JP 5823620 B2 2015.11.25

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

特許第5823620号

（Ｐ５８２３６２０）

(45)発行日　平成27年11月25日(2015.11.25)　(24)登録日　平成27年10月16日(2015.10.16)

(51)Int.Cl.　　ＦＩ

Ｇ０６Ｆ　17/30　(2006.01)　　Ｇ０６Ｆ　17/30　１８０Ｚ

Ｇ０６Ｆ　13/00　(2006.01)　　Ｇ０６Ｆ　17/30　１７０Ａ

Ｇ０６Ｆ　17/30　２１０Ｄ

Ｇ０６Ｆ　13/00　５４０Ｂ

請求項の数11　（全23頁）

(21)出願番号　特願2014-532241(P2014-532241)

(86)(22)出願日　平成24年12月13日(2012.12.13)

(65)公表番号　特表2014-528136(P2014-528136A)

(43)公表日　平成26年10月23日(2014.10.23)

(86)国際出願番号　PCT/CN2012/086584

(87)国際公開番号　WO2013/087012

(87)国際公開日　平成25年6月20日(2013.6.20)

審査請求日　平成26年3月28日(2014.3.28)

(31)優先権主張番号　201110415356.8

(32)優先日　平成23年12月13日(2011.12.13)

(33)優先権主張国　中国(CN)

(73)特許権者　507231932

北大方正集▲団▼有限公司

ＰＥＫＩＮＧ　ＵＮＩＶＥＲＳＩＴＹ　ＦＯＵＮＤＥＲ　ＧＲＯＵＰ　ＣＯ．，　ＬＴＤ

中華人民共和国北京市▲海▼淀区成府路２９８号中▲関▼村方正大厦５▲層▼

５　Ｆｌｏｏｒ，　Ｚｈｏｎｇｇｕａｎｃｕｎ　Ｆｏｕｎｄｅｒ　Ｂｕｉｌｄｉｎｇ，　Ｎｏ．２９８，　Ｃｈｅｎｇｆｕ　Ｒｏａｄ，　Ｈａｉｄｉａｎ　Ｄｉｓｔｒｉｃｔ，　Ｂｅｉｊｉｎｇ　１００８７１，　Ｃｈｉｎａ

最終頁に続く

(54)【発明の名称】ネットデータの採集方法及びシステム

(57)【特許請求の範囲】

【請求項１】

ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータを採集するためのネットデータの採集方法であって、

プロセッサーが採取対象のネットデータのウェブサイトのリンクアドレスに対応する種類に応じて、前記Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータが位置するウェブサイトのリンクアドレスである、前記採取対象のネットデータのウェブサイトのリンクアドレスを、対応する種類のキューに設置するステップと、

プロセッサーが前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得するステップと、

プロセッサーが前記ウェブサイトソースコードに対応するユニフォームリソースロケーターＵＲＬ情報及びＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するステップと

前記Ｍ個の主題における各主題は一つの文学作品であり、前記方法は、

【数１】

$$N_{Deep}=\begin{cases}\text{第１閾値} & \text{「名称→巻→章節→内容」のような作品構成を示す}\\ \text{第２閾値} & \text{「名称→章節→内容」のような作品構成を示す}\\ \text{第３閾値} & \text{「章節→内容」のような作品構成を示す}\end{cases}$$

のように、ネット文学の構成に応じて前記ＵＲＬの採集深度値を設置するステップとを備え、

前記採取対象のネットデータのウェブサイトリンクアドレスに対応する前記種類は、主題名称ページ、リストページ、及び内容ページを備え、プロセッサーが主題名称を抽出するように前記主題名称ページを設置し、主題章節目録又は主題章節を抽出するように前記リストページを設置して、主題本文内容を抽出するように前記内容ページを設置し、

前記採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置する前記ステップは、

プロセッサーが種類が主題名称であるリンクアドレスを主題名称ページキューに設置し、

プロセッサーが種類が前記リストページであるリンクアドレスをリストページキューに設置し、

プロセッサーが種類が前記内容ページであるリンクアドレスを内容ページキューに設置することであり、

前記Ｍは正の整数であり、

前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得するステップは、

プロセッサーが前記主題名称ページのキューから前記主題名称ページのリンクアドレスに対応するウェブサイトソースコードを取得することであり、

前記ウェブサイトソースコードに対応するＵＲＬ情報及び前記ＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するステップは、

プロセッサーが採集深度値が第１閾値である場合、主題の名称及び前記名称に対応するＵＲＬを抽出し、かつ、前記名称に対応するＵＲＬの採集深度値を第２閾値としてマークして前記リストページキューに追加し、

プロセッサーが採集深度値が第２閾値である場合、主題の名称及び前記名称に対応するＵＲＬを抽出し、かつ、前記名称に対応するＵＲＬの採集深度値を第３閾値としてマークして前記リストページキューに追加する

ことを特徴とするネットデータの採集方法。

【請求項２】

ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータを採集するためのネットデータの採集方法であって、

プロセッサーが採取対象のネットデータのウェブサイトのリンクアドレスに対応する種類に応じて、前記Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータが位置するウェブサイトのリンクアドレスである、前記採取対象のネットデータのウェブサイトのリンクアドレスを、対応する種類のキューに設置するステップと、

プロセッサーが前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得するステップと、

プロセッサーが前記ウェブサイトソースコードに対応するユニフォームリソースロケーターＵＲＬ情報及びＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するステップと

前記Ｍ個の主題における各主題は一つの文学作品であり、前記方法は、

【数２】

$$N_{Deep}=\begin{cases}\text{第１閾値} & \text{「名称→巻→章節→内容」のような作品構成を示す}\\ \text{第２閾値} & \text{「名称→章節→内容」のような作品構成を示す}\\ \text{第３閾値} & \text{「章節→内容」のような作品構成を示す}\end{cases}$$

のように、ネット文学の構成に応じて前記ＵＲＬの採集深度値を設置するステップとを備え、

前記採取対象のネットデータのウェブサイトリンクアドレスに対応する前記種類は、主

題名称ページ、リストページ、及び内容ページを備え、プロセッサーが主題名称を抽出するように前記主題名称ページを設置し、主題章節目録又は主題章節を抽出するように前記リストページを設置して、主題本文内容を抽出するように前記内容ページを設置し、

前記採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置する前記ステップは、

プロセッサーが種類が主題名称であるリンクアドレスを主題名称ページキューに設置し、

プロセッサーが種類が前記リストページであるリンクアドレスをリストページキューに設置し、

プロセッサーが種類が前記内容ページであるリンクアドレスを内容ページキューに設置することであり、

前記Ｍは正の整数であり、

前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得するステップは、

プロセッサーが前記リストページキューから前記リストページのリンクアドレスに対応するウェブサイトソースコードを取得することであり、

前記ウェブサイトソースコードに対応するＵＲＬ情報及び前記ＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するステップは、

プロセッサーが採集深度値が第２閾値である場合、主題の章節目録及び前記章節目録に対応するＵＲＬを抽出し、かつ、前記章節目録に対応するＵＲＬの採集深度値を第３閾値としてマークしてから前記リストページキューに追加し、

プロセッサーが採集深度値が第３閾値である場合、前記ウェブサイトソースコードに対応するＵＲＬには上位ＵＲＬが存在するか否かを判断し、

プロセッサーが存在すると判断する場合、主題の章節タイトル及び前記章節タイトルに対応する章節のＵＲＬを抽出し、かつ、前記章節のＵＲＬを前記内容ページキューに追加し、

プロセッサーが存在しないと判断する場合、主題の名称、主題の章節タイトル、及び前記章節タイトルに対応する章節のＵＲＬを抽出し、かつ、前記章節のＵＲＬを前記内容ページキューに追加する

ことを特徴とするネットデータの採集方法。

【請求項３】

ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータを採集するためのネットデータの採集方法であって、

プロセッサーが採取対象のネットデータのウェブサイトのリンクアドレスに対応する種類に応じて、前記Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータが位置するウェブサイトのリンクアドレスである、前記採取対象のネットデータのウェブサイトのリンクアドレスを、対応する種類のキューに設置するステップと、

プロセッサーが前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得するステップと、

プロセッサーが前記ウェブサイトソースコードに対応するユニフォームリソースロケーターＵＲＬ情報及びＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するステップと

前記Ｍ個の主題における各主題は一つの文学作品であり、前記方法は、

【数３】

$$N_{Deep}=\begin{cases}\text{第１閾値} & \text{「名称→巻→章節→内容」のような作品構成を示す}\\ \text{第２閾値} & \text{「名称→章節→内容」のような作品構成を示す}\\ \text{第３閾値} & \text{「章節→内容」のような作品構成を示す}\end{cases}$$

のように、ネット文学の構成に応じて前記ＵＲＬの採集深度値を設置するステップとを備

え、
前記採取対象のネットデータのウェブサイトリンクアドレスに対応する前記種類は、主題名称ページ、リストページ、及び内容ページを備え、プロセッサーが主題名称を抽出するように前記主題名称ページを設置し、主題章節目録又は主題章節を抽出するように前記リストページを設置して、主題本文内容を抽出するように前記内容ページを設置し、
前記採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置する前記ステップは、
プロセッサーが種類が主題名称であるリンクアドレスを主題名称ページキューに設置し、
プロセッサーが種類が前記リストページであるリンクアドレスをリストページキューに設置し、
プロセッサーが種類が前記内容ページであるリンクアドレスを内容ページキューに設置することであり、
前記Ｍは正の整数であり、
前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得するステップは、
プロセッサーが前記内容ページキューから前記内容ページのリンクアドレスに対応するウェブサイトソースコードを取得することであり、
前記ウェブサイトソースコードに対応するＵＲＬ情報及び前記ＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するステップは、
プロセッサーが前記ウェブサイトソースコードから主題の章節タイトル及び章節本文内容を抽出し、かつ、前記ウェブサイトソースコードに対応するＵＲＬから前記章節タイトルに対応する章節の章節ＩＤを抽出する
ことを特徴とするネットデータの採集方法。
【請求項４】
プロセッサーが前記ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントの更新頻度に応じてリフレッシュ時間間隔を設置するステップと、
プロセッサーが前記リフレッシュ時間間隔に基づいて前記採取対象のネットデータのウェブサイトリンクアドレスをリフレッシュするステップと
を更に備えることを特徴とする請求項１ないし請求項３のいずれか１つに記載の方法。
【請求項５】
プロセッサーが前記章節本文内容にはページングが存在する場合、次のページのリンクアドレスを抽出するとともに現在ページのページ番号及び次のページのページ番号をマークし、かつ、次のページのリンクアドレスを前記内容ページキューに追加して採集待機する
ことを特徴とする請求項３に記載の方法。
【請求項６】
プロセッサーが前記章節本文内容の第１ページのリンクを唯一のキー値として、前記ページングの内容を格納して、最後の１ページを採集終了する際に終了フラグを付与することを特徴とする請求項５に記載の方法。
【請求項７】
プロセッサーが抽出した全てのページングの内容を合併して、前記章節タイトルを結合して出力することを特徴とする請求項６に記載の方法。
【請求項８】
ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータを採集するためのネットデータの採集システムであって、
採取対象のネットデータのウェブサイトリンクアドレスに対応する種類に応じて、前記Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータが位置するウェブサイトのリンクアドレスである、採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置することに用いられる設置モジュールと、

前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトリンクアドレスに対応するウェブサイトソースコードを取得することに用いられるウェブサイト取得モジュールと、
前記ウェブサイトソースコードに対応するユニフォームリソースロケーターＵＲＬ情報及びＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するデータ抽出モジュールと
を備え、
前記Ｍは正の整数であり、
前記採取対象のネットデータのウェブサイトリンクアドレスに対応する種類は、主題名称ページ、リストページ、及び内容ページを備え、
前記設置モジュールは、主題名称を抽出するように前記主題名称ページを設置し、主題章節目録又は主題章節を抽出するように前記リストページを設置し、主題内容を抽出するように前記内容ページを設置することに用いられるウェブサイト設置モジュールと、
前記採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置するためのキュー設置モジュールを備え、
前記キュー設置モジュールは、
種類が前記主題名称であるリンクアドレスを主題名称ページキューに設置する第１設置手段と、
種類が前記リストページであるリンクアドレスをリストページキューに設置する第２設置手段と、
種類が前記内容ページであるリンクアドレスを内容ページキューに設置する第３設置手段とを備え、
前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得することは、
プロセッサーが前記主題名称ページのキューから前記主題名称ページのリンクアドレスに対応するウェブサイトソースコードを取得することであり、
前記ウェブサイトソースコードに対応するＵＲＬ情報及び前記ＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出することは、
プロセッサーが採集深度値が第１閾値である場合、主題の名称及び前記名称に対応するＵＲＬを抽出し、かつ、前記名称に対応するＵＲＬの採集深度値を第２閾値としてマークして前記リストページキューに追加し、
プロセッサーが採集深度値が第２閾値である場合、主題の名称及び前記名称に対応するＵＲＬを抽出し、かつ、前記名称に対応するＵＲＬの採集深度値を第３閾値としてマークして前記リストページキューに追加する
ことを特徴とするネットデータの採集システム。
【請求項９】
ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータを採集するためのネットデータの採集システムであって、
採取対象のネットデータのウェブサイトリンクアドレスに対応する種類に応じて、前記Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータが位置するウェブサイトのリンクアドレスである、採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置することに用いられる設置モジュールと、
前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトリンクアドレスに対応するウェブサイトソースコードを取得することに用いられるウェブサイト取得モジュールと、
前記ウェブサイトソースコードに対応するユニフォームリソースロケーターＵＲＬ情報及びＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するデータ抽出モジュールと
を備え、
前記Ｍは正の整数であり、

前記採取対象のネットデータのウェブサイトリンクアドレスに対応する種類は、主題名称ページ、リストページ、及び内容ページを備え、
前記設置モジュールは、主題名称を抽出するように前記主題名称ページを設置し、主題章節目録又は主題章節を抽出するように前記リストページを設置し、主題内容を抽出するように前記内容ページを設置することに用いられるウェブサイト設置モジュールと、
前記採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置するためのキュー設置モジュールを備え、
前記キュー設置モジュールは、
種類が前記主題名称であるリンクアドレスを主題名称ページキューに設置する第１設置手段と、
種類が前記リストページであるリンクアドレスをリストページキューに設置する第２設置手段と、
種類が前記内容ページであるリンクアドレスを内容ページキューに設置する第３設置手段とを備え、
前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得することは、
プロセッサーが前記リストページキューから前記リストページのリンクアドレスに対応するウェブサイトソースコードを取得することであり、
前記ウェブサイトソースコードに対応するＵＲＬ情報及び前記ＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出することは、
プロセッサーが採集深度値が第２閾値である場合、主題の章節目録及び前記章節目録に対応するＵＲＬを抽出し、かつ、前記章節目録に対応するＵＲＬの採集深度値を第３閾値としてマークしてから前記リストページキューに追加し、
プロセッサーが採集深度値が第３閾値である場合、前記ウェブサイトソースコードに対応するＵＲＬには上位ＵＲＬが存在するか否かを判断し、
プロセッサーが存在すると判断する場合、主題の章節タイトル及び前記章節タイトルに対応する章節のＵＲＬを抽出し、かつ、前記章節のＵＲＬを前記内容ページキューに追加し、
プロセッサーが存在しないと判断する場合、主題の名称、主題の章節タイトル、及び前記章節タイトルに対応する章節のＵＲＬを抽出し、かつ、前記章節のＵＲＬを前記内容ページキューに追加する
ことを特徴とするネットデータの採集システム。
【請求項１０】
ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータを採集するためのネットデータの採集システムであって、
採取対象のネットデータのウェブサイトリンクアドレスに対応する種類に応じて、前記Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータが位置するウェブサイトのリンクアドレスである、採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置することに用いられる設置モジュールと、
前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトリンクアドレスに対応するウェブサイトソースコードを取得することに用いられるウェブサイト取得モジュールと、
前記ウェブサイトソースコードに対応するユニフォームリソースロケーターＵＲＬ情報及びＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するデータ抽出モジュールと
を備え、
前記Ｍは正の整数であり、
前記採取対象のネットデータのウェブサイトリンクアドレスに対応する種類は、主題名称ページ、リストページ、及び内容ページを備え、
前記設置モジュールは、主題名称を抽出するように前記主題名称ページを設置し、主題

章節目録又は主題章節を抽出するように前記リストページを設置し、主題内容を抽出するように前記内容ページを設置することに用いられるウェブサイト設置モジュールと、
前記採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置するためのキュー設置モジュールを備え、
前記キュー設置モジュールは、
種類が前記主題名称であるリンクアドレスを主題名称ページキューに設置する第１設置手段と、
種類が前記リストページであるリンクアドレスをリストページキューに設置する第２設置手段と、
種類が前記内容ページであるリンクアドレスを内容ページキューに設置する第３設置手段とを備え、
前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得することは、
プロセッサーが前記内容ページキューから前記内容ページのリンクアドレスに対応するウェブサイトソースコードを取得することであり、
前記ウェブサイトソースコードに対応するＵＲＬ情報及び前記ＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出することは、
プロセッサーが前記ウェブサイトソースコードから主題の章節タイトル及び章節本文内容を抽出し、かつ、前記ウェブサイトソースコードに対応するＵＲＬから前記章節タイトルに対応する章節の章節ＩＤを抽出する
ことを特徴とするネットデータの採集システム。
【請求項１１】
前記システムは、
前記ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントの更新頻度に応じてリフレッシュ時間間隔を設置し、かつ、前記リフレッシュ時間間隔に基づいて、前記採取対象のネットデータのウェブサイトリンクアドレスをリフレッシュすることに用いられるリフレッシュモジュールを備える
ことを特徴とする請求項８ないし１０のいずれか１つに記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本出願は、２０１１年１２月１３日に中国特許局に提出し、出願番号が２０１１１０４１５３５６．８であり、発明名称が「ネットデータの採集方法及びシステム」との中国特許出願を基礎とする優先権を主張し、その全文の内容を引用することにより本出願に取り込む。
本発明は情報検索及びデータ集積の技術分野に関し、特に、ネットデータの採集方法及びシステムに関する。
【背景技術】
【０００２】
インターネットの現れ及び普及に伴い、インターネットは数億のネットワークのユーザに様々な文学資料情報を提供した。その同時に、伝統文学特徴と異なるネット文学は、新しい文学媒体とし、ネットワークのユーザを読書対象とし、盛んになっている。
【０００３】
ネット文学は、近頃に現われた、ネットワークを展示台として、ハイパーテキストリンクとマルチメディアプレゼンテーション等の手段により表現される文学作品、文学と類似する類似文学作品、及び一部が文学要素が含まれるネットワーク芸術を意味する。そのうち、オリジナルネットワーク作品を中心としている。
ネット文学は、以下のような３種類に分けられてもよい。
第１種類のネット文学は、既に公表した文学作品を電子走査技術又はマニュアル入力により形成されたデジタルリソースである。

第２種類のネット文学は、直接にインターネットで「公開発表」した文学作品である。
第３種類のネット文学は、コンピュータにより作成されたか、又はコンピュータソフトウェアにより生成された文学作品がインターネットで発表され、インターネット開放性に基づいて、数人、数十人乃至数百人の作家により、協力で作成した「リレー小説」等である。

【発明の概要】
【発明が解決しようとする課題】
【０００４】
ネット文学の発展に伴う出版権、文学作成内容等の問題を直面しなければならない。ネット文学の関連データの支援がいないため、ネット文学の最新内容を簡単且つ集中的にブラウズできないし、ネット文学に対する検索又はモニタリングを実現することができない。
【０００５】
本発明は、最新ネットデータをリアルタイムに採集できるネットデータの採集方法及びシステムを提供することに目的とする。
【課題を解決するための手段】
【０００６】
本発明１仕様によれば、ウェブサイトで公開した、Ｍ個（Ｍが正の整数である）の主題とそれぞれ関連するオンライン・ドキュメントのデータを採集する、ネットデータ採取方法が提供されている。
当該公表方法は、採取対象のネットデータのウェブサイトリンクアドレスに対応する種類に応じて、前記採取対象のネットデータのウェブサイトリンクアドレスに対応する種類のキューに設置するステップと、前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトリンクアドレスに対応するウェブサイトソースコードを取得するステップと、前記ウェブサイトソースコードに対応するユニフォームリソースロケーター（ＵＲＬ）情報及びＵＲＬの採集深度値（Ｃｏｌｌｅｃｔｉｏｎ ｄｅｐｔｈ ｖａｌｕｅｓ ｏｆ ｔｈｅ ＵＲＬｓ．）に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するステップと、を備える。
前記採取対象のネットデータのウェブサイトリンクアドレスは、前記Ｍ個の主題とそれぞれ関連するオンライン・ドキュメントのデータが所在するウェブサイトアドレスである。
【０００７】
好ましくは、前記ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントの更新頻度に応じてリフレッシュ時間間隔を設置し、また、前記リフレッシュ時間間隔に基づいて前記採取対象のネットデータのウェブサイトリンクアドレスをリフレッシュする。
【０００８】
好ましくは、前記Ｍ個の主題のうちのいずれも文学作品であり、前記方法は、前記ネット文学の構成に応じて前記ＵＲＬの採集深度値を設置するステップを更に備える。具体的に、以下のように示されている。
【０００９】
【数１】

$$N_{Deep}=\begin{cases}\text{第１閾値} & \text{「名称→巻→章節→内容」のような作品構成を示す}\\ \text{第２閾値} & \text{「名称→章節→内容」のような作品構成を示す}\\ \text{第３閾値} & \text{「章節→内容」のような作品構成を示す}\end{cases}$$

【００１０】
好ましくは、前記採取対象のネットデータのウェブサイトリンクアドレスに対応する種類は、主題名称ページ、リストページ、及び内容ページを備える。主題名称を抽出できる

ように前記主題名称ページを設置し、主題の章節目録又は主題章節を抽出できるように前記リストページを設置して、主題の本文内容を抽出できるように前記内容ページを設置する。

【００１１】

好ましくは、前記採取対象のネットデータのウェブサイトリンクアドレスを対応種類のキューに設置する前記ステップは、具体的には、種類が主題名称であるリンクアドレスを主題名称ページキューに追加し、種類が前記リストページであるリンクアドレスをリストページキューに追加し、種類が前記内容ページであるリンクアドレスを内容ページキューに追加することを含む。

【００１２】

好ましくは、前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトリンクアドレスに対応するウェブサイトソースコードを取得するステップは、具体的には、前記主題名称ページのキューから前記主題名称ページリンクアドレスに対応するウェブサイトソースコードを取得する。

【００１３】

好ましくは、前記ウェブサイトソースコードに対応するＵＲＬ情報及び前記ＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するステップは、具体的には、採集深度値が第１閾値である場合、主題の名称及び前記名称に対応するＵＲＬを抽出し、かつ、前記名称に対応するＵＲＬの採集深度値を第２閾値としてマークして前記リストページキューに追加し、採集深度値が第２閾値である場合、主題の名称及び前記名称に対応するＵＲＬを抽出し、かつ、前記名称に対応するＵＲＬの採集深度値を第３閾値としてマークして前記リストページキューに追加する。

【００１４】

好ましくは、前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトリンクアドレスに対応するウェブサイトソースコードを取得するステップは、具体的には、前記リストページキューから前記リストページリンクアドレスに対応するウェブサイトソースコードを取得する。

【００１５】

好ましくは、前記ウェブサイトソースコードに対応するＵＲＬ情報及び前記ＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するステップは、具体的には、採集深度値が第２閾値である場合、主題の章節目録及び前記章節目録に対応するＵＲＬを抽出し、かつ、前記章節目録に対応するＵＲＬの採集深度値を第３閾値としてマークして前記リストページキューに追加し、また、採集深度値が第３閾値である場合、前記ウェブサイトソースコードに対応するＵＲＬには上級ＵＲＬが存在するか否かを判断して、存在すると判断する場合、主題の章節タイトル及び前記章節タイトルに対応する章節のＵＲＬを抽出し、かつ、前記章節のＵＲＬを前記内容ページキューに追加して、存在しないと判断する場合、主題の名称、主題の章節タイトル、及び前記章節タイトルに対応する章節のＵＲＬを抽出し、かつ、前記章節のＵＲＬを前記内容ページキューに追加する。

【００１６】

好ましくは、前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクのアドレスに対応するウェブサイトソースコードを取得するステップは、具体的には、前記内容ページキューから前記内容ページのリンクアドレスに対応するウェブサイトソースコードを取得する。

【００１７】

好ましくは、前記ウェブサイトソースコードに対応するＵＲＬ情報及び前記ＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出するステップは、具体的には、前記ウェブサイトソースコードから主題の章節タイトル及び章節本文内容を抽出し、かつ、前記ウェブサイトソースコードに対応するＵＲＬから前記章節タイトルに対応する章節の章節ＩＤを抽出する。

【００１８】
好ましくは、前記章節本文内容がページングされているか否かことを判断して、ページングされたと判断された場合、次のページのリンクアドレスを抽出するとともに現在ページのページ番号及び次のページのページ番号をマークし、かつ、次のページのリンクアドレスを前記内容ページキューに追加して採集を待機する。
【００１９】
好ましくは、前記章節の本文内容の第１ページのリンクを唯一のキー値として、前記ページングの内容を格納して、最後の１ページの採集了の際に終了フラグ（Ｅｎｄ　ｆｌａｇ）を付ける。
【００２０】
好ましくは、抽出した全てのページングの内容を合併して、前記章節のタイトルを結合して出力する。
【００２１】
本発明の他方側面では、ウェブサイトに公表された、Ｍ個（Ｍが正の整数である）の主題のそれぞれに関連するオンライン・ドキュメントのデータを採集するためのネットデータの採集システムを提供している。前記システムは、設置モジュール、ウェブサイト取得モジュール、及びデータ抽出モジュールを備える。前記設置モジュールは、採取対象のネットデータのウェブサイトリンクアドレスに対応する種類に応じて、前記Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータが位置するウェブサイトのリンクアドレスである、採取対象のネットデータのウェブサイトリンクアドレスを、対応する種類のキューに設置することに用いられる。ウェブサイト取得モジュールは、前記対応する種類のキューにおける、前記採取対象のネットデータのウェブサイトリンクアドレスに対応するウェブサイトソースコードを取得することに用いられる。データ抽出モジュールは、前記ウェブサイトソースコードに対応するユニフォームリソースロケーターＵＲＬ情報及びＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出する。
【００２２】
好ましくは、前記システムは更にリフレッシュモジュールを備える。前記リフレッシュモジュールは、前記ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントの更新頻度に応じてリフレッシュ時間間隔を設置し、かつ、前記リフレッシュ時間間隔に基づいて、前記採取対象のネットデータのウェブサイトリンクアドレスをリフレッシュすることに用いられる。
【００２３】
好ましくは、前記採取対象のネットデータのウェブサイトリンクアドレスに対応する種類は、主題名称ページ、リストページ、及び内容ページを備える。前記設置モジュールは、主題名称を抽出するように前記主題名称ページを設置し、主題章節目録又は主題章節を抽出するように前記リストページを設置し、主題内容を抽出するように前記内容ページを設置することに用いられるウェブサイト設置モジュールを備える。
【００２４】
好ましくは、前記設置モジュールは、更に、前記採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置するためのキュー設置モジュールを備える。前記キュー設置モジュールは更に、種類が前記主題名称であるリンクアドレスを、主題名称ページキューに設置第１設置手段と、種類が前記リストページであるリンクアドレスをリストページキューに設置第２設置手段と、種類が前記内容ページであるリンクアドレスを内容ページキューに設置第３設置手段と、を備える。
【発明の効果】
【００２５】
本発明の有益な効果は以下のとおりである。
本発明に係る１つの実施例は、ネットデータの採集システムでネットデータの採集を行い、システムがネットデータのリンクアドレスを取得してリンクアドレスの種類を設置し

、かつ、リンクアドレスの種類に応じてリンクアドレスを対応するキューに追加する。キューからリンクアドレスに対応するソースコードを取得し、ソースコードにおける対応するＵＲＬ情報及びＵＲＬの採集深度値に応じて、ネットデータの情報を抽出することにより、リアルタイムのネットデータを採集する技術効果が得られる。

【００２６】

さらに、本発明は同一の主題に属するオンライン・ドキュメントを合併することができる内容合併モジュールを利用するため、リアルタイムのネットデータを採集する上に、便利にまとめにブラウジングする効果が得られる。

【図面の簡単な説明】

【００２７】

【図１】本発明に係る１つの実施例における採集方法のフローチャートである。

【図２】本発明の図１における採集方法のフローチャートである。

【図３】本発明に係る第１実施例の採集システムの構成図である。

【図４】本発明に係る１つの実施例における設置モジュールの構成図である。

【図５】本発明に係る１つの実施例におけるウェブサイト取得モジュールの構成図である。

【図６】本発明に係る１つの実施例におけるデータ抽出モジュールの構成図である。

【図７】本発明に係る第２の実施例の採集システムの構成図である。

【図８】本発明に係る第３の実施例の採集システムの構成図である。

【図９】本発明に係る第４の実施例の採集システムの構成図である。

【発明を実施するための形態】

【００２８】

以下、当業者が本発明をもっと明瞭かつ完全に理解できるように、図面を結合しながら本発明を詳細に説明する。

【００２９】

本発明に係る１つの実施例は、ウェブサイトで公表した、Ｍ個（Ｍが正の整数である）の主題のそれぞれに関連するオンライン・ドキュメントのデータを採集するためのネットデータの採集方法を提供している。図面１に示すように、図面１は本発明に係る１つの実施例における採集方法のフローチャートである。図面１に示すように、当該データの採集方法は、ステップ１１と、ステップ１２と、ステップ１３とを備える。

【００３０】

ステップ１１：採取対象のネットデータのウェブサイトのリンクアドレスに対応する種類に応じて、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータが位置するウェブサイトのリンクアドレスである採取対象のネットデータのウェブサイトのリンクアドレスを、対応する種類のキューに設置する。

【００３１】

ステップ１２：前記対応する種類のキューにおける前記採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得する。

【００３２】

ステップ１３：前記ウェブサイトソースコードに対応するユニフォームリソースロケーター（Ｕｎｉｆｏｒｍ　Ｒｅｓｏｕｒｃｅ　Ｌｏｃａｔｏｒ，　ＵＲＬ）情報及び前記ＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するオンライン・ドキュメントのデータを抽出する。

【００３３】

ステップ１１において、ウェブサイトで公表したＭ個の主題は、Ｍ個のネット文学作品であってもよい。本発明を理解し易くするために、以下の実施例はネット文学を例としているが、ネット文学に限られない。ネット文学は、例えばネットニュース等の主題と異なる公表構成を有し、普通なネットニュースは単なる短文であるが、ネット文学作品は一般には２つの形態にてウェブサイトで公表される。その１つは小説閲読ウェブサイトの「文学名称－＞章節目録ページ－＞具体的なある章節のネット文学内容ページ」に類似するも

のであり、その他方は普通なニュースウェブサイトの内容目録ホームページに類似するものである。異なる文学作品の章節は交互混合して同一ページに配置される場合もあるが、タイトルに「文学作品名称（５）」のように明記することにより、同一作品における異なる章節を区別する。

【００３４】

異なる構成のネット文学内容のオンライン・ドキュメントに対して採集するには、まず、オンライン・ドキュメントのデータが位置するホームページのリンクアドレスを取得すべきである。本実施例は、ネット文学内容がウェブサイトにおいて公表された構成に基づいて採集する。オンライン・ドキュメントのデータは、一般には、オンライン・ドキュメントが属するネット文学作品の名称と、オンライン・ドキュメントが属するネット文学作品におけるボリューム及び／又は章節の名称と、オンライン・ドキュメントの本文の内容と、を備える。それに対応して、オンライン・ドキュメントのデータが位置するホームページのリンクアドレスに対応する種類は、オンライン・ドキュメントが属するネット文学作品の名称を抽出するための主題名称ページと、ネット文学作品におけるネット文学のボリューム目録と章目録を備える章節目録のリンク及び章節のリンクを抽出するためのリストページと、主題本文の内容を抽出するための内容ページと、を備える。

【００３５】

本実施例において、Ｍ個のネット文学のデータが位置するホームページのリンクアドレスはそれの種類に応じてそれぞれに異なるキューに追加する。具体的には、種類が主題名称ページであるリンクアドレスは主題名称ページキューに設置され、種類がリストページであるリンクアドレスはリストページキューに設置され、種類が内容ページであるリンクアドレスは内容ページキューに設置される。例えば、Ａホームページには、三つのネット文学作品が公表され、それぞれがＡ１、Ａ２、Ａ３である。その中に、Ａ１のホームページＡにおける公表構成は、「文学名称－＞ボリューム目録－＞章目録－＞具体的なある章節のネット文学内容ページ」である。Ａ２のホームページＡにおける公表構成は、「文学名称－＞章目録－＞具体的なある章節のネット文学内容ページ」である。Ａ３のホームページＡにおける公表構成は、「章名称－＞具体的なある章節のネット文学内容ページ」である。Ａ３の章名称はＡ３の作品名称と章番号の組み合わせものである。例えば、Ａ３の第１章の章名称はＡ３（一）であり、Ａ３の第５章の章名称はＡ３（五）である。ホームページＡに対する毎回採集の開始の際に、Ａ１作品の名称を有するホームページのリンクアドレスＢ１を、主題名称ページキューに追加し、Ａ２作品の名称を有するホームページのリンクアドレスＢ２を主題名称ページキューに追加し、Ａ３作品の名称を有するホームページのリンクアドレスＢ３を主題名称ページキューに追加して、データ採集を待機する。一方、内容ページキューについて、採集開始の際に、採取対象のリンクアドレスの追加は、行わない。

【００３６】

実際の採集過程において、オンライン・ドキュメントは定期的にアップデートされるが、アップデートの頻度はネットニュースとフォーラム情報のように高くないため、定期的にリフレッシュする対策を採用することができ、もちろん、自己適応にリフレッシュする対策を採用することもできる。即ち、ホームページは、異なるネット文学作品の公表頻度に応じて、リフレッシュ間隔を自動的に調整する。ネット文学作品のリフレッシュ間隔時間になることを検出した場合に、リフレッシュした、採取対象のネットデータのホームページのリンクアドレスを、対応する種類のキューに追加する。

【００３７】

ステップ１２において、各キューにおける採取対象のネットデータのウェブサイトのリンクアドレスに対応するウェブサイトソースコードを取得することは、具体的には、システムにより設定されたＵＲＬの取得対策に応じて、例えば、システム運転の状況又は各キューの状況に応じて、当業者は実際に操作する際に必要な時間によってＵＲＬの取得対策を設定して、各キューから１つの採取対象のリンクアドレスを取得する。そして、システムはＨｔｔｐ請求によってウェブサイトソースコードを取得する。本実施例では、例えば

、ホームページＡ上の三つのネット文学作品に対する採集を開始する際に、主題名称ページキューから採取対象のネットデータのウェブサイトリンクアドレスＢ１，Ｂ２を抽出して、システムにより設定されたＵＲＬの取得対策に応じて、Ｂ１に対応するウェブサイトソースコード及びＢ２に対応するウェブサイトソースコードをそれぞれに取得し、リストページキューから採取対象のネットデータのウェブサイトリンクアドレスＢ３を抽出し、かつ、システムにより設定されたＵＲＬの取得対策に応じて、そのウェブサイトソースコードを取得する。

【００３８】

ステップ１３において、ウェブサイトソースコードに対応するＵＲＬ情報は、ネット文学作品名称、章節目録と章節リンク、及び本文内容のリンクを備える。ＵＲＬの採集深度値はネット文学作品の構成設置に応じて設置される。具体的には、以下のように

【００３９】

【数２】

$$N_{Deep}=\begin{cases}\text{第１閾値} & \text{「名称→巻→章節→内容」のような作品構成を示す}\\ \text{第２閾値} & \text{「名称→章節→内容」のような作品構成を示す}\\ \text{第３閾値} & \text{「章節→内容」のような作品構成を示す}\end{cases}$$

【００４０】

本実施例では、第１閾値を３とし、第２閾値を２とし、第３閾値を１とするが、当業者は他の数値又はマークで異なる閾値を示してもよい。

以下、本発明を説明し易くするために、第１閾値を３とし、第２閾値を２とし、第３閾値を１とする例を挙げて説明する。そして、ネット文学作品の構成設置の採集深度値に従って、ウェブサイトに公表されたＡ１、Ａ２、Ａ３を結合して本発明を理解することができる。

主題名称ページキューからリンクアドレスを取得した後、Ｂ１に対応するソースコードに応じて対応するＵＲＬ（ＵＲＬ－Ａ１）を取得するが、Ａ１の構成が「文学名称－＞ボリューム目録－＞章目録－＞具体的なある章節のネット文学内容ページ」であるため、ＵＲＬ－Ａ１の採集深度値は３であるべきである。

同様に、Ａ２の構成は「文学名称－＞章目録－＞具体的なある章節のネット文学内容ページ」であるため、Ｂ２に応じて取得したソースコードに対応するＵＲＬ（ＵＲＬ－Ａ２）の採集深度値は２である。

Ａ３の構成は「章目録－＞具体的なある章節のネット文学内容ページ」であるため、Ｂ３に応じて取得したソースコードに対応するＵＲＬ（ＵＲＬ－Ａ３）の採集深度値は３である。

【００４１】

詳しくは、ステップ１３は、ステップ１３１と、ステップ１３２と、ステップ１３３とを備える（図３を参照）。

【００４２】

ステップ１３１：主題名称ページキューから取得した、主題名称ページのリンクアドレスに対応するウェブサイトソースコードに対応するＵＲＬ情報及びＵＲＬ採集深度値に応じて、ＵＲＬに対応するオンライン・ドキュメントのデータを抽出する。

【００４３】

ステップ１３２：リストページキューから取得した、リストページのリンクアドレスに対応するウェブサイトソースコードに対応するＵＲＬ情報及びＵＲＬ採集深度値に応じて、ＵＲＬに対応するオンライン・ドキュメントのデータを抽出する。

【００４４】

ステップ１３３：内容ページキューから取得した、内容ページのリンクアドレスに対応するウェブサイトソースコードに対応するＵＲＬに応じて、ウェブサイトソースコードから主題の章節タイトルと章節本文内容を抽出し、かつ、ウェブサイトソースコードに対応

するＵＲＬから前記章節タイトルに対応する章節の章節ＩＤを抽出する。
【００４５】
上記ステップ１３１、１３２、１３３は、実行の際に順番が特定されていない。各キューにおいて採集する必要があるリンクアドレスがあれば、採取対象のリンクアドレスに対して採集を行い、採取対象のネットワークのホームページリンクアドレスに対応するウェブサイトソースコードを取得し、かつ、ウェブサイトソースコードに対応するＵＲＬ情報及びＵＲＬ採集深度値に応じて、ＵＲＬに対応するオンライン・ドキュメントのデータを抽出する。以下、各ステップにおいてオンライン・ドキュメントのデータに対する抽出する過程を詳細に説明する。
【００４６】
ステップ１３１においてＵＲＬに対応するオンライン・ドキュメントのデータを抽出することは、以下で具体的に説明する。
【００４７】
ＵＲＬの採集深度値が３である場合、主題の名称及び該名称に対応するＵＲＬを抽出し、かつ、該名称に対応するＵＲＬの採集深度値を第２閾値としてマークしてリストページキューに追加する。
【００４８】
ＵＲＬの採集深度値が２である場合、主題の名称及び該名称に対応するＵＲＬを抽出し、かつ、該名称に対応するＵＲＬの採集深度値を１としてマークしてリストページキューに追加する。
【００４９】
本実施例において、主題名称ページキューから抽出したリンクアドレスは、Ａ１のリンクアドレスＢ１及びＡ２のリンクアドレスＢ２である。Ｂ１に対応するソースコードに対応するＵＲＬ－Ａ１の採集深度値は３であるため、抽出すべきＡ１の主題名称を、「名称Ａ１」で示す。さらに、「名称Ａ１」に対応するＵＲＬも抽出すべき、「ＵＲＬ－Ａ１１」で示し、「ＵＲＬ－Ａ１１」の採集深度値を２にマークしてリストページキューに追加し、これにより、ＵＲＬ－Ａ１１における作品Ａ１に属する他の情報を抽出する。
リンクアドレスＢ２は、ＵＲＬ－Ａ２の採集深度値が２であるため、抽出すべきＡ２の主題名称を、「名称Ａ２」で示す。さらに、「名称Ａ２」に対応するＵＲＬも抽出しべき、「ＵＲＬ－Ａ２１」で示し、「ＵＲＬ－Ａ２１」の採集深度値を１にマークしてリストページキューに追加し、これにより、ＵＲＬ－Ａ２１における作品Ａ２に属する他の情報を抽出する。
【００５０】
ステップ１３２において、ＵＲＬに対応するオンライン・ドキュメントのデータを抽出することは、以下で、詳細に説明する。
【００５１】
ＵＲＬの採集深度値が２である場合、主題の章節目録及び該章節目録に対応するＵＲＬを抽出するとともに、該章節目録に対応するＵＲＬの採集深度値を１にマークしてリストページキューに追加する。
【００５２】
ＵＲＬの採集深度値が１である場合、ウェブサイトソースコードに対応するＵＲＬに上級ＵＲＬが存在するか否かを判断する。
【００５３】
存在すると判断する場合、主題の章節タイトル及び該章節タイトルに対応する章節のＵＲＬを抽出し、かつ、該章節のＵＲＬを内容ページキューに追加する。
【００５４】
存在しないと判断する場合、主題の名称、主題の章節タイトル、及び該章節タイトルに対応する章節のＵＲＬを抽出し、かつ、該章節のＵＲＬを内容ページキューに追加する。
【００５５】
本実施例では、リストページキューには、ステップ１３１により、採取対象のＵＲＬ－

Ａ１１及びＵＲＬ－Ａ２１が既に格納されている。また、ホームページＡ１に対するネット文学採集開始の際に作品Ａ３に対応するリンクアドレスＢ３をリストページキューに、既に追加した。

【００５６】

ＵＲＬ－Ａ１１は、その採集深度値が２であるため、Ａ１の章節目録及び該章節目録に対応するＵＲＬを抽出して、ＵＲＬ－Ａ１２で示す。ＵＲＬ－Ａ１２の採集深度値を１にマークしてリストページキューに追加する。

【００５７】

ＵＲＬ－Ａ２１は、その採集深度値が１であり、かつ上級ＵＲＬ（及びＵＲＬ－Ａ２１）を含む場合、Ａ２の章節タイトル及び該章節タイトルに対応する章節のＵＲＬを抽出して、ＵＲＬ－Ａ２２で示し、かつ、ＵＲＬ－Ａ２２を内容リストキューに追加する。

【００５８】

リストページキューＢ３は、Ｂ３に対応するソースコードに対応するＵＲＬ－Ａ３の採集深度値が１であり、かつ、上級ＵＲＬを含めないため、Ａ３の名称を抽出して、章節タイトルを「名称Ａ３」で示す。さらに、章節タイトルに対応するＵＲＬも抽出すべきであり、「ＵＲＬ－Ａ31」で示し、ＵＲＬ－Ａ３１を内容ページキューに追加する。

【００５９】

ステップ１３３において、章節の本文はページングされる場合、次のページのリンクアドレスを抽出すべきであり、同時に現在ページのページ番号及び次のページのページ番号をマークし、かつ、次のページのリンクアドレスを内容ページキューに追加して採集待機する。

【００６０】

さらに、章節の本文内容の第１ページのリンクを唯一のキー値とし、ページングの内容を格納する。最後の１ページを採集終了する際に、終了フラグを付ける。

【００６１】

さらに、抽出した全てのページングの内容を合併して、章節のタイトルを結合して出力してもよい。

【００６２】

さらに、ホームページ、主題の名称、主題の章節タイトル、章節ＩＤ、及び章節本文内容をデータベースにアップロードする。また、章節本文の内容を添付ファイルの形態にてファイルサーバに記憶して、ファイル記憶パスをデータベースに記録してもよい。

【００６３】

本実施例において、ネットデータに対する採集と合併の方法は、ネット文学を書籍形態であらわすことができる。さらに、採集データを自動的にリフレッシュすることで、データのリアルタイムの採集を実現することができるため、本実施例は、リアルタイムで、ネット文学作品を便利かつ集中的にブラウジングするという有益な効果が得られる。

【００６４】

本発明の第１実施例はウェブサイトで公表した、Ｍ個（Ｍが正の整数である）の主題のそれぞれに関連するオンライン・ドキュメントのデータを採集するためのネットデータの採集方法を提供している。

図３に示すように、データを採集するシステムは、設置モジュール３１、ウェブサイト取得モジュール３２、及びデータ抽出モジュール３３を備える。

設置モジュール３１は、採取対象のネットデータのウェブサイトリンクアドレスに対応する種類に応じて、採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置することに用いられる。

採取対象のネットデータのウェブサイトリンクアドレスは、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントのデータが位置するウェブサイトのリンクアドレスである。

【００６５】

ウェブサイト取得モジュール３２は、対応する種類のキューにおける採取対象のネット

データのウェブサイトリンクアドレスに対応するウェブサイトソースコードを取得することに用いられる。

データ抽出モジュール３３はウェブサイトソースコードに対応するＵＲＬ情報及びＵＲＬの採集深度値に応じて、ＵＲＬに対応するオンライン・ドキュメントのデータを抽出する。

【００６６】

本実施例では、採取対象のネットデータのウェブサイトリンクアドレスに対応する種類は、主題名称ページ、リストページ、及び内容ページを備える。

図４にしめすように、設置モジュール３１は、主題名称を抽出するように主題名称ページを設置し、主題章節目録又は主題章節を抽出するようにリストページを設置し、主題内容を抽出するように内容ページを設置するためのウェブサイト設置モジュール３１１を備える。

【００６７】

図４を続いて参照すると、設置モジュール３１は、更に、前記採取対象のネットデータのウェブサイトリンクアドレスを対応する種類のキューに設置するためのキュー設置モジュール３１２を備える。

キュー設置モジュール３１２は、更に、種類が主題名称であるリンクアドレスを主題名称ページキューに設置する第１設置手段３１２１と、種類がリストページであるリンクアドレスをリストページキューに設置する第２設置手段３１２２と、種類が内容ページであるリンクアドレスを内容ページキューに設置する第３設置手段３１２２と、を備える。

【００６８】

本実施例では、ウェブサイト取得モジュール３２は、主題名称ページキューから主題名称ページのリンクアドレスに対応するウェブサイトリソースを取得するための第１取得手段３２１と、リストキューからリストページのリンクアドレスに対応するウェブサイトリソースを取得するための第２取得手段３２２と、内容ページキューから内容ページのリンクアドレスに対応するウェブサイトリソースを取得するための第３取得手段３２３と、を備える。

図５を参照してください。

【００６９】

更に、本実施例では、データ抽出モジュール３３は、

ウェブサイトソースコードに対応するＵＲＬの採集深度値が第１閾値である時に、主題の名称及び名称に対応するＵＲＬを抽出し、かつ、名称に対応するＵＲＬの採集深度値を第２閾値にマークして第２設置手段に配送する第１抽出手段３３１と、

ウェブサイトソースコードに対応するＵＲＬの採集深度値が第２閾値である時に、主題の名称及び名称に対応するＵＲＬを抽出し、かつ、名称に対応するＵＲＬの採集深度値を第３閾値にマークして第２設置手段３１２２に配送する第２抽出手段３３２と、

ウェブサイトソースコードに対応するＵＲＬの採集深度値が第２閾値である時に、主題の章節目録及び章節目録のＵＲＬを抽出し、かつ、章節目録のＵＲＬの採集深度値を第３閾値にマークして第２設置手段３１２２に配送する第３抽出手段３３３と、

ウェブサイトソースコードに対応するＵＲＬには上級ＵＲＬが存在するか否か判断することに用いられ、存在すると判断する場合、主題の章節タイトル及び章節タイトルに対応する章節のＵＲＬを抽出し、かつ、章節のＵＲＬを第３設置手段３１２３に配送して、判断結果が存在しない場合、主題の名称、章節タイトル、及び章節タイトルに対応する章節のＵＲＬを抽出し、かつ、章節のＵＲＬを第３設置手段３１２３に配送する第４抽出手段３３４と、

ウェブサイトソースコードから主題の章節タイトル及び章節本文の内容を抽出し、かつ、ウェブサイトソースコードに対応するＵＲＬから章節タイトルに対応する章節の章節ＩＤを抽出することに用いられる第５抽出手段３３５と、

章節本文の内容にはページングが存在するか否か判断することに用いられ、章節本文の内容にはページングが存在する場合、第５抽出手段３３５が次のページのリンクアドレス

を抽出する同時に、現在ページのページ番号及び次のページのページ番号をマークするとともに次のページのリンクアドレスを第３設置手段３１２３に配送することにも用いられるページング判断手段３３６と、
章節の本文の第１ページのリンクを唯一のキー値としてページングの内容を格納するとともに、最後の１ページを採集終了する際に終了フラグを付けることに用いられるページング格納手段３３７と
を備える。
図６を参照してください。
【００７０】
第１実施例と異なるところは、第２実施例には、システムがリフレッシュモジュール３４を更に備える点である。
リフレッシュモジュール３４は、前記ウェブサイトで公表した、Ｍ個の主題のそれぞれに関連するオンライン・ドキュメントの更新頻度に応じてリフレッシュ時間間隔を設置し、かつ、前記リフレッシュ時間間隔に基づいて、前記採取対象のネットデータのウェブサイトリンクアドレスをリフレッシュすることに用いられる。
第２実施例について図７を参照してください。
【００７１】
第１、第２実施例と異なるところは、第３実施例には、システムは更に内容合併モジュール３５を備える点である。
内容合併モジュール３５は、抽出した全てのページングの内容を合併して、章節のタイトルを結合して出力することに用いられる。第３実施例に対しては図面８を参照してください。
【００７２】
本実施例において第２実施例におけるリフレッシュモジュールと組合せて採集作業を行ってもよいが、ここでは、発明の詳細な説明記載の簡潔さのために、組合せて使用するシステムに対する詳細の紹介を行っていない。
【００７３】
第１、第２、第３実施例と異なるところは、第４実施例では、システムは更に第１データ記憶モジュール３６及び第２データ記憶モジュール３７を備える点である。
第１データ記憶モジュール３６は、ホームページ、主題の名称、主題の章節タイトル、章節ＩＤ、及び章節本文内容をデータベースにアップロードすることに用いられる。
また、第２データ記憶モジュール３７は、章節本文の内容がより多くのデータベーススペースを占める可能性がある場合、該データベースを選択して、ウェブサイト、主題の名称、主題の章節タイトル、章節ＩＤ、及び章節本文内容の格納パスをデータベースにアップロードすることに用いられる。
ここで、章節本文の内容の格納パスは、章節本文内容を添付ファイルの形態にてファイルサーバに記憶するパスを意味する。
第４実施例について、図９を参照してください。
【００７４】
本実施例において第２実施例におけるリフレッシュモジュールと組合せて採集作業を行ってもよいが、ここでは、発明の詳細な説明記載の簡潔さのために、組合せて使用するシステムに対する詳細の紹介を行っていない。
【００７５】
上記の第１、第２、第３、及び第４実施例のシステムは、本発明が提供したネットデータの採集方法の実施例における方法及びその色々な変更形態に対して行った記述に基づいて実施することができる。ここでは、詳細の説明を行っていない。
【００７６】
本発明に係る１つの実施例では、ネットデータの採集システムを応用してネットデータの採集を行い、システムはネットデータのリンクアドレスを取得してリンクアドレスの種類を設置し、かつ、リンクアドレスの種類に応じてリンクアドレスを対応するキューに追

加する。そして、キューからリンクアドレスに対応するソースコードを取得して、ソースコードにおける対応するＵＲＬ情報及びＵＲＬの採集深度値に応じて、ネットデータの情報を抽出することにより、リアルタイムのネットデータを採集する技術効果が得られる。

さらに、本発明の実施例では、内容合併モジュールも採用して、同一の主題に属するオンライン・ドキュメントを合併することができるため、リアルタイムのネットデータを採集することにより、便利かつ集中的にブラウジングする効果が得られる。

【００７７】

当業者にとって理解すべきのは、本発明の実施形態が方法、システム、又はコンピュータプログラム製品で提供されることができる。従って、本発明は完全ハードウェア実施形態、完全ソフトウェア実施形態、又はソフトウェアとハードウェアの合わせの実施形態を用いることができる。かつ、本発明は１つ又は複数のその中にコンピュータ利用可能なプログラムコードを含むコンピュータ利用可能な記憶媒介（磁気メモリ、ＣＤ－ＲＯＭ、光学メモリ等を含むがこれらに限られない）で実施するコンピュータプログラム製品の形式を用いることができる。

【００７８】

本発明は本発明の実施形態による方法、装置（システム）、及びコンピュータプログラム製品のフロー図及び／又はブロック図を参照して説明したものである。理解すべきのは、コンピュータプログラムコマンドによりフロー図及び／又はブロック図の中の各流れ及び／又はブロック、及びフロー図及び／又はブロック図の中の流れ及び／又はブロックの合わせを実現できる。これらのコンピュータプログラムコマンドを通用コンピュータ、専用コンピュータ、埋め込みプロセッサ又はその他のプログラム可能なデータ処理装置のプロセッサに提供して１つの機器を生じ、コンピュータ又はその他のプログラム可能なデータ処理装置のプロセッサが実行するコマンドはフロー図の１つの流れ又は複数の流れ及び／又はブロック図の１つのブロック又は複数のブロックに指定する機能を実現するための装置を生じるようになる。

【００７９】

これらコンピュータプログラムコマンドはコンピュータ又はその他のプログラム可能なデータ処理装置を引導して所定の方式で動作させるコンピュータ読み取る可能なメモリに記憶されてもよく、該コンピュータ読み取る可能なメモリに記憶されるコマンドはコマンド装置を備える製品を生じるようになり、該コマンド装置がフロー図の１つの流れ又は複数の流れ及び／又はブロック図の１つのブロック又は複数のブロックに指定する機能を実現する。

【００８０】

これらコンピュータプログラムコマンドはコンピュータ又はその他のプログラム可能なデータ処理装置にロードしてもよく、コンピュータ又はその他のプログラム可能な装置で一連動作ステップを実行してコンピュータが実現する処理を生じ、このようにして、コンピュータ又はその他のプログラム可能な装置で実行するコマンドがフロー図の１つの流れ又は複数の流れ及び／又はブロック図の１つのブロック又は複数のブロックに指定する機能を実現するステップを提供する。

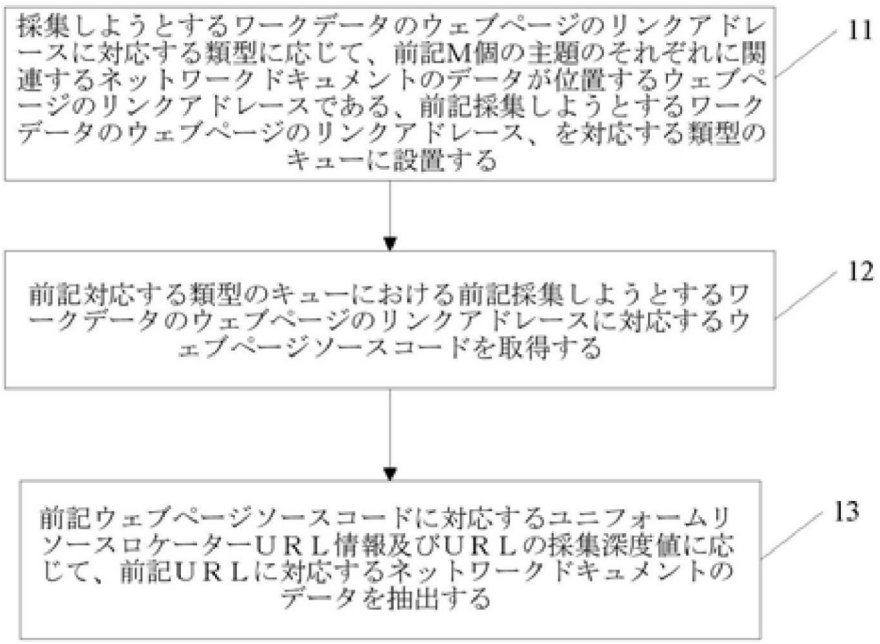
【００８１】

本発明の好適な実施形態を説明したが、当業者は基本的な創造性概念を知ると、これら実施形態に対して様々な変更と修正を行うことができる。従って、添付したクレームは好適な実施形態及び本発明範囲に落ちるすべての変更と修正を含む意図する。

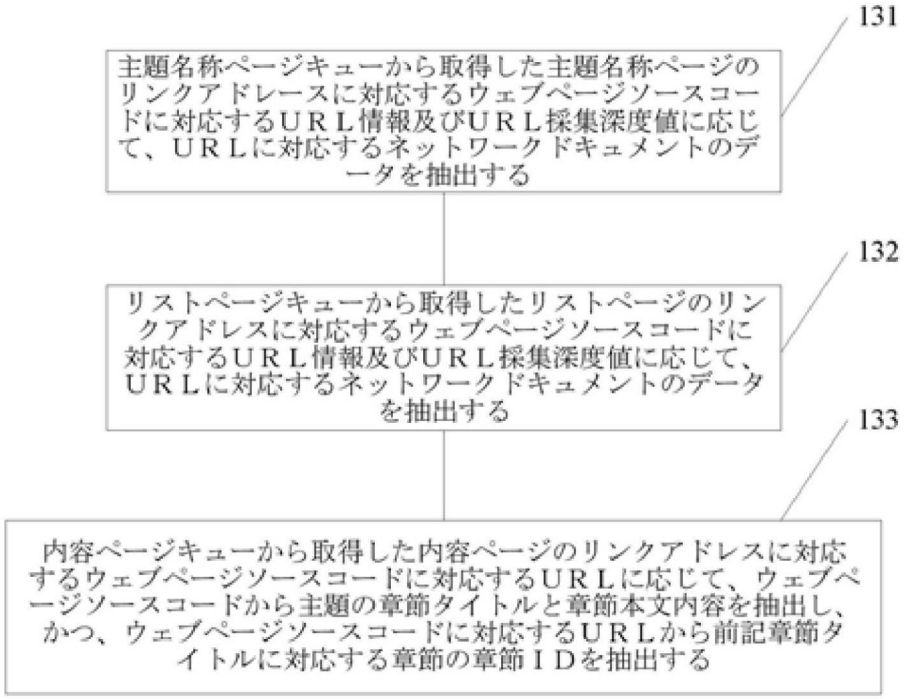
【００８２】

当然、当業者は本発明の実施形態に対して様々な変更と変形を行うことができるが、本発明の実施形態の精神と範囲を逸脱しない。このようにして、本発明の実施形態のこれら修正と変形が本発明のクレーム及びその同等技術の範囲に含まれれば、本発明はこれら修正と変形を含む意図する。

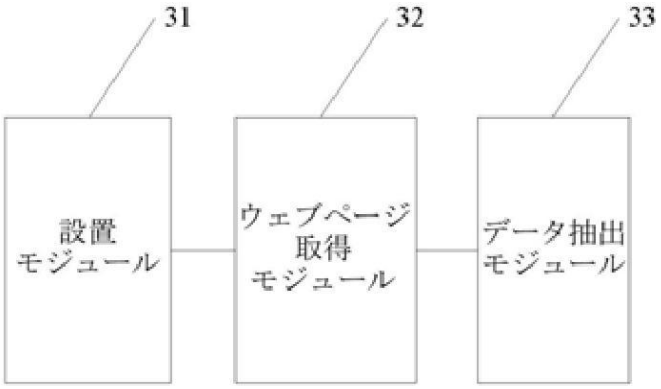
【図１】


採集しようとするワークデータのウェブページのリンクアドレースに対応する類型に応じて、前記M個の主題のそれぞれに関連するネットワークドキュメントのデータが位置するウェブページのリンクアドレースである、前記採集しようとするワークデータのウェブページのリンクアドレース、を対応する類型のキューに設置する
11
前記対応する類型のキューにおける前記採集しようとするワークデータのウェブページのリンクアドレースに対応するウェブページソースコードを取得する
12
前記ウェブページソースコードに対応するユニフォームリソースロケーターＵＲＬ情報及びＵＲＬの採集深度値に応じて、前記ＵＲＬに対応するネットワークドキュメントのデータを抽出する
13

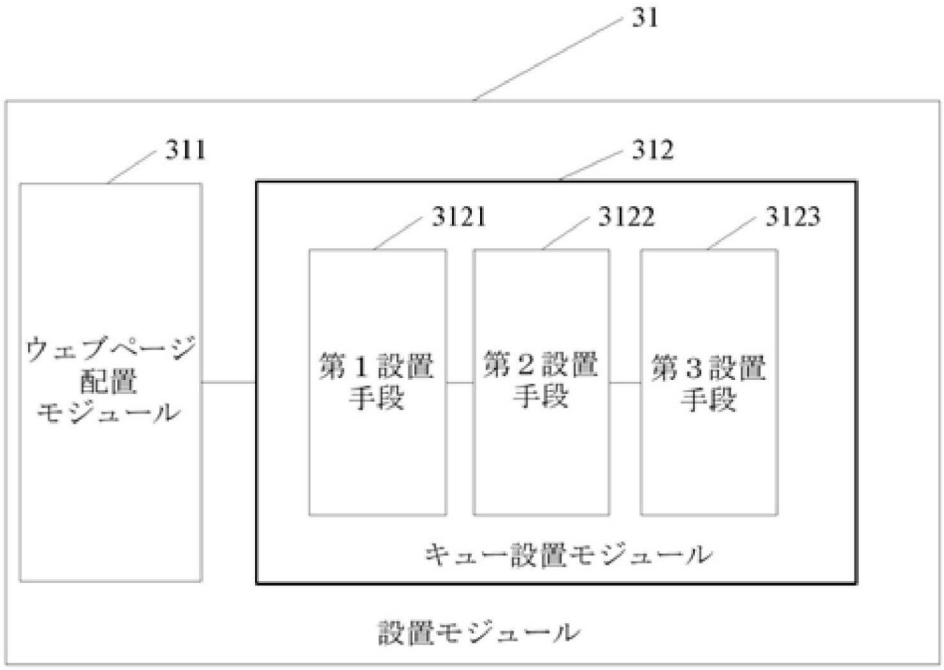
【図２】


主題名称ページキューから取得した主題名称ページのリンクアドレースに対応するウェブページソースコードに対応するＵＲＬ情報及びＵＲＬ採集深度値に応じて、ＵＲＬに対応するネットワークドキュメントのデータを抽出する
131
リストページキューから取得したリストページのリンクアドレスに対応するウェブページソースコードに対応するＵＲＬ情報及びＵＲＬ採集深度値に応じて、ＵＲＬに対応するネットワークドキュメントのデータを抽出する
132
内容ページキューから取得した内容ページのリンクアドレスに対応するウェブページソースコードに対応するＵＲＬに応じて、ウェブページソースコードから主題の章節タイトルと章節本文内容を抽出し、かつ、ウェブページソースコードに対応するＵＲＬから前記章節タイトルに対応する章節の章節ＩＤを抽出する
133

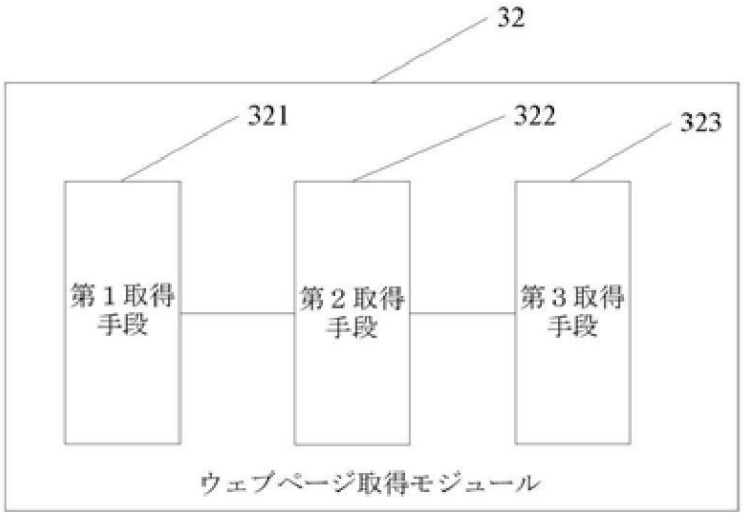
【図３】


31
32
33
設置
モジュール
ウェブページ
取得
モジュール
データ抽出
モジュール

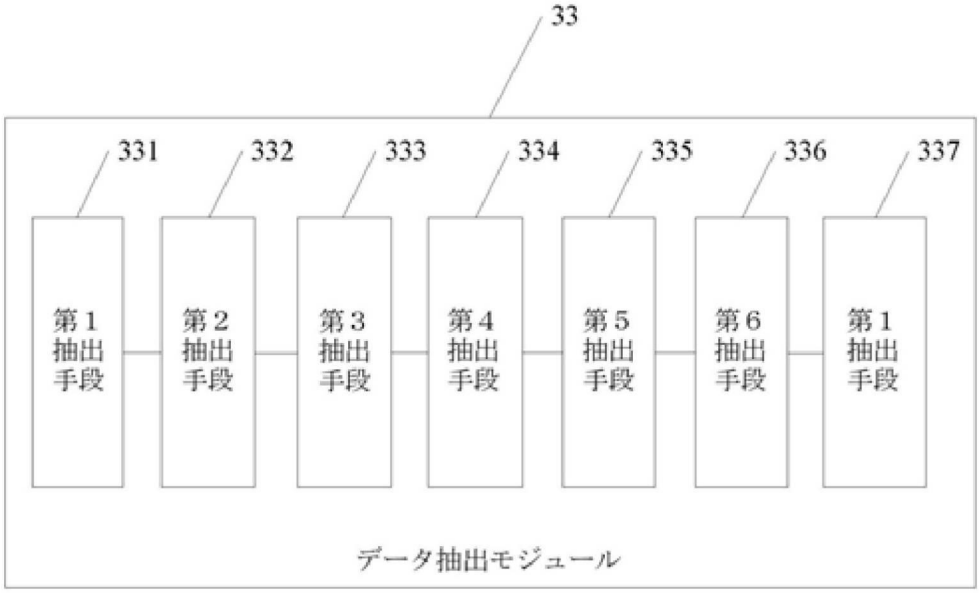
【図４】


31
311
312
3121
3122
3123
ウェブページ
配置
モジュール
第１設置
手段
第２設置
手段
第３設置
手段
キュー設置モジュール
設置モジュール

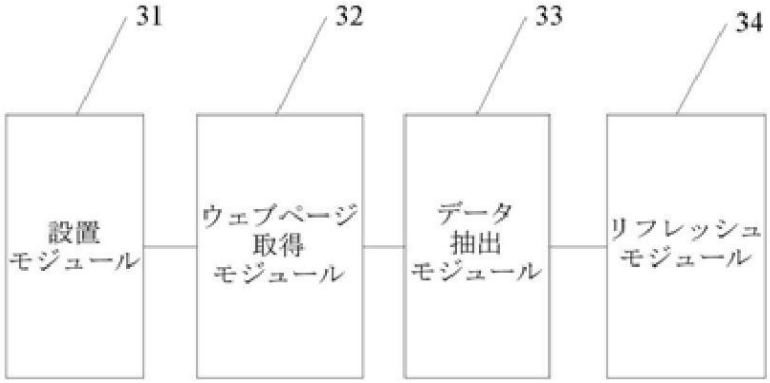
【図５】


32
321
322
323
第１取得
手段
第２取得
手段
第３取得
手段
ウェブページ取得モジュール

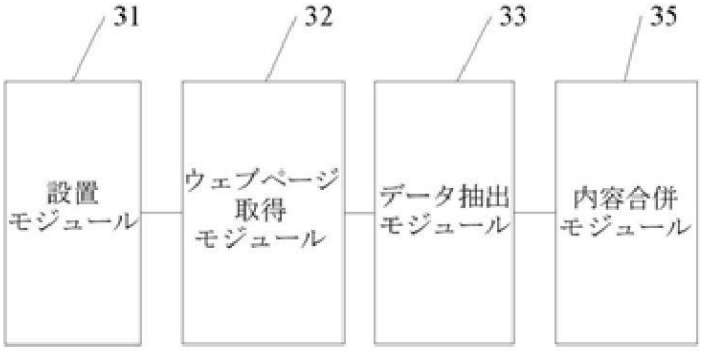
【図６】


33
331
332
333
334
335
336
337
第１抽出手段
第２抽出手段
第３抽出手段
第４抽出手段
第５抽出手段
第６抽出手段
第１抽出手段
データ抽出モジュール

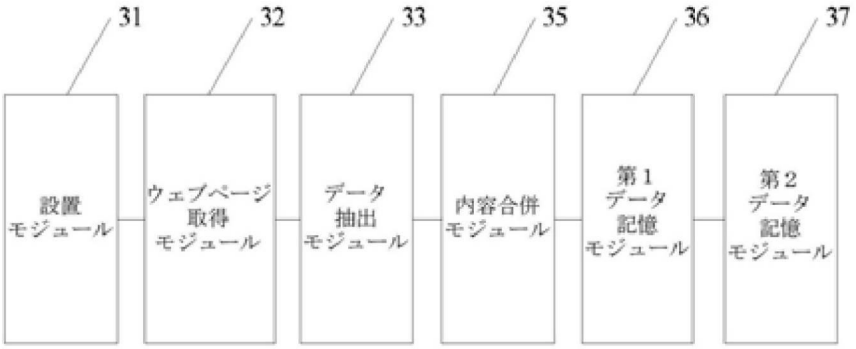
【図７】


31
32
33
34
設置モジュール
ウェブページ取得モジュール
データ抽出モジュール
リフレッシュモジュール

【図８】


31
32
33
35
設置モジュール
ウェブページ取得モジュール
データ抽出モジュール
内容合併モジュール

【図９】


31
32
33
35
36
37
設置
モジュール
ウェブページ
取得
モジュール
データ
抽出
モジュール
内容合併
モジュール
第１
データ
記憶
モジュール
第２
データ
記憶
モジュール

フロントページの続き

(73)特許権者 507232478
北京大学
PEKING UNIVERSITY
中華人民共和国北京市▲海▼淀区▲頤▼和▲園▼路５号
No. 5, Yiheyuan Road, Haidian District, Beijing 100871, China
(73)特許権者 507232456
北京北大方正▲電▼子有限公司
BEIJING FOUNDER ELECTRONICS CO., LTD.
中華人民共和国北京市▲海▼淀区上地五街９号方正大厦
Founder Building, No. 9, Shangdiwu Street, Haidian District, Beijing 100085, China
(74)代理人 110001243
特許業務法人 谷・阿部特許事務所
(72)発明者 ウー シンリー
中華人民共和国 １００８７１ ベイジン ハイディアン チェンフー ロード ナンバー ２９８ ジョングアンツン ファウンダー ビルディング ５ フロア
(72)発明者 ヤン ジエンウー
中華人民共和国 １００８７１ ベイジン ハイディアン チェンフー ロード ナンバー ２９８ ジョングアンツン ファウンダー ビルディング ５ フロア

審査官 齊藤 貴孝

(56)参考文献 特開２０１１－２１５９１２（ＪＰ，Ａ）
特開２００４－１１８４１５（ＪＰ，Ａ）
国際公開第２０１０／０４１５１７（WO，Ａ１）
特開２００６－２３５７２９（ＪＰ，Ａ）
特開２００６－０５８９６６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０
Ｇ０６Ｆ １３／００