明　細　書

発明の名称 ：

情報取得プログラム、情報取得方法および情報取得装置

技術分野

[0001] 本発明は、情報取得プログラム、情報取得方法および情報取得装置に関する。

背景技術

[0002] インターネット上に公開されている情報を収集するためのツールとして、クローラツールが知られている。クローラツールは、インターネット上のホームページを巡回して、ＵＲＬ（Uniform Resource Locator）単位、すなわちページ単位で内容を保存する。また、ホームページの内容のうち、ユーザの求めるレイアウトフォーム内の項目に必要なデータを抽出して集約データを生成し、集約データをレイアウトフォームに合わせて出力することが提案されている。さらに、レイアウトフォーム上のデータ入力部に空欄が残ったときに、再度、データの抽出を行うことが提案されている。

[0003] また、構造情報を用いて構造化された構造化データから、予め指定した構造情報に対応したデータを取得して、テンプレートに配置することが提案されている。さらに、構造情報に対応するデータの少なくとも一部が存在していない場合に、データの補充に関する情報を送信することが提案されている。

先行技術文献

特許文献

[0004] 特許文献1：特開２００２－１６９８３６号公報

特許文献2：特開２００８－２８２１１４号公報

発明の概要

発明が解決しようとする課題

[0005] しかしながら、例えば、抽出条件を変更して再度データを抽出したり、デ

ータの補充に関する情報を送信しても、レイアウトフォームやテンプレートの欄に対応する所望のデータが得られない場合がある。このため、取得したデータを用いた文書を生成する場合に、データ、すなわち情報の欠落が少ない文書を生成することが困難である。

[0006] 一つの側面では、本発明は、文書に記載すべき情報を補完できる情報取得プログラム、情報取得方法および情報取得装置を提供することにある。

課題を解決するための手段

[0007] 一つの態様では、情報取得プログラムは、あるサイトの文書中における特定のデータの位置を特定する情報と、該文書中における該特定のデータとの間で所定の関連を有する他のデータの位置を特定する情報とを記憶する記憶部を参照して、前記あるサイトの文書中における特定のデータの位置を特定する情報を用いて前記特定のデータを取得する処理をコンピュータに実行させる。また、情報取得プログラムは、前記記憶部を参照して、前記文書中における前記特定のデータに関連する前記他のデータの位置を用いて前記他のデータを取得するとともに、取得した前記他のデータを用いて前記他のデータと前記所定の関連を有するデータを取得する処理をコンピュータに実行させる。

発明の効果

[0008] 文書に記載すべき情報を補完できる。

図面の簡単な説明

[0009] [図1]図１は、データ取得装置の構成の一例を示すブロック図である。

[図2]図２は、対象記憶部の一例を示す図である。

[図3]図３は、項目記憶部の一例を示す図である。

[図4]図４は、ページ記憶部の一例を示す図である。

[図5]図５は、抽出データ記憶部の一例を示す図である。

[図6]図６は、抽出対象部分の受付画面の一例を示す図である。

[図7]図７は、定義生成処理の一例を示すフローチャートである。

[図8]図８は、クロール処理の一例を示すフローチャートである。

[図9]図９は、クロール処理の他の一例を示すフローチャートである。

[図10]図１０は、情報取得プログラムを実行するコンピュータの一例を示す図である。

発明を実施するための形態

[0010] 以下、図面に基づいて、本願の開示する情報取得プログラム、情報取得方法および情報取得装置の実施例を詳細に説明する。なお、本実施例により、開示技術が限定されるものではない。また、以下の実施例は、矛盾しない範囲で適宜組み合わせてもよい。

実施例

[0011] 図１は、データ取得装置の構成の一例を示すブロック図である。図１に示すデータ取得装置１００は、例えば、ネットワークＮを介してインターネットに接続され、管理者に指定されたインターネット上のホームページ（以下、サイトともいう）を巡回し、所定のデータを取得してデータベースに蓄積する。データ取得装置１００は、例えば、ある地域の観光情報を取得するために、観光スポットのサイトや都道府県が設けた観光情報サイトを巡回して、各観光スポットの住所、電話番号、説明文等のデータを取得する。このとき、各観光スポットのサイトや観光情報サイトでは、各種データのフォーマットが統一されていない場合が多い。このため、データ取得装置１００は、取得するデータ項目の定義を予め生成し、定義に基づいて各サイトからデータを取得する。また、ある観光スポットの情報は、１つのサイトでは、所望のデータの全てについて取得できない場合がある。このため、データ取得装置１００は、他のサイトの情報等を用いてある観光スポットの情報を補完する。

[0012] すなわち、データ取得装置１００は、あるサイトの文書中における特定のデータの位置を特定する情報と、該文書中における該特定のデータとの間で所定の関連を有する他のデータの位置を特定する情報とを記憶部１２０に記憶する。データ取得装置１００は、記憶部１２０を参照して、あるサイトの文書中における特定のデータの位置を特定する情報を用いて特定のデータを

取得する。また、データ取得装置１００は、記憶部１２０を参照して、文書中における特定のデータに関連する他のデータの位置を用いて他のデータを取得するとともに、取得した他のデータを用いて他のデータと所定の関連を有するデータを取得する。これにより、データ取得装置１００は、各種データ、つまり文書に記載すべき情報の一部が不足した文書について、文書に記載すべき情報を補完できる。

[0013] ここで、文書としては、例えば、マークアップ言語で記述された文書が挙げられ、例えばＨＴＭＬ（HyperText Markup Language）文書、ＸＭＬ（Extensible Markup Language）文書等が挙げられる。なお、以下の説明では、一例として、ＨＴＭＬ文書を用いたホームページを巡回し、不足するデータを補完する場合について説明する。

[0014] 次に、データ取得装置１００の構成について説明する。図１に示すように、データ取得装置１００は、入力部１０１と、出力部１０２と、通信部１１０と、記憶部１２０と、制御部１３０とを有する。なお、データ取得装置１００は、情報取得装置の一例である。また、データ取得装置１００は、図１に示す機能部以外にも既知のコンピュータが有する各種の機能部を有することとしてもかまわない。

[0015] 入力部１０１は、例えば、キーボードやマウス等の入力デバイスであり、データ取得装置１００の管理者から各種情報の入力を受け付ける。例えば、入力部１０１は、データ取得装置１００の管理者により、巡回するサイトのＵＲＬ、取得するデータ項目等が入力され、入力結果を制御部１３０に出力する。また、入力部１０１は、例えば、ＳＤ（Secure Digital）メモリカード等のリーダライタであってもよい。入力部１０１は、例えば、ＳＤメモリカードから読み込んだ、巡回するサイトのＵＲＬ、取得するデータ項目等を制御部１３０に出力する。なお、入力部１０１は、入力デバイスとＳＤメモリカード等のリーダライタとの双方を有してもよい。

[0016] 出力部１０２は、例えば、各種情報を表示するための表示デバイスである。出力部１０２は、例えば、表示デバイスとして液晶ディスプレイ等によっ

て実現される。また、出力部１０２は、ＳＤメモリカード等のリーダライタであってもよい。出力部１０２は、制御部１３０から出力データが入力されると、出力データについて表示またはメモリカードへの書き込みを行う。なお、入力部１０１および出力部１０２は、一体化されてもよく、例えば、ＳＤメモリカード等のリーダライタのように、双方の機能を有するデバイスであってもよい。また、出力部１０２は、例えば、表示デバイスとＳＤカードリーダライタの双方を有してもよい。

[0017] 通信部１１０は、例えば、ＮＩＣ（Network Interface Card）等によって実現される。通信部１１０は、ネットワークＮを介して、例えばインターネットと有線または無線で接続され、インターネット上の各種サイトのサーバとの間で情報の通信を司る通信インタフェースである。通信部１１０は、インターネット上の各種サイトからページ内容、例えば、ＨＴＭＬ文書、画像ファイルや、問合せ情報に対する応答情報等を受信する。通信部１１０は、受信したページ内容および応答情報を制御部１３０に出力する。また、通信部１１０は、制御部１３０から入力されたページ要求、問合せ情報等をインターネット上の各種サイトに送信する。

[0018] 記憶部１２０は、例えば、ＲＡＭ（Random Access Memory）、フラッシュメモリ（Flash Memory）等の半導体メモリ素子、ハードディスクや光ディスク等の記憶装置によって実現される。記憶部１２０は、対象記憶部１２１と、項目記憶部１２２と、ページ記憶部１２３と、抽出データ記憶部１２４とを有する。また、記憶部１２０は、制御部１３０での処理に用いる情報を記憶する。

[0019] 対象記憶部１２１は、データを取得するクロール処理の対象となるサイトのＵＲＬ（以下、対象ＵＲＬという）と、ＨＴＭＬ文書における抽出対象部分の位置特定情報とを対応付けて記憶する。すなわち、対象記憶部１２１は、対象ＵＲＬの定義を記憶する。図２は、対象記憶部の一例を示す図である。図２に示すように、対象記憶部１２１は、「ＵＲＬＩＤ」、「対象ＵＲＬ」、「抽出対象部分の位置特定情報」といった項目を有する。また、「抽出

対象部分の位置特定情報」は、「タイトル」、「住所」といった項目を有する。なお、抽出対象部分の位置特定情報は、図示はしないが、他にも、施設名称、電話番号、更新日、位置情報、説明文といった項目を有する。対象記憶部１２１は、例えば、１つの対象ＵＲＬごとに１レコードとして記憶する。

[0020] 「ＵＲＬＩＤ」は、対象ＵＲＬを識別する。「対象ＵＲＬ」は、クロール処理でアクセスする対象となるＨＴＭＬ文書のＵＲＬを示す。対象ＵＲＬは、例えば、管理者によって入力部１０１の入力デバイスにより入力される。「抽出対象部分の位置特定情報」は、対象ＵＲＬのＨＴＭＬ文書内における抽出対象部分の位置を特定するための情報を示す。「タイトル」は、対象となるＨＴＭＬ文書内のタイトルについて、タグの名称、タグの文書内における順番、および、タグの階層構造のうち１つ以上を組み合わせて、タグの階層構造上の位置を示す。「住所」は、対象となるＨＴＭＬ文書内の住所について、タグの名称、タグの文書内における順番、および、タグの階層構造のうち１つ以上を組み合わせて、タグの階層構造上の位置を示す。

[0021] 図２の１行目の例では、ＵＲＬＩＤが「１」の対象ＵＲＬ「http://aaaa.bbb.ccc/ddd/eee/001.html」のＨＴＭＬ文書内における、タイトルおよび住所の位置特定情報を示す。タイトルの位置特定情報は、例えば、「<DIV class="title"> </DIV>,順番：1,/title/」と表現される。「<DIV class="title"> </DIV>」は、例えば、ＣＳＳ（Cascading Style Sheets）セレクタを用いて抽出したタイトルを示すタグの名称を示す。「順番：1」は、当該ＨＴＭＬ文書内のタイトルを示すタグのうち、１番目のタグを示す。「/title/」は、当該ＨＴＭＬ文書のタイトルを示すタグの階層構造を示す。なお、当該ＨＴＭＬ文書からタイトルとして抜き出されるデータは、ＤＩＶタグに囲まれた部分となる。

[0022] 同様に、住所の位置特定情報は、例えば、「<DIV class="address"> </DIV>,順番：1,/info/address/」と表現される。「<DIV class="address"> </DIV>」は、例えば、ＣＳＳセレクタを用いて抽出した住所を示すタグの名称

を示す。「順番：1」は、当該ＨＴＭＬ文書内の住所を示すタグのうち、１番目のタグを示す。「/info/address/」は、当該ＨＴＭＬ文書の住所を示すタグの階層構造を示す。なお、当該ＨＴＭＬ文書から住所として抜き出されるデータは、ＤＩＶタグに囲まれた部分となる。また、抽出対象部分の位置特定情報は、タグの名称、タグの順番、および、タグの階層構造のうち１つ以上を用いて特定してもよい。

[0023] また、タグの名称は、正規表現を用いて表してもよい。図２の２行目の例では、住所を示すタグの名称を「/<DIV.*>(.+)</DIV>/ /住所：(.+)$/」と表現している。正規表現では、ＤＩＶタグに囲まれた箇所、または、「住所：」の後ろに続く箇所が、住所として抜き出されるデータとなる。さらに、抽出対象部分の位置特定情報は、ＣＳＳセレクタと正規表現を組み合わせてもよい。

[0024] また、図２の３行目の例のように、抽出対象部分の位置特定情報は、切り出し手法を用いて表現してもよい。この場合には、タイトルの位置特定情報は、例えば、ＣＳＳセレクタを用いて「div#left h2,順番：3,/tps/table/」と表現される。また、住所の位置特定情報は、例えば、ＣＳＳセレクタと正規表現とを用いて「#infoContent @<h3>所在地</h3>¥s+? (.+?) @is,順番：5,/info/address/」と表現される。

[0025] 図１の説明に戻って、項目記憶部１２２は、対象ＵＲＬのページ内容から抽出するデータ項目の定義を記憶する。図３は、項目記憶部の一例を示す図である。図３に示すように、項目記憶部１２２は、「項目ＩＤ」、「データ名」、「データ型」、「切り出し手法」、「関連」といった項目を有する。項目記憶部１２２は、例えば、１つのデータ名ごとに、１レコードとして記憶する。

[0026] 「項目ＩＤ」は、データ項目、すなわちデータ名を識別する。「データ名」は、抽出するデータの名前を示す。データ名は、例えば、タイトル、住所（住所情報）、施設名称（施設名称情報）、電話番号、更新日、位置情報（位置座標情報）、説明文といったデータが挙げられる。「データ型」は、抽

出したデータを抽出データ記憶部１２４に記憶する際の当該データの型を示す。データ型は、例えば、文字、数字、日付、緯度経度といった型が挙げられる。「切り出し手法」は、対象ＵＲＬのページ内容からデータを切り出す、つまり抜き出す手法を示す。切り出し手法は、例えば、ＣＳＳセレクタ、正規表現といった手法が挙げられる。「関連」は、データ名のうち、関連のあるデータ名を示す。関連は、例えば、住所と関連するデータ名として、施設名称、電話番号、位置情報といったデータ名が挙げられる。

[0027] 図１の説明に戻って、ページ記憶部１２３は、対象ＵＲＬについて、クロール処理でアクセスして取得したページ内容、すなわち、ＨＴＭＬ文書、画像ファイル等を記憶する。図４は、ページ記憶部の一例を示す図である。図４に示すように、ページ記憶部１２３は、「ＵＲＬＩＤ」、「対象ＵＲＬ」、「記憶領域」といった項目を有する。ページ記憶部１２３は、例えば、１つの対象ＵＲＬごとに１レコードとして記憶する。

[0028] 「ＵＲＬＩＤ」は、対象ＵＲＬを識別する。「対象ＵＲＬ」は、クロール処理でアクセスしたＨＴＭＬ文書のＵＲＬを示す。「記憶領域」は、取得したＨＴＭＬ文書や画像ファイル等を記憶した記憶領域を示す。記憶領域は、例えば、記憶部１２０のファイルシステムのディレクトリを記憶し、対応するディレクトリにＨＴＭＬ文書や画像ファイル等を記憶する。なお、ページ記憶部１２３は、記憶領域に、取得したＨＴＭＬ文書や画像ファイルを直接記憶するようにしてもよい。

[0029] 図１の説明に戻って、抽出データ記憶部１２４は、ＨＴＭＬ文書から抽出された、抽出対象部分のデータを記憶する。すなわち、抽出データ記憶部１２４は、クロール処理によって収集されたデータを格納するデータベースである。図５は、抽出データ記憶部の一例を示す図である。図５に示すように、抽出データ記憶部１２４は、「ＵＲＬＩＤ」、「タイトル」、「住所」、「施設名称」、「電話番号」、「更新日」、「位置情報」、「説明文」といった項目を有する。抽出データ記憶部１２４は、例えば、１つのＵＲＬＩＤごとに１レコードとして記憶する。

[0030] 「ＵＲＬＩＤ」は、対象ＵＲＬを識別する。「タイトル」は、対象ＵＲＬのＨＴＭＬ文書から抽出されたデータ項目の１つであり、対象ＵＲＬのＨＴＭＬ文書のタイトルを示す。「住所」は、対象ＵＲＬのＨＴＭＬ文書から抽出されたデータ項目の１つであり、対象ＵＲＬのＨＴＭＬ文書内に記載された住所を示す。「施設名称」は、対象ＵＲＬのＨＴＭＬ文書から抽出されたデータ項目の１つであり、対象ＵＲＬのＨＴＭＬ文書内に記載された施設の名称を示す。なお、タイトルと施設名称が同一である場合には、施設名称としてタイトルと同じデータを用いてもよい。

[0031] 「電話番号」は、対象ＵＲＬのＨＴＭＬ文書から抽出されたデータ項目の１つであり、対象ＵＲＬのＨＴＭＬ文書内に記載された電話番号を示す。「更新日」は、対象ＵＲＬのＨＴＭＬ文書から抽出されたデータ項目の１つであり、対象ＵＲＬのＨＴＭＬ文書内に記載された更新日を示す。「位置情報」は、緯度経度を示す。緯度経度は、対象ＵＲＬのＨＴＭＬ文書から抽出された住所に基づいて、例えば、外部のＡＰＩ（Application Programming Interface）サービスを利用することで取得される。なお、位置情報は、ＨＴＭＬ文書内に緯度経度の記載があれば、当該緯度経度であってもよい。「説明文」は、対象ＵＲＬのＨＴＭＬ文書から抽出されたデータ項目の１つであり、例えば、対象ＵＲＬのＨＴＭＬ文書が観光スポットに関する文書であれば、文書内の観光スポットに関する説明文を示す。なお、住所等の各項目は、ＨＴＭＬ文書内に記載がない場合には、例えば、タイトルに記載された観光スポット名を用いて、外部のＡＰＩサービスを利用することで取得された住所等であってもよい。

[0032] 図１の説明に戻って、制御部１３０は、例えば、ＣＰＵ（Central Processing Unit）やＭＰＵ（Micro Processing Unit）等によって、内部の記憶装置に記憶されているプログラムがＲＡＭを作業領域として実行されることにより実現される。また、制御部１３０は、例えば、ＡＳＩＣ（Application Specific Integrated Circuit）やＦＰＧＡ（Field Programmable Gate Array）等の集積回路により実現されるようにしてもよい。制御部１３０

は、登録部１３１と、クロール部１３２と、抽出部１３３と、補完部１３４と、出力制御部１３５とを有し、以下に説明する情報処理の機能や作用を実現または実行する。なお、制御部１３０の内部構成は、図１に示した構成に限られず、後述する情報処理を行う構成であれば他の構成であってもよい。

[0033] 登録部１３１は、対象ＵＲＬの定義およびデータ項目の定義を登録する。登録部１３１は、例えば、管理者が入力部１０１を操作することにより、抽出対象部分とするデータ名、データ型、切り出し手法および関連の入力を受け付ける。登録部１３１は、受け付けたデータ名、データ型、切り出し手法および関連を対応付けて、データ項目の定義を生成する。登録部１３１は、生成したデータ項目の定義を項目記憶部１２２に記憶する。つまり、登録部１３１は、生成したデータ項目の定義を項目記憶部１２２に登録する。なお、関連の項目は、入力されたデータ名に基づいて、登録部１３１が自動で割り当てるようにしてもよい。

[0034] 登録部１３１は、対象ＵＲＬに対応するＨＴＭＬ文書のソースを出力部１０２に出力して表示させる。登録部１３１は、例えば、管理者が入力部１０１を操作することにより、表示させた対象ＵＲＬに対応するＨＴＭＬ文書のソース上で、抽出対象部分の選択を受け付ける。なお、登録部１３１は、対象ＵＲＬのＨＴＭＬ文書を表示させて、ＨＴＭＬ文書上で抽出対象部分の選択を受け付けるようにしてもよい。

[0035] 登録部１３１は、受け付けた抽出対象部分に対応するタグの階層構造上の位置を特定する。登録部１３１は、特定した階層構造上の位置を抽出対象部分の位置特定情報とする。また、登録部１３１は、抽出対象部分に対応するタグの名称、および、タグの文書内における順番を、特定した階層構造上の位置とともに抽出対象部分の位置特定情報とする。登録部１３１は、対象ＵＲＬのＨＴＭＬ文書内の各データ項目について、抽出対象部分の選択を受け付けて、タグの階層構造上の位置を特定する。また、登録部１３１は、対象ＵＲＬが複数ある場合には、それぞれの対象ＵＲＬに対応するＨＴＭＬ文書について、同様に抽出対象部分に対応するタグの階層構造上の位置を特定す

る。登録部１３１は、対象ＵＲＬと抽出対象部分の位置特定情報とを対応付けて、対象ＵＲＬの定義を生成する。登録部１３１は、生成した対象ＵＲＬの定義を対象記憶部１２１に記憶する。つまり、登録部１３１は、生成した対象ＵＲＬの定義を対象記憶部１２１に登録する。

[0036] ここで、図６を用いて、抽出対象部分の受付画面について説明する。図６は、抽出対象部分の受付画面の一例を示す図である。図６に示すように、受付画面２１は、ＨＴＭＬ文書のソースを表示する領域２２と、抽出対象部分の選択を受け付ける領域２３とを有する。登録部１３１は、例えば、抽出対象部分として住所の選択を受け付ける場合には、領域２３内の抽出対象部分の選択欄で、住所が選択される。登録部１３１は、住所が選択されると、住所に対応するデータ項目の定義を項目記憶部１２２から読み出して抽出定義欄２４に表示する。なお、抽出定義欄２４は、編集可能なテキストとして表示されるようにしてもよい。

[0037] 登録部１３１は、抽出定義欄２４のＣＳＳセレクタおよび正規表現のうち１つ以上に対応する部分を、領域２２に表示されたソース上に抽出対象部分２５として、例えばバックグラウンドを着色して表示させる。登録部１３１は、管理者により抽出対象部分２５が確認され、例えば、図示しないユーザインタフェース上の選択ボタンが押下されることで、抽出対象部分２５の選択を受け付ける。また、登録部１３１は、例えば、管理者のマウス操作によって領域２２の抽出対象部分２５が選択され、選択された抽出対象部分２５を受け付けるようにしてもよい。

[0038] さらに、登録部１３１は、抽出対象部分２５に対して、不要な文字を削る変換処理をしてもよい。図６の例では、登録部１３１は、管理者によって設定された変換処理欄２６の変換定義を用いて、抽出対象部分２５の文字列に対して変換処理を行う。登録部１３１は、例えば、変換結果２７を抽出対象部分２５の下に挿入して、バックグラウンドを抽出対象部分２５と異なる色に着色して表示させる。登録部１３１は、変換処理を行った場合には、変換結果２７を抽出対象部分として選択させて受け付けることができる。

[0039] 図１の説明に戻って、クロール部１３２は、対象記憶部１２１を参照して、対象ＵＲＬを含むホームページ、例えば、ある観光情報サイトのトップページにアクセスする。すなわち、クロール部１３２は、ある観光情報サイトのサーバに対して通信部１１０を介してページ要求を送信し、当該サーバから通信部１１０を介してページ内容を受信する。クロール部１３２は、例えば、定期的または不定期に、つまり予め管理者によって指定された間隔または任意のタイミングで、対象ＵＲＬを含むホームページにアクセスする。指定された間隔は、例えば、１日、１週間、１ヶ月等のように任意の間隔とすることができる。クロール部１３２は、対象記憶部１２１を参照して、ホームページ内の全リンクのうち、ページ内容を取得する対象ＵＲＬを選定する。クロール部１３２は、例えば、観光スポットごとのページの対象ＵＲＬを選定する。クロール部１３２は、選定した対象ＵＲＬからページ内容を取得する。クロール部１３２は、取得したページ内容をページ記憶部１２３に記憶する。また、クロール部１３２は、ページ内容の取得が完了したことを示す取得完了情報を抽出部１３３に出力する。

[0040] 抽出部１３３は、クロール部１３２から取得完了情報が入力されると、対象記憶部１２１の抽出対象部分の位置特定情報を参照して、ページ記憶部１２３に記憶された対象ＵＲＬのページ内容から、抽出対象部分のデータ項目のデータを抽出する。抽出部１３３は、抽出したデータをＵＲＬＩＤと対応付けて、項目記憶部１２２のデータ項目の定義に従って、抽出データ記憶部１２４に記憶する。

[0041] 抽出部１３３は、抽出したデータのうち、特定のデータが抽出できたか否かを判定する。抽出部１３３は、特定のデータが抽出できた場合には、対象記憶部１２１および項目記憶部１２２を参照して、抽出した特定のデータに関連する他のデータの位置特定情報に基づき他のデータを抽出する。抽出部１３３は、例えば、特定のデータが施設名称であり、施設名称が抽出できた場合に、施設名称に関連する他のデータとして、例えば、住所を抽出する。抽出部１３３は、抽出した他のデータをＵＲＬＩＤと対応付けて、抽出デー

タ記憶部１２４に記憶すると、補完部１３４に抽出完了情報を出力する。なお、抽出部１３３は、抽出した他のデータが既に抽出データ記憶部１２４に記憶されている場合には、他のデータを新たに記憶せずに補完部１３４に抽出完了情報を出力する。

[0042] 抽出部１３３は、特定のデータが抽出できない場合には、対象記憶部１２１および項目記憶部１２２を参照して、特定のデータに関連する他のデータの位置特定情報に基づき他のデータを抽出する。抽出部１３３は、例えば、特定のデータが施設名称であり、施設名称が抽出できない場合に、施設名称に関連する他のデータとして、例えば、住所を抽出する。抽出部１３３は、抽出した他のデータをＵＲＬＩＤと対応付けて、抽出データ記憶部１２４に記憶すると、他のデータと所定の関連を有するデータの取得を要求する補完要求情報を補完部１３４に出力する。なお、他のデータと所定の関連を有するデータには、特定のデータが含まれ、上述の例では、施設名称が含まれる。なお、抽出部１３３は、抽出した他のデータが既に抽出データ記憶部１２４に記憶されている場合には、他のデータを新たに記憶せずに補完部１３４に補完要求情報を出力する。

[0043] 抽出部１３３は、抽出対象部分のデータ項目のデータを抽出する場合に、項目記憶部１２２の切り出し手法で指定された手法を用いて抽出する。抽出部１３３は、例えば、住所を示すタグの階層が「/info/address/」で定義され、例えば「.address」と記述されたＣＳＳセレクタを用いることで住所を抽出する。この場合には、抽出部１３３は、例えば、タグ内に「address」を含む項目を、住所として切り出すことができる。

[0044] また、抽出部１３３は、例えば、１行目に「.info」と記述され、２行目に「/<DIV.*>(.+)</DIV>/」と記述され、３行目に「/住所：(.+)$/」と記述された正規表現を用いることで住所を抽出する。この場合には、抽出部１３３は、例えば、ＤＩＶタグのクラスが「info」であるタグに含まれる階層から、「住所：」の文字列の後に続く文字列を住所として切り出すことができる。

[0045] 補完部１３４は、抽出部１３３から抽出完了情報が入力されると、項目記憶部１２２および抽出データ記憶部１２４を参照して、抽出した他のデータを用いて他のデータと所定の関連を有するデータを、例えば、インターネット上のＡＰＩサービスから取得する。補完部１３４は、例えば、他のデータが住所である場合には、住所を含む問合せ情報を、住所を受信することで緯度経度を示す位置情報を返信するＡＰＩサービスを提供するサーバに、通信部１１０およびネットワークＮを介して送信する。補完部１３４は、当該サーバから問合せ情報に対する応答情報を、通信部１１０およびネットワークＮを介して受信することで、他のデータと所定の関連を有するデータとして、位置情報を取得する。補完部１３４は、取得した他のデータと所定の関連を有するデータ、例えば住所に対応する位置情報を、対象ＵＲＬのＵＲＬＩＤと対応付けて抽出データ記憶部１２４に記憶する。

[0046] 補完部１３４は、抽出部１３３から補完要求情報が入力されると、項目記憶部１２２および抽出データ記憶部１２４を参照して、抽出した他のデータを用いて他のデータと所定の関連を有するデータを、例えば、インターネット上のＡＰＩサービスから取得する。補完部１３４は、例えば、他のデータが住所である場合には、住所を含む問合せ情報を、住所を受信することで施設名称を返信するＡＰＩサービスを提供するサーバに、通信部１１０およびネットワークＮを介して送信する。補完部１３４は、当該サーバから問合せ情報に対する応答情報を、通信部１１０およびネットワークＮを介して受信することで、他のデータと所定の関連を有するデータとして、施設名称を取得する。すなわち、補完部１３４は、ＨＴＭＬ文書から抽出できなかった施設名称について、補完することができる。

[0047] また、補完部１３４は、例えば、他のデータが住所である場合には、住所を含む問合せ情報を、住所を受信することで緯度経度を示す位置情報を返信するＡＰＩサービスを提供するサーバに、通信部１１０およびネットワークＮを介して送信する。補完部１３４は、当該サーバから問合せ情報に対する応答情報を、通信部１１０およびネットワークＮを介して受信することで、

他のデータと所定の関連を有するデータとして、位置情報を取得する。補完部１３４は、取得した他のデータと所定の関連を有するデータ、例えば、住所に対応する施設名称および位置情報を、対象ＵＲＬのＵＲＬＩＤと対応付けて抽出データ記憶部１２４に記憶する。補完部１３４は、他のデータと所定の関連を有するデータを、抽出データ記憶部１２４に記憶すると、出力情報を出力制御部１３５に出力する。

[0048] すなわち、抽出部１３３は、あるサイトの文書中における特定のデータの位置を特定する情報を用いて特定のデータを取得する処理が実行できない場合に、文書中における特定のデータに関連する他のデータの位置を用いて他のデータを取得する。また、補完部１３４は、取得した他のデータを用いて他のデータと所定の関連を有するデータを取得する。補完部１３４は、取得した他のデータと所定の関連を有するデータを対象ＵＲＬのＵＲＬＩＤと対応付けて抽出データ記憶部１２４に記憶する。補完部１３４は、取得した他のデータと所定の関連を有するデータを抽出データ記憶部１２４に記憶すると、出力情報を出力制御部１３５に出力する。

[0049] 補完部１３４は、ＡＰＩサービスとして、上述した住所と位置情報との関連を返信するサービスの他に、電話番号と施設名称との関連を相互に返信するサービスを用いることができる。また、補完部１３４は、ＡＰＩサービスとして、例えば、タイトル、施設名称、住所、電話番号、位置情報、公式サイト等を相互に関連付けたデータベースを有し、これらのうち１つ以上の情報を受信すると他の情報を返信するＡＰＩサービスを用いてもよい。例えば、住所と緯度経度との関連情報を提供するＡＰＩサービスとしては、住所および緯度経度について相互に関連を有する情報を提供できるＧｏｏｇｌｅ Ｍａｐｓ ＡＰＩ等が挙げられる。また、例えば、電話番号および施設名称について相互に関連を有する情報を提供できるＡＰＩサービスとしては、Ｆｏｕｒｓｑｕａｒｅ ＡＰＩ等が挙げられる。言い換えると、補完部１３４は、他のデータを所定の関連に応じた検索エンジンに提供することで、他のデータと所定の関連を有するデータを取得する。

[0050] また、補完部１３４は、他のデータと所定の関連を有するデータとして、所定の関連を有する複数のデータを、インターネット上のＡＰＩサービスから取得するようにしてもよい。補完部１３４は、例えば、ＨＴＭＬ文書から施設名称が抽出できない場合に、施設名称に関連する住所をＨＴＭＬ文書から抽出し、住所に関連する位置情報についてＡＰＩサービスを用いて取得する。また、補完部１３４は、例えば、施設名称に関連する電話番号を抽出し、電話番号に関連する施設名称についてＡＰＩサービスを用いて取得する。なお、補完部１３４は、複数のデータについて、１つ目のデータが取得できない場合に、２つ目のデータを取得するといった処理を行ってもよいし、複数のデータをそれぞれ取得するような処理を行ってもよい。

[0051] すなわち、抽出部１３３は、あるサイトの文書中における特定のデータの位置を特定する情報を用いて特定のデータを得る処理が実行できない場合に、文書中における特定のデータに関連する第一のデータの位置を用いて第一のデータを取得する。また、補完部１３４は、取得した第一のデータを用いて第一のデータと第一の関連を有するデータを取得する。さらに、抽出部１３３は、第一のデータと第一の関連を有するデータが取得できない場合に、文書中における特定のデータに関連する第二のデータの位置を用いて第二のデータを取得する。また、補完部１３４は、取得した第二のデータを用いて第二のデータと第二の関連を有するデータを取得する。補完部１３４は、取得した第一のデータと第一の関連を有するデータ、または、第二のデータと第二の関連を有するデータを対象ＵＲＬのＵＲＬＩＤと対応付けて抽出データ記憶部１２４に記憶する。補完部１３４は、取得した第一のデータと第一の関連を有するデータ、または、第二のデータと第二の関連を有するデータを抽出データ記憶部１２４に記憶すると、出力情報を出力制御部１３５に出力する。

[0052] 出力制御部１３５は、補完部１３４から出力情報が入力されると、抽出データ記憶部１２４を参照して、抽出および取得したデータを出力データとして出力部１０２に出力して表示させる。また、出力制御部１３５は、抽出し

たデータの出力時に、過去のクロール処理または補完処理によって取得して抜き出したデータと、今回のクロール処理または補完処理によって取得して抜き出したデータとが異なる場合には、例えば表示色を変更するようにしてもよい。なお、出力制御部１３５は、出力部１０２がＳＤメモリカード等のリーダライタである場合には、抽出および取得したデータを出力データとして出力部１０２に出力して、ＳＤメモリカード等に記憶させる。

[0053] 次に、実施例のデータ取得装置１００の動作について説明する。まず、クロール処理の対象ＵＲＬの定義と、抜き出すデータ項目の定義とを生成する定義生成処理について説明する。

[0054] 図７は、定義生成処理の一例を示すフローチャートである。登録部１３１は、例えば、管理者が入力部１０１を操作することにより、抽出対象部分とするデータ名、データ型、切り出し手法および関連の入力を受け付ける（ステップＳ１）。登録部１３１は、受け付けたデータ名、データ型、切り出し手法および関連を対応付けて、データ項目の定義を生成する。登録部１３１は、生成したデータ項目の定義を項目記憶部１２２に登録する（ステップＳ２）。

[0055] 登録部１３１は、対象ＵＲＬに対応するＨＴＭＬ文書のソースを出力部１０２に出力して表示させる（ステップＳ３）。登録部１３１は、例えば、管理者が入力部１０１を操作することにより、表示させた対象ＵＲＬに対応するＨＴＭＬ文書のソース上で、抽出対象部分の選択を受け付ける（ステップＳ４）。登録部１３１は、受け付けた抽出対象部分に対応するタグの階層構造上の位置を特定する（ステップＳ５）。登録部１３１は、特定した階層構造上の位置を抽出対象部分の位置特定情報とする（ステップＳ６）。また、登録部１３１は、抽出対象部分に対応するタグの名称、および、タグの文書内における順番を、特定した階層構造上の位置とともに抽出対象部分の位置特定情報とする。なお、登録部１３１は、対象ＵＲＬのＨＴＭＬ文書内に複数のデータ項目がある場合には、それぞれ抽出対象部分の選択を受け付けて、タグの階層構造上の位置を特定する。

[0056] 登録部１３１は、対象ＵＲＬと抽出対象部分の位置特定情報とを対応付けて、対象ＵＲＬの定義を生成する。登録部１３１は、生成した対象ＵＲＬの定義を対象記憶部１２１に登録する（ステップＳ７）。これにより、データ取得装置１００は、データ項目の定義と対象ＵＲＬの定義を登録することができる。

[0057] 続いて、クロール処理について説明する。図８は、クロール処理の一例を示すフローチャートである。クロール部１３２は、対象記憶部１２１を参照して、対象ＵＲＬを含むホームページにアクセスする（ステップＳ１１）。クロール部１３２は、対象記憶部１２１を参照して、ホームページ内の全リンクのうち、ページ内容を取得する対象ＵＲＬを選定する（ステップＳ１２）。

[0058] クロール部１３２は、選定した対象ＵＲＬからページ内容を取得する（ステップＳ１３）。クロール部１３２は、取得したページ内容をページ記憶部１２３に記憶する。また、クロール部１３２は、ページ内容の取得が完了したことを示す取得完了情報を抽出部１３３に出力する。

[0059] 抽出部１３３は、クロール部１３２から取得完了情報が入力されると、対象記憶部１２１の抽出対象部分の位置特定情報を参照して、ページ記憶部１２３に記憶された対象ＵＲＬのページ内容から、抽出対象部分のデータ項目のデータを抽出する（ステップＳ１４）。

[0060] 抽出部１３３は、抽出したデータをＵＲＬＩＤと対応付けて、抽出データ記憶部１２４に記憶する（ステップＳ１５）。抽出部１３３は、抽出したデータのうち、特定のデータが抽出できたか否かを判定する（ステップＳ１６）。抽出部１３３は、特定のデータが抽出できた場合には（ステップＳ１６：肯定）、対象記憶部１２１および項目記憶部１２２を参照して、抽出した特定のデータに関連する他のデータの位置特定情報に基づき他のデータを抽出する（ステップＳ１７）。抽出部１３３は、抽出した他のデータをＵＲＬＩＤと対応付けて、抽出データ記憶部１２４に記憶すると、補完部１３４に抽出完了情報を出力する。

[0061] 補完部１３４は、抽出部１３３から抽出完了情報が入力されると、項目記憶部１２２および抽出データ記憶部１２４を参照して、抽出した他のデータを用いて他のデータと所定の関連を有するデータを取得する（ステップＳ１８）。補完部１３４は、他のデータと所定の関連を有するデータを、例えば、インターネット上のＡＰＩサービスから取得する。

[0062] 抽出部１３３は、特定のデータが抽出できない場合には（ステップＳ１６：否定）、対象記憶部１２１および項目記憶部１２２を参照して、特定のデータに関連する他のデータの位置特定情報に基づき他のデータを抽出する（ステップＳ１９）。抽出部１３３は、抽出した他のデータをＵＲＬＩＤと対応付けて、抽出データ記憶部１２４に記憶すると、他のデータと所定の関連を有するデータの取得を要求する補完要求情報を補完部１３４に出力する。

[0063] 補完部１３４は、抽出部１３３から補完要求情報が入力されると、項目記憶部１２２および抽出データ記憶部１２４を参照して、抽出した他のデータを用いて他のデータと所定の関連を有するデータを取得する（ステップＳ２０）。補完部１３４は、他のデータと所定の関連を有するデータを、例えば、インターネット上のＡＰＩサービスから取得する。

[0064] 補完部１３４は、取得した他のデータと所定の関連を有するデータを、ＵＲＬＩＤと対応付けて抽出データ記憶部１２４に記憶する（ステップＳ２１）。補完部１３４は、他のデータと所定の関連を有するデータを、抽出データ記憶部１２４に記憶すると、出力情報を出力制御部１３５に出力する。出力制御部１３５は、補完部１３４から出力情報が入力されると、抽出データ記憶部１２４を参照して、抽出および取得したデータを出力データとして出力部１０２に出力して表示させる（ステップＳ２２）。これにより、データ取得装置１００は、不足するデータ、つまり情報についてＡＰＩサービスを用いて補完できるので、文書に記載すべき情報を補完できる。

[0065] 次に、クロール処理の他の一例について説明する。図９は、クロール処理の他の一例を示すフローチャートである。図９に示すクロール処理は、図８のクロール処理と比べて、特定のデータに関連する他のデータの抽出、およ

び、他のデータと関連を有するデータの取得を複数のデータについて行う点が異なる。

[0066] クロール部１３２は、対象記憶部１２１を参照して、対象ＵＲＬを含むホームページにアクセスする（ステップＳ１１）。クロール部１３２は、対象記憶部１２１を参照して、ホームページ内の全リンクのうち、ページ内容を取得する対象ＵＲＬを選定する（ステップＳ１２）。

[0067] クロール部１３２は、選定した対象ＵＲＬからページ内容を取得する（ステップＳ１３）。クロール部１３２は、取得したページ内容をページ記憶部１２３に記憶する。また、クロール部１３２は、ページ内容の取得が完了したことを示す取得完了情報を抽出部１３３に出力する。

[0068] 抽出部１３３は、クロール部１３２から取得完了情報が入力されると、対象記憶部１２１の抽出対象部分の位置特定情報を参照して、ページ記憶部１２３に記憶された対象ＵＲＬのページ内容から、抽出対象部分のデータ項目のデータを抽出する（ステップＳ１４）。

[0069] 抽出部１３３は、抽出したデータをＵＲＬＩＤと対応付けて、抽出データ記憶部１２４に記憶する（ステップＳ１５）。抽出部１３３は、抽出したデータのうち、特定のデータが抽出できたか否かを判定する（ステップＳ１６）。抽出部１３３は、特定のデータが抽出できた場合には（ステップＳ１６：肯定）、対象記憶部１２１および項目記憶部１２２を参照して、抽出した特定のデータに関連する第一のデータの位置特定情報に基づき第一のデータを抽出する（ステップＳ３１）。抽出部１３３は、抽出した第一のデータをＵＲＬＩＤと対応付けて、抽出データ記憶部１２４に記憶すると、補完部１３４に抽出完了情報を出力する。

[0070] 補完部１３４は、抽出部１３３から抽出完了情報が入力されると、抽出した第一のデータを用いて第一のデータと第一の関連を有するデータを、例えば、インターネット上のＡＰＩサービスから取得する（ステップＳ３２）。

[0071] 抽出部１３３は、特定のデータが抽出できない場合には（ステップＳ１６：否定）、対象記憶部１２１および項目記憶部１２２を参照して、特定のデ

ータに関連する第一のデータの位置特定情報に基づき第一のデータを抽出する（ステップＳ３３）。抽出部１３３は、抽出した第一のデータをＵＲＬＩＤと対応付けて、抽出データ記憶部１２４に記憶すると、第一のデータと第一の関連を有するデータの取得を要求する補完要求情報を補完部１３４に出力する。

[0072] 補完部１３４は、抽出部１３３から補完要求情報が入力されると、抽出した第一のデータを用いて第一のデータと第一の関連を有するデータを、例えば、インターネット上のＡＰＩサービスから取得する（ステップＳ３４）。

[0073] 抽出部１３３は、補完部１３４が第一のデータと第一の関連を有するデータが取得できたか否かを判定する（ステップＳ３５）。抽出部１３３は、補完部１３４が第一のデータと第一の関連を有するデータが取得できた場合には（ステップＳ３５：肯定）、ステップＳ３８に進む。抽出部１３３は、補完部１３４が第一のデータと第一の関連を有するデータが取得できない場合には（ステップＳ３５：否定）、特定のデータに関連する第二のデータの位置特定情報に基づき第二のデータを抽出する（ステップＳ３６）。すなわち、抽出部１３３は、対象記憶部１２１および項目記憶部１２２を参照して、特定のデータに関連する第二のデータの位置特定情報に基づき第二のデータを抽出する。抽出部１３３は、抽出した第二のデータをＵＲＬＩＤと対応付けて、抽出データ記憶部１２４に記憶すると、第二のデータと第二の関連を有するデータの取得を要求する補完要求情報を補完部１３４に出力する。

[0074] 補完部１３４は、抽出部１３３から補完要求情報が入力されると、抽出した第二のデータを用いて第二のデータと第二の関連を有するデータを、例えば、インターネット上のＡＰＩサービスから取得する（ステップＳ３７）。

[0075] 補完部１３４は、取得した、第一のデータと第一の関連を有するデータ、または、第二のデータと第二の関連を有するデータを、ＵＲＬＩＤと対応付けて抽出データ記憶部１２４に記憶する（ステップＳ３８）。補完部１３４は、第一のデータと第一の関連を有するデータ、または、第二のデータと第二の関連を有するデータを、抽出データ記憶部１２４に記憶すると、出力情

報を出力制御部１３５に出力する。出力制御部１３５は、補完部１３４から出力情報が入力されると、抽出データ記憶部１２４を参照して、抽出および取得したデータを出力データとして出力部１０２に出力して表示させる（ステップＳ３９）。これにより、データ取得装置１００は、１つ目の関連するデータが取得できない場合に、２つ目の関連するデータを取得することができ、より文書に記載すべき情報を補完できる。

[0076] このように、データ取得装置１００は、あるサイトの文書中における特定のデータの位置を特定する情報と、該文書中における該特定のデータとの間で所定の関連を有する他のデータの位置を特定する情報とを記憶部１２０に記憶する。データ取得装置１００は、記憶部１２０を参照して、あるサイトの文書中における特定のデータの位置を特定する情報を用いて特定のデータを取得する。また、データ取得装置１００は、記憶部１２０を参照して、文書中における特定のデータに関連する他のデータの位置を用いて他のデータを取得するとともに、取得した他のデータを用いて他のデータと所定の関連を有するデータを取得する。その結果、文書に記載すべき情報を補完できる。

[0077] また、データ取得装置１００は、他のデータを所定の関連に応じた検索エンジンに提供することで他のデータと所定の関連を有するデータを取得する。その結果、より文書に記載すべき情報を補完できる。

[0078] また、データ取得装置１００は、あるサイトの文書中における特定のデータの位置を特定する情報を用いて特定のデータを取得する処理が実行できない場合に、文書中における特定のデータに関連する他のデータの位置を用いて他のデータを取得する。また、データ取得装置１００は、取得した他のデータを用いて他のデータと所定の関連を有するデータを取得する。その結果、文書から特定のデータを抽出できない場合であっても、他のデータを用いて特定のデータを補完することができる。

[0079] また、データ取得装置１００は、文書として、マークアップ言語を用いた文書を用いる。その結果、タグの階層構造上の位置を用いてデータの位置を

特定できる。

[0080] また、データ取得装置１００は、所定の関連として、住所情報と位置座標情報、または、電話番号と施設名称情報を用いる。その結果、より文書に記載すべき情報を補完できる。

[0081] また、データ取得装置１００は、あるサイトの文書中における特定のデータの位置を特定する情報と、該文書中における該特定のデータとの間で第一の関連を有する第一のデータの位置を特定する情報とを記憶部１２０に記憶する。また、データ取得装置１００は、該文書中における該特定のデータとの間で第二の関連を有する第二のデータの位置を特定する情報を記憶部１２０に記憶する。データ取得装置１００は、記憶部１２０を参照して、あるサイトの文書中における特定のデータの位置を特定する情報を用いて特定のデータを得る処理が実行できない場合に、文書中における特定のデータに関連する第一のデータの位置を用いて第一のデータを取得する。また、データ取得装置１００は、記憶部１２０を参照して、第一のデータと第一の関連を有するデータが取得できない場合に、文書中における特定のデータに関連する第二のデータの位置を用いて第二のデータを取得する。また、データ取得装置１００は、取得した第二のデータを用いて第二のデータと第二の関連を有するデータを取得する。その結果、１つ目の関連するデータが取得できない場合に、２つ目の関連するデータを取得することができ、より文書に記載すべき情報を補完できる。

[0082] なお、上記実施例では、特定のデータに関連する他のデータの抽出、および、他のデータと関連を有するデータの取得を複数のデータについて行う場合に、２つのデータについて説明したが、これに限定されない。例えば、特定のデータに関連する他のデータの抽出数、および、他のデータと関連を有するデータの取得数は、任意の数のデータについて行うようにしてもよい。

[0083] また、上記実施例では、住所と位置情報、電話番号と施設名称とが所定の関連を有し、相互に補完する場合を一例として説明したが、これに限定されない。例えば、観光スポットの説明文が取得できない場合には、観光スポッ

トの名称または施設名称と、当該観光スポットまたは施設の住所とに基づいて、インターネット上のＡＰＩサービスや観光スポットの公式サイト等から説明文を取得するようにしてもよい。

[0084] また、上記実施例では、インターネット上のサイトから取得したＨＴＭＬ文書からデータを抽出したが、これに限定されない。例えば、ＳＤメモリカードに格納したＨＴＭＬ文書を入力部１０１を介して取得して、取得したＨＴＭＬ文書からデータを抽出し、不足するデータをインターネット上のＡＰＩサービスや観光スポットの公式サイト等から取得して補完するようにしてもよい。

[0085] また、図示した各部の各構成要素は、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、各部の分散・統合の具体的形態は図示のものに限られず、その全部または一部を、各種の負荷や使用状況等に応じて、任意の単位で機能的または物理的に分散・統合して構成することができる。例えば、クロール部１３２と、抽出部１３３とを統合してもよい。

[0086] さらに、各装置で行われる各種処理機能は、ＣＰＵ（またはＭＰＵ、ＭＣＵ（Micro Controller Unit）等のマイクロ・コンピュータ）上で、その全部または任意の一部を実行するようにしてもよい。また、各種処理機能は、ＣＰＵ（またはＭＰＵ、ＭＣＵ等のマイクロ・コンピュータ）で解析実行されるプログラム上、またはワイヤードロジックによるハードウェア上で、その全部または任意の一部を実行するようにしてもよいことは言うまでもない。

[0087] ところで、上記の実施例で説明した各種の処理は、予め用意されたプログラムをコンピュータで実行することで実現できる。そこで、以下では、上記の実施例と同様の機能を有するプログラムを実行するコンピュータの一例を説明する。図１０は、情報取得プログラムを実行するコンピュータの一例を示す図である。

[0088] 図１０が示すように、コンピュータ２００は、各種演算処理を実行するＣＰＵ２０１と、データ入力を受け付ける入力装置２０２と、モニタ２０３と

を有する。また、コンピュータ２００は、記憶媒体からプログラム等を読み取る媒体読取装置２０４と、各種装置と接続するためのインタフェース装置２０５と、他の情報処理装置等と有線または無線により接続するための通信装置２０６とを有する。また、コンピュータ２００は、各種情報を一時記憶するＲＡＭ２０７と、ハードディスク装置２０８とを有する。また、各装置２０１～２０８は、バス２０９に接続される。

[0089] ハードディスク装置２０８には、図１に示した登録部１３１、クロール部１３２、抽出部１３３、補完部１３４および出力制御部１３５の各処理部と同様の機能を有する情報取得プログラムが記憶される。また、ハードディスク装置２０８には、対象記憶部１２１、項目記憶部１２２、ページ記憶部１２３、抽出データ記憶部１２４、および、情報取得プログラムを実現するための各種データが記憶される。入力装置２０２は、入力部１０１と同等の機能を有し、例えば、コンピュータ２００の管理者から、対象ＵＲＬ、定義、管理情報等の各種情報の入力を受け付ける。モニタ２０３は、出力部１０２と同等の機能を有し、例えば、コンピュータ２００の管理者に対して管理情報の画面、受付画面、データ表示画面等の各種画面を表示する。インタフェース装置２０５は、例えば、印刷装置等が接続される。通信装置２０６は、例えば、図１に示した通信部１１０と同様の機能を有しネットワークＮと接続され、インターネット上のサイトと各種情報をやりとりする。

[0090] ＣＰＵ２０１は、ハードディスク装置２０８に記憶された各プログラムを読み出して、ＲＡＭ２０７に展開して実行することで、各種の処理を行う。また、これらのプログラムは、コンピュータ２００を図１に示した登録部１３１、クロール部１３２、抽出部１３３、補完部１３４および出力制御部１３５として機能させることができる。

[0091] なお、上記の情報取得プログラムは、必ずしもハードディスク装置２０８に記憶されている必要はない。例えば、コンピュータ２００が読み取り可能な記憶媒体に記憶されたプログラムを、コンピュータ２００が読み出して実行するようにしてもよい。コンピュータ２００が読み取り可能な記憶媒体は

、例えば、ＣＤ－ＲＯＭやＤＶＤディスク、ＵＳＢ（Universal Serial Bus）メモリ等の可搬型記録媒体、フラッシュメモリ等の半導体メモリ、ハードディスクドライブ等が対応する。また、公衆回線、インターネット、ＬＡＮ等に接続された装置にこの情報取得プログラムを記憶させておき、コンピュータ２００がこれらから情報取得プログラムを読み出して実行するようにしてもよい。

符号の説明

[0092] １００　データ取得装置

１０１　入力部

１０２　出力部

１１０　通信部

１２０　記憶部

１２１　対象記憶部

１２２　項目記憶部

１２３　ページ記憶部

１２４　抽出データ記憶部

１３０　制御部

１３１　登録部

１３２　クロール部

１３３　抽出部

１３４　補完部

１３５　出力制御部

Ｎ　ネットワーク

請求の範囲

[請求項1] あるサイトの文書中における特定のデータの位置を特定する情報と、該文書中における該特定のデータとの間で所定の関連を有する他のデータの位置を特定する情報とを記憶する記憶部を参照して、前記あるサイトの文書中における特定のデータの位置を特定する情報を用いて前記特定のデータを取得し、

前記記憶部を参照して、前記文書中における前記特定のデータに関連する前記他のデータの位置を用いて前記他のデータを取得するとともに、取得した前記他のデータを用いて前記他のデータと前記所定の関連を有するデータを取得する

処理をコンピュータに実行させることを特徴とする情報取得プログラム。

[請求項2] 前記他のデータと前記所定の関連を有するデータを取得する処理は、前記他のデータを前記所定の関連に応じた検索エンジンに提供することで前記他のデータと前記所定の関連を有するデータを取得することを特徴とする請求項１に記載の情報取得プログラム。

[請求項3] 前記他のデータと前記所定の関連を有するデータを取得する処理は、前記あるサイトの文書中における特定のデータの位置を特定する情報を用いて前記特定のデータを取得する処理が実行できない場合に、前記文書中における前記特定のデータに関連する前記他のデータの位置を用いて前記他のデータを取得するとともに、取得した前記他のデータを用いて前記他のデータと前記所定の関連を有するデータを取得することを特徴とする請求項１に記載の情報取得プログラム。

[請求項4] 前記文書は、マークアップ言語を用いた文書であることを特徴とする請求項１に記載の情報取得プログラム。

[請求項5] 前記所定の関連は、住所情報と位置座標情報、または、電話番号と施設名称情報であることを特徴とする請求項１に記載の情報取得プログラム。

[請求項6] あるサイトの文書中における特定のデータの位置を特定する情報と、該文書中における該特定のデータとの間で第一の関連を有する第一のデータの位置を特定する情報と、該文書中における該特定のデータとの間で第二の関連を有する第二のデータの位置を特定する情報とを記憶する記憶部を参照して、前記あるサイトの文書中における特定のデータの位置を特定する情報を用いて前記特定のデータを得る処理が実行できない場合に、前記文書中における前記特定のデータに関連する前記第一のデータの位置を用いて前記第一のデータを取得するとともに、取得した前記第一のデータを用いて前記第一のデータと前記第一の関連を有するデータを取得し、

前記記憶部を参照して、前記第一のデータと前記第一の関連を有するデータが取得できない場合に、前記文書中における前記特定のデータに関連する前記第二のデータの位置を用いて前記第二のデータを取得するとともに、取得した前記第二のデータを用いて前記第二のデータと前記第二の関連を有するデータを取得する

処理をコンピュータに実行させることを特徴とする情報取得プログラム。

[請求項7] あるサイトの文書中における特定のデータの位置を特定する情報と、該文書中における該特定のデータとの間で所定の関連を有する他のデータの位置を特定する情報とを記憶する記憶部を参照して、前記あるサイトの文書中における特定のデータの位置を特定する情報を用いて前記特定のデータを取得し、

前記記憶部を参照して、前記文書中における前記特定のデータに関連する前記他のデータの位置を用いて前記他のデータを取得するとともに、取得した前記他のデータを用いて前記他のデータと前記所定の関連を有するデータを取得する

処理をコンピュータが実行することを特徴とする情報取得方法。

[請求項8] 前記他のデータと前記所定の関連を有するデータを取得する処理は

、前記他のデータを前記所定の関連に応じた検索エンジンに提供することで前記他のデータと前記所定の関連を有するデータを取得することを特徴とする請求項７に記載の情報取得方法。

[請求項9] 前記他のデータと前記所定の関連を有するデータを取得する処理は、前記あるサイトの文書中における特定のデータの位置を特定する情報を用いて前記特定のデータを取得する処理が実行できない場合に、前記文書中における前記特定のデータに関連する前記他のデータの位置を用いて前記他のデータを取得するとともに、取得した前記他のデータを用いて前記他のデータと前記所定の関連を有するデータを取得することを特徴とする請求項７に記載の情報取得方法。

[請求項10] 前記文書は、マークアップ言語を用いた文書であることを特徴とする請求項７に記載の情報取得方法。

[請求項11] 前記所定の関連は、住所情報と位置座標情報、または、電話番号と施設名称情報であることを特徴とする請求項７に記載の情報取得方法。

[請求項12] あるサイトの文書中における特定のデータの位置を特定する情報と、該文書中における該特定のデータとの間で第一の関連を有する第一のデータの位置を特定する情報と、該文書中における該特定のデータとの間で第二の関連を有する第二のデータの位置を特定する情報とを記憶する記憶部を参照して、前記あるサイトの文書中における特定のデータの位置を特定する情報を用いて前記特定のデータを得る処理が実行できない場合に、前記文書中における前記特定のデータに関連する前記第一のデータの位置を用いて前記第一のデータを取得するとともに、取得した前記第一のデータを用いて前記第一のデータと前記第一の関連を有するデータを取得し、

前記記憶部を参照して、前記第一のデータと前記第一の関連を有するデータが取得できない場合に、前記文書中における前記特定のデータに関連する前記第二のデータの位置を用いて前記第二のデータを取

得するとともに、取得した前記第二のデータを用いて前記第二のデータと前記第二の関連を有するデータを取得する

処理をコンピュータが実行することを特徴とする情報取得方法。

[請求項13] あるサイトの文書中における特定のデータの位置を特定する情報と、該文書中における該特定のデータとの間で所定の関連を有する他のデータの位置を特定する情報とを記憶する記憶部と、

前記あるサイトの文書中における特定のデータの位置を特定する情報を用いて前記特定のデータを得る処理、および、前記文書中における前記特定のデータに関連する前記他のデータの位置を用いて前記他のデータを取得するとともに、取得した前記他のデータを用いて前記他のデータと前記所定の関連を有するデータを取得する処理の双方を実行可能な制御部と

を有することを特徴とする情報取得装置。

[請求項14] 前記制御部は、前記他のデータを前記所定の関連に応じた検索エンジンに提供することで前記他のデータと前記所定の関連を有するデータを取得する

ことを特徴とする請求項１３に記載の情報取得装置。

[請求項15] 前記制御部は、前記あるサイトの文書中における特定のデータの位置を特定する情報を用いて前記特定のデータを得る処理が実行できない場合に、前記文書中における前記特定のデータに関連する前記他のデータの位置を用いて前記他のデータを取得するとともに、取得した前記他のデータを用いて前記他のデータと前記所定の関連を有するデータを取得する処理を実行する

ことを特徴とする請求項１３に記載の情報取得装置。

[請求項16] 前記文書は、マークアップ言語を用いた文書であることを特徴とする請求項１３に記載の情報取得装置。

[請求項17] 前記所定の関連は、住所情報と位置座標情報、または、電話番号と施設名称情報であることを特徴とする請求項１３に記載の情報取得装

置。

[請求項18] あるサイトの文書中における特定のデータの位置を特定する情報と、該文書中における該特定のデータとの間で第一の関連を有する第一のデータの位置を特定する情報と、該文書中における該特定のデータとの間で第二の関連を有する第二のデータの位置を特定する情報とを記憶する記憶部と、

前記あるサイトの文書中における特定のデータの位置を特定する情報を用いて前記特定のデータを得る処理が実行できない場合に、前記文書中における前記特定のデータに関連する前記第一のデータの位置を用いて前記第一のデータを取得するとともに、取得した前記第一のデータを用いて前記第一のデータと前記第一の関連を有するデータを取得する処理を実行し、前記第一のデータと前記第一の関連を有するデータが取得できない場合に、前記文書中における前記特定のデータに関連する前記第二のデータの位置を用いて前記第二のデータを取得するとともに、取得した前記第二のデータを用いて前記第二のデータと前記第二の関連を有するデータを取得する処理を実行可能な制御部と

を有することを特徴とする情報取得装置。