(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4724357号
(P4724357)
(45)発行日　平成23年7月13日(2011.7.13)　(24)登録日　平成23年4月15日(2011.4.15)

(51)Int.Cl.　F I
G06F 17/28 (2006.01)　G06F 17/28 C

請求項の数 7　(全 17 頁)

(21)出願番号　特願2003-125929(P2003-125929)
(22)出願日　平成15年4月30日(2003.4.30)
(65)公開番号　特開2003-345796(P2003-345796A)
(43)公開日　平成15年12月5日(2003.12.5)
審査請求日　平成18年4月26日(2006.4.26)
審判番号　不服2008-6010(P2008-6010/J1)
審判請求日　平成20年3月10日(2008.3.10)
(31)優先権主張番号　10/137,456
(32)優先日　平成14年4月30日(2002.4.30)
(33)優先権主張国　米国(US)

(73)特許権者　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100077481
弁理士　谷　義一
(74)代理人　100088915
弁理士　阿部　和夫
(74)復代理人　100115624
弁理士　濱中　淳宏
(74)復代理人　100084191
弁理士　合田　潔

最終頁に続く

(54)【発明の名称】コンピュータ可読媒体及び単語情報を得るコンピュータ実行方法並びに単語情報を格納する方法

(57)【特許請求の範囲】
【請求項１】
テキストアナライザとして動作するコンピュータにより実行され、複数の異なる自然言語処理において使用される複数のレキシコンからまとめて単語情報を取得するコンピュータ実行方法であって、
前記コンピュータは、
コンピュータ記憶媒体と、
プロセッサと、
前記コンピュータ記憶媒体に格納され、前記プロセッサ上で実行可能なプログラムとを備え、
前記プログラムは、複数の異なる自然言語処理を実行するために、前記プロセッサによりアクセス可能なコンピュータ記憶媒体に格納されたレキシコンから単語情報を取得する命令を含み、
各レキシコンは、
複数の単語を格納する単語リストセクションと、
複数の組のデータセクションであって、各組のデータセクションは、前記単語リストセクションの各単語に対応し、各組のデータセクションの各データセクションは、前記各単語について選択された実質的に異なる情報を格納するデータセクションと、
前記複数の組のデータセクションと分離した、前記単語リストセクションの各単語についての複数のポインタを格納する索引セクションであって、各複数のポインタは、ある自

然言語処理に関連付けられた第１の組のポインタと、異なる自然言語処理に関連付けられた第２の組のポインタとを有し、前記第１の組のポインタは、前記第２の組のポインタと異なり、各ポインタは、前記複数の組のデータセクションのデータをポイントする索引セクションとを備え、
前記プロセッサは、前記コンピュータ記憶媒体から前記プログラムを読み出し、前記プログラムを実行し、
前記方法は、
前記プロセッサが、類似した情報を有する前記各レキシコンの前記複数の組のデータセクションに選択的にアクセスするステップと、
前記プロセッサが、実行される特定の自然言語処理に応じて前記第１または第２の組のポインタを使用し、アクセスしたデータセクションから情報を取得するステップと
を含むことを特徴とするコンピュータ実行方法。
【請求項２】
前記選択的にアクセスするステップは、類似する情報を有する前記各レキシコンの少なくとも２つのデータセクションの単語情報を組み合わせるステップを含むことを特徴とする請求項１に記載のコンピュータ実行方法。
【請求項３】
前記選択的にアクセスするステップは、
第１のレキシコンのデータセクションから単語情報を取得するステップと、
第２のレキシコンのデータセクションから単語情報を取得するステップと、
前記第２のレキシコンのデータセクションの情報だけを使用するステップと
を含むことを特徴とする請求項１に記載のコンピュータ実行方法。
【請求項４】
前記選択的にアクセスするステップは、停止インディケータが見つかるまで前記各レキシコンの類似するデータセクションから単語情報を取得するステップを含むことを特徴とする請求項１に記載のコンピュータ実行方法。
【請求項５】
前記データセクションに選択的にアクセスするステップは、選択された順序で前記複数のレキシコンに順次アクセスするステップを含み、
前記プログラムは、各レキシコンからの単語情報を読み出すか否かの第１の命令を備え、
前記プログラムは、前記複数のレキシコンの２つ以上からの単語情報を組み合わせる第２の命令を備えることを特徴とする請求項１に記載のコンピュータ実行方法。
【請求項６】
前記データセクションに選択的にアクセスするステップは、
アクセスする各レキシコンについて、
特定の自然言語処理に応じて、対応する索引セクションのポインタ識別を確定するために、所与の単語に応じて対応する単語リストセクションにアクセスするステップと、
前記対応する索引セクションのポインタを得るために、前記ポインタ識別を使用するステップと、
前記複数のデータセクションのうちのどの対応するデータセクションに前記単語についての情報があるか、及び前記情報が前記対応するデータセクションのどこに位置するかを確定するために、前記ポインタを使用するステップと
を含むことを特徴とする請求項１に記載のコンピュータ実行方法。
【請求項７】
前記選択的にアクセスするステップは、選択された順序で前記レキシコンに順次アクセスするステップを含むことを特徴とする請求項１に記載のコンピュータ実行方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】

本発明は、コンピュータ可読媒体及び単語情報を得るコンピュータ実行方法並びに単語情報を格納する方法に関する。より詳細には、言語またはテキストの処理に関連し、レキシコン（lexicon）を格納するための改良されたデータ構造、及びそのデータ構造を使用する方法に係る、レキシコンを有するコンピュータ可読媒体、単語情報を得るコンピュータ実行方法、単語情報を格納する方法及びそれらの方法を実施する命令を有するコンピュータ可読媒体に関する。

【０００２】

【従来の技術】

言語処理またはテキスト処理には多くのタイプのシステムが含まれる。例えば、パーサ、スペルチェッカ、文法チェッカ、ワードブレーカ、自然言語のプロセッサまたは理解システム、機械翻訳システムは、この広い範疇に該当するシステムのタイプのいくつかに過ぎない。

【０００３】

多くの言語またはテキスト処理システムに共通する重要なコンポーネントがレキシコンである。一般に、レキシコンは単語についての情報を含むデータ構造である。例えば、レキシコンは構文的情報及び意味的情報の指示を格納することができる。この例には、その単語が名詞であるか、動詞、形容詞であるかなどが挙げられる。また、異なるタイプの言語情報もレキシコンに格納することができる。しばしば、構文解析の助けとなる単語についての情報を格納するなど、特定タイプの言語処理に有用な他の情報を格納しておくことも有用である。さらに他のレキシコンでは、その単語が固有名詞か、地理的な場所かなどについての指示が有用であることもある。

【０００４】

動作の際は、単語の入力文字列を受け取ると、言語またはテキスト処理システムはレキシコンにアクセスして、各単語についての格納された情報を得る。入力文字列中の各単語についての情報を集めると、言語またはテキスト処理システムはその入力文字列を処理するが、これには存在する可能性のあるあいまい性をその単語の情報に基づいて解消することが含まれる場合がある。例えば、自然言語処理システムでは、レキシコンは入力文字列中の各単語に品詞を割り当てる。次いで構文パーサがどの品詞の割り当てが適切であるかを判断し、入力文字列から構造を構築し、それを解釈のために意味コンポーネントに渡すことができる。

【０００５】

いくつかの文献に上述のような従来の技術に関連した技術内容が開示されている（例えば、非特許文献１参照）。

【０００６】

【非特許文献１】

P.T.Sato著「A COMMON PARSING SCHEME FOR LEFG-AND RIGHT-BRANCHING LANGUAGES」Computational Linguistics、Vol1.14、No.1、１９８８年冬、ｐ．２０－３０

【０００７】

【発明が解決しようとする課題】

一般に、レキシコンの各項目は単一の大きなバイナリオブジェクトからなる。このフォーマットでは、情報へのアクセスは可能であるが、項目全体を読み込まずに、一般に使用される語彙情報への局所的なアクセスを容易には行うことができない。ある単語項目に関連するすべての情報をレキシコンから読み込まなければならない場合、特にその単語項目の情報のうちわずかな部分だけを必要とする場合には、より多くのメモリと処理時間が必要とされる。

【０００８】

語彙情報の変更または追加も難しい。具体的には、レキシコンを修正する、あるいはさらに情報を追加するために、レキシコンの作成者は、非常に複雑なデータ構造の整合性と編成を保ちながら、各項目中のすべてのビット、属性またはその他の情報を複製してから、所望の情報を変更するか、あるいは情報を追加しなければならない。

【０００９】
このため、上述の不都合点の１つ、一部、あるいはすべてに対処する改良したレキシコンデータ構造が必要とされる。
【００１０】
本発明は、このような課題に鑑みてなされたもので、その目的とするところは、レキシコンを格納するための高い柔軟性と効率を有する改良されたデータ構造を提供しそのデータ構造を使用可能とする、コンピュータ可読媒体及び単語情報を得るコンピュータ実行方法並びに単語情報を格納する方法を提供することにある。
【００１１】
【課題を解決するための手段】
本発明の一態様は、言語処理システムで使用するために適合された単語情報を有する、コンピュータ可読媒体に格納された単語レキシコンである。このレキシコンは、複数の単語を記憶する単語リストセクションと、それら複数の単語の単語情報を記憶する複数のデータセクションとを含む。複数のデータセクションは互いと単語リストセクションとから分離されている。単語情報にアクセスするために索引セクションが提供され、このセクションには複数のデータセクション中のデータをポイントするポインタが格納される。使用するポインタの識別は、単語リストセクション中の対応する単語に応じて決まる。
【００１２】
この改良されたレキシコン構造は、これまで得ることのできなかった柔軟性と効率を提供する。索引セクションと複数のデータセクションは、利用可能なコンピュータリソースなど言語処理システムの必要性に合わせて、レキシコンを適合することを可能にする。さらなる実施形態では、このレキシコン構造は、分類に基づいて単語情報を分類またはグループ化することを可能にする。例えば、この分類は、その単語項目が名詞か、動詞か、形容詞かなど、単語項目の品詞に基づくことができる。これにより、分類に応じて単語情報に選択的にアクセスすることができる。例示的実施形態では、対応する単語情報の分類を示す表示がポインタ中に提供される。
【００１３】
本発明の他の態様は、複数のデータセクションに単語情報を格納し、索引セクションにポインタ情報を格納し、単語リストセクションに単語リストを格納する、コンピュータによって実施される方法を含み、単語リストは、選択された単語に関連付けられた対応するポインタを識別する情報を有する。同様に、別の態様は、上述で提供されるレキシコンのデータ構造を使用して単語情報にアクセスすることである。
【００１４】
上述のレキシコン構造は、いくつかのレキシコンから情報を得ることが望ましい場合に特に有用であり、これは本発明のさらに別の態様である。一般に、各レキシコンのデータセクションに選択的にアクセスすることにより、特定の単語項目についての複数のレキシコンのデータを必要に応じて組み合わせる、無視する、または選択することができる。
【００１５】
【発明の実施の形態】
以下、図面を参照して本発明の実施形態を詳細に説明する。
【００１６】
図１に、通例はテキスト文字列の形で言語入力１２を受け取り、言語入力１２を処理して、通例は同じくテキスト文字列の形で言語出力１４を提供する言語またはテキスト処理システム１０を概略的に表す。数例を挙げると、言語処理システム１０は、例えばスペルチェッカ、文法チェッカ、あるいは自然言語プロセッサとして実施することができる。当業者には理解されるように、言語処理システム１０は、スタンドアロンアプリケーション、または別のシステムからのアクセスが可能な、あるいは別のシステムに含まれるモジュールやコンポーネントとすることができる。
【００１７】
一般に、言語処理システムはテキストアナライザ２０及びレキシコン２２を含む。テキス

トアナライザ２０は、入力１２を受け取り、レキシコン２２にアクセスしてレキシコン２２から情報を得、単語情報を処理して出力１４を提供するコンポーネントまたはモジュールを図式的に表している。本発明の一態様は、その利用例によって必要とされる可能性のある必要な情報を効率的にテキストアナライザ２０に提供するためのレキシコン２２の改良されたデータ構造である。レキシコン２２は、多くの言語処理システムと多くの形態のテキストアナライザに使用できる独立したコンポーネントであることを考慮して、テキストアナライザ２０とレキシコン２２の一般的な相互作用について説明するが、本発明の理解に必要でないため、各種形態のテキストアナライザに関する具体的な詳細については述べない。

【００１８】

本発明のさらなる詳細な説明に入る前に、動作環境の概要を見ておくと有用であろう。図２は、本発明の実施が可能な適切なコンピューティングシステム環境５０の一例である。コンピューティングシステム環境５０は適切なコンピューティング環境の一例に過ぎず、本発明の使用または機能性の範囲について何らの制限を示唆するものではない。またコンピューティングシステム環境５０は、その例示的動作環境に示す構成要素の任意の１つまたは組み合わせに関連する依存関係または必要性を有するものとも解釈すべきでない。

【００１９】

本発明は、数多くの他の汎用または特殊目的のコンピューティングシステム環境または構成で動作することができる。本発明に使用するのに適している可能性があるよく知られるコンピューティングシステム、環境、及び／または構成の例には、これらに限定しないが、パーソナルコンピュータ、サーバコンピュータ、ハンドヘルドまたはラップトップデバイス、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化製品、ネットワークＰＣ（personal computer）、ミニコンピュータ、メインフレームコンピュータ、上述のシステムまたはデバイスのいずれかを含む分散コンピューティング環境などがある。

【００２０】

本発明は、コンピュータによって実行されるプログラムモジュールなどのコンピュータ実行可能命令の一般的な状況で説明することができる。一般に、プログラムモジュールには、特定タスクを行うか、あるいは特定の抽象データ型を実装するルーチン、プログラム、オブジェクト、コンポーネント、データ構造などが含まれる。本発明は、通信ネットワークを通じてリンクされたリモートの処理装置によってタスクを行う分散コンピューティング環境で実施することもできる。分散コンピューティング環境では、メモリ記憶装置を含むローカル及びリモート両方のコンピュータ記憶媒体にプログラムモジュールを置くことができる。これらのプログラム及びモジュールによって行われるタスクについて、以下で図面を用いて説明する。当業者は、以下の説明及び図を、任意形態のコンピュータ可読媒体に書き込むことのできるプロセッサ実行可能命令として実行することができる。

【００２１】

図２を参照すると、本発明を実施する例示的システムは、コンピュータ６０の形態の汎用コンピューティングデバイスを含む。コンピュータ６０の構成要素には、これらに限定しないが、プロセッサ７０、システムメモリ８０、及びシステムメモリを含む各種のシステムコンポーネントをプロセッサ７０に結合するシステムバス７１が含まれる。システムバス７１は、各種のバスアーキテクチャの任意のものを使用したメモリバスまたはメモリコントローラ、ペリフェラルバス、及びローカルバスを含む数タイプのバス構造のいずれでもよい。例として、このようなアーキテクチャには、ＩＳＡ（Ｉｎｄｕｓｔｒｙ Ｓｔａｎｄａｒｄ Ａｒｃｈｉｔｅｃｔｕｒｅ）バス、ＭＣＡ（Ｍｉｃｒｏ Ｃｈａｎｎｅｌ Ａｒｃｈｉｔｅｃｔｕｒｅ）バス、ＥＩＳＡ（Ｅｎｈａｎｃｅｄ ＩＳＡ）バス、ＶＥＳＡ（Ｖｉｄｅｏ Ｅｌｅｃｔｒｏｎｉｃｓ Ｓｔａｎｄａｒｄｓ Ａｓｓｏｃｉａｔｉｏｎ）バス、及びメザニンバスとも称されるＰＣＩ（Ｐｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔｓ）バスが含まれるが、これらに限定しない。

【００２２】

コンピュータ６０は通例各種のコンピュータ可読媒体を含む。コンピュータ可読媒体は、コンピュータ６０がアクセスできる任意の利用可能媒体でよく、揮発性及び不揮発性の媒体、リムーバル及びノンリムーバル媒体が含まれる。これに限定しないが、例としてコンピュータ可読媒体は、コンピュータ記憶媒体及び通信媒体を含むことができる。コンピュータ記憶媒体は、コンピュータ可読命令、データ構造、プログラムモジュール、またはその他のデータなどの情報を記憶するための任意の方法または技術で実現された不揮発性及び不揮発性の媒体、リムーバル及びノンリムーバルの媒体を含む。コンピュータ記憶媒体には、これらに限定しないが、ＲＡＭ（random access memory）、ＲＯＭ（read only memory）、ＥＥＰＲＯＭ（electrically erasable PROM）、フラッシュメモリまたは他のメモリ技術、ＣＤ（compact disc [disk]）－ＲＯＭ、デジタル多用途ディスク（ＤＶＤ）または他の光ディスクストレージ、磁気カセット、磁気テープ、磁気ディスクストレージまたは他の磁気記憶装置、あるいは、所望の情報の記憶に用いることができ、コンピュータ６０によるアクセスが可能な任意の他の媒体が含まれる。

【００２３】

通信媒体は、通例、搬送波または他の搬送機構などの、変調データ信号にコンピュータ可読命令、データ構造、プログラムモジュール、または他のデータを統合し、任意の情報伝達媒体を含む。用語「変調データ信号」とは、信号中に情報を符号化するような方式でその特性の１つまたは複数を設定または変化させた信号を意味する。例として、通信媒体には、有線ネットワークまたは直接配線接続などの有線媒体と、音響、ＲＦ（radio frequencies）、赤外線、及び他の無線媒体などの無線媒体が含まれるが、これらに限定しない。上記の媒体のいずれの組み合わせもコンピュータ可読媒体の範囲に含めるべきである。

【００２４】

システムメモリ８０には、ＲＯＭ８１及びＲＡＭ８２など、揮発性及び／または不揮発性メモリの形態のコンピュータ記憶媒体が含まれる。起動時などにコンピュータ６０内の要素間の情報転送を助ける基本ルーチンを含むＢＩＯＳ（Basic Input/Output System）８３は、通例ＲＯＭ８１に記憶される。ＲＡＭ８２は通例、プロセッサ７０から即座にアクセス可能な、かつ／または現在プロセッサ７０によって操作中のデータ及び／またはプログラムモジュールを含む。これらに限定しないが、例として、図２にはオペレーティングシステム８４、アプリケーションプログラム８５、他のプログラムモジュール８６、及びプログラムデータ８７を示している。

【００２５】

コンピュータ６０は、他のリムーバル／ノンリムーバル、揮発性／不揮発性のコンピュータ記憶媒体も含むことができる。単なる例として、図２には、ノンリムーバル、不揮発性の磁気媒体の読み取りまたは書き込みを行うハードディスクドライブ９１、リムーバル、不揮発性の磁気ディスク１０２の読み取りまたは書き込みを行う磁気ディスクドライブ１０１、及びＣＤ－ＲＯＭや他の光媒体などのリムーバル、不揮発性の光ディスク１０６の読み取りまたは書き込みを行う光ディスクドライブ１０５を示す。例示的動作環境で使用できるこの他のリムーバル／ノンリムーバル、揮発性／不揮発性のコンピュータ記憶媒体には、これらに限定しないが、磁気テープカセット、フラッシュメモリカード、デジタル多用途ディスク、デジタルビデオテープ、ソリッドステートＲＡＭ、ソリッドステートＲＯＭなどが含まれる。ハードディスクドライブ９１は通例、インタフェース９０などのノンリムーバルのメモリインタフェースを通じてシステムバス７１に接続され、磁気ディスクドライブ１０１及び光ディスクドライブ１０５は通例、インタフェース１００などリムーバルなメモリインタフェースによってシステムバス７１に接続される。

【００２６】

上記で説明し、図２に示したドライブとそれに関連付けられたコンピュータ記憶媒体は、コンピュータ可読命令、データ構造、プログラムモジュール、及びコンピュータ６０のその他のデータの記憶を提供する。例えば図２では、ハードディスクドライブ９１にオペレーティングシステム９４、アプリケーションプログラム９５、他のプログラムモジュール９６、及びプログラムデータ９７を記憶している。これらのコンポーネントは、オペレー

ティングシステム８４、アプリケーションプログラム８５、他のプログラムモジュール８６、及びプログラムデータ８７と同じものでも、異なるものでもよいことに留意されたい。ここではオペレーティングシステム８４、アプリケーションプログラム８５、他のプログラムモジュール８６、及びプログラムデータ８７には、それらが少なくとも異なるコピーであることを表すために異なる参照符号をつけている。

【００２７】

ユーザは、キーボード１１２、マイクロフォン１１３、手書きタブレット１１４、及びマウス、トラックボール、タッチパッドなどのポインティングデバイス１１１などの入力装置を通じてコンピュータ６０にコマンドと情報を入力することができる。他の入力装置（図示せず）には、ジョイスティック、ゲームパッド、衛星放送受信アンテナ、スキャナなどがある。これら及び他の入力装置は、システムバスに結合されたユーザ入力インタフェース１１０を通じてプロセッサ７０に接続することが多いが、パラレルポート、ゲームポート、あるいはユニバーサルシリアルバス（ＵＳＢ）など他のインタフェース及びバス構造によって接続することも可能である。モニタ１４１または他タイプの表示装置も、ビデオインタフェース１４０などのインタフェースを介してシステムバス７１に結合される。コンピュータは、モニタに加えて、スピーカ１４７やプリンタ１４６など他の周辺出力装置も含むことができ、それらは出力周辺インタフェース１４５を通じて接続することができる。

【００２８】

コンピュータ６０は、リモートコンピュータ１３０など１つまたは複数のリモートコンピュータへの論理接続を使用するネットワーク環境で動作することができる。リモートコンピュータ１３０はパーソナルコンピュータ、ハンドヘルドデバイス、サーバ、ルータ、ネットワークＰＣ、ピアデバイス、あるいはその他の一般的なネットワークノードでよく、通例はコンピュータ６０との関連で上記で挙げた要素の多くまたはすべてを含む。図２に示す論理接続には、構内ネットワーク（ＬＡＮ）１２１と広域ネットワーク（ＷＡＮ）１２３が含まれるが、他のネットワークを含むことも可能である。このようなネットワーキング環境は、オフィス、企業内のコンピュータネットワーク、イントラネット、及びインターネットに一般的に見られる。

【００２９】

ＬＡＮネットワーキング環境で使用する場合、コンピュータ６０はネットワークインタフェースまたはアダプタ１２０を通じてＬＡＮ１２１に接続される。ＷＡＮネットワーキング環境で使用する場合、コンピュータ６０は通例、インターネットなどのＷＡＮ１２３を通じて通信を確立するためのモデム１２２またはその他の手段を含む。モデム１２２は内蔵型でも外付け型でもよく、ユーザ入力インタフェース１１０または他の適切な機構を介してシステムバス７１に接続することができる。ネットワーク環境では、コンピュータ６０との関連で図示したプログラムモジュール、またはその一部はリモートのメモリ記憶装置に格納することができる。これに限定しないが、例として図２ではリモートアプリケーションプログラム１３５がリモートコンピュータ１３０に常駐している。図のネットワーク接続は例示的なものであり、コンピュータ間に通信リンクを確立する他の手段を使用できることは理解されよう。

【００３０】

テキストアナライザ２０は、コンピュータ６０、またはリモートコンピュータ１３０などコンピュータ６０と通信する任意のコンピュータに常駐できることを理解されたい。同様に、レキシコン２２は、コンピュータ６０の上述の記憶装置の任意のものに常駐するか、または適切な通信リンクを通じてアクセス可能にすることができる。

【００３１】

図３は、レキシコン２２の図式表現である。図の例示的実施形態では、レキシコン２２は、ヘッダセクション１６０、単語リストセクション１６２、索引テーブルセクション１６４、索引セクション１６６、２つ以上のレキシコンデータセクション１６８（ここでは例として１６個のセクション、１６８ａ、１６８ｂ、１６８ｃ、１６８ｄ、１６８ｅ、１６

８ｆ、１６８ｇ、１６８ｈ、１６８ｉ、１６８ｊ、１６８ｋ、１６８ｌ、１６８ｍ、１６８ｎ、１６８ｏ、１６８ｐ）、及び文字列ヒープセクション１７０を含む。

【００３２】

ヘッダセクション１６０は一般に、レキシコン２２の構造についての情報を格納する。ヘッダセクション１６０は、例えば、レキシコンの名前やバージョンについての情報を含むことができる。ヘッダセクション１６０はまた、メモリオフセット及び各セクション１６２、１６４、１６６、１６８ａ～１６８ｐ、及び１７０のサイズについての情報も含むことができる。セクション１６２は、レキシコン２２の単語リストを含む。セクション１６２に単語リストを実施するには、任意の適切なフォーマットを用いることができる。特に有用なフォーマットの１つは、よく知られるデータ構造技術である「トライ（ｔｒｉｅ）」構造で単語リストを格納するものである。このフォーマットの利点としては、特定の接頭辞で始まる可能性のある単語がいくつあるかを容易に判定できることが挙げられ、これは例えば手書き認識や、ユーザが特定の文字を書いた可能性を確かめる必要がある際に有用である。このフォーマットではまた、トラバース（traversal）のパスを前方向と後ろ方向の両方で知ることができる。上記のように、セクション１６２には他の形態の単語リスティングを使用することができる。例えば単純なテーブルやリストを使用することができる。さらに別の実施形態では「差分」技術を使用して単語リストを格納することができ、この場合は連続した単語の記号または文字の違いを格納する。

【００３３】

セクション１６４を説明する前に、セクション１６６と、複数のセクション１６８とのその関係を初めに説明しておくと有用であろう。「従来の技術」の項で述べたように、現在のレキシコンでは、必要とするのが情報の一部だけであっても、特定の単語項目に関連付けられたすべての情報を読み出すことが必要とされる。セクション１６８ａ～１６８ｐは、レキシコン中の各単語項目のデータを所望の方式で編成することを可能にし、関連するレキシコン情報を概ねともにグループ化することができる。例えば、セクション１６８ａ～１６８ｐの１つを使用してスペルチェックに関連する情報を格納し、別のセクションに標準的な言語分類に関連する情報を格納することができる。索引セクション１６６は一般に、単語リストセクション１６２の単語項目に応じて、セクション１６８ａ～１６８ｐに格納されたデータへのポインタ（例えばセットとしてグループ化した）を提供する。すなわち、単語リストセクション１６２（例えばトライ構造）は、索引セクション１６６へのアクセスポイント（オフセット）を直接または間接的に決定する。一般に、単語情報を得るこの方法は、所与の単語に応じて単語リストセクションにアクセスして、索引セクションのポインタ識別を確定することを含む。このポインタ識別を使用して、索引セクションで単語のポインタを得る。次いでこのポインタを使用して複数のデータセクションのうちどのデータセクションに所与の単語についての情報があるか、そしてその情報がそのデータセクションのどこに位置するかを確定する。したがって、セクション１６２にある特定の単語項目に対して、索引セクション１６６を通じて、セクション１６８ａ～１６８ｐに格納されたその単語の対応するレキシコンデータに選択的にアクセスすることができ、それにより所与の単語のすべての単語情報を処理するか、または少なくとも読み出す必要がない。

【００３４】

特に有用な一実施形態では、セクション１６２にある各単語項目についてのセクション１６８ａ～１６８ｐの索引セクション１６６中のポインタまたはポインタのセットを、その単語項目が名詞か、動詞か、形容詞かなど、その品詞（「ＰＯＳ（part of speech）」）によって分類する。したがって、ある単語項目のＰＯＳについてのデータは、セクション１６８ａ～１６８ｐのＰＯＳ情報への一連のポインタとなる。したがって、２つのＰＯＳを持つ単語項目には、セクション１６６に２つの別個のポインタセットがある。セットの１つは、第１のＰＯＳ（例えばその項目の名詞形）についての情報の位置を示し、第２のポインタセットは、もう一方のＰＯＳ（例えばその項目の動詞形）についての情報の位置を示す。このように索引セクション１６６は、単語項目のＰＯＳに基づいた、あるレベル

のレキシコンデータ分類を提供する。ここで、レキシコン２２が対象とする言語に応じて、ＰＯＳ以外の他の分類形態を使用できることを理解されたい。例えば日本語や中国語には、品詞の代わりに屈折または声調による分類を使用することができる。ここでは索引セクション１６６がＰＯＳ分類を提供するものと例示しているが、この機能は制限的あるいは必須とみなすべきでない。

【００３５】

また、中国語や日本語のような言語で使用する際には、ここで用いる意味の単語「単語（ｗｏｒｄ）」の使用には、記号、表意文字、語標（ｌｏｇｏｇｒａｍ）なども含まれることにも留意されたい。したがって、本発明の態様を使用してこれらの言語のレキシコンも構築することができ、そのレキシコンは特に断らない限りは特許請求の範囲に包含されるものとする。

【００３６】

例示的実施形態では、各ポインタは、そのポインタがセクション１６８ａ～１６８ｐのうちどれをポイントするのかに関する情報、ＰＯＳの種類に関連付けられた情報、識別されたセクション１６８ａ～１６８ｐ中で関連するデータを見つけられるオフセット値を含む。セクション１６２の所与の単語項目に関連付けられたポインタは固定することができるが、例示的実施形態では、各単語項目のポインタ数は単語項目ごとに異なってよい。このようにして、本質的な制限が常に伴うことなく、索引セクション１６６をより小型かつ柔軟にすることができる。

【００３７】

ある単語項目のセクション１６６の例示的ポインタの概略表現を次に示す。

X_1：X_2：X_3：X_4

ここでX_1は単語項目のポインタセットの最後を示すフラグであり、X_2はセクション１６８の１つを識別する情報であり、X_3はＰＯＳまたは他の分類を識別する情報であり、X_4は、X_2によって識別されるレキシコンデータのオフセットを示す値である。このフォーマットを使用して、所与の単語についてのすべての情報のポインタを連続して格納することができ、単語リストセクションに応じて直接あるいは間接的に第１のポインタが識別され、最後のポインタのフラグX_1をセットして所与の単語のポインタリストの最後を示す。一実施形態では、索引セクション１６６は大きなＤＷＯＲＤ配列（迅速なアクセスのために４バイトの量、４バイトワード配置）である。この実施形態では、１バイトの内訳は、ポインタセットの最後を示すX_1の１ビットフラグ、セクション１６８ａ～１６８ｐを示す４ビットのX_2、及びＰＯＳの種類を示す３ビットのX_3である。そしてX_4に３バイトを使用して、データがセクション１６８ａ～１６８ｐのどこに格納されているかを示す２４ビットのオフセット値を提供する。このフォーマットは一例に過ぎず、他のフォーマットも使用できることを理解されたい。同様に、この例は必須あるいは制限的と解釈すべきでない。一般に、索引セクション１６６のポインタのフォーマットは、複数のセクション１６８中のデータの位置、及び必要な場合は単語情報の１つまたは複数の分類を示すために選択される。

【００３８】

ここで、セクション１６６のポインタ項目のオフセット部分に入るのに十分な小ささのデータは、別個のセクション１６８ａ～１６８ｐではなく、直接索引セクション１６６に符号化できることにも留意されたい。この種のデータの例にはスペリング情報や単語項目の確率及び頻度データが含まれるが、これらはいずれも多くの場合はデータオフセット値に割り当てられたビットに容易に格納することができる。

【００３９】

上記のように、索引セクション１６６へのエントリは、セクション１６２の単語項目に応じて決まる。セクション１６２と１６６間の移行には各種の技術を使用することができる。第１の実施形態では、セクション１６２の各単語項目は必要とされるセクション１６６へのオフセットを含むことができる。ただしセクション１６２がトライ構造を備える場合は、トライの葉ノード構造の修正が必要となる場合がある。あるいは、トライ構造中のノ

ードのオフセットを索引セクション１６６へのオフセットとして使用することができる。例示的実施形態では、これは、ある単語項目についてのＰＯＳ索引のセットに４０バイト（１０個のＰＯＳポインタ）を割り当てることを意味する。さらなる実施形態では、オフセット値をセクション１６６の単語項目の最後に付することができる。

【００４０】

さらに別の実施形態では、索引テーブル１６４をレキシコン２２の構造中に含める。索引テーブル１６４は単語項目とセクション１６６の索引とのマッピングを可能にし、これは、セクション１６６中のポインタの数が単語項目ごとに異なる可能性がある場合に特に有用である。ただし、関連付けられた単語項目ごとに、索引セクション１６６で固定サイズ数のポインタを使用することが可能である。この構造のセクション１６６を使用すると索引テーブルセクション１６４が不要になる。この代替実施形態では、セクション１６６の索引ポインタの固定数を超えるポインタを有する単語項目が許された場合は、オーバーフローテーブルを使用することができる。

【００４１】

ここで、セクション１６２から索引セクション１６６へのオフセット、より具体的にはセクション１６８ａ～１６８ｐ中のデータをポイントするセクション１６６のポインタは、レキシコン２２からデータを検索する際の効率と速度を提供するように編成することができることに留意されたい。例えば、頻繁に使用される単語の他の情報の隣にセクション１６８ａ～１６８ｐのレキシコン情報を配置するようにオフセットポインタを編成することができ、あるいは、必要な場合は、セクション１６８ａ～１６８ｐ中の関連付けられた情報同士をより近くに編成して、ハードディスク、フロッピー（登録商標）などのコンピュータ記憶装置に格納された際に情報検索時間を短縮することができる。

【００４２】

セクション１６８ａ～１６８ｐのデータはその中に存在することができ、即ち、必要な場合は、同じセクション１６８ａ～１６８ｐに含まれる参照データへの、他のセクション１６８ａ～１６８ｐに含まれる参照データへの、セクション１６２のもとの単語項目へのポインタを提供でき、及び例証的実施形態ではまた文字列ヒープ１７０へのポインタを提供することができる。文字列ヒープ１７０は、そのデータをセクション１６８ａ～１６８ｐ中に複数の出現として格納する必要がある選択された文字列に単一の記憶位置を提供するために使用される。文字列ヒープ１７０は単一のセクションであるか、あるいはセクション１６８ａ～１６８ｐと同様のサブセクションを含むことができる。セクション１６８中の情報の他の形態には、決定木中のブールフラグ、値、単語リストなどがある。

【００４３】

複数のセクション１６８ａ～１６８ｐを使用した単語項目データの編成により、それを実装するコンピュータのメモリを多量消費することなく、特定の用途の必要性を満たすようにレキシコン２２を容易に適合することが可能になる。例えば、レキシコン２２はＲＡＭなどの高速アクセスメモリに読み込むことができるが、レキシコン中の特定タイプのデータが必要でない場合は、複数のセクション１６８ａ～１６８ｐのうちそのセクションを省略することができる。索引セクション１６６のポインタは、存在するセクション１６８ａ～１６８ｐだけを反映するように変更することができるが、さらなる実施形態では、セクション１６８ａ～１６８ｐが存在すれば情報が得られ、一方セクションが存在しなければ情報が求められないので、変更は不必要である。エラーが生じないように、レキシコンに存在するセクションは例えばヘッダ１６０に記録することができる。

【００４４】

ここに記載するレキシコン構造の際立った利点は、入力１２（図１）が単語を含む場合にレキシコンのユーザまたは作成者が後の検索のためにその単語についてのどのようなタイプの情報でも入れることができる点である。さらに、ユーザによって定義された情報はレキシコンに含まれる他の情報と混在させる必要がなく、複数のセクション１６８ａ～１６８ｐの専用のセクションに格納することができる。

【００４５】

以下に挙げるのは、セクション１６８ａ～１６８ｐに適したセクションに編成されたレキシコンデータのいくつかの例である。これらは単なる例に過ぎず、レキシコン２２のデータは、利便性または理解のために任意の所望の方式で編成できることに留意されたい。ここで述べるセクションは特に有用であることが判明しているが、必須あるいは制限的なものと見なすべきではない。

【００４６】

形態データセクション－このような情報は、発音ならびに様々な単語の時制についてのその単語の他の形を含むことができる。

【００４７】

標準作成者データセクション－この情報には、その単語項目が単数形、複数形であるか否か、あるいはその単語が有生か無生かを示すデータを含むことができる。このセクションの単語項目に関連する情報は、一般にはその単語項目についてのよく知られた情報であり、素人でも作成することができる。このように、この情報はユーザの要件に合わせて用意に変更または修正することができる。

【００４８】

標準言語データセクション－この情報には単語項目の言語学的情報が含まれる。このような情報は一般の素人にはあまり知られないが、言語学者はこの情報を容易に理解し、必要に応じて修正することができる。

【００４９】

構文解析データセクション－この情報には、自然言語の構文解析に役立つ情報が含まれる。

【００５０】

領域／主題データセクション－この情報は領域または主題のコードに関連する。例えば、この情報により、対応する単語が物理、数学、地理、食物などに関連するものであることを示すことができる。

【００５１】

スペリングデータセクション－この情報はスペルチェック、例えば方言マーキング、制限マーキングなどに関連する。制限マーキングは、卑語、頭字語、古い語など許容されるがスペルチェックの際には提案されない単語を示す。

【００５２】

複数語表現データセクション－この情報は、イディオム、固有名、本や映画の題名、オフィスの名称、地名など、複数の単語を単独に識別する必要がある際に有用である。通例、各単語項目について格納されるデータは、複数単語表現でその単語の前にくる、かつ／または後に来る単語である。

【００５３】

例えば、複数のセクション１６８ａ～１６８ｐの１つが、そのレキシコン項目の作成者のみによって見つけられる、任意の階層的な名前値の対を含むことができる。例えば、作成者が複数語表現（上記）についての固有表現（ＮＥ；Name Entity）情報を追加したい場合には、名前文字列に基づく値の対のセットをそのセクションに追加することができ、これをＸＭＬ形式で表すと次のようになる。

```
〈named-entity〉
  〈app-ne-id〉movieFinder::the_longest_day〈/app-ne-id〉
〈semantic-type〉movieFinder::movieTitle〈/semanticType〉
  〈genre〉Drama〈/genre〉
  〈URL〉http://www.movieFinder.com/fetch-movie-info/
        the_longest_day〈/URL〉
  〈movie-info〉
        〈date〉Jan.30, 1969〈/date〉
        〈running-time〉137 min.〈/running-time〉
        〈studio〉20th Century Fox〈/studio〉
  〈/movie-info〉
  〈non-rated/〉
〈/named-entity〉
```

このように、このセクションは、任意の入れ子構造になった、文字列に基づく単純な値の名前の対を表すことができる。この形式はＸＭＬのタグ属性をサポートせず、作成者がこのセクションの別個の下位要素としてその属性を符号化する。上の例では、映画の題名についてのデータは、必要な場合には格納することのできる利用例固有のデータの混合を含んでいる。

【００５４】

このレキシコン２２の構造は、レキシコンの各セクションはそのすぐ後のセクションと連続する必要がないという事実を利用することにより書き込み可能なレキシコンに対応する。すなわち、セクションは将来行われる拡張のために余分の未使用スペースを取っておくことができる。レキシコンへの更新操作は、該当する位置に新しい値を書き込むことによって行う。レキシコン２２がＤＤＬ（ダイナミックリンクライブラリ）ベースのレキシコン、あるいは予備スペースを備えない事前にコンパイルされた（静的な）ファイルベースのレキシコンとして実施される場合は、単純なフリーリスト実装により、先頭一致（ｆｉｒｓｔ－ｆｉｔ）アルゴリズムに基づいて空いている項目スペースを見つける。

【００５５】

一般に、レキシコン２２に単語情報を格納する方法は、複数のデータセクション１６８に単語情報を格納し、各データセクションは単語リスト中の単語について実質的に異なる選択された情報を格納することと、複数のデータセクション１６８と分離した索引セクション１６６にポインタ情報を格納し、各ポインタは複数のデータセクション１６８中の選択されたデータをポイントすることと、複数のデータセクション１６８及び索引セクション１６６と分離した単語リストセクション１６２に単語リストを格納し、単語リストは選択された単語に関連付けられた対応するポインタを識別する情報を有することとを含む。必要な場合は、識別値を索引テーブルセクション１６４に格納することができ、この場合には各識別値が単語リストセクション１６２の単語に対応し、索引セクション１６６のポインタと関連付けられる。同様に、ポインタに分類の表示を含めて単語情報を分類することもできる。

【００５６】

このレキシコン２２の構造は、いくつかのレキシコンから情報を得ることが望ましい場合に特に有用である。一般には、特定の単語項目についての複数のレキシコンの情報を必要に応じて組み合わせる、無視する、あるいは選択することができる。いくつかのレキシコンからのレキシコン情報を組み合わせる例は、核または基礎となるレキシコンに単語項目

についての第１の量の情報が含まれ、第２のレキシコンに特定領域についての単語項目についての第２の量の情報が含まれ、第３のレキシコンにユーザが決定した単語項目についての第３の量の情報が含まれる実装に見られる。

【００５７】

図４に、特定の単語項目についての情報を複数のレキシコンから得る方式を図式的に示している。図４で、レキシコン（データセクションだけによって表しているが、本来は図３に示すセクションの一部またはすべてを備える）は行に編成され、これを１８０、１８１、１８２、及び１８３で示す。図４では個々のデータセクション（１６８に対応する）を縦に表しており、この例証的実施形態では、４つのレキシコン１８０～１８３を通じて最大６個のデータセクション１９０、１９１、１９２、１９３、１９４、及び１９５にアクセスすることができる。各レキシコン１８０～１８３がデータセクション１９０～１９５すべてを含むことは必須でなく、多くの実際の事例では、すべてのレキシコンのすべてのデータセクション間にそのような対応関係は存在しないことに留意されたい。

【００５８】

図４では符号X_yを使用してレキシコンセクションのデータを示しており、Ｘはデータセクション１９０～１９５の１つを現し、Ｙはレキシコン１８０～１８３を表す。例えばレキシコン１８０は、データセクション190_{180}、193_{180}、及び195_{180}を備える。

【００５９】

レキシコン１８０～１８３のデータは同じタイプの内容を有するセクション１９０～１９５に編成されているので、レキシコン１８０～１８３にまたがって情報を容易に組み合わせる、あるいは選択することができる。所与の単語項目について、第１のレキシコン１８０の情報を調べ、次いで必要に応じて他のレキシコン１８１～１８３の同じデータセクションに進むことにより情報を得ることができる。一実施形態では、検索するデータは、実行時に定義される所望のセクションタイプのセットによって制御する。１つの変数で、レキシコンのある項目のデータを読み出すか読み出さないかを決定する。第２の変数で、調べた他のレキシコンの対応するセクションから先に読み出された項目のデータと組み合わせる、あるいはそのデータに上書きするかどうかを決定する。概略的には、レキシコンは「スタック」されており、スタック中の最上位のレキシコンのデータセクション１９０～１９５を読み取り、次いでスタックを順次下に進み、読み出すか読み出さないか、選択するか、無視するか、上書きするか、あるいは組み合わせるかについての規則に従うことによって情報を得ると考えることができる。図３に示すレキシコン構造は、実装者が、所与のセクションタイプのデータを他のレキシコンの同じセクション中のデータと組み合わせる、あるいは上書きする方式を選択することを可能にする。

【００６０】

図４で、レキシコン１８０～１８３から得た情報１８６は、セクション190_{180}、191_{181}、192_{183}、193_{180}、194_{183}、及び$195_{180+181+182}$に対応する情報を含む。この例では、セクション１９０、１９１、１９２、１９３、及び１９４のデータは、単にデータセクションの１つに停止インディケータが見つかるまで、レキシコン１８０～１８３をセクションごとに順に調べることによって得る。例えば、レキシコン１８０及び１８３はともにセクション１９０に情報を有するが、停止インディケータがセクション190_{180}で見つかるのでレキシコン１８０の情報だけが取り出される。実行時に、これによりセクション190_{183}の情報が無視される。これに対してセクション195_{182}を調べるまでに停止インディケータが見つからないので、セクション195_{180}、195_{181}、及び195_{182}の情報は組み合わせて情報$195_{180+181+182}$を形成する。必要な場合は、テキストアナライザ２０、またはテキストアナライザ２０の要求に基づいてレキシコン２２にアクセスするインタフェースモジュール（図示せず）によって実施される規則に基づいて、すべてのレキシコンのセクションにわたる情報を組み合わせる、無視する、あるいはその他の形で選択することができる。そのような規則は、例えば、他のレキシコンの対応するセクションに情報がある

かどうかに関係なく、特定のレキシコンの特定のセクションを常に使用することを指定することができる。これを図４に表しており、ここではレキシコン１８２のセクション１９２に情報があり、この情報はスタック中で上方にあるので少なくとも最初は調べられるが、レキシコン１８３のセクション１９２の情報が得られる。ただしデータの選択は、例えば上述の要領で停止ポインタを使用して単語項目ごとにしてもよい。

【００６１】

要約すると、以前には得られなかった柔軟性と効率性を提供する改良したレキシコン構造について述べた。索引セクション及び複数のデータセクションにより、テキスト処理システム及び／または利用可能なコンピュータリソースの必要性に合わせてレキシコンを適合することが可能になる。この改良されたデータ構造により、複数のレキシコンのデータに選択的にアクセスし、かつ／または必要に応じて組み合わせることも可能になる。

【００６２】

本発明について好適実施形態を参照して説明したが、当業者は、本発明の趣旨及び範囲から逸脱せずに形態及び詳細に変更を加えられることを認識されよう。

【００６３】

【発明の効果】

以上説明したように本発明によれば、レキシコンを格納するための高い柔軟性と効率を有する改良されたデータ構造を提供しそのデータ構造を使用可能とすることができる。

【図面の簡単な説明】

【図１】本発明の実施形態の言語またはテキスト処理システムのブロック図である。

【図２】本発明の実施形態の例示的環境のブロック図である。

【図３】本発明の実施形態のレキシコンの図式表現の図である。

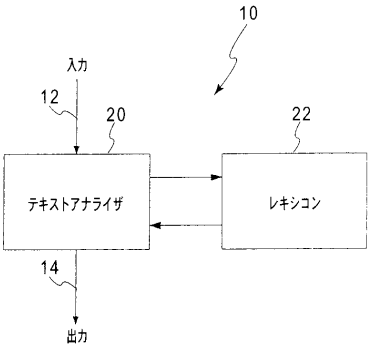
【図４】本発明の実施形態の複数のレキシコンにわたって情報を検索する、または情報にアクセスする図式表現の図である。

【符号の説明】

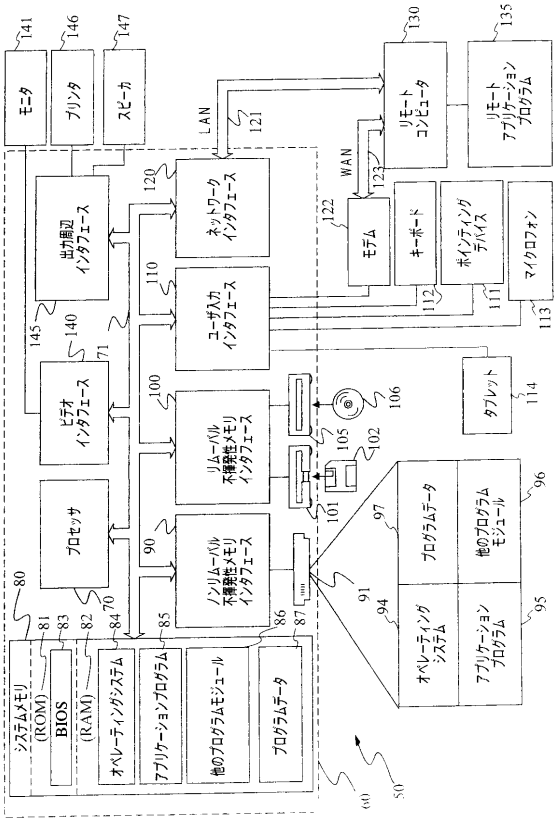
１０　言語処理システム
１２　入力
１４　出力
２０　テキストアナライザ
２２　レキシコン
５０　コンピューティングシステム環境
６０、１３０　コンピュータ
７０　プロセッサ
７１　システムバス
８０　システムメモリ
８１　ＲＯＭ
８２　ＲＡＭ
８３　ＢＩＯＳ
８４、９４　オペレーティングシステム
８５、９５、１３５　アプリケーションプログラム
８６、９６　プログラムモジュール
８７、９７　プログラムデータ
９０、１００　インタフェース
９１　ハードディスクドライブ
１０１　磁気ディスクドライブ
１０２　磁気ディスク
１０５、１０６　光ディスクドライブ
１１０　ユーザ入力インタフェース
１１１　ポインティングデバイス
１１２　キーボード

１１３　マイクロフォン
１１４　手書きタブレット
１２０　ネットワークインタフェース
１２１　ＬＡＮ
１２２　モデム
１２３　ＷＡＮ
１４０　ビデオインタフェース
１４１　モニタ
１４６　プリンタ
１４７　スピーカ
１６０、１６２、１６４、１６６、１６８、１７０
１６８ａ～１６８ｐ　セクション
１８０、１８１、１８２、１８３　レキシコン
１９０、１９１、１９２、１９３、１９４、１９５　データセクション

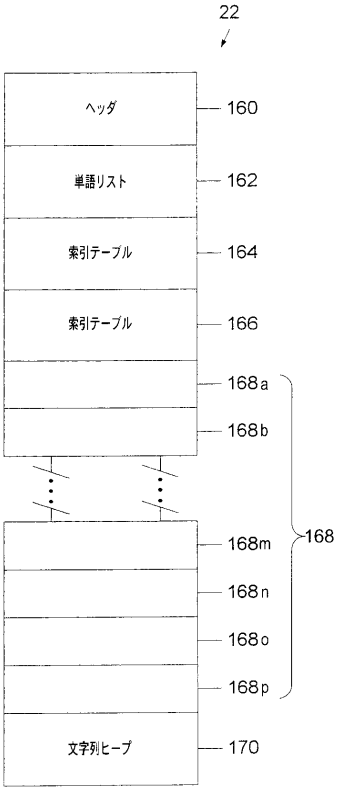
【図１】



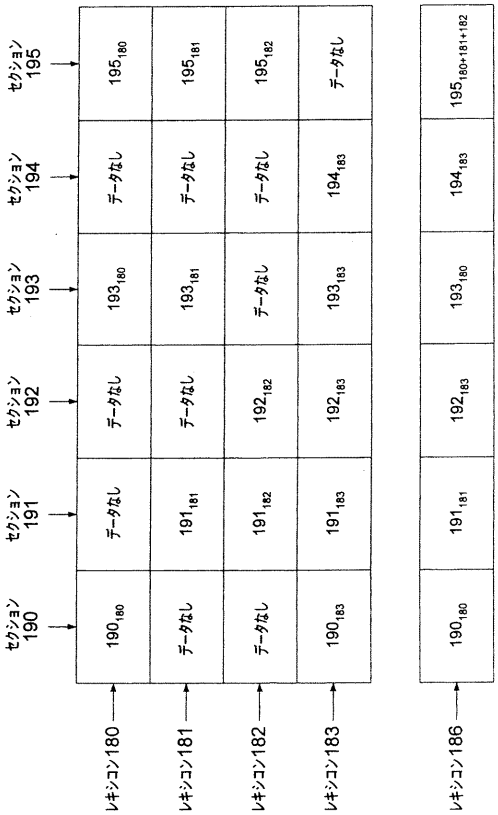
【図２】



【図３】


22
ヘッダ
160
単語リスト
162
索引テーブル
164
索引テーブル
166
168a
168b
168m
168n
168o
168p
168
文字列ヒープ
170

【図４】


セクション 195
セクション 194
セクション 193
セクション 192
セクション 191
セクション 190
データなし
レキシコン180
レキシコン181
レキシコン182
レキシコン183
レキシコン186

フロントページの続き

(72)発明者 ジェームズ ピー. フィニガン
アメリカ合衆国 ９８００７ ワシントン州 ベルビュー サウスイースト ６ １４３５１ アパートメント ０２０１
(72)発明者 カーティス イー. ハッテンハウアー
アメリカ合衆国 ９８００７ ワシントン州 ベルビュー ザ レイクス ノースイースト ４２ プレイス １４４４２ アパートメント ７０９
(72)発明者 ダグラス ダブリュ. ポッター
アメリカ合衆国 ９８１３３ ワシントン州 シアトル ノース １２８ ストリート ２１５５
(72)発明者 ケビン アール. パウエル
アメリカ合衆国 ９８０３４ ワシントン州 カークランド ノースイースト １３７ プレイス １３１０４

合議体
審判長 長島 孝志
審判官 久保 正典
審判官 飯田 清司

(56)参考文献 特開平７－１５２７５６（ＪＰ，Ａ）
特開平５－２０４９６２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl., ＤＢ名)
G06F17/28