

(19) 日本国特許庁(JP)

(12) 特許公報(B2)

(11) 特許番号

特許第5702474号
(P5702474)

(45) 発行日 平成27年4月15日(2015.4.15)

(24) 登録日 平成27年2月27日(2015.2.27)

(51) Int. Cl.		F I			
G06F 17/30	(2006.01)	G06F 17/30	180Z		
G06F 13/00	(2006.01)	G06F 13/00	560A		

請求項の数 17 (全 19 頁)

(21) 出願番号	特願2013-545030 (P2013-545030)	(73) 特許権者	507231932
(86) (22) 出願日	平成23年12月22日(2011.12.22)		北大方正集▲団▼有限公司
(65) 公表番号	特表2014-506355 (P2014-506355A)		PEKING UNIVERSITY F
(43) 公表日	平成26年3月13日(2014.3.13)		OUNDER GROUP CO., L
(86) 国際出願番号	PCT/CN2011/084457		TD
(87) 国際公開番号	W02012/083870		中華人民共和国北京市▲海▼淀区成府路2
(87) 国際公開日	平成24年6月28日(2012.6.28)		98号中▲関▼村方正大厦5▲層▼
審査請求日	平成26年4月28日(2014.4.28)		5 Floor, Zhongguanc
(31) 優先権主張番号	201010618393.4		un Founder Building
(32) 優先日	平成22年12月22日(2010.12.22)		, No. 298, Chengfu R
(33) 優先権主張国	中国 (CN)		oad, Haidian Distri
早期審査対象出願			ct, Beijing 100871,
			China

最終頁に続く

(54) 【発明の名称】 電子掲示板リプライ増加量の採集方法及びシステム

(57) 【特許請求の範囲】

【請求項 1】

判定装置と、抽出装置とを備えたシステムにおけるコンピューターによって実行される電子掲示板のリプライを増分的に採集する方法であって、

投稿のトップページURLと投稿のリプライ数情報に基づいて、採集が必要な全ての電子掲示板のリストページに新規投稿、及び/又は新規リプライがなされた投稿が存在するかどうかを周期的に判定する第1ステップ(1)と、

新規投稿が存在すると判定されたときに、新規投稿からメインポスト及びリプライ情報を抽出し、また、新規リプライがなされた投稿が存在すると判定されたときに、新規リプライ起点と新規リプライ数を計算し、新規リプライ起点と新規リプライ数に基づいて、新規リプライがなされた投稿から新規リプライ情報を抽出する第2ステップ(2)と、を含むことを特徴とする電子掲示板のリプライを増分的に採集する方法。

【請求項 2】

採集が必要な全ての電子掲示板のリストページに新規投稿及び/又は新規リプライがなされた投稿を周期的に判定する前記第1ステップ(1)において、

採集が必要な全ての電子掲示板のリストページURLを取得するステップ(a)と、

前記各リストページURLに対して、当該リストページURLに対応するウェブページ内容を取得し、前記ウェブページ内容から各投稿のトップページURLと現在のリプライ数を抽出するステップ(b)と、

前記投稿のトップページURLに基づいて、採集した投稿の情報リストに各投稿が存在

10

20

するかどうかを判定し、存在すると判定されたときに、当該投稿の現在のリプライ数が、採集した投稿情報に記録している今回のリプライ数より大きいかどうかを判定し、当該投稿の現在のリプライ数が今回のリプライ数より大きい場合に、当該投稿に新規リプライがあるとして、採集した投稿の情報リストに当該投稿の前回のリプライ数と今回のリプライ数を更新し、また、採集した投稿の情報リストに当該投稿がないと判定されたときに、当該投稿が新規投稿として、当該投稿のトップページURLと現在のリプライ数を、採集した投稿の情報リストに追加するステップ(c)と、
をさらに含むことを特徴とする請求項1に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項3】

前記採集が必要な全ての電子掲示板リストページURLを取得するステップ(a)において、

前記採集が必要な各電子掲示板リストページに対していずれも採集時間間隔を設け、各リストページの採集時間間隔をモニターし、あるリストページが採集時間間隔に達した時に、当該リストページURLをリストページ採集キューに追加し、

前記リストページ採集キューを定期的に走査し、前記リストページ採集キューが空でなければ、FIFOの順番で前記リストページ採集キューからリストページURLを順に抽出する

ことを特徴とする請求項2に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項4】

前記採集時間間隔は、リストページが属する電子掲示板の更新頻度によりダイナミックに調整され、

電子掲示板の更新頻度が速ければ速いほど、採集時間間隔が短くなり、

電子掲示板の更新頻度が遅ければ遅いほど、採集時間間隔が長くなる

ことを特徴とする請求項3に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項5】

前記リストページ採集キューから抽出されるリストページURLは、当該リストページURLが属するウェブサイトの正当のアクセスの条件を満たすことが必要である

ことを特徴とする請求項3に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項6】

新規投稿からメインポストとリプライを抽出し、新規リプライがなされた投稿から新規リプライの情報を抽出する前記第2ステップ(2)においては

新規投稿のトップページURLと新規リプライがなされた投稿のURLを内容ページ採集キューに追加するステップ(i)と、

前記内容ページ採集キューを定期的に走査するステップ(ii)と、

前記内容ページ採集キューが空でない場合、前記内容ページ採集キューから各URLを抽出するステップ(iii)と、

抽出されたURLに対応するウェブページ内容を取得し、前記ウェブページ内容からメインポスト及び/又はリプライ情報及び/又はページターニングURLを抽出し、ページターニングURLを前記内容ページ採集キューに追加するステップ(iv)と

を含むことを特徴とする請求項2乃至請求項5のいずれか1項に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項7】

前記新規投稿のトップページURLと、新規リプライがなされた投稿のURLとを内容ページ採集キューに追加するステップ(i)においては、

新規投稿については、当該投稿のトップページURLが前記内容ページ採集キューに存在する場合、当該投稿のトップページURLを抽出し、採集した投稿の情報リストに記録した当該投稿の今回のリプライ数を現在のリプライ数で置換し、当該投稿のトップページURLを前記内容ページ採集キューに挿入するが、当該投稿のトップページURLが前記内容ページ採集キューに存在しない場合、当該投稿のトップページURLを直接に前記内

10

20

30

40

50

容ページ採集キューに追加し、

新規リプライがなされた投稿については、当該投稿が属する電子掲示板のページターニングモードが計算ページターニングの場合、新規リプライがなされた投稿のトップページURLを直接に前記内容ページ採集キューに追加するが、当該投稿が属する電子掲示板のページターニングモードが次ページターニングの場合、ページターニングURL情報リストを調べて、前記ページターニングURL情報リストにおける最後のページターニングURLを前記内容ページ採集キューに追加する

ことを特徴とする請求項6に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項8】

前記内容ページ採集キューが空でない場合、前記内容ページ採集キューから各URLを抽出するステップ(iii)においては、

FIFOの順番で前記内容ページ採集キューからURLを順に抽出し、且つ、当該URLが属するウェブサイトの正当のアクセスの条件を満たす

ことを特徴とする請求項6に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項9】

前記抽出されたURLに対応するウェブページ内容を取得し、前記ウェブページ内容からメインポスト及び/又はリプライ情報及び/又はページターニングURLを抽出し、ページターニングURLを前記内容ページ採集キューに追加するステップ(iv)においては、

当該URLが投稿のトップページURLであって、初めて採集されるものである場合、当該URLに対応するウェブページ内容からメインポストとリプライ情報を抽出し、

当該URLが投稿のトップページURLであるが、初めて採集されるものではない場合、以下の式により新規リプライ起点 S'_{From} と新規リプライ数 $C'_{ParseCount}$ を確定し、新規リプライ起点 S'_{From} から $C'_{ParseCount}$ 個の新規リプライの情報を抽出し、

【数1】

$$S'_{From} = \begin{cases} R_{PreNum}, & N_{PerPage} \text{ は } \underline{\text{メインポストを含む}} \\ R_{PreNum} + 1, & N_{PerPage} \text{ は } \underline{\text{メインポストを含まない}} \end{cases}$$

$$C'_{ParseCount} = R_{CurNum} - R_{PreNum}$$

ここで、前記 R_{PreNum} は当該投稿の前回採集時のリプライ数を示し、前記 R_{CurNum} は当該投稿の現在のリプライ数を示し、前記 $N_{PerPage}$ は当該投稿のページ毎のリプライの数を示し、

当該URLが投稿のトップページURLではない場合、当該投稿に対応するページ番号が現在抽出すべきページのページ番号と同一であるかどうか判断することにより、オーバーラップページURLであるかどうかを判定し、

現在抽出すべきページのページ番号の計算式は以下の通りであり、

【数2】

$$P_{Begin} = \begin{cases} \text{ceil}\left(\frac{R_{PreNum} + 1}{N_{PerPage}}\right), & N_{PerPage} \text{ は } \underline{\text{メインポストを含む}} \\ \text{ceil}\left(\frac{R_{PreNum}}{N_{PerPage}}\right), & N_{PerPage} \text{ は } \underline{\text{メインポストを含まない}} \end{cases}$$

ここで、前記 P_{Begin} は現在抽出すべきページのページ番号を示し、前記 ceil はラウンドアップ演算を示し、

当該URLがオーバーラップページである場合、以下の式により新規リプライ起点 S'_{From} と新規リプライ数 $C'_{ParseCount}$ を計算し、新規リプライ起点 S'_{From} から $C'_{ParseCount}$ 個の新規リプライ情報を抽出し、

10

20

30

40

【数3】

$$S_{From}'' = \begin{cases} R_{PreNum} \% N_{PerPage} + 1, & N_{PerPage} \text{ はメインポストを含む} \\ R_{PreNum} \% N_{PerPage}, & N_{PerPage} \text{ はメインポストを含まない} \end{cases}$$

$$C_{ParseCount}'' = \begin{cases} R_{CurNum} - R_{PreNum}, & \text{このページは終了ページである} \\ N_{PerPage} - S_{From}'', & \text{このページは終了ページではない} \end{cases}$$

ここで、前記%は剰余演算を示し、

当該URLが投稿のトップページURLでもなく、オーバーラップページURLでもない場合、以下の式により、新規リプライ起点 S''_{From} と新規リプライ数 $C''_{ParseCount}$ を計算し、

【数4】

$$S_{From}''' = 0$$

$$C_{ParseCount}''' = \begin{cases} \begin{cases} R_{CurNum} \% N_{PerPage} + 1, & N_{PerPage} \text{ はメインポストを含む} \\ (R_{CurNum} - 1) \% N_{PerPage} + 1, \text{もし } R_{CurNum} - 1 > 0 \\ 0, & \text{もし } R_{CurNum} - 1 = 0 \end{cases} \text{ならば } N_{PerPage} \text{ はメインポストを含まない} & \text{このページは終了ページである} \\ N_{PerPage}, & \text{このページは終了ではない} \end{cases}$$

新規リプライ起点 S'''_{From} から $C'''_{ParseCount}$ 個の新規リプライの情報を抽出する

ことを特徴とする請求項6に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項10】

前記抽出されたURLに対応するウェブページ内容を取得し、前記ウェブページ内容からメインポスト及び/又はリプライ情報及び/又はページターニングURLを抽出し、ページターニングURLを前記内容ページ採集キューに追加するステップ(iv)においては、

掲示板が計算ページターニングモードであって、URLが投稿のトップページURLである場合、以下の式によりページターニングの開始ページ番号 P_{Begin} と終了ページ番号 P_{End} を計算し、

【数5】

$$P_{Begin} = \begin{cases} \text{ceil}\left(\frac{R_{PreNum} + 1}{N_{PerPage}}\right), & N_{PerPage} \text{ はメインポストを含む} \\ \text{ceil}\left(\frac{R_{PreNum}}{N_{PerPage}}\right), & N_{PerPage} \text{ はメインポストを含まない} \end{cases}$$

$$P_{End} = \begin{cases} \text{ceil}\left(\frac{R_{CurNum} + 1}{N_{PerPage}}\right), & N_{PerPage} \text{ はメインポストを含む} \\ \text{ceil}\left(\frac{R_{CurNum}}{N_{PerPage}}\right), & N_{PerPage} \text{ はメインポストを含まない} \end{cases}$$

$$S_{From}''' = \begin{cases} R_{PreNum} \% N_{PerPage} + 1, & N_{PerPage} \text{ はメインポストを含む} \\ R_{PreNum} \% N_{PerPage}, & N_{PerPage} \text{ はメインポストを含まない} \end{cases}$$

もし $S_{From} = 0$ 且つ $R_{PreNum} > 0$ であれば、 $S_{From} = N_{PerPage}$ 、 $P_{Begin} = P_{Begin} + 1$ とし、

ここで、前記 S_{From} は新規リプライ起点を示し、上記の式によりページターニングの開始ページ番号と終了ページ番号を算出してから、予め決められたページターニングURL規則に基づいて全てのページターニングURLを合成させ、

掲示板のページターニングモードが次ページターニングである場合、ウェブページ内容からページターニングURLを抽出する

ことを特徴とする請求項9に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項11】

前記ページターニングURL規則において、ページターニングURLは第1部分と、第2部分と第3部分と三つの部分に分けられており、前記第1部分と前記第3部分は変化しない部分であり、夫々 $strBeforePage$ と $strAfterPage$ と記しており、前記第2部分は変化する部分であり、 $nPageUp$ と記しており、前記ページターニングURLの合成方法は以下の通りであり、

【数6】

$$nPageNo = i + nFirstPostPageIndex - 1$$

$$nPageUp = (nPageNo \times nPageUsBaseNum)$$

$$strPostPageUrl = strBeforePage + nPageUp + strAfterPage$$

ここで、前記 i はターニングページ番号を示し、 P_{Begin} 、 i 、 P_{End} 、前記 $nPageNo$ は新規リプライが位置するページの番号を示し、前記 $nFirstPostPageIndex$ は投稿のトップページの番号を示し、前記 $nFirstPostPageIndex$ の値は0又は1となり、前記 $nPageUp$ は合成されるURL内に記録した、ページターニングを示すページ番号の値であり、前記 $nPageUsBaseNum$ はページターニング基数を示し、前記 $strPostPageUrl$ は合成されたURLを示す

ことを特徴とする請求項10に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項12】

ウェブページ内容より前記ページターニングURLを抽出してから、前記内容ページ採集キューに追加する前には、更に、前記ページターニングURLに対して重複除去処理を行う

ことを特徴とする請求項10に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項13】

前記重複除去処理においては、

前記ページターニングURL情報で当該ページターニングURLが属する投稿にページターニングURL情報リストが存在するかどうかを調べ、

前記ページターニングURL情報リストが存在しない場合、当該ページターニングURLが属する投稿のページターニングURL情報リストを立て、当該ページターニングURLをページターニングURL情報リストと前記内容ページ採集キューに追加し、

ページターニングURL情報リストが存在する場合、当該ページターニングURLのページ番号が当該ページターニングモードが属する投稿のページターニングURLのページ番号より大きいかどうかを判定し、大きい場合に、当該ページターニングURLが属する投稿のページターニングURL情報リストを更新し、ページターニングURLを内容ページ採集ジョブキューに追加し、大きくない場合、直接に当該ページターニングURLを削除する

ことを特徴とする請求項12に記載の電子掲示板のリプライを増分的に採集する方法。

【請求項14】

投稿のトップページURLと投稿のリプライ数情報に基づいて、採集が必要な全ての電子掲示板のリストページに新規投稿、及び/又は新規リプライがなされた投稿が存在するかどうかを周期的に判定する判定装置(11)と、

10

20

30

40

50

新規投稿に対しては、当該新規投稿からメインポストとリプライの情報を抽出し、また、新規リプライがなされた投稿に対しては、新規リプライ起点と新規リプライ数に基づいて、新規リプライがなされた投稿から新規リプライの情報を抽出する抽出装置(12)と、を備えることを特徴とする電子掲示板のリプライの増分的採集システム。

【請求項15】

前記判定装置(11)は、

採集が必要な全ての電子掲示板リストページURLをリストページ採集キューに追加する第1キュー手段(111)と、

リストページ採集キューから各リストページURLを抽出する第1取得手段(112)と、

抽出された各リストページURLに対して、このリストページURLに対応するウェブページ内容を取得し、上記ウェブページ内容から各投稿のトップページURLと現在のリプライ数を抽出するリストページ抽出手段(113)と、

投稿のトップページURLに基づいて、採集した投稿の情報リストに各投稿が存在するかどうかを判定し、存在すると判定されたときに、当該投稿の現在のリプライ数が、採集した投稿情報に記録している今回のリプライ数より大きいかどうかを判定し、当該投稿の現在のリプライ数が今回のリプライ数より大きい場合に、当該投稿に新規リプライがあるとして、採集した投稿の情報リストに当該投稿の前のリプライ数と今回のリプライ数を更新し、また、採集した投稿の情報リストに当該投稿がないと判定されたときに、当該投稿が新規投稿として、当該投稿のトップページURLと現在のリプライ数を、採集した投稿

の情報リストに追加する判定手段(114)と、を備えることを特徴とする請求項14に記載の電子掲示板のリプライの増分的採集システム。

【請求項16】

前記抽出装置(12)は、

新規投稿のトップページURLと、新規リプライがなされた投稿のURLとを前記内容ページ採集キューに追加する第2キュー手段(121)と、

前記内容ページ採集キューを定期的に走査する走査手段(122)と、

前記内容ページ採集キューから各URLを抽出する第2取得手段(123)と、

URLに対応するウェブページ内容を取得し、前記ウェブページ内容からメインポスト及び/又はリプライ及び/又はページターニングURLを抽出する内容ページ抽出手段(124)と、

を備えることを特徴とする請求項14に記載の電子掲示板のリプライの増分的採集システム。

【請求項17】

前記抽出装置(12)は、電子掲示板のページターニングモードが次ページターニングモードである場合、ウェブページ内容からページターニングURLを抽出して、重複除去処理を行う重複除去手段(125)を更に備え、

前記第2キュー手段(121)は、重複除去処理後のページターニングURLを前記内容ページ採集キューに追加することを特徴とする請求項16に記載の電子掲示板のリプライの増分的採集システム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、インターネット情報の採集技術に関し、更に詳しくは、電子掲示板のリプライ(reply)を増分的に採集する方法及びシステムに関する。

【背景技術】

【0002】

インターネットが出現し、特に、インターネット掲示板・インターネットコミュニティが広く設けられるにつれて、世界中の人々は一緒に自由に様々な考え方を発表し交流でき

10

20

30

40

50

ることになっている。中国でのインターネットの電子掲示板は一百万以上に達している。そして、80%のウェブサイトは独立的な電子掲示板を持っている。時々インターネットの電子掲示板を見るユーザー数は一億以上になっている。他の形式と異なり、インターネットの電子掲示板は、スピードが速くて、範囲が広い特徴を持っている。注目されている話題は、短い期間で一万以上のインターネット利用者にリプライされて検討されるので、リプライの情報も何百何千ページに達することが可能である。その場合、インターネット利用者は、話題が立てる人の言論であるメインポストの内容だけでなく、他のインターネット利用者のこの話題に対するリプライを見たがっている。ところで、通常の検索エンジンでは、ページめくりのリプライ情報を検索することが困難である。偶に見つかっていても、時効性の問題があるので、望ましいものではなく、データ遅延の問題もある。現在の電子掲示板採集システムも、スレッドのトップページの情報しか採集できず、メインポストへのリプライの情報を採集していない。

10

【発明の概要】**【発明が解決しようとする課題】****【0003】**

本発明は従来技術の上記した欠陥に鑑みながらなされたものであり、迅速に正確に完全に電子掲示板における投稿の全てのメインポスト・リプライの情報を採集することができる電子掲示板のリプライを増分的に採集する方法及びシステムを提供することを目的とする。本発明は、従来の検索エンジンが投稿のページターニングのリプライ情報を採集する時に存在する検索漏れや検索不能などの問題を解決し、従来の電子掲示板採集システムにおいてスレッドのトップページ情報が採集されるが、リプライ情報が採集されない問題を解決することができる。

20

【課題を解決するための手段】**【0004】**

前記の目的を達成するために、本発明は電子掲示板のリプライを増分的に採集する方法であって、投稿のトップページURLと投稿のリプライ数情報に基づいて、採集が必要な全ての電子掲示板のリストページに新規投稿、及び/又は新規リプライがなされた投稿が存在するかどうかを周期的に判定するステップと、新規投稿が存在すると判定されたときに、新規投稿からメインポスト及びリプライ情報を抽出し、また、新規リプライがなされた投稿が存在すると判定されたときに、新規リプライ起点と新規リプライ数を計算し、新規リプライ起点と新規リプライ数に基づいて、新規リプライがなされた投稿から新規リプライ情報を抽出するステップとを含むことを特徴とする。

30

また、前記の目的を達成するために、本発明は電子掲示板のリプライを増分的採集システムであって、投稿のトップページURLと投稿のリプライ数情報に基づいて、採集が必要な全ての電子掲示板のリストページに新規投稿、及び/又は新規リプライがなされた投稿が存在するかどうかを周期的に判定する判定装置と、新規投稿に対しては、当該新規投稿からメインポスト及びリプライ情報を抽出し、また、新規リプライがなされた投稿に対しては、新規リプライ起点と新規リプライ数に基づいて、新規リプライがなされた投稿から新規リプライの情報を抽出する抽出装置とを備えることを特徴とする。

【発明の効果】

40

【0005】

本発明の上記構成の方法とシステムは、電子掲示板のリストページを周期的にモニターすることによって、適時にリストページから新規リプライ及び、新規リプライがなされた投稿の情報を取得することができ、URL標識とリプライ数情報に基づいて重複除去処理を快速に行うことによって、重複採集を回避することができ、それぞれ異なったページターニングリンクの抽出方式を使い分けることによって、迅速にページターニングのリプライを採集することができる。本発明により、迅速に正確に完全に投稿の全てのスレッド/リプライ情報を採集することができるようになる。本発明は、リプライの採集漏れ率が5%以下になり、リアルタイム性が分レベルに達することができる効果を奏している。

【図面の簡単な説明】

50

【 0 0 0 6 】

【図 1】具体的な実施形態における電子掲示板のリプライの増分的採集システムの構成を示すブロック図である。

【図 2】具体的な実施形態における電子掲示板のリプライを増分的に採集する方法を示すフローチャートである。

【図 3】具体的な実施形態においてリストページには新規投稿と、新規リプライがなされた投稿とが存在するかどうかを判定する方法を示すフローチャートである。

【図 4】具体的な実施形態において新規投稿からメインポスト及びリプライ情報を抽出し、新規リプライがなされた投稿から新規リプライ情報を抽出する方法を示すフローチャートである。

10

【発明を実施するための形態】

【 0 0 0 7 】

以下、図面を参照しながら具体的な実施例で本発明を詳しく説明する。

【 0 0 0 8 】

図 1 に示すように、本実施形態にかかる電子掲示板のリプライの増分的採集システムは、判定装置 1 1 と、判定装置 1 1 に接続される抽出装置 1 2 とを備えている。その中、判定装置 1 1 は、第 1 キュー (q u e u e) 手段 1 1 1 と、第 1 取得手段 1 1 2 と、リストページ抽出手段 1 1 3 と、判定手段 1 1 4 とを含んでいる。抽出装置 1 2 は、第 2 キュー手段 1 2 1 と、走査手段 1 2 2 と、第 2 取得手段 1 2 3 と、内容ページ抽出手段 1 2 4 と、重複除去手段 1 2 5 とを含んでいる。

20

【 0 0 0 9 】

判定装置 1 1 は、投稿のトップページ URL と投稿のリプライ数情報に基づいて、採集が必要な全ての電子掲示板のリストページに、新規投稿と、新規リプライがなされた投稿とが存在するかどうかを周期的に判定する。その中、第 1 キュー手段 1 1 1 は、採集が必要な全ての電子掲示板リストページ URL をリストページ採集キューに追加する。第 1 取得手段 1 1 2 はリストページ採集キューから各リストページ URL を抽出する。リストページ抽出手段 1 1 3 は、抽出された各リストページ URL に対して、このリストページ URL に対応するウェブページ内容を取得し、上記ウェブページ内容から各投稿のトップページ URL と現在のリプライ数とを抽出する。判定手段 1 1 4 は、投稿のトップページ URL に基づいて、採集した投稿の情報リストに各投稿が存在するかどうかを判定し、存在すると判定されたときに、当該投稿の現在のリプライ数が、採集した投稿情報に記録している今回のリプライ数より大きいかどうかを判定し、当該投稿の現在のリプライ数が今回のリプライ数より大きい場合に、当該投稿に新規リプライがあるとして、採集した投稿の情報リストに当該投稿の前のリプライ数と今回のリプライ数を更新し、また、採集した投稿の情報リストに当該投稿がないと判定されたときに、当該投稿が新規投稿として、当該投稿のトップページ URL と現在のリプライ数を、採集した投稿の情報リストに追加する。

30

【 0 0 1 0 】

抽出装置 1 2 は、新規投稿に対して、新規投稿からメインポスト及びリプライ情報を抽出し、新規リプライがなされた投稿に対して、新規リプライ起点と新規リプライ数を計算し、新規リプライ起点と新規リプライ数に基づいて、新規リプライがなされた投稿から新規リプライ情報を抽出する。その中、第 2 キュー手段 1 2 1 は、新規投稿のトップページ URL と新規リプライがなされた投稿の URL を内容ページ採集キューに追加する。走査手段 1 2 2 は、内容ページ採集キューを定期的に走査する。第 2 取得手段 1 2 3 は、内容ページ採集キューから各 URL を抽出する。内容ページ抽出手段 1 2 4 は URL に対応するウェブページ内容を取得し、上記ウェブページ内容からメインポスト及び / 又はリプライ及び / 又はページターニング URL を抽出する。重複除去手段 1 2 5 は、電子掲示板のページターニングモードが次ページターニングである場合に、ウェブページ内容からページターニング URL を抽出して、重複除去処理を行う。第 2 キュー手段 1 2 1 は重複除去処理後のページターニング URL を内容ページ採集キューに追加する。

40

50

【 0 0 1 1 】

図2に示すように、本実施形態において、図1に示すシステムに基づく電子掲示板のリプライを増分的に採集する方法は、以下の各ステップを含む。

【 0 0 1 2 】

(1) 判定装置11は、採集が必要な全ての電子掲示板のリストページに新規投稿、及び/又は新規リプライがなされた投稿を周期的に判定する。

図3に示すように、本実施形態に採用された判定方法は以下のステップを含む。

【 0 0 1 3 】

(a) 第1キュー手段111は、採集が必要な全ての電子掲示板リストページURLをリストページ採集キューに追加する。上記のリストページとは、電子掲示板において、全ての投稿のタイトル、URL (Uniform Resource Locator)、クリック数、リプライ数などの情報が含まれるページである。投稿の具体的な内容を含まない。例えば、SOHU (登録商標) 電子掲示板の経済総合チャンネルのリストページは、URLがhttp://club.business.sohu.com/1-enjoy-0-0-0-0.himlである。

10

【 0 0 1 4 】

また、例えば、人民網の強国コミュニティの国際電子掲示板チャンネルのリストページは、URLがhttp://bbs1.people.com.cn/boardList.do?action=postList&boardId=6である。

【 0 0 1 5 】

本実施形態において、採集が必要な各電子掲示板リストページに対して採集時間間隔を設け、例えば、5分おきに採集するとする。各リストページの採集時間間隔をモニターし、あるリストページが採集時間間隔に達すると、そのリストページURLをリストページ採集キューに追加する。

20

【 0 0 1 6 】

好ましくは、リフレッシュの時間間隔は、電子掲示板の更新頻度によりダイナミックに調整され、電子掲示板の更新頻度が速ければ速いほど、リフレッシュの時間間隔が短くなり、電子掲示板の更新頻度が遅ければ遅いほど、リフレッシュの時間間隔が長くなる。例えば、五分間おきに一回採集すると規定しておいているが、以降の採集において電子掲示板の更新頻度が速くなっていると分かったときに、リフレッシュの時間間隔を、3分間おきに、更に、1分間おきに若しくはもっと短い時間おきにと、短縮させても良い。

30

【 0 0 1 7 】

電子掲示板の更新頻度の計算方法は中国特許出願 (出願番号: 201010236363.7、発明の名称: ウェブページデータ情報の定方向採集方法及び装置) を参照してください。詳細な説明はここで省く。

【 0 0 1 8 】

(b) 第1取得手段112は、リストページ採集キューから各リストページURLを抽出する。

【 0 0 1 9 】

本実施形態において、リストページ採集キューからリストページURLを取得するために採用する方法としては、リストページ採集キューを定期的に走査し (ユーザーが具体的な応用に応じて走査間隔時間を設けてもいい)、リストページ採集キューが空いていなければ、FIFOの順番でリストページ採集キューからリストページURLを順に抽出し (キューからURLを抽出してから、そのURLが自動的にキューから削除される)、そして、このリストページURLが属するウェブサイトの正当のアクセスの条件を満たすことが必要である。あるリストページURLがこのリストページURLが属するウェブサイトの正当のアクセスの条件を満たしていないと、今回の走査においてこのリストページURLを無視してそれを次の走査の処理に回し、次のリストページURLの判定に移行する。ウェブサイトの正当のアクセスの条件としては、現在のアクセス数の制限と、アクセスの時間間隔の制限とを含む。ウェブサイトの正当のアクセスの条件を満たしているかどうか

40

50

を判断する方法は、中国特許出願（出願番号：201010546334.0、発明の名称：ウェブサイトにおける複数の異なるIDのサーバからウェブページを抽出する方法及びシステム）を参照してください。詳細な説明は省く。

【0020】

(c) リストページ抽出手段113は、抽出された各リストページURLに対応するウェブページ内容を取得する。その後、ウェブページから各投稿のトップページURLと現在のリプライ数を抽出する。

各リストページのURLにより、そのURLが属するウェブサイトに、そのURLに対応するウェブページ内容を取得する旨のHTTP請求を送信し、その後、送信してくるウェブページの内容を受信する。ウェブページ内容から投稿のトップページと現在のリプライ数を抽出する技術は従来の技術でもよく、詳細な説明は省く。

10

【0021】

(d) 判定手段113は、投稿のトップページURLに基づいてその投稿が採集した投稿情報に存在するかどうかを判定する。存在すると、この投稿が採集されたとして、この投稿の現在のリプライ数が採集した投稿情報リストに記録した今回のリプライ数より大きいかどうかを判定する。この投稿の現在のリプライ数が今回のリプライ数より大きい場合に、この投稿に新規リプライがあるとして、採集した投稿情報リストにその投稿の前のリプライ数と今回のリプライ数を更新し、即ち、採集した投稿情報リストにおけるその投稿の今回のリプライ数の数値で前のリプライ数の数値を置換し、この投稿の現在のリプライ数の数値で採集した投稿情報リストにおけるこの投稿の今回のリプライ数の数値を置換する。この投稿の現在のリプライ数は今回の数より大きくない場合に、この投稿に新規リプライがないとして、この投稿のURLを無視して処理を行わない。この投稿が採集した投稿の情報リストにない場合に、この投稿が新規投稿であるとして、この投稿のトップページURLと現在のリプライ数を、採集した投稿情報リストに追加する。この投稿の前のリプライ数が0、今回のリプライ数が現在のリプライ数である。

20

【0022】

採集した投稿情報リストには、採集した投稿のトップページURLと、採集した投稿の前のリプライ数と今回のリプライ数とが記憶されている。その構成は以下の表に示される。

【0023】

【表1】

採集した投稿のトップページURL	前のリプライ数	今回のリプライ数
URL ₁ (URL ₁ のMD5値)	a ₁	a ₂
URL ₂ (URL ₂ のMD5値)	b ₁	b ₂
...		...
URL _n (URL _n のMD5値)	n ₁	n ₂

30

40

【0024】

好ましくは、採集した投稿の情報リストに、投稿のトップページのURLの標識情報を記憶し、例えば、MD5コードである。標識情報を比較することにより、採集した投稿の情報リストに投稿のトップページURLが存在するかどうかを確定する。そうすれば、URLの比較効率を向上することができる。

【0025】

(2) 採集が必要な電子掲示板リストページには新規投稿と、及び/又は新規リプライがなされた投稿とが存在すると、抽出装置12は新規投稿からメインポストとリプライ情報を抽出し、新規リプライがなされた投稿から新規リプライ情報を抽出する。

50

【 0 0 2 6 】

面 4 に示すように、本実施形態に採用される抽出方法は以下のステップを含む。

【 0 0 2 7 】

(i) 第 2 キュー手段 1 2 1 は新規投稿のトップページ URL と、新規リプライがなされた投稿の URL とを内容ページ採集キューに追加する。

【 0 0 2 8 】

新規投稿については、当該投稿のトップページ URL が内容ページ採集キューに存在する場合、当該投稿のトップページ URL を抽出し、採集した投稿の情報リストに記録した当該投稿の今回のリプライ数を現在のリプライ数で置換し、当該投稿のトップページ URL を内容ページ採集キューに挿入するが、当該投稿のトップページ URL が内容ページ採集キューに存在しない場合、当該投稿のトップページ URL を直接に内容ページ採集キューに追加する。

10

【 0 0 2 9 】

新規リプライがなされた投稿については、当該投稿が属する電子掲示板のページターニングモードが計算ページターニングの場合、新規リプライがなされた投稿のトップページ URL を直接に内容ページ採集キューに追加するが、当該投稿が属する電子掲示板のページターニングモードが次ページターニングの場合、ページターニング URL 情報リストを調べて、前記ページターニング URL 情報リストにおける最後のページターニング URL を内容ページ採集キューに追加する。

【 0 0 3 0 】

いわゆる「計算ページターニングモード」とは、一ページあたりのリプライ数が決められたページターニングモードを指す。例えば、人民網の強国コミュニティの国際電子掲示板において、

`http://bbs1.people.com.cn/postDetail.do?boardId=6&view=1&id=91384467` という投稿は計算ページターニングモードとなる。

20

【 0 0 3 1 】

いわゆる「次ページターニングモード」とは、一ページあたりのリプライ数が決められていないページターニングモードを指す。例えば、天涯総合電子掲示板において、`http://www.tianya.cn/publicforum/content/free/1/1880805.shtml` という投稿は次ページターニングモードである。

30

【 0 0 3 2 】

(i i) 走査手段 1 2 2 は内容ページ採集キューを定期的に走査する。ユーザーは具体的な応用に応じて走査間隔時間を設けることができる。

【 0 0 3 3 】

(i i i) 内容ページ採集キューが空いていないと、第 2 取得手段 1 2 3 は内容ページ採集キューから各 URL を抽出する。内容ページ採集キューから抽出された URL は自動的にこの内容ページ採集キューから削除されるようになる。

【 0 0 3 4 】

本実施形態において、第 2 取得手段 1 2 3 が内容ページ採集キューから URL を抽出する方法は、第 1 取得手段 1 1 2 がリストページ採集キューから URL を取る方法と同じであるため、ここでその説明を省く。

40

【 0 0 3 5 】

(i v) 内容ページ抽出手段 1 2 4 は、抽出された URL に対応するウェブページ内容を取得して、上記ウェブページ内容から メインポスト 及び / 又はリプライ情報及び / 又はページターニング URL を抽出し、ページターニング URL を内容ページ採集キューに追加する。

【 0 0 3 6 】

本実施形態において、ウェブページ内容から メインポスト 及び / 又はリプライを抽出する具体的な方法は以下のとおりである。

50

【 0 0 3 7 】

当該URLは投稿のトップページURLであり、且つ初めて採集される場合に、この投稿を新規投稿として、このURLに対応するウェブページ内容からメインポストとリプライ情報を抽出する。具体的には、まず、この投稿においてメインポストとリプライが同じスタイルであるかどうかを確定する。同じスタイルであれば、同一の抽出方式にて逐一に情報を抽出し、抽出した第1の情報をメインポストとし、他の情報をリプライとする。同じスタイルでなければ、所定の規則に従ってメインポストの情報を抽出してから、各リプライの情報を抽出する。メインポストとリプライが同じスタイルであるかどうかについては、人工的に規定することができる。所定の規則は人工的に設けられたキーワード又は正規表現である。

10

【 0 0 3 8 】

当該URLは投稿のトップページURLであるが、初めて採集されるものではない場合に、即ち、新規リプライがなされた投稿である場合に、以下の式により新規リプライ起点 S'_{From} と新規リプライ数 $C'_{ParseCount}$ を確定し、新規リプライ起点 S'_{From} から $C'_{ParseCount}$ 個の新規リプライ情報を抽出する。

【 0 0 3 9 】

【 数 1 】

$$S'_{From} = \begin{cases} R_{PreNum} & N_{PerPage} \text{ は } \underline{\text{メインポストを含む}} \\ R_{PreNum} + 1 & N_{PerPage} \text{ は } \underline{\text{メインポストを含まない}} \end{cases}$$

$$C'_{ParseCount} = R_{CurNum} - R_{PreNum}$$

20

【 0 0 4 0 】

ここで、 R_{PreNum} はこの投稿の前回採集される時のリプライ数を示し、 R_{CurNum} はこの投稿の現在のリプライ数を示し、 $N_{PerPage}$ はこの投稿のページ毎のリプライ数を示す。

【 0 0 4 1 】

また、当該URLが投稿のトップページURLではない場合、当該投稿に対応するページ番号が現在抽出すべきページのページ番号と同一であるかどうか判断することにより、オーバーラップページURLであるかどうかを判定する。いわゆる「オーバーラップページ」とは、当該ページ内の情報が全部リプライ情報であって一部の情報が新規リプライ情報となるページのことをいう。現在抽出すべきページのページ番号の計算式は以下の通りである。

30

【 0 0 4 2 】

【 数 2 】

$$P_{Begin} = \begin{cases} \text{ceil}\left(\frac{R_{PreNum} + 1}{N_{PerPage}}\right) & N_{PerPage} \text{ は } \underline{\text{メインポストを含む}} \\ \text{ceil}\left(\frac{R_{PreNum}}{N_{PerPage}}\right) & N_{PerPage} \text{ は } \underline{\text{メインポストを含まない}} \end{cases}$$

40

【 0 0 4 3 】

ここで、 P_{Begin} は現在抽出すべきページのページ番号を示し、 ceil はラウンドアップ演算を示す。

【 0 0 4 4 】

また、当該URLがオーバーラップページである場合に、下の式にて新規リプライ起点 S''_{From} と新規リプライ数 $C''_{ParseCount}$ を計算し、新規リプライ起点 S''_{From} から $C''_{ParseCount}$ 個の新規リプライ情報を抽出する。

【 0 0 4 5 】

【数3】

$$S''_{From} = \begin{cases} R_{PreNum} \% N_{PerPage} + 1, & N_{PerPage} \text{ はメインポストを含む} \\ R_{PreNum} \% N_{PerPage}, & N_{PerPage} \text{ はメインポストを含まない} \end{cases}$$

$$C''_{ParseCount} = \begin{cases} R_{CurNum} - R_{PreNum}, & \text{このページは終了ページである} \\ N_{PerPage} - S''_{From}, & \text{このページは終了ページではない} \end{cases}$$

【0046】

ここで、「%」は剰余演算である。

10

【0047】

このURLが投稿のトップページURLでも、オーバーラップページURLでもない場合に、すなわち、このページ内容が全部新規リプライである場合に、以下の式により、新規リプライ起点 S'''_{From} と新規リプライ数 $C'''_{ParseCount}$ を計算し、新規リプライ起点 S'''_{From} から $C'''_{ParseCount}$ 個の新規リプライの情報を抽出する。

【0048】

$$S'''_{From} = 0$$

【0049】

【数4】

20

$$C'''_{ParseCount} = \begin{cases} \begin{cases} R_{CurNum} \% N_{PerPage} + 1, & N_{PerPage} \text{ はメインポストを含む} \\ \begin{cases} (R_{CurNum} - 1) \% N_{PerPage} + 1, & \text{もし } R_{CurNum} - 1 > 0 \\ 0, & \text{もし } R_{CurNum} - 1 = 0 \end{cases}, & \text{ならば } N_{PerPage} \text{ はメインポストを含まない} \end{cases} \end{cases} \begin{cases} \text{このページは終了ページである} \\ \text{このページは終了ではない} \end{cases}$$

【0050】

ウェブページ内容からページターニングURLを抽出する方法は以下の具体的な方法を採用している。

30

【0051】

掲示板が計算ページターニングモードであり、且つURLが投稿のトップページURLである場合に、まず、以下の式によりページターニングの開始ページ番号と終了ページ番号を計算する。即ち、新規リプライが位置する開始ページ番号と終了ページ番号である。URLが投稿のトップページではない場合に、ページターニングURLを抽出しない。

【0052】

【数5】

$$P_{Begin} = \begin{cases} \text{ceil}\left(\frac{R_{PreNum} + 1}{N_{PerPage}}\right), & N_{PerPage} \text{ はメインポストを含む} \\ \text{ceil}\left(\frac{R_{PreNum}}{N_{PerPage}}\right), & N_{PerPage} \text{ はメインポストを含まない} \end{cases}$$

$$P_{End} = \begin{cases} \text{ceil}\left(\frac{R_{CurNum} + 1}{N_{PerPage}}\right), & N_{PerPage} \text{ はメインポストを含む} \\ \text{ceil}\left(\frac{R_{CurNum}}{N_{PerPage}}\right), & N_{PerPage} \text{ はメインポストを含まない} \end{cases}$$

$$S_{From} = \begin{cases} R_{PreNum} \% N_{PerPage} + 1, & N_{PerPage} \text{ はメインポストを含む} \\ R_{PreNum} \% N_{PerPage}, & N_{PerPage} \text{ はメインポストを含まない} \end{cases}$$

10

【0053】

もし $S_{From} = 0$ 且つ $R_{PreNum} > 0$ ならば、 $S_{From} = N_{PerPage}$ 、 $P_{Begin} = P_{Begin} + 1$ とする。

【0054】

ここで、 P_{Begin} と P_{End} は夫々ページターニングの開始ページ番号と終了ページ番号であり、 S_{From} は新規リプライ起点を示し、この起点から終了ページ番号までのすべてのリプライは新規リプライである。上記式によりページターニングの開始ページ番号と終了ページ番号を計算してから、所定のページターニングURL規則に従って、全てのページターニングURLを合成する。

20

【0055】

具体的な合成方法は、配置されたページターニング規則と、ページターニングの開始ページ番号と、ページターニング基数と基づいてページターニングURLを合成する。本実施形態は、前記ページターニング規則において、ページターニングURLは第1部分と、第2部分と第3部分と三つの部分に分けられており、第1部分と第3部分は変化しない部分であり、夫々 $strBeforePage$ と $strAfterPage$ と記され、第2部分は変化する部分であり、 $nPageUp$ と記される。ページターニングURLの合成過程の擬似コードは以下の通りである。

30

```
for (int i = P_Begin; i < P_End; ++ )
{
    nPageNo = i + nFirstPostPageIndex - 1;
    nPageUp = (nPageNo * nPageUsBaseNum);
    strPostPageUrl = strBeforePage + nPageUp +
strAfterPage;
}
```

40

【0056】

ここで、 $nPageNo$ は新規リプライが位置するページの番号を示し、 $nFirstPostPageIndex$ は投稿のトップページの番号を示す。実際の電子掲示板において、 $nFirstPostPageIndex$ の値は0又は1とすることができる。即ち、投稿のページ番号が0から番号を付け、トップページの番号が0であり、又は、投稿のページ番号が1から番号を付け、投稿のトップページの番号が1である。 $nPageUp$ は合成されるURL内に記録した、ターニングページを示すページ番号の値であり、即ち、第2部分の数値である。 $nPageUsBaseNum$ はページターニング基数を示す。 $strPostPageUrl$ は合成されたURLを示す。

【0057】

50

例を挙げて以下のように説明する。

事例一

人民網の強国コミュニティの国際電子掲示板チャンネルにおける一つの投稿URLは `http://bbs1.people.com.cn/postDetail.do?boardId=6&view=1&id=91384467` となる。

【0058】

そのページターニングのリンク規則は

`/postDetail¥.do¥?id=¥d+&view=¥d+&pageNo=(¥d+)&boardId=6`

である。

10

【0059】

開始ページ番号の `nFirstPostPageIndex` は1であり、ページターニング基数は1であり、`nPageUsBaseNum` は20である。

【0060】

ページターニングのリンク規則により、ページターニングURLの第1部分と第3部分を抽出し、夫々「`/postDetail.do?id=91384467&view=1&pageNo=`」と「`&boardId=6`」となる。

【0061】

以上の情報により、この投稿を始めて採集する時に、この投稿に対して210個のリプライがなされており、合成して取得したページターニングURLは合計10個となる。すなわち、

20

`/postDetail.do?id=91384467&view=1&pageNo=2&boardId=6`

`/postDetail.do?id=91384467&view=1&pageNo=3&boardId=6`

`/postDetail.do?id=91384467&view=1&pageNo=4&boardId=6`

.....

`/postDetail.do?id=91384467&view=1&pageNo=11&boardId=6`

30

となる。

【0062】

事例2

百度投稿バーの投稿URLは `http://tieba.baidu.com/f?kz=919731090` となる。

そのページターニングのリンク規則は

`/f?z=919731090&ct=335544320&lm=0&sc=0&rn=30&tn=baiduPostBrowser&word=%B6%B7%C6%C6%B2%D4%F1%B7&pn=30` となる。

【0063】

40

開始ページ番号は0であり、ページターニング基数は30である。ページターニングのリンク規則により、抽出されたページターニングURLの第1部分は `/f?z=919731090&ct=335544320&lm=0&sc=0&rn=30&tn=baiduPostBrowser&word=%B6%B7%C6%C6%B2%D4%F1%B7` となる。

第3部分は内容無しである。

【0064】

`NPerPage` は30である。以上の情報により、たとえば、この投稿を始めて採集する時、この投稿に対して210個のリプライがなされた場合に、合成して取得されたページターニングURLは6個となり、すなわち、

50

/ f ? z = 9 1 9 7 3 1 0 9 0 & c t = 3 3 5 5 4 4 3 2 0 & l m = 0 & s c = 0 & r n
= 3 0 & t n = b a i d u P o s t B r o w s e r & w o r d = % B 6 % B 7 % C 6 % C
6 % B 2 % D 4 % F 1 % B 7 & p n = 3 0

/ f ? z = 9 1 9 7 3 1 0 9 0 & c t = 3 3 5 5 4 4 3 2 0 & l m = 0 & s c = 0 & r n
= 3 0 & t n = b a i d u P o s t B r o w s e r & w o r d = % B 6 % B 7 % C 6 % C
6 % B 2 % D 4 % F 1 % B 7 & p n = 6 0

/ f ? z = 9 1 9 7 3 1 0 9 0 & c t = 3 3 5 5 4 4 3 2 0 & l m = 0 & s c = 0 & r n
= 3 0 & t n = b a i d u P o s t B r o w s e r & w o r d = % B 6 % B 7 % C 6 % C
6 % B 2 % D 4 % F 1 % B 7 & p n = 9 0

.....

/ f ? z = 9 1 9 7 3 1 0 9 0 & c t = 3 3 5 5 4 4 3 2 0 & l m = 0 & s c = 0 & r n
= 3 0 & t n = b a i d u P o s t B r o w s e r & w o r d = % B 6 % B 7 % C 6 % C
6 % B 2 % D 4 % F 1 % B 7 & p n = 2 1 0 となる。

【 0 0 6 5 】

最後に、合成された全てのページターニングURLをドメイン名付きの完全なURLとする。更に、このような処理がなされた全てのページターニングURLを内容ページ採集キューに追加する。

【 0 0 6 6 】

電子掲示板のページターニングモードが次ページターニングである場合に、ウェブページ内容からページターニングURLを抽出する。ウェブページ内容にページターニングURLがない場合、このページは最後のページであるとして、ページターニングがない。

【 0 0 6 7 】

ページターニングモードが次ページターニングである場合は、ウェブページ内容からページターニングURLを抽出してから、内容ページ採集キューに追加する前に、重複除去手段125は、ページターニングURLの重複除去処理を行う。具体的な処理過程は以下の通りである。

【 0 0 6 8 】

ページターニングURL情報リストにこのページターニングURLが属する投稿が存在するかどうかを調べる。存在しないと、このページターニングURLが属する投稿のページターニング情報をページターニングURL情報リストに追加し、ページターニングURLを内容ページ採集キューに追加する。存在すると、この投稿の現在のページターニングのページ番号が、ページターニングURL情報リストに記録したこの投稿のターニングページ番号より大きいかどうかを判定する。大きい場合に、ページターニングURL情報リストにおいてこの投稿のターニングページ番号を現在のターニングページ番号に更新し、ページターニングURLを内容ページ採集ジョブキューに追加する。大きくなければ、ページターニングURL情報リストにおけるこの投稿のターニングページ番号の更新が必要なく、直接にこのページターニングURLを削除すればよい。

【 0 0 6 9 】

上記ページターニングURL情報リストに、投稿のトップページURL（或いは標識情報）と、現在採集したターニングページ番号と、現在採集したページにおける最後のライの位置と、現在採集したページターニングURLとが記憶される。この情報リストのヘッド構造は以下の通りである。

【 0 0 7 0 】

10

20

30

40

【表 2】

投稿のトップページURL (又は標識情報)	ページター ニングのペ ージ番号	最後のリプライの位置	ページターニングURL
--------------------------	------------------------	------------	-------------

【0071】

所属分野の技術者は、本発明の主旨及び特許請求の範囲から逸脱しない限り、様々な変更や変形を行うことができる。このような変更や変形は本発明の特許請求の範囲並びにこれと均等する範囲内に属すると理解すべきである。

10

【図 1】

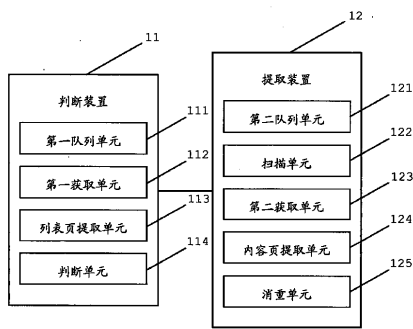
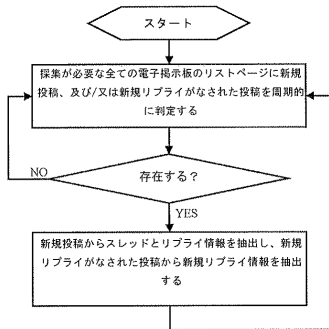
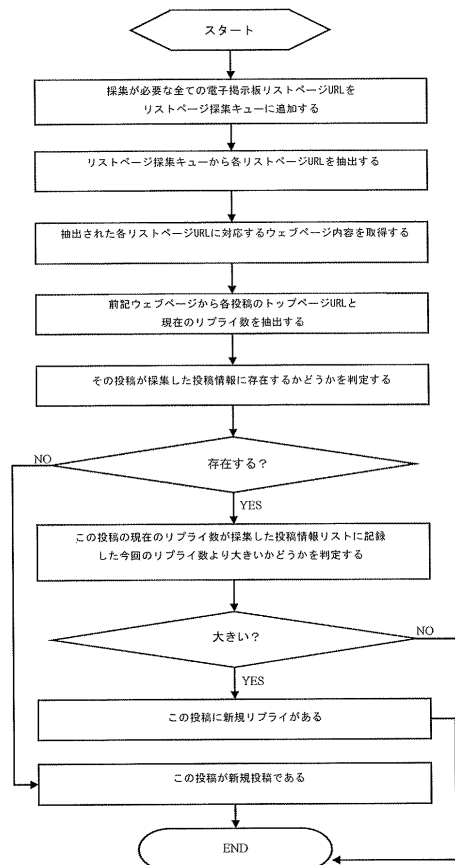


图 1

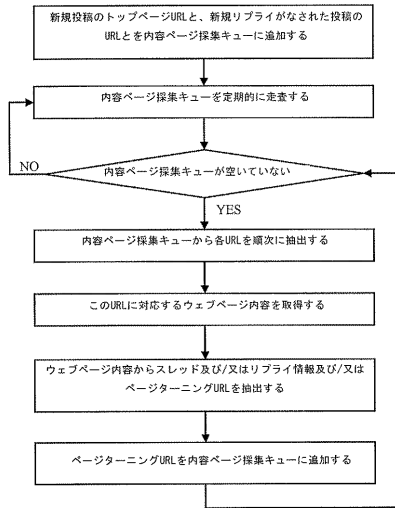
【图 2】



【图 3】



【 図 4 】



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

(73)特許権者 513157039

北京北大方正技 術 研究院有限公司

PEKING UNIVERSITY FOUNDER R & D CENTER

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4 層

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

中華人民共和国北京市 海 淀区上地五街9号方正大厦

(72)発明者 ヤン ジエンウー

中華人民共和国北京市 海 淀区上地五街9号方正大厦

審査官 野崎 大進

(56)参考文献 特開2004-246785(JP,A)

特開2009-230663(JP,A)

特開平08-330991(JP,A)

中国特許出願公開第101727486(CN,A)

中国特許出願公開第101193038(CN,A)

中国特許出願公開第101335639(CN,A)

(58)調査した分野(Int.Cl., DB名)

G06F 17/30

G06F 13/00

JSTPlus(JDreamIII)