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

特許第4908422号
(P4908422)

(45)発行日　平成24年4月4日(2012.4.4)　(24)登録日　平成24年1月20日(2012.1.20)

(51)Int.Cl.　　F I
G06F 13/00 (2006.01)　　G06F 13/00 540E
G06F 17/30 (2006.01)　　G06F 17/30 350C
　　G06F 17/30 340A
　　G06F 17/30 180Z

請求項の数 8　(全 19 頁)

(21)出願番号　特願2007-539077 (P2007-539077)
(86)(22)出願日　平成17年10月26日 (2005.10.26)
(65)公表番号　特表2008-519328 (P2008-519328A)
(43)公表日　平成20年6月5日 (2008.6.5)
(86)国際出願番号　PCT/US2005/038619
(87)国際公開番号　WO2006/049996
(87)国際公開日　平成18年5月11日 (2006.5.11)
審査請求日　平成20年7月17日 (2008.7.17)
(31)優先権主張番号　60/623,295
(32)優先日　平成16年10月28日 (2004.10.28)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　11/198,471
(32)優先日　平成17年8月4日 (2005.8.4)
(33)優先権主張国　米国 (US)

(73)特許権者　501438485
ヤフー！　インコーポレイテッド
アメリカ合衆国　カリフォルニア州　９４０８９　サニーヴェイル　ファースト　アヴェニュー　７０１
(74)代理人　100082005
弁理士　熊倉　禎男
(74)代理人　100067013
弁理士　大塚　文昭
(74)代理人　100086771
弁理士　西島　孝喜
(74)代理人　100109070
弁理士　須田　洋之
(74)代理人　100109335
弁理士　上杉　浩

最終頁に続く

(54)【発明の名称】リンクベースのスパム検出

(57)【特許請求の範囲】
【請求項１】
検索結果セットにおいて検索ヒットをランク付けする、コンピュータによって実現される方法であって、
ユーザからクエリを受取るステップと、
前記クエリに関連するヒットのリストを生成するステップと、を備え、
前記ヒットのリストの各々のヒットは前記クエリに関連し、
前記方法はさらに、少なくとも１つのヒットにメトリックを関連付けるステップを備え、前記メトリックは、前記少なくとも１つのヒットの前記クエリとの関連性を人為的に高める前記少なくとも１つのヒットへのリンクを含むブーストドキュメントの数を表わし、前記メトリックは、前記少なくとも１つのヒットについての信頼値及びリンクベースポピュラリティ値に少なくとも部分的に基づくものであり、前記方法はさらに、
前記少なくとも１つのヒットについてリンクベースのポピュラリティ値の第１の尺度を形成するステップであって、前記第１の尺度は、前記少なくとも１つのヒットのリンクポピュラリティを表わすステップと、
前記少なくとも１つのヒットについて信頼値の第２の尺度を形成するステップであって、前記第２の尺度は、前記少なくとも１つのヒットが評判のよいドキュメントである可能性を示すステップと、
前記第１の尺度および前記第２の尺度を備える組合せに少なくとも部分的に基づいて前記メトリックを形成するステップであって、前記組合せは、前記第１の尺度と前記第２の

尺度との間の差を表わすステップと、
前記メトリックが閾値より大きいかどうかを判断するステップと、
修正されたリストを形成するために前記ヒットのリストを処理するステップであって、前記メトリックが前記閾値より大きいと判断されたことに応答して、前記少なくとも１つのヒットを前記修正されたリストから除外するか、前記ヒットのリストにおける前記少なくとも１つのヒットに起因していた関連性よりも関連性が低い状態で前記修正されたリストの中の前記少なくとも１つのヒットを提示するかの１つを実行するステップと、
前記修正されたリストをユーザに伝送するステップと、を備える、方法。
【請求項２】
前記メトリックを前記形成するステップは、クエリを前記受取るステップの前に実行される、請求項１に記載の方法。
【請求項３】
前記第２の尺度を前記形成するステップは、
評判のよいドキュメントのシードセットを形成するステップを備え、評判のよいドキュメントの前記シードセットは他のドキュメントへのリンクを備え、前記第２の尺度を前記形成するステップはさらに、
前記シードセットにおけるドキュメントの各々に信頼値を割当てるステップと、
比例配分された信頼値を、前記シードセットにおけるドキュメントのうちの少なくとも１つによって指し示される複数のドキュメントの各々に割当てるステップとを備える、請求項１に記載の方法。
【請求項４】
シードセットを前記形成するステップは、
複数のドキュメントの各々ごとに、複数のドキュメントの各々に含まれるアウトリンクの数を表わすアウトリンクメトリックをそれぞれ求めるステップと、
アウトリンクメトリックを使用して複数のドキュメントをランク付けするステップと、
複数のドキュメントにおいて最も高くランク付けされたドキュメントのセットを識別するステップと、
最も高くランク付けされたドキュメントのセットの各々の質を評価するステップと、
最も高くランク付けされたドキュメントのセットから不適切であると考えられるドキュメントを除去することによって、最も高くランク付けされたドキュメントの修正されたセットを形成するステップと、
最も高くランク付けされたドキュメントの修正されたセットを使用してシードセットを形成するステップとを備える、請求項３に記載の方法。
【請求項５】
検索結果セットにおいて検索ヒットをランク付けするための命令を格納する、コンピュータによって実現されるコンピュータ可読記憶媒体であって、命令は、
ユーザからクエリを受取るステップと、
前記クエリに関連するヒットのリストを生成するステップとを実行するための命令を含み、
前記ヒットのリストの各々のヒットは前記クエリに関連し、
前記命令はさらに、少なくとも１つのヒットにメトリックを関連付けるステップを実行するための命令を含み、前記メトリックは、前記少なくとも１つのヒットの前記クエリとの関連性を人為的に高める前記少なくとも１つのヒットへのリンクを含むブーストドキュメントの数を表わし、前記メトリックは、前記少なくとも１つのヒットについての信頼値及びリンクベースポピュラリティ値に少なくとも部分的に基づくものであり、前記命令はさらに、
前記少なくとも１つのヒットについてリンクベースのポピュラリティ値の第１の尺度を形成するステップであって、前記第１の尺度は、前記少なくとも１つのヒットのリンクポピュラリティを表わすステップと、
前記少なくとも１つのヒットについて信頼度の第２の尺度を形成するステップであって

、前記第２の尺度は、前記少なくとも１つのヒットが評判のよいドキュメントである可能性を示すステップと、

前記第１の尺度および前記第２の尺度を備える組合せに少なくとも部分的に基づいて前記メトリックを形成するステップであって、前記組合せは、前記第１の尺度と前記第２の尺度との間の差を表わすステップと、

前記メトリックが閾値より大きいかどうかを判断するステップと、

修正されたリストを形成するために前記ヒットのリストを処理するステップであって、前記メトリックが前記閾値より大きいと判断されたことに応答して、前記少なくとも１つのヒットを前記修正されたリストから除外するか、前記ヒットのリストにおける前記少なくとも１つのヒットに起因していた関連性よりも関連性が低い状態で前記修正されたリストの中の前記少なくとも１つのヒットを提示するかの１つを実行するための命令を含み、前記命令はさらに、

前記修正されたリストをユーザに伝送するステップを実行するための命令を含む、コンピュータ可読記憶媒体。

【請求項６】

前記メトリックを前記形成するステップは、クエリを前記受取るステップの前に実行される、請求項５に記載のコンピュータ可読記憶媒体。

【請求項７】

前記第２の尺度を前記形成するステップは、

評判のよいドキュメントのシードセットを形成するステップを備え、評判のよいドキュメントの前記シードセットは他のドキュメントへのリンクを備え、前記第２の尺度を前記形成するステップはさらに、

前記シードセットにおけるドキュメントの各々に信頼値を割当てるステップと、

比例配分された信頼値を、前記シードセットにおけるドキュメントのうちの少なくとも１つによって指し示される複数のドキュメントの各々に割当てるステップとを備える、請求項５に記載のコンピュータ可読記憶媒体。

【請求項８】

シードセットを前記形成するステップは、

複数のドキュメントの各々ごとに、複数のドキュメントの各々に含まれるアウトリンクの数を表わすアウトリンクメトリックをそれぞれ求めるステップと、

アウトリンクメトリックを使用して複数のドキュメントをランク付けするステップと、

複数のドキュメントにおいて最も高くランク付けされたドキュメントのセットを識別するステップと、

最も高くランク付けされたドキュメントのセットの各々の質を評価するステップと、

最も高くランク付けされたドキュメントのセットから不適切であると考えられるドキュメントを除去することによって、最も高くランク付けされたドキュメントの修正されたセットを形成するステップと、

最も高くランク付けされたドキュメントの修正されたセットを使用してシードセットを形成するステップとを備える、請求項７に記載のコンピュータ可読記憶媒体。

【発明の詳細な説明】

【技術分野】

【０００１】

発明の背景

この発明は概して検索システムに関し、より詳細には、結果セットにおいて検索ヒットをランク付けする検索システムに関する。

【背景技術】

【０００２】

コーパス全体が吸収されることができず、所望の項目への厳密なポインタが存在しないまたは可能でない場合には、検索が有用である。概して、検索は、検索クエリを公式化または受入れ、ドキュメントのコーパスから一致するドキュメントのセットを求め、そのセ

ットまたはそのセットが大き過ぎる場合にはそのセットの何らかのサブセットを返すプロセスである。具体的な例において、この開示はその例に限定されないのだが、「ウェブ」と称されるハイパーリンクされたドキュメントのセットを検索することを考慮されたい。コーパスには、本明細書ではページと称され、またはより総称的にドキュメントと称される多くの検索可能な項目が入っている。検索エンジンは、典型的には検索クエリの受取に先立って生成される索引を使用して、検索クエリに一致するドキュメントをコーパスから識別する。「一致」とは多くのことを意味する可能性があり、検索クエリはさまざまな形態を有し得る。一般に、検索クエリは１つ以上の語または用語を含む文字列であり、ドキュメントが検索クエリ文字列からの語または用語のうち１つ以上（または、それらのすべて）を含むときに一致が発生する。各々の一致するドキュメントはヒットと称され、ヒットのセットは結果セットまたは検索結果と称される。コーパスは、データベースまたは他のデータ構造または非構造化データであり得る。ドキュメントはしばしばウェブページである。

【０００３】

ウェブページの典型的な索引は何十億ものエントリを含むため、一般的な検索は何百万ものページを含む結果セットを有するかもしれない。明らかに、このような状況では、クエリを行なう人（典型的には人間のコンピュータユーザであるが、そうである必要はない）に返されるものの大きさが妥当なものであるようにするために、検索エンジンは結果セットをさらに制約しなければならないかもしれない。セットを制約する１つのアプローチは、順序付けられた検索結果の中でより高いところに現われる少数のヒットのみをユーザは読むまたは使用すると仮定して、ある順序で検索結果を提示することである。

【０００４】

この仮定のために、多くのウェブページの作者は、順序付けられた検索結果の中で自分たちのページが高いところに現われることを望む。検索エンジンは、最高品質のページのみを選択し、返すために、関連するページのさまざまな特徴に依拠する。クエリ結果リストにおけるトップの位置（上位）がビジネス上の優位を与え得るので、あるウェブページの作者は、自分たちのページのランキングを故意にブーストしようとする。ランキングが人為的にブーストされたこのようなページは「ウェブスパム」ページと呼ばれ、総称して「ウェブスパム」として公知である。

【発明の開示】

【発明が解決しようとする課題】

【０００５】

ウェブスパムに関連付けられるさまざまな技術が存在する。１つの技術は、ウェブページを多くのクエリによって選択されるのにふさわしいものに人為的にするというものである。これは、本質的なコンテンツに無関係であり、小さなまたは目に見えないフォントで表現される莫大な数の用語でページを増大させることによって達成されることができる。このような増大によって、ページはより露出されるようになる（すなわち、より多くのクエリに関連する可能性がある）が、任意の特定のクエリについてその関連性を真に向上させることはない。この点に関して、スパムの作者は別の技術を使用する。その別の技術とは、他者によってより頻繁に参照されるページが概して検索エンジンによって好ましい（より高い関連性を有する）と考えられるという観察に基づいて、インリンク（inlink）とも呼ばれる多くの入ってくる（ハイパー）リンクをページに付加するというものである。価値が優れているために多くの他者によって参照される真に高品質のページと、多くのインリンクを有するウェブスパムとを区別することは困難である。

【０００６】

ウェブスパムページを識別することおよび検索結果リストにおいてウェブスパムページをその後格下げすることは、検索エンジンによってもたらされる回答の質を維持または向上させるために重要である。したがって、ウェブスパムの検出は検索エンジンにとって有用なタスクである。検索エンジンの索引に存在する多数のページを検証することによってウェブスパムを識別するためにヒューマンエディタ（human editor）がしばしば利用され

るが、それは実用的でないことが多い。
【０００７】
したがって、ウェブスパムを克服し、ドキュメントの作者の操作に従うのではなく、ユーザが欲するものにより従う検索結果を提供する改良された検索処理が必要である。
【課題を解決するための手段】
【０００８】
発明の簡単な概要
この発明の実施例は、検索結果セットを形成するヒットをランク付けすることを含む、検索要求を処理するためのシステムおよび方法を提供する。ヒットは、特定のページを指し示すスパムファーム（spam farm）の大きさの尺度である有効質量（effective mass）および他のパラメータを使用してランク付けされることができる。
【０００９】
一実施例では、この発明は、検索結果セットにおいて検索ヒットをランク付けする、コンピュータによって実現される方法を提供する。コンピュータによって実現される方法は、ユーザからクエリを受取ることと、クエリに関連するヒットのリストを生成することとを含み、ヒットの各々はクエリに関連し、ヒットは、ヒットを指し示す１つ以上のブーストリンクされたドキュメントを有し、ブーストリンクされたドキュメントは、ヒットの、クエリとの関連性に影響を及ぼす。上記方法は次いで、ヒットの少なくともサブセットの各々ごとにメトリック（metric）を関連付け、メトリックは、ヒットの少なくともサブセットの各々を指し示しかつヒットの関連性を人為的に膨らませる、ブーストリンクされたドキュメントの数を表わす。上記方法は次いで、ヒットを指し示すスパムファームの大きさを表わすメトリックを閾値と比較し、一部上記比較に基づいて、修正されたリストを形成するためにヒットのリストを処理し、修正されたリストをユーザに伝送する。
【００１０】
一局面では、メトリックは第１の尺度と第２の尺度との組合せである。ヒットについての第１の尺度はヒットのリンクポピュラリティを表わし、第２の尺度はヒットが評判のよいドキュメントである可能性の尺度である。
【００１１】
別の局面では、第２の尺度は評判のよいドキュメントのシードセット（seed set）を形成することによって生成され、評判のよいドキュメントのシードセットはリンクを張るドキュメントであり、上記第２の尺度はさらに、シードセットにおけるドキュメントの各々に信頼値を割当てることと、リンクを張るドキュメントによって指し示されるリンクされるドキュメントの各々にその信頼値を伝播させることと、比例配分された信頼値を、リンクされるドキュメントの各々に割当てることとによって生成される。
【００１２】
別の局面では、評判のよいドキュメントのシードセットは、複数のドキュメントの各々ごとに、ドキュメントの各々のアウトリンク（outlink）の数を表わすアウトリンクメトリックを求めることと、そのアウトリンクメトリックを使用して複数のドキュメントをランク付けすることと、最も高くランク付けされたドキュメントのセットを識別することと、最も高くランク付けされたドキュメントの質を評価することと、最も高くランク付けされたドキュメントから不適切であると考えられるドキュメントを除去することによってドキュメントの修正されたセットを形成することと、修正された保持されるセットを使用してシードセットを形成することとによって形成される。
【００１３】
添付の図面とともに以下の詳細な説明によって、この発明の性質および利点はよりよく理解されることになる。
【発明を実施するための最良の形態】
【００１４】
発明の詳細な説明
定義

別の方法で定義されない限り、本明細書において使用されるすべての技術的および科学的用語は、この発明が関係する当業者によって一般に理解されている意味を有する。本明細書において使用されるように、以下の用語は下記のとおりに定義される。

【００１５】

ページランク（PageRank）とは、検索エンジンによって索引付けされるハイパーリンクされたドキュメント（またはウェブページまたはウェブサイト）に数値で重み付けするための一群の周知のアルゴリズムである。ページランクはリンク情報を使用して、ウェブ上のドキュメントにグローバル重要性スコアを割当てる。ページランクプロセスは特許を受けており、米国特許第６，２８５，９９９号に記載される。ドキュメントのページランクは、ウェブ上のドキュメントのリンクベースのポピュラリティの尺度である。

【００１６】

トラストランク（TrustRank）とは、ページランクに関連するリンク分析技術である。トラストランクは、ウェブ上の評判のよい、優れたページをウェブスパムから分離するための方法である。トラストランクは、ウェブ上の優れたドキュメントが滅多にスパムにリンクしないという推測に基づいている。トラストランクは２つのステップを伴い、そのうちの１つはシード選択であり、別のステップはスコア伝播である。ドキュメントのトラストランクは、ドキュメントが評判のよい（すなわち、スパムのない）ドキュメントである可能性の尺度である。

【００１７】

リンクまたはハイパーリンクとは、別のページ、別のサイトまたは同一ページの別の部分に通常つながるウェブページ上のクリック可能なコンテンツを指す。したがって、クリック可能なコンテンツは、同一ページの他のページ／サイト／部分にリンクしていると言われている。スパイダは、ウェブサイトを索引付けするときに、リンクを使用して１つのページから次のページにゆっくり進む。

【００１８】

インバウンドリンクまたはインリンク／アウトバウンドリンクまたはアウトリンク。サイトＡがサイトＢにリンクするとき、サイトＡはアウトバウンドリンクを有し、サイトＢはインバウンドリンクを有する。インバウンドリンクは、リンクポピュラリティを求めるために数えられる。

【００１９】

ウェブまたはワールドワイドウェブ（「WWW」または単に「ウェブ」）とは、リソースと称される対象の項目が定型資源識別子（Uniform Resource Identifiers）（ＵＲＩ）と呼ばれるグローバル識別子によって識別される情報空間である。ウェブという用語はしばしばインターネットの同義語として使用されるが、ウェブは実際にはインターネット上で動作するサービスである。

【００２０】

ウェブページとは、通常ＨＴＭＬ／ＸＨＴＭＬ形式であり（ファイルの拡張子は典型的にはhtmまたはhtmlであり）、あるページまたはセクションから別のページまたはセクションへのナビゲーションを可能にするためにハイパーテキストリンクを有するワールドワイドウェブのページまたはファイルを指す。ウェブページはしばしば関連付けられるグラフィックスファイルを使用してイラストをもたらし、これらもクリック可能なリンクであり得る。ウェブページは、ウェブブラウザを使用して表示され、多くの場合モーション、グラフィックス、対話および音声をもたらすアプレット（ページ内で実行するサブプログラム）を利用するように設計されることができる。

【００２１】

ウェブサイトとは、単一のフォルダにまたはウェブサーバの関連するサブフォルダ内に格納されたウェブページの集まりを指す。ウェブサイトは概して、典型的にはindex.htmまたはindex.htmlと名付けられるトップページを含む。

【００２２】

ウェブホストは、独自のウェブサーバを持たない個人または企業によって管理されるウ

ェブサイトにサーバ空間、ウェブサービスおよびファイルメンテナンスを提供することに従事する。多くのインターネットサービスプロバイダ（Internet Service Provider）（ＩＳＰ）は、個人のウェブページのホストとして機能するように少量のサーバ空間を加入者に与えることになる。

【００２３】

スパムとは、大量に配信される、通常営利的な性質を有する不必要なドキュメントまたはｅメールを指す。

【００２４】

ウェブスパムとは、ウェブ上のスパムページを指す。ウェブスパムを作成する行為は、ウェブスパミングと称される。ウェブスパミングとは、値するいくつかのドキュメントにより高いランキングを与えるために検索エンジンを惑わせるように意図される行動を指す。ウェブ上のスパムページは、スパミングの何らかの形態の結果である。スパミングの一形態は、リンクスパミングである。

【００２５】

スパムページとは、ランキングスコアの大幅な違法のブーストを受け、したがって、検索結果が上位に現われる可能性を高くし、検索エンジンを惑わせるように意図されるウェブドキュメントである。

【００２６】

リンクスパミングとは、しばしば相互接続されてスパムファームと呼ばれるグループを形成するスパムドキュメントの作成を指し、スパムファームは、多数のブーストドキュメントが１つまたはいくつかのターゲットページのリンクベースの重要性ランキングを上げるであろうように構築される。

【００２７】

スパムファームとは、特定のターゲットページのリンクベースの重要性スコア（たとえば、ページランクスコア）をブーストするために作成される、相互にリンクされたスパムページのグループを指す。

【００２８】

概要

この発明の実施例は、リンクベースのスパムの検出のための方法およびシステムに向けられている。検索クエリに応答してもたらされる検索結果は、ヒットの有効質量を求めるために処理される。ヒットの有効質量は、ヒットを指し示すために作成され、したがって、ヒットの相対的な重要性を人為的にブーストするスパムファームの大きさの尺度である。この発明の実施例に従う方法およびシステムは、ヒットの有効質量を使用し、有効質量がリンクベースのスパムによってヒットを人為的にブーストする可能性を高くするそのヒットを格下げする。所与のウェブドキュメントについての有効質量を求めることは、所与のウェブドキュメントのリンクベースのポピュラリティ（たとえば、ページランク）と、所与のウェブドキュメントの信頼性（たとえば、トラストランク）との間の相違を部分的に査定する技術の組合せに依拠する。所与のウェブドキュメントの有効質量を求めるための技術について、以下でさらに詳細に説明する。

【００２９】

ネットワーク実現例

図１は、この発明の実施例を実施するために使用され得る１つ以上のクライアントシステム２０_{1-N}を含む情報抽出および通信ネットワーク１０の一般的な概要を示す。コンピュータネットワーク１０では、クライアントシステム２０_{1-N}は、インターネット４０または他の通信ネットワークを介して（たとえば、任意のローカルエリアネットワーク（local area network）（ＬＡＮ）または広域ネットワーク（wide area network）（ＷＡＮ）接続によって）任意の数のサーバシステム５０_{1}から５０_{N}に結合される。本明細書において記載されるように、クライアントシステム２０_{1-N}は、たとえば媒体コンテンツおよびウェブページなどの他の情報にアクセスし、それらを受取、抽出および表示するためにサーバシステム５０_{1}から５０_{N}のいずれかと通信するようこの発明に従って構成される。

【００３０】
図１に示されるシステムにおけるいくつかの要素は、ここで詳細に説明される必要がない従来型の周知の要素を含む。たとえば、クライアントシステム２０は、デスクトップパーソナルコンピュータ、ワークステーション、ラップトップ、パーソナルデジタルアシスタント（personal digital assistant）（ＰＤＡ）、携帯電話、または任意のＷＡＰ対応装置もしくはインターネットに直接的もしくは間接的に接続できる他の計算装置を含み得るであろう。クライアントシステム２０は典型的には、マイクロソフトのインターネットエクスプローラ（登録商標）ブラウザ、ネットスケープナビゲータ（登録商標）ブラウザ、モジラ（登録商標）ブラウザ、オペラ（登録商標）ブラウザ、アップルのサファリ（登録商標）、または携帯電話、ＰＤＡもしくは他の無線装置の場合のＷＡＰ対応ブラウザなどのブラウジングプログラムを実行し、クライアントシステム２０_{1-N}のユーザは、利用可能な情報およびページをインターネット４０によってサーバシステム５０_{1}から５０_{N}からアクセス、処理および閲覧できる。クライアントシステム２０は典型的には、ページ、形態、およびサーバシステム５０_{1}から５０_{N}または他のサーバによって与えられる他の情報とともに、ブラウザによってディスプレイ（たとえば、モニタスクリーン、ＬＣＤディスプレイなど）にもたらされるグラフィカルユーザインターフェイス（graphical user interface）（ＧＵＩ）と対話するための、キーボード、マウス、タッチスクリーン、ペンなどの１つ以上のユーザインターフェイス装置２２も含む。この発明はインターネットとともに使用するのに好適であり、インターネットとはネットワークの具体的なグローバルに関連するセットを指す。しかしながら、インターネットの代わりにまたはインターネットに加えて、イントラネット、エクストラネット、バーチャルプライベートネットワーク（virtual private network）（ＶＰＮ）、ＴＣＰ／ＩＰベースでないネットワーク、任意のＬＡＮまたはＷＡＮなどの他のネットワークが使用され得ることが理解されるべきである。

【００３１】
一実施例によれば、クライアントシステム２０およびその構成要素のすべては、インテルペンティアム（登録商標）プロセッサ、ＡＭＤアスロン（AMD Athlon）（登録商標）プロセッサ、アップルのパワーＰＣ（Power PC）などの中央演算処理装置または複数のプロセッサを使用して実行されるコンピュータソフトウェアを含むアプリケーションを使用して構成可能な演算子である。本明細書に記載されるデータおよび媒体コンテンツを通信、処理および表示するようにクライアントシステム２０を動作させ、構成するためのコンピュータソフトウェアは好ましくはハードディスクにダウンロードおよび格納されるが、全プログラムコードまたはその一部は、ＲＯＭもしくはＲＡＭなどの周知の他の揮発性または不揮発性メモリ媒体または装置にも格納される場合もあれば、コンパクトディスク（compact disk）（ＣＤ）媒体、デジタル多用途ディスク（digital versatile disk）（ＤＶＤ）媒体、フロッピー（登録商標）ディスクなどの、プログラムコードを格納できる任意の媒体に与えられる場合もある。さらに、全プログラムコードまたはその一部は、ソフトウェアソースから、たとえばサーバシステム５０_{1}から５０_{N}のうちの１つからクライアントシステム２０にインターネットによって伝送およびダウンロードされる場合もあれば、任意の通信媒体およびプロトコル（たとえば、ＴＣＰ／ＩＰ、ＨＴＴＰ、ＨＴＴＰＳ、イーサネット（登録商標）または他の従来型の媒体およびプロトコル）を使用して他のネットワーク接続（たとえば、エクストラネット、ＶＰＮ、ＬＡＮまたは他の従来型のネットワーク）によって伝送される場合もある。

【００３２】
この発明の局面を実現するためのコンピュータコードはＣ、Ｃ＋＋、ＨＴＭＬ、ＸＭＬ、Java（登録商標）、Java（登録商標）Scriptなどのコード、または他の好適なスクリプト言語（たとえば、VBScript）、またはクライアントシステム２０で実行されることができるかまたはクライアントシステム２０_{1-N}もしくはシステム２０_{1-N}で実行するようにコンパイルされることができる他の好適なプログラミング言語であり得ることが認識されるべきである。いくつかの実施例では、クライアントシステム２０にダウンロードされるコー

ドはなく、必要なコードはサーバによって実行され、またはクライアントシステム２０に既に存在するコードが実行される。

【００３３】

検索システム

図２は、この発明の実施例による媒体コンテンツを通信するための別の情報抽出および通信ネットワーク１１０を示す。示されるように、ネットワーク１１０は、クライアントシステム１２０と、１つ以上のコンテンツサーバシステム１５０と、検索サーバシステム１６０とを含む。ネットワーク１１０では、クライアントシステム１２０は、インターネット１４０または他の通信ネットワークを介して、サーバシステム１５０および１６０に通信可能に結合される。上述のように、クライアントシステム１２０およびその構成要素は、インターネット１４０または他の通信ネットワークによって、サーバシステム１５０および１６０ならびに他のサーバシステムと通信するよう構成される。

【００３４】

１．クライアントシステム

一実施例によれば、クライアントシステム１２０で実行する（モジュール１２５として表わされる）クライアントアプリケーションは、サーバシステム１５０および１６０と通信するため、ならびにそこから受取られたデータコンテンツを処理および表示するためにクライアントシステム１２０およびその構成要素を制御するための命令を含む。クライアントアプリケーション１２５は好ましくは、リモートサーバシステム（たとえば、サーバシステム１５０、サーバシステム１６０または他のリモートサーバシステム）などのソフトウェアソースからクライアントシステム１２０に伝送およびダウンロードされるが、クライアントアプリケーションモジュール１２５は、上述のように、フロッピー（登録商標）ディスク、ＣＤ、ＤＶＤなどの任意のソフトウェア記憶媒体に設けられることができる。たとえば、一局面では、クライアントアプリケーションモジュール１２５は、データを操作するためならびにさまざまなオブジェクト、フレームおよびウインドウでデータを表現するための、たとえば組込型Java（登録商標）ScriptまたはActiveX制御装置などのさまざまな制御装置を含むＨＴＭＬラッパーの状態でインターネット１４０によってクライアントシステム１２０に与えられてもよい。

【００３５】

さらに、クライアントアプリケーションモジュール１２５は、検索要求および検索結果データを処理するための検索モジュール１２６、テキストおよびデータフレームおよびアクティブウインドウ、たとえばブラウザウインドウおよびダイアログボックスの状態でデータおよび媒体コンテンツを表現するためのユーザインターフェイスモジュール１２７、ならびにクライアント１２０で実行するさまざまなアプリケーションと接続および通信するためのアプリケーションインターフェイスモジュール１２８などの、データおよび媒体コンテンツを処理するためのさまざまなソフトウェアモジュールを含む。アプリケーションインターフェイスモジュール１２８が好ましくは接続するよう構成される、クライアントシステム１２０で実行するさまざまなアプリケーションの例は、さまざまなｅメールアプリケーション、インスタントメッセージ（instant messaging）（ＩＭ）アプリケーション、ブラウザアプリケーション、ドキュメント管理アプリケーションなどを含む。さらに、インターフェイスモジュール１２７は、クライアントシステム１２０に構成されるデフォルトブラウザまたは異なるブラウザなどのブラウザを含んでもよい。

【００３６】

２．検索サーバシステム

一実施例によれば、検索サーバシステム１６０は、検索結果データおよび媒体コンテンツをクライアントシステム１２０に与えるよう構成される。コンテンツサーバシステム１５０は、たとえば検索サーバシステム１６０によって与えられる検索結果ページにおいて選択されたリンクに応答して、ウェブページなどのデータおよび媒体コンテンツをクライアントシステム１２０に与えるよう構成される。いくつかの変形例では、検索サーバシステム１６０は、コンテンツへのリンクおよび／または他の参照と同様に、またはその代わ

りに、コンテンツを返す。

【００３７】

一実施例における検索サーバシステム１６０は、たとえばページ、ページへのリンク、索引付けされたページのコンテンツを表わすデータなどが実装されたさまざまなページ索引１７０を参照する。ページ索引は、自動ウェブクローラ、スパイダなどを含むさまざまな収集技術、ならびに階層構造内でウェブページを分類およびランク付けするための手動または半自動分類アルゴリズムおよびインターフェイスによって生成されてもよい。これらの技術は、検索サーバシステム１６０で実現されてもよく、またはページ索引１７０を生成し、ページ索引１７０を検索サーバシステム１６０にとって利用可能なものにする別個のシステム（図示せず）において実現されてもよい。

【００３８】

検索サーバシステム１６０は、検索モジュール１２６からなど、クライアントシステムから受取られるさまざまな検索要求に応答するデータを提供するよう構成される。たとえば、検索サーバシステム１６０は、（たとえば、クエリにおける検索用語の発生のパターンによって測定される論理的な関連性の組合せ、コンテキスト識別子、ページスポンサーなどに基づいて）所与のクエリに対してウェブページを処理およびランク付けするための検索関連アルゴリズムで構成されてもよい。

【００３９】

リンクベースのスパム検出

図２に示されるように、検索サーバシステム１６０は、ウェブスパムページが格下げされるかまたはリストから除去された場合に修正された検索リストを返すリンクベースのスパム検出器１８０と組合せられて機能し、検索サーバシステム１６０の出力（結果、提案、媒体コンテンツなど）をリンクベースのスパム検出器１８０に与える。検索サーバシステム１６０は、この発明の実施例による検索エンジンを動作させるよう構成される。検索エンジンは、３つの部分、すなわち１つ以上のスパイダ１６２と、データベース１６３と、ツール／アプリケーション１６７とから成る。情報を集めるスパイダ１６２はインターネット中をゆっくり進み、データベース１６３にはスパイダが集めた情報および他の情報が入っており、ツール／アプリケーション１６７は、データベースを検索するためにユーザによって使用される検索ツール１６６などのアプリケーションを含む。データベース１６３には、検索ツールによって使用されるページ索引１７０が入っている。さらに、この発明の実施例による検索エンジンは、スパム検出器１８０を含む。スパム検出器１８０は、以下に記載されるさまざまなアルゴリズムを実行し、ページについてのウェブスパムメトリック１８１をページ索引１７０に格納する。上で説明されたように、この発明の実施例によるスパム検出器１８０は、ヒットの有効質量に一致するメトリックを推定し、検索ツール１６６およびページ索引１７０と組合せられて機能し、有効質量がリンクベースのスパムによってヒットを人為的にブーストする可能性を高くするそのヒットを格下げする。所与のウェブドキュメントについての有効質量を求めることは、所与のウェブドキュメントのリンクベースのポピュラリティ（たとえば、ページランク）と、所与のウェブドキュメントの信頼性（たとえば、トラストランク）との間の相違を部分的に査定する技術の組合せに依拠する。一実施例では、ウェブスパム検出器１８０は、ページ索引１７０におけるすべてのページを処理して、索引におけるページについてウェブスパムメトリック１８１を計算し、ウェブスパムメトリック１８１をデータベース１６３に格納する。メトリック１８１は、ドキュメントを検索結果に含ませる検索クエリから独立している。

【００４０】

所与のウェブドキュメントについてのスパムファームの有効質量をスパム検出器１８０によって求めることは、所与のウェブドキュメントのリンクベースのポピュラリティ（たとえば、ページランク）と、所与のウェブドキュメントの信頼性（たとえば、トラストランク）との間の差を見積ることに一部依拠する。所与のウェブドキュメントの信頼性を求めることは、信頼性があることが分かっているウェブドキュメント（すなわち、スパムのないドキュメント）の最初のシードセットから所与のページがどれくらい離れているかに

一部依拠する。したがって、この発明の実施例による検索エンジンは、信頼されるウェブドキュメントの最初のシードセット１８５を形成するためにページ索引１７０と組合せられて機能するシードセット生成器１８４も含む。ウェブスパムメトリック１８１を形成するスパム検出器１８０の動作、およびシードセット１８５を形成するシードセット生成器１８４の動作について、以下でさらに詳細に説明する。

【００４１】

スパムファーム、ページランクおよびトラストランク

このセクションでは、スパムファーム、（一般に「ページランク」と称される）インリンクページランキング、および信頼ランキングの概念について説明する。スパムファームとは、重要性をブーストするためにスパムターゲットページを指し示すページの人為的に作成されたセットである。信頼ランキング（「トラストランク」）とは、高品質のページのサブセットへの特別なテレポーテーション（teleportation）（すなわち、ジャンプ）を有するページランクの一形態である。本明細書に記載される技術を使用して、検索エンジンは自動的に不良なページ（ウェブスパムページ）を見つけることができ、より具体的には、人為的なスパムファーム（参照ページの集まり）を作成することによって、重要性をブーストするために作成されたウェブスパムページを見つけることができる。具体的な実施例では、均一なテレポーテーションを有するページランクプロセスおよび信頼ランキングプロセスが実行され、それらの結果は、ページまたはページの集まりの「スパム性」の試験の一部として比較される。さらに、信頼ランキングプロセスへの入力を構築する新規の方法について以下で説明する。

【００４２】

この発明の一局面は、スパムページを取巻くハイパーリンク構造の分析に基づいてスパムページ（の少なくともいくつか）を識別することに向けられている。特に、スパムファームの大きさを見積る新規のプロセスが使用される。スパムのないページは滅多にスパムを指し示すことはないので、トラストランクにおける特定オーソリティ分配は結果的に、スパムのないページとスパムページとの間をある程度分離する。高品質のスパムのないウェブページは、最高スコアをトラストランクによって割当てさせることが見込まれる。

【００４３】

トラストランクは、各ウェブページを指し示す他のページのスコアに応じて各ウェブページに数値スコアを割当てる周知のウェブ分析アルゴリズム、すなわちページランクに関連する。ページランクはテレポーテーションと呼ばれる技術を使用する。通常は均一な分配であるいわゆるテレポーテーション分配に従って、総スコアのうち特定の量がいくつかのページまたはすべてのページに送出される。均一なテレポーテーション分配を使用する代わりに、信頼ランキングは、信頼される（スパムのない）ウェブページのほんの小さなセット（すなわち、いわゆる「シードセット」）へのテレポーテーションを与える。これは、事実上、シードセットからのみ他のページにスコアを分配することになる。

【００４４】

以下の説明はウェブページを参照する。しかしながら、推論、実現例およびアルゴリズムは、（１）サイトのウェブ（ウェブコンテンツ／ページおよび単一のオーソリティに関連付けられる他のタイプのウェブドキュメントの論理的なグループ）、（２）ホスト間のグラフエッジ（graph edge）（たとえば、２つのホストがハイパーリンクによって接続される少なくとも１つのページを各々含む場合に２つのホストがリンクを有するホストグラフ、または他の試験）の何らかの定義を有する、ホストのウェブ（ホストランク）（Host Rank）によって表わされるサイトのウェブに近いもの、（３）他のウェブページのグラフの集約、および／または（４）参照の強さを反映する、関連付けられる重みを有するリンクの集まりに同様に適用可能である。

【００４５】

スパムファーム

スパムファームとは、重要性をブーストするためにスパムターゲットページを指し示すページの人為的に作成されたセット（または代替的には、ホスト）である。図３Ａ－図３

Ｂは、２つの単純なスパムファームを示す例示的な図である。

【００４６】

【数１】

図３Ａは、すべてがターゲットスパムページsを指し示すmページをスパムファームが有することを示している。スパムファームの大きさの優れた推定値を得ることを可能にするプロセスについて以下で説明する。ページiごとに数値M_iが計算され、ここで、数値M_iはページの「有効質量」と称される。ウェブスパムページでは、Mは、そのページをブーストしているスパムファームの大きさの優れた推定値の役割を果たす。

単純なスパムファームの場合、有効質量はmに近づく。たとえば図３Ｂに示されるスパムファームのようにより複雑なファームでは、有効質量Mはインジケータの役割を果たし、ここで、高いMの値はスパムファームであることを示す。この説明はウェブページについて言及するが、この概念はページ、ホストなどのグループにも適用され得ることが認識されるべきである。

【００４７】

ページランクおよびトラストランク

ページランクの概念は、ウェブページの分析に有用である。ページランクについての多くの可能な定義の中で、ページランキングの以下の線形システム定義が使用される。

【００４８】

【数２】

$$x = cT^T x + (1-c)v. \qquad (式\ 1)$$

【００４９】

【数３】

式１では、

Tは遷移行列であり、その要素は、ページiからページjを指し示すリンク$i \to j$が存在する場合には$T_{ij} = 1/outdeg(i)$であり、そうでなければ０である。ここで、$outdeg(i)$は、行列Tを確率的にするために標準化係数の役割を果たす、ページi上のアウトリンクの数である。

cはテレポーテーション定数であり、通常は０．７－０．９の範囲で選択される。

$x = (x_i)$はオーソリティベクトルであり、ここで、索引iはn個のページすべての範囲を動き、$i = 1:n$である（nはウェブページの数である）。

$v = (v_i)$はテレポーテーションベクトルであり、確率分布であると仮定され、$0 \le v_i \le 1, v_1 + \ldots + v_n = 1$である。

【００５０】

式１を解くための反復法が公知である。式１は、テレポーテーションベクトルに対して線形であるオーソリティベクトルを規定するという利点を有する。

【００５１】

【数４】

ページランクでは、pは、均一なテレポーテーションに対応して（すなわち、$v_i = 1/n$であるときに）式１の解をもたらすであろうオーソリティベクトルである。トラストランクでは、tは、特別なテレポーテーションに対応して（すなわち、vは、vのk要素が０でなく、残りが０であるようなものであり、０でない要素が、信頼されるセットの中に対応する索引iを有する場合に）式１の解をもたらすであろうオーソリティベクトルである。

【００５２】

有効質量の推定

ウェブページの有効質量は、インジケータとして使用されて、ウェブページがスパムページであるかどうかを判断するのを助ける。

【００５３】

推定値の構築

【００５４】

【数５】

潜在的なスパムページsに関して、任意のウェブページiの中で、以下のように数学的に示されることができる。

【００５５】

【数６】

$$p_s - t_s = p_s^{boost} + b \cdot p_s^{leak} + (1-c)/n, \qquad (式\ 2)$$

【００５６】

【数７】

ここで、等式の右側の第１項は、サポートするスパムファームからページに来るブーストに起因し（スパムのないページの場合、ファームは空であるかまたは存在せず）、第２項は、スパムページを時々誤って指し示すスパムのないページからのオーソリティリークに起因する。このリークは、残余のウェブから所与のページへの異なる偶発的なハイパーリンクを表わす破線の矢印として図３Ａ－図３Ｂに示されている。スパムページsでは、第１項が非常に優勢である。なぜなら、スパムファームを作成するスパム送信者の動機はsのページランクを高くすることであるためである。単純なファームでは以下のようになる。

【００５７】

【数８】

$$p_s^{boost} = \frac{m \cdot c(1-c)}{n}, \qquad (式\ 3)$$

【００５８】

同様の公式は、他の構造のファームについて有効である。たとえば、バックリンクを有するファームでは、

【００５９】

【数９】

$$p_s^{boost} = \frac{m \cdot c(1-c)}{(1-c^2)n}, \qquad (式\ 4)$$

【００６０】
である。
【００６１】
【数１０】

$$p_s^{leak} << p_s^{boost} \qquad (式\ 5)$$

【００６２】
【数１１】

という条件下で、単純なスパムファームの大きさmについての優れた推定値は、下記のとおりに式（２）および（３）から構築される。

【００６３】
【数１２】

$$M_s = \frac{n(p_s - t_s)}{c(1-c)}, \qquad (式\ 6)$$

【００６４】
【数１３】

式６は、任意のウェブページiについて計算され得る有効質量M_iを規定する。上述のように、iが単純なスパムファームによってブーストされたスパムページである場合、M_iは実際のファームの大きさmに近づき、他の構造のファームについては、式４によって示されたように、M_iは実際のファームの大きさから定数だけ異なるにすぎない。このような差は、実際のスパムファームが相当に大きい（たとえば、何百万ものブーストページが不正に作成される）という事実を考慮すると、重大ではない。

スパムのないページでは、M_iは、絶対項においてまたはp_iに対してはあまり大きくない何らかの数値になる。この発明の実施例によるリンクベースのスパム検出はこれを発見し、潜在的なウェブスパムページのようなページをM_iに基づいてインジケータとして指定することはない。

【００６５】
スパム検出プロセス
【００６６】

【数１４】

以下の例示的なプロセスは、リンクベースのスパムを検出するために使用される。このプロセスは、最も高い有効質量を有するページを見つけることを目的としている点で、素晴らしく単純で効果的である。しかしながら、式５が満たされる場合に限り、有効質量はスパムの大きさに非常に近くなり、信頼されるウェブページからポピュラリティを割当てることに起因するページのリンクベースのポピュラリティが、スパムページによる人為的なブーストに起因するページのリンクベースのポピュラリティよりも確実にはるかに小さくなるようにする。式５の条件下で、スパム検出プロセスは、合法的に人気のあるページと、リンクを張るスパムファームによって人気があるようにされたページとを区別できる。この発明の実施例による技術は、確実に式５の条件が満たされるようにする。これは、以下のステップＣにおいて実行され、ステップＣでは、$\eta > 1$は閾値の役割を果たすアルゴリズムパラメータである。Ｃにおいて比率が大きいことは、ページが式５を満たしていることに対応することを示し得る。概して、例示的なプロセスは以下を含む。

Ａ．リスト（たとえば、クエリに関連するヒットのリストまたはページ索引）におけるすべてのページ（ホストなど）iに関して、式（６）に従ってその有効質量M_iを見つける。

Ｂ．M_iの大きい順にページiをソートし、ソートされたリストのトップ部分を保持または識別する。代替的に、全リストが保存されてもよいが、必要なリソースが多すぎるかもしれず、したがって低いM_iのページを保持しないことがより効率的である。この識別および／または保持は、いずれのステップでなされてもよい。選択プロセスの一部は、M_iおよびM_i / p_iの両方が高いページを選択することに向けられている。

Ｃ．リストに保持されるすべてのページiについて比率M_i / p_iを見つける。

Ｄ．$M_i / p_i < \eta$を有するページiをリストから削除する。

Ｅ．保持されるページはスパムを構成する。

【００６７】

実験では、そのように検出されたスパムページは実際にはほとんどの場合に（人間の判断によって）スパムであることが確認された。これは、これらの技術を使用して偽陽性率が低くなる可能性があることを意味する。

【００６８】

シードセット

【００６９】

【数１５】

上述のプロセスは、いわゆるシードセットに関連付けられる特別なテレポーテーション分配を有するトラストランク、すなわち式（１）の解に依拠する。シードセットとは、スパムがないことが分かっているkの高品質のウェブページのセットである。この発明の実施例の局面は、信頼できる（すなわち、スパムのない）ページまたはサイトの適切なシードセットを見つけることに向けられている。信頼されるウェブページのシードセットを識別する一方法は、ヒューマンエディタの判断に基づいて特定のウェブページを指定することである。しかしながら、人間の評価は費用および時間を要する。実行可能な選択肢としてシードセットを手動で選択するオプションを保持するが、半自動的にシードセットを構築する別の技術について以下で説明する。

【００７０】

シード選択プロセスは、シードページが２つの重要な特徴を有するはずであるという観察に依拠する。２つの重要な特徴とはすなわち、１）多数の他のページは、シードページ

から始まり、遭遇したウェブページ上のアウトリンクを反復して辿って、到達可能であるはずであり、すなわち、シードページは高い適用範囲をもたらすはずであること、および２）シードページの品質は非常に高いはずであり、そのため、スパムのないページからスパムページへのリンクに遭遇するチャンスは最小限のはずであることである。

【００７１】

第１の特徴を確保するために、すべてのページ（すなわち、ページ索引におけるページ）のランキングがもたらされる。このために、式７によって示される以下の線形システムが使用される。

【００７２】

【数１６】

$$y = cU^T y + (1-c)v, \qquad (式7)$$

【００７３】

【数１７】

このシステムでは、

・Uは逆遷移行列であり、その要素は、リンク$j \to i$が存在する場合には$U_{ij} = 1/indeg(i)$であり、そうでなければ０である。ここで、$indeg(i)$は、行列Uを確率的にするために標準化係数の役割を果たす、ページiへのインリンクの数である。

・cはテレポーテーション定数であり、通常は０．７－０．９の範囲で選択される。

・$y = (y_i)$はオーソリティベクトルであり、ここで、索引iはn個のページすべての範囲を動き、$i = 1:n$である。

・$v = (v_i)$はテレポーテーションベクトルであり、確率分布であると仮定され、$0 \le v_i \le 1, v_1 + \ldots + v_n = 1$である。

式７が通常の遷移行列Tの代わりに逆遷移行列Uを使用する以外は、式７によって記載されたシステムは式１のシステムと類似していることに注目されたい。逆遷移行列は、リンクの方向性を逆にしたウェブグラフに対応する。これに関して、均一なテレポーテーションを有する式７の解yは逆ページランク（Inverse PageRank）と称される。逆ページランクは、ページ上のアウトリンクを辿ることによってそのページからどれだけのウェブに到達できるかについての尺度である。

【００７４】

シードページの第２の特徴を確保するために、最も高い逆ページランクを有するページはさらにヒューマンエディタによって処理される。ヒューマンエディタは、どの候補（逆ページランクによって測定されるように、高い適用範囲をもたらすページ）が実際に高品質のスパムのないページであるかを選択する。ヒューマンエディタによって選択されたページは次いで、上述のように、シードセットの中に含まれ、トラストランク計算において使用される。

【００７５】

【数１８】

例示的なシードセット構築プロセスは下記のとおりに要約される。
Ａ．すべてのページ（ホストなど）iに関して、式（７）に従ってその逆ページランクy_iを見つける。
Ｂ．y_iの大きい順にページiをソートし、ソートされたリストのトップを保持する。またはそうでなければ、最も高くランク付けされたページのセットを識別および保持する。
Ｃ．ヒューマンエディタを使用して、リストに保持されるページの質を評価する。
Ｄ．エディタによって不適切であると考えられたページをリストから削除する。
Ｅ．保持されるページはシードセットを構成する。

【００７６】
結果として生じるシードセットは、ページランクおよびトラストランクから導き出される質量推定に基づくトラストランク計算ならびにスパム検出に好適であることを実験結果は示してきた。
【００７７】
本明細書に記載される実施例は、ウェブサイト、リンク、およびワールドワイドウェブ（またはそのサブセット）が検索コーパスとして機能する場合に特有の他の専門用語について言及してもよい。本明細書に記載されるシステムおよびプロセスは（電子データベースまたはドキュメント収納庫などの）異なる検索コーパスとともに使用するために適合されることができ、結果はコンテンツおよびコンテンツが見つけられ得る場所へのリンクまたは参照を含み得ることが理解されるべきである。
【００７８】
このように、この発明は具体的な実施例に関して記載されてきたが、この発明は特許請求の範囲内にすべての修正例および等価物を包含するように意図されることが認識されるであろう。
【図面の簡単な説明】
【００７９】
【図１】この発明の実施例を実施するために使用され得る情報抽出および通信ネットワークの例示的なブロック図である。
【図２】この発明の実施例による情報抽出および通信ネットワークの例示的なブロック図である。
【図３Ａ】単純なスパムファームの例示的な図である。
【図３Ｂ】単純なスパムファームの例示的な図である。

【図１】

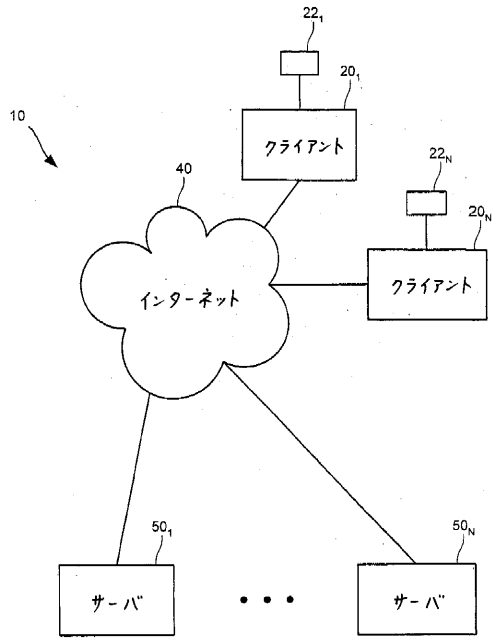

FIG. 1

【図２】

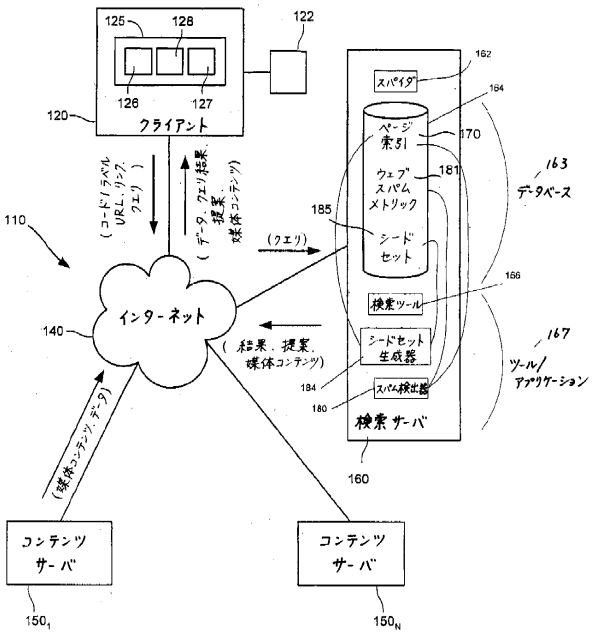

FIG. 2

【図３Ａ】

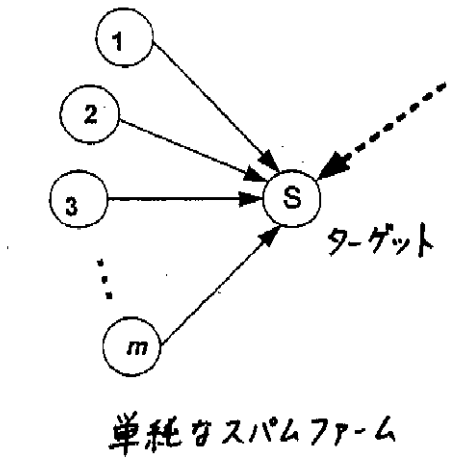

FIG. 3A

【図３Ｂ】

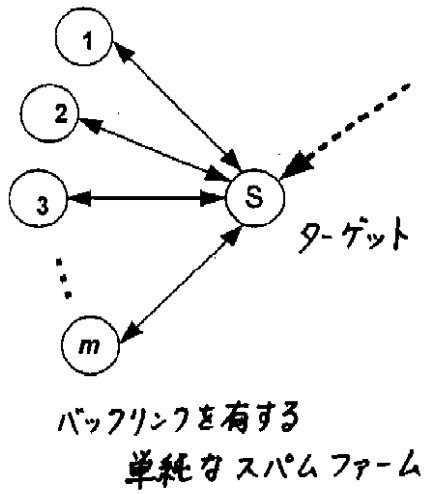

FIG. 3B

フロントページの続き

(74)代理人 100120525
弁理士 近藤 直樹
(72)発明者 バーキン，ペイベル
アメリカ合衆国、９４０８７ カリフォルニア州、サニーベール、ターンストーン・ウェイ、１３７８
(72)発明者 ギョンギィ，ゾルタン・アイ
アメリカ合衆国、９４３０５ カリフォルニア州、スタンフォード、オルムステッド・ロード、２７、アパートメント・１０１
(72)発明者 ペダーセン，ジャン
アメリカ合衆国、９４０２２ カリフォルニア州、ロス・アルトス・ヒルズ、ジョゼファ・レーン、２５７５０

審査官 千本 潤介

(56)参考文献 米国特許第０７５３３０９２（ＵＳ，Ｂ１）
米国特許出願公開第２００５／０２１６５３３（ＵＳ，Ａ１）
Zoltan Gyongyi, Hector Gracia-Molina, Jan Pedersen, Combating Web Spam with TrustRank, In VLDB, 米国, Morgan Kaufmann, ２００４年 3月１１日, p576-587
原田 昌紀 MASANORI HARADA, サーチエンジンにおける検索結果のランキング, ｂｉｔ Ｖｏｌ．３２ Ｎｏ．８, 日本, 共立出版株式会社, ２０００年 7月１４日, 第32巻第8号, 8-14ページ
成 凱 Kai CHENG, 多基準意思決定に基づくウェブ情報検索機能の改良 Improving IR Functions by Multicriteria Decision-Making Methods, 第１５回データ工学ワークショップ（ＤＥＷＳ２００４）論文集 ［ｏｎｌｉｎｅ］, 日本, 電子情報通信学会データ工学研究専門委員会, ２００４年 6月１８日
原田 昌紀 Masanori HARADA, WWWサーチエンジンの技術動向 Trends in Research on Web Information Retrieval, 電子情報通信学会技術研究報告 Ｖｏｌ．１００ Ｎｏ．６７１ IEICE Technical Report, 日本, 社団法人電子情報通信学会 The Institute of Electronics,Information and Communication Engineers, ２００１年 3月 1日, 第100巻第671号, 17-22ページ
風間 一洋 Kazuhiro Kazama, Ｗｅｂシステムにおける情報獲得支援技術 Supporting Technologies for Web-Based Information Acquisition System, 人工知能学会誌 第１６巻 第４号 Journal of Japanese Society for Artificial Intelligence, 日本, （社）人工知能学会 Japanese Society for Artificial Intelligence, ２００１年 7月 1日, 第16巻第4号, 503-508ページ

(58)調査した分野(Int.Cl., ＤＢ名)
G06F 13/00
G06F 17/30