(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6960274号
(P6960274)
(45)発行日　令和3年11月5日(2021.11.5)　(24)登録日　令和3年10月13日(2021.10.13)

(51)Int.Cl.　F I
G06F 16/951 (2019.01)　G06F 16/951

請求項の数 9　(全 15 頁)

(21)出願番号　特願2017-160210(P2017-160210)
(22)出願日　平成29年8月23日(2017.8.23)
(65)公開番号　特開2019-40297(P2019-40297A)
(43)公開日　平成31年3月14日(2019.3.14)
審査請求日　令和2年3月5日(2020.3.5)

(73)特許権者　319013263
ヤフー株式会社
東京都千代田区紀尾井町１番３号
(74)代理人　100149548
弁理士　松沼　泰史
(74)代理人　100154852
弁理士　酒井　太一
(74)代理人　100181124
弁理士　沖田　壮男
(74)代理人　100194087
弁理士　渡辺　伸一
(72)発明者　川崎　将平
東京都千代田区紀尾井町１番３号　ヤフー株式会社内

最終頁に続く

(54)【発明の名称】データ収集装置、データ収集方法、およびプログラム

(57)【特許請求の範囲】
【請求項１】
ネットワークを介してアクセス可能な装置からデータを収集する収集部と、
前記収集部によって収集されたデータに含まれる、前記ネットワークを介してアクセス可能な装置に格納されたデータを参照するための参照情報が、所定の条件を満たす場合に、前記参照情報の一部を構成して複数の参照情報の群を特定する所属情報に対する前記収集部による収集を抑制する抑制部と
を備え、
前記所定の条件は、前記参照情報が、他の参照情報への転送を指示する情報であることである、
データ収集装置。
【請求項２】
前記所定の条件は、前記参照情報と対応するデータが、画像データまたは動画データであることである、
請求項１に記載のデータ収集装置。
【請求項３】
前記収集部により収集されたデータに基づき、前記所属情報に関するデータ収集の優先順位を決定する決定部をさらに備える、
請求項１または２に記載のデータ収集装置。
【請求項４】

前記決定部は、前記参照情報が、他の参照情報への転送を指示する情報である場合、前記他の参照情報と対応するデータに基づき、前記優先順位を決定する、
請求項３に記載のデータ収集装置。
【請求項５】
前記決定部により決定された前記所属情報に対するデータ収集の優先順位に基づき、データ収集が優先される所属情報のリストを生成する生成部をさらに備える、
請求項３に記載のデータ収集装置。
【請求項６】
前記決定部は、前記参照情報が、所定の条件を満たす場合に、前記所属情報に対するデータ収集の優先順位を下げる、
請求項３に記載のデータ収集装置。
【請求項７】
前記所属情報が、予め定義されたデータ収集が優先される所属情報のリストに含まれているか否かを判定し、前記所属情報が前記リストに含まれていると判定した場合、前記参照情報と対応するデータを前記収集部に収集させる判定部
をさらに備える、
請求項１から６のうちいずれか一項に記載のデータ収集装置。
【請求項８】
コンピュータが、
ネットワークを介してアクセス可能な装置からデータを収集し、
前記収集されたデータに含まれる、前記ネットワークを介してアクセス可能な装置に格納されたデータを参照するための参照情報が、所定の条件を満たす場合に、前記参照情報の一部を構成して複数の参照情報の群を特定する所属情報に対する収集を抑制する
データ収集方法であって、
前記所定の条件は、前記参照情報が、他の参照情報への転送を指示する情報であることである、
データ収集方法。
【請求項９】
コンピュータに、
ネットワークを介してアクセス可能な装置からデータを収集させ、
前記収集されたデータに含まれる、前記ネットワークを介してアクセス可能な装置に格納されたデータを参照するための参照情報が、所定の条件を満たす場合に、前記参照情報の一部を構成して複数の参照情報の群を特定する所属情報に対する収集を抑制させる
プログラムであって、
前記所定の条件は、前記参照情報が、他の参照情報への転送を指示する情報であることである、
プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データ収集装置、データ収集方法、およびプログラムに関する。
【背景技術】
【０００２】
従来、ウェブから文章や画像等のデータを収集し、収集したデータを自動的にデータベース化するクローラが知られている（例えば、特許文献１および２参照）。クローラは、ウェブページ中のリンクを辿って、様々なドメインのウェブページからデータを収集する。クローラによって収集されたデータは、ウェブ情報データベースに蓄積される。ウェブ情報データベースに蓄積されたデータは、ウェブページの検索サービス等に利用される。
【先行技術文献】
【特許文献】

【０００３】
【特許文献１】特開２０１２－６９１７１号公報
【特許文献２】特開平９－３２５９６８号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
検索サイト等の利便性向上のため、より多くのデータを効率的に収集することが期待されている。例えば、検索クエリに対して、単に検索クエリに対応するウェブページを提供するだけでなく、検索クエリに応じた回答を提供するシステムにおいては、より多くの有益なデータ（知識）を収集する必要がある。
【０００５】
しかしながら、クロールの対象となる膨大なウェブページの中から有益なデータを効率的に収集することは容易ではない。例えば、従来の幅優先探索（ＢＦＳ：Breadth First Search）を用いたクロールを行う場合、不要なデータを多く含む価値の低いホストについてもクロールの対象となってしまい、リソースが無駄に消費されている場合があった。一方で、有効なデータを多く含む有益なホストに対するクロールが進まず、データ収集に時間を要してしまう場合があった。
【０００６】
本発明は、このような事情を考慮してなされたものであり、データ収集の効率を向上させることができるデータ収集装置、データ収集方法、およびプログラムを提供することを目的の一つとする。
【課題を解決するための手段】
【０００７】
本発明の一態様は、ネットワークを介してアクセス可能な装置からデータを収集する収集部と、前記収集部によって収集されたデータに含まれる、前記ネットワークを介してアクセス可能な装置に格納されたデータを参照するための参照情報が、所定の条件を満たす場合に、前記参照情報の一部を構成して複数の参照情報の群を特定する所属情報に対する前記収集部による収集を抑制する抑制部とを備えるデータ収集装置である。
【発明の効果】
【０００８】
本発明の一態様によれば、データ収集の効率を向上させることができる。
【図面の簡単な説明】
【０００９】
【図１】実施形態のクロールサーバ１０の使用環境を示す図である。
【図２】実施形態のクロールサーバ１０の構成を示す機能ブロック図である。
【図３】実施形態のホストランク情報Ｄ１の一例を示す図である。
【図４】実施形態のコンテンツ第１情報Ｄ２の一例を示す図である。
【図５】実施形態のコンテンツ第２情報Ｄ３の一例を示す図である。
【図６】実施形態の注目ホストリストＤ４の一例を示す図である。
【図７】実施形態のホストランク決定部２０の構成を示す機能ブロック図である。
【図８】実施形態のクロール処理の流れの一例を示すフローチャートである。
【図９】実施形態のホスト選択処理の流れの一例を示すフローチャートである。
【図１０】実施形態のホストランク決定処理の流れの一例を示すフローチャートである。
【図１１】実施形態の注目ホストリスト生成処理の流れの一例を示すフローチャートである。
【発明を実施するための形態】
【００１０】
［概要］
以下、図面を参照して、データ収集装置、データ収集方法、およびプログラムの実施形態について説明する。本実施形態では、データ収集装置がクロールサーバであるものとし

て説明する。クロールサーバとは、インターネット等のネットワークを介してアクセス可能な複数の装置からデータを自動的に収集するサーバである。本実施形態のクロールサーバは、価値の高いデータを提供する有益なホストに集中してクロールを行う。クロールサーバは、１つのプロセッサによって実現されてもよく、複数のプロセッサが分散処理することで実現されてもよい。本実施形態において、ホストとは、ネットワークを介してアクセス可能な装置に格納されたデータを参照するための参照情報（例えば、ＵＲＬ（Uniform Resource Locator））の一部を構成して複数の参照情報の群を特定する所属情報をいう。以下、実施形態について説明する。

【００１１】

［全体構成］

図１は、本実施形態のクロールサーバ１０（データ収集装置）の使用環境を示す図である。クロールサーバ１０は、画像データおよびＨＴＭＬ（HyperText Markup Language）データの少なくとも一方を含むページデータ（コンテンツ）を、ネットワークＮＷを介してアクセス可能な複数の外部サーバＳ１（装置）から収集する。コンテンツは、外部サーバＳ１に格納されており、ブラウザによって閲覧可能なページ単位のデータである。ただし、コンテンツは、ブラウザに限らず、アプリケーションプログラムによって再生されるデータでもよい。ネットワークＮＷは、インターネットやＷＡＮ（Wide Area Network）、ＬＡＮ（Local Area Network）等を含む。

【００１２】

［クロールサーバの構成］

以下、クロールサーバ１０の構成について説明する。図２は、クロールサーバ１０の構成を示す機能ブロック図である。クロールサーバ１０は、例えば、データ収集部１２（収集部）と、解析部１４（判定部）と、バッチ処理部１６と、記憶部１８とを備える。バッチ処理部１６は、例えば、ホストランク決定部２０（抑制部、決定部）と、注目ホストリスト生成部２２（生成部）と、ホスト選択部２４とを備える。記憶部１８には、例えば、ホストランク情報Ｄ１、コンテンツ第１情報Ｄ２、コンテンツ第２情報Ｄ３、および注目ホストリストＤ４が記憶されている。

【００１３】

ホストランク情報Ｄ１には、データ収集の対象となるホストと、データ収集の優先順位を示す指標であるホストランクとが関連付けされたデータが含まれる。図３は、本実施形態のホストランク情報Ｄ１の一例を示す図である。このホストランク情報Ｄ１には、例えば、ホスト“ＡＡＡ”であり、ホストランク“３０”であるデータが含まれている。

【００１４】

コンテンツ第１情報Ｄ２には、ＵＲＬと、このＵＲＬの関連情報とが関連付けされたデータが含まれる。ＵＲＬの関連情報には、例えば、ＵＲＬに対応するコンテンツが取得済みであるか未取得であるかを示すステータス１、コンテンツの取得の成否（例えば、ＨＴＴＰステータスコード）を示すステータス２、およびデータ収集の優先度を示す指標であるスコアが含まれる。図４は、本実施形態のコンテンツ第１情報Ｄ２の一例を示す図である。このコンテンツ第１情報Ｄ２には、例えば、ＵＲＬが“ddd.ddd”であり、ステータス１が“取得済”であり、ステータス２が“３０１”であり、スコアが“５”であるデータが含まれている。尚、初期状態のコンテンツ第１情報Ｄ２には、クロールを開始するための基礎となる複数のＵＲＬが格納されている。また、初期状態のコンテンツ第１情報Ｄ２では、全てのＵＲＬに関して、ステータス１および２は「未取得」、スコアは「未付与」となっている。

【００１５】

コンテンツ第２情報Ｄ３には、ＵＲＬと、このＵＲＬと対応するコンテンツとが関連付けされたデータが含まれる。コンテンツには、ＨＴＭＬデータおよび画像データが含まれる。図５は、本実施形態のコンテンツ第２情報Ｄ３の一例を示す図である。このコンテンツ第２情報Ｄ３には、例えば、ＵＲＬが“aaa.aaa”であり、コンテンツが“ＨＴＭＬデータ１”であるデータが含まれている。

【００１６】
注目ホストリストＤ４には、データ収集の優先度が高いホストの一覧データが含まれる。図６は、本実施形態の注目ホストリストＤ４の一例を示す図である。この注目ホストリストＤ４には、例えば、データ収集の優先度が高いホストとして、“ＢＢＢ”、“ＥＥＥ”等が含まれている。
【００１７】
データ収集部１２（フェッチャー）は、複数の外部サーバＳ１からコンテンツを収集（フェッチ）する。収集されるコンテンツには、ＨＴＭＬデータおよび画像データが含まれる。データ収集部１２は、収集したコンテンツを記憶部１８のコンテンツ第１情報Ｄ２に記憶させる。
【００１８】
解析部１４（パーサー）は、データ収集部１２により収集されて記憶部１８に記憶されたＨＴＭＬデータを解析する。例えば、解析部１４は、ＨＴＭＬデータから、ヘッダ部分を除くテキストデータを抽出し、抽出したテキストデータのなかに、新しいＵＲＬが含まれているか否かを判定する。ここで、「新しいＵＲＬ」とは、記憶部１８に未登録のＵＲＬである。解析部１４は、抽出したテキストデータに新しいＵＲＬが含まれていると判定した場合、そのＵＲＬを新しいＵＲＬとして記憶部１８のコンテンツ第１情報Ｄ２（ステータス１および２は「未取得」、スコアは「未付与」）に記憶させる。
【００１９】
また、解析部１４は、注目ホストリストＤ４を参照し、上述の新しいＵＲＬに含まれるホストが注目ホストリストＤ４に含まれているか否かを判定する。解析部１４は、新しいＵＲＬに含まれるホストが注目ホストリストＤ４に含まれていると判定した場合、このホストが優先度の高い有益なホストであると判定する。そして、解析部１４は、この新しいＵＲＬを、データ収集部１２の収集対象のＵＲＬのリスト（キュー）に追加する。これにより、この新しいＵＲＬに対するデータ収集が行われる。
【００２０】
ホストランク決定部２０は、記憶部１８に記憶されたコンテンツを解析し、そのコンテンツに対応するＵＲＬが属するホストに対して、データ収集の優先順位を示す指標であるホストランクを決定する。ホストランク決定部２０は、日次、週次等、所定の時間間隔のバッチ処理によりホストランクを決定する。
【００２１】
ホストランク決定部２０は、コンテンツに予め定義された特定の情報が含まれる場合に、付与する優先度を高くする。例えば、ホストランク決定部２０は、コンテンツに、コンテンツの内容を示す特定の情報が含まれる場合に、付与する優先度を高くする。
【００２２】
図７は、本実施形態のホストランク決定部２０の構成を示す機能ブロック図である。図７に示すように、ホストランク決定部２０は、例えば、タグ情報検出部３０、語句検出部３２、スコア付与部３４、およびホストランク決定部３６を備える。
【００２３】
タグ情報検出部３０は、コンテンツのなかに、特定の情報として設定された特定のタグが含まれるか否かを検出する。「特定のタグ」は、例えば、ＯＧＰ（Open Graph Protocol）タグのようなコンテンツの内容を示すテキストを含むタグである。ＯＧＰタグは、リンク先を示すＵＲＬ、リンク先のコンテンツの言語、リンク先のウェブサイトの名前、リンク先のコンテンツのタイトル、リンク先のコンテンツに関する画像データのＵＲＬ、リンク先のコンテンツの概要を示すテキストデータ等がひと纏まりになった情報である。
【００２４】
タグ情報検出部３０は、コンテンツのなかに、ＯＧＰタグが含まれるか否かを検出する。例えば、タグ情報検出部３０は、コンテンツのなかにＯＧＰタグが含まれることを検出した場合、ＯＧＰタグのなかから、リンク先を示すＵＲＬ、リンク先のコンテンツのタイトル、リンク先のコンテンツに関する画像データのＵＲＬ、リンク先のコンテンツの概要

を示すテキストデータ等の情報を抽出する。また、タグ情報検出部３０は、コンテンツのなかにＯＧＰタグが含まれることを検出した場合、ＯＧＰタグが含まれることを示す情報と、ＯＧＰタグを含むデータに対応するＵＲＬとを対応付けてスコア付与部３４に入力する。

【００２５】

語句検出部３２は、コンテンツのなかに、特定の情報として設定された「特定の語句」が含まれるか否かを検出する。「特定の語句」は、ウェブページのメタタグに含まれる語句であって、コンテンツの内容を示すものとして予め登録された語句でもよい。例えば、語句検出部３２は、コンテンツに含まれるテキストデータに対して形態素解析を行い、予め登録された語句を検索することで、特定の語句が含まれるか否かを検出する。語句検出部３２は、検出対象の特定の語句を検出した場合、特定の語句が含まれることを示す情報と、その特定の語句を含むデータに対応するＵＲＬとを対応付けてスコア付与部３４に入力する。

【００２６】

スコア付与部３４は、タグ情報検出部３０による検出結果と、語句検出部３２による検出結果とに基づき、コンテンツに対応するＵＲＬに、データ収集の優先度を示すスコアを付与する。

【００２７】

本実施形態では、スコア付与部３４は、タグ情報検出部３０の検出結果に基づき、ＵＲＬに対して優先度として第１スコアを付与する。スコア付与部３４は、タグ情報検出部３０によってデータのなかに特定のタグが含まれることを検出した場合、データ収集の優先度が高くなるように第１スコアを高くする。また、本実施形態では、スコア付与部３４は、語句検出部３２の検出結果に基づき、コンテンツに対応するＵＲＬに対して優先度として第２スコアを付与する。スコア付与部３４は、語句検出部３２によってデータのなかに特定の語句が含まれることを検出した場合、データ収集の優先度が高くなるように第２スコアを高くする。

【００２８】

ホストランク決定部３６は、スコア付与部３４によって付与された第１スコアおよび第２スコアの少なくとも一方に基づいて、ホストに対して、データ収集の優先順位を設定する。例えば、ホストランク決定部３６は、ＵＲＬごとに第１スコアと第２スコアとの合計スコアを算出する。そして、ホストランク決定部３６は、ホストごとに、このホストに属する複数のＵＲＬにおける合計スコアの平均値を算出する。そして、ホストランク決定部３６は、この合計スコアの平均値が高い順に、ホストランクを決定する。

【００２９】

また、ホストランク決定部３６は、処理対象のホストに属するＵＲＬのなかで、所定のＵＲＬへの転送（リダイレクト）を指示するＵＲＬの割合が所定の閾値以上であると判定した場合、すなわち、処理対象のホストに属するＵＲＬの多くがリダイレクトを示すものであると判定した場合、ホストランクを所定の順位だけ下げる。これにより、ホストランク決定部３６は、コンテンツに対応するＵＲＬが所定の条件（ＵＲＬの多くがリダイレクトを示すものである）を満たす場合に、このＵＲＬが属するホストに対する収集を抑制する。リダイレクト用のＵＲＬであるか否かは、データ収集部１２によるデータ収集の際に取得したＨＴＴＰステータスコードが、リダイレクトを示す３００系であるか否かに基づいて判断される。

【００３０】

また、ホストランク決定部３６は、処理対象のホストに属するＵＲＬのなかで、コンテンツデリバリネットワーク（ＣＤＮ：Content Delivery Network）を用いて取得される画像データ、動画データ等を示すＵＲＬの割合が所定の閾値以上であると判定した場合、すなわち、処理対象のホストに属するＵＲＬの多くが画像データ等を示すものであると判定した場合、ホストランクを所定の順位だけ下げる。これにより、ホストランク決定部３６は、コンテンツに対応するＵＲＬが所定の条件（処理対象のホストに属するＵＲＬの多く

が画像データ等を示すものである）を満たす場合に、このＵＲＬが属するホストに対する収集を抑制する。画像用のＵＲＬであるか否かは、ＵＲＬの拡張子に基づいて判断される。画像用の拡張子には、例えば、“ jpg ”, “ png ”の拡張子が含まれる。

【００３１】

また、ホストランク決定部３６は、処理対象のホストに属するＵＲＬのなかで、データ収集部１２によるデータ収集の際に取得したＨＴＴＰステータスコードが、サーバエラーを示す５００系であると判定した場合や、認証エラーを示す４００系であると判定した場合に、ホストランクを所定の順位だけ下げるようにしてもよい。また、ホストランク決定部３６は、外部サーバＳ１からクロールを拒否する旨の情報を受け取っている場合（例えば、robot.txtに拒否ＵＲＬが指定されている場合）、この拒否ＵＲＬが属するホストを所定の順位だけ下げるようにしてもよい。

【００３２】

また、ホストランク決定部２０は、処理対象のコンテンツのテキストデータに含まれるＵＲＬのリンク先のコンテンツに基づいて、ホストランクを決定してもよい。例えば、ホストランク決定部２０は、処理対象のコンテンツのテキストデータに含まれるＵＲＬのリンク先のコンテンツを取得し、取得したコンテンツに対して上述の第１スコアに相当するスコア（以下、「第３スコア」という）および上述の第２スコアに相当するスコアを算出し（以下、「第４スコア」という）、第１から第４スコアに基づいて、ホストランクを決定する。

【００３３】

例えば、ホストランク決定部３６は、処理対象のコンテンツに対応するＵＲＬごとに、第１および第２スコアの合計スコアを算出する。さらに、ホストランク決定部３６は、処理対象のコンテンツに対応するＵＲＬごとに、この処理対象のコンテンツに含まれるＵＲＬのリンク先の第３および第４スコアの合計スコアの平均値を算出する。さらに、ホストランク決定部３６は、処理対象のコンテンツに対応するＵＲＬごとに、第１および第２スコアの合計スコアと、第３および第４スコアの合計スコアの平均値との２次合計スコアを算出する。そして、ホストランク決定部３６は、ホストごとに、このホストに属するＵＲＬの２次合計スコアの平均値を算出し、この２次合計スコアの平均値が高い順に、ホストランキングを決定する。

【００３４】

また、ホストランク決定部２０は、処理対象のコンテンツのテキストデータに含まれるＵＲＬがリダイレクトを示すＵＲＬである場合、リダイレクト先のＵＲＬが示すコンテンツに対して、上述の第３スコアおよび第４スコアを算出し、データ収集の優先順位を設定する。

【００３５】

注目ホストリスト生成部２２は、記憶部１８に記憶されたホストランク情報Ｄ１を参照し、複数のホストのなかから優先してデータを収集するホストを選出した注目ホストリストＤ４を生成する。例えば、注目ホストリスト生成部２２は、ホストランクが所定の順位以上のホスト（例えば、上位１００位）を注目ホストとして決定し、注目ホストリストＤ４に登録する。注目ホストリスト生成部２２は、日次、週次等、所定の時間間隔のバッチ処理により注目ホストを決定する。

【００３６】

ホスト選択部２４は、記憶部１８に記憶されたコンテンツ第１情報Ｄ２を参照し、未だコンテンツが取得されていないＵＲＬを含むホスト（ステータス１が“ 未取得 ”であるＵＲＬが属するホスト）を選択する。さらに、ホスト選択部２４は、選択したホストに属するＵＲＬのうち、コンテンツが未取得である少なくとも１つのＵＲＬをデータ収集部１２の収集対象のＵＲＬのリストに追加する。これにより、キューに追加されたＵＲＬを用いたデータ収集がデータ収集部１２により行われる。ホスト選択部２４は、日次、週次等、所定の時間間隔のバッチ処理によりホスト選択処理を行う。

【００３７】

クロールサーバ１０の構成要素は、例えば、コンピュータにおいて、ＣＰＵ（Central Processing Unit）等のハードウェアプロセッサがプログラム（ソフトウェア）を実行することにより実現される。また、これらの構成要素のうち一部または全部は、ＬＳＩ（Large Scale Integration）やＡＳＩＣ（Application Specific Integrated Circuit）、ＦＰＧＡ（Field-Programmable Gate Array）、ＧＰＵ（Graphics Processing Unit）等のハードウェア（回路部；circuitryを含む）によって実現されてもよいし、ソフトウェアとハードウェアの協働によって実現されてもよい。

【００３８】

クロールサーバ１０の記憶部１８は、例えば、ＲＡＭ（Random Access Memory）、ＲＯＭ（Read Only Memory）、ＨＤＤ（Hard Disk Drive）、フラッシュメモリ、またはこれらのうち複数が組み合わされたハイブリッド型記憶装置等により実現される。また、記憶部１８の一部または全部は、ＮＡＳや外部のストレージサーバ等、クロールサーバ１０がアクセス可能な外部装置であってもよい。

【００３９】

［クロール処理］

以下、クロールサーバ１０のクロール処理について説明する。図８は、本実施形態のクロール処理の流れの一例を示すフローチャートである。本フローチャートによる処理は、クロールサーバ１０によって一定時間以上に亘って継続的に繰り返し実行される。尚、本フローチャートは、１つのＵＲＬを起点として実施するクロール処理の流れを示す。

【００４０】

まず、データ収集部１２は、記憶部１８に記憶されたコンテンツ第１情報Ｄ２に含まれる複数のＵＲＬのなかから、データ収集に用いるＵＲＬを選出し、選出したＵＲＬを用いてコンテンツの格納先である外部サーバＳ１にアクセスし、コンテンツを収集する（Ｓ１０１）。例えば、データ収集部１２は、コンテンツ第１情報Ｄ２に含まれる複数のＵＲＬのなかから、クロールを開始するための基礎となるＵＲＬとして格納されたＵＲＬ（ステータス１および２が「未取得」、スコアが「未付与」）を選出する。データ収集部１２は、収集に用いたＵＲＬと、収集したコンテンツとを関連付けたデータを、記憶部１８に記憶されたコンテンツ第２情報Ｄ３に追加する。

【００４１】

次に、解析部１４は、記憶部１８に記憶されたコンテンツ第１情報Ｄ２から、データ収集部１２により新たに追加されたＵＲＬとコンテンツとの組を読み出し、読み出したコンテンツを解析する（Ｓ１０３）。ここで、解析部１４により読み出されるコンテンツは、ＨＴＭＬデータである。例えば、解析部１４は、読み出したＨＴＭＬデータから、ヘッダ部分を除くテキストデータを抽出し、抽出したテキストデータのなかに含まれるＵＲＬが、新しいＵＲＬであるか否かを判定する。解析部１４は、抽出したテキストデータに新しいＵＲＬが含まれる場合、そのＵＲＬを新しいＵＲＬ（ステータス１および２が「未取得」、スコアが「未付与」）としてコンテンツ第１情報Ｄ２に追加する（Ｓ１０５）。

【００４２】

次に、解析部１４は、注目ホストリストＤ４を参照し、上記の新しいＵＲＬが属するホストが注目ホストリストＤ４に含まれているか否かを判定する（Ｓ１０７）。解析部１４は、新しいＵＲＬが属するホストが注目ホストリストＤ４に含まれていると判定した場合、このホストが優先度の高い有益なホストであると判定する。そして、解析部１４は、この新しいＵＲＬを、データ収集部１２の収集対象のＵＲＬのリスト（キュー）に追加する（Ｓ１０９）。これにより、データ収集部１２は、キューに追加された新しいＵＲＬを用いて、再度データ収集を行い（Ｓ１０１）、解析部１４は、再度上述の解析処理を行う（Ｓ１０３～Ｓ１０９）。これにより、優先度の高い有益なホストに属するＵＲＬが新しく発見された場合、そのＵＲＬを用いてデータを迅速に収集することができる。

【００４３】

一方、解析部１４は、新しいＵＲＬに属するホストが注目ホストリストＤ４に含まれていないと判定した場合、このホストが優先度の低いホストであると判定する。この場合、

新たなＵＲＬを用いたデータ収集は行われず、本フローチャートの処理が終了する。

【００４４】

［ホスト選択処理］

以下、クロールサーバ１０のホスト選択処理について説明する。図９は、本実施形態のホスト選択処理の流れの一例を示すフローチャートである。本フローチャートによる処理は、日次、週次等、所定の時間間隔のバッチ処理として実行される。

【００４５】

まず、ホスト選択部２４は、記憶部１８に記憶されたコンテンツ第１情報Ｄ２を参照し、未だコンテンツが取得されていないＵＲＬを含むホスト（ステータス１が“未取得”であるＵＲＬが属するホスト）を選択する（Ｓ２０１）。ここで選択されるホストは、上述のクロール処理において、解析部１４により注目ホストリストＤ４に含まれていないと判定され、データ収集の対象とならなかったＵＲＬが属するホストを含む。

【００４６】

次に、ホスト選択部２４は、選択したホストに属するＵＲＬのうち、コンテンツ取得済みであるＵＲＬ（ステータス１が“取得済”であるＵＲＬ）の数が、所定の閾値以下であるか否かを判定する（Ｓ２０３）。例えば、ホストごとにフェッチしたＵＲＬの数（コンテンツ取得済みＵＲＬの数）を記憶部１８で管理しておき、ホスト選択部２４は、このコンテンツ取得済みＵＲＬの数を参照することで、上述の判定処理を行う。ホスト選択部２４は、コンテンツ取得済みであるＵＲＬの数が、閾値以下ではないと判定した場合、このホストに属するＵＲＬをデータ収集部１２の収集対象に設定しない。尚、ホスト選択部２４が、選択したホストに属するＵＲＬのうち、コンテンツ取得済みであるＵＲＬの割合と、所定の閾値とを比較するようにしてもよい。

【００４７】

一方、ホスト選択部２４は、コンテンツ取得済みであるＵＲＬの数が、閾値以下であると判定した場合、記憶部１８に記憶されたコンテンツ第１情報Ｄ２から、選択したホストに属するＵＲＬのうち、コンテンツが未取得である（ステータス１が“未取得”である）少なくとも１つのＵＲＬを取得する（Ｓ２０５）。次に、ホスト選択部２４は、取得したＵＲＬをデータ収集部１２の収集対象のＵＲＬのリスト（キュー）に追加する（Ｓ２０７）。これにより、キューに追加されたＵＲＬを用いたデータ収集がデータ収集部１２により行われる。

【００４８】

ホスト選択部２４は、取得したＵＲＬをデータ収集部１２の収集対象のＵＲＬのリストに追加した後、または上述の判定処理においてコンテンツ取得済みであるＵＲＬの数が閾値以下ではないと判定した場合、コンテンツ第１情報Ｄ２に含まれる全てのホストに対する処理が完了したか否かを判定する（Ｓ２０９）。ホスト選択部２４は、全てのホストに対する処理が完了していないと判定した場合、未処理のホストに対して上述のホスト選択処理（Ｓ２０１）以降の処理を繰り返す。一方、ホスト選択部２４は、全てのホストに対する処理が完了したと判定した場合、本フローチャートの処理を終了する。

【００４９】

尚、上記の実施形態においては、ホスト選択部２４が、未だコンテンツが取得されていないＵＲＬに関して、コンテンツの取得処理を行わせる例を説明した。しかしながら、ホスト選択部２４は、コンテンツが取得済みのＵＲＬに関して、再度、コンテンツの取得処理を行わせるようにしてもよい。これにより、コンテンツが更新された場合等に、コンテンツの最新のデータを取得することが可能である。

【００５０】

［ホストランク決定処理］

以下、クロールサーバ１０のホストランク決定処理について説明する。図１０は、本実施形態のホストランク決定処理の流れの一例を示すフローチャートである。本フローチャートによる処理は、日次、週次等、所定の時間間隔のバッチ処理として実行される。尚、本フローチャートは、１つのホストに対するホストランク決定処理の流れを示す。

【００５１】
まず、ホストランク決定部２０は、記憶部１８に記憶されたコンテンツ第１情報Ｄ２から、処理対象とするホストを選択し、選択したホストに属するＵＲＬを取得する（Ｓ３０１）。次に、ホストランク決定部２０は、取得したＵＲＬに対して上述したスコア付与を行う（Ｓ３０３）。次に、ホストランク決定部２０は、そのＵＲＬが属するホストのホストランクを決定する（Ｓ３０５）。
【００５２】
次に、ホストランク決定部２０は、記憶部１８に記憶されたコンテンツ第１情報Ｄ２から、取得したＵＲＬのＨＴＴＰステータスコード（「ステータス２」）を取得する（Ｓ３０７）。次に、ホストランク決定部２０は、取得したＵＲＬのＨＴＴＰステータスコードのうち、リダイレクトを示すＨＴＴＰステータスコードの割合（リダイレクトを示すＵＲＬの割合）を算出し、この割合が所定の閾値以上であるか否かを判定する（Ｓ３０９）。リダイレクトを示すＨＴＴＰステータスコードは、例えば、３００系のコードである。
【００５３】
ホストランク決定部２０は、リダイレクトを示すＵＲＬの割合が所定の閾値以上であると判定した場合、すなわち、処理対象のホストに属するＵＲＬの多くがリダイレクトを示すものであると判定した場合、ホストランクを所定の順位だけ下げる（Ｓ３１１）。一方、ホストランク決定部２０は、リダイレクトを示すＵＲＬの割合が所定の閾値以上ではないと判定した場合、すなわち、処理対象のホストに属するＵＲＬにリダイレクトを示すもの数が少ないと判定した場合、上述のホストランクを下げる処理を行わない。
【００５４】
次に、ホストランク決定部２０は、処理対象とするホストに属するＵＲＬのうち、画像データ、動画データ等を示すＵＲＬの割合を算出し、この割合が所定の閾値以上であるか否かを判定する（Ｓ３１３）。ホストランク決定部２０は、例えば、ＵＲＬの拡張子に基づいて、ＵＲＬが、画像等を示すものであるか否かを判定する。尚、ホストランク決定部２０は、ＵＲＬに対応するコンテンツのヘッダ情報に基づいて、ＵＲＬが、画像等を示すものであるか否かを判定してもよい。
【００５５】
ホストランク決定部２０は、画像等を示すＵＲＬの割合が所定の閾値以上であると判定した場合、すなわち、処理対象のホストに属するＵＲＬの多くが画像等を示すものであると判定した場合、ホストランクを所定の順位だけ下げる（Ｓ３１５）。一方、ホストランク決定部２０は、画像等を示すＵＲＬの割合が所定の閾値以上ではないと判定した場合、すなわち、処理対象のホストに属するＵＲＬに画像等を示すものの数が少ないと判定した場合、上述のホストランクを下げる処理を行わない。ホストランク決定部２０は、上述の処理により決定したホストランクを記憶部１８のホストランク情報Ｄ１に追加または更新する。以上により、本フローチャートの処理を終了する。
【００５６】
尚、上記の実施形態においては、ホストランク決定部２０が、リダイレクトを示すＵＲＬの割合が所定の閾値以上であると判定した場合や、画像等を示すＵＲＬの割合が所定の閾値以上であると判定した場合に、ホストランクを所定の順位だけ下げる例を説明した。しかしながら、ホストランク決定部２０は、上述の場合に、処理対象のホストを、ホストランクから除外するようにしてもよい。
【００５７】
尚、上記の実施形態においては、リダイレクトを示すＵＲＬに対する処理と、画像等を示すＵＲＬに対する処理との両方を実施する例を説明した。しかしながら、ホストランク決定部２０は、リダイレクトを示すＵＲＬに対する処理と、画像等を示すＵＲＬに対する処理とのいずれか一方を行うようにしてもよい。
【００５８】
［注目ホストリスト生成処理］
以下、クロールサーバ１０の注目ホストリスト生成処理について説明する。図１１は、

本実施形態の注目ホストリスト生成処理の流れの一例を示すフローチャートである。本フローチャートによる処理は、日次、週次等、所定の時間間隔のバッチ処理として実行される。

【００５９】

まず、注目ホストリスト生成部２２は、記憶部１８に記憶されたホストランク情報Ｄ１を取得する（Ｓ４０１）。次に、注目ホストリスト生成部２２は、ホストランク情報Ｄ１に含まれる複数のホストのなかから、優先してデータを収集するホストを選出した注目ホストリストＤ４を生成する（Ｓ４０３）。例えば、注目ホストリスト生成部２２は、ホストランクが所定の順位以上のホスト（例えば、上位１００位）を注目ホストとして決定し、注目ホストリストＤ４を生成する。以上により、本フローチャートの処理が終了する。

【００６０】

以上において説明した実施形態によれば、ネットワークを介してアクセス可能な装置からデータを収集する収集部と、前記収集部によって収集されたデータに含まれる、前記ネットワークを介してアクセス可能な装置に格納されたデータを参照するための参照情報が、所定の条件を満たす場合に、前記参照情報の一部を構成して複数の参照情報の群を特定する所属情報に対する前記収集部による収集を抑制する抑制部とを備えることで、データ収集の効率を向上させることができる。すなわち、不要なデータを多く含む価値の低いホストに対するクロールを抑制し、有効なデータを多く含む有益なホストに集中してクロールを行うことができる。これにより、データ収集に要する時間を短縮し、リソースを有効に活用することができる。

【００６１】

以上、本発明を実施するための形態について実施形態を用いて説明したが、本発明はこうした実施形態に何等限定されるものではなく、本発明の要旨を逸脱しない範囲内において種々の変形及び置換を加えることができる。

【符号の説明】

【００６２】

１０…クロールサーバ（データ収集装置）
１２…データ収集部（収集部）
１４…解析部
１６…バッチ処理部
１８…記憶部
２０…ホストランク決定部
２２…注目ホストリスト生成部
２４…ホスト選択部

【図１】

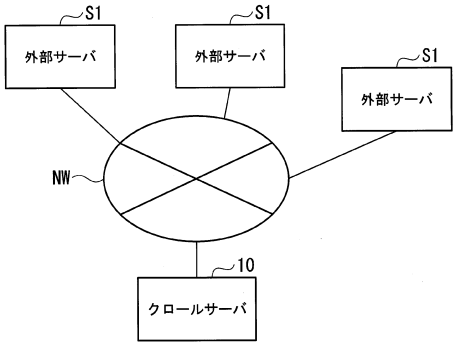

図１

【図２】

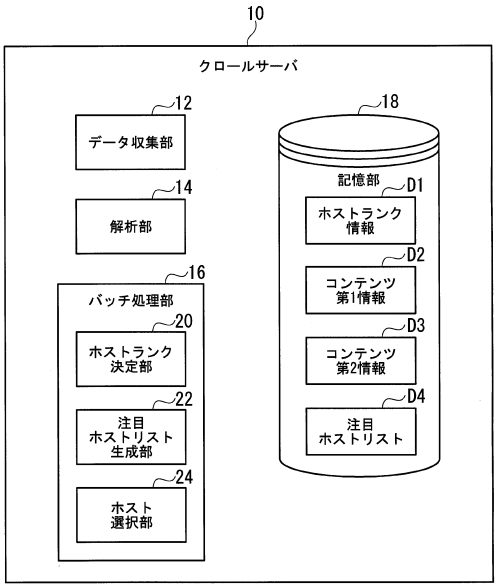

図２

【図３】

D1

ホスト	ホストランク
AAA	30
BBB	1
CCC	5
DDD	8
EEE	2

図３

【図４】

D2

URL	ステータス1	ステータス2	スコア
ddd.ddd	取得済	301	5
eee.eee	取得済	200	70
fff.fff	未取得	未取得	未付与
ggg.ggg	取得済	401	3
hhh.hhh	取得済	200	55
iii.iii	取得済	503	40

図４

【図５】

D3

URL	コンテンツ
aaa.aaa	HTMLデータ1
bbb.bbb	画像データ1
ccc.ccc	HTMLデータ5

図５

【図６】

D4

注目ホスト
BBB
EEE
QQQ
HHH
CCC
NNN

図６

【図７】

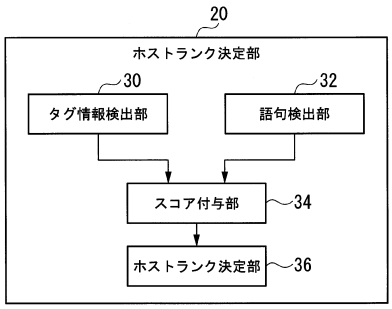

図7

【図８】

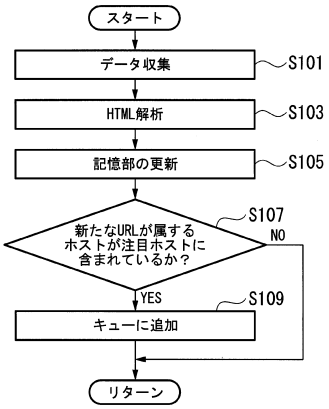

図8

【図９】

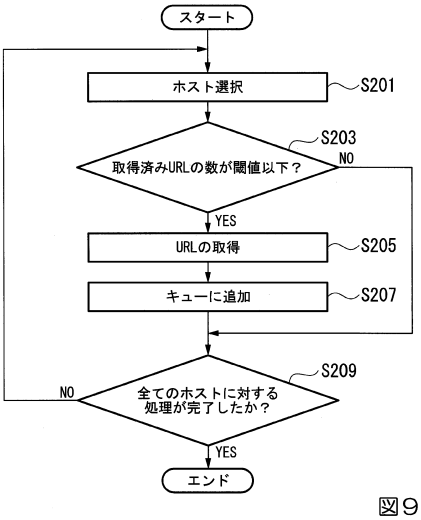

図9

【図１０】

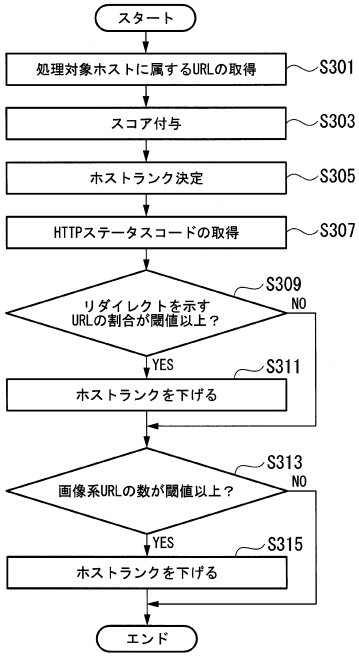

図１０

【図１１】

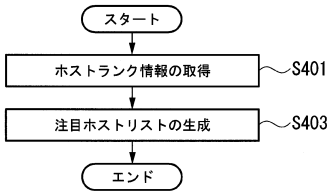

図１１

フロントページの続き

(72)発明者　和良品　友大
東京都千代田区紀尾井町１番３号　ヤフー株式会社内
(72)発明者　俵　雄貴
東京都千代田区紀尾井町１番３号　ヤフー株式会社内
(72)発明者　タウフィックラチマン
東京都千代田区紀尾井町１番３号　ヤフー株式会社内
(72)発明者　田中　康之
東京都千代田区紀尾井町１番３号　ヤフー株式会社内

審査官　鹿野　博嗣

(56)参考文献　特開２００６－２３５７２９（ＪＰ，Ａ）
特開平１０－２６０８９０（ＪＰ，Ａ）
特開２０１４－１８６７１９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１６／９５１