子ドキュメントへとユーザを詐欺的に転送して、無関連の電子ドキュメントをそのユーザに対して表示することなどが含まれる。その結果、検索エンジンは、クエリ（ｑｕｅｒｙ）を実行するユーザに対して、実際には関連していないこともある、高ランク付けされた電子ドキュメントを提供することになる。したがって、この検索エンジンは、このような作為的なランキング操作からユーザを保護をしていない。

【０００６】

既存の検索エンジンは、各スパム技法を別々に解析して操作された電子ドキュメントのパターンを識別することによって、検索エンジンスパムを防止しようと試みる。このような検索エンジンが識別されたパターンを有する電子ドキュメントを検出するとき、この検索エンジンは、この電子ドキュメントにスパムとしてラベルを付けて、この電子ドキュメントを検索結果の形でユーザに表示させず、または、その結果を格下げする。例えば、特定の検索エンジンでは、エンドユーザのためではなくその検索エンジンのために主として構築された電子ドキュメントに、検索エンジンスパムとしてラベルを付けることができる

。同様に、ある検索エンジンでは、電子ドキュメント中の隠しテキストおよび／または隠しリンクを検出し、この電子ドキュメントに検索エンジンスパムとしてラベルを付けることができる。一部の検索エンジンではまた、不必要な多数のホスト名（例えば、ｐｏｋｅｒ．ｆｏｏ．ｃｏｍ、ｂｌａｃｋｊａｃｋ．ｆｏｏ．ｃｏｍなど）を有し、過剰なクロスリンクを使用してそのウェブサイトの外見上の人気を人為的につり上げたウェブサイトを検出し、このウェブサイトに検索エンジンスパムとしてラベルを付けることができる。さらに、既存の検索エンジンでは、クローキング技法、またはウェブサイトが別のウェブサイトと相互リンクを交換して検索エンジンの最適化を増大させるリンクファーミング（ｌｉｎｋ　ｆａｒｍｉｎｇ）を使用したウェブサイトを検出することもできる。

【０００７】

検索エンジンスパムとは著しく違って、電子メール（またはＥメール）スパムは、通常、同時に多数の受信者に送信される迷惑電子メールメッセージである。電子メールスパムは、ジャンクメール（ｊｕｎｋ　ｍａｉｌ）の電子的な等価物である。ほとんどの場合に、電子メールスパムメッセージの内容は、その受信者の関心とは無関係である。したがって、電子メールスパムを作成することは、最小のコストで非常に多数の人々にメッセージを配信するインターネットの不正な使用である。

【発明の開示】

【発明が解決しようとする課題】

【０００８】

電子メールスパムは、いくつかの点で検索エンジンスパムから区別される。例えば、プログラムは、電子メールメッセージを自動的に生成して、電子メールスパムを多数の受信者に送信することができる。対照的に、検索エンジンスパムは、電子メールアドレス、送信者、または受信者を含んではいない。しかし、それにもかかわらず、検索エンジンスパムは、ある種の特徴を電子メールスパムと共有している。例えば、検索エンジンスパムも電子メールスパムも共に、それらが詐欺的にユーザを誘導して特定の製品またはサービスへと訪問させるために作成されているという点において、望ましいものではない。したがって、少なからず電子メールスパムの作成者は、検索エンジンスパムを生成して、製品もしくはサービスに関連した１つまたは複数の電子ドキュメントの露出を増大させようとすることもある。すなわち、スパマー（ｓｐａｍｍｅｒ）はしばしば、電子メールスパムも検索エンジンスパムも共に利用して製品またはサービスを売り込むことがある。したがって、一般的に、電子メールスパムと検索エンジンスパムの間には強い相関がある。それにもかかわらず、従来技術のシステムおよび方法では、電子メールスパムおよび検索エンジンスパムの可能性のあるソースの間におけるこのような相関が見逃されてきている。特に、従来技術においては、電子メールスパムおよび検索エンジンスパムを、まったく異なる解決方法を必要とする別々の問題として取り扱っている。

【０００９】

したがって、検索エンジンスパムを効果的に識別し防止する解決方法が、強く望まれている。

【課題を解決するための手段】

【００１０】

本発明の実施形態では、とりわけ外部ソースを使用して、検索に関連して望ましくないものである得る電子ドキュメントを検出し、それによってより良好な検索エンジンの結果を実現することにより、従来技術における１つまたは複数の欠陥を克服している。本発明の一実施形態によれば、電子メールスパム検出システムは、電子メールメッセージを電子メールスパムの可能性の高いものとして識別する。次いで、データベースなどのメモリ領域は、この電子メールメッセージ中に含まれるリンクのリストを記憶する。本発明の一実施形態は、このデータベースにアクセスし、このデータベースに記憶されたリンクが提供する電子ドキュメントについての信頼レベル（ｃｏｎｆｉｄｅｎｃｅ　ｌｅｖｅｌ）を決定する。この電子ドキュメントの信頼レベルは、この電子ドキュメントが検索エンジンスパムであるという可能性（ｌｉｋｅｌｉｈｏｏｄ）を示す。別の実施形態においては、本

発明は、電子メールスパムの可能性が高い電子メールの起源となるネットワークアドレスを識別する。次いで、このデータベースは、このネットワークアドレスを記憶する。このデータベースにアクセスすることにより、本発明の実施形態は、このネットワークアドレスに位置する電子ドキュメントについての信頼レベルを決定することができる。それによって検索エンジンスパムを、より良好に識別することができる。さらに、この電子メールスパム検出システムは、電子メールスパム中に頻繁に出現する用語（例えば、単語、単語の組合せ、フレーズ、ストリング、ｎ－グラム、バイナリデータなど）のリストを識別することができる。次いで、データベースは、この用語リストを記憶する。したがって、本発明の一実施形態は、この電子ドキュメントが検索エンジンスパムである可能性を示す１つまたは複数のこの記憶された電子メールスパム用語を含んでいる検索エンジンスパムに関して、電子ドキュメントについての信頼レベルを生成する。電子ドキュメントが、検索エンジンスパムであるという高い信頼レベルを有する場合には、本発明の実施形態は、ユーザに提供される検索結果中において、この電子ドキュメントを格下げすることができる。代わりに、本発明の実施形態は、この提供された検索結果からこの電子ドキュメントを除去することもできる。

【００１１】

１つまたは複数の他の実施形態によれば、本発明により、ユーザは電子ドキュメントの望ましさについての情報を提供できるようになる。ユーザは、電子メールスパムまたは検索エンジンスパムに応答して、この情報を提供することができる。このユーザ提供情報がこの電子ドキュメントを望ましくないとして特徴づける場合には、本発明の実施形態は、この電子ドキュメントの１つまたは複数の属性（ａｔｔｒｉｂｕｔｅ）を識別して、この電子ドキュメントについて格付けを生成する。この電子ドキュメントが高い格付けを有する場合には、この電子ドキュメントは、検索エンジンスパムであるという高い可能性を有する。したがって、本発明の実施形態は、検索結果におけるこの電子ドキュメントのランキングを調整して、正確な関連性ランキングをユーザに対して提供することができる。さらに、本明細書中で説明している本発明の実施形態の特徴は、経済的に実現可能であり、商業的に実用可能であり、また現在使用可能な技法に比べて実装するのがより簡単である。

【００１２】

簡単に説明すれば、検索に関連して、本発明の態様を使用した方法は、電子ドキュメントを評価する。この方法は、電子ドキュメントの第１の信頼レベルを決定するステップを含んでいる。この電子ドキュメントは、ユーザからの検索要求に応答して、検索エンジンによって検索が可能である。この第１の信頼レベルは、この検索エンジンの外部のソースが提供する情報に基づいて、この電子ドキュメントが望ましくないものである可能性を示す。この方法は、この電子ドキュメントの第２の信頼レベルを決定するステップも含んでいる。この第２の信頼レベルは、この電子ドキュメントの１つまたは複数の属性に基づいて、この電子ドキュメントがこの検索要求に関して不満足なものである可能性を示す。この方法は、この決定された第１の信頼レベルと決定された第２の信頼レベルの関数としてこの電子ドキュメントについての格付けを生成するステップをさらに含んでいる。この方法は、この電子ドキュメントの生成された格付けに基づいて、この電子ドキュメントを、この検索要求に関して不満足なものであるとして指定するステップも含んでいる。

【００１３】

本発明の他の実施形態においては、検索に関連して、本発明の態様を使用した方法は、電子ドキュメントを評価する。この方法は、電子ドキュメントに関してユーザ提供情報を受信するステップを含んでいる。この電子ドキュメントは、ユーザからの検索要求に応答して、検索エンジンによって検索が可能である。このユーザ提供情報は、この電子ドキュメントを望ましくないものとして特徴づける。この方法はまた、この受信されたユーザ提供情報の関数としてこの電子ドキュメントに関する格付けを生成するステップを含んでいる。この方法は、この電子ドキュメントの生成済みの格付けに従って、この検索要求に関して不満足であるものとして、この電子ドキュメントを指定するステップをさらに含んで

いる。

【００１４】

本発明のさらに他の実施形態においては、本発明の態様を使用したシステムは、検索に関連して、電子ドキュメントを評価する。このシステムは、ユーザからの検索要求を受信し、この受信された検索要求に基づいて電子ドキュメントを識別するためのプロセッサを含んでいる。このシステムはまた、この電子ドキュメントが望ましくないかどうかを評価する際に使用するための、このプロセッサの外部のソースが提供するデータを記憶するメモリ領域を含んでいる。このプロセッサは、この電子ドキュメントの第１の信頼レベルを決定するように構成されている。この第１の信頼レベルは、この外部のソースが提供するデータに基づいて、この電子ドキュメントが望ましくないものである可能性を示す。このプロセッサは、この電子ドキュメントの第２の信頼レベルを設定するようにも構成されている。この第２の信頼レベルは、この電子ドキュメントの１つまたは複数の属性に基づいて、この電子ドキュメントが検索に関連して不満足である可能性を示す。このプロセッサはさらに、この決定された第１の信頼レベルとこの設定された第２の信頼レベルの関数としてこの電子ドキュメントについての格付けを生成し、この電子ドキュメントの生成済みの格付けに基づいて、この受信された検索要求に関して不満足なものであるとして、この電子ドキュメントを分類するように構成される。

【００１５】

本発明のさらに他の実施形態においては、本発明の態様を使用したコンピュータ読取り可能媒体は、検索に関連して、電子ドキュメントを評価するためのコンピュータ実行可能コンポーネントを有している。このコンピュータ読取り可能媒体は、電子ドキュメントに関してユーザ提供情報を受信するためのインターフェースコンポーネント（ｉｎｔｅｒｆａｃｅ　ｃｏｍｐｏｎｅｎｔ）を含んでいる。この電子ドキュメントは、ユーザからの検索要求に応答して、検索可能である。このユーザ提供情報は、この電子ドキュメントを望ましくないとして特徴づける。このコンピュータ読取り可能媒体は、この受信されたユーザ提供情報の関数としてこの電子ドキュメントについての格付けを生成するための解析コンポーネント（ａｎａｌｙｚｉｎｇ　ｃｏｍｐｏｎｅｎｔ）も含んでいる。このコンピュータ読取り可能媒体は、この電子ドキュメントの生成済みの格付けに従って、この検索要求に関して不満足なものであるとして、この電子ドキュメントを分類するためのクエリコンポーネント（ｑｕｅｒｙ　ｃｏｍｐｏｎｅｎｔ）をさらに含んでいる。

【００１６】

本発明のさらに他の実施形態においては、本発明の態様を使用したコンピュータ読取り可能媒体は、検索に関連して、電子ドキュメントを評価するためのコンピュータ実行可能コンポーネントを有している。このコンピュータ読取り可能媒体は、ユーザからの検索要求を受信し、この受信された検索要求に基づいて電子ドキュメントを識別するクエリコンポーネントを含んでいる。このコンピュータ読取り可能媒体はまた、この電子ドキュメントが望ましくないかどうかを評価する際に使用するためのデータを提供する外部コンポーネントも含んでいる。このコンピュータ読取り可能媒体は、この電子ドキュメントの第１の信頼レベルを決定するための内部コンポーネントをさらに含んでいる。この第１の信頼レベルは、この外部コンポーネントが提供するデータに基づいてこの電子ドキュメントが望ましくないものである可能性を示す。この内部コンポーネントは、さらにこの電子ドキュメントの第２の信頼レベルを設定するように構成されている。この第２の信頼レベルは、この電子ドキュメントの１つまたは複数の属性に基づいて、この電子ドキュメントが検索に関連して不満足なものである可能性を示す。コンピュータ読取り可能媒体はまた、この決定された第１の信頼レベルとこの設定された第２の信頼レベルの関数としてこの電子ドキュメントについての格付けを生成する解析コンポーネントも含んでいる。クエリコンポーネントは、この電子ドキュメントの生成済みの格付けに基づいて、この受信された検索要求に関して不満足なものであるとして、この電子ドキュメントを分類するように構成されている。

【００１７】

検索に関連して不満足な電子ドキュメントを検出する方法を実施するためのコンピュータ実行可能命令を有するコンピュータ読取り可能媒体は、本発明のさらなる態様を実施している。

【００１８】

代わりに、本発明の実施形態は、他の様々な方法および装置を含むこともできる。

【００１９】

他の特徴については、本明細書中で以下において部分的には明白となり、また部分的には指摘もしている。以下、同一の参照文字は、図面全体を通して同一の要素を示している。

【発明を実施するための最良の形態】

【００２０】

望ましくない電子ドキュメントを検出するための例示的なネットワーク環境

最初に図１を参照すると、ブロック図が、本発明の実施形態を利用することができる適切なネットワーク環境の一実施例を示している。サーバコンピュータ１０２は、検索エンジン１０４などのプロセッサを含んでいる。この検索エンジン１０４は、クローラ（ｃｒａｗｌｅｒ）１０６をさらに含んでいる。クローラ１０６は、図１に示すリモートサーバコンピュータ１１０やリモートサーバコンピュータ１１２など、通信ネットワーク１０８に接続された１つまたは複数のコンピュータ上に分散された電子ドキュメントを検索する。通信ネットワーク１０８は、イントラネットなどのローカルエリアネットワーク、インターネットなどのワイドエリアネットワーク、またはサーバコンピュータ１０２が、それによってサーバコンピュータ１１０および１１２などのリモートコンピュータと直接的または間接的に情報をやりとりすることができるネットワークの組合せであってもよい。

【００２１】

クローラ１０６は、ネットワーク１０８に接続されたサーバコンピュータ１１０および１１２を検索し、サーバコンピュータ１１０上に記憶された電子ドキュメント１１４および１１６、並びに、サーバコンピュータ１１２上に記憶された電子ドキュメント１１８および１２０を見出す。これらのリモートサーバコンピュータ上に記憶された電子ドキュメントは、ウェブページ（例えば、ＨＴＭＬページおよびＸＭＬページ）、並びに、マルチメディアファイルを含むことができる。クローラ１０６は、これらの電子ドキュメントおよび関連付けられたデータを受信する。さらに、サーバコンピュータ１０２は、クローラ１０６によってアクセスされた電子ドキュメント１２２および１２４を含むことができる。

【００２２】

図１に示すように、検索エンジン１０４の外部にあるソースを構成する電子メールスパム検出システム１２６もまた、ネットワーク１０８に接続される。電子メールスパム検出システム１２６は、システム１２６のユーザに配信される電子メールスパムを検出するシステムである。特に、サーバ１１０および／またはサーバ１１２など１つまたは複数のリモートコンピュータは、システム１２６のユーザに対して電子メールメッセージを生成し送信することができる。次いで電子メールスパム検出システム１２６は、特定の電子メールメッセージが電子メールスパムであり得ることを検出し、そのユーザを保護するアクションを実施する。例えば、システム１２６は、ユーザのメールボックスから検出された電子メールスパムを締め出すことができ、もしくは、ユーザに特定の電子メールメッセージが電子メールスパムであり得ることを警告することができる。あるいは、システム１２６は、その受信者にそのメッセージを配信する前に、信用のけるユーザに対してそのメッセージが電子メールスパムではないことを確認するための電子メールメッセージを配信することもできる。

【００２３】

電子メールスパム検出システム１２６は、電子メールスパムを検出するためにいくつかの技法を利用することができる。１つの技法においては、システム１２６は、電子メールスパムのパターンを識別するようにトレーニングされた確率的分類機構（ｐｒｏｂａｂｉ

ｌｉｓｔｉｃ　ｃｌａｓｓｉｆｉｅｒ）を含んでいる。この確率的分類機構は、電子メールメッセージを分類するコンピュータ実行可能命令を含んでいる。一般に、この確率的分類機構は、電子メールスパムにおいて統計的に重要な属性（例えば、統計的に重要なキー用語および／またはコンテキスト情報）の組合せを識別する。迷惑電子メールメッセージは、しばしば一般的に共有されている一部の属性を含んでいる。このように一般的に共有され、したがって統計的に重要な属性の実施例には、製品またはサービスの非現実的な提案を記述するキー用語（例えば、無料の薬品、減量プログラム、またはクレジットカードの申し込み）が含まれている。さらに、このような属性は、電子メールスパムを送信していた判定される電子メールアドレスを含むこともある。特に、この確率的分類機構をトレーニングして、（例えば、電子メールスパムの「Ｆｒｏｍ：」行に基づいて）電子メールスパムの１人または複数の作成者のドメイン名を識別することができる。次いで、この確率的分類機構は、電子メールメッセージのこの「Ｆｒｏｍ：」行を解析することにより、この電子メールメッセージの送信者が既知の電子メールスパム作成者に対応するかどうかを決定することができる。

【００２４】

同様にこの確率的分類機構をトレーニングして、電子メールスパムが起源とするネットワークアドレスを認識することができる。電子メールスパマーは、しばしば勝手に電子メールスパムの「Ｆｒｏｍ：」行または他の情報を任意の値に設定することができる。しかし、この電子メールスパムの起源となるネットワークアドレス（例えば、ＩＰアドレス）を隠すことは難しい。したがって、電子メールスパムを特徴づけるために、この着信ＳＭＴＰ接続のネットワークアドレスは、この確率的分類機構をトレーニングするのに価値のある属性である。さらに、この確率的分類機構をトレーニングして、電子メールスパムに関連付けられた１つもしくは複数のリンクまたはＵＲＬを識別することもできる。すなわち、電子メールスパムの可能性の高い電子メールに含まれるＵＲＬを特に解析して、電子メールスパムを特徴づける属性を生成する。多数の電子メールメッセージは、埋め込まれたＵＲＬを含んでいる。これらのＵＲＬの存在は、これらの電子メールメッセージが電子メールスパムであることを示すことができる。例えば、これらのＵＲＬは、迷惑な製品やサービスを提案する１つまたは複数のウェブページへと、電子メール受信者を誘導することもある。一実施形態においては、ホスト名（例えば、アルファベットの、小数点付き１０進数の、１６進数の、または、８進符号のホスト名）が、これらのＵＲＬから抽出されて、電子メールスパムを特徴づける助けになる。したがって、組み合わされたＵＲＬが、＜ＵＲＬ１＞@＜ＵＲＬ２＞@．．．@＜ＵＲＬｎ＞の形式である場合には、この最後の@記号の後のＵＲＬ（すなわち、ＵＲＬｎ）が抽出すべきホスト名である。

【００２５】

電子メールスパマーは、このスパマーに関連しているホスト名がこの確率的分類機構によって抽出されないようにするためのリダイレクタ（ｒｅｄｉｒｅｃｔｏｒ）を、ＵＲＬ中に含めることもある。このリダイレクタがＵＲＬ中に含められることにより、この電子メールスパマーに関連するウェブサイトへとこの電子メール受信者を転送する。このようなシナリオにおいては、この確率的分類機構は、リダイレクトＵＲＬ中に隠された本物のホスト名を識別し、電子メールスパムを特徴づける属性としてこの本物のホスト名を使用するように構成される。

【００２６】

説明したように、この確率的分類機構を電子メールスパムの可能性の高い電子メール上でトレーニングして、この電子メールスパムの１つまたは複数の属性を認識することができる。電子メールスパム検出システム１２６は、この確率的分類機構をトレーニングするために、いくつかの技法を使用して電子メールスパムの可能性のある電子メールを識別することができる。１つの技法においては、電子メール受信者は、特定の電子メールメッセージが電子メールスパムであるかどうかを示すことができる。別の技法では、システム１２６は、電子メールスパムを捕捉するためのハニーポット（ｈｏｎｅｙｐｏｔ）を保持する。ハニーポットは、決して存在したことがないか、または、所与の期間に終了してしま

っている電子メールアカウントを表している。しかし、電子メールスパマーにとっては、ハニーポットは、通常の電子メールアカウントであるように見える。したがって、ハニーポットによって表される電子メールアカウントが、決して存在したことがないか、または、ある期間中に終了してしまっており、したがってこの電子メールアカウントが正当な電子メールを受信している理由はないと仮定すれば、ハニーポットに送信された電子メールメッセージは、電子メールスパムと考えることができる。

【００２７】

この確率的分類機構をトレーニングするために電子メールスパムの可能性のある電子メールを識別するさらに他の技法においては、電子メールスパム検出システム１２６は、着信電子メールに対するチャレンジ応答機構（ｃｈａｌｌｅｎｇｅ　ｒｅｓｐｏｎｓｅ）を実装することができる。すなわち、システム１２６は、着信電子メールの送信者に、この電子メールがマシン生成されたものでないことを確認するチャレンジを解決するように要求することができる。この送信者がこのチャレンジを解決できない場合には、システム１２６は、この確率的分類機構がその属性を抽出する電子メールスパムの可能性のあるメールとして、この電子メールを識別することができる。

【００２８】

広範な様々なトレーニング技法を利用して、この確率的分類機構をトレーニングすることができる。スパムとして識別される電子メールおよび非スパムとして識別される電子メールが、コンピュータ実行可能トレーニング命令に供給される。次いで、これらのコンピュータ実行可能トレーニング命令は、スパムとして識別される電子メール中に存在するが、非スパムとして識別される電子メール中には存在しない属性を認識する。したがって、これらの認識された属性は、電子メールスパム中において統計的に意味のあるものとして分類される。これらのコンピュータ実行可能トレーニング命令は、さらに統計的に意味のあるものとして分類される属性ごとに、重みを決定することができる。これらのトレーニング命令は、どれほど頻繁にその属性が電子メールスパム中に現れるかを含めて、いくつかのファクタに基づいて所与の属性についての重みを決定する。これらのコンピュータ実行可能トレーニング命令は、いくつかの異なるアーキテクチャとして実装することができる。例えば、これらのコンピュータ実行可能トレーニング命令は、ナイーブベイジアン分類機構、制限依存性ベイジアン分類機構、ベイジアンネットワーク分類機構、判定ツリー、サポートベクトルマシン、コンテンツマッチング分類機構、最大エントロピー分類機構、およびこれらの組合せなどとして実装することができる。

【００２９】

さらに、システム１２６の確率的分類機構は、パターン認識機構によってトレーニングして、キー用語マッチング技法では識別することができないこともある統計的に意味のある属性の組合せを識別することができる。特に、パターン認識機構がこの確率的分類機構をトレーニングするために使用するこれらの統計技法は、この確率的分類機構が所与の属性の変形を認識することができるように、トレーニングサンプルに基づいて属性を一般化することができる。例えば、この確率的分類機構は、「ｆｒｅｅ　ｓｔｅｒｅｏ　ｐｌ@ｙｅｒ」などの俗語的フレーズを電子メールスパムに関連したものとして認識することができる。対照的に、これらのキー用語マッチング技法では、このような俗語フレーズの変形または他のフレーズの変形を効果的に識別することができないこともある。それにもかかわらず、キー用語マッチングをパターン認識と同時に利用して、この確率的分類機構をトレーニングすることができる。

【００３０】

電子メールメッセージから抽出された属性のその解析に基づいて、この確率的分類機構は、この電子メールメッセージについての格付け（ｒａｔｉｎｇ）を生成する。例えば、確率的分類機構は、電子メールメッセージ中で識別される個々の属性（例えば、用語、ネットワークアドレス、ホスト名など）に対して絶対重みを割り当てることができる。前述のように、所与の属性についての重みは、この確率的分類機構のトレーニングプロセス中に決定される。次いで、この確率的分類機構は、この割り当てられた重みを（例えば、こ

れらの重みの合計を取る）数学的関数に適用することにより、この電子メールメッセージについての格付けを生成する。一実施形態においては、電子メールメッセージの格付けは、パーセンテージ（例えば、６０％）の形式とすることができる。電子メールメッセージの格付けが高くなればなるほど、この電子メールメッセージが電子メールスパムである可能性は高くなる。すなわち、電子メールメッセージの格付けは、この電子メールメッセージが、電子メールスパム中に現れる可能性の高い要素を含んでいる可能性を示す。別の実施形態においては、この確率的分類機構は、特定の属性がその電子メールメッセージ中に出現する頻度、並びに、この電子メールメッセージ中に存在する属性の組合せに基づいて、電子メールメッセージについての格付けを生成する。特に、それ自体では電子メールスパムであることを示すことができない属性は、電子メールメッセージが電子メールスパムを構成するコンテキスト情報または集約情報としての役割を果たすことができる。例えば、属性「クレジットカード」だけでは、電子メールメッセージが電子メールスパムであることを示唆しないこともある。しかし、属性「年会費無料」と組み合わされた属性「クレジットカード」は、この電子メールメッセージが、おせっかいな提案、つまり電子メールスパムを構成することを示唆することができる。

【００３１】

確率的分類機構は、さらに生成された格付けの関数としてこの電子メールメッセージを分類する。すなわち、確率的分類機構が電子メールメッセージについての格付けを生成した後に、確率的分類機構は、この電子メールメッセージがこの格付けに基づいて電子メールスパムを構成するかどうかを判定する。例えば、電子メールスパム検出システム１２６は、その上にしきい値レベル（例えば、７０％）を記憶していることもあり、このしきい値レベルは、所定の電子メールメッセージが望ましくないものである可能性を表す。次いで、確率的分類機構は、この電子メールメッセージの格付けをこのしきい値レベルと比較する。一実施形態においては、この電子メールメッセージの格付けがこのしきい値レベルよりも大きい（またはしきい値レベル以上である）場合には、確率的分類機構は、この電子メールメッセージを電子メールスパムとして分類する。管理者が、このしきい値レベルを変更することにより、電子メールスパム検出システム１２６の感受性を変更することができることに留意されたい。例えば、管理者は、より高いしきい値レベルを設定し、その結果、より少ない電子メールメッセージが電子メールスパムとして分類されるようにすることができる。

【００３２】

ある電子メールメッセージが電子メールスパムの可能性があるものとして分類される場合には、ネットワーク１０８に接続されたデータベース１２８などのメモリ領域に記憶するために、システム１２６はこの電子メールメッセージに関連付けられたある種の属性を抽出する。本発明の一実施形態によれば、システム１２６は、この電子メールメッセージに関連付けられた１つまたは複数のネットワークアドレス（例えば、ＩＰアドレス）を識別する。例えば、システム１２６は、この電子メールメッセージの起源となるネットワークアドレスを識別することができる。したがって、この電子メールメッセージが、サーバ１１０に由来する場合、システム１２６は、サーバ１１０のネットワークアドレスをデータベース１２８に記憶する。本発明の別の実施形態によれば、システム１２６は、さらに電子メールスパムとして分類される電子メールメッセージ中に含まれる１つまたは複数のリンクを識別する。次いでシステム１２６は、データベース１２８中に、この識別されたリンクのホスト名を記憶する。したがって、電子メールスパムとして分類された電子メールメッセージが電子ドキュメント１１４のＵＲＬを含む場合には、システム１２６は、このＵＲＬのホスト名をデータベース１２８に記憶する。さらに、システム１２６は、電子メールスパムに関連付けられた用語リスト（例えば、単語、単語の組合せ、フレーズ、ストリング、ｎ－グラム、２進データなど）を識別する。システム１２６は、この用語リストもまたデータベース１２８に記憶する。

【００３３】

データベース１２８中に記憶されたネットワークアドレス、ホスト名、または用語ごと

に、システム１２６は、さらにこのネットワークアドレス、ホスト名、または用語が電子メールスパムに関連付けられている信頼レベルを指定する。システム１２６は、このネットワークアドレス、ホスト名、または用語を含んでいる電子メールメッセージの格付けに基づいてこの信頼レベルを指定することができる。したがって、この確率的分類機構が特定の電子メールメッセージについて８０％の格付けを生成する場合には、この確率的分類機構は、この電子メールメッセージから識別されるネットワークアドレス、ホスト名、および／または用語について８０％の信頼レベルを指定する。このネットワークアドレス、ホスト名、および／または用語についての指定された信頼レベルは、同様にデータベース１２８に記憶される。

【００３４】

検索エンジン１０４のクローラ１０６が、ネットワーク１０８をナビゲートして、ネットワーク１０８上に配置された１つまたは複数の電子ドキュメントを収集し、検索エンジン１０４のインデックスビルダ（ｉｎｄｅｘ　ｂｕｉｌｄｅｒ）１２９が、この収集された電子ドキュメントを解析して、インデックス付けためのこれらの特徴を識別するときに、検索エンジン１０４は、収集された電子ドキュメントについての別の信頼レベルを設定して、この収集された電子ドキュメントが検索エンジンスパムである（すなわち、検索に関しては不満足なものである）可能性を示すことになる。詳細には、クローラ１０６は、この収集された電子ドキュメントの１つまたは複数のパターンを識別して、これらのパターンが検索エンジンスパムを特徴づけるパターンに対応するかどうかを決定する。例えば、クローラ１０６は、この収集された電子ドキュメントが、エンドユーザのためではなく検索エンジン１０４のために主として構築されているかどうかを識別することができる。クローラ１０６は、さらにこの収集された電子ドキュメントが検索エンジンスパムをしばしば特徴づける隠しテキストおよび／または隠しリンクを含んでいるかどうかを検出することができる。検索エンジンスパムを特徴づける他のパターンの一部には、非常に多数の不必要なホスト名、過剰なクロスリンク、リンクファーミング（ｌｉｎｋ　ｆａｒｍｉｎｇ）などを含んでいるものがある。収集された電子ドキュメントの識別済みのパターンに基づいて、検索エンジン１０４は、この収集された電子ドキュメントが検索エンジンスパムを構成する信頼レベルを生成することができる。

【００３５】

検索エンジン１０４は、さらにデータベース１２８にアクセスしてクローラ１０６によって収集された１つまたは複数の電子ドキュメントに関連した情報を抽出するように構成されている。一実施形態においては、検索エンジン１０４は、データベース１２８に記憶されたネットワークアドレスリストを取得する。検索エンジン１０４が、取得されたネットワークアドレスが収集された電子ドキュメントのロケーションに対応していると判定する場合には、この検索エンジンは、このネットワークアドレスに関連付けられた信頼レベルをデータベース１２８から抽出する。同様に、検索エンジン１０４は、データベース１２８からホスト名リストを取得し、取得されたホスト名が収集された電子ドキュメントを提供するホスト名に対応するかどうかを判定することができる。対応している場合には、検索エンジン１０４は、この取得されたホスト名に関連付けられた信頼レベルをデータベース１２８から抽出する。さらに、検索エンジン１０４は、このホスト名が提供する電子ドキュメントからリンクされた１つまたは複数の電子ドキュメントを、この信頼レベルを有するものとして指定することができる。データベース１２８に記憶された用語について、検索エンジンは、この用語が収集された電子ドキュメント中に現れるかどうかを判定する。この記憶された用語がこの収集された電子ドキュメント中に現れる場合には、検索エンジンは、この記憶された用語に関連付けられた信頼レベルをデータベース１２８から抽出する。

【００３６】

収集された電子ドキュメントが検索エンジンスパムを構成しているという可能性を示す検索エンジン１０４によって決定される信頼レベル、並びに、この収集された電子ドキュメントに関連付けられたネットワークアドレス、ホスト名、および／または用語の信頼レ

ベルに基づいて、検索エンジン１０４は、この収集された電子ドキュメントについての重み付けされた格付けを計算する。特に、ネットワーク１０８のクローリング中に検索エンジン１０４が決定する信頼レベルは、この収集された電子ドキュメントが検索に関して望ましくないものである可能性を示している。また、データベース１２８から取得された１つ（または複数）の信頼レベルは、この収集された電子ドキュメントが望ましくない電子メールメッセージ（すなわち、電子メールスパム）に関連付けられている可能性を示している。電子メールスパムと検索エンジンスパムとの間の所有者の関連性（すなわち、電子メールスパムの作成者は、検索エンジンスパムを生成する可能性が高い）のために、検索エンジン１０４は、これらの２つのタイプの信頼レベルを結合させて、この収集された電子ドキュメントが検索エンジンスパムであるかどうかを、高い信頼度で示す重み付けされた格付けを生成することができる。

【００３７】

電子ドキュメントが検索エンジンスパムを構成するというこの結合された可能性を、高い信頼度で決定する特定の一方法として、格付けを生成するため、様々なタイプの信頼レベルを重み付け平均する。例えば、この電子ドキュメントは検索エンジンスパムであるという６０％の信頼レベルを有し、この電子ドキュメントのネットワークアドレスは電子メールスパムに関連付けらているという８０％の信頼レベルを有し、この電子ドキュメント中に出現する用語は電子メールスパムに関連付けらているという７０％の信頼レベルを有する場合には、検索エンジン１０４は、これらの信頼レベルを平均してこの電子ドキュメントについて７０％の格付けを生成することができる。代わりに、この電子ドキュメントの格付けを検索エンジンスパムである信頼レベル、および、電子メールスパムに関連した信頼レベルの重み付け平均とすることもできる。したがって、上記の実施例において、この電子ドキュメントのネットワークアドレスが電子メールスパムに関連付けられたという８０％の信頼レベルは、この電子ドキュメント中に出現する用語が電子メールスパムに関連付けられたという７０％の信頼レベルと共に重み付けされて、この電子ドキュメントは電子メールスパムに関連しているという７５％の信頼レベルを生成することになる。次いで検索エンジン１０４は、この重み付けされた信頼レベルを検索エンジンスパムであるという６０％の信頼レベルと共に平均して、６７．５％の格付けを生成する。この６７．５％の格付けは、この電子ドキュメントが電子メールスパムを構成しているという重み付けされた確率を示している。

【００３８】

代わりに、これら２つの異なる信頼レベルは、電子ドキュメントがスパムに関連している可能性があるかどうかを判定する相異なるメカニズムを使用しているので、この電子ドキュメントが検索エンジンスパムを構成しているという結合された可能性は、これらのどちらのタイプの信頼レベルよりも高くなり得る。例えば、電子ドキュメントが検索エンジンスパムであることについて７０％の信頼レベルを有し、この電子ドキュメントのネットワークアドレスが電子メールスパムに関連付けられていることについて８０％の信頼レベルを有する場合には、この電子ドキュメントが検索エンジンスパムを構成しているこの結合された可能性は、９０％になることもある。したがって、電子メールスパムの可能性のある電子メールに対する電子ドキュメントの関連付けを考慮することにより、検索エンジン１０４は、この電子ドキュメントが検索エンジンスパムであるかどうかを正確に判定することができる。

【００３９】

検索エンジン１０４は、特定の電子ドキュメントが検索エンジンスパムを構成している可能性があると判定した後に（例えば、この電子ドキュメントの格付けが、あるしきい値レベルよりも大きいときには）、検索エンジン１０４のクエリプロセッサは、様々なアクションを実施して、検索結果においてこの電子ドキュメントをユーザに対して表示しないようにすることができる。したがって、ユーザによって送られた検索要求に基づいて、このクエリプロセッサは、検索エンジンスパムを構成していると判定された電子ドキュメントをこの送信された検索要求の「ヒット（ｈｉｔ）」として、識別することができる。こ

のようなシナリオにおいては、このクエリプロセッサは、このユーザに対して提供される検索結果中において、この電子ドキュメントを格下げすることができる。すなわち、この電子ドキュメントは検索エンジンスパムを構成している可能性があるので、検索エンジン１０４のクエリプロセッサは、この検索結果中においてこの電子ドキュメントのランキングを低下させる。代わりに、このクエリプロセッサは、このユーザに対して提供される検索結果からこの電子ドキュメントを取り除くこともできる。本発明の一実施形態においては、このクエリプロセッサによって取られるアクションを調整することが可能である。すなわち、電子ドキュメントが検索エンジンスパムであることがより確実な場合には、電子ドキュメントはさらに厳しいペナルティが課される。例えば、８５％よりも大きな格付けの電子ドキュメントは、ユーザに対して提供される検索結果から取り除くことができる。一方、６５％と８５％の間の格付けをもつ電子ドキュメントは、検索結果中において５０ランク分だけ格下げをすることができる。また、５０％と６５％の間の格付けをもつ電子ドキュメントは、２５ランク分だけ格下げすることができるが、一方で　５０％以下の格付けを有する電子ドキュメントは、ペナルティを受けないことになる。本発明の別の実施形態においては、電子ドキュメントの事前のランキングが、所定のランク（例えば、５番目のランク）よりも高い場合には、このクエリプロセッサは、検索結果中においてこの電子ドキュメントのランキングを保持する。すなわち、たとえその電子ドキュメントは検索エンジンスパムであると判定されるとしても、高い関連性のある電子ドキュメントは、ペナルティを受けないこともある。

【００４０】

次に図２を参照すると、ブロック図は、本発明の実施形態を利用することができる適切なネットワーク環境の別の実施例を示している。サーバコンピュータ２０２は、検索エンジン２０４を含んでいる。サーバコンピュータ２０２は、通信ネットワーク２０６に接続され、この通信ネットワークは、さらにリモートサーバコンピュータ２０８に接続される。通信ネットワーク２０６は、イントラネットなどのローカルエリアネットワーク、インターネットなどのワイドエリアネットワーク、または、このサーバコンピュータ２０２がリモートサーバコンピュータ２０８などのリモートコンピュータと直接または間接に通信を行うことができるネットワークの組合せであってもよい。リモートサーバコンピュータ２０８は、電子ドキュメント２１０および電子ドキュメント２１２を提供し、これらのドキュメントは、ウェブページまたはマルチメディアファイルとすることができる。さらに、リモートサーバコンピュータ２０８は、ネットワーク２０６に接続されるコンピュータを介して、１つまたは複数の電子メールメッセージをユーザ２１４に対して送信するように構成される。

【００４１】

ユーザ２１４が、サーバコンピュータ２０８から電子メールメッセージを受信した後、このユーザは、この受信された電子メールメッセージを電子メールスパムまたは非スパムとして識別する。次いでユーザ２１４は、電子メールスパム検出システム２１６のインターフェースに対する入力（または一般的にユーザが提供する情報）として、この受信された電子メールメッセージのユーザの識別を送信する。この入力を受信するのに応答して、電子メールスパム検出システム２１６は、この電子メールメッセージが電子メールスパムであることについての信頼レベルを設定する。さらに、システム２１６はこの電子メールメッセージについての複数の入力を複数のユーザから受信し、これらの入力が互いに矛盾する場合には、システム２１６は、この電子メールメッセージについての信頼レベルを設定しないように決定することもできる。他方、これらの入力が互いに一致する場合には、システム１２６は、この電子メールメッセージが電子メールスパムを構成するという信頼レベルを設定することができる。本発明の他の実施形態においては、システム２１６は、ある規則を実装して１つまたは複数の入力を判定することができる。すなわち、ある種の入力には、これらの入力を送信したユーザはより信頼がおけるので、より高い重み付けが行われる。この他の実施形態においては、システム２１６は、特定の電子メールメッセージを電子メールスパムとして報告するユーザのパーセンテージを決定する。これらの大多

数のユーザが、この電子メールメッセージは電子メールスパムであることに合意する場合には、少数のユーザからの入力は、あまり信頼することができない。すなわち、特定のユーザがある電子メールメッセージを電子メールスパムとして報告し、他の大多数のユーザがこの特定のユーザに同意する場合には、システム２１６は、このユーザは信頼がおけると判断することができる。他方、他の大多数のユーザが、この特定のユーザに同意しない場合には、システム２１６は、このユーザは信頼できないと判断することができる。したがって、システム２１６は、ユーザ提供からの入力の信頼性に少なくとも部分的に基づいて、電子メールメッセージについての信頼レベルを決定することができる。

【００４２】

電子メールスパム検出システム２１６が、特定の電子メールメッセージが電子メールスパムを構成すると判定する場合には、この電子メールスパム検出システムはこの電子メールメッセージを解析して、この電子メールメッセージの１つまたは複数の属性を識別して、この電子メールスパムのパターンを決定する。この電子メールメッセージが、あるイメージを含む場合には、システム２１６は、このイメージ中のフレッシュトーン（ｆｌｅｓｈ　ｔｏｎｅ）のレベルを検出することによってこれらの属性を識別する。一実施形態においては、システム２１６は、この電子メールスパムに関連付けられた１つまたは複数の用語を識別することができる。さらに、システム２１６は、この電子メールスパムが起源としているネットワークアドレス（例えば、サーバコンピュータ２０８のネットワークアドレス）を決定することができる。また、システム２１６は、この電子メールスパムに関連付けられたホスト名を識別することができる。例えば、電子ドキュメント２１０および／または電子ドキュメント２１２がこの電子メールスパムからリンクされる場合、システム２１６は、これらの電子ドキュメントのホスト名をこれらのリンクから抽出することができる。別の実施形態において、システム２１６は、ネットワーク２０６に接続されたデータベース２１６などのメモリ領域にこの電子メールスパムに関連したこの識別された属性を記憶する。

【００４３】

サーバコンピュータ２０２の検索エンジン２０４は、データベース２１８にアクセスして、これらの記憶された属性を取得する。これらの記憶された属性に基づいて、検索エンジン２０４は、特定のネットワークアドレスに配置され、もしくは、特定のホスト名によって提供される１つまたは複数の電子ドキュメントについての格付けを生成する。さらに、検索エンジン２０４は、データベース２１８に記憶された用語がネットワーク２０６上に配置される特定の電子ドキュメント中において出現するかどうかを判定して、この電子ドキュメントについての格付けを生成する。電子ドキュメントの格付けは、この電子ドキュメントが検索エンジンスパムである可能性を示す。次いで、検索エンジン２０４は、この電子ドキュメントの格付けがしきい値レベルを超えている場合に、この電子ドキュメントを検索エンジンスパムとして分類する。検索エンジン２０４のクエリプロセッサは、さらにアクションを実施して、（例えば、この検索結果中の電子ドキュメントを格下げし、この検索結果からこの電子ドキュメントを取り除くなどを行って）ユーザに対して正確な検索結果を提供する。

【００４４】

図３を参照すると、ブロック図は、本発明の実施形態が検索とともに電子ドキュメントを評価することができる適切なネットワーク環境のさらに他の実施例を示している。クライアントコンピュータ３０２は、ネットワーク３０６によってサーバコンピュータ３０４に接続される。この場合にも、ネットワーク３０６は、ローカルエリアネットワーク（例えば、イントラネット）、ワイドエリアネットワーク（例えば、インターネット）、または、複数のネットワークの組合せとすることができる。クライアントコンピュータ３０２は、電子ドキュメントの所在を突き止め、ユーザに対して表示する検索ユーザインターフェース３０８（例えば、ブラウザ）または他のマシンアクセス可能プログラミングインターフェース、あるいはプロトコルを含んでいる。

【００４５】

クライアントコンピュータ３０２のユーザが、１つまたは複数の電子ドキュメントを検索したいと望むときに、このユーザはクエリストリング３１０を検索ユーザインターフェース３０８に対して送信する。ユーザがクエリストリング３１０を送信した後に、クライアントコンピュータ３０２は、サーバコンピュータ３０４に配置された検索エンジン３１３のクエリプロセッサ３１２に対してクエリストリング３１０を送信して検索を要求する。この送信されたクエリストリング３１０に基づいて、クエリプロセッサ３１２は、リモートサーバコンピュータ３１６が提供する電子ドキュメント３１４をこの送信されたクエリストリング３１０の「ヒット」として識別する。リモートサーバコンピュータ３１６は、同様にネットワーク３０６に接続されている。次いで、クエリプロセッサ３１２は、クライアントコンピュータ３０２の検索ユーザインターフェース３０８に対して電子ドキュメント３１４、または電子ドキュメント３１４のネットワークロケーションを返す。ユーザが、この返されたネットワークロケーションにアクセスして電子ドキュメント３１４を取得した後、ユーザは、電子ドキュメント３１４を検索エンジンスパムまたは非スパムとして識別することができる。次いでユーザは、このユーザの識別を検索エンジン３１３への入力として送信する。

【００４６】

この入力を受信するのに応答して、検索エンジン３１３は、電子ドキュメント３１４が検索エンジンスパムであることについての信頼レベルを設定する。さらに、検索エンジン３１３は電子ドキュメント３１４についての複数の入力を複数のユーザから受信し、これらの入力が互いに矛盾する場合には、検索エンジン３１３は、電子ドキュメント３１４についての信頼レベルを設定しないように決定することができる。他方、これらの入力が互いに一致する場合には、検索エンジン３１３は、電子ドキュメント３１４が検索エンジンスパムを構成することについての信頼レベルを設定することができる。本発明の他の実施形態においては、検索エンジン３１３は、ある規則を実装して１つまたは複数の入力を判定することができる。すなわち、ある種の入力は、これらの入力を送信したユーザはより信頼がおけるので、より高い重み付けが行われる。この他の実施形態においては、検索エンジン３１３は、電子ドキュメント３１４を検索エンジンスパムとして報告するユーザのパーセンテージを決定する。大多数のこれらのユーザが、この電子ドキュメント３１４は検索エンジンスパムであると同意する場合には、少数のユーザからの入力は、あまり信頼することができない。すなわち、特定のユーザが電子ドキュメント３１４を検索エンジンスパムとして報告し、他の大多数のユーザがこの特定のユーザに同意する場合には、検索エンジン３１３は、このユーザは信頼がおけると判断することができる。他方、他の大多数のユーザが、この特定のユーザに同意しない場合には、検索エンジン３１３は、このユーザは信頼できないと判断することができる。したがって、検索エンジン３１３は、ユーザ提供入力の信頼性に少なくとも部分的に基づいて、特定の電子ドキュメントについての信頼レベルを決定することができる。

【００４７】

このユーザから提供される情報が電子ドキュメント３１４を検索エンジンスパムとして識別する場合には、検索エンジン３１３は電子ドキュメント３１４を解析して、検索エンジンスパムを特徴づける１つまたは複数の属性を検出する。電子ドキュメント３１４がイメージを含んでいる場合には、検索エンジン３１３は、そのイメージ中のフレッシュトーンのレベルを感知することによってこれらの属性を検出する。検索エンジン３１３は、電子ドキュメント３１４の１つまたは複数のパターンを識別して、これらのパターンが検索エンジンスパムを特徴づけるパターンに対応するかどうかを決定することになる。例えば、検索エンジン３１３は、電子ドキュメント３１４がエンドユーザのためにではなく検索エンジン３１３のために主として構築されているかどうかを、識別することができる。検索エンジン３１３は、さらに電子ドキュメント３１４が隠しテキストおよび／または隠しリンクを含んでいるかどうかを検出することができ、これらの隠しテキストおよび／または隠しリンクは、しばしば検索エンジンスパムを特徴づけている。検索エンジンスパムを特徴づける他の一部のパターンは、非常に多数の不必要なホスト名、過剰な相互リンク、

リンクファーミングなどを含んでいる。

【００４８】

この識別されたパターンまたは属性に基づいて、検索エンジン３１３は、電子ドキュメント３１４についての格付けを生成する。電子ドキュメント３１４の格付けは、電子ドキュメント３１４が検索エンジンスパムである可能性を示す。電子ドキュメント３１４の格付けがしきい値レベルを超える場合には、検索エンジン３１３は、電子ドキュメント３１４を検索エンジンスパムとして分類する。クエリプロセッサ３１２は、さらにアクションを実施して、（例えば、その検索結果中の電子ドキュメント３１４を格下げし、その検索結果から電子ドキュメント３１４を取り除くなどして）正確な検索結果をユーザに対して提供する。

【００４９】

望ましくない電子ドキュメントを検出する例示の方法

図４は、本発明の一実施形態による、検索に関する電子ドキュメントを評価するための例示的な方法を示している。４０２において、電子ドキュメントの第１の信頼レベルが決定される。この第１の電子ドキュメントは、ユーザからの検索要求に応答して検索エンジンによって検索可能である。この第１の信頼レベルは、この検索エンジンの外部のソースが提供する情報に基づいて、この電子ドキュメントが望ましくないものである可能性を示す。この外部ソースは、１つまたは複数の電子ドキュメントに関するデータを提供する電子メールスパム検出システムを含むこともある。例えば、この外部ソースは、望ましくないものであるという所定の可能性を有するものとしてこの外部ソースが識別する１つまたは複数の電子ドキュメントを表示するホスト名を、提供することができる。また、これらの電子ドキュメントからリンクされた電子ドキュメントについては、この第１の信頼レベルを指定することができる。この外部ソースは、望ましくないものであるという所定の可能性を有する１つまたは複数の電子ドキュメントが配置されたネットワークアドレスを、提供することもできる。この外部ソースは、さらに望ましくないものであるという所定の可能性を有する１つまたは複数の電子ドキュメント中に出現する用語を、提供することもできる。この電子ドキュメントについての第１の信頼レベルは、この所定の可能性に基づいて決定される。

【００５０】

４０４において、この電子ドキュメントの第２の信頼レベルが決定される。この第２の信頼レベルは、この電子ドキュメントの１つまたは複数の属性に基づいて、この電子ドキュメントがこの検索要求に関して不満足なものである可能性を示している。この電子ドキュメントの望ましくないパターンを特徴づけるこのような属性は、この電子ドキュメントを解析することによって識別される。代わりに、この電子ドキュメントに関連するユーザ提供情報を受信することもできる。このユーザ提供情報は、検索結果中においてこの電子ドキュメントを望ましくないものとして指定する。したがって、次いでこの電子ドキュメントの１つまたは複数の属性を識別して、望ましくないパターンを検出することができる。

【００５１】

４０６において、この電子ドキュメントについての格付けが、決定された第１の信頼レベルおよび決定された第２の信頼レベルの関数として生成される。４０８において、この電子ドキュメントは、この電子ドキュメントの生成された格付けに基づいて、この検索要求に関して不満足なものであるとして指定される。さらに、このユーザからの検索要求に応答して、検索結果をこのユーザに対して提供することもできる。この電子ドキュメントが、不満足なものであるとして指定される場合、この電子ドキュメントは、提供された検索結果から除外することができる。代わりに、この電子ドキュメントを、ユーザに提供される検索結果中で格下げすることもできる。この電子ドキュメントのランキングが検索結果中において所定のランクを超える場合には、この電子ドキュメントのランキングを保持することもできる。

【００５２】

図５は、本発明の一実施形態による、検索に関連して電子ドキュメントを評価するための別の例示的な方法を示している。５０２において、電子ドキュメントに関するユーザ提供情報が受信される。この電子ドキュメントは、ユーザからの検索要求に応答して検索エンジンによって検索可能である。このユーザ提供情報は、この電子ドキュメントを望ましくないものであるとして特徴づける。例えば、この受信されたユーザ提供情報は、この電子ドキュメントが、望ましくない電子メール（例えば、電子メールスパムの可能性のある電子メース）に関連していることを指定することができる。代わりに、この受信済されたユーザ提供情報は、この電子ドキュメントが、検索結果において望ましくないものである（例えば、検索エンジンスパムの可能性のあるドキュメントである）ことを指定することができる。５０４において、この電子ドキュメントについての格付けが、受信されたユーザ提供情報の関数として生成される。例えば、ユーザ提供情報によって望ましくないものであるとして特徴づけられた電子ドキュメントを解析して、この電子ドキュメントの１つまたは複数の属性を識別することができる。次いで、この識別された属性を確率的分類機構に適用して、この電子ドキュメントについての格付けを生成する。識別された属性が望ましいかどうかを認識するため、この確率的分類機構をトレーニングすることができる。また、確率的分類機構は、ナイーブベイジアン分類機構、制限依存性ベイジアン分類機構、ベイジアンネットワーク分類機構、判定ツリー、サポートベクトルマシン、コンテンツマッチング分類機構、最大エントロピー分類機構、および、これらの組合せなどとして実装することができる。

【００５３】

さらに、この受信されたユーザ提供情報の信頼性を決定することができる。また、この電子ドキュメントについての格付けをこの決定された信頼性の関数として生成することができる。一実施形態においては、この電子ドキュメントに関する他のユーザ提供情報を受信することができる。また、他のユーザ提供情報がこの受信されたユーザ提供情報に一致しているかどうかを判定することによって、この信頼性を決定することができる。５０６において、この電子ドキュメントは、この電子ドキュメントの生成された格付けに従って、この検索要求に関して不満足なものであるとして指定される。

【００５４】

例示的なコンピュータ読取り可能媒体

図６は、本発明の一実施形態による例示的なコンピュータ読取り可能媒体６００を示すブロック図である。図６に示すように、コンピュータ読取り可能媒体６００は、クエリコンポーネント６０２、外部コンポーネント６０４、内部コンポーネント６０６、および解析コンポーネント６０８を含んでいる。しかし、コンピュータ読取り可能媒体６００は、任意量のコンピュータ読取り可能媒体とすることができ、コンポーネント、および、各コンポーネントに関連付けられた機能の様々な組合せを含むことができることが企図されている。クエリコンポーネント６０２は、ユーザからの検索要求を受信し、この受信された検索要求に基づいて電子ドキュメントを識別する。外部コンポーネント６０４は、この電子ドキュメントが望ましくないものであるかどうかを評価する際に使用するためのデータを提供する。内部コンポーネント６０６は、この電子ドキュメントの第１の信頼レベルを決定するように構成される。この第１の信頼レベルは、外部コンポーネント６０４が提供するデータに基づいて、この電子ドキュメントが望ましくないものである可能性を示す。例えば、外部コンポーネント６０４が提供するデータは、１つまたは複数のホスト名を識別する。各ホスト名は、望ましくないものであるという所定の可能性を有する情報を提供する。内部コンポーネント６０６は、これら提供された名前のうちの１つが提供しているものとして、この電子ドキュメントを識別するように構成される。また、内部コンポーネント６０６はさらに、この電子ドキュメントを識別することに応答して、この電子ドキュメントについてのこの所定の可能性に基づいた第１の信頼レベルを、これらのホスト名のうちの１つが提供しているものとして指定するようにも構成されている。

【００５５】

同様に、外部コンポーネント６０４が提供するデータは、１つまたは複数のネットワー

クアドレスを識別することができる。外部コンポーネント６０４は、これらのネットワークアドレスのうちの１つに配置された１つまたは複数の電子ドキュメントを、望ましくないものであるという所定の可能性を有するものとして識別する。内部コンポーネント６０６は、これらのネットワークアドレスのうちの１つに配置されたものとして、これらの電子ドキュメントを識別するように構成される。また、内部コンポーネント６０６は、この電子ドキュメントを識別するのに応答して、この電子ドキュメントについてのこの所定の可能性に基づいた第１の信頼レベルを、これらのネットワークアドレスのうちの１つに配置されたものとして指定するように構成される。

【００５６】

さらに、外部コンポーネント６０４が提供するデータは、これらの用語のうちの少なくとも１つが出現する１つまたは複数の電子ドキュメントが、望ましくないものであるという所定の可能性を有するように、１つまたは複数の用語を識別することができる。内部コンポーネント６０６は、いつこれらの用語のうちの少なくとも１つがこの電子ドキュメント中に出現するかを決定するように、構成される。内部コンポーネント６０６はまた、これらの用語のうちの少なくとも１つがこの電子ドキュメント中に出現することを決定するのに応答して、この電子ドキュメントについての第１の信頼レベルを指定するようにも構成される。この第１の信頼レベルは、この所定の可能性に基づいている。

【００５７】

内部コンポーネント６０６はまた、この電子ドキュメントの第２の信頼レベルも設定する。この第２の信頼レベルは、この電子ドキュメントの１つまたは複数の属性に基づいて、この電子ドキュメントは検索に関連して不満足なものである可能性を示す。このような属性は、この検索に関するこの電子ドキュメントの望ましくないパターンを特徴づける。

【００５８】

解析コンポーネント６０８は、決定された第１の信頼レベルおよび設定された第２の信頼レベルの関数として、この電子ドキュメントについての格付けを生成する。クエリコンポーネント６０２は、この電子ドキュメントの生成された格付けに基づいて、この受信された検索結果に関連してこの電子ドキュメントを不満足なものであるとして分類するように構成される。クエリコンポーネント６０２はまた、この受信された検索要求に応答して検索結果をそのユーザに対して提供する。また、クエリコンポーネント６０２は、この提供された検索結果中において不満足なものであるとして分類された電子ドキュメントを格下げし、または、不満足なものであるとして分類された電子ドキュメントをこの提供される検索結果から除外することができる。代わりに、この電子ドキュメントのランキングが提供された検索結果中の所定のランクを越えるときには、クエリコンポーネント６０２は、提供された検索結果中のこの電子ドキュメントのランキングを保持することもできる。

【００５９】

図７は、本発明の一実施形態による例示的な別のコンピュータ読取り可能媒体７００を示すブロック図である。図６に示すように、コンピュータ読取り可能媒体７００は、インターフェースコンポーネント７０２、解析コンポーネント７０４、および、クエリコンポーネント７０６を含んでいる。しかし、コンピュータ読取り可能媒体７００は、任意量のコンピュータ読取り可能媒体とすることができ、コンポーネント、および各コンポーネントに関連付けられた機能の様々な組合せを含むことができることが企図されている。インターフェースコンポーネント７０２は、電子ドキュメントに関するユーザ提供情報を受信する。この電子ドキュメントは、ユーザからの検索要求に応答して検索可能である。ユーザ提供情報は、この電子ドキュメントを望ましくないものであるとして特徴づける。例えば、この受信されたユーザ提供情報は、この電子ドキュメントが望ましくない電子メールのソースに関連付けられることを指定することができる。この受信されたユーザ提供情報は、この電子ドキュメントが検索結果中で望ましくないものであることを指定することもできる。

【００６０】

解析コンポーネント７０４は、この受信されたユーザ提供情報の関数としてこの電子ド

キュメントについての格付けを生成する。一実施形態においては、解析コンポーネント７０４は、この電子ドキュメントを解析してこの電子ドキュメントの１つまたは複数の属性を識別する。解析コンポーネント７０４はさらに、この識別された属性が望ましくないかどうかを認識するためにトレーニングされる確率的分類機構に対してこの識別された属性を適用して、この電子ドキュメントについての格付けを生成する。別の実施形態においては、解析コンポーネント７０４は、この受信されたユーザ提供情報の信頼性を決定し、この電子ドキュメントについての格付けをこの決定された信頼性の関数として生成する。例えば、インターフェースコンポーネント７０２は、この電子ドキュメントに関する他のユーザ提供情報を受信することができる。次いで、解析コンポーネント７０４は、他のユーザ提供情報が受信されたユーザ提供情報に対応するかどうか検査して、受信された電子ドキュメントの信頼性を決定する。解析コンポーネント７０４がこの電子ドキュメントについての格付けを生成した後に、クエリコンポーネント７０６は、この電子ドキュメントの生成された格付けに従って、この検索要求に関してこの電子ドキュメントを不満足なものであるとして分類する。

【００６１】

例示の動作環境

図８は、コンピュータ１３０の形態における汎用コンピューティングデバイスの一実施例を示している。本発明の一実施形態においては、コンピュータ１３０などのコンピュータは、本明細書中において例示し説明している他の図において使用するために適している。コンピュータ１３０は、１つまたは複数のプロセッサまたは処理装置１３２と、システムメモリ１３４を有している。この例示的な実施形態において、システムバス１３６は、システムメモリ１３４を含めて様々なシステムコンポーネントをプロセッサ１３２に結合している。バス１３６は、メモリバスまたはメモリコントローラ、ペリフェラルバス、アクセラレーティドグラフィックスポート、および様々なバスアーキテクチャのうちのいずれかを使用したプロセッサバスまたはローカルバスを含めて、いくつかのタイプのバス構造のうちの１つまたは複数の任意のバス構造を表す。実施例として、限定するものではないが、かかるアーキテクチャには、ＩＳＡバス、ＭＣＡバス、ＥＩＳＡバス、ＶＥＳＡローカルバス、およびメザニンバスとしても知られているＰＣＩバスが含まれる。

【００６２】

コンピュータ１３０は、一般的に少なくとも一部の形態のコンピュータ読取り可能媒体を有する。コンピュータ読取り可能媒体は、コンピュータ１３０がアクセスすることができる使用可能な任意の媒体とすることができ、揮発性媒体も不揮発性媒体も、着脱可能媒体も着脱不能媒体も共に含んでいる。実施例として限定するものではないが、コンピュータ読取り可能媒体は、コンピュータストレージ媒体および通信媒体を含んでいる。コンピュータストレージ媒体は、コンピュータ読取り可能命令、データ構造、プログラムモジュール、他のデータなどの情報のストレージのための任意の方法または技術で実装される揮発性媒体および不揮発性媒体、着脱可能媒体および着脱不能媒体を含んでいる。例えば、コンピュータストレージ媒体には、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリまたは他のメモリ技術、ＣＤ－ＲＯＭ、ＤＶＤ（ｄｉｇｉｔａｌ　ｖｅｒｓａｔｉｌｅ　ｄｉｓｋデジタル多用途ディスク）または他の光ディスクストレージ、磁気カセット、磁気テープ、磁気ディスクストレージまたは他の磁気ストレージデバイス、あるいは所望の情報を記憶するために使用することができ、コンピュータ１３０がアクセスすることができる他の任意の媒体が含まれる。通信媒体は、一般的に搬送波や他の搬送メカニズムなどの変調データ信号の形のコンピュータ読取り可能命令、データ構造、プログラムモジュールまたは他のデータを実施し、任意の情報配信媒体を含んでいる。信号中の情報を符号化するようにして、その信号中の１つまたは複数の特性を設定または変更された被変調データ信号については、当業者ならよく理解されよう。有線ネットワークや直接配線接続などの有線媒体と、音響、ＲＦ、赤外線、他の無線媒体などの無線媒体は、通信媒体の実施例である。以上のうちの任意の組合せもまた、コンピュータ読取り可能媒体の範囲内に含められる。

【００６３】
システムメモリ１３４は、着脱可能および／または着脱不能な、揮発性および／または不揮発性のメモリの形態のコンピュータストレージ媒体を含んでいる。例示の実施形態においては、システムメモリ１３４は、ＲＯＭ１３８およびＲＡＭ１４０を含んでいる。起動中などコンピュータ１３０内のエレメント間で情報を転送する助けをする基本ルーチンを含むＢＩＯＳ（ｂａｓｉｃ　ｉｎｐｕｔ／ｏｕｔｐｕｔ　ｓｙｓｔｅｍ基本入出力システム）１４２は、一般的にＲＯＭ１３８に記憶される。ＲＡＭ１４０は、処理装置１３２にとって直ちにアクセス可能な、もしくは、処理装置１３２によって現在動作させられている、または、その両方が行われるデータおよび／またはプログラムモジュールを一般的に含んでいる。実施例として、限定するものではないが、図８は、オペレーティングシステム１４４、アプリケーションプログラム１４６、他のプログラムモジュール１４８、およびプログラムデータ１５０を示している。
【００６４】
コンピュータ１３０は、他の着脱可能／着脱不能な、揮発性／不揮発性のコンピュータストレージ媒体を含むこともできる。例えば、図８は、着脱不能な不揮発性磁気媒体から情報を読み取りまたはそれに情報を書き込むハードディスクドライブ１５４を示している。図８はまた、着脱可能な不揮発性磁気ディスク１５８から情報を読み取りまたはそれに情報を書き込む磁気ディスクドライブ１５６、およびＣＤ－ＲＯＭや他の光媒体など着脱可能な不揮発性の光ディスク１６２から情報を読み取りまたはそれに情報を書き込む光ディスクドライブ１６０を示している。この例示の動作環境中において使用することができる他の着脱可能／着脱不能な、揮発性／不揮発性のコンピュータストレージ媒体には、それだけには限定されないが、磁気テープカセット、フラッシュメモリカード、デジタル多用途ディスク、デジタルビデオテープ、ソリッドステートＲＡＭ、ソリッドステートＲＯＭなどが含まれる。ハードディスクドライブ１５４、ならびに磁気ディスクドライブ１５６および光ディスクドライブ１６０は、一般的にインターフェース１６６などの不揮発性メモリインターフェースによってシステムバス１３６に接続される。
【００６５】
前述したように、図８に示されるこれらのドライブまたは他のマスストレージデバイスおよびその関連するコンピュータストレージ媒体は、コンピュータ１３０についてのコンピュータ読取り可能命令、データ構造、プログラムモジュールおよび他のデータのストレージを提供する。図８において、例えばハードディスクドライブ１５４は、オペレーティングシステム１７０、アプリケーションプログラム１７２、他のプログラムモジュール１７４、およびプログラムデータ１７６を記憶するものとして示されている。これらのコンポーネントは、オペレーティングシステム１４４、アプリケーションプログラム１４６、他のプログラムモジュール１４８、およびプログラムデータ１５０と同じとすることもでき、また異なるものとすることもできることに留意されたい。オペレーティングシステム１７０、アプリケーションプログラム１７２、他のプログラムモジュール１７４、およびプログラムデータ１７６には、少なくともこれらが異なるコピーであることを示すために、ここでは異なる番号が付与されている。
【００６６】
ユーザは、キーボード１８０やポインティングデバイス１８２（例えばマウス、トラックボール、ペン、もしくはタッチパッド）などの入力デバイスまたはユーザインターフェース選択デバイスを介してコンピュータ１３０にコマンドおよび情報を入力することができる。他の入力デバイス（図示せず）は、マイクロフォン、ジョイスティック、ゲームパッド、サテライトディッシュ、スキャナなどを含むことができる。これらおよび他の入力デバイスは、システムバス１３６に結合されるユーザ入力インターフェース１８４を介して処理装置１３２に接続されるが、これらは、パラレルポート、ゲームポート、ＵＳＢなど他のインターフェースおよびバス構造によって接続することもできる。モニタ１８８または他のタイプのディスプレイデバイスもまた、ビデオインターフェース１９０などのインターフェースを介してシステムバス１３６に接続される。モニタ１８８に追加して、コ

ンピュータは、しばしばプリンタやスピーカなど他のペリフェラル出力デバイス（図示せず）を含んでおり、これらは、出力ペリフェラルインターフェース（図示せず）を介して接続することができる。

【００６７】

コンピュータ１３０は、リモートコンピュータ１９４など１つまたは複数のリモートコンピュータに対する論理接続を使用して、ネットワーク環境中で動作することができる。リモートコンピュータ１９４は、パーソナルコンピュータ、サーバ、ルータ、ネットワークＰＣ、ピアデバイス、または他の共通ネットワークノードとすることができ、また一般的にコンピュータ１３０に関連した前述のエレメントの多くまたはすべてを含んでいる。図８に示す論理接続は、ＬＡＮ１９６およびＷＡＮ１９８を含んでいるが、他のネットワークを含むこともできる。ＬＡＮ１３６および／またはＷＡＮ１３８は、有線ネットワーク、無線ネットワーク、これらの組合せなどとすることができる。このようなネットワーキング環境は、オフィス、企業規模のコンピュータネットワーク、イントラネット、およびグローバルコンピュータネットワーク（例えば、インターネット）においては、一般的なものである。

【００６８】

ローカルエリアネットワーキング環境中で使用する際には、コンピュータ１３０は、ネットワークインターフェースまたはアダプタ１８６を介してＬＡＮ１９６に接続される。ワイドエリアネットワーキング環境中で使用する際には、コンピュータ１３０は、一般的にインターネットなどのＷＡＮ１９８上で通信を確立するためのモデム１７８または他の手段を含んでいる。モデム１７８は、内蔵でも外付けでもよいが、ユーザ入力インターフェース１８４または他の適切なメカニズムを介してシステムバス１３６に接続される。ネットワーク環境においては、コンピュータ１３０に関連して示されるプログラムモジュール、またはその一部分は、リモートメモリストレージデバイス（図示せず）に記憶することができる。実施例として、限定するものではないが、図８は、リモートアプリケーションプログラム１９２をこのメモリデバイス上に存在するものとして示している。図に示すこのネットワーク接続は、例示的なものであり、コンピュータ間で通信リンクを確立する他の手段を使用することもできる。

【００６９】

一般に、コンピュータ１３０のデータプロセッサは、このコンピュータの様々なコンピュータ読取り可能ストレージ媒体に異なる時刻において記憶される命令の手段によってプログラムされる。プログラムおよびオペレーティングシステムは、一般的に例えばフロッピー（登録商標）ディスクまたはＣＤ－ＲＯＭ上に記憶されて配布される。そこから、これらは、コンピュータの二次メモリ中にインストールされ、またはロードされる。実行に際して、これらは、少なくとも部分的にこのコンピュータの一次電子メモリへとロードされる。このような媒体が、マイクロプロセッサまたは他のデータプロセッサに関連して以下で説明しているステップを実装するための命令またはプログラムを含んでいるときに、本明細書中で説明している本発明の実施形態は、これらおよび他の様々なタイプのコンピュータ読取り可能ストレージ媒体を含んでいる。本明細書中で説明している方法および技法に従ってプログラムされる際に、本発明の一実施形態はまた、コンピュータそれ自体も含んでいる。

【００７０】

図８では、このオペレーティングシステムなどプログラムおよび他の実行可能なプログラムコンポーネントは、本明細書中において個別のブロックとして示されている。しかし、かかるプログラムおよびコンポーネントは、様々な時刻にこのコンピュータの異なるストレージコンポーネント中に存在し、このコンピュータの１つ（または複数）のデータプロセッサによって実行されることが理解されよう。

【００７１】

コンピュータ１３０を含む例示的なコンピューティングシステム環境に関して説明しているが、本発明の一実施形態は、他の非常に多数の汎用または専用のコンピューティング

システム環境またはコンピューティングシステム構成と共に動作可能である。このコンピューティングシステム環境は、本発明の実施形態の使用または機能の範囲についてどのような限定を示唆することも意図してはいない。さらに、このコンピューティングシステム環境は、この例示的な動作環境中に示されるコンポーネントのどの１つもしくは組合せに関連したどのような依存性または必要要件を有するものとも解釈すべきではない。本発明の実施形態と共に使用するために適切とすることができるよく知られているコンピューティングシステム、コンピューティング環境、および／またはコンピューティングコンフィギュレーションの実施例には、それだけには限定されないが、パーソナルコンピュータ、サーバコンピュータ、ハンドヘルドデバイスまたはラップトップデバイス、マルチプロセッサシステム、マイクロプロセッサベースのシステム、セットトップボックス、プログラマブルな一般消費電子製品、移動電話、ネットワークＰＣ、ミニコンピュータ、メインフレームコンピュータ、以上のシステムもしくはデバイスのうちのいずれかを含んでいる分散コンピューティング環境などが含まれる。

【００７２】

本発明の実施形態は、１台もしくは複数台のコンピュータまたは他のデバイスによって実行されるプログラムモジュールなどのコンピュータ実行可能命令の一般的な文脈で説明することができる。一般に、プログラムモジュールには、それだけには限定されないが、特定のタスクを実施し、または特定の抽象データ型を実装するルーチン、プログラム、オブジェクト、コンポーネント、およびデータ構造が含まれる。本発明の実施形態は、タスクが、通信ネットワークを介してリンクされるリモート処理デバイスによって実施される分散コンピューティング環境中においても実行することができる。分散コンピューティング環境においては、プログラムモジュールは、メモリストレージデバイスを含めて、ローカルコンピュータストレージ媒体上にもリモートコンピュータストレージ媒体上にも配置することができる。

【００７３】

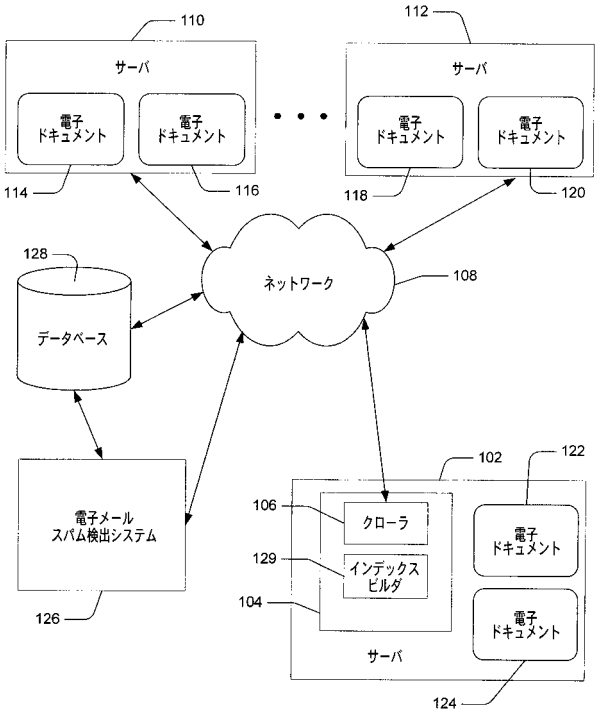
動作中には、コンピュータ１３０は、本明細書中で説明している命令などのコンピュータ実行可能命令を実行して、検索に関連して電子ドキュメントを評価する。コンピュータ実行可能命令は、電子ドキュメントの第１の信頼レベルを決定するように構成される。この電子ドキュメントは、ユーザからの検索要求に応答して検索エンジンによって検索可能である。第１の信頼レベルは、この検索エンジンの外部にあるソースが提供する情報に基づいて、この電子ドキュメントが望ましくない可能性を示す。コンピュータ実行可能命令はまた、この電子ドキュメントの第２の信頼レベルを決定するようにも構成される。この第２の信頼レベルは、この電子ドキュメントの１つまたは複数の属性に基づいてこの検索要求に関してこの電子ドキュメントが不満足なものである可能性を示す。コンピュータ実行可能命令は、さらにこの決定された第１の信頼レベルおよびこの決定された第２の信頼レベルの関数として、この電子ドキュメントについての格付けを生成するように構成される。コンピュータ実行可能命令はまた、この電子ドキュメントのこの生成された格付けに基づいて、この検索要求に関してこの電子ドキュメントを不満足なものであるものとして指定するようにも構成される。

【００７４】

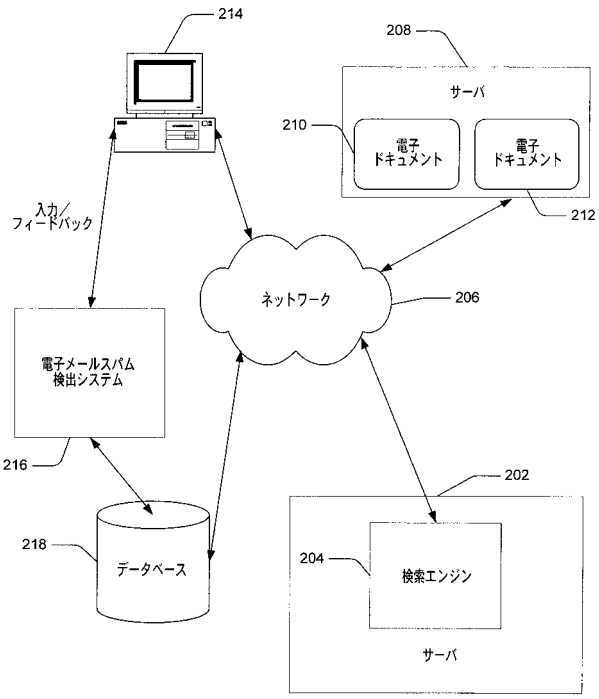
コンピュータ１３０はまた、本明細書中で説明している命令などコンピュータ実行可能命令を実行して、検索に関連して電子ドキュメントを評価する。コンピュータ実行可能命令は、電子ドキュメントに関するユーザ提供情報を受信するように構成される。この電子ドキュメントは、ユーザからの検索要求に応答して検索エンジンによって検索可能である。このユーザ提供情報は、この電子ドキュメントを望ましくないものであるとして特徴づける。コンピュータ実行可能命令はまた、この受信されたユーザ提供情報の関数として、この電子ドキュメントについての格付けを生成するようにも構成される。コンピュータ実行可能命令は、さらにこの電子ドキュメントの生成された格付けに従って、この検索要求に関してこの電子ドキュメントを不満足なものであるとして指定するようにも構成される。

【００７５】
本明細書中で示し、説明している方法の実行または実施の順序は、他に特に指定していない限り本質的なものではない。すなわち、これらの方法の要素は、他に特に指定していない限りどのような順序で実施することもでき、またこれらの方法は、本明細書中に開示されているよりも多い要素または少ない要素を含むこともできることを本発明者らは企図している。
【００７６】
本発明または本発明の実施形態の要素を導入するときに、冠詞「１つの（ａ）」、「１つの（ａｎ）」、「その（ｔｈｅ）」、および「前記（ｓａｉｄ）」は、１つまたは複数の要素が存在することを意味するように意図している。用語「含む（ｃｏｍｐｒｉｓｉｎｇ）」、「含む（ｉｎｃｌｕｄｉｎｇ）」、および「有する（ｈａｖｉｎｇ）」は、包含することを意図しており、これらのリストアップされた要素以外に追加の要素が存在し得ることを意味している。
【００７７】
以上を考慮すれば、本発明のいくつかの目的が達成され、他の有利な結果が実現されることが理解されよう。
【００７８】
本発明の実施形態の範囲を逸脱することなく、以上の構成および方法において様々な変更を行うことができる。以上の説明に含まれ、添付図面に示されるすべての事項は、例示的であり、限定的な意味ではないことが意図されている。
【図面の簡単な説明】
【００７９】
【図１】本発明の実施形態を利用することができる例示のネットワーク環境を示すブロック図である。
【図２】本発明の実施形態を利用することができる他の例示のネットワーク環境を示すブロック図である。
【図３】本発明の実施形態を利用することができるさらに他の例示のネットワーク環境を示すブロック図である。
【図４】検索に関連して電子ドキュメントを評価するための、本発明の一実施形態によるプロセスフローを示す流れ図の一例である。
【図５】検索に関連して電子ドキュメントを評価するための、本発明の一実施形態によるプロセスフローを示す流れ図の一例である。
【図６】本発明の一実施形態によるコンピュータ読取り可能媒体の一例を示すブロック図である。
【図７】本発明の一実施形態によるコンピュータ読取り可能媒体の他の一例を示すブロック図である。
【図８】本発明の一実施形態を実装することができる適切なコンピューティングシステム環境の実施形態の一例を示すブロック図である。
【符号の説明】
【００８０】
１０２、１１０、１１２、２０２、２０８、３０２、３０４、３１６　サーバ
１３０　コンピュータ
１５４　ハードディスクドライブ
１５６　磁気ディスクドライブ
１６０　光ディスクドライブ
２１４　ユーザ

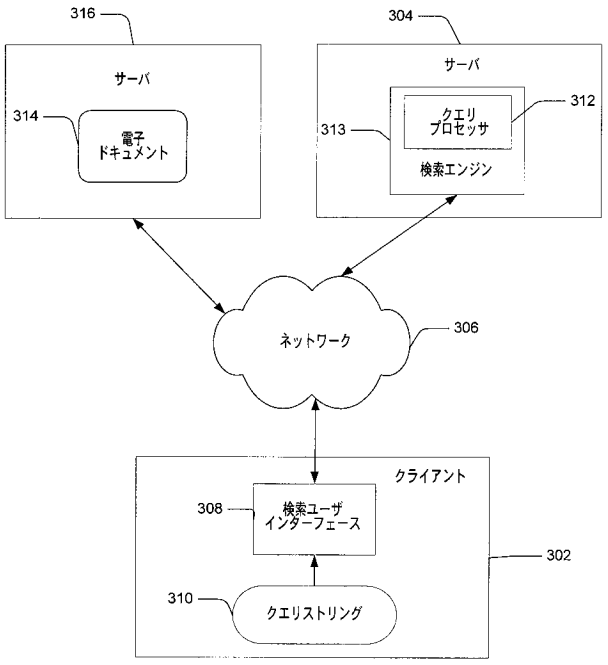
【図１】


110
サーバ
電子ドキュメント
電子ドキュメント
114
116
112
サーバ
電子ドキュメント
電子ドキュメント
118
120
128
データベース
ネットワーク
108
電子メールスパム検出システム
126
102
122
106
クローラ
129
インデックスビルダ
104
電子ドキュメント
電子ドキュメント
サーバ
124

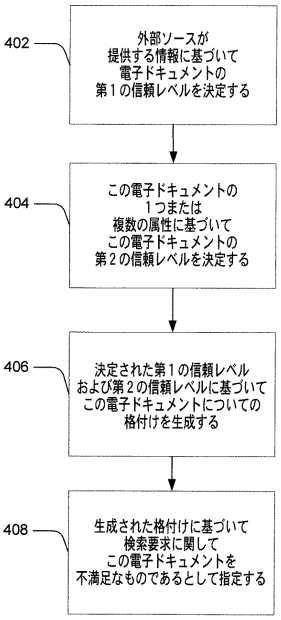
【図２】


214
208
サーバ
210
電子ドキュメント
電子ドキュメント
212
入力／フィードバック
ネットワーク
206
電子メールスパム検出システム
216
202
218
データベース
204
検索エンジン
サーバ

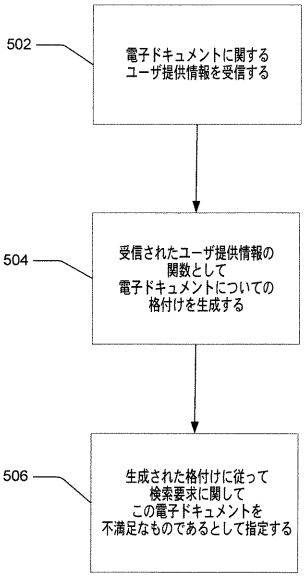
【図３】


316
サーバ
314
電子ドキュメント
304
サーバ
クエリプロセッサ
313
312
検索エンジン
ネットワーク
306
クライアント
308
検索ユーザインターフェース
302
310
クエリストリング

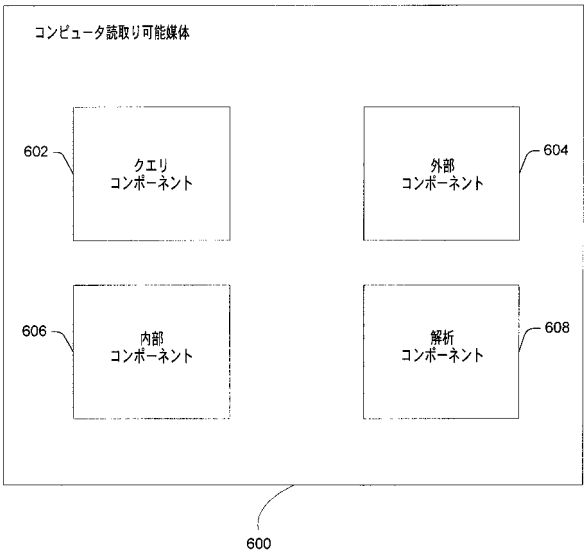
【図４】


402
外部ソースが提供する情報に基づいて電子ドキュメントの第１の信頼レベルを決定する
404
この電子ドキュメントの１つまたは複数の属性に基づいてこの電子ドキュメントの第２の信頼レベルを決定する
406
決定された第１の信頼レベルおよび第２の信頼レベルに基づいてこの電子ドキュメントについての格付けを生成する
408
生成された格付けに基づいて検索要求に関してこの電子ドキュメントを不満足なものであるとして指定する

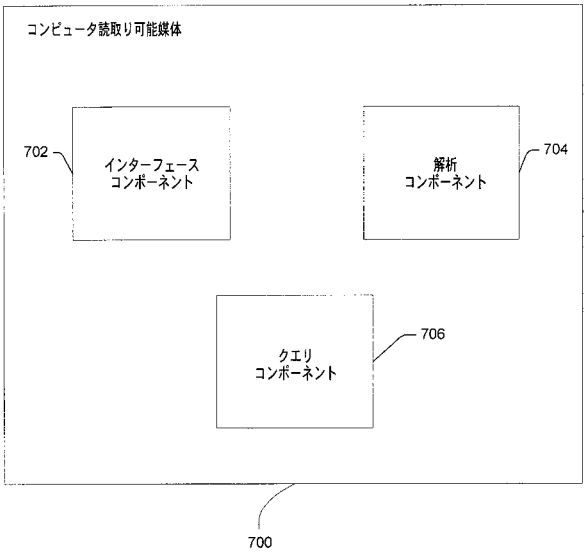
【図５】


502
電子ドキュメントに関する
ユーザ提供情報を受信する
504
受信されたユーザ提供情報の
関数として
電子ドキュメントについての
格付けを生成する
506
生成された格付けに従って
検索要求に関して
この電子ドキュメントを
不満足なものであるとして指定する

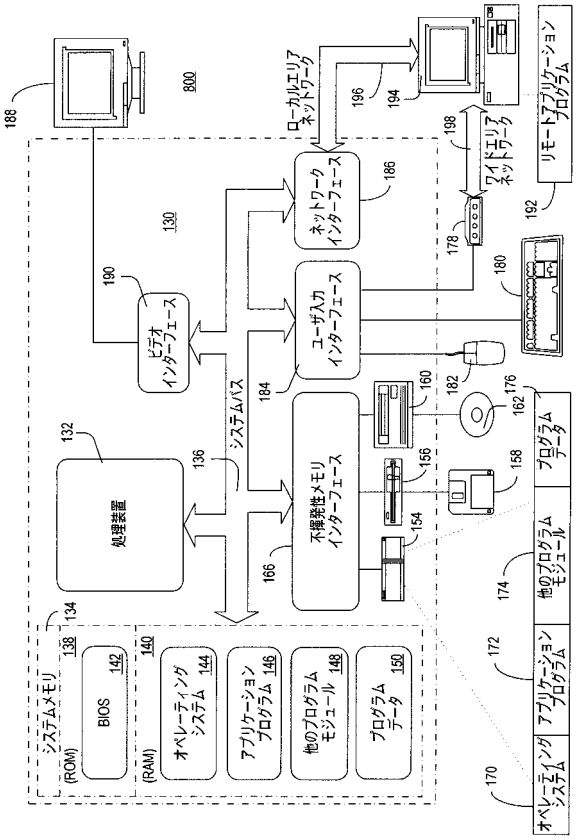
【図６】


コンピュータ読取り可能媒体
602
クエリ
コンポーネント
604
外部
コンポーネント
606
内部
コンポーネント
608
解析
コンポーネント
600

【図７】


コンピュータ読取り可能媒体
702
インターフェース
コンポーネント
704
解析
コンポーネント
706
クエリ
コンポーネント
700

【図８】


800
188
ローカルエリア
ネットワーク
196
194
198
ワイドエリア
ネットワーク
リモートアプリケーション
プログラム
192
130
ネットワーク
インターフェース
186
178
180
190
ビデオ
インターフェース
ユーザ入力
インターフェース
182
160
162
176
システムバス
184
132
処理装置
136
不揮発性メモリ
インターフェース
156
158
プログラム
データ
154
166
他のプログラム
モジュール
174
134
138
140
172
システムメモリ
(ROM)
BIOS
142
(RAM)
オペレーティング
システム
144
アプリケーション
プログラム
146
他のプログラム
モジュール
148
プログラム
データ
150
アプリケーション
プログラム
オペレーティング
システム
170

フロントページの続き

(72)発明者　エリック　ビー．ワトソン
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション内
(72)発明者　ジャニーン　ラス　クラム
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション内

審査官　長　由紀子

(56)参考文献　特表２００２－５３７７２７（ＪＰ，Ａ）
国際公開第２００４／０２５５１６（WO，Ａ１）
米国特許出願公開第２００２／０１９９０９５（ＵＳ，Ａ１）
米国特許出願公開第２００３／００３７０７４（ＵＳ，Ａ１）
米国特許出願公開第２００４／００９３３８４（ＵＳ，Ａ１）
福島　俊一，Ｗｅｂサーチエンジンの基本技術と最新動向（下），情報管理，日本，独立行政法人科学技術振興機構，２００３年１０月　１日，第46巻第7号，p.436-445

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０