(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5002631号
(P5002631)
(45)発行日　平成24年8月15日(2012.8.15)　(24)登録日　平成24年5月25日(2012.5.25)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 180Z
G06F 17/30 210A

請求項の数 8　(全 16 頁)

(21)出願番号　特願2009-204796(P2009-204796)
(22)出願日　平成21年9月4日(2009.9.4)
(65)公開番号　特開2011-54102(P2011-54102A)
(43)公開日　平成23年3月17日(2011.3.17)
審査請求日　平成22年6月8日(2010.6.8)

(73)特許権者　500257300
ヤフー株式会社
東京都港区赤坂９丁目７番１号
(74)代理人　110000637
特許業務法人樹之下知的財産事務所
(72)発明者　山本　芳郎
東京都港区赤坂九丁目７番１号　ヤフー株式会社内
(72)発明者　吉田　享晴
東京都港区赤坂九丁目７番１号　ヤフー株式会社内

審査官　鈴木　和樹

最終頁に続く

(54)【発明の名称】単語情報収集装置、単語情報収集方法および単語情報収集プログラム

(57)【特許請求の範囲】
【請求項１】
ネットワーク上のウェブページに含まれる単語に関する情報を収集し、収集した単語を用いて、検索キーに対してインデックス検索を実行するための検索用インデックスを生成する単語情報収集装置であって、
前記ネットワークを巡回してウェブページに関する情報とともに該ウェブページの更新日時を取得するページ情報取得手段と、
前記取得したウェブページを解析して単語候補を抽出するページ解析手段と、
前記抽出された単語候補と、取得済みの単語候補から予め生成された検索用インデックスとを比較し、前記単語候補が前記検索用インデックスに記憶されているか否かを判定する登録状況判定手段と、
前記判定の結果、前記検索用インデックスに記憶されていないと判定した場合に、前記単語候補と該ウェブページに関する情報とに前記更新日時を初出日時として関連付けて初出ワード記憶手段に記憶させる初出ワード登録手段と、を備える
ことを特徴とする単語情報収集装置。
【請求項２】
請求項１に記載の単語情報収集装置において、
前記初出ワード登録手段は、
前記登録状況判定手段により前記単語候補が前記検索用インデックスに記憶されていると判定した場合、前記単語候補に関連付けられて記憶された初出日時と前記取得した更新

日時とを比較し、前記更新日時が前記初出日時より古いと判定されると、該単語候補の初出日時を前記更新日時で更新する
ことを特徴とする単語情報収集装置。
【請求項３】
請求項１または請求項２に記載の単語情報収集装置において、
前記ネットワークを介して接続された端末装置に対して検索語の入力を要求し、入力された検索語を取得する検索語取得手段と、
前記取得した検索語と一致するキーワードを、前記検索用インデックスから検索し、該当するキーワードに関連付けられたウェブページに関する情報を取得するデータ検索手段と、
前記取得した検索語と一致する単語を、前記初出ワード記憶手段から検索し、該当する単語に関連付けられたウェブページに関する情報と初出日時とを取得する初出ワード検索手段と、
前記データ検索手段により取得したウェブページに関する情報と前記初出ワード検索手段により取得したウェブページに関する情報および初出日時とを表示させたウェブページを作成して配信する検索結果ページ提供手段と、をさらに備えた
ことを特徴とする単語情報収集装置。
【請求項４】
請求項１に記載の単語情報収集装置において、
前記抽出された単語候補と一致する単語が前記初出ワード情報記憶手段に記憶されているか否かを判定する初出ワード登録判定手段をさらに備え、
前記初出ワード登録手段は、前記登録状況判定手段により前記単語候補が前記検索用インデックスに記憶されていないと判定され、かつ、前記初出ワード登録判定手段により前記単語候補が前記初出ワード情報記憶手段に記憶されていないと判定された場合は、前記単語候補と該ウェブページに関する情報とに前記更新日時を初出日時として関連付けて初出ワード記憶手段に記憶させる
ことを特徴とする単語情報収集装置。
【請求項５】
請求項２に記載の単語情報収集装置において、
前記抽出された単語候補と一致する単語が前記初出ワード情報記憶手段に記憶されているか否かを判定する初出ワード登録判定手段をさらに備え、
前記初出ワード登録手段は、
前記登録状況判定手段により前記単語候補が前記検索用インデックスに記憶されていないと判定され、かつ、前記初出ワード登録判定手段により前記単語候補が前記初出ワード情報記憶手段に記憶されていないと判定された場合は、前記単語候補と該ウェブページに関する情報とに前記更新日時を初出日時として関連付けて初出ワード記憶手段に記憶させ、
前記初出ワード登録判定手段により前記単語候補が前記初出ワード情報記憶手段に記憶されていると判定された場合は、前記単語候補に関連付けられて記憶された初出日時と前記取得した更新日時とを比較し、前記更新日時が前記初出日時より古いと判定されると、該単語候補の初出日時を前記更新日時で更新する
ことを特徴とする単語情報収集装置。
【請求項６】
ネットワーク上のウェブページに含まれる単語に関する情報を収集し、収集した単語を用いて、検索キーに対してインデックス検索を実行するための検索用インデックスを生成する単語情報収集方法であって、
前記ネットワークを巡回してウェブページに関する情報とともに該ウェブページの更新日時を取得するページ情報取得ステップと、
前記取得したウェブページを解析して単語候補を抽出するページ解析ステップと、
前記抽出された単語候補と、取得済みの単語候補から予め生成された検索用インデック

スとを比較し、前記単語候補が前記検索用インデックスに記憶されているか否かを判定する登録状況判定ステップと、
前記判定の結果、前記検索用インデックスに記憶されていないと判定した場合に、前記単語候補と該ウェブページに関する情報とに前記更新日時を初出日時として関連付けて初出ワード記憶手段に記憶させる初出ワード登録ステップと、を備える
ことを特徴とする単語情報収集方法。
【請求項７】
請求項６に記載の単語情報収集方法において、
前記初出ワード登録ステップは、
前記登録状況判定ステップにより前記単語候補が前記検索用インデックスに記憶されていると判定した場合、前記単語候補に関連付けられて記憶された初出日時と前記取得した更新日時とを比較し、前記更新日時が前記初出日時より古いと判定されると、該単語候補の初出日時を前記更新日時で更新する
ことを特徴とする単語情報収集方法。
【請求項８】
請求項６または請求項７に記載の単語情報収集方法をコンピュータに実行させることを特徴とする単語情報収集プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ウェブページに含まれる単語に関する情報を収集する単語情報収集装置、単語情報収集方法および単語情報収集プログラムに関する。
【背景技術】
【０００２】
従来、ネットワーク上の検索エンジンによる検索結果ページには、指定された検索語に応じた検索結果の一覧のほかにも様々な情報が表示される。例えば、検索語に関連する広告情報や、検索語に関連のあるショッピングなどの特定のサービス情報などがある。このように、検索結果の一覧だけでなく、様々な情報を表示することでユーザに有益な情報を提供することができるため、検索結果ページにおけるコンテンツのさらなる充実化が求められている。
【０００３】
ここで、検索エンジンは、単語と該単語が含まれるウェブページに関する情報とを関連付けて記憶しており、これらの情報に基づいて、該単語のウェブページにおける出現頻度や重要度に基づいてインデクシング（索引化）した検索用インデックスを作成している。検索時には、この検索用インデックスを参照するため、検索結果ページにはインデクシングの高い（重みの高い）情報が上位に表示され、より有益で意味のある情報をユーザに提供している。
【０００４】
また、検索エンジンは、ネットワークを巡回してウェブページに関する情報を取得してデータベースに蓄積する処理を行う。ウェブページは日々更新されるため、これらの更新情報を速報することは、より有益な情報をユーザに提供することになり、ユーザにとって利便性が高い。このような情報提供を行う方式として、例えば、ホームページを定期、定時に巡回し、その都度ホームページ上の異同を検出、分析を行う技術が知られている（例えば、特許文献１参照）。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２００２－７３６４９号公報
【発明の概要】
【発明が解決しようとする課題】

【０００６】
ところで、検索語としては様々なものを入力でき、例えば、流行語のような珍しい単語を指定する場合がある。流行語は、それが初めて登場したウェブページを発端として流行が広まっている可能性が高く、流行の発端となったウェブページの情報を得たいと思うユーザもいる。
しかしながら、特許文献１に記載の方式では、各ウェブページの最新の情報は得られるものの、特定の単語がウェブページに最初に登場したときの情報を得ることはできない。
【０００７】
本発明の目的は、任意の単語が最初に登場したウェブページに関する情報を簡単に収集でき、収集した情報を用いてウェブページのコンテンツの充実化を図ることのできる単語情報収集装置、単語情報収集方法および単語情報収集プログラムを提供することである。
【課題を解決するための手段】
【０００８】
本発明の単語情報収集装置は、ネットワーク上のウェブページに含まれる単語に関する情報を収集し、収集した単語を用いて、検索キーに対してインデックス検索を実行するための検索用インデックスを生成する単語情報収集装置であって、前記ネットワークを巡回してウェブページに関する情報とともに該ウェブページの更新日時を取得するページ情報取得手段と、前記取得したウェブページを解析して単語候補を抽出するページ解析手段と、前記抽出された単語候補と、取得済みの単語候補から予め生成された検索用インデックスとを比較し、前記単語候補が前記検索用インデックスに記憶されているか否かを判定する登録状況判定手段と、前記判定の結果、前記検索用インデックスに記憶されていないと判定した場合に、前記単語候補と該ウェブページに関する情報とに前記更新日時を初出日時として関連付けて初出ワード記憶手段に記憶させる初出ワード登録手段と、を備えることを特徴とする。
【０００９】
本発明の単語情報収集装置は、ネットワーク上のウェブページに含まれる単語を収集し、収集した単語を用いて、検索キーに対してインデックス検索を実行するための検索用インデックスを生成する装置である。そのために、ページ情報取得手段は、ネットワークを巡回してウェブページに関する情報を取得する（この処理をクロール処理という。）。ここで、ウェブページに関する情報とは、ウェブページのＵＲＬ（Uniform Resource Locator）情報、ウェブページに表示される文章データ、および画像データ等である。このとき、ページ情報取得手段は、該ウェブページの更新日時も同時に取得する。ページ解析手段は、取得したウェブページの文章を解析して単語候補を抽出する。
【００１０】
登録状況判定手段は、抽出した単語候補が、検索用インデックスに登録済みであるか否かを判定する。検索用インデックスは、クロール処理が行われるたびに、それまでに取得した単語候補全体に対して作成されるものである。初出ワード登録手段は、登録状況判定手段により未登録と判定された単語候補を初出ワード記憶手段へ記憶させる。このとき、単語候補には、初出日時として、ページ情報取得手段により取得した該ウェブページの更新日時が関連付けられ、さらに該ウェブページのＵＲＬ情報や該ウェブページに表示された文章データや画像データ等のウェブページに関する情報が関連付けられて記憶される。
【００１１】
本発明では、ネットワークを通して収集した単語に対して検索用インデックスを作成するという通常の処理を行いながら、一方で、抽出した単語候補に関する初出情報を収集する。初出情報とは、任意の単語が最初にウェブページに登場したときの日時や該ウェブページに関する情報である。すなわち、単語情報収集装置が通常実施するクロール処理やインデックス作成処理を利用して、登録状況判定手段および初出ワード登録手段が同時に初出情報を収集する。
【００１２】
このように、クロール処理によりインデックスを作成しながら初出情報を収集すること

ができるので、初出情報を得るためだけの処理を実施する必要がなく、簡単かつ効率よく単語の初出情報を収集することができる。

また、このようにして収集された初出情報は、ウェブページに表示してユーザに提供することができる。例えば、ユーザが指定した検索語に応じた検索結果の一覧と共に初出情報を表示させることで、検索結果ページのコンテンツの充実化を図ることができる。

【００１３】

本発明の単語情報収集装置において、前記初出ワード登録手段は、前記登録状況判定手段により前記単語候補が前記検索用インデックスに記憶されていると判定した場合、前記単語候補に関連付けられて記憶された初出日時と前記取得した更新日時とを比較し、前記更新日時が前記初出日時より古いと判定されると、該単語候補の初出日時を前記更新日時で更新することが好ましい。

【００１４】

この発明では、初出ワード登録手段は、初出ワード記憶手段に記憶された単語データの更新処理を行う。更新処理を行うのは、登録状況判定手段により抽出した単語候補が検索用インデックスに登録済みであると判定された場合である。すなわち、検索用インデックスは、前回のクロール処理によって前回までに取得した単語候補全てに対して作成されているため、初出ワード登録手段に記憶されている単語は全て検索用インデックスに含まれている。したがって、単語候補の検索用インデックスへの登録の有無を判定することで、初出ワード記憶手段への登録の有無を判定できる。

【００１５】

更新処理は、クロール処理によって任意の単語が含まれるウェブページの情報を取得するたびに、初出ワード記憶手段に記憶された該単語の初出日時と、ウェブページの情報の取得と同時に取得したウェブページの更新日時と、を比較し、更新日時が初出日時よりも古い場合は、初出ワード記憶手段に記憶されている該単語に関連付けられている初出日時を更新日時で更新し、該単語に関連付けられているウェブページの情報を、新しく取得したウェブページの情報で更新する。すなわち、新しく取得したウェブページの更新日時が古いほど初出ワード記憶手段に記憶されることになる。このような処理が繰り返されることで、結果として該単語がウェブ上に登場した古いウェブページに関する情報を収集することができる。

本発明によれば、クロール処理が行われるたびに、初出ワード記憶手段に記憶された単語データが、より更新日時の古いウェブページの情報に更新されていくので、自動的に最も古い日時のウェブページに関する情報を簡単に収集することができる。したがって、通常のクロール処理を利用して効率よく単語の初出情報を収集することができる。

【００１６】

本発明の単語情報収集装置において、前記ネットワークを介して接続された端末装置に対して検索語の入力を要求し、入力された検索語を取得する検索語取得手段と、前記取得した検索語と一致するキーワードを、前記検索用インデックスから検索し、該当するキーワードに関連付けられたウェブページに関する情報を取得するデータ検索手段と、前記取得した検索語と一致する単語を、前記初出ワード記憶手段から検索し、該当する単語に関連付けられたウェブページに関する情報と初出日時とを取得する初出ワード検索手段と、前記データ検索手段により取得したウェブページに関する情報と前記初出ワード検索手段により取得したウェブページに関する情報および初出日時とを表示させたウェブページを作成して配信する検索結果ページ提供手段と、をさらに備えたことが好ましい。

【００１７】

この発明では、初出ワード記憶手段に収集した初出情報を、ユーザが指定した検索語に対する検索結果の一覧とともに表示させる。すなわち、通常利用されている検索エンジンと同様に、指定された検索語を取得し、検索用インデックスから該検索語のデータを取得し、検索結果ページに一覧表示する一方で、さらに初出ワード検索手段が、初出ワード記憶手段から該検索語のデータ（初出日時、ウェブページに関する情報）を取得し、検索結果ページ提供手段によりそのデータを検索結果ページに表示させて端末装置に送信する。

【００１８】
この発明によれば、ユーザは、指定した検索語に対する検索結果とは別の情報、すなわち検索語の初出情報を得ることができる。特に、検索語として流行後を指定した場合、この流行語に対する初出情報は流行の発端に関わる情報を得ることができ、ユーザにとって有益なものである。このようにして、検索結果ページのコンテンツの充実化を図ることができる。
【００１９】
本発明の単語情報収集装置において、前記抽出された単語候補と一致する単語が前記初出ワード情報記憶手段に記憶されているか否かを判定する初出ワード登録判定手段をさらに備え、前記初出ワード登録手段は、前記登録状況判定手段により前記単語候補が前記検索用インデックスに記憶されていないと判定され、かつ、前記初出ワード登録判定手段により前記単語候補が前記初出ワード情報記憶手段に記憶されていないと判定された場合は、前記単語候補と該ウェブページに関する情報とに前記更新日時を初出日時として関連付けて初出ワード記憶手段に記憶させることが好ましい。
【００２０】
また、本発明の単語情報収集装置において、前記抽出された単語候補と一致する単語が前記初出ワード情報記憶手段に記憶されているか否かを判定する初出ワード登録判定手段をさらに備え、前記初出ワード登録手段は、前記登録状況判定手段により前記単語候補が前記検索用インデックスに記憶されていないと判定され、かつ、前記初出ワード登録判定手段により前記単語候補が前記初出ワード情報記憶手段に記憶されていないと判定された場合は、前記単語候補と該ウェブページに関する情報とに前記更新日時を初出日時として関連付けて初出ワード記憶手段に記憶させ、前記初出ワード登録判定手段により前記単語候補が前記初出ワード情報記憶手段に記憶されていると判定された場合は、前記単語候補に関連付けられて記憶された初出日時と前記取得した更新日時とを比較し、前記更新日時が前記初出日時より古いと判定されると、該単語候補の初出日時を前記更新日時で更新することが好ましい。
【００２１】
この発明では、初出ワード登録手段による登録処理または更新処理を行う前に、初出ワード登録判定手段により、初出ワード記憶手段への該当単語候補の登録の有無を判定する。該当単語が初出ワード記憶手段へ登録済みの場合は更新処理を行い、未登録の場合は登録処理を行う。
これによれば、仮に初出ワード記憶手段に記憶された単語と検索用インデックスに記憶された単語が一致しない場合であっても、確実に登録処理または更新処理を行うことができる。
【００２２】
本発明の単語情報収集方法は、ネットワーク上のウェブページに含まれる単語に関する情報を収集し、収集した単語を用いて、検索キーに対してインデックス検索を実行するための検索用インデックスを生成する単語情報収集方法であって、前記ネットワークを巡回してウェブページに関する情報とともに該ウェブページの更新日時を取得するページ情報取得ステップと、前記取得したウェブページを解析して単語候補を抽出するページ解析ステップと、前記抽出された単語候補と、取得済みの単語候補から予め生成された検索用インデックスとを比較し、前記単語候補が前記検索用インデックスに記憶されているか否かを判定する登録状況判定ステップと、前記判定の結果、前記検索用インデックスに記憶されていないと判定した場合に、前記単語候補と該ウェブページに関する情報とに前記更新日時を初出日時として関連付けて初出ワード記憶手段に記憶させる初出ワード登録ステップと、を備えることを特徴とする。
【００２３】
この発明では、ネットワークを巡回してウェブページに関する情報を取得し、このウェブページを解析して単語候補を抽出し、これまでに抽出した単語に対して検索用インデックスを作成するという処理を行いながら、一方で、抽出した単語候補に関する初出情報を

収集する。本発明では、クロール処理により検索用インデックスを生成するという通常の処理を利用して、登録状況判定ステップおよび初出ワード登録ステップにより初出情報を収集する。

【００２４】

具体的には、抽出した単語候補が検索用インデックスに登録済みであるか否かを判定し、単語候補が検索用インデックスに未登録と判定されると、単語候補を初出ワード記憶手段へ登録する。登録の際、その単語候補には、初出日時としてページ情報取得ステップで取得された更新日時が関連付けられ、さらに該単語候補が含まれるウェブページに関する情報が関連付けられて記憶される。なお、検索用インデックスは、クロール処理が行われるたびに、それまでに取得した単語候補全体に対して作成されるものである。

【００２５】

このように、通常のクロール処理によりインデックスを作成しながら初出情報を収集することができるので、初出情報を得るためだけの処理を実施する必要がなく、簡単かつ効率よく単語の初出情報を収集することができる。

また、このようにして収集された初出情報は、ウェブページに表示することで該ウェブページのコンテンツの充実化を図ることができる。

【００２６】

本発明の単語情報収集方法において、前記初出ワード登録ステップは、前記登録状況判定ステップにより前記単語候補が前記検索用インデックスに記憶されていると判定した場合、前記単語候補に関連付けられて記憶された初出日時と前記取得した更新日時とを比較し、前記更新日時が前記初出日時より古いと判定されると、該単語候補の初出日時を前記更新日時で更新することが好ましい。

【００２７】

この発明では、登録状況判定ステップで抽出した単語候補が検索用インデックスに登録済みであると判定された場合に、初出ワード記憶手段に記憶されている該当単語データの更新を行う。更新処理は、クロール処理によって任意の単語が含まれるウェブページの情報を取得するたびに、初出ワード記憶手段に記憶された該当単語の初出日時と、ウェブページの情報の取得と同時に取得したウェブページの更新日時と、を比較し、更新日時が初出日時よりも古い場合は、初出ワード記憶手段に記憶されている該単語に関連付けられている初出日時を更新日時で更新し、該単語に関連付けられているウェブページの情報を、新しく取得したウェブページの情報で更新する。すなわち、新しく取得したウェブページの更新日時が古いほど初出ワード記憶手段に記憶されることになる。このような処理が繰り返されることで、結果として該単語が初出したと思われるウェブページに関する情報を収集することができる。

【００２８】

本発明によれば、クロール処理が行われるたびに、初出ワード記憶手段に記憶された単語データが、より更新日時の古いウェブページの情報に更新されていくので、自動的に最も古い日時のウェブページに関する情報を簡単に収集することができる。したがって、通常のクロール処理を利用して効率よく単語の初出情報を収集することができる。

【００２９】

本発明の単語情報収集プログラムは、前述の単語情報収集方法をコンピュータに実行させることを特徴とする。

この発明によれば、コンピュータに前述の単語情報収集方法を実行させるため、この単語情報収集プログラムをインストールするだけの簡単な構成で、前述と同様の作用効果を得ることができ、有用性が高い。

【図面の簡単な説明】

【００３０】

【図１】本発明の実施形態にかかる単語情報収集システムの概略構成を示すブロック図。

【図２】前記実施形態における単語情報収集装置の動作を示すフローチャート。

【図３】前記実施形態における単語情報収集装置が提供する検索結果ページを端末装置で

表示させた画面の概略図。
【発明を実施するための形態】
【００３１】
以下、本発明の実施形態を図面に基づいて説明する。本実施形態では、検索エンジンの機能を有する単語情報収集システムを例示して説明する。
［１．単語情報収集システムの構成］
図１に示すように、単語情報収集システム１は、単語情報収集装置１００と、インターネット２０を介して単語情報収集装置１００に接続された端末装置２００と、を備えている。
【００３２】
インターネット２０はＴＣＰ／ＩＰなどの汎用のプロトコルに基づくインターネットであるが、これに限られない。例えば、ＬＡＮ（Local Area Network）などのイントラネット、無線媒体により情報が送受信可能な複数の基地局がネットワークを構成する通信回線網や放送網などのネットワーク、さらには、データを直接受信するための媒体となる無線媒体自体など、データを送受信させるいずれの構成も利用できる。
【００３３】
単語情報収集装置１００は、検索エンジンの機能を有するとともに、単語の初出情報を収集するものである。ここで、初出情報とは、任意の単語がウェブページ上に最初に登場したときの情報であり、そのときの日時（初出日時）や該ウェブページのＵＲＬ情報、該ウェブページに表示される文章データおよび画像データ等の情報を含む。
単語情報収集装置１００としては、一般的に用いられているパーソナルコンピュータ（ＰＣ）が用いられ、各種情報を記憶する記憶手段と、各種演算を実施するＣＰＵ等の制御手段と、キーボードやマウス等の入力手段、ウェブページを画面表示として出力させる表示手段などを備えている。
【００３４】
単語情報収集装置１００は、図１に示すように、記憶手段として、検索用インデックスとしてのインデックスデータベース１０１と、初出ワード情報記憶手段としての初出ワードデータベース１０２と、を備えている。また、図示しないが、単語情報収集装置１００は、検索結果ページを作成するための各種フォームを記憶させたデータベースを備えている。
【００３５】
インデックスデータベース１０１は、例えば、以下の表１に示すように、単語ごとに該単語が含まれるウェブページのＵＲＬ（Uniform Resource Locator）情報およびランクが関連付けられて１つのレコードとして記憶されたテーブル構造となっている。なお、項目はここに列挙したものに限られず、検索結果として表示可能な情報、例えば単語に関連するイメージデータ等を適宜追加してもよい。
【００３６】
ランク付けとは、任意の単語を含む複数のウェブページに対して、単語とウェブページとの関連度を各種アルゴリズムにより算出し、該ウェブページに付与することである。ランク付けの方法として、例えば、該単語を含むウェブページ中で、該単語が該ウェブページの内容に占める頻度が多いほど重要度が高くランク付けされたり、ウェブページのタイトル中に該キーワードが含まれている場合は重要度が高くランク付けされたりする。また、キーワードを含むウェブページにどれだけ多くのリンクが張られているかに応じてランク付けする方法もある。
なお、このランク付けは、クロール処理を行われるたびに新しく収集した単語も含めた検索用インデックスが作成され、インデックスデータベース１０１に再登録される。
【００３７】

【表１】

キーワード	ＵＲＬ	ランク
ねこ鍋	http://b----------/	1
ねこ鍋	http://a----------/	2
ねこなべ	http://c----------/	3

【００３８】

初出ワードデータベース１０２は、例えば、以下の表２に示すように、単語ごとに該単語がウェブページ上に最初に登場した日時である初出日時、該単語が含まれるウェブページのＵＲＬ情報およびキャッシュが関連付けられて１つのレコードとして記憶されたテーブル構造となっている。キャッシュとは、ウェブページの内容を保存したものであり、該ウェブページが更新されてしまった場合でも、キャッシュを表示することによって更新前のウェブページを閲覧することができる。なお、項目はここに列挙したものに限られず、検索結果として表示可能な情報、例えば単語に関連するイメージデータ等を適宜追加してもよい。

【００３９】

【表２】

キーワード	初出日時	ＵＲＬ	キャッシュ
ねこ鍋	2005/10/25 15:24:33	http://a----------/	*******
ゲリラ豪雨	2007/05/14 02:55:01	http://nnn----------/	*******
負け犬	2002/03/08 19:00:12	http://rrr----------/	*******

【００４０】

単語情報収集装置１００は、演算処理手段として、ネットワーク上のウェブページから単語情報を収集する単語情報収集手段１１０と、指定された検索語に応じた検索結果を提供するウェブ検索手段１２０と、図示しないが、ネットワークを介して端末装置２００とデータの送受信を行う送受信手段と、を備えている。

【００４１】

単語情報収集手段１１０は、ネットワークから単語情報を収集するものであり、ページ情報取得手段１１１と、ページ解析手段１１２と、登録状況判定手段１１３と、初出ワード登録手段１１４と、検索用インデックス生成手段１１５と、初出ワード登録判定手段１１６と、を備えている。

ページ情報取得手段１１１は、ネットワーク内を巡回し、ネットワーク内に公開されているウェブページのＵＲＬ情報、文章データおよび画像データなどの情報（ウェブページに関する情報）を取得する。この処理は一般的にクロール処理と呼ばれ、前回作成された検索用インデックス、すなわちインデックスデータベース１０１に記憶されたウェブページのＵＲＬ情報に基づいて各ウェブページを巡回する。また、クロール処理の頻度は必要に応じて適宜調整することができる。

【００４２】

ページ解析手段１１２は、ページ情報取得手段１１１により取得したウェブページに含まれる文章（テキスト）を抽出し、該文章に対して形態素解析を実施する。形態素解析とは、文章を意味のある単語に区切り、各単語の品詞等を判別する処理である。ページ解析手段１１２は、形態素解析により得られる複数の単語のうち、名詞となり得るものを単語候補として取得する。

【００４３】

登録状況判定手段１１３は、ページ解析手段１１２により得られた単語候補が、インデックスデータベース１０１に登録済みであるか否かを判定する。インデックスデータベー

ス１０１には、前回のクロール処理までに取得した単語候補に対して作成したインデックスが記憶されている。インデックスデータベース１０１に登録済みである単語候補は、初出ワードデータベース１０２への更新対象となり、未登録である単語候補は登録対象となる。

【００４４】

初出ワード登録手段１１４は、取得した単語候補を初出ワードデータベース１０２に登録または更新の処理を行う。登録処理としては、登録対象となった単語候補に該単語候補が含まれるウェブページのＵＲＬ情報と該ウェブページの更新日時とを関連づけて初出ワードデータベース１０２に記憶させる。また、更新処理としては、更新対象となった初出ワードデータベース１０２内の単語データに対して、記憶されている初出日時と取得した更新日時とを比較し、更新日時が初出日時よりも古い場合は、初出日時を更新日時で更新する。

【００４５】

検索用インデックス生成手段１１５は、新しく収集した単語候補と、インデックスデータベース１０１に記憶されている単語情報と、に対して検索用インデックスを作成し、作成した検索用インデックスでインデックスデータベース１０１を更新する。

初出ワード登録判定手段１１６は、初出ワード登録手段１１４の登録または更新処理の前に、対象となる単語候補が初出ワードデータベース１０２に登録済みであるか否かを判定する。単語候補が初出ワードデータベース１０２に登録済みであると判定された場合は、該単語候補は更新処理の対象となる。一方、未登録であると判定された場合は、該単語候補は登録処理の対象となる。

【００４６】

ウェブ検索手段１２０は、端末装置２００で指定された検索語に応じた検索結果ページを提供するものであり、検索語取得手段１２１と、インデックス検索手段１２２と、初出ワード検索手段１２３と、検索結果ページ提供手段１２４と、を備えている。

検索語取得手段１２１は、端末装置２００からの要求に応じて、検索ページを端末装置２００に送信する。検索語を入力させるための欄などが表示された検索ページを端末装置２００に表示させることで、ユーザに検索語を入力させる。入力された検索語は、ユーザの要求により端末装置２００から単語情報収集装置１００に送信され、検索語取得手段１２１は、受信した検索語を取得する。

【００４７】

インデックス検索手段１２２は、取得した検索語をインデックスデータベース１０１から検索し、検索語に相当するキーワードと対応付けられたＵＲＬ情報を取得する。

初出ワード検索手段１２３は、取得した検索語を初出ワードデータベース１０２から検索し、検索語に該当する初出ワードと、該初出ワードに対応付けられた初出日時、ＵＲＬ情報、およびキャッシュ等を取得する。

【００４８】

検索結果ページ提供手段１２４は、検索結果ページを作成し、端末装置２００に送信する。端末装置２００の表示手段で表示される検索結果ページには、検索結果の一覧のほか、該検索語の初出情報が表示される。初出情報としては、検索語が初出したウェブページのタイトルが表示され、このタイトルにはウェブページへのリンクが張られている。タイトルをクリックするだけで該ウェブページを閲覧することができる。また、初出日時やキャッシュも表示される。キャッシュには、該ウェブページに関する情報を取得したときの内容が保存されているため、仮に該ウェブページが存在しない状況であったとしても、初出した当時のウェブページを閲覧することができる。

【００４９】

端末装置２００は、図示しないが、演算処理手段として、単語情報収集装置１００に対して検索サービスを要求し、要求した検索サービスのウェブページを受信する端末送受信手段と、ウェブページを画面表示として出力させる出力手段と、文字入力可能なマウスやキーボードなどの入力手段とを備えている。一方、記憶手段としては、各種フォームにか

かわるフォームデータを記憶するデータベースなどを備えている。端末装置２００としては特に限定されないが、例えば、携帯電話やノートパソコンなどが挙げられる。

【００５０】

［２．単語情報収集装置１００の動作］

次に、単語情報収集装置１００の動作について説明する。単語情報収集装置１００は、単語情報収集手段１１０による処理と、ウェブ検索手段１２０による処理と、が別々に動作する。

【００５１】

まず、単語情報収集手段１１０の動作について、図２に基づいて説明する。

ステップＳ１において、ページ情報取得手段１１１は、ネットワークに公開されているウェブページを巡回し、該ウェブページに関する情報と、該ウェブページの更新日時と、を取得する。ここで、ウェブページに関する情報とは、ウェブページのＵＲＬ情報、ウェブページに表示される文章データおよび画像データ等であり、更新日時とは、ウェブページが更新されたときに通常付与される日時のことである。

次に、ステップＳ２において、ページ解析手段１１２は、ページ情報取得手段１１１により取得したウェブページの文章データを抽出し、該文章データに対して形態素解析を実施する。形態素解析により得られる複数の単語のうち、名詞となり得るものを単語候補として取得する。

【００５２】

このようにしてウェブページから得られた単語候補のそれぞれに対して、以下の処理を実施する。

ステップＳ３において、登録状況判定手段１１３は、インデックスデータベース１０１を参照し、ページ解析手段１１２により得られた単語候補が記憶されているか否かを判定する。単語候補がインデックスデータベース１０１に記憶されている場合（Ｓ３：Ｙｅｓ）は、ステップＳ６へ進む。一方、単語候補がインデックスデータベース１０１に記憶されていない場合（Ｓ３：Ｎｏ）は、ステップＳ４へ進む。

【００５３】

ステップＳ４では、初出ワード登録判定手段１１６は、初出ワードデータベース１０２を参照し、ページ解析手段１１２により得られた単語候補が記憶されているか否かを判定する。単語候補が初出ワードデータベース１０２に記憶されている場合（Ｓ４：Ｙｅｓ）は、ステップＳ６へ進む。一方、単語候補が初出ワードデータベース１０２に記憶されていない場合（Ｓ４：Ｎｏ）は、ステップＳ５へ進む。

【００５４】

ステップＳ５では、初出ワード登録手段１１４は、ページ解析手段１１２により得られた単語候補に、該単語候補が含まれるウェブページの更新日時とＵＲＬ情報とを関連付けて、初出ワードデータベース１０２に記憶させてステップＳ８へ進む。

また、ステップＳ６では、初出ワード登録手段１１４は、ページ解析手段１１２により得られた単語候補と一致する単語を初出ワードデータベース１０２から検索し、該当単語に関連付けられた初出日時と、該単語候補が含まれるウェブページの更新日時と、を比較し、更新日時が初出日時よりも古いか否かを判定する。更新日時が初出日時よりも古い場合（Ｓ６：Ｙｅｓ）は、ステップＳ７へ進む。一方、更新日時が初出日時と同じか初出日時より新しい場合（Ｓ６：Ｎｏ）は、ステップＳ８へ進む。

【００５５】

ステップＳ７では、初出ワード登録手段１１４は、ページ解析手段１１２により得られた単語候補と一致する単語を初出ワードデータベース１０２から検索し、該当単語に関連付けられた初出日時を、該ウェブページの更新日時で更新し、さらの該当単語に関連付けられたウェブページのＵＲＬ情報およびキャッシュを該ウェブページのＵＲＬ情報およびキャッシュで更新して、ステップＳ８へ進む。

なお、ステップＳ３～Ｓ７までの処理は、単語候補の数に応じて複数回実施される。

【００５６】

ステップＳ８では、検索用インデックス生成手段１１５は、新しく収集した単語候補と、インデックスデータベース１０１に記憶されている単語情報と、に対して検索用インデックスを生成し、新しく生成した検索用インデックスでインデックスデータベース１０１を更新した後、処理を終了する。

【００５７】

次に、ウェブ検索手段１２０の動作について説明する。

まず、ユーザは、端末装置２００の入力手段を入力操作し、単語情報収集装置１００が提供する検索ページにアクセスするために、例えば、ウェブブラウザを起動させてアドレスを入力し、検索ページを要求する。

単語情報収集装置１００は、図示しない送受信手段により端末装置２００からの検索ページの要求を受信すると、検索語取得手段１２１は、図示しない記憶手段から検索ページ用のフォームを読み出し、これらの情報に基づいて検索ページを作成し、端末装置に送信する。

【００５８】

端末装置２００では、端末送受信手段により検索ページの情報を受信して、図示しない表示手段（ディスプレイ等）に画面表示させる。

ユーザは、画面表示にしたがって、入力手段を用いて検索したい単語（検索語）を入力し、単語情報収集装置１００へ送信する。

単語情報収集装置１００は、送受信手段で検索語を受信し、検索語取得手段１２１は検索語を取得する。

【００５９】

次に、インデックス検索手段１２２は、取得した検索語に相当する単語をインデックスデータベース１０１から検索し、該当する単語データを抽出する。

また、初出ワード検索手段１２３は、取得した検索語と一致する単語を初出ワードデータベース１０２から検索し、該当する単語データを抽出する。

次に、検索結果ページ提供手段１２４は、図３に示すような検索結果ページを作成し、端末装置２００に送信する。

【００６０】

図３において、検索結果ページ５は、検索語入力領域５１と、初出情報表示領域５２と、検索結果一覧表示領域５３を有している。

検索語入力領域５１は、ユーザが入力可能な検索語入力欄５１１と検索ボタン５１２を有する。検索語入力欄５１１にはユーザが入力した検索語が表示され、検索ボタン５１２は再検索の要求を単語情報収集装置１００へ送信するためのボタンである。

【００６１】

初出情報表示領域５２は、初出情報であることを示すタイトル欄５２１と、ウェブページのタイトルがテキスト表示されたＵＲＬ情報欄５２２と、初出日時が表示された初出日時欄５２３と、該ウェブページのキャッシュへのリンクが張られたキャッシュ欄５２４と、を有する。タイトル欄５２１には、指定された検索語が最初に登場したときのウェブページ情報を表示していることをユーザに理解させるためのタイトルが表示されればよい。例えば、検索語として「ねこなべ」が指定されている場合には「ねこなべの初出は！」というタイトルを表示することができる。ＵＲＬ情報欄５２２に表示されたテキストには、該ウェブページのＵＲＬへのリンクが張られており、該ＵＲＬ情報欄５２２をクリックするだけで、指定した検索語が初出したウェブページのＵＲＬへ移動しその内容を閲覧することができる。また、キャッシュ欄５２４をクリックすると、初出ワードデータベース１０２に保存した時（初出時）のウェブページの内容を閲覧することができる。

【００６２】

検索結果一覧表示領域５３は、インデックスデータベース１０１から抽出したデータが一覧表示される領域である。ウェブページのタイトルがテキスト表示されるとともに、該テキストにはウェブページのＵＲＬへのリンクが張られている。

【００６３】

ユーザは、端末装置２００の表示手段に画面表示された検索結果ページにより、指定した検索語に関連するウェブページの一覧を閲覧することができるだけでなく、指定した検索語が最初に登場したウェブページに関する情報も得ることができる。

【００６４】

［３．本実施形態の作用効果］

上述した実施形態では、以下に示す作用効果を奏することができる。

単語情報収集手段１１０において、ページ情報取得手段１１１がネットワークを巡回してウェブページに関する情報を取得し、ページ解析手段１１２が取得したウェブページから単語情報を取得し、検索用インデックス生成手段１１５が検索用インデックスを作成するという、いわゆる検索エンジンにおける通常の処理を行うとともに、登録状況判定手段および初出ワード登録手段１１４により取得した単語情報に関する初出情報を収集している。ページ情報取得手段１１１はウェブページに関する情報とともに、該ウェブページの更新日時を取得する。初出ワードデータベース１０２に記憶された単語には初出日時が関連付けられているので、この初出日時と取得した更新日時とを比較し、古いほうの日時を初出日時として再登録する。すなわち、取得するウェブページの更新日時が随時古い日時に更新されるので、結果として最も古いウェブページの情報を効率よく収集することができる。

このように、検索エンジンにおいて通常行われる処理を行いながら、簡単かつ効率よく初出情報を収集することができる。

【００６５】

また、ウェブ検索手段１２０では、ユーザが指定した検索語の検索結果の一覧とともに、収集した初出情報を検索結果ページに表示している。ユーザが指定する検索語としては、一般的な単語のほか、流行語のような単語もある。流行語は、あるウェブページに表示されたことが発端となって流行が広まることも多く、流行の発端となったウェブページに関する情報を得たいと思うユーザも多数いる。上記実施形態では、上述の単語情報収集手段１１０によって収集した初出情報を、ウェブ検索手段１２０が、例えば検出語が初出したウェブページのタイトルと、初出日時と、を表示させ、タイトルには該ウェブページのＵＲＬへのリンクを張った状態で検索結果ページに表示する。

したがって、ユーザは指定した検索語の初出情報を得ることができるとともに、初出したウェブページを閲覧することができる。このように、ユーザが知りたいと思う有益な情報を検索語の検索結果とともに提供することができ、検索結果ページのコンテンツの充実化を図ることができる。

【００６６】

さらに、上記実施形態では、検索結果ページの初出情報の一部にキャッシュを表示している。初出情報としてリンクが張られるウェブページは古く、その後更新されていることが多いため、初出時のウェブページを閲覧できない可能性が高い。しかしながら、初出時のウェブページの内容をキャッシュとして初出ワードデータベース１０２に保存し、検索結果ページにキャッシュとして表示させるので、仮に初出時のウェブページが存在しない場合でも、初出時のウェブページを閲覧することができる。したがって、ユーザにとって有益な情報を提供することができる。

【００６７】

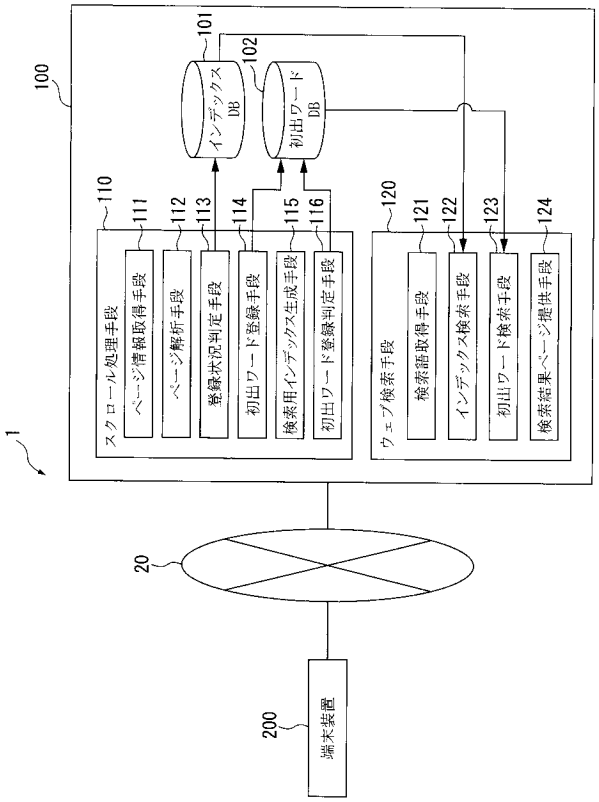
［４．変形例］

なお、本発明は、上述した実施形態に限定されるものではなく、本発明の目的を達成できる範囲で、以下に示される変形をも含むものである。

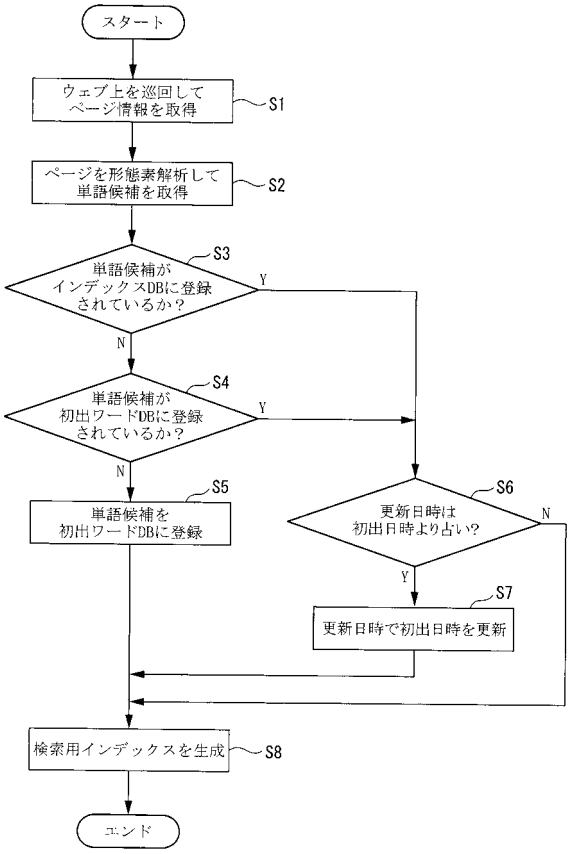
例えば、上記実施形態では、単語情報収集手段１１０の動作において、初出ワード登録判定手段１１６により、検索語が初出ワードデータベース１０２に登録済みであるか否かを判定する処理（Ｓ４）を行ったが、この処理は省略してもよい。これは、ステップＳ３において、登録状況判定手段１１３がインデックスデータベース１０１への登録状況を判定しているため、この判定結果に基づいて初出ワードデータベース１０２への登録の有無を判定することができるからである。これによれば、処理の高速化を図ることができる。

【００６８】
また、上記実施形態では、ページ解析手段１１２は、形態素解析により文章を単語候補に分解したが、単語候補を抽出する方法はこれに限られない。一般的に用いられる言語処理技術、例えばＮ－ｇｒａｍを用いて解析してもよい。
【００６９】
さらに、上記実施形態において、初出ワードデータベース１０２の項目として画像データを追加してもよい。任意の単語が含まれるウェブページから、該単語に関連する画像データを取得し、該単語にこの画像データを関連付けて初出ワードデータベース１０２に記憶させる。したがって、ウェブ検索手段１２０により初出情報を検索結果ページに表示させる際は、初出情報の一部としてこの画像データを表示させることができる。画像データは視覚的なものであるので、ユーザにとっては認識が容易である。すなわち、ユーザにわかりやすい情報提供を行うことができる。
【産業上の利用可能性】
【００７０】
本発明は、ネットワーク上のウェブページに含まれる単語情報を収集する単語情報収集装置として検索エンジン等に利用できる。
【符号の説明】
【００７１】
１００…単語情報収集装置
１０１…インデックスデータベース
１０２…初出ワードデータベース
１１０…単語情報収集手段
１１１…ページ情報取得手段
１１２…ページ解析手段
１１３…登録状況判定手段
１１４…初出ワード登録手段
１１５…検索用インデックス生成手段
１１６…初出ワード登録判定手段
１２０…ウェブ検索手段
１２１…検索語取得手段
１２２…インデックス検索手段
１２３…初出ワード検索手段
１２４…検索結果ページ提供手段
２００…端末装置

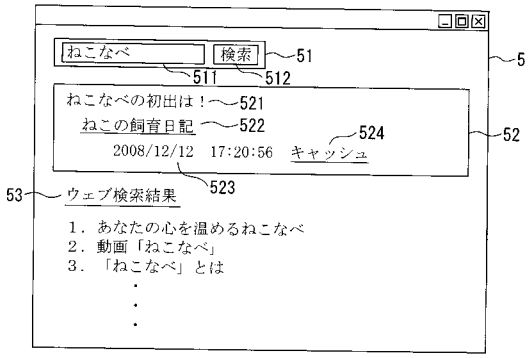
【図１】


100
101
インデックスDB
102
初出ワードDB
110
スクロール処理手段
111
ページ情報取得手段
112
ページ解析手段
113
登録状況判定手段
114
初出ワード登録手段
115
検索用インデックス生成手段
116
初出ワード登録判定手段
120
ウェブ検索手段
121
検索語取得手段
122
インデックス検索手段
123
初出ワード検索手段
124
検索結果ページ提供手段
1
20
200
端末装置

【図２】


スタート
ウェブ上を巡回して
ページ情報を取得
S1
ページを形態素解析して
単語候補を取得
S2
単語候補が
インデックスDBに登録
されているか？
S3
Y
N
単語候補が
初出ワードDBに登録
されているか？
S4
Y
N
単語候補を
初出ワードDBに登録
S5
更新日時は
初出日時より古い？
S6
N
Y
更新日時で初出日時を更新
S7
検索用インデックスを生成
S8
エンド

【図３】


ねこなべ
検索
51
511
512
5
ねこなべの初出は！
521
ねこの飼育日記
522
524
2008/12/12 17:20:56 キャッシュ
52
523
53
ウェブ検索結果
1．あなたの心を温めるねこなべ
2．動画「ねこなべ」
3．「ねこなべ」とは

フロントページの続き

(56)参考文献　特開２００９－１５７７３４（ＪＰ，Ａ）
特表２００７－５０７７９８（ＪＰ，Ａ）
特開２００６－１８５０２０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０