(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5394245号
(P5394245)
(45)発行日　平成26年1月22日(2014.1.22)　(24)登録日　平成25年10月25日(2013.10.25)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 350C
G06F 17/30 170B

請求項の数 16　(全 14 頁)

(21)出願番号　特願2009-533936(P2009-533936)
(86)(22)出願日　平成19年10月23日(2007.10.23)
(65)公表番号　特表2010-507856(P2010-507856A)
(43)公表日　平成22年3月11日(2010.3.11)
(86)国際出願番号　PCT/GB2007/004035
(87)国際公開番号　WO2008/050107
(87)国際公開日　平成20年5月2日(2008.5.2)
審査請求日　平成22年10月25日(2010.10.25)
(31)優先権主張番号　11/585,358
(32)優先日　平成18年10月23日(2006.10.23)
(33)優先権主張国　米国(US)

(73)特許権者　509116772
モンロ、ドナルド・マーティン
イギリス国、ビーエー１１・６アールエス、サマーセット、ベッキントン、グース・ストリート、ザ・レイズ　６
(74)代理人　100108855
弁理士　蔵田　昌俊
(74)代理人　100091351
弁理士　河野　哲
(74)代理人　100088683
弁理士　中村　誠
(74)代理人　100109830
弁理士　福原　淑弘
(74)代理人　100075672
弁理士　峰　隆司

最終頁に続く

(54)【発明の名称】ファジーデータベースマッチング

(57)【特許請求の範囲】
【請求項１】
サンプルレコードと、データベースの複数の記憶されているレコードとの間で可能性ある一致を識別する方法において、
前記記憶されているレコードのそれぞれから複数のインデックスコードを抽出し、前記インデックスコードは、インデックスコード空間内に含まれていることと、
サンプルレコードからサンプルインデックスコードを抽出することと
を含み、
前記方法は、
前記インデックスコード空間を規定するルックアップテーブルを維持し、前記ルックアップテーブルは、複数の行を持ち、それぞれの行は前記インデックスコード空間内の一意的なインデックスコードに対応していることと、
複数のレコード出現リストを維持し、前記レコード出現リストのそれぞれは、特定のインデックスコードに対応している前記ルックアップテーブル中の特定の行からリンクされており、前記レコード出現リストのそれぞれは、そこから前記特定のインデックスコードと、前記インデックスコード空間内で前記特定のインデックスコードに対する規定された近接性内にあるインデックスコードとが抽出された、記憶されているレコードを識別するものであることと、
前記サンプルインデックスコードをインデックスとして使用して、ルックアップテーブルをアドレス指定し、前記サンプルインデックスコードに関係する、対応する複数のレ

コード出現リストをルックアップすることと、
インデックスコードが抽出されるときに、前記記憶されているレコードによる一致を記録するヒストグラムを構築することと、
前記ヒストグラムからの前記レコード出現リスト内で識別されたそれぞれの記憶されているレコードの出現回数をカウントすることと、
所定の記憶されているレコードに対する前記ヒストグラムからの前記カウントが、要求されるしきい値を超える場合、前記サンプルとの可能性ある一致であるとして、前記所定の記憶されているレコードを識別することと
により特徴付けられる方法。
【請求項２】
前記規定された近接性は、規定されたハミング距離である、請求項１記載の方法。
【請求項３】
前記規定されたハミング距離は、ユーザ選択可能である、請求項２記載の方法。
【請求項４】
前記要求されるしきい値は、数値的しきい値である、請求項１記載の方法。
【請求項５】
前記要求されるしきい値は、記憶されているレコード毎の、レコード出現リストの平均数の関数である、請求項１記載の方法。
【請求項６】
前記複数のインデックスコードは、前記複数の記憶されているレコードから抽出された、前記インデックスコード空間内のすべてのインデックスコードを規定する、請求項１記載の方法。
【請求項７】
前記複数のインデックスコードは、あるサンプルレコードによって表示することができる前記インデックスコード空間内のすべての可能性あるインデックスコードを規定する、請求項１記載の方法。
【請求項８】
前記複数のインデックスコードは、前記記憶されているレコードに対して、ハッシュのような演算を適用することによって生成される、請求項１記載の方法。
【請求項９】
複数の規定された近接性を確立することと、それぞれのインデックスコードと近接性との組み合わせに対して、独立したレコード出現リストを維持することとを含む、請求項１記載の方法。
【請求項１０】
前記識別ステップは、ユーザ選択された規定された近接性に関連するリストを使用する、請求項１記載の方法。
【請求項１１】
前記サンプルレコードと、前記可能性ある一致のそれぞれとの間の関係をさらに解析する追加のステップを含む、請求項１記載の方法。
【請求項１２】
前記識別ステップは、複数の並列プロセッサの間で分けられ、それぞれが関係付けの結果をコンソリデータに送り、前記コンソリデータは、前記関係付けの結果に依存して、可能性ある一致として、記憶されているレコードを識別する、請求項１記載の方法。
【請求項１３】
サンプルレコードと、データベースの複数の記憶されているレコードとの間の可能性ある一致を識別するシステムにおいて、
前記記憶されているレコードから抽出された複数のインデックスコードを含むデータベースに結合されたコンピュータプロセッサを具備し、
前記システムは、前記サンプルレコードからサンプルインデックスコードを抽出するように構成されており、

前記システムは、
前記インデックスコード空間を規定するルックアップテーブルと、
複数のレコード出現リストと
により特徴付けられ、
前記ルックアップテーブルは、複数の行を持ち、それぞれの行は前記インデックスコード空間内の一意的なインデックスコードに対応しており、
前記レコード出現リストのそれぞれは、特定のインデックスコードに対応している前記ルックアップテーブル中の特定の行からリンクされており、前記レコード出現リストのそれぞれは、そこから前記特定のインデックスコードと、前記インデックスコード空間内で前記特定のインデックスコードに対する規定された近接性内にあるインデックスコードとが抽出された、記憶されているレコードを識別するものであることと、
前記システムは、
前記サンプルインデックスコードをインデックスとして使用して、ルックアップテーブルをアドレス指定し、前記サンプルインデックスコードに関係する、対応する複数のレコード出現リストをルックアップするようにと、
インデックスコードが抽出されるときに、前記記憶されているレコードによる一致を記録するヒストグラムを構築するようにと、
前記ヒストグラムからの前記レコード出現リスト内で識別されたそれぞれの記憶されているレコードの出現回数をカウントするようにと、
（ｃ）所定の記憶されているレコードに対する前記ヒストグラムからの前記カウントが、しきい値を超える場合、前記サンプルレコードとの可能性ある一致であるとして、前記所定の記憶されているレコードを識別するプロセッサと
を具備するシステム。
【請求項１４】
前記コンピュータプロセッサは、サンプルレコードからサンプルインデックスコードを抽出する第１のプロセッサと、前記サンプルレコードとの可能性ある一致であるとして、所定の記憶されているレコードを識別する第２のプロセッサとを備え、請求項１３記載のシステム。
【請求項１５】
前記第１のプロセッサは、前記第２のプロセッサから離れている、請求項１４記載のシステム。
【請求項１６】
前記第１のプロセッサは、複数の並列プロセッサを備え、それぞれが関係付けの結果をコンソリデータに送り、前記コンソリデータは、前記関係付けられた結果に依存して、可能性ある一致として、記憶されているレコードを識別する、請求項１３記載のシステム。
【発明の詳細な説明】
【発明の分野】
【０００１】
本発明は、データベースシステムの分野に関連する。より詳細には、本発明は、候補レコードをデータベース内のレコードに対して高い信頼度でファジーに一致させる、速度を改善する方法およびシステムに関連する。
【先行技術】
【０００２】
さまざまな分野において、特定のサンプルレコードが、大規模データベース内に既に存在するか否かを、非常にすばやく決定できること、そして、もし存在する場合、１つ以上の一致を識別することに対する増加している要望がある。１つの特定の分野は、バイオメトリクスであり、バイオメトリクスでの要求は、特定のバイオメトリックサンプルを提供した個人が、既にデータベース中に存在するか否かを決定することである。
【０００３】
説明したタイプのデータベースは、きわめて大規模であってもよく、サンプルレコード

と、データベース内のレコードのあらゆるものとの間で、完全一致解析を試みることは非実用的であるかもしれない。計算作業負荷を減少させるために、さまざまな事前選別プロセスが使用されているが、これらは、マッチングアルゴリズムの特色、または、マッチングされることになるデータの特色に依存することが多いので、これらの多くのものは、非常に制限された適用分野しか持たない。

【０００４】

他の応用においても起こることであるが、バイオメトリックデータのマッチングで特に起こる問題は、バイオメトリック測定値が、それらの本質によって、正確に再現されないことが多いことである。例えば、ある特定の個人の虹彩から繰り返して導出される、バイオメトリック測定値は、とりわけ、まぶたとまつ毛による虹彩のオクルージョンが画像ごとに異なることになるので、多少異なっていることが多い。結果として、バイオメトリックマッチングは、通常、完全一致というよりは、近似、または、“ファジー”一致の概念に依拠する。

【０００５】

一般的なシナリオは、特定の個人が、個々人の大規模データベース内に存在するか否かを決定する要望である。例えば、我々は、ある個人の虹彩スキャンを持っていてもよく、国家安全保障データベースが、同一の個人の１つ以上の虹彩スキャンを既に含んでいるか否かについて知りたいかもしれない。サンプルの虹彩スキャンと、記憶されている虹彩スキャンとは、あらゆる観点で、同一でないことが多いので、必要とされる“ファジー”一致を達成する１つの方法は、ある領域にわたって検索することである。サンプルと記憶されているレコードとの両方をコードに変換した後、何らかの予め規定されたプロトコルにしたがって、我々は、ある記憶されたレコードと、我々がサンプルコードに十分近いとして考える、ある領域内の任意のコードとの間の一致を見つけることを試みることができる。代わりに、我々は、サンプルコードと、記憶されているコードのうちの１つに十分近い、ある検索領域内の任意のコードとの間の一致を見つけようとしてもよい。何れのケースでも、ファジー一致を実行するときに、コードの領域全体にわたって検索する必要性は、マッチングプロセスをかなり遅延させるかもしれない。

【発明の概要】

【０００６】

本発明の第１の観点にしたがうと、サンプルレコードと、複数の記憶されているレコードとの間で可能性ある一致を識別する方法が提供され、方法は、

（ａ）記憶されているレコードから複数の特性を抽出し、前記特性は、特性空間内に含まれることと、

（ｂ）それぞれの前記特性に対する、記憶されているレコードのレコード出現リストを維持し、記憶されているレコードのレコード出現リストから、前記特性と、前記特性空間内の前記特性に対する規定された近接内の特性とが抽出されていることと、

（ｃ）サンプルレコードから特性を抽出することと、

（ｄ）所定の記憶されているレコードが、要求される数のレコード出現リスト中に現れる場合、サンプルとの可能性ある一致であるとして、所定の記憶されているレコードを識別することとを含む。

【０００７】

本発明のさらなる観点にしたがうと、特性空間内の複数の特性を使用して、サンプルレコードと、複数の記憶されているレコードとの間の可能性ある一致を識別するシステムにおいて、システムは、

（ａ）それぞれの特性に対する、記憶されているレコードのレコード出現リストから、前記特性と、前記特性空間内の前記特性に対する規定された近接内の特性とが抽出されている、記憶されているレコードのレコード出現リストと、

（ｂ）サンプルレコードから特性を抽出するプロセッサと、

（ｃ）所定の記憶されているレコードが、要求される数のレコード出現リスト中に出現する場合、サンプルとの可能性ある一致であるとして、所定の記憶されているレコードを識

別するプロセッサとを具備する。

【０００８】

このような方法は、データベース内に新しいレコードを登録するときに任意の追加の労力をかけて、非常に高速な候補マッチングを提供する。新しいレコードの登録の頻度に比べて、マッチングがより頻繁に行われるとき、このトレードオフは、十分に価値がある。

【０００９】

いくつかの実施形態では、サンプルレコードに対する特性のマッチングのために、また、可能性ある一致として、記憶されているレコードを識別するために、独立したプロセッサを使用してもよい。これらのプロセッサは、独立したコンピュータであってもよく、互いに離れていてもよい。

【００１０】

ある特定の例では、記憶されているレコードの完全なコレクションを含むメインデータリストが、特性リストとは別のものとして保持されてもよい。このことは、ローカルプロセッサが、ローカルに取得された虹彩スキャンのようなサンプルレコード上に初期解析を実行することを可能にする。一度、可能性ある一致のリストが識別されると、次に、このリストは、遠隔サーバに送られてもよく、遠隔サーバにおいて、サンプルと、可能性ある一致のそれぞれの完全な、エンコードされた虹彩スキャンとを比較することによって、より詳細な解析が実行される。

【００１１】

このアプローチは、システムの設計者が、エンコードされた虹彩スキャンの全体のデータベースの完全なコピーを多数のユーザに配信する必要がないというさらなる利点を持っている。代わりに、それぞれのユーザは、特性のリストを単に受け取るだけであり、ローカルに実行されることになる初期解析に対しては、このような特性のリストで十分である。１つ以上の可能性ある一致が見つかると、システムが、集中ロケーションに自動的にレポートしてもよく、集中ロケーションにおいて、完全なレコードに対して、さらなる解析が実行されてもよい。

【００１２】

本発明は、さまざまな方法で実行されてもよく、いくつかの特定の実施形態を、添付の図面を参照して、例として、ここで説明する。

【図面の簡単な説明】

【００１３】

【図１】図１は、本発明の実施形態にしたがった、データベース構造を示す。

【図２】図２は、マッチングプロセスを例示するヒストグラムである。

【図３】図３は、別の例示的なヒストグラムである。

【図４】図４は、いくつかの例示的なハードウェアを示す。

【発明の詳細な説明】

【００１４】

以下の詳細な説明では、多数の特定の詳細を述べて、特許請求されている主題の完全な理解をもたらす。しかしながら、特許請求されている主題は、これらの特定の詳細なしで実行されてもよいことが当業者によって理解されるだろう。他の例では、よく知られた方法、手続、構成部品、および／または、回路は、詳細に説明しない。

【００１５】

以下の詳細な説明のいくつかの部分は、コンピュータ、および／または、コンピューティングシステムメモリ内のような、コンピューティングシステム内に記憶される、アルゴリズム、ならびに／あるいは、データビットおよび／またはバイナリデジタル信号上の動作の象徴的表現に関して提示した。これらのアルゴリズム的記述および／または表現は、データ処理技術における当業者によって使用され、彼らの作業の内容を他の当業者に伝える。アルゴリズムは、ここで、および一般的に、演算の筋の通ったシーケンス、および／または、望ましい結果へと導く類似の処理であるとして考えられる。動作および／または処理は、物理的な量の物理的な操作を含んでもよい。一般的に、必ずしも必要ではないが

、これらの量は、記憶され、転送され、結合され、比較され、および／またはそうでなければ、操作されることができる、電子的および／または磁気的信号の形態をとってもよい。主に共通の利用のために、時として、これらの信号をビット、データ、値、エレメント、シンボル、キャラクタ、ターム、数字、数値、および／または類似物として呼ぶことが便利であることが分かっている。しかしながら、これらのすべてのものと、類似のタームとは、適切な物理的量に関係付けられることになり、単に便利なラベルにすぎないことを理解すべきである。そうではないと特に明記しない限り、以下の説明から明らかなように、この明細書の説明全体にわたって、“処理”、“演算”、“計算”、“決定”、および／または、類似物のようなタームを利用することは、コンピュータまたは類似の電子的コンピューティングデバイスのような、コンピューティングプラットフォームの動作および／または処理を指し、コンピューティングプラットフォームは、コンピューティングプラットフォームのプロセッサ、メモリ、レジスタ、ならびに／あるいは、他の情報記憶装置、送信、および／または表示デバイス内の物理的電子的および／または磁気的な量、ならびに／あるいは、他の物理的量として表されるデータを操作し、および／または、変換する。

【００１６】

簡潔さのために、以下の説明は、バイオメトリック分野の例示的な実施形態に向けられることとする。説明することになる実施形態では、特定の個人から虹彩スキャンが取られ、同一の個人の虹彩スキャンが、国家安全保障データベースのような大規模なデータベース内に既に存在しているか否かを決定することが必要である。

【００１７】

もちろん、この特定の例は、本発明の背景にある、一般的な原則を説明するために、単に使用されており、同じ技法を他の分野中でも等しく適用できることになることが理解されるだろう。その最も広い形態における本発明は、データベース内に保持される、任意の特定のクラスまたはタイプのデータに限定されることはなく、使用されるマッチングアルゴリズムの詳細に限定されることはない。

【００１８】

例示的な実施形態のデータベース構造を、図１に概念的に示した。特定の個人の詳細が、ケースリストまたはテーブル１６内に保持されており、ケースリストまたはテーブル１６のそれぞれの行１７は、特定の個人の虹彩スキャンを表す。理想的には、それぞれの個人が、単一の虹彩スキャンによって表現されることになるが、もちろん、一般的な国家安全保障データベースでは、実際には、少なくとも数人の個人の複数のスキャンがあることになる。それぞれの行または虹彩スキャンレコードは、列１８、２０、２２を含み、これは、それぞれ、システム内で使用するための一意的な虹彩スキャン参照番号と、知られている場合は、個人の名前と、国家安全保障または社会安全保障コードのような外部識別子とをそれぞれ保持する。

【００１９】

それぞれのレコードに対する完全な虹彩スキャンが、独立したデータリストまたはテーブル１０内で保持され、独立したデータリストまたはテーブル１０のそれぞれの行１１は、個人のスキャンを表す。このテーブルは、２つの列からなり、最初の列１２は、前述のような一意的な参照番号であり、第２の列１４は、特定の個人の虹彩スキャン、または他のバイオメトリックレコードを一意的に識別する、何らかのエンコードされた表現を保持するとして考えられてもよい。

【００２０】

それぞれの登録されたケース（虹彩スキャン）は、複数の属性、特性、またはコードにしたがって分類され、これらは、未処理の虹彩スキャン、または、より一般的にはエンコードされたスキャンデータ１４のいずれかから、抽出または導出される。

【００２１】

コードは、スキャンの人的識別可能な特徴を表現していてもよいが、必ずしもそうでなくてもよい。例えば、コードのいくつかのものは、眼の色を表現していてもよく、他のも

のは、虹彩内部の色素の量と、輝度変化のような特性を表現していてもよい。代わりに、エンコードされたスキャン１４は、データストリームに何らかの関数が単に適用された単純な結果としてのコードを備える、純粋なデータストリームとして取り扱われてもよい。前述のハッシュ関数以外の、さらなる可能性は、データストリーム内の特定のグループのビットの存在または、不在を検索することである。いずれにせよ、それぞれの個人のレコード１１から、複数のコードが一般的に抽出されることが理解されるだろう。

【００２２】

これらのコードをインデックスとして使用するのを容易にするために、（以下でより詳細に説明することになるように、）コードは一般的に数値であるとして制約され、特定の予め規定された範囲内にあることが多い。バイオメトリクスの応用では、例えば、コードは、１６バイナリビットによって規定されていてもよく、６５５３６個の可能性あるコードが起こることを可能にしている。好ましくは、所望のコードだけが利用可能であるように、未処理の、または、エンコードされたデータからこれらのコードを生成させる関数または演算は、これらの可能性ある出力範囲において制限されていてもよい。代わりに、出力の実際の範囲が、所望の範囲内の数値コードのリストに再マッピングされてもよい。要求される場合には、（表示していない）マッピングテーブルを使用してもよい。説明している例では、利用可能なコードP_nは、０から６５５３５の範囲内の整数であり、それぞれが１６ビットコードとして記憶されていることが仮定されるだろう。したがって、フル１６ビットまで、Ｐ１＝００１、Ｐ２＝０１０、Ｐ３＝０１１、Ｐ４＝１００、Ｐ５＝１０１等である。

【００２３】

その中でそれらが現れるケース（虹彩スキャン）にしたがって、コードを分類するために、複数のリストまたはテーブル２８が、可能性ある６５５３６個のコードのそれぞれに対して１つ、維持される。簡潔さのために、これらのリストの５つのものだけを図１に示した。理解されるように、コード値１に対するリスト４０は、単一の行だけを含み、この行は、名前Ａだけがこのコードを生成したことを示す。コード値２を表現するリスト４１は、この例では、登録された虹彩スキャンのどれもそのコードを生成していないので、何のデータも含んでいない。リスト４２、４３は、それぞれコード値３と４を表現しており、単一のスキャンだけに関連する。表４４は、名前ＡとＢに対する虹彩スキャンがそれぞれコード値５を生成することを示す。

【００２４】

本質的ではないが、それぞれのテーブルまたはリスト２８が、それぞれの行２９中に、該当するコードに対応する単一のレコードに対する、一意的な参照１８を単に含むことが一般的に好ましい。

【００２５】

リスト２８に加えて、第２の系列のリスト３０が維持され、これらのリストは、単に個別のコードに関連しているのではなく、ハミング距離のような何らかの所望のメトリックにしたがった、対応しているベースコードからの所定の距離であるコードに関連する。

【００２６】

図１によって与えられる例を参照すると、コード１から５の間のハミング距離は、表１で与えられる。ハミング距離は、２つのコードの間で異なっているビットの数である。例えば、コード１と２との間のハミング距離は、１のためのコード（００１）と、２のためのコード（０１０）との間で２ビットが変更されているので、２である。

【表１】

コード	1 (001)	2 (010)	3 (011)	4 (100)	5 (101)
1	0	2	1	2	1
2	2	0	1	2	3
3	1	1	0	3	2
4	2	2	3	0	1
5	1	3	2	1	0

表1　数1から5に対するバイナリコードの間のハミング距離

【００２７】

示した例において、テーブル３０は、テーブル２８の対応するベースコードから、ちょうど１のハミング距離を持っているコードを解決するケースに関連するデータを含む。したがって、例えば、テーブル５１は、Ｐ１（０００１）からちょうどＨ＝１距離であるすべてのコードに対するデータを含む。しかしながら、Ｈ＝１のコードが、既にテーブル４０中で出現している場合、これらは効率のために、テーブル５１から省略される。図１の例では、コード１は、コード３および５から、Ｈ＝１距離である。３に対するベースリストからの名前Ｄが適格となり、５に対するベースリストからの名前ＡとＢが適格となる。しかしながら、名前Ａは、コード１に対するベースリスト中で既に発生しているので、したがって、名前ＤとＢだけが、テーブル５１中に含まれる。同様に、テーブル５２が、Ｐ２（００１０）からちょうどＨ＝１距離であるすべてのコードに対するデータを含む、等である。しかしながら、Ｈ＝１のコードが、既にテーブル４１中で出現している場合、これらは効率のために、テーブル５２から省略される。例において、名前Ｄが３に対するベースリストからテーブル５２に追加されるように、コード３は、コード２からＨ＝１距離である。テーブル４１が空であるので、テーブル５２から省略すべきである、コード２からのＨ＝１のコードは何もない。

【００２８】

第３の系列のテーブル３２は、対応するベースコードから、Ｈ＝２距離であるコードを解決するケースの詳細を含む。要求される場合、（示していない）Ｈ＝３、Ｈ＝４等に対する、さらなる系列のテーブルもまた提供されてもよい。

【００２９】

実施形態を図示するのにハミング距離を使用したことと、他の任意の便利なメトリックを使用してもよいこととが理解されるだろう。取り扱う特定の応用にしたがって、要求されるメトリック（例えば、ハミング距離）を選択してもよく、さらに、固定されていても、ユーザ選択可能であってもよい。（示していない）より洗練された実施形態では、コードは、対応する多次元空間内で測定されている要求されるメトリック内の、多次元のものであってもよい。

【００３０】

データベース内で、新しい虹彩スキャンが登録されることになるときはいつでも、その詳細が、ケースリスト１６と、データリスト１０とに追加され、新しいスキャンに対する対応するコードが計算され、および／または、決定される。次に、スキャンの一意的な参照番号１８が、個別のリスト２８、３０、３２に対して、適切なように追加される。望ましい場合、１つ以上の新しいコードがコードリスト２４に追加されてもよく、この中で、個別のテーブル２８が自動的に作成され、データベース内のそれぞれの虹彩スキャンをチ

ェックして、その参照番号を、新しく作成されたテーブルの１つ以上に対して追加する必要があるか否かを決定する。

【００３１】

ここで、マッチングのタスク、または、言い換えると、ある未知の虹彩スキャンがデータベース内のスキャン１４のうちの１つに一致するか否かを決定することに戻る。スキャンをエンコードされたデータに対してマッチングさせることは、計算的に冗長であるかもしれないので、スキャンをエンコードされたデータに対してマッチングさせるというよりはむしろ、代わりに、サンプルを進めて、それから１つ以上のコード値を抽出する。登録されているスキャンに元々適用されたのと同じ関数を適用することによって、１つ以上のサンプルコードが生成される（もちろん、この例におけるコードは、すべて、整数であり、範囲０から６５５３６内にある）。

【００３２】

どのスキャンが、それぞれのサンプルコードに対応するかを見つけるために、それぞれのコードｎを、ルックアップテーブル２５に対するインデックス７０として使用し、このテーブルは、コード値１、２、３を保持する、メモリ中のそれぞれのエリアをポイントするポインタＰ１、Ｐ２、Ｐ３・・・を含む。それぞれのリストが、メモリ中で互いに引き続いて起こる、特定の公称コード値に集める場合、単一のポインタ（プラスオフセット）だけが要求されるだろう。代わりに、独立したポインタが、系列２８、系列３０、および系列３２内でそれぞれのリストに対して提供されてもよい。別の可能性は、リスト２８のそれぞれが、３０中の対応するリストをルックするポインタを持つこと等であるだろう。

【００３３】

一度、適切なテーブルが識別されると、システムは、次に、特定のハミング距離のすべてのテーブルにわたって、それぞれのケースの出現回数のヒストグラムを構築することによって、一致候補を識別するために進む。図２は、サンプルスキャンが、コード値１と３を生成し、１までのハミング距離を使用して、一致候補が識別されることになる例を図示する。関連するＨ＝１テーブル５１、５３とともに、コード１と３とに対して、ベーステーブル４０、４２中のレコードを見ることによって、このことが達成される。ベースコードテーブルは、２つのヒット、すなわちＡとＤを生成し、一方、Ｈ＝１テーブルは３つの追加ヒットＡ、Ｂ、およびＤを生成する。

【００３４】

カウントに対してしきい値が適用され、少なくともしきい値にスコア付けされる任意のレコードが、一致候補であるとして考えられる。ここで、しきい値を１ととる場合、一致候補は、スキャンＡ、Ｂ、およびＤである。２のしきい値において、候補はＡとＤである。

【００３５】

図３は、コード１と３を生成する、同一のサンプルに対するヒストグラムを示すが、今回は、２までのハミング距離に対して試みた。ここで、ベーステーブル４０、４２からのヒットは、ＡとＤであり、Ｈ＝１テーブル５１、５３からの追加のヒットは、Ａ、Ｂ、およびＤであり、Ｈ＝２テーブル６１、６３からの追加のヒットは、ＢおよびＣである。１のしきい値を適用することは、一致候補として、Ａ、Ｂ、Ｃ、およびＤを与え、他方、より高いしきい値２を適用することは、Ａ、Ｂ、およびＤを候補として戻す。

【００３６】

図２、３中で、カウントをヒストグラムとして示したが、他のカウント方法も等しく使用されてもよく、いずれにせよ、実際のヒストグラムは一般的に作図されないだろう。

【００３７】

（示していない）代わりの実施形態では、第２の系列のテーブル３０は、対応するベースコードから、ちょうどＨ＝１距離であるコードだけからのデータのみならず、最大その距離までのすべてのコードからのデータを含む。このような配置では、それぞれのＨ＝１テーブルは、対応するベーステーブルのすべてのデータを含み、それぞれのＨ＝２テーブルは、対応するＨ＝１テーブルのすべてのデータを含む等であるだろう。

【００３８】
しきい値および／またはハミング距離に対して、適切な値を選択することによって、応用にしたがって、システムの出力応答を調整してもよい。これらの値のいずれか、または、両方は、固定であってもよく、プログラミング可能に可変であってもよく、ユーザ可変のものであってもよい。任意のアプリケーションでは、ランタイムにおいて、これらのパラメータのいずれかまたは両方の適切な値をユーザが選択できることが便利であるかもしれない。

【００３９】
いくつかの応用では、より複雑なマッチングアルゴリズムが構想されてもよい。例えば、異なるハミング距離に対して、異なるしきい値を使用してもよい。システムは、さまざまなハミング距離において、候補を自動的に選択してもよく、異なる距離において、それぞれの選択を比較または組み合わせて、候補リストの改善された複合リストを生成してもよい。

【００４０】
全体のマッチングプロセスを高速化するために、事前選択プロセスが、多数のケースを考察から除外する必要の度合にしたがって、必要である場合、しきい値および／またはハミング距離の選択が決定されてもよい。単純なカウントと、固定されたしきい値の使用は、一致を不一致から分ける便利な方法であるが、他のアルゴリズムを等しく使用することもできる。１つの可能性あるアプローチは、例えば、すべてのケースにわたって集められた平均特性カウントよりも、ある固定パーセンテージ分より高い特性カウントを持っている、すべてのケースを一致候補として選択することであるだろう。

【００４１】
評価されることになるサンプルのサイズに依拠して、そのサンプルを全体的に使用する必要はないかもしれない。例えば、サンプルが、ある書籍のいくつかの章からなる場合、１ページのテキストだけに基づいて、事前選択を実行することが十分であるかもしれない。

【００４２】
誤った棄却の危険性が受入可能なまでに低くなるように、ほとんどのアプリケーションにおいて、特性の選択、マッチング基準、および、解析されることになるサンプルのサイズが、選択されることになるだろう。

【００４３】
一度、一致候補のリストが選択されると、前述の手続の内の１つを使用して、何らかの便利なマッチングアルゴリズムを使用して、可能性あるもののそれぞれに対して、より詳細な一致を実行してもよい。説明したテキストの例では、サンプルスキャンが、データベース内の候補に対して、より洗練されているが、より遅いアルゴリズムを使用して、比較されてもよい。

【００４４】
１つの実施形態では、データベース自体は、事前の、および／または最終のマッチングが発生するのと、同じコンピュータ、または同じロケーションに保持されていてもよい。代わりに、事前のマッチングがローカルコンピュータにおいて保持されているコードリストにしたがって実行され、事前の一致がリモートコンピュータに送られて、詳細なマッチングが発生するというように、プロセスが分散されていてもよい。このような配置は、（すべての記憶されているスキャンを表す完全なデータを含む）一次データリスト１０が、中央ロケーションにおいて保持されることを可能にし、ここで、ローカルマシンは、個別のケース出現リスト２８、３０、３２だけを保持すればよい。

【００４５】
図４に示した、別の実施形態では、本発明のプロセスは、並列で動作する、複数のコンピュータまたはプロセッサを使用することによって、さらに高速化されている。ユーザのコンピュータ３２は、マッチングタスクを制御装置３４に送り、制御装置３４は、これを分けて、複数のコンピュータまたはプロセッサ３６の間で分散させる。それぞれのプロセ

ッサ３６は、特定のコード、または、グループのコードを取り扱うように命令されていてもよい。代わりに、制御装置３４は、何らかの別の方法で、作業を分けてもよい。プロセッサ３６は、それらの結果をコンソリデータ３８に送り、コンソリデータ３８は、（例えば、図２、３で図示した手続を使用して、）可能性ある一致の選択を最終的に行う。次に、可能性あるもののリストが、要求されるように、詳細なマッチングを実行するコンピュータまたはプロセッサ４２に送られ、あるいは、参照番号４０で示すように、さらなる解析のためにユーザ３２に戻される。

【００４６】

もちろん、特定の実施形態を説明してきたが、特許請求されている主題は、特定の実施形態、または、実現に限定されていないことが理解されるだろう。例えば、１つの実施形態は、デバイスまたはデバイスの組み合わせ上で動作するように実現されるような、ハードウェア中のものであってもよく、他方、別の実施形態は、例えば、ソフトウェア中のものであってもよい。同様に、ある実施形態は、例えば、ファームウェア中で、または、ハードウェア、ソフトウェア、および／または、ファームウェアの任意の組み合わせとして実現されてもよい。同様に、特許請求されている主題は、この観点における範囲に制限されていないが、１つの実施形態は、記憶媒体のような１つ以上の物品を含んでもよい。例えば、１つ以上のＣＤ－ＲＯＭ、および／または、ディスクのような、この記憶媒体は、命令を記憶してもよく、命令は、コンピュータシステム、コンピューティングプラットフォーム、または、他のシステムのようなシステムによって実行されるときに、例えば、前述した実施形態のうちの１つのような、特許請求されている主題にしたがった方法の実施形態が実行されることに帰結してもよい。１つの可能な例として、コンピューティングプラットフォームは、１つ以上の処理ユニット、または、プロセッサ、ディスプレイ、キーボードおよび／またはマウスのような１つ以上の入出力デバイス、ならびに／あるいは、スタティックランダムアクセスメモリ、ダイナミックランダムアクセスメモリ、フラッシュメモリ、および／またはハードドライブのような１つ以上のメモリを含んでもよい。

【００４７】

これまでの説明において、特許請求されている主題のさまざまな観点を説明した。特許請求されている主題の完全な理解をもたらすために、説明の目的で、特定の数、システム、および／または構成を述べた。しかしながら、特許請求されている主題は、これらの特定の詳細なしで実行されてもよいことが、この開示の利益を受ける当業者にとって明らかになるだろう。他の例では、特許請求されている主題をあいまいにしないように、よく知られた特徴を省略し、および／または簡潔化した。ある特徴をここで図解および／または説明したが、当業者にとって、さまざまな修正、置換、変更、および／または、均等物が、ここで明らかになるだろう。したがって、添付の特許請求の範囲は、このような修正および／または変更のすべてが、特許請求されている主題の本来の精神の範囲内におさまることを意図している。

以下に、本願出願の当初の特許請求の範囲に記載された発明を付記する。

［１］サンプルレコードと、複数の記憶されているレコードとの間で可能性ある一致を識別する方法において、

（ａ）前記記憶されているレコードから複数の特性を抽出し、前記特性は、特性空間内に含まれていることと、

（ｂ）それぞれの前記特性に対する、記憶されているレコードのレコード出現リストを維持し、記憶されているレコードのレコード出現リストから、前記特性と、前記特性空間内の前記特性に対する規定された近接性内の特性とが抽出されていることと、

（ｃ）サンプルレコードから特性を抽出することと、

（ｄ）所定の記憶されているレコードが、要求される数のレコード出現リスト中に現れる場合、前記サンプルとの可能性ある一致であるとして、前記所定の記憶されているレコードを識別することと

を含む方法。

［２］前記規定された近接性は、規定されたハミング距離である、上記［１］の方法。

［３］前記規定されたハミング距離は、ユーザ選択可能である、上記［２］の方法。
［４］前記要求される数は、数値的しきい値である、上記［１］の方法。
［５］前記要求される数は、記憶されているレコード毎の、レコード出現リストの平均数の関数である、上記［１］の方法。
［６］前記複数の特性は、前記複数の記憶されているレコードから抽出される前記特性空間内のすべての特性を規定する、上記［１］の方法。
［７］前記複数の特性は、あるサンプルレコードによって表示することができる前記特性空間内のすべての可能性ある特性を規定する、上記［１］の方法。
［８］前記複数の特性は、前記記憶されているレコードに対して、ハッシュのような演算を適用することによって生成される、上記［１］の方法。
［９］前記サンプルレコードに演算を適用して、１つ以上のサンプル出力を生成することと、前記サンプル出力を使用して、ルックアップテーブルをアドレス指定することとを含み、前記ルックアップテーブル中のそれぞれの行はレコード出現リストをポイントする、上記［１］の方法。
［１０］特性が抽出されるときに、記憶されているレコードによる一致を記録するヒストグラムが構築され、前記ヒストグラムから、可能性ある一致として、レコードを識別する、上記［１］の方法。
［１１］複数の規定された近接性を確立することと、それぞれの特性と近接性との組み合わせに対して、独立したレコード出現リストを維持することとを含む、上記［１］の方法。
［１２］前記識別ステップは、ユーザ選択された規定された近接性に関連するリストを使用する、上記［１１］の方法。
［１３］前記サンプルレコードと、前記可能性ある一致のそれぞれとの間の関係をさらに解析する追加のステップを含む、上記［１］の方法。
［１４］前記識別ステップは、複数の並列プロセッサの間で分けられ、それぞれが関係付けの結果をコンソリデータに送り、前記コンソリデータは、前記関係付けの結果に依存して、可能性ある一致として、記憶されているレコードを識別する、上記［１］の方法。
［１５］特性空間内の複数の特性を使用して、サンプルレコードと、複数の記憶されているレコードとの間の可能性ある一致を識別するシステムにおいて、
（ａ）それぞれの特性に対する、記憶されているレコードのレコード出現リストから、前記特性と、前記特性空間内の前記特性に対する規定された近接性内の特性とが抽出されている、前記記憶されているレコードのレコード出現リストと、
（ｂ）前記サンプルレコードから特性を抽出するプロセッサと、
（ｃ）所定の記憶されているレコードが、要求される数のレコード出現リスト中に出現する場合、前記サンプルとの可能性ある一致であるとして、前記所定の記憶されているレコードを識別するプロセッサと
を具備するシステム。
［１６］前記抽出するプロセッサと、前記識別するプロセッサとは、共通のプロセッサからなる、上記［１５］のシステム。
［１７］前記抽出するプロセッサは、前記識別するプロセッサから離れている、上記［１５］のシステム。
［１８］前記抽出するプロセッサは、複数の並列プロセッサを備え、それぞれが関係付けの結果をコンソリデータに送り、前記コンソリデータは、前記関係付けられた結果に依存して、可能性ある一致として、記憶されているレコードを識別する、上記［１５］のシステム。

【図１】

図１

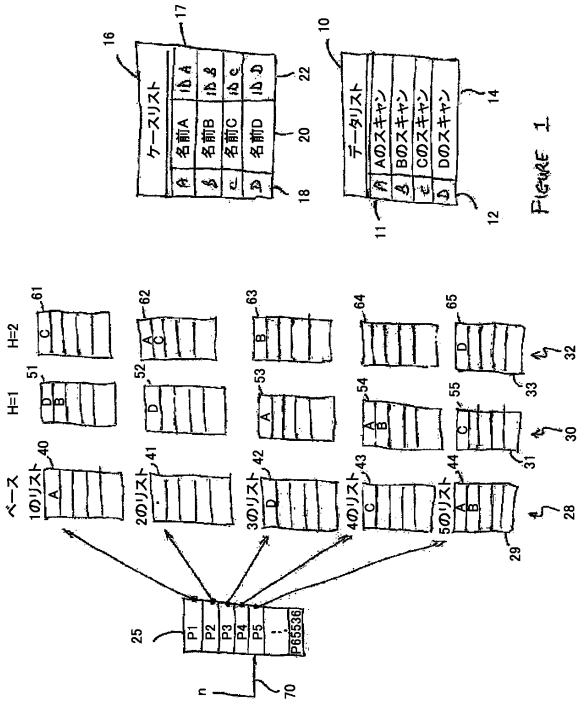

FIGURE 1

【図２】

図２

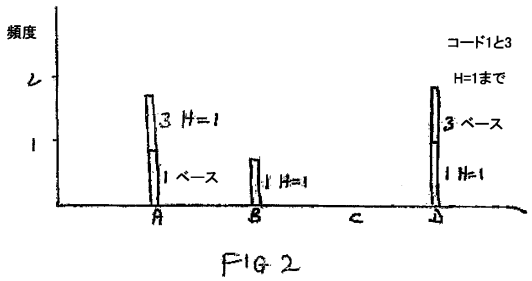

FIG 2

【図３】

図３

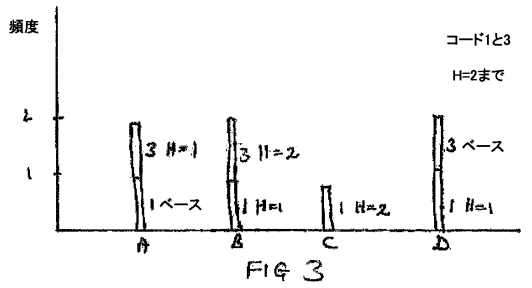

FIG 3

【図４】

図４

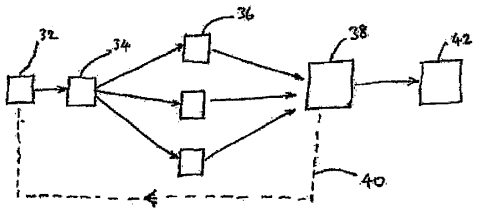

Figure 4

フロントページの続き

(74)代理人 100095441
弁理士 白根 俊郎
(74)代理人 100084618
弁理士 村松 貞男
(74)代理人 100103034
弁理士 野河 信久
(74)代理人 100140176
弁理士 砂川 克
(74)代理人 100100952
弁理士 風間 鉄也
(72)発明者 モンロ、ドナルド・マーティン
イギリス国、ビーエー１１・６アールエス、サマーセット、ベッキントン、グース・ストリート、
ザ・レイズ ６

審査官 佐藤 実

(56)参考文献 米国特許出願公開第２００３／００６１２３３（ＵＳ，Ａ１）
特表平０８－５０４９７９（ＪＰ，Ａ）
米国特許出願公開第２００６／０１０４４９３（ＵＳ，Ａ１）
米国特許出願公開第２００４／０２０２３５５（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０