

(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2004-178605

(P2004-178605A)

(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 360Z	5B075
G06F 3/00	G06F 17/30 350C	5E501
	G06F 3/00 651A	

審査請求 未請求 請求項の数 16 O L (全 27 頁)

(21) 出願番号	特願2003-398160 (P2003-398160)	(71) 出願人	593081408
(22) 出願日	平成15年11月27日 (2003.11.27)		ソニー・ユナイテッド・キングダム・リミテッド
(31) 優先権主張番号	0227692.1		Sony United Kingdom Limited
(32) 優先日	平成14年11月27日 (2002.11.27)		イギリス国 サリー, ウェブブリッジ, ブルックランズ, ザ ハイツ (番地なし)
(33) 優先権主張国	英国 (GB)	(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

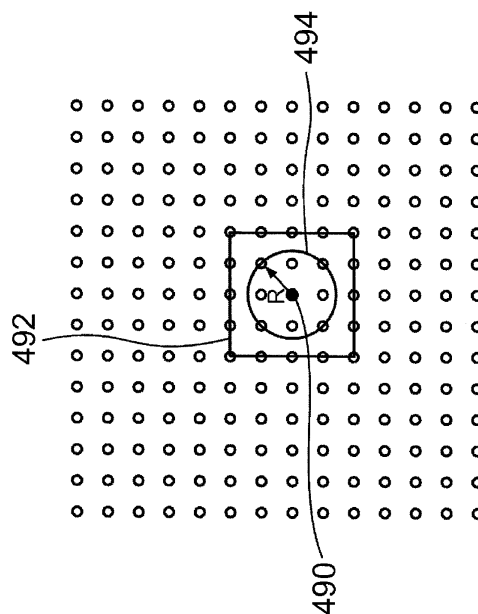
(54) 【発明の名称】 情報検索装置及びその方法

(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、相対的に短い検索「ヒット (hits)」の一覧を得ることは困難である。

【解決手段】 情報項目の集合を検索するための情報検索装置であり、マッピングプロセッサと、グラフィカルユーザインタフェースと、ユーザ制御装置と、検索プロセッサとを備える。マッピングプロセッサは、一組の情報項目から情報項目のマップを表すデータを生成するように動作可能である。マップは、情報項目の相互類似性に応じて、配列内の位置に対する情報項目を与え、同様の情報項目は配列中の同様の位置にマッピングする。

【選択図】 図 2 0



【特許請求の範囲】**【請求項 1】**

情報項目の集合から情報項目のマップを表し、前記マップが情報項目の相互類似性に従って配列中の位置に対して識別された情報項目を提供し、類似の情報項目が配列中の類似の位置にマッピングされるデータを生成するように動作可能なマッピングプロセッサと、前記情報項目の少なくとも幾つかの表現を表示するように動作可能なグラフィカルユーザインタフェースと、

ユーザが選択した情報項目に対応する配列位置の近隣位置である配列中の位置に対応する情報項目を識別することによって、ユーザが選択した情報項目に対する関連検索を行うように動作可能である検索プロセッサとを有した情報項目を選択するためのユーザ制御装置とを備える情報検索装置。

10

【請求項 2】

前記検索プロセッサは、検索クエリに従って情報項目の集合を検索し、その検索クエリに対応する情報項目を判別するように動作可能であり、前記マッピングプロセッサは、検索クエリに基づく検索の結果として前記検索プロセッサにより判別された情報項目のマップデータを発生するように動作可能であ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】

前記グラフィカルユーザインタフェースは、判別された情報項目に対応する配列の位置の少なくとも幾つかの表現を、表示領域内の表示点の n 次元表示配列として表示するように動作可能なことを特徴とする請求項 1 又は 2 に記載の情報検索装置。

20

【請求項 4】

前記次元の数 n は 2 であり、配列における位置は x y 座標で定義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】

ユーザの選択による情報項目に対応する配列位置から半径内の位置である配列における位置に対応する情報項目を判別することによって前記検索プロセッサはユーザの選択による情報項目に関連して関連検索を実行するように動作可能なことを特徴とする請求項 4 に記載の情報検索装置。

【請求項 6】

前記ユーザ制御装置が、興味ある配列位置について、関連検索で前記の検索プロセッサに検索される情報項目の相対類似性に従って位置の半径を特定する機能をユーザに与えるように動作することを特徴とする請求項 1 乃至 5 のいずれか 1 項に記載の情報検索装置。

30

【請求項 7】

情報項目の集合を検索する情報項目検索方法において、

前記情報項目の集合から情報項目のマップを表し、情報項目の相互類似性に従って配列中の位置に対して識別された情報項目を提供し、類似の情報項目が配列中の類似の位置にマッピングされるデータを生成し、

前記情報項目の少なくとも幾つかの表現を表示し、

前記情報項目を選択し、ユーザの選択による情報項目に対応する配列位置に対応し、近傍位置である配列における位置に対応する情報項目を判別することによってユーザの選択による情報項目に関連して関連検索を実行するステップを有する情報項目検索方法。

40

【請求項 8】

検索クエリに従って情報項目を検索し、

前記検索クエリに対応する情報項目を判別し、

前記検索クエリに基づく検索の結果として前記検索プロセッサによって判別された情報項目のマップデータを生成するステップを有する請求項 7 に記載の情報項目検索方法。

【請求項 9】

前記判別された情報項目に対応する配列の位置の少なくとも幾つかの表現を、表示領域内の表示点の n 次元表示配列として表示するステップを有する請求項 8 に記載の情報項目

50

検索方法。

【請求項 1 0】

前記次元の数 n は 2 であり、配列における位置は $x y$ 座標で定義されることを特徴とする請求項 9 に記載の情報項目検索方法。

【請求項 1 1】

前記関連検索を実行するステップは、情報項目を選択し、ユーザの選択による情報項目に対応する配列位置に対応し、近傍位置である配列における位置に対応する情報項目を判別することによってユーザの選択による情報項目に関連して関連検索を実行するステップを有することを特徴とする請求項 1 0 に記載の情報項目検索方法。

【請求項 1 2】

ユーザ制御装置が、興味ある配列位置について、関連検索で検索プロセッサに検索される情報項目の相対類似性に従って位置の半径を特定する機能をユーザに与えるように動作することを特徴とする請求項 7 乃至 1 2 のいずれか 1 項に記載の情報項目検索方法。

10

【請求項 1 3】

請求項 7 乃至 1 2 のいずれか 1 項に記載の情報項目検索方法を実行するためのプログラムコードを有するコンピュータソフトウェア。

【請求項 1 4】

請求項 1 3 に記載のプログラムコードを提供するプログラム提供媒体。

【請求項 1 5】

前記媒体が蓄積媒体であることを特徴とする請求項 1 4 に記載のプログラム提供媒体。

20

【請求項 1 6】

前記媒体が伝送媒体であることを特徴とする請求項 1 4 に記載のプログラム提供媒体。

【発明の詳細な説明】

【技術分野】

【0 0 0 1】

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】

【0 0 0 2】

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。

30

【0 0 0 3】

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、2 4 3 0 0 0 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。

40

【0 0 0 4】

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

50

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 I E E E 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。

10

【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、各単語の大辞書が発生する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。

20

【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 ($n \times m$) 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。

30

【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、800 MB 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。

40

【発明の開示】

【発明が解決しようとする課題】

【0009】

先に述べたように、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、相対的に短い検索「ヒット (hits)」の一覧を得ることは困難である。特に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想される。そのようなヒット一覧を精査することには、非常に時間

50

がかかる可能性が大である。

【0010】

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。

【課題を解決するための手段】

【0011】

本発明の様々な態様及び特徴は、添付の請求項において定義される。

【0012】

本発明の1つの態様によると、情報項目の集合を検索を行うための情報検索装置が提供される。この装置は、情報項目の集合から情報項目のマップを表すデータを生成するように動作可能なマッピングプロセッサを備える。このマップは、情報項目の相互類似性に従って配列中の位置に対して識別された情報項目を提供し、類似の情報項目を配列中の類似の位置にマッピングする。グラフィカルユーザインタフェースは、情報項目の少なくとも幾つかの表現を表示するように動作可能であり、ユーザ制御装置は、識別された情報項目を選択するために設けられる。検索プロセッサは、ユーザが選択した情報項目に対応する配列位置の近隣位置である配列中の位置に対応する情報項目を識別することによって、ユーザが選択した情報項目に対する関連検索を行うように動作可能である。検索プロセッサは、キーワードのような幾つかの特徴付け情報特徴について情報項目を検索するのではなく、配列から検索されている情報項目を識別するように構成されるので、対象となる情報項目についての検索は複雑さが低減されて行われることが可能である。

【発明の効果】

【0013】

本発明の実施の形態によって提供される利点は、ユーザが情報項目の集合から対象となる情報項目を識別した場合に、強く相関された情報項目がユーザに与えられることが可能なことである。相関された情報項目は、対象となる情報項目に対応する配列中の位置からの所定数の位置である配列内の位置から項目を識別することによって提供される。

【0014】

他の実施の形態において、検索プロセッサは、検索クエリに従って情報項目を検索し、かつ、検索クエリに対応する情報項目を識別するように動作可能であることが可能である。マッピングプロセッサは、検索クエリによる検索結果として、検索プロセッサによって識別された情報項目のマップデータを生成させるように動作可能であり得る。したがって、検索プロセッサは最初の検索を行って、特定の検索クエリに対応する情報項目を識別し得る。検索の結果、ユーザは対象となる情報項目を識別し得る。したがって、本発明の1つの実施の形態により、対象となる情報項目に対応する配列位置の所定数の位置内の配列位置に対応する情報項目を識別する「探索関連 (find related)」オプションに従ってユーザが検索を行う機能が提供される。この目的のために、ユーザ制御装置によって、探索関連検索を開始する機能が提供され得る。

【0015】

グラフィカルユーザインタフェースは、表示領域内の表示点のn次元表示配列として、識別された情報項目に対応する配列の位置の少なくとも幾つかの表現を表示するように動作可能である。有利とするため、目視及びナビゲーションを容易にするために、配列の次元数は2である。したがって、配列中の位置はx及びy座標によって定義される。したがって、幾つかの実施の形態において、検索プロセッサは、ユーザが選択した情報項目に対

応する配列位置から、位置半径を有する円内に存在する配列中の位置に対応する情報項目を識別することによって、関連検索を行うように動作可能である。

【0016】

必要とされる関連項目の相対的範囲に従って関連検索を開始するために、関連検索において検索プロセッサによって検索される情報項目の相対的類似性に従って、位置の半径を指定する機能をユーザに与えるように、ユーザ制御装置が構成され得る。

【発明を実施するための最良の形態】

【0017】

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。

【0018】

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。

【0019】

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。

【0020】

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。

【0021】

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。

【0022】

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を介して情報検索及びシステム(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

10

20

30

40

50

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。

【0023】

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。 10

【0024】

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、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的である場合、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。 20

【0025】

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。 30

【0026】

プロセスは、特定数の情報項目に限定されない。

【0027】

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。

【0028】

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。 40

【0029】

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。

【0030】

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。

【0031】

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。

【0032】

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（-2）値よりも低い、又は平均出現頻度に2標準偏差を足した（+2）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。

【0033】

次いで、特徴ベクトルがステップ130で生成される。

【0034】

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。

【0035】

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。

【0036】

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。

【0037】

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。

【0038】

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。

【0039】

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。

【0040】

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。

【0041】

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。

【0042】

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。

【0043】

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。

【0044】

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。

【0045】

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。

【0046】

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。

【0047】

10

20

30

40

50

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。

【0048】

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。 10

【0049】

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更す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。

【0050】

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。 20

【0051】

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。

【0052】

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。 30

【0053】

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。

【0054】

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。 40

【0055】

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。 50

【 0 0 5 6 】

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 1 1 0 から 1 4 0）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 1 5 0 及び 1 6 0 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。

10

【 0 0 5 7 】

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が初めて作成されて以降、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 1 5 0 及び 1 6 0 全体を再適用する。

20

【 0 0 5 8 】

図 7 は、表示画面 6 0 上の表示を概略的に図示する。表示は、検索クエリ領域 2 5 0、結果一覧領域 2 6 0 及び SOM 表示領域 2 7 0 を示す。

【 0 0 5 9 】

動作中は、始めはSOM表示領域 2 7 0 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 2 5 0 に入力する。次いで、ユーザは、例えば、キーボード 7 0 上の復改キーを押すことによって、又はマウス 8 0 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 2 5 0 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 2 6 0 中の各々の見出し項目 2 8 0 として示される。次いで、SOM表示領域 2 7 0 は、各結果項目に対応する表示点を表示する。

30

【 0 0 6 0 】

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 2 9 0 などのクラスタが生じる傾向にある。ここで、SOM表示領域 2 7 0 上の各点が、結果一覧領域 2 6 0 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 2 7 0 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。

40

【 0 0 6 1 】

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 8 0 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 2 7 0 中に表示される一組の表示点を囲む矩形ボックス 3 0 0 である。結果一覧領域 2 6 0 において、境界線 3 0 0 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。

【 0 0 6 2 】

結果一覧領域 2 6 0 は、境界線 3 0 0 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 2 5 0 中の検索基準を満たしたこと

50

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。

【0063】

図 9 は、本発明の 1 つの実施の形態を図示する。

【0064】

図 9 を参照すると、ステップ 920 で自己組織化マップ SOM が作成される時、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 及びノ又は図 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に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。

10

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。

20

【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。それによって、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。

30

【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。キーワードを受け取られると、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。

40

50

【0068】

1つの実施の形態において、検索プロセッサ404は、情報項目を検索し、かつ、検索クエリに対応した、情報項目を識別する検索結果を生じさせるように構成されることが可能である。次いで、マッピングプロセッサ412は、検索クエリに対応する情報項目を識別する検索結果を表すデータを受け取ることが可能である。次いで、マッピングプロセッサ412は、識別された情報項目に対応する配列内位置のx及びy座標を作成する。

【0069】

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。

【0070】

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。

【0071】

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。

【0072】

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。

【0073】

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。

【0074】

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。

10

【0075】

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。

【0076】

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。

20

30

【0077】

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

40

50

れる。ユーザコントロ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。

【0078】

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。

10

【0079】

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「群衆の騒音(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。

20

30

【0080】

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。

40

【0081】

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。

50

・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。

【0082】

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。

【0083】

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。

【0084】

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。

【0085】

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。

【0086】

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

10

20

30

40

50

する。

【0087】

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。

【0088】

「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。

【0089】

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。

【0090】

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。

【0091】

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。

10

【0092】

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、トニーブレイア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せられて、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。

20

【0093】

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。

30

【0094】

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隣接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。

40

【0095】

図19は、検索プロセッサ404がどのように「探索関連(find related)」検索を行うことが可能であるかを概略的に示す。ユーザは、特定の情報項目を対象とすると考えることが可能である。例えば、図19は、図7、図8及び図9に示されるグラフィカルユー

50

ザインタフェースの表現を再現するものである。以前の検索の結果によって、前記のように黒点によって図示される配列中の識別された位置の配置が明らかになると考えると、ユーザは、配列内の位置490に対応する特定の情報項目が対象となるものであることを見出している。関連検索を行うためには、ユーザはマウスポインタを対象とする位置490上に位置付けるように配置し、例えば、自動的に現れることが可能であるメニューオブションを通じて関連検索を行う。関連検索が行われると、検索プロセッサ404は、対象位置490から所定数の近隣位置内の配列位置に対応する情報項目を識別する。例えば、検索プロセッサ404は、図20に示すように、x及びy方向のプラス及びマイナス2の位置から形成される正方形ボックス492内の配列位置に対応する情報項目を識別することが可能である。あるいは、検索プロセッサ404は、対象とする選択された情報項目の配列位置490から、対角線上の1つの位置の所定半径Rを有する円内の配列位置に対応する情報項目を識別してもよい。

【0096】

関連配列位置に対応する情報項目を識別した後、各識別された情報項目についての特徴付け情報特徴が、図19に示される結果一覧領域260内に表示されることが可能である。

【0097】

幾つかの実施の形態においては、それに対する関連情報項目が識別されることが可能である配列位置の数は、関連検索が実行される相対的な感度に応じて、ユーザ制御装置を用いてユーザによって変えられることが可能である。したがって、関連検索において識別される所定の近隣位置の数は変化し得る。これは、円494の半径R又は正方形ボックス492のサイズを変えることによって行われ得る。

【0098】

キーワード検索などのように幾つかの特徴付け情報特徴について情報項目を検索することによってではなく、配列に基づいて関連検索を行うことによって、キーワードでの検索と比較してコンピュータの複雑さを低減することが可能な、対象とする情報項目の検索機能が提供される。配列を用いた関連検索動作は、配列内での同様の位置に同様の情報項目を位置付ける傾向にあるSOMの特性によって容易になる。したがって、対象とする情報項目に対応する位置に対する配列内の近隣位置を有する情報項目は、その情報項目に対して相関される。したがって、これらの近隣位置に対応する情報項目を検索することによって、ユーザの検索要件に一致する可能性がより高い項目に対する限定した検索が示される。

【0099】

関連検索の要約流れ図

検索プロセッサが関連検索を行うときの動作を要約する流れ図を、図21に示す。関連検索プロセスのステップは、以下のように要約される。

S2：動作中の第1のステップによって、ユーザ指定の検索クエリに従って情報項目の集合から情報項目のマップを表すデータを生成することが可能であるが、ユーザによって識別された情報項目から関連検索が行われる場合は、ステップS2～S10を省略してもよい。マップは、情報項目の相互類似性に従って、配列内の位置に対する情報項目を与え、類似した情報項目は、配列中の類似した位置にマッピングされる。

S4：情報項目は、検索において識別されるx及びy配列位置から、又はマッピングプロセッサによって二次元配列内の位置にマッピングされる。

S8：表示のために、配列内の配列位置のx及びy位置からマップデータが生成される。

S10：マップデータに従って、情報項目の少なくとも一部の表現が二次元配列として表示される。

S12：対象となる情報項目をユーザが選択する。

S14：ユーザは、関連検索が行われる条件を指定することが可能である。ユーザは、近隣位置の数又は対象とする情報項目の配列位置からの半径を識別することが可能である。

S16：ユーザが関連検索に対して特定の要件を指定しなかった場合、検索プロセッサは

所定の近隣位置の数を自動的に識別し、これらの位置に対応する情報項目を戻す。

S 1 8 : ユーザが関連検索に対して特定の要件を指定している場合、検索プロセッサは、ユーザ指定に従って近隣位置を識別し、これらの位置に対応する情報項目を戻す。

【 0 1 0 0 】

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。本発明の様々な局面及び特徴は、添付の請求項に定義される。

【 図面の簡単な説明 】

【 0 1 0 1 】

【 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 (S O M) の生成を示す概略的なフローチャートである。 10

【 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 】 S O M 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 】 S O M 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 】 S O M 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 】 S O M 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。 20

【 図 1 0 】 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 1 1 】 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 1 2 】 図 1 1 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 1 3 】 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 1 4 】 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 1 5 】 ブール A N D 演算による複合特徴ベクトルの形成の概略図である。 30

【 図 1 6 】 ブール O R 演算子による 2 つの特徴ベクトルと、ブール N O T 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 1 7 】 図 1 6 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 1 8 】 (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。

【 図 1 9 】 図 7 ~ 図 9 において図示されるユーザインタフェースに対応する、 S O M によって表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。 40

【 図 2 0 】 二次元の位置配列を例示的に表現する図であり、この位置に対する関連検索が図示される。

【 図 2 1 】 検索プロセッサによって行われる関連検索の動作を示す流れ図である。

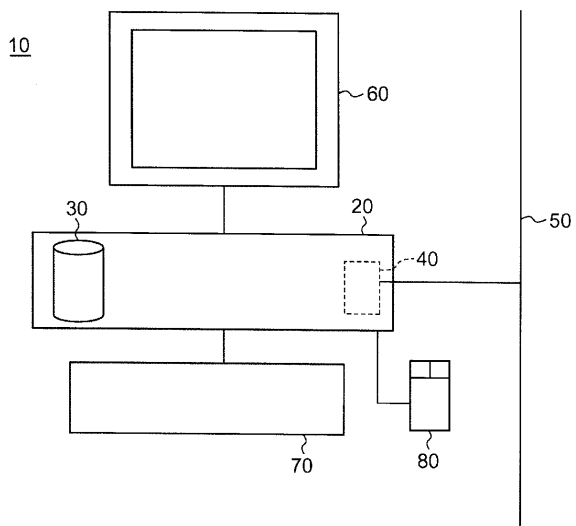
【 符号の説明 】

【 0 1 0 2 】

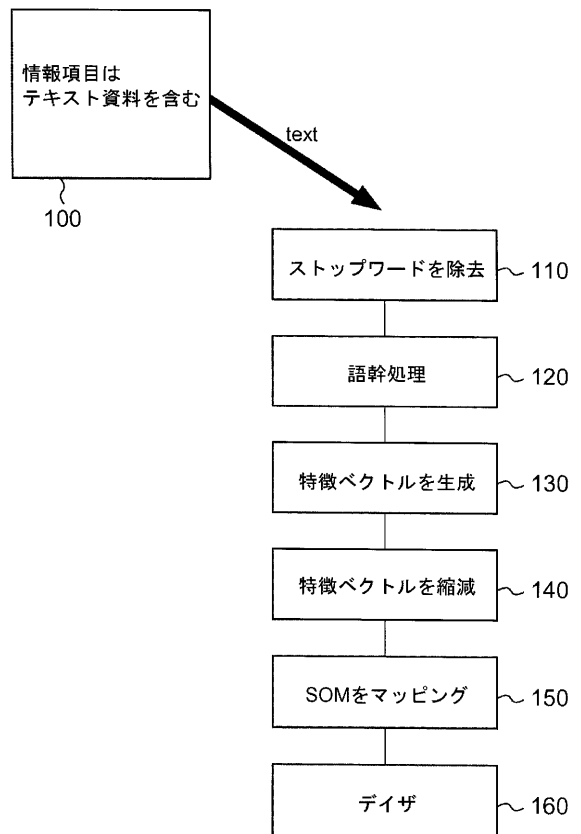
1 0 汎用コンピュータ、 2 0 プロセッサユニット、 3 0 ディスク記憶装置、 4 0 ネットワークインタフェースカード、 5 0 ネットワーク、 6 0 陰極線管表示装置、 7 0 キーボード、 8 0 マウス、 4 0 0 データ格納装置、 4 0 4 検索プロセッサ、 4 1 0 通信ネットワーク、 4 1 2 マッピングプロセッサ、 4 1 4 ユーザ制御装置、 4 1 6 ディスプレイプロセッサ、 4 1 8 グラフィカルユーザインタフェース (G U I) 50

、 4 2 0 ディスプレイ

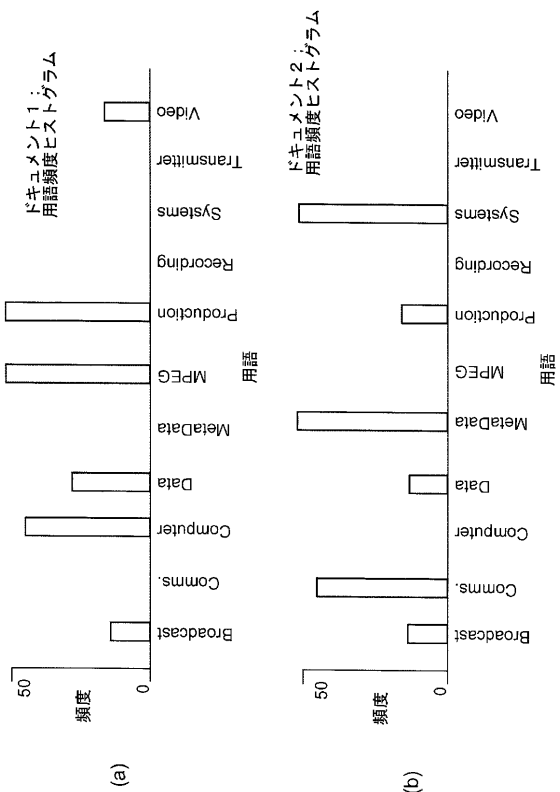
【 図 1 】



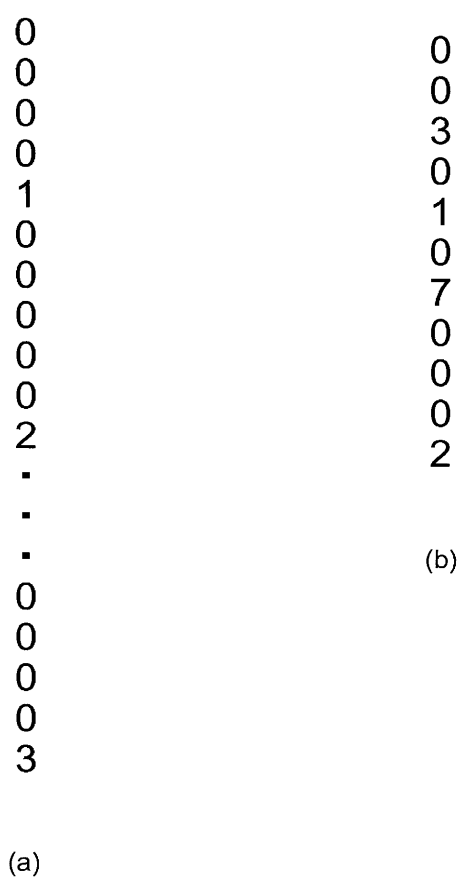
【 図 2 】



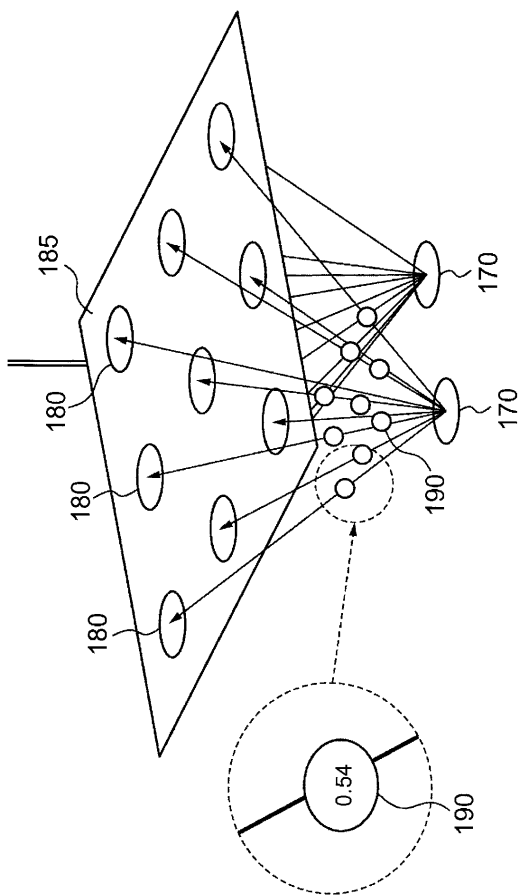
【図3】



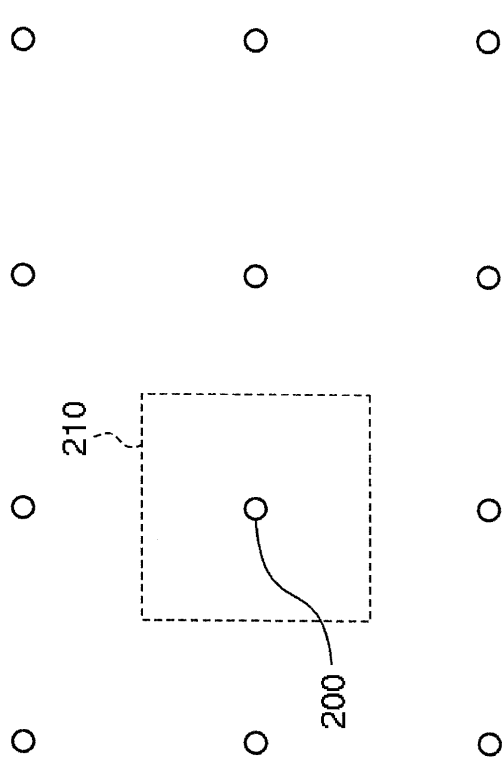
【図4】



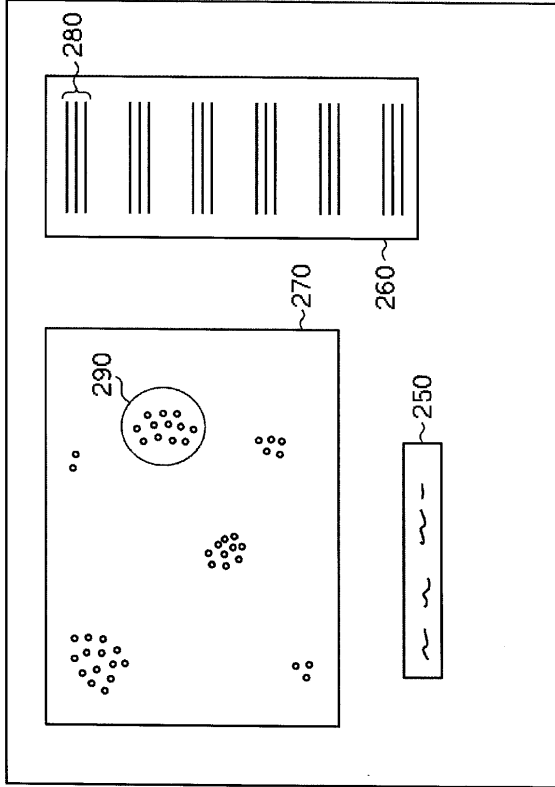
【図5】



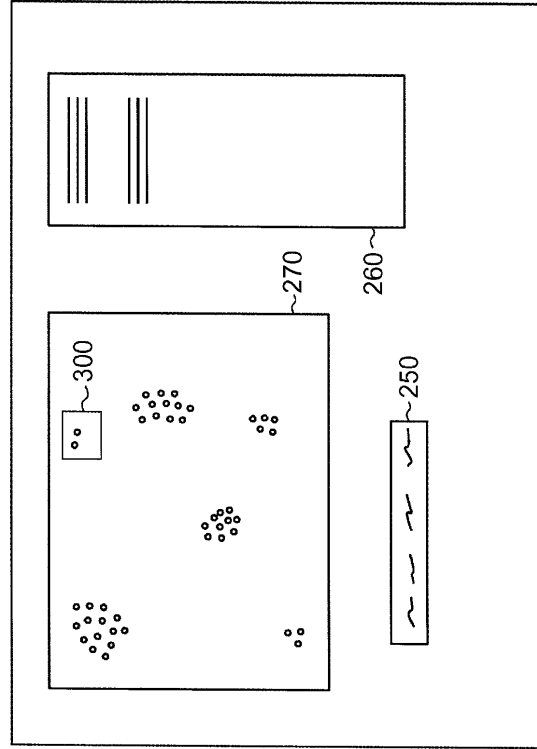
【図9】



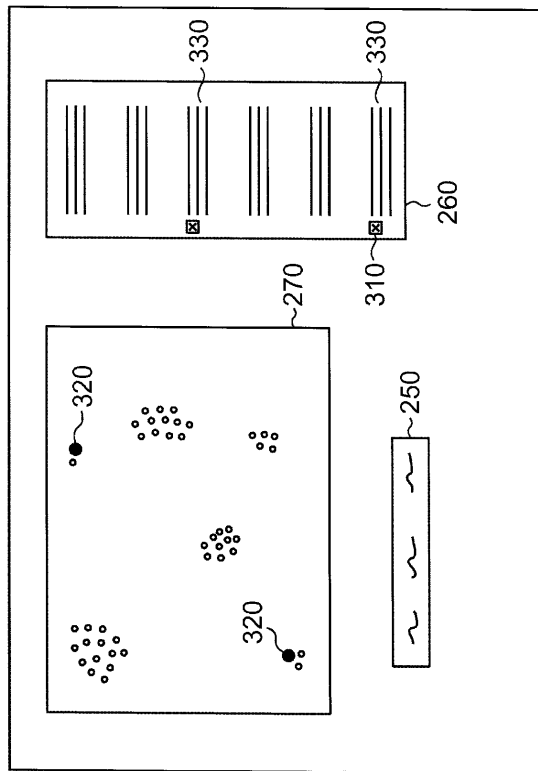
【 図 7 】



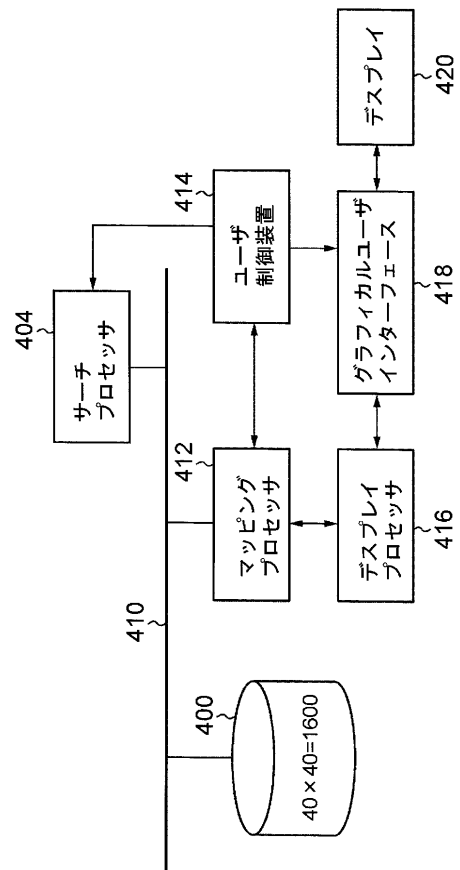
【 図 8 】



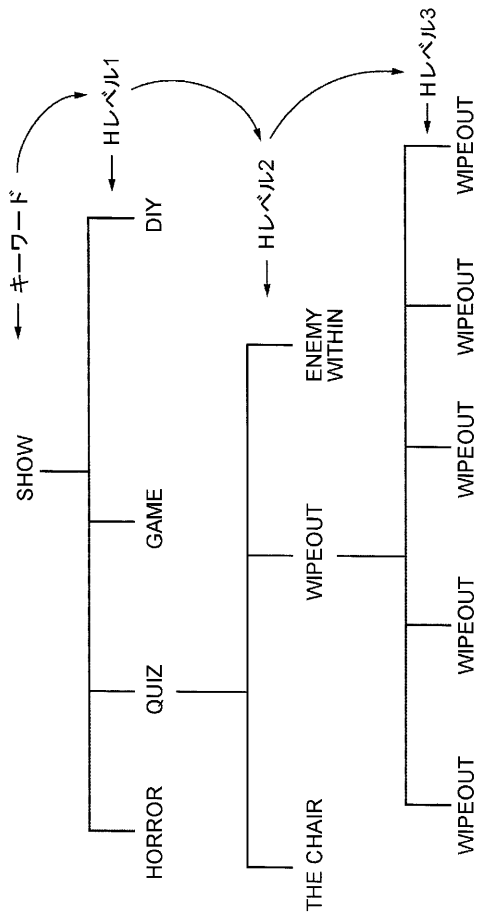
【 図 9 】



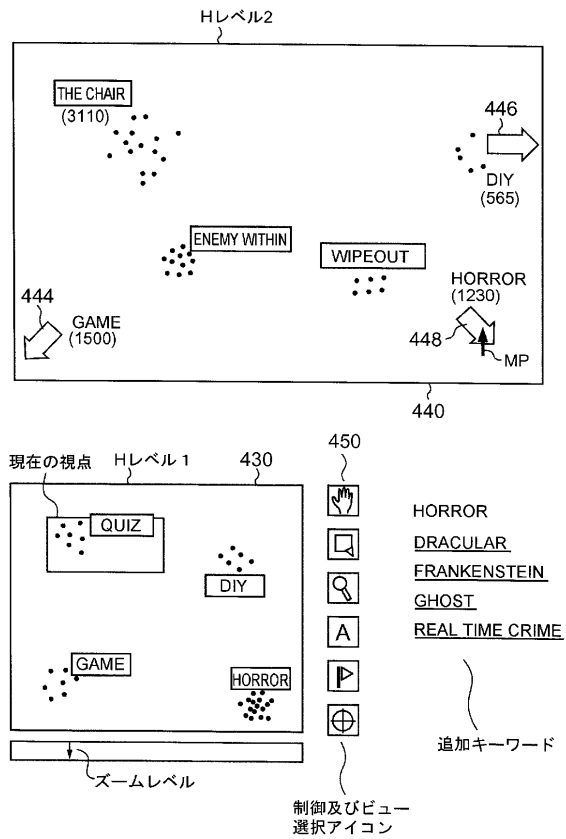
【 図 10 】



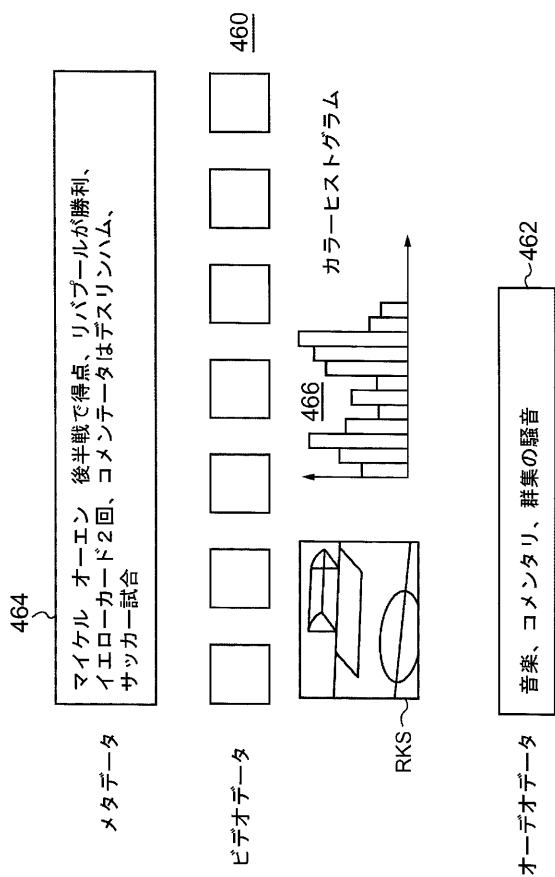
【 図 1 1 】



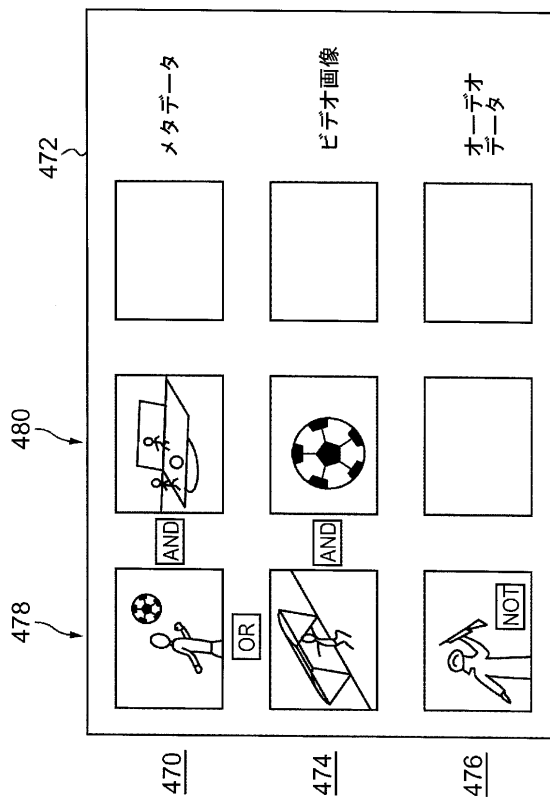
【 図 1 2 】



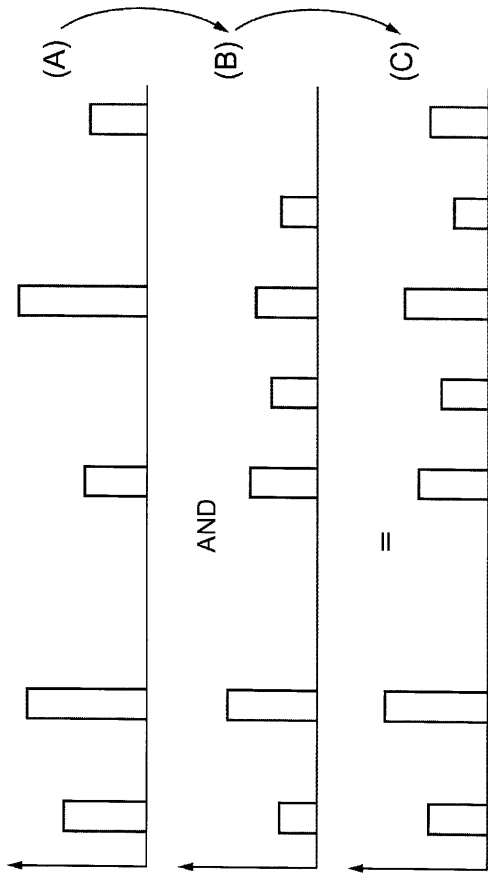
【 図 1 3 】



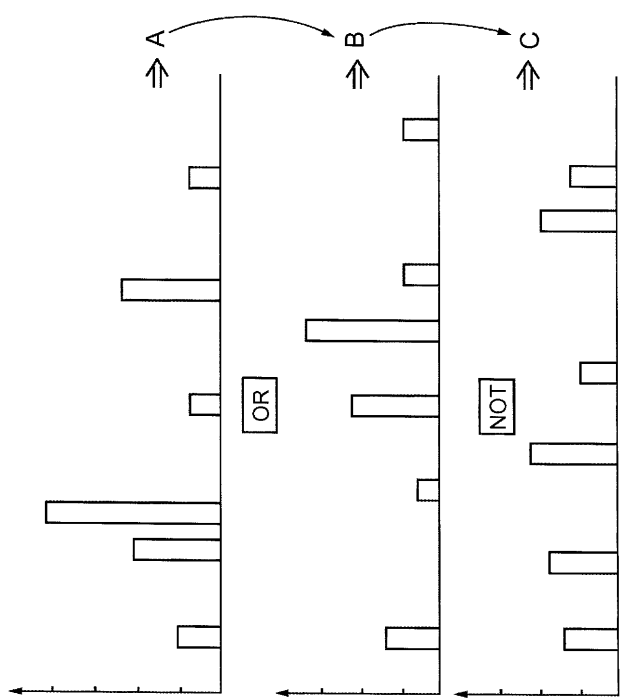
【 図 1 4 】



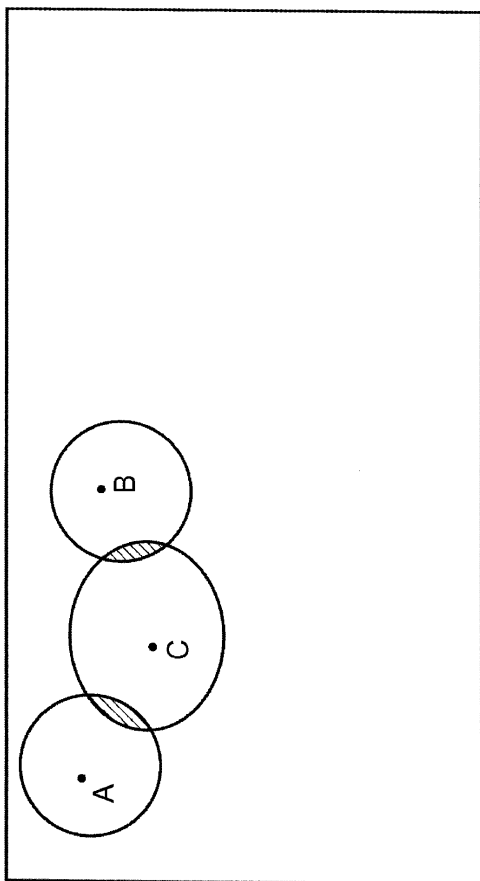
【 図 15 】



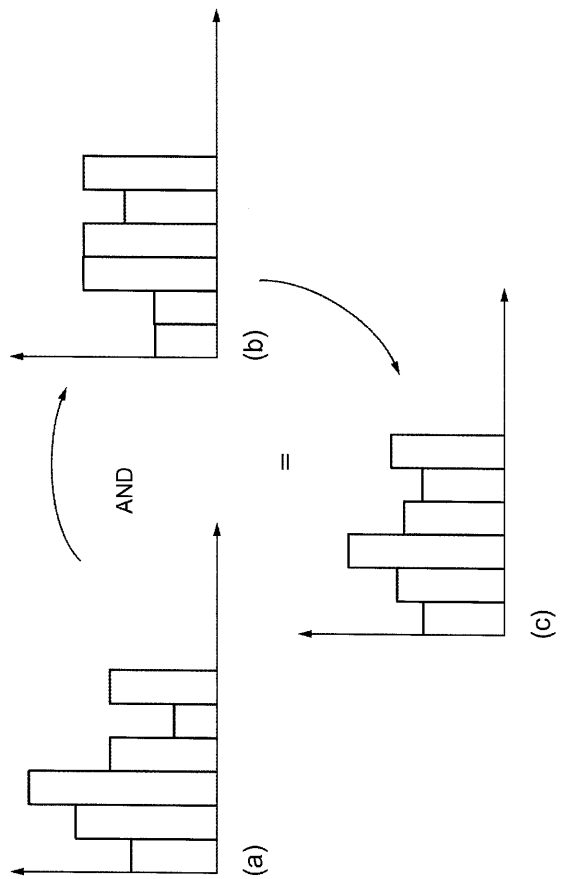
【 図 16 】



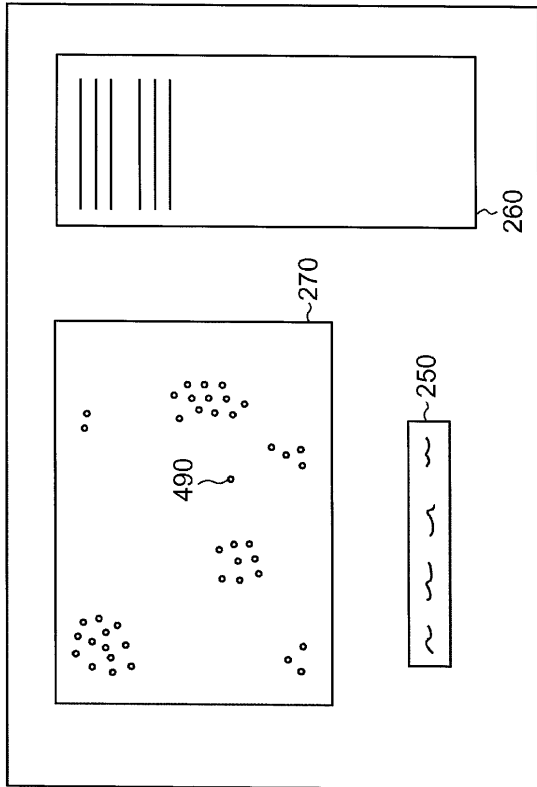
【 図 17 】



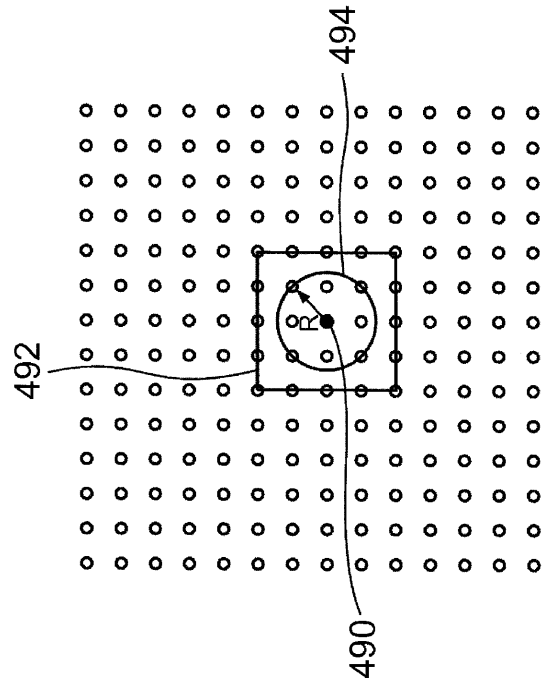
【 図 18 】



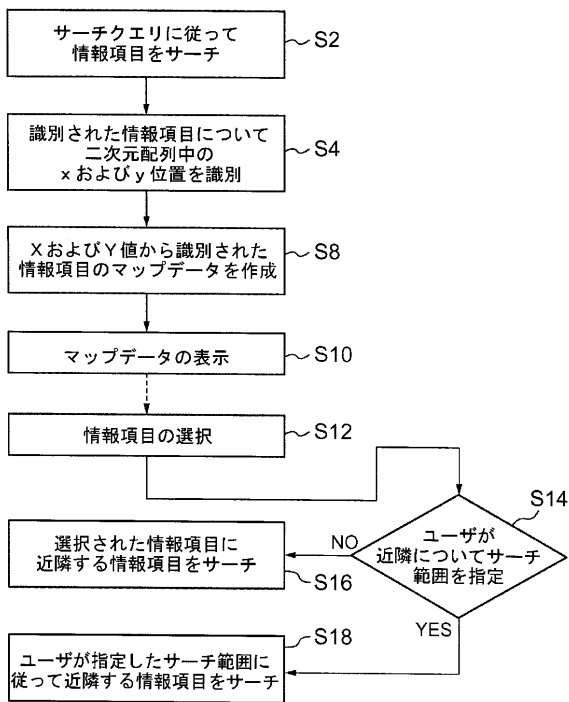
【図19】



【図20】



【図21】



フロントページの続き

(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、ウェイブリッジ、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 PQ02 PQ13 QM05

5E501 AA01 AA03 AA04 AA20 AC15 AC23 AC33 BA05 CA03 CB02

CB09 DA17 EA05 EA10 EB07 FA03 FA13 FA14 FA44 FA48