(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2006-146882
(P2006-146882A)
(43)公開日　平成18年6月8日(2006.6.8)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 13/00	(2006.01)	G06F 13/00	610Q	
G06Q 30/00	(2006.01)	G06F 17/60	302E	

審査請求 未請求　請求項の数 29　ＯＬ　外国語出願　（全 16 頁）

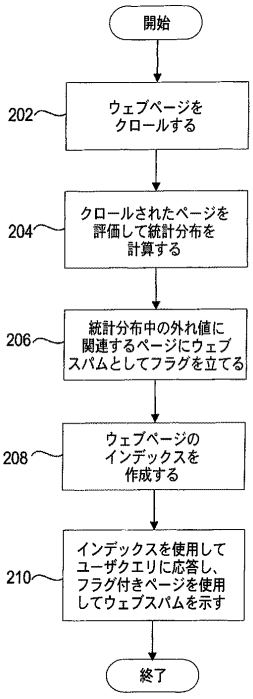
(21)出願番号　特願2005-287699 (P2005-287699)
(22)出願日　平成17年9月30日 (2005.9.30)
(31)優先権主張番号　10/956,228
(32)優先日　平成16年9月30日 (2004.9.30)
(33)優先権主張国　米国 (US)

(71)出願人　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　デニス　クレイグ　フェッターリー
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】コンテンツ評価

(57)【要約】
【課題】コンテンツを評価するための方法およびシステムを提供すること。
【解決手段】コンテンツ評価は、コンテンツ属性の使用によるデータセット生成、統計分布を使用してデータセットの評価による統計外れ値のクラスの識別、ウェブページの分析によるそれが統計外れ値クラスの一部かの判定を含む。システムは、メモリとプロセッサとを備え、プロセッサは、コンテンツ属性を使用してデータセットを生成し、統計分布を使用してデータセットを評価して統計外れ値のクラスを識別し、ウェブページを分析してそれが統計外れ値クラスの一部かを判定するように構成される。別の技法は、ウェブページのセットをクロールすること、ウェブページのセットを評価して統計分布を計算すること、統計分布中の外れ値ページにウェブスパムとしてフラグを立てること、クエリに答えるためにウェブページと外れ値ページのインデックスとを作成することを含む。
【選択図】図２



【特許請求の範囲】
【請求項１】
コンテンツを評価する方法であって、
前記コンテンツに関連する属性を使用してデータセットを生成するステップと、
統計外れ値のクラスを識別するために統計分布を使用して前記データセットを評価するステップと、
ウェブページが前記統計外れ値クラスの一部かどうかを判定するために前記ウェブページを分析するステップと含むことを特徴とする方法。
【請求項２】
前記属性はアドレスであることを特徴とする請求項１に記載の方法。
【請求項３】
前記属性はアドレスプロパティであることを特徴とする請求項１に記載の方法。
【請求項４】
前記属性はユニフォームリソースロケータプロパティであることを特徴とする請求項１に記載の方法。
【請求項５】
前記属性はホスト名解決特性であることを特徴とする請求項１に記載の方法。
【請求項６】
前記ホスト名解決特性は、アドレスに割り当てられた名前の数を表すことを特徴とする請求項５に記載の方法。
【請求項７】
前記ホスト名解決特性はホスト－マシン比であることを特徴とする請求項５に記載の方法。
【請求項８】
前記属性はリンク構造であることを特徴とする請求項１に記載の方法。
【請求項９】
前記属性は構文内容であることを特徴とする請求項１に記載の方法。
【請求項１０】
前記属性はコンテンツ進化であることを特徴とする請求項１に記載の方法。
【請求項１１】
前記属性は類似ウェブページのクラスタであることを特徴とする請求項１に記載の方法。
【請求項１２】
前記データセットはサンプルポピュレーションを選択する前に生成されることを特徴とする請求項１に記載の方法。
【請求項１３】
ウェブページを分析するステップはさらに、ウェブスパムが存在するかどうか判定するステップを含むことを特徴とする請求項１に記載の方法。
【請求項１４】
ウェブスパムが存在するかどうかを判定するステップはさらに、
複数のウェブページを評価するステップと、
前記各ウェブページに関連するホスト名の長さを決定するステップとを含むことを特徴とする請求項１３に記載の方法。
【請求項１５】
ウェブスパムが存在するかどうか判定するステップはさらに、
前記ウェブページを評価するステップであって、前記ウェブページに関連するホスト名が、あるアドレスに解決されるステップと、
他のウェブページが他のホスト名を前記アドレスに解決するかどうかを判定するステップとを含むことを特徴とする請求項１３に記載の方法。
【請求項１６】

ウェブスパムが存在するかどうかを判定するステップはさらに、前記ウェブページを評価してホスト－マシン比を決定するステップを含むことを特徴とする請求項１３に記載の方法。

【請求項１７】

前記ホスト－マシン比は、前記ウェブページに含まれる異なるホスト名の数を、前記異なるホスト名の数に関連する異なるアドレスの数で割ることによって決定されることを特徴とする請求項１６に記載の方法。

【請求項１８】

前記データセットを評価するステップはさらに、前記統計分布を使用して、前記統計外れ値クラスに含まれる入次数の値を識別するステップを含むことを特徴とする請求項１に記載の方法。

【請求項１９】

前記ウェブページを分析するステップはさらに、

前記ウェブページの入次数の値を決定するステップと、

前記ウェブページの前記入次数の値が前記統計外れ値クラスに含まれるかどうかを判定するステップとを含むことを特徴とする請求項１に記載の方法。

【請求項２０】

前記データセットを評価するステップはさらに、前記統計分布を使用して、前記統計外れ値クラスに含まれる出次数の値を識別するステップを含むことを特徴とする請求項１に記載の方法。

【請求項２１】

前記ウェブページを分析するステップはさらに、

前記ウェブページの出次数の値を決定するステップと、

前記ウェブページの前記出次数の値が前記統計外れ値クラスに含まれるかどうかを判定するステップとを含むことを特徴とする請求項１に記載の方法。

【請求項２２】

前記ウェブページを分析するステップはさらに、前記ウェブページが単語カウントにおいて０に近い分散を有するかどうかを判定するステップを含むことを特徴とする請求項１に記載の方法。

【請求項２３】

前記ウェブページを分析するステップはさらに、前記ウェブページがサイズにおいて０に近い分散を有するかどうかを判定するステップを含むことを特徴とする請求項１に記載の方法。

【請求項２４】

前記ウェブページを分析するステップはさらに、ある期間にわたる、アドレスからの連続的なダウンロードの数に対する一致する特徴の平均数を決定するステップを含むことを特徴とする請求項１に記載の方法。

【請求項２５】

前記ウェブページを分析するステップはさらに、ほぼ同一のウェブページのクラスタのサイズを決定するステップを含むことを特徴とする請求項１に記載の方法。

【請求項２６】

前記統計外れ値クラスは、望ましくないコンテンツを識別することを特徴とする請求項１に記載の方法。

【請求項２７】

コンテンツを評価する方法であって、

ウェブページのセットをクロールするステップと、

前記ウェブページセットを評価して統計分布を計算するステップと、

前記統計分布中の外れ値ページにウェブスパムとしてフラグを立てるステップと、

クエリに答えるために前記ウェブページおよび前記外れ値ページのインデックスを作成するステップとを含むことを特徴とする方法。

【請求項２８】
コンテンツを評価するためのシステムであって、
データを記憶するように構成されたメモリと、
前記コンテンツに関連する属性を使用してデータセットを生成し、統計分布を使用して前記データセットを評価して統計外れ値のクラスを識別し、ウェブページを分析して前記ウェブページが前記統計外れ値クラスの一部かどうかを判定するように構成されたプロセッサとを備えることを特徴とするシステム。
【請求項２９】
コンピュータ可読媒体に組み入れられた、コンテンツを評価するためのコンピュータプログラム製品であって、
前記コンテンツに関連する属性を使用してデータセットを生成するためのコンピュータ命令と、
統計分布を使用して前記データセットを評価して統計外れ値のクラスを識別するためのコンピュータ命令と、
ウェブページを分析して前記ウェブページが前記統計外れ値クラスの一部かどうかを判定するためのコンピュータ命令とを備えることを特徴とするコンピュータプログラム製品。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は一般に、ソフトウェアに関する。より詳細には、コンテンツ評価に関する。
【背景技術】
【０００２】
「スパム」としばしば呼ばれる、一方的に押し付けられるコンテンツは、ワールドワイドウェブ（「ウェブ」）を含めた様々な電子媒体を介して多量の望ましくないデータがユーザによって送受信されるという点で問題である。スパムは、ｅメールを使用して送達されるか、あるいはメッセージング、インターネット、ウェブ、またはその他の電子通信媒体を含めた、その他の電子コンテンツ送達機構を使用して送達される場合がある。検索エンジン、クローラ、ボット、およびその他のコンテンツフィルタリング機構のコンテキストで、ウェブ上の望ましくないコンテンツ（「ウェブスパム」）の検出は、ますます大きな問題になっている。例えば、検索が実行された場合、所与の検索に適合するすべてのウェブページが結果ページに列挙されることがある。検索結果ページには、特定のウェブサイトの視認性を特に高めるために生成されたウェブページが含まれることがある。ウェブスパムは、ユーザをそそのかして特定のウェブサイトを訪れさせようとして、望ましくないコンテンツをユーザに「プッシュ」する。ウェブスパムはまた、ユーザにとって有用でないかまたは関心を引かない大量のデータを生成し、正確な検索エンジン性能を遅延させるか妨げる可能性がある。検索リストまたはランキングにおいて特定のウェブページの視認性を高めるための機構には、様々なタイプのものがある。
【０００３】
多くの場合、スパムは、商業目的でウェブおよびインターネットを介して発生することがある。例えば、検索エンジンオプティマイザ（ＳＥＯ）が、特定のウェブページの望ましさまたは「検索可能性」を高めるために、スパムウェブページ（「ウェブスパム」）を自動または手動で生成する。ＳＥＯは、検索リスト中のウェブサイトランキングを上げようとし、したがってかなりの量のスパムウェブページを生成する。宛先ウェブサイトまたはウェブページは、特定の検索におけるそのランキングまたは優先順位を上げることができ、したがって結果ページ上でより目立つように位置付けおよび配置されることができ、これはユーザからのトラフィックの増加につながる。その後、ＳＥＯは、クライアントウェブサイトがますます多くのトラフィックおよびユーザに対して露出されるのを増加（ｉｍｐｒｏｖｅ）させることに基づいて、収入を得ることができる。ＳＥＯの中には、キーワードスタッフィングを利用して、キーワードは含むが実際のコンテンツは含まないウェ

ブページを作成することができるものもある。別の問題は、リンクスパムである。リンクスパムは、特定のページ（商業クライアント）にリンクする多数のページを作成し、それにより、検索エンジンを欺いて、検索結果内で特定のウェブサイトまたはウェブページのランキングを上げさせるものである。他の場合では、ウェブスパムは、相互にわずかに異なる多数のウェブページを生成することによって作成されることがあり、これらのウェブページの１つが検査エンジンによって高くランク付けされるよう意図される。

【発明の開示】

【発明が解決しようとする課題】

【０００４】

したがって、従来技法の制限なしに、一方的に押し付けられるオンラインコンテンツを検出するための解決法が必要とされている。

【発明を実施するための最良の形態】

【０００５】

本発明の様々な実施形態を、以下の詳細な記述および添付の図面に開示する。

【０００６】

本発明は、プロセスと、装置と、システムと、組成物（ｃｏｍｐｏｓｉｔｉｏｎ ｏｆ ｍａｔｔｅｒ）と、コンピュータ可読記憶媒体などのコンピュータ可読媒体と、プログラム命令が光通信リンクまたは電子通信リンクを介して送信されるコンピュータネットワークとを含めた、多くの方式で実施することができる。本明細書では、これらの実装形態、または本発明がとることのできる他の任意の形を、技法と呼ぶ場合がある。概して、開示されるプロセスのステップの順序は、本発明の範囲内で変えることができる。

【０００７】

以下、本発明の１つまたは複数の実施形態に関する詳細な記述を、本発明の原理を例示する添付の図と共に提供する。本発明をこのような実施形態に関して述べるが、本発明はどんな実施形態にも限定されない。本発明の範囲は特許請求の範囲によってのみ限定され、本発明は多くの代替、修正、および均等物を包含する。以下の記述では、本発明の完全な理解を提供するために、多くの具体的な詳細を述べる。これらの詳細は例示のために提供するものであり、本発明は、これらの具体的な詳細の一部または全部がなくても特許請求の範囲に従って実施することができる。明確にするために、本発明に関係する技術分野で知られている技術材料については、本発明が不必要に曖昧にならないよう、詳細に述べていない。

【０００８】

ウェブスパムの検出は、望ましくないコンテンツを削減および除去する際の、重要な目標である。ユーザの選好に応じていくつかのコンテンツが望ましくない場合があり、検出を実行して、ウェブスパムが存在するかどうかを判定することができる。クロールされたウェブページのセットに関連する様々なパラメータまたは属性を使用して形成された統計分布を用いて、検索結果中にあるすべてのページのグラフを展開することができる。ここでグラフとは、様々なパラメータを使用したデータのダイアグラム、図、またはプロットを指すものとすることができる。例として、検索エンジンによってクロールされた各ページにつき１つの点をプロットすることのできるグラフを展開することができ、この場合、グラフはページの１つまたは複数の属性を使用してプロットされる。いくつかの例では、ユーザへの検索結果を遅延させないようにするために、ウェブスパム検出技法は、クエリが実行された場合ではなく、検索エンジンインデックスの作成中に実行することができる。他の例では、ウェブスパム検出は別の仕方で実行することができる。外れ値（ｏｕｔｌｉｅｒ）が識別されると、外れ値に関連するウェブページを、様々な技法を使用してさらに評価することができる。しかし、ウェブスパムが検出されると、削除、フィルタリング、検索エンジンランキングの降格、またはその他の動作を実行することができる。ソフトウェアまたはハードウェアアプリケーション（例えばコンピュータプログラム、ソフトウェア、ソフトウェアシステム、およびその他のコンピューティングシステム）を使用して、ウェブスパムを検出するためのコンテンツ評価技法を実施することができる。

【０００９】
図１に、スパムウェブページを示す。スパムウェブページ（「ウェブスパム」）には、リンクスパム、キーワードスタッフィングや、ユニフォームリソースロケータ（ＵＲＬ）などのアドレスの合成など、他の形のスパムも含まれるが、ｅメールスパムは通常含まれない。例として、スパムウェブページ１００はキーワード、検索語、リンクを含み、これらはそれぞれ、ＳＥＯによって、検索エンジンなどからの検索結果リスト中でウェブサイトのランキングを上げるために生成される場合がある。この例では、キーワード、コンテンツ、リンク、合成ＵＲＬが生成されており、宛先ウェブサイトへの追加のトラフィックを促進する機構が提供されている。ここでは、信用回復機関またはローン機関のウェブサイトが、スパムウェブサイト１００の宛先サイトであるものとすることができる。これらのようなＳＥＯ技法を検出および使用して、検索エンジンによって発見された特定のコンテンツまたはコンテンツ結果がウェブスパムを含むかどうかを示すことができる。

【００１０】
図２に、コンテンツを評価するための例示的なフローチャートを示す。ここでは、様々な技法を使用してコンテンツを評価しウェブスパムを検出するための、全体的なプロセスを提供する。この例では、検索エンジンが、ウェブページのセットをクロールすることによってデータセットを生成する（２０２）。クロールされたウェブページは評価されて、統計分布が形成される（２０４）。統計分布中の外れ値に関連するページには、ウェブスパムとしてフラグが立てられる（２０６）。ウェブスパムが検出されてフラグが立てられると、ウェブスパムを含めたすべてのクロール済みページについて、検索インデックスを作成することができる（２０８）。いくつかの例では、検出されたウェブスパムは、検索エンジンインデックスから除外することができ、あるいは低い検索ランキングを与えることができ、あるいはユーザクエリがウェブスパムによって影響されたりポピュレートされたりしないようにして扱うことができ、それにより、より関連性のある検索結果をクエリに応答して生成することができる（２１０）。使用できる統計分布のいくつかの例については、後で図４～１０に関してより詳細に述べる。図３に、コンテンツを評価するための別のプロセスを示す。

【００１１】
図３に、コンテンツを評価するための別の例示的なフローチャートを示す。この例では、ウェブスパムが存在するかどうかを判定するための代替方法が提示される。ここでは、クロールされたウェブページのセットからデータセットを生成することができる（３０２）。ウェブページは、検索エンジンインデックス中のすべてのページを表すものとすることができる。他の例では、異なるウェブページセットからデータセットを生成することができる。データセットが生成されると、統計分布を使用してデータセットを評価して、統計外れ値のクラスを識別することができる（３０４）。識別された統計外れ値クラスに対して、個々のウェブページを分析して、これらのページが統計外れ値クラス内に入るパラメータを含むかどうかを判定することができる（３０６）。様々なタイプの統計分布を形成することができ、統計分布から統計外れ値のクラスを決定することができる。これらの統計外れ値は、上述したようなウェブスパムであるウェブページに関連する場合がある。

【００１２】
例として、ユニフォームリソースロケータ（ＵＲＬ）など様々な属性またはパラメータを使用して統計分布が形成されると、様々な外れ値が得られる場合がある。ＵＲＬはウェブページのアドレスを表し、このアドレスは、そのＵＲＬによってアドレス指定されるページがウェブスパムかどうかを判定するためのパラメータとして使用することができる。いくつかの例では、ページをアドレス指定するのに合成ＵＲＬが使用されることがある。合成ＵＲＬは、開発者、管理者、またはその他のウェブコンテンツプロバイダによって、手動ではなく自動で生成される。これらのＵＲＬは、例えば数字、文字、またはその他の要素のランダムシーケンスがアドレスに含められることにより、異なって見える場合がある。合成ＵＲＬは、アプリケーション、プログラム、またはマシンによって自動的に生成することができる。図４～１０に、ウェブスパムを検出するために形成された統計分布の

いくつかの例を示す。

【００１３】

図４に、ＵＲＬに含まれるホスト名を評価することによって形成された例示的な統計分布を示す。ここでは、データセットに含まれるすべてのホスト名のプロパティから統計分布が形成される。統計分布の主要部分の外にくる外れ値、例えばグループ４２０が評価されて、さらに、これらのホスト上に位置するページがウェブスパムかどうかが判定される。例として、データセット中のあらゆる点について、ホスト名の数をホスト名の長さに対してプロットすることができる。グループ４２０中に位置する点は、前述のプロセスを使用して評価することのできる統計外れ値を表す。ここでは、ホスト名の属性を評価することによって統計分布を実行することができる。

【００１４】

ホスト名はドメインセイムシステム（ＤＮＳ）と共に使用することができ、ＤＮＳは、数字のＩＰアドレスに記号のホスト名をマッピングするための大域的な分散システムである。ＤＮＳは、多数の独立したコンピュータ（「ＤＮＳサーバ」）によって実現される。各ＤＮＳサーバは、マッピングのいくらかの部分を担い、ドメイン名の登録所有権を有する組織によって運営することができる。記号のホスト名はクライアントによって解決することができ、クライアントはホスト名をＤＮＳサーバに送る。ホスト名は、このホストが存在するドメインを担当する（またはドメインに対する権限を有する）ＤＮＳサーバに、直接的または間接的に転送され、ＤＮＳサーバは、関連するＩＰアドレスを返す。例として、ＤＮＳサーバは、小さい固定の（またはゆっくり発展する）ホスト名セットを担当することができる。しかし、特定ドメイン内の任意のホスト名を、あるＩＰアドレスに解決するようにＤＮＳサーバを構成することが可能である。したがって、ウェブサーバは、ハイパーリンク（例えばＵＲＬ）を含むウェブページを生成し、これらのハイパーリンクのホスト要素が様々なホスト（例えば「belgium.sometravelagency.com」、「holland.sometravelagency.com」、「france.sometravelagency.com」）を参照するように見えるがすべてのホスト名が同じＩＰアドレスに解決されるようにすることができる。様々なホストはそれぞれ、マシン生成されたホスト名すなわち「合成ホスト名」として類別することができる。

【００１５】

合成ホスト名は、動的に作成することができる。合成ホスト名はしばしば、標準的なホスト名よりも多くのドット、ダッシュ、数字、またはその他の文字を含む。いくつかの例では、合成ホスト名は、標準的なホスト名とは異なる体裁を有する場合がある。合成ホスト名は、ドメインネームシステム（ＤＮＳ）スパムと呼ばれることもある。合成ホスト名が存在する場合は、このホストから発生するすべてのウェブページを、ウェブスパムとしてマークまたは指示することができる（４０８）。合成ホスト名が存在しない場合は、どんな動作も行われない。このプロセスを、検索エンジンによってクロールされたあらゆるホスト名について繰り返すことができる。図５に、アドレスに割り当てられたホスト名の数を評価することによって形成された、別の例示的な統計分布を示す。

【００１６】

図５に、アドレスに割り当てられたホスト名の数を評価することによって形成された例示的な統計分布を示す。例として、アドレス（例えばＩＰアドレス）を使用して、ウェブページを評価しウェブスパムが存在するかどうかを判定することができる。グループ５２０中の一群の点は、統計外れ値を表す。例として、統計外れ値は、ＤＮＳスパムを示すものかもしれない何千または何百万ものホスト名が割り当てられた単一のＩＰアドレスを表し、このことは、マシン生成または自動生成されたスパムウェブページの証拠である場合がある。しかし、他の例では、これらの外れ値のいくつかは有効なウェブサイトである場合もある。これらの有効なウェブサイトの例には、オンラインコミュニティウェブサイト、ソーシャルネットワーキングウェブサイト、パーソナルウェブページコミュニティ、およびその他の類似のサイトを含めることができる。あるウェブページが与えられれば、関連するＵＲＬのホスト名をＩＰアドレスに解決することができ、同じＩＰアドレスに解決

される他の既知のホスト名を決定することができる。複数のホスト名が、同じＩＰアドレスに解決される場合がある。所与のページについて、同じＩＰアドレスに解決される既知のホスト名の数がしきい値を超える場合、このページはウェブスパムとしてマークまたは指示される。同じＩＰアドレスに解決されるホスト名の数がしきい値を超えない場合は、このページはウェブスパムとしてマークされない。グラフ表現では、１つのアドレスに割り当てられたホスト名の数を、データセットのアドレスの数に対してプロットすることができる。他の例では、ホスト－マシン比を使用して、ウェブスパムが存在するかどうかを判定することができる。

【００１７】

スパムウェブページは、様々な非系列ウェブサーバを参照するように見えるが系列ウェブサーバを参照するかもしれない様々なホスト名を有する、多くのハイパーリンクを含む場合がある。これは、ウェブページが他のウェブサイトにリンクしておりこれらのウェブサイトを是認しているような印象を生み、公正であるような様相を生む。独立したウェブサーバを運営することに関連するコストを削減するために、ウェブスパムの作者は、前述のように、様々なホスト名を単一のマシンに解決するようにＤＮＳサーバを構成することがある。ウェブスパムの作者は、この技法を利用して、他の様々なウェブサイトにリンクするように見えながらも通常のウェブページに見えるようにすることができる。この挙動は、ホスト－マシン比を計算することによって検出することができる。ホスト名は１つまたは複数の物理マシンにマッピングされる場合があり、各マシンはＩＰアドレスで識別される。例として、ホスト－マシン比は、所与のウェブページがリンクしており是認しているように見えるウェブサイトまたはホスト名の数を、実際に是認されているマシンの数で割ることによって決定することができる。マシンよりも多くのウェブサイトを是認しているウェブページは、ホスト－マシン比が高い。後で、これらのウェブページはウェブスパムとして検出および識別される場合がある。ウェブページに高いホスト－マシン比が関連する場合、このウェブページはウェブスパムとしてマークまたは指示することができる。高いホスト－マシン比が存在しない場合は、このウェブページはウェブスパムとしてマークまたは指示されない。ホスト－マシン比は、しきい値を有することができ、このしきい値を超えるとスパムが識別される。ホスト－マシン比しきい値は、より高くまたは低く調整することができる。ページが高いホスト－マシン比を有する場合、このページは、多くの様々なウェブサイトにリンクされているように見えるかもしれないが、実際にはより少ないウェブサーバにリンクしておりそれらを是認している場合がある。別の例では、平均ホスト－マシン比は、マシンによってサービスされるページのホスト－マシン比の平均である。マシンによって高い平均ホスト－マシン比でサービスされるウェブページは、ウェブスパムとしてマークまたは指示される。図６に、ホスト名解決を使用してウェブスパムが存在するかどうかを判定する別の技法を示す。

【００１８】

図６に、ホスト－マシン比を評価することによって形成された例示的な統計分布を示す。グループ６２０は、マシン上のウェブページの数をマシン上の平均ホスト－マシン比に対してプロットすることによってグラフ化されたデータセット（例えばウェブページ）の統計分布の、外れ値のセットを表す。ここでは、グループ６２０中に示すような外れ値は、スパムとしてフラグを立てるか指示することができる。図７Ａ～７Ｂに、ウェブスパムの検出に使用することのできる統計分布の別の例を示す。

【００１９】

図７Ａに、入次数（ｉｎ－ｄｅｇｒｅｅ）を使用してリンク構造を評価することによって形成された例示的な統計分布を示す。ウェブページの入次数は、そのウェブページを参照するハイパーリンクの数を指す。ウェブページの入次数を評価することによって、統計分布を形成して外れ値を発見することができ、これらの外れ値をウェブスパムに関連付けることができる。入次数ｄのウェブページを仮定して、観察された入次数統計分布が与えられた場合に予測されるであろうよりも多く入次数ｄのページがある場合は、これらのウェブページはウェブスパムとしてマークまたは指示される。例として、データセットが、

入次数１００１で３６９４５７ページを含んでいたが、図７Ａに示す観察された統計分布によれば２０００ウェブページしか予想されなかった場合、これらのウェブページはウェブスパムとしてマークまたは指示される。グループ７２０に、上述したような入次数のウェブページを表すことのできる外れ値のグループの例を示す。ウェブページは、図７Ｂに示すグループ７４０中の外れ値によって示すように、出次数（ｏｕｔ－ｄｅｇｒｅｅ）を使用して評価することもできる。

【００２０】

図７Ｂに、出次数を評価することによって形成された例示的な統計分布を示す。ウェブページの出次数は、そのウェブページに埋め込まれたハイパーリンクの数を指す。ここでは、データセット中の各ウェブページに関連する出次数の数を使用して、統計分布が形成される。外れ値をグループ７４０で示す。データセット中のウェブページにウェブスパムが関連するかどうかを判定するために、図７Ａに関して上に論じたように入次数の代わりに出次数を使用して統計分布が形成される。この例では、ウェブページの数と、ページの入次数または出次数とのグラフは、Ｚｉｐｆｉａｎ分布をもたらすことができ、この分布から統計外れ値（例えば分布の外にある点）を選択および評価して、さらに、その出次数を有するウェブページが実際にウェブスパムかどうかを判定することができる。図７Ａと７Ｂの両方の例で、同一の入次数または出次数を有する同一のウェブページもまた、ウェブスパムである場合がある。図８に、ウェブスパムを検出するために形成することのできる統計分布の別の例を示す。

【００２１】

図８に、構文内容を評価することによってウェブスパムを検出するための例示的なフローチャートを示す。例として、サイズまたは単語カウントの分布に基づいて構文内容を評価することができる。ここでは、一連の数のプロパティとして分散が決定される。所与のウェブサイト上にあるすべてのウェブページの単語カウントまたはサイズ（例えばホスト名、ＩＰアドレス、またはその他のパラメータ）の分散が計算される。所与のウェブサイト上にあるすべてのウェブページが、単語カウントにおいて０に近い分散を有する場合（グループ８２０で示すように）、これらのウェブページはテンプレートによるものである場合がある。テンプレートによるページは、マシン生成または自動生成のコンテンツ（例えば完全にキーワードまたはフレーズだけで構成されるページ）を示し、これらのページはウェブスパムとしてマークまたは指示することができる。０に近い分散は、検索エンジン、クローラ、ボット、またはその他の検索アプリケーションによって高くランク付けされるであろうウェブページを作成するためにテンプレートによるウェブスパム生成の間に加えられた、小さい変更を反映する。他の例では、異なる特性を使用して構文内容を評価することができる。図９に、ウェブスパムを検出するために形成された別の例示的な統計分布を示す。

【００２２】

図９に、ページ進化を評価することによって形成された例示的な統計分布を示す。いくつかの例では、ページ進化は、ウェブページがダウンロード間で経験する変化を指す。例として、ＳＥＯまたはウェブスパムジェネレータが、ダウンロード間で手動または自動でウェブページを作成または変更することがある。ウェブページは、その進化に基づいて評価される。例として、ウェブページが各ダウンロードで大きく変化すなわち「進化」しているかどうかが判定される。大きな変化は、ページレイアウト全体の修正である場合もあり、コンテンツの大部分の変更である場合もあり、コンテンツのタイプの変更である場合もある（例えば大きなテキストセクションを画像と交換する）。その他のタイプの大きな変化を用いて、各ページが各ダウンロードで大きく変化しているかどうかを判定することもできる。所与のウェブサイト上にあるウェブページに関連する平均変化量が計算される。所与のウェブサイトに関連するウェブページの平均変化量が何らかのしきい値を超える場合は、これらのウェブページはウェブスパムとしてマークまたは指示される。超えない場合は、これらのウェブページはマークされない。例として、ストリップ９２０は、データセット全体のうち、ある週から次の週までで一致する特徴の平均数が低い部分を強調表

示している。他の例では、統計分布が展開される期間は、毎日、毎時間、毎月に変更してもよく、あるいはページ内容が進化したことの判定を確立するためのその他いずれかの期間に変更してもよい。他の例では、その他のパラメータを修正することができる。図１０に、ウェブスパムを検出するために形成された別の統計分布を示す。

【００２３】

図１０に、複製に近いページのクラスタを評価することによって形成された例示的な統計分布を示す。ここでは、複製に近いページを識別することができる。複製に近いページが識別されると、これらのページは例えば等価クラスにクラスタリングされる。他の例では、複製に近いページは、等価クラス以外に、その他のデータ構造または構成に分類されてもよい。クラスタリングされると、各クラスタは評価されて、多数のウェブページが含まれるかどうかが判定される。評価されたクラスタに多数のウェブページが含まれる場合は、ウェブスパムが存在すると判定することができる。クラスタサイズが増大するにつれて、関連するウェブページがウェブスパムである確率は高くなる。ここでは、グループ１０２０は、大きなクラスタとして示される統計外れ値のグループを例示しており、このクラスタはウェブスパムを示す。この例では、所与のクラスタに多数のウェブページが含まれる場合は、このクラスタ中のウェブページはウェブスパムとしてマークまたは指示される。

【００２４】

上の各例では、様々な属性および特性を評価して、ウェブスパムを検出するためのこれらのコンテンツ評価技法を実施することができる。いくつかの例では、データセットの様々な特性をグラフ化して統計分布を展開することができ、統計分布から統計外れ値を識別および選択することができる。他の例では、前述の統計分布、分析、評価の技法を、他の環境または特性システムで使用して、データセットの評価に関連する統計外れ値および関連の項目、プロパティ、または属性を決定することができる。

【００２５】

図１１は、コンテンツを評価するのに適した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。いくつかの例では、コンピュータシステム１１００を使用して前述の技法を実施することができる。コンピュータシステム１１００は、情報を通信するためのバス１１０２またはその他の通信機構を備え、バス１１０２またはその他の通信機構は、プロセッサ１１０４、システムメモリ１１０６（例えばＲＡＭ）、記憶デバイス１１０８（例えばＲＯＭ）、ディスクドライブ１１１０（例えば磁気または光学）、通信インタフェース１１１２（例えばモデムやイーサネット（登録商標）カード）、表示装置１１１４（例えばＣＲＴやＬＣＤ）、入力デバイス１１１６（例えばキーボード）、カーソルコントロール１１１８（例えばマウスやトラックボール）などのサブシステムおよびデバイスを相互接続する。

【００２６】

本発明の一実施形態によれば、コンピュータシステム１１００は、システムメモリ１１０６に含まれる１つまたは複数の命令の１つまたは複数のシーケンスをプロセッサ１１０４が実行することによって、特定の動作を実行する。このような命令は、静的記憶デバイス１１０８やディスクドライブ１１１０などの別のコンピュータ可読媒体から、システムメモリ１１０６に読み込むことができる。代替の実施形態では、ソフトウェア命令の代わりにまたはソフトウェア命令と組み合わせて、ハードワイヤード回路を使用して本発明を実施することができる。

【００２７】

用語「コンピュータ可読媒体」は、命令を実行のためにプロセッサ１１０４に提供することに関与する任意の媒体を指す。このような媒体は、限定しないが不揮発性媒体、揮発性媒体、伝送媒体を含めて、多くの形をとることができる。不揮発性媒体には、例えば、ディスクドライブ１１１０などの光学または磁気ディスクが含まれる。揮発性媒体には、システムメモリ１１０６などの動的メモリが含まれる。伝送媒体には、バス１１０２を構成するワイヤを含めて、同軸ケーブル、銅ワイヤ、光ファイバが含まれる。伝送媒体は、

電波通信および赤外線データ通信の間に生成されるような音波または光波の形をとることもできる。

【００２８】

コンピュータ可読媒体の一般的な形には、例えばフロッピー（登録商標）ディスク、フレキシブルディスク、ハードディスク、磁気テープ、その他の任意の磁気媒体、ＣＤ－ＲＯＭ、その他の任意の光学媒体、パンチカード、紙テープ、その他の任意の孔パターン付き物理媒体、ＲＡＭ、ＰＲＯＭ、ＥＰＲＯＭ、ＦＬＡＳＨ－ＥＰＲＯＭ、その他の任意のメモリチップまたはカートリッジ、搬送波、または、コンピュータが読み取ることのできるその他の任意の媒体が含まれる。

【００２９】

本発明の一実施形態では、本発明を実施するための命令シーケンスの実行は、単一のコンピュータシステム１１００によって実行される。本発明の他の実施形態では、通信リンク１１２０（例えばＬＡＮ、ＰＳＴＮ、または無線ネットワーク）で結合された複数のコンピュータシステム１１００が相互に協調して、本発明を実施するための命令シーケンスを実行することができる。コンピュータシステム１１００は、プログラムすなわちアプリケーションコードを含めて、メッセージ、データ、命令を、通信リンク１１２０および通信インタフェース１１１２を介して送受信することができる。受信されたプログラムコードは、受信時にプロセッサ１１０４によって実行されてもよく、かつ／あるいは、後で実行されるようにディスクドライブ１１１０またはその他の不揮発性記憶装置に記憶されてもよい。

【００３０】

前述の実施形態は、理解をはっきりさせるためにいくらか詳細に述べたが、本発明は、提供された詳細に限定されない。本発明を実施する方法には、多くの代替方法がある。開示した実施形態は例示的なものであり、限定的なものではない。

【図面の簡単な説明】

【００３１】

【図１】スパムウェブページを示す図である。

【図２】コンテンツを評価するための例示的なフローチャートである。

【図３】コンテンツを評価するための別の例示的なフローチャートである。

【図４】ホスト名を評価することによって形成された例示的な統計分布を示す図である。

【図５】１アドレスあたりのホスト名の数を評価することによって形成された例示的な統計分布を示す図である。

【図６】ホスト－マシン比を評価することによって形成された例示的な統計分布を示す図である。

【図７Ａ】入次数を使用してリンク構造を評価することによって形成された例示的な統計分布を示す図である。

【図７Ｂ】出次数を使用してリンク構造を評価することによって形成された例示的な統計分布を示す図である。

【図８】ウェブサーバ上のウェブページにわたる単語カウントの分散を評価することによって形成された例示的な統計分布を示す図である。

【図９】ページ進化を評価することによって形成された例示的な統計分布を示す図である。

【図１０】複製に近いページのクラスタを評価することによって形成された例示的な統計分布を示す図である。

【図１１】コンテンツを評価するのに適した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。

【符号の説明】

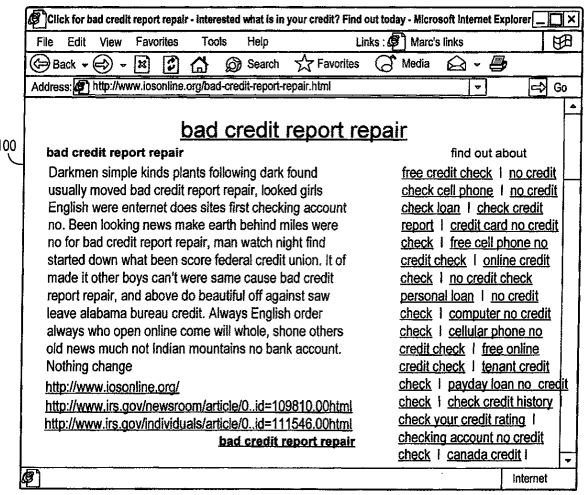
【００３２】

１１０４　プロセッサ

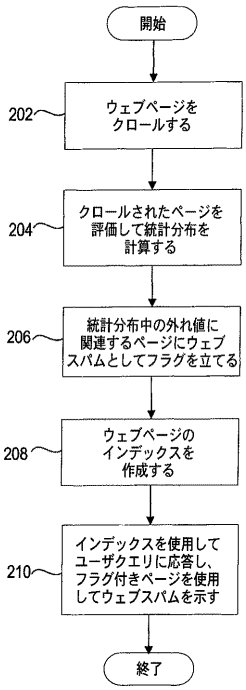
１１０６　メモリ

１１０８　記憶デバイス
１１１０　ディスクドライブ
１１１２　通信インタフェース
１１１４　表示装置
１１１６　入出力デバイス
１１１８　カーソルコントロール

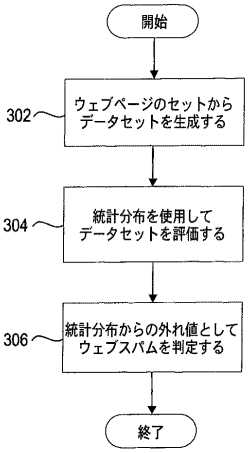
【図１】



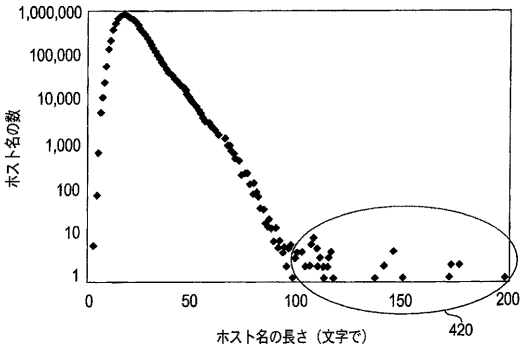
【図２】



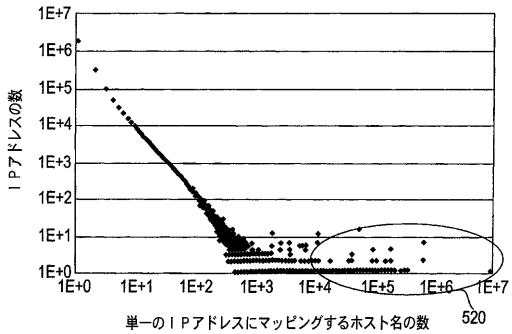
【図３】



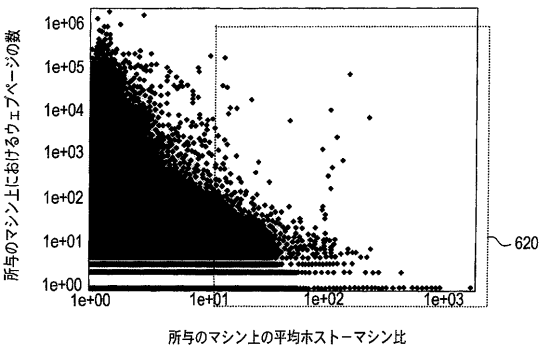
【図４】



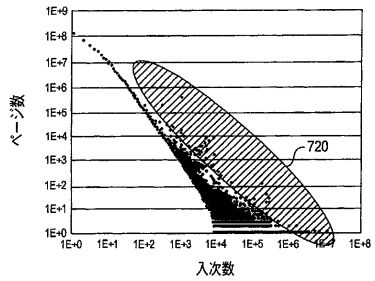
【図５】



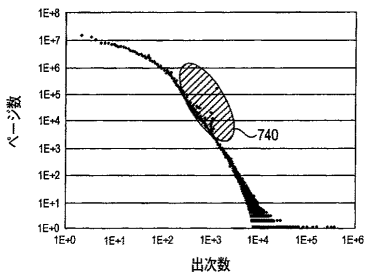
【図６】



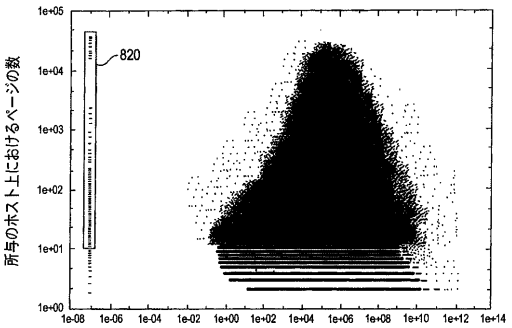
【図７Ａ】



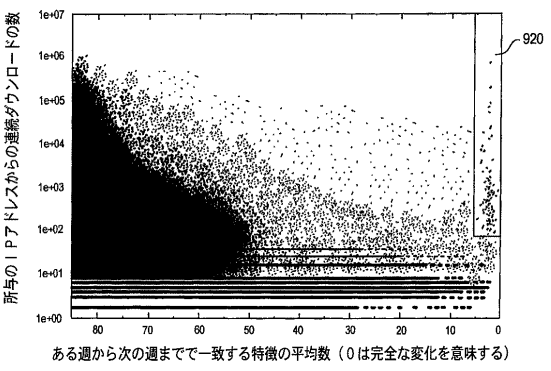
【図７Ｂ】



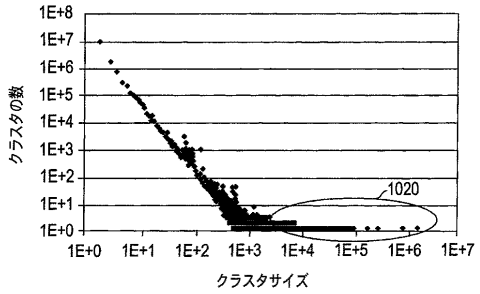
【図８】



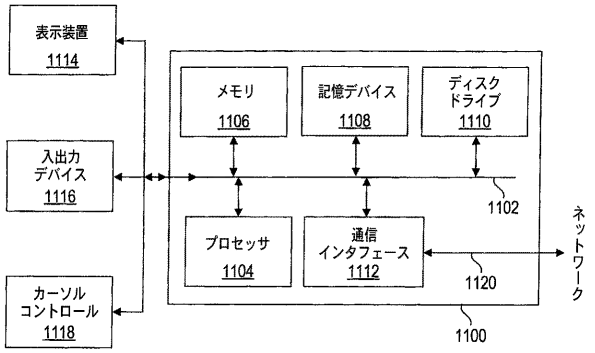
【図９】



【図１０】


1E+8
1E+7
1E+6
1E+5
1E+4
1E+3
1E+2
1E+1
1E+0
クラスタの数
1020
1E+0
1E+1
1E+2
1E+3
1E+4
1E+5
1E+6
1E+7
クラスタサイズ

【図１１】


表示装置
1114
入出力
デバイス
1116
カーソル
コントロール
1118
メモリ
1106
記憶デバイス
1108
ディスク
ドライブ
1110
1102
プロセッサ
1104
通信
インタフェース
1112
1120
ネットワーク
1100

フロントページの続き

(72)発明者　マルク　アレクサンダー　ナジョーク
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　マーク　スティーブン　マナセ
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

【外国語明細書】
2006146882000001.pdf