(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2004-178604
(P2004-178604A)
(43)公開日　平成16年6月24日(2004.6.24)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	210D	5B075
	G06F 17/30	360Z	
	G06F 17/30	414A	

審査請求 未請求　請求項の数 32　OL　（全 26 頁）

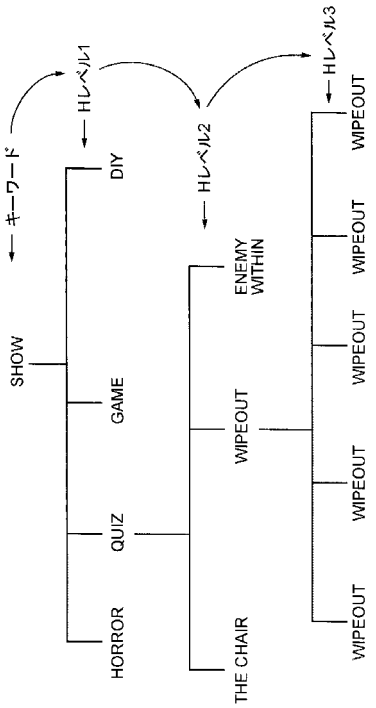
(21)出願番号　特願2003-398159(P2003-398159)
(22)出願日　平成15年11月27日(2003.11.27)
(31)優先権主張番号　0227683.0
(32)優先日　平成14年11月27日(2002.11.27)
(33)優先権主張国　英国(GB)

(71)出願人　593081408
ソニー・ユナイテッド・キングダム・リミテッド
Ｓｏｎｙ　Ｕｎｉｔｅｄ　Ｋｉｎｇｄｏｍ　Ｌｉｍｉｔｅｄ
イギリス国　サリー，ウェーブリッジ，ブルックランズ，ザ　ハイツ（番地なし）
(74)代理人　100067736
弁理士　小池　晃
(74)代理人　100086335
弁理士　田村　榮一
(74)代理人　100096677
弁理士　伊賀　誠司

最終頁に続く

(54)【発明の名称】情報検索装置及びその方法

(57)【要約】
【課題】　大容量階層化情報記憶装置及び検索装置並びにその方法を提供する。
【解決手段】　情報検索装置は検索で識別された情報項目集合から情報項目のマップを表すデータを受け取るマッピングプロセッサを備え、そのマップは情報項目の類似性に基づく配列内の位置関係で識別される情報項目を提供する。マップデータは類似の情報項目を配列内の類似の位置に配置する。マッピングプロセッサはマップデータを処理して情報項目を第１の階層分割レベルと１つ以上のその下位の階層にクラスタ分割する。その階層構造により情報間のナビゲーションとその表示が容易になる。各情報項目に特徴付ける情報特徴を与えることにより各階層間での識別が容易になる。
【選択図】　図１１



【特許請求の範囲】
【請求項１】
検索において識別された情報項目の集合から情報項目のマップを表すデータを受け取り、そのマップは情報項目の相互類似性に基づく配列内の位置関係でその識別された情報項目を提供し、類似する情報項目はその配列内で類似の位置にマッピングされ、
前記マップデータを処理して情報項目の階層クラスタ分割を形成し、その階層クラスタ分割は情報項目の第１のレベルのクラスタ分割と、その第１のレベルのクラスタ内の情報項目クラスタに対する少なくともう１つのクラスタ分割レベルを提供することを特徴とするマッピングプロセッサよりなる情報検索装置。
【請求項２】
前記情報項目は複数の特徴付ける情報特徴を含み、各情報項目のその特徴付ける情報特徴は各情報項目に対する特徴ベクトルを形成するのに使われ、その特徴ベクトルは配列内の位置にその情報項目をマッピングすることによりそのマップデータを形成するのに使われることを特徴とする請求項１に記載の情報検索装置。
【請求項３】
前記マッピングプロセッサはその第１のクラスタ分割レベルの情報項目にその第１のレベルのクラスタのそれぞれと関連する特徴付ける情報特徴を与え、その第１のレベルのクラスタ内の情報項目のクラスタに対する特徴付ける情報特徴を他の階層レベルにおいて提供することを特徴とする請求項１又は請求項２に記載の情報検索装置。
【請求項４】
各第１のレベルのクラスタと関連する特徴付ける情報特徴とその他の情報項目クラスタ分割レベル内の各クラスタと関連する特徴付ける情報特徴とは、それぞれのクラスタと関連する情報項目内に存在する最も共通的な特徴付ける情報特徴から生成されることを特徴とする請求項１乃至３のいずれか１項に記載の情報検索装置。
【請求項５】
ある下位レベルクラスタ内の情報項目のクラスタ同士は相互に関連するが、一方、第１のレベルのクラスタ同士はその下位レベルクラスタ内の情報項目に関する追加情報項目クラスタであることを特徴とする請求項１乃至４のいずれか１項に記載の情報検索装置。
【請求項６】
各クラスタと関連する特徴付ける情報項目は各クラスタ内の情報項目のそれぞれと関連するテキスト情報の最も共通的なワードであることを特徴とする請求項３乃至５のいずれか１項に記載の情報検索装置。
【請求項７】
前記情報項目はテキスト情報からなり、その特徴付ける情報特徴はワードであり、情報項目に対する特徴ベクトルはその情報項目内におけるワードグループそれぞれの、出現頻度の集合を表すことを特徴とする請求項１乃至６のいずれか１項に記載の情報検索装置。
【請求項８】
前記情報項目はテキスト情報を含み、その特徴付ける情報特徴はワードであり、配列内の位置がそのテキスト情報の少なくとも部分的な相互類似性によりマッピングされることを特徴とする請求項７に記載の情報検索装置。
【請求項９】
情報項目の集合中で出現頻度が閾値以上の頻度を有するテキスト情報内のワードを除外するマッピングに対する情報項目の前処理を行うことを特徴とする請求項７又は８に記載の情報検索装置。
【請求項１０】
情報項目の集合中で出現頻度が閾値以下の頻度を有するテキスト情報内のワードを除外するマッピングに対する情報項目の前処理を行うことを特徴とする請求項７乃至９のいずれか１項に記載の情報検索装置。
【請求項１１】
図形表示装置の表示領域内のｎ次元表示配列の表示点として、識別された情報項目に対

応する配列位置の少なくとも幾つかの表示を表示する、グラフィカルユーザインタフェース（ＧＵＩ）と組み合わせたディスプレイプロセッサを備えることを特徴とする請求項１乃至１０のいずれか１項に記載の情報検索装置。

【請求項１２】

前記表示領域は少なくとも２つの領域を含み、その１つの領域は第１の階層レベルのクラスタのｎ次元表示を提供し、その他の領域はその他の階層レベルのクラスタのｎ次元表示を提供し、ｎは整数であることを特徴とする請求項１１に記載の情報検索装置。

【請求項１３】

前記情報項目のワード検索を実行する検索プロセッサを備え、前記検索プロセッサとそのグラフィカルユーザインタフェースは識別された情報項目に対応する表示点だけを表示するように協調するよう構成されていることを特徴とする請求項１１又は１２に記載の情報検索装置。

【請求項１４】

前記ディスプレイプロセッサは、それがグラフィカルユーザインタフェース上に表示されたときに、階層レベルの１つにおける第１のクラスタを見ているユーザに、その階層レベル内の別のクラスタの位置のｎ次元空間の相対方向を提供する指示を表すデータを生成することを特徴とする請求項１１乃至１３のいずれか１項に記載の情報検索装置。

【請求項１５】

前記ディスプレイプロセッサは、その他のクラスタ内の情報項目の数を表すデータを生成し、前記情報項目の数はその第１のクラスタに関係するその他のクラスタのｎ次元空間内の相対方向の指示と関連することを特徴とする請求項１４に記載の情報検索装置。

【請求項１６】

前記ディスプレイプロセッサはグラフィカルユーザインタフェースと組合わさってその図形表示装置の第１の領域内のその他のクラスタの相対位置の指示を表示し、そのクラスタ内の情報項目の数を表すデータがその指示により表示可能であることを特徴とする請求項１４又は１５に記載の情報検索装置。

【請求項１７】

ユーザ制御ポインタを使ってそのｎ次元空間内の情報項目或いは情報項目のクラスタを選択するユーザ制御装置を備え、その指示上に置かれるそのポインタに応答して、相対方向の指示に関する情報項目の数が表示されることを特徴とする請求項１６に記載の情報検索装置。

【請求項１８】

次元数は２であることを特徴とする請求項１２乃至１７のいずれか１項に記載の情報検索装置。

【請求項１９】

前記情報項目はテキスト情報を有するビデオデータを含むことを特徴とする請求項１乃至１８のいずれか１項に記載の情報検索装置を備えるビデオ収集及び／又は処理装置。

【請求項２０】

情報項目を含む記憶装置と、

前記記憶装置を情報検索装置と接続するデータ通信ネットワークとを備えることを特徴とする請求項１９に記載のビデオ収集及び／又は処理装置。

【請求項２１】

前記情報項目はその情報項目からの代表画像を提供する代表キースタンプを含むことを特徴とする請求項１９に記載のビデオ収集及び／又は処理装置。

【請求項２２】

クラスタと関連する共通特徴付ける情報特徴はそのクラスタに共通する代表キースタンプを含むことを特徴とする請求項２１に記載のビデオ収集及び／又は処理装置。

【請求項２３】

検索で識別された情報項目の集合から情報項目のマップを表すデータを受け取るステップと、ここでそのマップはその情報項目の相互類似性に基づく配列内の位置に関してその

識別された情報項目を提供し、類似の情報項目はその配列内の類似の位置にマッピングされ、

第１のクラスタ分割レベルの情報項目及びその第１のレベルのクラスタ内の情報項目クラスタに対する少なくともう１つのクラスタ分割レベルの情報項目を提供する階層分割された情報項目を構成するようにそのマップデータを処理するステップとからなることを特徴とする情報検索及び表示方法。

【請求項２４】

前記マップデータ処理ステップは、第１のクラスタ分割レベルの情報項目に、その第１のレベルのクラスタの情報項目のそれぞれと関連する特徴付ける情報特徴を与えるステップと、その第１のレベルのクラスタ内の情報項目のクラスタに対する特徴付ける情報特徴をその他の階層レベルにおいて提供するステップとを含むことを特徴とする請求項２３に記載の情報検索及び表示方法。

【請求項２５】

上記情報項目は複数の特徴付ける情報特徴を含み、各情報項目のその特徴付ける情報特徴は各情報項目の特徴ベクトル形成に使用され、その特徴ベクトルはその情報項目を配列内の位置にマッピングに用いられることを特徴とする請求項２４に記載の情報検索及び表示方法。

【請求項２６】

図形表示装置上の表示領域内に２次元表示配列の表示点として少なくともその配列の幾つかの位置の表示を表示するステップを含むことを特徴とする請求項２３乃至２５のいずれか１項に記載の情報検索及び表示方法。

【請求項２７】

前記表示領域は少なくとも２つの領域を含み、その１つの領域は第１の階層レベルのクラスタのｎ次元表示を提供し、他方の領域はその他の階層レベルのクラスタのｎ次元表示を提供することを特徴とする請求項２６に記載の情報検索及び表示方法。

【請求項２８】

次元ｎの数は２であることを特徴とする請求項２７に記載の情報検索及び表示方法。

【請求項２９】

請求項２３乃至２７のいずれか１項に記載された情報検索及び表示方法を実行するためのプログラムコードを備える計算機ソフトウェア。

【請求項３０】

請求項２９に記載のプログラムコードを提供するための提供媒体。

【請求項３１】

前記媒体は記憶媒体であることを特徴とする請求項３０に記載の提供媒体。

【請求項３２】

前記媒体が伝送媒体であることを特徴とする請求項３０に記載の提供媒体。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、大量のコンテンツを扱う情報検索装置及びその方法に関する。

【背景技術】

【０００２】

キーワードに基づく検索によって情報（例えば、文書、画像、電子メール、特許、音声又は映像コンテンツなどのインターネットコンテンツ又はメディアコンテンツ）を探し出す定着した多数のシステムがある。その具体例として、「グーグル（Google）（登録商標）」や「ヤフー（Yahoo）（登録商標）」を始めとするインターネットサーチエンジンがあり、キーワードで実行される検索により、サーチエンジンによって認識され、関連度の順にランク付けされた結果の一覧が得られる。

【０００３】

しかし、大量コンテンツコレクションと称されることが多い大量のコンテンツを包含す

るシステムにおいては、効果的な検索クエリを策定し、比較的短い検索「ヒット（hits）」の一覧を得ることは困難である。例えば、本願作成時点に行った、キーワード「大量コンテンツコレクション（massive document collection）」に対するグーグルでの検索では、２４３０００件のヒットが引き出された。インターネットを通じて蓄積されるコンテンツ量は概して経時的に増大するので、検索がこの後繰り返される場合、このヒット数は増大することが予期される。そのようなヒット一覧を精査することには、非常に時間がかかる可能性が大である。

【０００４】

概して、大量コンテンツコレクションが良好に利用されない理由には、以下の点が問題である。

・利用者が関連するコンテンツの存在を知らない。
・利用者は関連するコンテンツの存在を知っているが、そのコンテンツが置かれている場所を知らない。
・利用者はコンテンツの存在を知っているが、それが関連性のあるものかどうかを知らない。
・利用者は関連コンテンツの存在及びその見出し方を知っているが、コンテンツを見つけるのに長い時間がかかる。

【０００５】

論文「大量文書コレクションの自己組織化（Self Organization of a Massive Document Collection）」、コホネン（Kohonen）他、ニューラルネットワークに関するＩＥＥＥトランザクション（IEEE Transactions on Neural Networks）、第１１巻、第３号、２０００年５月、第５７４～５８５頁には、所謂「自己組織化マップ（Ｓelf Organizing Map（ＳＯＭ））」を用いた技術が開示されている。これらの自己組織化マップは、各文書の特性を表す「特徴ベクトル（feature vectors）」がＳＯＭのノード上にマッピングされる、所謂非管理型自己学習ニューラルネットワークアルゴリズム（unsupervised self-learning neural network algorithm）を利用する。

【０００６】

コホネン等の論文において、第１のステップは、文書テキストを前処理し、次いで、各前処理された文書から特徴ベクトルが導かれることである。１つの形態において、これは、各単語の大辞書での出現頻度を示すヒストグラムであり得る。ヒストグラム中の各データ値（すなわち、各々の辞書単語の各出現頻度）は、ｎ値ベクトル中の値となるが、ここでｎは辞書中の候補単語の総数である（この論文において記載されている例においては４３２２２）。重み付けがｎ値ベクトルに与えられ得、それによって、ある幾つかの単語の増大した関連度又は改善された区別が強調されることになる可能性がある。

【０００７】

次いで、ｎ値ベクトルは、より大きさが小さいベクトル、すなわち、ｎよりも実質的に小さい数ｍ（この論文における例では５００）を有するベクトルにマッピングされる。マッピングは、乱数配列からなる（ｎ×ｍ）の「射影行列」でベクトルを乗算することによって達成される。この技術は、いずれか２つの縮小された大きさのベクトルが、２つの各々の入力ベクトルと同等のベクトル内積（dot product）を有する、より小さい大きさのベクトルを生じさせることが示されている。このベクトルマッピングプロセスは、論文「ランダムマッピングによる次元圧縮：クラスタリングのための高速類似性演算（Dimensionality Reduction by Random Mapping : Fast Similarity Computation for Clustering）」、カスキ（Kaski）、Ｐｒｏｃ．　ＩＪＣＮＮ、第４１３～４１８頁、１９９８年に記載されている。

【０００８】

次いで、次元が圧縮されたベクトルは、各ベクトルを「モデル（model）」（別のベクトル）で乗算するプロセスによってＳＯＭ上のノード（ニューロンとも称される）上にマッピングされる。モデルは、ＳＯＭ上への相互類似性によって自動的にモデルを配列する学習プロセスによって作成され、ＳＯＭはノードの二次元グリッドとして通常表される。

これは簡単な処理ではなく、コホネンらはこれに、７００万を丁度下回る数の文書の文書データベースのために、８００ＭＢのメモリを有する６プロセッサのコンピュータで６週間かかった。最後に、ユーザがマップの複数の領域にズームしてノードを選択できる状態にＳＯＭを形成するノードのグリッドが表示されるが、これによってユーザインタフェースがそのノードにリンクされた文書を含むインターネットのページへのリンクを提供する。

【発明の開示】

【発明が解決しようとする課題】

【０００９】

本発明は情報項目の大規模データベースから情報項目の検索結果を提供する実際的かつ管理可能な方法を見出す技術課題を取り扱う。

【課題を解決するための手段】

【００１０】

本発明の一態様によれば、検索中に確認された情報項目集合から情報項目のマップを表すデータを受け取り動作可能なマッピングプロセッサからなる情報検索装置が提供される。そのマップは識別された情報項目をその情報項目の相互類似性に応じた配列中における位置として提供する。マッピングデータは類似の情報項目はその配列中で類似の位置をマッピングするように構成される。そのマッピングプロセッサは、第１のクラスタ分割レベルの情報項目とその第１のレベルのクラスタ内の情報項目クラスタに対する少なくともも う１つのクラスタ分割レベルの情報項目とを提供する階層的クラスタ分割情報項目を形成するようにマップデータを処理できる。情報項目を階層構造のクラスタに分割形成することにより、情報項目のナビゲーションや表示が容易になる。

【００１１】

更にマッピングプロセッサは第１のクラスタ分割レベルの情報項目に対し第１のレベルのクラスタの情報項目のそれぞれと関連付けられる特徴付ける情報特徴を与えるようにしてもよい。マッピングプロセッサはそれに対応して第１のレベルのクラスタ内の情報項目クラスタに対する特徴付ける情報特徴を別の階層レベルで与えるようにしてもよい。特徴付ける情報特徴は１つのクラスタを別のクラスタから容易に区別する手段を提供する。

【００１２】

一般に、情報項目は複数の特徴付ける情報特徴を含んでおり、各情報項目の特徴付ける情報特徴は各情報項目に対する特徴ベクトルを形成するのに用いられ、その特徴ベクトルはその情報項目を配列内の１つの位置にマッピングするのに使われる。このように、各第１のレベルのクラスタと関連付けられた特徴付ける情報特徴とその他の情報項目レベル内の各クラスタと関連付けられた特徴付ける情報特徴とは各クラスタと関係する情報項目内に存在する最も共通的な特徴付ける情報特徴から形成される。

【００１３】

例えば、情報項目はテキスト情報であってもよく、特徴付ける情報特徴はワードであり、情報項目に対する特徴ベクトルはその情報項目のワードグループのそれぞれの出現頻度の集合を現している。この例に関しては、各クラスタと関連する特徴付ける情報特徴が各クラスタ内の情報項目のそれぞれと関連するテキスト情報の最も共通的なワードである。

【００１４】

情報検索装置は図形表示装置上の表示領域にｎ次元の表示配列の表示点として少なくともその配列の幾つかの点を表す表示を行うことのできるグラフィカルユーザインタフェースと組み合わせたディスプレイプロセッサからなってもよい。その表示領域は少なくとも２つの領域を含んでもよく、１つの領域は第１の階層レベルのクラスタのｎ次元表示領域を提供し、もう一方の領域はもう一方の階層レベルのクラスタのｎ次元表示領域を提供するものである。ｎ次元の次数は整数であればよく、限定するものではないがその次数は２であってもよい。もちろん１でも３でもよいことは当然理解されるべきである。

【発明の効果】

【００１５】

本発明の実施の形態によって提供される利点は、１つ以上の表示領域部分を備え、異なる階層レベルの情報項目を表示するのに便利な手段を提供する。例えば、第１のレベルの情報項目を一方の領域に表示しながら第１の領域から選択されたクラスタに現れる情報項目を第２の領域に表示することができる。したがって、検索によって疎母集団の配列であることが分かれば、第１の領域で明らかにされた異なるクラスタ間の相対的なナビゲーションがより簡単に管理できるようになり、より詳細な情報項目の表示が第２の領域内で選択表示されたクラスタで提供できる。

【発明を実施するための最良の形態】

【００１６】

添付図面を参照して、本発明の実施の形態を例としてのみ説明する。

【００１７】

図１は、プログラム及びデータ用のディスク記憶装置３０を備えたプロセッサユニット２０と、イーサネット（登録商標）ネットワーク又はインターネットなどのネットワーク５０に接続されたネットワークインタフェースカード４０と、陰極線管装置６０などの表示装置と、キーボード７０と、マウス８０などのユーザ入力装置とを有する汎用コンピュータ１０をベースとする情報記憶及び検索システムの概略図である。情報記憶及び検索システムはプログラム制御下で動作し、プログラムは、ディスク記憶装置３０上に記憶され、例えば、ネットワーク５０、着脱式ディスク（図示せず）又はディスク記憶装置３０上へのプリインストールによって与えられる。

【００１８】

この情報記憶及び検索システムは、２つの一般的な動作モードで動作する。第１のモードにおいては、一組の情報項目（例えば、テキスト情報項目）がディスク記憶装置３０又はネットワーク５０を介して接続されたネットワークディスクドライブ上で編集され、検索動作に備えて分類及び索引付けされる。第２の動作モードは、索引付け及び分類されたデータに対して実際に検索を行うことである。

【００１９】

実施の形態は多くの種類の情報項目に適用可能である。適切な種類の情報を全て網羅するものではないが、この一覧には、特許、映像素材、電子メール、プレゼンテーション、インターネットコンテンツ、放送コンテンツ、商用レポート、音声素材、グラフィック及びクリップアート、写真など、又はこれらのいずれもの組合せ又は合成を含む。この説明においては、テキスト情報項目に言及する。テキスト情報項目は、非テキスト項目と関連付けられても、又はリンクされてもよい。したがって、例えば、音声及び／又は映像素材は、テキスト用語においてその素材を定義するテキスト情報項目である「メタデータ（MetaData）」と関連付けられることが可能である。

【００２０】

情報項目は、従来の方法でディスク記憶装置３０にロードされる。好ましくは、これらの情報項目は、項目の検索及び索引付けをより容易にすることを可能にするデータベース構造の一部として記憶されるが、これは絶対的ではない。情報及び項目が一旦このように記憶されると、検索を行うためにこれらを配置するために用いられるプロセスは図２に概略的に示される。

【００２１】

索引付けされた情報項目は、ディスク記憶装置３０上に記憶される必要がないことが理解されるであろう。情報項目は、ネットワーク５０を介して情報記憶及び検索システム（汎用コンピュータ）１０に接続される外付けのリモートドライブ上に記憶されることが可能である。あるいは、情報は、例えば、インターネット中の様々なサイトに分散されて記憶されてもよい。情報が異なるインターネット又はネットワークサイトに記憶される場合、情報記憶の第２のレベルは、遠隔情報への「リンク（link）」（例えば、ユニバーサルリソースインジケータ：ＵＲＩ）をローカルに記憶するために用いられることが可能であり、そのリンクに関連付けられた関連した概要、要約又はメタデータを有する可能性がある。したがって、ユーザが関連リンクを選択しない（例えば、以下に説明する結果一覧領

域２６０から）限り、遠隔的に保持された情報はアクセスされないが、以下の技術的な説明のために、遠隔的に保持された情報又は要約／概要／メタデータあるいはリンク／ＵＲＩは、「情報項目（information item）」として考慮することが可能である。

【００２２】

言い換えれば、「情報項目（information item）」の形式的な定義は、特徴ベクトルが導かれ処理されて（以下を参照）、ＳＯＭへのマッピングを提供する項目である。結果一覧領域２６０（以下を参照）に示されるデータは、ユーザが検索する実際の情報項目（これがローカルに保持され、好都合な表示を行うのに十分短い場合）であっても、又は１つ又はそれ以上のメタデータ、ＵＲＩ、要約、一組のキーワード、代表的なキースタンプ画像などの情報項目を表現する及び／又は指示するデータであってもよい。これは、常にではないが、一組の項目を表現するデータの一覧表示を含むことが多い、動作「一覧（list）」に固有である。

【００２３】

別の例において、情報項目は、研究チーム又は法律事務所などのネットワーク化された作業グループを通じて記憶されることが可能である。複合的な手法は、ローカルに記憶された幾つかの情報項目及び／又はローカルエリアネットワークに亘って記憶された幾つかの情報項目及び／又は広域ネットワークに亘って記憶された幾つかの情報項目を包含し得る。この場合、情報検索及び検索システムは、例えば、大規模な多国間研究開発組織における、他人による同様の作業の位置指定において有用であることが可能であり、同様な研究作業は、ＳＯＭ（以下を参照）中の同様な出力ノードにマッピングされる傾向にある。あるいは、新しいテレビ番組が計画中である場合、この技術は、同様の内容を有する以前のプログラムを検出することによってその独自性をチェックするためにも用いられることが可能である。

【００２４】

図１の情報記憶及び検索システム（汎用コンピュータ）１０は、索引付けされた情報項目を有することが可能なシステムの一例でしかないことも理解されるであろう。初期（索引付け）段階は、適度に強力なコンピュータ、最も可能性が高くは、非ポータブルコンピュータによって実行され、情報へアクセスするというその後の段階は、「パーソナルディジタルアシスタント（personal digital assistant：ＰＤＡ）（概して片手に入る、表示装置及びユーザ入力装置を有するデータ処理装置）」などのポータブルマシン、ラップトップコンピュータなどのポータブルコンピュータ、あるいは更には携帯電話、ビデオ編集装置又はビデオカメラなどの装置で実行されることが可能である。概して、実際には、ディスプレイを有するいずれもの装置が、動作の情報アクセス段階のために使用可能である。

【００２５】

プロセスは、特定数の情報項目に限定されない。

【００２６】

情報項目の自己組織化マップ（ＳＯＭ）表現を生成させるプロセスを、図２～図６を参照して説明する。図２は、所謂「特徴抽出（feature extraction）」プロセスに次いでＳＯＭマッピングプロセスを図示する概略的なフローチャートである。

【００２７】

特徴抽出は、生データを抽象表現に変換するプロセスである。次いで、これらの抽象表現は、パターン分類、クラスタリング及び認識などのプロセスに用いられる。このプロセスにおいて、所謂「特徴ベクトル（feature vector）」が生成されるが、これは、文書内で用いられる用語の頻度を表す抽象表現である。

【００２８】

特徴ベクトルの作成による視覚化形成プロセスは、以下を含む。

・用語の「文書データベース辞書（document database dictionary）」の作成

・「文書データベース辞書」に基づく各個々の文書についての「用語頻度ヒストグラム（term frequency histogram）」作成

・ランダムマッピングを用いた「用語頻度ヒストグラム（term frequency histogram）」の縮小
・情報空間の二次元視覚化の作成
これらのステップをより詳細に検討すると、各文書（情報項目）１００が順に開かれる。ステップ１１０で、全ての「ストップワード（stop word）」が文書から除去される。ストップワードとは、「a」、「the」、「however」、「about」、「and」及び「the」などの、前もって作成された一覧にある非常に一般的な単語である。これらの単語は非常に一般的であるので、これらは、概して、十分な長さの全ての文書において同様の頻度で出現する傾向にある。このため、これらの単語は特定の文書の内容を特徴付ける試みにおいてほとんど効果がなく、したがって、除去されるべきである。

【００２９】
ストップワードの除去後、ステップ１２０で残りの単語の語幹分析がされるが、これは単語の変形の共通語幹を見出すことである。例えば、「thrower」、「throws」及び「throwing」は、共通語幹「throw」を有する。

【００３０】
文書中に出現する語幹分析された単語（「ストップ（stop）」ワードを除く）の「辞書（dictionary）」が維持される。新たな単語に遭遇すると、この単語は辞書に加えられ、文書コレクション全体（情報項目の集合）においてその単語が出現した回数の実行カウントも記録される。

【００３１】
結果として、集合内の中の全ての文書において用いられる用語をそれらの用語が現れる頻度と共に示した一覧が得られる。余りにも高い又は低い頻度で現れる単語は度外視され、これはすなわち、これらの単語が辞書から除去され、続いて行われる分析には加わらないということである。余りにも低い頻度で現れる単語は綴り間違いであるか、造語であるか、あるいは文書の集合によって表される分野に関連しないかである可能性がある。余りにも高い頻度で現れる単語は、集合の中の文書を区別するためには余り適切ではない。例えば、用語「News」は、放送に関連する文書の試験集合中の総文書の約３分の１の率で用いられるが、用語「football」は、その試験集合中の文書の約２％でしか用いられない。したがって、「football」は「News」よりも文書内容を特徴付けるためにより良い用語であると仮定することができる。逆に、単語「fottball」（「football」の綴り間違い）は、文書の集合全体において一度しか現れず、したがって、出現が余りにも少ないために度外視される。このような単語は、平均出現頻度から２標準偏差を引いた（－２σ）値よりも低い、又は平均出現頻度に２標準偏差を足した（＋２σ）値よりも高い出現頻度を有する単語として定義されることが可能である。

【００３２】
次いで、特徴ベクトルがステップ１３０で生成される。

【００３３】
これを行うために、集合中の各文書について用語頻度ヒストグラムが作成される。用語頻度ヒストグラムは、辞書（その文書の集合に属したもの）に存在する単語が個々の文書内で出現する回数をカウントすることによって構成される。辞書中の用語の大半が１つの文書中に存在することはないために、これらの用語は頻度ゼロを有する。２つの異なる文書についての用語頻度ヒストグラムの概略的な例を、図３ａ及び図３ｂに示す。

【００３４】
この例から、ヒストグラムが文書内容をどのように特徴付けるかがわかる。これらの例を検討することによって、文書１では文書２よりも用語「MPEG」及び「Video」の出現回数が多く、文書２自体は用語「MetaData」の出現がより多い。対応する単語が文書中に存在しないので、ヒストグラム中の見出し項目の多くはゼロである。

【００３５】
現実の例においては、実際の用語頻度ヒストグラムは、例におけるよりも大幅に多い数の用語を有する。代表的には、ヒストグラムは５００００を超える異なる用語の頻度をプ

ロットし得、５００００を超える大きさをヒストグラムに与える。このヒストグラムの大きさは、ＳＯＭ情報空間の構成に用いられる場合には、大幅に縮小される必要がある。

【００３６】

用語頻度ヒストグラム中の各見出し項目は、その文書を表す特徴ベクトル中の対応する値として用いられる。このプロセスの結果として、文書コレクション中の各文書についての辞書によって特定される全ての用語の頻度を含む（５００００×１）ベクトルが得られる。値の大半は代表的にはゼロであり、その他の値の大半が代表的には１などの非常に小さい数であるために、ベクトルは「スパース（sparse）」と称され得る。

【００３７】

特徴ベクトルのサイズ、したがって、用語頻度ヒストグラムの大きさは、ステップ１４０で縮小される。ヒストグラムの大きさを縮小するプロセスには、２つの方法が提案される。

【００３８】

ｉ）ランダムマッピング：ヒストグラムが乱数行列によって乗算される技術である。これは、計算上安価なプロセスである。

【００３９】

ｉｉ）潜在意味的索引付け：文書内に同時に出現する可能性が高い用語のグループを探すことによって、ヒストグラムの大きさを縮小する技術である。次いで、これらの単語グループは、単一のパラメータに縮小されることが可能である。これは、計算上高価なプロセスである。

【００４０】

本実施の態様における用語頻度ヒストグラムの大きさを縮小するために選択された方法は、上記で参照したカスキ（Kaski）の論文において詳細に説明されているような、「ランダムマッピング（random mapping）」である。ランダムマッピングは、乱数行列でヒストグラムを乗算することによって、ヒストグラムの大きさの縮小を達成する。

【００４１】

上述のように、「生（raw）」の特徴ベクトル（図４ａに概略的に図示）は、代表的には、５００００個の値の領域におけるサイズを有するスパースベクトルである。これは約２００のサイズ（図４ｂの概略図を参照）に縮小されることが可能であり、特徴ベクトルの相対的直交特性、すなわち、他の同様に処理された特徴ベクトルとの相対角度（ベクトル内積）などの関係を保持している。特定の直交ベクトル数は限られているが、略直交ベクトルの数は大幅に多いので、これは良好に働く。

【００４２】

実際に、ベクトルの大きさが増大するに従って、ランダムに生成されたベクトルの任意の集合は互いにほぼ直交する。この特性は、この乱数行列によって乗算されたベクトルの相対方向が保持されることを意味する。これは、それらの内積を調べることによりランダムマッピングの前後のベクトルの類似性を示すことによって表されることが可能である。

【００４３】

５００００個の値から２００個の値にスパースベクトルを縮小することによって、それらの相対的類似性が保持されることを経験的に示すことができる。しかし、このマッピングは完全なものではないが、文書の内容を簡潔に特徴付けるという目的のためには十分である。

【００４４】

特徴ベクトルが文書コレクションについて生成されて、コレクションの情報空間を規定すると、これらの特徴ベクトルはステップ１５０で二次元ＳＯＭに投影されて、意味マップが作成される。以下の節では、コホネンの自己組織化マップを用いた特徴ベクトルのクラスタリングによる二次元へのマッピングのプロセスを説明する。説明するに当たり、図５も参照される。

【００４５】

コホネンの自己組織化マップは、各文書について生成された特徴ベクトルをクラスタリ

ング及び組織化するために用いられる。

【００４６】

自己組織化マップは、入力ノード１７０と、二次元平面１８５として図示されるノードの二次元配列又はグリッド中の出力ノード１８０とからなる。マップを調整するために用いられる特徴ベクトル中に存在する値と同数の入力ノードが存在する。マップ上の各出力ノードは、重み付けされた結合１９０（各結合について１つの重み）によって入力ノードに結合されている。

【００４７】

初めに、これらの各重みが乱数に設定され、次いで、対話式プロセスによって重みが「調整（trained）」される。マップは、各特徴ベクトルをマップの入力ノードに与えることによって調整される。「最も近接した（closest）」出力ノードが、入力ベクトルと、各出力ノードに関連付けられた重みとの間のユークリッド距離を演算することによって算出される。

【００４８】

入力ベクトルと、そのノードに関連付けられた重みとの間の最小ユークリッド距離によって識別される最も近接したノードは「勝者（winner）」と称され、このノードの重みは、入力ベクトルに「近接して（closer）」移動するように、重みの値をわずかに変えることによって調整される。勝利ノードに加えて、勝利ノードの近隣にあるノードも調整され、入力ベクトルにわずかに近づいて移動させられる。

【００４９】

マップが一旦調整されると、ノードの二次元マップ内の入力空間のトポロジーの多くを保持することを可能にするのは、１つのノードの重みのみではなくマップ上のノード領域の重みも調整するこのプロセスである。

【００５０】

マップが一旦調整されると、各文書がマップに与えられて、その文書についての入力特徴ベクトルにどの出力ノードが最も近接しているかを見ることが可能になる。重みが特徴ベクトルと同一である可能性は低く、特徴ベクトルとマップ上のその最も近接したノードとの間のユークリッド距離はその「量子化誤差（quantisation error）」として知られている。

【００５１】

各文書についての特徴ベクトルをマップに与えて、それがどこに存在するかを見ることによって、各文書についてｘ及びｙマップ位置が生じる。これらのｘ及びｙ位置は、文書ＩＤと共にルックアップテーブルに入力されると、文書間の関係を視覚化するために用いられることができる。

【００５２】

最後に、ディザ成分がステップ１６０で付加されるが、これを以下で図６を参照して説明する。

【００５３】

上述したプロセスに起こる可能性のある問題は、２つの同一又は実質的に同一の情報項目が、ＳＯＭのノード配列中の同一ノードにマップされる可能性があることである。これによってデータの取扱いが困難になることはないが、これは表示画面（以下で説明する）上でのデータの視覚化を行う補助とはならない。特に、データが表示画面上で視覚化されると、複数の非常に類似した項目が特定のノードにある１つの項目に対して区別可能になるために有用であることがわかっている。したがって、各情報項目がマップされているノード位置に「ディザ（dither）」成分が付加される。ディザ成分は、ノード分離の±２分の１を無作為に付加することである。したがって、図６を参照すると、それについてマッピングプロセスが出力ノード２００を選択する情報項目は、実際には、図６において点線で境界付けられた領域２１０内のノード２００の周囲のいずれものマップ位置にマッピングされてもよいように付加されたディザ成分を有する。

【００５４】

したがって、情報項目は、ＳＯＭプロセスの「出力ノード（output node）」以外のノード位置で図６の平面上の位置へマッピングすると考えられ得る。

【００５５】

いずれもの時点で、上記で概説したステップ（すなわち、ステップ１１０から１４０）に従い、次いで、「前もって調整された（pre-trained）」ＳＯＭモデル、すなわち、マップの自己組織化作成の結果として生じるＳＯＭモデルの集合に、結果として得られる縮小された特徴ベクトルを適用することによって、新しい情報項目がＳＯＭに付加されることが可能になる。したがって、新たに付加された情報項目については、マップは概して「再調整（retrained）」されないが、その代わりに、全てのＳＯＭモデルが修正されていない状態でステップ１５０及び１６０が用いられる。新しい情報項目が付加される毎にＳＯＭを再調整するのは計算上高価であり、マップ中の共通してアクセスされる情報項目の相対位置に慣れていく可能性があるユーザに幾分使いにくいものでもある。

【００５６】

しかし、再調整プロセスが適切である時点も同様にあり得る。例えば、ＳＯＭが初めに作成されて以降、新しい用語（恐らくは、新しいものの新しい項目又は新しい技術分野）が辞書に入力される場合、それらの用語が出力ノードの既存の集合に特に良好にはマップしないこともあり得る。これは、新たに受け取られた情報項目の既存のＳＯＭへのマッピングの間に検出される、所謂「量子化誤差（quantisation error）」の増加として検出される可能性がある。本実施の形態においては、量子化誤差は閾値誤差量と比較される。量子化誤差が閾値誤差よりも大きい場合、（ａ）元の情報項目の全て及びＳＯＭが作成されてから付加されたいずれもの項目を用いて、ＳＯＭが自動的に再調整されるか、（ｂ）好都合な時間に再調整プロセスを開始するようにユーザが促されるか、のいずれかが行われる。再調整プロセスは、全ての関連する情報項目の特徴ベクトルを用い、ステップ１５０及び１６０全体を再適用する。

【００５７】

図７は、表示画面６０上の表示を概略的に図示する。表示は、検索クエリ領域２５０、結果一覧領域２６０及びＳＯＭ表示領域２７０を示す。

【００５８】

動作中は、始めはＳＯＭ表示領域２７０は空白である。ユーザは、キーワード検索クエリを検索クエリ領域２５０に入力する。次いで、ユーザは、例えば、キーボード７０上の復改キーを押すことによって、又はマウス８０を用いて表示画面の「ボタン（button）」を選択して検索を開始する。次いで、標準キーワード検索技術を用いて、検索クエリ領域２５０内のキーワードがデータベース中の情報項目と比較される。これによって結果一覧が生成され、この各々が結果一覧領域２６０中の各々の見出し項目２８０として示される。次いで、ＳＯＭ表示領域２７０は、各結果項目に対応する表示点を表示する。

【００５９】

ＳＯＭ表現を生成するために用いられる分類プロセスは、ＳＯＭ中の相互類似情報項目をグループ化する傾向にあるので、検索クエリの結果として、概して、クラスタ２９０などのクラスタが生じる傾向にある。ここで、ＳＯＭ表示領域２７０上の各点が、結果一覧領域２６０内の結果の１つと関連付けられたＳＯＭ内の各々の見出し項目に対応することと、ＳＯＭ表示領域２７０内で点が表示されている位置が、ノード配列内のこれらのノードの配列位置に対応することとが特筆される。

【００６０】

図８は、「ヒット（hits）」（結果一覧中の結果）数を低減させるための技術を概略的に図示する。ユーザはマウス８０を利用して境界線を引き、この境界線は、本例においては、ＳＯＭ表示領域２７０中に表示される一組の表示点を囲む矩形ボックス３００である。結果一覧領域２６０において、境界線３００内の点に対応する結果のみが表示される。これらの結果が対象となるものではないことがわかると、ユーザは、表示点の異なる集合を囲む別の境界線を引く。

【００６１】

結果一覧領域２６０は、境界線３００内に表示される表示点についての結果のための一覧見出し項目を表示し、これは単語の検索クエリ領域２５０中の検索基準を満たしたことが特筆される。境界線３００は、ノード配列中の母集団化されたノードに対応する他の表示位置を囲むことが可能であるが、これらが検索基準を満たさなかった場合、これらの表示位置は表示されず、したがって、結果一覧領域２６０内に示される結果のサブセットの一部を形成することはない。

【００６２】

図９は、本発明の１つの実施の形態を図示する。

【００６３】

図９を参照すると、ステップ９２０で自己組織化マップＳＯＭが作成されるとき、これはラベルを有さない（コホネンのＳＯＭとは異なる）。ユーザは、マップを探求するための誘導を与えるためにラベルを必要とする、本発明の実施の形態においては、ラベルは、ユーザの特定の必要性を満たすために自動的に作成される。ユーザは、図７及び／又は図８を参照して説明したように、検索の結果一覧を作成する。ラベルは、結果に従って自動的かつ動的に作成され、ＳＯＭ表示領域２７０内の表示点のクラスタをラベル付けするために用いられる。

【００６４】

クロスクラスタ関連付け／補助キーワード検索

本発明の実施の形態の例を図１０、図１１及び図１２を参照して説明する。

【００６５】

図１０において、情報項目のデータベースを含むデータ格納装置４００は、データ通信ネットワーク４１０によって検索プロセッサ４０４及びマッピングプロセッサ４１２に接続される。マッピングプロセッサ４１２は、ユーザ制御装置４１４及びディスプレイプロセッサ４１６に接続される。ディスプレイプロセッサ４１６の出力は、グラフィカルユーザインタフェース４１８に供給され、グラフィカルユーザインタフェース４１８はディスプレイ４２０に接続されている。ディスプレイプロセッサ４１６は、表示画面上で表示を行うために、マッピングプロセッサ４１２からのデータを処理するように動作可能である。

【００６６】

データ格納装置４００は、マッピングプロセッサ４１２とは別に配置されることが可能である。それに従って、検索プロセッサ４０４は、データ格納装置４００、マッピングプロセッサ４１２、並びにディスプレイプロセッサ４１６、グラフィカルユーザインタフェース４１８及びディスプレイ４２０である、情報を表示するために用いられる、図１０に示される構成要素とは別に配置されることが可能である。あるいは、マッピングプロセッサ４１２、検索プロセッサ４０４及びディスプレイプロセッサ４１６は、図１に示されるような汎用コンピュータ１０上で実行するために、ソフトウェアモジュールの形態で実施されてもよい。したがって、マッピングプロセッサ４１２、検索プロセッサ４０４及びディスプレイプロセッサ４１６は別々に製造及び配置されることが可能であることが理解されるであろう。

【００６７】

図１０に示される実施の形態は、図７、図８及び図９における図と組み合わせられた、図１に示されるような情報記憶及び検索システムと実質的に同様に動作する。図７、図８及び図９は、検索クエリに対してどのように情報項目が検索されるか及び検索結果がどのように表示されるかの図示例を提供する。したがって、図１０に示される実施の形態は、検索クエリ、例えば、ユーザ制御装置４１４からキーワードを受け取るように構成される。キーワードが受け取られると、検索プロセッサ４０４によって検索が実行されて、検索結果として識別される情報項目に対応する配列中のｘ及びｙ位置の組をマッピングプロセッサ４１２との組合せで識別する。例えば、ノードの４０×４０の配列については、正方形の二次元配列中に１６００個の位置が存在する。上記で説明したように、検索プロセッサ４０４は、検索クエリに従って情報項目を検索する。検索プロセッサ４０４による検索

によって、検索クエリに対応するものとして検索プロセッサ４０４によって識別された情報項目についてのｘ及びｙ位置の組が得られる。検索結果のｘ及びｙ位置は、マッピングプロセッサ４１２によって受け取られる。

【００６８】

マッピングプロセッサ４１２は、ｋ平均（ｋ-means）クラスタリングプロセスを行うことによって、第１の大域レベル（global level）での情報項目のクラスタを識別するように動作可能である。ｋ平均クラスタリングプロセスは、クラスタ及び配列内のクラスタの位置を識別する。ｋ平均クラスタリングプロセスは、クリストファー・エム・ビショップ（Christopher M. Bishop）による「パターン認識のためのニューラルネットワーク（Neural Networks for Pattern Recognition）」と題された書籍、第１８７～１８８頁、オックスフォード大学出版（Oxford University Press）に開示されている。ｋ平均クラスタリングアルゴリズムの更なる開示は、ウェブアドレスhttp://cne.gmu.edu/modules/dau/stat/clustgalgs/clust５bdy.htmlに開示されている。

【００６９】

図１１に図示されているように、キーワード「show」についての検索の結果によって、それらのメタデータの一部として単語「show」を有する情報項目に対応する配列中の位置が識別されることが可能である。したがって、配列にｋ平均クラスタリングアルゴリズムを行った結果、例えば、「quiz」「game」及び「DIY」である情報項目の３つのクラスタが識別される。情報項目のこれらのクラスタは、第１の階層レベルＨレベル１を形成する。ディスプレイプロセッサ４１６は、第１の階層レベルＨレベル１の情報項目のクラスタリングに対応するデータをマッピングプロセッサ４１２から受け取る。ディスプレイプロセッサ４１６は、この第１の階層レベルＨレベル１の二次元表示を表すデータを提供するように、データの第１の階層レベルを処理する。ディスプレイプロセッサ４１６によって生成されたデータは、図１２に示されるように、ディスプレイ４２０上の第１の表示領域４３０において表示を行うためにグラフィカルユーザインタフェース４１８に与えられる。

【００７０】

幾つかの実施の形態においては、ｋ平均クラスタリングアルゴリズムを用いてクラスタの識別を更に精密にするために、マッピングプロセッサ４１２によって更なる動作が行われることが可能である。更なる動作は、「ｋ平均クラスタリング及び剪定（k-means clustering and pruning）」と称される。公知のｋ平均クラスタリングプロセスは、類似した情報項目を示す検索結果において識別される情報項目について、配列位置のグループを識別する。次いで、結果項目のｘ及びｙ位置の隣接するサブクラスタが同一のメインクラスタの一部であるかを決定する更なる剪定プロセスが行われる。２つのサブクラスタの中心間の距離が閾値よりも小さい場合、これらの２つのサブクラスタは、同一のメインクラスタの一部であると考えられる。剪定は、クラスタが安定するまで、公知の方法で対話式に行われる。

【００７１】

マッピングプロセッサ４１２は、第１の階層レベルＨレベル１で識別された情報項目の各クラスタの更なる分析を行うように動作する。情報項目のクラスタを個々に検討し、かつ、それらの情報項目内で更なるクラスタを識別する機能をユーザに提供するために、マッピングプロセッサ４１２は更なる階層レベルを形成する。したがって、情報項目の各クラスタについて、情報項目のその第１の階層レベル内の更なるクラスタを識別するために、ｋ平均クラスタリングアルゴリズムがそのクラスタについて行われる。したがって、例えば、図１１に図示されるように、ｋ平均クラスタリングアルゴリズムが「quiz」クラスタに行われると、３つの更なるクラスタが第２の階層レベルＨレベル２で識別される。

【００７２】

第１の階層レベルについて図示されたように、各クラスタはキーワードに従ってラベル付けされる。キーワードは、クラスタ内の各情報項目が有する、その情報項目と関連付けられたメタデータ内に存在する最も共通する単語を見出すことによって識別される。した

がって、例えば、第１の階層レベルにおいて、単語「quiz」、「game」及び「DIY」によって３つのクラスタが識別される。

【００７３】

第１の階層レベルHレベル１のクラスタのラベル付けに対応した方法で、第２の階層レベルHレベル２における各クラスタについてキーワードが識別される。したがって、これらの３つのクラスタは、「the chair」「wipeout」及び「enemy within」とラベル付けされる。これらの３つのクラスタの各々が、quiz showの異なるエピソードを含む。

【００７４】

理解されるように、各クラスタの分析の更なる反復を行うことができる。これは、第２の階層レベルHレベル２で識別される各クラスタにｋ平均クラスタリングアルゴリズムを行うことによって達成される。図１１に図示されるように「wipeout」情報クラスタは、ｋ平均クラスタリングアルゴリズムを用いて更に分析される。しかし、第３の階層レベルHレベル３では、個別情報項目のみが明らかにされるために、図１１に図示されるように、第３の階層レベルHレベル３は、「wipeout」の個々のエピソードを識別する。

【００７５】

したがって、マッピングプロセッサ４１２は、異なる階層レベルで情報項目のクラスタを識別するように動作可能である。各階層レベルを表すデータが、ディスプレイプロセッサ４１６に与えられる。したがって、グラフィカルユーザインタフェース４１８と組み合わせられると、例えば、第２の階層レベルHレベル２に対応する可能性がある第２の領域がディスプレイ４２０上に表示されることが可能である。したがって、ズームコントロールを用いて、ユーザは第１の階層レベルHレベル１で表示されるクラスタにズームし得る。ズームコントロールは、ユーザ制御装置４１４を用いて動作させられることが可能である。したがって、特定のクラスタへズームすることで、情報項目の第２の階層レベルHレベル２を現す効果を有することができる。あるいは、第１の表示領域４３０内の「現在の目視」領域を選択するためにユーザ制御装置４１４を用いてもよい。したがって、第１の表示Hレベル１において示される第１の階層レベルで識別される「quiz」クラスタ内で識別されるクラスタに対して、第２の表示が行われる。

【００７６】

本発明の実施の形態によって提供される更なる利点は、第２の又はそれに続く領域において表示される第２の又はそれに続くレベルに、他のクラスタの標識が与えられ得る構成である。標識は、より低い階層レベルで目視されるクラスタと関連付けられたキーワードに対する代替的なクラスタにユーザを導く。したがって、第２の表示領域４４０内でより低い階層レベルで図示されているクラスタは、目視されているクラスタに対する代替的なクラスタを有する。例えば、図１２において、第１の表示領域４３０内で、第１の階層レベルは、「quiz」、「game」及び「DIY」の３つのクラスタを示す。ズームコントロールは「quiz」クラスタにズームするために用いられるので、第２の表示領域４４０は、「the chair」、「enemy within」及び「wipeout」である、「quiz」クラスタ内のクラスタの表示を与える。しかし、「quiz」クラスタに対する代替的なキーワードは、第１の表示領域４３０において図示されるように「DIY」、「horror」及び「game」である。したがって、矢印４４４、４４６及び４４８は、第２の表示領域４４０において表示されている「quiz」クラスタと同一の階層レベルにある情報項目のクラスタにユーザを導くために与えられる。したがって、次いでユーザが第１の階層レベルから異なるクラスタを閲覧して、第２の階層レベルにおけるクラスタを現すことを望む場合、ユーザは第１の階層レベル内の代替的なクラスタにナビゲートするために矢印を使用することができる。さらに、有利なことに、矢印は、第１の階層レベルで現れるクラスタについてのキーワードラベルでラベル付けされる。他の実施の形態において、クラスタ内の相対数の項目の図示をユーザに与えるために、この数は、方向を指示する矢印と関連付けられたキーワードと並んで示される。ユーザコンロトール及びディスプレイは、マウスポインタＭＰが指示矢印上を通過する、又はその上に位置付けられると、この数を指すように配置されることが可能である。

【００７７】
幾つかの実施の形態の更なる有利な特徴は、付加的なキーワードの一覧、すなわち、第１のレベルのクラスタ内の第２のレベルのクラスタと関連付けられたキーワードを提供することである。クラスタリングについて図１２において図示されるように、「horror」の更なる第１のレベルのクラスタを提供することによって、マウスポインタＭＰが「horror」と関連付けられた矢印上に位置付けられると、その第１のレベルのクラスタ「horror」内の第２のレベルのクラスタに対応する付加的な単語が生じる。その結果、ユーザには、第１のレベルのクラスタを第２の表示領域４４０内で目視する必要なく、これらのクラスタと関連付けられた情報項目の内容の非常に有効な図示が与えられる。図１２に図示されるように、表示領域は、第１の表示領域４３０内に出現する情報項目を検覧するため、及びそれらの周囲をナビゲートするための両方に用いられる、概して４５０で示されるコントロールアイコンを更に含むことが可能である。

【００７８】
マルチモード絞込み検索
本発明の別の実施の形態の例を、図１３～図１７と組み合わせて図１０を参照して説明する。図１３は、情報項目と関連付けられて記憶されている特徴付け情報特徴のタイプを図示したものを示す。例えば、情報項目は、テレビ番組からの音声／映像データの一部であることが可能である。本例においては、番組はサッカーの試合のハイライトを提供する。したがって、データ項目は、映像データ４６０及び音声データを含む。音声データと関連付けられているのは、ボックス４６２内に図示されている音声メタデータである。音声メタデータは、映像データと関連付けられた音声信号の内容及びタイプを示す。本例については、音声データは「音楽（music）」、「コメンタリ（commentary）」及び「群集の騒音（crowd noise）」を含むが、音声信号のタイプを示すメタデータの１つ又はそれ以上の他のタイプを含むことが可能である。映像データ及び音声データに加えて、情報項目は、映像及び音声データの内容又は属性を記載する他のメタデータも含むことが可能である。本例については、メタデータは、ボックス４６４内に図示されており、映像番組の内容の説明を含むことが示されている。ＳＯＭが作成される元となる特徴ベクトルを構築するために用いられるのは、このメタデータに含まれる単語である。しかし、本発明の他の実施の形態において、データ格納装置４００に含まれる情報項目の集合に、音声メタデータ４６２である音声データに対する、又は映像データに対する検索が行われることが可能である。この目的のために、映像データ４６０のフレームから代表キースタンプが生成されることが可能である。

【００７９】
代表キースタンプＲＫＳは、映像データの各フレームのカラーヒストグラムを形成することによって生成される。全ての又は選択された映像フレームについてのカラーヒストグラムは組み合わせられ、次いで正規化されて、図１３において棒グラフ４６６として代表的な形態で図示される、複合カラーヒストグラムが作成される。次いで、複合カラーヒストグラムは、各映像フレームについてのカラーヒストグラムと比較される。各映像フレームについての各列の複合ヒストグラムの対応する列に対する距離を加算することによって、各フレームについてのカラーヒストグラムと複合カラーヒストグラムとの距離が決定される。複合カラーヒストグラムに対して最小距離を有するカラーヒストグラムを有する代表キースタンプＲＫＳが選択される。次いで、したがって、サッカーの試合を表す番組については、作成された代表キースタンプは、サッカーの競技場の一部の映像画像である可能性が最も高く、これは図１３に示される代表キースタンプによって図示される。

【００８０】
他の実施の形態において、ＲＫＳは、以下の方法のいずれかによって、各情報項目について映像フレームから作成されることが可能である。
・ユーザは、情報項目の内容全体に対応する最も代表的なフレームであると考えられるフレームを選択することが可能である。情報項目を主観的に表す映像フレームが選択されることをユーザが確実にするので、この方法によって信頼性が改善され得る。しかし、この

方法にはより時間がかかる。
・ユーザは、情報項目内の第１のフレーム又は無作為のフレームを選択することが可能である。これは、適切なＲＫＳを選択するのには信頼性が低い方法である可能性がある。
・画像フレームの内容に基づいて映像フレームを処理し、ＲＫＳを選択する他の方法も考えられる。
【００８１】
本発明の実施の形態によって、選択された特徴付け情報特徴に基づいて絞込み検索を生じさせる機能が提供され得る。１つの実施の形態において、検索プロセッサ４０４は、メタデータの項目、映像画像又は音声データのいずれかと関連付けられた一回目の検索において識別されたこれらの情報項目を検索するように動作可能である。代替的な実施の形態においては、検索は、メタデータのみ、映像データのみ、又は音声データのみ、あるいはそれらのいずれもの組合せに対して行われることが可能である。検索クエリの形成を容易にするために、図１０に示されるディスプレイ４２０は、図１４に示されるグラフィカルユーザインタフェース４１８によって与えられる更なるグラフィカルディスプレイを含んでいてもよい。
【００８２】
図１４において、表示領域４７２内の第１の行４７０は、メタデータに基づいてクエリ情報を選択する機能をユーザに与える。したがって、情報項目からの画像代表キースタンプがこの行のウィンドウ内に配置される場合、この情報項目と関連付けられたメタデータ（図１３に図示されるように）が検索クエリに付加される。したがって、異なる情報項目からの１つ又はそれ以上の代表キースタンプが、タイプメタデータの特徴付け情報特徴についての検索クエリに導入されることが可能である。それに従って、第２の行４７４において、ユーザによって選択された映像フレームが導入され、検索クエリの一部が形成される。例えば、ユーザは映像データの特定の項目をブラウズし、対象となるフレームを選択することが可能である。次いで、ユーザは行４７４中にこの画像フレームを配置し、検索クエリの一部を形成させることが可能である。ユーザは、１つ又はそれ以上の映像フレームを導入することが可能である。
【００８３】
ユーザは、検索される情報項目を、その情報項目内の音声データに従って選択することも可能である。したがって、表示領域４７６内の第３の行は、その情報項目の代表画像を導入して、検索クエリが、検索クエリにおけるその情報項目に対応する音声データを含むものであることを音声データについての行内で識別する機能をユーザに与える。
【００８４】
特徴付け情報特徴のタイプに従って検索される情報項目を選択することに加えて、本発明の実施の形態は、選択された情報項目間でブール演算子に従って検索を行う機能も提供する。図１４に図示されるように、メタデータ検索について選択される情報項目は、初めの２列４７８及び４８０の間に示されるような「ＡＮＤ」演算子に従って検索されるべきである。しかし、検索クエリにおける第１のメタデータと第１の映像画像項目検索クエリとの間の検索クエリは、「ＯＲ」演算子によって結合される。映像画像データについて検索される２つの項目は、「ＡＮＤ」演算子によって結合される。音声データに従って検索される情報項目もまた、「ＮＯＴ」演算子に従って検索クエリにおいて検索されるものである。
【００８５】
検索クエリを構築した後、検索プロセッサ４０４は、ユーザによって行われた選択によって構築された、図１４に図示される検索クエリに従って、キーワード検索から識別された情報項目を検索するように動作可能である。検索プロセッサは、以下の節で説明されるように、選択された特徴付け情報特徴のタイプに依存して異なった方法で情報項目を検索する。
【００８６】
メタデータなどの特徴付け情報特徴についての検索の例については、いずれもの情報項

目についても、メタデータから生成されるその情報項目についての特徴ベクトルが、その特徴ベクトルに対応する二次元配列内の点を識別するために使用されることが可能である。したがって、配列内のその識別された位置の所定距離内にある情報項目は、検索クエリの結果として戻されることが可能である。しかし、１つを超える情報項目がメタデータ検索行内で選択された場合、選択されたブール演算子に従ってこれらの項目の両方を検索するように、検索クエリが構築されなければならない。

【００８７】

「ＡＮＤ」ブール演算子の例については、各情報項目についての特徴ベクトルが組み合わされて、図１５に図示されるような複合特徴ベクトルを形成する。この目的のために、メタデータ内の各単語と関連付けられた値が加算され、正規化されて複合特徴ベクトルが作成される。したがって、図１５に図示されるように、行４７０、列４７８～４８０並びにメタデータ検索クエリライン４７０で図示されるそれらの代表キースタンプを有する、ユーザが選択したメタデータと関連付けられた２つの特徴ベクトルＡと特徴ベクトルＢとが組み合わされて、特徴ベクトルＣが形成される。次いで、検索プロセッサは特徴ベクトルＣを取り上げ、これをＳＯＭと比較することが可能である。複合特徴ベクトルＣに対応する配列内の最も近い位置を識別した後、配列内のその識別された位置から配列内の所定数の位置内にある情報項目が検索クエリの結果として戻される。

【００８８】

対応するメタデータ検索のブール「ＯＲ」演算子の例については、第１の特徴ベクトルＡ及び第２の特徴ベクトルＢについて、これらの特徴ベクトルについての配列内の対応する位置が識別される。このように、検索クエリの結果として、配列内のこれらの識別された各点の所定数の位置内の全ての情報項目を戻すこととなる。これは図１６及び図１７に図示される。図１７において、二次元配列内の、特徴ベクトルＡに対応する位置及び特徴ベクトルＢに対応する位置が識別される。図１７に示されるように、特徴ベクトルＡ及びＢについての配列位置の所定半径内の配列内の位置は、次いで、検索クエリの結果として識別されたものとして戻されることが可能である。しかし、更なる特徴ベクトルＣが検索クエリで識別され、「ＮＯＴ」ブール演算子がこの更なる特徴ベクトルについて指定される場合、特徴ベクトルＣに対応する配列中の位置がここでも識別される。したがって、特徴ベクトルＣからの配列位置の所定半径内の情報項目がここでも識別されることが可能である。しかし、「ＮＯＴ」演算子の結果として、特徴ベクトルＣ並びに特徴ベクトルＡ及びＢについての配列位置からの半径間で識別されるいずれもの相互的に包括的な配列位置が検索結果から排除される。したがって、検索プロセッサは、特徴ベクトルＣからではなく特徴ベクトルＡ又はＢから作成された配列内の位置に対応する情報項目を戻すように構成される。

【００８９】

検索の特徴付け特徴である映像画像データに対応する検索クエリ中の第２列目について、検索プロセッサは、選択されたユーザ映像画像に対応する代表キースタンプについての映像データを検索するように動作可能である。この目的のために、ユーザが選択した映像画像と関連付けられたカラーヒストグラムは、情報項目と関連付けられた各代表キースタンプについてのカラーヒストグラムと比較される。各情報項目の代表キースタンプのカラーヒストグラムと、ユーザ指定の映像画像のカラーヒストグラムとの間の距離が算出される。これは、その画像の色成分を表す各列の間の距離を算出し、各列についてこれらの距離を合算することによって行われる。ユーザ選択映像画像のカラーヒストグラムと、その配列位置に対応する代表キースタンプのカラーヒストグラムとの間の距離が最小である情報項目に対応する配列位置が識別される。ここでもまた、クエリの結果として、識別された配列位置からの所定数の位置内の配列位置を有する情報項目が戻される。

【００９０】

ブール演算子の場合について、ここでもまた、ブール「ＡＮＤ」演算子について選択及び指定された２つの画像についてのカラーヒストグラムを組み合わせることによって、カラーヒストグラムが形成されることが可能である。複合カラーヒストグラムの形成プロセ

スは、図１８に図示される。図１４に図示される表示領域内の映像画像検索クエリ行の行４７４並びに列４７８及び４８０において与えられる第１及び第２のユーザ選択画像についてのカラーヒストグラムは、カラーヒストグラムの各列内の値を平均化することによって組み合わせられる。したがって、図１８ａ及び図１８ｂに図示される２つのカラーヒストグラムは組み合わせられて、図１８ｃにおいて形成されるカラーヒストグラムを形成する。検索される情報項目の代表キースタンプに対して検索されるのは、このカラーヒストグラムである。

【００９１】

音声データの例については、検索プロセッサは、選択された情報項目と関連付けられた音声メタデータから特徴ベクトルを形成することが可能である。例えば、音声メタデータは、音声信号中に存在する高調波、スピーチデータ、又は音声メタデータによって表される音声信号内に音楽が存在するかを識別することが可能である。さらに、メタデータは、トニーブレアーなどの特定の話し手又はジョンモトソンなどの特定の解説者が音声信号上に存在するかを識別することが可能である。したがって、ここでもまた、特に音声データと関連付けられる他の特徴ベクトルに対して検索されることが可能である選択された音声データから、特徴ベクトルが生成されることが可能である。上記の説明に対応した方法で、ブール演算子が、１つを超える音声メタデータタイプについての検索を組み合わせるために用いられることが可能である。「ＡＮＤ」演算子の例については、音声メタデータ項目が組み合わされて、複合メタデータ項目が作成されることが可能である。この複合項目に最も近い特徴ベクトルを有する対応する情報項目を検索することによって、情報項目が識別される。次いで、「ＯＲ」演算子が指定されると、検索プロセッサは、両方のメタデータ項目について配列内の所定数の位置の中にある情報項目を回復させることが可能である。ここでもまた、「ＮＯＴ」ブール演算子は、検索クエリの結果から、一致する音声データを有する戻された情報項目を排除する機能を有する。

【００９２】

識別された情報項目からの検索の絞込みについて、本発明の実施の形態が与えられた。しかし、他の実施の形態において、図１４において図示されるディスプレイによって形成される検索クエリ、並びにメタデータ、映像データ及び音声データに対するその検索クエリの用途は、データ格納装置４００内の情報の集合全体を検索するために与えられることが可能であることが理解されるであろう。

【００９３】

関連検索

本発明の実施の形態の一例に従って上記で説明したように、図１４に示されるグラフィカルユーザインタフェースを用いて構築された検索クエリによる情報項目は、検索クエリによって識別された特定の配列位置に近隣する項目を識別することによって検索されることが可能である。しかし、他の実施の形態例においては、どのような理由のためであっても、識別された情報項目から関連検索が行われることが可能である。しかし、代表的には、特定のキーワードによる検索によって、識別された情報項目の集合が得られる。これらの情報項目から、ユーザは、これらのうちの１つが特に対象となるものであることを決定することが可能である。次いで、関連検索によって、ＳＯＭによるこの情報と幾分かの相関を有する項目が与えられることが可能である。これは、例えば、対象とする情報項目に対応する配列位置から所定半径内にある、配列位置に対応する情報項目を識別することによって達成される。

【００９４】

本発明の範囲から逸脱することなく、上述の実施の形態に様々な改変を行うことが可能である。本発明の様々な態様及び特徴は、添付の請求項に定義される。

【図面の簡単な説明】

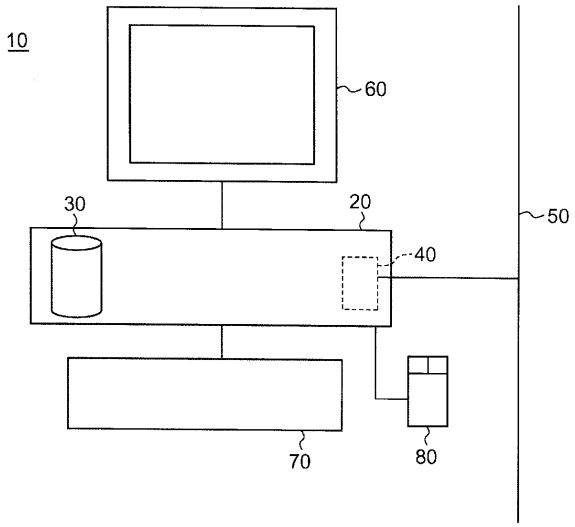
【００９５】

【図１】情報記憶及び検索システムを概略的に示す図である。

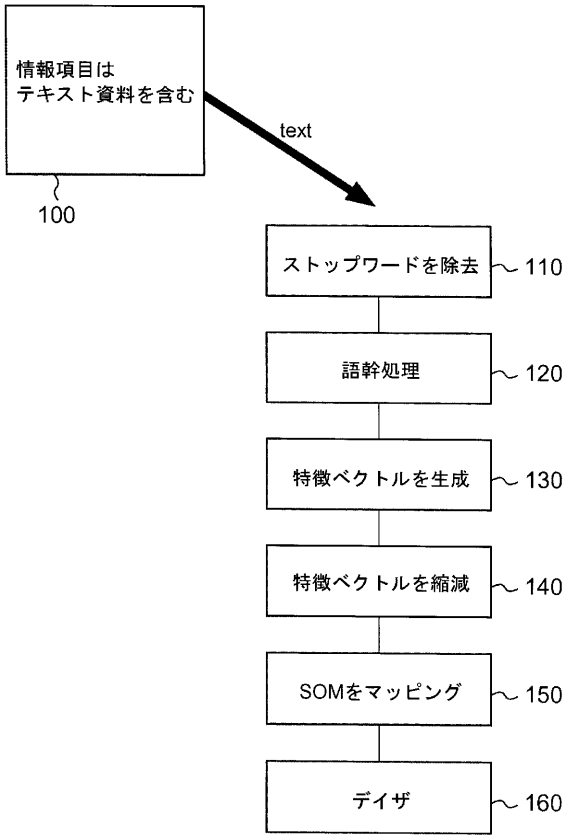
【図２】自己組織化マップ（ＳＯＭ）の生成を示す概略的なフローチャートである。

【図３】（ａ）及び（ｂ）は、用語頻度ヒストグラムを概略的に示す図である。
【図４】（ａ）は生の特徴ベクトルを概略的に示す図であり、（ｂ）は縮小された特徴ベクトルを概略的に示す図である。
【図５】ＳＯＭを概略的に示す図である。
【図６】ディザ処理を概略的に示す図である。
【図７】ＳＯＭによって表現される情報へのアクセスのためのユーザインタフェースを提供する表示画面を概略的に示す図である。
【図８】ＳＯＭによって表現される情報へのアクセスのためのユーザインタフェースを提供する表示画面を概略的に示す図である。
【図９】ＳＯＭによって表現される情報へのアクセスのためのユーザインタフェースを提供する表示画面を概略的に示す図である。
【図１０】本発明の実施の形態による情報検索装置の概略的なブロック図である。
【図１１】検索において識別された情報項目の階層配置を図示したものである。
【図１２】図１１において示される階層の異なるレベルを表示する２つの領域を提供する表示画面を概略的に示す図である。
【図１３】情報項目の例について３つのタイプの特徴付け情報特徴を図示したものである。
【図１４】本発明の実施の形態の例による検索クエリを形成するためのグラフィカルユーザインタフェースを概略的に示す図である。
【図１５】ブールＡＮＤ演算による複合特徴ベクトルの形成の概略図である。
【図１６】ブールＯＲ演算子による２つの特徴ベクトルと、ブールＮＯＴ演算子による第３の特徴ベクトルとの組合せを示す図である。
【図１７】図１６のブール演算子及び特徴ベクトルによる検索結果を示す識別された情報項目の二次元マップの一部を概略的に示す図である。
【図１８】（ａ）及び（ｂ）は、検索クエリを形成する２つの映像画像についてのカラーヒストグラムの２つの例を与える例示的な棒グラフであり、（ｃ）は、（ａ）及び（ｂ）のカラーヒストグラムを組み合わせることによって作成される例示的な棒グラフである。
【符号の説明】
【００９６】
１０　汎用コンピュータ、２０　プロセッサユニット、３０　ディスク記憶装置、４０　ネットワークインタフェースカード、５０　ネットワーク、６０　陰極線管表示装置、７０　キーボード、８０　マウス、４００　データ格納装置、４０４　検索プロセッサ、４１０　通信ネットワーク、４１２　マッピングプロセッサ、４１４　ユーザ制御装置、４１６　ディスプレイプロセッサ、４１８　グラフィカルユーザインタフェース（ＧＵＩ）、４２０　ディスプレイ

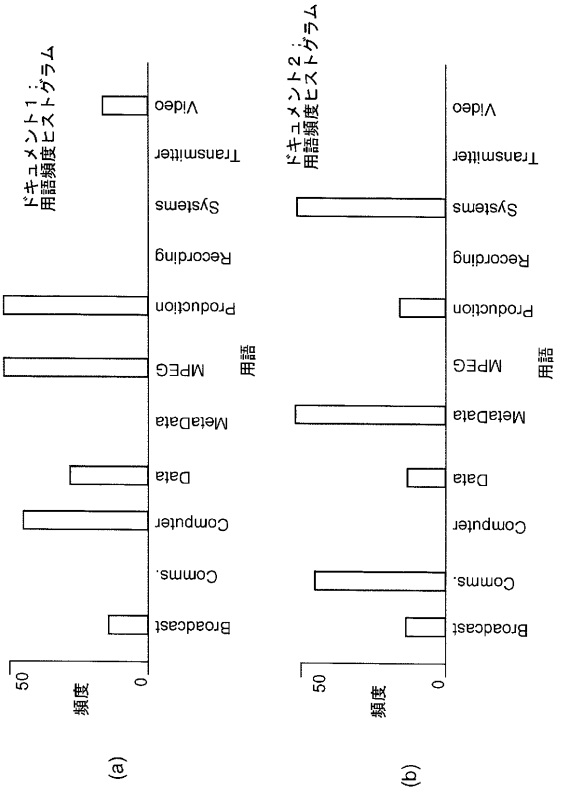
【図１】


10
60
30
20
50
40
70
80

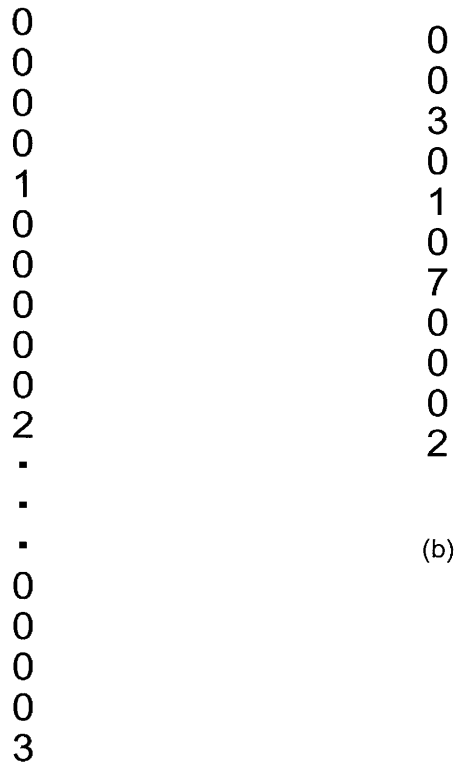
【図２】


情報項目は
テキスト資料を含む
text
100
ストップワードを除去
110
語幹処理
120
特徴ベクトルを生成
130
特徴ベクトルを縮減
140
SOMをマッピング
150
デイザ
160

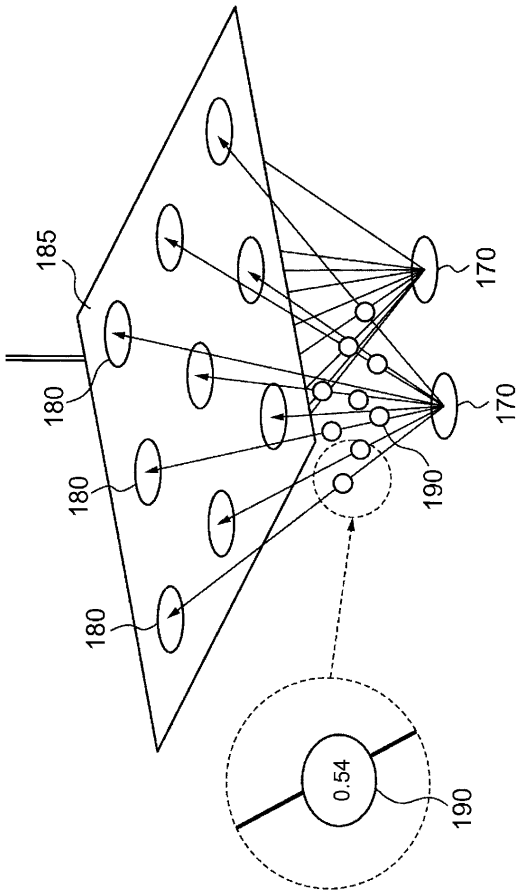
【図３】


ドキュメント1：
用語頻度ヒストグラム
ドキュメント2：
用語頻度ヒストグラム
Video
Transmitter
Systems
Recording
Production
MPEG
MetaData
Data
Computer
Comms.
Broadcast
用語
頻度
50
0
(a)
(b)

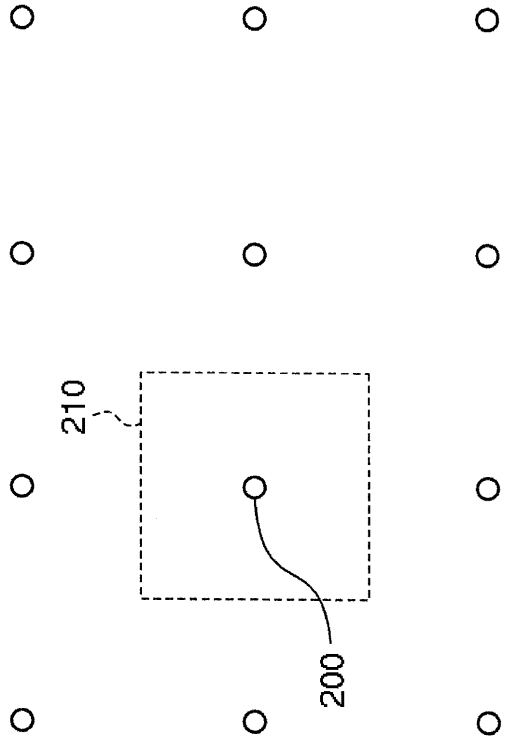
【図４】


0
0
0
0
1
0
0
0
0
0
2
.
.
.
0
0
0
0
3
0
0
3
0
1
0
7
0
0
0
2
(b)
(a)

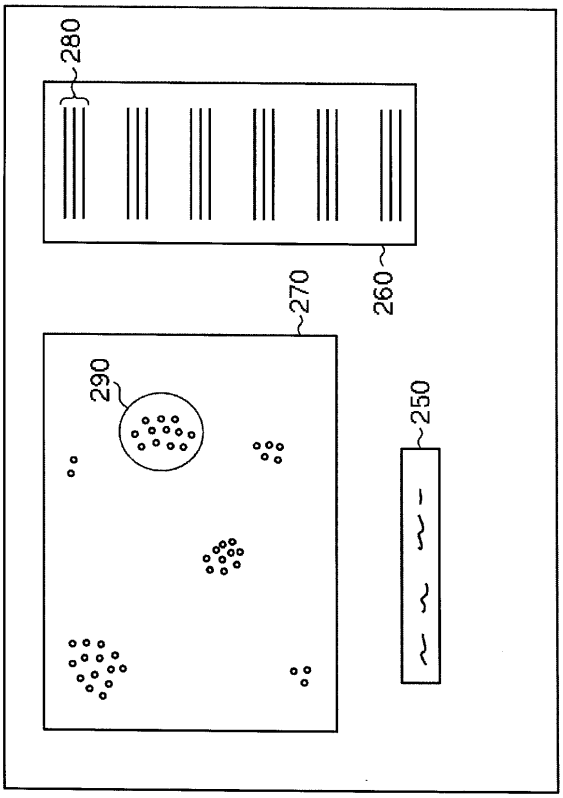
【図５】


185
180
180
180
170
170
190
190
0.54

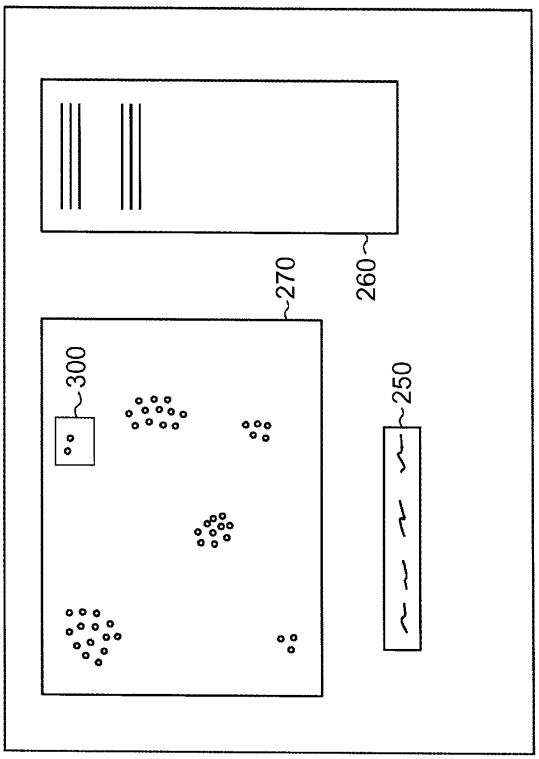
【図６】


210
200

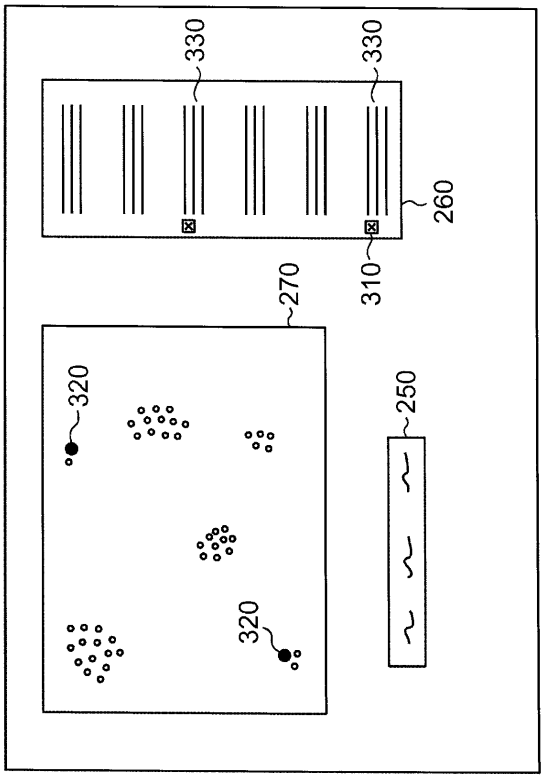
【図７】


280
270
260
290
250

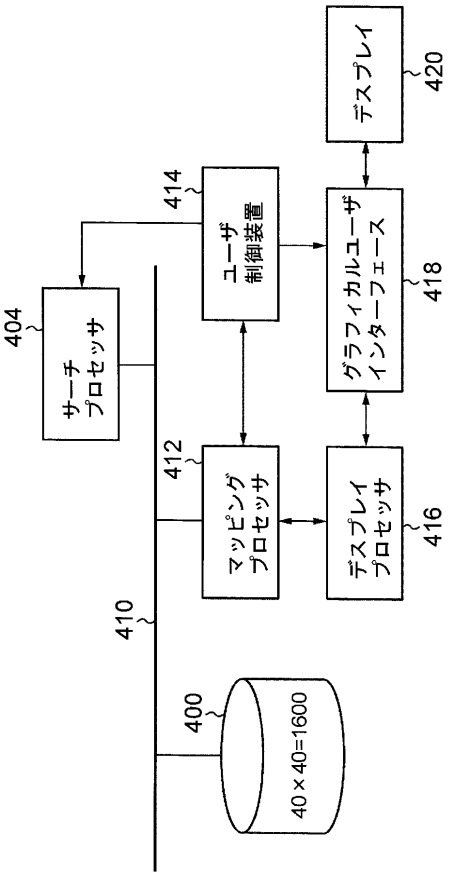
【図８】


270
260
300
250

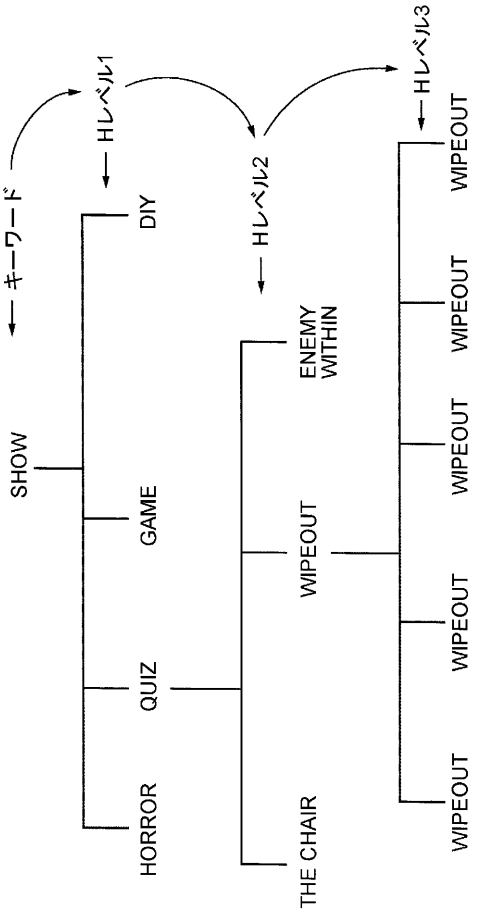
【図９】



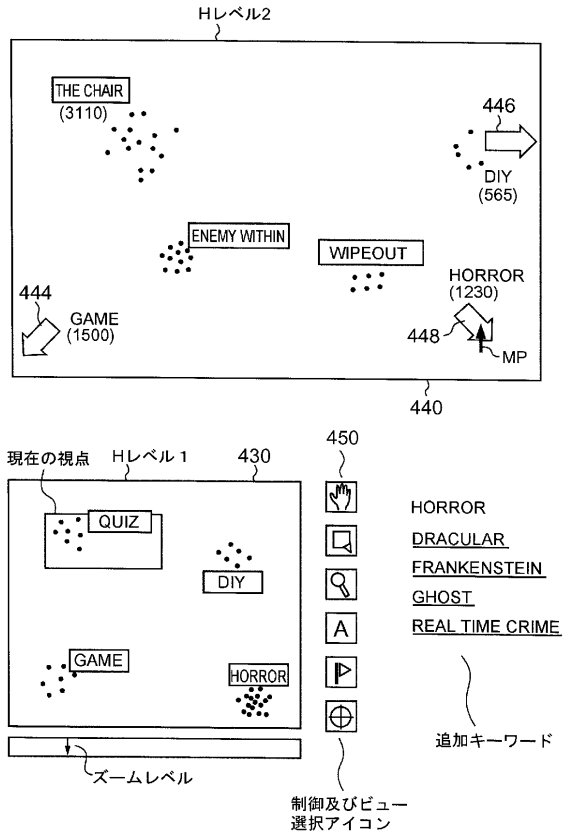
【図１０】



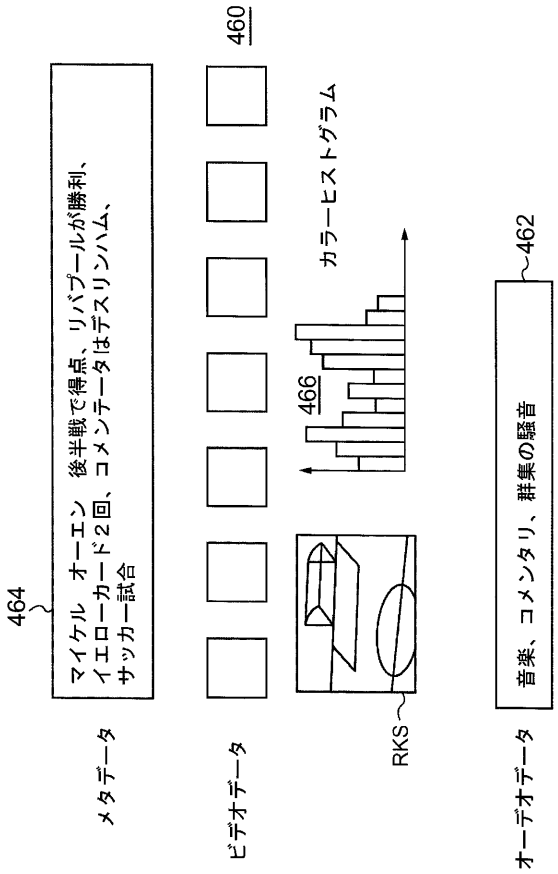
【図１１】



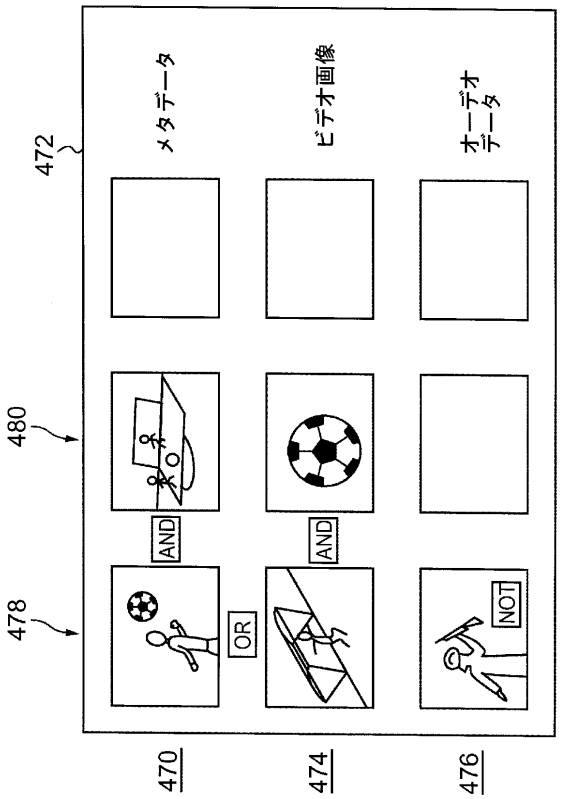
【図１２】



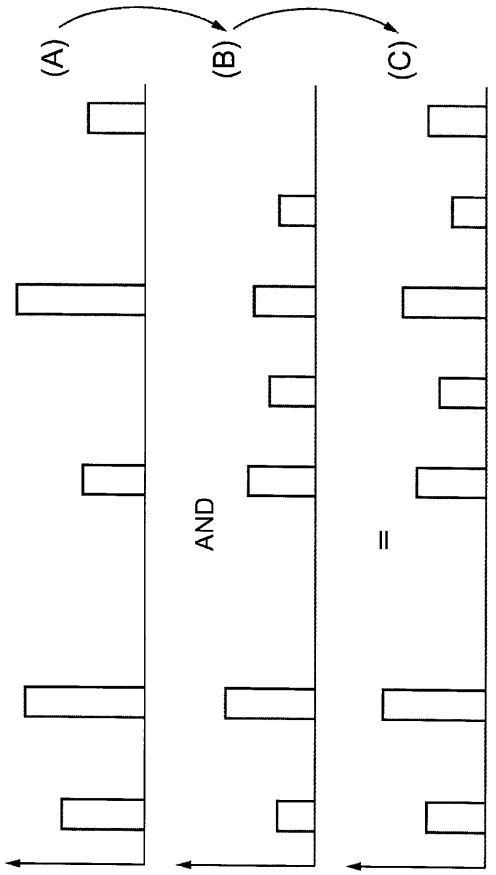
【図１３】



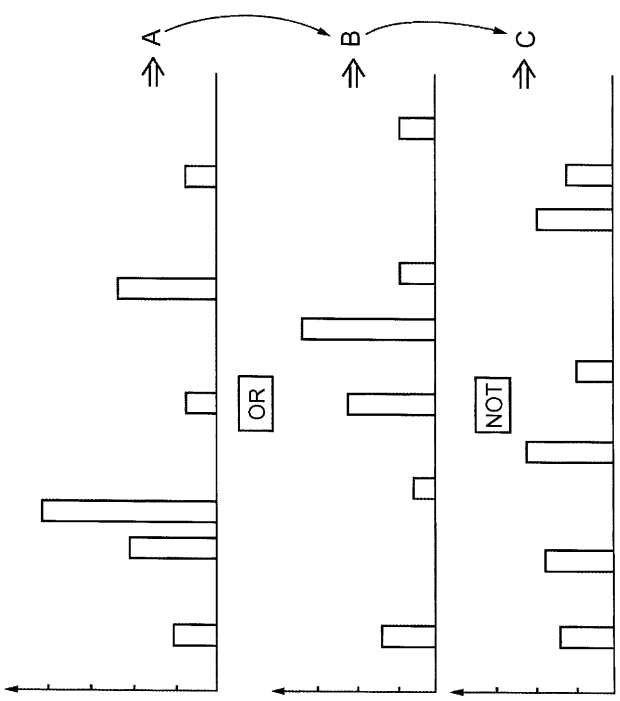
【図１４】



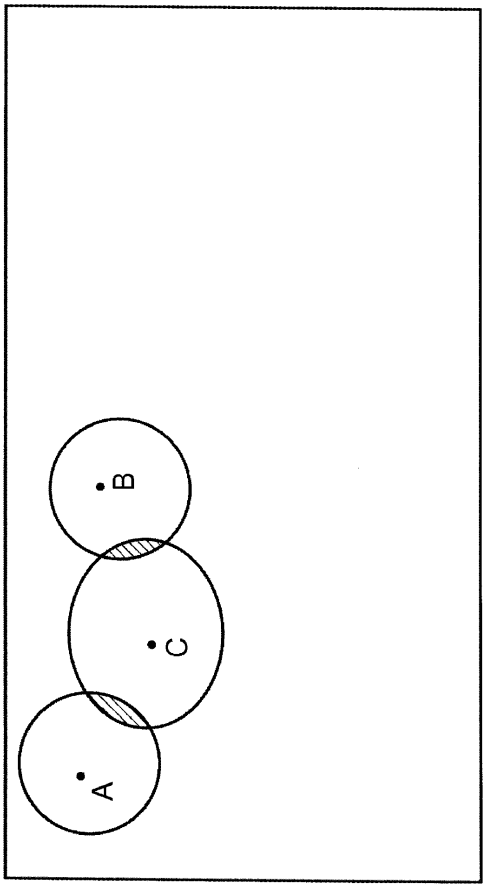
【図１５】



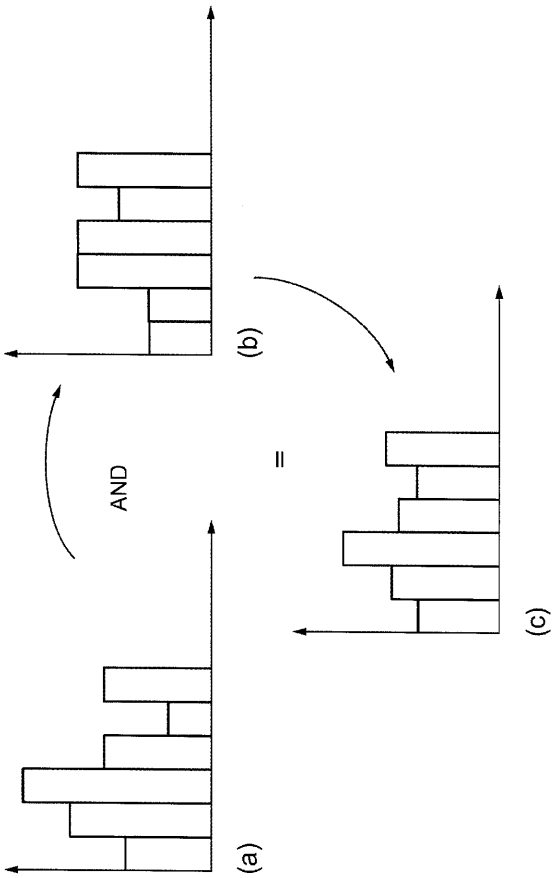
【図１６】



【図１７】



【図１８】



フロントページの続き

(72)発明者　トレペス、デヴィッド　ウィリアム
イギリス国　ＫＴ１３　ＯＸＷ　サリー、ウエィブリッジ、ブルックランズ、ザ　ハイツ　（番地無し）　ソニー　ユナイテッド　キングダム　リミテッド内
(72)発明者　ソープ、ジョナサン　リチャード
イギリス国　ＫＴ１３　ＯＸＷ　サリー、ウエィブリッジ、ブルックランズ、ザ　ハイツ　（番地無し）　ソニー　ユナイテッド　キングダム　リミテッド内

Ｆターム(参考) 5B075 NK43　NR12　PQ02　PQ13