

明 細 書

発明の名称：

データ検証プログラム、データ検証方法及びデータ検証装置

技術分野

[0001] 本発明は、データ検証プログラム、データ検証方法及びデータ検証装置に関する。

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、地図データについて、名称や住所をインターネット上で検索し、検索結果の件数に応じて誤りか否かを判定することが提案されている。

[0003] また、インターネット上の風説情報を収集して、風説情報を取得したサイトごとに、新規性、信憑性、影響度、危険度等のスコアを求め、風説情報の利用者に提供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2010-231560号公報

特許文献2：特開2008-165284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、難読地名等であれば、誤った読み方を記載したサイトの件数が、正しい読み方を記載したサイトの件数よりも多くなる場合がある。このため、検索結果の件数により一律に正誤を判定することは困難である。また、風説情報は、様々な情報を含むため、所望の情報の正確性について検証するには、不要な情報が膨大となり効率が低下する場合がある。

[0006] 一つの側面では、本発明は、文書に記載すべき情報の信頼度を高くして出力できるデータ検証プログラム、データ検証方法及びデータ検証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、データ検証プログラムは、特定の情報の取得元として登録された複数のサイトの内、前記特定の情報として情報が取得されたサイトの信頼度を、前記複数のサイトのそれぞれに対応付けて信頼度を記憶する記憶部の記憶内容に基づいて特定する処理をコンピュータに実行させる。また、データ検証プログラムは、前記特定の情報として取得された前記情報を前記特定した前記信頼度とともに出力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の信頼度を高くして出力できる。

図面の簡単な説明

[0009] [図1]図1は、データ検証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、ページ記憶部の一例を示す図である。

[図4]図4は、抽出データ記憶部の一例を示す図である。

[図5]図5は、信頼度記憶部の一例を示す図である。

[図6]図6は、グループ記憶部の一例を示す図である。

[図7]図7は、検証結果の表示画面の一例を示す図である。

[図8]図8は、検証結果の表示画面の他の一例を示す図である。

[図9]図9は、第1の検証処理の一例を示すフローチャートである。

[図10]図10は、第2の検証処理の一例を示すフローチャートである。

[図11]図11は、第3の検証処理の一例を示すフローチャートである。

[図12]図12は、データ検証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、本願の開示するデータ検証プログラム、データ検証方法及びデータ検証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 1

[0011] 図1は、データ検証装置の構成の一例を示すブロック図である。図1に示すデータ検証装置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。データ検証装置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、同一の項目のデータ（値）が異なる場合がある。このため、データ検証装置100は、各サイトから取得した同一の項目のデータについて信頼度を算出し、項目のデータごとやサイトごとにデータの信頼度を出力する。

[0012] すなわち、データ検証装置100は、特定の情報の取得元として登録された複数のサイトの内、特定の情報として情報が取得されたサイトの信頼度を、複数のサイトのそれぞれに対応付けて信頼度を記憶する記憶部の記憶内容に基づいて特定する。また、データ検証装置100は、特定の情報として取得された情報を特定した信頼度とともに出力する。これにより、データ検証装置100は、項目のデータ（情報）を信頼度とともに出力できるので、信頼度の高い情報を、文書に記載すべき情報とすることができる。つまり、データ検証装置100は、文書に記載すべき情報の信頼度を高くして出力できる。

[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、取得したデータを検証する場合について説明する。

[0014] 次に、データ検証装置100の構成について説明する。図1に示すように、データ検証装置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。また、データ検証装置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、データ検証装置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、データ検証装置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又はメモリカードへの書き込みを行う。なお、入力部101及び出力部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又は無線で接続され、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、画像ファイル等を受信する。通信部 110 は、受信したページ内容を制御部 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、ページ記憶部 122 と、抽出データ記憶部 123 と、信頼度記憶部 124 と、グループ記憶部 125 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)、及び、サイトの信頼度を URL ID と対応付けて記憶する。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、「URL ID」、「対象 URL」、「サイトの信頼度」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] 「URL ID」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。

「サイトの信頼度」は、対象 URL の HTML 文書内の情報に基づく対象 URL のサイトの信頼度を示す。サイトの信頼度は、例えば、HTML 文書内の各項目の信頼度の合計値とすることができる。

[0021] 図 1 の説明に戻って、ページ記憶部 122 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 3 は、ページ記憶部の一例を示す図である。図 3 に示すように、ページ記憶部 122 は、「URL ID」、「対象 URL」、「記憶領域」といった項目を有する。ページ記憶部 122 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。

[0022] 「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122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。

[0023] 図1の説明に戻って、抽出データ記憶部123は、HTML文書から抽出された情報の、項目及びその値を記憶する。すなわち、抽出データ記憶部123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4は、抽出データ記憶部の一例を示す図である。図4に示すように、抽出データ記憶部123は、「URLID」、「項目」、「値」といった項目を有する。抽出データ記憶部123は、例えば、1つのURLIDごとに、対応するHTML文書内の項目及びその値を1つ以上記憶する。なお、抽出データ記憶部123に記憶されたデータは、任意のフォーマットの文書の各項目に、容易に登録できる。

[0024] 「URLID」は、対象URLを識別する。「項目」は、取得された情報の項目を示す。「値」は、取得された情報の値を示す。図4の1行目の例では、URLID「1」の1つ目の情報は、項目は「名称」であり、値は「三内丸山遺跡」となる。

[0025] 図1の説明に戻って、信頼度記憶部124は、例えば、観光スポットの項目ごとに、対応する値と、値の信頼度（スコア）とを記憶する。図5は、信頼度記憶部の一例を示す図である。図5に示すように、信頼度記憶部124は、「名称」、「項目」、「URLID」、「値」、「信頼度の要素」、「値の信頼度」といった項目を有する。信頼度記憶部124は、例えば、URLIDごとに1レコードとして記憶する。

[0026] 「名称」は、例えば、観光スポット等の名称を示す。「項目」は、名称に

対応する各項目を示す。項目は、例えば、観光スポットに対応する住所や電話番号が挙げられる。「URL ID」は、項目に対応するデータ、つまり項目の値を取得した対象URLを識別する。「値」は、項目に対応するデータを示す。値は、例えば、電話番号について、URL IDが「1」のサイトでは「017-781-xxxx」、URL IDが「2」のサイトでは「017-781-pppp」といった場合が挙げられる。

[0027] 「信頼度の要素」は、さらに「情報提供者」、「情報鮮度」、「情報量」といった項目を有し、値ごとの信頼度のスコアを示す。「情報提供者」は、例えば、対象URLのサイトを運営している組織に基づくスコアを示す。「情報鮮度」は、例えば、最終更新日に基づくスコアを示す。「情報量」は、例えば、対象URLの対応するHTML文書における情報量を示す。「値の信頼度」は、URL IDごとの値について、信頼度の要素のスコアを合計した値を示す。なお、各信頼度（スコア）は、値が大きいほど、信頼度が高くなる。

[0028] 図1の説明に戻って、グループ記憶部125は、項目の値について、同一の内容を有する値をグループにまとめ、グループごとに当該値を有するURL ID、すなわちサイトの数を対応付けて記憶する。図6は、グループ記憶部の一例を示す図である。図6に示すように、グループ記憶部125は、「名称」、「項目」、「値のグループ」、「サイトの数」、「各サイトの値の信頼度の合計」といた項目を有する。

[0029] 「名称」は、例えば、観光スポット等の名称を示す。「項目」は、名称に対応する各項目を示す。項目は、例えば、観光スポットに対応する住所や電話番号が挙げられる。「値のグループ」は、同一の内容を有する値のグループを示す。「サイトの数」は、同一の内容を有する値を取得したサイト、すなわち対象URLの数を示す。「各サイトの値の信頼度の合計」は、例えば、ある電話番号の値について、当該値を取得したサイトの値の信頼度を、当該値を有するサイトの数の分だけ加算した値である。なお、各サイトの値の信頼度の合計は、値が大きいほど、信頼度が高くなる。図6では、例えば、

名称「三内丸山遺跡」の項目「電話番号」について、値のグループ「017-781-xxxx」は、当該値を有するサイトが9つあり、9つのサイトの電話番号の値「017-781-xxxx」の信頼度の合計が「63」であることを示す。これに対し、値のグループ「017-781-pppp」は、当該値を有するサイトが4つあり、4つのサイトの電話番号の値「017-781-pppp」の信頼度の合計が「16」であることを示す。

[0030] 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、取得部131と、算出部132と、分類部133と、特定部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又は実行する。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。

[0031] 取得部131は、クロール処理で巡回する対象URLを受け付ける。取得部131は、例えば、管理者によって入力部101の入力デバイスが操作されることにより、対象URLの入力を受け付ける。取得部131は、受け付けた対象URLを対象記憶部121に記憶する。

[0032] 取得部131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、取得部131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。取得部131は、例えば、定期的又は不定期に、つまり予め管理者によって指定された間隔又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。取得部131は、対象

記憶部 1 2 1 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。取得部 1 3 1 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。取得部 1 3 1 は、選定した対象 URL からページ内容を取得する。取得部 1 3 1 は、取得したページ内容をページ記憶部 1 2 2 に記憶する。つまり、取得部 1 3 1 は、対象 URL を含むホームページに対してクロール処理を実行し、取得したページ内容をページ記憶部 1 2 2 に記憶する。

[0033] 取得部 1 3 1 は、ページ記憶部 1 2 2 に記憶された対象 URL のページ内容から、項目及びその値を抽出する。取得部 1 3 1 は、項目及びその値として、例えば、互いの間を所定のタグ又は記号で区切られた項目及びその値を抽出する。取得部 1 3 1 は、対象 URL の URL ID ごとに、項目及びその値を対応付けて、抽出データ記憶部 1 2 3 に記憶する。取得部 1 3 1 は、抽出した項目及びその値を抽出データ記憶部 1 2 3 に記憶すると、算出部 1 3 2 に、抽出完了情報を出力する。

[0034] 算出部 1 3 2 は、取得部 1 3 1 から抽出完了情報が入力されると、抽出データ記憶部 1 2 3 を参照して、対象 URL の値ごとに信頼度の要素に基づいて信頼度を算出する。図 5 の 1 行目を例にとると、まず、URL ID 「1」の対象 URL のサイトは、名称「三内丸山遺跡」に係るサイトであり、項目「住所」の値が「青森市三内字丸山 x x x」である。算出部 1 3 2 は、例えば、当該サイトが公式ページであり、情報提供者が、例えば、青森県である場合には、信頼度が高いとして、情報提供者のスコアを「3」とする。また、算出部 1 3 2 は、例えば、当該サイトの最終更新日が半年以内であれば、信頼度が高いとして、情報鮮度のスコアを「3」とする。さらに、算出部 1 3 2 は、例えば、当該サイトの情報量が、担当部署まで記載されている等、詳細に記載されていれば、信頼度が高いとして、情報量のスコアを「2」とする。

[0035] また、算出部 1 3 2 は、算出した信頼度の各要素のスコアの合計値を、その項目、例えば住所の信頼度とする。図 5 の 1 行目を例では、項目「住所」

のURLID「1」の値の信頼度は、 $3 + 3 + 2 = 8$ となる。算出部132は、算出した信頼度を項目、URLID及び値と対応付けて信頼度記憶部124に記憶する。算出部132は、算出した信頼度を信頼度記憶部124に記憶すると、算出完了情報を分類部133及び特定部134に出力する。

[0036] 分類部133は、算出部132から算出完了情報が入力されると、信頼度記憶部124を参照して、各項目の値について、同一の内容を有する値をグループに分類する。分類部133は、例えば、名称「三内丸山遺跡」の項目「電話番号」について、値が「017-781-xxxx」であるサイトの数を計数し、値のグループ「017-781-xxxx」のサイトの数とする。分類部133は、同様に、例えば、値が「017-781-pppp」であるサイトの数を計数し、値のグループ「017-781-pppp」のサイトの数とする。

[0037] また、分類部133は、値のグループの各サイトの値の信頼度の合計を算出する。分類部133は、例えば、図6に示す値のグループ「017-781-xxxx」では、当該値を有するサイトが9つある場合に、各サイトの値の信頼度が、例えば、8、7、6、・・・であり、合計値である「63」を算出する。分類部133は、項目ごとに、値のグループと、サイトの数と、各サイトの値の信頼度の合計とを対応付けて、分類結果としてグループ記憶部125に記憶する。分類部133は、分類結果をグループ記憶部125に記憶すると、分類完了情報を出力制御部135に出力する。

[0038] なお、分類部133は、最も高い信頼度が割り当てられた第1のサイトを除く他のサイトの値をグループに分類してもよい。この場合には、分類部133は、他のサイトの値について、項目ごとに、値のグループと、サイトの数と、各サイトの値の信頼度の合計とを対応付けて、分類結果としてグループ記憶部125に記憶する。分類部133は、分類結果をグループ記憶部125に記憶すると、分類完了情報を出力制御部135に出力する。

[0039] 特定部134は、算出部132から算出完了情報が入力されると、信頼度記憶部124を参照して、対象URLの信頼度を特定する。特定部134は

、信頼度記憶部 124 の対象 URL の URL ID に対応する値の信頼度を取得する。特定部 134 は、URL ID が同一である各項目の値の信頼度の和を算出してサイトの信頼度を特定する。特定部 134 は、特定したサイトの信頼度を対象記憶部 121 に記憶するとともに、出力制御部 135 に特定情報を出力する。

[0040] 出力制御部 135 は、特定部 134 から特定情報が入力されると、対象記憶部 121 及び信頼度記憶部 124 を参照して、各項目及びその値と、各項目及びその値が取得されたサイトの信頼度とを出力データとして出力部 102 に出力して表示させる。すなわち、出力制御部 135 は、取得部 131 で抽出した項目及びその値を、特定した信頼度とともに出力データとして出力部 102 に出力して表示させる。

[0041] ここで、表示画面の一例について説明する。図 7 は、検証結果の表示画面の一例を示す図である。図 7 に示す表示画面 21 は、例えば、観光スポットである「三内丸山遺跡」の住所「青森市三内字丸山 x x x」を取得したサイトの信頼度が「27」であることを示す。ここで、住所のデータ（値）を取得したサイトは、例えば、対象記憶部 121 に記憶された URL ID が「1」の対象 URL のサイトである。また、表示画面 21 は、例えば、電話番号のデータを取得したサイトも URL ID 「1」である場合には、住所と同様に、電話番号の信頼度は「27」となる。

[0042] 出力制御部 135 は、分類部 133 から分類完了情報が入力されると、グループ記憶部 125 を参照して、値のグループのうちサイト数が最も多いグループに属する値を、各サイトの値の信頼度の合計とともに出力データとして出力部 102 に出力して表示させる。すなわち、出力制御部 135 は、グループの構成要素の数が最も多いグループの値を、グループの構成要素の数に応じた信頼度とともに出力データとして出力部 102 に出力して表示させる。

[0043] ここで、表示画面の他の一例について説明する。図 8 は、検証結果の表示画面の他の一例を示す図である。図 8 に示す表示画面 22 は、例えば、観光

スポットである「三内丸山遺跡」の住所「青森市三内字丸山×××」の信頼度が「73」であることを示す。ここで、住所のデータ（値）は、値のグループ「青森市三内字丸山×××」に属する値であり、他の値のグループと比べて最もサイトの数が多かったものである。また、表示画面22は、例えば、電話番号のデータ（値）について、最もサイトの数が多かった「017-781-××××」が、信頼度「63」とともに表示される。

[0044] また、出力制御部135は、分類部133で、最も高い信頼度が割り当てられた第1のサイトを除く他のサイトの値をグループに分類した場合には、第1のサイトと他のサイトとで信頼度を比較する。すなわち、出力制御部135は、他のサイトにおけるグループの構成要素の数が最も多い特定のグループの構成要素の数に応じた信頼度と、第1のサイトの信頼度とを比較する。出力制御部135は、比較の結果、特定のグループの構成要素の数に応じた信頼度が、第1のサイトの信頼度より高いか否かを判定する。

[0045] 出力制御部135は、特定のグループの構成要素の数に応じた信頼度が、第1のサイトの信頼度より高い場合には、項目及び特定のグループに属する値を出力データとして出力部102に出力して表示させる。出力制御部135は、特定のグループの構成要素の数に応じた信頼度が、第1のサイトの信頼度以下である場合には、第1のサイトの項目及びその値を出力データとして出力部102に出力して表示させる。

[0046] 次に、実施例のデータ検証装置100の動作について説明する。まず、第1の検証処理について説明する。

[0047] 図9は、第1の検証処理の一例を示すフローチャートである。取得部131は、クロール処理で巡回する対象URLを受け付ける（ステップS1）。取得部131は、受け付けた対象URLを対象記憶部121に記憶する。取得部131は、例えば、定期的又は不定期に、つまり予め管理者によって指定された間隔又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（ステップS2）。取得部131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S3）。

[0048] 取得部131は、選定した対象URLからページ内容を取得する（ステップS4）。取得部131は、取得したページ内容をページ記憶部122に記憶する。取得部131は、ページ記憶部122に記憶された対象URLのページ内容から、項目及びその値を抽出する（ステップS5）。取得部131は、対象URLのURLIDごとに、項目及びその値を対応付けて、抽出データ記憶部123に記憶する（ステップS6）。取得部131は、抽出した項目及びその値を抽出データ記憶部123に記憶すると、算出部132に、抽出完了情報を入力する。

[0049] 算出部132は、取得部131から抽出完了情報が入力されると、抽出データ記憶部123を参照して、対象URLの値ごとに信頼度の要素に基づいて信頼度を算出する（ステップS7）。算出部132は、算出した信頼度を項目、URLID及び値と対応付けて信頼度記憶部124に記憶する（ステップS8）。算出部132は、算出した信頼度を信頼度記憶部124に記憶すると、算出完了情報を特定部134に出力する。

[0050] 特定部134は、算出部132から算出完了情報が入力されると、信頼度記憶部124を参照して、対象URLの信頼度を特定する（ステップS9）。特定部134は、特定したサイトの信頼度を対象記憶部121に記憶するとともに、出力制御部135に特定情報を出力する。

[0051] 出力制御部135は、特定部134から特定情報が入力されると、対象記憶部121及び信頼度記憶部124を参照して、取得部131で抽出した項目及びその値を、特定した信頼度とともに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S10）。これにより、データ検証装置100は、項目のデータを信頼度とともに出力できるので、信頼度の高い情報を、文書に記載すべき情報とすることができる。

[0052] 次に、第2の検証処理について説明する。図10は、第2の検証処理の一例を示すフローチャートである。

[0053] 取得部131は、クロール処理で巡回する対象URLを受け付ける（ステ

ップS 2 1)。取得部 1 3 1 は、受け付けた対象 URL を対象記憶部 1 2 1 に記憶する。取得部 1 3 1 は、例えば、定期的又は不定期に、つまり予め管理者によって指定された間隔又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（ステップ S 2 2）。取得部 1 3 1 は、対象記憶部 1 2 1 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 S 2 3）。

[0054] 取得部 1 3 1 は、選定した対象 URL からページ内容を取得する（ステップ S 2 4）。取得部 1 3 1 は、取得したページ内容をページ記憶部 1 2 2 に記憶する。取得部 1 3 1 は、ページ記憶部 1 2 2 に記憶された対象 URL のページ内容から、項目及びその値を抽出する（ステップ S 2 5）。取得部 1 3 1 は、対象 URL の URL ID ごとに、項目及びその値を対応付けて、抽出データ記憶部 1 2 3 に記憶する（ステップ S 2 6）。取得部 1 3 1 は、抽出した項目及びその値を抽出データ記憶部 1 2 3 に記憶すると、算出部 1 3 2 に、抽出完了情報を出力する。

[0055] 算出部 1 3 2 は、取得部 1 3 1 から抽出完了情報が入力されると、抽出データ記憶部 1 2 3 を参照して、対象 URL の値ごとに信頼度の要素に基づいて信頼度を算出する（ステップ S 2 7）。算出部 1 3 2 は、算出した信頼度を項目、URL ID 及び値と対応付けて信頼度記憶部 1 2 4 に記憶する（ステップ S 2 8）。算出部 1 3 2 は、算出した信頼度を信頼度記憶部 1 2 4 に記憶すると、算出完了情報を分類部 1 3 3 に出力する。

[0056] 分類部 1 3 3 は、算出部 1 3 2 から算出完了情報が入力されると、信頼度記憶部 1 2 4 を参照して、各項目の値について、同一の内容を有する値をグループに分類する（ステップ S 2 9）。また、分類部 1 3 3 は、値のグループの各サイトの値の信頼度の合計を算出する。分類部 1 3 3 は、項目ごとに、値のグループと、サイトの数と、各サイトの値の信頼度の合計とを対応付けて、分類結果としてグループ記憶部 1 2 5 に記憶する。分類部 1 3 3 は、分類結果をグループ記憶部 1 2 5 に記憶すると、分類完了情報を出力制御部 1 3 5 に出力する。

- [0057] 出力制御部 135 は、分類部 133 から分類完了情報が入力されると、グループの構成要素の数が最も多いグループの値を、グループの構成要素の数に応じた信頼度とともに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 S30）。これにより、データ検証装置 100 は、グループの構成要素の数が最も多いグループの値を信頼度とともに出力できるので、多種類の情報のうち、掲載するサイト数が最も多い情報を信頼度の高い情報として、文書に記載すべき情報とすることができる。
- [0058] 続いて、第 3 の検証処理について説明する。図 11 は、第 3 の検証処理の一例を示すフローチャートである。
- [0059] 取得部 131 は、クロール処理で巡回する対象 URL を受け付ける（ステップ S41）。取得部 131 は、受け付けた対象 URL を対象記憶部 121 に記憶する。取得部 131 は、例えば、定期的又は不定期に、つまり予め管理者によって指定された間隔又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（ステップ S42）。取得部 131 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 S43）。
- [0060] 取得部 131 は、選定した対象 URL からページ内容を取得する（ステップ S44）。取得部 131 は、取得したページ内容をページ記憶部 122 に記憶する。取得部 131 は、ページ記憶部 122 に記憶された対象 URL のページ内容から、項目及びその値を抽出する（ステップ S45）。取得部 131 は、対象 URL の URL ID ごとに、項目及びその値を対応付けて、抽出データ記憶部 123 に記憶する（ステップ S46）。取得部 131 は、抽出した項目及びその値を抽出データ記憶部 123 に記憶すると、算出部 132 に、抽出完了情報を出力する。
- [0061] 算出部 132 は、取得部 131 から抽出完了情報が入力されると、抽出データ記憶部 123 を参照して、対象 URL の値ごとに信頼度の要素に基づいて信頼度を算出する（ステップ S47）。算出部 132 は、算出した信頼度を項目、URL ID 及び値と対応付けて信頼度記憶部 124 に記憶する（ス

テップS 4 8)。算出部 1 3 2 は、算出した信頼度を信頼度記憶部 1 2 4 に記憶すると、算出完了情報を分類部 1 3 3 に出力する。

[0062] 分類部 1 3 3 は、最も高い信頼度が割り当てられた第 1 のサイトを除く他のサイトの値をグループに分類する（ステップS 4 9）。分類部 1 3 3 は、他のサイトの値について、項目ごとに、値のグループと、サイトの数と、各サイトの値の信頼度の合計とを対応付けて、分類結果としてグループ記憶部 1 2 5 に記憶する。分類部 1 3 3 は、分類結果をグループ記憶部 1 2 5 に記憶すると、分類完了情報を出力制御部 1 3 5 に出力する。

[0063] 出力制御部 1 3 5 は、分類部 1 3 3 から分類完了情報が入力されると、グループの構成要素の数が最も多い特定のグループの構成要素の数に応じた信頼度と、第 1 のサイトの信頼度とを比較する（ステップS 5 0）。出力制御部 1 3 5 は、比較の結果、特定のグループの構成要素の数に応じた信頼度が、第 1 のサイトの信頼度より高いか否かを判定する（ステップS 5 1）。

[0064] 出力制御部 1 3 5 は、特定のグループの構成要素の数に応じた信頼度が、第 1 のサイトの信頼度より高い場合には（ステップS 5 1：肯定）、項目及び特定のグループに属する値を出力データとして出力部 1 0 2 に出力して表示させる（ステップS 5 2）。出力制御部 1 3 5 は、特定のグループの構成要素の数に応じた信頼度が、第 1 のサイトの信頼度以下である場合には（ステップS 5 1：否定）、第 1 のサイトの項目及びその値を出力データとして出力部 1 0 2 に出力して表示させる（ステップS 5 3）。これにより、データ検証装置 1 0 0 は、例えば、信頼度は高いが更新頻度が低い第 1 のサイトと、信頼度は第 1 のサイトより低いが更新頻度が高い他のサイトとを比較することで、より信頼度の高い情報を出力できる。

[0065] すなわち、データ検証装置 1 0 0 は、例えば、ある観光スポットの電話番号が変更となった場合に、観光スポットの公式サイトであるが更新が遅い第 1 のサイトでは、電話番号の変更が反映されていない場合がある。データ検証装置 1 0 0 は、例えば、電話番号の更新頻度が高い他のサイトから新しい電話番号を取得することで、当該観光スポットの電話番号について、変更が

反映された、より信頼度の高い情報を出力することができる。

[0066] このように、データ検証装置 100 は、特定の情報の取得元として登録された複数のサイトの内、特定の情報として情報が取得されたサイトの信頼度を、複数のサイトのそれぞれに対応付けて信頼度を記憶する記憶部 120 の記憶内容に基づいて特定する。また、データ検証装置 100 は、特定の情報として取得された情報を特定した信頼度とともに出力する。その結果、文書に記載すべき情報の信頼度を高くして出力できる。

[0067] また、データ検証装置 100 は、特定の情報の取得元として登録された複数のサイトから特定の情報として情報を取得する。また、データ検証装置 100 は、取得した情報を情報の内容の一致性に応じてグルーピングした場合に、グループの構成要素の数が最も多いグループに属する情報を、グループの構成要素の数に応じた信頼度とともに出力する。その結果、多種類の情報のうち、掲載するサイト数が最も多い情報を信頼度の高い情報として、文書に記載すべき情報とすることができる。

[0068] また、データ検証装置 100 は、特定の情報の取得元として登録された複数のサイトから特定の情報として情報を取得する。また、データ検証装置 100 は、特定の情報として情報が取得された複数のサイトのうち最も高い信頼度が割り当てられた第 1 のサイトの信頼度に対して、複数のサイトのうち少なくとも第 1 のサイトを除く他のサイトから取得した情報を情報の内容の一致性に応じてグルーピングする。また、データ検証装置 100 は、グループの構成要素の数が最も多い特定のグループの構成要素の数に応じて算出した信頼度が、より高い信頼性を示すことを検出すると、特定のグループに属する情報を出力する。その結果、情報の変更が反映された、より信頼度の高い情報を出力することができる。

[0069] なお、上記実施例では、対象 URL のホームページを受け付けて巡回し、対象 URL、各項目及びその値の信頼度を算出したが、これに限定されない。例えば、定期的に対象 URL のホームページを巡回して対象 URL、各項目及びその値の信頼度を算出し、データ検証装置 100 の管理者の要求に応

じて、対象URL、各項目及びその値の信頼度を出力部102に表示させるようにしてもよい。

[0070] 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文書から各項目及びその値を抽出し、それぞれの信頼度を出力部102に表示させるようにしてもよい。

[0071] 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、その全部又は一部を、各種の負荷や使用状況等に応じて、任意の単位で機能的又は物理的に分散・統合して構成することができる。例えば、算出部132と、分類部133とを統合してもよい。

[0072] さらに、各装置で行われる各種処理機能は、CPU（又はMPU、MCU（Micro Controller Unit）等のマイクロ・コンピュータ）上で、その全部又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（又は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、又はワイヤードロジックによるハードウェア上で、その全部又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。

[0073] 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12は、データ検証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。

[0074] 図12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又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 RAM 207 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。

[0075] ハードディスク装置 208 には、図 1 に示した取得部 131、算出部 132、分類部 133、特定部 134 及び出力制御部 135 の各処理部と同様の機能を有するデータ検証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、ページ記憶部 122、抽出データ記憶部 123、信頼度記憶部 124、グループ記憶部 125、及び、データ検証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、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、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。

[0076] CPU 201 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、RAM 207 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 に示した取得部 131、算出部 132、分類部 133、特定部 134 及び出力制御部 135 として機能させることができる。

[0077] なお、上記のデータ検証プログラムは、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、LA

N等に接続された装置にこのデータ検証プログラムを記憶させておき、コンピュータ200がこれらからデータ検証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

- [0078] 100 データ検証装置
- 101 入力部
- 102 出力部
- 110 通信部
- 120 記憶部
- 121 対象記憶部
- 122 ページ記憶部
- 123 抽出データ記憶部
- 124 信頼度記憶部
- 125 グループ記憶部
- 130 制御部
- 131 取得部
- 132 算出部
- 133 分類部
- 134 特定部
- 135 出力制御部
- N ネットワーク

請求の範囲

- [請求項1] 特定の情報の取得元として登録された複数のサイトの内、前記特定の情報として情報が取得されたサイトの信頼度を、前記複数のサイトのそれぞれに対応付けて信頼度を記憶する記憶部の記憶内容に基づいて特定し、
- 前記特定の情報として取得された前記情報を前記特定した前記信頼度とともに出力する
- 処理をコンピュータに実行させることを特徴とするデータ検証プログラム。
- [請求項2] 特定の情報の取得元として登録された複数のサイトから前記特定の情報として情報を取得し、
- 取得した情報を情報の内容の一致性に応じてグルーピングした場合に、グループの構成要素の数が最も多いグループに属する情報を、グループの構成要素の数に応じた信頼度とともに出力する
- 処理をコンピュータに実行させることを特徴とするデータ検証プログラム。
- [請求項3] 特定の情報の取得元として登録された複数のサイトから前記特定の情報として情報を取得し、
- 前記特定の情報として情報が取得された前記複数のサイトのうち最も高い信頼度が割り当てられた第1のサイトの信頼度に対して、前記複数のサイトのうち少なくとも前記第1のサイトを除く他のサイトから取得した情報を情報の内容の一致性に応じてグルーピングし、グループの構成要素の数が最も多い特定のグループの構成要素の数に応じて算出した信頼度が、より高い信頼性を示すことを検出すると、前記特定のグループに属する情報を出力する
- 処理をコンピュータに実行させることを特徴とするデータ検証プログラム。
- [請求項4] 特定の情報の取得元として登録された複数のサイトの内、前記特定

の情報として情報が取得されたサイトの信頼度を、前記複数のサイトのそれぞれに対応付けて信頼度を記憶する記憶部の記憶内容に基づいて特定し、

前記特定の情報として取得された前記情報を前記特定した前記信頼度とともに出力する

処理をコンピュータが実行することを特徴とするデータ検証方法。

[請求項5]

特定の情報の取得元として登録された複数のサイトから前記特定の情報として情報を取得し、

取得した情報を情報の内容の一致性に応じてグルーピングした場合に、グループの構成要素の数が最も多いグループに属する情報を、グループの構成要素の数に応じた信頼度とともに出力する

処理をコンピュータが実行することを特徴とするデータ検証方法。

[請求項6]

特定の情報の取得元として登録された複数のサイトから前記特定の情報として情報を取得し、

前記特定の情報として情報が取得された前記複数のサイトのうち最も高い信頼度が割り当てられた第1のサイトの信頼度に対して、前記複数のサイトのうち少なくとも前記第1のサイトを除く他のサイトから取得した情報を情報の内容の一致性に応じてグルーピングし、グループの構成要素の数が最も多い特定のグループの構成要素の数に応じて算出した信頼度が、より高い信頼性を示すことを検出すると、前記特定のグループに属する情報を出力する

処理をコンピュータが実行することを特徴とするデータ検証方法。

[請求項7]

特定の情報の取得元として登録された複数のサイトの内、前記特定の情報として情報が取得されたサイトの信頼度を、前記複数のサイトのそれぞれに対応付けて信頼度を記憶する記憶部の記憶内容に基づいて特定する特定部と、

前記特定の情報として取得された前記情報を前記特定した前記信頼度とともに出力する出力制御部と

を有することを特徴とするデータ検証装置。

[請求項8]

特定の情報の取得元として登録された複数のサイトから前記特定の
情報として情報を取得する取得部と、

取得した情報を情報の内容の一致性に応じてグルーピングした場合
に、グループの構成要素の数が最も多いグループに属する情報を、グ
ループの構成要素の数に応じた信頼度とともに出力する出力制御部と
を有することを特徴とするデータ検証装置。

[請求項9]

特定の情報の取得元として登録された複数のサイトから前記特定の
情報として情報を取得する取得部と、

前記特定の情報として情報が取得された前記複数のサイトのうち最
も高い信頼度が割り当てられた第1のサイトの信頼度に対して、前記
複数のサイトのうち少なくとも前記第1のサイトを除く他のサイトか
ら取得した情報を情報の内容の一致性に応じてグルーピングし、グ
ループの構成要素の数が最も多い特定のグループの構成要素の数に応じ
て算出した信頼度が、より高い信頼性を示すことを検出すると、前記
特定のグループに属する情報を出力する出力制御部と

を有することを特徴とするデータ検証装置。