(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2008-508575
(P2008-508575A)
(43)公表日　平成20年3月21日(2008.3.21)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	110C	5B075
		G06F	17/30	240A	

審査請求 未請求　予備審査請求 未請求　（全 23 頁）

(21)出願番号	特願2007-519340 (P2007-519340)
(86)(22)出願日	平成17年6月24日 (2005.6.24)
(85)翻訳文提出日	平成19年2月26日 (2007.2.26)
(86)国際出願番号	PCT/US2005/022777
(87)国際公開番号	WO2006/004680
(87)国際公開日	平成18年1月12日 (2006.1.12)
(31)優先権主張番号	60/584,613
(32)優先日	平成16年6月30日 (2004.6.30)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	11/157,491
(32)優先日	平成17年6月20日 (2005.6.20)
(33)優先権主張国	米国 (US)

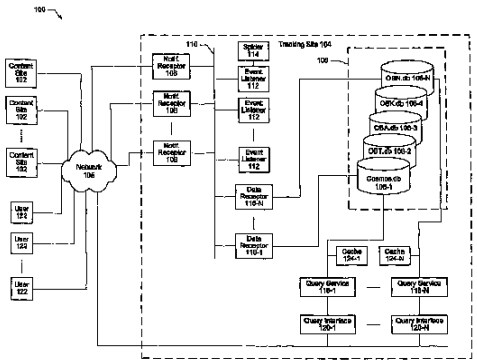
(71)出願人　505006910
テクノラティ，インコーポレーテッド
アメリカ合衆国，カリフォルニア州　９４１０７，サンフランシスコ，スウィート２０７，サードストリート　６６５
(74)代理人　110000028
特許業務法人明成国際特許事務所
(72)発明者　シフリィ・デイビッド・エル.
アメリカ合衆国　カリフォルニア州９４１２１　サン・フランシスコ，フルトン・ストリート，６０３４

Fターム(参考)　5B075 KK02　NS10　UU24

最終頁に続く

(54)【発明の名称】エコシステムを使用した集約および検索の方法、並びに、それらの関連技術

(57)【要約】
【課題】エコシステムを使用した集約および検索の方法、ならびに関連の技術を提供する。
【解決手段】ネットワークにおけるデータ集約を実現するための技術が説明される。ネットワーク上の複数のコンテンツサイトから、イベント通知が受信される。各イベント通知は、これらのコンテンツサイトのうちの対応する一コンテンツサイトに関連付けられたイベントの発生を示す。各イベントは、対応する一コンテンツサイトに関連付けられたコンテンツの変更に対応する。各イベント通知に応えて、イベントメタデータが生成される。イベントメタデータは、そのイベントのためのタイムスタンプと、そのコンテンツ変更に対応する変更データとを含む。イベントメタデータは、タイムスタンプを含む複数のインデックスメトリックを参照して、複数のデータベースにインデックス付けされる。イベントごとのイベントメタデータのインデックス付けは、対応するイベント通知の受信から数日以内、数時間以内、またはひいては数分以内に生じるので、インデックスメトリックのうちの任意のメトリックを使用して、複数のデータベー



【特許請求の範囲】
【請求項１】
コンピュータを使用した、ネットワークにおけるデータ集約の方法であって、
前記ネットワーク上の複数のコンテンツサイトからイベント通知を受信する工程であって、各イベント通知は、前記コンテンツサイトのうちの対応する一コンテンツサイトに関連付けられたイベントの発生を示し、各イベントは、前記対応する一コンテンツサイトに関連付けられたコンテンツの変更に対応する、工程と、
各イベント通知に応えて、前記イベントのためのタイムスタンプと、前記コンテンツの変更に対応する変更データとを含むメタデータを生成する工程と、
前記タイムスタンプを含む複数のインデックスメトリックを参照して、前記イベントメタデータを複数のデータベースにインデックス付けする工程と、
を備え、各イベントについての前記イベントメタデータのインデックス付けが前記対応するイベント通知の受信から７日以内に生じることによって、前記インデックスメトリックのうちの任意のメトリックを使用して前記複数のデータベースから前記コンテンツの変更を取得することを可能にする、方法。
【請求項２】
請求項１に記載の方法であって、
前記イベント通知は、コンテンツパブリッシングコードに関連付けられたイベント通知コードによって生成される、方法。
【請求項３】
請求項２に記載の方法であって、
前記コンテンツパブリッシングコードおよび前記イベント通知コードは、いずれも、前記コンテンツサイトのうちの選択されたコンテンツサイト上にある、方法。
【請求項４】
請求項３に記載の方法であって、
前記イベント通知コードは、前記選択されたコンテンツサイトの少なくとも一部では、前記コンテンツパブリッシングコードに統合されている、方法。
【請求項５】
請求項２に記載の方法であって、
前記コンテンツパブリッシングコードおよび前記イベント通知コードの少なくとも一部は、前記コンテンツサイトのうちの選択されたコンテンツサイトに関連付けられた、前記コンテンツサイトのうちの選択されたコンテンツサイトとは別の、第三者サービスプロバイダネットワーク上にある、方法。
【請求項６】
請求項２に記載の方法であって、
前記コンテンツパブリッシングコードは、ウェブログパブリッシングツール、ウィキウェブページ編集ツール、ソーシャルネットワークプロファイル編集ツール、コンテンツマネージメントシステムツール、およびパーソナルパブリッシングツールのうちの任意のものを含む、方法。
【請求項７】
請求項１に記載の方法であって、更に、
集約された変更データを第三者サービスプロバイダサイトから受信してインデックス付けする工程を備え、
前記集約された変更データは、前記第三者サービスプロバイダサイトに関連付けられた追加のコンテンツサイトに対応する経時的な追加のコンテンツ変更に対応する、方法。
【請求項８】
請求項１に記載の方法であって、更に、
前記イベント通知のうちの選択されたイベント通知に応えて、前記対応するコンテンツサイトから前記イベントメタデータの少なくとも一部を受信する工程を備える方法。
【請求項９】

請求項８に記載の方法であって、
前記イベントメタデータの少なくとも一部を受信する前記工程は、前記対応するサイトをスパイダリングすること、フィードのための既知のメタデータ位置を推測すること、およびプロファイルページのための既知のメタデータ位置を推測することのうちの任意を含む、方法。
【請求項１０】
請求項１に記載の方法であって、
前記イベント通知のうちの選択されたイベント通知のための前記イベントメタデータは、少なくとも部分的には、前記選択されたイベント通知を参照して決定される、工程。
【請求項１１】
請求項１に記載の方法であって、
各イベントのための前記イベントメタデータは、更に、前記イベントに対応する前記コンテンツサイトに関連付けられたＵＲＬ、前記コンテンツの変更に関連付けられたパーマリンク、コンテンツ作成者のＩＤ、イベントタイプのＩＤ、イベントのＩＤ、前記コンテンツ、前記コンテンツの変更、前記コンテンツに関連する少なくとも１つのキーワード、前記コンテンツへのインバウンドリンク、前記コンテンツからのアウトバウンドリンク、および前記イベントに関連付けられた任意の構造化データまたはメタデータ、のうちの任意のものに関連する、方法。
【請求項１２】
請求項１に記載の方法であって、
選択されたイベントのための前記変更データは、前記対応するコンテンツサイトに投稿された新しいコンテンツを含む、方法。
【請求項１３】
請求項１に記載の方法であって、
前記複数のデータベースは、全てのイベントに対応する前記イベントメタデータがインデックス付けされているグローバルデータベースを含む、方法。
【請求項１４】
請求項１に記載の方法であって、
前記複数のデータベースは、全てのイベントに対応する前記メタデータが主として前記タイムスタンプを参照してインデックス付けされている時間順のデータベースを含む、方法。
【請求項１５】
請求項１に記載の方法であって、
前記イベントのうちの選択されたイベントのための前記イベントメタデータは、前記選択されたイベントに関連付けられた前記コンテンツを指し示すリンクに関連するインバウンドリンクデータを含み、前記複数のデータベースは、前記選択されたイベントに対応する前記イベントメタデータが主として前記インバウンドリンクデータを参照してインデックス付けされている権威順のデータベースを含む、方法。
【請求項１６】
請求項１に記載の方法であって、
前記イベントのうちの選択されたイベントのための前記イベントメタデータは、前記選択されたイベントに関連付けられた前記コンテンツ内で識別されるキーワードに関連するキーワードデータを含み、前記複数のデータベースは、前記選択されたイベントに対応する前記イベントメタデータが主として前記キーワードデータを参照してインデックス付けされているキーワード順のデータベースを含む、方法。
【請求項１７】
請求項１に記載の方法であって、
前記イベントのうちの第１群の選択されたイベントのための前記イベントメタデータは、前記第１群の選択されたイベントに関連付けられた前記コンテンツを指し示すリンクに関連するインバウンドリンクデータを含み、前記イベントのうちの第２群の選択されたイ

ベントのための前記イベントメタデータは、前記第２群の選択されたイベントに関連付けられた前記コンテンツ内で識別されるキーワードに関連するキーワードデータを含み、前記複数のデータベースは、全てのイベントに対応する前記イベントメタデータがインデックス付けされているグローバルデータベースと、全てのイベントに対応する前記メタデータが主として前記タイムスタンプを参照してインデックス付けされている時間順のデータベースと、前記第１群の選択されたイベントに対応する前記イベントメタデータが主として前記インバウンドリンクデータを参照してインデックス付けされている権威順のデータベースと、前記第２群の選択されたイベントに対応する前記イベントメタデータが主として前記キーワードデータを参照してインデックス付けされているキーワード順のデータベースと、を含む、方法。

【請求項１８】

請求項１に記載の方法であって、

前記複数のデータベースの各自は、１つの関連するインデックスを有しており、前記イベントメタデータを前記複数のデータベースの各自にインデックス付けする前記工程は、

新しく受信されたイベントメタデータを、複数の増分式インデックスサイズのうちのより小さい方に相当する第１の複数の増分式インデックスにインデックス付けすることと、

前記第１の複数の増分式インデックスを統合し、前記複数の増分式インデックスサイズのうちのより大きい方に相当する第２の複数の増分式インデックスの１つを作成することと、

前記複数の増分式インデックスサイズの各自について、前記インデックス付けおよび前記統合を繰り返すことによって、全ての前記イベントメタデータを増分式に統合して前記インデックスを作成することと、

を含む、方法。

【請求項１９】

請求項１に記載の方法であって、更に、

クエリに応えて前記複数のデータベースから前記イベントメタデータの部分を取得する工程を備える方法。

【請求項２０】

請求項１９に記載の方法であって、

前記複数のデータベースの各自は、前記データベースのマスタコピーと、前記データベースの複数のスレーブコピーとを含み、前記クエリに応えて前記イベントメタデータの部分を取得する前記工程は、前記スレーブコピーから前記イベントメタデータを取得することを含む、方法。

【請求項２１】

請求項１に記載の方法であって、

前記イベントのうちの選択されたイベントのための前記イベントメタデータは、前記選択されたイベントに関連付けられた前記コンテンツを指し示すリンクに関連するインバウンドリンクデータを含み、前記複数のデータベースは、前記選択されたイベントに対応する前記イベントメタデータが主として前記インバウンドリンクデータを参照してインデックス付けされている権威順のデータベースを含み、前記方法は、更に、前記権威順のデータベースを参照して、複数のカテゴリの各自について権威ある個人を識別する工程を備える方法。

【請求項２２】

請求項２１に記載の方法であって、

各カテゴリについて前記権威ある個人を識別する前記工程は、前記権威ある個人によって生成された前記カテゴリ内の特定のコンテンツを、前記特定のコンテンツに関連付けられたインバウンドリンクの数を参照して識別することを含む、方法。

【請求項２３】

請求項２２に記載の方法であって、

各カテゴリについて前記権威ある個人を識別する前記工程は、更に、前記インバウンドリンクのリンク元である関連コンテンツに関連付けられた権威メトリックを参照して、前記インバウンドリンクの数に重み付けすることを含む、方法。

【請求項２４】

請求項２１に記載の方法であって、

各カテゴリについて前記権威ある個人を識別する前記工程は、前記権威ある個人に関連付けられた第２の特定のコンテンツが含むアウトバウンドリンクのリンク先である前記カテゴリ内の第１の特定のコンテンツを識別することを含む、方法。

【請求項２５】

請求項２４に記載の方法であって、

前記権威ある個人を識別する前記工程は、前記第２の特定のコンテンツに関連付けられた前記タイムスタンプを参照してなされる、方法。

【請求項２６】

請求項１に記載の方法であって、更に、

前記複数のデータベースを参照して、特定のイベントに対する複数の個人のレスポンスを追跡する工程を備える方法。

【請求項２７】

請求項２６に記載の方法であって、

前記レスポンスを追跡する前記工程は、前記個人によって生成され前記複数のデータベースにインデックス付けされた応答性のコンテンツを参照してなされる、方法。

【請求項２８】

請求項２６に記載の方法であって、更に、

ライブメディアへのフィードとして提示されるように動作可能である前記レスポンスの表現を生成する工程を備える方法。

【請求項２９】

請求項２８に記載の方法であって、

前記表現は、前記応答における経時的な変更を表すように動作可能である、方法。

【請求項３０】

請求項１に記載の方法であって、

各イベントについての前記イベントメタデータのインデックス付けは、１日以内、１２時間以内、６時間以内、３時間以内、２時間以内、１時間以内、１０分以内、および２分以内のうちのいずれかで生じる、方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、インターネット等のネットワークにおける動的コンテンツの監視に関するものである。本発明は、より具体的には、このようなコンテンツのほぼリアルタイムな監視およびインデックス付けを促進する技術に関するものである。

【背景技術】

【０００２】

ウェブ上およびインターネット上におけるユーザ生成コンテンツの公開が、多岐にわたる様々なソフトウェアソリューションによって促進されている。ソリューションは、ホストによって提供されるものもあれば、ユーザのマシンまたはサーバから操作されるものもある。また、ユーザによるカスタマイズが可能なソースコードを提供する高度に設定可能なものもある。このようなソリューションの、いわゆる「ウェブログ」の枠内における現在および過去の例として、Radio UserLand、Movable Type、Word Press、Live Journal、B2、Grey Matter、Blossom、Blogger、Blogspot、Type Pad、Xanga、Diaryland、Nifty等などが挙げられる。一般に、これらのツールおよびソリューションの大部分は、「パーソナルパブリッシング」を促進する比較的簡単なコンテンツマネージメントシステムとみなすことができる。これらのツールが利用可能になった結果、個人によってウェブ上に作成

されるのが通常であるウェブログ、すなわち「ブログ」が急増した。

【０００３】

一般的なブログは、「ブロガー」、すなわち投稿されるコンテンツの作者による、１つまたは複数のトピックに関する一連の投稿を含むことができる。投稿は、また、例えば、議論されている時事問題の関連記事へのリンク、ブロガーがコメントしているもしくは応答している別のブログへのリンク、または投稿の主題分野の権威へのリンクなども含むことができる。ブログは、また、通常の投稿の外側に、ブロガーが関心を持っているサイトもしくは文書を指し示すリンク、または他のブログを指し示すリンク（すなわちブログロール）を含むことも可能である。ブログは、また、これまでの投稿のアーカイブへのリンクを伴うカレンダーを含むことも多い。言うまでもなく、これらは、単にブログの代表的特性に過ぎず、ブログが比較的構造化された情報表示形態を有するという事実を指摘するのに有用である。また、ブログは、電子ネットワーク上においてコンテンツを動的に公開するためのメカニズムの一例に過ぎない。重要なのは、ウェブ上およびインターネット上で、他のコンテンツおよび情報へのリンクを含む膨大な量のコンテンツが動的に生成され公開されている点である。これらは、進行中の「対話」と見なすことができる。

【０００４】

そして、インターネット上で明言されているように、これらの進行中の相互に繋がれた対話は、市場と見なすことができる（例えば、「The Cluetrain Manifesto」を参照せよ）。これは、主に取引について市場を定義している従来の市場モデルと対照的である。取引に関する情報に主に依存して市場を監視または評価しようとすると、恐らく、監視または評価の対象である市場に関して最も重要な情報を見落とすことになる。このような従来のアプローチは、文書の内容ではなく文書中の句読点の傾向に焦点を当てることに喩えられる。そして、単なる取引データではなく、特定の市場に関する対話の内容に焦点を当てようとすると、今度は、これらの対話を有意義に且つ適時に追跡することが課題になる。

【０００５】

残念ながら、ウェブ上およびインターネット上で現在利用可能なツールの大部分は、このような作業を行うのに不十分である。例えば、インターネット上の検索エンジンの大部分は、その動作方式ゆえに、常に変化し続けるウェブ上のコンテンツを識別しカタログ化するのに数週間または数ヶ月の遅れをとっている。つまり、通常の検索エンジンは、ウェブを周期的に「クロール」することによって、ウェブのコピーであることを本質とする巨大なデータベースを構築している。ウェブの大きさを考慮すると、これらのクロールは、完了まで数週間程度が必要だと考えられる。しかも、このようなクロール技術が見ているのは、実際は、ウェブ上で利用可能なコンテンツの１０％未満に過ぎないという主張も多い。いずれにせよ、文書が識別されると、キーワードの語彙を使用して逆インデックスが作成され、これらのキーワードに応じて文書が記録される。次いで、これら全ての情報が、キーワード検索に応答するクエリサーバに転送される。

【０００６】

これら全ての作業を実施するために必要とされる時間を考慮すると、従来の検索エンジンは、ウェブ上にある２週間未満の新しさのものを識別するには特に有効なわけではないことが明らかになる。また、検索エンジンは、一般に、文書が作成された、または変更された時間に関しては一切不可知であるので、特定の時間範囲内に作成されたコンテンツを見つける、または任意の時間関連のメトリック（計量値）を参照してコンテンツを見つけるには特に有用なわけではない。

【０００７】

以上からわかるように、ウェブ上およびインターネット上の動的コンテンツを実質的にリアルタイムでインデックス付けする、監視する、および評価するためのメカニズムを提供することが必要とされている。

【発明の開示】

【０００８】

本発明にしたがって、ネットワークにおけるデータ集約を実現するための技術が提供さ

れる。ネットワーク上の複数のコンテンツサイトから、イベント通知が受信される。各イベント通知は、コンテンツサイトのうちの対応する一コンテンツサイトに関連付けられたイベントの発生を示す。各イベントは、対応する一コンテンツサイトに関連付けられたコンテンツの変更に対応する。各イベント通知に応答して、イベントメタデータが生成される。イベントメタデータは、そのイベントのためのタイムスタンプと、そのコンテンツ変更に対応する変更データとを含む。イベントメタデータは、タイムスタンプを含む複数のインデックスメトリックを参照して複数のデータベースにインデックス付けされる。各イベントについてイベントメタデータをインデックス付けする操作は、対応するイベント通知の受信から７日以内に生じるので、インデックスメトリックのうちの任意のインデックスメトリックを使用して、複数のデータベースからほぼリアルタイムでコンテンツ変更を取得することが可能になる。様々な実施形態によると、イベントメタデータのインデックス付けは、イベント通知の受信から１日以内、数時間以内、または、ひいては数分以内に生じる。

【０００９】

明細書の残りの部分および図面を参照することによって、本発明の特性および利点について更に理解することが可能になる。

【発明を実施するための最良の形態】

【００１０】

以下では、本発明を実施するのに最良であると発明者らが考える最良の形態を含む、特定の実施形態が詳細に説明される。これらの特定の実施形態は、添付の図面に例示されている。以下では、これら特定の実施形態との関連のもとで、本発明の説明が行われる。しかしながら、これは、当然のことながら、説明される実施形態に本発明を限定することを意図したものではなく、むしろ、添付の特許請求の範囲に定められた本発明の趣旨および範囲に含まれるものとして、あらゆる置換形態、変更形態、および等価形態を網羅することを意図している。以下の説明では、本発明の完全な理解を可能にするために、多くの詳細が特定される。しかしながら、本発明は、これらの一部または全部の詳細を特定しなくても実施可能である。また、本発明を不必要に不明瞭になる事態を避けるため、周知の特徴の詳細な説明は省略される。

【００１１】

本発明にしたがって、イベント駆動式の、データ集約およびインデックス付けの方法が提供される。本発明の様々な実施形態によると、本発明は、動的情報を、適時に、すなわちほぼリアルタイムで追跡し、インデックス付けし、検索することを可能にする。一部の実施形態によると、このような技術は、ウェブ上に公開されるコンテンツの半構造化特性を上手く活用することによって、そのようなコンテンツに関する関連情報を数週間以内ではなく数分間以内に追跡する。

【００１２】

本発明の具体的な実装形態は、参照される機能ブロックを、エコシステム内の他のサービスと対話する様々なサービス（すなわち、明確に定められたインターフェースを伴うソフトウェアオブジェクト）と見なす、「サービス指向型アーキテクチャ（ＳＯＡ）」を利用する。サービス指向型アーキテクチャ（ＳＯＡ）は、全ての機能、すなわち全てのサービスが、１つの記述言語を使用して記述され且つ処理の実行のために呼び出される呼び出し可能なインターフェースを有するアプリケーションアーキテクチャである。各相互作用は、他の全ての相互作用から独立し、通信機器（すなわち、通信システムを決定する基盤コンポーネント）の相互接続プロトコルは、インターフェースから独立している。インターフェースは、プラットフォーム独立であるので、どの通信機器からのクライアントでも、使用している言語およびオペレーティングシステムに依らずにサービスを使用することができる。

【００１３】

しかしながら、本明細書で説明される機能および処理は、他にも様々な方式で実装可能である。また、本明細書で説明される様々な機能ブロックは、それぞれ、ネットワーク内

の１つまたは複数の計算機プラットフォームに対応することができる。つまり、本明細書で説明されるサービスおよび処理は、個々のマシン上にあってもよいし、またはネットワーク内の複数のマシン間で、もしくはひいては複数のネットワーク間で分散されてもよい。したがって、本発明は、本発明の範囲から逸脱することなしに、多岐にわたる様々なハードウェア、ネットワーク構成、オペレーティングシステム、計算機プラットフォーム、プログラミング言語、サービス指向型アーキテクチャ（ＳＯＡ）、および通信プロトコル等などを使用して実装可能である。

【００１４】

以下の実施例の一部では、ブログの作成および管理のためのツールとして、コンテンツパブリッシングツールおよびコンテンツマネージメントツールにしばしば言及している。したがって、本発明の具体的な実施形態は、ブログの追跡について説明したものである。しかしながら、本発明による技術は、電子ネットワーク内においてコンテンツを生成および公開するための任意のツールに関するものであってよく、したがって、ブログへの言及によって制限されないことが理解されるべきである。このような他のツールの例としては、ウィキウェブページ編集ツール、ソーシャルネットワークプロファイル編集ツール、または他の任意の汎用もしくは特定用途向けのコンテンツマネージメントシステム（ＣＭＳ）もしくはパーソナルパブリシングツールが非限定的に挙げられる。より一般的に言うと、本発明によるデータ集約およびインデックス付けの技術は、本明細書で説明されるように、イベントとして特徴付け且つフラグ立てすることができるネットワーク上の情報のあらゆる状態変化によってトリガすることが可能である。

【００１５】

本発明にしたがって設計されたエコシステム１００を、図１を参照して説明する。ウェブ上には、様々なコンテンツサイト１０２が存在する。各コンテンツサイトでは、例えば上述されたブログツールなどの様々なコンテンツパブリッシングのツールおよびメカニズムを使用してコンテンツが生成され公開される。このようなパブリッシングのメカニズムは、コンテンツと同じサーバ上またはプラットフォーム上にあってもよいし、またはホストサービスであってもよい。

【００１６】

トラッキングサイト１０４が用意される。トラッキングサイト１０４は、任意のコンテンツサイト１０２でコンテンツの投稿または変更が発生するたびに、インターネット等の広域ネットワーク１０５を介してピング等のイベント通知を受信する。したがって、例えば、もしコンテンツがＴｙｐｅ　Ｐａｄを使用して変更されるブログであるならば、コンテンツの作成者が変更を公開すると、そのパブリッシングツールに関連付けられたコードがトラッキングサイト１０４と接続し、そのブログの名前およびＵＲＬを識別する例えばＸＭＬ遠隔手続き呼び出し（ＸＭＬ－ＲＰＣ）等を送信する。同様に、もしニュースサイトに新しい記事が投稿されたならば、イベント通知（例えばＸＭＬ－ＲＰＣ）が生成される。トラッキングサイト１０４は、すると、そのＵＲＬに「クローラ」を送信し、そこで見つけた情報を構文解析（パース）する。これは、データベース１０６内において、そのブログに関する情報をインデックス付けする、更新する、またはその両方を行うことを目的としている。具体的にブログに関する実施形態によると、ブログ内の情報の構文解析は、ブログの大部分が類似の構成を持つという事実、またはブログの大部分が周知のブログツールによって提供される一般的なアーキタイプもしくはテンプレートのいずれかにしたがう半構造化フォーマットを有するという事実によって促進される。一部の実施形態によると、ブログのスパイダリングおよび構文解析は、とりわけ、ブログの明示的および暗黙的な代替表現（例えばフィード）、外部メタデータ（例えばロボット、サイトマップ、および連絡先情報ファイル）、並びにブログアーカイブなどの使用によっても促進されうる。

【００１７】

一部の実装形態によると、トラッキングサイト１０４は、集約された変更情報を定期的に受信することができる。例えば、トラッキングサイト１０４は、他の「ピング」サービ

スから変更情報を取得することができる。つまり、複数のサイト上での変更に関する情報を蓄積させる、例えばＢｌｏｇｇｅｒなどの他のサービスが存在し、蓄積された変更情報をピングとして直接送信する。これらの変更は集約され、ｃｈａｎｇｅｓ．ｘｍｌファイルなどの形でトラッキングサイト１０４上において使用可能になる。このようなファイルは、上述されたピングと同様の情報を有するのが通常であるが、識別されたコンテンツが変更された時間、コンテンツが更新される頻度、ＵＲＬ、および類似のメタデータ等の情報を含むことも可能である。トラッキングサイト１０４は、この情報を、５分置きまたは１０分置きなどのように定期的に取得し、もしそのファイルを以前に取得したことがないならば、示されたサイトへとクローラを送信し、本明細書で説明されるように、そこで見つけた関連情報をインデックス付けして記録する。

【００１８】

また、トラッキングサイト１０４（または密接に関連付けられたデバイスまたはサービス）それ自体において同様の変更ファイルを蓄積させ、ピングを受信する毎ではなく定期的にデータベースに取り込むことも可能である。いずれにせよ、本発明の実施形態は、様々な技術を任意に組み合わせた任意の手法によって変更情報を取得するものと見なされる。

【００１９】

明らかなように、ピング等のイベント通知メカニズムは、多岐にわたる様々な形で実装されてよく、一般に、動的コンテンツの状態変化をシステムに通知するためのメカニズムとして特徴付けられる。このようなメカニズムは、パブリッシングツール（例えばブログツール）、またはＰＣ上もしくはウェブサーバ上のバックグラウンドアプリケーションに統合された、または関連付けられたコードに対応していてもよい。

【００２０】

様々な具体的な実施形態によると、トラッキングサイト１０４に対してピングを生成するメカニズムは、公開中のコンテンツの作者によって使用されているパブリッシングツールに何らかの形で統合されている。作者が、コンテンツの公開または投稿を（例えば画面上の「投稿および公開」オブジェクトを選択するなどによって）決定すると、パブリッシングツールに関連付けられたコードが、サイト１０４との間に特定のＵＲＬでＨＴＴＰ接続を確立し、「取得」または「投稿」のＨＴＴＰが、ＸＭＬ遠隔手続き呼び出し（ＲＰＣ）の形で伝送される。このコードは、トラッキングサイト１０４によって提供されてよく、パブリッシングツールに単純に関連付けられる、またはパブリッシングツールに不可欠な要素を構成することが可能である。

【００２１】

本発明の具体的な一実施形態によると、３種類の異なるピングが使用され、本明細書において、それぞれ標準ブログピング、拡張ブログピング、および非ブログピングと称される。標準ピングは、投稿サイト、すなわちウェブログの名称およびＵＲＬの２つの引数を有する。拡張ピングも、任意の関連のＲＳＳフィードを識別する。ブログサイト上の情報の相対的統一性および半構造化特性を考慮すると、ほとんどのブログサイトは標準ピングで大体十分である。非ブログピングは、より従来型のパブリッシャを対象としたものであり、メインサイトのＵＲＬと、新たに公開された文書の新しいＵＲＬとを含む。このピングは、パブリッシャによって自己選択された幾つかのカテゴリと、例えば作者などの任意のメタデータとを識別することができる。この情報は、このようなサイトに送信されたクローラが、ブログの場合の半構造化情報とは対照的に、任意のＨＴＭＬ文書をクロールするという意味で有用である。もちろん、本発明の範囲から逸脱することなく、その他のピングおよびイベント通知メカニズムも使用可能である。

【００２２】

図２のフローチャートも参照される。１つまたは複数の、例えばピングサーバなどの通知レセプタ１０８は、様々な場所から入ってくる様々なコンテンツ変更および状態変更に関する全てのイベント通知を取り込むイベントマルチプレクサとして機能する（２０２）。各通知レセプタ１０８は、これらのイベントについて、２つの非常に重要な事柄、すな

わち時間および発信元を理解する。つまり、通知レセプタ１０８は、入ってくる１つ１つのイベントにタイムスタンプを付け、そのタイムスタンプをイベントの発信元のＵＲＬに関連付ける（２０４）。通知レセプタ１０８は、次いで、そのイベントを、複数のイベントリスナ１１２が存在するバス１１０に乗せる（２０６）。

【００２３】

イベントリスナ１１２は、プレスリリース、ブログ投稿、求人一覧、任意のウェブページ更新、論評、カレンダー、関係、位置情報等など様々なタイプのイベントを探索する。イベントリスナの一部は、スパイダ１１４を含む、またはスパイダ１１４に関連付けられることが可能であり、スパイダ１１４は、特定タイプのイベントの認識に応答して関連のＵＲＬをクロールし、その通知を発生させた状態変更を識別する。別のイベントリスナは、受信された全タイプまたは特定タイプのイベントの数をカウントする単純なカウンタであることが可能である。

【００２４】

あるイベントリスナは、認識するように設計された各イベントを、やはり同様に設計された幾つかのピアへと再ブロードキャストする再ブロードキャスト機能を含む、または同機能に関連付けられることが可能である。その結果、事実上、イベントリスナの連合体が形成され、これは、例えば、特定タイプのイベントについて負荷平衡化スキームを実現することができる。

【００２５】

本明細書において「クチコミ」リスナ（”buzz” listener）と称される別のイベントリスナは、（例えばブログ投稿のコンテンツから決定されるなど）現在よく使用されているキーワードを、現在人々が話しているトピックの表れとして聴取し追跡するように構成することが可能である。更に別のイベントリスナは、イベントに関連付けられた任意のテキストを観察し、文字の種類や頻度などのメトリックを使用して言語を識別する。上述されるように、イベントリスナは、関心の対象となる事実上どのメトリックでも探索し追跡するように構成することが可能である。

【００２６】

イベントが認識され（２０８）、例えばスパイダなどの何らかのメカニズムを通してイベントデータが取得され（２１０）ると、イベントリスナは、ＵＲＬ（すなわちパーマリンク）、タイムスタンプ、イベントのタイプ、イベントのＩＤ、コンテンツ（必要に応じて）、並びに、例えばタグ、地理的情報、人、および事件などイベントに関連付けられた他の任意の構造的データまたはメタデータを非限定的に含む、各イベントのためのメタデータセットを出力する（２１２）。例えば、ＵＲＬは、そのイベントがニューヨークタイムズのウェブサイトで発生したものであることを示してよく、タイムスタンプは、そのイベントの時間を示してよく、イベントのタイプは、ブログ投稿を示してよく、イベントＩＤは、投稿ＩＤを示してよく、コンテンツは、任意のリンクを含むブログ投稿の内容を含んでよい。これらのメタデータは、ＵＲＬそれ自体から入手可能な情報から得られてもよいし、または、前述の言語決定アルゴリズムなどの何らかの形式の人口知能を使用して生成されてもよい。イベントメタデータは、スパイダリングだけでなく、例えばフィードまたはプロファイルページのための既知のメタデータ位置を推測することを含む様々な手段によって生成されてよい。

【００２７】

従来のクローラが比較的自律型で、且つ通常は特定のＵＲＬを対象としないという意味で言うと、本発明の具体的な実施形態によって使用される「クローラ」は、必ずしも従来の意味でのクローラでなくてよい。反対に、本発明によって使用されるクローラは、例えばサイトマップまたはｃｈａｎｇｅｓ．ｘｍｌファイルに列挙されるなどの特定のＵＲＬまたは特定のＵＲＬセットを対象としている。これらのクローラは、パーサを使用することができる。パーサは、クロールされている情報を分析し、例えば投稿などの関連部分を、エコシステムデータベース（例えばデータベース１０６）によって使用されているデータモデルに入れるように動作可能である。

【００２８】
一部の実施形態によると、サイト１０４は、新しい情報のみを確実にインデックス付けして記録するために、例えば前の投稿のハッシュなどの情報を維持する。これは、ＨＴＭＬ文書中の各パーツを別々に「経年」させられる非常に大規模なバージョン管理システムを可能にする。すなわち、ＨＴＭＬ文書中の、あらゆるリンクを含むあらゆる個別パーツの作成日を１つ１つ追跡することができる。
【００２９】
具体的な一実施形態によると、コンテンツは、例えば各コンテンツを観察し、アウトバウンドリンクおよび特異なフレーズを識別するなどによって、確立されたトピックディレクトリへのリンク、すなわちオントロジーに基づいて分類することができる。すると、オントロジー（例えばＤＭＯＺ（http://dmoz.org/を参照せよ）または他の任意の適切なオントロジー）に照らしてアウトバウンドリンクがチェックされ、コンテンツは、リンクのパターンに基づいて、自動的に、その特定のカテゴリ内のものとしてタグ付けされる。次いで、カテゴリ内における作者の相対的権威（後述される）と、カテゴリ内の文書へのインバウンドリンクとを参照して、その文書に適切な重みが割り当てられる。この重みは、ブログおよび投稿の自己カテゴリ化（例えば「タグ」）を更に盛り込むことが可能である。
【００３０】
イベントメタデータが格納される複数のデータベース１０６が維持される。具体的な一実施形態によると、各イベントリスナ、またはそれに関連付けられたスパイダ、またはそれらの両方は、あるイベントに関するメタデータが既に格納されているか否かを決定するために、そのイベントメタデータをデータベースに照らしてチェックするように動作可能である。これは、複数の通知を生成されたイベントの重複格納を回避する。新しいイベントが既に受信され、データベースに格納されているか否かを決定するためには、様々な発見的手法が使用可能である。例えば、上述されたように、メタデータのハッシュを、その特定のＵＲＬについて受信された他のイベントのメタデータのハッシュとすばやく比較することができる。しかしながら、これは、全てのコンテンツ変更を格納するには望ましくないという意味で、不十分である可能性がある。
【００３１】
ブログ投稿を一例にとると、わかりやすいであろう。もし、ブログ内の新しい投稿に対応するイベントのみを格納することが目的であるならば、受信されたイベントが新しい投稿に対応するか否か、または例えばスポーツの最新の得点など、ウェブページに組み込まれた何らかの外来情報に対応するか否かを決定可能であることが重要である。ブログのパブリッシングツールは、新しいｈｔｍｌの横にメタデータフィード（例えばＲＳＳフィードまたはＡｔｏｍフィード）を作成するのが通常である。発見的手法は、イベントが新しい投稿に対応するか否かを決定するために、これらのフィード（例えばリンクタグの代替物をサイトマップとして使用している）を参照することができる。これは、例えば、このフィード内で識別されるパーマリンクを参照することによって実現可能である。パーマリンクは、コンテンツに関連付けられた永続的なリンクであり、例えば新しいサイトのホームページなどの特定のＵＲＬにもうコンテンツが含まれていなくてもそのコンテンツを見つけることができるリンクである。
【００３２】
イベントメタデータが生成または検索され（２１２）、イベントがまだデータベースに格納されていないと決定され（２１４）ると、そのイベントは、再びバス１１０に乗せられる（２１６）。バス上には様々なデータレセプタ１１６（１－Ｎ）が配されている。これらは、例えばブログ投稿などの特定タイプのイベントをフィルタに掛けて検出するように、そして、認識された各イベントのメタデータを１つまたは複数のデータベースに格納することを促進するように構成される（２２０）。
【００３３】
具体的な一実装形態によると、各データレセプタは、それぞれ特定のデータベースへの

イベントの格納を促進するように構成される。第１のレセプタセット１１６－１は、本明細書においてコスモスデータベース（コスモスＤＢ）と称され且つシステムによって「時間の始まり」から記録されてきた全イベントについてのメタデータを含むデータベースへの、イベントの格納を促進するように構成される。つまり、コスモスＤＢは、エコシステム１００に関連付けられたデータ世界の「真の姿」を表すシステムのデータウェアハウスである。エコシステム１００の他の全てのデータベースは、このデータウェアハウスから派生する、またはこのデータハウスからデータの追加を受けることが可能である。

【００３４】

別のレセプタセット１１６－２は、時間順のデータベースＯＢＴ．ｄｂ　１０６－２へのイベントの格納を促進する。具体的な一実施形態によると、このデータベース内の情報は、マシンごとに一定の量ずつ順次格納されている。つまり、ある１つのマシンに一定量（例えば１日など一定期間、または例えば４ＧＢのＲＡＭベースのインデックスなど一定格納量におおよそ対応する）が格納されると、時間順ＤＢにデータ入力しているデータレセプタは、次のマシンに移動する。これは、日付と時間とに基づいた効率良い情報検索を可能にする。例えば、ユーザは、特定の日付に誰（または特定の人物）が何を話していたか、または所定の期間内に世界でどのような大事件が起きていたかを知りたいと考える可能性がある。

【００３５】

別のデータレセプタセット１１６－３は、権威順のデータベースＯＢＡ．ｄｂ　１０６－３へのイベントデータの格納を促進する。具体的な一実施形態によると、このデータベース内の情報は、個人ごとにインデックス付けされ、各個人の権威、すなわち影響力に応じて並べられる。権威は、例えば個人のブログにリンクしている人の数など、各個人にリンクしている人の数によって決定することが可能である。各個人に対するリンク数の変動とともに、権威順ＤＢ内における順番も変化する。このような手法は、権威順ＤＢを複数のセグメントに区切り、それぞれを各マシンに対応させることによって、最も効率的な情報検索を実現することができる。例えば、権威ある個人に対応する情報は、高速にアクセス可能な小さいデータベースセグメントに格納されてよく、一方で、ほとんどリンクされていない個人に関する情報は、より大きくより低速なデータベースセグメントに格納されてよい。

【００３６】

権威は、個人が書いている特定のカテゴリまたは主題に基づいて決定およびインデックス付けすることも可能である。例えば、もしある個人が主に米国の選挙制度について書いているとすると、この個人の権威は、他の人がどれだけリンクしているかだけでなく、自身を政治論評家としている人がどれだけリンクしているかにも基づいて決定することができる。また、権威の決定は、リンクしている個人の権威レベルを使用して更に改良することも可能である。一部の実施形態によると、特定の個人の権威レベルに関わるカテゴリまたは主題は、必ずしも、その個人によって明示されたカテゴリまたは主題によって限定される、すなわち決定されるとは限らない。つまり、例えば、もしある個人が自身を政治分野のブロガーと見なし、しかし主にスポーツについて書いているならば、その個人はスポーツに分類されると考えられる。これは、例えばキーワード、またはリンク先（例えばESPN.comへのリンク）、またはその両方など、その個人の投稿の内容を参照して決定することができる。

【００３７】

更に別のデータレセプタセット１１６－４は、キーワード順のデータベースＯＢＫ．ｄｂ　１０６－４へのイベントデータの格納を促進する。これらのデータレセプタは、イベントメタデータ内のキーワードを取り出して、定期的（例えば１分ごと）に構築される増分式キーワードインデックスのために使用する。具体的な一実施形態によると、これらのデータレセプタは、Ｌｕｃｅｎｅ（テキストのインデックス付けおよび検索のためのオープンソースＪａｖａツールキット(Ｊａｖａは登録商標です。)）に基づくもので、高速で且つほぼリアルタイムのキーワードインデックス付けを可能にするように調整されてい

る。従来のキーワードインデクサの大部分は、インデックスの作成に数日または数週間を費やす可能性がある。つまり、従来のキーワードインデクサは、データセットを作成し、そのデータセット全体をインデックス付けした後、そのデータセット全体を記録する。これに対して、本発明で使用されるキーワードインデクサは、キーワードインデックスを増分的に構築する。

【００３８】

具体的な一実施形態によると、キーワード検索が極めて並列的に実施可能であるという事実が活用される。新しいインデックス情報の極薄「スライス」が、既存のインデックスの上に「積層」され、時間の経過とともにメインインデックスに組み込まれる。したがって、例えば、キーワードデータレセプタは、先立つ１分間にインデックス付けされた情報を、１分ごとに既存のインデックスの上に追加する。これらの１分間スライスが何枚か、例えば５枚蓄積されると、これらのスライスは合体されて１つの５分間スライスになる。これは、何枚か（例えば４枚）の５分間スライスが蓄積されるまで繰り返され、次いで、蓄積された５分間スライスが合体されて１つの２０分間スライスになる。より厚いスライスを順に形成していくこのような合体は、基礎をなす元のインデックスの大きさのスライスが形成されるまで繰り返され、今度は、基礎をなす元のインデックスとの合体がなされる。この手法は、ウェブ上またはインターネット上に情報が投稿されてから文字通り数分以内、ひいては数秒以内に、その情報に対する構造的クエリを可能にする。なお、本段落で言及されているキーワードインデックス付けの手法は、例示を目的としたものに過ぎず、上述された増分式インデックス作成技術に限定するものでは無いことに注意すべきである。反対に、上述された増分式インデックス作成技術は、新しいインデックス情報をあらゆるタイプのインデックスに組み入れるために使用することができることが理解されるべきである。

【００３９】

エコシステムの各メインデータベース（すなわち、コスモスＤＢ、時間順ＤＢ、権威順ＤＢ、およびキーワード順ＤＢ）は、実質的に重複した情報セットを含む。しかしながら、各データベースは、応答時間に対応してどのように情報をインデックス付けかにおいて、他のデータベースとそれぞれ異なっている。

【００４０】

例えばＭＰ３のタイトルなど、何らかの任意のインデックスの順に並べられる新しいデータベースが作成されるときは、新しいデータベースへのイベントのインデックス付けを促進するように、新しいデータレセプタが構成される。新しいデータベースは、前述のように、最初は、コスモスＤＢ内の情報、すなわち「時間の始まり」までさかのぼるＭＰ３に関する情報を使用して構築することができる。なお、何をインデックスとして用いるか次第では、コスモスＤＢ内に表示される情報の世界全体を含まないデータベースも存在する。

【００４１】

データベースレセプタによって、特定のデータベースのための新しいスライスが生成されるとともに、これらのスライスは、エコシステムの各データベース（例えば時間順ＤＢや権威順ＤＢ）のマスタデータベースにコピーされる。後ほど詳述されるように、各マスタデータベースには、それぞれ幾つかのスレーブデータベースコピーも関連付けられている。これらのスレーブは、マスタデータベースと同様に更新され、検索クエリに対して応答をサービスする。つまり、各データベースのスレーブは、１つまたは複数のクエリサービス１１８によってアクセスされ、これら１つまたは複数のクエリサービス１１８には、その特定のデータベースに適したクエリを探索して提示するクエリインターフェース１２０に関連付けられている。具体的な実施形態によると、各スレーブは、データベースのコピー全体をシステムのＲＡＭ内に維持しているので、長期記憶装置内のデータベースは、少なくとも実行時間中は、書き込み専用である。したがって、データベースの読み出しに長期記憶装置へのアクセスが必要である場合と比べて、より速くクエリにサービスすることが可能になる。もちろん、この最適化は、本発明を実装するために必ずしも必要ではな

い。例えば、他の実施形態によれば、マスタデータベースの各セグメントをそれぞれ異なるスレーブに存在させることが可能である。一実施例では、一群のスレーブの各スレーブに、ブログおよびニュースサイトからの一週間分に値する投稿および記事を格納することができる。なお、各スレーブにどのようにデータを格納するまたは区分けするかは、本発明の範囲から逸脱することなく様々に可変である。

【００４２】

データベース内にインデックス付けされたメタデータは、ユーザ１２２によるクエリにサービスするクエリサービスによって利用可能になる。通常の検索エンジンで用いられる手法と異なり、このプロセスに必要な時間は、一般に、１分未満である。つまり、ウェブ上に投稿された変更は、その投稿から１分以内に、クエリサービス１１８を介して利用可能になる。したがって、本発明の実施形態によれば、あらゆる主題に関する対話を実質的にリアルタイムで追跡することが可能になる。

【００４３】

一部の実施形態にしたがって、クエリサービスとデータベースの間に、キャッシュサブシステム１２４（クエリサービスの一部を構成する、またはクエリサービスに関連付けられることが可能である）が用意される。キャッシュサブシステムは、データベースより小さい高速なメモリに格納され、特定の情報を求める要求内のスパイクをシステムによって取り扱うことを可能にする。情報は、様々な周知の技術のうちの任意の技術にしたがってキャッシュサブシステムに格納することができるが、エコシステムのリアルタイム特性ゆえに、情報がキャッシュ内に留まることができる時間は、例えば数分程度など比較的短期間に限定されることが望ましい。具体的な一実施形態によると、キャッシュサブシステムは、周知のオープンソースソフトウェア、Ｍｅｍｃａｃｈｅｄに基づく。情報は、削除される時間を示す有効期限、または「ダーティ」としてマーク付けされる時間を示す有効期限付きで、キャッシュに挿入される。キャッシュは、満杯になると、例えば「最長未使用時間」（ＬＲＵ）アルゴリズムなど、様々な周知の技術のうちの任意の技術にしたがって動作し、どの情報を削除するべきかを決定する。

【００４４】

本発明によるエコシステムは、データがどのように集約され検索可能にされるかについての根本的なパラダイムシフトを表している。単純にデータベースの片側からデータを挿入して反対側から取り出す従来のパラダイムの代わりに、インターネット上およびウェブ上のデータの世界は、情報「ストリーム」として概念化し監視することができる。特定の情報ストリームを単に探索し獲得するだけの、非常に単純で且つ非常に高速なアプリケーション（例えばイベントリスナおよびデータレセプタ）が構築される。得られた情報ストリームは、ほぼリアルタイムでインデックス付けされ、格納され、検索可能にされる。これらのアプリケーションは、全て並列的に動作するので、どの所定の「ストリーム」に関する情報であろうと、利用可能にされる前に先ず何らかの大きなデータウェアハウスから取り出す必要がなくなる。

【００４５】

様々な実施形態によると、上述されたイベントリスナおよびデータレセプタは、例えばＬｉｎｕｘ、Ａｐａｃｈｅ、ＭｙＳＱＬ、Ｐｙｔｈｏｎ、Ｐｅｒｌ、ＰＨＰ、Ｊａｖａ（登録商標）、およびＬｕｃｅｎｅを含む、様々なオープンソースソフトウェアおよびプロプライエタリソフトウェアから構築可能である。具体的な一実施形態によると、メッセージバスは、Ｓｐｒｅａｄとして知られるオープンソースソフトウェアに基づく。Ｓｐｒｅａｄは、外部ネットワークまたは内部ネットワークにおける障害に対して高い抵抗力を持つハイパフォーマンスのメッセージサービスを提供するツールキットである。Ｓｐｒｅａｄは、分散アプリケーションのための統一メッセージバスとして機能し、高度に調整されたアプリケーションレベルマルチキャストおよびグループ通信のサポートを提供する。

【００４６】

様々な具体的な実施形態によると、トラッキングサイト１０４によって蓄積された情報へのアクセスは、様々な方式で提供することができる。エコシステム内にインデックス付

けされた関心のある情報をユーザが獲得できるようにするために、多岐にわたる様々なメカニズムが使用可能である。例えば、ユーザがキーワード、フレーズ、ＵＲＬ等などを入力できる、テキストボックスを含む従来の検索インターフェースを使用することができる。例えばブール式の構築を可能にするなどの、より高度な検索ツールを用意することもできる。

【００４７】

使用される検索インターフェースの種類に寄らず、エコシステム内の各データベース（例えばコスモスＤＢ、時間順ＤＢ、権威順ＤＢ、キーワード順ＤＢ等など）に対応するクエリサービス１１８は、（クエリインターフェース１２０を介して）入ってくる検索クエリを観察し、例えばそのクエリのテキストがスペースやドット（例えばドットコム（.com））などを含むかなどのクエリの構文構造または意味構造を参照し、例えばキーワード検索なのかＵＲＬ検索なのかのようにクエリのタイプを決定する。サービス指向型アーキテクチャ（ＳＯＡ）を使用する実施形態によると、これらのクエリサービスは、実質的にリアルタイムで、ステートレスにクエリを処理するために、アーキテクチャ内に配される。

【００４８】

自身のデータベースに対応する検索クエリを認識したクエリサービスは、任意の適切な負荷平衡化スキームにしたがって、または／およびデータがスレーブ間でどのように編成されているかにしたがって、そのデータベースの１つまたは複数のスレーブにクエリを提示する。例えば、各スレーブに特定の週の投稿または記事が格納されている上記の例において、特定の主題に関する最新２０件の投稿を求めるクエリが認識された場合は、時間順データベースに関連付けられたクエリサービスが、同データベースに関連付けられたスレーブのうち最新の週に対応する幾つかのスレーブに接続する。同様に、ニューヨークタイムズの特定の記事に言及している最も権威ある２０件のブログ投稿を求めるクエリが認識された場合は、権威順データベースに関連付けられたクエリサービスが、同データベースに関連付けられた幾つかのスレーブに接続する。もし、クエリサービスが最初に接続したスレーブによってクエリが満足された場合は、これ以上他のスレーブを調べる必要はない。他方で、もし、要求された数の結果が第１群のスレーブから返されなかった場合は、クエリサービスは、更に他のスレーブに接続する必要がある。

【００４９】

特定の主題または問題に関する対話を識別するには、キーワード検索が使用可能である。「コスモス」検索は、リンク関係の識別を可能にすると考えられる。この能力を使用すれば、例えば、ブロガーは、誰が自分のブログにリンクしているかを見つけ出すことができる。この能力は、ブログの集約的特性を考えるときに、特に強力であると考えられる。

【００５０】

つまり、ブロガーの集団コミュニティは、本質的に、ウェブ上の情報世界に対して非常に大きな協調フィルタリングとして機能する。ブロガーによって作成されるリンクは、特定の情報に対する関連性、重要性、またはその両方に票を投じることに相当する。そして、ブログの半構造化特性は、関連情報を獲得してインデックス付けするための系統的なアプローチを可能にする。この協調的プロセスによってもたらされる、情報の関連部分に対する系統的で且つ適時なアクセスによって、ユーザは、関心ある物事に関連して現存する秩序体系を識別することが可能になる。

【００５１】

特定のコンテンツへのリンクを追跡可能であることによって、本発明の実施形態は、２つの重要な統計情報へのアクセスを可能にする。第１に、多数の人々が対話している主題を識別することができる。そして、この情報の取得およびインデックス付けの適時性は、これらの対話に、その主題に関する「市場」または「経済」の現在の状態を確実に反映させる。第２に、特定の主題に関する「権威者」または「影響者」と見なされうる作者を、これらの作者によって生成されたコンテンツにリンクしている人の数を追跡することによって識別することができる。

【００５２】

また、本発明の実施形態は、特定の個人がリンクしているまたは書いている主題を経時的に追跡するように動作可能である。つまり、ある文書群の作成者のプロファイルを経時的に作成し、それをその人物の嗜好および関心の表れとして使用することができる。これらのカテゴリにしたがって個人をインデックス付けすることによって、特定の主題に関する権威または影響力として特定の個人を識別することが可能になる。つまり、例えば、特定の個人がデジタル音楽プレーヤに関する多量のコンテンツを投稿している場合は、デジタル音楽プレーヤに関するその個人の権威（または影響力）レベルは、やはりデジタル音楽プレーヤに関心を持っているまたはその権威である他の個人（投稿およびリンクを通して追跡される）がどれだけ前者の個人にリンクしているかを識別することによって決定することができる。これは、オントロジー内のあらゆるトピックについて、各作者の相対的権威レベルをそのカテゴリ内の文書を作成している他の作者からのインバウンドリンクの数に基づいて豊富に且つ詳細に分析することを可能にする。

【００５３】

そして、エコシステムは、投稿、リンク、およびフレーズなどの各コンテンツがいつ作成されたかを「理解している」ので、この情報を、任意のデータ解析への追加入力として使用することができる。例えば、ある文書（またはその文書を作成した作者）の持つ影響力に対する理解を高めるために、時間を使用して、ある文書群に対するインバウンドリンクのパターンを観察すれば、ある文書に対して誰のリンクが早かったか、または遅かったかを素早く決定することができる。もし、ある人物が、関心を引く文書に対して常に早くリンクするならば、その人物は、その分野の専門家である、または少なくともその分野について専門的に発言できる可能性が最も高いと考えられる。

【００５４】

特定の主題における権威の識別および追跡は、従来の検索エンジンの手法では不可能な一部の機能を可能である。例えば、検索エンジンによってインデックス付けされる場合の新しい文書は、それが新しく、まだ誰もリンクしていないため、その関連性は全く不確定である。これに対して、本発明の実施形態は、所定の主題領域における特定の作者の影響力を追跡するので、その作者からの新しい投稿を、その作者の影響力に基づいて即時に記録することができる。つまり、文書作成における時間および性格に関して新しく得られた理解を使用すれば、まだ広くリンクされていない新しい文書でも、即時に記録することができる。なぜなら、（ａ）新しい文書または更新された文書に含まれるものがわかっており、したがって、分類法を使用してそのトピックを決定することができるから、そして、（ｂ）上述されたトピック領域における作者の相対的権威レベルがわかっているからである。したがって、従来の検索エンジンと異なり、本発明は、関連コンテンツの大部分に対して実質的に即時のアクセスを提供することができる。

【００５５】

また、本発明による技術は、特定の主題領域で最も影響力のある作者らによって現在議論されている、その主題領域内のサブトピックを追跡するために使用することができる。例えば、特定の主題領域で最も影響力のある１０人の作者によって現在議論されているトピックについて、データベースに対して照会することができる。

【００５６】

前述のように、個人（とりわけ権威ある個人）によるコンテンツの投稿および同コンテンツへのリンクを経時的に追跡すると、結果として、任意の主題またはトピックに関する協調フィルタリング効果を得られる。したがって、ウェブ上で利用可能なニュースソースのうちどれが現在重要であるかを編集者の選択に基づいて通知する代わりに、本発明によって可能になる協調フィルタリングを使用すれば、何が重要であるか、そしてそれは何故であるかについて様々な広い視点を提供することができる。

【００５７】

例えば、本発明は、ブロガーらが主要なニュースサイトのどの記事に現在リンクしているかを追跡するために使用することができる。つまり、トラッキングサイトによって取得されたデータをどのように編成するかによって、主題または作者による検索（すなわち「

ディープ」検索）だけでなく、時間による検索（すなわち「ワイド」検索）も可能になる。したがって、例えば、一部または全部のブロガーがコンテンツを投稿している最多リンク数のニュース記事（および／または書籍、映画等など）を識別するために、過去３時間の（ひいては、移動してゆく時間窓内の）ブログ投稿を全て評価することができる。この情報は、次いで、ブログのコミュニティによって現在最も重要だと見なされているトピックとして、ウェブページ上に公表される。そして、ウェブのグローバル性を考慮して、その重要なトピックの展開を、移動してゆく時間窓を使用して地球の回転とともに観察することができる。一部または全部のブロガーがリンクしている特定のニュース記事（および／または書籍、映画等など）の識別およびランク付けを向上させるため、移動する時間窓は、例えば１２時間など（または２４時間、４８時間、７２時間、７日間等など）に任意に拡張することが可能である。

【００５８】

様々な実施形態によると、本発明にしたがって収集されたデータに基づいて、様々なサービスが提供可能である。例えば、主要なニュースサービスに、そのサイトにリンクしている個人からなるコミュニティがそのニュースサービスおよびそのニュースサービスによって投稿された特定の記事について何を発言しているかを提供することができる。また、この「関心あるコミュニティ」の他の特徴に関する情報も、ニュースサービスに提供することができる。つまり、もし、このニュースサービスが明らかに上記のコミュニティの個人の注目を集めている場合は、このコミュニティが他に何について話しているかを識別すると有意義であると考えられる。これは、ある意味、ニュースサービスの編集上の決定に、ほぼ即時的な専用のフォーカスグループがあることに喩えられる。この情報は、ニュースサービスに配信され、例えば、記事の着想を得るため（すなわち、これが我々の読者が関心を持っているものだ）、書き出しを採用するため（すなわち、我々の読者の多くがカンザス州トピーカのコラムニストにリンクしている）、またはひいては、ウェブ上で何らかの形で直接公表するため（すなわち、我々の読者コミュニティはこのようなことを発言している）などを含む、多岐にわたる様々な方法で使用することができる。なお、このようなデータセットをもとに関心ある情報を提供するために、様々な洗練されたデータ解析技術が使用可能である。

【００５９】

明らかなように、このような関心あるコミュニティは、あらゆるウェブサイトについて識別することができる。実際、様々なウェブサイト、公開内容、および主題領域などごとにそれぞれ異なる関心あるコミュニティを識別し、（例えばウェブサイト上に）公表することによって、ユーザは、例えばスポーツニュース、技術ニュース、右翼政治に関するニュース、左翼政治に関するニュース等など任意の特定の公開内容またはトピックについて、何が話されているかを摂取することが可能になる。

【００６０】

したがって、時間および個人に関する理解を通して、本発明の実施形態は、カテゴリ化と権威、そして特定のカテゴリ内の権威を見分けることが可能である。そして、この情報のデータ解析は、様々なメトリックを「軸にする」ことができるので、「ディープ」検索および「ワイド」検索の両方を実現することによって、従来の検索技術の機能では及ばなかった様々な関心ある情報を得ることが可能になる。

【００６１】

更に、本明細書で説明される、エコシステムを使用した集約および検索の方法は、多岐にわたる様々な状況に適用可能である。例えば、エコシステムは、インターネット上における個人による物の販売を追跡するように実装することができる。したがって、例えば、もしある個人がオークションサイトに出品した場合は、このイベントは、クローラの伝送を生じさせるピングまたは他の通知メカニズムの生成をトリガすることができる。クローラは、オークションサイトのこの新しい出品を、上述されたのと類似の方式で構文解析し、インデックス付けし、記録する。別の一実施例は、大型の小売サイトにおける最新刊の発売である。実際、ウェブ上またはインターネット上で公開されるいかなるタイプのコン

テンツも、この方法でインデックス付けし記録することができる。別の一実施例は、ＰＲ　Ｎｅｗｓｗｉｒｅにおけるプレスリリースの公開である。

【００６２】

明らかにように、このような公開の適時な獲得は、様々な付加サービスを可能にする。例えば、デジタル音楽プレーヤの市場に大きな影響力を持つ人物を、容易に識別することができるので、このような人物は、広告協力者になって、自身のサイトに、訪問者に合わせて選ばれた特定タイプのイベント（例えばオークションにおけるデジタル音楽プレーヤの出品など）の通知を投稿しようと考える可能性がある。このような個人は、また、自身の関心分野または専門分野に関連する公開のイベントを通知する「生の」フィードを、パブリッシャおよびサイトから得ようと考える可能性もある。

【００６３】

同様に、企業は、企業ニュースをＰＲ　Ｎｅｗｓｗｉｒｅに転送する代わりに、自身のサイトにそのニュースを投稿し、トラッキングサイトに対してピングを打つ、すなわち変更情報を送信することができる。すると、トラッキングサイトは、様々な任意の使用方法に備えてその情報を取得し、インデックス付けし、記録する。例えば、特定のトピックに関連する投稿を知らせてくれるフィルタに、個人が加入（サブスクライブ）することができる。

【００６４】

別の一実施例では、トラッキングサイトは、雇用関係のサイトに新しい履歴が投稿されたときに通知を受けることができる。その履歴は、適切なフィルタに登録している雇用主が、投稿されたその履歴が自分の基準を満たしているか否かについて通知を受けることができるように、インデックス付けされ記録される。履歴情報の構文解析を容易にするため、履歴は、標準化フォーマットを有することができ、例えば、テンプレートされたＸＭＬ文書であってよい。このアプローチは、また、コンテンツを公開した人、すなわち求職者が自身のデータに対してある程度の管理力を保持できるようにする。つまり、例えば履歴などのコンテンツは、そのコンテンツの作成者のサイト上で公開されるのが通常であるので、コンテンツの作成者は、情報の取り下げを含む編集権を行使し続けることができる。

【００６５】

明らかなように、本発明によるイベント駆動式のエコシステムは、ワールドワイドウェブ（ＷＷＷ）に対して従来の検索技術とは異なる見方を有している。つまり、本明細書で説明されるデータ集約および検索のアプローチは、適時性（例えば２週間ではなく２分間）と、時間（すなわちいつ作成されたか）と、人および対話（すなわち文書ではなく）とを理解している。したがって、本発明によるエコシステムは、以前は不可能であった様々なアプリケーションを可能にする。例えば、本発明によるエコシステムは、ウェブ上の動的コンテンツに対する高度なソーシャルネットワーク分析を可能にする。エコシステムは、「何が」発言されているかだけでなく、「誰が」発言しているか、およびそれは「いつ」であるかも追跡することができる。このようなアプローチを使用すれば、ウェブ上で何かを最初に明言したのが誰であるかを識別することができる。また、発想がウェブ上をどのように伝搬するのか、そして、影響力を持つ、権威である、もしくは評判がよいのは誰であるかを（例えば何人の人がこの人物にリンクしているかによって）決定することができる。また、特定の人物が「いつ」リンクされたかを決定することもできる。この種の情報は、これまで実施不可能であった多くの更なる分析を可能にするために使用することができる。

【００６６】

例えば、ブログ界は、多くの場合、最新の記事またはニュース報道に反応して特定のトピック（例えば、大統領の国家警備隊のスキャンダルや、ＭａｃＷｏｒｌｄ　Ｅｘｐｏ（マックの祭典）でのｉＰｏｄ　ｍｉｎｉの発表等など）をめぐる議論を「加熱」させる。つまり、多くのブロガーは、主だったメディアにおけるニュースの発表に反応してそのトピックについて「対話」し始める。本発明は、これらの対話の追跡を可能にするだけでなく、そのニュースの発表「前」にそのトピックについて話していた個人の識別も可能にす

る。明らかなように、このような「対話を始めた人」、または特定のトピックに影響力のある人を識別する能力は、多くの観点からみて非常に有益である。

【００６７】

他の実施形態によると、本発明によるエコシステムは、ピーアール（ＰＲ）活動のために投資収益率（ＲＯＩ）を有意義に追跡することを可能にする。このための従来の技術は、あまり有益な情報を提供できないという意味で非効果的である。例えば、あるアプローチは、例えば通常３０～９０日間など、ある期間にわたって企業によって言及された任意の記事を単にスクラップブックにまとめあげるだけである。この情報は、頻度以外に、ＰＲ費用が有意義に使用されたか否かを決定するために企業が容易に使用できる他の定性的情報または定量的情報を、ほとんど提供しない。実際、今まで、ＰＲ費用の有効性を決定するための、一貫性のある、または信頼性の高い技術は、実質的に存在してこなかった。

【００６８】

反対に、本発明によるエコシステムは、特定のマーケティングキャンペーンに関する対話をリアルタイムで追跡することを可能にする。このような対話は、例えば、そのキャンペーンについて誰が発言しているか、およびそれらの発言者がそのキャンペーンについて実際に何を話しているかを含む。したがって、企業は、自社製品に関する「クチコミ」を作るためのベストの方法を特定できるだけでなく、そのクチコミを追跡することもでき、そして、それを、動的コンテンツへの適時なアクセスを通じて、費やされるＰＲ費用に直接つなげることができる。

【００６９】

ＰＲクライシスも、また、本発明によるエコシステムを使用して追跡および管理することができる。例えば、もし、欠陥商品に関するニュース記事など企業の評判にダメージを与える可能性があるイベントが発生した場合は、その危機に対処する適切な方策を考え出すために、影響力ある個人が参加しているそのイベントに関する対話を追跡することが可能である。

【００７０】

報道発信源（例えばニュース配信組織）は、エコシステムアーキテクチャを様々な方式で活用することができる。例えば、エコシステムは、ニュースサイトにおいて、そのサイトの記事に人々がどのように反応しているかを理解するために使用することができる。つまり、このような報道発信源は、その公開システムにイベント通知機能を組み入れることによって、記事が公開されるたびに、エコシステムへとピングを打ち、上述されたようなインデックス付けを得ることができる。すると、報道発信源は、これらの記事に関して誰が対話しているか、これらの記事に誰がリンクしているか、そしてそれらの人々が何を発言しているかを知ることができる。

【００７１】

同様に、ニュースサイトのオペレータは、そのサイト上に公開された記事のうち、過去１２時間で最も人気のあった記事を、各記事へのリンクの数によって調べることができる。また、記事に関するこの「クチコミ」を、経時的に追跡する、または、競合サイトによる同じトピックに関する記事によって発生したクチコミと比較することができる。また、ピングの時間（記事の最初の投稿に相当する）がデータベースに保存されるので、「スクープ」保護に対する何らかの措置も保障される。

【００７２】

更に、ニュースサイトは、クチコミを追跡できるだけでなく、他の人々がその記事について何を発言しているかを読者が知ることができるように、例えばリアルタイムの「投書欄」のように、追跡された情報の一部をニュースサイト上の元の記事に組み入れることもできる。より一般的に言うと、データベースから得られるほぼリアルタイムの情報を表現したもの（例えば、グラフおよびチャートに組み込んだ表現、またはひいては生データの表現）を、様々なメディアを介して生で提示することが可能である。例えば、このような情報フィードは、特定のトピックに関連したテレビ番組に供給する、またはテレビ番組（例えば、ニュース、バラエティ、トークショー、タレントサーチなど）に対するリアルタ

イムのフィードバックとして供給することが可能である。

【００７３】

報道発信源は、また、ニュース記事のソースとして有用だと考えられる、または新しい従業員として採用する魅力があると考えられる、権威ある個人を識別するために、エコシステムデータベースを検索することができる。より一般的に言うと、データベースは、権威順に情報をインデックス付けしているので、理由の如何を問わず、任意の所定の主題領域で最も影響力のあるまたは最も権威ある人物を検索することが可能である。

【００７４】

図１のデータベース１０６は、本発明の範囲から逸脱しない範囲内で、特定の実装形態に適するように様々に構成可能である。また、明らかなように、多数の挿入、更新、および削除が多数の選択と同時に実施される１つまたは複数のリレーショナルデータベースを維持することは、とりわけ、データベースの大きさが増すという理由で困難である。

【００７５】

したがって、本発明の具体的な一実施形態によると、ＳＱＬデータベースを多重化することによって、シャード化されクラスタ化された拡張可能なデータベースシステム３００を作成する方法が提供される。ネットワーク化されたハードウェア上で実行される個々のＳＱＬ実装形態はリンクされており、データベースは断片化されて個々のクラスタ上にある。データベースの各断片は、データベース全体のなかの一シャード上で実行される。各データベースシャードは、複数のマシンからなる１つのクラスタで構成される。複数のマシンのうち、１つは読み書きマスタ３０２として機能し、残りはマスタのコピーを含有する読み出し専用スレーブ２０４として機能する。読み出し要求は、任意多数のスレーブ２０４に配信される。各シャードは、例えばブログＩＤなど単一の固有のインデックスにインデックス付けされ、そのインデックスに関連付けられた全ての情報を格納される。このアプローチは、全てのデータをＲＡＭ内に維持することを可能にするので、その結果、適度に速い応答時間が得られる。

【００７６】

特定のインデックスに関連する情報をどのシャードが含有しているかに関する情報を維持するために、独立したデータベース（ソースＤＢ　２０６）が使用される。各マルチプレクサ（ｍｕｘ　２０８）は、データベースの局所的なメモリ内コピーを含有しており、スキームのみを含み、データを含まない。このデータベースのメモリ内コピーは、マックスＤＢと称される。各マルチプレクサは、各シャード上のデータベースに接続しており、入ってくる各クエリを構文解析し、そのクエリの内容（選択／挿入／更新／削除等など）およびそのクエリがシャード化されたインデックスを軸にしているか否かに応じてクエリを多重化し、１つまたは複数のシャードへと送り出す。クエリが、シャード化されたインデックスを軸（ピボット）にしている場合は、マルチプレクサは、問題になっているインデックスに関するデータを含有するシャードのみと通信する。クエリが、シャード化されたインデックスを直接軸にしていない場合は、そのクエリは、全てのシャードに送信され、次いで、それらの応答が、マルチプレクサによって並列的に収集される。そして、データの最終的な照合が行われ、マックスＤＢ内に、元のクエリを表の定義として使用した新しい表が作成される。そして、シャード化された並列な各クエリの結果が挿入される。こうして、元のクエリは、大幅に小さい新しいメモリ内のテーブル上の実行になる。

【００７７】

マルチプレクサは、並列化によって大きな冗長システムを構成することもできる。また、もしクエリが時間切れになる場合は、マルチプレクサは、クラスタ内の不良データベースを除去または修復できるように、クエリの送信先のシャードにマークを付ける。最後に、クラスタ化されたインデックスの大きさの増大とともに、新しいシャードをシステムに追加することができ、システム全体に対する読み出しクエリの数の増加とともに、各シャードにより沢山のスレーブを追加することができる。これは、たとえシステム全体にかかる負荷が増大する場合でも、リアルタイム結果の配信を可能にし、たとえシステム内のデータ量が増大する場合でも、システム全体に対する入出力スループットの増大を可能にす

る。

【００７８】

一実装形態によると、グリッド内のコンピュータ、それらの役割、シャードＩＤ、および状態（オンライン、オフライン、ビジー、修復等など）のリストを維持するコンピュータＤＢが維持される。マルチプレクサは、この情報を、フェイルオーバーのために、そしてグリッドを高い信頼性で動的に拡張または縮小するために使用する。

【００７９】

以上では、発明の具体的な実施形態に基づいて、本発明を例示および説明してきた。しかしながら、当業者ならば明らかなように、開示された実施形態の形式および詳細は、本発明の範囲及び趣旨から逸脱することなしに変更可能である。また、本明細書では、様々な実施形態に基づいて、本発明の様々な利点、態様、および目的を議論してきた。しかしながら、本発明の範囲は、このような利点、態様、および目的への言及によって限定されることは望ましくなく、むしろ、添付された特許請求の範囲を参照して決定されることが望ましい。

【図面の簡単な説明】

【００８０】

【図１】本発明の具体的な一実施形態にしたがった、データ集約および検索のための代表的なシステムの簡易ネットワーク図である。

【図２】本発明の具体的な一実施形態にしたがった、ネットワーク環境におけるデータ集約のための技術を示したフローチャートである。

【図３】本発明の様々な実施形態で使用するための、データベースアーキテクチャの簡易ブロック図である。

【図１】

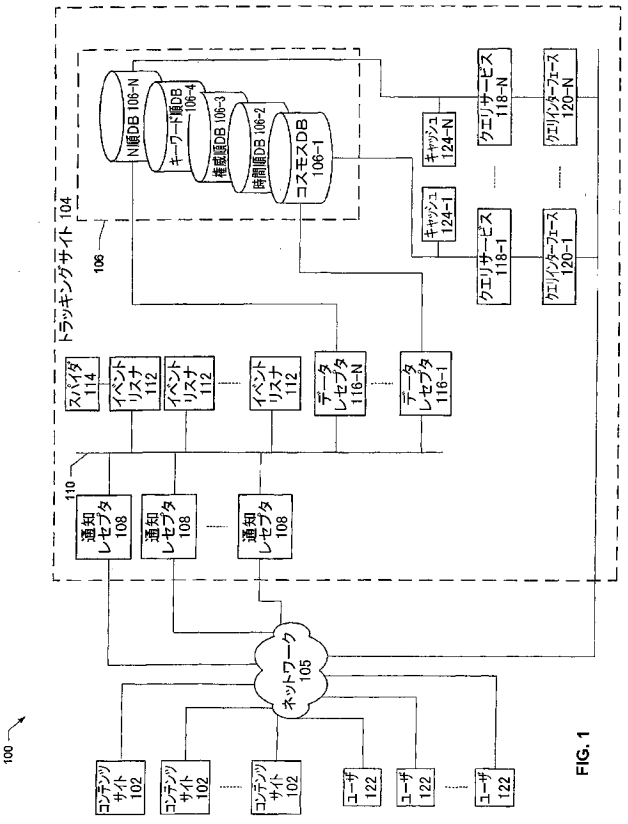


FIG. 1

【図２】

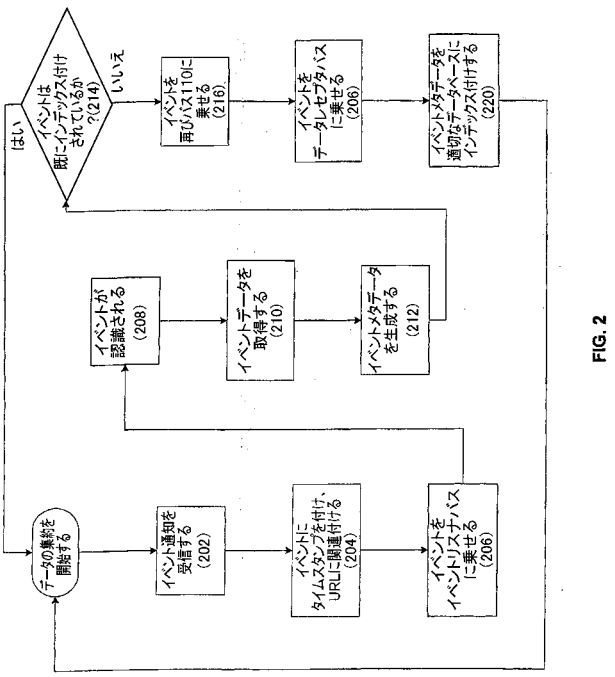


FIG. 2

【図３】

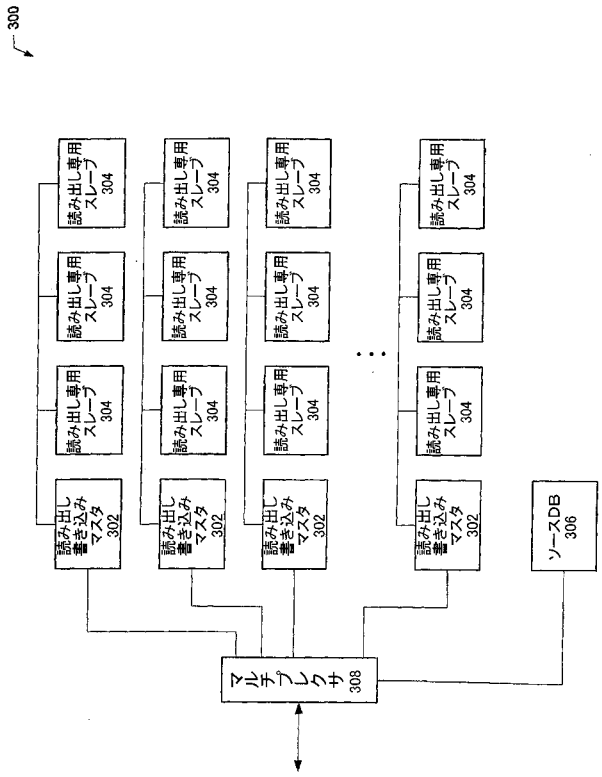

300
読み出し専用スレーブ 304
読み出し専用スレーブ 304
読み出し専用スレーブ 304
読み出し書き込みマスタ 302
読み出し専用スレーブ 304
読み出し専用スレーブ 304
読み出し専用スレーブ 304
読み出し書き込みマスタ 302
読み出し専用スレーブ 304
読み出し専用スレーブ 304
読み出し専用スレーブ 304
読み出し書き込みマスタ 302
読み出し専用スレーブ 304
読み出し専用スレーブ 304
読み出し専用スレーブ 304
読み出し書き込みマスタ 302
ソースDB 306
マルチプレクサ 308

FIG. 3

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LS,MW,MZ,NA,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IS,IT,LT,LU,MC,NL,PL,PT,RO,SE,SI,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BW,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,HR,HU,ID,IL,IN,IS,JP,KE,KG,KM,KP,KR,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,MG,MK,MN,MW,MX,MZ,NA,NG,NI,NO,NZ,OM,PG,PH,PL,PT,RO,RU,SC,SD,SE,SG,SK,SL,SM,SY,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,YU,ZA,ZM,ZW

（特許庁注：以下のものは登録商標）
１．Ｌｉｎｕｘ

【要約の続き】
スからほぼリアルタイムでコンテンツ変更を取得することが可能になる。
【選択図】図１