(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4874413号
(P4874413)
(45)発行日　平成24年2月15日(2012.2.15)　(24)登録日　平成23年12月2日(2011.12.2)

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

請求項の数 3　(全 46 頁)

(21)出願番号	特願2010-140222 (P2010-140222)
(22)出願日	平成22年6月21日 (2010.6.21)
(62)分割の表示	特願2000-16705 (P2000-16705)の分割
原出願日	平成12年1月26日 (2000.1.26)
(65)公開番号	特開2010-250849 (P2010-250849A)
(43)公開日	平成22年11月4日 (2010.11.4)
審査請求日	平成22年6月21日 (2010.6.21)
(31)優先権主張番号	117462
(32)優先日	平成11年1月26日 (1999.1.26)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	421416
(32)優先日	平成11年10月19日 (1999.10.19)
(33)優先権主張国	米国 (US)

(73)特許権者　596170170
ゼロックス　コーポレイション
ＸＥＲＯＸ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国、コネチカット州　０６８５６、ノーウォーク、ピーオーボックス　４５０５、グローバー・アヴェニュー　４５
(74)代理人　100079049
弁理士　中島　淳
(74)代理人　100084995
弁理士　加藤　和詳
(72)発明者　フランシーン　アール．チェン
アメリカ合衆国　９４０２５　カリフォルニア州　メンロ　パーク　シャーマン　アベニュー　９７５

最終頁に続く

(54)【発明の名称】オブジェクト間類似度計算方法

(57)【特許請求の範囲】
【請求項１】
コンピュータが実行する、オブジェクトのコレクション内の２つのオブジェクトの間の類似度を計算するオブジェクト間類似度計算方法であって、
各オブジェクトは、少なくとも第１の特徴ベクトルと第２の特徴ベクトルに関連付けられ、
前記第１の特徴ベクトルと前記第２の特徴ベクトルはそれぞれ、複数の次元のベクトルであり、
前記第１の特徴ベクトルは、前記オブジェクトの第１の特徴を表し、
前記第２の特徴ベクトルは、前記オブジェクトの第２の特徴を表し、
前記第１の特徴は、テキスト特徴、ＵＲＬ特徴、インリンク特徴、及びアウトリンク特徴を含むマルチモードの特徴の第１のセット内の１つであり、
前記第２の特徴は、画像特徴であり、
前記オブジェクト間類似度計算方法は、
第１のオブジェクトの前記第１の特徴ベクトルと第２のオブジェクトの前記第１の特徴ベクトルとを識別するステップと、
前記第１のオブジェクトの前記第１の特徴ベクトルと前記第２のオブジェクトの前記第１の特徴ベクトルとの間の第１の距離メトリックを計算するステップと、
テキスト情報の参照なく、前記第１のオブジェクトの前記第２の特徴ベクトルと前記第２のオブジェクトの前記第２の特徴ベクトルとを識別するステップと、

前記第１のオブジェクトの前記第２の特徴ベクトルと前記第２のオブジェクトの前記第２の特徴ベクトルとの間の第２の距離メトリックを計算するステップと、
前記第１の距離メトリックと前記第２の距離メトリックとの合計を計算するステップと、
を備えたオブジェクト間類似度計算方法。
【請求項２】
各オブジェクトは、ドキュメントのコレクション内の１つのドキュメントに対応する請求項１記載のオブジェクト間類似度計算方法。
【請求項３】
前記第１の特徴は、(i)前記テキスト特徴、(ii)ＵＲＬ特徴、(iii)インリンク特徴、及び(iv)アウトリンク特徴の少なくとも１つであり、
前記第１の距離メトリックは、前記第１のオブジェクトの前記第１の特徴ベクトルと前記第２のオブジェクトの前記第１の特徴ベクトルとの間のコサイン類似度である
請求項２記載のオブジェクト間類似度計算方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、オブジェクト間類似度計算方法、ドキュメント間類似度計算方法、及びユーザ特性間類似度計算方法にかかり、より詳細には、コンピュータが実行する、オブジェクトのコレクション内の２つのオブジェクトの間の類似度を計算するオブジェクト間類似度計算方法、コンピュータが実行する、ドキュメントのコレクション内の２つのドキュメントの間の類似度を計算するドキュメント間類似度計算方法、及び、コンピュータが実行する、ドキュメントコレクションのユーザ集団内の二人のユーザの特性間の類似度を計算するユーザ特性間類似度計算方法に関する。
【背景技術】
【０００２】
コンピュータユーザは、求めているドキュメントコレクションを発見することが益々困難になってきている。理由は、そのようなコレクションのサイズが増加しているためである。例えば、インターネット上でワールドワイドウエブ（WWW）は、数百万の個々のページを含む。また、多くの会社の内部イントラネットは、しばしば数千ものドキュメントを含むレポジトリを含む。
【０００３】
ワールドワイドウエブ上及びイントラネットレポジトリ中のドキュメントは、あまりうまくインデックス付けされていないということもしばしば真実である。その結果、特定のドキュメントのアイデンティティや位置や特徴が公知でない場合、そのような大きなコレクション（収集物）中で必要な情報を検出することは、望みの無い捜しものをするようなものである。
【０００４】
ワールドワイドウエブは、インターネットを介して分散されたサーバ上に位置するドキュメント（大部分がテキストと画像である）の緩く相互リンクされたコレクションである。一般的に言えば、各ドキュメントは、例示のフォーム"ｈｔｔｐ：／／ｗｗｗ．ｓｅｒｖｅｒ．ｎｅｔ／ｄｉｒｅｃｔｏｒｙ／ｆｉｌｅ．ｈｔｍｌ"において、アドレス即ちユニフォームリソースロケータ（ＵＲＬ）を有する。その表記法において、"ｈｔｔｐ："は、そのドキュメントが引き渡されるべきプロトコルを指定し、この場合、"ＨｙｐｅｒＴｅｘｔ Ｔｒａｎｓｐｏｒｔ Ｐｒｏｔｏｃｏｌ（ハイパーテキスト転送プロトコル"である。"ｗｗｗ．ｓｅｒｖｅｒ．ｎｅｔ"は、そのドキュメントが常駐するコンピュータやサーバの名前を指定し、"ｄｉｒｅｃｔｏｒｙ"は、そのドキュメントが常駐するサーバ上のディレクトリやフォルダを指し、"ｆｉｌｅ．ｈｔｍｌ"は、そのファイルのネームを指定する。
【０００５】

ウエブ上の多くのドキュメントは、ＨＴＭＬ（ハイパーテキストメークアップ言語）フォーマットであり、とりわけ、フォーマッティングのそのドキュメントへの適用、外部コンテント（画像及び他のマルチメディアデータタイプのような）のそのドキュメント内への導入、及び他のドキュメントへの"ホットリンク"や"リンク"のそのドキュメント内への配置を可能とする。"ホットリンキング"によって、ユーザは、ページ内の対象となるアイテムを選択することによって、簡単にウエブ上のドキュメント同士間をナビゲートすることが出来る。例えば、再グラフィック技術についてのウエブページは、ゼロックスコーポレートサイトへのホットリンクを持ち得る。ホットリンクを選択することによって（しばしば、マークされたワードや画像や領域をマウスのようなポインティングデバイスでクリックすることによって）、ユーザのウエブブラウザは、（通常、そのホットリンクに関連する、しばしばユーザには見えないＵＲＬを介して）ホットリンクに従って、異なるドキュメントを読み出すよう指示される。

【０００６】

インターネット上の各及び全てのドキュメントに対するＵＲＬや、好ましいドキュメントの同様のコレクション中のこれらのドキュメントでさえユーザが思い出すことが期待できないことは明白である。従って、ナビゲーションの助けは、役に立つのみならず必要である。

【０００７】

従って、ユーザがユーザのブックマークコレクションにはまだ表れていないインターネット（又は他の大きなネットワーク）上で情報を見つけることを希望する場合、ユーザは、しばしばその情報を検出するために"サーチエンジン"をオンにする。サーチエンジンは、インターネット上に格納されたコンテントへのインデックスとして働く。

【０００８】

サーチエンジンには、二つのカテゴリがあり、一つは、サブジェクトマターカテゴリの階層を形成するために解析され且つ使用されるドキュメント及びウエブサイトを含むカテゴリ（例えば、ヤフー（登録商標））であり、他は、用語のサーチ可能データベースを確立するためにウエブやドキュメントコレクションを"クロール"してページコンテント上のキーワードサーチを可能とするカテゴリー（とりわけ、ＡｌｔａＶｉｓｔａ、Ｅｘｃｉｔｅ、及びＩｎｆｏｓｅｅｋ）である。

【０００９】

また、ユーザにより或いは一つの好ましいドキュメントへの比較により提供される基準に基づきウエブサイト推奨を行うことが出来る推奨システムが知られている。

【００１０】

前述のシステム及びサービスは、伝統的なサーチエンジンと推奨システム能力をある程度組合せるが、それらの何れもが全体としてインターネットよりも小さなグループのプリファランスを考慮すると共にサーチエンジンのような能力を提供するために現在適応できない。特に、コミュニティやクラスタベースの推奨（レコメンデーション）をインターネットや他のドキュメントのコレクションから未知のドキュメントを検索出来るシステムへ組合せることは有利である。

【００１１】

従って、ドキュメントの大きなコレクション、即ちコーパスを処理する場合、それらのコンテント（内容）に基づき、これらのドキュメントをサーチし、ブラウジングし、検索し、そして見ることが出来る事が有用である。しかしながら、そのドキュメントの制限のために、これは多くの場合困難である。例えば、ワールドワイドウエブ上には、ドキュメントとファイルの一般的コレクションンで利用できる多くの種類の情報がある。テキストファイル、テキストと画像の両方を含むＨＴＭＬ（ハイパーテキストメイクアップ言語）、画像、サウンドファイル、マルチメディアファイル、及び他のタイプのコンテントがある。

【００１２】

画像を容易にブラウジングし検索するために、コレクション中の各画像は、理想的には

、その画像中のオブジェクト及びその画像の記述を含む記述情報でラベル付けされるべきである。しかしながら、ウエブ上のオブジェクトのような、画像の無制限のコレクション中のオブジェクトの識別は困難なタスクである。オブジェクトの自動的識別方法は、一般的に、マシーンパーツのような特定のドメインに制限される。人間に各画像を識別させることは、ウエブ上におけるように、面倒な仕事であり、しばしば不可能である。

【００１３】

情報検索における多くのサーチは、コンテントに基づくテキストドキュメントの検索や、可視特徴に基づく画像ドキュメントに基づく検索に焦点があてられている。更に、ウエブ及び会社のイントラネット上の情報の急激な増加で、ユーザは、特定の情報をサーチする時、ヒント攻めに遭う。実際に求められる物を発見するために結果を分類するタスクは、しばしば冗長であり、時間がかかる。最近、多くのサーチエンジンは、ユーザがメタデータ（例えば、Ｈｏｔｂｏｔ、Ｉｎｆｏｓｅｅｋ）を介して伝統的なキーワードエントリから質問を拡張することを可能とする機能性を追加している。メタデータは、言語、日付、サイトの位置、或いは画像、ビデオ又はオーディオ等のモダリティが存在するか否かのような、種々のフォームを取ることが出来る。

【００１４】

しかしながら、最近、検索のために使用マルチモード特徴についてのあるリサーチがある。コレクション中のドキュメントのマルチモード特徴並びにユーザのブラウジングの習癖の間の類似度に基づいて、ユーザが希望の情報を検出できる幾つかのアプローチがここで示される。

【００１５】

ユーザが画像とその画像と関連するテキストの両方、及び使用のような、ドキュメントに関連する他のタイプの情報を使用してサーチを反復的に狭めるドキュメントブラウジング及び検索へのアプローチがここで記述される。テキスト、画像特徴及び使用のような、異なるタイプの情報は、"モダリティ"と呼ばれる。オブジェクトに関連する幾つかのモダリティからのデータを有するオブジェクトのグループ化は、ここでは、マルチモードクラスタリングである。

【００１６】

画像を囲む或いはそれと関連するテキストは、しばしばそのコンテキストの指示を提供する。ここで提案される方法は、（ここで述べられる例示のケースでは、画像の）ブラウジング及び検索を実行するために、テキストと画像の特徴のようなマルチモード情報の使用を可能とする。この方法は、より一般的には、コレクションの要素（例えば、ドキュメント、フレーズ或いは画像）が複数の特性、即ち特徴によって述べられ得る他の用途へ適用出来る。

【００１７】

サーチ及びブラウジングにおける複数の特徴の使用での一つの困難なことは、異なる特徴から情報を組合せることである。これは、一般には、ユーザによって設定され得る、各特徴（通常、カラーヒストグラム、テクスチャー及び形状）と関連する重みを有することによって画像検索タスクで扱われる。重みが変更される毎に、新たなサーチが実行されなければならない。しかしながら、異質セットのマルチモード特徴の使用において、しばしば異なる特徴の重要度へ重みを割り当てることが困難である。メタデータを使用するシステムにおいて、有限で離散的な値を有し、特定の値を含むか排除するブールシステム（系）が使用され得る。この概念を離散的リードではないマルチモード特徴へ拡張することは、特徴の組合せ方の問題を一層悪化する。

【先行技術文献】

【特許文献】

【００１８】

【特許文献１】特開平１０-２７１２５号公報

【発明の開示】

【発明が解決しようとする課題】

【００１９】
従って、種々のコンテキスト及び用途において、フレキシブルにマルチモード情報を処理できるシステムが必要である。照会を実行すると共に、直接テキストコンテント以外の特徴、即ち、画像特徴及び間接特徴によってサーチ結果を詳細化し調節できることは有用である。また、個人ユーザがアクセスするドキュメントの特徴を介してその個人の情報アクセス習性を追跡でき、それによってユーザが類似クラスタへ割り当てられる推奨システムを可能とすることは有用である。
【課題を解決するための手段】
【００２０】
この開示は、マルチモードブラウジングとクラスタリングに対するフレームワークを記述し、ドキュメントのコレクション中のコンテントのブラウジング、サーチ、検索及び推奨を向上するためにそのフレームワークを有利に使用するシステムを記述する。
【００２１】
大きなデータセットのクラスタリングは、探索データ解析、可視化、統計的汎用化、及び推奨システムとって重要である。大部分のクラスタリングアルゴリズムは、オブジェクト間の類似度に依存する。このプロトコルは、データ表示モードとマルチモードデータに対する関連する類似度メトリックを記述する。このアプローチは、各オブジェクトがそれと関連する幾つかの異なるタイプの情報を有するデータセットと関連し、モダリティと呼ばれる。そのようなデータセットの例は、ワールドワイドウエブ（ここで、モダリティは、テキスト、インリンク、アウトリンク、画像特徴、テキストジャンル、等）を含む。
【００２２】
本発明の主な特徴は、その新規のデータ表示モードにある。各ドキュメント内の各モダリティは、ここでは、ｎ－次元ベクトルによって記述され、コレクション内の関連の量的解析を容易にする。
【００２３】
本発明の一用途は、情報をブラウジング及び検索するために異なる空間（即ち、異なるモダリティ）におけるドキュメント特徴を連続して使用するための方法が記述される。この方法の一実施の形態は、画像のブラウジング及び検索のために画像とテキスト特徴を使用するが、本方法は、一般に、あらゆるセットの個々の特徴に適用される。本方法は、ユーザが対象となる項目（アイテム）を指定できる複数の方法の利点を有する。例えば、画像において、そのテキストと画像のモダリティから特徴がその画像を記述するために使用され得る。この方法は、米国特許第５，４４２，７７８号、及び１９９２年の第１５周年Int'l SIGIR'92の会報のD. Cutting、ドクターKarger, J.O.、Pedersen 及び J.W. Tukeyによる"Scatter／Gather: A Cluster-based approach to browsing largedocument collections"（"Scatter/Gather"）に開示された方法と類似する。この"Scatter/Gather"では、クラスタの選択及びその後に続く選択されたクラスタの再クラスタリングが反復的に行われる。それは、少なくとも二つのこと、即ち、各クラスタリングが異なる特徴（例えば、周りのテキスト、画像ＵＲＬ、画像カラーヒストグラム、その周りのジャンル）に実行できること、及び"マップ"機能が指定された特徴に関する最も類似するクラスタを識別すること、に関して、Scatter/Gatherパラダイムを拡張する。後者の機能は、特徴値の欠落に起因して除外されている追加の類似画像の識別を可能とする。画像クラスタは、各クラスタから少数の代表的画像を選択することによって表示される。
【００２４】
本発明の他の用途では、異なるモダリティにおける種々のドキュメント特徴は、全体の類似度を表すクラスタを形成するために、適切に重み付けされ組み合わされる。
【００２５】
また、本発明の種々の代替えの実施の形態は、一つ又はそれより多くの特徴に従ってユーザ及びドキュメントのクラスタリング、ユーザクラスタの従来のブラウジング動作に基づくドキュメントの推奨、及び図形及びテキストでのドキュメント又はユーザのクラスタの可視表示を可能とする。

【００２６】
最初に、ベクトル空間におけるユーザ及びドキュメントの表示を行い、ウエブ画像のコレクション及びＨＴＭＬページ上の関連テキストのブラウジング及び検索を実行するシステムが記述される。ブラウジングは、ユーザによる情報のコレクション又はコーパスの対象となる部分の検出を助けるために、このコーパスに良好にマッチする照合を公式化する必要無く、検索と組み合わされる。画像を囲むテキスト及び幾つかの単純な画像特徴の形態で、マルチモード情報がこの処理で使用される。このシステムを使用して、ユーザは、コレクションを少数の対象となる要素に徐々に狭める。これは、マルチモード特徴を使用するために拡張されることを除いてテキストブラウジングのために開発されたScatter/Gatherシステムと類似する。上述のように、幾つかのコレクションの要素は、幾つかの特徴に対して未知の又は未定義の値を有してもよい。一つの方法は、これらの要素を結果としてのセットへ結合する方法が示される。また、この方法は、サーチが二つのクラスタ同士間の境界近くの空間の一部へ狭められるケースを扱う方法を提供する。多くの例が提供される。

【００２７】
種々のメタデータフィールドと有するデータベースと類似して、本発明のコレクションにおけるドキュメントは、多くの異なる特徴、即ち、多くが非構成ドキュメントのコンテンツから導出される（多分、非直交）"次元"によって特徴付けられる。

【００２８】
マルチモード特徴は、ユーザ情報、テキストジャンル、又は画像解析等の多くのフォームを取り得る。本発明において使用される特徴は、メタデータが一般的に手作業で割り当てられる現在の画像サーチシステムとは異なり、データ（例えば、テキストと画像）及びそのコンテキストから導出される且つ自動的又は半自動的に割り当てられるメタデータのフォームであると考えられ得る。表１は、幾つかの可能な特徴（それらの全ては、より詳細に後述される）を示す。種々の他の特徴及びモダリティもまた本発明において使用でき、且つ表１の特徴は、例示に過ぎない事が理解される。

【表１】

特徴	モダリティ
Text Vector	テキスト
Subject	テキスト
URLs	テキスト
Inlinks	ハイパーリンク
Outlinks	ハイパーリンク
Genre	ジャンル
Page Usage	ユーザ情報
Color Histogram	画像
Complexity	画像

【００２９】
ここでは、豊富な"マルチモード"特徴を組合せてユーザに情報のニーズを満足させる方法が提供される。一方では、これは、アドホック検索（画像に適用される）を含み、ユーザのニーズに関連する情報への簡単で迅速なアクセスを提供する。他方では、これは、解析されるドキュメントコレクションとそれらのユーザを含む。共通のシナリオは、ワールドワイドウエブであり、それは、多くの大量のドキュメントコレクションに典型的な種類の未構成ドキュメントから成る。

【００３０】
従って、この明細書は、ウエブ画像のコレクション及びＨＴＭＬページ上の関連するテ

キストに対する情報アクセスの方法を提供する。この方法は、画像及びそれらの画像に関連するドキュメント又はドキュメント領域のブラウジング及び検索を実行するために、テキストと画像の特徴のようなマルチモード情報の使用を可能とする。記述されたアプローチにおいて、画像の内容を指示する、その画像を囲む又はそれと関連するテキストから導出されるテキスト特徴は、画像特徴と共に使用される。このアプローチの新規性は、ユーザに見えるテキスト及び画像特徴を作る方法に依存し、ユーザが対象となるユーザのサーチを連続的に狭めることを可能とする。これは、特にユーザがコーパスに良好にマッチする照合を公式化することが困難な時、特にコーパスで使用されるボキャブラリや画像記述子が未知である場合にウエブのような一般的でない又は異質コーパスと共に働く時に有用である。

【００３１】

ここで開示される方法は、ドキュメント（とユーザ）特徴が多次元ベクトル空間内に埋め込まれている有利なデータ表示モデルを前提とする。このデータ表示モデルは、一貫性があり対称的類似度尺度の使用を容易とし、それは、以下で詳細に説明される。ここで記述されるデータ表示と類似度モデルによって、ユーザ（即ち、コレクション使用データ）によってアクセスされるドキュメントのコンテンツと特徴に基づいて、ユーザとユーザのクラスタを表示することが出来、それによって、それらの類似度に従ってユーザをクラスタリングする能力を改良する。

【００３２】

更に、マルチモードユーザクラスタに基づく推奨システムは、後述されるように、マルチモードコレクション使用データのコレクションで可能とされる。一セットのクラスタは、ユーザのトレーニングセットから誘導される。推奨を望むユーザは、最も近いクラスタへ割り当てられ、このクラスタの好ましいドキュメントがユーザへ推奨される。

【００３３】

最後に、ここでの開示は、ドキュメントのクラスタ及びユーザのクラスタを可視表示する改良された方法を示す。ドキュメントがしばしば階層的に格納されて、階層可視表示を可能とするが、それは、ユーザに対して通常は真であるとは限らない。従って、本発明は、適切なユーザによってアクセス又はアクセスされ得るドキュメントの階層的ビューを介してユーザデータを見ることが可能となる。ドキュメント及びドキュメントのクラスタは、クラスタの"顕著な次元（ディメンション）"を介して類似的に且つテキスト的に可視化され得る。

【００３４】

画像検索においてクラスタリングの使用は新しいことではないが、クラスタリングは、データベース母集団化ステージの間に人を助けるために、或いは画像をオフラインでクラスタリングして照合同士間の距離サーチがクラスタ内で実行されるように、事前処理のために使用されている。本発明において、反復的なクラスタリング及びクラスタサブセットの選択は、対象となる画像の識別を助けることが出来る。クラスタリングは、反復サーチ及び提示のために使用され、適合フィードバックは、クラスタのユーザの選択において内在する。また、ユーザが個々の画像ではなくてクラスタを扱っているので、フィードバックステップは、実行がより容易である。

【００３５】

ここで述べられるマルチモードクラスタリングの種々のフォームは、情報アクセスのため、即ち、ドキュメントを見つけるためにコレクションをブラウジングするため、ユーザにとって新しいコレクションを理解するため、及び"なにも検出されない"場合（クラスタリングは、ユーザがユーザの照合をコレクションに適するボキャブラリに公式化することによってユーザの照合を再公式化することを助ける事が出来る）を扱うために、使用され得る。

【発明を実施するための最良の形態】

【００３６】

概説的に上記されたように、コレクション中のドキュメントを効率的にブラウジングし

サーチするために本発明のシステム及び方法の能力は、矛盾の無いデータ表示モデルの存在に大きく依存する。具体的には、ドキュメント同士間の定量的類似度メトリックを定義するために、マルチ次元ベクトル空間にドキュメントをマッピングすることが有用であることが判った。従って、ここで記述されるアプローチは、全てのモダリティに対してデータ表示モデルを定義し、ここで各ドキュメントは、R^nとして表される。このモデルは、図１に最も良く図示されている。

【００３７】

図１に示されるように、コレクション１２０から選択された各ドキュメント（例えば、ＨＴＭＬドキュメント１１０）は、各モダリティ（例えば、テキストベクトル１１４とＵＲＬベクトル１１６）毎に一つ、一セットの特徴ベクトル１１２へマッピングする。

【００３８】

特徴ベクトル１１２は、ドキュメントコレクション１２０と通信ネットワーク１２４（インターネットや会社のイントラネットのような）の両方へのアクセスを有するプロセッサ１２２によって計算される。本発明の一実施の形態において、コレクション１２０は、ネットワーク１２４へも接続された一つ又はそれより多くのサーバによってホストされる。各ドキュメント毎の特徴ベクトル１１２は、データベース１２６へ格納され、そこでそれらの特徴ベクトルは、対応するドキュメントと照合される。ネットワークへ接続された複数のユーザ端末１２８、１３０及び１３２は、システムをアクセスするために使用される。

【００３９】

これらの特徴ベクトルは、ドキュメントが初めにそのコレクション１２０に追加された時又はその後に、システムによって発生される。本発明のこの実施の形態において、コレクション１２０は、本発明のシステムによって今までに処理された全ての既知のドキュメントから成ることが観察されるべきである。しかしながら、サーチエンジンの照合の結果に対して即時にコレクションを発生することもまた可能である。次に、極端に大きなグループのドキュメント（ワールドワイドウエブのような）に対してより実践的である、このアプローチは、オリジナルサーチ結果を編成、ブラウジング、ビュー或いは扱うために使用できる。

【００４０】

ドキュメントをコレクション１２０へ追加するこの動作は、図２に示されるように、実行される。初めに、新たなドキュメントが検出される（ステップ２１０）。そのドキュメントは、特徴ベクトル１１２を計算するために処理される（ステップ２１２）。次にそのドキュメントは、本発明で利用できるコレクション即ちコーパスへ追加され得る（ステップ２１４）。もはやドキュメントが無い時（ステップ２１６）、処理は終了する（ステップ２１８）。そうでない場合、他のドキュメントが検出され（ステップ２１０）、処理が繰り返される。

【００４１】

そのシステムのここで示される好ましい演算ベクトルは、八つの可能なドキュメントベクトル特徴、即ち、テキストコンテント、ドキュメントリンク、インリンク、アウトリンク、テキストジャンル、画像カラーヒストグラム、及び画像複雑さ、を使用できる。リストアップされた特徴の最初の二つは、テキストベースであり、インリンクとアウトリンクは、ハイパーリンクベースであり、テキストジャンルは、確率ベースであり、そして最後の二つの特徴（画像カラーヒストグラムと画像複雑さ）は、画像ベースである。これらの特徴は、それらの単純性と理解可能性のために、本発明と共に使用するために選択された。選択された特徴は、情報アクセスにおいて画像とテキストモダリティを使用及び組合せるための本発明方法を図示するように働く。しかしながら、多くの他のドキュメントメトリック（異なる画像領域に対するローカルカラーヒストグラム、画像セグメンテーション、及びテクスチャー特徴から２，３ではあるがネームまで）もまた可能であり、本発明のシステム又は方法の範囲内で開発可能であることが理解される。

【００４２】

本発明の一実施の形態において、これらの特徴ベクトルが図３に示されるように導出される。新たなドキュメント（それらは、テキストドキュメント、画像又は他のタイプの情報であってもよい）コンテンツが分離された（ステップ３１０）後、本発明の方法は、特徴ベクトルを導出するために、種々の情報源を使用する。テキストは、ドキュメントから抽出され（ステップ３１２）、対応するテキストベクトルを生成し（ステップ３１４）、対応するＵＲＬベクトルを生成する（ステップ３１６）ために使用される。

【００４３】

他方（同時に或いはそれに続いて）、全てのアウトリンク（他の場所を指すドキュメント内のハイパーリンク）が抽出され（ステップ３１８）、対応するアウトリンクベクトルを生成する（ステップ３２０）ために使用される。インリンク（主題のドキュメントを指すコレクション内のドキュメント）が抽出され（ステップ３２２）、対応するインリンクベクトルを生成する（ステップ３２４）ために使用される。テキストジャンルが識別され（ステップ３２６）、対応するジャンルベクトルを生成する（ステップ３２８）ために使用される。

【００４４】

その新たなドキュメントが少なくとも一つの画像である或いはそれを含む場合、次に、カラーが画像から抽出され（ステップ３３０）、対応するカラーヒストグラムベクトルを生成する（ステップ３３２）ために使用される。単一のカラー（又は一セットの類似のカラー）の水平及び垂直ランもまた画像から抽出され（ステップ３３４）、カラー複雑さベクトルを生成する（ステップ３３６）ために使用される。

【００４５】

最後に、ドキュメントに対する参照が使用ログから抽出され（ステップ３３８）、ユーザのページアクセスベクトルを更新する（ステップ３４０）ために使用される。

【００４６】

次に、コンテントベクトルの全てがデータベースに格納される（ステップ３４２）。

【００４７】

上述の異なる特徴ベクトルタイプを計算するための本発明の方法が以下に詳細に説明される。

【００４８】

しかしながら、ある特徴を有するドキュメントを既存のコレクションに追加することは、コレクション中の全てのドキュメントに対する全セットの特徴ベクトルを訂正することが必要である場合があることを理解すべきである。例えば、特異なワードを含むドキュメントを追加することは、そのワードが余分の用語を各ドキュメントのテキストベクトルへ追加することを必要とするので、そのコレクション中の全てのドキュメントに対するテキストベクトルに影響を及ぼす。従って、実質的に大きなグループのドキュメント中のコレクションを更新することは、新たなドキュメントが利用出来るようになる毎に増分的に更新するよりも演算的により効率的である。このような考察及びそのセットのベクトルを演算的に最適化する方法が導入されるが、その詳細は、本発明にとって重要ではない。

【００４９】

本発明の一実施の形態において、各特徴が別々に使用され、最も適切な距離メトリックが各特徴に適用され得る。本発明の他の実施の形態において、それらの特徴は、そのドキュメントを表す単一のコンテントテキストベクトルへ組み合わされ、そして単一の距離メトリックがそれらのドキュメントをクラスタリングし比較するために使用される。これらの他の実施の形態は、以下により詳細に記述される。

【００５０】

ドキュメント情報のベクトル空間表示

各タイプの特徴ベクトルの計算は、以下でより詳細に記述される。しかしながら、以下に見られるように、幾つかの一般的な特徴が全ての表示に当てはまる。

【００５１】

テキスト特徴は、図４に示されるように計算される。テキスト特徴は、用語ベクトルで

あり、そこで、そのベクトルの要素は、ドキュメント自体で使用される用語を表す。本発明のこの実施の形態において、全てがテキスト又はＨＴＬＭドキュメント（又はテキストを実際に含む他のドキュメントタイプ）に対して、テキストベクトルは、ドキュメントの全体のテキストコンテントに基づく。そのドキュメントが画像（又は実際のテキストを含まない他のタイプのドキュメント）の場合、テキストベクトルを公式化するために使用されるテキストは、"ホスト"ＨＴＭＬページ中の画像を囲むテキストから導出される。この囲むテキストの範囲は、画像位置の前又は後の８００文字に制限される。水平ルール、ヘッディング又は他の画像がその制限に達する前に生じると、その範囲は、そのルール、ヘッディング又は画像で終わる。"ストップリスト"は、冠詞、前置詞及び接続詞のような殆ど内容の無い共通の用語のインデックス付けを防止するために使用される。

【００５２】

従って、ここで記述される本発明の目的のために、テキストドキュメント、画像ドキュメント及びマルチメディアドキュメントは、全て総称的用語"ドキュメント"の特別のケースであり、これらの特別なケースの各々に対して、ここで記述されるモダリティの幾つか又は全てが適用出来る。例えば、上述されたように、画像は、必ずテキストを含む必要は無いが、それらを指すハイパーテキスト又はＵＲＬ中にテキストによって記述される。テキストを含む画像（ファクシミリビットマップ）は、既知のドキュメント画像デコーディング技術を介して抽出されたそれらのテキストを有してもよい。同様に、オーディオファイルは、またハイパーリンク及びＵＲＬ中のテキストにより参照されることが出来、且つ既知のスピーチ認識アルゴリズムを介して抽出可能なテキストを含むことが出来る。幾つかの用途において、テキスト（及び他の埋め込まれたモダリティ）を引き出すために画像及び他のタイプのデータファイルを処理することは有利であるが、それは本発明にとって必須ではないことが理解されるべきである。

【００５３】

上で提案されたように、ここで記述されるベクトル空間モデルにおいて、各テキストドキュメントｄ（又は、あらゆる種類のテキストを含むドキュメント）が本発明によってR^{n_t}に埋め込まれるが（ベクトル空間がn_t次元を含み、そこで各次元が実数で表される）、ここでn_tは、コレクション中の特異なワードの全数（n_tはテキスト要素の数を表す）である。ベクトル空間への埋め込みは、以下のように定義される。

$\phi_t(d)_i = tf_{di} icf_i$

ここで、ｄは、特定のドキュメントであり、ｉは、ワードのインデックスであり、$\phi_t(d)_i$は、ベクトル$\phi_t(d)$の要素ｉである。トークン頻度重み(tf)と逆コンテキスト頻度重み(icf)は、情報検索において使用される用語頻度重み及び逆ドキュメント頻度重みの汎用化である。それらは次のように定義される。

【数１】

$$tf_{ci} = \log(1+N_{ci}) \text{ and } icf = \log\frac{N}{N_i}$$

ここでN_{ci}は、コンテキストｃ中の要素ｉの発生回数であり、N_iは、ｉが発生するコンテキストの数であり、Ｎは、コンテクストの全数である。テキストモダリティ、ワードの対応する要素、及びドキュメントに対応するコンテキストの場合、この定義は、情報検索フィールドにおける用語頻度重み及び逆ドキュメント頻度重みに対する標準の定義と一致する。

【００５４】

従って、テキストベクトルは、上述のようにトークン頻度重みを最初に計算し（ステップ４１０）、次に、上述のように、逆コンテキスト頻度重みを計算し（ステップ４１２）、次にテキストコンテントベクトルを計算するためにこれら二つを乗算する（ステップ４１４）ことによって、計算される。

【００５５】

本発明によって使用される埋め込みのためのトークン頻度重み及び逆コンテキスト重みの使用は、以下の直感的記述の一致する。コンテキスト（例えば、ドキュメント）中の要素（例えば、又はワード）の各更なる発生は、記述的特徴として、その要素に対する重要性の増加されたレベルを反映する。しかしながら、この増加は、線形である必要はないが、幾分"減衰"される。従来、対数が減衰関数として使用されたが、それはこの用途に対しても満足出来るものであることが判っている。同様に、逆コンテキスト頻度重みは、全てのコンテキスト（一例はテキストドキュメント中のワード"ｔｈｅ"であるかもしれない）中に発生する要素に対する０から唯一つのコンテキスト中に発生する一つの要素に対する最大（ｌｏｇ Ｎ）に達する範囲に亘る。対数スケーリングに対する一つのモチベーションは、情報理論に基づく。ｌｏｇ Ｎ／$Ｎ_{i}$は、コンテキスト中の要素ｉの発生について学習する時にいくら多くの情報が得られるかの尺度として解釈され得る。ワード"ｔｈｅ"が一つのドキュメントに発生することが学習される場合、（それが全てのドキュメントで発生すると仮定されると）重要な情報は得られない。しかしながら、フレーズ"Ｈａｒｒｙ Ｔｒｕｍａｎ"が一つのドキュメントで発生することが学習される場合、（そのフレーズが２，３のドキュメントのみで発生すると仮定すれば）多くの情報が提供される。

【００５６】

逆コンテキスト頻度重みが乗算されるトークン頻度重みは、ベクトルをスケーリングするための有利な方法であることがわかったことが理解されるべきである。しかしながら、他の重み付けスキームもまた可能であり、他の利点を提供し得る。

【００５７】

従って、上述のように、テキストベクトルが計算されると、二つのテキストベクトル間の類似度が簡単なコサイン距離を介して計算され得る。

【数２】

$$\mathrm{sim}_t(d_1,d_2)=\frac{\sum_i \phi_t(d_1)_i \phi_t(d_2)_i}{\sqrt{(\sum_i \phi_t(d_1)_i^2)(\sum_i \phi_t(d_2)_i^2)}}$$

ここで、$ｄ_{1}$と$ｄ_{2}$は、二つの異なるドキュメントを表し、$φ_{t}(ｄ_{1})_{i}$は、ドキュメント$ｄ_{1}$を表すｉ番目の項を表す。以下に更に詳細に議論されるように、ドキュメントの対同士間のコサイン距離は、テキスト特徴単独に基づきドキュメントをクラスタリングするために使用されることが出来又は他の特徴と組合せて使用されることが出来る。

【００５８】

本発明の他の実施の形態において、上述のテキスト特徴は、異なる方法で、或いは別の独立した特徴として、計算され得る。この他のバージョンにおいて、タイトル、ヘディング及びキャプションからのテキストのみが$Ｒ^{ns}$（ここで、$ｎ_{s}$は、コレクション中のドキュメントのタイトル、ヘッダー及びキャプション中の特異のワードの全数である）中の"主題"のモダリティを定義するためにドキュメントから分離される。この他の（又は追加の）モダリティが上述されたテキストモダリティ（ドキュメントの完全なテキストのサブセットのみ殻のものを除き）と正確に同じ方法で引き出されるので、対応する特徴ベクトルと類似度を引出すために使用される公式は同じままである。

【数３】

$$\phi_s(d)_i = tf_{di} icf_i \quad \text{and} \quad \mathrm{sim}_s(d_1,d_2)=\frac{\sum_i \phi_s(d_1)_i \phi_s(d_2)_i}{\sqrt{(\sum_i \phi_s(d_1)_i^2)(\sum_i \phi_s(d_2)_i^2)}}$$

両実施の形態は、有用であることが検出され、必要ならば、交換的に又は共に使用され得

る。例えば、（例えば、タイトル中のワードの各発生を、それがそのテキスト中に２回又は３回発生したかのように、処理することによって）タイトル、ヘッディング及びキャプションテキストをドキュメント中の他のテキストとは異なるように重み付けすることも可能である。一般的な提案として、ドキュメント中の全てのテキストがテキストベースのモダリティの目的のために同じように処理される必要は無いことが認識されるべきである。調節と重み付けが可能であり、幾つかの用途で利点があり得る。

【００５９】

同様に、ベクトルは、ドキュメントのＵＲＬに対して計算され得る。上述の例を詳論すると、その例示のＵＲＬ"ｈｔｔｐ：／／ｗｗｗ．ｓｅｒｖｅｒ．ｎｅｔ／ｄｉｒｅｃｔｏｒｙ／ｆｉｌｅ．ｈｔｍｌ"は、７の用語"ｈｔｔｐ"、"ｗｗｗ"、"ｓｅｒｖｅｒ"、"ｎｅｔ"、"ｄｉｒｅｃｔｏｒｙ"、"ｆｉｌｅ"、及び"ｈｔｍｌ"を含む。テキスト特徴に関して、これらの用語の幾つかは、殆ど又は全く情報的値を含まない（この例では、"ｈｔｔｐ"、"ｗｗｗ"、"ｎｅｔ"及び"ｈｔｍｌ"）。従って、トークン頻度重み及び逆コンテキスト頻度重み埋め込みがここでは十分適切である。再度図４を参照のこと。

【００６０】

この結果、各ドキュメントｄは、R^{nu}（各ベクトル空間はn_u次元を有し、ここで、各次元は、実数によって表される）に埋め込まれ、ここで、n_uは、コレクション中の全てのドキュメントを識別する特異のＵＲＬの全数を表す（n_uは、"ＵＲＬ要素の数"を表す）。ベクトル空間への埋め込みは、以下のように表される。

$\phi_u(d)_i = tf_{di} icf_i$

ここで、ｄは、特定のドキュメントであり、ｉは、ワードのインデックスであり、$\phi_u(d)_i$は、ベクトル$\phi_u(d)$の要素ｉである。トークン頻度重み(tf)及び逆コンテキスト頻度重み(icf)は、情報検索で使用される用語頻度重み及び逆ドキュメント頻度重みである。それらは、以下のように定義される。

【数４】

$$tf_{ci} = \log(1+N_{ci}) \text{ and } icf = \log\frac{N}{N_i}$$

ここで、N_{ci}は、コンテキストｃ中の要素ｉの発生回数であり、N_iは、ｉが発生するコンテキストの数であり、Ｎは、コンテキストの全数である。ＵＲＬモダリティの場合、要素はＵＲＬ用語に対応し、コンテキストは、ドキュメントに対応する。

【００６１】

同様のベクトル埋め込みは、インリンクモダリティ($\phi_l(d)_i = tf_{di} icf_i$)及びアウトリンクモダリティ($\phi_o(d)_i = tf_{di} icf_i$)のために使用される。インリンクベクトルは、$R^{nl}$に存在し、ここで、$n_l$は、コレクション中に埋め込まれた個別のインリンクの全数（即ち、コレクション中のドキュメントの全数は、コレクション中の他のドキュメントに関連する）。アウトリンクベクトルは、R^{no}に存在し、そこで、n_oは、コレクション中に埋め込まれた個別のアウトリンクの全数（即ち、コレクション中又は外において、ドキュメントの総数は、コレクション中のドキュメントによって参照される）。コサイン類似度は、類似的に計算される。

【数５】

$$\text{sim}_t(d_1,d_2) = \frac{\sum_i \phi_t(d_1)_i \phi_t(d_2)_i}{\sqrt{(\sum_i \phi_t(d_1)_i^2)(\sum_i \phi_t(d_2)_i^2)}} \text{ and } \text{sim}_o(d_1,d_2) = \frac{\sum_i \phi_o(d_1)_i \phi_o(d_2)_i}{\sqrt{(\sum_i \phi_o(d_1)_i^2)(\sum_i \phi_o(d_2)_i^2)}}$$

【００６２】

本発明の他の実施の形態において、ＵＲＬ中の用語（上記のように定義されたＵＲＬ埋め込みで使用されるように）インリンク及びアウトリンクから抽出され、そのように使用

された。しかしながら、この他の方法で引出された印リンクおよびアウトリンク特徴に基づくクラスタリングは、類似のドキュメントのクラスタリングにおいて効果は少ないことが判った。

【００６３】

ドキュメントのテキストジャンルは、R^{ng}に埋め込まれ、ここでn_gは、既知のテキストジャンルの数である。ドキュメントジャンルは、ドキュメントの解釈をガイドする文化的に定義されたドキュメントカテゴリである。ジャンルは、ドキュメントテキストではなくて、より大きなドキュメント環境（例えば、ニューヨークタイムズからナショナルエンクイアリを一目で区別するように働く物理的メディア、ピクチャー、等のような）によって合図される。二つの異なるジャンル中に存在する同じ情報は、二つの異なる解釈を導くかもしれない。例えば、ロー"Ａｔ ｄａｗｎ ｔｈｅ ｓｔｒｅｅｔ ｗａｓ ｐｅａｃｅｆｕｌ．．．"から始まるドキュメントは、小説の読者とは異なるようにタイムマガジンの読者により解釈されるであろう。各ドキュメントタイプは、容易に認識され且つ文化的に定義されるジャンル構造を有し、この構造は、それが含む我々の情報の理解と解釈をガイドする。例えば、ニュースレポート、新聞の社説、カレンダー、新聞発表、短編小説は、可能なジャンルの全ての例である。ドキュメントの構造とジャンルは、しばしばドキュメント又はテキストの自動解析によって（少なくとも一部）決定され得る（ステップ５１０）。テキストジャンルは、常に決定出来得るとは限らないが、特に、ウエブページ（それはしばしば良好に定義されたジャンルを有するわけではないが）の場合、多くの既知の可能なジャンルに対して、確率スコアのベクトルを計算することが一般的に可能であり（ステップ５１２）、次に、そのベクトルは、テキスト用語ベクトルに関して上述された方法で、類似度を決定するために使用され得る（コサイン類似度計算を介して）。

【数６】

$$\mathrm{sim}_g(d_1,d_2)=\frac{\sum_i \phi_g(d_1)_i \phi_g(d_2)_i}{\sqrt{(\sum_i \phi_g(d_1)_i^2)(\sum_i \phi_g(d_2)_i^2)}}$$

【００６４】

ベクトル空間に埋め込まれた画像に対して、二つのモダリティ（カラーヒストグラムと複雑さ）が良好に使用された。カラーヒストグラム特徴に対して、画像ドキュメントは、R^{nh}に埋め込まれる。ここでn_hは、ヒストグラム中の"ビン"の数（本実施の形態では１２）である。好ましくは、一つの単一のカラーヒストグラムは、カラー特徴として使用される。この特徴空間は、ＨＳＶ（その色相、彩度、及び明度のカラーモデル）へ変換され、二つのビンは、各次元に割り当てられる（ステップ６１０）。従って、カラー空間に対して三つの次元があり、各カラー次元に対して二つのビン（四つの値）があり、その結果、好ましいベクトル空間には、全１２の次元がある。

【００６５】

次に、処理される画像中の各画素は、カテゴリに分けられ（ステップ６１２）、その色相、彩度及び明度は、各次元毎に４つのビンの一つに入り、対応するベクトル要素が増分される（ステップ６１４）。本発明の好適な実施の形態において、各ドキュメントに対するカラーヒストグラムは正規化され（ステップ６１６）、それによって、ビン値の全てが合計され、その結果がそのヒストグラムベクトルとして格納される（ステップ６１８）。テキスト（及び特定の他の）モダリティに対して行われるのが好ましいトークン頻度重みと逆コンテキスト頻度重み埋め込みを使用することは、このコンテキストでは意味がないので、適切ではない事を理解すべきである。しかしながら、ヒストグラムベクトル間の距離は、次のコサイン距離によって計算されることが好ましい。

【数７】

$$\mathrm{sim}_h(d_1,d_2)=\frac{\sum_i \phi_h(d_1)_i \phi_h(d_2)_i}{\sqrt{(\sum_i \phi_h(d_1)_i^2)(\sum_i \phi_h(d_2)_i^2)}}$$

【００６６】

本発明の他の実施の形態において、ヒストグラム間の距離は、最も大きなビン値による正規化で交点測定を介して計算され得る。

【数８】

$$\mathrm{sim}_h(d_1,d_2)=1.0-\frac{\sum_i \min(\phi_h(d_1)_i,\phi_h(d_2)_i)}{\sum_i \max(\phi_h(d_1)_i,\phi_h(d_2)_i)}$$

【００６７】

本発明の他の実施の形態において、複数のカラーヒストグラムは、各画像の複数の領域に対して決定され、複数のカラーヒストグラム特徴ベクトルを得る。例えば、４象限（左上部、右上部、左下部及び右下部）中のカラーヒストグラムと画像の中心は、別々に計算されることが出来、その結果、５つの別々のカラーヒストグラムベクトルを得、次に、ユーザによって必要ならば重み付けされ且つ組み合わされることが出来又は別々のベクトルとして残されても良い。或いは、上半分、下半分、左半分、右半分及び中心の矩形部分のような部分的に又は完全に重なる領域が使用され得る。効率化のためには、画像は、タイルに細分割され、ヒストグラムが各タイル毎に別々に計算され、次に、適切に領域へ組合せることが出来る。次に、画像の領域的類似度を介して画像を計算することが出来る。例えば、ブルースカイを有する全ての画像がそれらの"上部"カラーヒストグラムベクトルにおける類似度に基づき一緒にグループ化されてもよい。領域的画像類似度を求める他の実施の形態及び用途もまたここで記述される本発明のフレームワーク内で可能であることが理解されるべきである。

【００６８】

これらの距離メトリックは、二つの画像に関して対称的である。対称距離は、一つの画像と他の一つの画像又は図心との間の距離が簡単な検索目的ではなく、クラスタリング目的のために必要なため、本フレームワークにおいて必要とされる。

【００６９】

複雑さ特徴は、人間が画像同士間、一方では高いカラー均一性の領域を有する比較的少数のカラーから成る簡単なロゴと漫画と、他方では微細な陰影を有する比較的多数のカラーを有する写真との間、で行うことが出来る粗い意味的区別を捕獲するためである。この特徴は、画像内でカラーの水平及び垂直ラン長から引き出される。具体的には、同じカラーのラン（本実施の形態では、ステップ７１０において、上述のように、２ビットＨＳＶ値に粗く量子化される）は、ｘ方向（ステップ７１２）及びｙ方向（ステップ７１４）に識別される。ヒストグラムは、各方向毎に計算され（ステップ７１６）、そこで、各ビンは、画素の数（他の実施の形態では、全高さ又は幅の量子化された割合）を表し、ランは、ｘ方向及びｙ方向へ夫々スパンする。各ビンにおけるカウントは、その特定のラン長に属する画像中の画素数である。或いは、各ラン毎にビンへ追加される値は、そのランの長さによって重み付けされることが出来、より長いランへはより大きな重みを与えることが出来る。ヒストグラム中の要素の全数は、夫々画像の水平及び垂直次元における画素数である。従って、二つのベクトル（各ヒストグラムに対して一つ、水平方向及び垂直方向）が生成され（ステップ７１８と７２０）、画像複雑さに対する水平ベクトル及び垂直ベク

トルは、R^{nx}（ここで、n_xは、画像の最大水平画素次元である）とR^{ny}（ここで、n_yは、画像の最大水平画素次元である）に夫々埋め込まれる。

【００７０】

本発明のこの好適な実施の形態において、ラン長複雑さ情報は、少数のビン（従って、各ベクトル毎により少数の次元）に量子化される。これは、ベクトルの希薄性を減少し、画像間のより効率的及びより頑強な比較を可能とするために実行される。Ｎ個のビンとn_xの最大水平次元が与えられると、n_x／４よりも長いあらゆる水平方向ランは、N^{th}（又は最後）ビンに配される。より短いランr_xは、フロアー（r_x（Ｎ－１）／（n_x／４）＋１によってインデックス付けされたビンに配される（ここで、"フロアー"関数は、その項が最も近い整数に丸められる）。従って、ラン長は、Ｎ個のビンに線形に量子化され、n_x／４よりも大きな長さの全てのランは、最後のビンに入る。同様な操作が垂直方向ランに実行され、その結果、水平方向複雑さベクトルは、Ｎ次元を有し、垂直方向複雑さベクトルもまたＮ次元を有する。

【００７１】

以下で述べられるように使用されるコサイン距離メトリックで、ビンの合計を正規化する必要はない。

【数９】

$$\mathrm{sim}_c(d_1,d_2)=0.5\frac{\sum_i \phi_x(d_1)_i\phi_x(d_2)_i}{\sqrt{(\sum_i \phi_x(d_1)_i^2)(\sum_i \phi_x(d_2)_i^2)}}+0.5\frac{\sum_i \phi_y(d_1)_i\phi_y(d_2)_i}{\sqrt{(\sum_i \phi_y(d_1)_i^2)(\sum_i \phi_y(d_2)_i^2)}}$$

ここで、Φ_xとΦ_yは、夫々水平方向複雑さベクトルと垂直方向複雑さベクトルを表す。

【００７２】

或いは、二つのベクトル（水平方向と垂直方向）は、以下の標準コサイン距離メトリックを使用して、R^{nx+ny}（又は量子化された好適な実施の形態では、R^{2N}）中のより大きなベクトルへ追加され得る。

【数１０】

$$\mathrm{sim}_c(d_1,d_2)=\frac{\sum_i \phi_c(d_1)_i\phi_c(d_2)_i}{\sqrt{(\sum_i \phi_c(d_1)_i^2)(\sum_i \phi_c(d_2)_i^2)}}$$

ここで、Φ_cは、追加されたベクトルを表す。

【００７３】

カラー複雑さ特徴とカラーヒストグラム特徴の両方に対して、サブサンプリングがベクトル埋め込みの計算において生ずる演算費用を減少するために実行される。例えば、有用な結果を達成するように、画像中の端数（１／１０のような）又は画素の全数の制限された数（１０００のような）を選択することが可能であることが理解された。これらのサブサンプリングされた画素は、その画像全体に亘って均一に離間されることが好ましいが、他の実施の形態においては、ランダムに選択されてもよい。ヒストグラム特徴に対しては、サブサンプリングされた画素のみに対して適切なヒストグラムを計算するだけで十分である。複雑さ特徴に対しては、サブサンプリングされた画素が属する水平方向及び垂直方向のランの長さを決定することが必要である。本発明の好適な実施の形態において、これは、ロー及びコラムをサブサンプリングすることによって達成される。水平方向複雑さベクトルに対しては、最大５０個の略等しく分布されたローの画素が選択され（画像が高さ方向において５０画素よりも短いならば５０未満）、且つこれらのローのみのランがカウントされる。同様の処理が垂直方向の複雑さベクトルのコラムに対して行われる。そうで

ないものは、ベクトル埋め込みが同じままである。

【００７４】

最後に、（その他のベクトル埋め込みが指示されたように）ドキュメント間ではなくて、ユーザ母集団中のユーザ間の違いをハイライト可能な類似の特徴がある。例えば、ページ使用は、ユーザの情報探求優先度を指示することがわかった。ページ使用モダリティに対しては、ページアクセスは、最初に識別される（ステップ８１０）。好ましくは、トークン頻度重み（ステップ８１２）と逆コンテキスト頻度重み（ステップ８１４）が再び使用され、コンテキストは、各ユーザであり、且つトークンは、ユーザのページアクセスである。積はページ使用ベクトルとして格納される（ステップ８１６）。従って、ページ埋め込みは、$\phi_p(u)_i = tf_{di} icf_i$であり、そこで、ｕはユーザを表し、ｉはページを表す。その結果、埋め込みは、R^{np}になされ、ここでn_pはコレクション中のドキュメントの全数である。他の実施の形態において、各ユーザのページアクセスは、２値とみなされる。それは、ユーザがページにアクセスした、その場合その対応するユーザのベクトルは対応する要素で"１"を有し、或いは、ユーザがページにアクセスしなかった、その場合対応する要素は"０"である。何れの場合においても、以下のコサイン距離メトリックは、（ユーザのページレファランスに関して）ユーザ間の類似度を計算するために使用され得る。

【数１１】

$$\mathrm{sim}_p(u_1,u_2)=\frac{\sum_i \phi_p(u_1)_i \phi_p(u_2)_i}{\sqrt{(\sum_i \phi_p(u_1)_i^2)(\sum_i \phi_p(u_2)_i^2)}}$$

【００７５】

他のモダリティもまたユーザから導出され得る。例えば、ユーザ指定の人口統計情報（名前、年、ホビー、電話番号、ホームアドレス、選択されたグループのメンバーシップ等）及び他の種類のトラックされた情報（限定されるわけではないが、オンライン購買習慣、ソフトウエア使用及びドキュメントを見るために費やされた時間）もまた、スカラー又はベクトル空間に埋め込まれることが出来、（以下に述べられるように）数値距離メトリックの使用及びクラスタリングの実行が可能となる。例えば、ユーザのグループのメンバーシップは、既知のグループの数に等しい次元の数を有するベクトル空間に埋め込まれることが出来、ユーザのグループメンバーシップは、ユーザが対応するグループのメンバーであるか否かを表すブール（"０"又は"１"）を有する。これらの追加の例示のモダリティは、ここでは詳細には述べられない。しかしながら、本発明に係るシステムは、マッピングをベクトル空間に定義することによって、これらのモダリティや、ほぼあらゆるドキュメントベースの又はユーザベースの情報を組み入れるように容易に向上され得る。

【００７６】

各モダリティ毎のベクトル空間中の次元の数は、ファクタの数によって変化し得ることを理解すべくである。例えば、テキストモダリティに対しては、各テキストベクトルは、コレクション中の特異のワードの数に等しい数の次元を有し、画像の複雑さモダリティに対しては、各ベクトルは、コレクション中の最大水平方向と垂直方向画素次元に等しい数の次元を有し、そして、ページ使用モダリティに対しては、各ベクトルは、コレクション中のドキュメントの数に等しい数の次元を有する。従って、ドキュメントがコレクションに追加すると（且つユーザがユーザ母集団へ追加すると）、同じ特徴に対するベクトルの全てが同じ次元を有し、それにより上述の類似度メトリックの使用を可能とすることを確実にするために、特徴ベクトルの多くを再計算することが必要となる。従って、演算費用を減少するために、ある状況では、有意な数のドキュメント（又はユーザ）が追加されるまで、特徴ベクトルのデータベースの更新を遅延することが有利である。勿論、新たなドキュメント（及びユーザ）は、それらが追加され、対応する特徴ベクトルが計算されるま

で、本発明に係るシステムによって認識されない。

【００７７】

種々のモダリティの前述の表示は、本発明のシステムにおいてドキュメントとユーザ間の類似度をトラッキングするために有用であり効率的であることがわっかた。しかしながら、ドキュメント情報のベクトル空間への埋め込み及びドキュメント間の類似度の計算方法の種々の他の方法も可能であることが認識されるべきである。例えば、ドキュメントに対応するテキスト、ＵＲＬ、インリンクテキスト及びアウトリンクテキストを単一の支配的なテキストベクトルへ組合すことが可能である。このアプローチは、画像ドキュメントと関連するほんの僅かなテキストがある時、有用であり得る。また、上述のコサイン類似度メトリックは、一度に、単一の特徴又はモダリティに基づいて、二つのドキュメント間の類似度を計算することが理解されるべきである。また、二つのドキュメント間の集合類似度を計算することが、可能であり、状況下で好ましい。

【数１２】

$$\mathrm{sim}(d_1, d_2) = \sum_j w_j \, \mathrm{sim}_j(d_1, d_2)$$

ここで、ｊは上述の適用可能モダリティを表すと共にその範囲に亘り、w_jは各モダリティに対応する重み付けファクタ（好ましくは、単一ではあるが、必要に応じて調節可能である）を表す。次に、この集合類似度は、全ての可能な（又は実際の）モダリティの基づくドキュメント間の全体の類似度を表す。

【００７８】

全てのモダリティが全てのドキュメント中にあるわけではないことが前述より明らかである。例えば、ウエブ（又は、ウエブのようなイントラネットコレクション）上で、テキスト、画像又は全体として他の何かである、全てのドキュメントは、検索のためにドキュメントを識別するように働く一つの対応するＵＲＬを有する。しかしながら、ドキュメントの全てが画像ではなく、全てのドキュメントが画像ではなく、それにより、ヒストグラムと複雑さメトリックが幾つかのドキュメントに対して不能である。同様に、全てのドキュメントがテキストを含むわけで無いが、（上述のように）テキストは、ある場合（インリンクがある場合）ドキュメント参照から組み立てられる。

【００７９】

従って、集合類似度メトリックは、ある状況で、サブ最適であってもよく、必要な時、個々の類似度メトリック上に"フォールバックする"能力を有することが望ましい。

クラスタリング

【００８０】

集合類似度メトリックを含む、上述の類似度メトリックは、ドキュメントとユーザ（集合的にオブジェクト）をクラスタリングするためのベースを定義する。標準のクラスタリングアルゴリズムが使用される。本発明のこの好適な実施の形態において、"ｋ－平均"クラスタリングがｋ個の異なるクラスタへオブジェクトを割り当てるために使用される。

【００８１】

当該技術で公知のように、ｋ－平均クラスタリングは、通常クラスタセンターとしてｋ個のランダムに選択された複数のオブジェクトで始まる区分方法である。複数のオブジェクトは、クラスタセンター（オブジェクトが最も高い類似度を有するセンター）へ割り当てられる。次に、クラスタセンターはそれらのメンバーの平均として再計算される。オブジェクトの（再）割り当てと平均の再計算の処理は、それが収束するまで数回繰り返される。クラスタの数ｋはその方法のパラメータである。ｋ＝２０及びｋ＝５０の値は、それらの値が良好な結果を与えたので、種々の実施及び研究に使用されたが、との値もまたユーザの優先度に基づき等しい効果を得るために使用され得る。

【００８２】

本発明の他の実施の形態において、階層マルチモードクラスタリングもまた使用できる

が、ｋ－平均クラスタリングが満足な結果を提供することが確認された。

【００８３】

上述のように、ｋ‐平均クラスタリングの旧来のフォームは、クラスタリングされるべきオブジェクトからのランダム選択によって初期クラスタを選択する。初期クラスタを選択するための他の方法は、バックショットアルゴリズムを使用し、階層（しかし演算的に効果である）クラスタリングアルゴリズムをそれらのオブジェクトのサブセットに適用することによって初期センターを計算する。次に、ｋ‐平均クラスタリングに対する初期センターは、そのサブセットをクラスタリングすることにより発見されるクラスタのセンターである。

【００８４】

しかしながら、ランダム選択及び階層サブセットクラスタリングの両方は、マルチモードクラスタリングに対してはサブ最適であることが判った。ドキュメントコレクションでは典型的なベクトル空間は、しばしばその空間の一つの小さな領域に共に集中する大部分のオブジェクトと、他の領域に希薄に分布する他の有意な数ンオブジェクトとを有する。このタイプのデータに対しては、初期センターを識別するためにウエーブフロントクラスタリングが一層効率的であることが判った。このウエーブフロントアルゴリズムは、以下のように且つ図９に示されるように進む。

【００８５】

最初に、ｍ個（クラスタリングされるべきオブジェクトの総数Ｎよりもかなり小さい数）オブジェクトがランダムに選択される（ステップ９１０）。この数は、数ｋ（それは、最終的に計算されるクラスタの数である）から独立している。実験により、ｍの適切な値は、１０であることが判った。

【００８６】

次に、ｍ個のオブジェクトのベクトル図心aaaが計算される（ステップ９１２）。この図心は、当該技術で公知の方法によって、即ち、対象となるベクトルの対応する項を平均することによって、計算される。

【００８７】

次に、トータルでｋ個のオブジェクトbbbがクラスタリングされるべきＮ個のオブジェクトからランダムに選択される（ステップ９１４）。上述のように、ｋは、最終クラスタの望ましい数である。最後に、ｋ個の初期オブジェクトbbbの各々に対して、ｋ個の初期オブジェクトの各々への図心aaa回りにｋ個のクラスタセンターcccを計算する。これらのクラスタセンターは以下のように計算される（ステップ９１６）。

【数１３】

$$\vec{x}_i' = \alpha\vec{c} + (1-\alpha)\vec{x}_i$$

ここで、ｉ＝１．．．ｋに対して、αの適切な値は０．９であることが判った。他の値でも有効である。

【００８８】

この技術は、名称"ウエーブフロントクラスタリング"が与えられた。理由は、単純化された項において、"ウエーブ"は、図心aaaから送られ、且つこのウエーブによって第２のセットのランダムにピックアップされたオブジェクトへの途中でヒットされるオブジェクトが初期クラスタセンターとして選択されるからである。これらの初期センターは、多数のオブジェクトが一つの点に収束する場合に適切である。理由は、図心aaaは、その点へ近接している傾向があるからである。これらの初期センターは、集中する領域を効率的に区分するのに良好に当てはまる。

【００８９】

次に、標準のｋ－平均クラスタリングは、図１０に示されるように、各オブジェクトをその最も近いクラスタへ割り当てられることによって進む。最初に、図９に図示されているようにクラスタセンターを選択した（ステップ１０１０）後、割り当てられなかったオ

ブジェクトが選択される（ステップ１０１２）。その類似度は、上述の類似度メトリックの一つを使用して、各クラスタセンターに関して計算される（ステップ１０１４）。次に、そのオブジェクトは、最も近いクラスタセンターへ割り当てられる（１０１６）。割り当てのためのオブジェクトがもっとある場合、その処理を繰り返す（ステップ１０１８）。次に、クラスタセンターは、各クラスタセンターに対応する各クラスタの図心（平均）として再計算される（ステップ１０２０）。クラスタセンターが、例えば、十分に少数のオブジェクトがクラスタにスイッチされたか否かを決定することによって、十分に収束した（ステップ１０２２）時、クラスタリング処理が終了する（ステップ１０２４）。或いは、全てのオブジェクトは、すべたのクラスタから割り当てが解除され（ステップ１０２６）、処理は、再び新たに決定されたクラスタセンターで始まる。

【００９０】

用途

本発明のシステム及び方法を示すため、マルチモード特徴の二つの用途がここで考察される。（１）マルチモードブラウジング及び検索と呼ばれるシステムにおいて対象となるドキュメントをユーザが識別することを助けることと（２）ユーザのコレクションとの対話のマルチモード解析（コレクション使用解析、即ちＣＵＡ）である。

【００９１】

第１の用途において、上述されたように生成されたドキュメントのクラスタは、ドキュメントのサーチ、推奨及びブラウジングのためのシステムにおいて使用される。第１の用途の第１の実施の形態において、一つの特徴は、ユーザの指定のように、一度に考えられ、第２の実施の形態において、複数の特徴が同時に考えられる。

【００９２】

第２の用途において、上述のように生成されたユーザクラスタは、二つの別々の関数に適用される。第１に、ユーザクラスタは、調停を介して可視化に適するようにされ、これについては、以下でより詳細に記述される。第２に、マルチモードユーザクラスタが推奨を生成するために使用される。

【００９３】

以下、これら二つの用途において、マルチモード情報の使用が記述される。その方法は、このような情報を組合せ且つ例を介してそれらの利点を示す方法を含む。

【００９４】

シーケンシャルマルチモードブラウジング

一度に一つのタイプの特徴を使用してのマルチモードサーチとブラウジングは、図１１乃至２２に関連して最も良く示される。各特徴は、そのセットの画像をリファインするか対象となる画像の関連セットへマッピングするために使用される。従って、画像特徴は、テキスト特徴から独立して使用され、テキストがより適切に知覚される時のテキスト（例えば、セクションヘッディング、アブストラクトタイトル、画像アンカー中の"ＡＬＴ"タグ）の使用と、画像特徴がそのようにより多い時の画像特徴の使用との間を人間のユーザがナビゲート出来る複数のクラスタリングを生成する。

【００９５】

異なる特徴に基づきサーチを徐々に狭めることにの一つの起こり得る問題は、特徴値が欠落した画像が考慮から不注意で削除されるかもしれないことである。例えば、幾つかのドキュメントは、テキストに関係無い画像や、画像のコンテンツに関連しないテキストを画像を含む。具体的には、幾つかの画像は、テキストの無いページ上に存在する。他の場合、画像を囲むテキストはその画像の意味的コンテントに関連しない。サーチを徐々に狭めることの他の問題は、そのサーチが二つのクラスタ間の境界近くの空間の一部へ狭められるかもしれないことである。

【００９６】

特徴の使用は、ここで対象となるセットの要素への迅速な初期フォーカシング、次の編成及び類似の要素を含むための拡張を可能とし、類似の要素の幾つかは、不完全な特徴セットを含んでも良く、或いは他の一つのクラスタに生じてもよい。

【００９７】
ここで提案される方法の幾つかは、画像ブラウジングへの拡張として考えられ得る。理想的なブラウジングシステムは、ユーザが画像を含むドキュメントのブラウジングを可能とし、それらのドキュメントは、記述的注釈テキストを含んでも含まなくていもよく、且つテキスト特徴及び又は画像特徴を使用する。ユーザは、意味的コンテント（"その画像が何を示すか？"）か視覚的コンテント（"その画像がどのように見えるか？"）に基づいて画像コレクションがブラウジングすることを望むことが出来る。画像検索システムは、画像を意味的情報で自動的に注釈することは、現在不可能なタスクであるために、しばしば手作業によるキーワード注釈や画像特徴のマッチングに基づく。そうとしても、手作業によりラベル付けされた画像コレクションは、画像が有することが出来る全ての可能な意味的重要度を含むことは出来ない。
【００９８】
上述のように、ここで記述されたアプローチは、Ｃｕｔｔｉｎｇ等の論文で述べられたＳｃａｔｔｅｒ／Ｇａｔｈｅｒ方法といくつかの点で類似する。
【００９９】
Ｓｃａｔｔｅｒ／Ｇａｔｈｅｒは、元来はドキュメントから引出されるテキスト特徴と共に使用されるようにデザインされている。Ｓｃａｔｔｅｒ／Ｇａｔｈｅｒは、コレクションを少数のクラスタに"分散する"ことによってサーチを繰り返しリファインし、次に、ユーザは、再び分散するために対象となるクラスタを"集める"。Ｓｃａｔｔｅｒ／Ｇａｔｈｅｒ方法は、本発明によってテキスト特徴と画像特徴の両方を使用してテキストと画像を有するドキュメントのコレクションをナビゲートして、拡張される。また、ワーキングセットの外側からの要素がワーキングセットに組み込まれることが出来るように"拡張"（即ち、マッピング）機能がある。
【０１００】
マルチモードブラウジング、推奨及び可視化への本発明のアプローチにおいて、照会に対する正確な解答は、ユーザに依存する。従って、ブラウジングに関連する本発明の一態様において、ユーザは、各ステップで使用される特徴を選択する。ユーザは、現在のワーキングセットを見るのみである。マップ機能が使用されず、各オペレーション後に唯一つのクラスタが選択される時、これは、ユーザがデプスファーストサーチ中のツリーの唯一のノードを拡張することに等しい。クラスタを組合せるために選択することによって、格子が形成される。そして、マップ機能を使用することによって、ワーキングセットの外側からの要素は、ワーキングセットの一部となってもよく、それによって、ツリーも格子も生成されない。従って、本発明の方法は、決定ツリーとは全く異なっている。
【０１０１】
実際、初期テキスト照会は、対象となる候補画像を検出するために使用され得る。次に、対象となる画像を含む戻されたクラスタの幾つかは、更なる考察のためにユーザによって識別される。一つの画像特徴の類似度に基づく拡張によって、システムは、関連するテキストなしに又はユーザ指定の照会に対して十分には類似しないテキストと共に、初期に選択されたクラスタによって表示されるものに類似する画像クラスタを検出及び提示する。このように、拡張機能は、テキスト照会の結果として、オリジナルセット中にない等価画像が識別され且つ含まれることを許容する。拡張機能は、また対象となる特徴空間に近いが、前のステップでの区分に起因して他の一つのクラスタ中にある要素を考察のために識別できる。
【０１０２】
上述のように、本発明のマルチモードブラウジングと検索態様において、事前処理ステップがブラウジングの間に必要な情報を事前演算を行い、そのデータの初期編成を提供するために使用される。一セットの個々の特徴（多分、異なるモダリティからの）は、各ドキュメント毎に事前演算され、ベクトルとして格納される。本出願において、ウエブページ中の画像の特徴は、以下のように演算される。テキスト特徴は、各画像を囲みそれと関連するテキスト、その画像のＵＲＬ，ＡＬＴタグ、ハイパーリンクテキスト、及びテキス

トジャンル（後述される）を含む。画像特徴は、カラーヒストグラムとカラー複雑さのメトリックを含む。上記の表１を参照。ドキュメントは、特徴の各々に基づいてグループにクラスタリングされる。

【０１０３】

画像のサーチのために、ユーザはテキスト照会を入力することで開始する。仮説セッションが図１１に示されており、その図で、円形ノードは、クラスタ中のデータを表し、実線の矢印はノード中のデータの拡散（スキャターリング）又は集合を表し、鎖線は、拡張（マップ）機能におけるように、一つのノード中のサブセットのデータの他のノードへの移動を表わす。事前演算されたテキストクラスタは、コサイン距離を使用して照会用語への等価（即ち、類似度）に関してランク付けされ、最高にランク付けされるクラスタが戻される。これらは、第１のセットの結果１１１０に代表的テキスト又は画像としてディスプレイされ得る。次に、ユーザは、ユーザの対象に最も類似するクラスタを選択する。これは、全て又はサブセットのクラスタ１１１２を含み得る。次に、以下の二つの操作の一方が一般的に実行される。選択されたクラスタ中の画像は、選択された特徴に基づき再クラスタリングが行われ、結果として他のセットの結果１１１４を得る。又は選択されたクラスタが選択された特徴に基づき新たな類似クラスタ１１１６へマッピング（拡張）される。

【０１０４】

何時でも、ユーザは、新たなサーチを開始でき、又は新たな照会（初期テキスト照会のような）を実行することによって既存の結果セット上に動作することが出来る。次に、後者の照会の結果は、ユーザのオプションとして、その既存の結果セットをリファイン又はそれへ追加するために使用され得る。

【０１０５】

新たなクラスタは、選択された特徴がテキストデータから引出されたか又は画像データから引出されたかに依存して、代表的テキスト又は画像としてディスプレイされる。選択された特徴は、事前演算された特徴の何れであってもよい。再クラスタリングすることによって、ユーザは、そのセットの画像をリファインすることが出来る。マッピング又は拡張（即ち、先行するリファインが行われるか否かに拘らず同じ又は類似クラスタにおいて他の類似のドキュメントを識別すること）によって、指定された特徴において類似するが多分他の特徴において値が欠落している画像考察のためにそのセットの画像へ持ち込むことが出来る。

【０１０６】

上述のように、クラスタリングは、予め設定された数のクラスタを有する標準ｋ－平均クラスタリングを使用して実行される。上述の事前演算ステップにおいて、クラスタの数は、ユーザに提示されるクラスタの数よりも大きい。これは、サブセットのクラスタのみが初期テキストストリング照会に応答して提示されるためである。初期テキスト照会を有する本発明の一実施の形態において、１２個のクラスタは、最初に使用されるが、５個の最も類似するクラスタのみがその照会に基づいて戻される。次に、集めるためにユーザによって選択されたクラスタは、再クラスタリングされ、そこで、クラスタの数は、ディスプレイされるべきクラスタの数、本実施の形態では、５個、に等しい。各更なる集合とクラスタリング操作は、５個のクラスタを得る。各操作が実行されると、クラスタの結果は格納される。これは、一連の操作をバックアップすることを可能とし、それはまたマッピング又は拡張操作によって必要とされる。

【０１０７】

初期クラスタリングは、或いはカラーヒストグラム特徴のような一つの他の特徴に基づいてもよい。初期クラスタの適切な数は、特徴に依存してより少なくても良い。この実施の形態において、初期クラスタリングは、テキストに基づくが、何時でも、拡散と更なるクラスタリングは、テキスト特徴又は画像特徴に基づいてもよい。本発明の他の実施の形態において、非テキスト特徴に基づく初期クラスタリングが可能であり、且つそれはある状況において有用であり得る。

【０１０８】
上述のように、拡張／マップ機能は、異なる特徴に基づくサーチを徐々に狭めることの問題を扱い、ここにおいて、値が欠落した画像が考察から削除される。例えば、幾つかのドキュメントは、関連するテキストを有さない画像やその画像のコンテンツに関連しないテキストを含む画像を含む。他のケースでは、画像を囲むテキストは、画像の意味的コンテントと関連しない。サーチを徐々に狭めることの他の問題は、そのサーチが二つのクラスタ間の境界に近い空間の一部へ狭められてもよいことである。
【０１０９】
マッピング又は拡張操作は、画像又はクラスタを一つの特徴次元における類似度に基づき現在のセットへ追加する。唯一つの特徴が一度に考慮されるので、類似度確立に使用される距離メトリックが各特徴毎に異なっても良い事を理解すべきである。例えば、上述のように、コサイン距離は、テキスト特徴類似度のために使用されることが出来、ユークリッド距離又は正規化されたヒストグラム交差がヒストグラム類似度のために使用される。
【０１１０】
拡張操作は、幾つかの方法で実行され得る。一つの方法は、現在のクラスタの要素がマッピングされたセット中に残り、そのセットサイズが増加されることを確実にする。これは、選択された特徴に基づき（適切な距離メトリックを介して）現在のワーキングセットへ近い幾つかの要素をそのワーキングセットへ追加することによって達成される。この好適な実施の形態において、現在のワーキングセットに対して選択された特徴の平均が計算され、次に、この平均に最も近い、全体のデータベースから選択されたこれらの要素（ベクトルとして表される）が追加される。これはテキスト特徴に対して最も適切である。他のバージョンでは、ワーキングセット中の各表示された代表に近い要素が選択され追加される。この他のマッピング手順は、画像特徴により多く適用可能であり、そこで、クラスタは、テキストを表すために使用される要素の収集物の代わりに選択された画像によって表示される。しかしながら、テキストが選択されたドキュメントによって表示される時、マッピングの後者の方法は、また適切である。
【０１１１】
マッピングは、上述されたように、バックアップのためにセーブされた連続したワーキングセットの更に上に提示されたこれらの要素のみを考慮することによってスピードアップされ得る。即ち、マッピングのために選択された特徴がクラスタリングのために使用されるまで、バックアップの連続する操作を参照する。例えば、クラスタリングがカラーヒストグラム特徴に基づき実行され、ＵＲＬ特徴に基づき更なるクラスタリングが続いたと仮定する。カラー複雑さに基づくマップ操作が必要な場合、カラーヒストグラム（他の一つの画像特徴）に基づき選択されたクラスタからの要素が、全てのクラスタではなくて、使用され得る。
【０１１２】
最終の拡張は、特徴のためのデータを有さない要素の全てを含む各特徴毎の特別のクラスタを生成することを含む。マッピングが実行されるべき時、すでに使用された特徴を関連する特別のクラスタ中のこれらの要素のみが候補と考えられ、現在のワーキングセットへ追加される。
【０１１３】
図１１及び上述のカラーヒストグラム／ＵＲＬ特徴の例を参照して、マッピングのための他の一つの（より簡単な）方法は、カラーヒストグラム特徴に基づき最も類似するクラスタを識別することを含む。この方法では、関連するテキストを有さない画像は、それらが適切に関連するテキストを有する画像と類似する場合は、同一である。例えば、あるＵＲＬが情報を有さない（例えば、"ｈｔｔｐ：／／ｗｗｗ．ｃｏｍｐａｎｙ．ｃｏｍ／ｐｒｏｄｕｃｔｓ／ｓｐｅｃｉａｌ／ｉｍａｇｅ．ｊｐｇ"、それは、唯共通の用語"ｗｗｗ"、"ｃｏｍｐａｎｙ"、"ｃｏｍ"、"ｐｒｏｄｕｃｔｓ"、"ｓｐｅｃｉａｌ"、"ｉｍａｇｅ"、及び"ｊｐｇ"を含む）。最初にＵＲＬ特徴を有する画像を識別し、次に、他の一つの特徴において類似する画像へマッピングすることによって、より多くの数の画像がサーチ

を再開したり、特徴重みの使用を必要とすることなく識別され得る。

【０１１４】

Ｓｃａｔｔｅｒ／Ｇａｔｈｅｒのようなクラスタリングスキームを使用する時、ブラウジングセッションの間ユーザに対してクラスタをディスプレイ或いは表示することが必要である。テキストクラスタは、多くの方法で表示されることが出来、最も一般的には、ある意味で、クラスタの最も代表的であるワードのセットの選択とディスプレイである。画像クラスタが表示されることが必要な時、クラスタメンバーに共通する画像特徴を選択しそれらをディスプレイすることは殆ど意味がない。理由は、これらは、一般的に、ユーザにとって意味的な意義を持たないからである。従来のクラスタリング画像ブラウザは、画像をより低い（２）次元空間へマッピングしそのマップをディスプレイすることによって画像クラスタを表示していた。代わりに、本発明の好適な実施の形態は、クラスタの更なるクラスタリングを呼び出し、次に（ａ）そのクラスタの図心に最も近い三つの画像と（ｂ）そのクラスタのサブ領域を代表する三つの画像によってそのクラスタを表示する。その三つのサブ領域の代表は、その三つの最も中心の画像を上記（ａ）から除去し、三つのサブクラスタを計算し、そして各サブクラスタの図心に最も近い画像（適切な距離メトリックを介して測定される）を使用することによって計算される。この表示は、クラスタ図心の向きとクラスタ中の画像の範囲を提供する。また、代表的画像は、多次元スケーリングを使用して２次元ディスプレイ上に配されてもよいが、本発明の例では、それらの代表は、一コラムの三つの"図心"画像又は三つの"サブクラスタ"画像にディスプレイされる（例えば、図１４参照）。これにより、サムネル画像及びオリジナルの複数のコピーのような非常に類似する画像がより容易に識別され得る。

【０１１５】

２３１０個の画像を含むウエブのようなドキュメントのコレクションが、以下の例に対する例示のコーパスとして使用された。ウエブドキュメントは、ドキュメントの走査画像中に発見されることが出来、ドキュメントのコンテントやドキュメント中の構成要素を推理するために使用出来る同じタイプの"メタ情報"の多くを含む。ウエブドキュメントと共に働くことによって、画像中の構成要素及びレイアウトを識別するのに伴う問題が最小化され、且つ検索処理でメタデータを使用するための技術の開発が可能となる。

【０１１６】

ウエブ上に遍在するロゴとアイコンのような"興味のない"画像によってコーパスが支配されることを防止するために、画像が満足しなければならない幾つかの単純で幾分任意の基準がコーパスに含まれるように適用された。何れかの特定のクラスの全ての画像を含むことが必要でも実行される実験のゴールでもなく、ウエブ上で利用出来る興味あるコーパスをアセンブルすることのみが必要で、それによって高い拒絶閾値が意図的に使用された。画像は、少なくとも５０画素の高さと幅を有することが必要であり、少なくとも１００００個のトータル画素を含む事が必要であった。また、画像は、次の幾つかのカラーコンテントベースのテストをパスすることが必要であった。その画像の９０％以下が８カラーから成り、その画像の９５％以下が１６カラーから成り、画像の画素のＲＧＢカラー空間共分散マトリックスが非単一であった。質的に、これらの基準は、その画像が単純なライン図ではなく、詳細に上述されたように、カラー特徴によって十分に区別出来得るような十分に種々のカラーコンテントを含むことを保証する。同じ画像の複数のバージョンに対してスクリーニングが実行されなかった、それによってコーパスは同一の画像と、一つの画像とそのサムネル画像を含まない。

【０１１７】

異なるモダリティにおいて"分散"と"集合"の使用を示す三つのサンプルセッションが以下に述べられる。第１の例は、最初にコレクションを狭めるためにテキスト特徴を使用し次に結果を編成するために画像特徴を使用することを示す。最初に、図１２を参照して、ユーザがテキスト照会"ａｎｃｉｅｎｔ　ｃａｔｈｅｄｒａｌ（昔の大聖堂）"１２１０をタイピングし、"実行"ボタン１２１２を押すことによって開始する。ここで述べられるようにシステムとのユーザの対話が公知の方法で、例えば、実際の物理的ボタンとの相互作

用により、マウスのようなポインティングデバイスと有するボタンのスクリーン上の表示の操作により、あまり多くはないがネームへのボイスコマンドより起こる。本発明のこの好適な実施の形態では、ユーザは、本発明を実施するソフトウエアプログラムによってウインドウ１２１４として提供されるマルチモード画像ブラウザと対話する。

【０１１８】

５つの戻されたテキストクラスタ１２１６、１２１８、１２２０、１２２２及び１２２４をディスプレイするスクリーンのスナップショットが図１２の左半分に示されている。これらのクラスタは、照会用語に最も近いクラスタである。各クラスタにおける最頻度コンテント用語は、各クラスタを表示するようにディスプレイされる。ユーザは、テキストクラスタに対する追加の表示用語を見るために各テキストウインドウをスクロール出来る。ユーザは、テキストに基づいて、用語"ａｃｉｅｎｔ（昔の）"と"ｃａｔｈｅｄｒａｌ（大聖堂）"を含む第１のテクストクラスタを再び分散することを決定する。そうするために、ユーザは、望ましいクラスタの次のチェックボックス１２２６を選択し、引き続いて"テキストクラスタ"ボタン１２２８を押す。上述のように、これによって、そのシステムが既存の選択されたクラスタをより小さな別々のクラスタにリファインする。

【０１１９】

５つの結果のテキストクラスタ１３１０、１３１２、１３１４、１３１６及び１３１８をディスプレイするスクリーンのスナップショットが図１３の左半分に示されている。ユーザは、用語"ａｃｉｅｎｔ"、"ｃａｔｈｅｄｒａｌ"と"ｃｈｕｒｃｈ（教会）"を含む３つのクラスタを選択して、（対応するチェックボックス１３２０、１３２２及び１３２４を介して）集め且つ（"複雑さクラスタ"ボタン１３２６を押すことによって）分散のための特徴として複雑さを選択する。

【０１２０】

画像の複雑さに基づくクラスタリング後のスクリーンのスナップショットが図１４に示されている。図心ン最も近い表示画像がディスプレイされる。各画像クラスタの次の矢印（例えば、第１の画像クラスタ１４１４に対応する左矢印１４１０と右矢印１４１２）をクリックすることによって、ユーザは、図心とサブクラスタ表示ビューとの間を移動出来る。画像クラスタ１４１４、１４１６及び１４２０は、古い教会及び大聖堂を含む、"ａｎｃｉｅｎｔ"ビルディングとモニュメントを主とする画像を含む。画像クラスタは、１４１８は、ロゴを含み、画像クラスタ１４２２は、種々雑多なアイテムを含むように現れる。

【０１２１】

第２の例において、我々の仮説ユーザは、我々のコーパス中に多くのｐａｐｅｒ ｍｏｎｅｙ（紙幣）の画像を見つけ出すように試みている。図１５に示されるように、"ｐａｐｅｒ ｍｏｎｅｙ"の初期照会が与えられ、且つ結果としてのテキストクラスタ１５１０、１５１２、１５１４、１５１６及び１５１８がディスプレイされる。第１のテキストクラスタ１５１０は、ワード"ｍｏｎｅｙ（お金）"とワード"ｎｏｔｅ（紙幣）"を含む。第２のテキストクラスタ１５１２は、ワード"ｐａｐｅｒ（紙）"を含むが、それを囲むワードは望ましい意味のワードｐａｐｅｒが使用されていることを示さず、それによってこのクラスタは選択されない。お金が多くのカラーでプリントされるので、最初に画像特徴としてカラー複雑さメトリックが使用されるのが適切である。従って、第１のテキストクラスタ１５１０は、カラー複雑さ特徴に基づいて、第１のテキストクラスタ１５１０が分散され、結果としてのクラスタが図１６に示されている。画像クラスタ１６１４と１６１８は、紙幣の画像を含み、それによってそれらは、（両方のクラスタを選択することによって）集められ、次にこの時にカラーヒストグラムに基づいて分散される。他の画像クラスタ１６１０、１６１２及び１６１６は、対象となる画像を含むようには現れず、それによってユーザはこれらを選択しない。

【０１２２】

結果の画像クラスタは図１７に示されている。画像クラスタ１７１２は、１４個の画像を含み、且つ中央の代表例は、紙幣の全ての画像である。このクラスタは、ヒストグラム

特徴に基づき再び分散される。図１８に示されているように、それが紙幣の多くの画像を含むことが観察され得る。画像の幾つかは、複製であるように現れるが、この場合、それらは実際には一つのサムネル画像と一つのフルサイズ画像である。サブクラスタ表示の審査は、マネー（お金）を含まないサブクラスタ中の幾つかの画像を表すが、それらは、マネー画像と類似するカラーを有する。

【０１２３】

この例は、選択的にそのセットの画像を一セットの対象物に狭めるための一連のコンビネーションにおける異なる特徴の使用を示す。分散は、より大きなコレクションをより小さなサブセットへ編成することを助けるために使用される。集合することによって、異なるコレクションが組み合わされ且つ一緒に再編成される得る。

【０１２４】

最後の例において、図１９に初めに示されているように、ユーザは、照会"ｐｙｒａｍｉｄ ｅｇｙｐｔ（ピラミッドエジプト）においてピラミッドとタイプをサーチしている。戻されたテキストクラスタ１９１０、１９１２、１９１４、１９１６及び１９１８がディスプレイされる。ユーザは、複雑さ特徴に基づき分散されるべき第１のテキストクラスタ１９１０を選択し、結果としての画像クラスタからの代表画像が図２０に示されている。ユーザは、第２及び第４の画像クラスタ２０１２と２０１６に石を含むアウトドアシーンがあることに気付き、カラーヒストグラム特徴に基づき、更なるクラスタリングのためにこれらを選択する。結果としての画像クラスタが図２１に示されている。第１の画像クラスタ２１１０は、４つの画像を含み、且つ第１の画像はピラミッドの画像である。

【０１２５】

第１の画像クラスタ２１１０が、（第１の画像クラスタ２１１０を選択し、且つ"ヒストグラム拡張"ボタン２１２０を押すことによって）カラーに基づいて類似画像を含むように拡張されると、図２２に示されているように、ピラミッドの他の画像２２１０が識別される。この画像は、テキスト無しで且つ意味のないＵＲＬと共にウエブページ上に発生し、それによって、カラーヒストグラム特徴に基づいてそれが検索された。

【０１２６】

この例において、テキスト照会は、画像コレクションのサイズを減少するために使用され、且つ減少されたコレクションは、画像複雑さ特徴に基づきプレゼンテーションのために編成された。カラーヒストグラム特徴次元で類似する追加の画像が得られた。

【０１２７】

これらの例において、異なるモダリティにおける特徴は、コーパス中のサブセットの要素の"分散"と"集合"の技術を使用して、関連するテキストを有する一セットの画像をユーザがブラウジングするのを連続して助けるために使用される。セッションは、テキスト紹介で始まり、全体のコーパスよりも多くの焦点を当てられた初期セットで開始する。一つ又はそれを越える興味ある要素を含むことが観察されるクラスタは、それらのコンテントを審査するために分散されるか、又は全体のコレクションから類似の結果を検索するために拡張される。上述の例（図１２乃至図２２）は、三つの特徴タイプ、即ち、テキスト、画像ヒストグラム及び画像複雑さ、のみを使用したが、本発明の方法は、ここで記述される全８つのモダリティ、及び他にも等しく適用出来る。

【０１２８】

従って、本発明の一態様は、複数のモダリティを利用してコレクションをブラウジングするためのシステムを含む。クラスタを審査するために要素を"集合"且つ"分散"する繰り返し処理を介して、ユーザが対象となる画像のグループを検出出来る。"拡張"又は"マップ"機能は、一つ又はそれより多くの次元において値を欠落しているが対象となるある次元における他の要素に類似するコレクション中の要素の識別することを可能とする。

【０１２９】

集合マルチモードブラウジングまた、上で提案されたように、クラスタリングと拡張操作のために距離メトリックの種々の組合せを使用することが可能である。

【０１３０】

上述された例示のシステムと方法を使用してこれを実施するために、二つのドキュメント又はオブジェクト間の集合類似度ｓｉｍ（d_1，d_2）が、前述のセッションで記述された集合、分散及び拡張操作で使用され得る。図１２乃至２２に示されるユーザインターフェースへの僅かな変更は、この追加の特徴を許容する。例えば、"集合クラスタ"と"集合拡張"ボタンが、全ての可能なモダリティを同時に操作することを容易にするために追加出来、或いは可能なモダリティのリストアップ（テキスト、カラー複雑さ、カラーヒストグラム等）は、"クラスタ選択モダリティ"又は"拡張選択モダリティ"ボタンが起動されると、一つのモダリティが使用されるべきか、一度に複数のモダリティが使用されるべきかをユーザが指示することが出来るようにチェックボックス（及び任意ではあるが、ユーザ調節可能重み）が設けられ得る。次に、選択されたモダリティ上に集合類似度ｓｉｍ（d_1，d_2）が分散とマッピング機能のために使用される。

【０１３１】

マルチモードコレクション使用解析

ユーザをユーザの癖に従ってクラスタリング使用とすると困難がある。幾つかのケースにおいて、ウエブサイトのユーザをクラスタリングするために利用出来る唯一の直接情報は、ユーザが何れのページを如何なる頻度でアクセスするかということである。残念ながら、これは、類似度を決定するためには十分な情報がないので、しばしば相互に排他的なページビューでユーザをクラスタリングすることが不可能となる。

【０１３２】

このタイプの状況でマルチモードクラスタリングを可能とするために、媒介マルチモード表示がマトリックス操作によって計算される。例えば、Ｐがn_p個のロー（ページの全数）とn_u個のコラム（ユーザの全数）を有するページアクセスのマトリックスであるとする。各コラムは、関数ϕ_pによって発生されるベクトルに対応し、その微分は詳細に上述している。例えば、ユーザ番号５に対応する第５コラムは、$\phi_p(u_5)$である。Ｔをn_p個のコラム（ページの全数）とn_tコラム（ワードの全数）を有するテキストマトリックスとする。上述のように、各コラムは関数ϕ_tによって生成されるベクトルに対応する。例えば、ドキュメント番号７に対応する７番目のコラムは$\phi_t(d_7)$である。次に、ユーザのテキスト表示が以下のように計算される。

$$P_T = T \cdot Pn_t$$

ローとn_uコラムを有するマトリックスであるこのマトリックスの内積は、各ユーザがアクセスしたページのテキストコンテントの重み付け平均として解釈される。或いは、他の方法として、P_Tは、アクセスされたページのコンテンツへのページアクセスの補外として解釈されてもよい。

【０１３３】

このアプローチの有用性の例として、個人の複写機ＸＣ５４０を記述するページにアクセスした唯一のユーザの例を考える。モノモードクラスタリングがページアクセスに基づいてのみで実行される時、このユーザの類似度を他のユーザでアクセスすることは実際的ではない。理由は、このユーザは、このページをアクセスした唯一のユーザのためである。また、ユーザは、積P_T＝Ｔ・Ｐによって計算されるように、テキストモダリティに基づいて表される時、ユーザは、ＸＣ５４０ページ上に生じる"リーガルサイズ"又は"ペーパトレイ"のようなワードによってP_Tに表される。このユーザのテキスト表示（P_Tにおける単一のコラムによって定義されるベクトル）複写機ページをアクセスする他のユーザのテキスト表示に類似する。そして、上述のように、コサイン距離メトリックは、クラスタリング目的のために、P_Tにおけるユーザ間の類似度を決定するために使用され得る。この例は、媒介表示が類似度アセスメントとクラスタリングにおいて、どのように助けになるかを示す。

【０１３４】

更なる例によって、インリンク、アウトリンクとＵＲＬモダリティは、類似に計算される媒介によって表示可能である。ここで、マトリックス乗法は、Ｌ・Ｐ（インリンク）、Ｏ・Ｐ（アウトルック）及びＵ・Ｐ（ＵＲＬ）であり、ここで、Ｌ、Ｏ及びＵは、夫々イ

ンリンク、アウトリンク及びＵＲＬに対するマトリックスである。この概念は、ドキュメント当りベースに基づき計算されたあらゆる望ましいモダリティ又は特徴と共に、テキストジャンル、カラーヒスグラム及びカラー複雑さのような、その他のモダリティに対しても拡張され得る。

【０１３５】

従って、ユーザがドキュメントコレクションとどのように対話するかを解析するためのマルチモード技術が可能である。この処理は、コレクションユーズ解析（ＣＵＡ）と呼ばれる。ライブラリの編成及び解析において大きな文学があるが、これは、ディジタルコレクションのための調査中の領域である。大部分の既知の従来の作業において、コレクションは、ユーザニーズの特徴付け（例えば、汎用クラスタリングによる）を行うこと無しに編成される。このセッションにおいて、実際のコレクション使用の解析は、コレクションの編成がどのように改良され得るか及びコレクションの何れの部分がユーザ母集団の特定のセグメントに対して最も価値があるかのような問題をどのように通報するかを示す。

【０１３６】

これらの問題は、ウエブコレクションの豊富なハイパーリンク構造及びそれらの商用的重要性（これらの両方は、優れたコレクションデザインを必要とする）のために、ワールドワイドウエブのコンテクストにおいて特に重要である。表１（上記）にリストアップされたモダリティの内、以下の情報は、ページ及びユーザ：ＵＲＬ、アウトリンク、インリンク及び使用ログ、を特徴付けるために本発明の好適な実施の形態において使用される。上述のように、この情報の利用可能性は、ＣＵＡへのマルチモードアプローチに動機付けを行う。全ての可能なモダリティから利用出来る情報を利用し組合せることが出来ることが望ましい。

【０１３７】

ここで記述されるＣＵＡのために使用される主な技術は、ユーザのマルチモードクラスタリングであるが、これらのクラスタを解釈しようとする問題が残る。アブストラクトにおいて、クラスタのオブジェクトは、テキスト、使用、コレクショントポロジー（インリンクとアウトリンク）及びＵＲＬの特徴上のそれらのオブジェクト間の類似度によって特徴付けられる。オブジェクト間のこれらの特徴の類似度を明らかにするために、種々のユーザインターフェースと可視化技術が使用される。

【０１３８】

ディスクツリー（後述される図２３）は、ウエブサイトのページとハイパーリンクトポロジーを可視化するために使用されることが出来、且つ典型的にユーザの種々のクラスタに興味を持たせるサイトの部分を識別するのに有利であることが判った。また、ユーザのクラスタの興味の特徴を表すテキストとＵＲＬを要約するための技術は、本発明によって使用される。このような技術を組合せることによって、解析者には、自動的に識別されたクラスタの興味を特徴付けるテキスト、トポロジー及びＵＲＬの識別が提供され得る。

【０１３９】

以下に述べられる例の実行において使用されるテストベッドは、１９９８年５月の１７日と１８日の２４時間でのゼロックスウエブサイト（ｈｔｔｐ：／／ｗｗｗ．ｘｅｒｏｘ．ｃｏｍ）の完全なスナップショットより成った。約６４００人のユーザに対する丸１日の使用情報が収集された。ユーザは、ブラウザクッキーに基づいて識別された。更に、全体のテキストとハイパーリンクトポロジーが抽出された。スナップショット時に、サイトは、６０００以上のＨＴＭＬページと８０００の非ＨＴＭＬドキュメントから成った。

【０１４０】

テストベッドは、３つの主な要素から成った。即ち、モード情報を実数ベクトル（Ｒｎに埋め込まれた）にマッピングするマッピングプログラム、複数のセットのユーザをクラスタリングするクラスタリングプログラム及びウエブサイトの対話データ可視化を処理する可視化プログラムである。可視化プログラムは、ウエブサイトのディレクトリ構造を解析して、図２３に示されるように、ディスクツリーを構成出来た。図示されているように、ウエブサイトの各ディレクトリは、全てのサブディレクトリを有するツリーの一つのノ

ードに対応し、ディレクトリ中のファイルは、そのノードの子として表される。好ましくは、そのツリーのレイアウトは、幅優先ベースで実行される。

【０１４１】

従って、本発明の実施の形態において使用される可視化システムは、図２３に示されるように、ディスクツリーを構成し、ウエブサイトのベーシックトポロジーを表す。各ディレクトリは、一つのノードの子達として表されるディレクトリ中の全てのサブディレクトリとファイルを有するツリー中のそのノードに対応する。このツリーのレイアウトは、幅優先ベースで実行される。図２３中のディスクツリー２３１０は、ゼロックス"スプラッシュページ"（ｈｔｔｐ：／／ｗｗｗ．ｘｅｒｏｘ．ｃｏｍ／）から開始する、ゼロックスウエブサイトを示し、引き続くディレクトリは、ディスクの中心から延出する同心リンクとして描かれる。これは、非対称ディスクを生成する。

【０１４２】

このディスクツリーは、アナリストユーザにクラスタについてのトポロジー情報へアクセスする方法を提供する。

【０１４３】

ディスクツリー２３１０において、クラスタは、そのクラスタの数に対応する全てのセグメントを一つのカラーに着色することによって可視化される。例えば、本発明の好適な実施の形態において、クラスタ中のメンバーシップは、そのクラスタ中のドキュメントに対応するセグメント２３１２、２３１４及び２３１６をレッド（図中では、太線で指示されている）で着色されることによって指示され得る。更に、好ましいシステムは、複数のメンバーシップの可視化を可能とする。これらの場合、複数のメンバーシップは、例えば、"レッドクラスタ"と"ブルークラスタ"中の同時メンバーシップを指示するためにレッドとブルーのストライプでセグメントの一つのグループ２３２０を着色することによって、そのページが属する全てのクラスタのカラーをミックスすることによって簡単に指示される。

【０１４４】

また、ダイアログボックスインターフェース（図２４）を介して、本発明の好適な実施の形態のユーザは、何れのクラスタを表示するかを対話的に指定できる（本例では、同時に一つ又は二つのクラスタに制限される）。このダイアログボックスは、各クラスタのメンバーのテキスト表示をディスプレイする。各クラスタメンバーに対して、各モダリティの重みがリストアップされる。インリンク、アウトリンク、テキスト及び使用モダリティは、等しく重み付けされる（各々２５％）。"クラスタリングレポート"２４１０は、ユーザクラスタに対する全てのドキュメントに亘って最も特徴的なキーワード２４１２を含む。これによって、他の特性を見ると同時にこのモダリティのハイレベルアブストラクションへの素早いアクセスが可能となる。"ドキュメントレポート"２４１４は、ＵＲＬとそのクラスタ中の最も特徴的なドキュメントのテキストサマリー２４１８を提供する。多次元クラスタリングを有する経験は、幾つかの場合、クラスタリングレポートがそのクラスタの最良の特徴付けであり、また他の場合、ドキュメントレポートが最良の特徴付けを提供する。そのシステムとの対は、全体のクラスタ又はその最も代表的なドキュメントのサマリー又はその両方を容易にアクセス出来ることによって多いに促進される。

【０１４５】

マルチモードクラスタリングの結果は、各モダリティに対するクラスタの最も特徴的な次元のテキストのリストアップである。例えば、クラスタが、Ｘｅｒｏｘ ＨｏｍｅＣｅｎｔｒｅ ｐｒｏｄｕｃｔ（ゼロックスホームセンタープロダクト"について（ａｂｏｕｔ）"の場合、テキストモダリティに対する顕著な次元は、ワード"Ｈｏｍｅｃｅｎｔｒｅ（ホームセンター）"である。それがテキストベッドゼロックスウエブサイトに与えられると、２０から５０個のクラスタが生成した。各クラスタは、数百のユーザを含み、テキストフォームのクラスタ結果の識別、比較及び評価のタスクは驚くべきものである。この場合、ディスクツリー（上述の）は有用である。

【０１４６】

図２４に示されるように、クラスタレポートウインドウ２４１０は、ユーザクラスタに対する全てのドキュメントに及ぶ特徴的キーワード２４１２を含む。これらは、クラスタのテキスト図心（テキストベクトルが図心を表す）中の最も高く重み付けされたワードを選択することによって計算される。このようなサマリーは、ユーザに大きなクラスタのテキストの信頼できるアセスメントを提供することが判った。

【０１４７】

ドキュメントレポートウインドウ２４１４は、ＵＲＬ２４１６と最も特徴的なドキュメント（クラスタのテキスト図心に最も近いドキュメント）のテキストサマリー２４１８を提供する。それと共に、クラスタレポートウインドウ２４１０とドキュメントレポートウインド２４１４は、アナリストユーザに他のモダリティを見ると同時に、テキストモダリティのハイレベルアセスメントとＵＲＬを提供する。

【０１４８】

図２４のダイアログボックスインターフェースの残りの部分は、何れのクラスタが表示されるかを指定するために使用される。ダイアログボックスは、テキストを使用して、各クラスタのメンバーを表示する。各クラスタメンバーに対して、各モダリティの重み２４２０がリストアップされ（本図に示されるクラスタリングは、五つのモダリティの内の四つに対して行われた）、本発明の好適な実施の形態においては、ユーザによって調節され得る。例えば、図２４において、／ｉｎｖｅｓｔｏｒ／ｐｒ／ｉｒ９８０５１２．ｈｔｍｌは、クラスタ０のメンバーとして示される。インリンク、アウトリンク、テキスト及び使用モダリティは等しく重み付けされる（各２５％）。

【０１４９】

図２４のダイアログボックスと図２３のディスクツリーにユーザを直接示しの代わりにページを示す一つのモチベーションは、ユーザが、ページがされるのと同じ方法で、構造的且つ階層的に編成されない場合であり、その場合ユーザの直接可視化が困難となる。

【０１５０】

従って、クラスタを提示するための二つの方法が提案される。第１の方法は、そのクラスタの全てのメンバーの可視提示から成る。上述のディスクツリーに基づき組み立てる場合、この方法は、メンバーが埋め込まれている階層構造がある場合、直進的である。例えば、ページのクラスタは、そのクラスタのメンバーに対応するディスクツリー中の全てのノードを着色することによって示される。

【０１５１】

調停を介して表示されるオブジェクトのクラスタリングを等しく直進的に示す方法は無い。ディスクツリーとして可視化され得るユーザの直接階層的編成はない。従って、次に、技術的問題は、ウエブページベースの可視化においてユーザクラスタをどのように示すかである。この問題は、ランダムなユーザが望ましいクラスタから選択される場合、特定のページがアクセスされる確率を計算することによって解決される。確率Ｐ（ｐ｜ｕ）は、ページｐがユーザｕによってアクセスされる相対頻度として計算される。例えば、ユーザが三つのページをアクセスする場合、その三つのページの各々は、１／３の確率Ｐ（ｐ｜ｕ）を有する。次に、確率Ｐ（ｐ｜ｃ）、即ち、クラスタｃ中の何れかのユーザによってページｐがアクセスされる相対頻度は、以下のように、そのクラスタ中のユーザに対する確率Ｐ（ｐ｜ｕ）の平均として計算される。

【数１４】

$$P(p|c)=\sum_{u\in c}\frac{1}{|c|}P(p|u)$$

ここで、｜ｃ｜はクラスタｃ中のユーザの総数である。この可視化は、"密度プロット"と考えられ得る。直感的に、それは、このクラスタからの典型的なユーザが直面し得る問題に答える。本発明のこの好適な実施の形態において、全ての非０確率は、０．３から１．

０のスケールにマッピングされ、それによってクラスタ中のユーザによって２，３回しかアクセスされない偶数ページが明瞭に可視である。

【０１５２】

ユーザ母集団を解析するために、テストベッドの６４００人のユーザ全てが２０個のクラスタへクラスタリングされた。ユーザクラスタの内の９個は、ゼロックスプロダクト提供、例えば、Ｐａｇｉｓ走査、複写、ＸＳｏｆｔソフトウエア、ゼロックスソフトウエアライブラリ（プログラムをダウンロードするため）、ホームとデスクトップ製品、及びウインドウズ（登録商標）のためのＴｅｘｔＢｒｉｄｇｅ、への興味によって特徴付けられた。７個のユーザクラスタは、一つの単一ページ、例えば、ドライバのインデックスやゼロックスホームページ、をアクセスした。ユーザの一つのクラスタは、使用情報をアクセスした。一つのクラスタは、ゼロックスについてのプレスリリースとニュースのような投資情報への興味によって特徴付けられた。二つのクラスタは、その他のカテゴリーに何れにもうまく当てはまらないユーザを含むようにミックスされた。従って、再び、図２３を参照して、本発明の好適な実施の形態において、種々のセットのドキュメント２３１２、２３１４および２３１６は、カラーでハイライトされ、ユーザの特定のクラスタ（又は複数のクラスタ）がアクセスする見込みのあるドキュメントを指示する。

【０１５３】

クラスタを提示するための第２の方法において、テキストベースのクラスタサマリーは、各モダリティ毎に最も顕著な次元を提示することによって発生される。一例が、ゼロックスホームセンター（Ｘｅｒｏｘ ＨｏｍｅＣｅｎｔｒｅ）に興味のあるユーザのクラスタに対して表２に示されている。各モダリティ毎に、１０個の最も顕著な次元がリストアップされる。１０個の最も顕著なワード、このクラスタによってアクセスされたページを指す１０個の最も顕著なページ、アクセスされたページ上に生じる１０個の最も顕著なアウトリンク、アクセスされた１０個の最も顕著なページ、及び１０個の最も顕著なｕｒｌエレメントである。クラスタ（この場合、ユーザ）中にあるオブジェクトのみに基づいてクラスタを解釈し比較することは、大変なタスクである。顕著な次元によるテキストサマリーは、クラスタ及びユーザが同じクラスタに置かれた理由を理解するのを一層容易にする。

【表２】

テキスト		
0.504	8332	homecentre
0.221	14789	detachable
0.171	15270	artist
0.162	5372	slot
0.155	12010	mono
0.142	21335	photoenhancer
0.122	237	foot
0.121	4605	creative
0.113	3533	projects
0.109	21336	pictureworks

インリンク		
0.343	23856	products/dhc/index.htm
0.265	24144	products/dhc/06does.htm
0.259	17045	soho/whatsnew.html
0.257	24155	products/dhc/13inclu.htm
0.240	24151	products/dhc/07buser.htm
0.240	24152	products/dhc/07cuser.htm
0.235	24143	products/dhc/12more.htm
0.235	24157	products/dhc/15supp.htm
0.235	24156	products/dhc/14req.htm

アウトリンク		
0.527	24143	products/dhc/12more.htm
0.272	24156	products/dhc/14req.htm
0.272	24155	products/dhc/13inclu.htm
0.272	24157	products/dhc/15supp.htm
0.255	24149	products/dhc/11pagis.htm
0.248	31814	http://www.teamxrx.com/retailers.html
0.216	24145	products/dhc/07user.htm
0.216	24144	products/dhc/06does.htm
0.192	23856	products/dhc/index.htm
0.137	23857	products/dwc450c/index.htm

ページ		
0.557	37067	products/dhc
0.330	24143	products/dhc/12more.htm
0.303	19452	products/multiprd.htm
0.287	24144	products/dhc/06does.htm
0.274	24739	soho/dhc.html
0.233	24155	products/dhc/13inclu.htm
0.208	24156	products/dhc/14req.htm
0.191	24148	products/dhc/09scan.htm
0.184	24157	products/dhc/15supp.htm
0.176	24145	products/dhc/07user.htm

url		
0.791	15	products
0.583	2036	dhc
0.141	646	soho
0.057	2037	dwc450c
0.054	895	print
0.044	31	cgi-bin
0.042	603	supplies
0.036	1768	usa
0.027	91	xps
0.020	844	wwwwais

【０１５４】

所与のモダリティに対する顕著な次元は、Ｐ（ｐ｜ｃ）によって表される確率を使用して集合特徴ベクトルへ寄与するドキュメントを重み付けることによって計算される。次に、集合特徴ベクトル中の最も大きな項が、顕著な次元を表す。例えば、表２を参照すると、図示されたクラスタに対する集合テキスト特徴ベクトルは、ワード"ｈｏｍｅｃｅｎｔｒｅ"に対応し、２番目に大きな項は、ワード"ｄｅｔａｃｈａｂｌｅ"に対応する。集合

ＵＲＬ特徴ベクトルでは、最も重要なワードは、"ｐｒｏｄｕｃｔｓ"であり、"ｄｈｃ"が続く。

【０１５５】

アクセスされるコレクションの部分のこのような詳細な特徴付けは、適切なマテリアルの追加又は既存のマテリアルの改良のために使用され得る。例えば、唯一つの小さな投資家クラスタがあることを決定することは、驚くべきことである。これは、サイトに十分な投資情報が無い場合やそのレイアウトがより魅力的にするために改良されるべきである場合の証拠として解釈される。

【０１５６】

上述のように、幾つかのクラスタの顕著な特徴は、それらが唯一ページのみをアクセスするユーザからなることである。一例は、ＴｅｘｔＢｒｉｄｇｅ Ｐｒｏ ９８（光学的文字認識プログラム）のトライアルバージョンを要求するためのページのみにアクセスするクラスタである。これらのユーザは、明確に定義された情報の必要性があり、多分外部からのリンクに続いている。ユーザは、ユーザが必要とする情報（例えば、ゼロックスホームページ上のゼロックス社の株価）を得ると、直ちにユーザは離れる。

【０１５７】

他のクラスタは、俯角動作、ユーザが多数のページを介してブラウジングする時に、徐々に満足される多くのよりアモルファス情報の必要性によって特徴付けられる。一例は、小さなオフィスやホームオフィスに相応しいより小さなデバイス上で情報を得るドキュメントホームセンター（Ｄｏｃｕｍｅｎｔ Ｈｏｍｅｃｅｎｔｒｅ）と呼ばれるサブ階層をブラウジングするユーザのクラスタである。経験的な解析において、このクラスタからのユーザが一般に幾つかの異なるドキュメントホームセンター製品に対応するサブ階層の幾つかのページを見ることが判った。明らかなように、これらのユーザは、ゼロックスウエブサイトに入り、利用出来る製品の範囲及び比較的広い範囲の情報を見ることを求める処理について学習する。

【０１５８】

コレクションの使用のこの解析は、再びより良いデザインへ反映され得る。例えば、しばしば一緒にブラウジングされる一セットのページは、ブラウジングを容易にするために、ハイパーリンキングを介して一緒にリンクされるべきである。

【０１５９】

また、マルチモードユーザクラスタリングは、ウエブサイトのデザインを改良するために有用である。図２３のディスクツリー２３１０は、５０－クラスタのクラスタリングからの投資家のクラスタを示す。太い領域２３１２と２３１４によって指示される図面の上半分に強力なアクティビティの二つの領域がある。一つの領域２３１２は、サブ階層"ａｎｎｕａｌｒｅｐｏｒｔ（年報）"に対応し、他の領域２３１４は、サブ階層"ｆａｃｔｂｏｏｋ（ファクトブック）"に対応する。多くの投資家が両方を見るという事実は、コレクションが再編成されたて、これら二つのサブ階層が共に検出されるべきであることを提案する。

【０１６０】

本システムは、探索データ解析のためにマルチモードクラスタリングを使用することの例である。本システムは、１９９８年5月１７日のユーザ母集団を特徴付けるために使用された。６４００人のユーザの全員が２０個のクラスタに割り当てられた。９個のクラスタは、製品カテゴリーに対応する。Ｐａｇｉｓ走査、複写機、ＸＳｏｆｔソフトウエア、ゼロックスソフトウエアライブラリ（ページをダウンロードするため）、ホームとデスクトップ製品、ウインドウズのためのＴｅｘｔＢｒｉｄｇｅがある。７個のクラスタは、主に単一ページ、例えば、ドライバのインデックスやゼロックス社のホームページ、にアクセスするユーザに対応する。一つのクラスタは、使用情報をアクセスする投資家を含む。一つのクラスタは、投資家及びプレスリリースとゼロックス社についての他のニュースに興味がある他の投資家を含む。二つのクラスタは、その他のカテゴリーの何れにもうまく当てはまらないユーザを含む。このように、マルチモードクラスタリングは、アナリスト

がユーザ母集団の迅速な特徴付けを得る事を可能とする。

【０１６１】

ディスクツリーを含む多くの可視化は、スクリーン上に制限された数のノードのみを描くことが出来る。マルチモードクラスタリングは、ノードのグループをメタノードへのノード集合のために使用され得る。例えば、スクリーンのエッジ上にノードの１０００個のサブノードをディスプレイするのに必要なスクリーン領域が無い場合、これら１０００個のサブノードは、マルチモードクラスタリングを使用して、５個のメタノードへ集合され得る。次に、５個のメタノードのディスプレイは、全１０００個のサブノードをディスプレイするよりも小さい空間を取る。

【０１６２】

また、マルチモードクラスタリングは、データ最小化のために使用され得る。ユーザのクラスタがマルチモードアルゴリズムによって生成されると、ユーザは、顕著な特徴を自動的に検出することが出来る。例えば、顕著なワードとして"ｈｏｍｅｃｅｎｔｒｅ"を示す表２のＨｏｍｅＣｅｎｔｒｅクラスタのテキスト表示に基づいて、ユーザは、如何に上手に"ｈｏｍｅｃｅｎｔｒｅ"単独によって特徴付けられるかをテスト出来る。

【０１６３】

他の一つのデーだ最小化アプリケーションは、例外的なオブジェクトの発見である。例えば、訴訟の開示段階において、法律事務所は、型通りの記事を大部分含む類似のドキュメントの大きなグループにではなくて、孤立するドキュメントに興味を示すに過ぎないかもしれない。マルチモードクラスタリングは、大きなグループの類似ドキュメント（例えば、共有する型通りの記事）を識別する。次に、興味あるドキュメントは、大きなクラスタの図心から大きく離れたものの間にあるかもしれない。

【０１６４】

本発明のデータ最小化技術は、第１のグループに対してマルチモードクラスタリングを実行し、次に第２のグループを第１のグループのクラスタへ割り当てることによって二つのグループのオブジェクトを比較する。この解析技術は、ウエブサイトのゼロックスベースのユーザと非ゼロックスベースのユーザを比較するのに成功裏に使用され、ゼロックス従業員がゼロックス製品のユーザでありためにほんの僅かな差が発見された。それは、外部ゼロックスウエブサイト（ドライバをダウンロードるため、製品情報を参照するため等）へ行く主な理由の一つである。一つの差は、より高い比率のゼロックスユーザが一ページのみ、ゼロックスホームページ、を訪れることであった。その理由は、多分ゼロックス従業員の多くのブラウザが彼等／彼女等のデフォルトページとしてゼロックスホームページを有することであるので、ユーザは、ユーザのブラウザを開始すると自動的にゼロックスホームページに行き、次に異なるサイトのゼロックスホームページに行く。この例は、異なるユーザグループを比較するために、マルチモードクラスタリングのユーティリティを証明する。

【０１６５】

イントラネットを含む、大きなコレクションを編成するための益々重要な技術は、階層的クラスタリングである。その目的は、ヤフー上で（及び多くのイントラネット上で）見つけられ得るように、階層を自動的に発生することである。階層的マルチモードクラスタリングは、そのような階層を自動的に発生するために或いは類別する人に手作業の編集を可能する第１のカットを与えるために使用され得る。

【０１６６】

コレクション使用解析の基づく推奨

最後に、マルチモードユーザクラスタに基づく推奨システムは、上述されたように、マルチモードコレクション使用データのコレクションを有することが可能である。一セットのクラスタは、一トレーニングセットのユーザから導かれる。新たなユーザは、２，３回の初期ページアクセスに基づいてクラスタの一つへ割り当てられる。次に、割り当てられたクラスタ中のユーザによってアクセスされたページは、ユーザに推奨される。クラスタリングがマルチモード情報に基づいてなされるので、それは、有用な推奨を行うのに十分に

頑強である。

【０１６７】

本発明のマルチモード推奨システムは、図２５に示されている。最初に、ユーザのトレーニングセットが識別される（ステップ２５１０）。ユーザについて利用可能なあらゆるタイプの情報が収集される。ここで開示された実施の形態において、ページユーザアクセスに関する情報、並びにテキストコンテント、インリンク、アウトリンク、及びこれらのページのＵＲＬを収集するために有用であることが判った。また、リアルタイムドキュメントアクセスがこの（そのデータは使用ログから又は利用出来る時ユーザのセットのブラウザ"ｂｏｏｋｍａｒｋｓ"ですらからもであってもよい）ために使用される必要は無いことに注意すべきである。また、上述から気付かれるように、デモグラフィック情報及び他の種類の追跡された情報のような、このアプリケーションにおいて有用であり得るユーザへ適用可能な他のモダリティ（ページ使用を超えて）がある。

【０１６８】

次に、ユーザは、マルチモードクラスタリングに関連するセクションにおいて、上述のように、マルチモード情報を介してクラスタリングされる（ステップ２５１２）。ページ使用が、本発明の好適な実施の形態におけるように、ユーザについて収集された主情報である場合、上述のように、種々のドキュメント特徴によってユーザの調停された表示を介してユーザをクラスタリングすることが適切である。他の戦略のまた可能であることを認識すべきである。例えば、デモグラフィック情報が収集されると、デモクラフィっク情報に関して簡単にユーザをクラスタリングすることがより適切である。クラスタリングされるベースの選択は、本発明のシステムのデザイナーの判断に任される実施の詳細事項である。或いは、その選択は、ユーザに任されてもよい。

【０１６９】

新たなユーザが無い場合（ステップ３５１４）、その処理は終了される（ステップ２５１６）。或いは、新たなユーザが識別されると（ステップ２５１８）、ブラウジング情報が新たなユーザから収集され（ステップ２５２０）、ユーザが最も近い既存のクラスタに割り当てられる（ステップ２５２２）。本発明の好適な実施の形態において、ユーザは、上述のように、テキストコンテント、インリンク、アウトリンク及びＵＲＬに亘って計算された集合コサイン類似度に基づいて、割り当てられる。

【０１７０】

次に、最も近いクラスタ中の最もポピュラーなページが識別され（ステップ２５１４）、新たなユーザへ推奨される（ステップ２５２６）。本発明の他の実施の形態において、ネーム、イーメールアドレス、又は最も近いクラスタにおけるユーザのための他の識別情報（或いは、上述の集合コサイン類似度メトリックを介して識別される、その最も近いクラスタ中の少なくとも一つのユーザ）が、その新たなユーザに提供され、それによってその新たなユーザが望ましい領域中に"エキスパート"を識別出来る。

【０１７１】

このアルゴリズムは、他の推奨アルゴリズム以上に幾つかの利点を有する。このアルゴリズムは高速である。クラスタリングがコンパイルタイム動作であるので、唯一のランタイム動作がマルチモード情報の各モダリティのベクトル空間へのマッピングと各クラスタを有する集合コサイン類似度の演算である。これは効率的である。同じ利点を得る他の方法は、ユーザ母集団をサマリー化する一つの方法としてクラスタリングを見なすことである。これは、ユーザ母集団が大きい場合に重要である。例えば、百万人のユーザの追跡を維持しなければならない代わりに、推奨は、唯、即ち１０００人のユーザのみに基づいて、行われ、１０００個のクラスタを代表するものは、完全なユーザ母集団から導出される。

【０１７２】

ユーザ母集団からクラスタを導出することは、新たなユーザを割り当てるよりのより高価であるが、１日に数回或いはそれを越える頻度で行われることはなお十分に効率的である（クラスタリングは、クラスタリングされるべきオブジェクトの数に関してリニアであ

るので）。このように、推奨は、ユーザのニーズを素早く変化するために採用し得る。これは、図２６に示されるように実行される。そうすることが望ましい場合（例えば、定期的に或いは十分な数の新たなユーザがユーザプールに追加された場合）、サブセットのユーザが初めに識別される（ステップ２６１０）。上述のように、大きな母集団では、サブセットのユーザは、全体の母集団の特徴を非常に良好に表示出来る。次に、そのセットのユーザは、再クラスタリングされる（ステップ２６１２）。次に、各ユーザ毎に最もポピュラーなページが決定され（ステップ２６１４）、従って新たなユーザに推奨されたページが調節される（ステップ２６１６）。

【０１７３】

コレクション使用解析に基づくマルチモード推奨のためのここで記述されるアルゴリズムは、非常に正確で頑強であることが判った。他の推奨アルゴリズムは、新たなユーザの前のユーザとの比較に依存する。推奨がたまたま最も近い隣人である一人又は二人のユーザに基づく場合、孤立するものが推奨されたページに影響を及ぼし得るので、間違ったページが推奨されるかもしれない。クラスタベースの汎用化は、孤立するものの影響を減少する。更に、全ての利用可能情報が使用され且つ組み合わされるので、そのアルゴリズムは、情報の単一源に依存する推奨アルゴリズムよりも頑強である。

【０１７４】

以下に記述される例に対して、テストベッドユーザ（即ち、１９９８年５月１７日及び１８日のゼロックスウエブサイトンユーザ）のアクションがログされた。ユーザのブラウジング癖に基づいて、これらのユーザが２００個のクラスタに配された。

【０１７５】

クラスタベースのシステムによって行われる推奨の第１のタイプが表３に示されている。

【表３】

クラスタ35		
0.976277 probsum		
16406	0.088639	products/copiers.htm
37005	0.085385	http://www.xerox.com
19453	0.059099	products/cop_soho.htm
33739	0.051071	soho/xc0355.html
21231	0.040836	soho/xc1044.html
17033	0.039741	soho/xc0830.html
37025	0.036496	cgi-bin/wwwwais
19451	0.035938	products/cop_pers.htm
17029	0.034706	soho/xc0540.html
17010	0.028586	soho/5306.html
21232	0.026014	soho/xc1045.html

表３は、確率Ｐ（ｐ｜ｕ）（上記を参照：ページｐの確率は我々がクラスタ３５からのユーザｕを有することが与えられる）の計算に基づいて、ユーザクラスタ３５に対する最もポピュラーなページを示す。この情報は、ページ"ｐｒｏｄｕｃｔｓ／ｃｏｐｉｅｒｓ．ｈｔｍ"をアクセスするあらゆるユーザへそのクラスタ中のその他のページ、即ち最もポピュラーな複写機、を推奨することによって宣伝され得る。これらのリンクの幾つかは、ページ"ｐｒｏｄｕｃｔｓ／ｃｏｐｉｅｒｓ．ｈｔｍ"からアクセス可能である。このアルゴリズムは、ユーザが最も関連する可能性のあるリンクを選択することを容易とする。

【０１７６】

クラスタベースの汎用化によって可能とされる第２のタイプの推奨が表４に示されてい

る。

【表４】

クラスタ127		
1.000000 probsum		
24663	0.297222	employment/ressend.htm
37057	0.268162	employment
24666	0.079701	employment/resascii.htm
21384	0.076923	research/xrcc/jobopps.htm
37005	0.054701	http://www.xerox.com
37087	0.050000	cgi-bin/employment/xrxresume.cgi
24675	0.047436	employment/restip.htm
24664	0.023077	employment/college.htm
15355	0.012821	XBS/employmt.htm
24665	0.012821	employment/recruit.htm
34418	0.012821	employment/overview.htm
37025	0.012821	cgi-bin/wwwwais

この表は、ユーザクラスタ１２７に対する最も顕著なページを含む。このクラスタのコンテンツに基づいて、本システムは、種々の細分割のｅｍｐｌｏｙｍｅｎｔ（雇用）ページをジョブのために容易に適用するユーザへ推奨することが出来る。リストアップされたドキュメントは、中央のｅｍｐｌｏｙｍｅｎｔページ（数値識別子３７０５７を有するその表の第２ページ）から直接にはアクセス出来ないゼロックスのウエブサイト上の幾つかのｅｍｐｌｏｙｍｅｎｔページを含む。二つのこのような直接にはアクセス出来ないページは"ｒｅｓｅａｒｃｈ／ｘｒｃｃ／ｊｏｂｏｐｐｓ．ｈｔｍ"と"ＸＢＳ／ｅｍｐｌｏｙｍｔ．ｈｔｍ"である。このタイプの推奨によって、ユーザはユーザがそうでない場合（丁度時間を節約することとは反対である）は全く発見できないかもしれない何かを発見することが出来る。上述されたものと同じアルゴリズムは、これを、即ち、（幾つかの初期ページアクセス後の）新たなユーザをユーザに割り当てそしてそのユーザがアクセスしなかったクラスタのページ特性を推奨すること、完成するために使用される。

【０１７７】

表５は、ユーザクラスタ２５に対する最も顕著なページを含む。

【表５】

クラスタ25		
0.998387 probsum		
37057	0.661425	employment
37005	0.300403	http://www.xerox.com
34418	0.022581	employment/overview.htm
12839	0.004435	searchform.html
24675	0.004032	employment/restip.htm
37155	0.002688	scansoft/tbapi
37113	0.002016	factbook/1997
23465	0.000806	xbs

これらのユーザは、ブラウジングしており、多分ジョブへの適用は容易ではなく、ＸＢＳ

のような指定の分割のｅｍｐｌｏｙｍｅｎｔページはユーザに推奨されない。表４と表５との間のコントラストは、マルチモードクラスタリングによって発見される汎用化の例である。第１のクラスタのユーザは、ユーザの概要を提出することが多いにあり得る。ＸＢＳのような細分割のｅｍｐｌｏｙｍｅｎｔ（雇用）ページをユーザに推奨することは優れたアイデアである。理由は、ユーザは、ジョブを見つけることについて厳しいと思うからである。

【０１７８】

他方、第２のクラスタのユーザは、何らかの一般的ブラウジングを行う。ｅｍｐｌｏｙｍｅｎｔは、ユーザのブラウジングの目的であり、ユーザは、目的のジョブサーチを実行するようには思わない。これらのユーザは、ジョブの広告を有するページを見ようとはあまりせず、従って細分割のｅｍｐｌｏｙｍｅｎｔページが彼等／彼女等には推奨されない。

【０１７９】

本明細書中に記載されている３つの符号、aaa，bbb，cccは、下記の通り、便宜上置きかえたものである。

【外１】

【０１８０】

$$\mathrm{aaa} \rightarrow \vec{c}$$

$$\mathrm{bbb} \rightarrow \vec{x}_i$$

$$\mathrm{ccc} \rightarrow \vec{x}'_i$$

【図面の簡単な説明】

【０１８１】

【図１】本発明に従うシステムとの使用に適するネットワークに接続されたドキュメントコレクションを示すブロック図である。

【図２】コレクションに追加された新たなドキュメントを処理するために本発明の一実施の形態によって使用される処理を示すフロー図である。

【図３】種々のドキュメントとユーザを表す特徴ベクトルを計算するために本発明の一実施の形態によって使用される処理を示すフロー図である。

【図４】本発明の一実施の形態において、テキストベースの特徴ベクトルを計算するために使用される処理を示すフロー図である。

【図５】本発明の一実施の形態において、テキストジャンル特徴ベクトルを計算するために使用される処理を示すフロー図である。

【図６】本発明の一実施の形態において、カラーヒストグラム特徴ベクトルを計算するために使用される処理を示すフロー図である。

【図７】本発明の一実施の形態において、対応する対のカラー複雑さ特徴ベクトルを計算するために使用される処理を示すフロー図である。

【図８】本発明の一実施の形態において、ページ使用ベクトルを計算するために使用される処理を示すフロー図である。

【図９】本発明の一実施の形態において、初期クラスタセンタを識別するためにウェーブフロントクラスタリングで使用される処理を示すフロー図である。

【図１０】関連するオブジェクトをクラスタに割り当てるためにｋ平均クラスタリングで使用される処理を示すフロー図である。

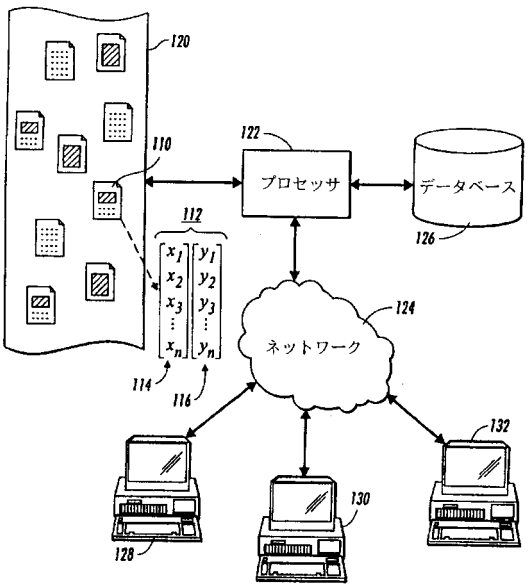
【図１１】異なるモダリティにおけるコレクションオブジェクトの分散収集の仮説セッションを示す図である。

【図１２】照会"ａｎｃｉｅｎｔ　ｃａｔｈｅｄｒａｌ（昔の大聖堂）"に応答して戻されるテキストクラスタの例示の可視ディスプレイである。

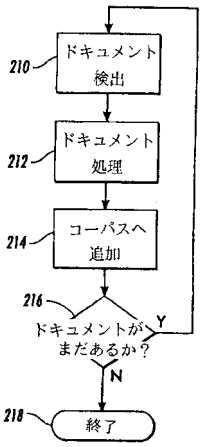
【図１３】図１２の第１のテキストクラスタを分散した後に戻されるテキストクラスタの

例示の可視ディスプレイである。
【図１４】複雑さ特徴に基づくクラスタリングの後に戻される画像クラスタの例示の可視ディスプレイである。
【図１５】照会"ｐａｐｅｒ　ｍｏｎｅｙ（紙幣）"に応答して戻されるテキストクラスタの例示の可視ディスプレイである。
【図１６】複雑さ特徴に基づく図１５の第１のテキストクラスタのクラスタリングの後に戻される画像クラスタの例示の可視ディスプレイである。
【図１７】カラーヒストグラム特徴に基づく図１６の第３及び第４画像クラスタをクラスタリングした後に戻される画像クラスタの例示の可視ディスプレイである。
【図１８】カラーヒストグラム特徴に基づく図１７の第２の画像クラスタをクラスタリングした後に戻される画像クラスタの例示の可視ディスプレイである。
【図１９】照会"ｐｙｒａｍｉｄ　ｅｇｙｐｔ（ピラミッドエジプト）"に応答して戻されたテキストクラスタの例示の可視ディスプレイである。
【図２０】複雑さ特徴に基づくクラスタリングの後に戻される画像クラスタの例示の可視ディスプレイである。
【図２１】カラーヒストグラム特徴に基づくクラスタリングの後に戻される画像クラスタの例示の可視ディスプレイである。
【図２２】図２１の画像のセットを拡張し、カラーヒストグラムの基づく結果をクラスタリングした後に戻されるテキストクラスタの例示の可視ディスプレイである。
【図２３】本発明によるクラスタの例示の間接可視化であり、一つのユーザクラスタは、そのクラスタのメンバーによって選択される可能性の高い全てのページを赤で（ここでは、矢印によって指示される）着色することによって図示される。
【図２４】本発明の一実施の形態において、クラスタとドキュメントのコンテンツをブラウジングし示すために使用されるインターフェースを示す例示の可視ディスプレイである。
【図２５】本発明に従う例示の推奨システムにおいてポピュラーなページをユーザに推奨するために使用される処理を示すフロー図である。
【図２６】本発明に従う例示の推奨システムにおいて推奨を再計算するために使用される処理を示すフロー図である。
【符号の説明】
【０１８２】
１２０　コレクション
１１０　ドキュメント
１１４　テキストベクトル
１１６　ＵＲＬベクトル
１１２　特徴ベクトル
１２４　通信ネットワーク
１２２　プロセッサ
１２４　ネットワーク
１２６　データベース
１２８、１３０、１３２　ユーザ端末

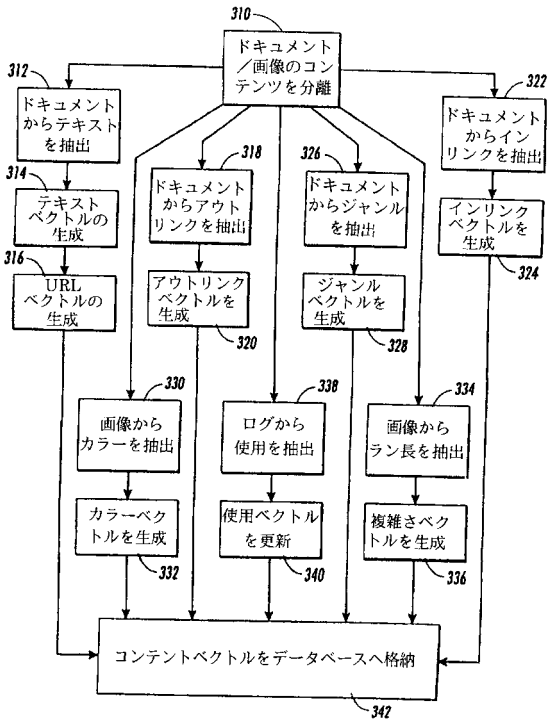
【図１】


120
110
122
プロセッサ
データベース
126
112
x_1 x_2 x_3 x_n
y_1 y_2 y_3 y_n
124
ネットワーク
114
116
132
130
128

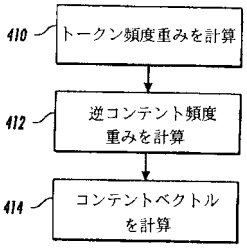
【図２】


210
ドキュメント検出
212
ドキュメント処理
214
コーパスへ追加
216
ドキュメントがまだあるか？
Y
N
218
終了

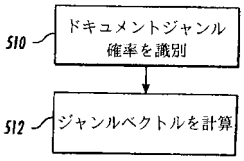
【図３】


310
ドキュメント／画像のコンテンツを分離
312
ドキュメントからテキストを抽出
314
テキストベクトルの生成
316
URLベクトルの生成
318
ドキュメントからアウトリンクを抽出
320
アウトリンクベクトルを生成
322
ドキュメントからインリンクを抽出
324
インリンクベクトルを生成
326
ドキュメントからジャンルを抽出
328
ジャンルベクトルを生成
330
画像からカラーを抽出
332
カラーベクトルを生成
338
ログから使用を抽出
340
使用ベクトルを更新
334
画像からラン長を抽出
336
複雑さベクトルを生成
コンテントベクトルをデータベースへ格納
342

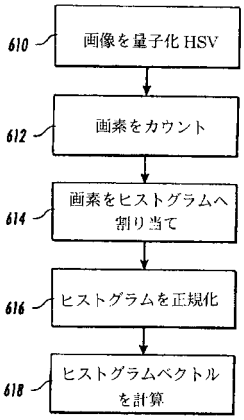
【図４】


410
トークン頻度重みを計算
412
逆コンテント頻度重みを計算
414
コンテントベクトルを計算

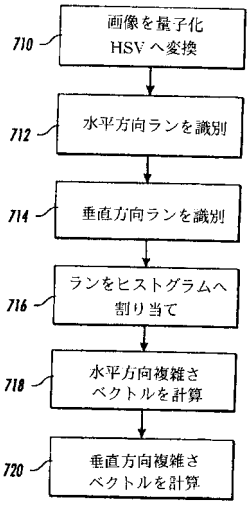
【図５】


510
ドキュメントジャンル確率を識別
512
ジャンルベクトルを計算

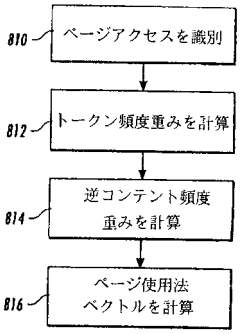
【図６】



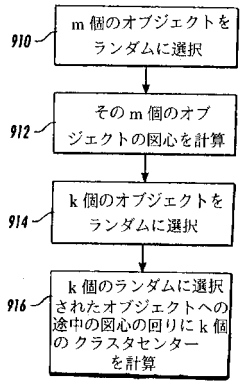
【図７】



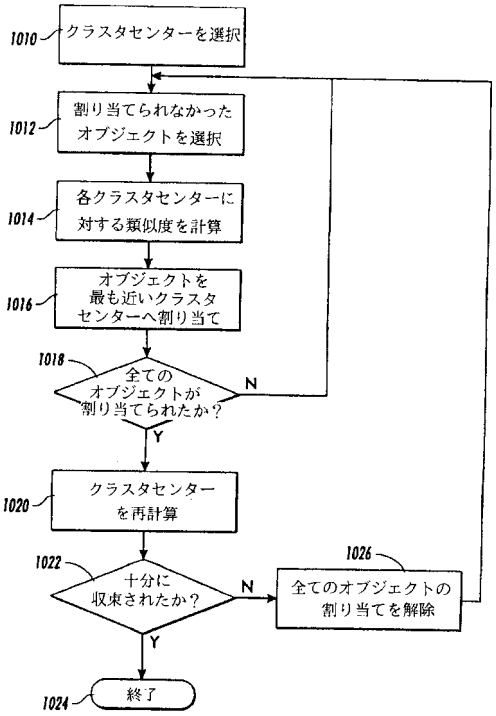
【図８】



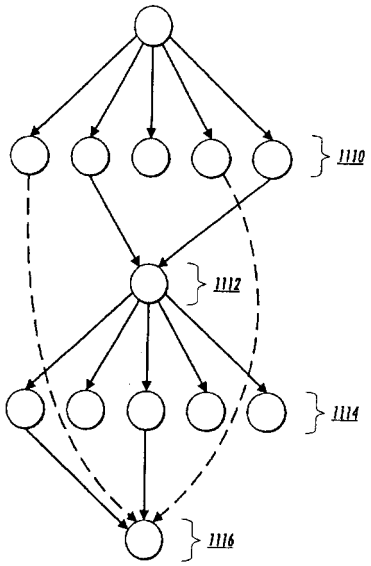
【図９】



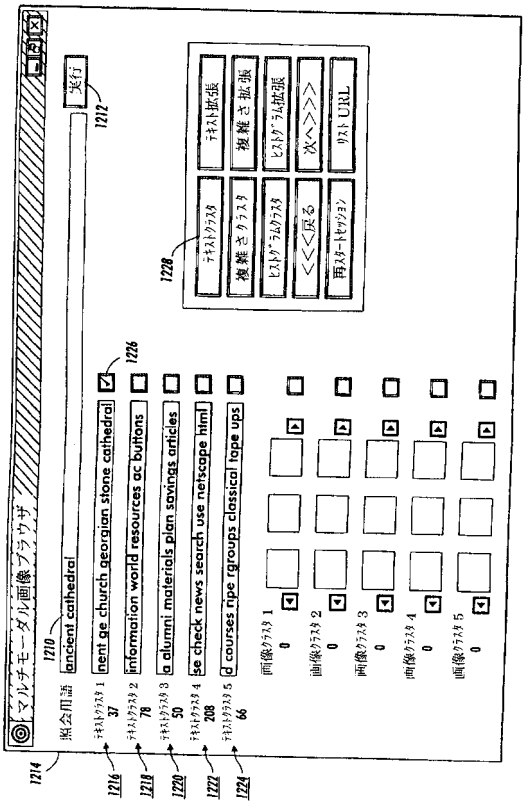
【図１０】



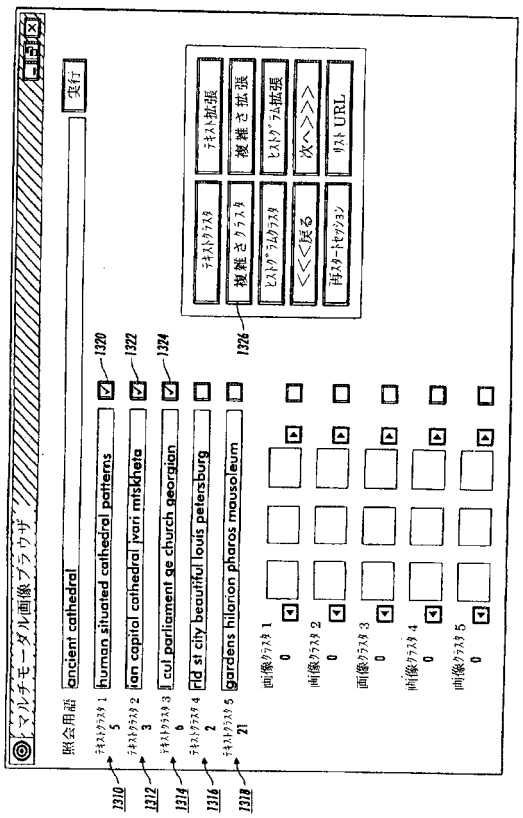
【図１１】



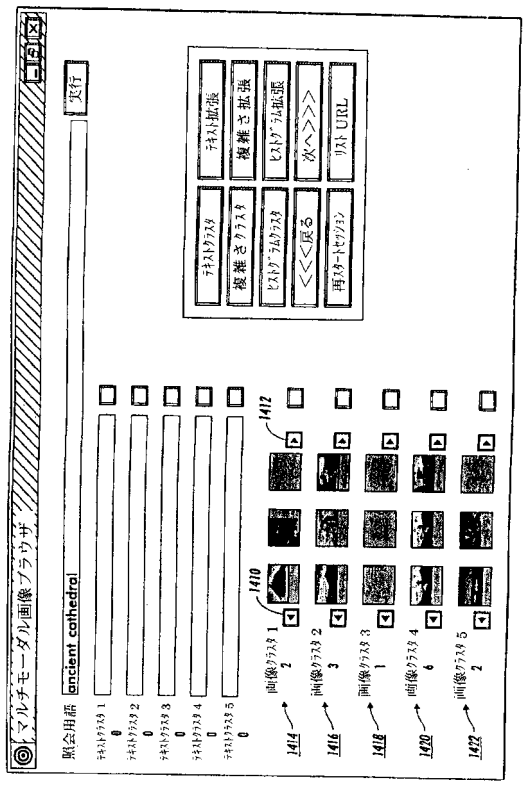
【図１２】



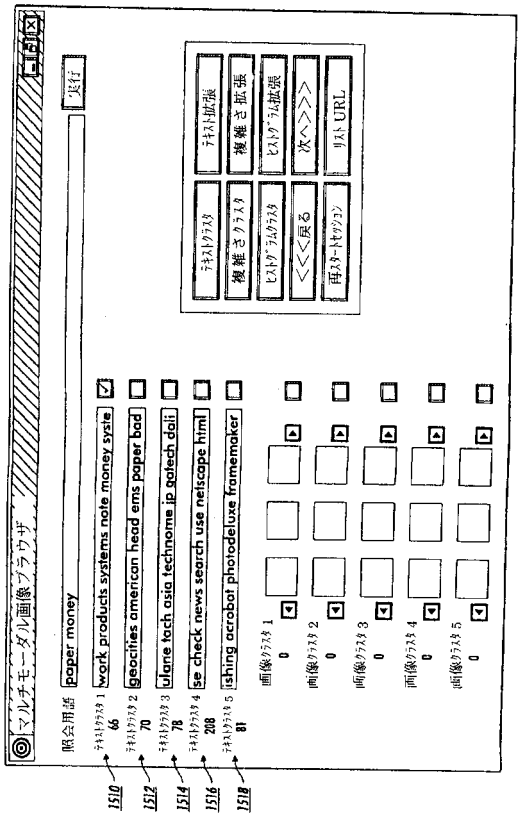
【図１３】



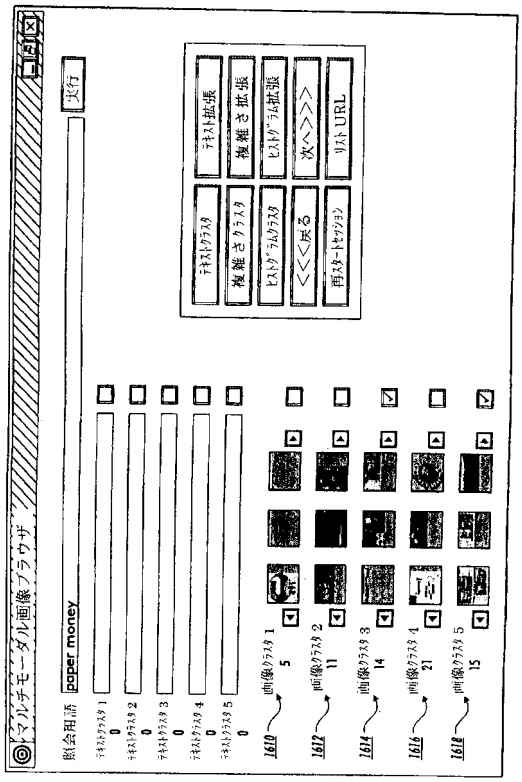
【図１４】



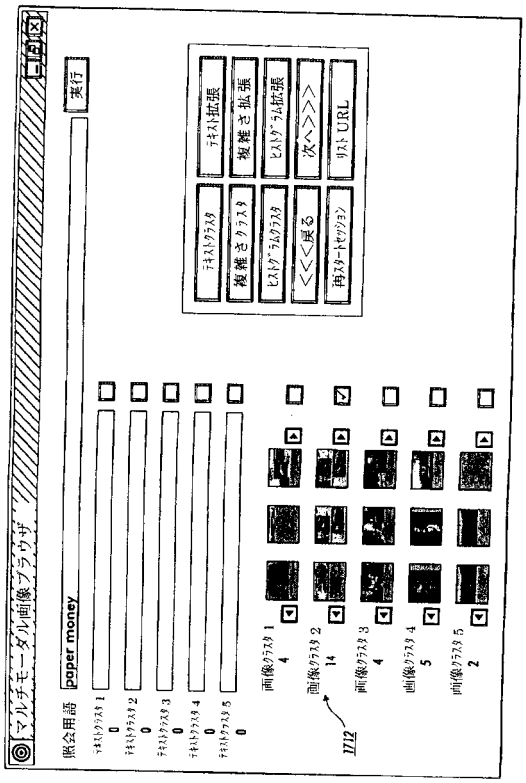
【図１５】



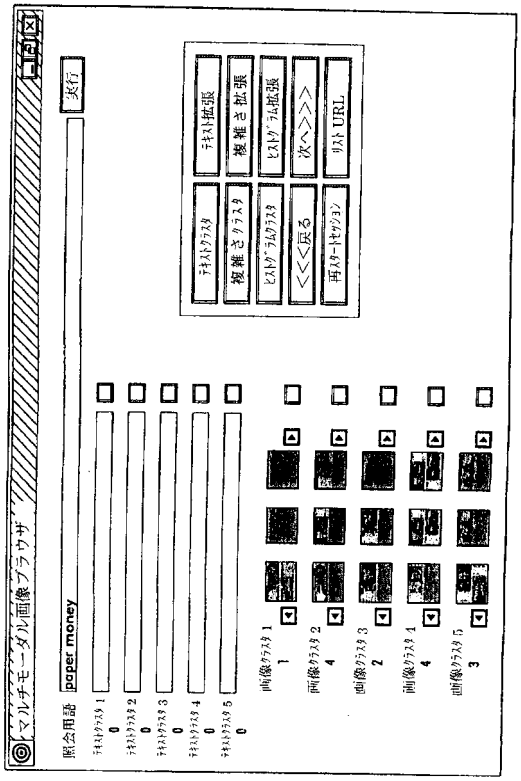
【図１６】



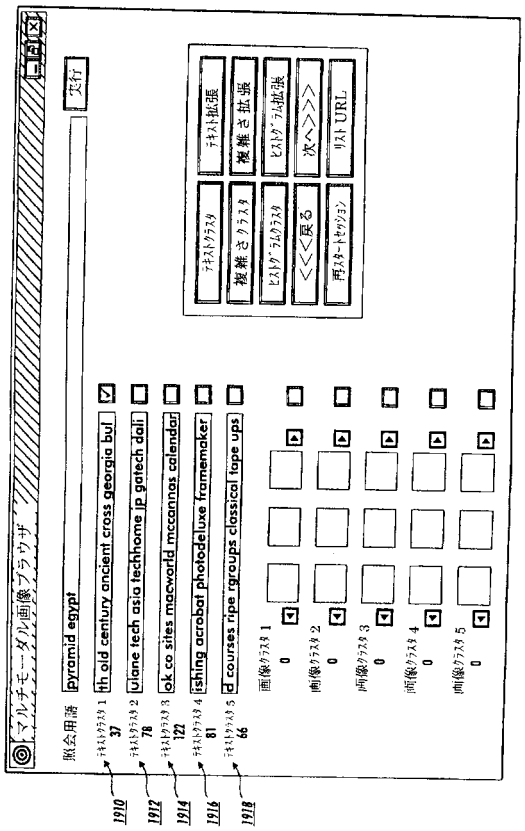
【図１７】



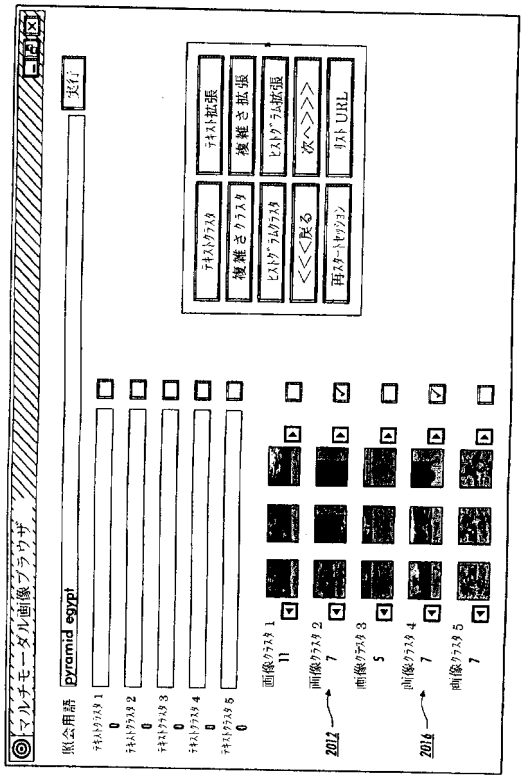
【図１８】



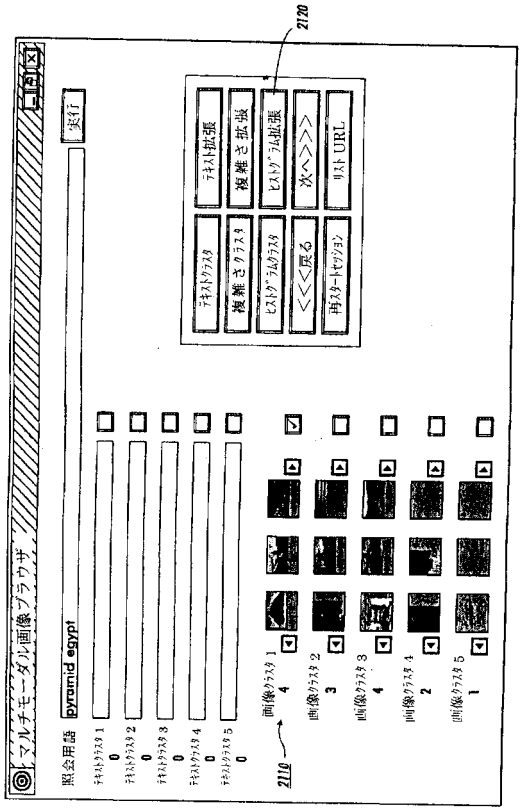
【図１９】



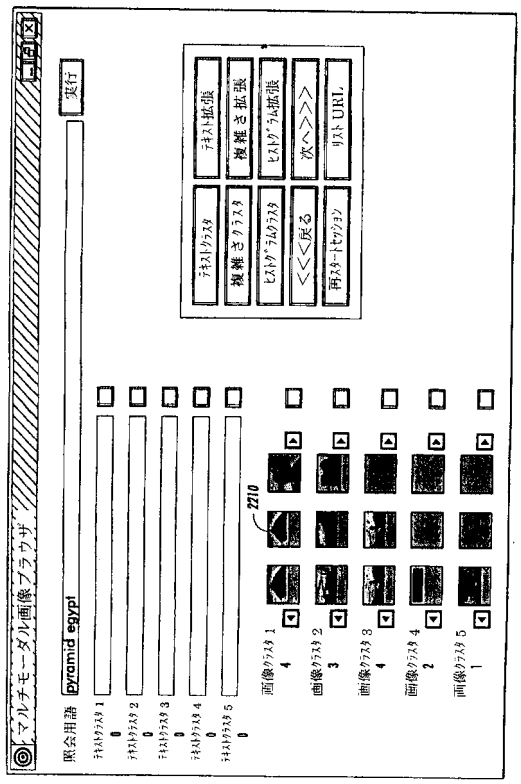
【図２０】



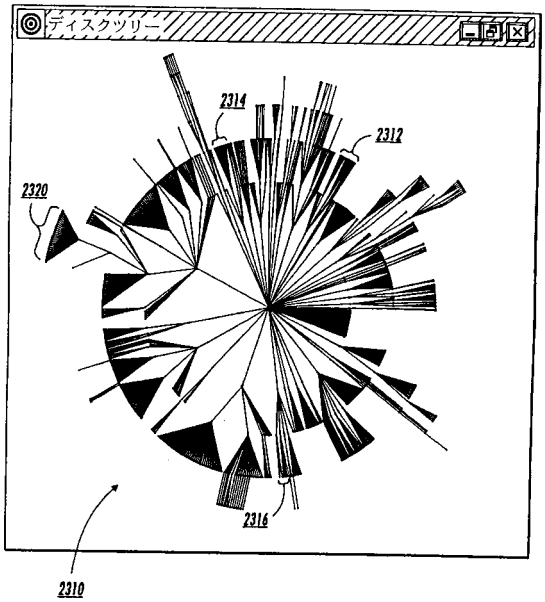
【図２１】



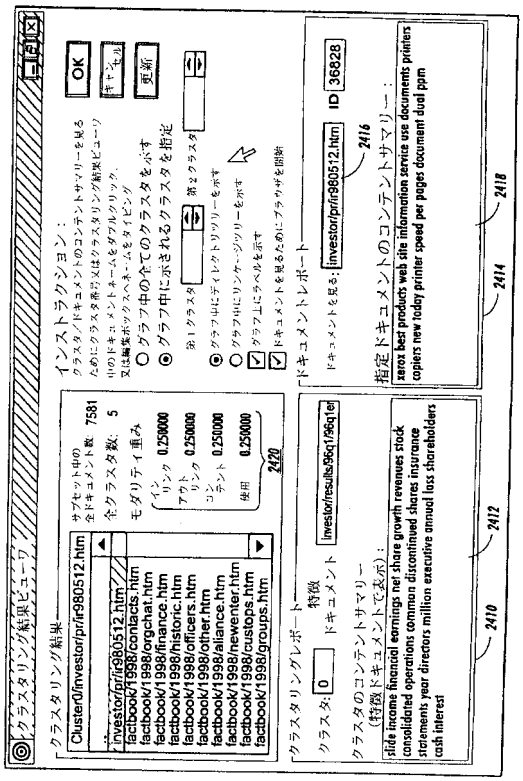
【図２２】



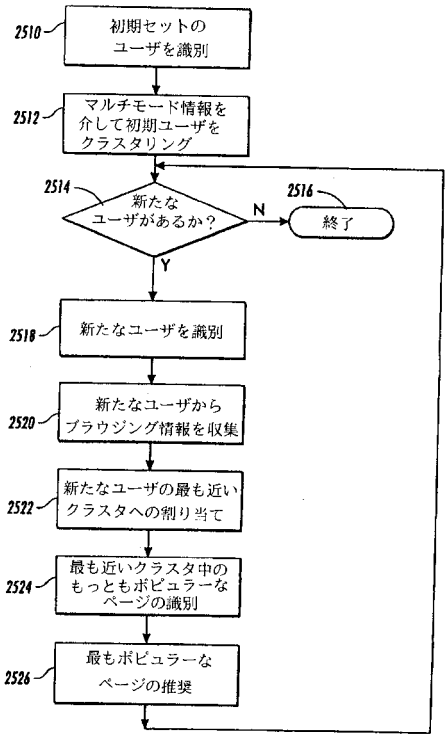
【図２３】



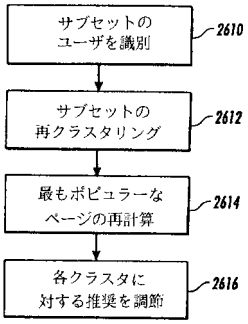
【図２４】



【図２５】



【図２６】



フロントページの続き

(31)優先権主張番号 421419
(32)優先日 平成11年10月19日(1999.10.19)
(33)優先権主張国 米国(US)
(31)優先権主張番号 421767
(32)優先日 平成11年10月19日(1999.10.19)
(33)優先権主張国 米国(US)
(31)優先権主張番号 421770
(32)優先日 平成11年10月19日(1999.10.19)
(33)優先権主張国 米国(US)
(31)優先権主張番号 425038
(32)優先日 平成11年10月19日(1999.10.19)
(33)優先権主張国 米国(US)
(31)優先権主張番号 425039
(32)優先日 平成11年10月19日(1999.10.19)
(33)優先権主張国 米国(US)

(72)発明者 ヒンリッヒ シュエッツェ
アメリカ合衆国 ９４１３１－１５５２ カリフォルニア州 サンフランシスコ ポートラ ドライブ １００ ナンバー１
(72)発明者 ウラス ガーギ
アメリカ合衆国 １６８０３ ペンシルベニア州 ステイト カレッジ ウエスト クリントン アベニュー ２３４
(72)発明者 ジェームズ イー．ピットカウ
アメリカ合衆国 ９４３０６ カリフォルニア州 パロ アルト エルスワース プレイス ７４２
(72)発明者 ピーター エル．ピロッリ
アメリカ合衆国 ９４１１６ カリフォルニア州 サンフランシスコ スロート ブルバード ２９５８
(72)発明者 エド エイチ．チ
アメリカ合衆国 ５５４１７－１９３７ ミネソタ州 ミネアポリス ショアビュー アベニュー サウス ５２４１
(72)発明者 ジュン リ
アメリカ合衆国 ４７４０１ インディアナ州 ブルーミントン イースト セカンド ストリート ２１０６ ナンバー７

審査官 鈴木 和樹

(56)参考文献 特開平１１－２２４２５６（ＪＰ，Ａ）
特開平０８－１９０５７８（ＪＰ，Ａ）
米国特許第０５８３５０８７（ＵＳ，Ａ）
吉田勝彦、外１名，インターネット上のマルチメディアデータの柔軟な検索方式の提案，電子情報通信学会技術研究報告（ＫＢＳＥ９７－２０～２４），日本，社団法人電子情報通信学会，１９９７年１１月２８日，第９７巻，第４１３号，ｐ．１７－２２
Rohini K. Srihari, Automatic Indexing and Content-Based Retrieval of Captioned images, Computer [online], １９９５年 ９月, 第２８巻, 第９号, ｐ．４９－５６, [DL from IEEE Xplore]
Marco La Cascia、外２名, Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web, Content-Based Access of Image and Video Libraries, 1998.

Proceedings. IEEE Workshop on [online], １９９８年 ６月２１日, ｐ. ２４－２８, [DL from IEEE Xplore]
Ron Weiss、外６名, HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering, HYPERTEXT '96 Proceedings of the seventh ACM conference on Hypertext [online], 米国, ACM, １９９６年, ｐ. １８０－１９３, [DL from ACM Digital Library]

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ １７／３０
ＩＥＥＥ Ｘｐｌｏｒｅ
ＴＨＥ ＡＣＭ ＤＩＧＩＴＡＬ ＬＩＢＲＡＲＹ