(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2014-506355
(P2014-506355A)
(43)公表日　平成26年3月13日(2014.3.13)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	180Z	5B084
G06F 13/00	(2006.01)	G06F	13/00	560A	

審査請求 未請求　予備審査請求 未請求　（全 23 頁）

(21)出願番号　特願2013-545030 (P2013-545030)
(86)(22)出願日　平成23年12月22日 (2011.12.22)
(85)翻訳文提出日　平成25年8月22日 (2013.8.22)
(86)国際出願番号　PCT/CN2011/084457
(87)国際公開番号　WO2012/083870
(87)国際公開日　平成24年6月28日 (2012.6.28)
(31)優先権主張番号　201010618393.4
(32)優先日　平成22年12月22日 (2010.12.22)
(33)優先権主張国　中国 (CN)

(71)出願人　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】電子掲示板リプライ増加量の採集方法及びシステム

(57)【要約】
本発明は迅速に正確に完全に一つの投稿の全てのスレッド・リプライの情報を採集し、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消することができる電子掲示板リプライの増加量の採集方法及びシステムを提供する。本発明は、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が存在するかどうかを周期的に判定し、存在すると判定されたときに、新規投稿からスレッドとリプライを抽出し、新規リプライがなされた投稿からリプライ情報を抽出する。

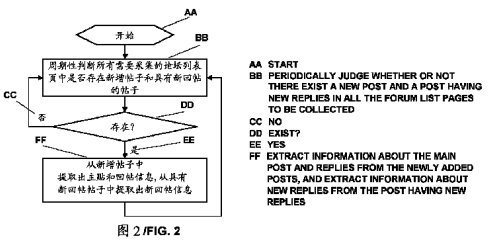

图 2/FIG. 2

【特許請求の範囲】
【請求項１】
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、新規投稿からスレッド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リプライ増加量の採集方法。
【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リプライ増加量の採集方法。
【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いていなければ、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リプライ増加量の採集方法。
【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リプライ増加量の採集方法。
【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リプライ増加量の採集方法。
【請求項６】
新規投稿からスレッド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いていないと、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スレッド及び／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リプライ増加量の採集方法。
【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リプライ増加量の採集方法。
【請求項８】
前記内容ページ採集キューが空いていないと、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リプライ増加量の採集方法。
【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スレッド及び／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スレッド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】

$$S'_{From}=\begin{cases}R_{PreNum}, & N_{PerPage}\ \text{はリプライを含む}\\ R_{PreNum}+1, & N_{PerPage}\ \text{はリプライを含まない}\end{cases}$$

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

$$P_{Begin}=\begin{cases} ceil\left(\dfrac{R_{PreNum}+1}{N_{PerPage}}\right), & N_{PerPage}\text{はリプライを含む} \\ ceil\left(\dfrac{R_{PreNum}}{N_{PerPage}}\right), & N_{PerPage}\text{はリプライを含まない} \end{cases}$$

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、

【数３】

$$S''_{From}=\begin{cases} R_{PreNum}\%N_{PerPage}+1, & N_{PerPage}\text{はリプライを含む} \\ R_{PreNum}\%N_{PerPage}, & N_{PerPage}\text{はリプライを含まない} \end{cases}$$

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

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】

$$C'''_{ParseCount}=\begin{cases} R_{CurNum}\%N_{PerPage}+1, & N_{PerPage}\text{はリプライを含む} \\ \begin{cases}(R_{CurNum}-1)\%N_{PerPage}+1, & \text{もし} R_{CurNum}-1>0 \\ 0, & \text{もし} R_{CurNum}-1=0\end{cases} & \text{ならば } NPerPage\text{はリプライを含まない，このページは終了ページである} \\ N_{PerPage}, & \text{このページは 終了ではない} \end{cases}$$

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リプライ増加量の採集方法。

【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スレッド及び／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】

$$P_{Begin}=\begin{cases} ceil\left(\frac{R_{PreNum}+1}{N_{PerPage}}\right), & N_{PerPage}はリプライを含む \\ ceil\left(\frac{R_{PreNum}}{N_{PerPage}}\right), & N_{PerPage}はリプライを含まない \end{cases}$$

$$P_{End}=\begin{cases} ceil\left(\frac{R_{CurNum}+1}{N_{PerPage}}\right), & N_{PerPage}はリプライを含む \\ ceil\left(\frac{R_{CurNum}}{N_{PerPage}}\right), & N_{PerPage}はリプライを含まない \end{cases}$$

$$S_{From}=\begin{cases} R_{PreNum}\%N_{PerPage}+1, & N_{PerPage}はリプライを含む \\ R_{PreNum}\%N_{PerPage}, & N_{PerPage}はリプライを含まない \end{cases}$$

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リプライ増加量の採集方法。

【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リプライ増加量の採集方法。

【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リプライ増加量の採集方法。
【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リプライ増加量の採集方法。
【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、
新規投稿に対しては、当該新規投稿からスレッド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リプライ増加量の採集システム。
【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リプライ増加量の採集システム。
【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スレッド及び／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リプライ増加量の採集システム。
【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リプライ増加量の採集システム。
【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、電子掲示板リプライ（ｒｅｐｌｙ）増加量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ト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スレッド（ｔｈｒｅａｄ）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、スレッドへのリプライの情報を採集していない。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
本発明は従来の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スレッド・リプライの情報を採集することができる電子掲示板リプライ増加量の採集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。
【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リプライ増加量の採集方法であって、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、新規投稿からスレッド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。
また、前記の目的を達成するために、本発明は電子掲示板リプライ増加量の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、当該新規投稿からスレッド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】
【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】
【０００６】
【図１】具体的な実施形態における電子掲示板リプライ増加量の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リプライ増加量の採集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スレッド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。
【発明を実施するための形態】
【０００７】
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。
【０００８】
図１に示すように、本実施形態にかかる電子掲示板リプライ増加量の採集システムは、判定装置１１と、判定装置１１に接続される抽出装置１２とを備えている。その中、判定装置１１は、第１キュー（ｑｕｅｕｅ）手段１１１と、第１取得手段１１２と、リストページ抽出手段１１３と、判定手段１１４とを含んでいる。抽出装置１２は、第２キュー手段１２１と、走査手段１２２と、第２取得手段１２３と、内容ページ抽出手段１２４と、重複除去手段１２５とを含んでいる。
【０００９】
判定装置１１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。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。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１１４は、投稿のトップページＵ

ＲＬ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。

【００１０】

抽出装置１２は、新規投稿に対して、新規投稿からスレッド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１２１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１２２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１２３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１２４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スレッド及び／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１２５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１２１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。

【００１１】

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リプライ増加量の採集方法は、以下の各ステップを含む。

【００１２】

（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。

【００１３】

（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／ｌ－ｅｎｊｏｙ－０－０－０－０．ｈｉｍｌである。

【００１４】

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。

【００１５】

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。

【００１６】

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。

【００１７】

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。

【００１８】

（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。

【００１９】

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。

【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回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回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がを、採集した投稿情報リストに追加する。この投稿の前回のリプライ数が０、今回のリプライ数が現在のリプライ数である。

【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、採集した投稿の前回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	前回のリプライ数	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

【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は新規投稿からスレッド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。
【００２６】
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。
【００２７】
（ｉ）第２キュー手段１２１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。
【００２８】
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。
【００２９】
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。
【００３０】
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。
【００３１】
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、ｈｔｔ

ｐ：／／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。

【００３２】

（ｉｉ）走査手段１２２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。

【００３３】

（ｉｉｉ）内容ページ採集キューが空いていないと、第２取得手段１２３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。

【００３４】

本実施形態において、第２取得手段１２３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１１２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。

【００３５】

（ｉｖ）内容ページ抽出手段１２４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スレッド及び／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。

【００３６】

本実施形態において、ウェブページ内容からスレッド及び／又はリプライを抽出する具体な方法は以下のとおりである。

【００３７】

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スレッドとリプライ情報を抽出する。具体的には、まず、この投稿においてスレッド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の情報をスレッドとし、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スレッドの情報を抽出してから、各リプライの情報を抽出する。スレッド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ワード又は正規表現である。

【００３８】

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。

【００３９】

【数１】

$$S'_{From}=\begin{cases}R_{PreNum}, & N_{PerPage}\text{はリプライを含む}\\ R_{PreNum}+1, & N_{PerPage}\text{はリプライを含まない}\end{cases}\quad A$$

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

【００４０】

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。

【００４１】

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。

【００４２】

【数２】

$$P_{Begin}=\begin{cases} ceil\left(\dfrac{R_{PreNum}+1}{N_{PerPage}}\right), & N_{PerPage}\ \text{はリプライを含む} \\ ceil\left(\dfrac{R_{PreNum}}{N_{PerPage}}\right), & N_{PerPage}\ \text{はリプライを含まない} \end{cases}$$

【００４３】

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。

【００４４】

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。

【００４５】

【数３】

$$S''_{From}=\begin{cases} R_{PreNum}\%N_{PerPage}+1, & N_{PerPage}\text{はリプライを含む} \\ R_{PreNum}\%N_{PerPage}, & N_{PerPage}\text{はリプライを含まない} \end{cases}$$

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

【００４６】

ここで、「％」は剰余演算である。

【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】

$$C'''_{ParseCount}=\begin{cases} R_{CurNum}\%N_{PerPage}+1, & N_{PerPage}\text{はリプライを含む} \\ \begin{cases} (R_{CurNum}-1)\%N_{PerPage}+1, & \text{もし} R_{CurNum}-1>0 \\ 0, & \text{もし} R_{CurNum}-1=0 \end{cases} & \text{ならば}\ NPerPage\text{はリプライを含まない}, \quad \text{このページは終了ページである} \\ N_{PerPage}, & \text{このページは 終了ではない} \end{cases}$$

【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な方法を採用している。

【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】

$$P_{Begin}=\begin{cases} ceil\left(\dfrac{R_{PreNum}+1}{N_{PerPage}}\right), & N_{PerPage}\text{はリプライを含む} \\ ceil\left(\dfrac{R_{PreNum}}{N_{PerPage}}\right), & N_{PerPage}\text{はリプライを含まない} \end{cases}$$

$$P_{End}=\begin{cases} ceil\left(\dfrac{R_{CurNum}+1}{N_{PerPage}}\right), & N_{PerPage}\ \text{はリプライを含む} \\ ceil\left(\dfrac{R_{CurNum}}{N_{PerPage}}\right), & N_{PerPage}\ \text{はリプライを含まない} \end{cases}$$

$$S_{From}=\begin{cases} R_{PreNum}\%N_{PerPage}+1, & N_{PerPage}\text{はリプライを含む} \\ R_{PreNum}\%N_{PerPage}, & N_{PerPage}\text{はリプライを含まない} \end{cases}$$

【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。

【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、ページターニングド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。

```
ｆｏｒ（ｉｎｔ　ｉ＝Ｐ$_{Begin}$；ｉ＜Ｐ$_{End}$；＋＋）
{
    ｎＰａｇｅＮｏ＝ｉ＋ｎＦｉｒｓｔＰｏｓｔＰａｇｅＩｎｄｅｘ－１；
    ｎＰａｇｅＵｐ＝（ｎＰａｇｅＮｏ×ｎＰａｇｅＵｓＢａｓｅＮｕｍ）；
    ｓｔｒＰｏｓｔＰａｇｅＵｒｌ＝ｓｔｒＢｅｆｏｒｅＰａｇｅ＋ｎＰａｇｅＵｐ＋ｓｔｒＡｆｔｅｒＰａｇｅ；
}
```

【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】

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。

【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、

／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

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】

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＆ｐｎ＝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、
／ｆ？ｚ＝９１９７３１０９０＆ｃｔ＝３３５５４４３２０＆ｌｍ＝０＆ｓｃ＝０＆ｒｎ＝３０＆ｔｎ＝ｂａｉｄｕＰｏｓｔＢｒｏｗｓｅｒ＆ｗｏｒｄ＝％Ｂ６％Ｂ７％Ｃ６％Ｃ６％Ｂ２％Ｄ４％Ｆ１％Ｂ７＆ｐｎ＝３０
／ｆ？ｚ＝９１９７３１０９０＆ｃｔ＝３３５５４４３２０＆ｌｍ＝０＆ｓｃ＝０＆ｒｎ＝３０＆ｔｎ＝ｂａｉｄｕＰｏｓｔＢｒｏｗｓｅｒ＆ｗｏｒｄ＝％Ｂ６％Ｂ７％Ｃ６％Ｃ６％Ｂ２％Ｄ４％Ｆ１％Ｂ７＆ｐｎ＝６０
／ｆ？ｚ＝９１９７３１０９０＆ｃｔ＝３３５５４４３２０＆ｌｍ＝０＆ｓｃ＝０＆ｒｎ＝３０＆ｔｎ＝ｂａｉｄｕＰｏｓｔＢｒｏｗｓｅｒ＆ｗｏｒｄ＝％Ｂ６％Ｂ７％Ｃ６％Ｃ６％Ｂ２％Ｄ４％Ｆ１％Ｂ７＆ｐｎ＝９０
……
／ｆ？ｚ＝９１９７３１０９０＆ｃｔ＝３３５５４４３２０＆ｌｍ＝０＆ｓｃ＝０＆ｒｎ＝３０＆ｔｎ＝ｂａｉｄｕＰｏｓｔＢｒｏｗｓｅｒ＆ｗｏｒｄ＝％Ｂ６％Ｂ７％Ｃ６％Ｃ６％Ｂ２％Ｄ４％Ｆ１％Ｂ７＆ｐｎ＝２１０となる。
【００６５】
最後に、合成された全てのページターニングＵＲＬをドメインネーム付きの完全な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。
【００６６】
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。
【００６７】
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。
【００６８】
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。
【００６９】
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リプライの位置と、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。
【００７０】

【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。

【図１】

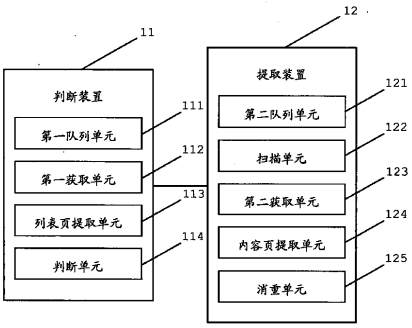

图 1

【図２】

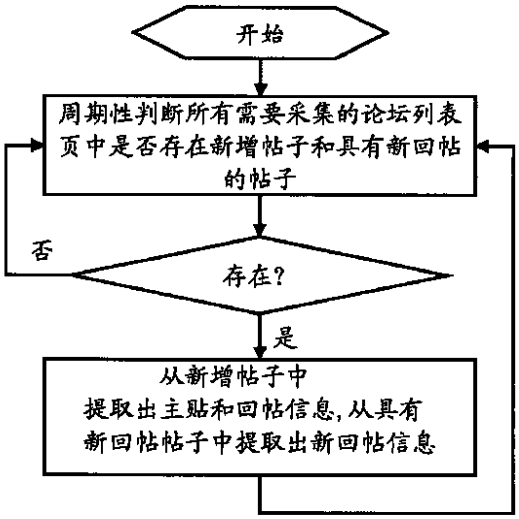

图 2

【図３】

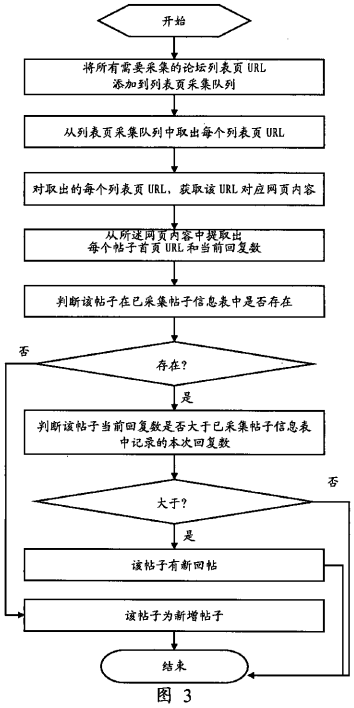
开始
将所有需要采集的论坛列表页 URL
添加到列表页采集队列
从列表页采集队列中取出每个列表页 URL
对取出的每个列表页 URL，获取该 URL 对应网页内容
从所述网页内容中提取出
每个帖子首页 URL 和当前回复数
判断该帖子在已采集帖子信息表中是否存在
否
存在？
是
判断该帖子当前回复数是否大于已采集帖子信息表
中记录的本次回复数
否
大于？
是
该帖子有新回帖
该帖子为新增帖子
结束

图 3

【図４】

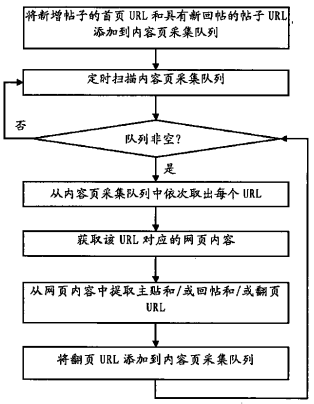
将新增帖子的首页 URL 和具有新回帖的帖子 URL
添加到内容页采集队列
定时扫描内容页采集队列
否
队列非空？
是
从内容页采集队列中依次取出每个 URL
获取该 URL 对应的网页内容
从网页内容中提取主贴和/或回帖和/或翻页
URL
将翻页 URL 添加到内容页采集队列

图 4

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2011/084457

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CPRSABS, CJFD, VEN: reply, gather, newly increased, forum, blog, BBS, replay, post, return, URL, collect, increase

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 101727486 A (THE PLA INFORMATION ENGINEERING UNIVERSITY), 09 June 2010 (09.06.2010), the whole document	1-17
A	CN 101193038 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.), 04 June 2008 (04.06.2008), the whole document	1-17
A	CN 101335639 A (WEN, Guihua et al.), 31 December 2008 (31.12.2008), the whole document	1-17

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 March 2012 (10.03.2012)	**05 April 2012 (05.04.2012)**
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer **TIAN, Zhigang** Telephone No.: (86-10) **62411705**

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2011/084457

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 101727486 A	09.06.2010	None	
CN 101193038 A	04.06.2008	CN 101193038 B	22.12.2010
CN 101335639 A	31.12.2008	None	

Form PCT/ISA/210 (patent family annex) (July 2009)

国际检索报告

国际申请号
PCT/CN2011/084457

A. 主题的分类

G06F 17/30 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CPRSABS,CJFD,VEN: 论坛,博客,BBS,URL,回帖,回复,答复,跟帖,采集,收集,增量,新增 forum, blog, BBS, replay, post, return, URL, collect, increase

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 101727486 A (中国人民解放军信息工程大学) 09.6 月 2010 (09.06.2010) 全文	1-17
A	CN 101193038 A (腾讯科技（深圳）有限公司) 04.6 月 2008 (04.06.2008) 全文	1-17
A	CN 101335639 A (文贵华 等) 31.12 月 2008 (31.12.2008) 全文	1-17

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型：
"A" 认为不特别相关的表示了现有技术一般状态的文件
"E" 在国际申请日的当天或之后公布的在先申请或专利
"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）
"O" 涉及口头公开、使用、展览或其他方式公开的文件
"P" 公布日先于国际申请日但迟于所要求的优先权日的文件
"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件
"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性
"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性
"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
10.3 月 2012 (10.03.2012)	**05.4 月 2012 (05.04.2012)**
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号： (86-10)62019451	受权官员 田志刚 电话号码：(86-10) **62411705**

PCT/ISA/210 表(第 2 页) (2009 年 7 月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2011/084457

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN 101727486 A	09.06.2010	无	
CN 101193038 A	04.06.2008	CN 101193038 B	22.12.2010
CN 101335639 A	31.12.2008	无	

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IS,JP,KE,KG,KM,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LT,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN

(71)出願人 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
(71)出願人 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 １００８５， Ｃｈｉｎａ
(71)出願人 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

Ｆターム(参考) 5B084 AA17 AA26 AB04 BB19 DB02 DC04 EA33