(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5661104号
(P5661104)

(45)発行日　平成27年1月28日(2015.1.28)　(24)登録日　平成26年12月12日(2014.12.12)

(51)Int.Cl.　　F I
G06F 17/30 (2006.01)　　G06F 17/30 414Z
G06F 12/00 (2006.01)　　G06F 17/30 210D
　　G06F 12/00 520A

請求項の数 12　(全 14 頁)

(21)出願番号　特願2012-509810 (P2012-509810)
(86)(22)出願日　平成22年5月6日 (2010.5.6)
(65)公表番号　特表2012-526320 (P2012-526320A)
(43)公表日　平成24年10月25日 (2012.10.25)
(86)国際出願番号　PCT/US2010/001359
(87)国際公開番号　WO2010/129063
(87)国際公開日　平成22年11月11日 (2010.11.11)
審査請求日　平成25年4月22日 (2013.4.22)
(31)優先権主張番号　12/800,015
(32)優先日　平成22年5月5日 (2010.5.5)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　200910136443.2
(32)優先日　平成21年5月8日 (2009.5.8)
(33)優先権主張国　中国 (CN)

(73)特許権者　510330264
アリババ・グループ・ホールディング・リミテッド
ＡＬＩＢＡＢＡ　ＧＲＯＵＰ　ＨＯＬＤＩＮＧ　ＬＩＭＩＴＥＤ
英国領，ケイマン諸島，グランド・ケイマン，ジョージ・タウン，ワン・キャピタル・プレイス，フォース・フロア，ピー．オー．ボックス　８４７
(74)代理人　110000028
特許業務法人明成国際特許事務所
(72)発明者　ヤン・ハンフェイ
中華人民共和国　ハンチョウ，ワーナー・ロード，ウエスト・レイク・インターナショナル・プラザ，１０階，ナンバー３９１
最終頁に続く

(54)【発明の名称】検索エンジンインデクシング及びインデックスを使用する検索のための方法とシステム

(57)【特許請求の範囲】
【請求項１】
コンピュータによって実行されるデータインデクシングのための方法であって、
データソースからデータを受信することと、
所定のデータ分類基準にしたがって、前記データを複数のカテゴリの１つに分類することと、
前記データの第１の部分に少なくとも一部基づいて前記データが分類された前記カテゴリに関連付けられている複数のインデックスの中から前記データを割り当てるインデックスを決定し、
前記データと、前記データが分類された前記カテゴリに関連付けられ既定の最大容量を有する決定された前記インデックスとの間に対応関係を確立することと、
前記データと前記決定されたインデックスとの間の前記関係を記録することと、前記決定されたインデックスは対応するインデックス書き込みデバイスによって独占的に書き込まれるよう構成されており、前記データと前記決定されたインデックスとの間の前記関係を記録することは、
少なくとも前記データの第２の部分を前記決定されたインデックスに関連付けられているファイルに格納することと、
前記決定されたインデックスに関連付けられている識別子および前記対応するインデックス書き込みデバイスの関連付けられている識別子を前記決定されたインデックスに関連付けられている前記ファイルの少なくとも一部に格納することとを含むこと、

を備える、方法。
【請求項２】
請求項１に記載の方法であって、更に、
検索クエリを受信することと、
前記受信された検索クエリに対応するインデックスを前記複数のインデックスのなかから決定することと、
前記インデックスに関連付けられた格納されたデータを出力することと、
を備える方法。
【請求項３】
請求項１に記載の方法であって、更に、
少なくとも一部には前記カテゴリの容量と前記既定の最大容量とに基づいて、前記複数のインデックスの数を決定することを備える方法。
【請求項４】
請求項１に記載の方法であって、
前記インデックスを決定することは、
前記複数のインデックスを連続する正の整数で表示することと、
前記データに固有の整数値を割り当てることと、
前記固有の整数値を前記複数のインデックスの数で割って余りを得ることと、
前記複数のインデックスの中で前記余りと同じ通し番号を有する１つのインデックスを前記決定されたインデックスとして決定することと、
を含む、方法。
【請求項５】
請求項１に記載の方法であって、
前記対応するインデックス書き込みデバイスは、複数のインデックス書き込みデバイスの１つであり、
前記データと前記決定されたインデックスとの間の前記関係を記録することは、
前記複数のインデックス書き込みデバイスを連続する正の整数で表示することと、
前記決定されたインデックスの通し番号を前記インデックス書き込みデバイスの数で割って余りを得ることと、
前記データと、前記余りと同じ通し番号を有するインデックス書き込みデバイスとを関連付けることと、
を含む、方法。
【請求項６】
請求項１に記載の方法であって、
前記データを前記複数のカテゴリの１つに分類することは、少なくとも一部には前記データの特性に基づく、方法。
【請求項７】
データインデクシングのためのシステムであって、
１つ又は２つ以上のプロセッサであって、
データソースからデータを受信し、
所定のデータ分類基準にしたがって、前記データを複数のカテゴリの１つに分類し、
前記データの第１の部分に少なくとも一部基づいて前記データが分類された前記カテゴリに関連付けられている複数のインデックスの中から前記データを割り当てるインデックスを決定し、
前記データと、前記データが分類された前記カテゴリに関連付けられ既定の最大容量を有する決定された前記インデックスとの間に対応関係を確立し、
前記データと前記決定されたインデックスとの間の前記関係を記録するように、
構成されている、１つ又は２つ以上のプロセッサと、
前記決定されたインデックスは対応するインデックス書き込みデバイスによって独占的に書き込まれるよう構成されており、前記データと前記決定されたインデックスとの間の

前記関係を記録することは、
　少なくとも前記データの第２の部分を前記決定されたインデックスに関連付けられているファイルに格納すること、
　前記決定されたインデックスに関連付けられている識別子および前記対応するインデックス書き込みデバイスの関連付けられている識別子を前記決定されたインデックスに関連付けられている前記ファイルの少なくとも一部に格納することを含み、
　前記１つ又は２つ以上のプロセッサに結合され、前記プロセッサに命令を提供するように構成されたメモリと、
　を備えるシステム。
【請求項８】
　請求項７に記載のシステムであって、
　前記１つ又は２つ以上のプロセッサは、更に、
　　検索クエリを受信し、
　　前記受信された検索クエリに対応するインデックスを前記複数のインデックスのなかから決定し、
　　前記インデックスに関連付けられた格納されたデータを出力するように、
　構成される、システム。
【請求項９】
　請求項７に記載のシステムであって、
　前記１つ又は２つ以上のプロセッサは、更に、少なくとも一部には前記カテゴリの容量と前記既定の最大容量とに基づいて、前記複数のインデックスの数を決定するように構成される、システム。
【請求項１０】
　請求項７に記載のシステムであって、
　前記インデックスを決定することは、
　　前記複数のインデックスを連続する正の整数で表示することと、
　　前記データに固有の整数値を割り当てることと、
　　前記固有の整数値を前記複数のインデックスの数で割って余りを得ることと、
　　前記複数のインデックスの中で前記余りと同じ通し番号を有する１つのインデックスを前記決定されたインデックスとして決定することと、
　を含む、システム。
【請求項１１】
　請求項７に記載のシステムであって、
　前記インデックス書き込みデバイスは、複数のインデックス書き込みデバイスの１つであり、
　前記データと前記決定されたインデックスとの間の前記関係を記録することは、
　　前記複数のインデックス書き込みデバイスを連続する正の整数で表示することと、
　　前記決定されたインデックスの通し番号を前記インデックス書き込みデバイスの前記数で割って余りを得ることと、
　　前記データと、前記余りと同じ通し番号を有するインデックス書き込みデバイスとを関連付けることと、
　を含む、システム。
【請求項１２】
　請求項７に記載のシステムであって、
　前記データを前記複数のカテゴリの１つに分類することは、少なくとも一部には前記データの特性に基づく、システム。
【発明の詳細な説明】
【技術分野】
【０００１】
［関連出願の相互参照］

本出願は、あらゆる目的のために参照によって本明細書に組み込まれる、発明の名称を「METHOD AND SYSTEM FOR IMPLEMENTING SEARCH SERVICE（検索サービスを実施するための方法及びシステム）」とする２００９年５月８日付けで出願の中国特許出願第２００９１０１３６４４３．２号の優先権を主張する。

【０００２】

本発明は、コンピュータ技術の分野に関し、特に、データ検索のための方法及びシステムに関する。

【背景技術】

【０００３】

情報技術における進歩は、生成される情報を次々に増加させている。例えば、多くの電子商取引用ウェブサイトは、そのユーザによって生成されるデータ量の急増に直面している。人々が必要な情報を見つけるのを助けるために、データソースの全文検索を可能にする検索サービスが提供されている。この検索は、ユーザによって提供される検索クエリに含まれるキーワード又は記述的情報に基づくことができる。検索結果は、ユーザに戻される。

【０００４】

検索サービスの実施には、データソースからデータを収集、解析、及び格納して高速で且つ正確な情報の読み出しを促すプロセスがよく使用されており、このようなプロセスは、検索エンジンインデクシングと呼ばれる。検索のためにユーザによって提供されるキーワードは、通常はテキスト形式であるので、キーワード検索のためのインデックスもやはり、通常はテキスト形式である。

【０００５】

インデックス対象とされるドキュメントは、検索エンジンによって提供されるウェブページスナップショット、又はウェブページスナップショットの一部分であってよい。ウェブページスナップショットは、様々な形式（フォーマット）を有してよい。検索エンジンのなかには、複数のドキュメント形式をサポートしているものがある。インデックスは、データソースからの様々な情報を含んでよく、例えば、もしデータソースのコンテンツの一部がテキストであるならば、インデックスは、そのようなテキストを含んでよく、もしデータソースファイルが画像、音声、又は映像の形式であるならば、インデックスは、例えば８＋ファイルのウェブアドレスを示すフィールドなどそのようなファイルのソースを示すフィールドを有してよい。

【０００６】

インデックスを管理するために、多くの場合、インデックスサーバが使用される。ユーザが検索を開始すると、ユーザによって提供されたクエリが検索サーバによって受信される。検索サーバは、ユーザによって求められているデータにどのインデックスがインデックスされているかを決定し、次いで、対応するインデックスにおいて（１つ又は２つ以上の）クエリ用語を探索し、そのインデックスから読み出された検索結果をユーザに提供する。

【０００７】

大量のデータをインデックスするために、ウェブサイトのオペレータは、多くの場合、幾つかのインデックスサーバを使用する。インデックスは、一連のインデックスデータアイテムを含んでよく、各データアイテムは、ドキュメントと称される。通常、各ドキュメントは、ソースデータの中の一記録に対応している。インデックスサーバは、一般に、ソースデータから抽出された記録をインデックスに変換する。インデックス管理における大きな課題は、並列計算プロセスの管理である。競合状態及び一貫性障害の機会は、数多くある。例えば、複数のインデックスサーバが、同じインデックスファイルを同時に書き込む必要があるかもしれない。従来の実装形態では、非一貫性障害を回避するために、複数のインデックスサーバの１つがインデックスファイルにデータを書き込んでいるときは、その他のインデックスサーバは、アイドル状態にあり、最初のインデックスサーバがデータの書き込みを終了した後にはじめてインデックスファイルにデータを書き込むことがで

きる。複数のインデックスサーバによる同じ共有リソース（例えばインデックスファイル）へのこのような書き込みの行為は、書き込み共有コンフリクトと称される。したがって、従来の方法は、インデックシングプロセス中に、低パフォーマンス及び共有コンフリクトを生じることがあった。

【０００８】

更に、クエリ検索プロセス中に、インデックスのサイズは、検索効率に影響を及ぼすことがある。もしインデックスが大きすぎると、データの探索に時間がかかる恐れがあり、もしインデックスが小さすぎると、多くのインデックスにアクセスする必要があるであろう。

【０００９】

したがって、より効率的な検索エンジンインデクシング及び検索の方法又はシステムが必要とされている。

【図面の簡単な説明】

【００１０】

以下の詳細な説明及び添付の図面において、発明の様々な実施形態が開示される。

【００１１】

【図１】検索エンジンインデクシングシステムの一実施形態を配備したネットワークを示す概観図である。

【００１２】

【図２Ａ】検索エンジンインデクシングプロセスの一実施形態を示したフローチャートである。

【００１３】

【図２Ｂ】インデックスされたデータの一例を示した図である。

【００１４】

【図３】検索エンジンインデクシングシステムの一実施形態を示したブロック図である。

【発明を実施するための形態】

【００１５】

発明は、プロセス、装置、システム、合成物、コンピュータ読み取り可能な記録媒体に実装されるコンピュータプログラム製品、並びに／又は結合先のメモリに格納された命令及び／若しくは結合先のメモリによって提供される命令を実行するように構成されたプロセッサなどのプロセッサを含む、数々の形態で実装することができる。本明細書では、これらの実装形態、又は発明がとりえるその他のあらゆる形態を技術と称してよい。総じて、開示されたプロセスのステップの順序は、発明の範囲内で変更されてよい。別途明記されない限り、タスクを実施するように構成されているものとして説明されるプロセッサ又はメモリなどのコンポーネントは、所定時にタスクを実施するように一時的に構成された汎用コンポーネントとして、又はタスクを実施するように製造された特殊コンポーネントとして実装されてよい。本明細書において、「プロセッサ」という用語は、コンピュータプログラム命令などのデータを処理するように構成された１つ又は複数の、デバイス、回路、及び／又は処理コアを言う。

【００１６】

発明の原理を示した添付の図面とともに、以下で、発明の１つ又は２つ以上の実施形態の詳細な説明が提供される。発明は、このような実施形態に関連して説明されているが、いかなる実施形態にも限定されず、発明の範囲は、特許請求の範囲によってのみ限定され、発明は、数々の代替形態、変更形態、及び均等物を内包している。以下の説明では、発明の完全な理解を可能にするために、数々の詳細が特定されている。これらの詳細は、例示を目的として提供されたものであり、発明は、これらの詳細の一部又は全部を伴わずとも、特許請求の範囲にしたがって実施することができる。明瞭さを期するために、発明に関連した技術分野で知られている技工物は、発明が不必要に不明瞭にされないように、詳細な説明を省略されている。

【００１７】

図１は、検索エンジンインデクシングシステムの一実施形態を配備したネットワークを示す概観図である。図に示された例では、ユーザは、検索サーバへ検索クエリを送出するクライアントデバイス１０を介して検索サーバ１１にアクセスする。クエリは、一般に、ユーザによって求められているデータのタイプ、範囲、又は特性を示すためにユーザによって提供されるキーワードを含む。例えば、ユーザが電子メールサーバから自身の全ての電子メールを取得したい場合には、キーワードは、そのユーザの電子メールアドレスであってよい。ユーザが製品の情報を問い合わせたい場合には、キーワードは、その製品のいずれかのモデルの名前であってよい。一部の実施形態では、図１に示されるように、ネットワークのなかに複数のクライアントデバイス１０があってよい。

【００１８】

検索クエリを受信した後、検索サーバ１１は、キーワードを取得するために検索クエリを解析し、格納されたデータを探索するためにはどのインデックスが使用されるべきかを決定する。一部の実施形態では、インデックスは、ユーザの検索行為に一致させるように作成されたとある所定のルールの下で配付されたデータである。例えば、電子メールメッセージは、電子メールサーバに格納されてよい。電子メールのテキスト、及び電子メールの中の非テキスト形式のファイルのリンクアドレスを含む、電子メールメッセージの中のデータは、１つ又は２つ以上の電子メールアドレスに関連付けられた電子メールをそれぞれ含みえる幾つかの既定のインデックスに書き込まれてよい。更に、検索サーバは、各電子メールを探索するためにはどのインデックスが使用されるべきかの記録も含む。したがって、検索サーバは、ユーザの電子メールを検索するための、電子メールアドレスを含む検索クエリを受信すると、電子メールアドレス及び記録にしたがって、どのインデックスの下にその電子メールアドレスが配されているかを決定することが可能である。

【００１９】

図に示された例では、インデックスは、インデックスサーバ１３に接続されたインデックスストレージデバイス１２に格納されている。インデックスサーバ１３は、所定のルールにしたがって、データストレージデバイス１４からのデータに基づいてインデックスを生成するために使用される。一部の実施形態では、複数のインデックスサーバが使用される。生成されたインデックスは、インデックスストレージデバイス１２に格納される。データストレージデバイス１４のなかのデータは、時間とともに変化を受けるであろう。例えば、格納されている電子メールの数は、時間とともに増加するであろう。本実施形態では、効率的な検索を行うために、各インデックスの容量が、とある所定の範囲内に制限されている。これは、もし１つのインデックスのサイズが大きすぎると、又は非常に小さいインデックスが多数あると、１つのインデックスにおけるデータ検索に比較的長い時間がかかる恐れがあるからである。一部の実施形態では、各インデックスの容量にしたがって上限が設定される。一部の実施形態では、特定のアプリケーションにおけるデータ特性及び／又は検索サーバ１１のパフォーマンスなどの基準にしたがって、インデックスストレージデバイス１２のデータソースとしてデータストレージデバイス１４からのデータの一部が選択される。データソースは、容量に限界があるので、データソースの容量限界をインデックスの容量の上限で割った値の端数を切り上げることによって正の整数を得て、この値によって、データソースがその最大容量までデータを格納する場合の各データソースのためのインデックスの最大数を示すことができる。同様に、データストレージデバイス１４のなかのその他のデータについても、各データソースのためのインデックスの最大数を計算することができる。

【００２０】

図２Ａは、検索エンジンインデクシングプロセスの一実施形態を示したフローチャートである。インデックスは、このプロセスの完了後に生成される。インデックスは、データソースのなかのデータに関する情報を含んでおり、したがって、インデックスのコンテンツにしたがってクエリ結果をユーザに提供することが可能である。ユーザから検索クエリが受信されると、検索サーバによってインデックスが決定される。インデックスにしたがって、データが読み出され、ユーザに戻される。ユーザに送信されるデータは、データソ

ースのなかのテキストデータ又はその他のデータ（画像、映像、音声など）のウェブアドレスであってよい。

【００２１】

ステップ２１では、データソースからデータが受信され、少なくとも一部にはその受信データの特性に基づいて、カテゴリに分類される。一部の実施形態では、受信データの分類に、１つ又は２つ以上の所定のルールが採用される。これらのルールは、データソースのなかのデータの特性に基づいて設定されてよい。例えば、一部の電子商取引用プラットフォームでは、注文情報及び電子メール情報を、それらの形式の相違に基づいて容易に識別することができる。したがって、注文情報及び電子メールメッセージは、別々のカテゴリに分類される。一部の実施形態では、カテゴリは、更に、下位カテゴリに分類される。例えば、複数ユーザの電子メールを含むデータソースの場合は、データは、ユーザの電子メールアドレスにしたがって下位カテゴリに分類されてよい。各カテゴリのなかのデータのサイズは、設定可能であり、一部の実施形態では、データ成長の予測に基づいて決定される。

【００２２】

ステップ２２では、データと、カテゴリに関連付けられたインデックス一式との間に、対応関係が確立される。各データカテゴリに対応するインデックスの数は、事前に設定されたカテゴリのサイズを一インデックスの最大容量で割ることによって予め設定される。大きいインデックスの場合は、そのサイズが大きいほど、そのインデックスにおけるデータ探索に必要な時間が長くなる恐れがあることを意味する。小さいインデックスの場合も、その数が多いほど、インデックスを開くのに要する時間が増すのでデータの探索にかかる時間が長くなる恐れがある。したがって、一部の実施形態では、要求／応答時間、要求処理の速さ、これらのパフォーマンス測定基準における揺らぎなどの、システムの様々なパフォーマンス測定基準にしたがって、一インデックスの最大容量が決定される。異なる検索技術を用いた様々な実施形態では、例えば、一インデックスの最大容量は、１～４ギガバイトの範囲である。

【００２３】

データソースに格納可能な一カテゴリのデータの総容量がＴであり、一インデックスの最大容量がＭであると決定されたとすると、そのカテゴリに対応するインデックスの数は、ＴをＭで割った値の端数を切り上げることによって計算される。インデックスの数が決定された後、各インデックスは、通し番号を割り振られてよく、これらの通し番号は、インデックスファイルに又はインデックスファイルのファイル名に記録されてよい。インデックスされたデータの一例を示した図２Ｂを参照する。この例では、２つのデータカテゴリ、すなわち注文カテゴリと電子メールカテゴリとがある。一インデックスの最大容量は、４ギガバイトであると決定される。注文カテゴリ及び電子メールカテゴリの容量は、２０ギガバイト及び１０ギガバイトにそれぞれ設定される。したがって、注文カテゴリには５つのインデックスが、電子メールカテゴリには３つのインデックスが関連付けられる。

【００２４】

受信データを適切なインデックスに割り当てるために、様々なインデクシング技術を使用することができる。例えば、一部の実施形態では、HASHをベースにした技術が使用される。図２Ａを再び参照すると、分類された受信データは、固有の整数を割り当てられる。一部の実施形態では、受信データのための固有な整数は、そのカテゴリのデータにおいて見いだされる固有なフィールドから決定される。例えば、インデックス対象とされる受信電子メールメッセージは、固有の電子メールアドレスを含み、これは、抽出され、アドレスをパラメータとして表すASCIIコードを使用するマッピング関数などの関数にしたがって、数字にマッピングされる。取得された数字に対し、HASH関数又はその他の適切な技術を使用して算術演算を実施することによって、整数値も得られる。一部の実施形態では、受信データについて得られた固有の整数値は、Ｈとして示され、受信データが属するカテゴリのインデックスの数は、Ｎとして示される。現カテゴリのための一インデックスは、Ｈ％Ｎから得られる値に等しく、ここで、Ｈ％Ｎは、ＨをＮで割った後に得られる余りを

言う。こうして、受信データとインデックスとの間に対応関係が確立される。一例として、図２Ｂに戻る。インデックス対象とされる２つの受信電子メールメッセージがそれぞれ整数値３５及び５７にマッピングされるとすると、電子メールカテゴリに対応するインデックスは３つであるので、メッセージは、2002及び2003にそれぞれ関連付けられる。

【００２５】

図２Ａに戻ると、ステップ２３では、受信データとインデックスとの間の関係が記録される。一部の実施形態では、関係は、インデックスに含まれるインデックスエントリとして記録され、ここで、各エントリは、識別子や格納されたコンテンツなどのフィールドを含む。例えば、電子メールメッセージのインデックスエントリは、その送信者アドレス、件名、メッセージに含まれる画像若しくはその他のリソースのリンク、メッセージのテキストなどに基づいた、１つ又は２つ以上のHASH値を含んでよい。

【００２６】

一部の実施形態では、インデックスに対応するデータの総量がそのインデックスの既定の最大容量を超えているかどうかを判定するために、随意にチェックが実施される。インデックスの容量を超えている場合には、新しいインデックスが追加され、そのカテゴリ専用のストレージ領域が増やされる。

【００２７】

一部の実施形態では、図１のインデックスサーバ１３などのインデックス書き込みデバイスによって、受信データ及び／又はインデックスが保存される。インデックスの書き込みにおける共有コンフリクトを回避するために、各インデックスは、１つのインデックス書き込みデバイスにのみ関連付けられ、したがって、その１つのインデックス書き込みデバイスによってのみ書き込みされる。

【００２８】

各カテゴリのデータが同じインデックス書き込みデバイスに割り振られることを保証するために、ステップ２２と同様なプロセスが使用されてよい。例えば、インデックス書き込みデバイスを連続する正の整数（すなわち通し番号）で表示し、各インデックスの通し番号をインデックス書き込みデバイスの数によって割った余りを得て、その余りに等しい通し番号を有するインデックス書き込みデバイスに、インデックスに対応するデータを割り振ってよい。図２Ｂに、一例が示されており、ここでは、異なるデータカテゴリのインデックスを記録するために、３つのデバイス（１、２、及び３）が使用される。各データカテゴリが１つのインデックス書き込みデバイスに関連付けられる限り、データをインデックス書き込みデバイスに関連付けるためのその他の技術が使用されてもよい。

【００２９】

一部の実施形態では、インデックス書き込みデバイスは、以下のように、インデックスのなかの情報を書き込む。先ず、データソースからデータが抽出され、データソースから抽出されたデータの中のテキスト、及び抽出されたデータの中のその他のデータ形式のファイルのリンクアドレスが、データストレージデバイスに格納されたテキストファイルに保存される。次に、データに関連付けられたカテゴリと、それらのカテゴリの対応するインデックスとの間の関係に基づいて、同じインデックスに対応するデータは、同じファイルに格納され、インデックス書き込みデバイスの識別子と、テキストファイルに対応するインデックスの識別子とが、ファイル又はそのファイル名に記録される。一部の実施形態では、インデックスの識別子は、図２Ａのステップ２２において説明されたように、インデックスの通し番号であり、インデックス書き込みデバイスの識別子は、ステップ２３において説明されたように、インデックス書き込みデバイスの通し番号である。ここで、データソースからのファイル及びデータは、ともに、同じストレージデバイスに格納される。引き続き、抽出されたデータファイルのリストが、各ファイルのファイル名及びステータスを格納するファイルステータステーブルに記録される。データファイルのステータスは、「処理済み」又は「未処理」などのように、そのデータがインデックスに書き込まれたかどうかを示す。一部の実施形態では、データファイルのファイル名は、{DATA PREFIX}_yyyy_mm_dd_hh_MM_ss_k.txtの形式である。DATA PREFIXは、データの記述を含み、yyyy

_mm_dd_hh_MM_ssは、データが抽出されたときの年、月、日、時、分、及び秒を示し、kは、インデックス書き込みデバイスの通し番号を示す。ファイルステータステーブルは、各インデックス書き込みデバイスに格納されてよい。データインデクシングを実施するために、インデックス書き込みデバイスは、ファイルステータステーブルを探索し、自身の通し番号に対応するファイルを、ファイルを格納しているストレージデバイスから読み出し、ファイル又はファイル名に記録されているインデックスの識別子にしたがって、対応するインデックスにファイルのなかのデータを書き込む。インデックス書き込みデバイスは、ファイルを時系列で読み出す。データがインデックスに書き込まれた後、ファイルのステータスは、「処理済み」としてファイルステータステーブルに記録される。

【００３０】

検索エンジンインデクシングのためのシステムの実施形態の実装形態が、以下で説明される。システムは、幾つかのモジュール又はユニットを含むものとして説明される。モジュール／ユニットは、１つ又は２つ以上のプロセッサ上で実行されるソフトウェアコンポーネントとして、プログラマブルロジックデバイス及び／若しくは特定の機能を実施するように設計された特定用途向け集積回路などのハードウェアとして、又はそれらの組み合わせとして実装することができる。一部の実施形態では、モジュール／ユニットは、本発明の実施形態で説明される方法をコンピュータデバイス（パソコン、サーバ、ネットワーク機器など）に実行させるための幾つかの命令を含み且つ不揮発性のストレージ媒体（光ディスク、フラッシュストレージデバイス、モバイルハードディスクなど）に格納可能であるソフトウェア製品の形で具現化することができる。モジュール／ユニットは、１つのデバイス上に実装されてよい、又は複数のデバイスに跨って分散されてよい。モジュール／ユニットの機能は、互いに合体されてよい、又は更に複数のサブモジュール／サブユニットに分割されてよい。

【００３１】

図３は、検索エンジンインデクシングシステムの一実施形態を示したブロック図である。検索サービスシステム３０は、検索サービスを実施するために使用され、このシステムは、分類モジュール３１と、インデックス書き込みモジュール３２と、インデックスストレージモジュール３３とを含む。検索サービスシステム３０は、図３に示されるように、１つ又は２つ以上のインデックス書き込みモジュール３２を含んでよい。分類モジュール３１は、データソースから受信されたデータを、図２Ａにおいて説明された所定のデータ分類方式にしたがってカテゴリ分けするように構成される。各データカテゴリと既定のインデックスとの間の対応関係が格納される。ここで、各データカテゴリは、１つのインデックスのみに対応する。インデックス書き込みモジュール３２は、図２Ａのステップ２３において説明されたインデックス書き込みデバイスの全ての機能を有する。すなわち、インデックス書き込みモジュール３２は、各カテゴリのデータを、分類モジュール３１に格納された関係にしたがってインデックスに書き込んでよい。インデックスストレージモジュール３３は、インデックスを格納するように構成されてよい。

【００３２】

検索エンジンインデクシングシステム３０は、随意として、各データカテゴリが１つのインデックス書き込みモジュールのみに割り振られるようにインデックス書き込みデバイスへの各データカテゴリの割り振りを行うように構成された割り振りモジュールを含んでよい。したがって、インデックス書き込みモジュール３２は、更に、分類モジュール３１に格納された関係にしたがって各カテゴリのデータをそれに対応するインデックスに書き込むように構成されてよい。また、検索エンジンインデクシングシステム３０は、更に、データソースにデータを格納するように構成されたソースデータストレージモジュールを含んでよい。検索エンジンインデクシングシステム３０は、更に、検索クエリを受信し、受信された検索クエリにしたがってインデックスストレージモジュール３３において検索するべきインデックスを決定し、そのインデックスを使用して検索結果を返すように構成された検索モジュールを含んでよい。通常、検索クエリは、ユーザが操作する端末から発せられる。

【００３３】
一部の実施形態では、分類モジュール３１は、更に、インデックス数決定ユニットと、インデックス番号振りユニットと、分類ユニットと、特性値割り振りユニットと、インデックス対応付けユニットとを含んでよい。インデックス番号決定ユニットは、データソースの容量と一インデックスの既定容量とにしたがって既定のインデックスの数を決定するように構成される。インデックス番号振りユニットは、既定のインデックスを連続する正の整数で表示するように構成される。分類ユニットは、既定のデータ分類方法にしたがってデータソースの中のデータを分類するように構成される。特性値割り振りユニットは、分類ユニットによる分類から得られた各データカテゴリに固有の整数値を割り振るように構成される。インデックス対応付けユニットは、各データカテゴリに割り振られた整数値を既定のインデックスの数で割ることによって分類した余りを得て、そのデータカテゴリと、得られた余りに等しい通し番号を有するインデックスとの間に対応関係を確立するように構成される。
【００３４】
分類モジュール３１のインデックス対応付けユニットは、更に、各データカテゴリに対応するインデックスの通し番号を記録するように構成されてよい。したがって、データ割り振りユニットは、インデックス決定サブユニットと、書き込みサブユニットとを含んでよい。インデックス決定サブユニットは、インデックスの通し番号と、インデックス対応付けユニットに記録された確立された関係とにしたがって、割り振られた各カテゴリのデータに対応するインデックスを決定するように構成され、書き込みサブユニットは、割り振られた各カテゴリのデータにしたがって、インデックス決定サブユニットによって決定されたインデックスにデータを書き込むように構成される。
【００３５】
分類モジュール３１が、上述のような構造を有する場合には、割り振りモジュールは、デバイス番号振りユニットと、データ割り振りユニットとを含んでよい。デバイス番号振りユニットは、連続する正の整数を使用してインデックス書き込みモジュールを表示するように構成される。データ割り振りユニットは、各インデックスの通し番号をインデックス書き込みモジュール３２の数によって分類した余りを得て、その余りに等しい通し番号を有するインデックス書き込みモジュール３２にデータを割り振るように構成される。
【００３６】
本出願の実施形態にしたがうと、データソースのなかのデータが分類され、分類によって得られた各カテゴリのデータは、インデックスにマッピングされ、このような対応関係にしたがってインデックスに書き込まれる。したがって、データの書き込みにおける共有コンフリクトが回避されえる。更に、とある所定のカテゴリのデータに関する情報を得るためには、そのデータのカテゴリに対応する１つのインデックスでのみ探索が行われればよい。したがって、検索の効率も向上されえる。更に、インデックスの容量は、大きすぎでも小さすぎでもないように選択され、したがって、過度に大きいインデックスにおける探索ゆえの効率の低下も、過度に多い小容量インデックスを開く必要も回避されえる。その結果、検索効率が向上されえて、ゆえに、検索サービスの質も向上されえる。
【００３７】
説明を容易にするために、上記のシステムは、機能にしたがって各種のモジュールに分割され、それぞれ説明される。ただし、各モジュールの機能は、本出願の実施では、１つ又は２つ以上のソフトウェア及び／又はハードウェアとして実装されてよい。
【００３８】
当業者ならば、本出願の実施形態が、方法、システム、又はコンピュータ製品として提供されえることを理解するべきである。したがって、本出願は、完全なるハードウェアの実施形態、完全なるソフトウェアの実施形態、又はそれらの組み合わせの形態であってよい。更に、本出願は、コンピュータによって使用可能なプログラムコードを含む１つ又は２つ以上のコンピュータ可用ストレージ媒体（非限定的な例として磁気ディスクストレージ、ＣＤ－ＲＯＭ、フラッシュ、及び光ストレージを含む）に実装されるコンピュータプ

ログラム製品の形態であってよい。これらのプログラムコードは、上述された方法の実施形態の全部又は一部をコンピュータ装置に実行させるための命令を含む。

【００３９】

本出願のそれぞれの実施形態は、１つずつ説明されており、これらの実施形態の間で同じ分及び類似の部分は、参照をなされ、各実施形態では、そのほかの実施形態との違いが強調されている。具体的には、システムの実施形態は、方法の実施形態との類似性ゆえに簡単に説明され、システムの実施形態の関連モジュールは、方法の実施形態を参照される。

【００４０】

本出願は、パソコン、サーバコンピュータ、ハンドセット機器すなわち携帯用機器、フラットパネルデバイス、マルチプロセッサシステム、マイクロプロセッサベースのシステム、セットトップボックス、プログラム可能な家庭用電子機器、ネットワークＰＣ、ミニコンピュータ、大型コンピュータ、上記のシステム若しくは機器の任意の１つを含む分散コンピューティングシステムなどの、多くの汎用又は専用のコンピューティングシステム環境又はコンピューティングシステム構成に適用されてよい。

【００４１】

本出願は、本出願の実施形態にしたがった方法、システム、及びコンピュータプログラム製品の、フローチャート及び／又はブロック図を参照にして説明されている。フローチャート及び／又はブロック図における各フロー及び／又はブロック、並びにフローチャート及び／又はブロック図におけるフロー及び／又はブロックの組み合わせは、コンピュータプログラム命令として実現されえることを理解されるべきである。実際、本出願全体が、例えばプログラムモジュールなどコンピュータによって実行されるコンピュータ実行可能命令を一般的な背景として説明されてよい。一般に、プログラムモジュールは、指定のタスクを実行するための、又は指定の抽象データ型を実装するための、ルーチン、プログラム、オブジェクト、コンポーネント、データ構造などを含む。或いは、本発明は、通信ネットワークを通じて接続されたリモート処理機器がタスクを実行する分散コンピューティング環境に実装されてよい。分散コンピューティング環境では、プログラムモジュールは、ストレージ機器を含むローカル又はリモートのコンピュータストレージ媒体内に配されてよい。

【００４２】

以上の実施形態は、理解を明瞭にする目的で幾らか詳細に説明されてきたが、本発明は、提供された詳細に限定されない。本発明を実行に移すには、多くの代替的手法がある。開示された実施形態は、例示的であって限定的ではない。

適用例１：データインデクシングのための方法であって、データソースからデータを受信することと、所定のデータ分類基準にしたがって、前記データを複数のカテゴリの１つに分類することと、前記データと、前記データに関連付けられ既定の最大容量を有するインデックスとの間に対応関係を確立することと、前記データと前記インデックスとの間の前記関係を記録することと、を備え、前記インデックスは、複数のインデックスの１つであり、前記複数のインデックスの各インデックスは、一インデックス書き込みデバイスによって独占的に書き込まれる、方法。

適用例２：適用例１に記載の方法であって、更に、検索クエリを受信することと、前記受信された検索クエリに対応するインデックスを前記複数のインデックスのなかから決定することと、前記インデックスに関連付けられた格納されたデータを出力することと、
を備える方法。

適用例３：適用例１に記載の方法であって、更に、少なくとも一部には前記カテゴリの容量と前記インデックスの前記既定の最大容量とに基づいて、前記複数のインデックスの数を決定することを備える方法。

適用例４：適用例１に記載の方法であって、前記データと前記インデックスとの間に対応関係を確立することは、　前記複数のインデックスを連続する正の整数で表示することと、前記データに固有の整数値を割り当てることと、前記固有の整数値を前記複数のイン

デックスの数で割って余りを得ることと、前記データの前記カテゴリと、前記複数のインデックスの中で前記余りと同じ通し番号を有する１つのインデックスとの間に前記対応関係を確立することと、を含む、方法。

適用例５：適用例１に記載の方法であって、前記インデックス書き込みデバイスは、複数のインデックス書き込みデバイスの１つであり、前記データと前記インデックスとの間の前記関係を記録することは、前記複数のインデックス書き込みデバイスを連続する正の整数で表示することと、前記インデックスの通し番号を前記インデックス書き込みデバイスの数で割って余りを得ることと、前記データと、前記余りと同じ通し番号を有するインデックス書き込みデバイスとを関連付けることと、を含む、方法。

適用例６：適用例１に記載の方法であって、前記データを前記複数のカテゴリの１つに分類することは、少なくとも一部には前記データの特性に基づく、方法。

適用例７：適用例１に記載の方法であって、前記データと前記インデックスとの間の前記関係を記録することは、前記データの少なくとも一部分をファイルに格納することと、前記ファイル及び前記データをストレージデバイスに格納することとを含む、方法。

適用例８：データインデクシングのためのシステムであって、１つ又は２つ以上のプロセッサであって、データソースからデータを受信し、所定のデータ分類基準にしたがって、前記データを複数のカテゴリの１つに分類し、前記データと、前記データに関連付けられ既定の最大容量を有するインデックスとの間に対応関係を確立し、前記データと前記インデックスとの間の前記関係を記録するように、構成され、前記インデックスは、複数のインデックスの１つであり、前記複数のインデックスの各インデックスは、一インデックス書き込みデバイスによって独占的に書き込まれる、１つ又は２つ以上のプロセッサと、前記１つ又は２つ以上のプロセッサに結合され、前記プロセッサに命令を提供するように構成されたメモリと、を備えるシステム。

適用例９：適用例８に記載のシステムであって、前記１つ又は２つ以上のプロセッサは、更に、検索クエリを受信し、前記受信された検索クエリに対応するインデックスを前記複数のインデックスのなかから決定し、前記インデックスに関連付けられた格納されたデータを出力するように、構成される、システム。

適用例１０：適用例８に記載のシステムであって、前記１つ又は２つ以上のプロセッサは、更に、少なくとも一部には前記カテゴリの容量と前記インデックスの前記既定の最大容量とに基づいて、前記複数のインデックスの数を決定するように構成される、システム。

適用例１１：適用例８に記載のシステムであって、前記データと前記インデックスとの間に対応関係を確立することは、前記複数のインデックスを連続する正の整数で表示することと、前記データに固有の整数値を割り当てることと、前記固有の整数値を前記複数のインデックスの数で割って余りを得ることと、前記データの前記カテゴリと、前記複数のインデックスの中で前記余りと同じ通し番号を有する１つのインデックスとの間に前記対応関係を確立することと、を含む、システム。

適用例１２：適用例８に記載のシステムであって、前記インデックス書き込みデバイスは、複数のインデックス書き込みデバイスの１つであり、前記データと前記インデックスとの間の前記関係を記録することは、前記複数のインデックス書き込みデバイスを連続する正の整数で表示することと、前記インデックスの通し番号を前記インデックス書き込みデバイスの前記数で割って余りを得ることと、前記データと、前記余りと同じ通し番号を有するインデックス書き込みデバイスとを関連付けることと、を含む、システム。

適用例１３：適用例８に記載のシステムであって、前記データを前記複数のカテゴリの１つに分類することは、少なくとも一部には前記データの特性に基づく、システム。

適用例１４：適用例８に記載のシステムであって、前記データと前記インデックスとの間の前記関係を記録することは、前記データの少なくとも一部分をファイルに格納することと、前記ファイル及び前記データをストレージデバイスに格納することとを含む、システム。

【図１】

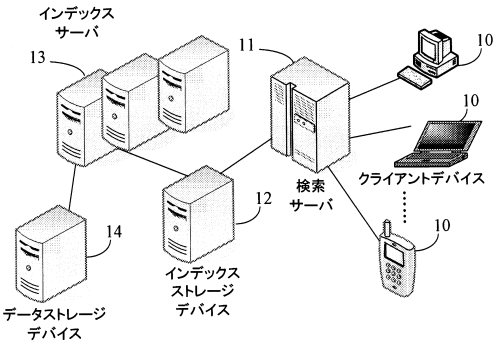

FIG. 1

【図２Ａ】

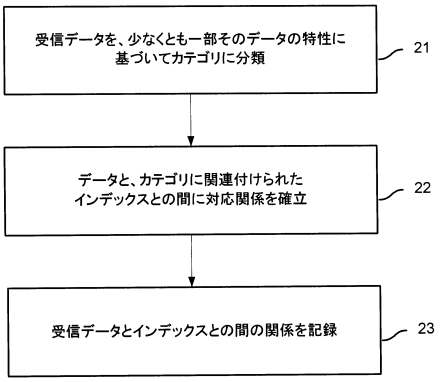

FIG. 2A

【図２Ｂ】

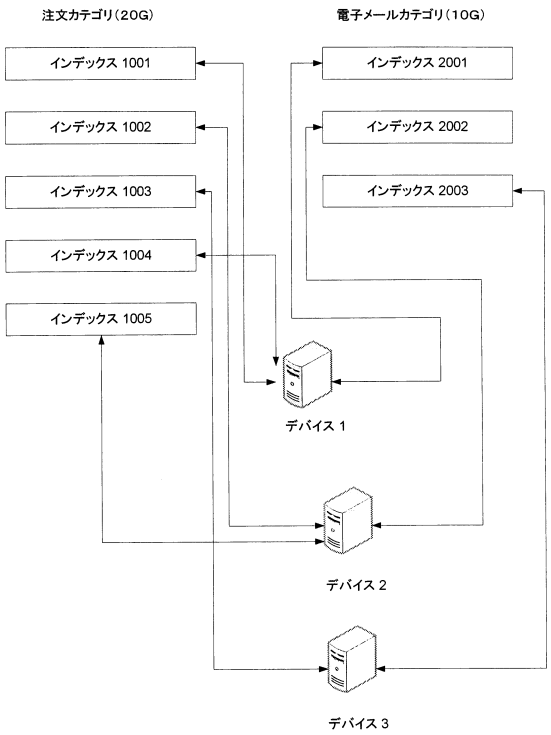

FIG. 2B

【図３】

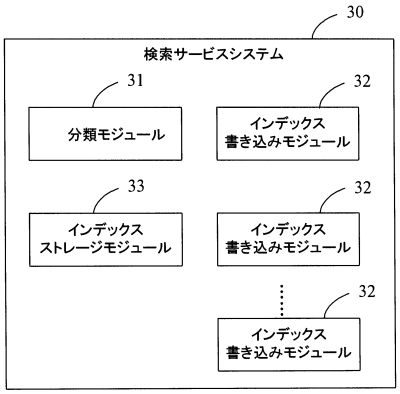

FIG. 3

フロントページの続き

審査官　梅本　達雄

(56)参考文献　特開２００８－０５２５１２（ＪＰ，Ａ）
特開２００２－２４５０３９（ＪＰ，Ａ）
国際公開第２００７／０８７６２９（WO，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０
Ｇ０６Ｆ　１２／００