(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2004-213626
(P2004-213626A)
(43)公開日　平成16年7月29日(2004.7.29)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	360Z	5B075
G06F 12/00	G06F 12/00	513D	5B082

審査請求 未請求　請求項の数 39　OL　外国語出願　（全 47 頁）

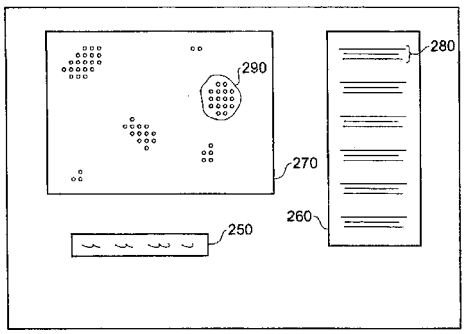
(21)出願番号　特願2003-398038(P2003-398038)
(22)出願日　平成15年11月27日(2003.11.27)
(31)優先権主張番号　0227658.2
(32)優先日　平成14年11月27日(2002.11.27)
(33)優先権主張国　英国(GB)

(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】情報の格納及び検索

(57)【要約】　（修正有）
【課題】検索結果情報をユーザにわかりやすく表示し、検索を支援する。
【解決手段】異なる性質の情報アイテムの組内の各情報アイテムが情報アイテムの相互の類似性に基づいてノードのアレー内の各ノードにマッピングされ、類似する情報アイテムが、ノードのアレー内の類似する位置におけるノードにマッピングされる情報検索装置は、情報アイテムを選択するための検索条件を定義するユーザ操作手段と、ノードのアレー内で、選択された情報アイテムに対応する位置を検出する検出手段と、選択された情報アイテムに対応するノードのアレー内の位置を代表する表示点を表示するグラフィックユーザインタフェースと、検索条件によって定義された選択された情報アイテムに応じて、選択された情報アイテムの情報コンテンツを代表する１以上の表現を提供するプロセッサとを備える。
【選択図】図７



【特許請求の範囲】
【請求項１】
異なる性質の情報アイテムの組内の各情報アイテムが該情報アイテムの相互の類似性に基づいてノードのアレー内の各ノードにマッピングされ、類似する情報アイテムが、該ノードのアレー内の類似する位置におけるノードにマッピングされる情報検索装置において、
上記情報アイテムを選択するための検索条件を定義するユーザ操作手段と、
上記ノードのアレー内で、上記選択された情報アイテムに対応する位置を検出する検出手段と、
選択された情報アイテムに対応するノードのアレー内の位置を代表する表示点を表示するグラフィックユーザインタフェースと、
上記検索条件によって定義された選択された情報アイテムに応じて、該選択された情報アイテムの情報コンテンツを代表する１以上の表現を提供するプロセッサとを備える情報検索装置。
【請求項２】
上記グラフィックユーザインタフェースは、上記表示点の２次元表示アレーを表示することを特徴とする請求項１記載の情報検索装置。
【請求項３】
上記情報アイテムとノードの間のマッピングにディザ成分が適用され、実質的に同一の情報アイテムが、上記アレーにおいて、近接しているが異なるノードにマッピングされるようにする請求項２記載の情報検索装置。
【請求項４】
上記情報アイテムは、各情報アイテムから導かれた特徴ベクトルに基づいて、上記アレー内のノードにマッピングされることを特徴とする請求項２記載の情報検索装置。
【請求項５】
上記情報アイテムから導かれた特徴ベクトルは、該情報アイテム内における、情報の特徴のグループ内の各特徴の出現頻度の組を表すことを特徴とする請求項４記載の情報検索装置。
【請求項６】
上記情報アイテムは、テキスト情報を含み、上記情報アイテムから導かれた特徴ベクトルは、該情報アイテム内における、単語のグループ内の各単語の出現頻度の組を表すことを特徴とする請求項５記載の情報検索装置。
【請求項７】
上記情報アイテムは、テキスト情報を含み、上記ノードは、上記テキスト情報の少なくとも一部の相互の類似性に基づいてマッピングされることを特徴とする請求項１記載の情報検索装置。
【請求項８】
上記情報アイテムは、上記情報アイテムの組における頻度の閾値を超える頻度で出現する単語を除外することによって、マッピングのための前処理が施されることを特徴とする請求項６記載の情報検索装置。
【請求項９】
上記情報アイテムは、上記情報アイテムの組における頻度の閾値を下回る頻度で出現する単語を除外することによって、マッピングのための前処理が施されることを特徴とする請求項６記載の情報検索装置。
【請求項１０】
上記ユーザ操作手段は、上記情報アイテムの単語に関する検索を実行する検索手段を有し、
上記検索手段及び上記グラフィックユーザインタフェースは、協働して、上記検索によって選択された情報アイテムに対応する表示点のみが表示されるように動作することを特徴とする請求項１記載の情報検索装置。

【請求項１１】
上記プロセッサは、類似する情報アイテムのクラスタを検出し、各クラスタの情報アイテムの代表的な表現を提供することを特徴とする請求項１記載の情報検索装置。
【請求項１２】
上記プロセッサは、上記表現又は各表現をユーザディスプレイに、該表現が代表している情報アイテムに対応する表示点のラベルとして表示することを特徴とする請求項１記載の情報検索装置。
【請求項１３】
上記ラベルは、単語又は単語の組であることを特徴とする請求項１２記載の情報検索装置。
【請求項１４】
上記プロセッサは、ラベルを付すべき情報アイテムの組に関して、上記選択された情報アイテムに対応する情報アイテム内で最も頻繁に用いられている単語又は単語の組を判定し、該単語又は単語の組をラベルとして適用することを特徴とする請求項１１記載の情報検索装置。
【請求項１５】
上記情報アイテムは、少なくとも画像アイテムに関連し、上記プロセッサは、上記検索条件によって定義された選択された情報アイテムの情報コンテンツを代表する１以上の画像を提供することを特徴とする請求項１記載の情報検索装置。
【請求項１６】
上記プロセッサは、上記画像アイテムの組から、所定の選択条件に基づいて、該画像アイテムの組を代表する画像アイテムを選択することを特徴とする請求項１５記載の情報検索装置。
【請求項１７】
上記プロセッサは、該画像アイテムの組における同じプロパティの平均に最も近いプロパティを有する画像アイテムを選択することを特徴とする請求項１５記載の情報検索装置。
【請求項１８】
上記１以上の代表する画像は、該画像によって代表される情報アイテムに対応する表示点のラベルとして適用されることを特徴とする請求項１５記載の情報検索装置。
【請求項１９】
請求項１記載の情報検索装置を備える携帯型データ処理装置。
【請求項２０】
請求項１記載の情報検索装置を備えるビデオ取得及び／又は処理装置。
【請求項２１】
異なる性質の情報アイテムの組内の各情報アイテムが該情報アイテムの相互の類似性に基づいてノードのアレー内の各ノードにマッピングされ、類似する情報アイテムが、該ノードのアレー内の類似する位置におけるノードにマッピングされる情報検索方法において、
上記情報アイテムを選択するための検索条件を定義するステップと、
ノードのアレー内で、上記選択された情報アイテムに対応する位置を検出するステップと、
選択された情報アイテムに対応するノードのアレー内の位置を代表する表示点を表示するステップと、
上記検索条件によって定義された選択された情報アイテムに応じて、該選択された情報アイテムの情報コンテンツを代表する１以上の表現を提供するステップとを有する情報検索方法。
【請求項２２】
上記表示点を表示するステップは、上記表示点の２次元表示アレーを表示するステップを有することを特徴とする請求項２１記載の情報検索方法。

【請求項２３】
情報アイテムの検索を実行するステップと、
上記検索によって選択された情報アイテムに対応する表示点のみをユーザディスプレイに表示するステップとを有する請求項２１記載の情報検索方法。
【請求項２４】
類似する情報アイテムのクラスタを検出し、各クラスタの情報アイテムの代表的な表現を提供するステップを有する請求項２１記載の情報検索方法。
【請求項２５】
上記表現又は各表現をユーザディスプレイに、該表現が代表している情報アイテムに対応する表示点のラベルとして表示するステップを有する請求項２１記載の情報検索方法。
【請求項２６】
上記ラベルは、単語又は単語の組であることを特徴とする請求項２５記載の情報検索方法。
【請求項２７】
上記情報アイテムは、少なくとも画像アイテムに関連し、上記検索条件によって定義された選択された情報アイテムの情報コンテンツを代表する１以上の画像を提供するステップを有することを特徴とする請求項２１記載の情報検索方法。
【請求項２８】
上記画像アイテムの組から、所定の選択条件に基づいて、該画像アイテムの組を代表する画像アイテムを選択するステップを有する請求項２７記載の情報検索方法。
【請求項２９】
該画像アイテムの組における同じプロパティの平均に最も近いプロパティを有する画像アイテムを選択するステップを有する請求項２８記載の情報検索方法。
【請求項３０】
請求項２１に示す情報検索方法を実行する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３１】
請求項３０記載のコンピュータソフトウェアを提供する提供媒体。
【請求項３２】
当該提供媒体は、記録媒体であることを特徴とする請求項３１記載の提供媒体。
【請求項３３】
当該提供媒体は、伝送媒体であることを特徴とする請求項３１記載の提供媒体。
【請求項３４】
異なる性質の情報アイテムの組内の各情報アイテムが該情報アイテムの相互の類似性に基づいてノードのアレー内の各ノードにマッピングされ、類似する情報アイテムが、該ノードのアレー内の類似する位置におけるノードにマッピングされるユーザインタフェースにおいて、
上記情報アイテムを選択するための検索条件を定義するユーザ操作手段と、
選択された情報アイテムに対応するノードのアレー内の位置を代表する表示点を表示するとともに、上記検索によって選択された情報アイテムの情報コンテンツを代表する１以上の表現を表示するグラフィックユーザインタフェースとを備えるユーザインタフェース。
【請求項３５】
上記ユーザ操作手段は、上記情報アイテムの単語に関する検索を実行する検索手段を有し、
上記検索手段及び上記グラフィックユーザインタフェースは、協働して、上記検索によって選択された情報アイテムに対応する表示点のみが表示されるように動作することを特徴とする請求項３４記載のユーザインタフェース。
【請求項３６】
上記グラフィックユーザインタフェースは、類似する情報アイテムを含む各クラスタの

情報アイテムの代表的な表現を提供することを特徴とする請求項３４記載のユーザインタフェース。
【請求項３７】
上記グラフィックユーザインタフェースは、上記表現又は各表現をユーザディスプレイに、該表現が代表している情報アイテムに対応する表示点のラベルとして表示することを特徴とする請求項３４記載のユーザインタフェース。
【請求項３８】
上記ラベルは、単語又は単語の組であることを特徴とする請求項３７記載のユーザインタフェース。
【請求項３９】
上記表現は、画像アイテムであり、該画像アイテムは、該画像アイテムによって代表される情報アイテムに対応する表示点に表示されるラベルとして適用されることを特徴とする請求項３４記載のユーザインタフェース。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、情報の格納及び検索に関する。
【背景技術】
【０００２】
キーワードに基づいて情報（例えば文書、画像、電子メール、特許情報、例えばオーディオ／ビデオコンテンツ等のインターネットコンテンツやメディアコンテンツ等）を検索するための多くの方式が確立されている。この検索の具体例としては、「グーグル（Google：商標）」、「ヤフー（Yahoo：商標）」等から提供されるインターネット検索「エンジン」も含まれ、これらにおいては、キーワードに基づく検索が実行され、検索エンジンによって適合度の順にランク付けされた検索結果のリストが得られる。
【０００３】
しかしながら、多くの場合大量のコンテンツコレクション（massive content collection）と呼ばれる大量のコンテンツを含むシステムにおいては、比較的短い「ヒット」した結果のリストが得られる有効な検索クエリを定式化することは困難となることがある。例えば、本出願を準備している時点では、グーグルにおいて、キーワード「massive document collection」を用いた検索を行うと、ヒット数は、２４３０００件となる。インターネット全体に蓄積されているコンテンツの量は、通常、時間と共に増加すると考えられるので、後に同じ検索を行った場合には、このヒット数は更に増加すると予想される。このようなヒットリストに基づいて、ヒットした全てのコンテンツに目を通す（Reviewing）ことは、現実的には殆ど不可能な程に時間がかかる作業である。
【０００４】
大量のコンテンツコレクションを活用することが困難である理由は、一般的には、以下のようなものがある。
・ユーザは、関連するコンテンツが存在することを知らない。
・ユーザは、関連するコンテンツが存在することを知っているが、そのコンテンツがどこにあるかを知らない。
・ユーザは、コンテンツが存在していることを知っているが、そのコンテンツが関連するコンテンツであることを知らない。
・ユーザは、関連するコンテンツが存在し、そのコンテンツを見つけ出す手法も知っているが、コンテンツを見つけ出すには時間がかかる。
【０００５】
「大量のドキュメントコレクションの自己組織化（Self Organisation 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）、Ｖｏｌ１１、Ｎｏ．３、２０００年５月、５７４～５８５頁には、所謂「自己組織化マップ（self-organis

ing maps：以下、ＳＯＭという。）」を用いた技術が開示されている。ＳＯＭは、各ドキュメントのプロパティを表す「特徴ベクトル（feature vector）」をＳＯＭのノードにマッピングする、所謂教師なき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】

コホネン他の論文に開示されている手法では、まず、文書テキストを前処理し、前処理された各文書から特徴ベクトルが導かれる。この特徴ベクトルは、大きな単語の辞書における各単語の出現頻度を示すヒストグラムの形式で表してもよい。ヒストグラム内の各データ値（すなわち、辞書中のそれぞれの単語の各出現頻度）は、辞書内の候補となる単語の総数（この論文に記述されている具体例では、４３２２２個）をｎとして、ｎ値のベクトルにおける値となる。例えば、ある単語の適合度を増加させ又は特異性を強調するために、ｎ個のベクトル値に重み付けを適用してもよい。

【０００７】

ｎ値のベクトルは、次元がより小さいベクトル（すなわちｎに比べてかなり小さな値の数ｍ（論文内の具体例では５００）の値を有するベクトル）にマッピングされる。これは、乱数配列により構成されるｎ×ｍの「プロジェクション行列（projection matrix）」をベクトルに乗算することによって実現される。この手法により、次元が小さくされたどの２つのベクトルをとっても、それぞれ対応する２つの入力ベクトルと略同じ内積を有する、次元がより小さいベクトルが生成される。このベクトルマッピング処理は、論文「ランダムマッピングによる次元縮退：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）、ＰｒｏｃＩＪＣＮＮ、４１３～４１８頁、１９９８年に記載されている。

【０００８】

次元が低減されたベクトルは、各ベクトルを１つの「モデル（他のベクトル）」に乗算する処理によって、ＳＯＭ上のノード（ニューロンともいう）にマッピングされる。これらのモデルは、学習プロセスによって生成される。この学習プロセスでは、モデルを相互の類似性に基づいてＳＯＭ上に順序付けし、ＳＯＭは、通常、ノードの２次元グリッドとして表される。これは、膨大な演算を必要とする処理であり、コホネン他は、この処理のために、７００万の文書に満たない文書データベースに対して、８００ＭＢのメモリを有する６プロセッサのコンピュータで６週間を費やしている。最終的には、ＳＯＭを構成するノードのグリッドが表示され、ユーザはマップの領域をズームしてノードを選択することができ、これにより、ユーザインタフェースは、そのノードにリンクされた文書が含まれるインターネットページへのリンクを提供する。

【０００９】

コホネンの論文には、マップ内の情報を検索するためのガイダンスとなるラベルを用いる手法が開示されている。これらのラベルは、キーワードを選択することによって自動的に生成される。この手法は、ケー・ラグス（K.Lagus）及びエス・ラスキ（S.Laski）著、テキスト文書マップを特徴付けるキーワード選択法（Keyword selection method for characterising text document maps）、Ｐｒｏｃ ＩＣＡＮＮ９９、Ｎｉｎｔｈ Ｉｎｔ．Ｃｏｎｆ．Ａｒｔｉｆｉｃｉａｌ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋｓ，ｖｏｌ．１、１９９９年、ｐｐ．３７１～３７６に開示されている。

【発明の開示】

【課題を解決するための手段】

【００１０】

本発明の一側面として、本発明は、ユーザによる検索に関係する情報アイテムの表現を提供し、これによりユーザを補助する。

【００１１】

本発明に係る情報検索装置は、異なる性質の情報アイテムの組内の各情報アイテムが情報アイテムの相互の類似性に基づいてノードのアレー内の各ノードにマッピングされ、類

似する情報アイテムが、ノードのアレー内の類似する位置におけるノードにマッピングされる情報検索装置において、情報アイテムを選択するための検索条件を定義するユーザ操作手段と、ノードのアレー内で、選択された情報アイテムに対応する位置を検出する検出手段と、選択された情報アイテムに対応するノードのアレー内の位置を代表する表示点を表示するグラフィックユーザインタフェースと、検索条件によって定義された選択された情報アイテムに応じて、選択された情報アイテムの情報コンテンツを代表する１以上の表現を提供するプロセッサとを備える。

【００１２】

これにより、情報検索装置は、ユーザによって定義された検索に関連する表示点と、検索によって定義された情報アイテムを代表する表現の両方を表示し、ユーザを補助する。

【００１３】

一具体例においては、ユーザ操作手段は、情報アイテムの単語に関する検索を実行する検索手段を有し、検索手段及びグラフィックユーザインタフェースは、協働して、検索によって選択された情報アイテムに対応する表示点のみが表示されるように動作する。

【００１４】

一具体例においては、プロセッサは、表現又は各表現をユーザディスプレイに、表現が代表している情報アイテムに対応する表示点のラベルとして表示する。一具体例においては、ラベルは、単語又は単語の組である。

【００１５】

他の具体例においては、情報アイテムは、少なくとも画像アイテムに関連し、プロセッサは、検索条件によって定義された選択された情報アイテムの情報コンテンツを代表する１以上の画像を提供する。好ましくは、プロセッサは、画像アイテムの組から、所定の選択条件に基づいて、画像アイテムの組を代表する画像アイテムを選択する。更に、プロセッサは、画像アイテムの組における同じプロパティの平均に最も近いプロパティを有する画像アイテムを選択してもよい。選択された画像は、この画像によって代表される情報アイテムに対応する表示点のラベルとして適用できる。

【００１６】

グラフィックユーザインタフェースは、好ましくは、表示点の２次元表示アレーを表示する。但し、表示される表示アレーは、１次元表示アレーであっても、３次元表示アレーであってもよい。３次元表示アレーの第３の次元は、時間であってもよい。

【００１７】

また、本発明に係る情報検索方法は、異なる性質の情報アイテムの組内の各情報アイテムが情報アイテムの相互の類似性に基づいてノードのアレー内の各ノードにマッピングされ、類似する情報アイテムが、ノードのアレー内の類似する位置におけるノードにマッピングされる情報検索方法において、情報アイテムを選択するための検索条件を定義するステップと、ノードのアレー内で、選択された情報アイテムに対応する位置を検出するステップと、選択された情報アイテムに対応するノードのアレー内の位置を代表する表示点を表示するステップと、検索条件によって定義された選択された情報アイテムに応じて、選択された情報アイテムの情報コンテンツを代表する１以上の表現を提供するステップとを有する。

【００１８】

本発明の更なる側面及び特徴は、添付の請求の範囲において定義されている。

【発明を実施するための最良の形態】

【００１９】

図１は、情報格納及び検索システム（information storage and retrieval system）を概略的に示す図であり、この情報格納及び検索システムは、汎用コンピュータ１０によって実現されており、汎用コンピュータ１０は、プログラム及びデータを格納するディスク記憶装置３０を含むプロセッサユニット２０と、イーサネット（Ethernet network）（登録商標）又はインターネット等のネットワーク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、例えば陰極線管等の表示装置６０と、キーボード７０及びマウス８０

等のユーザ入力装置とを備える。このシステムは、プログラム制御の下に動作し、このプログラムは、ディスク記憶装置３０に格納され、及び、例えばネットワーク５０又はリムーバブルディスク（図示せず）から、若しくはディスク記憶装置３０にプレインストールされて提供される。

【００２０】

情報格納及び検索システムは、２つの通常の動作モードで動作する。第１のモードでは、第１の情報アイテム（例えば、テキスト情報アイテム）の組がディスク記憶装置３０に又はネットワーク５０を介して接続されたネットワークディスクドライブに集められ（assembled）、検索動作のために予めソートされるとともに、インデックスが付けられる。第２のモードの動作では、ソートされ、インデックスが付されたデータに対して実際の検索が行われる。

【００２１】

この具体例は、多くの種類の情報アイテムに適用できる。適用される情報アイテムとしては、以下に限定されるものではないが、例えば、特許情報、ビデオマテリアル、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、営業報告書、オーディオマテリアル、グラフィクス及びクリップアート、写真、これらの組合せ又は混合された情報等がある。ここでは、テキスト情報アイテム、又は少なくともテキストコンテンツを含む又は関連した情報について説明する。例えば、オーディオ及び／又はビデオマテリアル等の放送コンテンツは、そのマテリアルをテキスト用語（textual terms）で定義する関連したメタデータを有することができる。テキスト情報アイテムは、非テキスト情報アイテムに関連付けられ、又はリンクされていてもよい。例えば、オーディオ及び／又はビデオマテリアルは、そのマテリアルを言葉により定義するテキスト情報であり「メタデータ」に関連付けられていてもよい。

【００２２】

情報アイテムは、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、情報アイテムは、アイテムに索引を付けてより容易に検索できるようにデータベース構造の一部として格納されるが、このことは必ずしも必要なわけではない。情報及びアイテムが一旦格納されると、検索のために格納された情報を整理する、図２に示す処理が行われる。

【００２３】

なお、インデックスが付されたデータは、必ずしもローカルのディスクドライブであるディスク記憶装置３０に格納する必要はない。情報アイテムは、ネットワーク５０を介して汎用コンピュータ１０に接続されたリモートのドライブに格納することもできる。これに代えて、情報は、分散方式で、例えばインターネット全域の様々サイトに格納することもできる。情報が異なるインターネットサイト又はネットワークサイトに格納されている場合、リモートの情報への「リンク」（例えば、汎用リソース識別子：universal resource identifier：ＵＲＩ）を、例えば関連した要約（summary）、概要（abstract）、又はこのリンク関するメタデータと共に格納する第２のレベルの情報格納を用いることができる。リモートに格納された情報は、ユーザが（例えば、後述する結果リスト２６０から）関連するリンクを選択しない限り、アクセスされないが、以下に記述する技術的な説明においては、リモートに格納された情報、又は要約、概要、メタデータ、若しくはリンク／ＵＲＩも「情報アイテム」とみなすことができる。

【００２４】

換言すると、「情報アイテム」とは、（後述のように）特徴ベクトルを導出及び処理して、ＳＯＭへのマッピングを行うことができるアイテムと形式的に定義することができる。後述する結果リスト２６０内のデータは、（情報アイテムがローカルに格納され、容易に表示できる程に十分短い場合）ユーザが検索する実際の情報アイテム自体であってもよく、及び／又は、例えばメタデータ、ＵＲＩ、概要、キーワードのセット、代表的なキースタンプ画像等のうちの１つ以上である、情報アイテムを表現及び／又は指示するデータであってもよい。これは、必ずではないが、多くの場合、１組のアイテムを表すデータを

一覧化する「リスト」動作に特有なものである。

【００２５】

更なる具体例においては、情報アイテムは、例えば研究チームや法律事務所等のネットワーク化されたワークグループ全体に格納してもよい。また、ハイブリッド法は、ローカルに格納された情報アイテム及び／又はローカルエリアネットワーク全体に格納された情報アイテム及び／又はワイドエリアネットワーク全体に格納された情報アイテムを含んでいてもよい。この場合、本発明に基づくシステムは、他者による類似した業績を検索するのに有用であり、例えば、後述するように、複数の国に亘る大規模な研究開発組織において、類似した研究業績をＳＯＭにおける類似した出力ノードにマッピングすることができる。または、新たなテレビジョン番組を企画する場合、本発明に基づく技術を用いて、類似するコンテンツを有する以前の番組を検索することにより、企画中の番組の独創性を確認することもできる。

【００２６】

図１に示すシステム１０は、インデックスが付された情報アイテムを用いる可能なシステムの一例にすぎない。初期の（インデックス付与）段階は、相当の演算能力を有するコンピュータ、可能性としては携帯型ではないコンピュータによって実行されるが、後の段階では、例えば、携帯情報端末（personal digital assistant：ＰＤＡ）（ディスプレイとユーザ入力装置とを備えた、通常片手で持てる程度の寸法のデータ処理装置を指す）、ラップトップコンピュータ等の携帯型コンピュータによって、情報のアクセスを実行してもよく、更に携帯電話、ビデオ編集装置、ビデオカメラ等の機器で行ってもよい。包括的に言えば、情報にアクセスする動作段階では、ディスプレイを有するあらゆる機器を用いることができる。

【００２７】

この処理は、特定の数の情報アイテムに対する処理に限定されるものではない。

【００２８】

情報アイテムの自己組織化マップ（ＳＯＭ）表現を生成する処理について、図２～図６を用いて説明する。図２は、ＳＯＭマッピング処理の前に行われる、所謂「特徴抽出（feature extraction）」を説明するフローチャートである。

【００２９】

特徴抽出は、生のデータ（raw data）を抽象表現（abstract representation）に変換する処理である。そして、この抽象表現は、パターン分類、クラスタリング、認識等の処理に用いることができる。この処理により、所謂「特徴ベクトル」が生成される。特徴ベクトルは、文書内で使用される用語の頻度の抽象表現である。

【００３０】

特徴ベクトルを生成することにより、ビジュアライゼーション（visualisation）を形成する処理は、以下のステップを含む。

・用語の「文書データベース辞書（document database dictionary）」を作成する。
・「文書データベース辞書」に基づいて、個々の各文書について、「用語頻度ヒストグラム（term frequency histogram）」を作成する。
・ランダムマッピングを用いて、「用語頻度ヒストグラム」の次元を低減する。
・情報空間の２次元ビジュアライゼーションを作成する。

【００３１】

以下、これらのステップをより詳細に説明する。各文書（情報アイテム）１００は、順番に開かれる。ステップ１１０において、文書から全ての「ストップワード（stop word）」を取り除く。ストップワードとは、例えば「ａ」、「ｔｈｅ」、「ｈｏｗｅｖｅｒ」、「ａｂｏｕｔ」、「ａｎｄ」等、予め準備されたリストに挙げられている極めてありふれた単語である。これらの単語は、極めてありふれており、平均的に、十分な長さを有するあらゆる文書において、同様の頻度で出現する可能性が高い。したがって、これらの単語は、特定の文書のコンテンツを特徴付ける目的には殆ど貢献せず、このため、これらのストップワードを削除する。

【００３２】
ストップワードを削除した後、ステップ１２０において、残っている単語の語幹を調べる。これは、単語の活用形に対する共通の原形を見出す処理を含む。例えば「ｔｈｒｏｗｅｒ」、「ｔｈｒｏｗｓ」、「ｔｈｒｏｗｉｎｇ」は、共通の語幹又は原形「ｔｈｒｏｗ」を有している。
【００３３】
文書内に出現する（ストップワードを除く）単語の語幹又は原形の辞書は、保守される。すなわち、新たな単語が出現すると、この単語は辞書に追加され、文書コレクション（情報アイテム）全体の中でその単語が出現した回数も記録される。
【００３４】
この結果、その組内の全ての文書内で使用された用語が、その出現頻度とともに登録されたリストが作成される。出現頻度が高すぎる又は低すぎる単語は除外（discount）され、すなわち、これらの単語は辞書から削除され、後の分析には使用しない。出現頻度が低すぎる単語は、綴り間違いのある単語や、造語（made up）、又はその文書の組によって表されている分野に関係がない単語である可能性がある。一方、出現頻度が高すぎる単語は、その組内の文書を区別するために用いるには不適切である。例えば、「Ｎｅｗｓ」という単語が試験的な放送関連の文書の組全体の内、３分の１の文書において使用されており、一方、「ｆｏｏｔｂａｌｌ」という単語がこの試験的な文書の組全体の内、約２％しか使用されていないとする。この場合、「ｆｏｏｔｂａｌｌ」は、「Ｎｅｗｓ」より、文書のコンテンツを特徴付けるために適切な用語であると考えることができる。一方、「ｆｏｔｔｂａｌｌ（ｆｏｏｔｂａｌｌの綴り間違い）」が文書の組全体の内、１回しか出現していないとすると、この単語は、出現頻度が低すぎるとして除外される。これらの単語は、平均出現頻度に対して２標準偏差小さい出現頻度を有する単語、又は平均出現頻度に対して２標準偏差大きな出現頻度を有する単語と定義してもよい。
【００３５】
そして、ステップ１３０において、特徴ベクトルを生成する。
【００３６】
特徴ベクトルを生成するために、組内の各文書に対して用語頻度ヒストグラムを作成する。用語頻度ヒストグラムは、辞書（文書の組に関連付けられている）内に存在する単語が、個々の文書内で出現する頻度を計数することによって作成される。辞書内の大半の用語は、単一の文書内には存在せず、したがって、これらの用語の出現頻度は０である。図３ａ及び図３ｂは、２つの異なる文書についての用語頻度ヒストグラムの具体例を示している。
【００３７】
この具体例から、用語頻度ヒストグラムが、どのようにして文書のコンテンツを特徴付けているかがわかる。この具体例の場合、文書１は、文書２に比べて、「ＭＰＥＧ」、「Ｖｉｄｅｏ」という用語の出現頻度が高く、一方、文書２は、文書１に比べて、「ＭｅｔａＤａｔａ」という用語の出現頻度が高い。用語頻度ヒストグラム内の多くの項目（entry）は、対応する単語がその文書内に存在しないため、値が０となっている。
【００３８】
現実的には、実際の用語頻度ヒストグラムは、この具体例に示す用語頻度ヒストグラムより更に多数の用語を含んでいる。多くの場合、用語頻度ヒストグラムは、５００００個以上の異なる用語の頻度をプロットし、したがって、用語頻度ヒストグラムの次元は、５００００以上となる。この用語頻度ヒストグラムの次元は、ＳＯＭ情報空間の構築に使用するためには、かなり低減する必要がある。
【００３９】
用語頻度ヒストグラム内の各項目は、その文書を表現する特徴ベクトルにおける対応する値として使用される。この処理の結果、文書コレクション内の各文書について、辞書によって特定された全ての用語の頻度を含む（５００００×１）ベクトルが生成される。このベクトルでは、殆どの値が０であり、更にこの他の値の大半が１程度の非常に低い値を

取るため、このベクトルは、「スパース（sparse）」ベクトルと呼ばれる。

【００４０】

ステップ１４０において、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次元を小さくする。用語頻度ヒストグラムの次元を小さくする手法としては、例えば以下のような２つの手法がある。

１）ランダムマッピング法：用語頻度ヒストグラムに乱数の行列を乗算する手法。これは、計算機的には安い処理である。

２）潜在的意味抽出法（Latent Semantic Indexing）：文書内で同時に出現する可能性が高い用語のグループを調べることによって用語頻度ヒストグラムの次元を小さくする手法。これにより得られた単語のグループは、単一のパラメータにすることができる。これは、計算機的には高価な処理である。

【００４１】

ここで説明する具体例においては、用語頻度ヒストグラムの次元を低減するために、上述したカスキ（Kaski）の論文に開示されているランダムマッピング法を用いる。ランダムマッピング法では、用語頻度ヒストグラムに乱数の行列を乗算することによって、用語頻度ヒストグラムの次元を低減する。

【００４２】

上述のように、「生の」特徴ベクトル（図４ａに示す）は、５００００個もの値を有するスパースベクトル（sparse vector）である。このベクトルは、約２００個の値を有するサイズ（図４ｂに示す）にまで縮小されるが、それでもなお、特徴ベクトルの相対的特徴を保持しており、すなわち、同様に処理された他の特徴ベクトルに対する相対角（ベクトル内積）等の関係を保っている。これは、特定の次元の直交ベクトルの数が限られていても、略直交するベクトルの数が非常に多いので、有効である。

【００４３】

実際、ベクトルの次元が増加すると、ランダムに生成されたあらゆるベクトルの組が、互いに略直交するようになる。このような性質は、ベクトルの相対的な向きは、そのベクトルに乱数の行列を乗算しても保たれることを意味する。この性質は、ランダムマッピングの前後のベクトルの内積の類似性によって示される。

【００４４】

スパースベクトルの値を５００００個から２００個に低減しても、これらのベクトルの相対的な類似性が保たれることが経験的に確認されている。このマッピングは、完全ではないが、文書のコンテンツを簡潔に特徴付ける目的としては十分である。

【００４５】

文書コレクションについて、特徴ベクトルを作成し、文書コレクションの情報空間を定義した後、ステップ１５０において、意味論的なマップを生成するために、特徴ベクトルを２次元ＳＯＭに射影する。以下、図５を参照して、コホネン（Kohonen）の自己組織化マップを用いて、特徴ベクトルをクラスタリングすることによって２次元にマッピングする処理を説明する。

【００４６】

コホネンの自己組織化マップは、各文書に対して作成された特徴ベクトルをクラスタ化し、組織化するために使用される。

【００４７】

自己組織化マップは、入力ノード１７０と、２次元平面１８５として示されるノードの２次元アレー又はグリッド内の出力ノード１８０とを含む。自己組織化マップをトレーニングするために使用される特徴ベクトル内の値と同じ数の入力ノード１７０が存在する。自己組織化マップ上の各出力ノード１８０は、重み付けされたコネクション（weighted connection）１９０（１つのコネクションにつき、１つの重み）によって入力ノード１７０に関連付けられている。

【００４８】

まず、各重みは、ランダムな値に設定され、繰り返し処理により、これらの重みが「ト

レーニング」される。自己組織化マップは、このマップの入力ノード１７０に各特徴ベクトルを供給することによってトレーニングされる。各出力ノードの入力ベクトルと重みとの間のユークリッド距離を算出することにより、「最も近い」出力ノードが算出される。

【００４９】

最も近い出力ノードは、「勝者（winner）」と呼ばれ、このノードの重みは、このノードが入力ベクトルにより近づくように重みの値を若干変更することによってトレーニングされる。勝者のノードに加えて、この勝者のノードに隣接するノードもトレーニングされ、入力ベクトルに若干近づけられる。

【００５０】

単一のノードを重み付けするだけではなく、マップ上のノードの領域を重み付けするこの処理によって、マップは、一旦トレーニングされれば、ノードの２次元マップ内に入力空間のトポロジの大部分を保存することができる。

【００５１】

マップが一旦トレーニングされると、各文書をマップにマッピングし、どの出力ノードが、その文書について入力された特徴ベクトルに最も近いかを確認することができる。重みが特徴ベクトルと等しいことは希であり、特徴ベクトルとこの特徴ベクトルに最も近いノードとの間のユークリッド距離は、「量子化誤差」として知られる。

【００５２】

各文書についての特徴ベクトルをマップに与えて、この特徴ベクトルがどこに存在するかを確かめることにより、各文書について、（ｘ，ｙ）座標で表されるマップ上の位置が得られる。これらの（ｘ，ｙ）座標で表されるマップ上の位置は、文書ＩＤとともにルックアップテーブルで照合することにより、文書間の関係を視覚化するために使用することができる。

【００５３】

更に、ステップ１６０においてディザ成分（dither component）を加算する。ディザ成分については、図６を用いて後に説明する。

【００５４】

上述の手法では、２つの同一の又は実質的に同一の情報アイテムが、ＳＯＭのノードアレー内の同じノードにマッピングされてしまうことがある。これは、データを取り扱う上では問題にならないが、表示画面（後に説明する）上にデータを視覚化して表示する際に問題となる。特に、データを表示画面上に視覚化して表示する場合、多数の非常に似ているアイテムを特定のノードにおいて単一のアイテム上で区別できることが有用であると認められる。したがって、各情報アイテムがマッピングされるノード位置にディザ成分を加算する。ディザ成分は、ノード間隔の±１／２をランダムに加算するものである。例えば、図６に示すように、マッピング処理により出力ノード２００が選択された情報アイテムにはディザ成分が加算され、この情報アイテムは、実際には、図６に点線枠で示す領域２１０内の任意のノード位置にマッピングされる。

【００５５】

このように、情報アイテムは、図６の面上において、ＳＯＭ処理の「出力ノード」とは異なるノード位置にマッピングされることもある。

【００５６】

いかなる時点においても、ＳＯＭに新たな情報アイテムを追加することができる。この処理は、上述したステップ（すなわち、ステップ１１０～ステップ１４０）を実行し、これにより得られた縮小された特徴ベクトルを「先にトレーニングされている」ＳＯＭモデル、すなわち、マップの自己組織化処理によって得られたＳＯＭモデルの組に適用することによって行われる。したがって、新たに追加された情報アイテムについては、マップは、通常、「再トレーニング」されない。これに代えて、変更されていない全てのＳＯＭモデルとともに、ステップ１５０及びステップ１６０を実行する。新たな情報アイテムが追加される毎にＳＯＭを再トレーニングすると、演算コストが高くなり、また、ユーザは、マップ内においてよくアクセスする情報アイテムの相対的位置に慣れてくるので、再トレ

ーニングは、ユーザを困惑させる場合もある。

【００５７】

しかしながら、再トレーニング処理を行う方が望ましい場合もある。例えば、最初にＳＯＭが作成された後に、新たな用語（例えば、ニュース又は新たな技術分野における新たな用語）が辞書に追加されたとすると、これらの新たな用語は、既存の出力ノードの組には、うまくマッピングできない場合がある。これは、新たに受け取った情報アイテムの既存のＳＯＭへのマッピングの際に検出される所謂「量子化誤差」の増大によって検出される。この具体例では、量子化誤差を誤差量の閾値と比較する。量子化誤差が閾値を超えている場合、（ａ）ＳＯＭに元から含まれる情報アイテム及びＳＯＭの作成後に加えられたあらゆる情報アイテムを用いてＳＯＭを自動的に再トレーニングし、又は（ｂ）適切なときに再トレーニング処理を開始するようユーザに促す。再トレーニング処理は、関連する全ての情報アイテムの特徴ベクトルを用いて、ステップ１５０及びステップ１６０を全ての特徴アイテムを用いて再び実行することによって行われる。

【００５８】

図７は、検索作業のために、ＳＯＭに格納されたデータを視覚的に表示する表示装置６０に表示される画面を示している。この画面には、検索文字列入力枠２５０、結果リスト２６０、ＳＯＭ表示領域２７０が表示されている。

【００５９】

実際の作業では、表示領域２７０は、最初はブランクになっている。ユーザは、検索文字列入力枠２５０にキーワード検索文字列を入力する。次にユーザは、キーボード７０のエンターボタンを押圧操作し、又はマウス８０を用いて画面上の検索開始ボタンを選択することにより検索を開始する。キーワード検索文字列入力枠２５０に入力されたキーワードは、標準的なキーワード検索技術を用いて、データベース内の情報アイテムと比較される。この検索の結果は、結果リスト２６０内の各エントリ２８０として表示される。ＳＯＭ表示領域２７０は、検索の結果として得られたアイテムに対応する表示点を表示する。

【００６０】

ＳＯＭ表現を作成するために用いたソート処理により、互いに類似する情報アイテムがＳＯＭ内でグループ化されるため、文字列の検索結果は、例えばクラスタ２９０のように、クラスタ内に集中しやすい。ここで、ＳＯＭ表示領域２７０内の各点は、結果リスト２６０内の関連する結果の１つに対応しており、ＳＯＭ表示領域２７０内の各点が表示されている位置は、ノードアレー内におけるこれらのノードのアレー位置に対応している。

【００６１】

ヒット（結果リスト内の結果）の数を少なくする手法を図８を用いて説明する。この具体例では、ユーザは、マウス８０を用いて、興味があるノードに対応する表示点の組の周辺を長方形状のボックス（境界線）３００で囲む。結果リスト２６０内には、ボックス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、ユーザの求めるものと違うことが判明した場合、ユーザは、新たなボックスで他の表示点の組を囲む。

【００６２】

なお、結果リスト２６０は、検索文字列入力枠２５０に入力されたキーワードによる検索条件を満たし、且つ、表示点がボックス３００内にある結果に対応するエントリを表示している。ボックス３００は、ノードアレー内に含まれる（populated）ノードに対応する他の表示点を囲むこともできるが、これらのノードが検索条件を満たしていない場合、これらのノードは表示されず、したがって、結果リスト２６０内の結果には含まれない。

【００６３】

本発明の具体例を図９を用いて説明する。まず、図９におけるステップ９２０において、ラベルを有していない（この点がコホネンのＳＯＭと異なる）自己組織化マップＳＯＭを生成する。ユーザは、自己組織化マップ内を検索するためのガイダンスとなるラベルを要求する。本発明の具体例においては、ユーザの特定の要求（ニーズ）に合わせて、ラベルが自動的に生成される。ユーザは、図７及び／又は図８を用いて説明したように、検索の結果を示すリストを作成する。ラベルは、この結果に基づいて自動的且つ動的に生成さ

れ、表示領域２７０の表示点のクラスタ（clusters）を区別するために用いられる。

【００６４】

ステップ９２１において、ユーザは、検索操作を行う。検索操作は、この具体例では、図７を用いて説明したように、キーワード検索文字列を用いて行われる。この検索の結果得られた例えば文書等の多くのアイテムは、ステップ９２２において、マップ上の個々の（ｘ，ｙ）座標によって示される。ステップ９２１及びステップ９２２に続いて、ステップ９２３において、Ｋ－ｍｅａｎｓ法に基づくクラスタリング及び枝刈り（"k-means clustering and pruning"）が実行される。ステップ９２３における処理は、ステップ９２１における検索操作の結果として得られたアイテムの組に対応するマップ上の（ｘ，ｙ）座標に対して実行される。このステップ９２３においては、周知のＫ－ｍｅａｎｓクラスタリング法により、類似するアイテムを示すアイテムのグループが識別される。この処理は、検索結果のアイテムに対応する隣接する（ｘ，ｙ）座標のサブクラスタ（sub-cluster）が同じメインクラスタ（main cluster）の一部であるか否かを判定する処理（ステップ９２３における枝刈り（pruning）処理）を含む。２つのサブクラスタの中心間の距離が閾値より小さい場合、２つのサブクラスタは、同じメインクラスタの一部であるとみなされる。この枝刈り処理は、ステップ９２４において、クラスタ処理が安定する（stable）まで、周知の手法に基づいて繰り返し行われる。

Ｐ１５

ステップ９２６において、クラスタ内のアイテム及びキーワードが特定される。各アイテムについて、上位２０個のキーワード及びその出現頻度が予め算出される。クラスタラベルを算出する処理は、各アイテムについて予め導出された全てのキーワードの出現頻度の合計を算出し、最も出現頻度が高いキーワードを選択する処理を含む。選択されたキーワードは、クラスタのラベルとして適用される。このように、キーワードを予め算出することにより、クラスタのラベルを作成するために必要な時間を著しく短縮することができる。

【００６５】

クラスタ内に１つのアイテムしかない場合、そのアイテムの「タイトル」をラベルとして用いる。

【００６６】

このように、クラスタの決定及びクラスタに対するラベルの付与は、ユーザによる検索に基づいて行われる。

【００６７】

上述したサブクラスタの中心間の距離に関する閾値は、ユーザが選択してもよい。これに代えて、閾値を「拡大率（zoom）」に基づいて定めてもよい。ＳＯＭの一部のスケールを拡大して捉えてもよい。マップを小さなスケールで捉えると、隣接するクラスタは、より大きな１つのクラスタに見えるが、マップをより大きなスケールで捉えると、これらのクラスタは、それぞれ独立して見える。したがって、閾値は、拡大率に基づいて決定される。

【００６８】

アイテムは、文書でも画像でも映像でもよい。アイテムが画像や映像等のメディアアイテムである場合、キーワードは、画像や映像に関連するメタデータにおいて用いられているキーワードであってもよい。タイトルもメタデータの一例である。メタデータのこの他の例については、当分野について周知である。

【００６９】

更に、映像等のメディアアイテムを表すために、代表キースタンプ（Representative Key Stamp：ＲＫＳ）を作成する手法も知られている。本発明の他の具体例においては、図７、又は図７及び図８、並びに上述したステップ９２０～ステップ９２４に基づいて検索を実行し、表示領域２７０における表示点に対応する検索結果のリストを作成し、安定したクラスタを特定する。そして、ステップ９２８～ステップ９３１として示すように、クラスタ２９０のラベルとして、ＲＫＳが選択される。例えば、本発明の一具体例において

は、ステップ９２８において、クラスタ２０のメディアアイテムのＲＫＳが選択され、ステップ９２９において、これらのＲＫＳの全てのカラーヒストグラムが算出される。ステップ９３０において、カラーヒストグラムの平均値が算出され、ステップ９３１において、この平均値に最も近いカラーヒストグラムを有するＲＫＳが選択される。このように平均に近いＲＫＳがクラスタを代表するＲＫＳとなる。これにより、クラスタに対して、そのクラスタを代表するＲＫＳをラベルとして付すことができる。個々のＲＫＳは、そのクラスタ又は各クラスタ内の個々のメディアアイテムを代表している。

【００７０】

検索結果のリスト全体を代表するＲＫＳと、各クラスタを代表するＲＫＳと、検索リストの結果における個々のアイテムを代表する個々のＲＫＳとからなるＲＫＳの階層構造を生成してもよい。

【００７１】

単一のメディアアイテムのＲＫＳは、周知のいかなる手法で選択してもよい。以下に、ＲＫＳを選択する手法の一例を説明する。

【００７２】

１）単一のメディアアイテムのＲＫＳを判定するために、メディアアイテム内の各フレームについてメトリックを算出し、フレーム画像内に顔があるか、その顔が誰の顔であるか、及び音声チャンネルに人間の声が含まれているかを判定する。そして、メディアアイテム内に最も多く登場する個人の顔を含み、更にその個人の声を含むフレームを、代表フレームとして選択することができる。例えば、「パーキンソン（Parkinson）：英国で有名なテレビジョン番組司会者」は、ゲストの登場するフレームより、パーキンソン自身がカメラに向かって話しているフレームを好む。

【００７３】

メディアアイテムのグループのＲＫＳ（例えば、リスト又はクラスタ）は、周知のいかなる手法で選択してもよい。この幾つかの具体例を以下に示す。

【００７４】

２）メディアアイテムのグループのＲＫＳを決定するために、全てのメディアアイテム内の全てのフレームに対して、１）と同様の処理を行ってもよい。例えば、５分のアイテムが１０個ある場合、これを単一の５０分のアイテムとみなして、１）と同様の処理を行ってもよい。

【００７５】

３）これに代えて、メディアアイテムのグループについて、１）と同様の手法で、各アイテムのＲＫＳを判定することもできる。次に、検索結果リストが、検索問い合わせ文字列への関連性に基づいて各検索結果をランク付けする関係メトリック（relevancy metric）を含んでいると仮定する。これにより、メディアアイテムのグループにおいて最も関連性が高いアイテムのＲＫＳをメディアアイテムのグループのＲＫＳとすることができる。

【００７６】

４）上述した３）の手法の拡張例として、任意の手法を用いて検索結果をランク付けし、検索結果のクラスタのうち、最も高いランクが高いアイテムを判定し、そのアイテムのＲＫＳによってグループを代表させてもよい。

【００７７】

５）ＲＫＳを決定する単純な手法として、グループ内の全てのメディアアイテムを構成する全てのフレームの組から、自動的にランダムにフレームを選択し、又は、単純に、アイテムのグループのうち、最初に処理することになったフレームを自動的に選択してもよく、ユーザが最も適当と考えるフレームを代表的フレームとして選択してもよい。

【００７８】

図１０は、ビデオ取得及び／又は処理装置（video acquisition and/or processing apparatus）の具体例として、カムコーダ５００の構成を示す図である。カムコーダ５００は、撮像装置５１０と、撮像装置５１０に取り付けられたレンズ５２０と、データ／信号プロセッサ５３０と、テープ状記録媒体５４０と、ディスク状又はランダムアクセス記録

媒体５５０と、ユーザ操作子５６０と、表示装置５７０と、表示装置５７０に取り付けられた接眼レンズ５８０とを備える。周知のカムコーダ又は他の代替物におけるこの他の特徴（例えば、異なる記録媒体又は異なる表示画面構成）は、当業者にとって明らかである。実際の使用時には、捕捉されたビデオマテリアルに関連するメタデータがディスク状又はランダムアクセス記録媒体５５０に保存され、この保存されたデータに関連するＳＯＭが表示装置５７０に表示され、ユーザ操作子５６０を用いて、上述のように制御される。

【００７９】

図１１は、携帯可能なデータ表示装置の具体例として、携帯情報端末（personal digital assistant：以下、ＰＤＡという。）６００の構成を示す図である。ＰＤＡ６００は、表示領域６２０及びユーザ操作子として機能するタッチセンサ領域６３０を有する表示画面６１０と、データ処理部及びデータ記録部（図示せず）とを備える。ここでも、この分野における当業者は、変形例を容易に想到できる。ＰＤＡ６００は、図１を用いて説明したシステムと同様に使用することができる。

【００８０】

添付の図面を参照して本発明を詳細に説明したが、本発明は上述の実施の形態の詳細に限定されるものではなく、当業者は、添付の請求の範囲に定義された本発明の思想及び範囲から逸脱することなく、上述の実施の形態を様々に変更及び修正することができる。

【図面の簡単な説明】

【００８１】

【図１】情報保存及び検索システムの構成を示す図である。

【図２】自己組織化マップ（ＳＯＭ）の作成の手順を説明する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示す図である。

【図４】ａは、生の特徴ベクトルを示し、ｂは、縮小された特徴ベクトルを示す図である。

【図５】ＳＯＭの構造を示す図である。

【図６】ディザ処理を説明する図である。

【図７】ＳＯＭによって表現された情報にアクセスするためのユーザインタフェースを提供する表示画面を示す図である。

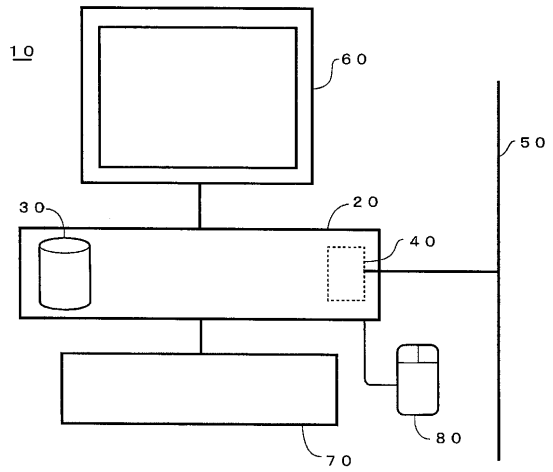
【図８】ＳＯＭによって表現された情報にアクセスするためのユーザインタフェースを提供する表示画面を示す図である。

【図９】ラベルの自動的な生成処理を説明するフローチャートである。

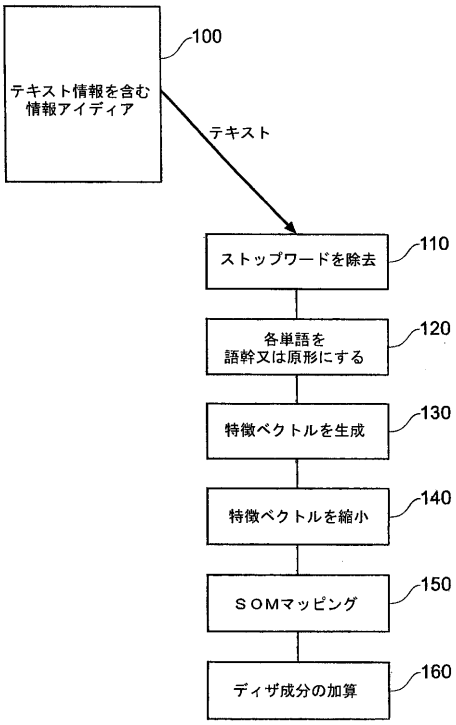
【図１０】ビデオ取得及び／又は処理装置の具体例として、カムコーダの構成を示す図である。

【図１１】携帯型データ処理装置の具体例として、携帯情報端末の構成を示す図である。

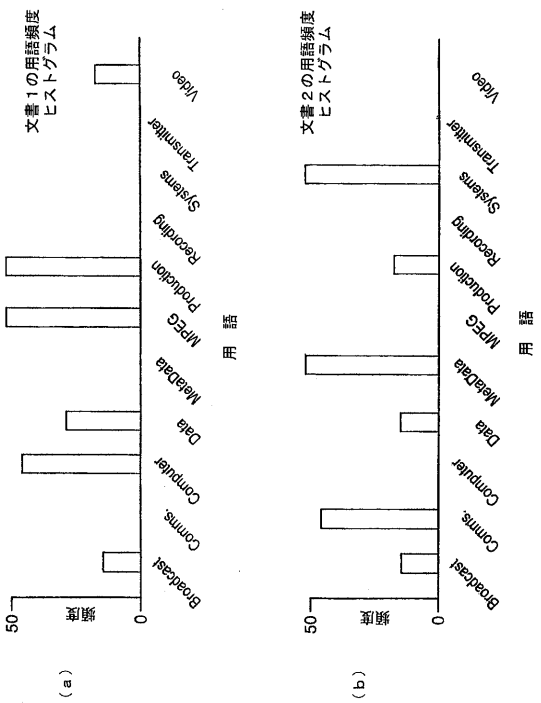
【図１】



【図２】



【図３】

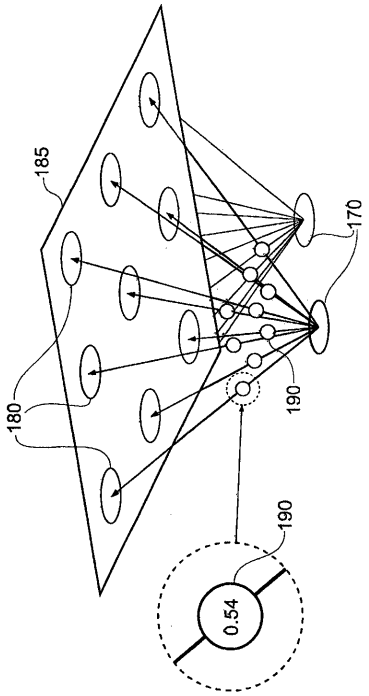


【図４】

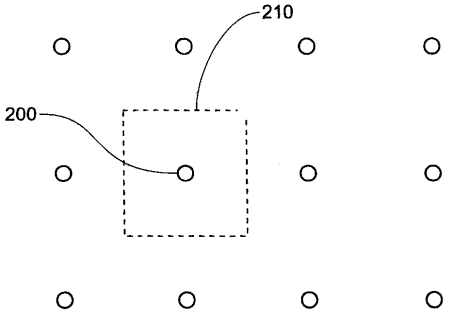
0	0
0	0
0	3
0	0
1	1
0	0
0	7
0	0
0	0
0	0
2	2
.	
.	(b)
.	
0	
0	
0	
0	
3	

(a)

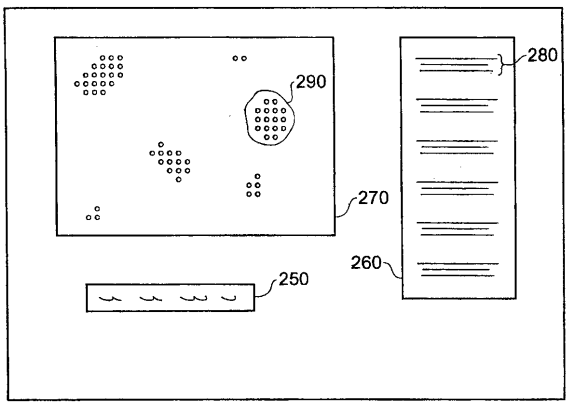
【図５】


185
170
180
190
190
0.54

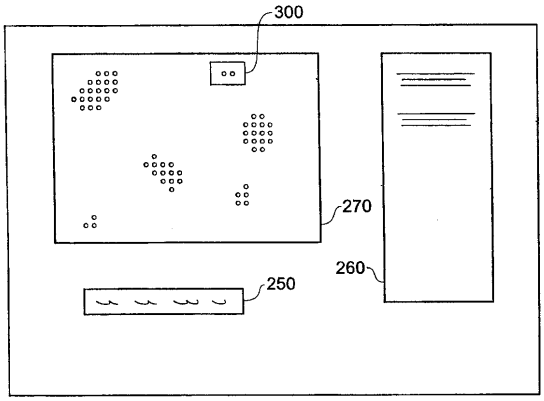
【図６】


210
200

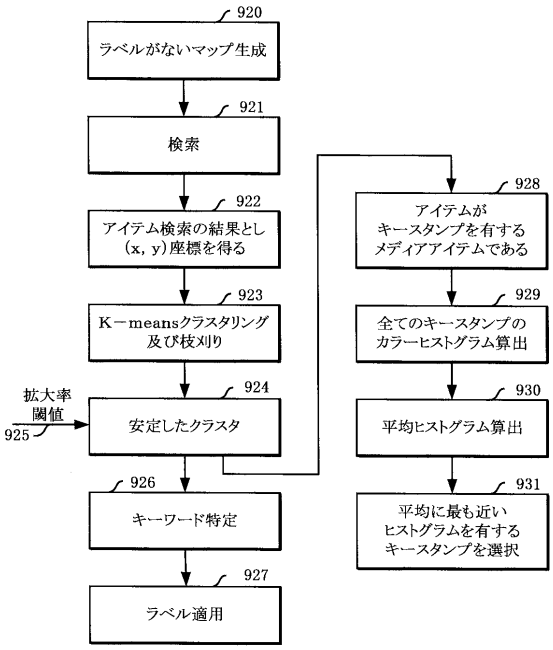
【図７】


280
290
270
260
250

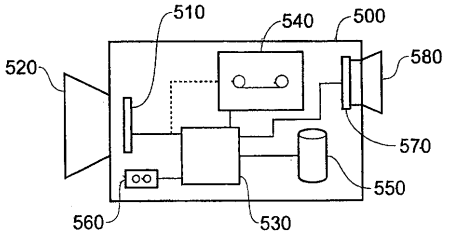
【図８】


300
270
260
250

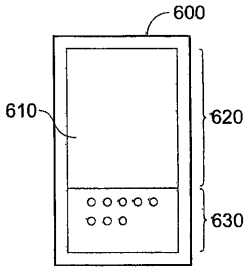
【図９】


920
ラベルがないマップ生成
921
検索
922
アイテム検索の結果とし
(x, y)座標を得る
923
K－meansクラスタリング
及び枝刈り
924
安定したクラスタ
拡大率
閾値
925
926
キーワード特定
927
ラベル適用
928
アイテムが
キースタンプを有する
メディアアイテムである
929
全てのキースタンプの
カラーヒストグラム算出
930
平均ヒストグラム算出
931
平均に最も近い
ヒストグラムを有する
キースタンプを選択

【図１０】


510
540
500
520
580
570
550
560
530

【図１１】


600
610
620
630

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 PQ02　PQ13
5B082 GA08

【外国語明細書】

Title of Invention

INFORMATION STORAGE AND RETRIEVAL

ABSTRACT OF THE DISCLOSURE

An information retrieval system in which a set of distinct information items map to respective nodes in an array of nodes by mutual similarity of the information items, so that similar information items map to nodes at similar positions in the array of nodes; the system comprising:

a user control for defining a search criterion for selecting information items;

a detector for detecting those positions within the array of nodes corresponding to the selected information items;

a graphical user interface for displaying a two-dimensional display array of display points representing those positions within the array of nodes corresponding to the selected information items; and

a processor responsive to the selected information items, for providing a representation which is generally representative of the information content of the selected information items.

BACKGROUND OF THE INVENTION

Field of the invention

This invention relates to information storage and retrieval.

Description of the Prior Art

There are many established systems for locating information (e.g. documents, images, emails, patents, internet content or media content such as audio/video content) by searching under keywords. Examples include internet search "engines" such as those provided by "Google" ™ or "Yahoo" ™ where a search carried out by keyword leads to a list of results which are ranked by the search engine in order of perceived relevance.

However, in a system encompassing a large amount of content, often referred to as a massive content collection, it can be difficult to formulate effective search queries to give a relatively short list of search "hits". For example, at the time of preparing the present application, a Google search on the keywords "massive document collection" drew 243000 hits. This number of hits would be expected to grow if the search were repeated later, as the amount of content stored across the internet generally increases with time. Reviewing such a list of hits can be prohibitively time-consuming.

In general, some reasons why massive content collections are not well utilised are:

- a user doesn't know that relevant content exists
- a user knows that relevant content exists but does not know where it can be located
- a user knows that content exists but does not know it is relevant
- a user knows that relevant content exists and how to find it, but finding the content takes a long time

The paper "Self Organisation of a Massive Document Collection", Kohonen et al, IEEE Transactions on Neural Networks, Vol 11, No. 3, May 2000, pages 574-585 discloses a technique using so-called "self-organising maps" (SOMs). These make use of so-called unsupervised self-learning neural network algorithms in which "feature vectors" representing properties of each document are mapped onto nodes of a SOM.

In the Kohonen et al paper, a first step is to pre-process the document text, and then a feature vector is derived from each pre-processed document. In one form, this may be a histogram showing the frequencies of occurrence of each of a large dictionary of words. Each data value (i.e. each frequency of occurrence of a respective dictionary

word) in the histogram becomes a value in an n-value vector, where n is the total number of candidate words in the dictionary (43222 in the example described in this paper). Weighting may be applied to the n vector values, perhaps to stress the increased relevance or improved differentiation of certain words.

The n-value vectors are then mapped onto smaller dimensional vectors (i.e. vectors having a number of values m (500 in the example in the paper) which is substantially less than n. This is achieved by multiplying the vector by an (n x m) "projection matrix" formed of an array of random numbers. This technique has been shown to generate vectors of smaller dimension where any two reduced-dimension vectors have much the same vector dot product as the two respective input vectors. This vector mapping process is described in the paper "Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering", Kaski, Proc IJCNN, pages 413-418, 1998.

The reduced dimension vectors are then mapped onto nodes (otherwise called neurons) on the SOM by a process of multiplying each vector by a "model" (another vector). The models are produced by a learning process which automatically orders them by mutual similarity onto the SOM, which is generally represented as a two-dimensional grid of nodes. This is a non-trivial process which took Kohonen et al six weeks on a six-processor computer having 800 MB of memory, for a document database of just under seven million documents. Finally the grid of nodes forming the SOM is displayed, with the user being able to zoom into regions of the map and select a node, which causes the user interface to offer a link to an internet page containing the document linked to that node.

The paper by Kohonen discloses the use of labels in the map to provide guidance to the exploration of the map. The labels are automatically generated, by a method of selecting keywords. The method is described in the paper "K.Lagus and S. Laski, Keyword selection method for characterising text document maps, in Proc ICANN99, Ninth Int. Conf. Artificial Neural Networks, vol. 1 1999, pp 371 to 376.

Summary of the Invention

One aspect of the present invention seeks to provide representations of information items which are relevant to a particular search made by a user so as to assist the user.

One aspect of the present invention provides an information retrieval system in which a set of distinct information items map to respective nodes in an array of nodes by mutual similarity of the information items, so that similar information items map to nodes at similar positions in the array of nodes; the system comprising:

a user control for defining a search criterion for selecting information items;

a detector for detecting those positions within the array of nodes corresponding to the selected information items;

a graphical user interface for displaying display points representing those positions within the array of nodes corresponding to the selected information items; and

a processor, responsive to the selected information items defined by the search criterion, for providing one or more representations representative of the information content of the selected information items.

Thus the system provides both display points which are relevant to a search of defined by the user and representations of information items defined by the search; thereby assisting the user.

In an embodiment, the said user control comprises: search means for carrying out a search of the information items; the search means and the graphical user interface being arranged to co-operate so that only those display points corresponding to information items selected by the search are displayed on the user display.

In an embodiment of the system, the processor is operable to provide the said representation(s) on the user display as a label of the display points corresponding to the selected information items. In an embodiment, the label is a word or set of words.

In another embodiment, the information items are at least associated with image items, and the processor is responsive to the selected information items, for providing an image item representing the information content of the selected information items. Preferably, the said processor is operable to select, from the set of image items, an image item which is representative of the selected image items as a whole according to a predetermined selection criterion. Furthermore, the processor may be operable to select the image item a property of which is nearest to the average of the same property of the said set of image items. The selected image item may be applied as a label on display to the display points corresponding to information items represented by the image item.

The graphical user interface preferably displays a two dimensional array of the said display points. However, a one dimensional array could be displayed or a three dimensional array could be displayed. The third dimension of the three dimensional arra.y could be time

According to another aspect of the invention, there is provided an information retrieval method in which a set of distinct information items map to respective nodes in an array of nodes by mutual similarity of the information items, so that similar information items map to nodes at similar positions in the array of nodes; the method comprising the steps of:

defining a search criterion for selecting information items;

detecting those positions within the array of nodes corresponding to the selected information items;

displaying at least display points which are at positions representing those positions within the array of nodes corresponding to the selected information items; and

in response to the selected information items defined by the search criterion, providing one or more representations representative of the information content of the selected information items.

Further respective aspects and features of the invention are defined in the appended claims.

BRIEF DESCRIPTION OF THE DRAWINGS

The above and other objects, features and advantages of the invention will be apparent from the following detailed description of illustrative embodiments which is to be read in connection with the accompanying drawings, in which:

Figure 1 schematically illustrates an information storage and retrieval system;

Figure 2 is a schematic flow chart showing the generation of a self-organising map (SOM);

Figures 3a and 3b schematically illustrate term frequency histograms;

Figure 4a schematically illustrates a raw feature vector;

Figure 4b schematically illustrates a reduced feature vector;

Figure 5 schematically illustrates an SOM;

Figure 6 schematically illustrates a dither process;

Figures 7 and 8 schematically illustrate display screens providing a user interface to access information represented by the SOM;

Figure 9 is a schematic flow diagram illustrating the automatic generation of labels.

Figure 10 schematically illustrates a camcorder as an example of a video acquisition and/or processing apparatus; and

Figure 11 schematically illustrates a personal digital assistant as an example of portable data processing apparatus.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

Figure 1 is a schematic diagram of an information storage and retrieval system based around a general-purpose computer 10 having a processor unit 20 including disk storage 30 for programs and data, a network interface card 40 connected to a network 50 such as an Ethernet network or the Internet, a display device such as a cathode ray tube device 60, a keyboard 70 and a user input device such as a mouse 80. The system operates under program control, the programs being stored on the disk storage 30 and provided, for example, by the network 50, a removable disk (not shown) or a pre-installation on the disk storage 30.

The storage system operates in two general modes of operation. In a first mode, a set of information items (e.g. textual information items) is assembled on the disk storage 30 or on a network disk drive connected via the network 50 and is sorted and indexed ready for a searching operation. The second mode of operation is the actual searching against the indexed and sorted data.

The embodiments are applicable to many types of information items. A non-exhaustive list of appropriate types of information includes patents, video material, emails, presentations, internet content, broadcast content, business reports, audio material, graphics and clipart, photographs and the like, or combinations or mixtures of any of these. In the present description, reference will be made to textual information items. The textual information items may be associated with, or linked to, non-textual items. So, for example, audio and/or video material may be associated with "MetaData" which is a textual information item defining that material in textual terms.

The information items are loaded onto the disk storage 30 in a conventional manner. Preferably, they are stored as part of a database structure which allows for easier

retrieval and indexing of the items, but this is not essential. Once the information and items have been so stored, the process used to arrange them for searching is shown schematically in Figure 2.

It will be appreciated that the indexed information items need not be stored on the local disk drive 30. The information items could be stored on a remote drive connected to the system 10 via the network 50. Alternatively, information may be stored in a distributed manner, for example at various sites across the internet. If the information is stored at different internet or network sites, a second level of information storage could be used to store locally a "link" (e.g. a URI –universal resource identifier) to the remote information, perhaps with an associated summary, abstract or MetaData associated with that link. So, the remotely held information would not be accessed unless the user selected the relevant link (e.g. from the results list 260 to be described below), although for the purposes of the technical description which follows, the remotely held information, or the abstract/summary/MetaData, or the link/URI could be considered as the "information item".

In other words, a formal definition of the "information item" is an item from which a feature vector is derived and processed (see below) to provide a mapping to the SOM. The data shown in the results list 260 (see below) may be the actual information item which a user seeks (if it is held locally and is short enough for convenient display) or may be data representing and/or pointing to the information item, such as one or more of MetaData, a URI, an abstract, a set of key words, a representative key stamp image or the like. This is inherent in the operation "list" which often, though not always, involves listing *data representing* a set of items.

In a further example, the information items could be stored across a networked work group, such as a research team or a legal firm. A hybrid approach might involve some information items stored locally and/or some information items stored across a local area network and/or some information items stored across a wide area network. In this case, the system could be useful in locating similar work by others, for example in a large multi-national research and development organisation, similar research work would tend to be mapped to similar output nodes in the SOM (see below). Or, if a new television programme is being planned, the present technique could be used to check for its originality by detecting previous programmes having similar content.

It will also be appreciated that the system 10 of Figure 1 is but one example of possible systems which could use the indexed information items. Although it is envisaged that the initial (indexing) phase would be carried out by a reasonably powerful computer, most likely by a non-portable computer, the later phase of accessing the information could be carried out at a portable machine such as a "personal digital assistant" (a term for a data processing device with display and user input devices, which generally fits in one hand), a portable computer such as a laptop computer, or even devices such as a mobile telephone, a video editing apparatus or a video camera. In general, practically any device having a display could be used for the information-accessing phase of operation.

The processes are not limited to particular numbers of information items.

The process of generating a self-organising map (SOM) representation of the information items will now be described with reference to Figures 2 to 6. Figure 2 is a schematic flow chart illustrating a so-called "feature extraction" process followed by an SOM mapping process.

Feature extraction is the process of transforming raw data into an abstract representation. These abstract representations can then be used for processes such as pattern classification, clustering and recognition. In this process, a so-called "feature vector" is generated, which is an abstract representation of the frequency of terms used within a document.

The process of forming the visualisation through creating feature vectors includes:

- Create "document database dictionary" of terms
- Create "term frequency histograms" for each individual document based on the "document database dictionary"
- Reduce the dimension of the "term frequency histogram" using random mapping
- Create a 2-dimensional visualisation of the information space.

Considering these steps in more detail, each document (information item) 100 is opened in turn. At a step 110, all "stop words" are removed from the document. Stop-words are extremely common words on a pre-prepared list, such as "a", "the", "however", "about", "and", and "the". Because these words are extremely common they

are likely, on average, to appear with similar frequency in all documents of a sufficient length. For this reason they serve little purpose in trying to characterise the content of a particular document and should therefore be removed.

After removing stop-words, the remaining words are stemmed at a step 120, which involves finding the common stem of a word's variants. For example the words "thrower", "throws", and "throwing" have the common stem of "throw".

A "dictionary" of stemmed words appearing in the documents (excluding the "stop" words) is maintained. As a word is newly encountered, it is added to the dictionary, and running count of the number of times the word has appeared in the whole document collection (set of information items) is also recorded.

The result is a list of terms used in all the documents in the set, along with the frequency with which those terms occur. Words that occur with too high or too low a frequency are discounted, which is to say that they are removed from the dictionary and do not take part in the analysis which follows. Words with too low a frequency may be misspellings, made up, or not relevant to the domain represented by the document set. Words that occur with too high a frequency are less appropriate for distinguishing documents within the set. For example, the term "News" is used in about one third of all documents in a test set of broadcast-related documents, whereas the word "football" is used in only about 2% of documents in the test set. Therefore "football" can be assumed to be a better term for characterising the content of a document than "News". Conversely, the word "fottball" (a misspelling of "football") appears only once in the entire set of documents, and so is discarded for having too low an occurrence. Such words may be defined as those having a frequency of occurrence which is lower than two standard deviations less than the mean frequency of occurrence, or which is higher than two standard deviations above the mean frequency of occurrence.

A feature vector is then generated at a step 130.

To do this, a term frequency histogram is generated for each document in the set. A term frequency histogram is constructed by counting the number of times words present in the dictionary (pertaining to that document set) occur within an individual document. The majority of the terms in the dictionary will not be present in a single document, and so these terms will have a frequency of zero. Schematic examples of term frequency histograms for two different documents are shown in Figures 3a and 3b.

It can be seen from this example how the histograms characterise the content of the documents. By inspecting the examples it is seen that document 1 has more occurrences of the terms "MPEG" and "Video" than document 2, which itself has more occurrences of the term "MetaData". Many of the entries in the histogram are zero as the corresponding words are not present in the document.

In a real example, the actual term frequency histograms have a very much larger number of terms in them than the example. Typically a histogram may plot the frequency of over 50000 different terms, giving the histogram a dimension of over 50000. The dimension of this histogram needs to be reduced considerably if it is to be of use in building an SOM information space.

Each entry in the term frequency histogram is used as a corresponding value in a feature vector representing that document. The result of this process is a (50000 x 1) vector containing the frequency of all terms specified by the dictionary for each document in the document collection. The vector may be referred to as "sparse" since most of the values will typically be zero, with most of the others typically being a very low number such as 1.

The size of the feature vector, and so the dimension of the term frequency histogram, is reduced at a step 140. Two methods are proposed for the process of reducing the dimension of the histogram.

i) Random Mapping – a technique by which the histogram is multiplied by a matrix of random numbers. This is a computationally cheap process.

ii) Latent Semantic Indexing – a technique whereby the dimension of the histogram is reduced by looking for groups of terms that have a high probability of occurring simultaneously in documents. These groups of words can then be reduced to a single parameter. This is a computationally expensive process.

The method selected for reducing the dimension of the term frequency histogram in the present embodiment is "random mapping", as explained in detail in the Kaski paper referred to above. Random mapping succeeds in reducing the dimension of the histogram by multiplying it by a matrix of random numbers.

As mentioned above, the "raw" feature vector (shown schematically in Figure 4a) is typically a sparse vector with a size in the region of 50000 values. This can be reduced to size of about 200 (see schematic Figure 4b) and still preserve the *relative*

characteristics of the feature vector, that is to say, its relationship such as relative angle (vector dot product) with other similarly processed feature vectors. This works because although the number of orthogonal vectors of a particular dimension is limited, the number of *nearly* orthogonal vectors is very much larger.

In fact as the dimension of the vector increases any given set of randomly generated vectors are nearly orthogonal to each other. This property means that the relative direction of vectors multiplied by this a matrix of random numbers will be preserved. This can be demonstrated by showing the similarity of vectors before and after random mapping by looking at their dot product.

It can be shown experimentally that by reducing a sparse vector from 50000 values to 200 values preserves their relative similarities. However, this mapping is not perfect, but suffices for the purposes of characterising the content of a document in a compact way.

Once feature vectors have been generated for the document collection, thus defining the collection's information space, they are projected into a two-dimensional SOM at a step 150 to create a semantic map. The following section explains the process of mapping to 2-D by clustering the feature vectors using a Kohonen self-organising map. Reference is also made to Figure 5.

A Kohonen Self-Organising map is used to cluster and organise the feature vectors that have been generated for each of the documents.

A self-organising map consists of input nodes 170 and output nodes 180 in a two-dimensional array or grid of nodes illustrated as a two-dimensional plane 185. There are as many input nodes as there are values in the feature vectors being used to train the map. Each of the output nodes on the map is connected to the input nodes by weighted connections 190 (one weight per connection).

Initially each of these weights is set to a random value, and then, through an iterative process, the weights are "trained". The map is trained by presenting each feature vector to the input nodes of the map. The "closest" output node is calculated by computing the Euclidean distance between the input vector and weights of each of the output nodes.

The closest node is designated the "winner" and the weights of this node are trained by slightly changing the values of the weights so that they move "closer" to the

input vector. In addition to the winning node, the nodes in the neighbourhood of the winning node are also trained, and moved slightly closer to the input vector.

It is this process of training not just the weights of a single node, but the weights of a region of nodes on the map, that allow the map, once trained, to preserve much of the topology of the input space in the 2-D map of nodes.

Once the map is trained, each of the documents can be presented to the map to see which of the output nodes is closest to the input feature vector for that document. It is unlikely that the weights will be identical to the feature vector, and the Euclidean distance between a feature vector and its nearest node on the map is known as its "quantisation error".

By presenting the feature vector for each document to the map to see where it lies yields an x, y map position for each document. These x, y positions when put in a look up table along with a document ID can be used to visualise the relationship between documents.

Finally, a dither component is added at a step 160, which will be described with reference to Figure 6 below.

A potential problem with the process described above is that two identical, or substantially identical, information items may be mapped to the same node in the array of nodes of the SOM. This does not cause a difficulty in the handling of the data, but does not help with the visualisation of the data on display screen (to be described below). In particular, when the data is visualised on a display screen, it has been recognised that it would be useful for multiple very similar items to be distinguishable over a single item at a particular node. Therefore, a "dither" component is added to the node position to which each information item is mapped. The dither component is a random addition of ±½ of the node separation. So, referring to Figure 6, an information item for which the mapping process selects an output node 200 has a dither component added so that it in fact may be mapped to any map position around a node 200 within the area 210 bounded by dotted lines on Figure 6.

So, the information items can be considered to map to positions on the plane of Figure 6 at node positions other than the "output nodes" of the SOM process.

At any time, a new information item can be added to the SOM by following the steps outlined above (i.e. steps 110 to 140) and then applying the resulting reduced

feature vector to the "pre-trained" SOM models, that is to say, the set of SOM models which resulted from the self-organising preparation of the map. So, for the newly added information item, the map is not generally "retrained"; instead steps 150 and 160 are used with all of the SOM models not being amended. To retrain the SOM every time a new information item is to be added is computationally expensive and is also somewhat unfriendly to the user, who might grow used to the relative positions of commonly accessed information items in the map.

However, there may well come a point at which a retraining process is appropriate. For example, if new terms (perhaps new items of news, or a new technical field) have entered into the dictionary since the SOM was first generated, they may not map particularly well to the existing set of output nodes. This can be detected as an increase in a so-called "quantisation error" detected during the mapping of newly received information item to the existing SOM. In the present embodiments, the quantisation error is compared to a threshold error amount. If it is greater than the threshold amount then either (a) the SOM is automatically retrained, using all of its original information items and any items added since its creation; or (b) the user is prompted to initiate a retraining process at a convenient time. The retraining process uses the feature vectors of all of the relevant information items and reapplies the steps 150 and 160 in full.

Figure 7 schematically illustrates a display on the display screen 60. The display shows a search enquiry 250, a results list 260 and an SOM display area 270.

In operation, initially, the display area 270 is blank. The user types a key word search enquiry into the enquiry area 250. The user then initiates the search, for example by pressing enter on the keyboard 70 or by using the mouse 80 to select a screen "button" to start the search. The key words in the search enquiry area 250 are then compared with the information items in the database using a standard keyword search technique. This generates a list of results, each of which is shown as a respective entry 280 in the list area 260. Then the display area 270 displays display points corresponding to each of the result items.

Because the sorting process used to generate the SOM representation tends to group mutually similar information items together in the SOM, the results for the search enquiry generally tend to fall in clusters such as a cluster 290. Here, it is noted that each

point on the area 270 corresponds to the respective entry in the SOM associated with one of the results in the result list 260; and the positions at which the points are displayed within the area 270 correspond to the array positions of those nodes within the node array.

Figure 8 schematically illustrates a technique for reducing the number of "hits" (results in the result list). The user makes use of the mouse 80 to draw a boundary, which in this example is a rectangular box, 300 around a set of the display points displayed in area 270. In the results list area 260, only those results corresponding to points within the boundary 300 are displayed. If these results turn out not to be of interest, the user may draw another boundary encompassing a different set of display points.

It is noted that the results area 260 displays list entries for those results for which display points are displayed within the boundary 300 **and** which satisfied the search criteria in the word search area 250. The boundary 300 may encompass other display positions corresponding to populated nodes in the node array, but if these did not satisfy the search criteria they will not be displayed and so will not form part of the subset of results shown in the list 260.

Figure 9 illustrates an embodiment of the present invention.

Referring to Figure 9, step 920, when the Self Organising Map SOM is generated it has no labels, (unlike the SOM of Kohonen). Users require labels to give guidance for exploring the map. In embodiments of the invention the labels are automatically generated to match the particular needs of the users. Users generate a list of results of a search as described with reference to Figure 7 and/or Figure 8. A label is automatically dynamically generated according to the results and used to label the clusters of display points in the area 270.

At step 921, a user initiates a searching operation. The search is initiated in this example using a key word search enquiry as describe above with reference to Figure 7. This results in many result items, for example documents, denoted by individual X-Y positions on the map as indicated by step 922. Steps 921 and 922 are followed by step 923 "k-means clustering and pruning". Step 923 operates on the X-Y positions on the map, corresponding to the set of result items resulting from the search performed in step 921. At step 923, the known process "k-means clustering" identifies groups of result items which denote similar items. The process includes a process (denoted pruning in

step 923) of determining whether adjacent sub-clusters of X-Y positions of result items are part of the same main cluster. If the distance between the centres of two sub-clusters is less than a threshold value, then the two sub-clusters are deemed to be part of the same main cluster. The pruning is performed iteratively in known manner until the clustering is stable as indicated at step 924.

In step 926 the items which are in a cluster are identified and the keywords are identified. The top 20 keywords, and their frequency are pre-computed for each of the items. Calculating the cluster labels involves summing totals for all of the keywords that have been pre-computed for each item, and selecting the most frequent. The selected most frequent keyword is applied as the label of a cluster. The pre-computation of keywords significantly reduces the time required to generate a label for a cluster.

If there is only a single item in the cluster then the 'title' of that item is used as the label.

Thus the determination of clusters and the labelling of the clusters is dependent on the search enquiry made by the user.

The threshold, mentioned above, of the distance between sub-clusters may be chosen by the user. Alternatively, the threshold may be dependent on "zoom". A part of the SOM may be viewed on an enlarged scale. When the map is viewed on a small scale, adjacent sub-clusters form one larger cluster yet when viewed on an enlarged scale they are separate. Thus the threshold is made dependent on the zoom factor.

The items may be documents, images or video. If the items are media items such as images or video the keywords may be keywords used in metadata relating to the images or video. Titles are an example of metadata. Other examples of metadata are known in the art.

It is known to provide Representative Key Stamps (RKS) representing media items such as video. In accordance with another embodiment of the invention, a search is performed according to Figure 7 or Figures 7 and 8 and steps 920 to 924 as described above to produce a results list, corresponding display points in the display area 270, and to identify stable clusters. As shown in steps 928 to 931 an RKS is selected as a label for a cluster 290. By way of example, in one embodiment of the invention, at step 928, the RKSs of the media items of a cluster 290 are obtained, and at step 929, colour histograms of all those RKSs are calculated. At step 930, the average colour histogram is calculated,

and at step 931, the RKS having a colour histogram which most closely matches the average histogram is selected. The average RKS is a representation of the cluster. In this way, clusters may be labelled with RKSs representative of the clusters. Individual RKSs represent individual media items in the or each cluster.

A heirarchy of RKSs may be provided in which an RKS is chosen to represent a whole results list, an RKS is chosen to represent each cluster, and individual RKSs represent individual media items in the results list.

An RKS of a single media item may be chosen in any known way. One example is as follows:

i) For determining an RKS for a single media item, metrics can be calculated for each frame in the item to determine whether there are faces in the image, who the face belongs to and whether there is speech in the audio. The representative frame can then be selected as being a frame containing the face of the individual featured most within the media item, that also contains speech. For instance the media item 'Parkinson' (a television presenter well known in the UK) would favour a frame containing Parkinson speaking to camera over frames of his guests.

An RKS of a group of media items (e.g. a list or a cluster) may be chosen any known way. Some examples are as follows:

ii) For determining an RKS for a group of media items, one can take all the frames in all the media items and process them in the same way as we do in(i). For example, if we have 10 x 5 minute items, we process this as in (i) as if it were a single 50 minute item.

iii) Alternatively for a group of media items, one can determine the RKS for each individual item as in (i). Next we assume that our search result list includes some relevancy metric, which ranks each result according to its relevancy to the search query. The RKS for the group of media items can now be determined as being the same as the RKS of the most relevant item in the group of media items.

iv) As an extension to (iii), any method used to rank search results can be used to determine the highest ranked hit in any cluster of results, and simply using the RKS of that item to represent the group.

v) Simple methods for obtaining an RKS include: automatically selecting a frame at random from the set of all frames that make up all the media items in the group; simply

automatically selecting the frame, from the group of items, that happens to be to processed first; and the user choosing a frame they think is most representative.

Figure 10 schematically illustrates a camcorder 500 as an example of a video acquisition and/or processing apparatus, the camcorder including an image capture device 510 with an associated lens 520; a data/signal processor 530; tape storage 540; disk or other random access storage 550; user controls 560; and a display device 570 with eyepiece 580. Other features of conventional camcorders or other alternatives (such as different storage media or different display screen arrangements) will be apparent to the skilled man. In use, MetaData relating to captured video material may be stored on the storage 550, and an SOM relating to the stored data viewed on the display device 570 and controlled as described above using the user controls 560.

Figure 11 schematically illustrates a personal digital assistant (PDA) 600, as an example of portable data processing apparatus, having a display screen 610 including a display area 620 and a touch sensitive area 630 providing user controls; along with data processing and storage (not shown). Again, the skilled man will be aware of alternatives in this field. The PDA may be used as described above in connection with the system of Figure 1.

Although illustrative embodiments of the invention have been described in detail herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various changes and modifications can be effected therein by one skilled in the art without departing from the scope and spirit of the invention as defined by the appended claims.

1. An information retrieval system in which a set of distinct information items map to respective nodes in an array of nodes by mutual similarity of the information items, so that similar information items map to nodes at similar positions in the array of nodes; the system comprising:

a user control for defining a search criterion for selecting information items;

a detector for detecting those positions within the array of nodes corresponding to the selected information items;

a graphical user interface for displaying display points representing those positions within the array of nodes corresponding to the selected information items; and

a processor, responsive to the selected information items defined by the search criterion, for providing one or more representations representative of the information content of the selected information items.

2. A system according to claim 1, wherein the graphical user interface is operable to display a two-dimensional display array of the said display points.

3. A system according to claim 2, in which a dither component is applied to the mapping between information items and nodes in the array so that substantially identical information items tend to map to closely spaced but different positions in the displayed array.

4. A system according to claim 2, in which the information items are mapped to nodes in the array on the basis of a feature vector derived from each information item.

5. A system according to claim 4, in which the feature vector for an information item represents a set of frequencies of occurrence, within that information item, of each of a group of information features.

6. A system according to claim 5, in which the information items comprise textual information, the feature vector for an information item represents a set of frequencies of occurrence, within that information item, of each of a group of words.

7. A system according to claim 1, in which the information items comprise textual information, the nodes being mapped by mutual similarity of at least a part of the textual information.

8. A system according to claim 6, in which the information items are pre-processed for mapping by excluding words occurring with more than a threshold frequency amongst the set of information items.

9. A system according to claim 6, in which the information items are pre-processed for mapping by excluding words occurring with less than a threshold frequency amongst the set of information items.

10. A system according to claim 1, wherein the said user control comprises:

search means for carrying out a search of the information items;

the search means and the graphical user interface being arranged to co-operate so that only those display points corresponding to information items selected by the search are displayed on the user display.

11. A system according to claim 1, wherein the said processor is operable to detect clusters of similar information items and to provide representations representative of the information content of the respective clusters.

12. A system according to claim 1, wherein the processor is operable to provide the or each said representation on the user display as a label of the display points corresponding to the information items represented thereby.

13. A system according to claim 12, wherein the label is a word or set of words.

14. A system according to claim 11, wherein the processor determines, in respect of a set of information items with which a label is to be associated, the most frequently used word or set of words in the information items corresponding to the selected information items and applies that word or that set of words as the label.

15. An information retrieval system according to claim 1, in which the information items are at least associated with image items, and

wherein the processor is responsive to the selected information items, for providing one or more image items representative of the information content of the selected information items defined by the search criterion.

16. A system according to claim 15, wherein the said processor is operable to select, from the set of image items, an image item which is representative of the set of image items as a whole according to a predetermined selection criterion.

17. A system according to claim 15, wherein the processor is operable to select the image item a property of which is nearest to the average of the same property of the said set of image items.

18. A system according to claim 15, wherein the said one or more representative image items are applied as labels to the display points corresponding to the information items represented thereby.

19. A portable data processing device comprising a system according to claim 1.

20. Video acquisition and/or processing apparatus comprising a system according to claim 1.

21. An information retrieval method in which a set of distinct information items map to respective nodes in an array of nodes by mutual similarity of the information items, so that similar information items map to nodes at similar positions in the array of nodes; the method comprising the steps of:

defining a search criterion for selecting information items;

detecting those positions within the array of nodes corresponding to the selected information items;

displaying at least display points which are at positions representing those positions within the array of nodes corresponding to the selected information items; and

in response to the selected information items defined by the search criterion, providing one or more representations representative of the information content of the selected information items.

22. A method according to claim 21, wherein the displaying step displays a two-dimensional display array of the said display points.

23. A method according to claim 21, comprising:

carrying out a search of the information items;

displaying on the display that only those display points corresponding to information items selected by the search are displayed on the user display.

24. A method according to claim 21, comprising detecting clusters of similar information items and providing representations representative of the information content of the respective clusters.

25. A method according to claim 21, comprising providing the or each said representation on the user display as a label of the display points corresponding to the information items represented thereby.

26. A method according to claim 25, wherein the label is a word or set of words.

27. A method according to claim 21, in which the information items are at least associated with image items, and

comprising providing one or more image items representative of the information content of the selected information items defined by the search criterion.

28. A method according to claim 27, comprising selecting, from the set of image items, an image item which is representative of the set of image items as a whole according to a predetermined selection criterion.

29. A method according to claim 28, comprising selecting the image item a property of which is nearest to the average of the same property of the said set of image items.

30. Computer software having program code for carrying out a method according to claim 21.

31. A providing medium for providing program code according to claim 30.

32. A medium according to claim 31, the medium being a storage medium.

33. A medium according to claim 31, the medium being a transmission medium.

34. A user interface of an information retrieval system in which a set of distinct information items map to respective nodes in an array of nodes by mutual similarity of the information items, so that similar information items map to nodes at similar positions in the array of nodes; the interface comprising:

a user control for defining a search criterion for selecting information items; and

a graphical user interface arranged to displaying display points representing those positions within the array of nodes corresponding to the selected information items and to display one or more representations representative of the information content of the information items. selected by the search criterion.

35. A user interface according to claim 34, wherein the said user control comprises:

search means for carrying out a search of the information items;

the search means and the graphical user interface being arranged to co-operate so that only those display points corresponding to information items selected by the search are displayed on the user display.

36. An interface according to claim 34, wherein the graphical user interface is arranged to display representations representative of the information content of respective clusters of similar information items.

37. An interface according to claim 34, wherein graphical user interface is operable to provide the or each said representation as a label of the display points corresponding to the information items represented thereby.

38. An interface according to claim 37, wherein the label is a word or set of words.

39. An interface according to claim 34, wherein the said representations are image items which are applied as labels to the display points corresponding to the information items represented thereby.

ABSTRACT

INFORMATION STORAGE AND RETRIEVAL

An information retrieval system in which a set of distinct information items map to respective nodes in an array of nodes by mutual similarity of the information items, so that similar information items map to nodes at similar positions in the array of nodes; the system comprising:

a user control for defining a search criterion for selecting information items;

a detector for detecting those positions within the array of nodes corresponding to the selected information items;

a graphical user interface for displaying a two-dimensional display array of display points representing those positions within the array of nodes corresponding to the selected information items; and

a processor responsive to the selected information items, for providing a representation which is generally representative of the information content of the selected information items.

Representative Drawing

Figure 8.

【図１】

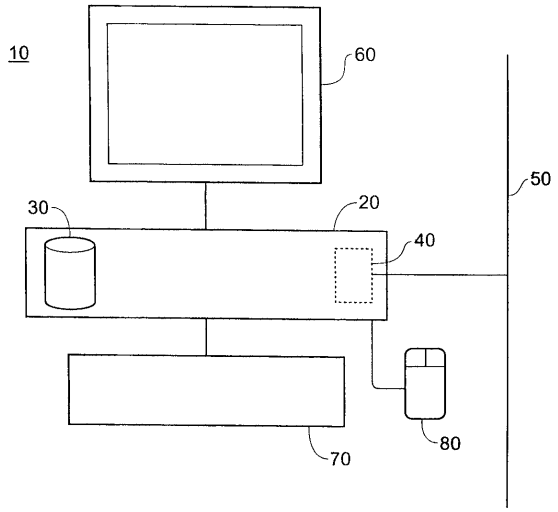


Fig. 1

【図２】

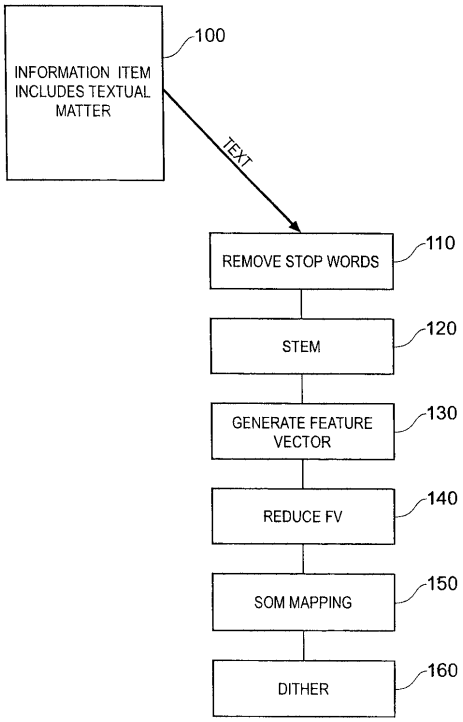


Fig. 2

【図３】

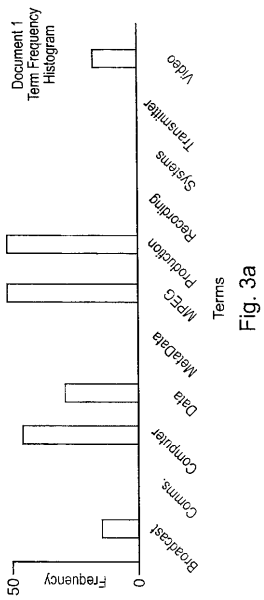


Fig. 3a

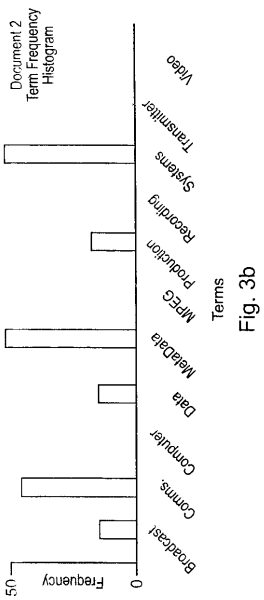


Fig. 3b

【図４】

0	0
0	0
0	3
0	0
1	1
0	0
0	7
0	0
0	0
0	0
2	2
.	
.	
.	
0	
0	
0	
0	
3	

Fig. 4a

Fig. 4b

【図５】

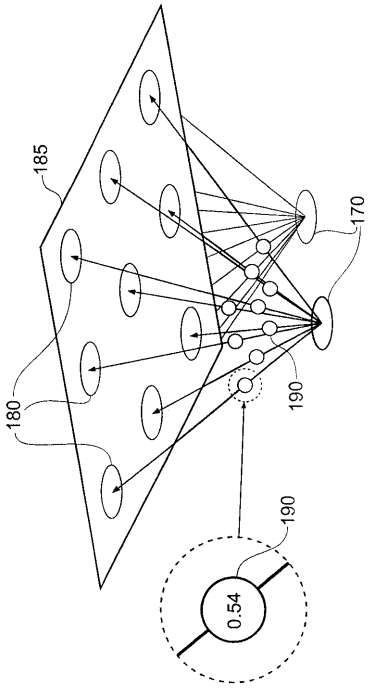


Fig. 5

【図６】

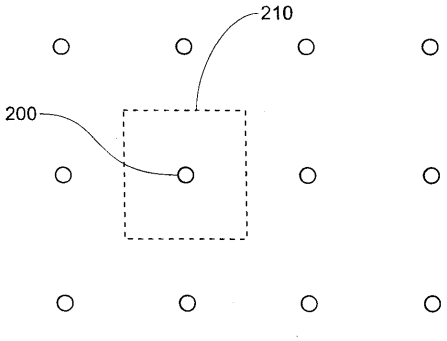


Fig. 6

【図７】

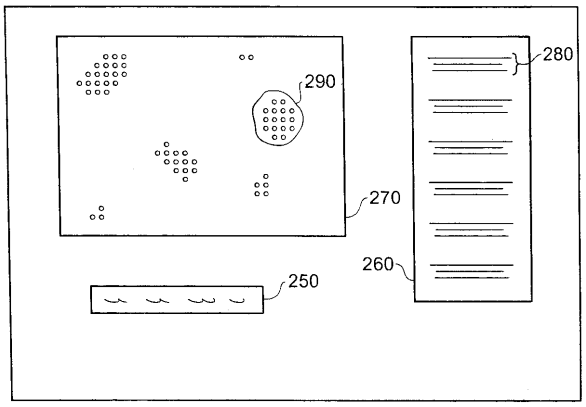


Fig. 7

【図８】

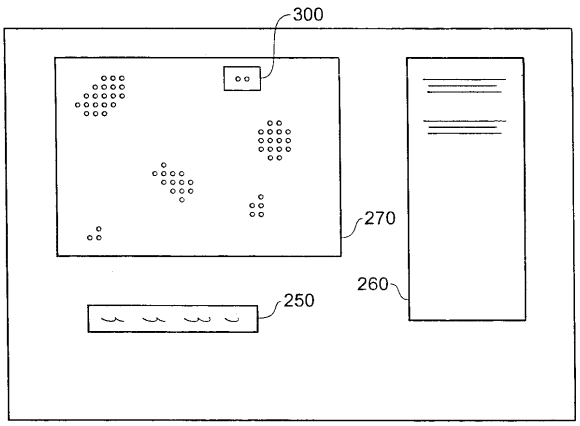


Fig. 8

【図９】

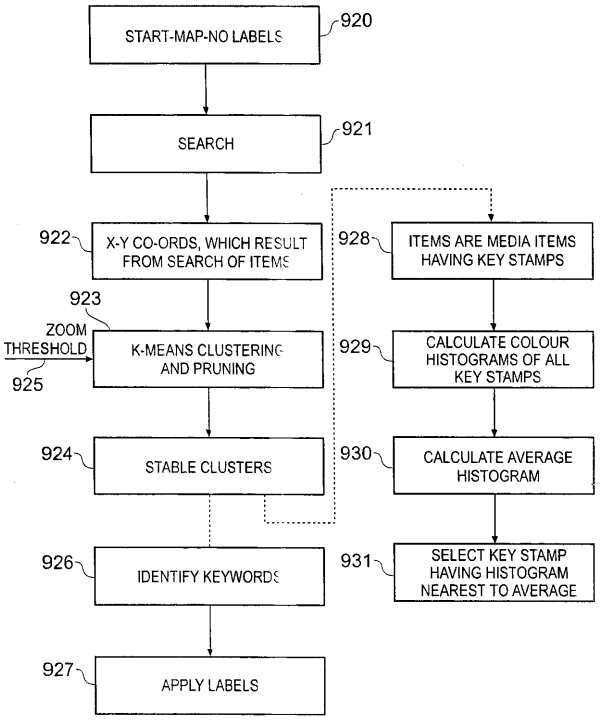


Fig. 9

【図１０】

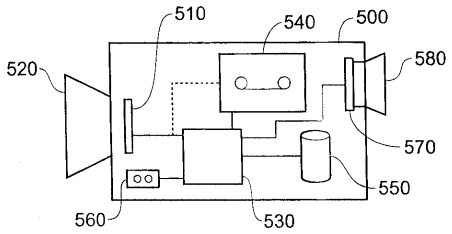


Fig. 10

【図１１】

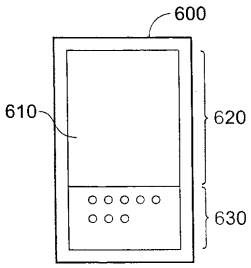


Fig. 11