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

特許第5007405号
(P5007405)

(45)発行日　平成24年8月22日(2012.8.22)　(24)登録日　平成24年6月8日(2012.6.8)

(51)Int.Cl.　　F I
G06F 13/00 (2006.01)　　G06F 13/00 550A
G06F 17/21 (2006.01)　　G06F 17/21 501T

請求項の数 12　(全 50 頁)

(21)出願番号　特願2008-42722(P2008-42722)
(22)出願日　平成20年2月25日(2008.2.25)
(65)公開番号　特開2009-199512(P2009-199512A)
(43)公開日　平成21年9月3日(2009.9.3)
審査請求日　平成22年9月27日(2010.9.27)

（出願人による申告）平成１９年度独立行政法人情報通信研究機構、研究テーマ「軽度脳障害者のための情報セラピーインタフェースの研究開発」に関する委託研究、産業技術力強化法第１９条の適用を受ける特許出願

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100115749
弁理士　谷川　英和
(72)発明者　米澤　朋子
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　光永　法明
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　宮下　敬宏
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

最終頁に続く

(54)【発明の名称】情報処理装置、およびプログラム

(57)【特許請求の範囲】
【請求項１】
１以上のウェブページを格納し得るウェブページ格納部と、
分野を示す情報である分野情報と、ページ内のレイアウトに関する情報であるレイアウト情報との組を、２組以上格納し得るレイアウト情報格納部と、
前記１以上のウェブページから、電子ブックのページ単位の情報であるページ情報を、２以上取得するページ情報取得部と、
前記１以上のウェブページの内容から、当該１以上のウェブページの分野を決定し、分野情報を取得する分野情報決定部と、
前記ページ情報取得部が取得した２以上のページ情報を出力するページ情報出力部を具備し、
前記ページ情報出力部は、
前記分野情報決定部が取得した分野情報と対になるレイアウト情報を取得するレイアウト情報取得手段と、
前記レイアウト情報に従って、前記ページ情報取得部が取得した２以上のページ情報をレイアウトするレイアウト手段と、
前記レイアウト手段がレイアウトした２以上のページ情報を出力するページ情報出力手段とを具備する情報処理装置。
【請求項２】
前記ページ情報取得部は、

前記１以上のウェブページから情報を取得するための情報であるページ取得ルール情報を、２以上格納しているページ取得ルール情報格納手段と、
前記１以上のウェブページに対応するページ取得ルール情報、または前記分野情報決定部が取得した分野情報と対になるページ取得ルール情報を用いて、前記１以上のウェブページから、ページ単位の情報であるページ情報を、２以上取得するページ情報取得手段とを具備する請求項１記載の情報処理装置。
【請求項３】
前記分野情報決定部は、
分野情報と、当該分野情報が示す分野に対応する１以上の用語を含む用語情報を１以上格納している用語格納手段と、
前記１以上のウェブページから、前記１以上の用語の出現の度合いに関する情報である出現度情報を取得する出現度情報取得手段と、
前記出現度情報を用いて、前記１以上のウェブページの分野を決定し、分野情報を取得する分野情報決定手段を具備する請求項１または請求項２記載の情報処理装置。
【請求項４】
前記分野情報決定部は、
前記１以上のウェブページから、１以上のデータタイプのデータ量またはデータサイズに関する情報であるデータ情報を取得するデータ情報取得手段と、
前記データ情報を用いて、前記１以上のウェブページの分野を決定し、分野情報を取得する分野情報決定手段を具備する請求項１または請求項２記載の情報処理装置。
【請求項５】
１以上のウェブページを格納し得るウェブページ格納部と、
前記１以上のウェブページから、電子ブックのページ単位の情報であるページ情報を、２以上取得するページ情報取得部と、
前記ページ情報取得部が取得した２以上のページ情報を出力するページ情報出力部を具備し、
前記ページ情報取得部は、
前記１以上のウェブページの各々に対して、ウェブページの構造またはタグまたは内容を用いて、主となる情報を取得する主情報取得手段と、
前記主情報取得手段が取得した情報から、ページ情報を取得するページ情報取得手段を具備する情報処理装置。
【請求項６】
１以上のウェブページを格納し得るウェブページ格納部と、
前記１以上のウェブページから、電子ブックのページ単位の情報であるページ情報を、２以上取得するページ情報取得部と、
前記ページ情報取得部が取得した２以上のページ情報を出力するページ情報出力部を具備し、
前記ページ情報取得部は、
２以上のウェブページの中の一のウェブページを決定し、当該一のウェブページと、他の１以上の各ウェブページの組のうち、１以上の組に対して、内容の共通の度合いを示す情報である共通度を、１以上、取得する共通度取得手段と、
前記２以上のウェブページから２以上のページ情報を取得するページ情報取得手段を具備し、
前記ページ情報取得手段は、
前記共通度に応じて、前記他の１以上の各ウェブページからの情報の取得方法が異なる情報処理装置。
【請求項７】
前記共通度取得手段は、
前記２以上の各ウェブページのデータ量を算出し、データ量が最も大きいウェブページに対する、他の各ウェブページの共通度を取得し、

前記ページ情報取得手段は、
前記共通度が予め格納されている第一の閾値より小さい場合には、データ量が最も大きいウェブページに対して、前記他のウェブページを結合し、
前記共通度が前記第一の閾値より大きい場合には、各ウェブページを構成する一まとまりの情報であるブロックを取得し、前記データ量が最も大きいウェブページのブロック毎に、前記他のウェブページのブロックのうちで、予め格納されている第二の閾値より大きい共通度を有するブロックを抽出し、当該抽出したブロックを、前記データ量が最も大きいウェブページの前記ブロックと次のブロックの間に挿入し、出力する全ページ情報を構成し、
前記全ページ情報を、１ページ単位に区切り、２以上のページ情報を取得する請求項６記載の情報処理装置。
【請求項８】
前記ページ情報取得手段は、
前記共通度が前記第一の閾値より大きい場合、前記他のウェブページのブロック中で、前記データ量が最も大きいウェブページの全ブロックのいずれに対しても、前記第二の閾値より大きな閾値を有しないブロックについて、その前に配置されているブロックで、前記データ量が最も大きいウェブページのいずれかのブロックに対して前記第二の閾値より大きな閾値を有するとして、前記データ量が最も大きいウェブページ内に挿入された箇所の直後に挿入する処理をさらに行い、出力する全ページ情報を構成し、前記全ページ情報を、１ページ単位に区切り、２以上のページ情報を取得する請求項７記載の情報処理装置。
【請求項９】
前記ページ情報取得部は、
ユーザから受け付けたキーワードを用いて、１以上のサーバ装置からウェブページを検索し、当該検索したウェブページから、上位のｎ（ｎは２以上の自然数）のウェブページを取得する請求項５から請求項８いずれか記載の情報処理装置。
【請求項１０】
記憶媒体に、
分野を示す情報である分野情報と、ページ内のレイアウトに関する情報であるレイアウト情報との組を、２組以上格納しており、
コンピュータを、
１以上のウェブページから、電子ブックのページ単位の情報であるページ情報を、２以上取得するページ情報取得部と、
前記１以上のウェブページの内容から、当該１以上のウェブページの分野を決定し、分野情報を取得する分野情報決定部と、
前記ページ情報取得部が取得した２以上のページ情報を出力するページ情報出力部として機能させるためのプログラムであって、
前記ページ情報出力部は、
前記分野情報決定部が取得した分野情報と対になるレイアウト情報を取得するレイアウト情報取得手段と、
前記レイアウト情報に従って、前記ページ情報取得部が取得した２以上のページ情報をレイアウトするレイアウト手段と、
前記レイアウト手段がレイアウトした２以上のページ情報を出力するページ情報出力手段とを具備するものとして、コンピュータを機能させるプログラム。
【請求項１１】
コンピュータを、
１以上のウェブページから、電子ブックのページ単位の情報であるページ情報を、２以上取得するページ情報取得部と、
前記ページ情報取得部が取得した２以上のページ情報を出力するページ情報出力部として機能させるプログラムであって、
前記ページ情報取得部は、

前記１以上のウェブページの各々に対して、ウェブページの構造またはタグまたは内容を用いて、主となる情報を取得する主情報取得手段と、
前記主情報取得手段が取得した情報から、ページ情報を取得するページ情報取得手段を具備するものとして、コンピュータを機能させるプログラム。
【請求項１２】
コンピュータを、
１以上のウェブページから、電子ブックのページ単位の情報であるページ情報を、２以上取得するページ情報取得部と、
前記ページ情報取得部が取得した２以上のページ情報を出力するページ情報出力部として機能させるプログラムであって、
前記ページ情報取得部は、
２以上のウェブページの中の一のウェブページを決定し、当該一のウェブページと、他の１以上の各ウェブページの組のうち、１以上の組に対して、内容の共通の度合いを示す情報である共通度を、１以上、取得する共通度取得手段と、
前記２以上のウェブページから２以上のページ情報を取得するページ情報取得手段を具備し、
前記ページ情報取得手段は、
前記共通度に応じて、前記他の１以上の各ウェブページからの情報の取得方法が異なるものとして、コンピュータを機能させるプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、１以上のＷｅｂページから、ページ単位の情報に分割した電子ブックのコンテンツを得る情報処理装置等に関するものである。
【背景技術】
【０００２】
従来、より幅広いユーザ層が快適にウェブページ等の閲覧対象情報を閲覧できるように配慮した情報閲覧システムがあった（例えば、特許文献１参照）。本システムにおいて、ウェブページをＴＶモニタに表示させるために、見出しとなる表示要素のサイズを拡大させるなどの処理を行っていた。
【特許文献１】特開２００６－２７９８８７号公報（第１頁、第１図等）
【発明の開示】
【発明が解決しようとする課題】
【０００３】
しかしながら、従来の情報閲覧システムにおいては、ウェブページから、複数ページの集合である電子ブックのコンテンツを自動構成することはできなかった。
【課題を解決するための手段】
【０００４】
本第一の発明の情報処理装置は、１以上のウェブページを格納し得るウェブページ格納部と、前記１以上のウェブページから、ページ単位の情報であるページ情報を、２以上取得するページ情報取得部と、前記ページ情報取得部が取得した２以上のページ情報を出力するページ情報出力部を具備する情報処理装置である。
【０００５】
かかる構成により、ウェブページから、複数ページの集合である電子ブックのコンテンツを自動構成することができる。
【０００６】
また、本第二の発明の情報処理装置は、第一の発明に対して、分野を示す情報である分野情報と、ページ内のレイアウトに関する情報であるレイアウト情報との組を、２組以上格納しているレイアウト情報格納部と、前記１以上のウェブページの内容から、当該１以上のウェブページの分野を決定し、分野情報を取得する分野情報決定部とをさらに具備し

、前記ページ情報出力部は、前記分野情報決定部が取得した分野情報と対になるレイアウト情報を取得するレイアウト情報取得手段と、前記レイアウト情報に従って、前記ページ情報取得部が取得した２以上のページ情報をレイアウトするレイアウト手段と、前記レイアウト手段がレイアウトした２以上のページ情報を出力するページ情報出力手段とを具備する情報処理装置である。

【０００７】

かかる構成により、ウェブページの内容に適したレイアウトで、電子ブックのコンテンツを自動構成することができる。

【０００８】

また、本第三の発明の情報処理装置は、第二の発明に対して、前記ページ情報取得部は、分野情報と、前記１以上のウェブページから情報を取得するための情報であるページ取得ルール情報との組を、２以上格納しているページ取得ルール情報格納手段と、前記分野情報決定部が取得した分野情報と対になるページ取得ルール情報を用いて、前記１以上のウェブページから、ページ単位の情報であるページ情報を、２以上取得するページ情報取得手段とを具備する情報処理装置である。

【０００９】

かかる構成により、ウェブページの分野を適切に判断した上で、ウェブページの内容に適したレイアウトで、電子ブックのコンテンツを自動構成することができる。

【００１０】

また、本第四の発明の情報処理装置は、第二、第三いずれかの発明に対して、前記分野情報決定部は、分野情報と、当該分野情報が示す分野に対応する１以上の用語を含む用語情報を１以上格納している用語格納手段と、前記１以上のウェブページから、前記１以上の用語の出現の度合いに関する情報である出現度情報を取得する出現度情報取得手段と、前記出現度情報を用いて、前記１以上のウェブページの分野を決定し、分野情報を取得する分野情報決定手段を具備する情報処理装置である。

【００１１】

かかる構成により、例えば、ニュース記事のウェブページから、新聞風のレイアウトのコンテンツが構成でき、ブログのウェブページから、日記風のレイアウトのコンテンツが構成できる。

【００１２】

また、本第五の発明の情報処理装置は、第二、第三いずれかの発明に対して、前記分野情報決定部は、前記１以上のウェブページから、１以上のデータタイプのデータ量またはデータサイズに関する情報であるデータ情報を取得するデータ情報取得手段と、前記データ情報を用いて、前記１以上のウェブページの分野を決定し、分野情報を取得する分野情報決定手段を具備する情報処理装置である。

【００１３】

かかる構成により、例えば、子供向けのウェブページから、絵本風のレイアウトのコンテンツが構成できる。

【００１４】

また、本第六の発明の情報処理装置は、第一の発明に対して、前記ページ情報取得部は、前記１以上のウェブページの各々に対して、ウェブページの構造またはタグまたは内容を用いて、主となる情報を取得する主情報取得手段と、前記主情報取得手段が取得した情報から、ページ情報を取得するページ情報取得手段を具備する情報処理装置である。

【００１５】

かかる構成により、ウェブページに記載されている情報のうち、重要と思われる情報のみを用いて、電子ブックのコンテンツを自動構成することができる。

【００１６】

また、本第七の発明の情報処理装置は、第一の発明に対して、前記ページ情報取得部は、２以上のウェブページの各々から、内容の共通の度合いを示す情報である共通度を取得する共通度取得手段と、前記２以上のウェブページから２以上のページ情報を取得する手

段であり、前記共通度に応じて、取得方法が異なるページ情報取得手段を具備する情報処理装置である。

【００１７】

かかる構成により、複数のウェブページの内容を用いて、優れた電子ブックのコンテンツを自動構成することができる。

【００１８】

また、本第八の発明の情報処理装置は、第七の発明に対して、前記共通度取得手段は、前記２以上の各ウェブページのデータ量を算出し、データ量が最も大きいウェブページに対する、他の各ウェブページの共通度を取得し、前記ページ情報取得手段は、前記共通度が予め格納されている第一の閾値より小さい場合には、データ量が最も大きいウェブページに対して結合し、前記共通度が前記第一の閾値より大きい場合には、各ウェブページを構成する一まとまりの情報であるブロックを取得し、前記データ量が最も大きいウェブページのブロック毎に、前記他の各ウェブページのブロックのうちで、予め格納されている第二の閾値より大きい共通度を有するブロックを抽出し、当該抽出したブロックを、前記データ量が最も大きいウェブページの前記ブロックと次のブロックの間に挿入し、出力する全ページ情報を構成し、前記全ページ情報を、１ページ単位に区切り、２以上のページ情報を取得する情報処理装置である。

【００１９】

かかる構成により、複数のウェブページの内容を用いて、近似する内容の情報が近いところに配置され、整理された電子ブックのコンテンツを自動構成することができる。

【００２０】

また、本第九の発明の情報処理装置は、第八の発明に対して、前記ページ情報取得手段は、前記他の各ウェブページのブロック中で、前記データ量が最も大きいウェブページの全ブロックのいずれに対しても、前記第二の閾値より大きな閾値を有しないブロックについて、その前に配置されているブロックで、前記データ量が最も大きいウェブページのいずれかのブロックに対して前記第二の閾値より大きな閾値を有するとして、前記データ量が最も大きいウェブページ内に挿入された箇所の直後に挿入する処理をさらに行い、出力する全ページ情報を構成し、前記全ページ情報を、１ページ単位に区切り、２以上のページ情報を取得する情報処理装置である。

【００２１】

かかる構成により、複数のウェブページの内容を用いて、近似する内容の情報が近いところに配置され、整理された電子ブックのコンテンツを自動構成することができる。

【００２２】

また、本第十の発明の情報処理装置は、第七から第十いずれかの発明に対して、前記ページ情報取得部は、ユーザから受け付けたキーワードを用いて、１以上のサーバ装置からウェブページを検索し、当該検索したウェブページから、上位のｎ（ｎは２以上の自然数）のウェブページを取得する情報処理装置である。

【００２３】

かかる構成により、ユーザが欲する情報を、複数のウェブページから、自動的に選択し、有用な電子ブックのコンテンツを自動構成することができる。

【発明の効果】

【００２４】

本発明による情報処理システムによれば、ウェブページから、複数ページの集合である電子ブックのコンテンツを自動構成することができる。

【発明を実施するための最良の形態】

【００２５】

以下、情報処理システム等の実施形態について図面を参照して説明する。なお、実施の形態において同じ符号を付した構成要素は同様の動作を行うので、再度の説明を省略する場合がある。

【００２６】

（実施の形態１）
本実施の形態において、１以上のウェブページから、ページ単位の情報に分割した電子ブックのコンテンツを得る情報処理システム等について説明する。また、本実施の形態において、ウェブページから分野（ニュース、子供向けなど）を決定し、当該分野に対応するテンプレート（後述するレイアウト情報と同意義）を取得し、当該テンプレートに従って、ウェブページを電子ブックのコンテンツに変換する情報処理システム等について説明する。また、ウェブページのＵＲＬによって、異なるテンプレートを取得し、当該テンプレートに従って、ウェブページを電子ブックのコンテンツに変換する情報処理システム等について説明する。また、本実施の形態において、ウェブページの分野やウェブページのＵＲＬによって、ページ情報の取得方法が異なる情報処理システム等について説明する。さらに、本実施の形態において、ウェブページの主要な情報（後述する主情報）を取得し、主情報のみを用いて、ウェブページから電子ブックのコンテンツを取得する情報処理システム等について説明する。

【００２７】
図１は、本実施の形態における情報処理システム１の概念図である。情報処理システム１は、情報処理装置１１、１以上のサーバ装置１２を具備する。情報処理装置１１は、ウェブページから電子ブックのコンテンツを得る処理を行う装置である。また、サーバ装置１２は、ウェブページを格納しており、情報処理装置１１等の他の装置からの要求に応じて、ウェブページを送信する装置である。

【００２８】
図２は、本実施の形態における情報処理システム１のブロック図である。情報処理装置１１は、指示受付部１１０、ウェブページ格納部１１１、ウェブページ受信部１１２、ウェブページ蓄積部１１３、レイアウト情報格納部１１４、分野情報決定部１１５、ページ情報取得部１１６、ページ情報出力部１１７を具備する。

【００２９】
分野情報決定部１１５は、用語格納手段１１５１、出現度情報取得手段１１５２、分野情報決定手段１１５３を具備する。

【００３０】
ページ情報取得部１１６は、ページ取得ルール情報格納手段１１６１、主情報取得手段１１６２、ページ情報取得手段１１６３を具備する。

【００３１】
ページ情報出力部１１７は、レイアウト情報取得手段１１７１、レイアウト手段１１７２、ページ情報出力手段１１７３を具備する。

【００３２】
サーバ装置１２は、ウェブページ記憶部１２１、ウェブページ送信部１２２を具備する。

【００３３】
指示受付部１１０は、指示を受け付ける。この指示とは、例えば、ページ情報を出力する指示である。また、指示とは、他の指示でも良い。また、ページ情報出力指示は、電子ブックのコンテンツの元になるウェブページを特定する情報（ＵＲＬやＵＲＩなど）を含んでも良いし、元になるウェブページを検索するためのキーワードなどを含んでも良い。また、受付とは、ユーザからの入力の受け付け、他の装置からの受信などを含む概念である。指示の入力手段は、キーボードやマウスやメニュー画面によるもの等、何でも良い。指示受付部１１０は、キーボード等の入力手段のデバイスドライバーや、メニュー画面の制御ソフトウェア等で実現され得る。

【００３４】
ウェブページ格納部１１１は、１以上のウェブページを格納し得る。ウェブページは、通常、他の装置から取得されたウェブページである。ウェブページ格納部１１１は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。ウェブページ格納部１１１にウェブページが記憶される過程は問わない。例えば、記録媒体を介してウェ

ブページがウェブページ格納部１１１で記憶されるようになってもよく、通信回線等を介して送信されたウェブページがウェブページ格納部１１１で記憶されるようになってもよく、あるいは、入力デバイスを介して入力されたウェブページがウェブページ格納部１１１で記憶されるようになってもよい。

【００３５】

ウェブページ受信部１１２は、ウェブページをサーバ装置１２から受信する。ウェブページ受信部１１２は、通常、無線または有線の通信手段で実現されるが、放送を受信する手段で実現されても良い。

【００３６】

ウェブページ蓄積部１１３は、ウェブページ受信部１１２が受信したウェブページを、ウェブページ格納部１１１に蓄積する。ウェブページ蓄積部１１３は、通常、ＭＰＵやメモリ等から実現され得る。ウェブページ蓄積部１１３の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００３７】

レイアウト情報格納部１１４は、分野を示す情報である分野情報と、電子ブックのページ内のレイアウトに関する情報であるレイアウト情報との組を、２組以上格納している。また、レイアウト情報格納部１１４は、ウェブページを特定する情報であるウェブページ特定情報（ＵＲＬやＵＲＩなど）とレイアウト情報との組を、２組以上格納していても良い。また、分野情報とレイアウト情報との組、またはウェブページ特定情報とレイアウト情報との組を、以下、適宜、格納レイアウト情報という。分野とは、ニュース、ブログ、子供向け、技術解説などである。分野情報とは、分野に対応する情報である。分野情報は、分野がニュースの場合は「１」、ブログの場合は「２」、子供向けの場合は「３」、技術解説の場合は「４」などである。レイアウト情報とは、ページ情報をレイアウトするための、いわゆる定義データでも良いし、ページ情報をレイアウトするプログラムまたはプログラム名などでも良い。レイアウト情報格納部１１４は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。レイアウト情報格納部１１４にレイアウト情報が記憶される過程は問わない。例えば、記録媒体を介してレイアウト情報がレイアウト情報格納部１１４で記憶されるようになってもよく、通信回線等を介して送信されたレイアウト情報がレイアウト情報格納部１１４で記憶されるようになってもよく、あるいは、入力デバイスを介して入力されたレイアウト情報がレイアウト情報格納部１１４で記憶されるようになってもよい。

【００３８】

分野情報決定部１１５は、１以上のウェブページの内容から、当該１以上のウェブページの分野を決定し、分野情報を取得する。なお、ウェブページの内容とは、例えば、ｈｔｍｌやｘｍｌなどの記述言語で記述されたウェブページの情報のことである。分野情報決定部１１５は、例えば、ニュース記事に現れる用語を多数含むウェブページの分野を「ニュース」と決定する。また、分野情報決定部１１５は、例えば、ウェブページのタイトルに「ニュース」を含むウェブページの分野を「ニュース」と決定する。また、分野情報決定部１１５は、例えば、口語調の文が多い（口語調の文に出現する単語（例えば、「おれ」「～だし」「だね」など）を多数含む）ウェブページ、日付が先頭から所定の範囲に記載されているウェブページ、主語がない文が多いなどのウェブページの分野を「ブログ」と決定する。また、分野情報決定部１１５は、例えば、ひらがなの割合が多いウェブページの分野を「子供向け」と決定する。また、分野情報決定部１１５は、例えば、画像の数や割合が多いウェブページの分野を「子供向け」と決定する。分野情報決定部１１５は、通常、ＭＰＵやメモリ等から実現され得る。分野情報決定部１１５の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００３９】

用語格納手段１１５１は、分野情報と、当該分野情報が示す分野に対応する１以上の用

語を含む用語情報を１以上格納している。用語情報は、例えば、（ニュース，「事件」「総理」「政治」「経済」・・・）、（ブログ，「天気」「晴れ」「曇り」「雨」・・・）などである。なお、ここでの用語情報のデータ構造は（分野情報，１以上の用語）である。用語格納手段１１５１は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。用語格納手段１１５１に用語情報が記憶される過程は問わない。例えば、記録媒体を介して用語情報が用語格納手段１１５１で記憶されるようになってもよく、通信回線等を介して送信された用語情報が用語格納手段１１５１で記憶されるようになってもよく、あるいは、入力デバイスを介して入力された用語情報が用語格納手段１１５１で記憶されるようになってもよい。

【００４０】

出現度情報取得手段１１５２は、１以上のウェブページから、１以上の用語の出現の度合いに関する情報である出現度情報を取得する。出現度情報は、用語の出現数でも良いし、出現の割合（出現数／全用語数）でも良いし、タイトルの用語として使われているか否でも良いし、ｔｆ・ｉｄｆにより取得される情報である。その他、出現度情報は、用語の出現の度合いに関する情報であれば良い。出現度情報取得手段１１５２は、ある用語Ａを、１以上のウェブページに対して、パターンマッチングを行い、出現度情報を取得しても良いし、形態素解析を行い、形態素を抽出し、その品詞などを用いて、出現度情報を取得しても良い。出現度情報を取得するアルゴリズムは問わない。出現度情報取得手段１１５２は、通常、ＭＰＵやメモリ等から実現され得る。出現度情報取得手段１１５２の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００４１】

分野情報決定手段１１５３は、出現度情報を用いて、１以上のウェブページの分野を決定し、分野情報を取得する。分野情報決定手段１１５３は、分野情報が示す分野に対応する１以上の用語の出現度情報が予め決められた条件を満たす（例えば、閾値以上、など）場合は、当該分野情報を取得する。分野情報決定手段１１５３は、ウェブページのタイトルで使われている用語に対応する分野情報を取得しても良い。分野情報決定手段１１５３は、通常、ＭＰＵやメモリ等から実現され得る。分野情報決定手段１１５３の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００４２】

ページ情報取得部１１６は、１以上のウェブページから、ページ単位の情報であるページ情報を、２以上取得する。ページ情報とは、電子ブック用のページの情報である。例えば、１ページに入るデータ量が決まっており、ページ情報取得部１１６は、１以上のウェブページから、その１ページ分のデータごとに、２ページ以上、情報を区切って取得する。また、ページ情報の取得とは、１以上のウェブページから、取得した情報を用いて、ページ単位の情報を構成することである、とも言える。ページ情報取得部１１６の具体的な処理アルゴリズムの例は、後述する。ページ情報取得部１１６は、通常、ＭＰＵやメモリ等から実現され得る。ページ情報取得部１１６の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００４３】

ページ取得ルール情報格納手段１１６１は、分野情報と、１以上のウェブページから情報を取得するための情報であるページ取得ルール情報との組を、２以上格納している。ページ取得ルール情報格納手段１１６１は、ウェブページ特定情報とページ取得ルール情報との組を、２以上格納していても良い。ページ取得ルール情報とは、取得する情報に対応するタグ、取得する情報のデータタイプ（静止画、テキスト列など）、除く情報（タグ、データ）を指定する情報などにより構成されても良い。ページ取得ルール情報は、プログラムに埋め込まれていても、プログラム自体でも良い。ページ取得ルール情報は、正規表現で記載されていても、特定の構造、文言からなる自然言語で記載されていても良い。ペ

ージ取得ルール情報の記載方法は問わない。ページ取得ルール情報の具体例については、後述する。ページ取得ルール情報格納手段１１６１は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。ページ取得ルール情報格納手段１１６１にページ取得ルール情報が記憶される過程は問わない。例えば、記録媒体を介してページ取得ルール情報がページ取得ルール情報格納手段１１６１で記憶されるようになってもよく、通信回線等を介して送信されたページ取得ルール情報がページ取得ルール情報格納手段１１６１で記憶されるようになってもよく、あるいは、入力デバイスを介して入力されたページ取得ルール情報がページ取得ルール情報格納手段１１６１で記憶されるようになってもよい。

【００４４】

主情報取得手段１１６２は、１以上のウェブページの各々に対して、ウェブページの構造またはタグまたは内容を用いて、主となる情報（以下、適宜、主情報という。）を取得する。主情報取得手段１１６２は、例えば、右側にレイアウトされるビットマップをバナー広告であると判断し、主情報から除く処理を行う。また、主情報取得手段１１６２は、例えば、タグ「＜div class="topicsindex"＞」に対応するテキスト列を、主情報として取得する。その他、主情報の取得方法は、種々あり得る。なお、主情報の取得方法を、ページ取得ルール情報としても良い。かかる場合、主情報取得手段１１６２は、ページ情報取得手段１１６３と同様の処理となる。主情報取得手段１１６２は、通常、ＭＰＵやメモリ等から実現され得る。主情報取得手段１１６２の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００４５】

ページ情報取得手段１１６３は、分野情報決定部１１５が取得した分野情報と対になるページ取得ルール情報を用いて、１以上のウェブページから、ページ単位の情報であるページ情報を、２以上取得する。また、ページ情報取得手段１１６３は、ウェブページに対応するページ取得ルール情報を用いて、１以上のウェブページから、ページ単位の情報であるページ情報を、２以上取得しても良い。また、ページ情報取得手段１１６３は、主情報取得手段１１６２が取得した情報から、２以上のページ情報を取得しても良い。ページ情報取得手段１１６３は、通常、ＭＰＵやメモリ等から実現され得る。ページ情報取得手段１１６３の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００４６】

ページ情報出力部１１７は、ページ情報取得部１１６が取得した２以上のページ情報を出力する。ここで、出力とは、ディスプレイへの表示、プロジェクターを用いた投影、プリンタへの印字、音出力、外部の装置（例えば、電子ブック装置）への送信、記録媒体（例えば、電子ブック装置に装着して用いられる着脱可能な可搬型の記録媒体）への蓄積、他の処理装置や他のプログラム等への処理結果の引渡し等を含む概念である。ページ情報出力部１１７は、ディスプレイやスピーカー等の出力デバイスを含むと考えても含まないと考えても良い。ページ情報出力部１１７は、出力デバイスのドライバーソフトまたは、出力デバイスのドライバーソフトと出力デバイス等で実現され得る。

【００４７】

レイアウト情報取得手段１１７１は、分野情報決定部１１５が取得した分野情報と対になるレイアウト情報を、レイアウト情報格納部１１４から取得する。また、レイアウト情報取得手段１１７１は、ページ情報を構成する元となる処理対象のウェブページに対応するレイアウト情報を、レイアウト情報格納部１１４から取得しても良い。レイアウト情報取得手段１１７１は、通常、ＭＰＵやメモリ等から実現され得る。レイアウト情報取得手段１１７１の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００４８】

レイアウト手段１１７２は、レイアウト情報取得手段１１７１が取得したレイアウト情

報に従って、ページ情報取得部１１６が取得した２以上のページ情報をレイアウトする。レイアウト手段１１７２が、ページ情報をレイアウトした結果、電子ブックの２以上のページが構成される。レイアウト手段１１７２は、通常、ＭＰＵやメモリ等から実現され得る。レイアウト手段１１７２の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００４９】

ページ情報出力手段１１７３は、レイアウト手段１１７２がレイアウトした２以上のページ情報を出力する。ページ情報出力手段１１７３は、ディスプレイやスピーカー等の出力デバイスを含むと考えても含まないと考えても良い。ページ情報出力手段１１７３は、出力デバイスのドライバーソフトまたは、出力デバイスのドライバーソフトと出力デバイス等で実現され得る。

【００５０】

ウェブページ記憶部１２１は、１以上のウェブページを格納している。ウェブページ記憶部１２１は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。

【００５１】

ウェブページ送信部１２２は、ウェブページ記憶部１２１のウェブページを、情報処理装置１１に送信する。ウェブページ送信部１２２は、通常、情報処理装置１１からの要求を受け付け、当該要求に対応して、ウェブページを、情報処理装置１１に送信する。ウェブページ送信部１２２は、通常、無線または有線の通信手段で実現されるが、放送手段で実現されても良い。

【００５２】

次に、情報処理システム１の動作について説明する。まず、情報処理装置１１の動作例について、図３から図５のフローチャートを用いて説明する。

【００５３】

（ステップＳ３０１）ウェブページ受信部１１２は、ウェブページを受信したか否かを判断する。ウェブページを受信すればステップＳ３０２に行き、ウェブページを受信しなければステップＳ３０３に行く。

【００５４】

（ステップＳ３０２）ウェブページ蓄積部１１３は、ステップＳ３０１で受信されたウェブページを、ウェブページ格納部１１１に蓄積する。ステップＳ３０１に戻る。

【００５５】

（ステップＳ３０３）指示受付部１１０は、ページ情報出力指示を受け付けたか否かを判断する。ページ情報出力指示を受け付ければステップＳ３０４に行き、ページ情報出力指示を受け付けなければステップＳ３０１に戻る。なお、ページ情報出力指示には、１以上のウェブページ特定情報を含む。１以上のウェブページ特定情報は、例えば、トップページのＵＲＬであり、例えば、トップページからリンクされているウェブページも含めたウェブページ群から、ページ情報が構成され得る。

【００５６】

（ステップＳ３０４）分野情報決定部１１５は、ウェブページ格納部１１１から、ページ情報出力指示により特定される１以上のウェブページを読み出す。

【００５７】

（ステップＳ３０５）分野情報決定部１１５は、ステップＳ３０４で読み出された１以上のウェブページから、分野情報を取得する。分野情報を取得する処理（以下、適宜「分野決定処理」という）の詳細例について、図４のフローチャートを用いて説明する。

【００５８】

（ステップＳ３０６）ページ情報取得部１１６は、ステップＳ３０４で読み出された１以上のウェブページから、ページ情報を取得する処理を行う。ページ情報取得処理の詳細例について、図５のフローチャートを用いて説明する。

【００５９】
（ステップＳ３０７）レイアウト情報取得手段１１７１は、ステップＳ３０５で取得された分野情報に対応するレイアウト情報を、レイアウト情報格納部１１４から読み出す。
【００６０】
（ステップＳ３０８）レイアウト手段１１７２は、ステップＳ３０６で取得されたページ情報（ここでは、レイアウトされる前の情報）を、ステップＳ３０７で読み出されたレイアウト情報に従ってレイアウトし、ページ情報をメモリ上に得る。ここで、レイアウト手段１１７２は、例えば、ウェブページ内のタグを解釈し、ページ単位に区切る（例えば、改ページコードを挿入する）処理を行う。ウェブページ内のタグを解釈する処理は、いわゆるウェブブラウザの処理である。
【００６１】
（ステップＳ３０９）ページ情報出力手段１１７３は、ステップＳ３０８でレイアウトし、構成された２以上のページ情報を出力する。ステップＳ３０１に戻る。
【００６２】
なお、図３のフローチャートにおいて、分野情報を自動取得したが、分野情報は、ユーザが入力し、指示受付部１１０が受け付けても良い。
【００６３】
また、図３のフローチャートにおいて、分野情報に応じたレイアウト情報を用いて、ページ情報を構成したが、ウェブページに対応するレイアウト情報を用いて、ページ情報を構成しても良い。
【００６４】
さらに、図３のフローチャートにおいて、電源オフや処理終了の割り込みにより処理は終了する。
【００６５】
次に、ステップＳ３０５の分野決定処理の詳細例について、図４のフローチャートを用いて説明する。
【００６６】
（ステップＳ４０１）出現度情報取得手段１１５２は、カウンタｉに１を代入する。
【００６７】
（ステップＳ４０２）出現度情報取得手段１１５２は、ｉ番目の分野情報が、用語格納手段１１５１に格納されているか否かを判断する。ｉ番目の分野情報が格納されていればステップＳ４０３に行き、格納されていなければステップＳ４１１に行く。
【００６８】
（ステップＳ４０３）出現度情報取得手段１１５２は、ｉ番目の分野情報に対応する１以上の用語を、用語格納手段１１５１から読み出す。
【００６９】
（ステップＳ４０４）出現度情報取得手段１１５２は、カウンタｊに１を代入する。
【００７０】
（ステップＳ４０５）出現度情報取得手段１１５２は、ステップＳ４０３で取得した用語の中に、ｊ番目の用語が存在するか否かを判断する。ｊ番目の用語が存在すればステップＳ４０６に行き、ｊ番目の用語が存在しなければステップＳ４０９に行く。
【００７１】
（ステップＳ４０６）出現度情報取得手段１１５２は、ステップＳ４０５で取得したｊ番目の用語に対する出現度情報を取得する。出現度情報取得手段１１５２は、例えば、ｊ番目の用語の、読み出されている１以上のウェブページにおける出現頻度を取得する。また、出現度情報取得手段１１５２は、例えば、ｊ番目の用語の、読み出されている１以上のウェブページにおける出現割合（出現頻度／全ワード数）を取得する。また、出現度情報取得手段１１５２は、例えば、ｊ番目の用語についてのｔｆ・ｉｄｆを取得する。ｔｆ・ｉｄｆは、ｔｆ（用語の出現頻度）とｉｄｆ（逆出現頻度）の二つの指標で計算される。ｔｆ・ｉｄｆは、公知技術であるので詳細な説明を省略する。

【００７２】
（ステップＳ４０７）出現度情報取得手段１１５２は、ステップＳ４０６で取得された出現度情報を、バッファに一時格納する。
【００７３】
（ステップＳ４０８）出現度情報取得手段１１５２は、カウンタｊを１、インクリメントする。ステップＳ４０５に戻る。
【００７４】
（ステップＳ４０９）出現度情報取得手段１１５２は、ステップＳ４０７でバッファに一時格納された１以上の出現度情報を用いて、ｉ番目の分野情報についてのスコアを算出し、当該算出結果を、ｉ番目の分野情報と対応付けて、メモリ上に一時格納する。出現度情報取得手段１１５２は、例えば、ステップＳ４０７でバッファに一時格納された１以上の出現度情報の和、または、平均値を算出し、ｉ番目の分野情報についてのスコアとする。
【００７５】
（ステップＳ４１０）出現度情報取得手段１１５２は、カウンタｉを１、インクリメントする。ステップＳ４０２に戻る。
【００７６】
（ステップＳ４１１）分野情報決定手段１１５３は、ステップＳ４０９で算出した、ｉ番目の分野情報についてのスコアのうちで、最も大きなスコアに対応する分野情報を取得し、メモリ上に配置する。上位処理にリターンする。
【００７７】
なお、図４のフローチャートにおいて、分野情報ごとにスコアを算出するアルゴリズムが異なっても良い。例えば、分野が「ニュース」（分野情報が、例えば「１」）の場合、ウェブページのタイトルで使われている用語の中に「ニュース」が存在すれば、スコアを非常に大きな数にしても良い。かかる場合、ウェブページのタイトルで使われている用語の中に「ニュース」が存在すれば、当該ウェブページの分野は、「ニュース」である、と判断されることを示す。また、例えば、分野が「ブログ」（分野情報が、例えば「２」）の場合、「天気」「晴れ」「曇り」「雨」などの用語の出現頻度を取得し、出現頻度の合計をスコアとしても良い。かかる場合、用語格納手段１１５１に、例えば、分野情報と、スコアを算出するアルゴリズムの情報（例えば、スコア算出処理を行うプログラム名や関数名など）が対で格納されている。
【００７８】
次に、ステップＳ３０６のページ情報取得処理の詳細例について、図５のフローチャートを用いて説明する。
【００７９】
（ステップＳ５０１）主情報取得手段１１６２は、ステップＳ３０４で読み込まれた１以上のウェブページの各々に対して、ウェブページの構造またはタグまたは内容を用いて、主情報を取得する。主情報を取得するアルゴリズムは種々あり、公知技術である。また、その例は、上述した。
【００８０】
（ステップＳ５０２）ページ情報取得手段１１６３は、ステップＳ３０５で取得された分野情報と対になるページ取得ルール情報を、ページ取得ルール情報格納手段１１６１から読み出す。
【００８１】
（ステップＳ５０３）ページ情報取得手段１１６３は、ステップＳ５０２で読み出したページ取得ルール情報を用いて、１以上のウェブページ（正確には、ここでは、ステップＳ５０１で取得された主情報）から、ページ単位の情報であるページ情報を、２以上取得する。
【００８２】
なお、図５のフローチャートのステップＳ５０３における処理は、ページ取得ルール情

報に従った種々の処理が考えられえる。その処理例については後述する。

【００８３】

また、図５のフローチャートにおいて、ステップＳ５０１の主情報取得処理は、必須ではない。

【００８４】

また、図５のフローチャートのステップＳ５０２において、ウェブページ識別情報と対になるページ取得ルール情報を、ページ取得ルール情報格納手段１１６１から読み出しても良い。

【００８５】

さらに、図５のフローチャートにおいて、上記の情報処理装置１１の処理において、ページ情報を構成する元の情報をすべて取得してから、レイアウトした。しかし、一つのまとまりのある情報（サブ情報）を取得した後、レイアウトし、その後さらに、次のサブ情報を取得しても良い。つまり、サブ情報ごとに、取得とレイアウトを繰り返すアルゴリズムでも良い。かかる場合、レイアウトに応じた、また、空きスペースに応じたサブ情報の取得が可能となる。

【００８６】

次に、サーバ装置１２の動作について説明する。サーバ装置１２のウェブページ送信部１２２は、情報処理装置１１からの要求に対応して、ウェブページ記憶部１２１から１以上のウェブページを読み出し、情報処理装置１１に送信する。

【００８７】

以下、本実施の形態における情報処理システム１の具体的な動作について説明する。情報処理システム１の概念図は図１である。

【００８８】

今、情報処理装置１１のレイアウト情報格納部１１４には、図６に示すレイアウト情報管理表（レイアウト情報管理表Ａ）と、図７に示すレイアウト情報管理表（レイアウト情報管理表Ｂ）が格納されている。

【００８９】

レイアウト情報管理表Ａは、分野情報がニュースのウェブページに対応するレイアウト情報である。レイアウト情報管理表Ａは、「ＩＤ」「主情報の種類」「ページ数」「領域」「文字サイズ」の属性値を有するレコードを１以上格納している。「ＩＤ」は、レコードを識別する属性であり、表管理のために存在する。「主情報の種類」は、ページ情報取得部１１６が取得した主情報の種類を示す情報が格納される。「ページ数」は、レイアウトされた場合のページ番号に関する情報が格納される。「領域」は、ページ内の領域を示す情報が格納される。「領域」内の情報のタイプには、ここでは、２種類ある。一つは、座標情報の集合である。「領域」の値が座標情報の集合である場合、領域は、それらを結ぶ矩形であることを示す。他は、始点と幅、高さの情報で定義される。かかる場合、領域は、始点と幅、高さの情報により示される長方形である。また、対応する情報がテキストの場合、「文字サイズ」の属性値により、文字サイズが示される。なお、レイアウト情報格納部１１４には、別途、分野情報がニュースの場合に、例えば、段組数「３」、文字列は縦書き、など、他の情報も格納されている、とする。また、図６において、「ＩＤ＝６、７」の領域の情報と、「ＩＤ＝８，９」の領域の情報は、交互に適用され、２つの記事で、２ページ目以降の各ページを構成することとなる。さらに、図６において、「ＩＤ＝６」の領域の情報「始点（０，０）」に下線が付与されている（フラグが付与されている、ことと同意義）のは、「始点（０，０）」は、前の記事がはみ出してきた場合に、下方にずれることを意味する。「ＩＤ＝７」の領域の情報「（０，０）（０，２９７）」も同様である。かかることにより、後述する図８（ｂ）の８１の領域が出来る場合がある。図８の８１の領域が出来る場合とは、図８（ａ）の記事（１ページ目の記事）が２ページ目にはみ出してきた場合である。

【００９０】

レイアウト情報管理表Ｂは、分野がブログのウェブページに対応するレイアウト情報で

ある。レイアウト情報管理表Ｂは、「ＩＤ」「主情報の種類」「配置情報」「文字サイズ」「フォント」「文字色」の属性値を有するレコードが１以上格納される。「配置情報」は、ここでは、各主情報の種類ごとの情報の配置順序を示している。また、レイアウト情報格納部１１４には、別途、分野情報がブログの場合に、例えば、段組数「１」、文字列は縦書き、など、他の情報も格納されている、とする。

【００９１】

なお、レイアウト情報管理表Ａを用いて、ニュースのウェブページをレイアウトした場合、図８に示すようなレイアウトとなる。図８において、数値は、レイアウト情報管理表Ａのレコードの「ＩＤ」に対応する。また、図８（ａ）は、１ページ目のページ情報のレイアウトの概要図、図８（ｂ）は、２ページ目以降のページ情報のレイアウトの概要図である。

【００９２】

また、レイアウト情報管理表Ｂを用いて、ブログのウェブページをレイアウトした場合、図９に示すようなレイアウトとなる。図９において、数値は、レイアウト情報管理表Ｂのレコードの「ＩＤ」に対応する。また、図９（ａ）は、１ページ目のページ情報のレイアウトの概要図、図９（ｂ）は、２ページ目のページ情報のレイアウトの概要図である。なお、図９において、「１」「２」「３」に対応する情報が順に横に、縦書きで配置されることを示しており、そのデータ量に応じて、横（左）にずれていく、ことは言うまでもない。

【００９３】

図１０、図１１は、ページ取得ルール情報格納手段１１６１に格納されているページ取得ルール情報管理表である。ページ取得ルール情報管理表は、「分野情報」「ＩＤ」「ページ取得ルール」の属性値を有するレコードを１以上格納している。「ページ取得ルール」は、「ルール」と「主情報の種類」の属性を有する。「ルール」は、ここでは、自然言語で記載しているが、プログラミング言語で記載しても良いし、プログラムの実行モジュール名などでも良い。「ルール」がプログラムの実行モジュール名である場合、実行モジュールは、ページ取得ルール情報格納手段１１６１に格納されている、とする。

【００９４】

図１０、図１１におけるルールは、例えば、ＨＴＭＬやＸＭＬなどのウェブページに記載されている情報の中の、特定のタグや、属性や、タグまたは属性に対応する値が条件となっている。また、例えば、ＨＴＭＬやＸＭＬなどのウェブページに記載されている情報の中の、アンカーを辿っていき、さらに下位のウェブページを取得して、当該下位のウェブページから、特定のタグや、属性や、タグまたは属性に対応する値を条件として情報を取得するルールである。さらに、図１０、図１１におけるルールは、例えば、情報が配置される領域の大きさと、配置される候補となるデータのサイズ（データ量）とを比較して、合致する（収まりが良い場合で、完全一致するとは限らない）候補のデータを取得するルールである。さらに、図１０、図１１において、主情報の種類ごとに取得ルールを設けているが、主情報に対して一つのルールでも良い。

【００９５】

また、用語格納手段１１５１は、図１２に示す用語情報管理表を格納している。用語情報管理表は、「分野情報」「用語」「スコア」の属性値を有するレコードを１以上格納している。「用語」は、単なる単語と、出現領域（例えば、"ｔｉｔｌｅ"）を特定した単語と、用語のタイプ（例えば、「＄日付タイプ」）などがある。「＄日付タイプ」とは、「２００８／１／２９」「２００８年１月２９日」「１月２９日（火）」「１／２９（火）」「平成２０年１月２９日」などの日付の表記となる文字列を言う。図１２において、「"ニュース"　ｉｎ　ｔｉｔｌｅ」のレコードは、＜title＞タグ内に「ニュース」という文字列を含む場合に、ヒットするレコードである。「スコア」とは、ヒットした場合（用語を含むなど）に、カウントされる数値である。

【００９６】

かかる状況において、ニュースのウェブページから電子ブックのコンテンツを自動生成

する具体例１と、ブログのウェブページから電子ブックのコンテンツを自動生成する具体例２の、２つの具体例について説明する。

（具体例１）

【００９７】

ニュースのウェブページから電子ブックのコンテンツを自動生成する具体例について説明する。まず、情報処理装置１１の指示受付部１１０は、あるウェブページの出力指示をユーザから受け付けた、とする。すると、情報処理装置１１は、サーバ装置１２に対して、そのウェブページの送信を要求する。そして、サーバ装置１２のウェブページ送信部１２２は、受け付けた要求に対応するウェブページを、ウェブページ記憶部１２１から読み出し、情報処理装置１１に送信する。

【００９８】

情報処理装置１１のウェブページ受信部１１２は、要求したサイトのウェブページを受信する。そして、ウェブページ蓄積部１１３は、受信されたウェブページを、ウェブページ格納部１１１に蓄積する。なお、ウェブページ格納部１１１は、ここでは、主メモリでも良い。そして、情報処理装置１１の図示しない処理部は、受信されたウェブページを解釈し、例えば、図１３に示すようなサイトを画面上に表示する。図１３において、１３１は、電子ブックのコンテンツを自動生成するツールバーである。

【００９９】

次に、ユーザは、ツールバー１３１をマウスで押下した、とする。すると、指示受付部１１０は、ページ情報出力指示を受け付ける。つまり、ツールバー１３１の押下は、ページ情報出力指示である。

【０１００】

次に、分野情報決定部１１５は、現在表示中のウェブページのスクリプトを取得する。かかる、スクリプト（ここでは、ＨＴＭＬで記述された文字列）を、図１４に示す。

【０１０１】

分野情報決定部１１５は、図１４の情報を用いて、以下のように、分野情報を取得する。つまり、出現度情報取得手段１１５２は、図１２の用語情報管理表の分野情報「ニュース」を取得する。そして、出現度情報取得手段１１５２は、「ニュース」に対応する１番目の用語「トピックス」を取得し、図１４の情報から、用語「トピックス」を検索し、用語「トピックス」の出現回数「１」を取得する。そして、出現度情報取得手段１１５２は、出現回数「１」と、用語情報管理表のスコア「１」を乗算し、「１」を、用語「トピックス」のスコアとして得る。同様に、出現度情報取得手段１１５２は、用語「経済」のスコアとして「１」を得る。また、同様に、用語「社会」のスコアとして「１」を得る。また、同様に、用語「スポーツ」のスコアとして「１」を得る。さらに、出現度情報取得手段１１５２は、用語「"ニュース" ｉｎ ｔｉｔｌｅ」から、＜title＞タグに対応する情報「○×サイトニュース」を取得する。そして、出現度情報取得手段１１５２は、情報「○×サイトニュース」に用語「ニュース」が含まれていると判断し、「"ニュース" ｉｎ ｔｉｔｌｅ」に合致するとする。そして、出現度情報取得手段１１５２は、「"ニュース" ｉｎ ｔｉｔｌｅ」と対になるスコア「１０」を得る。以上の処理を繰り返し、出現度情報取得手段１１５２は、「用語」のすべての属性値に対するスコアを算出し、それらの和を取得する、とする。ここで、出現度情報取得手段１１５２は、分野情報「ニュース」に対するスコア「１８」を得た、とする。

【０１０２】

次に、図１４の情報について、分野情報「ブログ」に対するスコアを算出する。つまり、出現度情報取得手段１１５２は、用語「＄日付タイプ」を用いて、図１４の情報の中に日付タイプの情報が存在するか否かを判断する。ここで、図１４の情報の中に日付タイプの情報が存在するとして、出現度情報取得手段１１５２は、用語「＄日付タイプ」に対するスコア「２」を得る。次に、図１４の情報の中に用語「俺」が存在するか否かを判断する。ここで、図１４の情報の中に用語「俺」が存在しない、とする。そして、次に、図１４の情報の中に用語「おれ」が存在するか否かを判断する。ここで、図１４の情報の中に

用語「おれ」も存在しない、とする。用語が存在しない場合、加算されるスコアは「０」である。以上の処理を繰り返し、出現度情報取得手段１１５２は、分野情報「ブログ」に対するスコア「２」を得た、とする。

【０１０３】

そして、次に、分野情報決定手段１１５３は、上記で算出した、１番目、２番目の分野情報についてのスコアのうちで、最も大きなスコア「１８」に対応する分野情報「ニュース」を取得し、メモリ上に配置する。以上の処理により、図１４のウェブページの分野情報がニュースである、と決定された。

【０１０４】

次に、ページ情報取得部１１６は、図１４のウェブページから、以下のように、ページ情報を取得する。つまり、ページ情報取得手段１１６３は取得された分野情報「ニュース」と対になるすべてのページ取得ルール情報を、図１０、図１１のページ取得ルール情報管理表から読み出す。そして、ページ情報取得手段１１６３は、以下のように、ページ取得ルール情報に従って、１以上のページ情報を取得する。

【０１０５】

まず、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝１」（１）のルール「「ｃｌａｓｓ＝ｔａｂ？　ｏｎ」に対応するｉｄの値を取得」に従って、タグ「＜li class="tab0 on"＞」を検索し、当該タグに対応するｉｄの値"topics"を、タグ「＜a id="topics" href="r/ttp" hidefocus="true"＞」から取得する。

【０１０６】

次に、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝１」（２）のルールに従い、ｉｄの値"topics"を含む＜div＞内の4以上の項目「＜li＞＜a href="f/topics/top/１/?１20１573096"＞XX県知事にABC氏が当選へ＜/a＞＜img src="http://photo2.gif" alt="[photo]"＞＜/li＞」「＜li＞＜a href="f/topics/top/2/?１20１5771１3"＞サッカー　日本対ブラジル＜/a＞＜/li＞」「＜li＞＜a href="f/topics/top/3/?１20１574581"＞女優YYさん　結婚＜/a＞＜img src="http://photo2.gif" alt="[photo]"＞＜/li＞」「＜li＞＜a href="f/topics/top/7/?１20１5755１5"＞NY株　反発＜/a＞＜/li＞」・・・を、読み出す。

【０１０７】

次に、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝１」（３）のルールに従い、「＜img src="?*" alt="[photo]"＞」に合致する項目の文字列「XX県知事にABC氏が当選へ」を取得する。そして、ページ情報取得手段１１６３は、主情報の種類「メイン記事の見出し」とその情報「XX県知事にABC氏が当選へ」を対にして、バッファに一時格納する。

【０１０８】

次に、レイアウト情報取得手段１１７１は、取得された分野情報「ニュース」に対応するレイアウト情報（図６）を、レイアウト情報格納部１１４から読み出す。そして、その中で、主情報の種類「メイン記事の見出し」と対になる、ページ数「１」、領域「（０，０）（０，２５０）（４０，０）（４０，２５０）」、文字サイズ「２４」を取得する。

【０１０９】

そして、レイアウト手段１１７２は、取得されたページ情報「XX県知事にABC氏が当選へ」を、取得されたレイアウト情報に従ってレイアウトし、ページ情報をメモリ上に得る。ここで、メイン記事の見出しは、図１５の１５１のようにレイアウトされる、こととなる。

【０１１０】

次に、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝２」（１）のルール「メイン記事の見出しに対応するアンカーを取得」に対応するアンカーの値「f/topics/top/１/?１20１573096」を取得する。

【０１１１】

次に、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝２」

（２）のルール「アンカーの先のウェブページを取得」に従って、当該アンカー「f/topics/top/１/?１20１573096」が示すウェブページを、サーバ装置１２から取得する。

【０１１２】

次に、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝２」（３）のルール「［記事全文］に対応するウェブページを取得」に従って、［記事全文］に対応するアンカーが示すウェブページを、サーバ装置１２から取得する。

【０１１３】

次に、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝２」（４）のルール「ウェブページ中の＜div＞に対応する文字列を取得」に従って、取得したウェブページから、記事の本文となる文字列「任期満了に伴うXX県知事選は２７日投票、即日開票の結果、無所属ＡＢＣ氏が・・・・・・・・・・・・・ＸＹＺ氏は、あと少しのところで届かなかった。」を取得し、主情報の種類「メイン記事の内容」と対応付けて、バッファに格納する。

【０１１４】

次に、レイアウト情報取得手段１１７１は、取得された分野情報「ニュース」であり、主情報の種類「メイン記事の内容」と対になる、ページ数「１」、領域「（４０，０）（４０，２５０）（０，２５０）（０，２９７）（１１０，２９７）（１１０，０）」、文字サイズ「１２」を取得する。

【０１１５】

そして、レイアウト手段１１７２は、取得されたページ情報「任期満了に伴うXX県知事選は２７日投票、即日開票の結果、無所属ＡＢＣ氏が・・・・・・・・・・・・・ＸＹＺ氏は、あと少しのところで届かなかった。」を、取得されたレイアウト情報に従ってレイアウトし、ページ情報をメモリ上に得る。ここで、メイン記事の見出しは、図１５の１５２のようにレイアウトされる、こととなる。

【０１１６】

次に、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝３」（１）のルール「＜img src="　　"＞のgifデータを取得」に従って、「＜img src=http://photo.gif」を用いて、「photo.gif」を、サーバ装置１２から取得する。そして、ページ情報取得手段１１６３は、画像「photo.gif」を、主情報の種類「メイン画像」と対応付けて、バッファに格納する。

【０１１７】

次に、レイアウト情報取得手段１１７１は、取得された分野情報「ニュース」であり、主情報の種類「メイン画像」と対になる、ページ数「１」、領域「（１１０，２３０）（１１０，２９７）（２１０，２３０）（２１０，２９７）」を取得する。

【０１１８】

そして、レイアウト手段１１７２は、取得された画像「photo.gif」を、１ページ目の領域「（１１０，２３０）（１１０，２９７）（２１０，２３０）（２１０，２９７）」に配置する。ここで、メイン画像は、図１５の１５３のようにレイアウトされる、こととなる。

【０１１９】

次に、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝４」のルールに従って、ページ情報の１ページ目の残る領域に最も合致する（近い）項目を決定する。つまり、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝４」の（１）から（５）により、既に出力している項目「＜li＞＜a href="f/topics/top/１/?１20１573096"＞XX県知事にABC氏が当選へ＜/a＞＜img src="http://photo2.gif" alt="[photo]"＞＜/li＞」を除き、３つ以上の項目「＜li＞＜a href="f/topics/top/2/?１20１577１１3"＞サッカー　日本対ブラジル＜/a＞＜/li＞」「＜li＞＜a href="f/topics/top/3/?１20１57458１"＞女優YYさん　結婚＜/a＞＜img src="http://photo2.gif" alt="[photo]"＞＜/li＞」「＜li＞＜a href="f/topics/top/7/?１20１5755１5"＞NY株反発＜/a＞＜/li＞」・・・、から、それぞれの項目のアンカーの先のウェブページの

記事の情報（＜div＞に対応する文字列）を、それぞれ取得する。

【０１２０】

次に、分野情報「ニュース」に対応する「ＩＤ＝４」の（６）に従い、ページ情報取得手段１１６３は、各項目に対応する情報のデータ量（例えば、文字数）を取得する。そして、ページ情報取得手段１１６３は、１ページ目の残る領域のサイズを取得する。残る領域のサイズは、ページのサイズ（例えば、Ａ４、など）と、既に配置されている情報のデータ量から、取得可能である。次に、ページ情報取得手段１１６３は、「ＩＤ＝４」（６）のルールに従い、１ページ目の残りの領域サイズに最も近い項目の文字列（ここでは、例えば、「女優YYさん　結婚」）を取得する。次に、ページ情報取得手段１１６３は、文字列「女優YYさん　結婚」と、主情報の種類「サブ記事１の見出し」とを対にして、バッファに格納する。

【０１２１】

次に、レイアウト情報取得手段１１７１は、取得された分野情報「ニュース」であり、主情報の種類「サブ記事１の見出し」と対になる、ページ数「１」、領域「（１１０，０）（１１０，２３０）（１４０，２３０）（１４０，０）」、文字サイズ「２０」を取得する。

【０１２２】

そして、レイアウト手段１１７２は、取得された文字列「女優YYさん　結婚」を、１ページ目の領域「（１１０，０）（１１０，２３０）（１４０，２３０）（１４０，０）」に、文字サイズ「２０」として、配置する。ここで、メイン画像は、図１５の１５４のようにレイアウトされる、こととなる。

【０１２３】

次に、分野情報「ニュース」に対応する「ＩＤ＝５」（１）に従い、ページ情報取得手段１１６３は、サブ記事１のタグ＜div＞に対応する文字列「女優YYさんは、・・・・・・・・・・・・・・」を、取得する。なお、この文字列「女優YYさんは、・・・・・・・・・・・・・・」は、項目「＜li＞＜a href="f/topics/top/3/?120157458１"＞女優YYさん　結婚＜/a＞＜img src="http://photo2.gif" alt="[photo]"＞＜/li＞」のアンカー"f/topics/top/3/?120157458１"を用いて、サーバ装置１２から取得されたウェブページの＜div＞に対応する文字列である。

【０１２４】

以上の処理により、電子ブックの１ページ目のページ情報が構成できた。次に、情報処理装置１１は、２ページ目以降のページ情報の構成を行う。

【０１２５】

つまり、分野情報「ニュース」に対応する「ＩＤ＝６」（１）に従い、ページ情報取得手段１１６３は、他の項目「＜li＞＜a href="f/topics/top/2/?120157７１１３"＞サッカー　日本対ブラジル＜/a＞＜/li＞」に対応する文字列「サッカー　日本対ブラジル」を取得する。そして、ページ情報取得手段１１６３は、文字列「サッカー　日本対ブラジル」と、主情報の種類「一般記事の見出し」を対にして、バッファに格納する。

【０１２６】

次に、レイアウト情報取得手段１１７１は、取得された分野情報「ニュース」であり、主情報の種類「一般記事の見出し１」と対になる、ページ数「２～」、領域「始点（０，０）　（ｗ，ｈ）＝（２０，１５０）」、文字サイズ「１８」を取得する。

【０１２７】

そして、レイアウト手段１１７２は、取得された文字列「サッカー　日本対ブラジル」を、２ページ目の「始点（０，０）　（ｗ，ｈ）＝（２０，１５０）」で特定される領域に、文字サイズ「１８」として、配置する。ここで、文字列「サッカー　日本対ブラジル」は、図１６の１６１のようにレイアウトされる、こととなる。

【０１２８】

次に、分野情報「ニュース」に対応する「ＩＤ＝７」（１）に従い、ページ情報取得手段１１６３は、一般記事の見出し「サッカー　日本対ブラジル」に対応するアンカー「"f

/topics/top/2/?１20１577１１3"」を取得する。そして、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝７」（２）に従い、当該アンカー「"f/topics/top/2/?１20１577１１3"」の先のウェブページをサーバ装置１２から取得する。

【０１２９】

次に、ページ情報取得手段１１６３は、分野情報「ニュース」に対応する「ＩＤ＝７」（３）に従い、当該ウェブページの＜div＞に対応する文字列「サッカー　日本対ブラジルは、２９日１９時から・・・・・」を取得する。そして、ページ情報取得手段１１６３は、文字列「サッカー　日本対ブラジルは、２９日１９時から・・・・・」と、主情報の種類「一般記事の内容」を対にして、バッファに格納する。

【０１３０】

次に、レイアウト情報取得手段１１７１は、取得された分野情報「ニュース」であり、主情報の種類「一般記事の内容」、特に、「一般記事の内容１」と対になる、ページ数「２～」、領域「（０，０）（０，２９７）（２１０，２９７）（２１０，１９８）（９０，１９８）（９０，０）」、文字サイズ「１２」を取得する。

【０１３１】

そして、レイアウト手段１１７２は、取得された文字列「サッカー　日本対ブラジルは、２９日１９時から・・・・・」を、２ページ目の領域「（０，０）（０，２９７）（２１０，２９７）（２１０，１９８）（９０，１９８）（９０，０）」に、文字サイズ「１２」として、配置する。ここで、文字列「サッカー　日本対ブラジルは、２９日１９時から・・・・・」は、図１６の１６２のようにレイアウトされる、こととなる。

【０１３２】

同様に、ページ情報取得手段１１６３、レイアウト情報取得手段１１７１、レイアウト手段１１７２の処理により、タイトルの文字列「ＮＹ株　反発」、および文字列「週明けの米株式市場は、・・・・・」が、図１６の１６３、１６４のように配置される。なお、文字列「ＮＹ株　反発」、および文字列「週明けの米株式市場は、・・・・・」の配置のために、図６の「ＩＤ＝８」「ＩＤ＝９」のレイアウト情報が用いられる。

【０１３３】

次に、ページ情報出力手段１１７３は、レイアウトし、構成された２以上のページ情報（図１５、図１６など）を出力する。ここでの出力は、例えば、図示しない電子ブック装置への送信である。また、出力は、例えば、情報処理装置１１に装着されている着脱可能な記憶媒体である。

（具体例２）

【０１３４】

ブログから電子ブックのコンテンツを自動生成する具体例について説明する。まず、情報処理装置１１の指示受付部１１０は、あるウェブページの出力指示をユーザから受け付けた、とする。すると、情報処理装置１１は、サーバ装置１２に対して、そのウェブページの送信を要求する。そして、サーバ装置１２のウェブページ送信部１２２は、受け付けた要求に対応するウェブページを、ウェブページ記憶部１２１から読み出し、情報処理装置１１に送信する。

【０１３５】

情報処理装置１１のウェブページ受信部１１２は、要求したサイトのウェブページを受信する。そして、ウェブページ蓄積部１１３は、受信されたウェブページを、ウェブページ格納部１１１に蓄積する。なお、ウェブページ格納部１１１は、ここでは、主メモリでも良い。そして、情報処理装置１１の図示しない処理部は、受信されたウェブページを解釈し、例えば、図１７に示すようなサイトを画面上に表示する。

【０１３６】

次に、ユーザは、ツールバー１７１をマウスで押下した、とする。すると、指示受付部１１０は、ページ情報出力指示を受け付ける。つまり、ツールバー１７１の押下は、ページ情報出力指示である。また、ユーザは、ツールバー１７１の押下により、出力されるダイアログ（図示しない）に対して、分野情報「ブログ」を入力した、とする。なお、具体

例１においては、分野情報を自動決定したが、本例のように、分野情報が入力されても良い。そして、情報処理装置１１の指示受付部１１０は、分野情報「ブログ」を受け付け、メモリ上に配置する。

【０１３７】

次に、ページ情報取得部１１６は、現在表示中のウェブページのスクリプトを取得する。かかる、スクリプト（ここでは、ＨＴＭＬで記述された文字列）を、図１８に示す。

【０１３８】

次に、ページ情報取得部１１６は、図１８のウェブページから、以下のように、ページ情報を取得する。つまり、ページ情報取得手段１１６３は、入力された分野情報「ブログ」と対になるすべてのページ取得ルール情報を、図１１のページ取得ルール情報管理表から読み出す。そして、ページ情報取得手段１１６３は、以下のように、ページ取得ルール情報に従って、１以上のページ情報を取得する。

【０１３９】

まず、ページ情報取得手段１１６３は、分野情報「ブログ」に対応する「ＩＤ＝８」（１）のルール「＜h2＞タグに対応する文字列を取得」に従って、タグ「＜h2＞」を検索し、当該タグに対応する値「2008年１月27日(日)」を取得する。そして、ページ情報取得手段１１６３は、文字列「2008年１月27日(日)」と、主情報の種類「日付」とを対応付けて、バッファに格納する。

【０１４０】

次に、ページ情報取得手段１１６３は、分野情報「ブログ」に対応する「ＩＤ＝９」（１）のルール「＜h3＞タグに対応する文字列を取得」に従って、文字列「スキー場にて」を取得する。そして、ページ情報取得手段１１６３は、文字列「スキー場にて」と、主情報の種類「見出し」とを対応付けて、バッファに格納する。

【０１４１】

また、ページ情報取得手段１１６３は、分野情報「ブログ」に対応する「ＩＤ＝１０」（１）のルール「＜div class="entry-body-text"＞タグに対応する文字列を取得」に従って、文字列「ＸＸスキー場の雪は最高。・・・・・・・・・・・・・・・・・・・・・・・」を取得する。そして、文字列「ＸＸスキー場の雪は最高。・・・・・・・・・・・・・・・・・・・・・・・」と、主情報の種類「本文」とを対応付けて、バッファに格納する。

【０１４２】

さらに、ページ情報取得手段１１６３は、分野情報「ブログ」に対応する「ＩＤ＝８」（１）のルール「＜h2＞タグに対応する文字列を取得」に従って、タグ「＜h2＞」を検索し、当該タグに対応する値「2008年１月26日(土)」を取得する。そして、ページ情報取得手段１１６３は、文字列「2008年１月26日(土)」と、主情報の種類「日付」とを対応付けて、バッファに格納する。

【０１４３】

次に、ページ情報取得手段１１６３は、分野情報「ブログ」に対応する「ＩＤ＝９」（１）のルール「＜h3＞タグに対応する文字列を取得」に従って、文字列「バスでの移動」を取得する。そして、文字列「バスでの移動」と、主情報の種類「見出し」とを対応付けて、バッファに格納する。

【０１４４】

次に、ページ情報取得手段１１６３は、分野情報「ブログ」に対応する「ＩＤ＝１０」（１）のルール「＜div class="entry-body-text"＞タグに対応する文字列を取得」に従って、文字列「ＸＸスキー場に行くために、バスで深夜から移動した。とても寒く、・・・・・・・・・・・・・・・」を取得する。そして、文字列「ＸＸスキー場に行くために、バスで深夜から移動した。とても寒く、・・・・・・・・・・・・・・・」と、主情報の種類「本文」とを対応付けて、バッファに格納する。

【０１４５】

以上の処理を繰り返し、例えば、バッファ内に図１９の情報を得る。なお、上記は、説明の簡素化のためにルールを、直ちに、ウェブページの情報に適用したが、通常、ページ

情報取得手段１１６３は、ウェブページの中のブロック（一つの区切り）単位の情報を取得し、当該情報に対して、順にルールを適用し、合致するルールに対応する主情報を、前記情報と対にして格納する。

【０１４６】

次に、レイアウト情報取得手段１１７１は、分野情報「ブログ」に対応するレイアウト情報（図７）を、レイアウト情報格納部１１４から読み出す。

【０１４７】

そして、レイアウト手段１１７２は、図７のレイアウト情報を用いて、日付、見出し、本文を順に配置し、かつ、それぞれの主情報の種類に応じた文字サイズ、フォント、文字色で、図１９の情報を配置する。そして、レイアウト手段１１７２は、図２０のページ情報を得て、メモリ上に配置する。そして、２ページ以降のページ情報も、１ページ目と同様に、日付、見出し、本文を順に配置する。

【０１４８】

以上、本実施の形態によれば、ウェブページから、複数ページの集合である電子ブックのコンテンツを自動構成することができる。

【０１４９】

また、本実施の形態によれば、ウェブページの内容に適したレイアウトで、電子ブックのコンテンツを自動構成することができるため、非常に読みやすいコンテンツを提供できる。例えば、ニュース記事のウェブページから、新聞風のレイアウトのコンテンツが構成でき、ブログのウェブページから、日記風のレイアウトのコンテンツが構成できる。

【０１５０】

また、本実施の形態によれば、ウェブページに記載されている情報のうち、重要と思われる情報のみを用いて、電子ブックのコンテンツを自動構成することができる。例えば、ウェブページに表示される広告などの情報を除いて、電子ブックのコンテンツを自動構成することができる。

【０１５１】

なお、本実施の形態によれば、レイアウト情報の定義方法、ページ取得ルール情報の定義情報は問わない。これらの情報は、データとして定義されていても良いし、プログラムにより実現されていても良い。かかることは、他の実施の形態においても同様である。

【０１５２】

また、本実施の形態によれば、主情報取得手段１１６２を有さなくても良い。かることも、他の実施の形態においても同様である。

【０１５３】

また、本実施の形態において、ウェブページの分野やウェブページ特定情報に応じて、異なるレイアウトで、電子ブックのコンテンツを構成した。しかし、かかる分野情報やウェブページ特定情報を用いずに、常に同一のレイアウトで、電子ブックのコンテンツを構成しても良い。かることも、他の実施の形態においても同様である。

【０１５４】

また、本実施の形態において、出力中のウェブページを元として、電子ブックのコンテンツを構成した。しかし、情報処理装置１１が、電子ブックのコンテンツを構成する元となるウェブサイトのＵＲＬを１以上、記憶媒体に格納しており、例えば、予め設定された時刻（例えば、毎日７：００）に、情報処理装置１１が動作を開始するようにしても良い。かかる場合、情報処理装置１１は、記憶媒体に格納された１以上のＵＲＬ（ニュース記事が記載されたウェブページのＵＲＬ）から取得されるウェブサイトから、電子ブックのコンテンツを構成し、例えば、電子ブックプレーヤーに送信する、などの処理を行っても良い。かかる処理を行うことにより、ユーザは、自分の電子ブックプレーヤーに、毎朝、出勤前に、ニュースの情報を新聞に類似したレイアウトのコンテンツを入手でき、通勤中に、見ることができる。かかることは、他の実施の形態においても同様である。

【０１５５】

さらに、本実施の形態における処理は、ソフトウェアで実現しても良い。そして、この

ソフトウェアをソフトウェアダウンロード等により配布しても良い。また、このソフトウェアをＣＤ－ＲＯＭなどの記録媒体に記録して流布しても良い。なお、このことは、本明細書における他の実施の形態においても該当する。なお、本実施の形態における情報処理装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、コンピュータを、１以上のウェブページから、電子ブックのページ単位の情報であるページ情報を、２以上取得するページ情報取得部と、前記ページ情報取得部が取得した２以上のページ情報を出力するページ情報出力部として機能させるためのプログラム、である。

【０１５６】

また、上記プログラムにおいて、前記ページ情報出力部は、ページ内のレイアウトに関する情報であり、前記１以上のウェブページに応じたレイアウト情報を、記憶媒体から取得するレイアウト情報取得手段と、前記レイアウト情報に従って、前記ページ情報取得部が取得した２以上のページ情報をレイアウトするレイアウト手段と、前記レイアウト手段がレイアウトした２以上のページ情報を出力するページ情報出力手段を具備するものとして、コンピュータを機能させることは好適である。

【０１５７】

また、上記プログラムにおいて、コンピュータを、前記１以上のウェブページの内容から、当該１以上のウェブページの分野を決定し、分野情報を取得する分野情報決定部をさらに具備し、前記レイアウト情報格納部は、分野を示す情報である分野情報と、ページ内のレイアウトに関する情報であるレイアウト情報との組を、２組以上格納しており、前記ページ情報出力部のレイアウト情報取得手段は、前記分野情報決定部が取得した分野情報と対になるレイアウト情報を取得するものとして機能させることは好適である。

【０１５８】

また、上記プログラムにおいて、前記ページ情報取得部は、前記１以上のウェブページから情報を取得するための情報であるページ取得ルール情報を、２以上格納しているページ取得ルール情報格納手段と、前記１以上のウェブページに対応するページ取得ルール情報、または前記分野情報決定部が取得した分野情報と対になるページ取得ルール情報を用いて、前記１以上のウェブページから、ページ単位の情報であるページ情報を、２以上取得するページ情報取得手段とを具備するものとして、コンピュータを機能させることは好適である。

【０１５９】

また、上記プログラムにおいて、前記分野情報決定部は、分野情報と、当該分野情報が示す分野に対応する１以上の用語を含む用語情報を１以上格納している用語格納手段と、前記１以上のウェブページから、前記１以上の用語の出現の度合いに関する情報である出現度情報を取得する出現度情報取得手段と、前記出現度情報を用いて、前記１以上のウェブページの分野を決定し、分野情報を取得する分野情報決定手段を具備するものとして、コンピュータを機能させることは好適である。

【０１６０】

また、上記プログラムにおいて、前記ページ情報取得部は、前記１以上のウェブページの各々に対して、ウェブページの構造またはタグまたは内容を用いて、主となる情報を取得する主情報取得手段と、前記主情報取得手段が取得した情報から、ページ情報を取得するページ情報取得手段を具備するものとして、コンピュータを機能させることは好適である。

【０１６１】

（実施の形態２）

本実施の形態において、１以上のウェブページから、ページ単位の情報に分割した電子ブックのコンテンツを得る情報処理システム等について説明する。また、本実施の形態において、ウェブページから分野（ニュース、子供向けなど）を決定し、当該分野に対応するテンプレートを取得し、当該テンプレートに従って、ウェブページを電子ブックのコンテンツに変換する情報処理システム等について説明する。また、ウェブページのＵＲＬに

よって、異なるテンプレートを取得し、当該テンプレートに従って、ウェブページを電子ブックのコンテンツに変換する情報処理システム等について説明する。また、本実施の形態において、ウェブページの分野やＵＲＬによって、ページ情報の取得方法が異なる情報処理システム等について説明する。さらに、本実施の形態において、特定のデータタイプ（例えば、静止画）のウェブページ内での割合から、ウェブページの分野やレイアウトなどを決定する情報処理システム等について説明する。なお、例えば、静止画の割合が多い場合、分野を「子供向け」と決定し、絵本のページ情報を自動生成する。

【０１６２】

本実施の形態における情報処理システム２の概念図は図１である。図２１は、本実施の形態における情報処理システム２のブロック図である。情報処理システム２は、情報処理装置２１、サーバ装置１２を具備する。

【０１６３】

情報処理装置２１は、指示受付部１１０、ウェブページ格納部１１１、ウェブページ受信部１１２、ウェブページ蓄積部１１３、レイアウト情報格納部１１４、分野情報決定部２１５、ページ情報取得部１１６、ページ情報出力部１１７、を具備する。

【０１６４】

分野情報決定部２１５は、データ情報取得手段２１５１、分野情報決定手段２１５２を具備する。

【０１６５】

データ情報取得手段２１５１は、１以上のウェブページから、１以上のデータタイプのデータ量またはデータサイズに関する情報であるデータ情報を取得する。データタイプとは、例えば、静止画、動画、テキスト、グラフィックス、ひらがなのテキスト、漢字のテキストなどである。データ情報とは、データの割合、データ数などである。なお、上記の、１以上のウェブページとは、ページ情報を構成する元になるウェブページであるが、例えば、実施の形態１で述べた主情報取得手段１１６２が処理した後に取得された主情報でも良い。データ情報取得手段２１５１は、通常、ＭＰＵやメモリ等から実現され得る。データ情報取得手段２１５１の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０１６６】

分野情報決定手段２１５２は、データ情報を用いて、１以上のウェブページの分野を決定し、分野情報を取得する。分野情報決定手段２１５２は、例えば、データタイプ、データ情報、分野情報を有する分野情報決定ルールを１以上保持しており、データ情報取得手段２１５１が取得したデータタイプとデータ情報に対応する分野情報を取得する。分野情報決定手段２１５２は、通常、ＭＰＵやメモリ等から実現され得る。分野情報決定手段２１５２の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０１６７】

次に、情報処理システム２の動作について、情報処理システム１の動作と比較すると、分野決定処理が異なる。そこで、情報処理システム２の分野決定処理について、図２２のフローチャートを用いて説明する。なお、図２２のフローチャートにおいて、図４のフローチャートと同一の処理について、説明を省略する。

【０１６８】

（ステップＳ２２０１）データ情報取得手段２１５１は、ｉ番目の分野情報に対応するデータタイプの情報を取得する。データタイプの情報とは、例えば、データタイプが静止画なら「１」、動画なら「２」、ひらがなのテキストなら「３」などである。データタイプの情報は、ｉ番目の分野情報に対応付けて、予め記憶媒体に格納されている。

【０１６９】

（ステップＳ２２０２）データ情報取得手段２１５１は、ステップＳ２２０１で取得したデータタイプの情報に対応するデータタイプのデータを、ページ情報の変換元となるウ

ェブページから取得し、当該取得したデータから、データ情報（例えば、割合）を取得する。

【０１７０】

（ステップＳ２２０３）データ情報取得手段２１５１は、データ情報を用いてスコアを算出し、当該スコアを、バッファに一時格納する。

【０１７１】

以下、本実施の形態における情報処理システム２の具体的な動作について説明する。情報処理システム２の概念図は図１である。

【０１７２】

今、情報処理装置２１のレイアウト情報格納部１１４には、図２３に示すレイアウト情報管理表（レイアウト情報管理表Ｃ）が格納されている。レイアウト情報管理表Ｃは、分野情報「絵本」に対応するレイアウト情報の管理表である。また、レイアウト情報管理表Ｃは、「ＩＤ」「主情報の種類」「レイアウト情報」の属性値を有するレコードを１以上格納している。「レイアウト情報」は、ここでは「配置情報」「属性」を有する。「配置情報」は、ここでは、ページ情報内における、情報の配置を決定するための情報であり、「始点＝（１０，１０）」は、（１０，１０）を始点として、対応する情報が配置される、ことを示す。「配置情報」が「全面」とは、対応する情報が、ページ全体に配置されることを示す。かかる場合、対応する情報は、拡大、または縮小され得る。「配置情報」が「下づめ」とは、対応する情報が、ページ内の、下側に詰めて配置されることを示す。「属性」は、文字の属性であり、ここでは、文字サイズのみを規定している。ただし、文字色やフォントなどの他の文字属性を規定しても良い。なお、レイアウト情報管理表Ｃには、規定されていないが、文字は、横書きで配置される、とする。

また、ページ取得ルール情報格納手段１１６１は、図２４に示すページ取得ルール情報管理表を格納している。

【０１７３】

さらに、分野情報決定部２１５は、「ウェブページ内のビットマップデータの割合が３０％以上の場合は、分野情報が「絵本」である」と判断する、分野情報決定判断ルールを格納している、とする。

【０１７４】

かかる状況において、ユーザが、情報処理装置２１に、図２５のウェブページを表示させた、とする。図２５のウェブページを表示させるための処理については、実施の形態１で述べた。

【０１７５】

次に、ユーザは、電子ブックのコンテンツを自動生成するツールバーを押下した、とする。すると、指示受付部１１０は、ページ情報出力指示を受け付ける。つまり、ツールバーの押下は、ページ情報出力指示である。

【０１７６】

次に、分野情報決定部２１５は、現在表示中のウェブページのスクリプトを取得する。かかる、スクリプト（ここでは、ＨＴＭＬで記述された文字列）を、図２６に示す。

【０１７７】

分野情報決定部２１５は、図２６の情報を用いて、以下のように、分野情報を取得する。つまり、分野情報決定部２１５は、図２６のスクリプトを解析し、ビットマップのサイズ（Ｓ１）を取得する。また、ページ全体のサイズ（Ｓ２）を取得する。そして、分野情報決定部２１５は、「Ｓ１／Ｓ２＞＝０．３」を満たすか否かを判断する。ここで、分野情報決定部２１５は、「Ｓ１／Ｓ２＞＝０．３」を満たす、と判断する。そして、分野情報決定部２１５は、上記の分野情報決定判断ルールを適用し、分野情報が「絵本」である、と判断する。

【０１７８】

次に、ページ情報取得部１１６は、図２６のウェブページから、以下のように、ページ情報を取得する。つまり、ページ情報取得手段１１６３は取得された分野情報「絵本」と

対になるすべてのページ取得ルール情報を、図２４のページ取得ルール情報管理表から読み出す。そして、ページ情報取得手段１１６３は、以下のように、ページ取得ルール情報に従って、１以上のページ情報を取得する。

【０１７９】

まず、ページ情報取得手段１１６３は、分野情報「絵本」に対応する「ＩＤ＝１」（１）のルール「＜title＞タグに対応する文字列を取得」に従って、タグ「＜title＞」を検索し、当該タグに対応する文字列「元気なポチくん」を取得する。そして、ページ情報取得手段１１６３は、図２４の主情報の種類「タイトル」と対応付けて、取得した文字列「元気なポチくん」を、バッファに格納する。

【０１８０】

次に、ページ情報取得手段１１６３は、分野情報「絵本」に対応する「ＩＤ＝１」（２）のルール「＜frame src="　　"＞に従ってデータを取得」に従って、犬のビットマップデータを取得する。つまり、ページ情報取得手段１１６３は、「src="　　"」で指定されたデータ取得先からデータを取得する。そして、ページ情報取得手段１１６３は、図２４の主情報の種類「絵」と対応付けて、取得したビットマップデータを、バッファに格納する。

【０１８１】

次に、ページ情報取得手段１１６３は、分野情報「絵本」に対応する「ＩＤ＝１」（３）のルール「＜body＞タグ内の文字列を取得」に従って、文字列「うちの愛犬のポチは、ほんとうに元気です。・・・・・・・・・・・・・」を取得する。そして、ページ情報取得手段１１６３は、図２４の主情報の種類「文章」と対応付けて、取得した文字列を、バッファに格納する。以上の処理により、図２７に示す情報が、バッファに格納された、こととなる。

【０１８２】

次に、レイアウト情報取得手段１１７１は、取得された分野情報「絵本」に対応するレイアウト情報（図２３）を、レイアウト情報格納部１１４から読み出す。

【０１８３】

次に、レイアウト手段１１７２は、バッファ上に取得された文字列「元気なポチくん」を、主情報の種類「タイトル」に従ったレイアウト（「始点＝（１０，１０）」「属性＝１４ｐｔ」）で配置する。その結果、１４ｐｔの文字サイズで、文字列「元気なポチくん」がページ内の左上に配置される。なお、ここでは、ページは、左上が原点（０，０）であり、右側に行けばｘの値が＋となり、下側に行けばｙの値がプラスとなる、とする。

【０１８４】

また、レイアウト手段１１７２は、バッファ上のビットマップを、主情報の種類「絵」に従ったレイアウトで配置する。その結果、ページの全面に、ビットマップが拡大され、配置される。

【０１８５】

さらに、レイアウト手段１１７２は、バッファ上の文字列「うちの愛犬のポチは、ほんとうに元気です。・・・・・・・・・・・・・」を、主情報の種類「文章」に従ったレイアウト（下づめ」「属性＝２０ｐｔ」）で配置する。

【０１８６】

また、レイアウト手段１１７２は、ビットマップ（絵）が最背面にくるように、主情報を配置する。さらに、レイアウト手段１１７２は、ビットマップの色に対して、文字が映える文字色を決定し、当該文字色に変換して、配置することは好適である。つまり、レイアウト手段１１７２は、以下の文字色変換手段（図示しない）を有する。文字色変換手段は、文字列を配置する際に、文字列を配置する領域のビットマップの情報を取得する。そして、文字色変換手段は、取得したビットマップの領域の色の平均値を算出する。次に、文字色変換手段は、当該平均値を用いて、その補色となる色を算出する。次に、文字色変換手段は、補色の色に文字列の色属性を変更する。以上の処理により、見やすい絵本が自動生成できる。

【０１８７】
なお、上記の処理において、ページ情報取得手段１１６３が取得したビットマップは、文字列を有するビットマップである場合もある。かかる場合に対応して、ページ情報取得手段１１６３は、ビットマップを取得した際には、文字認識処理を行い、内部の文字列を取得することは好適である。そして、ページ情報取得手段１１６３は、取得できた文字列が一定以上の割合である場合、当該ビットマップを文字列と判断することは好適である。
【０１８８】
そして、ページ情報出力手段１１７３は、構成されたページ情報を出力する。以上の処理により、情報処理装置２１は、図２８に示すページ情報が出力される。図２８に示すページ情報は、例えば、絵本の第１ページである。なお、ここでの出力態様は、上述したように、電子ブックプレーヤーへの送信、記憶媒体への蓄積など、種々、考えられる。
【０１８９】
以上、本実施の形態によれば、ウェブページから、複数ページの集合である電子ブックのコンテンツを自動構成することができる。
【０１９０】
また、本実施の形態によれば、ウェブページの内容に適したレイアウトで、電子ブックのコンテンツを自動構成することができるため、非常に読みやすいコンテンツを提供できる。例えば、子供向けのウェブページから、絵本のレイアウトのコンテンツが構成できる。
【０１９１】
また、本実施の形態によれば、ウェブページに記載されている情報のうち、重要と思われる情報のみを用いて、電子ブックのコンテンツを自動構成することができる。例えば、ウェブページに表示される広告などの情報を除いて、電子ブックのコンテンツを自動構成することができる。
【０１９２】
さらに、本実施の形態における情報処理装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、コンピュータを、１以上のウェブページから、電子ブックのページ単位の情報であるページ情報を、２以上取得するページ情報取得部と、前記ページ情報取得部が取得した２以上のページ情報を出力するページ情報出力部として機能させるためのプログラム、である。
【０１９３】
また、上記プログラムにおいて、前記ページ情報出力部は、ページ内のレイアウトに関する情報であり、前記１以上のウェブページに応じたレイアウト情報を、記憶媒体から取得するレイアウト情報取得手段と、前記レイアウト情報に従って、前記ページ情報取得部が取得した２以上のページ情報をレイアウトするレイアウト手段と、前記レイアウト手段がレイアウトした２以上のページ情報を出力するページ情報出力手段を具備するものとして、コンピュータを機能させることは好適である。
【０１９４】
また、上記プログラムにおいて、前記分野情報決定部は、前記１以上のウェブページから、１以上のデータタイプのデータ量またはデータサイズに関する情報であるデータ情報を取得するデータ情報取得手段と、前記データ情報を用いて、前記１以上のウェブページの分野を決定し、分野情報を取得する分野情報決定手段を具備するものとして、コンピュータを機能させることは好適である。
【０１９５】
（実施の形態３）
本実施の形態において、１以上のウェブページから、ページ単位の情報に分割した電子ブックのコンテンツを得る情報処理システム等について説明する。また、本実施の形態において、複数のウェブページから情報を取得し、１以上のページ情報を構成する処理について説明する。かかる処理は、他のウェブページとの共通度合いに応じて、合成されるアルゴリズムが異なる場合についても説明する。

【０１９６】
本実施の形態における情報処理システム３の概念図は図１である。図２９は、本実施の形態における情報処理システム３のブロック図である。情報処理システム３は、情報処理装置３１、サーバ装置１２を具備する。
【０１９７】
情報処理装置３１は、ウェブページ格納部１１１、ウェブページ受信部１１２、ウェブページ蓄積部１１３、レイアウト情報格納部１１４、ページ情報取得部３１６、ページ情報出力部１１７、を具備する。
【０１９８】
ページ情報取得部３１６は、ウェブページ取得手段３１６１、共通度取得手段３１６２、ページ情報取得手段３１６３を具備する。
【０１９９】
ウェブページ取得手段３１６１は、ページ情報を構成する元になる、２以上のウェブページを、１以上のサーバ装置１２から取得する。具体的には、例えば、ウェブページ取得手段３１６１は、ユーザから入力されたキーワードを取得する。そして、ウェブページ取得手段３１６１は、キーワードを用いて、関連キーワードを１以上、上位からｍ個取得する。なお、キーワードに対する関連キーワードを取得する処理は公知技術であるので、説明を省略する。次に、キーワードおよび１以上（ｍ個）の関連キーワードをキーとして、ウェブページを検索する、かかる技術は、いわゆるウェブの検索エンジンの技術である。そして、次に、ウェブページ取得手段３１６１は、検索されたウェブページから、上位ｎ個のＵＲＬを取得する。次に、ウェブページ取得手段３１６１は、上位ｎ個のＵＲＬを用いて、ｎ個のウェブページを取得する。ウェブページ取得手段３１６１は、通常、ＭＰＵやメモリ等から実現され得る。ウェブページ取得手段３１６１の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。
【０２００】
共通度取得手段３１６２は、２以上のウェブページの各々から、内容の共通の度合いを示す情報である共通度を取得する。ここで、２以上のウェブページは、通常、ウェブページ取得手段３１６１が取得したウェブページであるが、予め決められていても良い。共通度取得手段３１６２は、さらに具体的には、２以上の各ウェブページのデータ量を算出し、データ量が最も大きいウェブページに対する、他の各ウェブページの共通度を取得しても良い。また、共通度取得手段３１６２は、データ量が最も大きいウェブページに１以上のウェブページが挿入された後の、挿入対象のウェブページに対する、他の各ウェブページの共通度を取得しても良い。かかる場合、「データ量が最も大きいウェブページ」には、挿入対象のウェブページも含む、とする。なお、共通度とは、２つのウェブページの記載内容の共通度合いに関する情報である。共通度は、例えば、２つのウェブページに、共通して出現する用語（単語）の数、２つのウェブページに、共通して出現する用語（単語）の、全用語に対する割合などである。その他、共通度の算出方法は何でも良い。共通度取得手段３１６２は、通常、ＭＰＵやメモリ等から実現され得る。共通度取得手段３１６２の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。
【０２０１】
ページ情報取得手段３１６３は、２以上のウェブページから２以上のページ情報を取得する。なお、ページ情報取得手段３１６３は、共通度に応じて、取得アルゴリズムが異なることは好適である。例えば、ページ情報取得手段３１６３は、他の各ウェブページの共通度が予め格納されている第一の閾値より小さい場合には、データ量が最も大きいウェブページに対して、前記他の各ウェブページを結合する。なお、「結合」とは、データ量が最も大きいウェブページの後ろ、または前（通常、後ろ）に、他の１以上のウェブページを追記することを言う。この１以上の他のウェブページは、データ量が最も大きいウェブページに対する、共通度が予め格納されている第一の閾値より小さいウェブページである

。そして、ページ情報取得手段３１６３は、共通度が第一の閾値より大きい場合には、各ウェブページを構成する一まとまりの情報であるブロックを取得し、データ量が最も大きいウェブページ（挿入対象のウェブページも含む。）のブロック毎に、他の各ウェブページのブロックのうちで、予め格納されている第二の閾値より大きい共通度を有するブロックを抽出し、当該抽出したブロックを、データ量が最も大きいウェブページの前記ブロック（共通度が大きいブロック）と次のブロックの間に挿入し、出力する全ページ情報を構成する。そして、ページ情報取得手段３１６３は、全ページ情報を、１ページ単位に区切り、２以上のページ情報を取得する。なお、ページ情報取得手段３１６３は、１ページのデータ量（行数など）を保持している、とする。

【０２０２】

さらに、ページ情報取得手段３１６３は、以下のような処理を加えても良い。つまり、ページ情報取得手段３１６３は、他の各ウェブページのブロック中で、データ量が最も大きいウェブページの全ブロックのいずれに対しても、第二の閾値より大きな閾値を有しないブロックについて、その前に配置されているブロックで、データ量が最も大きいウェブページのいずれかのブロックに対して前記第二の閾値より大きな閾値を有するとして、前記データ量が最も大きいウェブページ内に挿入された箇所の直後に挿入する処理を行う。そして、ページ情報取得手段３１６３は、出力する全ページ情報を構成し、全ページ情報を、１ページ単位に区切り、２以上のページ情報を取得する。ページ情報取得手段３１６３は、通常、ＭＰＵやメモリ等から実現され得る。ページ情報取得手段３１６３の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【０２０３】

次に、情報処理システム３の動作について図３０のフローチャートを用いて説明する。図３０のフローチャートにおいて、図３のフローチャートと同一の処理について、説明を省略する。

【０２０４】

（ステップＳ３００１）ページ情報取得部３１６は、ページ情報を構成する元になる、２以上のウェブページを、１以上のサーバ装置１２から取得する。かかるウェブページ取得処理の動作例の詳細について、図３１のフローチャートを用いて説明する。

【０２０５】

（ステップＳ３００２）ページ情報取得部３１６は、ページ取得処理を行う。ページ取得処理の動作例の詳細について、図３２のフローチャートを用いて説明する。

【０２０６】

なお、図３０のフローチャートにおいて、電源オフや処理終了の割り込みにより処理は終了する。

【０２０７】

次に、ステップＳ３００１のウェブページ取得処理の動作例の詳細について、図３１のフローチャートを用いて説明する。

【０２０８】

（ステップＳ３１０１）ウェブページ取得手段３１６１は、ユーザから入力されたキーワードを取得する。

【０２０９】

（ステップＳ３１０２）ウェブページ取得手段３１６１は、キーワードを用いて、関連キーワードを１以上、上位からｍ個取得する。なお、ｍは、１以上の自然数である。なお、関連キーワードを取得する処理は、概念辞書を用いて、キーワードの同義語などを取得しても良いし、過去にユーザが検索エンジンに入力したキーワードセットから取得しても良い。かかる関連キーワードを取得する処理は、公知技術であるので、詳細な説明を省略する。

【０２１０】

（ステップＳ３１０３）ウェブページ取得手段３１６１は、キーワードおよび１以上（

ｍ個）の関連キーワードをキーとして、ウェブページを検索する。なお、ウェブページ取得手段３１６１は、例えば、公知の検索エンジンの構成を有する。

【０２１１】

（ステップＳ３１０４）ウェブページ取得手段３１６１は、ステップＳ３１０３で検索されたウェブページのＵＲＬの中で、上位ｎ個のＵＲＬを取得する。上位ｎ個のＵＲＬとは、いわゆる検索エンジンが出力した、上位ｎ個のウェブページのＵＲＬである。

【０２１２】

（ステップＳ３１０５）ウェブページ取得手段３１６１は、ステップＳ３１０４で取得した上位ｎ個のＵＲＬを用いて、ｎ個のウェブページを取得する。上位処理にリターンする。

【０２１３】

次に、ステップＳ３００２のページ取得処理の動作例の詳細について、図３２のフローチャートを用いて説明する。

【０２１４】

（ステップＳ３２０１）共通度取得手段３１６２は、カウンタｉに１を代入する。

【０２１５】

（ステップＳ３２０２）共通度取得手段３１６２は、ｉ番目のウェブページが存在するか否かを判断する。ｉ番目のウェブページが存在すればステップＳ３２０３に行き、ｉ番目のウェブページが存在しなければステップＳ３２０８に行く。

【０２１６】

（ステップＳ３２０３）共通度取得手段３１６２は、ｉ番目のウェブページのデータ量を算出し、メモリ上に配置する。

【０２１７】

（ステップＳ３２０４）共通度取得手段３１６２は、ｉ番目のウェブページの全用語を取得する。共通度取得手段３１６２は、例えば、以下のようにウェブページの全用語を取得する。つまり、共通度取得手段３１６２は、例えば、ウェブページからタグ（例えば、＜HTML＞、＜HEAD＞、 ＜/HEAD＞など）を消去し、制御文字（例えば、\、"、｛、｝など｝をスペースに置き換えるなどの処理をする。次に、共通度取得手段３１６２は、日本語の文字列を形態素解析し、名詞や動詞などの自立語のみを取得し、各形態素を用語とする。また、英文字列や数値列などは、区切り（スペース、リターンなど）ごとに用語として取得する。

【０２１８】

（ステップＳ３２０５）共通度取得手段３１６２は、ステップＳ３２０４で取得した１以上の用語の集合である用語群から、（用語、出現回数）の組を１組以上取得する。なお、共通度取得手段３１６２は、ステップＳ３２０４で取得した１以上の用語の集合をソートし、（用語、出現回数）の組を１組以上取得することは好適である。

【０２１９】

（ステップＳ３２０６）共通度取得手段３１６２は、ｉ番目のウェブページをまとまりのある１以上のブロックに分割し、ブロックごとにメモリ上に配置する。なお、共通度取得手段３１６２は、例えば、ウェブページの区切りとなるタグ（例えば、＜HR＞、＜DIV＞、＜SPAN＞など）の情報を予め保持しており、当該区切りとなるタグを用いて（パターンマッチングして）、ブロックの区切りを認識し、ウェブページから１以上のブロックを取り出す。

【０２２０】

（ステップＳ３２０７）共通度取得手段３１６２は、カウンタｉを１、インクリメントする。ステップＳ３２０２に戻る。

【０２２１】

（ステップＳ３２０８）ページ情報取得手段３１６３は、２以上のウェブページの中で、データ量が最大のウェブページを決定し、取得する。

【０２２２】

（ステップＳ３２０９）ページ情報取得手段３１６３は、カウンタｊに１を代入する。

【０２２３】

（ステップＳ３２１０）ページ情報取得手段３１６３は、ｊ番目の他のウェブページが存在するか否かを判断する。ｊ番目の他のウェブページが存在すればステップＳ３２１１に行き、ｊ番目の他のウェブページが存在しなければステップＳ３２１６に行く。なお、他のウェブページとは、データ量が最大のウェブページを除くウェブページのことである。

【０２２４】

（ステップＳ３２１１）ページ情報取得手段３１６３は、データ量が最大のウェブページの（用語、出現回数）の組の集合と、ｊ番目の他のウェブページの（用語、出現回数）の組の集合とを用いて、２つのウェブページの共通度を取得する。ここで、共通度は、共通する用語の数でも良いし、共通する用語に対応する出現回数の和（１つの共通する用語に対して、２つの出現回数が加算される）でも良いし、共通する用語の全用語数に対する割合（用語の種類だけで見ても良いし、用語の出現回数を考慮しても良い）などでも良い。なお、データ量が最大のウェブページは、挿入対象のウェブページである、ことは好適である。挿入対象のウェブページは、元のデータ量が最大のウェブページに対して、１以上のウェブページの内容が挿入されたウェブページである。

【０２２５】

（ステップＳ３２１２）ページ情報取得手段３１６３は、共通度が予め格納されている閾値（第一の閾値）より大きいか否かを判断する。共通度が第一の閾値より大きい場合ステップＳ３２１３に行き、大きくない場合ステップＳ３２１５に行く。

【０２２６】

（ステップＳ３２１３）ページ情報取得手段３１６３は、ウェブページの合成処理を行う。合成処理について、図３３のフローチャートを用いて説明する。

【０２２７】

（ステップＳ３２１４）ページ情報取得手段３１６３は、カウンタｊを１、インクリメントする。ステップＳ３２１０に戻る。

【０２２８】

（ステップＳ３２１５）ページ情報取得手段３１６３は、ｊ番目の他のウェブページを、挿入対象のウェブページの最後に付加する。ステップＳ３２１４に行く。

【０２２９】

（ステップＳ３２１６）ページ情報取得手段３１６３は、構成したページ情報に対して、１ページごとに分割する処理を行う。分割処理とは、単に、１ページごとに区切る（例えば、改ページのコードを挿入など）ことである。分割処理は、いわゆるワードプロセッサのページ単位に区切る処理と同様でも良い。上位処理にリターンする。

【０２３０】

次に、ステップＳ３２１３の合成処理について、図３３のフローチャートを用いて説明する。

【０２３１】

（ステップＳ３３０１）ページ情報取得手段３１６３は、カウンタｉに１を代入する。

【０２３２】

（ステップＳ３３０２）ページ情報取得手段３１６３は、ブロックの情報の挿入対象のウェブページについて、ｉ番目のブロックが存在するか否かを判断する。ｉ番目のブロックが存在すればステップＳ３３０３に行き、ｉ番目のブロックが存在しなければステップＳ３３１３に行く。なお、ここで、ブロックの情報の挿入対象のウェブページとは、元の最大のデータ量のウェブページに対して、１以上の他のウェブページのブロックが挿入されたウェブページ、または元の最大のデータ量のウェブページである。

【０２３３】

（ステップＳ３３０３）ページ情報取得手段３１６３は、ブロックの情報の挿入対象のウェブページの中の、ｉ番目のブロックの情報を取得する。

【０２３４】
（ステップＳ３３０４）ページ情報取得手段３１６３は、カウンタｊに１を代入する。
【０２３５】
（ステップＳ３３０５）ページ情報取得手段３１６３は、処理対象の他のウェブページについて、ｊ番目のブロックが存在するか否かを判断する。ｊ番目のブロックが存在すればステップＳ３３０６に行き、ｊ番目のブロックが存在しなければステップＳ３３１２に行く。
【０２３６】
（ステップＳ３３０６）ページ情報取得手段３１６３は他のウェブページの中の、ｊ番目のブロックの情報を取得する。
【０２３７】
（ステップＳ３３０７）ページ情報取得手段３１６３は、ステップＳ３３０３で取得したブロックの情報と、ステップＳ３３０６で取得したブロックの情報とを比較し、共通度を取得する。なお、この共通度は、上述したウェブページの共通度を取得する処理と同様でも良い。また、他のアルゴリズムで共通度を取得しても良い。
【０２３８】
（ステップＳ３３０８）ページ情報取得手段３１６３は、ステップＳ３３０７で取得した共通度が閾値（第二の閾値）より大きいか否かを判断する。大きい場合はステップＳ３３０９に行き、大きくない場合はステップＳ３３１１に行く。
【０２３９】
（ステップＳ３３０９）ページ情報取得手段３１６３は、他のウェブページのｊ番目のブロックの情報を、挿入対象のウェブページのｉ番目のブロックと（ｉ＋１）番目のブロックの間に挿入する。
【０２４０】
（ステップＳ３３１０）ページ情報取得手段３１６３は、他のウェブページからｊ番目のブロックの情報を削除する。
【０２４１】
（ステップＳ３３１１）ページ情報取得手段３１６３は、カウンタｊを１、インクリメントする。ステップＳ３３０５に戻る。
【０２４２】
（ステップＳ３３１２）ページ情報取得手段３１６３は、カウンタｉを１、インクリメントする。ステップＳ３３０２に戻る。
【０２４３】
（ステップＳ３３１３）ページ情報取得手段３１６３は、他のウェブページのブロックのうち、挿入されずに残ったブロックの情報を、挿入対象のウェブページの最後に挿入する。上位処理にリターンする。
【０２４４】
以下、本実施の形態における情報処理システム３の具体的な動作について説明する。情報処理システム３の概念図は図１である。
【０２４５】
ユーザは、Ｌａｔｅｘのｔａｂｕｌａｒ環境について勉強するために、情報処理装置３１に対して、「Ｌａｔｅｘ　ｔａｂｕｌａｒ」のキーワードを有するページ情報出力指示を入力した、とする。すると、ページ情報取得部３１６は、以下のように３つのウェブページを取得する。
【０２４６】
つまり、ウェブページ取得手段３１６１は、ユーザから入力されたキーワード「Ｌａｔｅｘ　ｔａｂｕｌａｒ」を取得する。次に、ウェブページ取得手段３１６１は、キーワードを用いて、関連キーワードを検索し、上位から２個「ｔｅｘ　ｌａｔｅｘコマンド」を取得する。そして、ウェブページ取得手段３１６１は、キーワードおよび１以上（ｍ個）の関連キーワード「Ｌａｔｅｘ　ｔａｂｕｌａｒ　ｔｅｘ　ｌａｔｅｘコマンド」をキー

として、ウェブページを検索する。

【０２４７】

次に、ウェブページ取得手段３１６１は、検索されたウェブページのＵＲＬの中で、上位３個のＵＲＬを取得する。そして、ウェブページ取得手段３１６１は、取得した上位３つのＵＲＬを用いて、３つのウェブページを取得する。かかるウェブページを、図３４の（ａ）から（ｃ）に示す。かかるウェブページは、出力されている場合のイメージである。

【０２４８】

次に、ページ情報取得部３１６は、以下のように３つのウェブページを合成する、ページ取得処理を行う。

【０２４９】

まず、共通度取得手段３１６２は、１番目のウェブページ（図３４（ａ））について、データ量「１００６５ｂｙｔｅ」を算出する。次に、共通度取得手段３１６２は、１番目のウェブページの全用語（「表組」「TeX」「表」・・・「表組」・・・「tabular」・・・）を取得する。

【０２５０】

次に、共通度取得手段３１６２は、取得した１以上の用語の集合である用語群から、（用語、出現回数）の組を１組以上取得する。ここでは、共通度取得手段３１６２は、（表組，２）（Tex，３）（tabular，３４）（表，１７）・・・、を取得した、とする。

【０２５１】

次に、共通度取得手段３１６２は、１番目のウェブページ（図３４（ａ））をまとまりのある１以上のブロックに分割し、ブロックごとにメモリ上に配置する。ここで、共通度取得手段３１６２は、ウェブページの区切りとなるタグ（例えば、＜HR＞、＜DIV＞、＜SPAN＞など）の情報を予め保持しており、当該区切りとなるタグを用いて（パターンマッチングして）、ブロックの区切りを認識し、ウェブページから１以上のブロックを取り出す。

【０２５２】

同様に、共通度取得手段３１６２は、２番目のウェブページ（図３４（ｂ））について、データ量「２３５５ｂｙｔｅ」を算出する。次に、共通度取得手段３１６２は、２番目のウェブページの全用語（「LaTeX」「以下」・・・「開発」・・・「tabular」・・・）を取得する。

【０２５３】

次に、共通度取得手段３１６２は、取得した１以上の用語の集合である用語群から、（用語、出現回数）の組、（LaTeX，１）（以下，１）・・・（tabular，２）・・・、を取得した、とする。

【０２５４】

次に、共通度取得手段３１６２は、２番目のウェブページ（図３４（ｂ））をまとまりのある１以上のブロックに分割し、ブロックごとにメモリ上に配置する。

【０２５５】

同様に、共通度取得手段３１６２は、３番目のウェブページ（図３４（ｃ））について、データ量「７５２２ｂｙｔｅ」を算出する。次に、共通度取得手段３１６２は、２番目のウェブページの全用語（「shortstack」・・・・「tabular」・・・）を取得する。

【０２５６】

次に、共通度取得手段３１６２は、取得した１以上の用語の集合である用語群から、（用語、出現回数）の組、（shortstack，１）・・・（tabular，３２）・・・、を取得した、とする。

【０２５７】

次に、共通度取得手段３１６２は、３番目のウェブページ（図３４（ｃ））をまとまりのある１以上のブロックに分割し、ブロックごとにメモリ上に配置する。

【０２５８】

ページ情報取得手段３１６３は、２以上のウェブページの中で、データ量が最大のウェブページを１番目のウェブページ（図３４（ａ））と決定し、取得する。

【０２５９】

次に、ページ情報取得手段３１６３は、１番目の他のウェブページ（図３４（ｃ））を取得する。なお、ここでは、他のウェブページについて、データ量が多い方から処理する、とする。

【０２６０】

次に、ページ情報取得手段３１６３は、データ量が最大のウェブページの（用語、出現回数）の組の集合と、１番目の他のウェブページ（図３４（ｃ））の（用語、出現回数）の組の集合とを用いて、２つのウェブページの共通度を、「１９８」と取得した、とする。なお、共通度は、共通する用語と対になるすべての出現回数の和である、とする。

【０２６１】

次に、ページ情報取得手段３１６３は、共通度が予め格納されている閾値（第一の閾値）「１００」より大きいか否かを判断する。ここで、２つのウェブページの共通度「１９８」は、第一の閾値「１００」より大きいので、ページ情報取得手段３１６３は、図３３のフローチャートを用いて説明したように、１番目の他のウェブページ（図３４（ｃ））の各ブロックを、共通度が第二の閾値より大きい、データ量が最大のウェブページのブロックの直後に挿入していく。

【０２６２】

そして、ページ情報取得手段３１６３は、１番目の他のウェブページ（図３４（ｃ））のブロックのうち、データ量が最大のウェブページに挿入されなかったブロックを、データ量が最大のウェブページの最後に挿入していく。

【０２６３】

次に、ページ情報取得手段３１６３は、以下のように、２番目の他のウェブページ（図３４（ｂ））の合成処理を行う。つまり、ページ情報取得手段３１６３は、挿入対象のウェブページの（用語、出現回数）の組の集合と、２番目の他のウェブページ（図３４（ｂ））の（用語、出現回数）の組の集合とを用いて、２つのウェブページの共通度を、「６８」と取得した、とする。

【０２６４】

次に、ページ情報取得手段３１６３は、共通度が予め格納されている閾値（第一の閾値）「１００」より大きいか否かを判断する。ここで、２つのウェブページの共通度「６８」は、第一の閾値「１００」より小さいので、ページ情報取得手段３１６３は、図３２のフローチャートを用いて説明したように、１番目の他のウェブページ（図３４（ｃ））の各ブロックを、挿入対象のウェブページ（図３４（ａ）と（ｃ）のブロックが合成された後のウェブページ）の最下位に付加する。

【０２６５】

以上の処理により、３つのウェブページが合成された。合成後のイメージ図は、図３５である。図３５において、「１」は図３４（ａ）のウェブページを構成する１以上のブロック、「２」は図３４（ｃ）のウェブページを構成する１以上のブロック、「３」は図３４（ｂ）のウェブページ（全ブロック）を示す。そして、図３６は、合成後のウェブページである。

【０２６６】

次に、図３６のウェブページに対して、ページ情報取得手段３１６３は、１ページごとに分割する処理を行う。ページ情報取得手段３１６３は、例えば、ＨＴＭＬで記述された図３６のウェブページをフラットなｔｅｘｔファイル（ビットマップなどを含み得る）にして、ページ単位に区切る処理を行う。そして、ページ情報取得手段３１６３は、２以上のページ情報を得て、メモリ上に配置する。

【０２６７】

次に、レイアウト情報取得手段１１７１は、レイアウト情報を、レイアウト情報格納部１１４から読み出す。そして、レイアウト手段１１７２は、取得されたページ情報（図３

６の情報）を、読み出されたレイアウト情報に従ってレイアウトし、１以上のページ情報をメモリ上に得る。次に、ページ情報出力手段１１７３は、メモリ上の１以上のページ情報を出力する。ここでの出力は、例えば、電子ブックプレーヤーへの送信や、着脱可能な可搬型の記憶媒体への蓄積などである。

【０２６８】

以上、本実施の形態によれば、複数のウェブページの内容を用いて、電子ブックのコンテンツを自動構成することができる。また、本実施の形態によれば、複数のウェブページの内容を用いて、近似する内容の情報が近いところに配置され、整理された電子ブックのコンテンツを自動構成することができる。さらに、本実施の形態によれば、かかる構成により、ユーザが欲する情報を、複数のウェブページから、自動的に選択し、有用な電子ブックのコンテンツを自動構成することができる。

【０２６９】

なお、本実施の形態における情報処理装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、コンピュータを、１以上のウェブページから、電子ブックのページ単位の情報であるページ情報を、２以上取得するページ情報取得部と、前記ページ情報取得部が取得した２以上のページ情報を出力するページ情報出力部として機能させるためのプログラム、である。

【０２７０】

また、上記プログラムにおいて、前記ページ情報出力部は、ページ内のレイアウトに関する情報であり、前記１以上のウェブページに応じたレイアウト情報を、記憶媒体から取得するレイアウト情報取得手段と、前記レイアウト情報に従って、前記ページ情報取得部が取得した２以上のページ情報をレイアウトするレイアウト手段と、前記レイアウト手段がレイアウトした２以上のページ情報を出力するページ情報出力手段を具備するものとして、コンピュータを機能させることは好適である。

【０２７１】

また、上記プログラムにおいて、前記ページ情報取得部は、２以上のウェブページの中の一のウェブページを決定し、当該一のウェブページと、他の１以上の各ウェブページの組のうち、１以上の組に対して、内容の共通の度合いを示す情報である共通度を、１以上、取得する共通度取得手段と、前記２以上のウェブページから２以上のページ情報を取得するページ情報取得手段を具備し、前記ページ情報取得手段は、前記共通度に応じて、前記他の１以上の各ウェブページからの情報の取得方法が異なるものとして、コンピュータを機能させるためのプログラム、であることは好適である。

【０２７２】

また、上記プログラムにおいて、前記共通度取得手段は、前記２以上の各ウェブページのデータ量を算出し、データ量が最も大きいウェブページに対する、他の各ウェブページの共通度を取得し、前記ページ情報取得手段は、前記共通度が予め格納されている第一の閾値より小さい場合には、データ量が最も大きいウェブページに対して、前記他のウェブページを結合し、前記共通度が前記第一の閾値より大きい場合には、各ウェブページを構成する一まとまりの情報であるブロックを取得し、前記データ量が最も大きいウェブページのブロック毎に、前記他のウェブページのブロックのうちで、予め格納されている第二の閾値より大きい共通度を有するブロックを抽出し、当該抽出したブロックを、前記データ量が最も大きいウェブページの前記ブロックと次のブロックの間に挿入し、出力する全ページ情報を構成し、前記全ページ情報を、１ページ単位に区切り、２以上のページ情報を取得するものとして、コンピュータを機能させるためのプログラム、であることは好適である。

【０２７３】

また、上記プログラムにおいて、前記ページ情報取得手段は、前記共通度が前記第一の閾値より大きい場合、前記他のウェブページのブロック中で、前記データ量が最も大きいウェブページの全ブロックのいずれに対しても、前記第二の閾値より大きな閾値を有しないブロックについて、その前に配置されているブロックで、前記データ量が最も大きいウ

ェブページのいずれかのブロックに対して前記第二の閾値より大きな閾値を有するとして、前記データ量が最も大きいウェブページ内に挿入された箇所の直後に挿入する処理をさらに行い、出力する全ページ情報を構成し、前記全ページ情報を、１ページ単位に区切り、２以上のページ情報を取得するものとして、コンピュータを機能させるためのプログラム、であることは好適である。

【０２７４】

また、上記プログラムにおいて、前記ページ情報取得部は、ユーザから受け付けたキーワードを用いて、１以上のサーバ装置１２からウェブページを検索し、当該検索したウェブページから、上位のｎ（ｎは２以上の自然数）のウェブページを取得するものとして、コンピュータを機能させるためのプログラム、であることは好適である。

【０２７５】

また、図３７は、本明細書で述べたプログラムを実行して、上述した実施の形態の情報処理装置等を実現するコンピュータの外観を示す。上述の実施の形態は、コンピュータハードウェア及びその上で実行されるコンピュータプログラムで実現され得る。図３７は、このコンピュータシステム３４０の概観図であり、図３８は、コンピュータシステム３４０の内部構成を示す図である。

【０２７６】

図３７において、コンピュータシステム３４０は、ＦＤ（Ｆｌｏｐｐｙ　Ｄｉｓｋ(登録商標)）ドライブ３４１１、ＣＤ－ＲＯＭ（Ｃｏｍｐａｃｔ　Ｄｉｓｋ　Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）ドライブ３４１２を含むコンピュータ３４１と、キーボード３４２と、マウス３４３と、モニタ３４４とを含む。

【０２７７】

図３８において、コンピュータ３４１は、ＦＤドライブ３４１１、ＣＤ－ＲＯＭドライブ３４１２に加えて、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）３４１３と、ＣＰＵ３４１３、ＣＤ－ＲＯＭドライブ３４１２及びＦＤドライブ３４１１に接続されたバス３４１４と、ブートアッププログラム等のプログラムを記憶するためのＲＯＭ３４１５と、ＣＰＵ３４１３に接続され、アプリケーションプログラムの命令を一時的に記憶するとともに一時記憶空間を提供するためのＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）３４１６と、アプリケーションプログラム、システムプログラム、及びデータを記憶するためのハードディスク３４１７とを含む。ここでは、図示しないが、コンピュータ３４１は、さらに、ＬＡＮへの接続を提供するネットワークカードを含んでも良い。

【０２７８】

コンピュータシステム３４０に、上述した実施の形態の情報処理装置等の機能を実行させるプログラムは、ＣＤ－ＲＯＭ３５０１、またはＦＤ３５０２に記憶されて、ＣＤ－ＲＯＭドライブ３４１２またはＦＤドライブ３４１１に挿入され、さらにハードディスク３４１７に転送されても良い。これに代えて、プログラムは、図示しないネットワークを介してコンピュータ３４１に送信され、ハードディスク３４１７に記憶されても良い。プログラムは実行の際にＲＡＭ３４１６にロードされる。プログラムは、ＣＤ－ＲＯＭ３５０１、ＦＤ３５０２またはネットワークから直接、ロードされても良い。

【０２７９】

プログラムは、コンピュータ３４１に、上述した実施の形態の情報処理装置等の機能を実行させるオペレーティングシステム（ＯＳ）、またはサードパーティープログラム等は、必ずしも含まなくても良い。プログラムは、制御された態様で適切な機能（モジュール）を呼び出し、所望の結果が得られるようにする命令の部分のみを含んでいれば良い。コンピュータシステム３４０がどのように動作するかは周知であり、詳細な説明は省略する。

【０２８０】

なお、上記プログラムにおいて、情報を送信する送信ステップや、情報を受信する受信ステップなどでは、ハードウェアによって行われる処理、例えば、送信ステップにおける

モデムやインターフェースカードなどで行われる処理（ハードウェアでしか行われない処理）は含まれない。

【０２８１】

また、上記プログラムを実行するコンピュータは、単数であってもよく、複数であってもよい。すなわち、集中処理を行ってもよく、あるいは分散処理を行ってもよい。

【０２８２】

また、上記各実施の形態において、一の装置に存在する２以上の通信手段（端末情報送信部、端末情報受信部など）は、物理的に一の媒体で実現されても良いことは言うまでもない。

【０２８３】

また、上記各実施の形態において、各処理（各機能）は、単一の装置（システム）によって集中処理されることによって実現されてもよく、あるいは、複数の装置によって分散処理されることによって実現されてもよい。

【０２８４】

本発明は、以上の実施の形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。

【産業上の利用可能性】

【０２８５】

以上のように、本発明にかかる情報処理システムは、ウェブページから、複数ページの集合である電子ブックのコンテンツを自動構成することができる、という効果を有し、情報処理装置、電子ブックプレーヤー等として有用である。

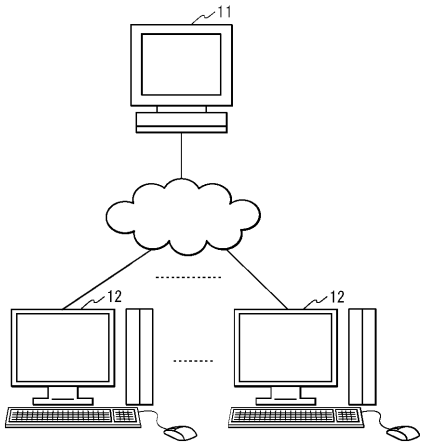
【図面の簡単な説明】

【０２８６】

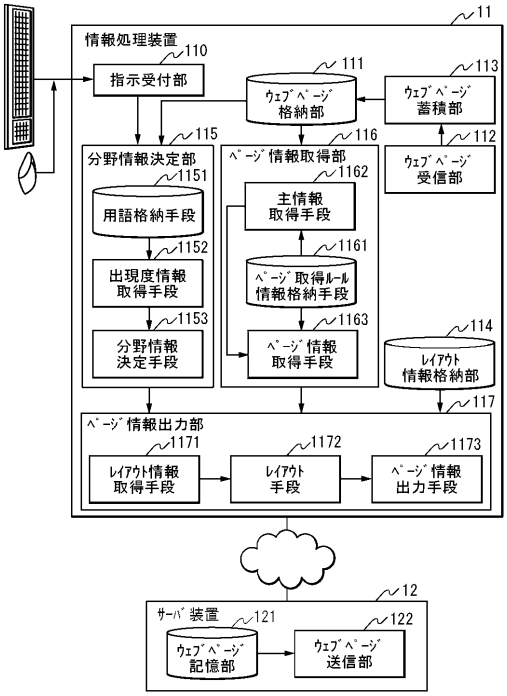
【図１】実施の形態１における情報処理システム概念図
【図２】同情報処理システムのブロック図
【図３】同情報処理装置の動作について説明するフローチャート
【図４】同分野決定処理の動作について説明するフローチャート
【図５】同ページ情報取得処理の動作について説明するフローチャート
【図６】同レイアウト情報管理表を示す図
【図７】同レイアウト情報管理表を示す図
【図８】同レイアウトイメージを示す図
【図９】同レイアウトイメージを示す図
【図１０】同ページ取得ルール情報管理表を示す図
【図１１】同ページ取得ルール情報管理表を示す図
【図１２】同用語情報管理表を示す図
【図１３】同ウェブページの例を示す図
【図１４】同ウェブページのスクリプトを示す図
【図１５】同レイアウト結果を示す図
【図１６】同レイアウト結果を示す図
【図１７】同ウェブページの例を示す図
【図１８】同ウェブページのスクリプトを示す図
【図１９】同バッファ上のデータ例を示す図
【図２０】同レイアウト結果を示す図
【図２１】実施の形態２における情報処理システムのブロック図
【図２２】同分野情報決定処理について説明するフローチャート
【図２３】同レイアウト情報管理表を示す図
【図２４】同ページ取得ルール情報管理表を示す図
【図２５】同ウェブページの例を示す図
【図２６】同ウェブページのスクリプトを示す図
【図２７】同バッファ上のデータ例を示す図

【図２８】同レイアウト結果を示す図
【図２９】実施の形態３における情報処理システムのブロック図
【図３０】同情報処理システムの動作について説明するフローチャート
【図３１】同ウェブページ取得処理の動作について説明するフローチャート
【図３２】同ページ取得処理の動作について説明するフローチャート
【図３３】同合成処理の動作について説明するフローチャート
【図３４】同ウェブページの例を示す図
【図３５】同合成後のイメージ図
【図３６】同合成後のウェブページのイメージ図
【図３７】同コンピュータシステムの外観図
【図３８】同コンピュータシステムのブロック図
【符号の説明】
【０２８７】
１、２、３　情報処理システム
１１、２１、３１　情報処理装置
１２　サーバ装置
１１０　指示受付部
１１１　ウェブページ格納部
１１２　ウェブページ受信部
１１３　ウェブページ蓄積部
１１４　レイアウト情報格納部
１１５、２１５　分野情報決定部
１１６、３１６　ページ情報取得部
１１７　ページ情報出力部
１２１　ウェブページ記憶部
１２２　ウェブページ送信部
１１５１　用語格納手段
１１５２　出現度情報取得手段
１１５３、２１５２　分野情報決定手段
１１６１　ページ取得ルール情報格納手段
１１６２　主情報取得手段
１１６３、３１６３　ページ情報取得手段
１１７１　レイアウト情報取得手段
１１７２　レイアウト手段
１１７３　ページ情報出力手段
２１５１　データ情報取得手段
３１６１　ウェブページ取得手段
３１６２　共通度取得手段

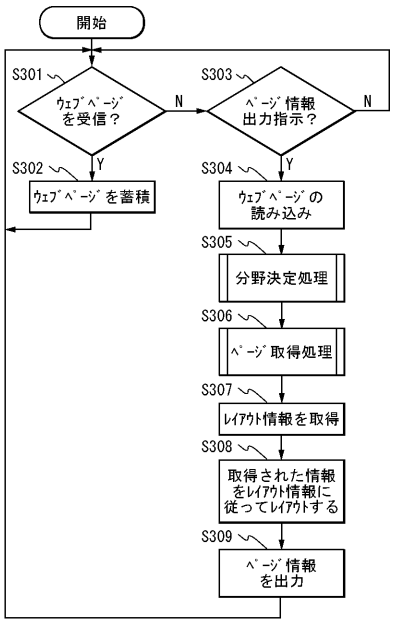
【図１】


11
12
12

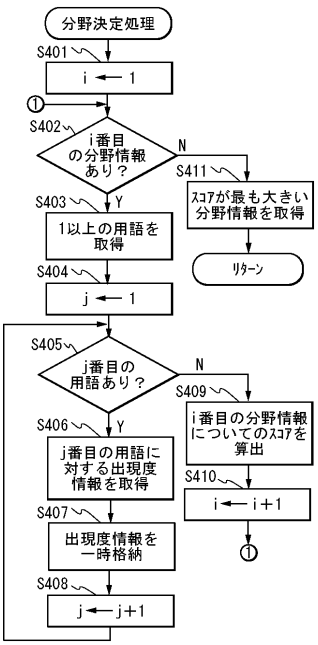
【図２】


11
情報処理装置
110
指示受付部
111
ｳｪﾌﾞﾍﾟｰｼﾞ
格納部
113
ｳｪﾌﾞﾍﾟｰｼﾞ
蓄積部
112
ｳｪﾌﾞﾍﾟｰｼﾞ
受信部
115
分野情報決定部
1151
用語格納手段
1152
出現度情報
取得手段
1153
分野情報
決定手段
116
ﾍﾟｰｼﾞ情報取得部
1162
主情報
取得手段
1161
ﾍﾟｰｼﾞ取得ﾙｰﾙ
情報格納手段
1163
ﾍﾟｰｼﾞ情報
取得手段
114
ﾚｲｱｳﾄ
情報格納部
117
ﾍﾟｰｼﾞ情報出力部
1171
ﾚｲｱｳﾄ情報
取得手段
1172
ﾚｲｱｳﾄ
手段
1173
ﾍﾟｰｼﾞ情報
出力手段
12
ｻｰﾊﾞ装置
121
ｳｪﾌﾞﾍﾟｰｼﾞ
記憶部
122
ｳｪﾌﾞﾍﾟｰｼﾞ
送信部

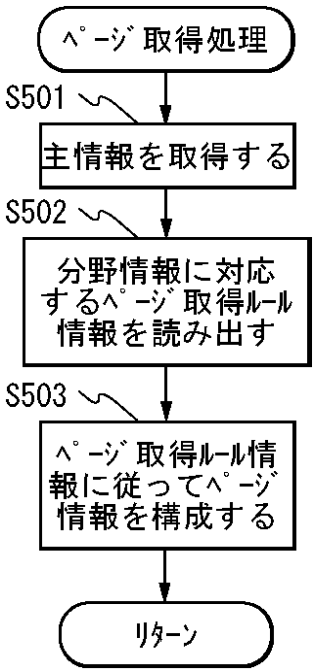
【図３】


開始
S301
ｳｪﾌﾞﾍﾟｰｼﾞ
を受信？
N
Y
S302
ｳｪﾌﾞﾍﾟｰｼﾞを蓄積
S303
ﾍﾟｰｼﾞ情報
出力指示？
N
Y
S304
ｳｪﾌﾞﾍﾟｰｼﾞの
読み込み
S305
分野決定処理
S306
ﾍﾟｰｼﾞ取得処理
S307
ﾚｲｱｳﾄ情報を取得
S308
取得された情報
をﾚｲｱｳﾄ情報に
従ってﾚｲｱｳﾄする
S309
ﾍﾟｰｼﾞ情報
を出力

【図４】


分野決定処理
S401
i ← 1
①
S402
i番目
の分野情報
あり？
N
Y
S411
ｽｺｱが最も大きい
分野情報を取得
ﾘﾀｰﾝ
S403
1以上の用語を
取得
S404
j ← 1
S405
j番目の
用語あり？
N
Y
S409
i番目の分野情報
についてのｽｺｱを
算出
S410
i ← i＋1
①
S406
j番目の用語に
対する出現度
情報を取得
S407
出現度情報を
一時格納
S408
j ← j＋1

【図５】



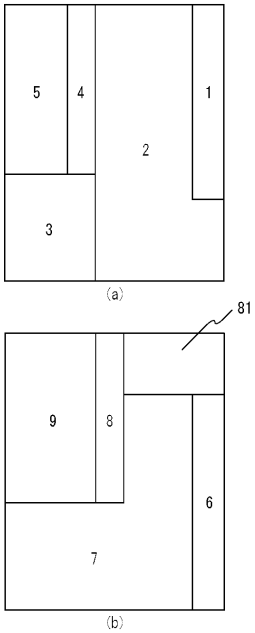
【図６】

ID	主情報の種類	ﾍﾟｰｼﾞ数	領域	文字ｻｲｽﾞ
1	ﾒｲﾝ記事の見出し	1	(0, 0) (0, 250) (40, 0) (40, 250)	24
2	ﾒｲﾝ記事の内容	1	(40, 0) (40, 250) (0, 250) (0, 297) (110, 297) (110, 0)	12
3	ﾒｲﾝ画像	1	(110, 230) (110, 297) (210, 230) (210, 297)	――
4	ｻﾌﾞ記事1の見出し	1	(110, 0) (110, 230) (140, 230) (140, 0)	20
5	ｻﾌﾞ記事1の内容	1	(140, 0) (140, 230) (210, 230) (210, 0)	12
6	一般記事の見出し1	2～	始点 (0, 0) (w, h) = (20, 150)	18
7	一般記事の内容1	2～	(0, 0) (0, 297) (210, 297) (210, 198) (90, 198) (90, 0)	12
8	一般記事の見出し2	2～	始点 (90, 0) (w, h) = (20, 198)	18
9	一般記事の内容2	2～	(110, 0) (110, 198) (210, 198) (210, 0)	12

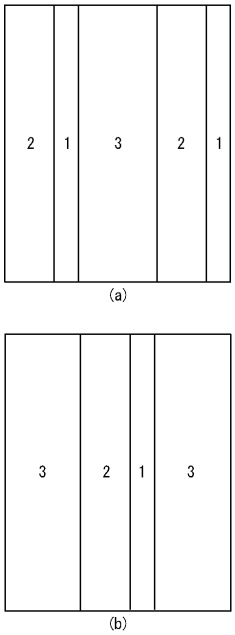
【図７】

ID	主情報の種類	配置情報	文字ｻｲｽﾞ	ﾌｫﾝﾄ	文字色
1	日付	1	14	ｺﾞｼｯｸ	赤
2	見出し	2	28	明朝・太	青
3	本文	3	16	明朝	黒

【図８】



【図９】



【図１０】

分野情報	ID	ﾍﾟｰｼﾞ取得ﾙｰﾙ	
		ﾙｰﾙ	主情報の種類
ﾆｭｰｽ	1	(1)「class＝“tab? on”」に対応するidの値を取得 (2)idの値の文字列を含む 内の項目(に対応)を取得 (3)<img src＝“?*” alt＝“[photo]”>が付与されている文字列を取得	ﾒｲﾝ記事の見出し
	2	(1)ﾒｲﾝ記事の見出しに対応するｱﾝｶｰを取得 (2)ｱﾝｶｰの先のｳｪﾌﾞﾍﾟｰｼﾞを取得 (3)[記事全文]に対応するｳｪﾌﾞﾍﾟｰｼﾞを取得 (4)ｳｪﾌﾞﾍﾟｰｼﾞ中 に対応する文字列を取得	ﾒｲﾝ記事の内容
	3	(1)<img src＝“　　”>のgifﾃﾞｰﾀを取得	ﾒｲﾝ画像
	4	(1)「class＝“tab? on”」に対応するidの値を取得 (2)idの文字列を含む 内の項目であり、ﾒｲﾝ記事の見出しを除く項目を取得 (3)各項目に対応するｱﾝｶｰを取得 (4)ｱﾝｶｰの先のｳｪﾌﾞﾍﾟｰｼﾞを取得 (5)ｳｪﾌﾞﾍﾟｰｼﾞ中の に対応する文字列を取得 (6)文字列中で、そのｻｲｽﾞが1ﾍﾟｰｼﾞ目の残り領域ｻｲｽﾞに最も合致する項目(文字列)を取得	ｻﾌﾞ記事1の見出し
	5	(1)ｻﾌﾞ記事1の に対応する文字列を取得	ｻﾌﾞ記事1の内容

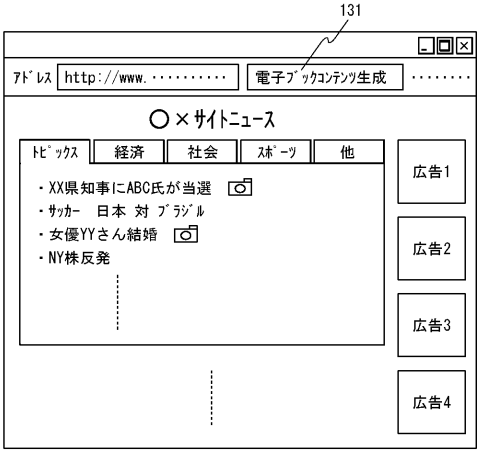
【図１１】

分野情報		ﾍﾟｰｼﾞ取得ﾙｰﾙ	
		ﾙｰﾙ	主情報の種類
ﾆｭｰｽ	6	(1)他の項目に対応する文字列を取得	一般記事の見出し
	7	(1)一般記事の見出しに対応するｱﾝｶｰを取得 (2)ｱﾝｶｰの先のｳｪﾌﾞﾍﾟｰｼﾞを取得 (3)ｳｪﾌﾞﾍﾟｰｼﾞ中の に対応する文字列を取得	一般記事の内容
ﾌﾞﾛｸﾞ	8	(1)<h2>ﾀｸﾞに対応する文字列を取得	日付
	9	(1)<h3>ﾀｸﾞに対応する文字列を取得	見出し
	10	(1)<div class＝“entry-body-text”>ﾀｸﾞに対応する文字列を取得	本文

【図１２】

分野情報	用語	ｽｺｱ
ﾆｭｰｽ	ﾄﾋﾟｯｸｽ	1
	経済	1
	社会	1
	ｽﾎﾟｰﾂ	1
	“ﾆｭｰｽ” in title	10
	⋮	⋮
ﾌﾞﾛｸﾞ	$日付ﾀｲﾌﾟ	2
	俺	1
	おれ	1
	⋮	⋮

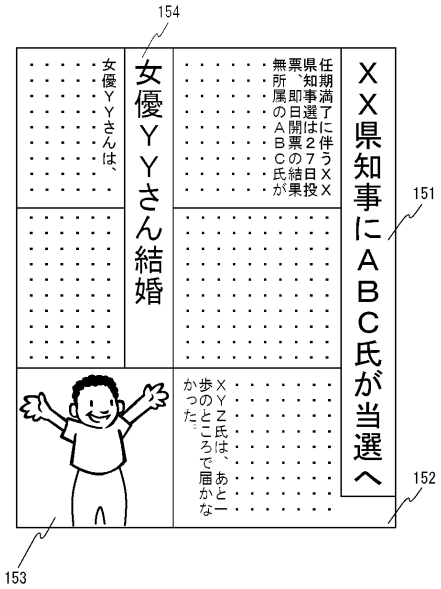
【図１３】



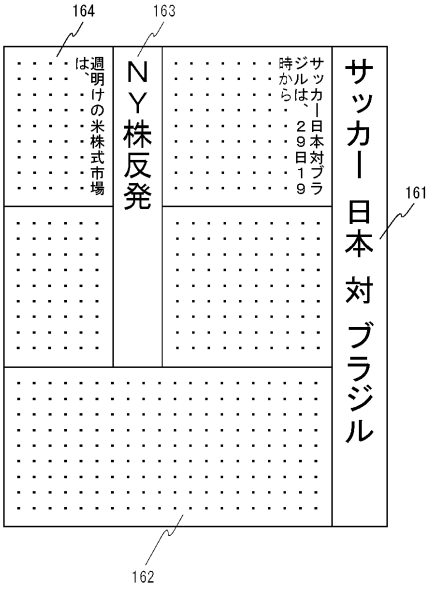
【図１４】

```
<html lang="ja">
<head>
                                        :
<title>○×サイトニュース</title>
                                        :
<hr class="separate">
<div id="division">
<div id="main">
<div id="topicsbox" class="bx">
<div class="hd">
<ul class="tab on0">
<li class="tab0 on"><span><a id="topics" href="r/ttp" hidefocus="true">
トピックス</a></span></li>
<li class="tab1"><span><a id="economy" href="r/teco"
hidefocus="true">経済</a>< span>< li>
<li class="tab2"><span><a id="social" href="r/tent" hidefocus="true">
社会</a></span></li>
<li class="tab3"><span><a id="sports" href="r/tspo" hidefocus="true">
スポーツ</a></span></li>
<li class="tab4 last"><span><a id="others" href="r/toth"
hidefocus="true">他</a></span>< li>
</ul>
</div>
<div id="topicsboxbd">
<div id="topicsfb" class="current">
<div class="topicsindex">
<ul class="emphasis">
<li><a href="f/topics/top/1/?1201573096">XX県知事にABC氏が当選
へ</a> <img src="http://photo.gif" alt="[photo]"> </li>
<li><a href="f/topics/top/2/?1201577113">サッカー　日本対ブラジル
</a></li>
<li><a href="f/topics/top/3/?1201574581">女優YYさん　結婚</a><img
src="http://photo2.gif" alt="[photo]">< li>
<li><a href="f/topics/top/7/?1201575515">NY株　反発</a></li>
                                        :
</ul>
                                        :
                                        :
</html>
```

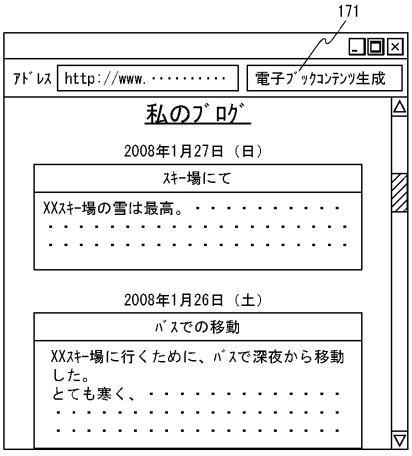
【図１５】



【図１６】



【図１７】



【図１８】

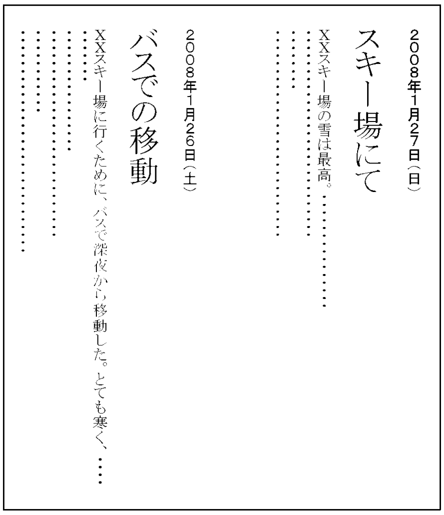
```
<html xmlns="・・・・・・・・・・・・・・・・・・・・">
<head>
                                                    :
<title>私のブログ日記</title>
                                                    :
<h2>2008年1月27日(日)</h2>
                                                    :
<h3>スキー場にて</h3>
                                                    :
<div class="entry-body">
<div class="entry-body-text">
  XXスキー場の雪は最高。・・・・・・・・・・・・・・・・・・・・
</div>
</div>

<h2>2008年1月26日(土)</h2>
<h3>バスでの移動</h3>
                                                    :
<div class="entry-body">
<div class="entry-body-text">
  XXスキー場に行くために、バスで深夜から移動した。とても寒く、・・・
・・・・・・・・・
</div>
</div>
                                                    :
</body>
</html>
```

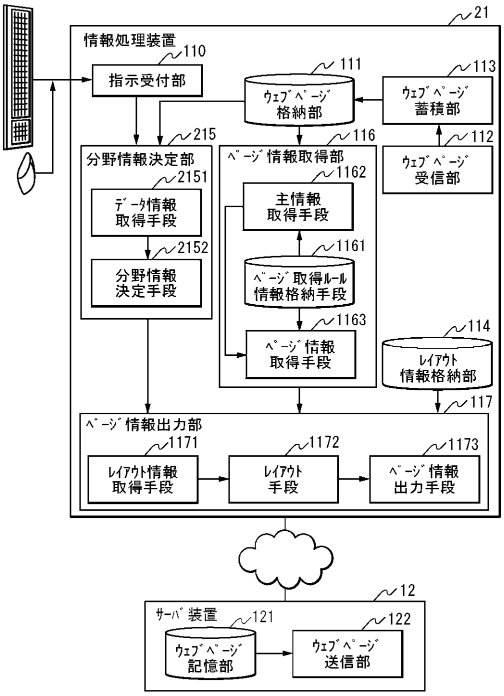
【図１９】

ID	主情報の種類	ﾃﾞｰﾀ
1	日付	2008年1月27日（日）
2	見出し	ｽｷｰ場にて
3	本文	XXｽｷｰ場は最高・・・・・・・・・・・・・・・・・
4	日付	2008年1月26日（土）
5	見出し	ﾊﾞｽでの移動
6	本文	XXｽｷｰ場に行くために、ﾊﾞｽで深夜から移動した。とても寒く、・・・・・・・・・・・・・・・・・・・
⋮	⋮	⋮

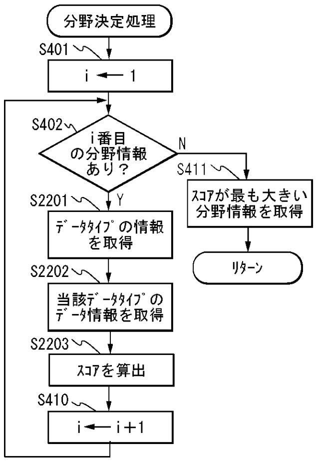
【図２０】



【図２１】



【図２２】



【図２３】

ID	主情報の種類	ﾚｲｱｳﾄ情報	
		配置情報	属性
1	ﾀｲﾄﾙ	始点＝(10, 10)	14pt
2	絵	全面	――
3	文章	下づめ	20pt
⋮	⋮	⋮	⋮

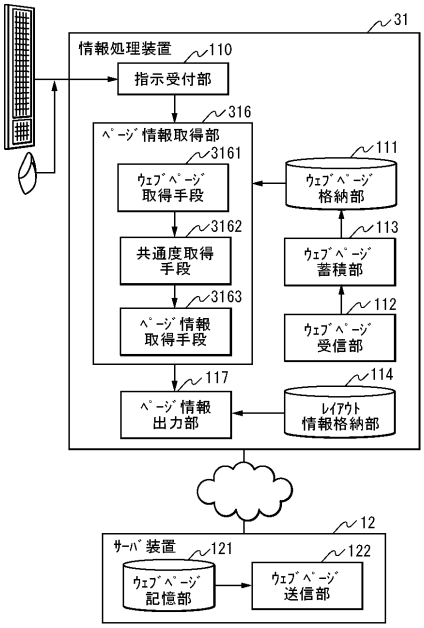
【図２４】

ID	分野情報	ﾍﾟｰｼﾞ取得ﾙｰﾙ		
		ﾙｰﾙ		主情報の種類
1	絵本	(1)	<title>ﾀｸﾞに対応する文字列を取得	ﾀｲﾄﾙ
		(2)	<frame src=" " >内のﾃﾞｰﾀを取得	絵
		(3)	<body>ﾀｸﾞ内の文字列を取得	文章
⋮	⋮	⋮	⋮	⋮

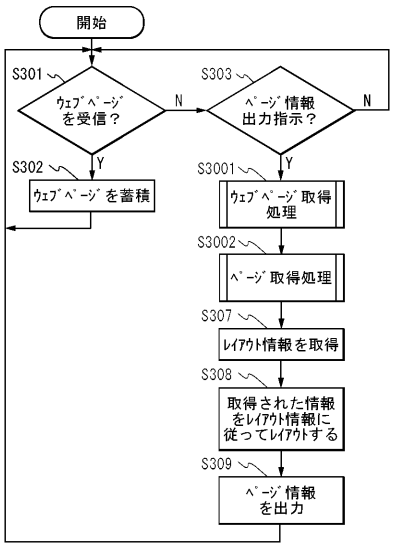
【図２６】

```
<html>
<head>
<title>元気なポチくん</title>
</head>
<meta http-equiv="Content-Type" content="text/html;
charset=EUC-JP">
<frameset rows="48,*" frameborder="1" border="1"
framespacing="1">
<frame
src="kghead.php?SY=2&MD=0&FM=0&BL=&TP=http://www.e-
hon.jp/kushar/kusj1.htm" scrolling="no" marginwidth="0"
marginheight="0" name="kidsgoo_header">
<frame
src="kgbody.php?SY=2&MD=0&FM=0&BL=&TP=http://www.e-
hon.jp/kushar/kusj1.htm" scrolling="auto" name="_kgbody">
</frameset>
<noframes>
<body bgcolor="#FFFFFF">
うちの愛犬のポチは、ほんとうに元気です。・・・・・・・・・・・・
</body>
</noframes>
</html>
```

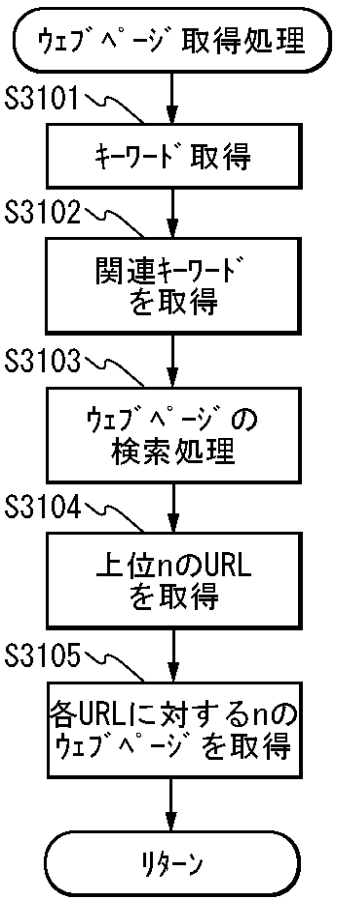
【図２９】



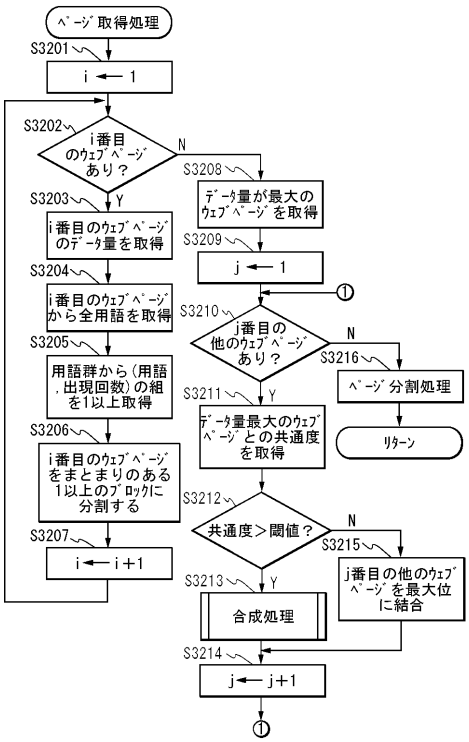
【図３０】



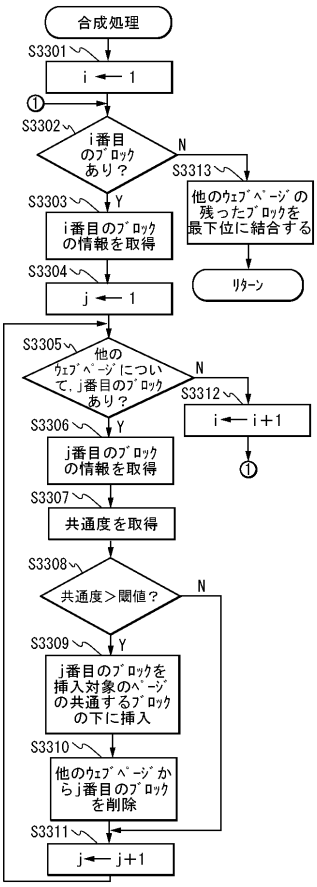
【図３１】



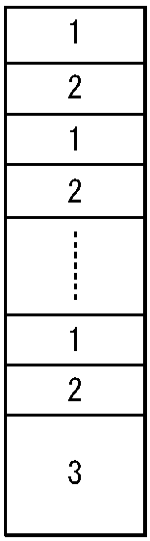
【図３２】



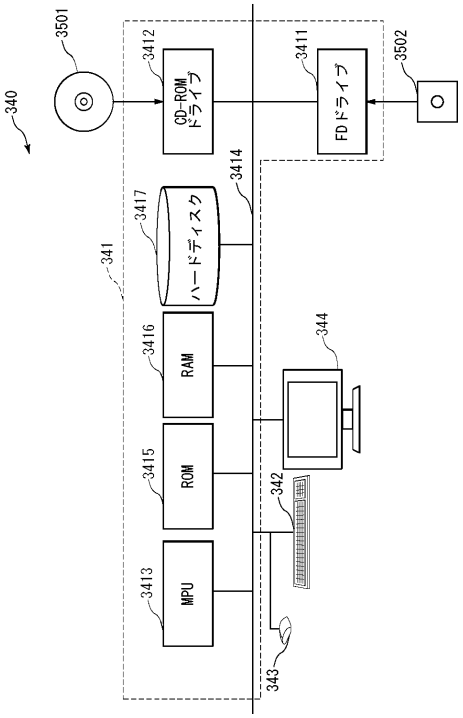
【図３３】



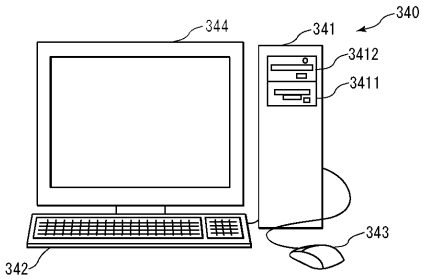
【図３５】



【図３８】



【図３７】



【図２５】

【図２７】

主情報の種類	情報
ﾀｲﾄﾙ	元気なポチくん
絵	
文章	うちの愛犬のポチは、ほんとうに元気です。・・・

【図２８】

【図３４】

:
表組

・¥begin{tabular}{l}:
　　項目を左揃え。
・¥begin{tabular}{c}:
　　項目を中央揃え。
・¥begin{tabular}{r}:
　　項目を右寄せ。
:
:

(a)

:
LaTeXで以下のような表を作ろうと思ったとき，容易に作成できるソフトウェアを開発しました。

理論値	実験値	
	室内	室外
0	0	0.5
10	10.5	9.5
100	90	110

:
「VisualTabular」といいます。

(b)

:
¥shortstack を使いまして
:
¥begin{tabular}{rrr}
... & ¥shortstack{line1¥¥ line2}
& ...
¥end{tabular}
:
:
:

(c)

【図３６】

:

表組

・¥begin{tabular}{l}:
　　項目を左揃え。
・¥begin{tabular}{c}:
　　項目を中央揃え。
・¥begin{tabular}{r}:
　　項目を右寄せ。

¥shortstack を使いまして

:

¥begin{tabular}{rrr}
... & ¥shortstack{line1¥¥ line2}
& ...
¥end{tabular}

:
:
:
:
:
:

LaTeXで以下のような表を
作ろうと思ったとき，容易
に作成できるソフトウェア
を開発しました。

理論値	実験値	
	室内	室外
0	0	0.5
10	10.5	9.5
100	90	110

:

「VisualTabular」といい
ます。

フロントページの続き

(72)発明者　安部　伸治
京都府相楽郡精華町光台二丁目２番地２　株式会社国際電気通信基礎技術研究所内

審査官　木村　雅也

(56)参考文献　特開２００６－２７９８８７（ＪＰ，Ａ）
山根　康宏，まるごとＮｏｋｉａ　Ｅ６１　初版，日本，株式会社技術評論社　片岡　巖，２００７年　9月１０日，第1版，第186頁

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１３／００
Ｇ０６Ｆ　１７／２１