

(19) 日本国特許庁(JP)

再公表特許(A1)

(11) 国際公開番号

W02016/098739

発行日 平成29年11月2日 (2017.11.2)

(43) 国際公開日 平成28年6月23日 (2016.6.23)

(51) Int.Cl.	F I	テーマコード (参考)
<b>G06F 17/22 (2006.01)</b>	G06F 17/22 647	5B109
<b>G06F 17/30 (2006.01)</b>	G06F 17/30 210A	
	G06F 17/30 140	
	G06F 17/30 170A	
	G06F 17/30 350C	

審査請求 未請求 予備審査請求 未請求 (全 24 頁) 最終頁に続く

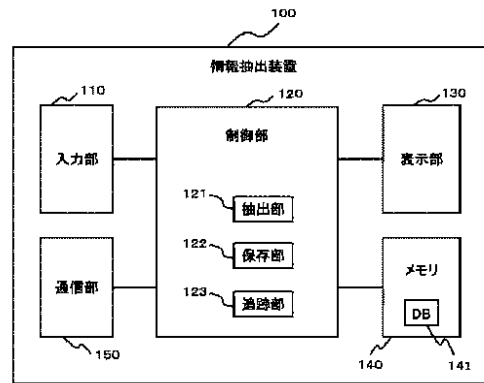
出願番号 特願2016-564846 (P2016-564846)	(71) 出願人 504202472 大学共同利用機関法人情報・システム研究機構 東京都立川市緑町10番3号
(21) 国際出願番号 PCT/JP2015/084974	
(22) 国際出願日 平成27年12月14日 (2015.12.14)	
(31) 優先権主張番号 特願2014-253058 (P2014-253058)	(74) 代理人 100100158 弁理士 鮫島 睦
(32) 優先日 平成26年12月15日 (2014.12.15)	(74) 代理人 100131808 弁理士 柳橋 泰雄
(33) 優先権主張国 日本国 (JP)	(72) 発明者 坂本 一憲 東京都千代田区一ツ橋二丁目1番2号 大学共同利用機関法人情報・システム研究機構 国立情報学研究所内

最終頁に続く

(54) 【発明の名称】 情報抽出装置、情報抽出方法、及び情報抽出プログラム

(57) 【要約】

構造化された文書の仕様が変更した場合であっても、仕様変更前に抽出した特定情報を仕様変更後も容易且つ確実に抽出することが可能な情報抽出装置を提供する。情報抽出装置(100)は、構造化された複数の文書間で異なる部分を可変要素として抽出すると共に、各可変要素から所定範囲内にある要素を周辺情報として抽出する制御部(120)と、可変要素のうち少なくとも1つを抽出対象とし、少なくとも抽出対象について可変要素と周辺情報を格納する記憶部(140)と、を有し、制御部は、構造化された複数の文書から可変要素と周辺情報を再抽出し、再抽出した可変要素及び周辺情報と記憶部に格納されている可変要素及び周辺情報とに基づいて、再抽出前後の可変要素及び周辺情報の類似度を計算し、計算した類似度に基づいて、抽出対象に対応する可変要素を再抽出後の可変要素の中から特定する。



100 Information extraction apparatus  
110 Input unit  
120 Control unit  
121 Extraction unit  
122 Storage unit  
123 Tracking unit  
130 Display unit  
140 Memory  
141 DB  
150 Communication unit

**【特許請求の範囲】****【請求項 1】**

構造化された複数の文書を取得し、取得した複数の文書間で異なる部分を可変要素として抽出すると共に、各可変要素から所定範囲内にある要素を周辺情報として抽出する、制御部と、

前記可変要素のうち少なくとも1つを抽出対象とし、少なくとも前記抽出対象について前記可変要素と前記周辺情報を格納する記憶部と、

を有し、

前記制御部は、前記構造化された複数の文書を再度取得して、再度取得した複数の文書間で異なる部分を可変要素として再抽出すると共に、再抽出した各可変要素から所定範囲内にある要素を周辺情報として再抽出し、再抽出した前記可変要素及び前記周辺情報と前記記憶部に格納されている前記可変要素及び前記周辺情報とに基づいて、再抽出前後の前記可変要素及び前記周辺情報の類似度を計算し、計算した前記類似度に基づいて、前記抽出対象に対応する前記可変要素を再抽出後の前記可変要素の中から特定する、

情報抽出装置。

**【請求項 2】**

再抽出後の前記可変要素の中から、前記抽出対象の可変要素に対する類似度が最も高い可変要素を特定する、請求項 1 に記載の情報抽出装置。

**【請求項 3】**

再抽出した前記可変要素と前記記憶部に格納されている前記可変要素の類似度を計算し、且つ再抽出した前記周辺情報と前記記憶部に格納されている前記周辺情報の類似度を計算し、前記可変要素同士の類似度と前記周辺情報同士の類似度とに基づいて、前記抽出対象に対応する可変要素を再抽出後の前記可変要素の中から特定する、請求項 1 に記載の情報抽出装置。

**【請求項 4】**

再抽出した前記可変要素と前記記憶部に格納されている前記可変要素とにそれぞれ含まれる数字部分と文字部分を、前記数字部分と前記文字部分に分割し、前記数字部分同士の類似度と前記文字部分同士の類似度とに基づいて、前記可変要素の類似度を決定する、請求項 1 に記載の情報抽出装置。

**【請求項 5】**

前記構造化された複数の文書の差分を計算することにより、前記可変要素を抽出する、請求項 1 に記載の情報抽出装置。

**【請求項 6】**

抽出された前記可変要素を表示する表示部と、

表示された前記可変要素の中からユーザにより選択された前記抽出対象を入力する入力部と、

をさらに有する、請求項 1 に記載の情報抽出装置。

**【請求項 7】**

対象とする文書を複数回取得し、複数回取得した文書間で所定回数異なった部分を除外要素として、前記可変要素から除外する、請求項 1 に記載の情報抽出装置。

**【請求項 8】**

構造化された複数の文書を取得するステップと、

取得した複数の文書間で異なる部分を可変要素として抽出するステップと、

各可変要素から所定範囲内にある要素を周辺情報として抽出するステップと、

前記可変要素のうち少なくとも1つを抽出対象とし、少なくとも前記抽出対象について前記可変要素と前記周辺情報を記憶部に格納するステップと、

前記構造化された複数の文書を再度取得するステップと、

再度取得した複数の文書間で異なる部分を可変要素として再抽出するステップと、

再抽出した各可変要素から所定範囲内にある要素を周辺情報として再抽出するステップと、

10

20

30

40

50

再抽出した前記可変要素及び前記周辺情報と前記記憶部に格納されている前記可変要素及び前記周辺情報とに基づいて、再抽出前後の前記可変要素及び前記周辺情報の類似度を計算するステップと、

計算した前記類似度に基づいて、前記抽出対象に対応する可変要素を再抽出後の前記可変要素の中から特定するステップと、

を含む、情報抽出方法。

【請求項 9】

再抽出後の前記可変要素の中から、前記抽出対象の可変要素に対する類似度が最も高い可変要素を特定する、請求項 8 に記載の情報抽出方法。

【請求項 10】

再抽出した前記可変要素と前記記憶部に格納されている前記可変要素の類似度を計算し、且つ再抽出した前記周辺情報と前記記憶部に格納されている前記周辺情報の類似度を計算し、前記可変要素同士の類似度と前記周辺情報同士の類似度とに基づいて、前記抽出対象に対応する可変要素を再抽出後の可変要素の中から特定する、請求項 8 に記載の情報抽出方法。

【請求項 11】

再抽出した前記可変要素と前記記憶部に格納されている前記可変要素にそれぞれ含まれる数字部分と文字部分を、前記数字部分と前記文字部分に分割し、前記数字部分同士の類似度と前記文字部分同士の類似度とに基づいて、前記可変要素の類似度を決定する、請求項 8 に記載の情報抽出方法。

【請求項 12】

前記構造化された複数の文書の差分を計算することにより、前記可変要素を抽出する、請求項 8 に記載の情報抽出方法。

【請求項 13】

抽出された前記可変要素を表示するステップと、  
表示された前記可変要素の中からユーザにより選択された前記抽出対象を入力するステップと、  
をさらに含む、請求項 8 に記載の情報抽出方法。

【請求項 14】

対象とする文書を複数回取得し、複数回取得した文書間で所定回数異なった部分を除外要素として、前記可変要素から除外する、請求項 8 に記載の情報抽出方法。

【請求項 15】

構造化された複数の文書を取得するステップと、  
取得した複数の文書間で異なる部分を可変要素として抽出するステップと、  
各可変要素から所定範囲内にある要素を周辺情報として抽出するステップと、  
前記可変要素のうち少なくとも 1 つを抽出対象とし、少なくとも前記抽出対象について前記可変要素と前記周辺情報を記憶部に格納するステップと、  
前記構造化された複数の文書を再度取得するステップと、  
再度取得した複数の文書間で異なる部分を可変要素として再抽出するステップと、  
再抽出した各可変要素から所定範囲内にある要素を周辺情報として再抽出するステップと、

再抽出した前記可変要素及び前記周辺情報と前記記憶部に格納されている前記可変要素及び前記周辺情報とに基づいて、再抽出前後の可変要素及び周辺情報の類似度を計算するステップと、

計算した前記類似度に基づいて、前記抽出対象に対応する可変要素を再抽出後の前記可変要素の中から特定するステップと、

をコンピュータに実行させるための情報抽出プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

10

20

30

40

50

本発明は、構造化された文書から特定情報を抽出する情報抽出装置、情報抽出方法、及び情報抽出プログラムに関する。

【背景技術】

【0002】

従来の情報抽出装置（特許文献1参照）は、同一構造を持つWebページ間の差分を検出し、差分が検出された場所（タグ）を差分領域として特定し、その差分領域に記載されている情報を差分データとして抽出し、差分領域と差分データとを紐付けして特定情報として記憶している。例えば、「郵便番号」のタグと、実際の郵便番号（例えば、100-1000）とを対応付けて記憶する。この情報抽出装置によれば、例えば、AさんとBさんの英語学習記録のWebページの差分を計算することにより、ユーザごとに内容の異なる箇所（ユーザの名前、単語学習時間、文法学習時間など）を個人情報と捉えて抽出することができる。

10

【0003】

また、別の情報抽出装置（特許文献2参照）は、複数のWebページのツリー構造の各々に共通する部分からデータを抽出する抽出ルールを自動で作成すると共に、その抽出ルールが適用されるWebページのURLを特定する特定ルールを自動で作成している。この情報抽出装置は、作成したWebページのURLを特定するための特定ルールとWebページからデータを抽出するための抽出ルールとを対応付けて記憶している。抽出対象となるWebページからデータ（特定情報）を抽出する際、情報抽出装置は、抽出対象となるWebページのURLが特定される特定ルールを選択し、選択された特定ルールに対応付けられている抽出ルールを選択し、選択された抽出ルールに基づいて抽出対象のWebページからデータ（特定情報）を抽出している。

20

【0004】

さらに別の情報抽出装置（特許文献3参照）は、複数の個人領域が混在する単一のWebページ（掲示板など）から、個人領域に該当する箇所を特定して抽出し、各個人領域に対応付けている情報を特定する機能を実現している。例えば、掲示板のページにおいて、ユーザが書き込んだ箇所を特定し、ユーザごとに書き込み内容を抽出する。

【0005】

さらに別の情報抽出装置（非特許文献1参照）は、Webアプリケーションに対する機能テストにおいて、仕様変更があった際に、抽出対象となる特定要素の抽出プログラムを修正しなくても、「contextual clues」と呼ばれる周囲の情報を参考にして特定の要素を抽出するルールの記述に関する手法を実現している。例えば、英語学習記録のWebページから単語学習時間と文法学習時間を抽出する際に、「単語学習時間は“単語”という文言付近に存在」及び「文法学習時間は“文法”という文言付近に存在」というルールを用いることにより、特定の情報を継続的且つロバストに抽出している。

30

【先行技術文献】

【特許文献】

【0006】

【特許文献1】特開2012-098855号公報

【特許文献2】特開2012-059212号公報

【特許文献3】特開2012-168892号公報

40

【非特許文献】

【0007】

【非特許文献1】Rahulkrishna Yandrapally, Suresh Thummalapenta, Saurabh Sinha, Satish Chandra, "Robust Test Automation Using Contextual Clues", IBM Research Report, 2014.

【発明の概要】

【発明が解決しようとする課題】

【0008】

Webページは、一般に、その仕様（例えば、ページのデザイン、ページ内の情報の配

50

置、及びページのツリー構造)が頻繁に変更される場合がある。しかし、上述した従来の情報抽出装置は、後述するように、構造化された文書(例えば、Webページ)の仕様が変更した場合に、仕様変更前に抽出した特定情報(例えば、個人情報)を仕様変更後は容易且つ確実に抽出することができない。

【0009】

特許文献1の情報抽出装置は、仕様変更前後における抽出情報の追跡を行っていない。そのため、例えば、ある時点で英語学習記録のWebページから単語学習時間と文法学習時間を抽出できたとしても、仕様変更後に抽出した情報が単語学習時間か文法学習時間を区別できなくなる場合がある。

【0010】

特許文献2の情報抽出装置は、Webページの構造変化を検出した場合、手動又は自動で抽出ルール及び特定ルールを再作成している。すなわち、特許文献2の場合は、Webページに仕様変更があった場合、抽出ルールや特定ルールを再度作成しなおす必要がある。また、特許文献2において抽出される情報は、複数のWebページの共通部分に限定される。

【0011】

特許文献3の情報抽出装置は、Webページのデザインや構成が変化した場合の変更前後における抽出情報の追跡を行っていない。

【0012】

非特許文献1の情報抽出装置は、抽出対象となる要素を抽出する際に使用する周辺情報をユーザが選択する必要がある。また、その周辺情報は特定の情報(例えば、“文法”という文言付近)に限定されるため、その周辺情報がWebページの仕様変更により消失した場合に、抽出対象の要素を抽出できなくなる。

【0013】

以上のように、従来の情報抽出装置は、構造化された文書(例えば、Webページ)の仕様が変更した場合に、仕様変更前に抽出した特定情報を仕様変更後は容易且つ確実に抽出することができない。

【0014】

本発明は、構造化された文書(例えば、Webページ)の仕様が変更した場合であっても、仕様変更前に抽出した特定情報を仕様変更後も容易且つ確実に抽出することが可能な情報抽出装置、情報抽出方法、及び情報抽出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】

【0015】

本発明の情報抽出装置は、構造化された複数の文書(具体的には、構造が等しくコンテキストが異なる複数の文書)を取得し、取得した複数の文書間で異なる部分を可変要素として抽出すると共に、各可変要素から所定範囲内にある要素を周辺情報として抽出する制御部と、可変要素のうち少なくとも1つを抽出対象とし、少なくとも抽出対象について可変要素と周辺情報を格納する記憶部と、を有し、制御部は、構造化された複数の文書を再度取得して、再度取得した複数の文書間で異なる部分を可変要素として再抽出すると共に、再抽出した各可変要素から所定範囲内にある要素を周辺情報として再抽出し、再抽出した可変要素及び周辺情報と記憶部に格納されている可変要素及び周辺情報とに基づいて、再抽出前後の可変要素及び周辺情報の類似度を計算し、計算した類似度に基づいて、抽出対象に対応する可変要素を再抽出後の可変要素の中から特定する。

【0016】

本発明の情報抽出方法は、構造化された複数の文書を取得するステップと、取得した複数の文書間で異なる部分を可変要素として抽出するステップと、各可変要素から所定範囲内にある要素を周辺情報として抽出するステップと、可変要素のうち少なくとも1つを抽出対象とし、少なくとも抽出対象について可変要素と周辺情報を記憶部に格納するステップと、構造化された複数の文書を再度取得するステップと、再度取得した複数の文書間で異なる部分を可変要素として再抽出するステップと、再抽出した各可変要素から所定範囲

10

20

30

40

50

内にある要素を周辺情報として再抽出するステップと、再抽出した可変要素及び周辺情報と記憶部に格納されている可変要素及び周辺情報とに基づいて、再抽出前後の可変要素及び周辺情報の類似度を計算するステップと、計算した類似度に基づいて、抽出対象に対応する可変要素を再抽出後の可変要素の中から特定するステップと、を含む。

【0017】

本発明の情報抽出プログラムは、上記情報抽出方法の各ステップをコンピュータに実行させる。

【発明の効果】

【0018】

本発明の情報抽出装置は、構造化された複数の文書間で異なる部分（例えば、氏名、体重、身長などの個人情報）を可変要素として抽出すると共に、各可変要素から所定範囲内にある要素（例えば、テキスト、HTMLタグ、及び属性等）を周辺情報として抽出し、可変要素のうち少なくとも1つを抽出対象（特定情報）とし、少なくとも抽出対象について可変要素と周辺情報を記憶している。本発明の情報抽出装置によれば、再度、可変要素とその周辺情報を抽出したときに、記憶されている可変要素及び周辺情報と再度抽出された可変要素及び周辺情報の類似度を計算し、その結果に基づいて、抽出対象に対応する可変要素を再抽出後の可変要素の中から特定している。これにより、構造化された文書（例えば、Webページ）の仕様が変更した場合であっても、仕様変更前に抽出した特定情報を仕様変更後も容易且つ確実に抽出する又は追跡することができる。

10

【図面の簡単な説明】

20

【0019】

【図1】本発明の実施形態1の情報抽出装置の構成図

【図2】本発明の実施形態1における可変要素及び周辺情報の抽出を示すフローチャート

【図3】本発明の実施形態1におけるWebページに関する具体例であって、(a)はURL、(b)はHTML文書、(c)は可変要素の画面表示をそれぞれ示す図

【図4】本発明の実施形態1におけるメモリに記憶される抽出情報の例

【図5】本発明の実施形態1における特定情報の抽出を示すフローチャート

【図6】本発明の実施形態1におけるWebページの仕様変更前後の例

【図7】本発明の実施形態1における類似度の例

【図8】本発明の実施形態2における対象者のWebページの1分経過前後の例

30

【図9】本発明の実施形態2における除外候補の抽出及び除外を示すフローチャート

【発明を実施するための形態】

【0020】

以下、本発明の実施形態について、図面を参照しながら説明する。

【0021】

<実施形態1>

本実施形態の情報抽出装置は、構造化された複数の文書（本実施形態において、Webページ）間で異なる部分を可変要素として抽出すると共に、各可変要素から所定範囲内にある要素を周辺情報として抽出し、可変要素のうち少なくとも1つを抽出対象（特定情報）とし、少なくとも抽出対象について可変要素と周辺情報を記憶する。情報抽出装置は、再度、可変要素とその周辺情報を抽出したときに、記憶されている可変要素及び周辺情報と再度抽出された可変要素及び周辺情報の類似度を計算し、その結果に基づいて、抽出対象に対応する可変要素を再抽出後の可変要素の中から特定している。これにより、構造化された文書の仕様が変更した場合であっても、仕様変更前に抽出した特定情報を仕様変更後も容易且つ確実に抽出する、すなわち仕様変更前後において特定情報を追跡することができる。本実施形態によれば、仕様変更前後で抽出箇所の位置を追跡することにより、機械的且つ定常的に特定情報の抽出を行うことができる。以下、構造化された文書がWebページである場合を例にして説明する。

40

【0022】

1-1. 情報抽出装置の構成

50

図 1 に、本発明の実施形態の情報抽出装置の構成を示す。情報抽出装置 100 は、パーソナルコンピュータなどで実現できる。情報抽出装置 100 は、ユーザからの入力を受け付ける入力部 110、情報抽出装置 100 全体を制御する制御部 120、表示部 130、メモリ 140、及び通信部 150 を有する。

【0023】

入力部 110 は、例えば、構造化された文書の場所を示す情報（本実施形態において、Web ページの URL）を入力するのに使用される。入力部 110 は、また、複数の Web ページ間で異なる部分である可変要素の中の少なくとも 1 つを抽出対象となる特定情報（抽出要素）として選択するのに使用される。入力部 110 は、例えばキーボード又はタッチパネルである。

10

【0024】

制御部 120 は、複数の Web ページ間で異なる部分である可変要素及びその周辺情報を抽出する抽出部 121、抽出した可変要素及びその周辺情報をメモリ 140 に書き込む保存部 122、及びメモリ 140 に書き込まれた可変要素及びその周辺情報を使用して、抽出要素を追跡する追跡部 123 を有する。

【0025】

抽出部 121 は、対象の Web ページを含む複数の Web ページの各々の構成情報（本実施形態において、HTML (Hyper Text Markup Language) 文書）を対応する URL に基づいて取得し、取得した構成情報に基づいて複数の Web ページ間で異なる部分を可変要素として抽出する。本実施形態においては、複数の Web ページの差分を計算することにより、可変要素を抽出する。可変要素は、例えば個人情報（氏名、体重、身長など）に該当する。さらに、抽出部 121 は、対象ページ内の全ての可変要素から所定範囲内にある要素（テキスト、HTML タグ、及び属性等）を周辺情報として、対象ページ内から抽出する。

20

【0026】

表示部 130 は、抽出部 121 が抽出した可変要素を表示する。表示部 130 は、ディスプレイなどで実現できる。ユーザは表示部 130 に表示された可変要素の中から、抽出したい要素を選択し、入力部 110 に入力する。

【0027】

保存部 122 は、図 4 に示すような抽出情報をメモリ 140 内のデータベース (DB) 141 に記録する。抽出情報は、対象ページ内の全ての可変要素と、その周辺情報と、ユーザによる抽出対象としての選択の有無を含む。さらに、保存部 122 は、入力された URL をメモリ 140 に格納する。メモリ 140 は、例えばハードディスクである。なお、メモリ 140 は、ハードディスクに限らず、光ディスクなどの記憶装置、フラッシュメモリなどの半導体メモリ素子、又は RAM などであっても良い。

30

【0028】

追跡部 123 は、抽出対象として選択された可変要素（特定情報）を追跡する。具体的には、追跡部 123 は、抽出部 121 により再度抽出された現在の Web ページの可変要素及びその周辺情報と、データベース 141 の抽出情報とを用いて、再抽出前後の可変要素間の対応関係を復元する。本実施形態において、対応関係の復元は、新たに抽出した可変要素に関する情報とデータベース 141 に記録済みの可変要素に関する情報との類似度を計算し、類似度の高い可変要素同士を対応付けることにより、行う。より具体的には、類似性の計算は、可変要素自身の類似度と、周辺情報の類似度との両方を総合的に判断することによって行う。これにより、再抽出後の可変要素の中から、以前にユーザが抽出対象として指定した要素を特定する。

40

【0029】

通信部 150 は、インターネットなどのネットワークに接続される。抽出部 121 は、通信部 150 を介して、URL に対応する HTML 文書を取得する。また、通信部 150 を介して、ユーザによる抽出要素の選択を行っても良い。さらに、追跡された抽出要素を通信部 150 を介して外部機器に出力しても良い。

50

## 【 0 0 3 0 】

## 1 - 2 . 情報抽出装置の動作

図 2 に、情報抽出装置 1 0 0 による可変要素及び周辺情報の抽出のフローチャートを示す。図 3 ( a ) は URL、図 3 ( b ) は HTML 文書、図 3 ( c ) は抽出後の可変要素の画面表示の例をそれぞれ示している。図 3 ( b ) の左側が、本実施形態において抽出対象となる Web ページ、その右側が抽出対象の Web ページとコンテキスト ( アカウント、日時など ) が異なる Web ページを示す。図 3 ( b ) の例では、HTML 文書は、ユーザごとに、名前、現在の体重、一ヶ月前の体重、及び身長 の 4 種類の情報を含む。図 4 は、メモリ 1 4 0 に格納される抽出情報の DB 1 4 1 の例を示している。

## 【 0 0 3 1 】

以下、図 4 に示すように、「 5 5 k g ( 坂本さんの今月の体重 ) 」を抽出対象として選択する場合を例にして、説明する。

## 【 0 0 3 2 】

図 2 のフローにおいて、まず、入力部 1 1 0 は、図 3 ( a ) に示すような複数の Web ページの URL を入力する ( ステップ S 2 0 1 ) 。具体的には、抽出対象の Web ページの URL 及び抽出対象の Web ページとレイアウト及び構造が等しくコンテキストが異なる 1 以上の他の Web ページの URL を入力する。保存部 1 2 2 は、入力した URL をメモリ 1 4 0 に格納する。抽出部 1 2 1 は、複数の Web ページの URL に対応する構成情報 ( HTML 文書 ) を、通信部 1 5 0 を介して、取得する ( ステップ S 2 0 2 ) 。

## 【 0 0 3 3 】

抽出部 1 2 1 は、取得したページ構成情報に基づいて、抽出対象の Web ページ内におけるその他の Web ページと異なる部分を可変要素として抽出する ( ステップ S 2 0 3 ) 。例えば、図 3 ( b ) に示すような個人情報が掲載されている Web ページから、ユーザごとに異なる個人情報 ( 「 5 5 k g 」 、 「 5 4 k g 」 、 「 1 7 1 c m 」 、 「 坂本 」 ) を可変要素として抽出する。本実施形態においては、可変要素の抽出は、抽出対象の Web ページとその他の Web ページ間の差分を計算することにより、行う。差分計算として、例えば、既存アルゴリズム ( X D i f f : Wang, Yuan, David J. DeWitt, and J-Y. Cai. " X-Diff: An effective change detection algorithm for XML documents." IEEE 19th International Conference on Data Engineering, pp. 519-530, 2003. ) を使用できる。なお、差分計算は、このアルゴリズムに限定されない。個人情報が偶然同じ内容である場合 ( 例えば、坂本と佐藤が同じ体重又は同じ身長の場合 ) 、その個人情報を可変要素として抽出できなくなる。そのため、抽出対象のページと比較するための他の Web ページを複数用意することにより、偶然同じ情報を有する可能性を十分に下げることができ、より正確に可変要素を抽出することができる。

## 【 0 0 3 4 】

抽出部 1 2 1 は、可変要素から所定範囲内 ( 例えば、可変要素の周囲 1 0 0 文字以内 ) にある要素である周辺情報を Web ページの構成情報 ( HTML 文書 ) の中から抽出する ( ステップ S 2 0 4 ) 。具体的には、周辺情報として、HTML タグ名、属性名、属性値及びテキスト、から構成されるトークン列を抽出する。例えば、図 3 ( b ) 及び図 4 に示すように、可変要素 “ 5 5 k g ” に対して、テキスト ( “ あなたの体重は ” 、 “ 。 ” ) 、HTML タグ ( d i v 、 s p a n ) 、属性名 ( i d ) 及び属性値 ( “ h i g h t ” ) を抽出する ( 例えば、 “ あなたの体重は ” 、 span、id、 “ bw ” 、 /span、 “ 。 ” ) 。

## 【 0 0 3 5 】

抽出部 1 2 1 は、図 3 ( c ) に示すように、抽出した可変要素を表示部 1 3 0 に表示する ( ステップ S 2 0 5 ) 。これにより、ユーザは対象の Web ページ内の可変要素を視認でき、可変要素の中から抽出対象 ( 追跡したい要素 ) を選択することが可能になる。例えば、ユーザは、図 3 ( c ) に示される可変要素から「 5 5 k g ( 現在の体重 ) 」を定常的に抽出する情報として選択する。入力部 1 1 0 は、その選択を入力する ( ステップ S 2 0 6 ) 。保存部 1 2 2 は、図 4 に示すように、抽出対象となる Web ページ内の全ての可変要素とその周辺情報、及び入力部 1 1 0 を介して取得した抽出対象としての選択の有無を

10

20

30

40

50



含む抽出情報をメモリ 140 内のデータベース 141 に記憶する (ステップ S207)。

【0036】

以上のようにして、抽出対象として選択された特定情報 (抽出要素) の追跡に必要な情報記録が完了する。抽出要素の追跡は、データベース 141 に記録された抽出情報を用いて行う。これにより、Web ページの仕様変更によってデザインや構成が変わっても、抽出要素を追跡することを可能にする。

【0037】

図 5 に、情報抽出装置 100 による特定情報 (抽出要素) の追跡のフローチャートを示す。図 6 に、Web ページの仕様変更前後の HTML 文書の例を示す。図 7 に、記録済みと再抽出後の可変要素の類似度を示す。

【0038】

図 5 において、情報抽出装置 100 は、所定の周期 (例えば、月 1 回) 又はユーザの指定により、特定情報 (抽出要素) の追跡を行う。特定情報 (抽出要素) を追跡する際、まず、情報抽出装置 100 の抽出部 121 は、メモリ 140 に格納されている URL を使用して、図 2 のステップ S202 及び S203 と同様の方法で、再度、複数の Web ページの構成情報 (HTML 文書) を取得し (ステップ S501)、現在の Web ページの可変要素を抽出する (ステップ S502)。例えば、図 6 に示すように、Web ページの仕様変更が起こり、さらに、月が変わり体重が 1 kg 増加したことを想定する。この場合、対象となる Web ページの可変要素として、「坂本」、「56 kg」、「55 kg」、「171 cm」が抽出される。その後、抽出部 121 は、図 2 のステップ S204 と同様の方法で、可変要素の周辺情報を再度抽出する (ステップ S503)。具体的には、可変要素の周囲 100 文字から、HTML タグ名、属性名、属性値、及びテキストから構成されるトークン列を抽出する (例えば、div、“体重:”、span、id、“bw”、/span、/div)。

【0039】

追跡部 123 は、再抽出した可変要素とデータベース 141 に記録済みの可変要素とを用いて、可変要素同士の類似度を計算する (ステップ S504)。さらに、追跡部 123 は、再抽出した周辺情報とデータベース 141 に記録済みの周辺情報とを用いて、周辺要素の類似度を計算する (ステップ S505)。このように計算された可変要素自身の類似度と、その周辺情報の類似度とを総合的に判断し、最も類似度の高い組合せが同一の可変要素であるとして類似度の高い可変要素同士を対応付けて、可変要素の対応関係を復元する。これにより、抽出要素を特定する (ステップ S506)。すなわち、抽出対象となる特定情報を追跡する。

【0040】

任意の計算方法が、可変要素と周辺情報 (周囲の構造化された文字列) の類似度の計算方法として利用できる。例えば、可変要素と周辺情報の類似度の計算において、レーベンシュタイン距離を使用することができる。本実施形態においては、0 以上 1.0 以下で正規化された実数を用いて、類似度を計算する。具体的には、類似度を以下のように定義する。

類似度 = 「可変要素の類似度 (S1) × 係数 A」 + 「周辺情報の類似度 (S2) × 係数 B」

(ここで、係数 A と係数 B は 0 以上の実数、且つ、係数 A + 係数 B = 1.0)

係数 A と係数 B はパラメータであり、値を変更して、適用先に応じて類似度計算の精度を調整できる。

【0041】

「可変要素の類似度 (S1 = 0.0 ~ 1.0)」は、以下のように定義される。

可変要素の類似度 (S1) = 「数字部の類似度 (S3) × 係数 C」 + 「文字部の類似度 (S4) × 係数 D」

(ここで、係数 C と係数 D は 0 以上の実数、且つ、係数 C + 係数 D = 1.0)

よって、可変要素の類似度において、まず、可変要素のテキストを数字部と文字部に分

10

20

30

40

50

解する。例えば、「55kg」「55」と「kg」、「56kg」「56」と「kg」、「171cm」「171」と「kg」。

【0042】

次に、可変要素における数字部と文字部の類似度を以下のように計算する。可変要素の数字部の類似度(S3)において、まず、抽出要素に対し、再抽出後の可変要素を数字部分の差の絶対値(例えば、|55-55|、|56-55|、|171-55|)で小さい順に並べ、再抽出後の可変要素の順位を決定する。数字部分が無い場合は、差の絶対値を無限大として設定する。その後、「類似度=(差の絶対値の種類数-順位)×1/(差の絶対値の種類数-1)」により、数字部の類似度を求める。例えば、図6上段の抽出要素「55kg」の数字部分「55」に対する再抽出後の可変要素の数字部の類似度(S3)は以下ようになる。

10

再抽出後の可変要素	55kg (1位)	56kg (2位)	171cm (3位)	坂本 (4位)
差の絶対値	0	1	116	∞
差の絶対値の種類数	4			
記録済みの可変要素との類似度	1.0	0.66	0.33	0

【0043】

可変要素の文字部(文字列)の類似度(S4)において、まず、可変要素の文字列に対して、最長共通部分列(LCS)の長さを用いる。「文字部の類似度=LCSの長さ/仕様変更前の文字列長」により、文字部の類似度(S4)を求める。例えば、抽出要素「55kg」の文字部分「kg」に対する再抽出後の可変要素の文字部の類似度(S4)は以下ようになる。

20

再抽出後の可変要素	55kg	56kg	171cm	坂本
LCSの長さ	2	2	0	0
仕様変更前の文字列長	2			
記録済みの可変要素との類似度	1.0	1.0	0	0

30

【0044】

以上のように、可変要素の数字部と文字部のそれぞれの類似度から、可変要素全体の類似度を求める。次に、周辺情報(周囲の文字列同士)の類似度(S2)を計算する。例えば、周囲の構造化された文字列「あなたの体重は<span id="bw">55kg</span>。先月は54kg!」に対応する周辺情報の類似度を計算する。

【0045】

まず、HTML文書の構造に着目して、トークン列を生成する。例えば、可変要素を除いて、HTMLタグ名、属性名、属性値、及びテキストをそれぞれ1トークンと見なして列にする(「<div>名前:<span id="name">坂本</span></div>」から「"div"、"名前:"、"span"、"id"、"name"、"/span"、"/div"」を生成)。次に、可変部分の前後X個(Xは任意)のトークンをそれぞれ周囲の文字列として抽出する(「<div>名前:<span id="name">坂本</span></div>」から、前後2個(X=2)であれば、「"id"、"name"、"/span"、"/div"」を抽出。「<div>あなたの体重は<span id="bw">55kg</span>。先月は54kg!</div>」から、前後2個(X=2)であれば、「"id"、"bw"、"/span"、"。先月は54kg!"」を抽出)。その後、抽出後に各トークンに対して形態素解析を行い、単語列に変換する(「"id"、"name"、"/span"、"/div"」は変化なし。「"id"、"bw"、"/span"、"。先月は54kg!"」は「

40

50

“ id ”，“ bw ”，“ /span ”，“ 。 ”，“ 先月 ”，“ は ”，“ 5 4 k g ”，“ ! ”」に変換）。

【 0 0 4 6 】

このようにして得られる単語列は、例えば前後 2 トークンを抽出する場合、以下のようになる。

仕様変更前の 5 5 k g の周辺情報の単語列は「 “ id ”，“ bw ”，“ /span ”，“ 。 ”，“ 先月 ”，“ は ”，“ 5 4 k g ”，“ ! ”」となる。

仕様変更後の単語列は、

坂本の周辺情報 ( 1 ) 「 “ id ”，“ name ”，“ /span ”，“ /div ”」、

5 6 k g の周辺情報 ( 2 ) 「 “ id ”，“ bw ”，“ /span ”，“ /div ”」、

5 5 k g の周辺情報 ( 3 ) 「 “ id ”，“ lbw ”，“ /span ”，“ /div ”」、

1 7 1 c m の周辺情報 ( 4 ) 「 “ id ”，“ height ”，“ /span ”，“ /div ”」となる。

10

【 0 0 4 7 】

得られた単語列同士で比較して、類似度を計算する。具体的には、周辺情報の類似度 ( S 2 ) は、「周辺情報の類似度 = 仕様変更前後で共通する単語数 / ( 仕様変更前の単語数 + 仕様変更後の単語数 ) により求める。上述の例では、仕様変更前の単語数は 8 、仕様変更後の単語数は 4 である。仕様変更前後で共通する単語数は重複も含めて、仕様変更前後の両方の数をカウントする ( 例えば、仕様変更前の「 5 5 k g 」の周辺情報と仕様変更後の「坂本」の周辺情報 ( 1 ) の場合、「 id 」と「 /span 」が仕様変更前後の両方に含まれているため、「 id 」× 2 と「 /span 」× 2 で「 4 」となる)。

【 0 0 4 8 】

20

このようにして求めた周辺情報の類似度 ( S 2 ) は以下のようになる。

再抽出後の 周辺情報	(1) 坂本	(2) 5 6 k g	(3) 5 5 k g	(4) 1 7 1 c m
共通する単語数	4	6	4	4
類似度	0. 3 3 3	0. 5	0. 3 3 3	0. 3 3 3

【 0 0 4 9 】

以上のようにしてそれぞれ求めた、可変要素の文字部の類似度 ( S 4 ) 及び数字部の類似度 ( S 3 ) 並びに周辺要素の類似度 ( S 2 ) から、係数 A , B , C , D の値をそれぞれ A = 0 . 2 、 B = 0 . 8 、 C = 0 . 5 、 D = 0 . 5 として、「類似度 = ( ( S 3 × C + S 4 × D ) × A + S 2 × B ) 」により得られた、抽出要素 ( この例では、現在の体重 ) である仕様変更前の「 5 5 k g 」に関する最終的な類似度は以下のようになる。

30

記録済\再抽出後	坂本	5 6 k g	5 5 k g	1 7 1 c m
5 5 k g (現在の体重)	0. 2 6 6 4	0. 5 6 6	0. 4 6 6 4	0. 2 9 9 4

【 0 0 5 0 】

また、「類似度 = ( ( S 3 × C + S 4 × D ) × A + S 2 × B ) 」により得られた、W e b ページの仕様変更前後の各可変要素のペアの類似度の例が図 7 に示されている。なお、図 7 では、前述の例の値とは異なるが、上記方法による計算結果により各数値が得られたと仮定している。図 7 の一番上の「 ( 記録済 ) 5 5 k g 」の行において、再抽出後の可変要素の中の「 5 6 k g 」が、抽出要素「 5 5 k g 」に対して、類似度が 0 . 4 と最も高い。よって、再抽出後の「 5 6 k g 」と抽出対象として記録済みの「 5 5 k g 」とに対応関係があるとみなす。すなわち、再抽出後の「 5 6 k g 」が抽出要素として特定される。

40

【 0 0 5 1 】

なお、記録済みの「 5 4 k g 」についても、再抽出後の「 5 6 k g 」に対して、類似度 0 . 3 と最も類似度が高い。しかし、「 5 5 k g ( 記録済み ) 」と「 5 6 k g ( 再抽出後 ) 」のペアは類似度 0 . 4 であるのに対し、「 5 4 k g ( 記録済み ) 」と「 5 6 k g ( 再抽出後 ) 」のペアは類似度 0 . 3 であるため、より類似度が高い「 5 5 k g ( 記録済み ) 」

50

」と「56kg（再抽出後）」のペアが対応関係を有するとして、対応関係を復元する。また、図7において、「坂本」と「171cm」は仕様変更前後で可変要素自身のテキストに変化がない。よって、再抽出前後の「坂本」の類似度は0.5と高く、再抽出前後の「171cm」の類似度も0.4と高い。このように、可変要素自身に変更がない場合は、類似度の高いペアが容易に見つかる。対応関係の復元は、類似度の数値が高いペアから順に決定する（0.5（坂本 - 坂本）、0.4（56kg - 55kg）、0.4（171cm - 171cm）、0.2（55kg - 54kg））。よって、「55kg（再抽出後）」については「54kg（記録済み）」とペアになる。なお、図7においては、全ての要素についてペアが成立する場合を例示しているが、ペアが作成できなかった要素がある場合（例えば、仕様変更後に、性別（男）が含まれている場合）は対応関係がないと判断する。

10

**【0052】**

なお、図7においては、類似度の計算を説明するために、対象ページ内の全ての可変要素（抽出要素以外の可変要素を含む。）と再抽出後の可変要素の対応関係を示しているが、抽出要素を特定するための類似度の計算（S504、S505）においては、少なくとも抽出対象として選択された可変要素のみについて類似度を計算しても良い（例えば、図7の一番上の「（記録済）55kg」の行のみ）。

**【0053】**

このように、再抽出後の各可変要素について、ユーザが選択した抽出要素（特定情報）に対する類似度を計算して、再抽出前後の可変要素の対応関係を復元することにより、抽出対象の特定情報を機械的且つ定常的に抽出する。

20

**【0054】**

## 1-3. まとめ

以上のようにして、情報抽出装置100は記憶している抽出情報（可変要素、周辺情報、及び抽出対象としての選択の有無）に基づいて、対象とするWebページの新たに取得した構成情報から、抽出対象の特定情報を抽出する。Webページは一般にデザインや構造などの仕様変更される頻度が高く、例えば図6のように仕様変更される場合がある。しかし、本発明によれば、可変要素及びその周辺情報を用いて特定情報を抽出するため、Webページの構成情報に変更があっても、ユーザが指定した特定情報を自動で抽出（追跡）することができる。また、ユーザが指定した特定情報自体が変更している場合がある。例えば、図6に示すように特定情報の数値（今月の体重の数値）が更新されている場合もある。しかし、本発明によれば、記憶している抽出情報を用いて特定情報を抽出するため、特定情報自体に変更があっても、ユーザが指定した特定情報を自動で抽出（追跡）することができる。

30

**【0055】**

本実施形態の情報抽出装置100によれば自動で特定情報を抽出（追跡）することができるため、情報抽出装置100を様々なサービスに利用することができる。例えば、情報抽出装置100が抽出した特定情報を利用して、ユーザが設定した目標に対する達成支援を行い、目標達成の結果に応じて報酬や罰金をユーザに対して行うような、目標達成支援システムに情報抽出装置100を利用しても良い。上述したように、本実施形態の情報抽出装置100によれば、Webページの構成や個人情報に変更があっても、その個人情報を自動で収集できるため、抽出した個人情報を利用したサービスに有用である。

40

**【0056】**

近年、Webアプリケーション及びウェアラブルデバイスの発達により、日々の活動や体重などの変動する個人情報を記録及び発信するためのWebサービス（ライフログサービス）が普及している。これらのWebサービスはそれぞれ異なる特徴を有するため、ユーザは複数のWebサービスを利用することになる。しかし、複数のWebサービスを利用した場合、各Webサービスから情報を集約して処理する際の集約コストが増大する。利用サービス数に比例して集約コストが増大するという問題を解決するためには、様々なライフログサービスから情報を抽出して、一括して個人情報を管理する仕組みが必要とな

50

る。本発明の情報抽出装置100を利用すれば、既存のライフログサービスを構成するユーザごとのWebページを解析して、情報を抽出することができる。ライフログは日々の活動を記録するため、抽出対象の情報の更新頻度が高い。そのため、Webページから定期的に情報を抽出する際、Webページのデザイン又は構成がライフログサービスの仕様変更に伴い発生した場合、従来の情報抽出装置においては、情報を抽出するメカニズムが機能しなくなる。しかし、本発明の情報抽出装置100によれば、Webページのデザイン又は構成が変化した場合であっても、Webページから機械的且つ定期的特定情報を抽出し続けることができる。よって、複数のライフログサービスなどから個人情報を収集して、収集した情報と以前収集した履歴を一括して管理する仕組みを実現できる。その結果、情報の集約及び管理コストを低減できる。集約した情報が、読書のページ数や英語の勉強時間などの数値を扱う場合、グラフなどを生成して可視化することが可能となる。また、過去と比べて値が大きく変動している場合は、動機付けのためのフィードバックを与える仕組みを構築することもできる。

10

#### 【0057】

コンテキストに応じて変化する情報は、個人情報である可能性が高い。よって、個人情報を定期的に収集する場合に、本発明は有用である。また、本発明は、複数のWebページを有するWebアプリに有用である。本発明は、ソフトウェア産業、主にWeb上の情報源を解析するようなソフトウェアを利用する産業において有効に機能する。

#### 【0058】

##### 1-4. 変形例

本実施形態において、周辺情報の類似度(S2)の計算は、可変要素を除いたトークン列を作成することにより行ったが、可変要素を含めたトークン列を作成して行っても良い(例えば、「<div>名前:<span id="name">坂本</span></div>」から「"div"、"名前:"、"span"、"id"、"name"、"坂本"、"/span"、"/div"」のトークン列を生成)。この場合、仕様変更前の単語数及び仕様変更後の単語数として、可変要素を含めてカウントしても良い(例えば、可変部分の前後2個のトークンを周囲の文字列として抽出した場合の、使用変更後の坂本の周辺情報(1)「"id"、"name"、"坂本"、"/span"、"/div"」の単語数は5である)。

20

#### 【0059】

本実施形態の情報抽出装置100は、Webページに限らず、構造化された文書に適用できる。また、可変要素の抽出方法は、差分計算に限らず、任意の方法で行っても良い。また、類似度の計算方法は、本実施形態の例に限らず、任意の方法で行っても良い。

30

#### 【0060】

上記実施形態においては、抽出部121は、入力部110に入力されたURLに対応するHTML文書を、通信部150を介して取得した。しかし、HTML文書の取得方法はこれに限らない。例えば、URLの入力をせず、通信部150は、ユーザからHTML文書を直接受信しても良い。このように受信したHTML文書はメモリ140に格納されても良い。

#### 【0061】

なお、本実施形態においては、1つのコンピュータにより情報抽出装置100を実現したが、情報抽出装置100の機能を複数の機器により実現しても良い。例えば、入力部110及び表示部130を他の携帯端末に設けても良い。また、抽出部121、保存部122、及び追跡部123は、異なる部品であっても良い。

40

#### 【0062】

##### <実施形態2>

本実施形態の情報抽出装置は、抽出対象の候補となる可変要素として、対象者に紐づく情報のみを抽出することができるようにする。具体的には、本実施形態の情報抽出装置は、対象者の文書(本実施形態において、Webページ)内で短期間(例えば、1分毎)に変化した部分(本実施形態においては、現在時刻)を可変要素から除外する。このように、可変要素として抽出したくない要素(本実施形態の場合、現在時刻などの対象者に紐づ

50

かない情報)を除外要素として、可変要素から除外することにより、周辺情報の抽出や類似度の計算の処理(例えば、図2のステップS204及び図5のステップS503~S506)が速くなると共に、必要な情報だけを可変要素としてユーザに提示できる(図2のステップS205)。さらに、類似度に基づいた対応関係の復元の精度が良くなる(図5のステップS506)。

#### 【0063】

##### 2-1. 情報抽出装置の構成

本実施形態の情報抽出装置は、図1に示される実施形態1と同一の構成を持つ。

#### 【0064】

##### 2-2. 情報抽出装置の動作

図8に、抽出対象のWebページ(対象者のWebページ)のURLに対応する、1分経過前後のHTML文書を示す。この例では、現在時刻が「11:59」から「12:00」に変化している。実施形態1の場合、複数のWebページを比較した結果、現在時刻が異なれば、その現在時刻が可変要素として抽出される。しかし、現在時刻は、図8に示されるように、対象者が同一の場合でも、変化する要素である。本実施形態では、対象者が同一の場合でも変化する要素を可変要素から除外する。

#### 【0065】

図9に、本発明の実施形態2における除外候補の抽出及び除外のフローチャートを示す。図9に示す除外候補の抽出及び除外の工程は、可変要素の抽出前(図2のステップS203の直前)に行っても良いし、可変要素の抽出後(図2のステップS203の直後)に行っても良い。なお、図9に示す除外候補の抽出及び除外の工程は、任意のタイミングで行っても良いが、可変要素の周辺情報を抽出する(図2のステップS204)前にすることが好ましい。本実施形態においては、可変要素を抽出した後且つその周辺情報を抽出する前(図2のステップS203とステップS204の間)に、図9に示すステップS901~S908を行う。

#### 【0066】

本実施形態の情報抽出装置100の抽出部121は、「変化の頻度」を表すカウンタ値を「0」に設定し、図9に示す処理を開始する。抽出部121は、ステップS202で対象者のページ構成情報(WebページのHTML文書)を取得した後、所定時間(例えば、1分)が経過したかどうかを判断する(ステップS901)。所定時間が経過していれば(ステップS901でYes)、抽出部121は、対象者のURLに対応するページ構成情報を、通信部150を介して、再度取得する(ステップS902)。抽出部121は、今回取得したページ構成情報と前回取得したページ構成情報とを比較する(ステップS903)。具体的には、今回取得したHTML文書と前回取得したHTML文書の差分を計算する。抽出部121は、比較した結果、変化した箇所があるかどうかを判断し(ステップS904)、変化した箇所があれば、その変化した箇所を除外候補として抽出する(ステップS905)。これにより、例えば、図8に示される現在時刻の「11:59」及び/又は「12:00」が抽出される。また、ステップS905において、抽出部121は、「変化の頻度」を表すカウンタ値を「+1」する。

#### 【0067】

抽出部121は、対象者のページ構成情報の比較(ステップS903)を所定回数(例えば、10回)行ったかどうかを判断する(ステップS906)。所定回数行っていなければ(ステップS906でNo)、ステップS901に戻り、対象者のページ構成情報の比較の処理を繰り返す。所定回数の比較が完了すれば(ステップS906でYes)、抽出部121は、除外候補として抽出した要素の変化の頻度を表すカウンタ値が所定数(例えば、9回)以上かどうかを判断する(S907)。変化の頻度を表すカウンタ値が所定数以上であれば(ステップS907でYes)、抽出部121は、除外候補が可変要素から除外したい除外要素であると判断して、その除外候補を可変要素から除外する(ステップS908)。変化の頻度を表すカウンタ値が所定数以上でなければ(ステップS907でNo)、除外候補を可変要素から除外しない。このような処理により、例えば、1分経

10

20

30

40

50

過する毎に対象者のページ構成情報の変化の有無を検出し、10回中9回以上変化した箇所があれば、その変化した箇所（現在時刻）は対象者に依存した値ではない（時間に依存した値である）と判断して、可変要素から除外する。

【0068】

### 2-3. まとめ

本実施形態によれば、複数回取得した対象者のページ構成情報を比較して、変化した箇所（本実施形態において、現在時刻）を可変要素から除外することにより、対象者に紐づく情報（本実施形態において、55kg、54kg、171cm、坂本）のみを可変要素として抽出することができる。

【0069】

類似度に基づいて対応関係を復元する際に（図5のステップS506）、候補が多ければ多いほど対応関係を誤って復元する可能性が生じる。例えば、「体重」、「身長」、「気温」が可変要素としてある場合、最初に取得した初期ページの「体重」の数値と、新たに取得した現時点のページの「気温」の数値に、対応関係があると誤って判断してしまう可能性があり、その場合は現時点の体重の情報を追跡することができなくなる。類似度の計算が上手く行えない（例えば、可変要素の周囲の文言が少ない）ケースでは、可変要素の種類の数が多いと、対応関係の復元の失敗に繋がるおそれが生じる。よって、可変要素から、不要な除外要素を事前に除外することで、対応関係の復元の精度が高まる。

【0070】

### 2-4. 変形例

なお、ステップS903では、今回取得したページ構成情報を前回取得したページ構成情報と比較（例えば、12:00と11:59に取得したHTML文書と比較、12:01と12:00に取得したHTML文書とを比較）したが、最初に取得したページ構成情報（例えば、11:59に取得したHTML文書）を新たに取得したページ構成情報（例えば、12:00、12:01、12:02、12:03・・・に取得したHTML文書）と比較しても良い。

【0071】

また、本実施形態においては、除外要素を抽出するために変化させるコンテキスト（すなわち、ステップS901で使用する判定基準）は、Webページの取得時間であったが、除外要素を抽出するために変化させるコンテキストは、任意に設定可能である。例えば、抽出部121が設定しても良いし、ユーザが入力部110を介して設定しても良い。可変要素として抽出したい情報が何のコンテキストに基づいているかを考慮することにより、そのコンテキストが変わった時のみ変化する情報を可変要素として抽出することができる。例えば、天気やアクセス元の地域などを、除外要素を抽出するために変化させるコンテキストとして設定しても良い。これにより、例えば、現在時刻だけでなく、広告バナーなどの個人に紐づかない情報を、可変要素から除外することができる。

【0072】

また、本実施形態では、ステップS901の所定時間を1分、ステップS906の所定回数を10回、ステップS907の所定数を9回として、1分経過毎に対象者のページ構成情報を比較して、10回中9回以上変化した場合に、除外候補を可変要素から除外したが、ステップS901の所定時間（判定基準）、ステップS906の所定回数、ステップS907の所定数は任意に設定可能である。例えば、抽出部121が設定しても良いし、ユーザが入力部110を介して設定しても良い。また、可変要素として抽出したい情報に応じて、及び/又は除外要素を抽出するために変化させるコンテキストに応じて、ステップS901の所定時間（判定基準）、ステップS906の所定回数、ステップS907の所定数を設定しても良い。

【0073】

例えば、個人の体重、身長、及び名前は1分毎に変化する可能性は少ないため、1分経過する毎に対象者のページ構成情報の変化の有無を検出して、3回中3回変化した箇所を除外要素（現在時刻）としても良い。また、例えば、除外要素（広告バナー）を抽出する

10

20

30

40

50

ために変化させるコンテキストが「アクセス元の地域」である場合、アクセス元の地域が変わる毎に対象者のページ構成情報の変化の有無を検出し、5回中5回変化した箇所を除外要素としても良い。なお、誤判定を防ぐためには、複数回、比較することが好ましく、比較回数が多いほど誤判定を防ぐことができる。

#### 【0074】

可変要素から除外要素を除外するその他の例について、さらに説明する。SNSサービス（Facebook、Twitterなど）において、「通知件数」の情報を抽出して追跡するケースについて説明する。SNSサービスでは、他のユーザが書き込みなどを行うと、対象者（自分自身）のページの内容も変化するため、大量の可変要素が存在することになる。そのため、抽出対象となる可変要素を絞り込むことが重要になる。この場合、他のユーザが書き込みする前後で、対象者のWebページを取得して、取得したページ間の差分を比較することによって、除外要素（抽出対象として不要な可変要素）を発見する。具体的には、抽出手法用に、機械が操作するアカウントを用意し、機械アカウントと抽出を行いたいユーザを、情報を共有できるフレンド状態にする。その後、機械アカウントが書き込みを行う前に、一度ページを保存し、さらに、機械アカウントが書き込みを行った直後に、もう一度ページを保存し直す。機械アカウントが書き込みを行った前後のページ間で差分を計算することにより、除外要素（抽出対象として不要な可変要素）を除去する。なお、機械アカウントが書き込み中に、除外したくない要素である「通知件数」が増えてしまう場合もあるため、「通知件数」を誤って除外するのを防ぐために、試行回数を十分多くして、必要な変化件数を高めにするのが好ましい。例えば、「変化件数/試行回数（アクセス頻度）= 19/20」とし、20回中19回変化した箇所を除外する。

10

20

#### 【0075】

次に、「今日の天気」の情報を除外したいケースについて説明する。例えば、今日の天気の情報除外するためには、天気情報変化するように1日ごとにアクセスすることが考えられる。一方、「毎日のランニング距離」と「今日の天気」が、同一ページに掲載されている場合、1日ごとにアクセスをすると、ランニング距離も変化してしまうため、「ランニング距離」と「今日の天気」の両方が除外要素となってしまう。そのため、1日ごとのアクセスでは、「今日の天気」の情報のみを除外することができない。このような場合、「今日の天気」を除外するために、例えば、利用者の位置情報を変更して、東京の天気と、大阪の天気のように、天気の情報のみが変化するようにして、同一ページに複数回アクセスをする。このように、アクセスの頻度や変化件数は、抽出する情報や除外する情報に応じて設定すると良い。欲しい情報（可変要素）が変化しないという条件を満たし、且つ、不要な情報（除外要素）が変化するという条件を満たすような頻度や回数を設定する。これにより、より精度良く、不要な情報のみを除外要素として抽出し除外できる。

30

#### 【0076】

なお、実施形態2の除外要素の抽出（図9）を実施する代わりに、実施形態1の可変要素の抽出（図2のステップS203及び図5のステップS502）において、可変要素を抽出する範囲を制限しても良い。例えば、可変要素の抽出をHTML文書のBODYタグの中身の部分だけから行うようにしても良い。また、Webページの上にあるメニューのみから可変要素を抽出するようにしても良い。このように、除外要素を可変要素から除外する代わりに、可変要素の抽出箇所を絞り込んでも良い。抽出箇所を絞り込むことで、不要な情報を可変要素として抽出してしまうことを防ぐことができる。また、可変要素の抽出範囲の制限を、実施形態2の除外要素の抽出（図9）の実施と共に行っても良い。

40

#### 【産業上の利用可能性】

#### 【0077】

本発明の情報抽出装置は、構造化された文書の仕様変更の有無にかかわらず、特定情報を抽出し続けることができるため、定期的に特定情報を抽出して抽出した特定情報を利用するようなサービスに有用である。

#### 【符号の説明】

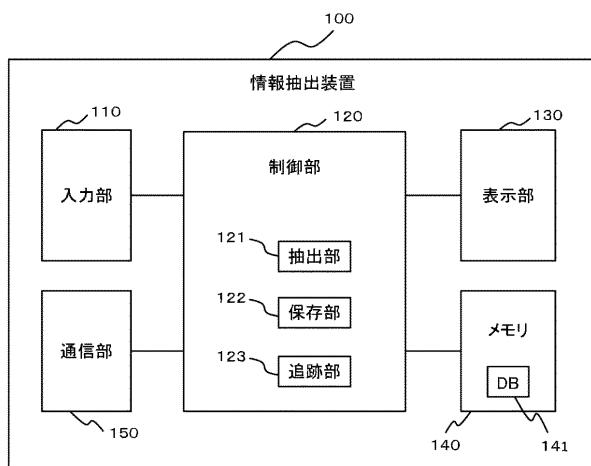
50



【 0 0 7 8 】

- 1 0 0 情報抽出装置
- 1 1 0 入力部
- 1 2 0 制御部
- 1 2 1 抽出部
- 1 2 2 保存部
- 1 2 3 追跡部
- 1 3 0 表示部
- 1 4 0 メモリ
- 1 4 1 データベース ( D B )
- 1 5 0 通信部

【 図 1 】



【 図 2 】



【 図 3 】

(a) `http://www.....profile-A` `http://www.....profile-B`

```

<div>あなたの体重は
<span id="bw">55kg</span>。
先月は54kg ! <div>
<div>あなたの身長は
<span id="height">171cm
</span>です !
</div>
<div>名前 : <span>坂本</span>
</div>
    
```

```

<div>あなたの体重は
<span id="bw">51kg</span>。
先月は52kg ! <div>
<div>あなたの身長は
<span id="height">173cm
</span>です !
</div>
<div>名前 : <span>佐藤</span>
</div>
    
```

(b)

(c)

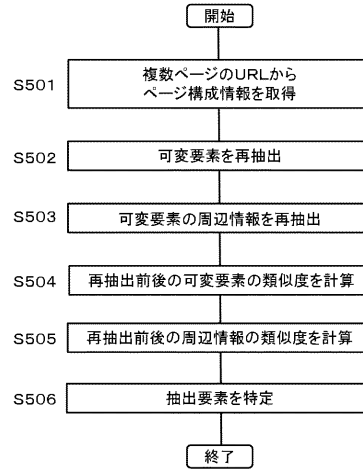
```

55kg
54kg
171cm
坂本
    
```

【 図 4 】

可変要素	周辺情報	選択の有無
55kg	“あなたの体重は”、span、id、“bw”、 /span、“ε”	○(抽出対象)
54kg	“先月は”、“!”	×
171cm	span、id、“height”、/span、“です!”	×
坂本	div、“名前:”、span、/span、/div	×

【 図 5 】



【 図 6 】

```

<div>あなたの体重は
<span id="bw">55kg</span>。
先月は54kg ! <div>
<div>あなたの身長は
<span id="height">171cm
</span>です !
</div>
<div>名前 : <span>坂本</span>
</div>
    
```

仕様変更

```

<div>名前 : <span id="name">
坂本</span></div>
<div>体重 : <span id="bw">
56kg</span></div>
<div>先月の体重 : <span id="lbw">
55kg</span></div>
<div>身長 : <span id="height">
171cm</span></div>
    
```

【 図 7 】

記録済 \ 再抽出後	坂本	56kg	55kg	171cm
55kg	0.1	<b>0.4</b>	0.3	0.2
54kg	0.2	0.3	<b>0.2</b>	0.2
171cm	0.2	0.2	0.2	<b>0.4</b>
坂本	<b>0.5</b>	0.1	0.1	0.1

【 図 8 】

```

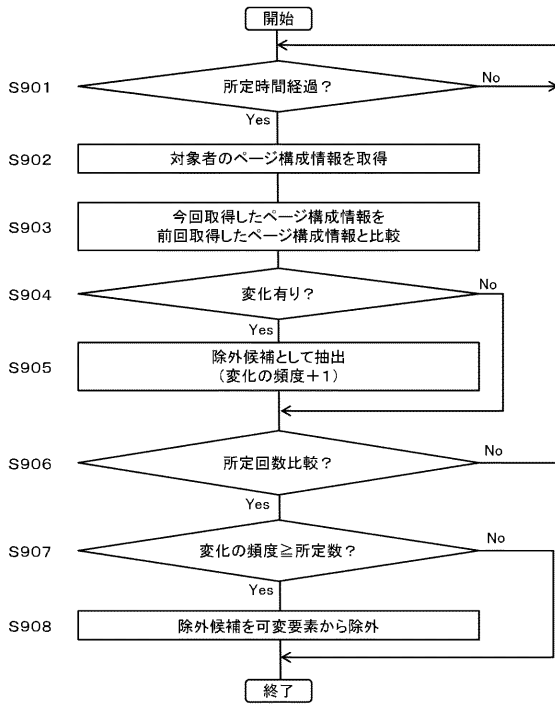
<div>現在時刻は 11:59 </div>
<div>あなたの体重は55kgです !
先月は54kgでした。 </div>
<div>あなたの身長は
<span id="height">171cm
</span>です !
</div>
<div>名前 : <span>坂本</span>
</div>
    
```

1分後

```

<div>現在時刻は 12:00 </div>
<div>あなたの体重は55kgです !
先月は54kgでした。 </div>
<div>あなたの身長は
<span id="height">171cm
</span>です !
</div>
<div>名前 : <span>坂本</span>
</div>
    
```

【 図 9 】



## 【 国際調査報告 】

INTERNATIONAL SEARCH REPORT		International application No. PCT/JP2015/084974
A. CLASSIFICATION OF SUBJECT MATTER G06F17/30(2006.01)i		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED		
Minimum documentation searched (classification system followed by classification symbols) G06F17/30		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2016 Kokai Jitsuyo Shinan Koho 1971-2016 Toroku Jitsuyo Shinan Koho 1994-2016		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2005-63332 A (Fujitsu Ltd.), 10 March 2005 (10.03.2005), entire text; all drawings (Family: none)	1-15
A	JP 2004-178426 A (Mitsubishi Electric Corp.), 24 June 2004 (24.06.2004), entire text; all drawings (Family: none)	1-15
A	JP 2007-293874 A (Digital Works Kabushiki Kaisha), 08 November 2007 (08.11.2007), entire text; all drawings (Family: none)	1-15
<input checked="" type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.		
* Special categories of cited documents:		
"A"	document defining the general state of the art which is not considered to be of particular relevance	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"E"	earlier application or patent but published on or after the international filing date	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means	"&" document member of the same patent family
"P"	document published prior to the international filing date but later than the priority date claimed	
Date of the actual completion of the international search 22 February 2016 (22.02.16)		Date of mailing of the international search report 01 March 2016 (01.03.16)
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan		Authorized officer  Telephone No.

**INTERNATIONAL SEARCH REPORT**

International application No.

PCT/JP2015/084974

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2013/0226944 A1 (MICROSOFT CORP.), 29 August 2013 (29.08.2013), entire text; all drawings (Family: none)	1-15
A	EP 2648115 A1 (SEEBURGER AG), 09 October 2013 (09.10.2013), entire text; all drawings (Family: none)	1-15
A	US 2014/0297670 A1 (ORACLE INTERNATIONAL CORP.), 02 October 2014 (02.10.2014), entire text; all drawings (Family: none)	1-15

国際調査報告		国際出願番号 PCT/J P 2 0 1 5 / 0 8 4 9 7 4	
A. 発明の属する分野の分類 (国際特許分類 (IPC)) Int.Cl. G06F17/30(2006,01)i			
B. 調査を行った分野 調査を行った最小限資料 (国際特許分類 (IPC)) Int.Cl. G06F17/30			
最小限資料以外の資料で調査を行った分野に含まれるもの 日本国実用新案公報 1922-1996年 日本国公開実用新案公報 1971-2016年 日本国実用新案登録公報 1996-2016年 日本国登録実用新案公報 1994-2016年			
国際調査で使用した電子データベース (データベースの名称、調査に使用した用語)			
C. 関連すると認められる文献			
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号	
A	JP 2005-63332 A (富士通株式会社) 2005.03.10, 全文、全図 (ファミリーなし)	1-15	
A	JP 2004-178426 A (三菱電機株式会社) 2004.06.24, 全文、全図 (ファミリーなし)	1-15	
A	JP 2007-293874 A (デジタル・ワークス株式会社) 2007.11.08, 全文、全図 (ファミリーなし)	1-15	
<input checked="" type="checkbox"/> C欄の続きにも文献が列挙されている。 <input type="checkbox"/> パテントファミリーに関する別紙を参照。			
* 引用文献のカテゴリー 「A」 特に関連のある文献ではなく、一般的技術水準を示すもの 「E」 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの 「L」 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す) 「O」 口頭による開示、使用、展示等に言及する文献 「P」 国際出願日前で、かつ優先権の主張の基礎となる出願			
国際調査を完了した日 22.02.2016		国際調査報告の発送日 01.03.2016	
国際調査機関の名称及びあて先 日本国特許庁 (ISA/J P) 郵便番号100-8915 東京都千代田区霞が関三丁目4番3号		特許庁審査官 (権限のある職員) 野崎 大進	5M 9289 電話番号 03-3581-1101 内線 3599

国際調査報告		国際出願番号 PCT/J P 2 0 1 5 / 0 8 4 9 7 4
C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	US 2013/0226944 A1 (MICROSOFT CORPORATION) 2013.08.29, 全文、全図 (ファミリーなし)	1 - 1 5
A	EP 2648115 A1 (SEEBURGER AG) 2013.10.09, 全文、全図 (ファミリーなし)	1 - 1 5
A	US 2014/0297670 A1 (ORACLE INTERNATIONAL CORPORATION) 2014.10.02, 全文、全図 (ファミリーなし)	1 - 1 5

## フロントページの続き

(51) Int.Cl. F I テーマコード(参考)  
G 0 6 F 17/22 6 1 1

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

(72) 発明者 本位田 真一  
東京都千代田区一ツ橋二丁目1番2号 大学共同利用機関法人情報・システム研究機構 国立情報学研究所内

Fターム(参考) 5B109 NH12 SA14

(注) この公表は、国際事務局(WIPO)により国際公開された公報を基に作成したものである。なおこの公表に係る日本語特許出願(日本語実用新案登録出願)の国際公開の効果は、特許法第184条の10第1項(実用新案法第48条の13第2項)により生ずるものであり、本掲載とは関係ありません。