

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2010-15202

(P2010-15202A)

(43) 公開日 平成22年1月21日(2010.1.21)

(51) Int.Cl.
G06F 17/30 (2006.01)F I
G O 6 F 17/30 2 1 0 Aテーマコード (参考)
5 B 0 7 5

審査請求 有 請求項の数 7 O L (全 18 頁)

(21) 出願番号 特願2008-171883 (P2008-171883)
(22) 出願日 平成20年6月30日 (2008. 6. 30)(71) 出願人 500257300
ヤフー株式会社
東京都港区赤坂9丁目7番1号
(74) 代理人 100106002
弁理士 正林 真之
(72) 発明者 山本 健一
東京都港区六本木六丁目10番1号 ヤフー株式会社内
Fターム(参考) 5B075 NK24 NK31 NR02 NR05

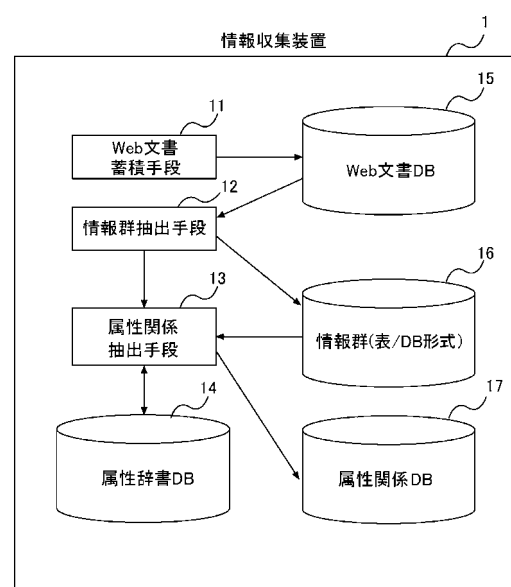
(54) 【発明の名称】 情報収集方法、装置及びプログラム

(57) 【要約】

【課題】 Web文書に含まれるタグに基づいて、項目、属性及び属性値の関係を有する情報を抽出することにより、項目に関する情報をWeb文書から自動的に収集する。

【解決手段】 情報収集装置1は、通信ネットワークNを介してアクセス可能なWeb文書から、Web文書に含まれるタグに基づいて表形式あるいはデータベース形式の情報を抽出する情報群抽出手段12と、抽出した表形式あるいはデータベース形式の情報から、前記タグが示す各情報間の従属関係に基づいて、所定の項目に対して従属する関係にある属性及び前記属性の内容を示す属性値の関係を有する情報を抽出し、抽出した前記項目、属性及び属性値の関係を有する情報を関連付けて記憶手段に記憶する属性関係抽出手段13と、を備える。

【選択図】 図2



【特許請求の範囲】**【請求項 1】**

情報収集装置が、通信ネットワークを介してアクセス可能なWeb文書から、前記Web文書に含まれるタグに基づいて表形式あるいはデータベース形式の情報を抽出するステップと、

抽出した前記表形式あるいはデータベース形式の情報から、前記タグが示す各情報間の従属関係に基づいて、所定の項目に対して従属する関係にある属性及び前記属性の内容を示す属性値の関係を有する情報を抽出するステップと、

抽出した前記項目、属性及び属性値の関係を有する情報を関連付けて記憶手段に記憶するステップと、を少なくとも実行することを特徴とする情報収集方法。

10

【請求項 2】

前記抽出するステップは、前記表形式の情報の直上方、直下方あるいは直左方に位置する情報を前記項目として抽出する請求項 1 に記載の方法。

【請求項 3】

前記抽出するステップは、前記表形式の情報において、上端行あるいは左端列に位置する情報を属性として、それぞれその下方あるいは右方に位置する情報を属性値として抽出する請求項 1 または請求項 2 に記載の方法。

【請求項 4】

前記抽出するステップは、前記データベース形式の情報の直左方に位置する情報を属性として、前記データベース形式の情報をその属性値として抽出する請求項 1 から請求項 3 のいずれかに記載の方法。

20

【請求項 5】

前記表形式あるいはデータベース形式の情報を抽出する際に基づく前記タグが、ブルダウニストを形成するHTMLタグ、XML文書に含まれるタグ又はHTMLのテーブルタグである請求項 1 から請求項 4 のいずれかに記載の方法。

【請求項 6】

請求項 1 から請求項 5 のいずれかに記載の方法をコンピュータに実行させることを特徴としたプログラム。

【請求項 7】

通信ネットワークを介してアクセス可能なWeb文書から、前記Web文書に含まれるタグに基づいて表形式あるいはデータベース形式の情報を抽出する情報群抽出手段と、

30

抽出した前記表形式あるいはデータベース形式の情報から、前記タグが示す各情報間の従属関係に基づいて、所定の項目に対して従属する関係にある属性及び前記属性の内容を示す属性値の関係を有する情報を抽出し、抽出した前記項目、属性及び属性値の関係を有する情報を関連付けて記憶手段に記憶する属性関係抽出手段と、を備えたことを特徴とする情報収集装置。

【発明の詳細な説明】**【技術分野】****【0001】**

本発明は、Web文書に記載された情報を収集する情報収集方法、装置及びプログラムに関する。

40

【背景技術】**【0002】**

従来、ある特定の項目に関しWeb文書から情報を得たい場合、Web文書を一つ一つ閲覧し、情報を収集しなければならなかった。

【0003】

そこで、パソコンに関する情報を調べたい場合には、非特許文献 1 に記載の技術によれば、一覧表等に表示する情報に基づいて、ユーザは、様々なメーカー、販売会社毎にパソコンに関する情報を探索することができるようになっている。

【非特許文献 1】[online]、[平成 20 年 6 月 18 日検索]、インターネット

50

URL : http : / / k a k a k u . c o m / p c / d e s k t o p - p c /

【発明の開示】

【発明が解決しようとする課題】

【0004】

しかしながら、非特許文献1に記載の技術によっても、パソコンに関する情報を人手により収集、蓄積して再編集しているために、その作業に要する時間と労力は膨大なものとなる。また、非特許文献1に記載の技術は、特定の商品について提供されるものであり、取り扱いのない商品や、商品以外のものについて調べようすると、蓄積情報のカバレッジの面で限界がある。

【0005】

10

そこで、本発明は、このような事情を考慮して提案されるものであり、Webページ上に分散して存在している共通の項目やその属性および属性値の関係にある情報を自動的に収集する情報収集方法、装置及びプログラムを提供することを目的とする。

【課題を解決するための手段】

【0006】

上記目的を達成するために、本発明者は、項目に関する情報を複数のWeb文書から自動的に収集する仕組みを見出し、本発明を想到するに至った。

【0007】

本発明に係る情報収集方法は、Web文書に含まれるタグに基づいて、項目、属性及び属性値の関係を有する情報を抽出することにより、項目に関する情報をWeb文書から自動的に収集するものである。

20

【0008】

(1) 情報収集装置が、通信ネットワークを介してアクセス可能なWeb文書から、前記Web文書に含まれるタグに基づいて表形式あるいはデータベース形式の情報を抽出するステップと、

抽出した前記表形式あるいはデータベース形式の情報から、前記タグが示す各情報間の従属関係に基づいて、所定の項目に対して従属する関係にある属性及び前記属性の内容を示す属性値の関係を有する情報を抽出するステップと、

抽出した前記項目、属性及び属性値の関係を有する情報を関連付けて記憶手段に記憶するステップと、を少なくとも実行することを特徴とする情報収集方法。

30

【0009】

(1)に記載の発明の構成によれば、Web文書に含まれるタグに基づいて表形式あるいはデータベース形式の情報を抽出し、抽出した表形式あるいはデータベース形式の情報から、タグが示す各情報間の従属関係に基づいて、所定の項目に対して従属する関係にある属性及び当該属性の内容を示す属性値の関係を有する情報を抽出する。

【0010】

このことにより、項目、属性及び属性値の関係を有する情報を複数のWeb文書から自動的に収集することが可能となる。

【0011】

ここで、前記表形式の情報とは、カンマ区切り、スペース区切りなどにより表と同等の表現がされた平文により示された情報も含む。

40

【0012】

また、所定の項目に対して、より多くのWeb文書において共通して従属する情報を、より大きなウェイトを掛けて属性及び属性値として抽出してもよい。また、所定の項目に対して、所定の閾値を超える数のWeb文書において共通して従属することがない情報は、当該属性及び属性値として抽出しないこととしてもよい。

【0013】

ここで、抽出して記憶した属性及び属性値は、項目に関するWeb文書を検索する際の検索クエリーの示唆等に活用することができる。

【0014】

50

(2) 前記抽出するステップは、前記表形式の情報の直上方、直下方あるいは直左方に位置する情報を前記項目として抽出する(1)に記載の方法。

【0015】

(2)に記載の発明の構成によれば、表形式の情報の直上方、直下方あるいは直左方に位置する情報を前記項目として抽出する。

【0016】

このことにより、Web文書において表形式の情報のタイトルが表示されることが多い位置に位置する情報を当該表形式の情報に対する項目として抽出することができる。

【0017】

ここで、上述の様に、当該関係を有する頻度に応じて、ウェイトを掛けたり、当該頻度が所定の閾値に達するまで、当該抽出を行なわないこととして、精度を上げてよい。

【0018】

(3) 前記抽出するステップは、前記表形式の情報において、上端行あるいは左端列に位置する情報を属性として、それぞれその下方あるいは右方に位置する情報を属性値として抽出する(1)または(2)に記載の方法。

【0019】

(3)に記載の発明の構成によれば、前記表形式の情報において、上端行あるいは左端列に位置する情報を属性として、それぞれその下方あるいは右方に位置する情報を属性値として抽出する。

【0020】

このことにより、Web文書において表形式の情報の属性が表示されることが多い位置に位置する情報を属性として、その属性の内容が表示されることが多い位置に位置する情報を属性値として抽出することができる。

【0021】

ここで、上述の様に、当該関係を有する頻度に応じて、ウェイトを掛けたり、当該頻度が所定の閾値に達するまで、当該抽出を行なわないこととして、精度を上げてよい。

【0022】

(4) 前記抽出するステップは、前記データベース形式の情報の直左方に位置する情報を属性として、前記データベース形式の情報をその属性値として抽出する(1)から(3)のいずれかに記載の方法。

【0023】

(4)に記載の発明の構成によれば、データベース形式の情報の直左方に位置する情報を属性として、前記データベース形式の情報をその属性値として抽出する。

【0024】

このことにより、Web文書においてデータベース形式の情報の属性が表示されることが多い位置に位置する情報を属性として、その属性の内容が表示されることが多いデータベース形式の情報を属性値として抽出することができる。

【0025】

ここで、上述の様に、当該関係を有する頻度に応じて、ウェイトを掛けたり、当該頻度が所定の閾値に達するまで、当該抽出を行なわないこととして、精度を上げてよい。

【0026】

(5) 前記表形式あるいはデータベース形式の情報を抽出する際に基づく前記タグが、ブルダウンリストを形成するHTML(HyperText Markup Language)タグ、XML(Extensible Markup Language)文書に含まれるタグ又はHTMLのテーブルタグである(1)から(4)のいずれかに記載の方法。

【0027】

(5)に記載の発明の構成によれば、ブルダウンリストを形成するHTMLタグ、XML文書に含まれるタグ又はHTMLのテーブルタグに基づいて表形式あるいはデータベース形式の情報と判断するので、Web文書に含まれるタグの一致を判定することにより自

10

20

30

40

50

動的に表形式あるいはデータベース形式の情報を抽出することができる。

【 0 0 2 8 】

(6) (1) から (5) のいずれかに記載の方法をコンピュータに実行させることを特徴としたプログラム。

【 0 0 2 9 】

(7) 通信ネットワークを介してアクセス可能な W e b 文書から、前記 W e b 文書に含まれるタグに基づいて表形式あるいはデータベース形式の情報を抽出する情報群抽出手段と、

抽出した前記表形式あるいはデータベース形式の情報から、前記タグが示す各情報間の従属関係に基づいて、所定の項目に対して従属する関係にある属性及び前記属性の内容を示す属性値の関係を有する情報を抽出し、抽出した前記項目、属性及び属性値の関係を有する情報を関連付けて記憶手段に記憶する属性関係抽出手段と、を備えたことを特徴とする情報収集装置。

【 発明の効果 】

【 0 0 3 0 】

この発明によれば、 W e b 文書に含まれるタグに基づいて、項目、属性及び属性値の関係を有する情報を抽出することにより、 W e b ページ上に分散して存在している共通の項目やその属性および属性値の関係にある情報を自動的に収集することができる。

【 発明を実施するための最良の形態 】

【 0 0 3 1 】

以下、本発明を実施するための最良の形態について図を参照しながら説明する。なお、これはあくまでも一例であって、本発明の技術的範囲はこれに限られるものではない。

[情報収集装置と関連要素の全体構成]

【 0 0 3 2 】

図 1 において、情報収集装置 1 は、通信ネットワークとしてのインターネット N を通じて複数の W e b サーバ装置 2 に接続される。インターネット N との接続は、有線であるか無線であるかを問わない。

【 0 0 3 3 】

情報収集装置 1 は、複数の W e b サーバ装置 2 から W e b 文書を取得する。各 W e b サーバ装置 2 は、情報収集装置 1 からのリクエストに応じて種々の W e b 文書を提供する。

[情報収集装置の機能構成]

【 0 0 3 4 】

図 2 は、本実施形態に係る情報収集装置 1 の機能構成の概要を示す図である。情報収集装置 1 は、 W e b 文書蓄積手段 1 1、情報群抽出手段 1 2 及び属性関係抽出手段 1 3 を備えている。また、 W e b 文書 D B 1 5、情報群記憶部 1 6 及び属性関係 D B 1 7 を有する (D B はデータベースの略) 。

【 0 0 3 5 】

W e b 文書蓄積手段 1 1 は、 W e b サーバ装置 2 から W e b 文書を取得し W e b 文書 D B 1 5 に格納する。情報群抽出手段 1 2 は、 W e b 文書 D B 1 5 に蓄積された W e b 文書を読み出し、読み出した当該 W e b 文書に含まれるタグに基づき、表形式又はデータベース形式の情報を抽出し情報群記憶部 1 6 に格納する。属性関係抽出手段 1 3 は、情報群記憶部 1 6 に格納された表形式又はデータベース形式の情報を読み出し、読み出した当該情報に含まれるタグに基づき、当該情報に含まれる項目、属性及び属性値の関係を有する情報を抽出し、抽出した当該情報を属性関係 D B 1 7 に登録する。 W e b 文書蓄積手段 1 1、情報群抽出手段 1 2 及び属性関係抽出手段 1 3 は、コンピュータがプログラムを実行することによって実現される。

【 0 0 3 6 】

また、 W e b 文書 D B 1 5、情報群記憶部 1 6 及び属性関係 D B 1 7 は、後述のハードウェアに含まれる記憶装置 4 1 0 の一領域に設けられている。

[各種データベースと関連要素の構成]

【 0 0 3 7 】

図 3 は、W e b 文 書 D B 1 5、情 報 群 記 憶 部 1 6 及 び 属 性 関 係 D B 1 7 と 関 連 要 素 の 構 成 の 概 要 を 示 す 図 で あ る。

【 0 0 3 8 】

図 3 (a) に 示 す よ う に、W e b 文 書 D B 1 5 は、文 書 I D と、通 信 ネットワーク上で配信されているW e b 文 書 の U R L 等 の ネットワーク上のアドレス及びこのW e b 文 書 の記述であるソースコードと、をそれぞれ対応付けて記憶している。

【 0 0 3 9 】

図 3 (b) 及 び (c) に 示 す よ う に、情 報 群 記 憶 部 1 6 に は、W e b 文 書 の ソースコードから抽出された表形式又はデータベース形式の情報が当該表形式又はデータベース形式を構成するタグと共に格納される。

10

【 0 0 4 0 】

図 3 (e) に 示 す よ う に、属 性 関 係 D B に は、表 形 式 又 は データベース形式の情報から抽出された項目、属性及び属性値の関係を有する情報が格納される。

【 0 0 4 1 】

図 3 (d) に 示 す よ う に、本 実 施 形 態 で は、表 形 式 又 は データベース形式の情報から項目、属性及び属性値の関係を有する情報を抽出する際に利用する属性辞書 1 4 を備えている。

【 0 0 4 2 】

図 3 (b) 及 び (c) は、W e b 文 書 から 抽 出 さ れ る 表 形 式 又 は データベース形式の情報の例である。例えば、P C 販 売 会 社 A の W e b ページ (W e b 文 書) が、図 3 (b) に示す表形式又はデータベース形式の情報を含んでおり、P C 販 売 会 社 B の W e b ページ (W e b 文 書) が、図 3 (c) に示す表形式又はデータベース形式の情報を含んでいるものとする。

20

【 0 0 4 3 】

当該情報は、ブルダウンリストを形成する一群のH T M L として記述されている場合、X M L 文 書 として記述されている場合、テーブルタグにより表を形成する一群のH T M L として記述されている場合など、種々考えられる。

【 0 0 4 4 】

ブルダウンリストを形成する一群のH T M L として記載されている場合は、例えば < S e l e c t > タグの開始タグと終了タグを判定し、当該開始タグおよび終了タグとその間にある要素の内容とを抽出することが考えられる。

30

【 0 0 4 5 】

また、X M L 文 書 として記載されている場合は、X M L インスタンスが情報の階層構造を持つので、例えば、X M L インスタンスの最上位の開始タグと終了タグとの間にある要素の内容を抽出することが考えられる。

【 0 0 4 6 】

また、テーブルタグにより表を形成する一群のH T M L として記述されている場合は、例えば、< T a b l e > タグの開始タグと終了タグを判定し、当該開始タグおよび終了タグとその間にある要素の内容とを抽出することが考えられる。

40

【 0 0 4 7 】

本 実 施 形 態 に お い て、図 3 (b) の 情 報 は、テーブルタグにより、1 列目に項目「ノートP C」が記述され、1 行目にノートP C の属性として「C P U」、「クロック」が記述され、「C P U x x x」、「C P U y y y」が属性「C P U」の属性値として記述され、「1 . 5 G H z」、「2 . 0 G H z」が属性「クロック」の属性値として記述されているものとする。

【 0 0 4 8 】

なお、項目、属性及び属性値の関係がこれらのW e b 文 書 の 基 礎 と な る データを格納しているデータベースサーバ等に蓄積され、C G I 等のプログラムを用いて取得する構造となっている場合には、これらのC G I 等のプログラムを実行することにより、これらの関

50

係を有する情報を収集する。

【 0 0 4 9 】

また、本実施形態において、図 3 (c) の情報は、テーブルタグにより形成されているものとし、< T a b l e > タグの要素内容として記述された< C a p t i o n > タグの要素内容の一部に、項目「ノートパソコン」が記述され、その直後の表の 1 行目に、項目「ノートパソコン」の属性として「C P U」、「クロック」が記述され、以降、属性「C P U」の属性値として、「C P U z z z」、「C P U p p p」が記述され、属性「クロック」の属性値として、「8 0 0 M H z」、「3 . 2 G H z」が記述されているものとする。

【 0 0 5 0 】

図 3 (d) は、属性関係抽出手段が利用する属性辞書 1 4 の例である。属性辞書 1 4 は、後述のハードウェアに含まれる記憶装置 4 1 0 に格納されている。本実施形態において、属性辞書 1 4 は、項目と属性とを関連付けている。例えば、項目には「ノートパソコン」のほか、その類義語である「ノート P C」等が登録されている。一方、項目に関連する属性として「C P U」「H D D」「バッテリー」「価格」等が登録されている。更に、属性（属性 1）に関する下位の属性（属性 2）が登録されている。例えば、属性 1「C P U」に関し、属性 2「クロック」「キャッシュ」等が登録されている。

【 0 0 5 1 】

図 3 (e) は、属性関係 D B に格納される項目、属性及び属性値の関係を有する情報の例を示している。ユニークな番号である属性関係 I D ごとに、項目、属性及び属性値を関連付けて保存している。例えば、ある項目「ノートパソコン」は、属性 1「C P U」の属性値が「C P U x x x」であり、かつ、属性 2「クロック」の属性値が「1 . 5 G H z」であるとして保存される。

[情報収集装置のハードウェア構成図]

【 0 0 5 2 】

図 4 は、本実施形態に係る情報収集装置 1 のハードウェア構成を示す図である。情報収集装置 1 は、制御部 3 0 0 を構成する C P U (C e n t r a l P r o c e s s i n g U n i t) 3 1 0 (マルチプロセッサ構成では C P U 3 2 0 等複数の C P U が追加されてもよい)、バスライン 2 0 0、通信 I / F (I / F : インタフェース) 3 3 0、メインメモリ 3 4 0、B I O S (B a s i c I n p u t O u t p u t S y s t e m) 3 5 0、I / O コントローラ 3 6 0、ハードディスク 3 7 0、光ディスクドライブ 3 8 0、並びに半導体メモリ 3 9 0 を備える。尚、ハードディスク 3 7 0、光ディスクドライブ 3 8 0、並びに、半導体メモリ 3 9 0 はまとめて記憶装置 4 1 0 と呼ばれる。

【 0 0 5 3 】

制御部 3 0 0 は、情報収集装置 1 を統括的に制御する部分であり、ハードディスク 3 7 0 (後述) に記憶された各種プログラムを適宜読み出して実行することにより、上述したハードウェアと協働し、本発明に係る各種機能を実現している。

【 0 0 5 4 】

通信 I / F 3 3 0 は、情報収集装置 1 が、インターネット N (図 1) を介して W e b サーバ装置 2 (# 1) ~ 2 (# N) 等 (図 1) と情報を送受信する場合のネットワーク・アダプタである。通信 I / F 3 3 0 は、モデム、ケーブル・モデム及びイーサネット (登録商標) ・アダプタを含んでよい。

【 0 0 5 5 】

B I O S 3 5 0 は、情報収集装置 1 の起動時に C P U 3 1 0 が実行するブートプログラムや、情報収集装置 1 がハードウェアに依存するプログラム等を記録する。

【 0 0 5 6 】

I / O コントローラ 3 6 0 には、ハードディスク 3 7 0、光ディスクドライブ 3 8 0、及び半導体メモリ 3 9 0 等の記憶装置 4 1 0 を接続することができる。

【 0 0 5 7 】

ハードディスク 3 7 0 は、本ハードウェアを情報収集装置 1 として機能させるための各

10

20

30

40

50

種プログラム、本発明の機能を実行するプログラム及び前述の各DB15、17、情報群記憶部16及び属性辞書14を記憶する。なお、情報収集装置1は、外部に別途設けたハードディスク（図示せず）を外部記憶装置として利用することもできる。

【0058】

光ディスクドライブ380としては、例えば、DVD-ROMドライブ、CD-ROMドライブ、DVD-RAMドライブ、CD-RAMドライブを使用することができる。この場合は各ドライブに対応した光ディスク400を使用する。光ディスク400から光ディスクドライブ380によりプログラムまたはデータを読み取り、I/Oコントローラ360を介してメインメモリ340またはハードディスク370に提供することもできる。

【0059】

なお、本発明でいうコンピュータとは、記憶装置、制御部等を備えた情報処理装置をいい、情報収集装置1は、記憶装置410、制御部300等を備えた情報処理装置により構成され、この情報処理装置は、本発明のコンピュータの概念に含まれる。

[Webサーバ装置のハードウェア構成]

【0060】

Webサーバ装置2も、上述の情報収集装置1と同様なハードウェア構成を持つ。

[本発明の実施形態に係るフローチャート]

【0061】

図5は、本発明の実施形態に係る情報収集処理のフローチャートを示している。

【0062】

S1: Web文書蓄積手段11は、ネットワークN上に配信されているWeb文書を任意のWebサーバ装置2からダウンロードし、このWeb文書のURLをネットワーク上のアドレスとして、自動生成した文書IDとこの通信アドレスとを対応付けてWeb文書DB15に記憶する。また、Web文書蓄積手段11は、このWeb文書の文書IDと、このWeb文書の記述であるソースコードとを関連付けて記憶する。

【0063】

S2: 情報群抽出手段12は、Web文書DB15に蓄積されたWeb文書のソースコードを解析し、当該ソースコードに含まれるタグの記述に基づいて表形式又はデータベース形式の情報の存在を判定し、表形式又はデータベース形式の情報を見つけた場合は当該情報を抽出して情報群記憶部16に格納する。

【0064】

表形式又はデータベース形式の情報をタグに基づいて判定する方法としては、前述したように、ブルダウンリストを構成するタグを見つける方法、XML宣言を判定する方法、テーブルタグを見つける方法等が考えられる。

【0065】

本実施形態においては、例えば、図3(b)に示すPC販売会社AのWebページに掲載されていた情報と、図3(c)に示すPC販売会社BのWebページに掲載されていた情報とを抽出できたとする。図3(b)及び(c)の情報は、前述したようにテーブルタグによって構成されているものとする。

【0066】

S3: 属性関係抽出手段13は、情報群記憶部16に格納された表形式又はデータベース形式の情報を解析し、当該情報に含まれる項目、属性及び属性値の関係にある情報を抽出する。

【0067】

抽出にあたり、属性関係抽出手段13は、項目、属性及び属性値の関係にある情報の所在を推定する。推定の方法は幾つか考えられる。

(1) 表形式又はデータベース形式の情報が、<Select>タグによりブルダウンリストを形成している場合、例えば<Select>タグのname属性の値を項目であると推定し、同<Select>タグの要素内容に列記された<Option>タグの要素内容を当該項目に関する属性及び属性値であると推定することが可能である。例えば、

10

20

30

40

50

```
<select name = "ノートPC">
<option>CPU xxx 1.5GHz</option>
<option>CPU yyy 2.0GHz</option>
</select>
```

上記において、項目「ノートPC」、属性「CPU」の属性値「CPU xxx」と推定することができる。また、属性「CPU」に続く属性は「クロック」と推定し、属性「クロック」の属性値「1.5GHz」と推定することができる。同様に、項目「ノートパソコン」、属性「CPU」の属性値「CPU yyy」、属性「クロック」の属性値「2.0GHz」と推定することができる。

(2) 表形式又はデータベース形式の情報が、XMLインスタンスの場合、階層構造を成している上位のタグ要素名を「項目」と推定し、その一つ下位のタグ要素名を「属性」と推定し、当該「属性」を示すタグ要素名の属性又は要素内容を「属性値」と推定することが可能である。例えば、

```
<ノートパソコン>
  <CPU type = "CPU zzz">
    <クロック>1.1GHz</クロック>
  </CPU>
  <CPU type = "CPU ppp">
    <クロック>3.2GHz</クロック>
  </CPU>
</ノートパソコン>
```

上記において、項目「ノートパソコン」、属性「CPU」の属性値「CPU zzz」、属性「クロック」の属性値「1.1GHz」を推定することができる。同様に、項目「ノートパソコン」、属性「CPU」の属性値「CPU ppp」、属性「クロック」の属性値「3.2GHz」を推定することができる。

(3) 表形式又はデータベース形式の情報がテーブルタグによって構成されている場合、例えば、表のタイトルを「項目」と推定し、1行目にある要素の列を各「属性」と推定し、2行目以降にある要素を同列の属性に対応する「属性値」と推定することが考えられる。例えば、

```
<table>
  <caption>ノートパソコン</caption>
  <tr>
    <td>CPU</td>
    <td>クロック</td>
  </tr>
  <tr>
    <td>CPU zzz</td>
    <td>1.1GHz</td>
  </tr>
  <tr>
    <td>CPU ppp</td>
    <td>3.2GHz</td>
  </tr>
</table>
```

上記において、項目「ノートパソコン」、属性「CPU」の属性値「CPU zzz」、属性「クロック」の属性値「1.1GHz」を推定することができる。同様に、項目「ノートパソコン」、属性「CPU」の属性値「CPU ppp」、属性「クロック」の属性値「3.2GHz」を推定することができる。

【0068】

(4) 属性辞書14を利用する方法も考えられる。属性関係抽出手段13は、属性辞書

10

20

30

40

50

14に登録された項目を参照し、情報群記憶部16に格納された情報Aに同一の項目が含まれているか判定する。同一の項目が含まれていたら、属性辞書14においてその項目に関連付けられている属性を参照し、同一の属性が情報Aに含まれているか判定する。同一の属性が含まれていたら、情報Aにおいて当該属性の例えば直後にある要素内容を当該属性についての属性値であると推定する。属性辞書14を利用すると、項目や属性を表すテキストが不要な語句を一部に含んでいても、当該不要な語句を無視して項目名や属性名を取得することができる。

【0069】

なお、属性関係抽出手段13は、表形式の情報の直上方、直下方あるいは直左方に位置する情報を前記項目として抽出してもよい。また、表形式の情報において、上端行あるいは左端列に位置する情報を属性として、それぞれその下方あるいは右方に位置する情報を属性値として抽出してもよい。さらに、前記データベース形式の情報の直左方に位置する情報を属性として、前記データベース形式の情報をその属性値として抽出してもよい。

【0070】

S4：属性関係抽出手段13は、S3において抽出した項目、属性及び属性値の関係を有する情報を属性関係DB17に登録する。本実施形態において、属性関係抽出手段13は、この登録の際に属性辞書14を参照し、項目の類義語を1種類に統一する。例えば、図3(b)に示す情報から得た項目「ノートPC」は、図3(d)に示した属性辞書の項目を参照し「ノートパソコン」に統一して属性関係DB17に登録する。このようにすると、項目が統一されることによって、生成された属性関係DB17の情報を活用しやすくなる。もっとも、属性辞書14とは別に類義語辞書を備え、この類義語辞書を参照することにより、項目や属性の類義語を統一してもよい。図3に示した情報の例によると、図3(b)及び(c)に示した表形式又はデータベース形式の情報から、図3(e)に示した属性関係DBを生成することができる。

【0071】

以上説明したように、情報群抽出手段12及び属性関係抽出手段13が、Web文書に含まれるタグに基づいて、項目、属性及び属性値の関係を有する情報を抽出するので、項目に関する情報をサイトの異なる複数のWeb文書から自動的に収集し、属性ごとに整理された情報として取得することができる。

【0072】

以上、本発明の実施形態について説明したが、本発明は上述した実施形態に限るものではない。例えば、「項目」は本実施形態の例示に限られず、「属性」及び「属性値」を伴うものはすべて「項目」になり得る。また、表形式データベース形式の情報を抽出する方法及び当該情報から項目、属性及び属性値の関係を有する情報を抽出する方法は、本実施形態の例示に限られるものではない。また、図6に示した各ステップは、1つのWeb文書を蓄積するごとに全てのステップを一通り実行する必要はない。各ステップが非同期でバッチ処理を行なってもよい。

【0073】

また、本発明の実施形態に記載された効果は、本発明から生じる最も好適な効果を列挙したに過ぎず、本発明による効果は、本発明の実施例に記載されたものに限定されるものではない。

【図面の簡単な説明】

【0074】

【図1】本実施形態に係る情報収集装置と関連要素の全体構成を示す図である。

【図2】本実施形態に係る情報収集装置の機能構成の概要を示す図である。

【図3】本実施形態に係るWeb文書データベース、情報群抽出手段及び属性関係データベースと関連要素の構成の概要を示す図である。

【図4】本実施形態に係る情報収集装置のハードウェア構成を示す図である。

【図5】本発明の実施形態に係る情報収集処理のフローチャートを示している。

【符号の説明】

10

20

30

40

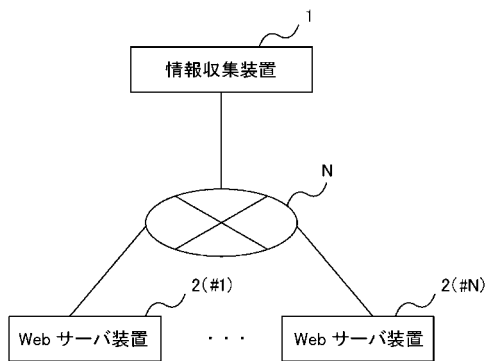
50

【 0 0 7 5 】

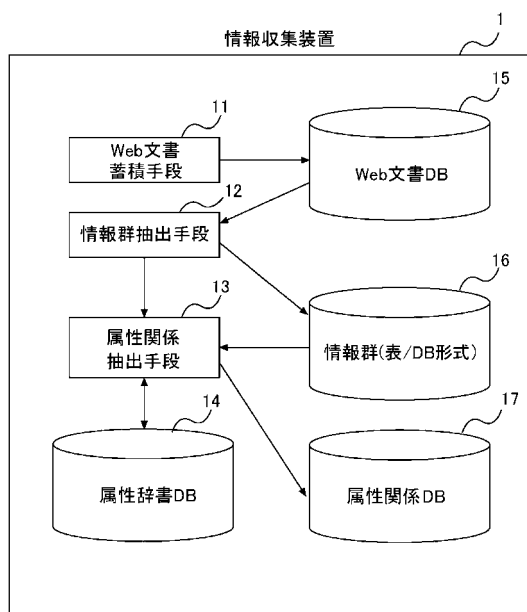
- 1 情報収集装置
- 2 Webサーバ装置
- 1 1 Web文書蓄積手段
- 1 2 情報群抽出手段
- 1 3 属性関係抽出手段
- 1 4 属性辞書
- 1 5 Web文書DB
- 1 6 情報群記憶部
- 1 7 属性関係DB

10

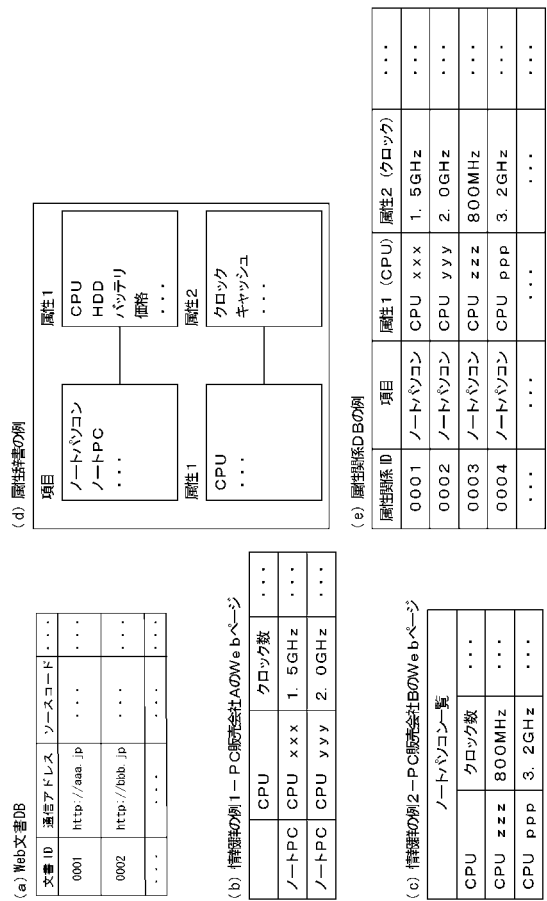
【 図 1 】



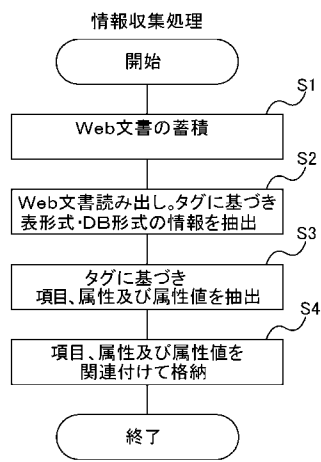
【 図 2 】



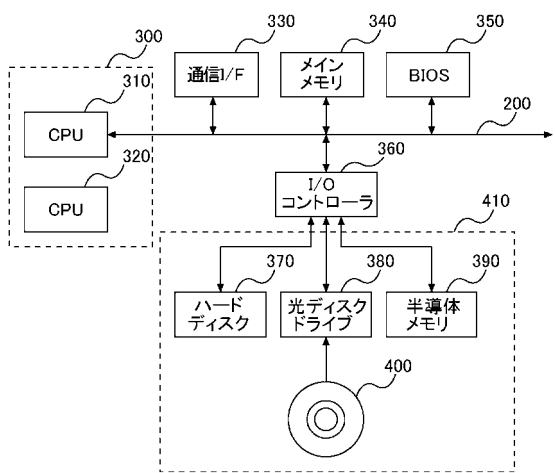
【図 3】



【図 5】



【図 4】



【手続補正書】

【提出日】平成21年9月11日(2009.9.11)

【手続補正 1】

【補正対象書類名】特許請求の範囲

【補正対象項目名】請求項 3

【補正方法】変更

【補正の内容】

【請求項 3】

前記抽出するステップは、前記表形式の情報において、上端行あるいは左端列に位置する情報を属性として、それぞれその下方あるいは右方に位置する情報を属性値として抽出する請求項 1 又は請求項 2 に記載の方法。

【手続補正 2】

【補正対象書類名】明細書

【補正対象項目名】0 0 0 5

【補正方法】変更

【補正の内容】

【0 0 0 5】

そこで、本発明は、このような事情を考慮して提案されるものであり、W e b ページ上に分散して存在している共通の項目やその属性及び属性値の関係にある情報を自動的に収集する情報収集方法、装置及びプログラムを提供することを目的とする。

【手続補正 3】

【補正対象書類名】明細書

【補正対象項目名】0 0 1 8

【補正方法】変更

【補正の内容】

【0 0 1 8】

(3) 前記抽出するステップは、前記表形式の情報において、上端行あるいは左端列に位置する情報を属性として、それぞれその下方あるいは右方に位置する情報を属性値として抽出する(1) 又は (2) に記載の方法。

【手続補正 4】

【補正対象書類名】明細書

【補正対象項目名】0 0 3 0

【補正方法】変更

【補正の内容】

【0 0 3 0】

この発明によれば、W e b 文書に含まれるタグに基づいて、項目、属性及び属性値の関係を有する情報を抽出することにより、W e b ページ上に分散して存在している共通の項目やその属性及び属性値の関係にある情報を自動的に収集することができる。

【手続補正 5】

【補正対象書類名】明細書

【補正対象項目名】0 0 3 1

【補正方法】変更

【補正の内容】

【0 0 3 1】

以下、本発明を実施するための最良の形態について図を参照しながら説明する。なお、これはあくまでも一例であって、本発明の技術的範囲はこれに限られるものではない。

【手続補正 6】

【補正対象書類名】明細書

【補正対象項目名】0 0 3 2

【補正方法】変更

【補正の内容】

【 0 0 3 2 】

〔 情報収集装置と関連要素の全体構成 〕

図 1 において、情報収集装置 1 は、通信ネットワークとしてのインターネット N を通じて複数の Web サーバ装置 2 に接続される。インターネット N との接続は、有線であるか無線であるかを問わない。

【手続補正 7】

【補正対象書類名】明細書

【補正対象項目名】 0 0 3 3

【補正方法】変更

【補正の内容】

【 0 0 3 3 】

情報収集装置 1 は、複数の Web サーバ装置 2 から Web 文書を取得する。各 Web サーバ装置 2 は、情報収集装置 1 からのリクエストに応じて種々の Web 文書を提供する。

【手続補正 8】

【補正対象書類名】明細書

【補正対象項目名】 0 0 3 4

【補正方法】変更

【補正の内容】

【 0 0 3 4 】

〔 情報収集装置の機能構成 〕

図 2 は、本実施形態に係る情報収集装置 1 の機能構成の概要を示す図である。情報収集装置 1 は、Web 文書蓄積手段 1 1、情報群抽出手段 1 2 及び属性関係抽出手段 1 3 を備えている。また、Web 文書 DB 1 5、情報群記憶部 1 6 及び属性関係 DB 1 7 を有する（DB はデータベースの略）。

【手続補正 9】

【補正対象書類名】明細書

【補正対象項目名】 0 0 3 6

【補正方法】変更

【補正の内容】

【 0 0 3 6 】

また、Web 文書 DB 1 5、情報群記憶部 1 6 及び属性関係 DB 1 7 は、後述のハードウェアに含まれる記憶装置 4 1 0 の一領域に設けられている。

【手続補正 1 0】

【補正対象書類名】明細書

【補正対象項目名】 0 0 3 7

【補正方法】変更

【補正の内容】

【 0 0 3 7 】

〔 各種データベースと関連要素の構成 〕

図 3 は、Web 文書 DB 1 5、情報群記憶部 1 6 及び属性関係 DB 1 7 と関連要素の構成の概要を示す図である。

【手続補正 1 1】

【補正対象書類名】明細書

【補正対象項目名】 0 0 3 8

【補正方法】変更

【補正の内容】

【 0 0 3 8 】

図 3 (a) に示すように、Web 文書 DB 1 5 は、文書 ID と、インターネット N 上で配信されている Web 文書の URL 等の インターネット N 上の通信 アドレス及びこの Web

b 文書の記述であるソースコードと、をそれぞれ対応付けて記憶している。

【手続補正 1 2】

【補正対象書類名】明細書

【補正対象項目名】0 0 4 0

【補正方法】変更

【補正の内容】

【0 0 4 0】

図 3 (e) に示すように、属性関係 DB 1 7 には、表形式又はデータベース形式の情報から抽出された項目、属性及び属性値の関係を有する情報が格納される。

【手続補正 1 3】

【補正対象書類名】明細書

【補正対象項目名】0 0 4 4

【補正方法】変更

【補正の内容】

【0 0 4 4】

プルダウンリストを形成する一群の HTML として記載されている場合は、例えば < S e l e c t > タグの開始タグと終了タグを判定し、当該開始タグ及び終了タグとその間にある要素の内容とを抽出することが考えられる。

【手続補正 1 4】

【補正対象書類名】明細書

【補正対象項目名】0 0 4 6

【補正方法】変更

【補正の内容】

【0 0 4 6】

また、テーブルタグにより表を形成する一群の HTML として記述されている場合は、例えば、< T a b l e > タグの開始タグと終了タグを判定し、当該開始タグ及び終了タグとその間にある要素の内容とを抽出することが考えられる。

【手続補正 1 5】

【補正対象書類名】明細書

【補正対象項目名】0 0 5 0

【補正方法】変更

【補正の内容】

【0 0 5 0】

図 3 (d) は、属性関係抽出手段が利用する属性辞書 1 4 の例である。属性辞書 1 4 は、後述のハードウェアに含まれる記憶装置 4 1 0 (図 4 参照) に格納されている。本実施形態において、属性辞書 1 4 は、項目と属性とを関連付けている。例えば、項目には「ノートパソコン」のほか、その類義語である「ノート PC」等が登録されている。一方、項目に関連する属性として「CPU」「HDD」「バッテリー」「価格」等が登録されている。更に、属性 (属性 1) に関する下位の属性 (属性 2) が登録されている。例えば、属性 1 「CPU」に関し、属性 2 「クロック」「キャッシュ」等が登録されている。

【手続補正 1 6】

【補正対象書類名】明細書

【補正対象項目名】0 0 5 1

【補正方法】変更

【補正の内容】

【0 0 5 1】

図 3 (e) は、属性関係 DB 1 7 に格納される項目、属性及び属性値の関係を有する情報の例を示している。ユニークな番号である属性関係 ID ごとに、項目、属性及び属性値を関連付けて保存している。例えば、ある項目「ノートパソコン」は、属性 1 「CPU」の属性値が「CPU x x x」であり、かつ、属性 2 「クロック」の属性値が「1 . 5 G

H z」であるとして保存される。

【手続補正 17】

【補正対象書類名】明細書

【補正対象項目名】0052

【補正方法】変更

【補正の内容】

【0052】

[情報収集装置のハードウェア構成図]

図4は、本実施形態に係る情報収集装置1のハードウェア構成を示す図である。

情報収集装置1は、制御部300を構成するCPU(Central Processing Unit)310(マルチプロセッサ構成ではCPU320等複数のCPUが追加されてもよい)、バスライン200、通信I/F(I/F:インタフェース)330、メインメモリ340、BIOS(Basic Input Output System)350、I/Oコントローラ360、ハードディスク370、光ディスクドライブ380及び半導体メモリ390を備える。尚、ハードディスク370、光ディスクドライブ380及び半導体メモリ390はまとめて記憶装置410と呼ばれる。

【手続補正 18】

【補正対象書類名】明細書

【補正対象項目名】0058

【補正方法】変更

【補正の内容】

【0058】

光ディスクドライブ380としては、例えば、DVD-ROMドライブ、CD-ROMドライブ、DVD-RAMドライブ及びCD-RAMドライブを使用することができる。この場合は各ドライブに対応した光ディスク400を使用する。光ディスク400から光ディスクドライブ380によりプログラム又はデータを読み取り、I/Oコントローラ360を介してメインメモリ340又はハードディスク370に提供することもできる。

【手続補正 19】

【補正対象書類名】明細書

【補正対象項目名】0059

【補正方法】変更

【補正の内容】

【0059】

なお、本発明でいうコンピュータとは、記憶装置、制御部等を備えた情報処理装置をいい、情報収集装置1は、記憶装置410、制御部300等を備えた情報処理装置により構成され、この情報処理装置は、本発明のコンピュータの概念に含まれる。

【手続補正 20】

【補正対象書類名】明細書

【補正対象項目名】0060

【補正方法】変更

【補正の内容】

【0060】

[Webサーバ装置のハードウェア構成]

Webサーバ装置2も、上述の情報収集装置1と同様なハードウェア構成を持つ。

【手続補正 21】

【補正対象書類名】明細書

【補正対象項目名】0061

【補正方法】変更

【補正の内容】

【0061】

[本発明の実施形態に係るフローチャート]

図 5 は、本発明の実施形態に係る情報収集処理のフローチャートを示している。

【 手続補正 2 2 】

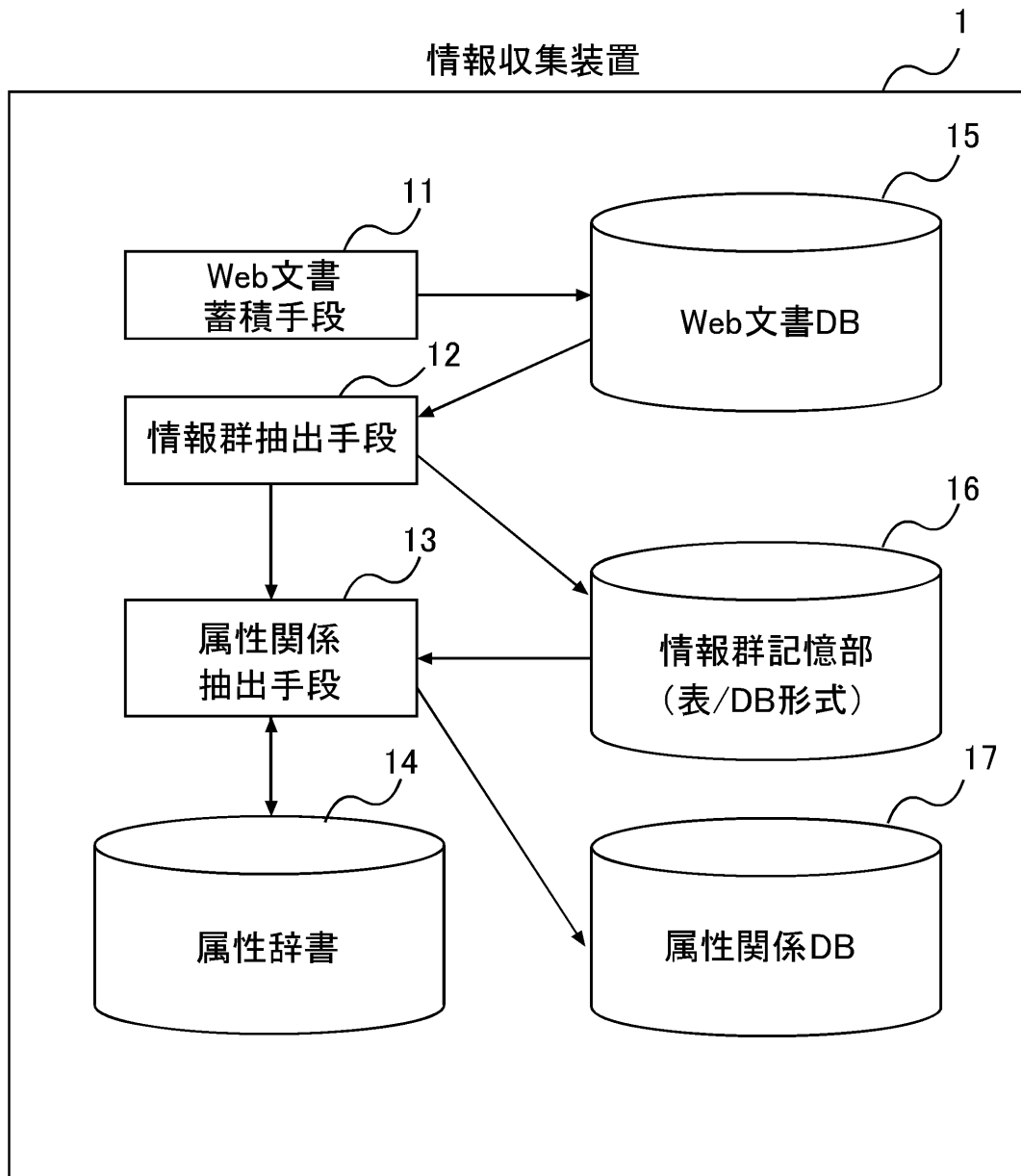
【 補正対象書類名 】 図面

【 補正対象項目名 】 図 2

【 補正方法 】 変更

【 補正の内容 】

【 図 2 】



【 手続補正 2 3 】

【 補正対象書類名 】 図面

【 補正対象項目名 】 図 4

【 補正方法 】 変更

【 補正の内容 】

【 図 4 】

1