(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2014-532220
(P2014-532220A)
(43)公表日　平成26年12月4日(2014.12.4)

(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	540F	

審査請求 有　予備審査請求 未請求　（全 21 頁）

(21)出願番号　特願2014-532240 (P2014-532240)
(86)(22)出願日　平成24年12月13日 (2012.12.13)
(85)翻訳文提出日　平成26年3月28日 (2014.3.28)
(86)国際出願番号　PCT/CN2012/086575
(87)国際公開番号　WO2013/087005
(87)国際公開日　平成25年6月20日 (2013.6.20)
(31)優先権主張番号　201110415749.9
(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)

(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】ネットコメントの収集方法およびシステム

(57)【要約】
本発明は、ネットコメントの収集方法およびシステム開示した。当該方法は、前記ウェブページのエントリーリンクアドレスに対応するウェブページにN個ネットコメントを有するか否かを判断する。ここで、Nは正の整数であるステップと；前記N個ネットコメントを有する場合、前記N個ネットコメントにN個ネットコメントに収集の条件を満たすM個ネットコメントを有するか否かを判断する。ここで、前記MはNより小さいまたは大きい正の整数であるステップと；前記収集の条件を満たすM個ネットコメントを有する場合、前記M個ネットコメントを収集するステップとを備える。

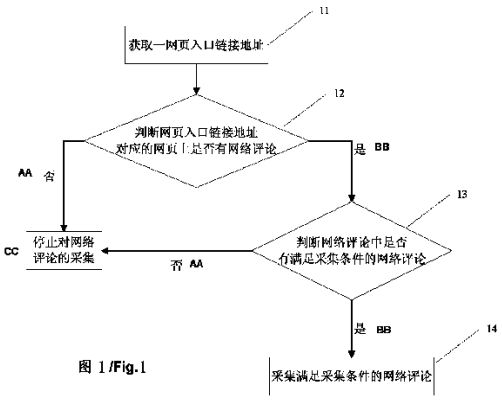


图 1/Fig.1

11 OBTAIN A WEB PAGE ENTRY LINK ADDRESS
12 DETERMINE WHETHER A WEB PAGE CORRESPONDING TO THE WEB PAGE ENTRY LINK ADDRESS HAS NETWORK COMMENTS
13 DETERMINE WHETHER ANY NETWORK COMMENT SATISFYING A COLLECTION CONDITION IS AMONG THE NETWORK COMMENTS
14 COLLECT THE NETWORK COMMENT SATISFYING THE COLLECTION CONDITION
AA NO
BB YES
CC STOP NETWORK COMMENT COLLECTING

【特許請求の範囲】
【請求項１】
ウェブページのエントリーリンクアドレスを取得するステップと、
前記ウェブページのエントリーリンクアドレスに対応するウェブページにＮ個ネットコメントを有するか否かを判断するステップと、
前記Ｎ個ネットコメントを有する場合、前記Ｎ個ネットコメントにＮ個ネットコメントに収集の条件を満たすＭ個ネットコメントを有するか否かを判断するステップと、
前記収集の条件を満たすＭ個ネットコメントを有する場合、前記Ｍ個ネットコメントを収集するステップとを備え，
前記Ｎは正の整数であり、
前記ＭはＮより小さいまたは大きい正の整数であることを特徴とするネットコメントの収集方法。
【請求項２】
前記ウェブページのエントリーリンクアドレスの取得は、
前記Ｎ個ネットコメントがコメントしたトピックがあるトピックページを取得するステップと、
前記トピックページの特性コードを取得するステップと、
前記トピックがあるチャネルの特性コードを取得するステップと、
前記トピックページの特性コードと前記チャネルの特性コードを接合し、ウェブページのエントリーリンクアドレスを取得するステップとを備えることを特徴とする請求項１に記載のネットコメントの収集方法。
【請求項３】
前記方法は、周期的に前記トピックページのエントリーリンクアドレスを更新することをさらに備えることを特徴とする請求項２に記載のネットコメントの収集方法。
【請求項４】
前記方法は、前記ウェブページでのネットコメントが予定時間を越えても更新がない場合、前記ウェブページのエントリーリンクアドレスを削除することをさらに備えることを特徴とする請求項１に記載のネットコメントの収集方法。
【請求項５】
請求項１に記載した前記Ｎ個ネットコメントに収集の条件を満たすＭ個ネットコメントを有するか否かの判断は、具体的に、ＮとＰの差分値を計算し、もしＮがＰより大きければ、新たに増えたネットコメントを有することを示すステップをさらに備ええ、
前記新たに増えたネットコメントの個数はＮとＰの差分値Ｍであり、Ｐは前回前記ページにアクセスした際のネットコメントの個数であることを特徴とする請求項１に記載のネットコメントの収集方法。
【請求項６】
前記方法は、前記ページの目下のページに備えるネットコメントの個数Ｌを計算し、もし前記ＬがＭより小さければ、ページングするページ数を計算し、かつ前記ページ数に対応するページングのリンクを抽出ことさらに備え、
前記Ｌは正の整数であることを特徴とする請求項５に記載のネットコメントの収集方法。
【請求項７】
前記方法は、前記Ｎ個ネットコメントでの各ネットコメントと前記Ｐ個ネットコメントでの各ネットコメントをそれぞれ比較し、もし比較結果が異なれば、前記比較結果が異なるＭ個ネットコメントを抽出することをさらに備えることを特徴とする請求項５に記載のネットコメントの収集方法。
【請求項８】
請求項１に記載した前記Ｎ個ネットコメントに収集の条件を満たすＭ個ネットコメントを有するか否かの判断は、具体的に、前記Ｎ個ネットコメントでの各ネットコメントと前記Ｐ個ネットコメントでの各ネットコメントをそれぞれ比較し、もし比較結果が異なれば

、比較結果が異なるＭ個ネットコメントは収集の条件を満たすネットコメントであることを確認するステップを備えることを特徴とする請求項１に記載のネットコメントの収集方法。
【請求項９】
前記方法は、抽出した前記Ｍ個ネットコメント内容を前記ウェブページと異なるストレージユニットに保存することをさらに備えることを特徴とする請求項１に記載のネットコメントの収集方法。
【請求項１０】
ウェブページのエントリーリンクアドレスを取得する、エントリーリンク取得コンポーネントと、
前記ウェブページのエントリーリンクアドレスに対応するウェブページにＮ個ネットコメントを有するか否かを判断する、第１判断コンポーネントと、
前記Ｎ個ネットコメントを有する場合、前記Ｎ個ネットコメントにＮ個ネットコメントに収集の条件を満たすＭ個ネットコメントを有するか否かを判断する、第２判断コンポーネントと
前記収集の条件を満たすＭ個ネットコメントを有する場合、前記Ｍ個ネットコメントを収集する、内容収集コンポーネントとを備え、
前記Ｎは正の整数であり、
前記ＭはＮより小さいまたは大きい正の整数であることを特徴とするネットコメントの収集システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は情報検索とデータ統合技術分野に関し、特にネットコメントの収集方法およびシステムに関する。
【背景技術】
【０００２】
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２０１１１０４１５７４９．９であり、発明名称が「ネットコメントの収集方法およびシステム」である中国特許出願を基礎である優先権を主張し、その開示の総てをここに取り込む。
【０００３】
現在、インターネット技術の急速な発展に伴い、インターネットはすでに世界で最大のデータベースとなり、人類のほぼ全ての領域を網羅し、人々の情報獲得、情報交換における重要なプラットフォームとなっている。人々が情報を閲覧しやすくするため、インターネットに基づいた情報検索技術もより深い研究と充足した発展を得た。さらに、ネットワーク世論分析、バーティカル検索の評価等のような情報検索に基づいた関連応用も出現した。これら応用技術は全て、まずウェブページをローカルにダウンロードし、不純物を取り除いて分析に必要な内容を抜き出し、最後にこの基礎の上でさらに分析を行う。
【０００４】
インターネットに公表する情報について、ネットワークユーザーが情報をブラウジング後に自身の考えを発表し、当該情報に対するコメントを形成する。目下のインターネットの普遍性、広汎性および即効性により、ネットコメントは大衆のある事柄への見方を一定程度代表すると言える。これは、世論分析に大きな意義と応用空間を持たせる。
【発明の概要】
【発明が解決しようとする課題】
【０００５】
よって、ネットコメントはすでに多様な応用の重要なデータソースの一つとなり、ネットコメントデータソースの収集は最も基本的な条件である。だが、従来技術において、ネットコメントの収集研究はほぼ白紙であり、ネットコメントに対する効率的、全面的な収集技術に欠ける。

【課題を解決するための手段】
【０００６】
本発明に係る実施形態は、効率的、全面的にネットコメントを収集するネットコメントの収集方法およびシステムを提供する。
【０００７】
一方で本発明に係るネットコメントの収集方法は、ウェブページのエントリーリンクアドレス（Ｅｎｔｒｙ ｌｉｎｋ ａｄｄｒｅｓ）を取得するステップと；前記ウェブページのエントリーリンクアドレスに対応するウェブページにＮ個ネットコメントを有するか否かを判断し、Ｎが正の整数であるステップと；前記Ｎ個ネットコメントを有する場合、前記Ｎ個ネットコメントにＮ個ネットコメントに収集の条件を満たすＭ個ネットコメントを有するか否かを判断し、前記ＭはＮより小さいまたは大きい正の整数であるステップと；前記収集の条件を満たすＭ個ネットコメントを有する場合、前記Ｍ個ネットコメントを収集するステップとを備える。
【０００８】
好ましくは、前記ウェブページのエントリーリンクアドレスの取得は、具体的に、前記Ｎ個ネットコメントがコメントしたトピックがあるトピックページを取得するステップと；前記トピックページの特性コードを取得するステップと；前記トピックがあるチャネルの特性コードを取得するステップと；前記トピックページの特性コードと前記チャネルの特性コードを接合し、ウェブページのエントリーリンクアドレスを取得するステップとを備える。
【０００９】
好ましくは、周期的に前記トピックページのエントリーリンクアドレスを更新する。
【００１０】
好ましくは、前記ウェブページでのネットコメントが予定時間を越えても更新がない場合、前記ウェブページのエントリーリンクアドレスを削除する。
【００１１】
好ましくは、上述の前記Ｎ個ネットコメントに収集の条件を満たすＭ個ネットコメントを有するか否かの判断は、具体的に、ＮとＰの差分値を計算し、もしＮがＰより大きければ、新たに増えたネットコメントを有することを示す。かつ、前記新たに増えたネットコメントの個数はＮとＰの差分値Ｍである。ここで、Ｐは前回前記ページにアクセスした際のネットコメントの個数であるステップを備える。
【００１２】
好ましくは、前記ページの目下のページに備えるネットコメントの個数Ｌを計算する。もし前記ＬがＭより小さければ、ページングするページ数を計算する。かつ前記ページ数に対応するページングのリンクを抜き出す。ここで、Ｌは正の整数である。
【００１３】
好ましくは、前記Ｎ個ネットコメントでの各ネットコメントと前記Ｐ個ネットコメントでの各ネットコメントをそれぞれ比較する。もし比較結果が異なれば、前記比較結果が異なるＭ個ネットコメントを抽出する。
【００１４】
好ましくは、上述の前記Ｎ個ネットコメントに収集の条件を満たすＭ個ネットコメントを有するか否かの判断は、具体的に、前記Ｎ個ネットコメントでの各ネットコメントと前記Ｐ個ネットコメントでの各ネットコメントをそれぞれ比較する。もし比較結果が異なれば、比較結果が異なるＭ個ネットコメントは収集の条件を満たすネットコメントであることを確認するステップを備える。
【００１５】
好ましくは、抽出した前記Ｍ個ネットコメント内容を前記ウェブページと異なるストレージユニットに保存する。
【００１６】
また一方で本発明に係るネットコメントの収集システムは、ウェブページのエントリー

リンクアドレスを取得する、エントリーリンク取得コンポーネントと；前記ウェブページのエントリーリンクアドレスに対応するウェブページにＮ個ネットコメントを有するか否かを判断する。ここで、Ｎは正の整数である、第１判断コンポーネントと；前記Ｎ個ネットコメントを有する場合、前記Ｎ個ネットコメントにＮ個ネットコメントに収集の条件を満たすＭ個ネットコメントを有するか否かを判断し、前記ＭはＮより小さいまたは大きい正の整数である、第２判断コンポーネントと；前記収集の条件を満たすＭ個ネットコメントを有する場合、前記Ｍ個ネットコメントを収集する、内容収集コンポーネントとを備える。

【発明の効果】

【００１７】

本発明の有益な効果を以下に記述する。

【００１８】

本発明に係る実施形態はネットコメント収集システムを用いてネットコメントを収集し、ネットコメントのエントリーリンクアドレスの取得および収集の条件の設定により、全面的にネットコメントを収集する技術効果を果たす。

【００１９】

さらに、比較コンポーネントを用いて、今回抽出した全コメントでの各コメントと前回抽出した全コメントでの各コメントの比較を実現できる。そして、内容抽出コンポーネントを用いて、比較結果が異なるコメントのみを抽出するため、全面的にネットコメントを収集する基礎において効率的な収集の効果を果たすことができる。

【図面の簡単な説明】

【００２０】

【図１】本発明の実施形態における収集方法のフロー図。

【図２】本発明の図１における収集方法の詳細なフロー図。

【図３】本発明の図１における収集方法の詳細なフロー図。

【図４】本発明の第１実施形態における収集システムアーキテクチャを示す図。

【図５】本発明の第２実施形態における収集システムアーキテクチャを示す図。

【図６】本発明の第３実施形態における収集システムアーキテクチャを示す図。

【図７】本発明の第４実施形態における収集システムアーキテクチャを示す図。

【図８】本発明の別の実施形態における収集システムアーキテクチャを示す図。

【発明を実施するための形態】

【００２１】

図１は、本発明に係るネットコメントを収集するネットコメントの収集方法はであり、以下のステップ１１からステップ１４を備える。

【００２２】

ステップ１１において、ウェブページのエントリーリンクアドレスを取得する。

【００２３】

ステップ１２において、ウェブページのエントリーリンクアドレスに対応するウェブページにＮ個ネットコメントを有するか否かを判断する。ここでＮは正の整数である。

【００２４】

ステップ１３において、Ｎ個ネットコメントを有する場合、Ｎ個ネットコメントにＮ個ネットコメントに収集の条件を満たすＭ個ネットコメントを有するか否かを判断する。ここで、前記ＭはＮより小さいまたは大きい正の整数である。

【００２５】

ステップ１４において、収集の条件を満たすＭ個ネットコメントを有する場合、Ｍ個ネットコメントを収集する。

【００２６】

ここで、図２に示すように、ステップ１１は具体的にさらに以下のステップ１１１からステップ１１４を備える。

【００２７】

ステップ１１１において、Ｎ個ネットコメントがコメントしたトピックがあるトピックページを取得する。

【００２８】

ステップ１１２において、トピックページの特性コードを取得する。

【００２９】

ステップ１１３において、トピックがあるチャネルの特性コードを取得する。

ステップ１１４において、トピックページの特性コードとチャネルの特性コードを接合（Ｓｐｌｉｃｉｎｇ）し、ウェブページのエントリーリンクアドレスを取得する。

【００３０】

本発明において、トピックページはニュースがあるページでも良く商品情報があるページでも良い。ここではニュースウェブページを例に挙げ、本実施形態を詳細に説明する。実際には、トピックページは他の情報があるページでも良いが、本発明ではこれを制限しない。

【００３１】

本実施形態において、ニュースにコメントするコメントページのエントリーリンクアドレスは、ニュースページのスクリプトにおける特性コードにより特定ルールに従い接合後に取得する。例えば、ニュースに対するネットコメントページのエントリーリンクアドレスは、ニュースページのスクリプトにより当該ニュースを識別する特性コード、当該ニュースがあるチャネル識別する特性コード、さらにドメイン名および一部他の要素（例えば目下の時間）を加えて接合してできる。前記特性コードを取得し、かつ個性的なルールを設定し、指定モデルに基づき、ネットコメントページのエントリーリンクアドレスをマッチングする。

【００３２】

引き続き図２に示すように、ステップ１１はさらに以下のステップ１１５を備える。

【００３３】

ステップ１１５において、周期的にウェブページのエントリーリンクアドレスを更新する。

【００３４】

ステップ１１５において、ニュースウェブページのホームページバックグラウンドはニュースを再編集する可能性があり、同じ内容のニュースウェブページリンクには変化が生じる。即ち、ニュースの識別およびニュースがあるチャネルの特性コードには変化が生じ、ネットコメントのエントリーリンクもこれに伴い変化する。新しいネットコメント内容は変化後のネットコメントのエントリーリンクによりロードする。さらに、これより前に抽出したネットコメントのエントリーリンクアドレスが指定するページには新しいコメントの更新は無いことを意味する。よって、もし元々記録したネットコメントのエントリーリンクを引き続き使用し、アクセスすれば、新たに更新したコメント内容を取得できない。故に当該状況において、周期的に目下記録したニュースページリンクを更新する。もしリンクアドレスが変化すれば、サイトは自動的に変化後のニュースウェブページにジャンプする。こうして、新たに獲得したニュースウェブページに基づき、ネットコメントのエントリーリンク改めて抽出し、引き続き収集できることは明らかである。即ち、ニュースウェブページのエントリーリンクアドレスが更新される場合、ステップ１１１にジャンプし、実行する。そうでなければ、本フローを終了する。

【００３５】

図３に示すように、ステップ１３の具体的なステップは、ステップ１３１からステップ１３３を備える。

【００３６】

ステップ１３１において、ウェブページから目下のネットコメントの個数Ｎを抽出し、ＮとＰの差分値Ｍを計算する。ここで、Ｐは前回アクセスした当該リンクが抽出したネットコメント個数である。

【００３７】

ステップ１３２において、Ｍが０より大きいか否かを判断する。

【００３８】

ステップ１３３において、ステップ１３２の結果が、Ｍが０より大きい場合、Ｍ個ネットコメントを抽出する。

【００３９】

ここで、ステップ１３１におけるウェブページからの目下のネットコメントの個数Ｎの抽出は、正規表現によりウェブページから抽出しても良く、他の方法を使用し、抽出しても良いが、本発明はこれを制限しない。最初にネットコメントを収集する場合、Ｐは０と等しい。

【００４０】

引き続き図３に示すように、ここでステップ１３３は具体的に以下のステップ１３３１からステップ１３３３を備える。

【００４１】

ステップ１３３１において、ページにおける目下のページに備えられるネットコメントの個数Ｌを計算する。ここで、ＬはＭより小さいまたは等しい正の整数である。

【００４２】

ステップ１３３２において、ＬがＭより小さいか否かを判断する。

【００４３】

ステップ１３３３において、ステップ１３３２の結果が、ＬがＭより小さい場合、ページングするページ数を計算する。かつ、ページ数に対応するページングのリンクを抽出する。

【００４４】

ここで、ステップ１３３３において、ページングの計算公式は：

【００４５】

【数１】

$$P_{count} = ceil(\frac{C_{Update} - C_{Current}}{N_{Perpage}})$$

【００４６】

ここで、P_{count}は、ページングするページ数を示し、P_{Update}（即ち、Ｍ）は、コメント更新数を示し、$C_{Current}$（即ち、Ｌ）は、目下のウェブページコメント個数を示し、$N_{Perpage}$は、単数のウェブページコメント数を示す。

【００４７】

引き続き図３に示すように、ステップ１３３はさらに以下のステップ１３３４およびステップ１３３５を備える。

【００４８】

ステップ１３３４において、Ｎ個ネットコメントでの各ネットコメントとＰ個ネットコメントでのネットコメントが同じか否かを判断する。

【００４９】

ステップ１３３５において、ステップ１３３４の結果が、Ｎ個ネットコメントでの各ネットコメントとＰ個ネットコメントでの各ネットコメントが同じである場合、比較結果が異なるＭ個ネットコメントを抽出する。

【００５０】

ステップ１３３５において、抽出したＭ個ネットコメント内容は、コメントウェブページの異なるストレージユニットに保存される。ストレージユニットに保存されたネットコメントは集中ブラウジングしやすく、ユーザー収集後のネットコメントを応用しやすい。

【００５１】

本実施形態において、ニュースには即効性があり、一定時間を越えたニュースは意味が

無いと認識される。同様に、ニュースの附属であるニュースコメントもニュースの失効に伴い失効する。前記原因に基づき、もしネットコメントが予定時間を越えても更新がない場合、当該ニュースコメントリンクを削除し、引き続いて更新はしない。こうして、システムリソースを節約し、より高い作業効率を有することができる。

【００５２】

別の実施形態において、Ｎ個ネットコメントに収集の条件を満たすＭ個ネットコメントを有するか否かを判断する場合、前記実施形態におけるＮとＰの差分値Ｍを計算する方法を用いなくても良い。つまり、Ｎ個ネットコメントでの各ネットコメントとＰ個ネットコメントでの各ネットコメントをそれぞれ直接比較する。もし比較結果が異なれば、前記比較結果が異なるＭ個ネットコメントを抽出する。このような収集方法を用いるのは、ニュースウェブページのホームページバックグラウンドが不定期にネットコメントを削除するためである。例えば、システムの最初の収集は１５ネットコメントを有し、２回目の収集感覚では、一部の原因によりホームページバックグラウンドは１５コメントを全て削除し、同時に３０の新しいコメントを加える。つまり１つのウェブページでは１５コメントしか表示できないため、ネットコメントの第１ページと第２ページのネットコメントは全て新しいと認識できる。収集周期に達する場合、今回収集した３０コメントと前回の１５コメントを比較する。こうして、比較の結果が今回収集した３０コメントと前回の１５コメント全てが異なる。故に、今回３０の新しいコメントを収集する。さらに、今回収集した３０ネットコメント内容はコメントウェブページの異なるストレージユニットに保存される。ストレージユニットに保存されたネットコメントは集中ブラウジングしやすく、ユーザー収集後のネットコメントを応用しやすい。

【００５３】

本発明の第１実施形態に係るネットコメントデータの収集システムは、図４に示すようにである。図４は本実施形態におけるシステムアーキテクチャであり、当該システムは、エントリーリンク取得コンポーネント１０、第１判断コンポーネント２０、第２判断コンポーネント３０および内容収集コンポーネント４０を備える。エントリーリンク取得コンポーネント１０は、ウェブページのエントリーリンクアドレスを取得する。第１判断コンポーネント２０は、ウェブページのエントリーリンクアドレスに対応するウェブページにＮ個ネットコメントを有するか否かを判断する。第２判断コンポーネント３０は収集の条件を満たすＭ個ネットコメントを有するか否かを判断する。内容収集コンポーネント４０は、ネットコメントを収集する。

【００５４】

ここで、エントリーリンク取得コンポーネント１０は、第１獲得ユニット１０１、第２獲得ユニット１０２、第３獲得ユニット１０３および接合ユニット１０４を備える。第１獲得ユニット１０１は、Ｎ個ネットコメントがコメントしたトピックがあるトピックページを取得する。第２獲得ユニット１０２は、トピックページの特性コードを取得する。第３獲得ユニット１０３は、トピックがあるチャネルの特性コードを取得する。接合ユニット１０４は、トピックページの特性コードとチャネルの特性コードを接合し、ウェブページのエントリーリンクアドレスを取得する。

【００５５】

第２判断コンポーネント３０による収集の条件を満たすＭ個ネットコメントを有するか否かの判断は具体的に、ウェブページからＮ個ネットコメントを抽出し、ＮとＰの差分値Ｍを計算する。ここで、Ｐは前回アクセスした当該リンクが抽出したネットコメント個数であるステップをさらに備える。さらに、Ｍが０より大きいか否かを判断する。もしＭが０より大きければ、Ｍ個ネットコメントは収集の条件を満たすコメントであることを示す。第２実施形態において、第１実施形態と異なる点は、システムが周期的にウェブページのエントリーリンクアドレスを更新する、エントリーリンクアドレス更新コンポーネント５０をさらに備えることである。本実施形態において、エントリーリンクアドレス更新コンポーネント５０は、エントリーリンク取得コンポーネント１０と共に運用でき、更新したネットコメントの速やかな収集を実現する。

【００５６】
第３実施形態において、第１、第２実施形態と異なる点は、システムがウェブページでのネットコメントの無更新が予定時間を越えているか否かを判断する。もし超えていれば、ウェブページのエントリーリンクアドレスを削除する、ネットコメントページ更新コンポーネント６０をさらに備えることである。本実施形態において、ネットコメントページ更新コンポーネント６０は、第１判断コンポーネント２０と共に運用でき、システム収集効率を高め、いまだ更新しないネットコメントは収集を放棄できる。
【００５７】
第２と第３実施形態はそれぞれ図５と図６に示すようにである。実際には、２つの実施形態を結合して使用でき、収集は全面的なネットコメントの収集を実現すると同時にシステムの収集効率を高める。第４実施形態において、第１、第２および第３実施形態と異なる点は、内容収集コンポーネント４０がページング抽出コンポーネント４０１、比較コンポーネント４０２、内容抽出コンポーネント４０３およびディスクＩ／Ｏコンポーネント４０４をさらに備えることである。ページング抽出コンポーネント４０１は、ページングするページ数を計算する。かつ、ページ数に対応するページングのリンクを抽出する。比較コンポーネント４０２は、前記Ｎ個ネットコメントでの各ネットコメントと前記Ｐ個ネットコメントでの各ネットコメントをそれぞれ比較する。内容抽出コンポーネント４０３は、比較結果が異なる場合、前記比較結果が異なるネットコメントを抽出する。ディスクＩ／Ｏコンポーネント４０４は、抽出したネットコメント内容をウェブページの異なるストレージユニットに保存する。本実施形態は図７に示すようにである。
【００５８】
本発明に係る別のネットコメントデータの収集システムは図８に示すようにである。図８は、本実施形態におけるシステムアーキテクチャである。
【００５９】
本実施形態と第１実施形態が異なる点は、本実施形態が比較コンポーネント４０２と内容抽出コンポーネント４０３を備えないことである。図８に示すように、本実施形態のシステムは、エントリーリンク取得コンポーネント８０、第１判断コンポーネント８１、第２判断コンポーネント８２および内容収集コンポーネント８３を備える。エントリーリンク取得コンポーネント８０は、ウェブページのエントリーリンクアドレスを取得する。第１判断コンポーネント８１は、ウェブページのエントリーリンクアドレスに対応するウェブページにネットコメントを有するか否かを判断する。第２判断コンポーネント８２は、収集の条件を満たすネットコメントを有するか否かを判断する。内容収集コンポーネント８３は、ネットコメントを収集する。
【００６０】
ここで、エントリーリンク取得コンポーネント８０は、第１獲得ユニット８０１、第２獲得ユニット８０２、第３獲得ユニット８０３および接合ユニット８０４を備える。第１獲得ユニット８０１は、Ｎ個ネットコメントがコメントしたトピックがあるトピックページを取得する。第２獲得ユニット８０２は、トピックページの特性コードを取得する。第３獲得ユニット８０３は、トピックがあるチャネルの特性コードを取得する。接合ユニット８０４は、トピックページの特性コードとチャネルの特性コードを接合し、ウェブページのエントリーリンクアドレスを取得する。
【００６１】
第２判断コンポーネント８２は、前記Ｎ個ネットコメントでの各ネットコメントと前記Ｐ個ネットコメントでの各ネットコメントをそれぞれ比較する。もし比較結果が異なれば、比較結果が異なるＭ個ネットコメントは収集の条件を満たすネットコメントであることを確認する。
【００６２】
内容収集コンポーネント８３は、ページング抽出コンポーネント８３１およびディスクＩ／Ｏコンポーネント８３２を備える。ページング抽出コンポーネント８３１は、ページングするページ数を計算する。かつ、ページ数に対応するページングのリンクを抽出する

。ディスクＩ／Ｏコンポーネント８３２は、抽出したネットコメント内容をウェブページの異なるストレージユニットに保存する。

【００６３】

本実施形態において、エントリーリンク取得コンポーネント８０は、第２実施形態におけるエントリーリンクアドレス更新コンポーネント８４と結合して共に応用でき、比較的全面的なネットコメントの収集を実現する。第１判断コンポーネント８１は、第３実施形態におけるネットコメントページ更新コンポーネント８５と結合して共に応用でき、全面的、効率的にネットコメントの収集を実現する。

【００６４】

前記第１、第２、第３、第４および別の実施形態におけるシステムは、本発明が提供したネットコメント収集方法の実施形態における方法およびその各種変化の形式の記述に基づき、実施できる。明細書を簡潔にするため、ここでは説明を繰り返さない。

【００６５】

本発明の実施形態は、ネットコメント収集システムを用いて、ネットコメントを収集し、ネットコメントのエントリーリンクアドレスの取得および収集の条件の設定により全面的なネットコメントを収集する技術効果を果たす。

【００６６】

さらに、比較コンポーネントを用いて、今回抽出した全コメントでの各コメントと前回抽出した全コメントでの各コメントの比較を実現できる。そして、内容抽出コンポーネントを用いて、比較結果が異なるコメントのみを抽出する。よって、全面的にネットコメントを収集する基礎の上に効率的な収集の効果を果たすことができる。

【００６７】

以上は本発明の実施形態の方法、装置（システム）、およびコンピュータプログラム製品のフロー図および／またはブロック図によって、本発明を記述した。理解すべきことは、コンピュータプログラム指令によって、フロー図および／またはブロック図における各フローおよび／またはブロックと、フロー図および／またはブロック図におけるフローおよび／またはブロックの結合を実現できる。プロセッサはこれらのコンピュータプログラム指令を、汎用コンピュータ、専用コンピュータ、組込み式処理装置、或いは他のプログラム可能なデータ処理装置設備の処理装置器に提供でき、コンピュータ或いは他のプログラム可能なデータ処理装置のプロセッサは、これらのコンピュータプログラム指令を実行し、フロー図における一つ或いは複数のフローおよび／またはブロック図における一つ或いは複数のブロックに指定する機能を実現する。

【００６８】

これらのコンピュータプログラム指令は又、コンピュータ或いは他のプログラム可能なデータ処理装置を特定方式で動作させるコンピュータ読取記憶装置に記憶できる。これによって、指令を含む装置は当該コンピュータ読取記憶装置内の指令を実行でき、フロー図における一つ或いは複数のフローおよび／またはブロック図における一つ或いは複数のブロックに指定する機能を実現する。

【００６９】

これらコンピュータプログラム指令はさらに、コンピュータ或いは他のプログラム可能なデータ処理装置設備に実装もできる。コンピュータプログラム指令が実装されたコンピュータ或いは他のプログラム可能設備は、一連の操作ステップを実行することによって、関連の処理を実現し、コンピュータ或いは他のプログラム可能な設備において実行される指令によって、フロー図における一つ或いは複数のフローおよび／またはブロック図における一つ或いは複数のブロックに指定する機能を実現する。

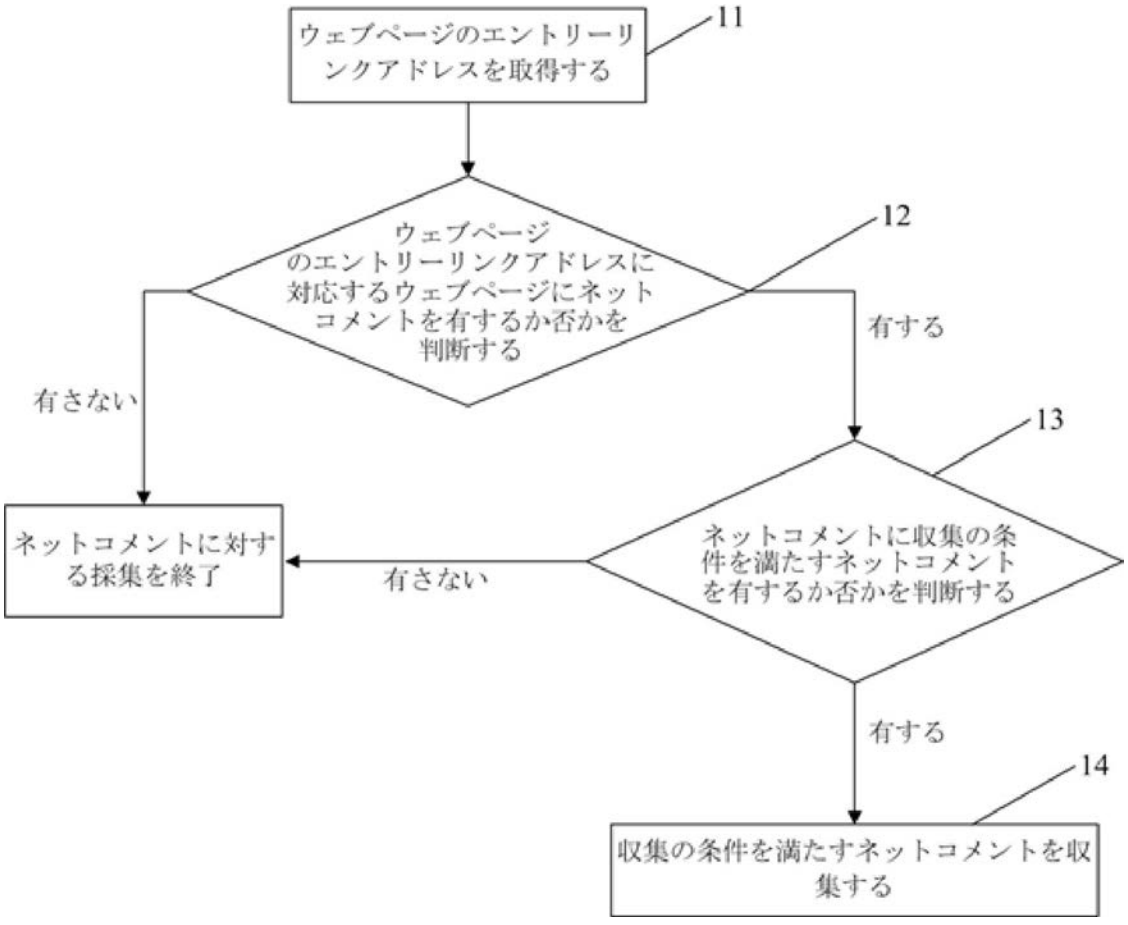
【００７０】

上述した実施形態に記述された技術的な解決手段を改造し、或いはその中の一部の技術要素を置換することもできる。そのような、改造と置換は本発明の各実施形態の技術の範囲から逸脱するとは見なされない。

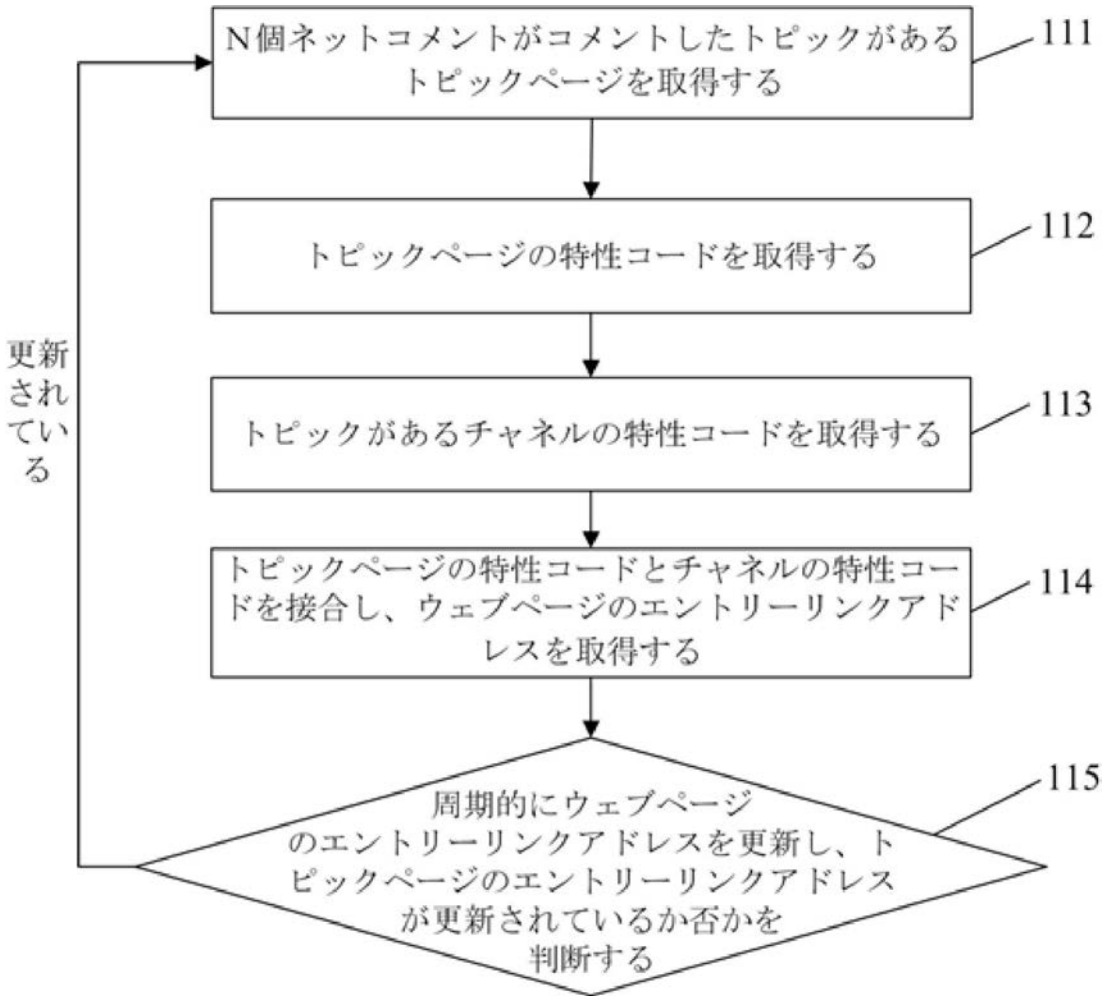
【００７１】

無論、当業者によって、上述した実施形態に記述された技術的な解決手段を改造し、或いはその中の一部の技術要素を置換することもできる。そのような、改造と置換は本発明の各実施形態の技術の範囲から逸脱するとは見なされない。そのような改造と置換は、すべて本発明の請求の範囲に属する。

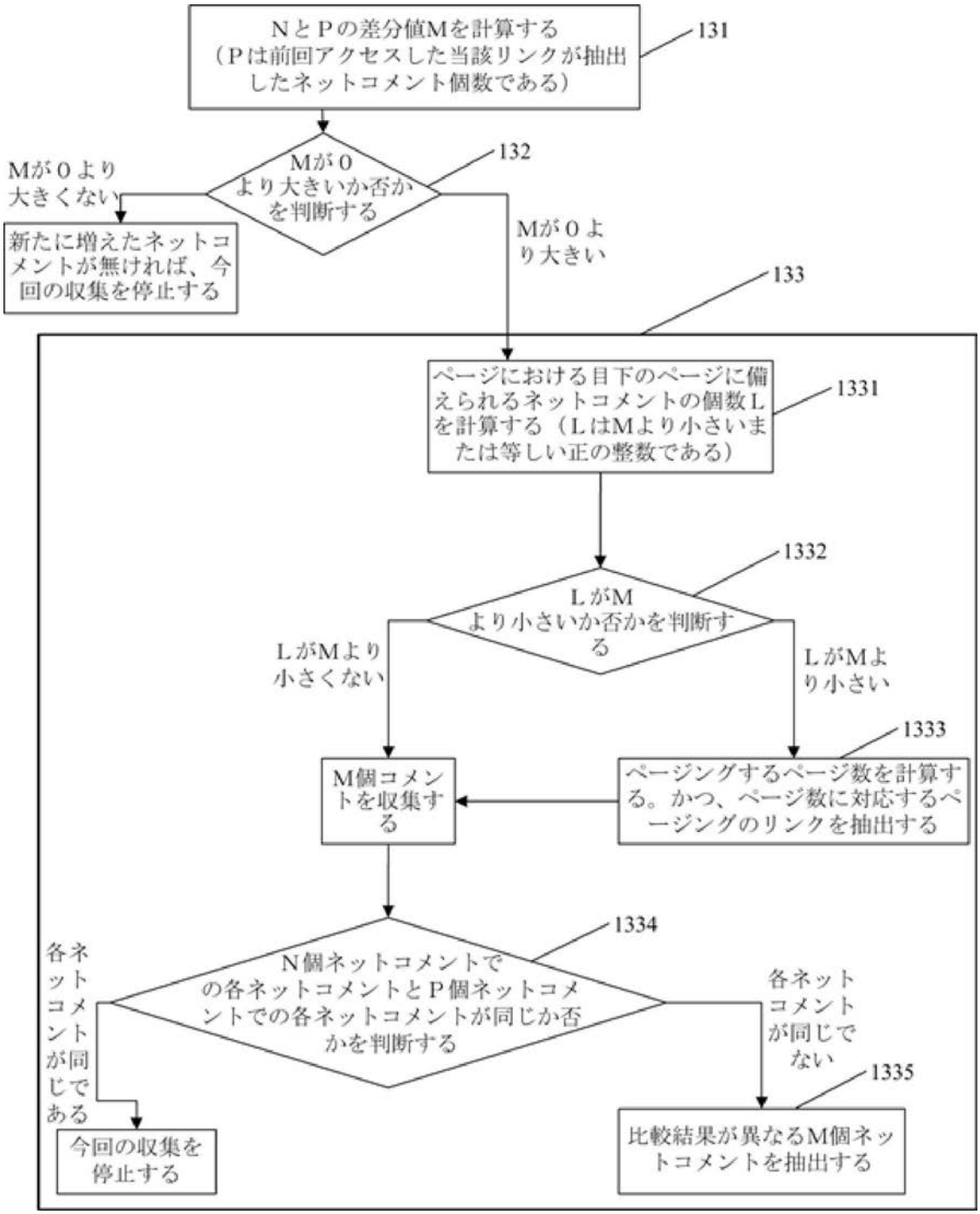
【図１】



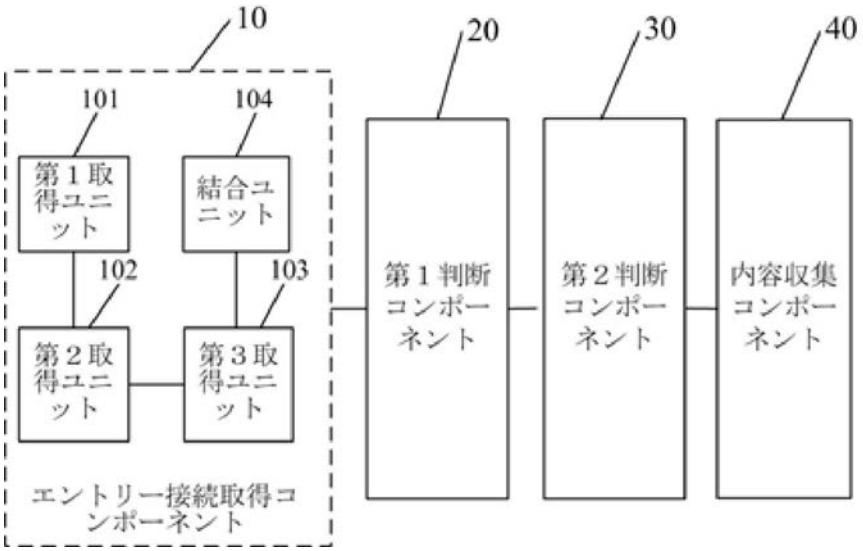
【図２】


N個ネットコメントがコメントしたトピックがあるトピックページを取得する
111
トピックページの特性コードを取得する
112
トピックがあるチャネルの特性コードを取得する
113
トピックページの特性コードとチャネルの特性コードを接合し、ウェブページのエントリーリンクアドレスを取得する
114
周期的にウェブページのエントリーリンクアドレスを更新し、トピックページのエントリーリンクアドレスが更新されているか否かを判断する
115
更新されている

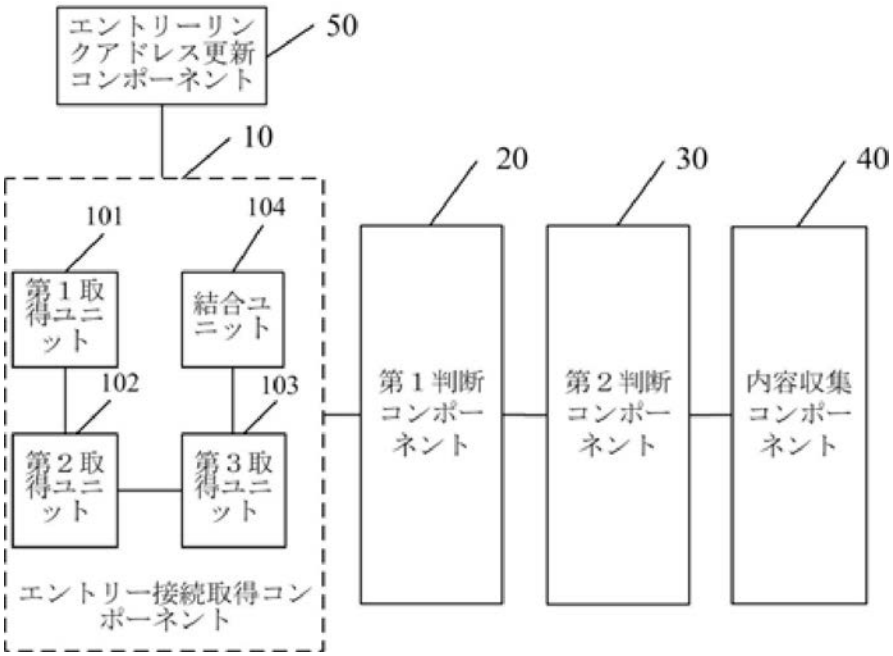
【図３】


ＮとＰの差分値Ｍを計算する（Ｐは前回アクセスした当該リンクが抽出したネットコメント個数である）
131
Ｍが０より大きいか否かを判断する
132
Ｍが０より大きくない
新たに増えたネットコメントが無ければ、今回の収集を停止する
Ｍが０より大きい
133
ページにおける目下のページに備えられるネットコメントの個数Ｌを計算する（ＬはＭより小さいまたは等しい正の整数である）
1331
ＬがＭより小さいか否かを判断する
1332
ＬがＭより小さくない
ＬがＭより小さい
Ｍ個コメントを収集する
ページングするページ数を計算する。かつ、ページ数に対応するページングのリンクを抽出する
1333
Ｎ個ネットコメントでの各ネットコメントとＰ個ネットコメントでの各ネットコメントが同じか否かを判断する
1334
各ネットコメントが同じである
今回の収集を停止する
各ネットコメントが同じでない
比較結果が異なるＭ個ネットコメントを抽出する
1335

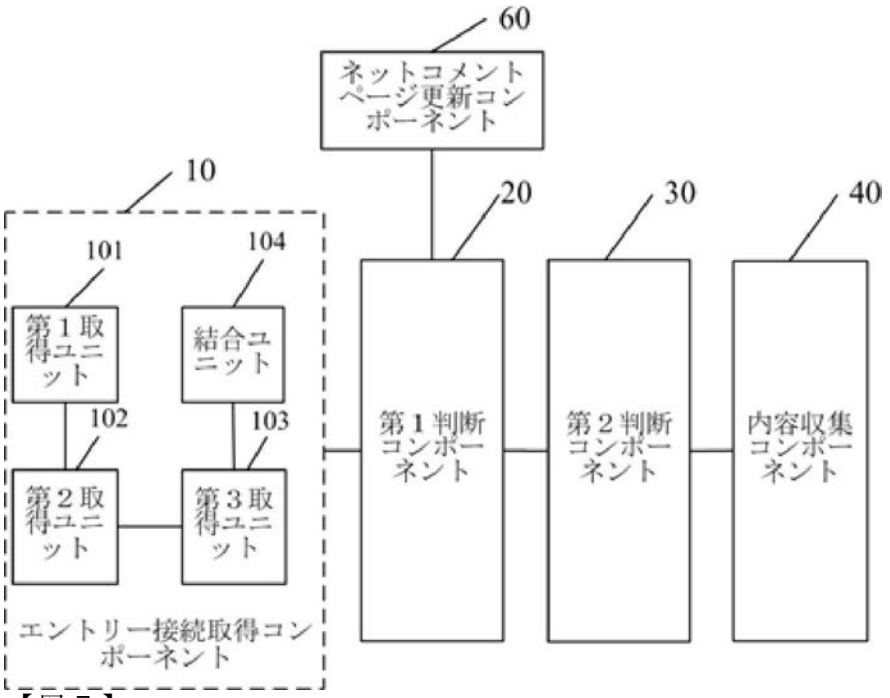
【図４】


10
101
104
第１取得ユニット
結合ユニット
102
103
第２取得ユニット
第３取得ユニット
エントリー接続取得コンポーネント
20
30
40
第１判断コンポーネント
第２判断コンポーネント
内容収集コンポーネント

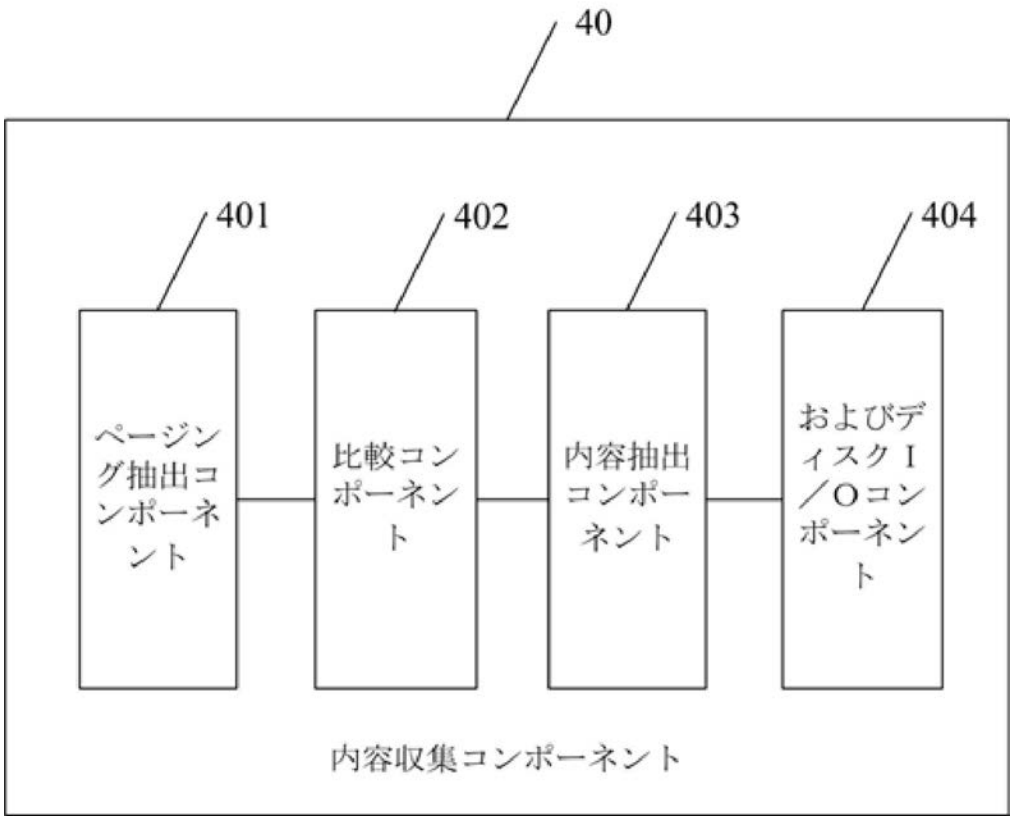
【図５】


エントリーリンクアドレス更新コンポーネント
50
10
101
104
第１取得ユニット
結合ユニット
102
103
第２取得ユニット
第３取得ユニット
エントリー接続取得コンポーネント
20
30
40
第１判断コンポーネント
第２判断コンポーネント
内容収集コンポーネント

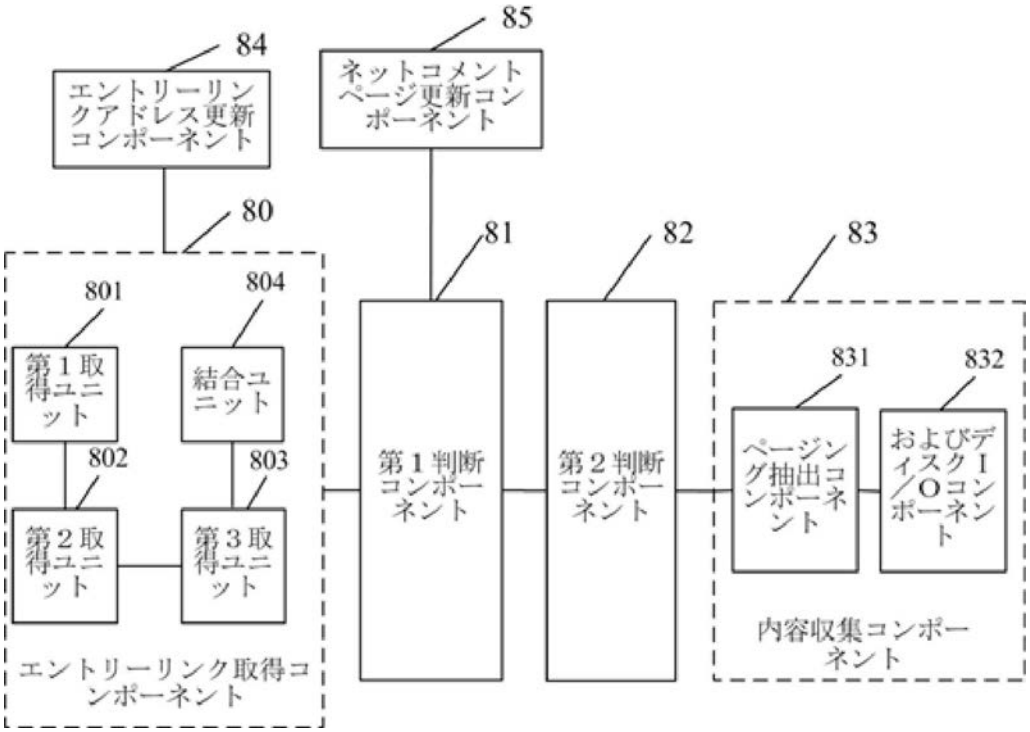
【図６】


60
ネットコメントページ更新コンポーネント
10
101
104
第１取得ユニット
結合ユニット
102
103
第２取得ユニット
第３取得ユニット
エントリー接続取得コンポーネント
20
30
40
第１判断コンポーネント
第２判断コンポーネント
内容収集コンポーネント

【図７】


40
401
402
403
404
ページング抽出コンポーネント
比較コンポーネント
内容抽出コンポーネント
およびディスクＩ／Ｏコンポーネント
内容収集コンポーネント

【図８】


84
エントリーリンクアドレス更新コンポーネント
85
ネットコメントページ更新コンポーネント
80
801
第１取得ユニット
804
結合ユニット
802
第２取得ユニット
803
第３取得ユニット
エントリーリンク取得コンポーネント
81
第１判断コンポーネント
82
第２判断コンポーネント
83
831
ページングコンテンツ抽出ポーネント
832
およびディスクＩ／Ｏコンポーネント
内容収集コンポーネント

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/086575

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: G06F17/-; G06F15/-; H04L12/-; H04M1/-; H04Q7/-; H04M3/-;

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)

CNABS, SIPOABS, DWPI: web, network, data, information, link, criticism, acquisition, review

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 101515269 A (CHINESE ACAD SCI AUTOMATION INST) 26 August 2009 (26.08.2009) description, page 2, line 8 to page 7, line 1,figure 1 and the abstract	1-2, 5-10
Y		3-4
Y	CN 101178713 A (TENCENT SCI&TECHNOLOGY SHENZHEN CO LTD) 14 May 2008 (14.05.2008) description, page 2, line 2, line 15 to line 23	3-4
A	US 6707470 B1 (NEC COPRORATION) 16 March 2004 (16.03.2004) , the whole document	1-10
A	CN 101436196 A (UNIV BEIJING POSTS&TELECOM) 20 May 2009 (20.05.2009) description, page 2, line 8 to page 8, line 24, figure 1 and the abstract	1-10

☐ 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
26 February 2013 (26.02.2013)	**21 March 2013 (21.03.2013)**
Name and mailing address of the ISA 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 GONG , JinLing Telephone No. (86-10) **62413492**

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

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2012/086575

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 101515269 A	26.08.2009	CN 101515269 B	31.08.2011
CN 101178713 A	14.05.2008	None	
US 6707470 B1	16.03.2004	JP 2000330856 A	30.11.2000
CN 101436196 A	20.05.2009	CN 101436196 B	08.12.2010

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

国际检索报告

国际申请号
PCT/CN2012/086575

A. 主题的分类

G06F17/30(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：G06F17/-,G06F15/-;H04L29/-;H04L12/-;H04M1/-;H04Q7/-;H04M3/-.

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）)

CNABS,SIPOABS,DWPI; 网络评论, 采集, 网页, 链接, 评论, web, network, data, information, link, Criticism, acquisition, review,

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
X	CN101515269A（中国科学院自动化研究所)26.8月2009(26.08.2009)，说明书摘要，说明书第2页第8行至说明书第7页第一行，附图1	1-2,5-10
Y		3-4
Y	CN101178713A（腾讯科技（深圳）有限公司)14.5月2008(14.05.2008) 说明书第 2 页第 2 行，第 15-23 行	3-4
A	US6707470B1(NEC CORPORATION)16.3 月 2004(16.03.2004),全文	1-10
A	CN101436196A(北京邮电大学),20.5 月 2009(20.05.2009)，说明书摘要，说明书第 2 页第 8 行至说明书第 8 页最后一行，附图 1	1-10

☐ 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
26.2 月 2013（26.02.2013）	**21.3 月 2013 (21.03.2013)**
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 龚锦玲 电话号码：(86-10) **62413492**

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/CN2012/086575

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN101515269A	26.08.2009	CN101515269B	31.08.2011
CN101178713A	14.05.2008	无	
US6707470B1	16.03.2004	JP2000330856A	30.11.2000
CN101436196A	20.05.2009	CN101436196B	08.12.2010

PCT/ISA/210 表(同族专利附件) (2009 年 7 月)

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,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,BN,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,PA,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

(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
(74)代理人 110001243
特許業務法人 谷・阿部特許事務所
(72)発明者 ジャン タオ
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 ユー シアオミン
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

Ｆターム(参考) 5B084 BB12