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

特許第4668567号
(P4668567)

(45)発行日　平成23年4月13日(2011.4.13)　(24)登録日　平成23年1月21日(2011.1.21)

(51)Int.Cl.		F I		
G06F 13/00	(2006.01)	G06F	13/00	540A
G06F 15/00	(2006.01)	G06F	15/00	310U
G06F 17/30	(2006.01)	G06F	17/30	180Z

請求項の数 20　外国語出願　(全 28 頁)

(21)出願番号	特願2004-239997 (P2004-239997)
(22)出願日	平成16年8月19日 (2004.8.19)
(65)公開番号	特開2005-135381 (P2005-135381A)
(43)公開日	平成17年5月26日 (2005.5.26)
審査請求日	平成19年8月20日 (2007.8.20)
(31)優先権主張番号	10/670,681
(32)優先日	平成15年9月25日 (2003.9.25)
(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　エリック　ディー．ブリル
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　１７２　コート　ノースイースト　３０４２

最終頁に続く

(54)【発明の名称】クライアントベースのウェブクローリングのためのシステムおよび方法

(57)【特許請求の範囲】

【請求項１】

ウェブクローラと前記ウェブクローラによって収集されたウェブページ情報に関する第１のデータセットを格納した第１のストレージとを含む第１のコンピュータと、

ブラウザまたはプロキシサーバから訪れたウェブページのウェブページ情報に関する第２のデータセットを格納した第２のストレージを含む少なくとも１つの第２のコンピュータと、

を備え、前記第１のコンピュータは前記第１のデータセットから第１のウェブページに関する第１のウェブページ情報の第１の表現を生成して、その第１の表現を前記第２のコンピュータに送信し、前記第２のコンピュータは前記第１の表現に対応する前記第１のウェブページのウェブページ情報をブラウザまたはプロキシサーバを用いて収集し、前記収集したウェブページ情報の第２の表現が前記第１の表現と異なる場合、前記第２のコンピュータは収集したウェブページ情報に基づき前記第２のデータセットを更新して、その更新された第２のデータセットを前記第１のコンピュータに送信し、前記第１のコンピュータは前記第２のデータセットに基づき第１のデータセットを更新することを特徴とするデータ分析システム。

【請求項２】

ウェブクローラと前記ウェブクローラによって収集されたウェブページ情報に関する第１のデータセットを格納した第１のストレージとを含む第１のコンピュータと、

ブラウザまたはプロキシサーバから訪れたウェブページのウェブページ情報に関する第

２のデータセットを格納した第２のストレージを含む少なくとも１つの第２のコンピュータと、
を備え、前記第１のコンピュータは前記第１のデータセットから第１のウェブページに関する第１のウェブページ情報の第１の表現を生成して、その第１の表現を前記第２のコンピュータに送信し、前記第２のコンピュータは前記第１の表現に対応する前記第１のウェブページのウェブページ情報をブラウザまたはプロキシサーバを用いて収集し、前記第２のコンピュータは収集したウェブページ情報に基づき前記第２のデータセットを更新して、その更新された第２のデータセットを前記第１のコンピュータに送信し、前記第１のコンピュータは、前記第２のコンピュータから受信したウェブページ情報の第２の表現が前記第１の表現と異なる場合、前記第２のデータセットに基づき第１のデータセットを更新することを特徴とするデータ分析システム。

【請求項３】
前記ウェブクローラは、インターネットウェブクローラを含むことを特徴とする請求項１又は２に記載のシステム。

【請求項４】
前記ウェブクローラは、イントラネットウェブクローラを含むことを特徴とする請求項１又は２に記載のシステム。

【請求項５】
前記第１のコンピュータは、前記少なくとも１つの第２のコンピュータからの前記第２のデータセットの受信を制御するためのスケジューリング機能を提供することを特徴とする請求項１又は２に記載のシステム。

【請求項６】
前記第２のコンピュータは、前記第１のデータセットと前記第２のデータセットを比較して、前記第１のコンピュータのウェブクローラによって検索されたなりすましデータを検出するのにさらに使用されることを特徴とする請求項１又は２に記載のシステム。

【請求項７】
前記第２のコンピュータは、前記第１のデータセットに関連するデータについての状況情報を生成するのにさらに使用され、前記状況情報は、前記第２のコンピュータが複数ある場合、少なくとも１つの他の第２のコンピュータに送信されることを特徴とする請求項１又は２に記載のシステム。

【請求項８】
前記状況情報は、前記第１のデータセットに関連する情報の鮮度を示す鮮度フラグを少なくとも一部は含むことを特徴とする請求項７に記載のシステム。

【請求項９】
前記状況情報は、前記第１のデータセットに関連する情報のコンテンツのハッシュを少なくとも一部は含むことを特徴とする請求項７に記載のシステム。

【請求項１０】
前記状況情報は、前記第１のデータセットに関連する情報のコピーを少なくとも一部は含むことを特徴とする請求項７に記載のシステム。

【請求項１１】
第１のコンピュータが備えるウェブクローラによってウェブページ情報に関する第１のデータセットを生成して前記第１のコンピュータが備える第１のストレージに前記第１のデータセットを格納するステップと、
前記第１のコンピュータが第１のウェブページに関する第１のウェブページ情報の第１の表現を前記第１のデータセットに基づき生成するステップと、
前記第１のコンピュータからブラウザまたはプロキシサーバから訪れたウェブページのウェブページ情報に関する第２のデータセットを格納した第２のストレージを含む少なくとも１つの第２のコンピュータに前記第１の表現を送信するステップと、
第２のコンピュータが前記第１の表現に対応する前記第１のウェブページのウェブページ情報を収集するステップと、

第２のコンピュータが前記第１の表現に基づき収集したウェブページ情報の第２の表現を生成するステップと、
第２のコンピュータが、前記第２の表現が前記第１の表現と異なる場合、前記収集したウェブページ情報に基づき第２のデータセットを更新して、その更新された前記第２のデータセットを前記第１のコンピュータに送信するステップと、
前記第１のコンピュータが受信した前記前記第２のデータセットに基づき前記第１のデータセットを更新するステップと
を有することを特徴とするデータ分析方法。
【請求項１２】
第１のコンピュータが備えるウェブクローラによってウェブページ情報に関する第１のデータセットを生成して前記第１のコンピュータが備える第１のストレージに前記第１のデータセットを格納するステップと、
前記第１のコンピュータが第１のウェブページに関する第１のウェブページ情報の第１の表現を前記第１のデータセットに基づき生成するステップと、
前記第１のコンピュータからブラウザまたはプロキシサーバから訪れたウェブページのウェブページ情報に関する第２のデータセットを格納した第２のストレージを含む少なくとも１つの第２のコンピュータに前記第１の表現を送信するステップと、
第２のコンピュータが前記第１の表現に対応する前記第１のウェブページのウェブページ情報を収集するステップと、
第２のコンピュータが、前記収集したウェブページ情報に基づき第２のデータセットを更新して、その更新された前記第２のデータセットを前記第１のコンピュータに送信するステップと、
前記第１のコンピュータが前記第２のデータセットに基づき、前記第１のウェブページに関するウェブページ情報の第２の表現を生成するステップと、
前記第１のコンピュータが、前記第２の表現が前記第１の表現と異なる場合、受信した前記前記第２のデータセットに基づき前記第１のデータセットを更新するステップと
を有することを特徴とするデータ分析方法。
【請求項１３】
前記ウェブクローラは、インターネットウェブクローラを含むことを特徴とする請求項１１又は１２に記載の方法。
【請求項１４】
前記ウェブクローラは、イントラネットウェブクローラを含むことを特徴とする請求項１１又は１２に記載の方法。
【請求項１５】
前記第１のコンピュータが、前記少なくとも１つの第２のコンピュータからの前記第２のデータセットの受信を制御するスケジューリングするステップをさらに有することを特徴とする方法。
【請求項１６】
前記第２のコンピュータが、前記第１のデータセットと前記第２のデータセットを比較して、前記第１のコンピュータのウェブクローラによって検索されたなりすましデータを検出するステップをさらに有することを特徴とする請求項１１又は１２に記載の方法。
【請求項１７】
前記第２のコンピュータが、前記第１のデータセットに関連するデータについての状況情報を生成するステップをさらに有し、前記状況情報は、前記第２のコンピュータが複数ある場合、少なくとも１つの他の第２のコンピュータに送信されることを特徴とする請求項１１又は１２に記載の方法。
【請求項１８】
前記状況情報は、前記第１のデータセットに関連する情報の鮮度を示す鮮度フラグを少なくとも一部は含むことを特徴とする請求項１７に記載の方法。
【請求項１９】

前記状況情報は、前記第１のデータセットに関連する情報のコンテンツのハッシュを少なくとも一部は含むことを特徴とする請求項１７に記載の方法。
【請求項２０】
前記状況情報は、前記第１のデータセットに関連する情報のコピーを少なくとも一部は含むことを特徴とする請求項１７に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、概してデータ分析に関し、より詳細には、分散型ウェブクローラを使って、ネットワーク接続されたシステムから情報を取得するシステムおよび方法に関する。
【背景技術】
【０００２】
高コスト、低性能のデータ処理システムから、低コスト、高性能の通信システム、問題解決システム、および娯楽システムへの、コンピュータおよびネットワーク技術の発展により、書簡のやり取り、請求書の支払い、買物、予算の立案、および情報収集など、日常業務を実施するための負担を軽減する、コスト効率が高く時間を節約する手段がもたらされた。たとえば、有線または無線技術を介してインターネットとインターフェイスをとる計算機システムは、世界中に位置するウェブサイトおよびサーバのリポジトリからの大量の情報に、ユーザが指一本で、ほぼ瞬時に近いアクセスをするためのチャネルをユーザに提供する。
【０００３】
一般に、ウェブサイトおよびサーバを介して利用可能な情報は、ウェブクライアント（たとえばコンピュータ）上で実行されるウェブブラウザを介してアクセスされる。たとえば、ウェブユーザは、ウェブブラウザを展開し、ウェブサイトのＵＲＬ（Ｕｎｉｆｏｒｍ Ｒｅｓｏｕｒｃｅ Ｌｏｃａｔｏｒ）（たとえば、ウェブアドレスおよび／またはインターネットアドレス）をウェブブラウザのアドレスバーに入力し、キーボード上のエンターキーを押下しまたはマウスで「ｇｏ」ボタンをクリックすることによって、ウェブサイトにアクセスすることができる。ＵＲＬは通常、アクセスを容易にする４つの情報を含む。すなわち、情報交換のための規則および標準の集合を示すプロトコル（互いに通信するためのコンピュータ用言語）と、ウェブサイトまでの位置指定と、ウェブサイトを維持する組織名と、組織のタイプを識別する添字（たとえば、ｃｏｍ、ｏｒｇ、ｎｅｔ、ｇｏｖ、およびｅｄｕ）である。
【０００４】
場合によっては、ユーザは、自分がアクセスしたいと望むサイトもしくはサーバの名称、および／またはサイトもしくはサーバへのＵＲＬを事前に知っている。このような状況において、ユーザは、上述したように、アドレスバーにＵＲＬを入力しサイトに接続することによって、サイトにアクセスすることができる。しかし、ほとんどの場合、ユーザは、ＵＲＬもサイト名も知らない。代わりに、ユーザは、検索エンジンを利用して、自分が提供したキーワードに基づいてサイトの発見を容易にする。概して、検索エンジンは、キーワードを求めてウェブサイトおよびサーバのコンテンツを検索するとともに、キーワードが見つかったウェブサイトおよびサーバへのリンクの一覧を返す、実行可能なアプリケーションまたはプログラムからなる。基本的に、検索エンジンは、関連づけられたＵＲＬとしてできるだけ多くの文書を検索するウェブ「クローラ（ｃｒａｗｌｅｒ）」（別名、「スパイダー」または「ロボット」）を組み込む。この情報は次いで、インデクサ（ｉｎｄｅｘｅｒ）が、検索されたデータを処理することができるように格納される。インデクサは、文書を読み出し、各文書に含まれるキーワードおよび文書の他の属性に基づいて、優先順位をつけられた索引を作成する。それぞれの検索エンジンは一般に、所有権のあるアルゴリズムを利用して、クエリに対して有意義な結果が返されるように、索引を作成する。
【０００５】

したがって、ウェブクローラは、検索エンジンの動作にとって重要である。現在および最新の検索結果を提供するために、クローラは、ウェブを絶えず検索して、新しいウェブページを見つけ、古いウェブページ情報をアップデートし、消去されたページを削除しなければならない。インターネット上で見られるウェブページの数は天文学的である。したがって、ウェブクローラは、極度に高速であることが要求される。ほとんどのウェブクローラは、ウェブページを提供するサーバにポーリングを行うことによってデータを集めるので、クローラは、ある特定のサーバにアクセスするとき、できるだけ控えめにもしなければならない。そうでないと、クローラは、サーバの資源すべてを非常に高速に吸収し、サーバをシャットダウンさせてしまう場合がある。一般に、クローラは、サーバのウェブページにアクセスする前に、サーバに対してそれ自体を識別し、アクセス許可を求める。この時点で、サーバは、サーバの資源すべてを盗む不正クローラに対して、アクセスを拒否することができる。サーバをホスティングするウェブページは一般に、検索エンジンにより、ユーザがウェブページをより容易にみつけることが可能になるので、検索エンジンから利益を受ける。したがって、ほとんどのサーバは、サーバの資源すべてを使い果たさない限り、クローラを歓迎し、そうすることによって、サーバのコンテンツは、ユーザによってより便利に活用することができるようになる。

【発明の開示】

【発明が解決しようとする課題】

【０００６】

サーバに対してそれ自体を識別するクローラの欠点の１つは、サーバがクローラに対して「なりすまし（ｓｐｏｏｆ）」を行うことができることである。サーバは通常、広範なインターネットに対して公開したくない、保護された領域を有する。クローラがそれ自体を識別するとき、クローラは、どの領域にアクセスすることができないかも知らされる。クローラは、その特定のサーバとの仕事上の関係を維持したい場合、サーバの要求を遵守する。しかし、サーバは、その本当のコンテンツについてなりすましを行いまたは偽りたい場合、そのサーバの本当のＵＲＬを模倣しているが「代替」コンテンツを含むページ領域にクローラを向けさせることができる。したがって、猫に関する情報のみを通常は提供するサーバが、ウェブクローラのみがアクセスすることができるセクション中に、犬に関する情報を有するＵＲＬを設定することができる。これは、ユーザが「犬」を検索するとき、猫に関するサーバのウェブページが検索エンジンによって示されるように行われる。通常、なりすましは、サーバのコンテンツが世間によっていかがわしいとみなされているが、サーバが、通常の「キーワード」の範囲を超えてそのコンテンツを広めたいと思う場合に使用される。このようにして、いかがわしい素材が、たとえば花、犬、猫、天気など、一般的な言葉を使うことによって、検索エンジンの一覧で返される場合がある。なりすましは、正確さも、なりすましを受けたウェブクローラのデータを利用する検索エンジンの評判も低下させる。

【課題を解決するための手段】

【０００７】

以下では、本発明のいくつかの態様の基本的な理解をもたらすために、本発明の簡略な要約を提示する。この要約は、本発明の包括的な概要ではない。本発明の主要な／重大な要素を明らかにすることも、本発明の範囲を詳述することも意図していない。後で提示するより詳細な説明の前置きとして、本発明のいくつかの概念を簡略な形で提示することだけを目的としている。

【０００８】

本発明は、一般にデータ分析に関し、より詳細には、分散型ウェブクローラを使って、ネットワーク接続されたシステムから情報を取得するシステムおよび方法に関する。サーバのクライアントの、分散される性質は、高速かつ正確なウェブクローリングデータを提供するために利用される。サーバのウェブクローラによって集められた情報は、クローラのデータをアップデートするために、サーバのクライアントによって検索されたデータと比較される。本発明の一例では、データの比較は、検索エンジンの結果ページを介して広

められる情報を使うことによって遂行される。本発明の別の例では、データの妥当性確認は、ウェブクローラのデータを要約する、サーバから生じる、クライアントの辞書によって遂行される。本発明の一態様では、「弱インジケータ（ｗｅａｋ　ｉｎｄｉｃａｔｏｒ）」関数の集合の、ある弱インジケータ関数が、クライアントにランダムに送られる。こうした弱インジケータ関数は、サーバのウェブクローラによって見つけられるすべてのＵＲＬの全一覧より著しく小さく、したがって、サーバとクライアントの間の通信トラフィックを大幅に削減する。このことは、サーバとクライアントの間のインターフェイスの簡素化を容易にするとともに、ウェブクローラのデータの正確さを最適に保つ。

【発明の効果】

【０００９】

本発明は、ウェブクローラがなりすましを受けないよう阻止する手段を提供し、データの正確さを向上することによって、データ分析も容易にする。本発明を利用するサーバは、ウェブクローラのデータを、クライアントによって提供されるデータと比較することによって、なりすましを阻止することができる。このことは、サーバが、その検索エンジンから、なりすましたデータを除去することを可能にし、より高品質の検索エンジン結果を可能にする。この能力は、特に害のない検索において通常は返されない、いかがわしい素材をフィルタリングして取り除くことを容易にし、検索エンジンのクライアントに対して、よりユーザフレンドリーな体験をもたらす。

【００１０】

上記の目的および関連する目的を達成するために、本発明の例示的な態様を、本明細書において、以下の記述および添付の図面に関連して説明する。ただし、こうした態様は本発明の原理を利用することができる様々な方法のごくわずかを示すに過ぎず、本発明は、このようなすべての態様およびその等価物を含むことを意図する。本発明の他の利点および新規の特徴は、本発明の以下の詳細な説明を図面と併せ読むことにより、明らかになるであろう。

【発明を実施するための最良の形態】

【００１１】

ここで図面を参照して本発明を説明するが、同じ参照番号は、全体を通して同じ要素を指すのに使われる。以下の記述では、説明のために、多くの具体的な詳細を、本発明の完全な理解をもたらすために述べる。ただし、こうした具体的な詳細なしでも本発明を実施できることが明らかになるであろう。他の例では、本発明の説明を容易にするために、公知の構造および装置をブロック図の形で示す。

【００１２】

本願において使う「コンポーネント」という用語は、コンピュータ関連のエンティティ、すなわちハードウェア、ハードウェアおよびソフトウェアの組合せ、ソフトウェア、または実行中のソフトウェアのいずれかを指すことを意図している。たとえば、コンポーネントは、プロセッサ上で実行されている処理、プロセッサ、オブジェクト、実行ファイル、実行スレッド、プログラム、および／またはコンピュータでよいが、それに限定されない。例として、サーバ上で実行されているアプリケーションおよびそのサーバ両方がコンピュータコンポーネントとなることができる。１つまたは複数のコンポーネントが実行の処理をし、および／またはスレッド中に常駐することができ、コンポーネントは、１台のコンピュータに配置することも、および／または２台以上のコンピュータの間に分散することもできる。「スレッド」とは、オペレーティングシステムのカーネルが実行のためにスケジューリングする処理におけるエンティティである。当該分野において公知であるように、各スレッドは、スレッドの実行に関連づけられた揮発性データである、関連する「コンテキスト」を有する。スレッドのコンテキストは、システムレジスタのコンテンツおよびスレッドの処理に属する仮想アドレスを含む。したがって、スレッドのコンテキストを含む実際のデータは、実行時に変化する。

【００１３】

本発明は、ウェブ文書の索引を維持する、改良されたシステムおよび方法を提供する。

索引は、他のタイプの情報のデータを検索し、維持するのにも利用されることができる。従来のウェブクローラは、本発明によって軽減されるある欠点を有する。各クライアント（たとえば、ウェブにアクセスする任意の人のマシン）は、ローカルな情報を格納するので、したがって、クライアントが最後に訪れたときから、ウェブページが変更されているか否かを知ることができる。変更されている場合、クライアントは、この情報を検索エンジンに伝達することができる。同様に、サーバは、クライアントが訪れたウェブページについての情報を用いて、現時点でサーバにとって未知であるページを見つけることができる。文書を効率よく見つけ、そうした文書についての現時点での知識を維持することは、イントラネットおよびインターネット検索両方にとって、非常に重要なタスクである。本発明は、イントラネット検索などの状況でも利用することができ、その場合、ページをクロールし、サーバ上でページ情報を新鮮に保つことは、さらに重要な挑戦課題である。

【００１４】

検索エンジンの、（インターネット、イントラネット、またはそれ以外にとって）重要なコンポーネントは、データクローラまたは文書クローラである。文書クローラは、２つの主要なタスクを実施する。すなわち、検索エンジンによって索引を付けられるべき未知の文書をみつけること、および、その文書が、既知の各文書についての最新の知識を有することを保証しようと試みることである。こうしたタスクは両方とも、困難であり、（ページランクの質とともに）検索エンジンにおいて最も重要であり、目に見える品質の差別化要因に属する。文書クローラは一般に、サーバモデルに基づく。検索エンジンは、トポロジ検索によってウェブをクロールする。既知のウェブページからなるシードセット（seed set）から始まり、クローラは、そうしたページからのリンクをたどり、そうすることによって、シードセットからのパス（ＵＲＬ参照の集合）を介してつながれているすべてのウェブページを見つけることができる。検索エンジンが文書の集合体についての最新の知識をもつことを保証するために、クロールは、頻繁に繰り返されなければならない。クローラは、クロールを行う度にウェブページを再訪するので、どの程度頻繁にページ（または部分グラフ）が変更されるかを知ることができ、ページの過去の変更頻度に基づいて、特定のページを他のページより頻繁に再クロールする。

【００１５】

現在の、サーバベースのクローリングパラダイムには、いくつかの脆弱性がある。第１に、クローラは、シード文書の１つで始まる後続リンクによって到達することができるページしか発見することができない。最近の研究によると、大部分のウェブページは、どの検索エンジンによっても現時点で索引を付けられていないことがわかっている。第２に、検索エンジンは、クローラがページを再訪したときに、文書に対する変更（たとえば、コンテンツの変更や、既に存在しないページ）を知ることしかできない。

【００１６】

本発明は、上述した脆弱性を改善するやり方で、文書（たとえばデータ）を効率的に見つけ、既知の文書についての最新の知識を保持するシステムおよび方法を提供する。このことは、分散型の、クライアントベースのクロールによって達成される。各クライアント（たとえば、ウェブにアクセスする任意の人のマシン）は、ローカルな情報を格納し、したがって、クライアントが最後に訪れたときから、ウェブページが変更されているかどうかを知ることができる。変更されている場合、クライアントは、この情報を検索エンジンに伝達することができる。同様に、サーバは、クライアントが訪れたウェブページについての情報を用いて、現時点でサーバにとって未知であるページを見つけることができる。

【００１７】

図１に、本発明の一態様によるデータ分析システム１００のブロック図を示してある。本発明のこの事例において、データ分析システム１００は、１から「Ｎ」（Ｎは、１から無限大のどの数も表す）の番号がつけられたクライアント１０２～１０６、通信システム１０８、検索サーバ１１０、およびウェブページサーバ１１２からなる。クライアント１０２～１０６は、検索サーバ１１０用のウェブページ情報の「分散資源」群を備える。こうしたクライアントは一般に、新しいＵＲＬ、およびウェブページの変更などを、通信シ

ステム１０８を介して検索サーバ１１０に提供するように機能する。通信システム１０８は、インターネットおよび／またはイントラネットなどからなる。通信システム１０８は、検索サーバ１１０とクライアント１０２～１０６の間の通信用アクセス手段を提供する。通信システム１０８は、ウェブページ情報を集めるために、クライアント１０２～１０６と他のウェブページサーバ１１２、および／または検索サーバ１１０と他のサーバの間の通信も可能にする。本質的に、ウェブクローラの機能性は、検索サーバ内で機能するだけでなく、検索サーバ１１０およびクライアント１０２～１０６に分散される。検索サーバ１１０は、クライアント１０２～１０６を使って、ウェブページサーバ１１２から情報を取得し、検索サーバ自体の情報の洗練を容易にする。この機能を分散させることによって、本発明は、検索エンジンがそのデータを使用することができる、より最新の、頑強な、なりすましを防ぐデータセットを提供する。

【００１８】

図２に移ると、本発明の態様によるデータ分析システム２００の別のブロック図を示してある。データ分析システム２００は、クライアント２０２およびサーバ２０４からなり、それらの間で相互運用可能な通信手段を有する。通常の動作の間、サーバ２０４は、ウェブクローラをホスティングし、そのウェブクローラは、ウェブページをホスティングする他のサーバを求めて、インターネットなどの通信ネットワークを検索する。クローラは、ウェブページの検索エンジンでの使用のために、こうしたウェブページについての情報ソースを編集する。サーバ２０４は次いで、このウェブページ情報の表現をクライアント２０２に送る。この表現により、クライアント２０２は、その特定のウェブページをホスティングするサーバにアクセスするときに、独立してウェブページ情報を検証することが可能になる。クライアント２０２は、サーバ２０４にとって未知であるウェブページを検出することもできる。これにより、クライアント２０２は、既知および未知のウェブページについての、変更／状況および／または新しい情報を編集することが可能になる。この情報は次いで、サーバ２０４に送信される。サーバ２０４は、この情報を使って、サーバの、クローラ側の元のウェブページデータを改良する。分散資源を有することにより、サーバ２０４は、それ自体の直接資源（たとえば、プロセッサの使用、記憶空間など）に負荷をかけることなく、そのクローラ機能を拡張する。さらに、ウェブクローラは通常、アクセスする各サーバに対してそれ自体を識別するので、そのサーバ上の誤ったデータに宛先変更されるという危険がある。サーバも、ウェブクローラがサーバの資源に対して課し得るアクセス量および時間を制限することができる。サーバにアクセスするクライアントは一般に、こうした制限をもたず、誤ったデータに宛先変更されない。したがって、クライアントのウェブページデータは、ウェブクローラによって編集された、誤ったデータを修正するのに使うことができる。本発明のこの態様は、後でより詳しく説明する。

【００１９】

図３を参照すると、本発明の態様によるデータ分析システム３００のさらに別のブロック図を示してある。データ分析システム３００は、クライアントシステムコンポーネント３０２およびサーバシステムコンポーネント３０４からなり、その間で相互運用可能な通信システム（ＣＳ）３０６を有する。本発明のこの事例において、クライアントシステムコンポーネント３０２は、ＣＳインターフェイスコンポーネント３０８、クライアント制御コンポーネント３１０、およびデータ記憶コンポーネント３１２からなる。ＣＳ　ＧＵＩ（グラフィカルユーザインターフェイス）コンポーネント３０８は、通常は、利用される通信システムのタイプに特有のインターフェイスをユーザに提供する。このようなインターフェイスの一例は、少なくともワールドワイドウェブ上の情報を図表によって中継するために使われるウェブブラウザである。ウェブブラウザは、一企業内で供給されるウェブページなどのイントラネットを「サーフィンする」のにも使うことができる。本発明の他の例では、同様の情報を、グラフィカルユーザインターフェイスではなく、テキストベースのインターフェイスなどを使って中継することができる。一般に、ユーザが、通信システム３０６に接続されたリモートサーバ上に常駐する検索エンジンにおいて検索クエリを実行することを可能にするのは、このコンポーネント３０８である。したがって、ＣＳ

ＧＵＩコンポーネント３０８は、通信システム３０６から情報を受信し、かつ／または送信する。クライアント制御コンポーネント３１０は、ウェブクローリングを容易にすることに関するクライアントの制御を提供する。クライアント制御コンポーネント３１０は、たとえばウェブページなどの情報に関するデータを受信し、かつ／または送信する。クライアント制御コンポーネント３１０は、アルゴリズムを処理し、データの変更および状況を追跡し、かつ／またはデータ分析システム３００内のクライアント用のローカルデータ記憶を制御する。コンポーネント３１０は、ＣＳ　ＧＵＩコンポーネント３０８からの情報を、ウェブクローラから受け取った情報を用いて分析して、違いなどを判別することもできる。クライアント制御コンポーネント３１０は、クライアントが、ウェブクローラなどのための「分散資源」として関与することも可能にする。コンポーネント３１０は、格納されたデータにアクセスし、ＣＳ　ＧＵＩコンポーネント３０８に情報を提供することもできる。本発明の一例では、ＣＳ　ＧＵＩコンポーネント３０８は、埋め込まれたクローラのデータを送信し、かつ／または受信する。したがって、クライアント制御コンポーネント３１０は、ＧＵＩコンポーネント３０８とインターフェイスをとって、埋め込まれたクローラ関連データを、このようにして受信し、かつ／または送信する。同様に、コンポーネント３１０は、サーバからの制御を、同じやり方で送信し、かつ／または受信することもできる。本発明の別の例では、クライアント制御コンポーネント３１０は、サーバと同様に振る舞い、ピアツーピア方式で他のクライアントに制御を提供することができる。クライアント制御コンポーネント３１０およびＣＳ　ＧＵＩコンポーネントの機能は、単一のコンポーネント内で組み合わせることができることが、当業者には理解されよう。クライアントを、ＣＳ　ＧＵＩコンポーネント３０８をもたない分散資源として使うことも可能である。本発明のこの場合の例は、別のクライアントを中継し、かつ／または制御するクライアントを含むが、それに限定されない。データ記憶コンポーネント３１２は、たとえば、サーバからのクローラのデータ、クライアントからのクローラのデータ、ウェブページの変更、新しいウェブページデータ、およびクライアント制御パラメータなどを格納するのに使われる。コンポーネント３１２は、利用される本発明の例に応じて、クライアント制御コンポーネント３１０および／またはＣＳ　ＧＵＩコンポーネント３０８と直接、インターフェイスをとることができる。データ記憶コンポーネント３１２は、ハードドライブ、ランダムアクセスメモリ、読出し専用メモリ、取外し可能媒体、およびＣＤ－ＲＯＭなどのデータ記憶装置でもよい。本発明のさらに別の例では、データ記憶コンポーネント３１２に格納された情報は、ＣＳ　ＧＵＩコンポーネント３０８ともクライアント制御コンポーネント３１０ともインターフェイスをとることなく、サーバによって直接アクセスすることができる。いくつかの例では、このことは、より高速なデータの検索を可能にする。

【００２０】

本発明の一例では、通信システム３０６は、「インターネット」などの相互接続ネットワークでよい。通信システム３０６は、ＷＡＮ（ワイドエリアネットワーク）および／またはＬＡＮ（ローカルエリアネットワーク）などのイントラネットシステムでもよい。通信システム３０６は、より従来型の通信手段、たとえば、電話システム、無線システム、光信号（光学）システム、およびサウンドシステムなどを利用することもできる。他のグローバルおよびローカルネットワーク構造を、本発明によって通信システム３０６として使うこともできることが当業者には理解されよう。

【００２１】

サーバシステムコンポーネント３０４は、検索エンジンコンポーネント３１４、分散資源制御コンポーネント３１６、クローラコンポーネント３１８、データ記憶コンポーネント３２０、および任意選択によるＣＳデータホスティングコンポーネント３２２からなる。本発明の一例では、クローラコンポーネント３１８は、サーバおよび／またはプロキシサーバにアクセスするのに通信システム３０６を使って、ウェブページ関連情報、たとえばウェブページのコンテンツ、古さ、サイズ、ＵＲＬ、および埋込みリンクなどを取得する。この情報は次いで、データ記憶コンポーネント３２０に格納される。データ記憶コン

ポーネント３２０は、ハードドライブ、ランダムアクセスメモリ、読出し専用メモリ、取外し可能媒体、およびＣＤ－ＲＯＭなどのデータ記憶装置でよい。検索エンジンコンポーネント３１４は、ウェブクローラ３１８によって示されるとともにデータ記憶コンポーネント３２０に格納されるすべてのウェブページ用の検索機能を提供する。検索エンジンコンポーネント３１４は、ユーザから検索要求／クエリを受け取り、データ記憶コンポーネント３２０上の情報にアクセスしてリンク一覧およびウェブページデータを編集して、ユーザに送信する。したがって、一般的なシステムでは、検索コンポーネント３１４は、クローラコンポーネント３１８によって取得される情報にのみ依拠することができる。しかし、本発明の例では、分散資源コントローラ３１６は、データ記憶コンポーネント３２０に格納された情報の編集を容易にし、情報がより頑強、最新、より包括的になるようにする。分散資源制御コンポーネント３１６は、たとえば、単一の分散型クローラ、すなわち「クライアントベースのウェブクローラ」として相互作用する、サーバのクライアントなどの分散資源に対する制御を提供する。コンポーネント３１６は、クライアントシステムコンポーネント３０２などの分散資源から受信したデータの分析、機能ならびにデータの割振りと割振りのタイミングの判定、既知のクローラのデータを判定するための分散資源へのアルゴリズムの提供、データのアップデートおよび／または追加の受信、データのアップデートおよび／または追加のデータ記憶コンポーネント３２０への格納、分散資源の最適化された利用の決定、ある特定の検索クエリに対する検索結果ページへのデータの埋込みを可能にするための検索エンジンコンポーネント３１４へのページデータの提供、埋込みページリンク情報を含むページを生成するためのページデータのインターネットのサービスプロバイダへの提供、カウント、タイプ、なりすましの割合、およびソースなどのデータ特性の追跡などの機能を提供する。本発明の別の例では、コンポーネント３１６が通信システム３０６に直接アクセスするのではなく、検索ページコンポーネント３１４が、分散資源制御コンポーネント３１６用の情報を送信し、かつ／または受信する。

【００２２】

本発明の例では、任意選択のＣＳデータホスティングコンポーネント３２２は、通信システム３０６および分散資源制御コンポーネント３１６両方とインターフェイスをとる。ＣＳデータホスティングコンポーネント３２２は、ウェブページのホスト機能を提供し、ユーザにウェブページへのアクセスを提供する。ＣＳデータホスティングコンポーネント３２２は、分散資源制御コンポーネント３１６と対話するので、ウェブページのリンク情報を受信し、コンポーネント３２２がホスティングするウェブページに情報を直接埋め込むことができる。本発明の他の例では、ＣＳデータホスティングコンポーネント３２２は、データ記憶コンポーネント３２０と直接インターフェイスをとって、ウェブページに埋め込むための情報にアクセスする。本発明のさらに別の例では、ＣＳデータホスティングコンポーネント３２２は、検索エンジンコンポーネント３１４とインターフェイスをとって、そのウェブページのリンクに埋め込むための情報にアクセスする。本発明のさらに別の例では、ＣＳデータホスティングコンポーネント３２２は、クライアントなどの分散資源に常駐することができる。コンポーネント３２２は、サーバシステムコンポーネント３０４へのアクセス権を有する別のサーバに常駐することもできる。この例では、クライアント（またはサーバ）は、事実上、ホスティングされるウェブページに対するサーバとなり、ウェブページのリンクに埋め込むための情報を、そのローカルストレージおよび／または他のローカル手段から供給する。

【００２３】

各コンポーネントをそれぞれ独立に説明したが、本発明の他の例におけるコンポーネントは、他のコンポーネントに関連づけられた機能を含むことができることが当業者には理解されよう。同様に、いくつかのコンポーネントは、本発明の範囲を変えることなく、削除することができる。

【００２４】

図４に移ると、本発明の態様によるデータ分析システム４００を示すさらに別のブロック図を示してある。データ分析システム４００は、クライアントシステムコンポーネント

４０２およびサーバシステムコンポーネント４０４からなり、その間で相互運用可能な通信システム４０６を有する。本発明のこの事例において、サーバシステムコンポーネント４０４は、分散資源制御コンポーネント４１４およびデータ記憶コンポーネント４１６からなる。サーバシステムコンポーネント４０４は、クライアントシステムコンポーネント４０２からのウェブページ情報の受信に関して、本発明の事例を強調するために省略してある。通常、情報は、通信システム４０６を介して、分散資源制御コンポーネント４１４へ、およびコンポーネント４１４から流れる。クライアントシステムコンポーネント４０２は、クライアント制御コンポーネント４０８、データ記憶コンポーネント４１０、および任意選択の通知コンポーネント４１２からなる。本発明のこの事例において、通知コンポーネント４１２は、クライアントシステムコンポーネント４０２からサーバシステムコンポーネント４０４に流れるデータを制御する。本発明の他の例では、コンポーネント４１２は、クライアントシステムコンポーネント４０２と他のクライアントシステムコンポーネントとの間のピアツーピア通信も制御する。具体的には、通知コンポーネント４１２は、いつ、および／またはどのデータが、クライアントシステムコンポーネント４０２から送信されるべきかを決定する。決定は、蓄積されたウェブページのデータのサイズ、サーバシステムコンポーネント４０４にとって未知であるリンクが見つかったかどうか、ウェブページに対する変更の重要度（たとえば５０％以上のコンテンツの変更および／または重要度の高いページの変更など）、アクセス許可時刻、および／または分散資源制御コンポーネント４１４によって設定される一般的なアクセス許可時間などに基づくことができる。通知コンポーネント４１２は、データ転送のために、アルゴリズムを用いて、独自の重要度の要素および／または独自のタイミングスケジュールを決定することもできる。通知コンポーネント４１２の機能は、クライアント制御コンポーネント４０８および／または図４に示さない他のクライアントシステムコンポーネントに常駐できることが、当業者には理解されよう。

【００２５】

本発明の完全な理解のために、動作例を説明する。本発明の一例では、分散型クライアントベースのクローラは、以下のように動作する。潜在的な新しいウェブページ、およびウェブページのコンテンツ／状況変更についての着信クライアントメッセージを受信するサーバ、ならびにサーバと通信するクライアントの集合が存在するものと仮定する。クライアントマシンは、ウェブブラウジング用に使われるパーソナルコンピュータ、またはパーソナルコンピュータにページを供給するプロキシサーバのいずれでもよい。クライアントは、（１）ウェブページに到達するのに使われるＵＲＬ、（２）ウェブページのコンテンツのハッシュ、（３）ウェブページのコンテンツ、および（４）訪れた時間を含むことができるがそれに限定されない、閲覧されるウェブページ上の情報を集めるように装備される。本発明のいくつかの例（たとえばプロキシサーバなど）において、この情報すべてを保持することは実現不可能であり、いくつかの情報は、ある程度の期間保持されるだけである。

【００２６】

本発明の別の例では、クライアントは、ある特定のブラウザまたはプロキシサーバから訪れたウェブページのＵＲＬを、一定の期間記録し、次いで、このＵＲＬの集合をサーバに送信する。サーバは次いで、どのＵＲＬがサーバにとって未知であったかを調べ、そうしたＵＲＬを、今後のクロール／ダウンロード／索引づけのために、既知のＵＲＬ一覧に追加する。こうすることにより、サーバに関連づけられた検索エンジンは、トポロジカルなクロールによって見つけることができなかったウェブページについて知ることができるようになる。

【００２７】

クライアントからサーバに送信される情報のボリュームを減らすために、クライアントは、ある特定のＵＲＬをすでにサーバに知らせてあるか否かという情報をローカルに保持することができ、まだ知らせていない場合は、サーバに情報を送信するだけでよい。２つのウェブページが同じであるかどうかを効率的に判定する公知の方法がある。ハッシュ関

数によって、各文書を整数にマッピングし、次いで、２つのハッシュ値が同じであるかを調べる。ＵＲＬに関連づけられたコンテンツの最新のハッシュが、そのＵＲＬに関連づけられたコンテンツの、以前のハッシュと異なる場合、そのコンテンツは変更されている。クライアントは、ウェブページを訪れる度に、そのページのハッシュ値を計算する。クライアントは、そのページを訪れたことがある場合、ハッシュ値が変わっているかを調べる。ハッシュ値が変わっている場合、クライアントは、クライアントが最後にそのページを訪れた後にウェブページが変更されたと判定し、サーバに知らせることができる。クライアントは、新しい＜ｕｒｌ、ハッシュ値＞のペアをローカルに記録する。

【００２８】

クライアントが、変更についてサーバに知らせるための、異なるいくつかの方法がある。最も簡単な方法は、ＵＲＬのコンテンツ／状況が変更されたというメッセージを単に送信するだけである。次いで、サーバは、そのページをできるだけすぐに再クロールするようにスケジュールすることができる。サーバがページを再訪する必要をなくすために、クライアントは、付加情報を送信することができる。クライアントは、最後に訪ずれたときの、ページをキャッシュしたコピーをもっている場合、古いバージョンと新しいバージョンの間の違いとともに、古いハッシュ値、および新しいハッシュ値を送信することができる。サーバは最初に、クライアントの古いハッシュ値が、そのページの、サーバの現在のハッシュ値と一致するかを調べる。一致する場合、サーバは、それに従ってページのコンテンツをアップデートすることができる。一部の文書変更は、他の変更より重要であることに留意されたい。たとえば、ある場合には、ページ全体が変更されているが、別の場合には、ただ１つのコンマがある文に追加されているだけである。クライアントは、変更の重要度を計算し、（ａ）この情報を使って、どのアップデートをサーバに送信するかという優先順位を決定するか、または（ｂ）他のページ情報とともに重要度の値をサーバに送信し、そうすることによって、サーバがページの再クロール／再索引づけに優先順位をつける際にこの情報を利用できるようにすることができる。変更重要度関数の例は、変更された文書の割合、変更の言語的／意味的重要度、および変更によって影響を受けるユーザ検索の割合の推定などのような項目を含むが、それに限定されない。重要度は、ページの人気の推定によって重みづけすることもできる。

【００２９】

上述した通信手段に伴う欠点の１つは、クライアントとサーバの間の、重大なトラフィックのオーバーヘッドを生じることである。たとえば、１００個のクライアントがすべて、ページ「Ｘ」を初めて訪れる場合、クライアントはそれぞれ、ページ「Ｘ」を発見したというメッセージをサーバに送信する。同様に、サーバは、ページ「Ｙ」が変更されたことを通知されると、それ以外のクライアントからその通知を受ける必要はない。したがって、クライアントとサーバの間の不必要な通信を大幅に減少させる、本発明のこれ以外の例を後で説明する。

【００３０】

図５を参照すると、本発明の態様による、ページ検索結果を使用するデータ分析システム５００の図を示してある。データ分析システム５００は、検索結果ページ５０６を有するクライアント５０２と、クライアント５０２からサーバ５０４へ送信し（５０８）受信する（５１０）ための通信手段を有する検索サーバ５０４とからなる。本発明の事例の第１の例において、クライアント５０２は、サーバ５０４に、変更されたウェブページを通知するが、いかなる付加情報も送信しない。ユーザが検索エンジンを使用すると、検索サーバ５０４は、検索結果ページ５０６中の各ウェブページに関して、コンテンツの、サーバ側バージョンのハッシュと、コンテンツが、新しくないと知られているか否かを示す鮮度フラグとを含む結果とを、クライアントに提供する。クライアント５０２は、検索結果ページ５０６にあるページの１つを訪れる場合、最初に、サーバ５０４が、ページが新しくないことを知っているか否かを調べ（たとえば、別のクライアントがサーバ５０４に知らせたが、サーバ５０４がそのページを更新していない）、ページのコンテンツのハッシュを計算し、検索エンジンが提供したハッシュと比較する。２つのハッシュが一致しない

場合、クライアント５０２は、そのＵＲＬに関連づけられたコンテンツが変更されたという通知をサーバ５０４に送る。サーバ５０４は、通知を受け取ると、鮮度フラグの状況を変更し、再クロールのために、優先待ち行列にそのＵＲＬを追加する。

【００３１】

この例は、クライアントが、サーバにページ差異情報（この情報は、サーバが、ウェブページについてのサーバ側の情報を、クロールせずにアップデートするため、および／またはサーバがウェブページをいつ再クロールすべきかという優先順位をつけるために使うことができる）を送るシナリオ用に拡張することができる。この拡張は、検索エンジンに、各検索結果を有する２つの追加フィールド、すなわち最新クライアント通知の時間、および最新クライアント通知からのページのハッシュ値を送信させることによって、遂行することができる。クライアントが、検索エンジンによって返されたページを訪れて、（ａ）ｋｎｏｗｎ－ｎｏｔ－ｆｒｅｓｈフラグが偽であるか、または（ｂ）ｋｎｏｗｎ－ｎｏｔ－ｆｒｅｓｈフラグが真であり、かつ最新クライアント通知からのハッシュ値が、このクライアントがページに対して計算したハッシュ値と異なる場合、クライアントは、サーバに通知を行う。ページ変更の周期を認識し、そうすることによって、ページがＡからＢ、Ｃ、Ａへと繰り返し変更される場合、本発明によりその変更を認識し、このページについてのクライアントによるアップデートを制限できるようにすることも可能である。

【００３２】

クライアントとサーバの間の不必要な通信の量の低下に加え、「検索エンジンの結果ページによるメッセージ通信」のそれ以外の利点の１つは、サーバが既に知っているウェブページについての情報のみをクライアントがサーバに送ることを保証することによって、秘密に関する起こり得るいくつかの問題を回避することである。このようにして、クライアントは、たとえば、クライアントが秘密にしておくことを期待しているページを訪れず、サーバにこのページの存在を知らせないことが保証される。

【００３３】

上述した本発明の例に伴う欠点の１つは、サーバが、ユーザの検索クエリを介してクライアントに返したウェブページについての情報しか知ることができないことである。この要件は、サーバに、どの検索エンジンを介してクライアントに返されたウェブページについても知らせることによって、緩和することができる。クライアントは、ユーザがいかなる検索エンジンを使っていることも認識するように装備される。クライアントは、検索結果を訪ねるとき、コンテンツのハッシュを計算する。クライアントは、このＵＲＬを訪れたことがある場合、コンテンツのハッシュをキャッシュしている。２つのハッシュが異なる場合、クライアントは、ＵＲＬおよび新しいハッシュを（前回の訪問からの経過時間、および他の情報に応じて）サーバにアップロードすることができる。クライアントは、そのＵＲＬを訪れたことがない場合、ＵＲＬおよび新しいハッシュをサーバにアップロードすることができる。

【００３４】

しかし、クライアントは、サーバにとって既知であるＵＲＬの詳細な一覧のローカルなコピーをもっている場合、新規なものである可能性のあるＵＲＬに遭遇すると、そのＵＲＬが既知のＵＲＬ一覧にあるかを単に調べ、一覧にない場合は、そのＵＲＬをサーバに送るだけである。同様に、クライアントは、サーバにとって既知であるすべてのＵＲＬに対して、＜ｕｒｌ、ハッシュ値＞のペアの完全な一覧のローカルなコピーをもっているとすると、情報がサーバにとって新しいものである場合にアップデート情報を送るだけでよい。このアイディアに伴う問題は、こうした一覧全体を各クライアントに渡すのは実現不可能であることである。たとえば、検索エンジンは、数ギガバイトものデータとなる何十億ものＵＲＬについて知っている場合がある。重大な帯域幅の問題に加え、各クライアントがこのような一覧のためにこれ程大量のローカルストレージを費やすことを期待するのは現実的でない。

【００３５】

あるいは、本発明の別の例では、重大な帯域幅の問題を排除する通信手段が提供される

。たとえば、アルファベットＳが与えられると仮定する。この場合、Ｓ＊は、Ｓ以降の文字からなる全文字列の集合である。辞書Ｄを、集合Ｓ＊中の文字列の部分集合であると定義する。辞書Ｄ用のインジケータ関数Ｉ、すなわちＩ：Ｓ＊→｛０，１｝は、ｄ∈Ｄ（ｄ　ｅ　Ｄ）である場合、かつその場合に限り、Ｉ（ｄ）＝１というプロパティを有する。辞書Ｄ用の弱インジケータ関数$Ｉ_{w}$は、ｄがＤ中にないことを意味する$Ｉ_{w}$（ｄ）＝０（言い換えると、すべてのｄ∈Ｄに対して$Ｉ_{w}$（ｄ）＝１であり、$Ｉ_{w}$（ｄ）は、Ｄ中にない任意のｄに対して、０または１のいずれでもよい）というプロパティを有する関数である。最後に、弱インジケータ関数の適切な集合Ｉ＝｛$Ｉ_{w1}$，$Ｉ_{w2}$，．．．，$Ｉ_{wn}$｝を、Ｄ中にない任意のｄに対して、$Ｉ_{wi}$（ｄ）＝０であるような、少なくとも１つの$Ｉ_{wi}$∈Ｉ（$Ｉ_{wi}$　ｅ　Ｉ）が存在するというプロパティを有する弱インジケータ関数の有限集合であると定義する。

【００３６】

したがって、各クライアントは、Ｉからランダムに選ばれた弱インジケータ関数を受け取る。こうしたインジケータ関数は、ＵＲＬの集合全体より大幅に小さく、したがって、そうしたインジケータ関数をクライアントに送ることは現実的である。サーバによって知られているどのＵＲＬに対しても、インジケータ関数は、ＵＲＬが既知であると正しく判定する。サーバによって知られていないＵＲＬに対して、インジケータ関数は、既知であると誤ってラベル付けする場合があるが、その場合、クライアントは、何もしないか、または未知であると正しくラベルづけし、この場合、クライアントはサーバに知らせることができる。弱インジケータ関数の適切な集合の定義により、サーバにとって未知であるウェブサイトがクライアントによって訪問されるときはいつでも、クライアントのインジケータ関数がそのサイトを新しいものであると認識する確率がゼロでないことが保証される。

【００３７】

上で挙げた例をさらに簡略化するために、Ｓ＝｛ａ，ｂ，ｃ，ｄ｝であり、Ｓ＊中のすべての文字列は長さが４未満であり、辞書Ｄ＝｛ａｂｃ，ａｄｃ，ｂ，ｃｂｄ，ｄｄｄ）であると仮定する。この辞書用の弱インジケータ関数の例は、以下のようになる。

【００３８】

（第２の文字が｛ｂ，ｄ，ｎｕｌｌ｝の１つ）である場合、かつその場合に限り、Ｉ（文字列）＝１

弱インジケータ関数は、Ｄに対して以下のようにランダムに構成することができる。

【００３９】

（１）Ｄを、２つの重ならない部分辞書Ｄ’およびＤ”にランダムに区切る。

（２）「ｉ番目の文字が集合Ｓのメンバーである（Ｓは、Ｓの部分集合である）」という様式の１つまたは複数の項の結合からなる弱インジケータ関数Ｉ’を、Ｄ’に対してランダムに選ぶ。

（３）同じようにして、Ｄ”用の弱インジケータ関数Ｉ”をランダムに選ぶ。

（４）Ｉ’（ｘ）＝１またはＩ”（ｘ）＝１である場合、かつその場合に限り、関数Ｉ（ｘ）＝１を作成する。

【００４０】

このようなすべての弱インジケータ関数の集合は、弱インジケータ関数の適切な集合を生じる。クライアント辞書は、＜ｕｒｌ、ページのハッシュ値＞のペアからなる辞書をもつことによって、ページの最新性の検出という問題にも拡張することができる。

【００４１】

本発明の独自の一態様は、専用クローラの視点およびクライアントの視点から、クローラのデータを比較できることである。このことは、サーバの高度化が進む際、特に重要である。「より精密な」ソフトウェアを用いることによって、サーバは、サーバ中にあるデータの流れおよびアクセスをよりうまく制御することができる。これは、任意のまたはすべてのユーザが、サーバ上にある情報の一部または全てにアクセスするのを阻止できることを含む。異なるタイプのユーザに対して、サーバアクセス特権、さらにアクセス時間特

権に関して異なる「許可レベル」を与えることもできる。概して、こうした柔軟性の向上は、セキュリティ、有料アクセスの実施、および悪意のあるハッキングの防止など、建設的な目的のために利用される。しかし、サーバ上にあるウェブページの実コンテンツをマスクするのに利用されることも多い。図６に、本発明の態様による、ウェブクローラシステム６０２を伴うなりすまし処理６００のブロック図を示してある。処理６００は、ウェブクローラシステム６０２およびサーバ６０４を含み、その間で相互運用可能な通信システム６０６を有する。ウェブクローラシステム６０２は、クローラコンポーネント６０８およびデータ記憶コンポーネント６１０からなる。サーバ６０４は、サーバアクセス制御６１２、なりすましデータ６１４、および実データ６１６からなる。通常のクローラコンポーネント６０８がサーバ６０４にアクセスすると、コンポーネント６０８は、サーバ６０４に対してそれ自体をウェブクローラとして識別する。この識別は、「礼儀正しい」とみなされる。礼儀正しさは、サーバ規則を無視することによってサーバを悪用するウェブクローラが、一般に、将来にわたってサーバへのアクセスを拒否されるという点において、自己検閲である。サーバへのアクセスを拒否されることは、サーバアクセスに頼り、検索エンジンのユーザにコンテンツを提供する検索エンジンにとって、特に危機的である。したがって、クローラは通常、礼儀規則を遵守する。他の礼儀規則は、時間限定アクセス、サーバ資源の使用、およびデータの非破壊的な検索などを含む。この例では、サーバアクセス制御６１２は、クローラコンポーネント６０８を識別し、実データ６１６にアクセスを向けるのではなく、クローラコンポーネント６０８をなりすましデータ６１４に向ける。なりすましデータは通常、実データ６１６と同じＵＲＬ情報を含むが、異なるコンテンツを有する。これは、一般に、いかがわしいコンテンツをマスクするために行われる。一例として、サーバ６０４は、猫のおもちゃを検索している、熱心な猫の愛好者に、犬のコンテンツを有するＵＲＬを返すように、検索エンジンをだますことができる。サーバ６０４は、正しいＵＲＬを使ってなりすましデータ６１４を編集するが、コンテンツは、猫に関連する情報に変更されている。しかし、実データ６１６は、犬に関連する情報を含む。したがって、クローラコンポーネント６０８は、実際には、犬に関する場合でも、ＵＲＬが猫に関するものと思って、なりすましデータ６１４を検索する。クローラコンポーネント６０８は次いで、検索エンジンによってアクセス可能なデータ記憶コンポーネント６１０になりすましデータ６１４を格納する。その後、検索エンジンによる猫の検索も、犬の情報を含むＵＲＬを返す。この、犬／猫の例は害のないものと思われるであろうが、同じ技法が、たとえば広告、ポルノグラフィ、過激な文学、破壊活動集団、および他の主観的な攻撃的素材などをマスクするのにも利用することができる。

【００４２】

図７に、本発明の態様による、ウェブクローラシステム７０２を伴うなりすまし防止処理７００を示すブロック図を示してある。処理７００は、ウェブクローラシステム７０２、クライアント制御コンポーネント７０４、ウェブサーバ７０６、および相互接続性を提供する通信システム７０８を含む。ウェブクローラシステム７０２は、クローラコンポーネント７１０、なりすましデータ７１４を有するデータ記憶コンポーネント７１２、および比較コンポーネント７１８を有する分散資源制御コンポーネント７１６からなる。クローラコンポーネント７１０は、図６に関して示し、説明したようなウェブサーバ７０６からなりすましデータ７１４を検索する。なりすましデータ７１４は次いで、データ記憶コンポーネント７１２に格納される。この時点で、一般的な検索エンジン（図７に示さず）は、なりすましデータ７１４にアクセスし、本当のコンテンツを知らない検索エンジンのユーザにデータ７１４を広める。しかし、本発明を利用することにより、なりすましデータ７１４を除去することができる。これは、サーバは、クローラに対してなりすましを行うが、通常、サーバのウェブページにアクセスするユーザに対してはなりすましを行わないという事実のおかげで遂行することができる。本発明では、クライアント制御コンポーネント７０４などの分散資源を利用するので、コンポーネント７０４は、サーバ７０６にユーザとしてアクセスし、サーバ７０６から実データを検索することができる。クライアント制御コンポーネント７０４は次いで、実データ（すなわち「クライアントデータ」）

および／または実データの表現を、分散資源制御コンポーネント７１６に転送することができる。分散資源制御コンポーネント７１６内部の比較コンポーネント７１８は次いで、格納されているなりすましデータ７１４を検索し、クライアントから受け取った実データと比較することができる。データが異なる場合、分散資源制御コンポーネント７１６は、データ記憶コンポーネント７１２にあるなりすましデータ７１４を上書きすることができ、その不正確さを排除する。こうすることにより、検索エンジンは、そうしないと利用可能にならなかった正確なデータにアクセスできるようになる。

【００４３】

上で示し説明した例示的なシステムを念頭において、図８～１２のフロー図を参照すると、本発明によって実装することができる方法がよりよく理解されよう。説明を簡単にするために、この方法を一連のブロックとして示し、説明するが、本発明はブロックの順序に限定されないことを理解されたい。というのは、いくつかのブロックは、本発明によって、異なる順序で起こることもでき、かつ／または本明細書において示し説明する他のブロックと同時に起こることもできるからである。さらに、図示したすべてのブロックが、本発明による方法の実装に必要なわけではない。

【００４４】

本発明は、１つまたは複数のコンポーネント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という一般的な状況で説明することができる。概して、プログラムモジュールは、特定のタスクを実施しまたは特定の抽象データタイプを実装するルーチン、プログラム、オブジェクト、データ構造などを含む。一般に、プログラムモジュールの機能は、様々な実施形態において所望される場合には、組み合わせることも分散することもできる。

【００４５】

図８を参照すると、本発明の態様による、クライアントベースのウェブクローリングの方法８００のフロー図を示してある。方法８００は、８０２で始まり、８０４で、クライアントが、ウェブページを訪れたときに取得したウェブページ情報を記録する。本発明の単純な例において、情報は、訪れたウェブページのＵＲＬのみを含む。本発明のより複雑な例では、たとえば、ＵＲＬ、ウェブページのコンテンツのハッシュデータ、およびタイムスタンプなどを含むことができる。クライアントは次いで、８０６で、サーバにウェブページ情報を送る。本発明の一例では、クライアントが、他のクライアントにウェブページ情報を知らせることも可能である。やはり、単純な例では、情報は、ＵＲＬのみを含むことができ、または複雑な例では、情報は、ウェブページについてのいくつかの異なるタイプのデータを含むことができる。本発明の一例では、クライアントは、ウェブページ情報から派生した付加情報を生成する。このデータは、たとえば、ウェブページがアクセス可能な期間、アクセスの容易さ（過負荷、接続再試行など）、および埋込みリンク状況などを含むことができる。さらに、ウェブページ情報は、いつ情報が送られるかを制御するようにスケジュールすることができる。スケジューリングは、クライアントおよび／またはサーバによって開始することができる。情報を送るための基準は、時刻、時間の長さ、日付、集められたデータの量、および集められたデータのタイプ（たとえば、未知のデータまたは既知のデータ、発見されたなりすましデータ）などを含み得るが、それに限定されない。クライアントからサーバに送られる情報のボリュームを減らすために、本発明の一例では、クライアントが、ある特定のＵＲＬをサーバに知らせてあるかどうかに関する情報をローカルに保持し、まだ知らせていない場合のみ、サーバに情報を送ることができる。サーバは、ウェブページ情報を受け取ると、８０８で、情報を調べて、蓄積してあるデータと比較して新しいデータがあるかどうかを判定する。本発明の単純な例では、この判定は、サーバ上にすでに蓄積されたＵＲＬ一覧と比較して、ＵＲＬが新しいかどうか判定することを含む。未知の情報が見つかると、サーバは、８１０でサーバの蓄積データすなわち「既知の」データに未知の情報を追加し、８１２でフローが終わる。本発明の一例では、既知のデータは、サーバによる今後のウェブクローリング、ダウンロード、および／または索引づけなどに利用されるＵＲＬ一覧である。

【００４６】
図９に移ると、本発明の態様による、クライアントベースのウェブクローリングの方法９００の別のフロー図を示してある。方法９００では、最初に、９０４で、クライアントが、訪れたウェブページのコンテンツに対するハッシュ値を計算する。クライアントが、そのウェブページを複数回訪れている場合、以前のハッシュ値が、そのウェブページに対して計算され、格納されている。クライアントは次いで、９０６で、計算したばかりの、すなわち「最新の」ハッシュ値を、ウェブページに対する以前のハッシュ値と比較する。比較を行っているクライアントは、見つかった違いの重要度を設定することができる。たとえば、ある場合には、ページ全体が変更されているが、別の場合には、ただ１つのコンマがある文に追加されているだけである。クライアントは、変更の重要度を計算し、（ａ）この情報を使って、どの更新をサーバに送信するかという優先順位を決定するか、および／または（ｂ）他のウェブページ情報とともに重要度の値をサーバに送信し、そうすることによって、サーバがページの再クロール／再索引づけに優先順位をつける際にこの情報を利用できるようにすることができる。変更重要度の例は、変更された文書の割合、変更の言語的／意味的重要度、および変更によって影響を受けるユーザ検索の割合の推定などのような項目を含むが、それに限定されない。重要度は、ページの人気の推定によって重みづけすることもできる。通常、クライアントは、９０８で、ウェブページ状況情報をローカルに格納し、必要な場合は、格納されたこの情報をアップデートする。クライアントは次いで、９１０で、ウェブページ状況情報をサーバに知らせる。本発明の一例では、クライアントが、他のクライアントにウェブページ状況情報を知らせることも可能である。サーバおよび／またはクライアントが通知を受ける方法は、ＵＲＬのみ、ＵＲＬと新しいハッシュ、ならびに／またはＵＲＬと新しいハッシュおよび古いハッシュなどを含み得るが、それに限定されない。サーバ（または他のクライアント）は次いで、クライアントのウェブページ状況情報が、ＵＲＬの他に付加情報も含む場合、９１２で、その情報を、サーバ自体のウェブページ状況情報と比較する。サーバが、ウェブページのＵＲＬのみを状況変化として受信した場合、サーバは通常、サーバの以前のウェブページ状況情報と比較するための新しい状況情報を得るために、そのウェブページの再訪／クロールを開始する。サーバは次いで、９１４で、必要な場合は、サーバのウェブページ状況情報をアップデートし、９１６でフローが終わる。サーバがウェブページを再訪する必要をなくすために、クライアントは、付加情報を送信することができる。クライアントは、最後に訪ずれたときの、ページをキャッシュしたコピーをもっている場合、古いバージョンと新しいバージョンの間の違いとともに、古いハッシュ値、および新しいハッシュ値を送信することができる。サーバは最初に、クライアントの古いハッシュ値が、そのページの、サーバの現在のハッシュ値と一致するかを調べる。一致する場合、サーバは、それに従ってページのコンテンツをアップデートすることができる。

【００４７】
図１０に、本発明の態様による、クライアントベースのウェブクローリングの方法１０００のさらに別のフロー図を示してある。方法１０００は、１００２で始まり、１００４で、クライアントが、検索サーバ上で検索クエリを開始する。検索サーバは、１００６で、検索クエリを分解し、クエリに応答して検索結果の一覧を編集する。検索サーバは次いで、１００８で、埋込みウェブページリンク情報を有する検索結果ページを構成する。一般的な情報は、ウェブページのコンテンツの、サーバ側バージョンのハッシュ、および／または、コンテンツが、各ウェブページに対して新しくないと知られているかどうか（たとえば、別のクライアントが、ウェブページに対する新しいアップデートを検索サーバに知らせたが、検索サーバがそのページをまだ更新していない）を示すフラグを含み得るが、それに限定されない。したがって、「新しくないことが知られている」というフラグを有するウェブページは、検索サーバが、クライアントに、アップデート情報を送信するための検索を要求してほしくないウェブページである。検索サーバは次いで、１０１０で、検索を要求したクライアントに、埋込みリンクを有する検索結果ページを送信する。クライアントは、１０１２で、検索結果ページに列挙されているウェブページを訪れると、検

索サーバによって提供された、埋め込まれている鮮度フラグ（すなわち鮮度状況）を調べる。クライアントは、１０１４で、鮮度状況が「新鮮」であるとき、訪れたウェブページのコンテンツのハッシュを計算する。新鮮である状況は、検索サーバが、ウェブページの最近の、または最新のバージョンを所有していると思っていることを示す。したがって、クライアントは、ウェブページのコンテンツの新しいハッシュを計算し、１０１６で、このハッシュを、検索サーバによって提供された、埋め込まれているハッシュと比較する。クライアントは次いで、１０１８で、新しいハッシュと、検索サーバが提供したハッシュとの間に違いまたは差分が見つかったときは常に、検索サーバに通知する。検索サーバは次いで、１０２０で、通知を受信し、鮮度状況を「新しくないことが知られている」にアップデートし、再クロール用の一覧にもウェブページを追加し、１０２２でフローが終わる。本発明のこの事例において、再クロール一覧は、検索サーバが、「新しくないことが知られている」状況を有する、列挙されたウェブページのコンテンツの、サーバのハッシュをアップデートするのに用いるための方法である。検索サーバは、アップデートを完了するために、ウェブページを「再クロール」、すなわち再訪する。

【００４８】

別の例では、上記の方法は、最新クライアント通知の時間、および最新クライアント通知からのウェブページのハッシュ値を含むがそれに限定されない、各検索結果を有する追加フィールドを検索サーバに送信させることによって、クライアントが、検索サーバにページ差異情報（この情報は後で、検索サーバが、ウェブページについてのサーバ側の情報を、クロールせずにアップデートするため、および／または検索サーバがウェブページをいつ再クロールすべきかという優先順位づけを容易にするために使われる）を送信する方法を構成するように拡張される。クライアントが、検索サーバによって返されたページを訪れて、（ａ）「新鮮でないことが知られている」というフラグが偽であるか、または（ｂ）「新鮮でないことが知られている」が真であり、かつ最新のクライアント通知からのハッシュ値が、このクライアントがウェブページに対して計算したハッシュ値と異なる場合、クライアントは、検索サーバに通知を行う。ページ変更の周期を認識することによって、ページがＡからＢ、Ｃ、Ａへと繰り返し変更される場合、その変更を認識し、このページについてのクライアントによるアップデートを制限できるようにすることも可能である。

【００４９】

検索結果ページを使うことによって、クライアントとサーバの間のトラフィックは、本発明によって示したように、クライアントベースのウェブクローラにおいて徹底して削減される。さらに、クライアントの秘密は、サーバが検索結果ページ中で提供したウェブページのみがクライアントによってアップデートされることにより保たれる。このようにして、クライアントが、限定アクセスを有するウェブページを訪れた場合、この情報は、不注意によって検索サーバに送信されない。この方法の利点は、新しいウェブページが秘密でない場合でも、検索サーバが、クライアントを用いて、検索において利用するために既知のウェブページを拡張できないという点で、欠点でもある。

【００５０】

本発明の別の例では、方法（図示せず）は、検索サーバのウェブページ情報を利用するだけでなく、他の検索サーバのウェブページ情報も利用する。したがって、別の検索サーバによってクライアントに返された新しいウェブページは、新しいウェブページが存在することを検索サーバに通知するのに使うことができる。この方法もやはり、クライアントの秘密を保つ。というのは、この方法は、検索サーバが列挙していない、公に利用可能なウェブページのみを検索サーバに通知するからである。こうすることにより、検索サーバは、クライアントの信用を損なわずに未知のウェブページを追加できるようになる。通知は、ウェブページのＵＲＬ、ウェブページのコンテンツのハッシュ、ウェブページにアクセスしたときのタイムスタンプ、およびウェブページに対する以前のハッシュと比較した、新しいハッシュの差分などを含み得るが、それに限定されない。

【００５１】

図１１に移ると、本発明の態様による、クライアントベースのウェブクローリングの方法１１００のさらに別のフロー図を示してある。方法１１００は、１１０２で始まり、１１０４で、検索サーバが、弱（損失を伴う）インジケータ関数の集合を生成する。こうした関数を生成する方法は、後で説明する。検索サーバは、１１０６で、クライアントベースのウェブクローラを備えるクライアントに、ランダムに選択された弱インジケータ関数を送信する。クライアントは次いで、１１０８で、ランダムに選択された弱インジケータ関数によって既知でないと示されるウェブページ用のウェブページのデータを生成する。概して、既知でないウェブページのみが、弱インジケータ関数によって正確に表される。「既知の」ウェブページは、実際に既知であっても、既知でなくてもよい。クライアントは次いで、１１１０で、未知のウェブページのデータをサーバに送信する。サーバは次いで、１１１２で、このデータを使って、ウェブページに関するサーバ側の情報をアップデートし、１１１４でフローが終わる。

【００５２】

図１２を参照すると、本発明の態様による、クライアントベースのウェブクローリングのための弱インジケータ関数の適切な集合を生成する方法１２００のフロー図を示してある。方法１２００は、１２０２で始まり、１２０４で、検索サーバ上にあるウェブページ情報を表す辞書を、重なりのない部分辞書にランダムに分割する。一般に、部分辞書は、それぞれがウェブページ情報のグループ化における共通の特色を表すように選ばれる。弱（損失を伴う）インジケータ関数は次いで、１２０６で、各部分辞書が、ある特定の部分辞書にあるウェブページ情報を表すようにランダムに選ばれる。関数は次いで、１２０８で、少なくとも１つの部分辞書の弱関数が１に等しい場合、かつその場合に限り、Ｉ（ｘ）＝１となるように作成され、１２１０でフローが終わる。このようにして、弱インジケータ関数の「適切な集合」が生成される。したがって、たとえば、サーバによって知られているどのＵＲＬに対しても、インジケータ関数は、ＵＲＬが既知であると正しく判定する。サーバによって知られていないＵＲＬに対して、インジケータ関数は、既知であると誤ってラベルづけする可能性があるが、その場合、クライアントは、何もしないか、または未知であると正しくラベルづけする可能性があり、この場合、クライアントはサーバに知らせることができる。定義により、サーバにとって未知であるウェブサイトがクライアントによって訪問されるときはいつでも、クライアントのインジケータ関数がそのサイトを新しいものであると認識する確率がゼロでないことを、弱インジケータ関数の適切な集合が保証する。

【００５３】

本発明の様々な態様を実装するさらなる状況を提供するために、図１３および以下の説明では、本発明の様々な態様を実装することができる適切な計算機環境１３００の、簡潔で全般的な説明を提供することを意図している。これまでは、ローカルコンピュータおよび／またはリモートコンピュータを実行するコンピュータプログラムのコンピュータ実行可能命令という一般的な状況において本発明を説明したが、本発明は他のプログラムモジュールとの組合せにおいても実装できることが当業者には理解されよう。概して、プログラムモジュールは、特定のタスクを実施し、かつ／または特定の抽象データタイプを実装するルーチン、プログラム、コンポーネント、データ構造などを含む。さらに、発明性のある本方法は、他のコンピュータシステム構成とともに実施できることが当業者には理解されよう。他のコンピュータシステム構成には、シングルプロセッサコンピュータシステムまたはマルチプロセッサコンピュータシステム、ミニコンピュータ、メインフレームコンピュータ、ならびにパーソナルコンピュータ、ハンドヘルド計算装置、マイクロプロセッサベースの家電製品および／またはプログラム可能な家電製品などがあり、それぞれが１つまたは複数の関連する装置と動作可能に通信することができる。図示した本発明の態様は、通信ネットワークを介してリンクされるリモート処理装置によって特定のタスクが実施される分散型計算機環境でも実施することができる。ただし、すべてではなくともいくつかの本発明の態様は、独立型のコンピュータにおいても実施することができる。分散型計算機環境では、プログラムモジュールは、ローカルメモリ記憶装置および／またはリ

モートメモリ記憶装置内に配置することができる。

【００５４】

本アプリケーションで使用する「コンポーネント」という用語は、ハードウェア、ハードウェアおよびソフトウェアの組合せ、ソフトウェア、または実行中のソフトウェアのいずれかであるコンピュータ関連エンティティを指すことを意図している。たとえば、コンポーネントは、プロセッサで実行中の処理、プロセッサ、オブジェクト、実行ファイル、実行のスレッド、プログラム、およびコンピュータでよいが、それに限定されない。実例として、サーバ上で実行中のアプリケーションおよび／またはそのサーバがコンポーネントとなり得る。さらに、コンポーネントは、１つまたは複数の下位コンポーネントを含むことができる。

【００５５】

図１３を参照すると、本発明の様々な態様を実装する例示的なシステム環境１３００は、従来のコンピュータ１３０２を含み、このコンピュータは、処理装置１３０４、システムメモリ１３０６、およびシステムメモリを含む様々なシステムコンポーネントを処理装置１３０４に結合するシステムバス１３０８を含む。処理装置１３０４は、市販されているどのプロセッサでも、固有のどのプロセッサでもよい。さらに、この処理装置は、並列に接続することができるような複数のプロセッサから形成されるマルチプロセッサとして実装することができる。

【００５６】

システムバス１３０８は、従来の様々なバスアーキテクチャ、たとえばいくつか例を挙げると、ＰＣＩ、ＶＥＳＡ、マイクロチャネル、ＩＳＡ、およびＥＩＳＡのどれをも使用するメモリバスまたはメモリコントローラ、周辺バス、およびローカルバスなどいくつかのタイプのバス構造のいずれでもよい。システムメモリ１３０６は、ＲＯＭ（読出し専用メモリ）１３１０およびＲＡＭ（ランダムアクセスメモリ）１３１２を含む。ＢＩＯＳ（基本入出力システム）１３１４は、たとえば起動中にコンピュータ１３０２内部の要素間の情報の転送を助ける基本ルーチンを含み、ＲＯＭ１３１０に格納される。

【００５７】

コンピュータ１３０２は、たとえば、ハードディスクドライブ１３１６、たとえば取外し可能ディスク１３２０からの読出しまたはそこへの書込みを行うための磁気ディスクドライブ１３１８、および、たとえばＣＤ－ＲＯＭディスク１３２４または他の光学媒体からの読出しまたはそこへの書込みを行う光ディスクドライブ１３２２を含むこともできる。ハードディスクドライブ１３１６、磁気ディスクドライブ１３１８、および光ディスクドライブ１３２２は、それぞれハードディスクドライブインターフェイス１３２６、磁気ディスクドライブインターフェイス１３２８、および光ドライブインターフェイス１３３０によって、システムバス１３０８に接続される。ドライブ１３１６～１３２２およびそれに関連するコンピュータ可読媒体は、データ、データ構造、コンピュータ実行可能命令などを含む不揮発性記憶をコンピュータ１３０２に提供する。上記のコンピュータ可読媒体の説明では、ハードディスク、取外し可能な磁気ディスク、およびＣＤに言及したが、コンピュータ可読な他のタイプの媒体、たとえば磁気カセット、フラッシュメモリカード、デジタル映像ディスク、ベルヌーイカートリッジなども、例示的な動作環境１３００において使うことができ、さらに、このようなどの媒体も、本発明の方法を実施するコンピュータ実行可能命令を含むことができることが当業者には理解されよう。

【００５８】

オペレーティングシステム１３３２、１つまたは複数のアプリケーションプログラム１３３４、他のプログラムモジュール１３３６、およびプログラムデータ１３３８などいくつかのプログラムモジュールを、ドライブ１３１６～１３２２およびＲＡＭ１３１２に格納することができる。オペレーティングシステム１３３２は、適切などのオペレーティングシステムでも、オペレーティングシステムの組合せでもよい。一例として、アプリケーションプログラム１３３４およびプログラムモジュール１３３６が、本発明の態様によるクライアントベースのウェブクローリングを容易にすることを含むことができる。

【００５９】
ユーザは、キーボード１３４０およびポインティングデバイス（たとえばマウス１３４２）など１つまたは複数のユーザ入力装置を介して、コマンドおよび情報をコンピュータ１３０２に入力することができる。他の入力装置（図示せず）には、マイクロホン、ジョイスティック、ゲーム用パッド、衛星パラボラアンテナ、無線リモコン、スキャナなどがあり得る。こうしたおよび他の入力装置はしばしば、システムバス１３０８に結合されるシリアルポートインターフェイス１３４４を介して処理装置１３０４に接続されるが、他のインターフェイス、たとえばパラレルポート、ゲームポート、またはＵＳＢ（ユニバーサルシリアルバス）によって接続することもできる。モニタ１３４６または他のタイプの表示装置も、ビデオアダプタ１３４８などのインターフェイスを介してシステムバス１３０８に接続される。モニタ１３４６に加えて、コンピュータ１３０２は、他の周辺出力装置（図示せず）、たとえばスピーカ、プリンタなども含むことができる。

【００６０】
コンピュータ１３０２は、１つまたは複数のリモートコンピュータ１３６０への論理接続を使用してネットワーク接続された環境において動作できることを理解されたい。リモートコンピュータ１３６０は、ワークステーション、サーバコンピュータ、ルータ、ピア装置、または他の共通ネットワークノードでよく、通常、コンピュータ１３０２に関連して説明した要素の多くまたはすべてを含むが、簡潔にするために、メモリ記憶装置１３６２のみを図１３に示した。図１３に示した論理接続は、ＬＡＮ（ローカルエリアネットワーク）１３６４およびＷＡＮ（ワイドエリアネットワーク）１３６６を含むことができる。このようなネットワーク環境は、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、およびインターネットにおいてよく見られる。

【００６１】
ＬＡＮネットワーク環境において使われる場合、たとえば、コンピュータ１３０２は、ネットワークインターフェイスまたはアダプタ１３６８を介してローカルネットワーク１３６４に接続される。ＷＡＮネットワーク環境において使われる場合、コンピュータ１３０２は通常、モデム（たとえば、電話、ＤＳＬ、ケーブルなど）１３７０を含み、またはＬＡＮ上の通信サーバに接続され、あるいは、たとえばインターネットなどのＷＡＮ１３６６を介した通信を確立する他の手段を有する。モデム１３７０は、コンピュータ１３０２に対して内蔵型でも外付け型でもよく、シリアルポートインターフェイス１３４４を介してシステムバス１３０８に接続される。ネットワーク接続された環境では、プログラムモジュール（アプリケーションプログラム１３３４など）および／またはプログラムデータ１３３８は、リモートメモリ記憶装置１３６２に格納することができる。図示したネットワーク接続は例示的なものであり、本発明の態様を実施する際に、コンピュータ１３０２と１３６０の間の通信リンクを確立する他の手段（たとえば、有線または無線）も使うことができることが理解されよう。

【００６２】
他の指示がない限り、コンピュータプログラミングの当業者による実施に従って、コンピュータ、たとえばコンピュータ１３０２またはリモートコンピュータ１３６０によって実施される作用および象徴的に表した動作を参照して本発明を説明した。このような作用および動作は、コンピュータに実行されるものとして何度か言及した。こうした作用および象徴的に表した動作は、処理装置１３０４による、データビットを表す電気信号の処理を含み、その結果、電気信号表示の変換または減少、およびメモリシステム（システムメモリ１３０６、ハードドライブ１３１６、フロッピー（登録商標）ディスク１３２０、ＣＤ－ＲＯＭ１３２４、およびリモートメモリ１３６２など）内のメモリ位置でのデータビットの保持を引き起こし、そうすることによって、コンピュータシステムの動作、ならびに他の信号処理を再構成し、あるいは変更することが理解されよう。このようなデータビットが保持されるメモリ位置は、データビットに対応する特定の電気属性、磁気属性、または光学属性を有する物理的な位置である。

【００６３】

図１４は、本発明と相互作用する一例である計算機環境１４００の別のブロック図である。システム１４００はさらに、１つまたは複数のクライアント（群）１４０２を含むシステムを示す。クライアント（群）１４０２は、ハードウェアおよび／またはソフトウェア（たとえば、スレッド、処理、計算装置）でよい。システム１４００は、１つまたは複数のサーバ（群）１４０４も含む。サーバ（群）１４０４は、ハードウェアおよび／またはソフトウェア（たとえば、スレッド、処理、計算装置）でよい。サーバ１４０４は、たとえば、本発明を利用して変換を実施するためのスレッドを収容することができる。クライアント１４０２とサーバ１４０４の間の可能な１つの通信は、２つ以上のコンピュータ処理の間で伝送されるように適合されたデータパケットの形をとることができる。システム１４００は、クライアント（群）１４０２とサーバ（群）１４０４の間の通信を容易にするのに利用できる通信フレームワーク１４０８を含む。クライアント（群）１４０２は、クライアント（群）１４０２にローカルな情報を格納するのに利用できる、１つまたは複数のクライアントデータストア（群）１４１０に動作可能に接続される。同様に、サーバ（群）１４０４は、サーバ１４０４にローカルな情報を格納するのに利用できる、１つまたは複数のサーバデータストア（群）１４０６に動作可能に接続される。

【００６４】

本発明の一例では、ウェブクローリングを容易にする、２つ以上のコンピュータコンポーネントの間で伝送されるデータパケットは、少なくとも部分的に、ウェブクローリング用の分散型システムを少なくとも部分的に使用するウェブクローリングに関する情報からなる。

【００６５】

本発明の別の例では、ウェブクローリングを容易にするシステムの、コンピュータ実行可能なコンポーネントを格納するコンピュータ可読媒体は、少なくとも部分的には、ウェブクローリング用の分散システムによって編集される、ウェブページに関連する情報を少なくとも部分的に判定するウェブクローリングシステムからなる。

【００６６】

本発明のシステムおよび／または方法は、コンピュータコンポーネント、および非コンピュータ関連コンポーネントを同様に容易にするウェブクローリングシステムにおいて利用できることを理解されたい。さらに、本発明のシステムおよび／または方法は、有線および／または無線などでよい、コンピュータ、サーバ、および／またはハンドヘルド電子装置などを含むがそれに限定されない広範囲の電子関連技術において利用可能であることが当業者には理解されよう。

【００６７】

本発明は、サーバークライアントベースのクローリングシステムだけでなく、ピアツーピアのクローリングシステムにも利用できることも当業者には理解されよう。クライアントは、一般に「サーバ」の動作に関連づけられたタスクを実施することができ、したがって、本発明のいくつかの例において、サーバに関連付けられたいくつかの特性をクライアントに転送できることも可能である。本発明の一事例として、クライアントは、他のクライアントに対して「部分クロール」を実施して、サーバに送信するための情報を確認し、かつ／または検索する。この例は、たとえば、特定のクライアントとサーバの間のボトルネックを有するネットワークにおいて有益であり得る。データは、サーバへの最高のアクセス権を有するクライアントに転送することができる。本発明の他の例では、クライアントは、イントラネットシステムにおいて部分クロールを開始することによってサーバの動作を示すことができ、したがって、イントラネット上に存在する唯一の、および／または大幅に削減された数のクライアントから、サーバに情報を報告する。このようにして、検索サーバは、クライアントにおいて様々な部分クロールを開始して、サーバのクロール用資源を拡張することができる。

【００６８】

上記の説明内容は、本発明のいくつかの例を含む。当然ながら、本発明を説明するためのコンポーネントまたは方法のあらゆる組合せを説明することはできないが、本発明のさ

らに多くの組合せおよび入替えが可能であることが当業者には理解できよう。したがって、本発明は、添付の特許請求の範囲の精神および範囲内であるこのようなすべての変更形態、修正形態、および変形形態を包含することを意図したものである。さらに、詳細な説明または特許請求の範囲において「含む」という用語が使われている限りでは、「備える」は、使用される場合、請求項においては接続語として解釈されるが、「含む」のような用語は、「備える」という用語と同様に包括的であることを意図している。

【図面の簡単な説明】

【００６９】

【図１】本発明の態様によるデータ分析システムを示すブロック図である。

【図２】本発明の態様によるデータ分析システムを示す別のブロック図である。

【図３】本発明の態様によるデータ分析システムを示すさらに別のブロック図である。

【図４】本発明の態様によるデータ分析システムを示すさらに別のブロック図である。

【図５】本発明の態様による、ページ検索結果を使用するデータ分析システムを示す図である。

【図６】本発明の態様による、ウェブクローラシステムを伴うなりすまし処理を示すブロック図である。

【図７】本発明の態様による、ウェブクローラシステムを伴うなりすまし防止処理を示すブロック図である。

【図８】本発明の態様による、クライアントベースのウェブクローリングの方法を示すフロー図である。

【図９】本発明の態様による、クライアントベースのウェブクローリングの方法を示す別のフロー図である。

【図１０】本発明の態様による、クライアントベースのウェブクローリングの方法を示すさらに別のフロー図である。

【図１１】本発明の態様による、クライアントベースのウェブクローリングの方法を示すさらに別のフロー図である。

【図１２】本発明の態様による、クライアントベースのウェブクローリングのための弱インジケータ関数の適切な集合を生成する方法を示すフロー図である。

【図１３】本発明が機能することができる一例の動作環境を示す図である。

【図１４】本発明が機能することができる別の例の動作環境を示す図である。

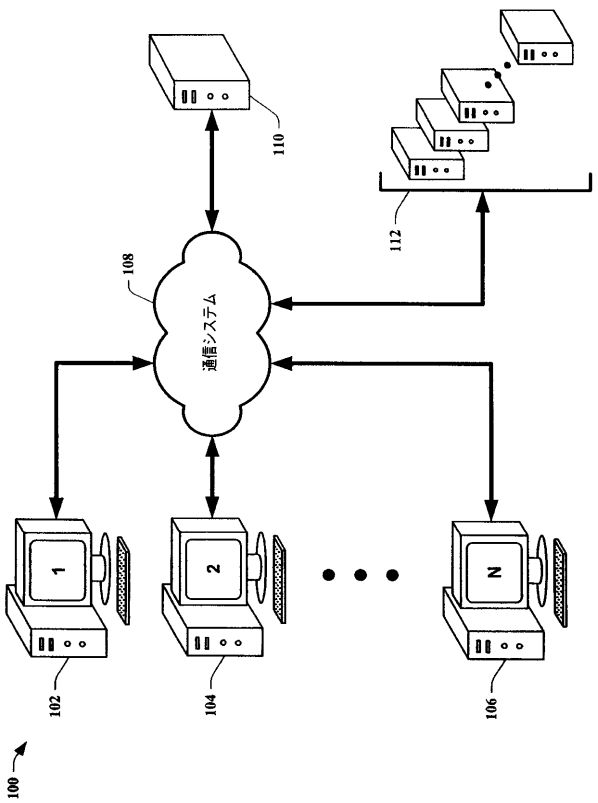
【符号の説明】

【００７０】

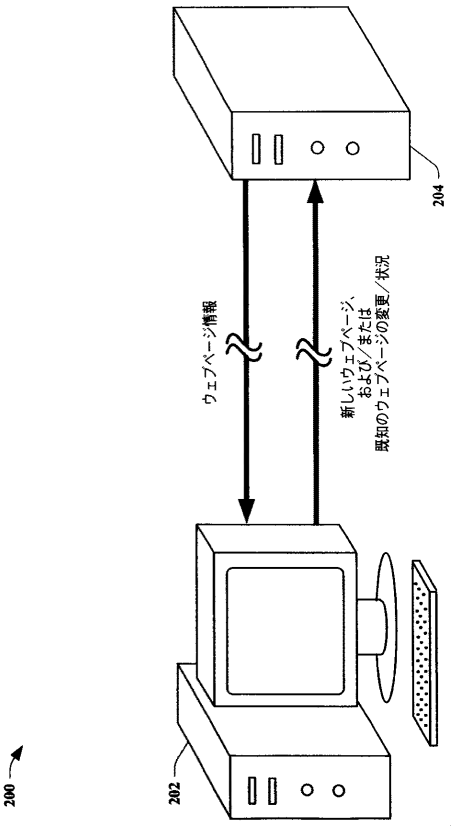
１００　データ分析システム
１０２～１０６　クライアント
１１０　検索サーバ
１１２　ウェブページサーバ
２００　データ分析システム
２０２　クライアント
２０４　サーバ
３００　データ分析システム
４００　データ分析システム
５００　データ分析システム
５０２　クライアント
５０４　サーバ
５０８　サーバへ送信
５１０　サーバから受信
６００　なりすまし処理
７００　なりすまし防止処理
７０６　ウェブサーバ
８００　クライアントベースのウェブクローリングの方法

９００　クライアントベースのウェブクローリングの方法
１０００　クライアントベースのウェブクローリングの方法
１１００　クライアントベースのウェブクローリングの方法
１２００　弱インジケータ関数の適切な集合を生成する方法
１３００　例示的なシステム環境
１３０２　従来のコンピュータ
１３１２　ＲＡＭ
１３１６　ハードディスクドライブ
１３１８　磁気ディスクドライブ
１３２０　取外し可能ディスク
１３２２　光ディスクドライブ
１３２４　ＣＤ－ＲＯＭディスク
１３４０　キーボード
１３４２　マウス
１３６２　メモリ記憶装置
１４００　計算機環境

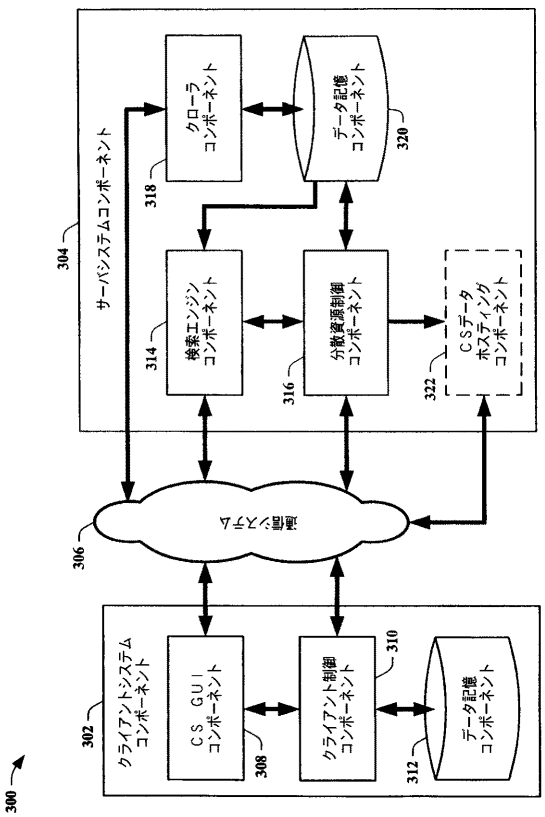
【図１】



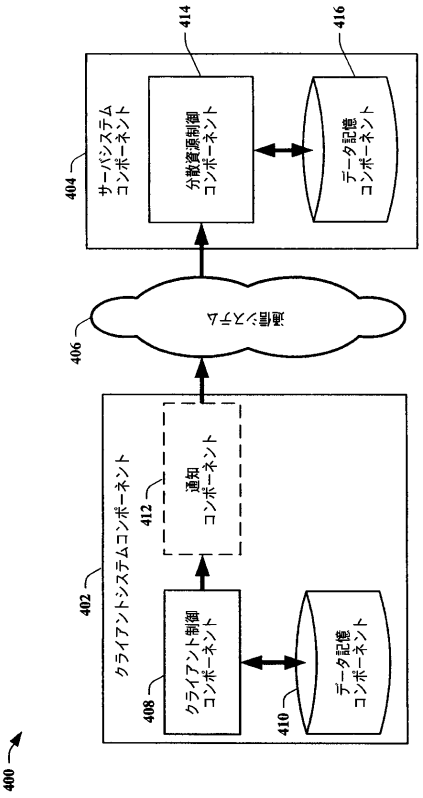
【図２】



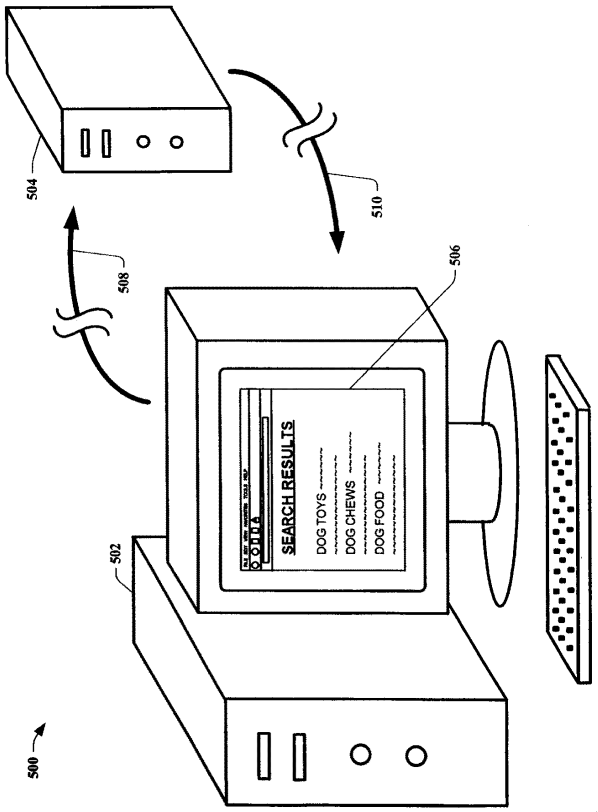
【図３】


300
302
クライアントシステムコンポーネント
CS GUIコンポーネント
308
クライアント制御コンポーネント
310
データ記憶コンポーネント
312
306
通信システム
304
サーバシステムコンポーネント
クローラコンポーネント
318
データ記憶コンポーネント
320
検索エンジンコンポーネント
314
分散資源制御コンポーネント
316
CSデータホスティングコンポーネント
322

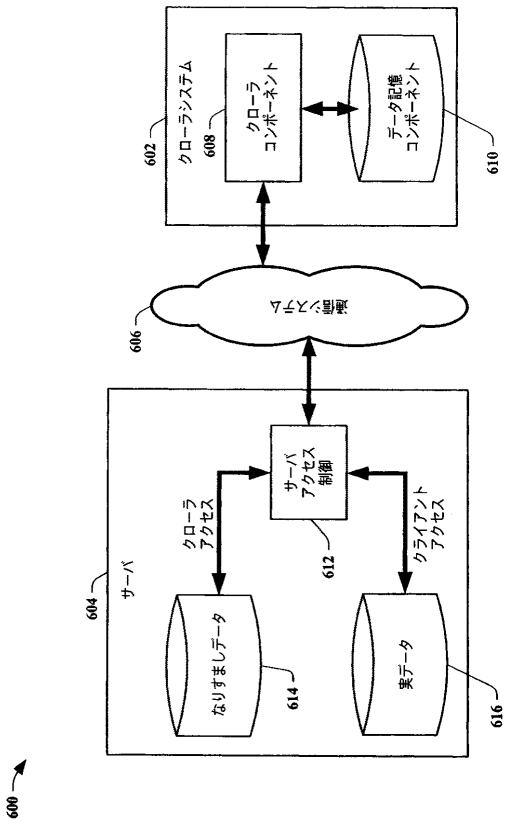
【図４】


400
402
クライアントシステムコンポーネント
クライアント制御コンポーネント
408
データ記憶コンポーネント
410
通知コンポーネント
412
406
通信システム
404
サーバシステムコンポーネント
分散資源制御コンポーネント
414
データ記憶コンポーネント
416

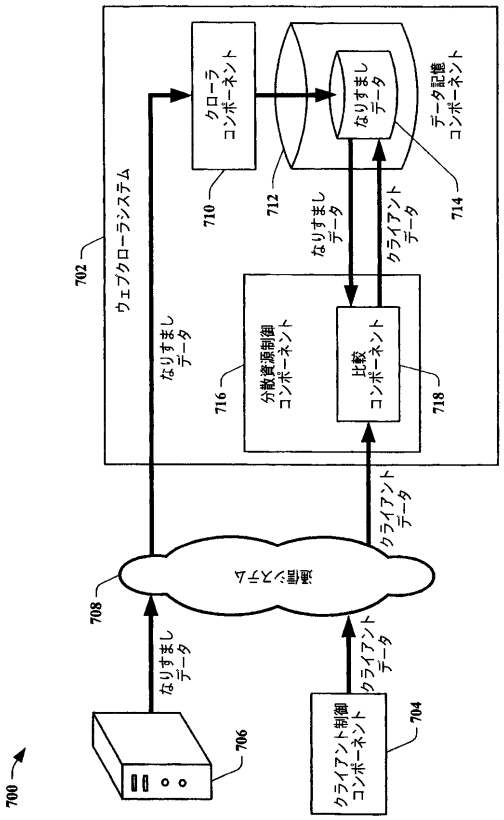
【図５】


500
504
508
510
506
502
SEARCH RESULTS
DOG TOYS
DOG CHEWS
DOG FOOD

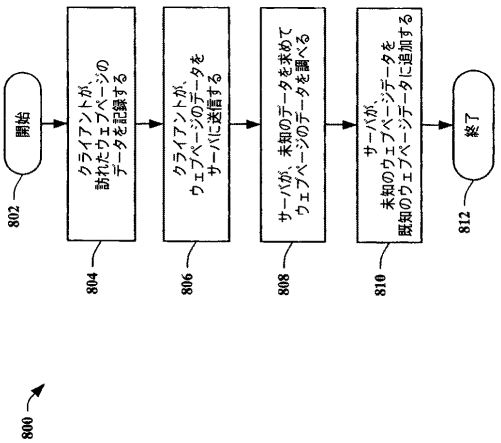
【図６】


600
602
クローラシステム
クローラコンポーネント
608
データ記憶コンポーネント
610
606
通信システム
604
サーバ
サーバアクセス制御
612
クローラアクセス
クライアントアクセス
なりすましデータ
614
実データ
616

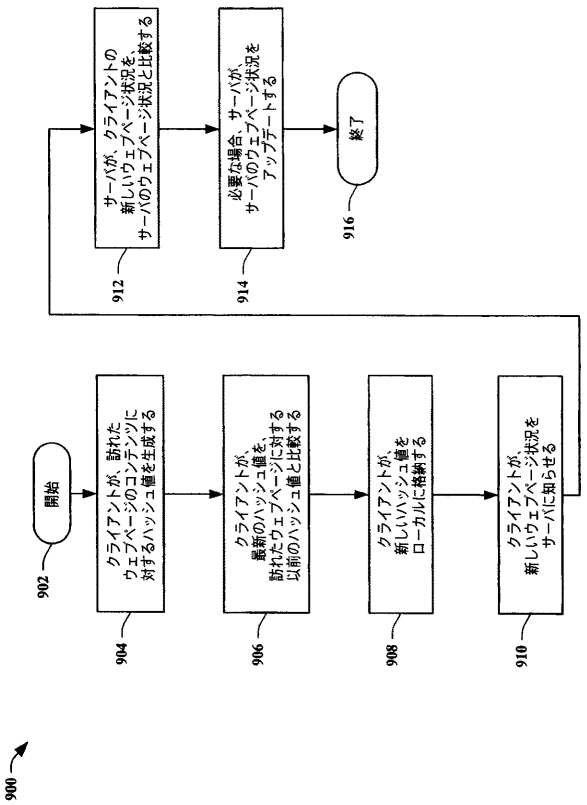
【図７】


700
ウェブクローラシステム
702
クローラ
コンポーネント
710
712
なりすまし
データ
データ記憶
コンポーネント
714
なりすまし
データ
クライアント
データ
分散資源制御
コンポーネント
716
比較
コンポーネント
718
なりすまし
データ
クライアント
データ
ネットワーク
708
なりすまし
データ
706
クライアント
データ
クライアント制御
コンポーネント
704

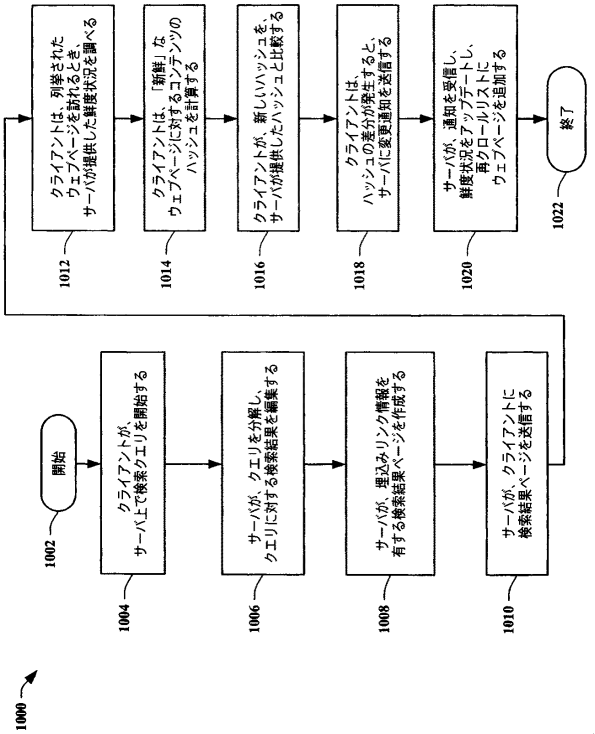
【図８】


800
開始
802
クライアントが、
訪れたウェブページの
データを記録する
804
クライアントが、
ウェブページのデータを
サーバに送信する
806
サーバが、未知のデータを求めて
ウェブページのデータを調べる
808
サーバが、
未知のウェブページデータを
既知のウェブページデータに追加する
810
終了
812

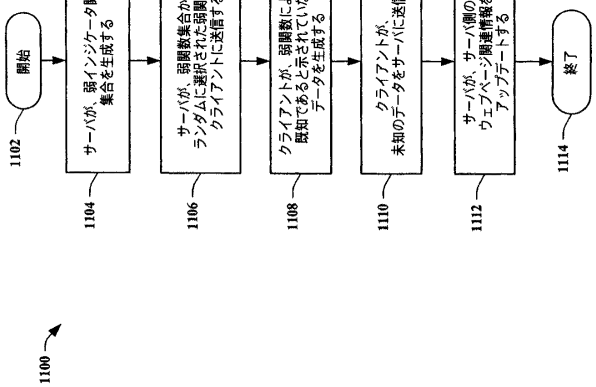
【図９】


900
開始
902
クライアントが、訪れた
ウェブページのコンテンツに
対するハッシュ値を生成する
904
クライアントが、
最新のハッシュ値を、
訪れたウェブページに対する
以前のハッシュ値と比較する
906
クライアントが、
新しいハッシュ値を
ローカルに格納する
908
クライアントが、
新しいウェブページ状況を
サーバに知らせる
910
サーバが、クライアントの
新しいウェブページ状況を、
サーバのウェブページ状況と比較する
912
必要な場合、サーバが、
サーバのウェブページ状況を
アップデートする
914
終了
916

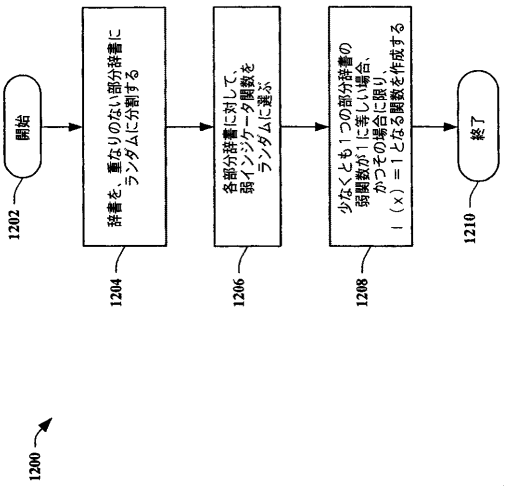
【図１０】


1000
開始
1002
クライアントが、
サーバ上で検索クエリを開始する
1004
サーバが、クエリを分解し、
クエリに対する検索結果を編集する
1006
サーバが、埋込みリンク情報を
有する検索結果ページを作成する
1008
サーバが、クライアントに
検索結果ページを送信する
1010
クライアントは、列挙された
ウェブページを訪れるとき、
サーバが提供した鮮度状況を調べる
1012
クライアントは、「新鮮」な
ウェブページに対するコンテンツの
ハッシュを計算する
1014
クライアントが、新しいハッシュを、
サーバが提供したハッシュと比較する
1016
クライアントは、
ハッシュの差分が発生すると、
サーバに変更通知を送信する
1018
サーバが、通知を受信し、
鮮度状況をアップデートし、
再クロールリストに
ウェブページを追加する
1020
終了
1022

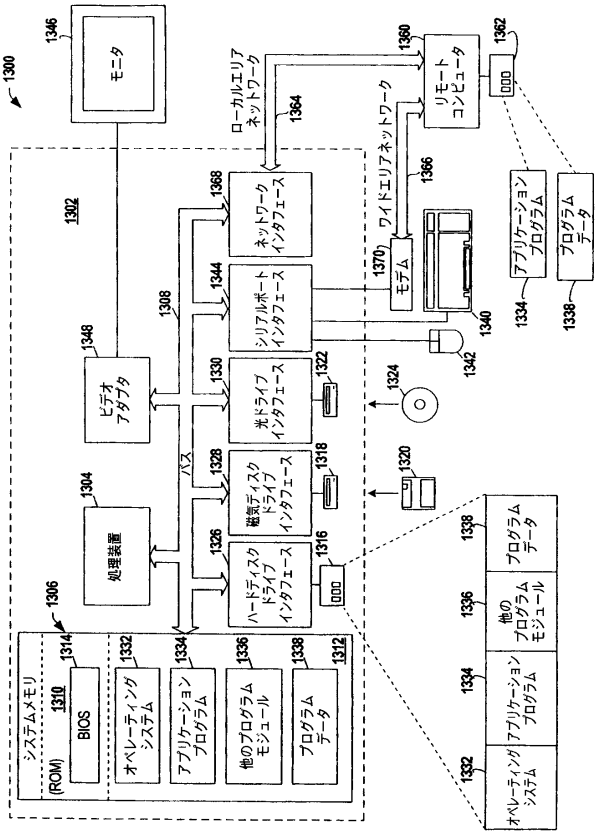
【図１１】



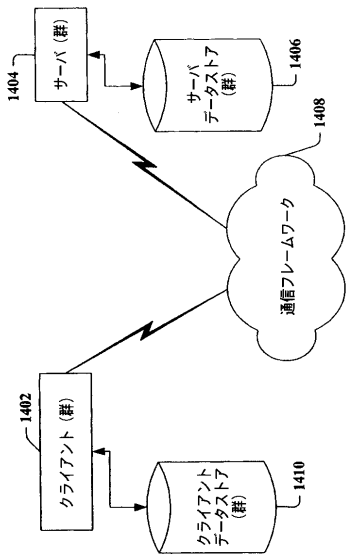
【図１２】



【図１３】



【図１４】



フロントページの続き

(72)発明者　クリストファー　エー．ミーク
アメリカ合衆国　９８０３３　ワシントン州　カークランド　ノースイースト　７１　ストリート　１２９３５

審査官　田上　隆一

(56)参考文献　特開２００２－３１２２８４（ＪＰ，Ａ）
特開２００２－１４０２５７（ＪＰ，Ａ）
岡崎　浩基，訪問者を確実に増やす　誰でもできるアクセス解析読み取り術　眠れるお宝を活用しろ！，月刊ＣＹＢｉＺ　ＳＯＨＯ　ｃｏｍｐｕｔｉｎｇ，日本，株式会社サイビズ，２００３年　5月　1日，第8巻　第5号

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１３／００