(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6562276号
(P6562276)

(45)発行日　令和1年8月21日(2019.8.21)　(24)登録日　令和1年8月2日(2019.8.2)

(51)Int.Cl.		F I	
G06F 16/80	(2019.01)	G06F	16/80
G06F 16/33	(2019.01)	G06F	16/33
G06F 16/383	(2019.01)	G06F	16/383
G06F 16/903	(2019.01)	G06F	16/903
G06F 16/908	(2019.01)	G06F	16/908

請求項の数 15　(全 19 頁)　最終頁に続く

(21)出願番号　特願2016-564846 (P2016-564846)
(86)(22)出願日　平成27年12月14日 (2015.12.14)
(86)国際出願番号　PCT/JP2015/084974
(87)国際公開番号　WO2016/098739
(87)国際公開日　平成28年6月23日 (2016.6.23)
審査請求日　平成30年6月27日 (2018.6.27)
(31)優先権主張番号　特願2014-253058 (P2014-253058)
(32)優先日　平成26年12月15日 (2014.12.15)
(33)優先権主張国・地域又は機関
日本国(JP)

(73)特許権者　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
(74)代理人　100100158
弁理士　鮫島　睦
(74)代理人　100131808
弁理士　柳橋　泰雄
(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)【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とを計算し、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、

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とを計算し、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】
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。
【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。

【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い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。

【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

【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か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際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。

【図面の簡単な説明】

【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
【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。
【００２２】
１－１．　情報抽出装置の構成
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。
【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。

【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。

【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。

【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。

【００３０】

１－２．　情報抽出装置の動作

図２に、情報抽出装置１００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１４０に格納される抽出情報のＤＢ１４１の例を示している。

【００３１】

以下、図４に示すように、「５５ｋｇ（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。

【００３２】

図２のフローにおいて、まず、入力部１１０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Ｓ２０１）。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１２２は、入力したＵＲＬをメモリ１４０に格納する。抽出部１２１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１５０を介して、取得する（ステップＳ２０２）。

【００３３】

抽出部１２１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Ｓ２０３）。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（「５５ｋｇ」、「５４ｋｇ」、「１７１ｃｍ」、「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（ＸＤｉｆｆ：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。

【００３４】

抽出部１２１は、可変要素から所定範囲内（例えば、可変要素の周囲１００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Ｓ２０４）。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“５５ｋｇ”に対して、テキスト（“あなたの体重は”、“。”）、ＨＴＭＬタグ（ｄｉｖ、ｓｐａｎ）、属性名（ｉｄ）及び属性値（“ｈｉｇｈｔ”）を抽出する（例えば、“あなたの体重は”、span、id、“bw”、/span、“。”）。

【００３５】

抽出部１２１は、図３（ｃ）に示すように、抽出した可変要素を表示部１３０に表示する（ステップＳ２０５）。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「５５ｋｇ（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１１０は、その選択を入力する（ステップＳ２０６）。保存部１２２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１１０を介して取得した抽出対象としての選択の有無を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１ｋｇ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」、「５６ｋｇ」、「５５ｋｇ」、「１７１ｃｍ」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）は以下のようになる。

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）は以下のようになる。

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

【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５５ｋｇ。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（「 名前：坂本 」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（「 名前：坂本 」から、前後２個（Ｘ＝２）であれば、「“id”、“name”、“/span”、“/div”」を抽出。「 あなたの体重は５５ｋｇ。先月は５４ｋｇ！ 」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「“id”,“bw”,“/span”,“。”,“先月”,“は”,“５４ｋｇ”,“！”」に変換）。

【００４６】

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５５ｋｇの周辺情報の単語列は「“id”,“bw”,“/span”,“。”,“先月”,“は”,“５４kg”,“！”」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”」、

５６ｋｇの周辺情報（２）「“id”,“bw”,“/span”,“/div”」、

５５ｋｇの周辺情報（３）「“id”,“lbw”,“/span”,“/div”」、

１７１ｃｍの周辺情報（４）「“id”,“height”,“/span”,“/div”」となる。

【００４７】

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（Ｓ２）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「５５ｋｇ」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）。

【００４８】

このようにして求めた周辺情報の類似度（Ｓ２）は以下のようになる。

再抽出後の 周辺情報	（１） 坂本	（２） ５６ｋｇ	（３） ５５ｋｇ	（４） １７１ｃｍ
共通する単語数	４	６	４	４
類似度	０．３３３	０．５	０．３３３	０．３３３

【００４９】
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（Ｓ４）及び数字部の類似度（Ｓ３）並びに周辺要素の類似度（Ｓ２）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＝（（Ｓ３×Ｃ＋Ｓ４×Ｄ）×Ａ＋Ｓ２×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「５５ｋｇ」に関する最終的な類似度は以下のようになる。

記録済＼再抽出後	坂本	５６ｋｇ	５５ｋｇ	１７１ｃｍ
５５ｋｇ （現在の体重）	０．２６６４	０．５６６	０．４６６４	０．２９９４

【００５０】
また、「類似度＝（（Ｓ３×Ｃ＋Ｓ４×Ｄ）×Ａ＋Ｓ２×Ｂ））」により得られた、Ｗｅｂ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）５５ｋｇ」の行において、再抽出後の可変要素の中の「５６ｋｇ」が、抽出要素「５５ｋｇ」に対して、類似度が０．４と最も高い。よって、再抽出後の「５６ｋｇ」と抽出対象として記録済みの「５５ｋｇ」とに対応関係があるとみなす。すなわち、再抽出後の「５６ｋｇ」が抽出要素として特定される。
【００５１】
なお、記録済みの「５４ｋｇ」についても、再抽出後の「５６ｋｇ」に対して、類似度０．３と最も類似度が高い。しかし、「５５ｋｇ（記録済み）」と「５６ｋｇ（再抽出後）」のペアは類似度０．４であるのに対し、「５４ｋｇ（記録済み）」と「５６ｋｇ（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「５５ｋｇ（記録済み）」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。
【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。

【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が変更される頻度が高く、例えば図６のように仕様が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。
【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。
【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常的に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。
【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、「 名前：坂本 」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）。

【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。

【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。

【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文書を比較、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】

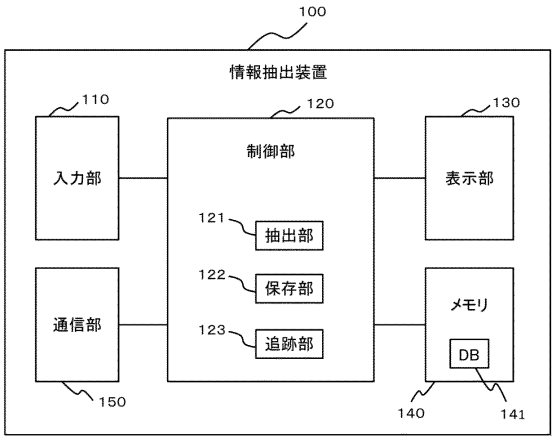
例えば、個人の体重、身長、及び名前は１分毎に変化する可能性は少ないため、１分経過する毎に対象者のページ構成情報の変化の有無を検出して、３回中３回変化した箇所を除外要素（現在時刻）としても良い。また、例えば、除外要素（広告バナー）を抽出するために変化させるコンテキストが「アクセス元の地域」である場合、アクセス元の地域が変わる毎に対象者のページ構成情報の変化の有無を検出し、５回中５回変化した箇所を除外要素としても良い。なお、誤判定を防ぐためには、複数回、比較することが好ましく、比較回数が多いほど誤判定を防ぐことができる。

【００７４】

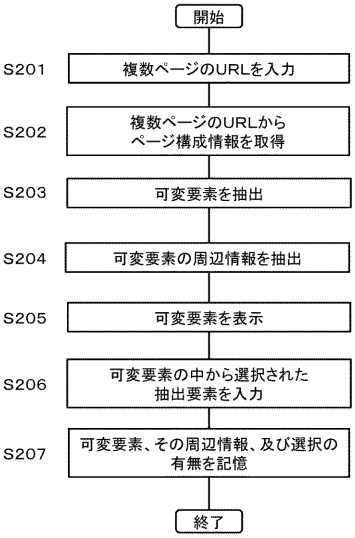
可変要素から除外要素を除外するその他の例について、さらに説明する。ＳＮＳサービス（Facebook、Twitterなど）において、「通知件数」の情報を抽出して追跡するケースについて説明する。ＳＮＳサービスでは、他のユーザが書き込みなどを行うと、対象者（自分自身）のページの内容も変化するため、大量の可変要素が存在することになる。そのため、抽出対象となる可変要素を絞り込むことが重要になる。この場合、他のユーザが書き込みする前後で、対象者のＷｅｂページを取得して、取得したページ間の差分を比較することによって、除外要素（抽出対象として不要な可変要素）を発見する。具体的には、抽出手法用に、機械が操作する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ことが好ましい。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。

【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、今日の天気の情報を除外するためには、天気情報が変化するように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。
【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ページの上部にあるメニューバ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。
【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】
【００７８】
１００　情報抽出装置
１１０　入力部
１２０　制御部
１２１　抽出部
１２２　保存部
１２３　追跡部
１３０　表示部
１４０　メモリ
１４１　データベース（ＤＢ）
１５０　通信部

【図１】



【図２】



【図３】

(a)

http://www.....profile-A	http://www.....profile-B

(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>
```

(c)

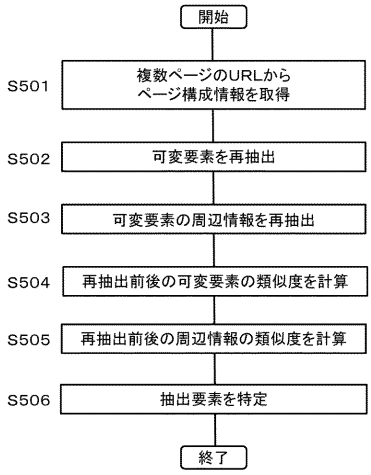
55kg
54kg
171cm
坂本

【図４】

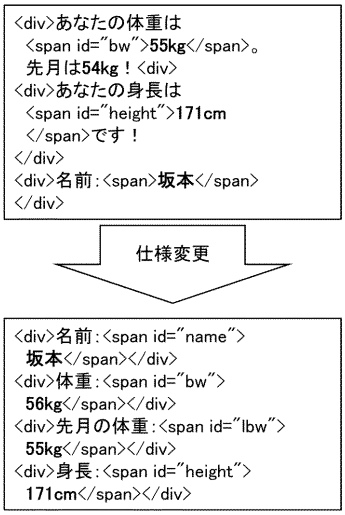
可変要素	周辺情報	選択の有無
55kg	“あなたの体重は”、span、id、“bw”、/span、“。”	○（抽出対象）
54kg	“先月は”、“！”	×
171cm	span、id、“height”、/span、“です！”	×
坂本	div、“名前：”、span、/span、/div	×

141

【図５】



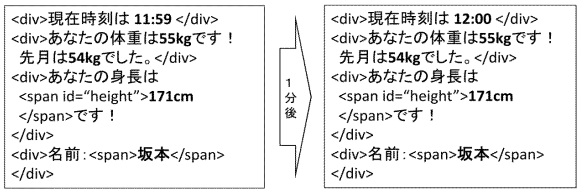
【図６】



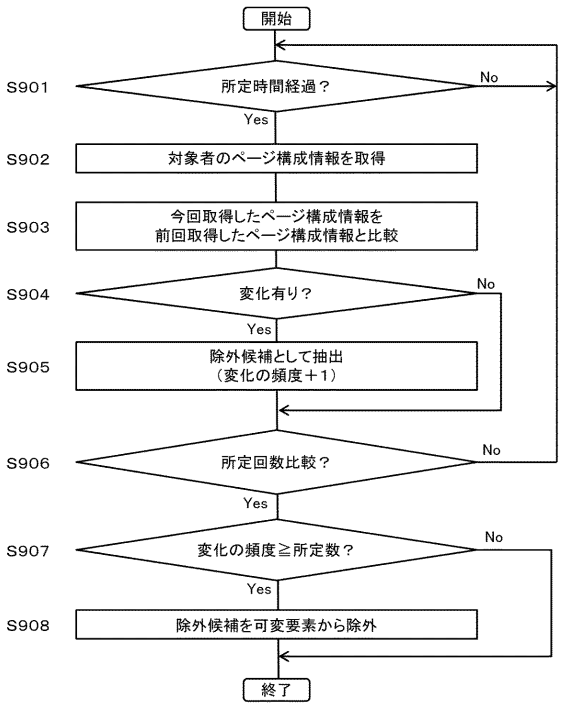
【図７】

記録済 ＼ 再抽出後	坂本	56kg	55kg	171cm
55kg	0.1	**0.4**	0.3	0.2
54kg	0.2	0.3	0.2	0.2
171cm	0.2	0.2	0.2	0.4
坂本	0.5	0.1	0.1	0.1

【図８】



【図９】



フロントページの続き

(51)Int.Cl. FI

G06F 17/22 (2006.01) G06F 17/22 647

G06F 17/22 611

(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

審査官 篠塚 隆

(56)参考文献 特開２００５－６３３３２（ＪＰ，Ａ）
特開２００４－１７８４２６（ＪＰ，Ａ）
特開２００７－２９３８７４（ＪＰ，Ａ）
米国特許出願公開第２０１３／０２２６９４４（ＵＳ，Ａ１）
欧州特許出願公開第２６４８１１５（ＥＰ，Ａ１）
米国特許出願公開第２０１４／０２９７６７０（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ１６／００
Ｇ０６Ｆ１７／２２