

(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.	F I				
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 W02013/087012
 (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 F
 OUNDER GROUP CO., L
 TD
 中華人民共和国北京市▲海▼淀区成府路2
 98号中▲関▼村方正大厦5▲層▼
 5 Floor, Zhongguanc
 un Founder Building
 , No. 298, Chengfu R
 oad, Haidian Distri
 ct, 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ーケータ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】

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

10

20

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】

$$N_{Deep} = \left\{ \begin{array}{l} \text{第1閾値} \quad \text{「名称} \rightarrow \text{巻} \rightarrow \text{章節} \rightarrow \text{内容」のような作品構成を示す} \\ \text{第2閾値} \quad \text{「名称} \rightarrow \text{章節} \rightarrow \text{内容」のような作品構成を示す} \\ \text{第3閾値} \quad \text{「章節} \rightarrow \text{内容」のような作品構成を示す} \end{array} \right\}$$

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

50

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ーケータ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】

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

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、

10

20

30

40

50

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、

10

20

30

40

50

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

10

20

30

40

50

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。

10

20

30

40

50

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著作権、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。

10

【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。

20

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（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。

30

【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。

40

【0009】

【数1】

$$N_{Deep} = \left\{ \begin{array}{l} \text{第1 閾値} \quad \text{「名称} \rightarrow \text{巻} \rightarrow \text{章節} \rightarrow \text{内容」 のような作品構成を示す} \\ \text{第2 閾値} \quad \text{「名称} \rightarrow \text{章節} \rightarrow \text{内容」 のような作品構成を示す} \\ \text{第3 閾値} \quad \text{「章節} \rightarrow \text{内容」 のような作品構成を示す} \end{array} \right\}$$

【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

50

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。

10

20

30

40

50

【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

10

20

30

40

50

、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、便利にまとめて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ーケータ（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つは小説閲覧ウェブサイトの「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

10

20

30

40

50

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。

10

【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。

20

30

【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は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、定期的には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。

40

【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

50

、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

10

【0039】

【数2】

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

【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。

20

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 -> ポリウム目録 -> 章節目録 -> 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 -> 章節目録 -> 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。

30

A3の構成は「章節目録 -> 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。

40

【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

50

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ソースコードを取得し、かつ、ウェブ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も抽出すべき、「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 -

10

20

30

40

50

A 1 1 及び URL - A 2 1 が既に格納されている。また、ホームページ A 1 に対するネット文学採集開始の際に作品 A 3 に対応するリンクアドレス B 3 をリストページキューに、既に追加した。

【 0 0 5 6 】

URL - A 1 1 は、その採集深度値が 2 であるため、A 1 の章節目録及び該章節目録に対応する URL を抽出して、URL - A 1 2 で示す。URL - A 1 2 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。

【 0 0 5 7 】

URL - A 2 1 は、その採集深度値が 1 であり、かつ上級 URL (及び URL - A 2 1) を含む場合、A 2 の章節目録及び該章節目録に対応する URL を抽出して、URL - A 2 2 で示し、かつ、URL - A 2 2 を内容リストキューに追加する。

10

【 0 0 5 8 】

リストページキュー B 3 は、B 3 に対応するソースコードに対応する URL - A 3 の採集深度値が 1 であり、かつ、上級 URL を含めないため、A 3 の名称を抽出して、章節目録を「名称 A 3 」で示す。さらに、章節目録に対応する URL も抽出すべきであり、「URL - A 3 1」で示し、URL - A 3 1 を内容ページキューに追加する。

【 0 0 5 9 】

ステップ 1 3 3 において、章節目録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。

20

【 0 0 6 0 】

さらに、章節目録の本文内容の第 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。

【 0 0 6 1 】

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

【 0 0 6 2 】

さらに、ホームページ、主題の名称、主題の章節目録、章節目録 ID、及び章節目録本文内容をデータベースにアップロードする。また、章節目録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。

30

【 0 0 6 3 】

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。

【 0 0 6 4 】

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。

40

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 3 1、ウェブサイト取得モジュール 3 2、及びデータ抽出モジュール 3 3 を備える。

設置モジュール 3 1 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。

【 0 0 6 5 】

ウェブサイト取得モジュール 3 2 は、対応する種類のキューにおける採取対象のネット

50

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 3 3 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。

【 0 0 6 6 】

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 3 1 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 3 1 1 を備える。

10

【 0 0 6 7 】

図 4 を続いて参照すると、設置モジュール 3 1 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 3 1 2 を備える。

キュー設置モジュール 3 1 2 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 3 1 2 1 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 3 1 2 2 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 3 1 2 2 と、を備える。

20

【 0 0 6 8 】

本実施例では、ウェブサイト取得モジュール 3 2 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 3 2 1 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 3 2 2 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 3 2 3 と、を備える。

図 5 を参照してください。

【 0 0 6 9 】

更に、本実施例では、データ抽出モジュール 3 3 は、

30

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 3 3 1 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 3 1 2 2 に配送する第 2 抽出手段 3 3 2 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 3 1 2 2 に配送する第 3 抽出手段 3 3 3 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 3 1 2 3 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 3 1 2 3 に配送する第 4 抽出手段 3 3 4 と、

40

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 3 3 5 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 3 3 5 が次のページのリンクアドレス

50

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。

10

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。

20

【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、

30

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。

40

【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

50

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。

10

【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。

20

【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。

30

【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】

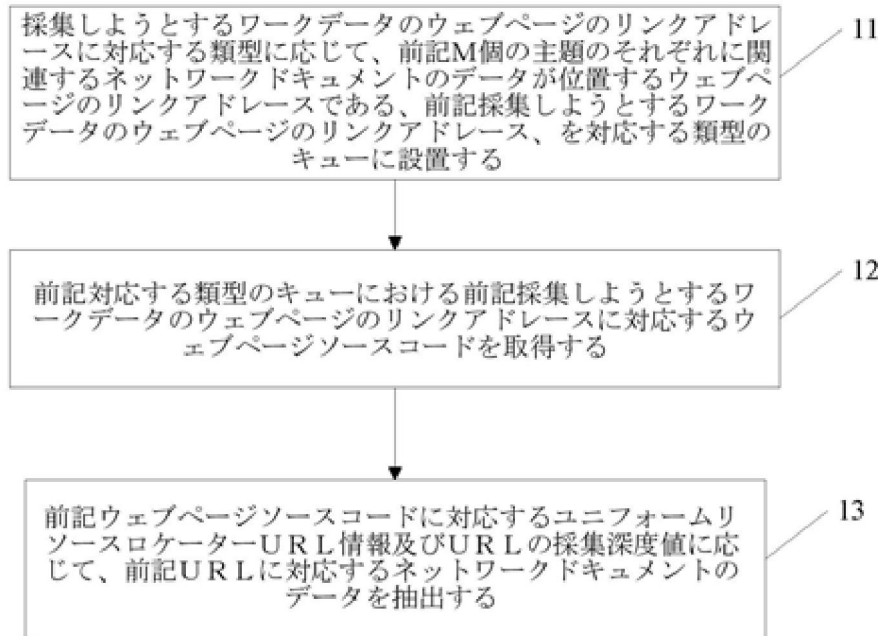
本発明の好適な実施形態を説明したが、当業者は基本的な創造性概念を知ると、これら実施形態に対して様々な変更と修正を行うことができる。従って、添付したクレームは好適な実施形態及び本発明範囲に落ちるすべての変更と修正を含む意図する。

40

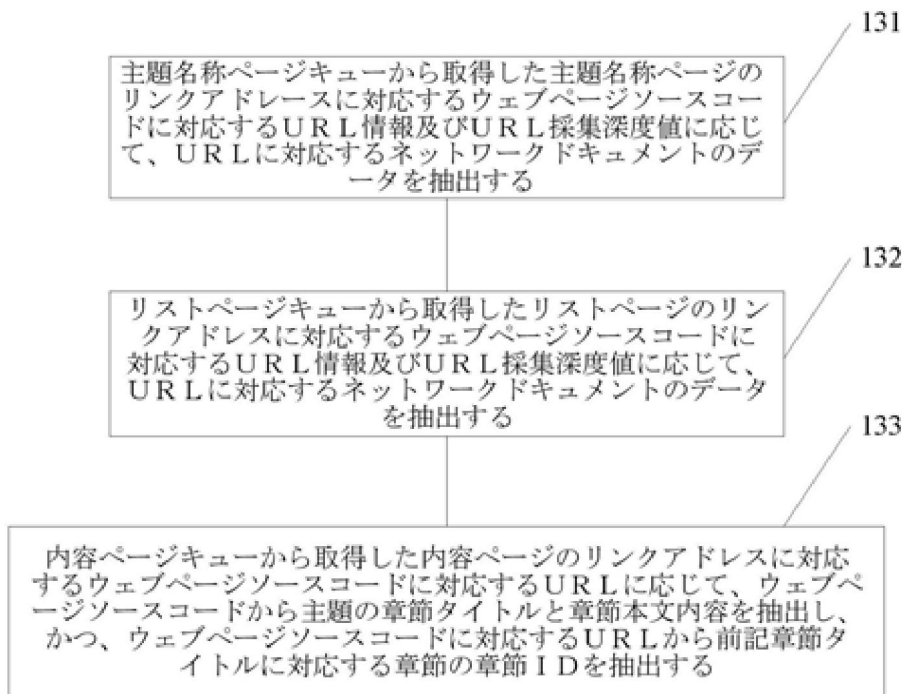
【0082】

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。

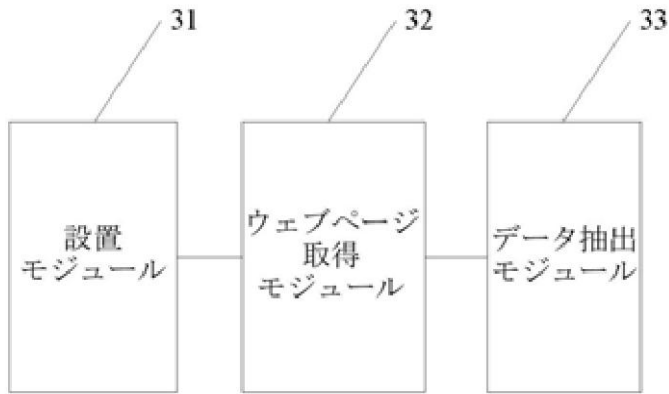
【図1】



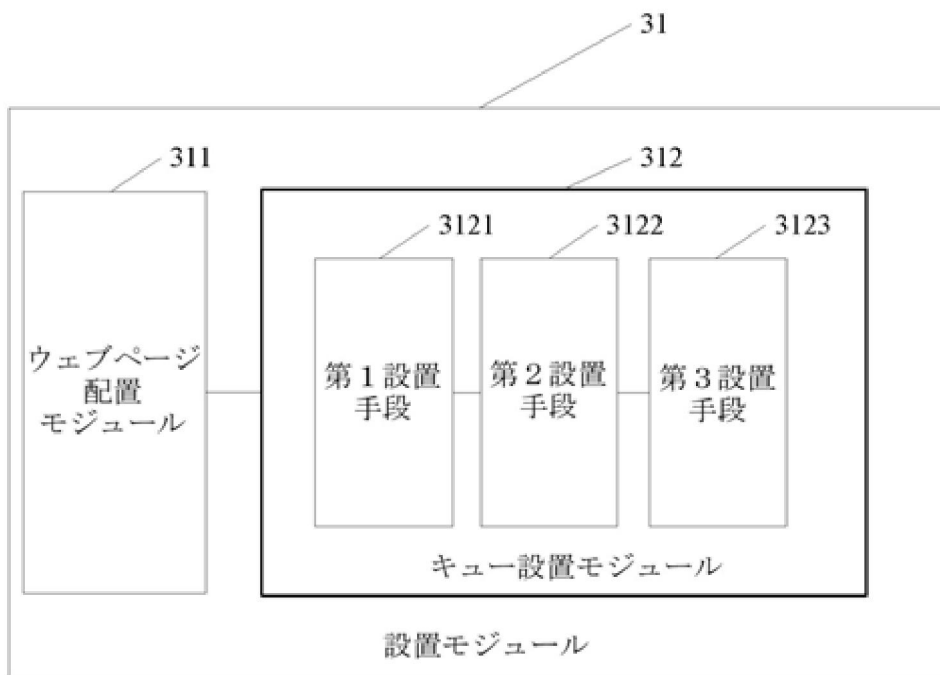
【図2】



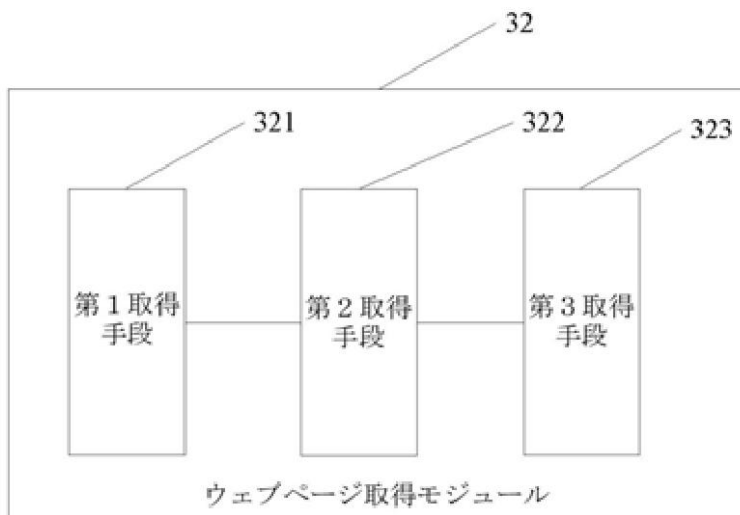
【図3】



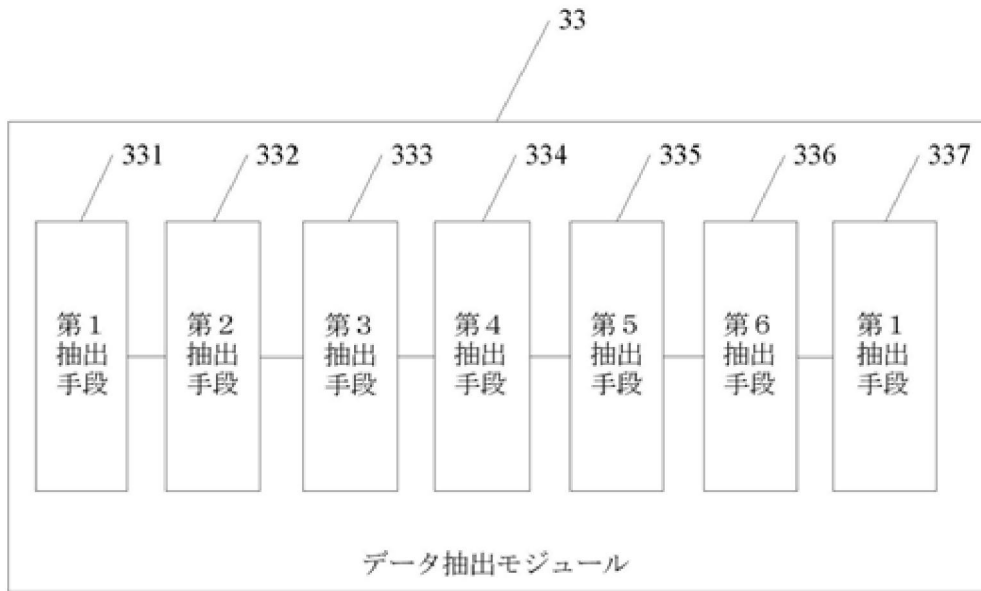
【図4】



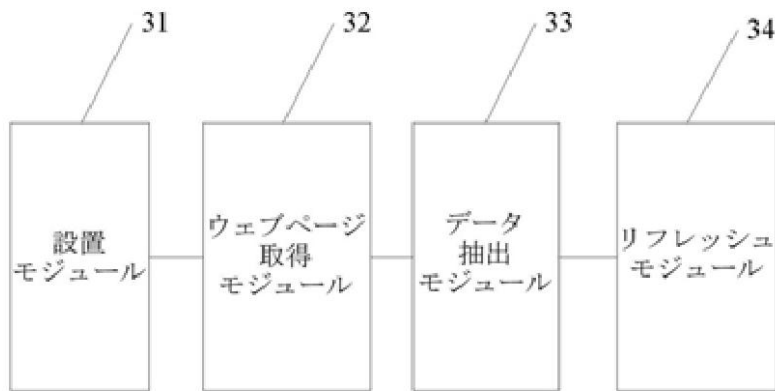
【図5】



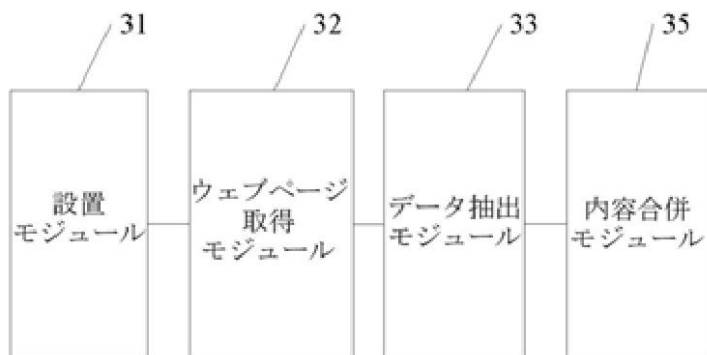
【図6】



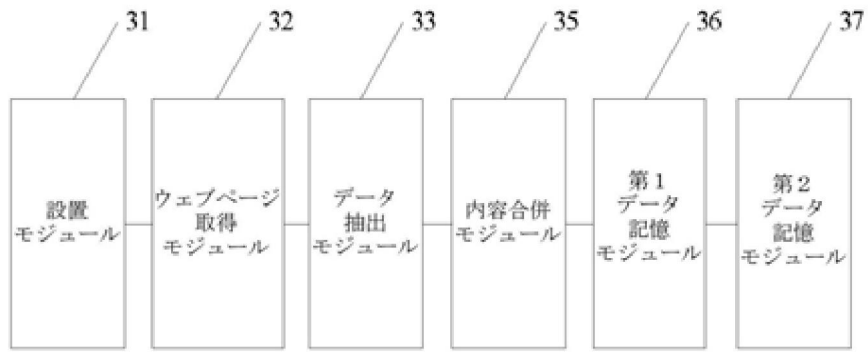
【図7】



【図8】



【図9】



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 チェンファー ロード ナンバー 29
8 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 チェンファー ロード ナンバー 29
8 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