(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2005-149493
(P2005-149493A)
(43)公開日　平成17年6月9日(2005.6.9)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	220B	5B075
	G06F 17/30	170B	
	G06F 17/30	210C	

審査請求 未請求　請求項の数 26　ＯＬ　（全 39 頁）

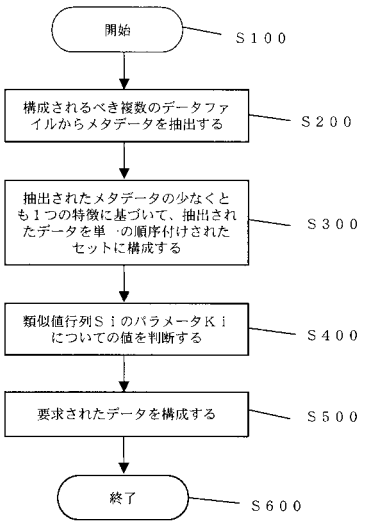
(21)出願番号　特願2004-314846 (P2004-314846)
(22)出願日　平成16年10月28日 (2004.10.28)
(31)優先権主張番号　60/515713
(32)優先日　平成15年10月31日 (2003.10.31)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　10/729915
(32)優先日　平成15年12月9日 (2003.12.9)
(33)優先権主張国　米国 (US)

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目１７番２２号
(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　マシュー　エル．クーパー
アメリカ合衆国　９４１１４　カリフォルニア州　サンフランシスコ　トウェンティーサード　ストリート　３９９８
(72)発明者　ジョナサン　ティー．フート
アメリカ合衆国　９４０２５　カリフォルニア州　メンロ　パーク　ローレル　ストリート　４５０

最終頁に続く

(54)【発明の名称】データファイルを構成するための方法、プログラム、及びシステム

(57)【要約】
【課題】複数のデータファイルを構成するための方法、プログラム、及びシステムを提供する。
【解決手段】データ構成システム及び方法は、データファイルの少なくともいくつかについて、関連したデータを抽出することによって、複数のデータファイルに関連するメタデータ又は他のデータを用いて、複数のデータファイルを構成し、抽出された関連したデータ及び入力パラメータ値に基づいて、抽出された関連したデータに構成して、データファイルを少なくともいくつかの群に分割する。
【選択図】　図１



【特許請求の範囲】
【請求項１】
少なくとも各データファイルと関連した少なくとも１つのメタデータ要素を有するメタデータを用いて、複数のデータファイルを構成するための方法であって、
前記データファイルの少なくともいくつかについて、前記データファイルと関連した少なくとも１つのメタデータ要素を抽出し、
前記抽出されたメタデータ要素についての値に基づいて、前記抽出されたメタデータ要素を要求された順序に構成し、
少なくとも１つのパラメータ値を入力し、
前記抽出されたメタデータ要素及び前記入力パラメータ値に基づいて、前記データファイルの少なくともいくつかを群に分割する、ことを含む
方法。
【請求項２】
前記少なくともいくつかのデータファイルを分割することが、前記少なくとも１つのパラメータ値の少なくとも１つの各々について、前記抽出されたメタデータ要素及びそのパラメータ値を用いて、前記複数のデータファイルの少なくとも２つについての類似性値を判断することを含む、請求項１に記載の方法。
【請求項３】
前記少なくとも１つの類似性値を判断することが、次の数式で少なくとも１つの類似性値を判断することを含み、

$$S_K(i,j) = \exp\left(-\frac{|t_i - t_j|}{K}\right)$$

$S_K(i,j)$がi^{th}データファイル及びj^{th}データファイルについての類似性値であり、
Kがパラメータ値であり、
t_i及びt_jが、前記i^{th}及びj^{th}データファイルについての前記少なくとも１つの抽出されたメタデータ要素の少なくとも一つのメタデータ要素の実際の値である、
請求項２に記載の方法。
【請求項４】
少なくとも１つの類似性値を判断することが、次の数式で前記少なくとも１つの類似性値を判断することを含み、

$$S_K(i,j) = \exp\left(\frac{1}{K}\left(\frac{\langle v_i, v_j \rangle}{|v_i||v_j|} - 1\right)\right)$$

$S_K(i,j)$がi^{th}データファイル及びj^{th}データファイルについての類似性値であり、
Kがパラメータ値であり、
v_i及びv_jが、前記i^{th}及びj^{th}データファイルから判断された実際のベクトル値である、
請求項２に記載の方法。
【請求項５】
少なくともいくつかのデータファイルの各々について、そのデータファイルについての及び多数の近くのデータファイルについての少なくとも１つの類似性値に基づいて、その

データファイルについての少なくとも１つの新規性値を判断することをさらに含む、請求項２に記載の方法。

【請求項６】

少なくとも１つの新規性値を判断することが、次の数式で少なくとも１つの新規性値を判断することを含み、

$$v_K(s) = \sum_{l,n=-5}^{5} S_K(s+l, s+n)\, g(l,n)$$

$v_K(s)$が新規性値であり、

ｇがガウシアンテーパの１１ｘ１１チェッカーボードカーネルである、

請求項５に記載の方法。

【請求項７】

前記データファイルの少なくともいくつかについて判断された前記少なくとも１つの新規性値に基づいて、前記複数のデータファイルの境界位置間の少なくとも１つの境界位置を判断することをさらに含む、請求項５に記載の方法。

【請求項８】

前記判断された境界位置の少なくともいくつかについて、その境界位置についての信頼値を判断することをさらに含む、請求項７に記載の方法。

【請求項９】

境界位置についての信頼値を判断することが、次の数式で信頼値を判断することを含み、

$$C(B_K) = \sum_{l=1}^{|B_K|-1} \frac{1}{(b_{l+1}-b_l)^2} \sum_{i,j=b_l}^{b_{l+1}} S_K(i,j) - \sum_{l=1}^{|B_K|-2} \frac{1}{(b_{l+1}-b_l)(b_{l+2}-b_{l+1})} \sum_{i=b_l}^{b_{l+1}} \sum_{j=b_{l+1}}^{b_{l+2}} S_K(i,j)$$

$C(B_K)$が$B_K{}^{th}$境界についての信頼値であり、

$S_K(i, j)$がi^{th}データファイル及びj^{th}データファイルについての類似性値であり、

ｂが前記入力パラメータＫレベルについての特定の値で検出された境界のインデックス値である、

請求項８に記載の方法。

【請求項１０】

前記判断された境界位置の少なくともいくつかについて、前記信頼値を最大にする少なくとも１つのパラメータ値の少なくとも１つを判断することをさらに含む、請求項８に記載の方法。

【請求項１１】

データファイルの対応するメタデータ要素と少なくとも関連した少なくとも１つのメタデータ要素を有するメタデータを用いて、複数のデータファイルを構成するための方法であって、

各メタデータがデータファイルに対応している、少なくとも１つのメタデータのセットを処理することと、

前記メタデータを分析するために要求されたパラメータ値を得ることと、

得られたパラメータ値を用いてメタデータ要素のセット内の構造を判断することであって、前記複数のデータファイルの少なくともサブセットについて、前記パラメータ値を互いに用いて前記メタデータの少なくともサブセットを比較することによって前記構造が判断される、構造を判断することと、を含む、

方法

【請求項１２】
前記メタデータ要素の前記判断された構造を用いて、前記データファイルを群にクラスタ化することをさらに含む、請求項１１に記載の方法。
【請求項１３】
データファイルの前記判断されたクラスタから境界を判断することをさらに含み、前記境界はデータファイルの前記判断されたクラスタ間に位置している、請求項１２に記載の方法
【請求項１４】
データファイルの１つのクラスタ内の前記メタデータ要素の少なくともいくつかを、データファイルの要素クラスタ内の前記メタデータ要素の少なくともいくつかと比較することによって、類似性値を判断することと、
データファイルの１つのクラスタ内の前記メタデータ要素の少なくともいくつかを、データファイルの他のクラスタ内の前記メタデータ要素の少なくともいくつかと比較することによって、非類似性値を判断することと、をさらに含む、
請求項１３に記載の方法。
【請求項１５】
前記類似性値と前記非類似性値との相違に基づいて、データファイルのクラスタの要求された群に対応するパラメータ値を判断することをさらに含む、請求項１４に記載の方法。
【請求項１６】
データ処理装置上で実行可能であると共に、少なくとも各データファイルと関連した少なくとも１つのメタデータ要素を有するメタデータを用いることによって複数のデータファイルを構成するために用いられ得る、プログラムであって、前記プログラムが、
前記データファイルの少なくともいくつかについて、そのデータファイルと関連した少なくとも１つのメタデータ要素を抽出するための命令と、
前記抽出されたメタデータ要素についての値に基づいて、前記抽出されたメタデータ要素を要求された順序に構成するための命令と、
パラメータ値を入力するための命令と、
前記抽出されたメタデータ要素及び前記入力パラメータ値に基づいて、前記データファイルの少なくともいくつかを群に分割するための命令と、を含む、
プログラム。
【請求項１７】
前記データファイルの少なくともいくつかを群に分割するための命令が、前記少なくとも１つのパラメータ値の少なくとも１つの各々について、前記抽出されたメタデータ要素の少なくともいくつか及びそのパラメータ値を用いて、前記複数のデータファイルの少なくとも２つについての類似性値を判断するための命令をさらに含む、請求項１６に記載のプログラム。
【請求項１８】
少なくともいくつかのデータファイルの各々について、そのデータファイルについての及び多数の近くのデータファイルについての少なくとも１つの類似性値に基づいて、そのデータファイルについての少なくとも１つの新規性値を判断するための命令をさらに含む、請求項１７に記載のプログラム。
【請求項１９】
前記少なくとも１つの類似性値を判断するための命令が、次の数式で前記少なくとも１つの類似性値を判断するための命令を含み、

$$S_K(i,j) = \exp\left(-\frac{|t_i - t_j|}{K}\right)$$

S_K（ｉ，ｊ）がｉthデータファイル及びｊthデータファイルについての類似性値であり、
Ｋがパラメータ値であり、
t_i及びt_jが、前記ｉth及びｊthデータファイルについての前記少なくとも１つの抽出されたメタデータ要素の少なくとも一つのメタデータ要素の実際の値である、
請求項１７に記載のプログラム。
【請求項２０】
少なくとも１つの類似性値を判断するための命令が、次の数式で前記少なくとも１つの類似性値を判断するための命令を含み、

$$S_K(i,j) = \exp\left(\frac{1}{K}\left(\frac{\langle v_i, v_j \rangle}{|v_i||v_j|} - 1\right)\right)$$

S_K（ｉ，ｊ）がｉthデータファイル及びｊthデータファイルについての類似性値であり、
Ｋがパラメータ値であり、
v_i及びv_jが、前記ｉth及びｊthデータファイルから判断された実際のベクトル値である、
請求項１７に記載のプログラム。
【請求項２１】
前記データファイルの少なくともいくつかについて判断された前記少なくとも１つの新規性値に基づいて、前記複数のデータファイルの境界位置間の少なくとも１つの境界位置を判断するための命令をさらに含む、請求項１８に記載のプログラム。
【請求項２２】
少なくとも１つの新規性値を判断するための命令が、次の数式で少なくとも１つの新規性値を判断する命令を含み、

$$v_K(s) = \sum_{l,n=-5}^{5} S_K(s+l, s+n)\, g(l,n)$$

v_K（ｓ）が新規性値であり、
ｇがガウシアンテーパの１１ｘ１１チェッカーボードカーネルである、
請求項１８に記載のプログラム。
【請求項２３】
前記判断された境界位置の少なくともいくつかについて、その境界位置についての信頼値を判断するための命令をさらに含む、請求項２１に記載のプログラム。
【請求項２４】
前記少なくとも１つの信頼値を判断するための命令が、次の数式でこのような信頼値の各々を判断するための命令を含み、

$$C(B_K) = \sum_{l=1}^{|B_K|-1} \frac{1}{(b_{l+1}-b_l)^2} \sum_{i,j=b_l}^{b_{l+1}} S_K(i,j) - \sum_{l=1}^{|B_K|-2} \frac{1}{(b_{l+1}-b_l)(b_{l+2}-b_{l+1})} \sum_{i=b_l}^{b_{l+1}} \sum_{j=b_{l+1}}^{b_{l+2}} S_K(i,j)$$

Ｃ（B_K）が$B_K$th境界についての信頼値であり、
S_K（ｉ，ｊ）が前記ｉthデータファイル及び前記ｊthデータファイルについての類似性値であり、

ｂがあるレベルの検出された境界である、
請求項２３に記載のプログラム。
【請求項２５】
前記判断された境界位置の少なくともいくつかについて、前記信頼値を最大にする少なくとも１つのパラメータ値の少なくとも１つを判断するための命令をさらに含む、請求項２３に記載のプログラム。
【請求項２６】
データファイルの対応するメタデータ要素と少なくとも関連した、少なくとも１つのメタデータ要素を有するメタデータを用いて、複数のデータファイルを構成するために用いられ得る、データファイル構成システムであって、
前記データファイルの少なくともいくつかについて、前記データファイルと関連した少なくとも１つのメタデータ要素を抽出する、メタデータ抽出回路、ルーティン、又はアプリケーションと、
前記抽出されたメタデータについての値に基づいて、前記抽出されたメタデータ要素を要求された順序に構成するための、メタデータ構成回路、ルーティン、又はアプリケーションと、
前記少なくとも１つのパラメータ値の少なくとも１つについて、前記抽出されたメタデータ要素の少なくともいくつか及びそのパラメータ値を用いて、前記複数のデータファイルの少なくとも２つについての類似性値を判断する、類似性値判断回路、ルーティン、又はアプリケーションと、
前記データファイルについての及び多数の近くのデータファイルについての少なくとも１つの類似性値に基づいて、そのデータファイルについての少なくとも１つの新規性値を判断する、新規性値判断回路、ルーティン、又はアプリケーションと、
前記データファイルの少なくともいくつかについて判断された前記少なくとも１つの新規性値に基づいて前記複数のデータファイルの境界位置間の少なくとも１つの境界位置を判断することによって、前記抽出されたメタデータ要素及び前記入力パラメータ値に基づいて、前記データファイルの少なくともいくつかを群に分割する、データ分割回路、ルーティン、又はアプリケーションと、
前記判断された境界位置の少なくともいくつかについて、その境界位置についての信頼値を判断し、前記判断された境界位置の少なくともいくつかについて、前記データ分割回路、ルーティン、又はアプリケーションが、前記信頼値を最大にする前記少なくとも１つのパラメータ値をさらに判断する、信頼値判断回路、ルーティン、又はアプリケーションと、を備える、
データファイル構成システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データの階層的クラスタリングによってデータを構成するためのシステム及び方法に関する。すなわち、複数のデータファイルを構成するための方法、プログラム、及びシステムに関する。
【背景技術】
【０００２】
データは、例えば、メディアデータとして、メディアファイル内等に、種々の方法によって保存されている。メディアデータは、オーディオ、ビデオ、グラフィック、及び／又はテキストストリーム又はファイル等の、ストリーム又はファイルでもよい。メディアデータの１つの例示的な形式は、デジタル写真である。高品質のデジタルカメラの値ごろ感は、デジタル写真を増加させ、多くの人がデジタル写真を容易に撮影して保存できるようにした。これらのデジタル写真は、しばしば、デジタル写真データファイルとして保存される。
【０００３】

メディアデータファイルは、通常、いくつかの異なる部分を含む。たとえば、１つのデジタル写真データファイルは、たとえばＪＰＥＧフォーマット等の特定のファイルフォーマットで記録された画像データを含んでいてもよい。画像データに加えて、画像データについてのある情報は、画像データと関連した得られるデジタル写真データファイル内に、メタデータとして典型的に保存されてもよい。関連したメタデータは、基礎となる画像データとは分離した、異なるデータである。１つの例示的なフォーマットは、Exif（エグジフ、Exchangeable Image File Format）であり、それは、ＪＰＥＧ画像データファイルの一部として保存されたヘッダ情報についてのフォーマットとして、しばしば用いられる。Exifフォーマットで保存されたメタデータの例は、ファイル名、データが作成された時間、画像ファイルに最後の変更が行なわれた時間等の１つ以上のタイムスタンプ、画像データの短い説明、又は画像データが得られた場所についてのＧＰＳ位置を含む。

【０００４】

デジタル写真データファイル及び他のこのように急速に蓄積するデータファイルを処理するために多くの技術が作成されてきた。単純なデータファイルについては、このような技術の１つは、このようなデータファイルの各々が関連した項目に応じて、このようなデータファイルを特定のフォルダ内に配置することを含む。他の技術は、ある人の連絡先の情報を、パーソナルコンピュータデータベース内の所与のファイルディレクトリ内に手動で構成することを含む。ユーザは、内容を検討し、ファイルディレクト内への特定の連絡先の情報の配置、及び友人、仕事関係（business contact）、学校関係（school contact）等の任意のサブカテゴリを判断する。

【０００５】

Microsoft Word（登録商標）で用いられるフォーマット等特定のフォーマットで書かれた連絡先の情報のような、単純なデータでさえ、二つの特徴を含む。データを識別するデータレコードの名前は、レコード内に含まれた情報を圧縮するスカラー特徴と呼び得る。連絡先の名前、連絡先（contact）の住所、又はその特定の連絡先に関する他のデータ等の記録の実際の内容は、より詳細であり、ベクトル特徴と呼び得る。

【０００６】

【特許文献１】米国特許第６，５４２，８６９　Ｂ１号
【非特許文献１】ヘッカーマン（Heckerman），「ベイズのネットワークを学ぶに当っての指導書」（“A Tutorial on Learning With Bayesian Networks”），マイクロソフトリサーチ（Microsoft Research），１９９５年３月，Ｐ．１～５７
【非特許文献２】フット（Foote），「オーディオの新規性の測定を用いる自動的オーディオ分割」（“Automatic Audio Segmentation Using a Measure of Audio Novelty”），ＦＸパロ・アルト研究所（FX Palo Alto Laboratory, Inc）
【非特許文献３】プラットら（Platt et al.），「写真ＴＯＣ：個人的な写真を閲覧するための自動的クラスタリング」（“Photo TOC: Automatic Clustering for browsing Personal Photographs”），マイクロソフト・リサーチ（Microsoft Research），２００２年２月，Ｐ．１～１９
【非特許文献４】スラニーら（Slaney et al.），「マルチメディア・エッジ：すべての次元における階層を見つける」（“Multimedia Edges: Finding Hierarchy in all Dimensions”），マルチメディアについての第９回ＡＣＭ国際会議の予稿集（Proceedings of the 9th ACM International Conference on Multimedia），Ｐ．１～１２
【非特許文献５】ルイら（Loui et al.），「アルバムに適用するための自動的な画像イベント分割及び品質選別」（“Automatic Image Event Segmentation and Quality Screening for Albuming Applications”），マルチメディアについてのＩＥＥＥ国際会議及び博覧会（IEEE International Conference on Multimedia and Expo），２０００年７月，ニューヨーク
【非特許文献６】チェンら（Chen et al.），「ベイズの情報基準による、話者、環境、及びチャンネル変化の検出及びクラスタリング」（“Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion”），I

ＢＭ　Ｔ．Ｊ．ワトソン研究センター（IBM T.J.Watson Research Center）
【非特許文献７】レナルら（Renals et al.），「会議室からのオーディオ情報アクセス」（“Audio Information Access From Meeting Rooms”），IEEE ICASSAP-2003，香港，Ｐ．１～４
【非特許文献８】グラハムら（Graham et al.），「個人的なデジタルライブラリによる写真閲覧用の要素としての時間」（“Time as Essence for Photo Browsing Through Personal Digital Libraries”），スタンフォード大学（Stanford University）
【非特許文献９】「デジタルスチルカメラ画像ファイルフォーマット基準」（“Digital Still Camera Image File Format Standard”），Version2.1，１９９８年６月１２日、日本電子工業振興協会（Japan Electronic Industry Development Association (JEIDA)），Ｐ．１～１６６
【発明の開示】
【発明が解決しようとする課題】
【０００７】
データファイルを構成するための１つの方法は、ユーザが各データファイルの内容及び／又はそのデータファイルの名前を実際に調査し、次に、適切な項目記述子が付けられたフォルダ等の、特定のファイルディレクトリ内の、データファイルの適切な位置を手動で判断することである。データファイルを特定の位置に配置して集めることは、データファイルを特定の関係に構成することである。しかしながら、たとえば、膨大な枚数の写真が編成されなければならない時には、写真の各データファイルを手動で編成・構成することは、ほとんど不可能になる。この困難さは、各データファイルの内容が複雑である場合、たとえば内容が画像データである場合に大きくなる。
【課題を解決するための手段】
【０００８】
（本発明の目的）
本発明は、データファイル内のメタデータ又は他の順序付けられた情報に基づいてデータを効率良く構成するためのシステム及び方法を提供する。
【０００９】
本発明は、データファイルを構成するメタデータに基づいて、関連したデータファイルをクラスタリングすることによって、データファイルを構成するための、システム及び方法を、別々に提供する。
【００１０】
本発明は、データファイルのメタデータを抽出するためのシステム及び方法を、別々に提供する。
【００１１】
本発明は、データファイルのメタデータに基づいてデータファイルを構成するためのシステム及び方法を、別々に提供する。
【００１２】
本発明は、閲覧及び／又は検索するために要求されたデータファイルを構成するためのシステム及び方法を、別々に提供する。
【００１３】
本発明によるシステム及び方法の種々の例示的な実施の形態において、データファイルの要求されたセットは、メタデータのセットを調査することによって構成され、ここでは、メタデータの各メタデータ要素は、特定のデータファイルから抽出されるか、又は少なくとも関連して来た。種々の実施の形態において、メタデータのセット内の構造は、メタデータ要素を分析するためのメタデータの要素の値の要求された範囲を得て、次に、データファイルのすべて又はサブセットについてメタデータの要素についての値を比較することによって、評価される。
【００１４】
種々の例示的な実施の形態において、メタデータのセットのメタデータ要素は、メタデ

ータのセットの評価された構造を用いて、クラスタ化される。メタデータのセットの構造は、メタデータ要素値の各クラスタを他のクラスタから線引きする境界を含む。種々の例示的な実施の形態において、比較されたデータファイル間の類似性又は非類似性を判断するために、あるデータファイルのあるメタデータ要素の値は、範囲の値に基づいて、クラスタ内の他のデータファイルのメタデータ要素の値と比較される。

【００１５】

種々の例示的な実施の形態において、データは、データのすべての可能な対間の比較又はデータのすべての可能な対のサブセットを用いて、構成される。種々の例示的な実施の形態において、比較された類似性又は非類似性には、メタデータ要素のクラスタ及びそれらの対応するデータファイルの配置に対応する数値が与えられる。種々の例示的な実施の形態において、より正確にするために、クラスタの配置が調べられる。種々の例示的な実施の形態において、コンテンツベースの類似性測定を作り出すことによってデータファイルは低レベルの特徴を発生する時よりも、より効率良くかつコンピュータによってより安く構成される。

【００１６】

本発明の第１の態様は、少なくとも各データファイルと関連した少なくとも１つのメタデータ要素を有するメタデータを用いて、複数のデータファイルを構成するための方法であって、データファイルの少なくともいくつかについて、データファイルと関連した少なくとも１つのメタデータ要素を抽出し、抽出されたメタデータ要素についての値に基づいて、抽出されたメタデータ要素を要求された順序に構成し、少なくとも１つのパラメータ値を入力し、抽出されたメタデータ要素及び入力パラメータ値に基づいて、データファイルの少なくともいくつかを群に分割する、ことを含む方法である。

【００１７】

本発明の第２の態様は、第１の態様において、少なくともいくつかのデータファイルを分割することが、少なくとも１つのパラメータ値の少なくとも１つの各々について、抽出されたメタデータ要素及びそのパラメータ値を用いて、複数のデータファイルの少なくとも２つについての類似性値を判断することを含む、方法である。

【００１８】

本発明の第３の態様は、第２の態様において、少なくとも１つの類似性値を判断することが、次の数式で少なくとも１つの類似性値を判断することを含み、

$$S_K(i,j) = \exp\left(-\frac{|t_i - t_j|}{K}\right)$$

S_K（ｉ，ｊ）がｉthデータファイル及びｊthデータファイルについての類似性値であり、Ｋがパラメータ値であり、t_i及びt_jが、ｉth及びｊthデータファイルについての少なくとも１つの抽出されたメタデータ要素の少なくとも一つのメタデータ要素の実際の値である、方法である。

【００１９】

本発明の第４の態様は、第２の態様において、少なくとも１つの類似性値を判断することが、次の数式で少なくとも１つの類似性値を判断することを含み、

$$S_K(i,j) = \exp\left(\frac{1}{K}\left(\frac{\langle v_i, v_j \rangle}{|v_i||v_j|} - 1\right)\right)$$

S_K（ｉ，ｊ）がｉthデータファイル及びｊthデータファイルについての類似性値であ

り、Ｋがパラメータ値であり、v_i及びv_jが、i^{th}及びj^{th}データファイルから判断された実際のベクトル値である、方法である。

【００２０】

本発明の第５の態様は、第２の態様において、少なくともいくつかのデータファイルの各々について、そのデータファイルについての及び多数の近くのデータファイルについての少なくとも１つの類似性値に基づいて、そのデータファイルについての少なくとも１つの新規性値を判断することをさらに含む、方法である。

【００２１】

本発明の第６の態様は、第５の態様において、少なくとも１つの新規性値を判断することが、次の数式で少なくとも１つの新規性値を判断することを含み、

$$v_K(s) = \sum_{l,n=-5}^{5} S_K(s+l, s+n)g(l,n)$$

v_K（ｓ）が新規性値であり、ｇがガウシアンテーパの１１ｘ１１チェッカーボードカーネルである、方法である。

【００２２】

本発明の第７の態様は、第５の態様において、データファイルの少なくともいくつかについて判断された少なくとも１つの新規性値に基づいて、複数のデータファイルの境界位置間の少なくとも１つの境界位置を判断することをさらに含む、方法である。

【００２３】

本発明の第８の態様は、第７の態様において、判断された境界位置の少なくともいくつかについて、その境界位置についての信頼値を判断することをさらに含む、方法である。

【００２４】

本発明の第９の態様は、第８の態様において、境界位置についての信頼値を判断することが、次の数式で信頼値を判断することを含み、

$$C(B_K) = \sum_{l=1}^{|B_K|-1} \frac{1}{(b_{l+1}-b_l)^2} \sum_{i,j=b_l}^{b_{l+1}} S_K(i,j) - \sum_{l=1}^{|B_K|-2} \frac{1}{(b_{l+1}-b_l)(b_{l+2}-b_{l+1})} \sum_{i=b_l}^{b_{l+1}} \sum_{j=b_{l+1}}^{b_{l+2}} S_K(i,j)$$

Ｃ（B_K）が$B_K{}^{th}$境界についての信頼値であり、S_K（ｉ，ｊ）がi^{th}データファイル及びj^{th}データファイルについての類似性値であり、ｂが入力パラメータＫレベルについての特定の値で検出された境界のインデックス値である、方法である。

【００２５】

本発明の第１０の態様は、第８の態様において、判断された境界位置の少なくともいくつかについて、信頼値を最大にする少なくとも１つのパラメータ値の少なくとも１つを判断することをさらに含む、方法である。

【００２６】

本発明の第１１の態様は、データファイルの対応するメタデータ要素と少なくとも関連した少なくとも１つのメタデータ要素を有するメタデータを用いて、複数のデータファイルを構成するための方法であって、各メタデータがデータファイルに対応している、少なくとも１つのメタデータのセットを処理することと、メタデータを分析するために要求されたパラメータ値を得ることと、得られたパラメータ値を用いてメタデータ要素のセット内の構造を判断することであって、複数のデータファイルの少なくともサブセットについて、パラメータ値を互いに用いてメタデータの少なくともサブセットを比較することによって構造が判断される、構造を判断することと、を含む、方法である。

【００２７】

本発明の第１２の態様は、第１１の態様において、メタデータ要素の判断された構造を用いて、データファイルを群にクラスタ化することをさらに含む、方法である。

【００２８】

本発明の第１３の態様は、第１２の態様において、データファイルの判断されたクラスタから境界を判断することをさらに含み、境界はデータファイルの判断されたクラスタ間に位置している、方法である。

【００２９】

本発明の第１４の態様は、第１３の態様において、データファイルの１つのクラスタ内のメタデータ要素の少なくともいくつかを、データファイルの要素クラスタ内のメタデータ要素の少なくともいくつかと比較することによって、類似性値を判断することと、データファイルの１つのクラスタ内のメタデータ要素の少なくともいくつかを、データファイルの他のクラスタ内のメタデータ要素の少なくともいくつかと比較することによって、非類似性値を判断することと、をさらに含む、方法である。

【００３０】

本発明の第１５の態様は、第１４の態様において、類似性値と非類似性値との相違に基づいて、データファイルのクラスタの要求された群に対応するパラメータ値を判断することをさらに含む、方法である。

【００３１】

本発明の第１６の態様は、データ処理装置上で実行可能であると共に、少なくとも各データファイルと関連した少なくとも１つのメタデータ要素を有するメタデータを用いることによって複数のデータファイルを構成するために用いられ得る、プログラムであって、プログラムが、データファイルの少なくともいくつかについて、そのデータファイルと関連した少なくとも１つのメタデータ要素を抽出するための命令と、抽出されたメタデータ要素についての値に基づいて、抽出されたメタデータ要素を要求された順序に構成するための命令と、パラメータ値を入力するための命令と、抽出されたメタデータ要素及び入力パラメータ値に基づいて、データファイルの少なくともいくつかを群に分割するための命令と、を含む、プログラムである。

【００３２】

本発明の第１７の態様は、第１６の態様において、データファイルの少なくともいくつかを群に分割するための命令が、少なくとも１つのパラメータ値の少なくとも１つの各々について、抽出されたメタデータ要素の少なくともいくつか及びそのパラメータ値を用いて、複数のデータファイルの少なくとも２つについての類似性値を判断するための命令をさらに含む、プログラムである。

【００３３】

本発明の第１８の態様は、第１７の態様において、少なくともいくつかのデータファイルの各々について、そのデータファイルについての及び多数の近くのデータファイルについての少なくとも１つの類似性値に基づいて、そのデータファイルについての少なくとも１つの新規性値を判断するための命令をさらに含む、プログラムである。

【００３４】

本発明の第１９の態様は、第１７の態様において、少なくとも１つの類似性値を判断するための命令が、次の数式で少なくとも１つの類似性値を判断するための命令を含み、

$$S_K(i,j) = \exp\left(-\frac{|t_i - t_j|}{K}\right)$$

S_K（ｉ，ｊ）がｉthデータファイル及びｊthデータファイルについての類似性値であり、Ｋがパラメータ値であり、t_i及びt_jが、ｉth及びｊthデータファイルについての少なくとも１つの抽出されたメタデータ要素の少なくとも一つのメタデータ要素の実際の値である、プログラムである。

【００３５】

本発明の第２０の態様は、第１７の態様において、少なくとも１つの類似性値を判断するための命令が、次の数式で少なくとも１つの類似性値を判断するための命令を含み、

$$S_K(i,j) = \exp\left(\frac{1}{K}\left(\frac{\langle v_i, v_j \rangle}{|v_i||v_j|} - 1\right)\right)$$

S_K（i，j）がithデータファイル及びjthデータファイルについての類似性値であり、Kがパラメータ値であり、v_i及びv_jが、ith及びjthデータファイルから判断された実際のベクトル値である、プログラムである。

【００３６】

本発明の第２１の態様は、第１８の態様において、データファイルの少なくともいくつかについて判断された少なくとも１つの新規性値に基づいて、複数のデータファイルの境界位置間の少なくとも１つの境界位置を判断するための命令をさらに含む、プログラムである。

【００３７】

本発明の第２２の態様は、第１８の態様において、少なくとも１つの新規性値を判断するための命令が、次の数式で少なくとも１つの新規性値を判断する命令を含み、

$$v_K(s) = \sum_{l,n=-5}^{5} S_K(s+l, s+n)\, g(l,n)$$

v_K（s）が新規性値であり、gがガウシアンテーパの１１ｘ１１チェッカーボードカーネルである、プログラムである。

【００３８】

本発明の第２３の態様は、第２１の態様において、判断された境界位置の少なくともいくつかについて、その境界位置についての信頼値を判断するための命令をさらに含む、プログラムである。

【００３９】

本発明の第２４の態様は、第２３の態様において、少なくとも１つの信頼値を判断するための命令が、次の数式でこのような信頼値の各々を判断するための命令を含み、

$$C(B_K) = \sum_{l=1}^{|B_K|-1} \frac{1}{(b_{l+1} - b_l)^2} \sum_{i,j=b_l}^{b_{l+1}} S_K(i,j) - \sum_{l=1}^{|B_K|-2} \frac{1}{(b_{l+1} - b_l)(b_{l+2} - b_{l+1})} \sum_{i=b_l}^{b_{l+1}} \sum_{j=b_{l+1}}^{b_{l+2}} S_K(i,j)$$

C（B_K）が$B_K$th境界についての信頼値であり、S_K（i，j）がithデータファイル及びjthデータファイルについての類似性値であり、bがあるレベルの検出された境界である、プログラムである。

【００４０】

本発明の第２５の態様は、第２３の態様において、本発明の判断された境界位置の少なくともいくつかについて、信頼値を最大にする少なくとも１つのパラメータ値の少なくとも１つを判断するための命令をさらに含む、プログラムである。

【００４１】

本発明の第２６の態様は、データファイルの対応するメタデータ要素と少なくとも関連した、少なくとも１つのメタデータ要素を有するメタデータを用いて、複数のデータファイルを構成するために用いられ得る、データファイル構成システムであって、データファイルの少なくともいくつかについて、データファイルと関連した少なくとも１つのメタデータ要素を抽出する、メタデータ抽出回路、ルーティン、又はアプリケーションと、抽出されたメタデータについての値に基づいて、抽出されたメタデータ要素を要求された順序に構成するための、メタデータ構成回路、ルーティン、又はアプリケーションと、少なくとも１つのパラメータ値の少なくとも１つについて、抽出されたメタデータ要素の少なく

ともいくつか及びそのパラメータ値を用いて、複数のデータファイルの少なくとも２つについての類似性値を判断する、類似性値判断回路、ルーティン、又はアプリケーションと、データファイルについての及び多数の近くのデータファイルについての少なくとも１つの類似性値に基づいて、そのデータファイルについての少なくとも１つの新規性値を判断する、新規性値判断回路、ルーティン、又はアプリケーションと、データファイルの少なくともいくつかについて判断された少なくとも１つの新規性値に基づいて複数のデータファイルの境界位置間の少なくとも１つの境界位置を判断することによって、抽出されたメタデータ要素及び入力パラメータ値に基づいて、データファイルの少なくともいくつかを群に分割する、データ分割回路、ルーティン、又はアプリケーションと、判断された境界位置の少なくともいくつかについて、その境界位置についての信頼値を判断し、判断された境界位置の少なくともいくつかについて、データ分割回路、ルーティン、又はアプリケーションが、信頼値を最大にする少なくとも１つのパラメータ値をさらに判断する、信頼値判断回路、ルーティン、又はアプリケーションと、を備える、データファイル構成システムである。

【発明を実施するための最良の形態】

【００４２】

本発明のこれらの又は他の特徴及び利点は、本発明による方法及び装置の種々の例示的な実施の形態の次の詳細な説明に記載されており、又はこれらから明らかである。

【００４３】

本発明の種々の例示的な実施の形態を、添付の図を参照して詳細に説明する。

【００４４】

本発明によるシステム及び方法の種々の例示的な実施の形態の次の詳細な説明は、データファイルに対応するメタデータの処理に基づいて要求されたデータを構成することを主眼にしている。しかしながら、当然のことながら、本発明は開示された例示的な実施の形態のみに限定されない。一般には、本発明は、対応するメタデータを用いて多数のデータを構成する、任意の方法又は装置に用いられ得る。なお、本発明は、コンピュータを使用して実施されてもよい。また、そのようなコンピュータが端末として複数備わるネットワークシステム上で実施されてもよい。本発明の実施の形態におけるコンピュータは、少なくとも演算処理を行うプロセッサ、データ及びインストラクションをユーザが入力するための入力手段、データ及び処理結果を出力する出力手段、及び、データ及びプログラムを記憶する記憶手段を備える。以下に説明する本発明の実施の形態のシステムまたは方法は、当該システムまたは方法を実行するためのプログラムを記憶手段に記憶し、当該記憶手段から当該プログラムを読み出し、該プロセッサにより実行するようにしてもよい。

【００４５】

図１は、本発明によるデータを構成するための方法の１つの例示的な実施の形態を概説する、フローチャートである。種々の例示的な実施の形態において、図１で概説された方法は、複数のデータファイル内の及び／又は関連したメタデータに基づいて、任意の要求されたデータの種類の複数のデータファイルを構成するために、用いられ得る。

【００４６】

図１で示されるように、本方法の動作は、ステップＳ１００で始まり、Ｓ２００に続き、各データファイルのメタデータの少なくとも１つの要素が、構成されるべき複数のデータファイルから抽出される。次に、ステップＳ３００では、抽出されたメタデータの要素は、抽出されたメタデータ要素についての１つ以上の値に基づいて、セットに構成され、たとえば、セット内での要求された順序及び識別という、指示が与えられる。動作は、次に、ステップＳ４００に続く。

【００４７】

ステップＳ４００では、パラメータＫについての値が選択される。次に、ステップＳ５００では、メタデータが、要求通りに階層的に構成される。動作は、次に、ステップＳ６００に続き、本方法の動作が終了する。

【００４８】

当然のことながら、種々の例示的な実施の形態において、たとえば、メタデータの少なくとも１つの抽出された要素がタイムスタンプ要素を含む場合には、抽出されたメタデータ要素は時間的順序で構成されてもよい。あるいは、メタデータの少なくとも１つの抽出された要素がファイル名又は他のテキストストリングを含む場合には、メタデータ要素はアルファベット順で構成されてもよい。他の種々の例示的な実施の形態において、メタデータの少なくとも１つの抽出されたメタデータ要素が数値データを含む場合には、メタデータ要素は数値順に構成されていてもよい。さらに他の種々の例示的な実施の形態において、メタデータの少なくとも１つの抽出されたメタデータ要素は、たとえばＧＰＳデータ等の位置を定義してもよい。当然のことながら、上述の、時間による、アルファベットによる、数値による、及び／又は位置によるメタデータ要素に加えて又はその代わりに、任意の他の適切なメタデータ要素が、構成する特徴として用いられ得る。これも当然のことであるが、選択されたメタデータ要素の値を順序付け又は構成する、任意の周知又は今後開発される方法が、データファイルを要求された順序に構成するために用いられてもよい。

【００４９】

種々の例示的な実施の形態において、各抽出されたメタデータ要素は要求された識別を与えられるか、又はインデックスが付される。その結果、このような例示的な実施の形態では、各データファイルは、時間、名前、又は位置によって構成するメタデータ構成要素の実際の値に基づいてではなく、データファイルのセット内のそのメタデータ要素の値の位置によって、このように識別される。いいかえれば、たとえば、データファイルのセットは、タイムスタンプメタデータ要素の値に基づいて、時間的順序で構成される。しかしながら、データファイルは、次に、タイムスタンプメタデータ要素の絶対的時間値によってではなく、タイムスタンプメタデータ要素の時間値の観点から、データファイルのセット内に位置した順序によって、識別されるか、又はインデックスが付される。それにもかかわらず、各データファイルについてのメタデータ要素は、その絶対値を保持し続け、後に比較し得る。

【００５０】

種々の例示的な実施の形態において、パラメータＫは、数値を有する。パラメータＫについての入力値は、デフォルト値又は要求値の場合がある。種々の例示的な実施の形態において、パラメータＫは、セット内のデータファイルの各対の選択されたメタデータ要素又はセット内のデータファイル対のサブセット間の対の二つ一組の比較を行なうために、クラスタリング感度を判断する値である。それゆえ、パラメータＫのより大きな値は、結果としてデータファイルのより粗いクラスタリングになる、比較を示す。言い換えれば、パラメータＫのより大きな値は、分離したクラスタに分類されるために、メタデータについての値が互いにより離れていることを要求する。他方、パラメータＫについてのより小さな値は、パラメータＫについてのより大きな又はより小さな値のいずれかにおいて、多かれ少なかれ明らかになるメタデータの特有の特徴を統合するか又は強調するために、調整され得る。

【００５１】

たとえば、パラメータＫについてのより小さな値は、典型的には、非常に細かく離れた値、又はより小さな差異でより明らかになるメタデータの特徴を有するメタデータ要素について、より適切である。対照的に、パラメータＫについてのより大きな値は、典型的には、非常に粗く離れた値又はより大きな差異でより明らかになるメタデータの特徴を有するメタデータ要素について、より適切である。その結果、パラメータＫに対する要求値は、メタデータの種類と、メタデータの間隔と、セット内のメタデータ要素の数とに従って、異なるであろう。それゆえ、種々の例示的な実施の形態において、パラメータＫについての複数の値は、メタデータを完全に分析して比較するために用いられる。このようにして、本発明による種々の例示的な実施の形態において、メタデータ要素の入力セットの先天的な分布に関して、仮定は行なわれない。パラメータＫについてのこのような値を用いて分類及び／又は比較され得るメタデータの種々の例示的な種類は、たとえば、低レベル

画像特徴、ＧＰＳデータ、時間、月、及び／又は年におけるタイムスタンプを含む。

【００５２】

図２は、ステップＳ５００の要求されたメタデータを階層的に構成するための方法の１つの例示的な実施の形態をより詳細に概説する、フローチャートである。種々の例示的な実施の形態において、図２で概説された方法は、データファイルの任意の要求されたセットが、そのメタデータを用いることによって構成するために、用いられ得る。

【００５３】

図２で示されたように、本方法の動作は、ステップＳ５００で始まり、ステップＳ５１０に続き、ここでパラメータＫについての値のリストが得られる。次に、ステップＳ５２０で、パラメータＫについての値のリストから、最初の又は次の値が選択される。動作は、次に、ステップＳ５３０に続く。

【００５４】

パラメータＫについての値のリストは、ステップＳ４００で選択されたパラメータＫについての値に対応する。種々の例示的な実施の形態において、パラメータＫについての複数の異なる値を含む、パラメータＫについての値のリストは、たとえばメタデータ値のクイックスキャンに基づいて無作為に自動的に生じ得るか、又は手動で入力され得るかのいずれかである。種々の例示的な実施の形態において、リスト内のパラメータＫについての値は、パラメータＫについての複数の値を含む。

【００５５】

ステップＳ５３０では、リスト内のインデックスが付されたメタデータ要素の各対について類似性値S_Kを得るために、リスト内のパラメータＫについての値の各々が用いられる。

$$S_K(i,j) = \exp\left(-\frac{|t_i - t_j|}{K}\right) \qquad (1)$$

ここで、S_K（ｉ，ｊ）はi^{th}及びj^{th}データファイルについての類似性値であり、

Ｋはパラメータの値であり、

t_i及びt_jは、i^{th}及びj^{th}データファイルの選択されたメタデータ要素の実際の値である。

【００５６】

パラメータＫについての特定の値を用いる、メタデータ要素の各比較された対についての類似性値S_Kの集合は、類似性行列として表現され得る。

【００５７】

言い換えれば、i^{th}及びj^{th}データファイルのメタデータ要素の値t_i及びt_jについての類似性値S_Kを得るために、i^{th}及びj^{th}データファイルについてのメタデータは、パラメータＫに基づいて比較し得る。ｔ値がメタデータの実際の値であるから、１つの例示的な実施の形態において、メタデータがタイムスタンプである場合には、ｔは分単位の時間であってもよい。

【００５８】

類似性値S_Kを得るために用いられ得るメタデータ要素の実際の値の種類は、時間等のスカラ値である必要はない。類似性値S_Kを得るために、他の種類のメタデータ要素が用いられ得る。種々の例示的な実施の形態において、コンテンツベースの特徴ベクトルが、メタデータと共に、又は代わりにも用いられ得る。この場合には、類似性値は次のとおりである。

$$S_K(i,j) = \exp\left(\frac{1}{K}\left(\frac{\langle v_i, v_j\rangle}{|v_i||v_j|} - 1\right)\right) \qquad (2)$$

ここで、v_i及びv_jは、i^{th}及びj^{th}データファイルの選択されたメタデータ要素についての実際のベクトルである。他の適切な種類の値及び数式が、種々の他の例示的な実施の形態において用いられてもよい。動作は、次に、ステップＳ５４０に続く。

【００５９】

ステップＳ５４０では、パラメータＫについての特定の値について生じた類似性行列S_Kの各要素について、新規性スコアv_Kが得られる。新規性シェアv_Kを得るための１つの方法は、類似性行列S_K（ｉ，ｊ）の主要な対角線Ｓ（ｉ，ｊ）に沿ってカーネルを相関させるために、適合したフィルタ技術を用いることである。すなわち、種々の例示的な実施の形態において、新規性スコアv_Kは、類似性行列S_Kの対角線に沿ってのみ判断される。メタデータの群間の実際の境界を見出すために、種々の例示的な実施の形態において、新規性スコアv_K（ｓ）を計算するために、ガウシアンテーパの１１ｘ１１チェッカーボードカーネルｇが、次のように用いられる。

$$v_K(s) = \sum_{l,n=-5}^{5} S_K(s+l, s+n)\, g(l,n) \qquad (3)$$

ここで、v_K（ｓ）は、パラメータＫについての特定の値についての類似性行列S_Kのi^{th}要素及びガウシアンテーパの１１ｘ１１チェッカーボードカーネルｇについての新規性スコアである。

【００６０】

数式（３）では、１１ｘ１１行列のゆえに、-５と+５との間のｌ及びｎの範囲についての値が用いられる。種々の例示的な実施の形態において、たとえば、９ｘ９行列等、他の寸法の行列が用いられてもよく、ここでは、-４と４との間のｊ及びｋの範囲についての値である。新規性スコアv_Kを得るために、任意の要求された寸法のチェッカーボードカーネルを用いてもよい。

【００６１】

チェッカーボードカーネルを用いることによって、完全な分析を実行する必要はない。むしろ、カーネルと同じ幅を有する主要な対角線の周囲のストリップのみを得る必要があり、それによって、データファイルの数に直線的に一致する、計算の複雑さを減らす。注目すべきことは、データのすべての可能な対ではなく、データの対のサブセットのみの比較が、任意の二つ一組の比較において用いられてもよいことである。一般には、すべての可能な対のサブセットのみを用いることが、最小の性能劣化を伴う実質的な計算の節約につながる。

【００６２】

パラメータＫの種々の値について新規性スコアv_Kが判断される時には、新規性スコア内にいくつかのピークが現れる。注目すべきことは、パラメータＫの異なる値について異なるピークが現れることである。パラメータＫについての値が構造の範囲を示すから、パラメータＫについての異なる値は、類似性行列S_Kが異なる解像度での構造を明らかにすることを可能にする。新規性スコアv_K内のピークは、次に、他の群と類似又はより近接したメタデータ要素値を有する、連接したデータ群間の境界の階層的なセット、すなわちクラスタを示す。それゆえ、新規性スコアv_K内のピークは、類似したメタデータ値を有する群間の境界であり、他のクラスタから分離可能なメタデータ値のクラスタを示す。それゆえ、メタデータの群間の境界である、新規性スコアv_K内のピークが得られる。動作は、次に、ステップＳ５５０に続く。

【００６３】
ステップＳ５５０では、最初に、パラメータＫの各値について新規性スコアv_K内にすべてのピークを配置し、次に、検出された境界上に階層的構造を適用することによって、パラメータＫの各異なる値についての境界リストが得られる。種々の例示的な実施の形態において、パラメータＫの値のリスト内の各値を用いて、より粗いスケールからより細かいスケールまで、すなわちパラメータＫについての値を減少して、境界リストを得るための分析が行なわれる。ピーク値又は境界の階層的なセットを作るために、検出されたすべての境界を含むであろう境界リストB_K＝[ｂ１，．．b_{nk}]を用いて、パラメータＫの各値についての新規性スコアv_K内のすべてのピークが、次に集められる。すなわち、粗いスケールすなわちパラメータＫのより大きい値で検出されたすべての境界は、すべてのより細かいスケールすなわちパラメータＫのより少ない値についての境界リスト内に含まれるであろう。より粗いスケールで得られたより遠く離れた群間の境界は、依然としてより細かいスケールで存在すると仮定される。

【００６４】
新規性スコアv_Kが、局所的な最大値にあって、類似性測定の最大及び類似性行列の主要な対角線に沿って相関されたカーネルから判断されるところに、境界が位置する。新規性スコアの最大又は最小を得る他の方法は、たとえば数式（３）の導関数を得ることである。動作は、次に、ステップＳ５６０に続く。

【００６５】
ステップＳ５６０では、パラメータＫの各値について、類似性値S_Kと、新規性スコアV_Kと、境界b_Kとを得ることによって、境界を判断するためにリスト内のパラメータＫについてのすべての値が用いられたかが判断される。否の場合には、動作はステップＳ５２０に戻る。さもなければ、動作は、ステップ５７０に続く。

【００６６】
ステップＳ５７０では、境界B_Kのリストによって表された検出された境界は、検出された境界の階層内の各レベルについてランクされて来たクラスタリングの結果を表わす、信頼スコアＣ（B_K）を得るために用いられる。信頼スコアＣ（B_K）は、次に数式によって示された平均的なクラス内の類似性とクラス間の非類似性とに基づく。

$$C(B_K)=\sum_{l=1}^{|B_K|-1}\frac{1}{(b_{l+1}-b_l)^2}\sum_{i,j=b_l}^{b_{l+1}}S_K(i,j)-\sum_{l=1}^{|B_K|-2}\frac{1}{(b_{l+1}-b_l)(b_{l+2}-b_{l+1})}\sum_{i=b_l}^{b_{l+1}}\sum_{j=b_{l+1}}^{b_{l+2}}S_K(i,j) \qquad (4)$$

ここで、Ｃ（B_K）は信頼スコアであり、
ｂは各レベルにおける検出された境界である。

【００６７】
上述のように、各クラスタ内のデータファイル間の平均的なクラス内の類似性を定量化する第１の合計と、隣接したクラスタ内のデータファイル間の平均的なクラス間の類似性を定量化する第２の合計とは、クラスタ間の非類似性を定量化するために、無効にされる。第１の合計及び第２の合計についての変化率は、パラメータＫの値に従って変化する。それゆえ、パラメータＫについての複数の値については、１つの値は、信頼スコアＣ（B_K）が最大化されることを可能にするであろう。その結果、動作はステップＳ５８０に続き、そこでは、信頼スコアＣ（B_K）を最大にするパラメータＫの値についての境界リストB_Kが得られる。次に、動作は、ステップＳ５９０に続き、ここで動作がステップＳ６００に戻る。ベイズ情報基準（ＢＩＣ）等の信頼スコアＣ（B_K）を得るために、他の種類の統計的測定が用いられ得る。ベイズ情報基準のいくつかの例が、Ｄ．ハッカーマン（D. Heckermann）著「ベイズのネットワークを学ぶに当っての指導書」（“A tutorial on learning with Bayesian networks”），技術レポート（Technical Report）ＭＳＲ－ＴＲ－９５－０６，マイクロソフト・リサーチ（Microsoft Research）、レッドモンド、ワシントン（１９９５年。１９９６年改定）（非特許文献１）、Ｓ．チェンら（S. Chen et al.）「ベイズの情報基準による、話者、環境、及びチャンネル変化の検出及びクラスタ

リング」（“Speaker, environmental and change detection and clustering via the Bayesian information criterion”）ＤＡＲＰＡ音声認識ワークショップ（DARPA Speech Recognition Workshop）（１９９８年）（非特許文献６）、及びＳ．レナルら（S. Renal et al.）著「会議室からのオーディオ情報アクセス」（“Audio Information Access from Meeting Rooms”）（２００３年４月）（非特許文献７）で述べられており、それらの各々は、本明細書にその全体を参照することによって組み込まれる。

【００６８】

本発明によるシステム及び方法の１つの例示的な使用は、階層的なクラスタリングによってデジタル写真をタイムベースのイベントに構成することを含む。デジタルカメラの急増と共に、パーソナルコンピュータ上に蓄積するデジタル写真の数が急速に増えている。典型的にはＪＰＥＧ画像ファイルフォーマットである個々のデジタル画像ファイルは、典型的にはExif（エグジフ、Exchangeable Image File Format）で保存されている、デジタルファイル内の大量のメタデータを含む。このようなメタデータは、いつ写真が撮影されたか又はその後にいつ再保存若しくは修正されたかを示すタイムスタンプを含む。それにもかかわらず、複数のメタデータは画像ファイルと共に記録されてもよいから、オリジナルのタイムスタンプ、又は任意のその後に修正されたタイムスタンプ等の情報は、メタデータとして別々に記録されてもよく、本発明によるシステム及び方法の種々の例示的な実施の形態を用いて、個別に抽出及び分析可能である。

【００６９】

１つの例示的な実施の形態において、５１２の写真のクラスタリングが用いられた。最初に、すべての写真がタイムスタンプ（メタデータ）を有していて、撮影者によって、意味のあるフォルダ、すなわち特有のイベント、の中に手動で置かれた。これらの写真のこの手動のクラスタリングは、グランドトルースクラスタリング（ground truth clustering）として、次の説明において参照されるであろう。

【００７０】

各写真についてのＥｘｉｆヘッダが、その写真についてのタイムスタンプを抽出するために、最初に処理された。抽出されたタイムスタンプは、最初に構成され、時間で順序付けられた。タイムスタンプは、分（minutes）等の任意の基本的な時間単位を用いて、時間順に順序付けられた。しかしながら、一旦タイムスタンプが時間順に順序付けられると、次に、各タイムスタンプ及びこのような各対応する写真は、インデックス又は時間順の番号若しくは値を付与され、引き続いてその後、タイムスタンプの絶対的な時間値によってではなく、このインデックスによって参照された。

【００７１】

タイムスタンプを抽出して写真を構成するための最初の処理の後に、タイムスタンプの集合の構造は、類似性行列S_Kを作ることによって評価された。図３は、グランドトルースクラスタリングから生じた類似性行列S_Kについて得られた結果を図示的に示す。図３で図示的に表現された類似性行列S_Kの要素についての値は、同一のフォルダからの写真の対については１であり、撮影者によって異なるフオルダ内に保存された写真の対については０である。写真は、上述のように、時間順にインデックスが付される。類似性行列S_Kの（ｉ，ｊ）要素についての値を判断するために、その中にi^{th}及びj^{th}写真が保存されたフォルダの名前が比較される。それらが同じ場合には、（ｉ，ｊ）要素は、１の値が割り当てられる。さもなければ、それは、０の値が割り当てられる。種々の例示的な実施の形態において、行列の主要な対角線に沿った類似性行列S_Kの要素のブロックは、各フォルダ内の写真の群と対応する。

【００７２】

図３で示された類似性行列S_Kの主要な対角線に沿ったチェッカーボードパターンは、すでに異なるイベントに分類された写真を含むフォルダ間の境界を示す。それゆえ、チェッカーボードパターンは、異なるイベントの写真の群間の時間的順序での境界の図示的な表現である。類似性行列のi^{th}及びj^{th}要素として写真が表現されたときには、チェッカーボードパターンは、写真が記述するイベントも時間においてまとまりがない一方、類似

性行列内で写真が隣接していることを示す。

【００７３】

図４は、グランドトルースクラスタリングについて生じた新規性スコアV_Kを示す。新規性スコアv_Kは、ガウシアンテーパの１１ｘ１１チェッカーボードカーネルｇを用いて得られる。図４は、図３で示されたチェッカーボードに対応する新規性スコアv_Kのピークを示す。たとえば、図３では、２つの黒い正方形によって示された２つの比較的大きな群は、インデックス値２１０の近くで分離される。２つの正方形は、インデックス値２１０の近くで接触するだけである。２つの正方形が単に接触する位置では、写真の２つの群間の境界を示す。図４では、この境界を示すインデックス値２１０の近くの新規性スコアv_Kに、対応するピークがある。

【００７４】

図５～１０は、グランドトルースクラスタリングでクラスタ化された写真を用いて、10^3分、10^4分、10^5分のパラメータＫの値について得られた、いくつかの類似性行列S_K及びそれらの対応する新規性スコアv_Kを示す。図５、７、９は、それぞれ、10^3分、10^4分、10^5分のパラメータＫの値についての類似性行列S_Kを示す。図６、８、１０は、それぞれ、10^3分、10^4分、10^5分のパラメータＫの値についての新規性スコアv_Kを示す。パラメータＫについての３つの異なる値は、３つの異なる解像度を表す。特に、パラメータＫについての値が少ない程、解像度は大きくなり、ここではタイムスタンプの群間のより細かい非類似性が明らかになる。

【００７５】

図５、７、９で示されたように、類似性行列S_Kは、異なる解像度での構造を明らかにする。それにもかかわらず、パラメータＫについてのより大きな値では、詳細は、パラメータＫについてのより小さい値について程容易には現れない。パラメータＫについての値を用いている極端な例が、図１２及び１３で示される。２つの異なる類似性行列で現れる例示的な写真のインデックスを用いて、図１２は、パラメータＫ（Ｋ＝１０）についての１０の値について得られた類似性行列の部位を示す。図１３は、パラメータＫ（Ｋ＝１，０００）についての１，０００の値について得られた類似性行列の部位を示す。図１２及び１３で示されたように、パラメータＫについてのより大きな値について得られるよりも、パラメータＫについてのより小さな値については、より良い境界の定義が得られる。これは、数式（１）によって、境界のいずれかの側におけるクラスタ内の写真は、パラメータＫの異なる値については異なるクラス内の類似性を示すことから生じる。これは、次に、チェッカーボードカーネルとの相関の強度を変化させる。それゆえ、類似性測定S_Kは、低レベル画像特徴、ＧＰＳデータ、又は他のメタデータ等の他の特徴を統合又は強調するために、加除可能である。

【００７６】

上記のように、異なる特徴は、パラメータＫの異なる値で、より明らかになる。対応する新規性スコアv_Kでは、境界点は、分析のスケール、すなわちパラメータＫの値に応じて著しく変化する。図６、８、１０では、パラメータＫの値の限られた数についての新規性スコアv_Kが示されている。しかしながら、図１１では、パラメータＫの値のより大きな数についての新規性スコアv_Kが示されている。図１１で示されたように、新規性スコアv_Kは、パラメータＫの値と共に大きく変化し、新規性スコアv_Kは、異なるスケールすなわちパラメータＫの値で、異なる境界ピークを示す。これは、異なるイベントが異なる時間範囲を有するから生じる。すなわち、休暇又は誕生パーティ等のイベントは、異なる時間範囲を有するであろう。たとえば、後者のイベントは、一般には前者のイベントの時間的範囲に比べてより短い時間範囲を有するであろう。

【００７７】

図１１では、最小の新規性スコアv_Kは、$S_{(K)}$内の高い自己類似性の、すなわち低い新規性の領域に対応する。このようにして、領域は、このような高い自己類似性の領域間に、優先的に位置している。境界は、パラメータＫの値を減らすことによって順序付けられ、検出された境界上に、階層的構造が与えられる。このような階層は、検出された境界上

に適用されてもよい。言い換えれば、より細かいスケールについての境界のセット内に、非常に粗いスケール（高いＫ値）からのすべての検出された境界が含まれているところでは、階層的境界のセットが作られても良い。この技術を用いると、より重要でない境界がさらに検出されるにつれて、より重要な境界が保持されることを可能にする。

【００７８】

本技術は、あるスケールでは、すなわちパラメータＫのある値については、検出されたイベント境界が最大の新規性スコアに近づくはずである、という仮定をもとにしている。パラメータＫの各値については、境界を示す新規性スコアv_K内のピークは、第１の相違の分析によって検出される。所与のしきい値スコアを用いると、たとえば、同じイベントである写真中の時間値内にある異常に長いギャップのゆえに、現れるかも知れない擬制のピークを検出することを回避できる。このような所与のしきい値スコアは、最小しきい値スコアとして用いられてもよい。たとえば、５よりも大きい新規性スコアが、各隣接範囲におけるピークとして選択され得る。

【００７９】

図１４は、数式（４）によって表されたように、各クラスタ内の選択されたメタデータ要素についての値間の平均的なクラス内の類似性と、隣接したクラスタ内の選択されたメタデータ要素についての値間の平均的なクラス間の類似性との相違である、推測されたクラスタ内の信頼を定量化する概念を示す。クラス内の類似性条件は、主要な対角線に沿った領域の条件全体の平均である。クラス間の類似性条件は、主要な対角線から離れた矩形領域の平均である。図１４は、信頼スコアの計算を図で示す。

【００８０】

この信頼測定Ｃ（B_K）は、検出されたクラスタの数とパラメータＫの値との両方に明らかに依存する。図１５～１７は、動作を表す。図１５～１７は、数式（４）で定義された信頼測定を形成するために平均化され、合計されたそれぞれの類似性行列S_Kの領域を示す。図１５は、パラメータＫ（Ｋ＝１７７８．２８）についての１７７８．２８の値についての行列を示す。図１６は、Ｋ＝１，０００についての行列を示す。最後に、図１７は、Ｋ＝５６２．３４についての行列を示す。図１５～１７で示された行列の表現では、Ｃ（B_K）に貢献しない要素は、行列内では０にセットされている。図１５～１７では、パラメータＫのより大きな値については、パラメータＫについてのより低い値についてよりも、より低い信頼スコアが得られる。たとえば、Ｋ＝１，０００（図１６）については、信頼スコアＣ（B_K）は２１．０９８８６であり、これはＫ＝１７７８．２８（図１５）についての１１．７８１４の信頼スコアＣ（B_K）よりも大きい。実際、図１６は、より少ない数のクラスタであって、比較的低い類似性についてのクラスタ化された領域を示す。他方、図１７のＫ＝５６２．３４についての行列は、図１６のＫ＝１，０００についての行列よりも多いクラスタを示す。しかし、パラメータＫの値がより小さいゆえに、低い類似性の領域がクラスタ化される。このようにして、当然のことながら、種々の例示的な実施の形態において、信頼測定によって、類似性分析についての１つの適切なスケールが強調される。

【００８１】

図１８は、本発明によるデータ構成システム１００の１つの例示的な実施の形態の構成図である。図１８で示されたように、データ構成システム１００は、１つ以上の制御及び／又はデータバス及び／又はアプリケーションプログラミングインタフェース１９５によって相互接続された、入力／出力インタフェース１１０、コントローラ１２０、メモリ１３０、メタデータ抽出回路、ルーティン、又はアプリケーション１４０、メタデータ構成回路、ルーティン、又はアプリケーション１５０、類似性値判断回路、ルーティン、又はアプリケーション１６０、新規性値判断回路、ルーティン、又はアプリケーション１７０、データ分割回路、ルーティン、又はアプリケーション１８０、及び信頼値判断回路、ルーティン、又はアプリケーション１９０を含む。

【００８２】

図１８で示されたように、表示装置１０２、１つ以上のユーザ入力装置１０６、データ

送信装置２００、及びデータ受信装置２２０が、リンク１０４、１０８、２１０、２３０によって、それぞれデータ構成システム１００に接続されている。

【００８３】

一般には、図１８で示されたデータ送信装置２００は、データファイル及びそれらの対応するメタデータをデータ構成システム１００に供給することができる、任意の周知又は今後開発される装置であってもよい。一般には、図１８で示されたデータ受信装置２２０は、データ構成システム１００からの任意のデータを受信できる、任意の周知又は今後開発される装置であってもよい。

【００８４】

データ送信装置２００及び／又はデータ受信装置２２０は、データ構成システム１００と一体化されていてもよい。さらに、データ構成システム１００は、捉えられた写真を自動的にフォルダ内に構成するデジタルカメラ等、多機能を実行するより大きなシステム内で、データ送信装置２００及び／又はデータ受信装置２２０に加えて追加の機能を与える装置と一体化されていてもよい。

【００８５】

それぞれの１つ以上のユーザ入力装置１０６の各々は、キーボード、マウス、ジョイスティック、トラックボール、タッチパッド、タッチスクリーン、ペンベースのシステム、マイクロホン、及び関連した音声認識ソフトウェア、又はデータ構成システム１００にデータ及び／又はユーザコマンドを入力するための任意の他の周知又は今後開発される装置等の、複数の入力装置の１つ又は任意の組合せであってもよい。当然のことながら、図１８の１つ以上のユーザ入力装置１０６は、同じ種類の装置である必要はない。

【００８６】

表示装置１０２、１つ以上のユーザ入力装置１０６、データ送信装置２００、及びデータ受信装置２２０をデータ構成システム１００に接続しているリンク１０４、１０８、２１０、２３０のリンクの各々は、信号線、直接ケーブル接続、モデム、ローカルエリアネットワーク、広域ネットワーク、イントラネット、インターネット、任意の他の分散型の処理ネットワーク、又は任意の他の周知又は今後開発される接続装置又は構造であってもよい。当然のことながら、これらのリンク１０４、１０８、２１０、２３０のいずれも、有線又は無線の部位を含んでもよい。一般には、リンク１０４、１０８、２１０、２３０の各々は、それぞれの装置をデータ構成システム１００に接続するために使用可能な、任意の周知又は今後開発される接続システム又は構造を用いて実行され得る。当然のことながら、リンク１０４、１０８、２１０、２３０は、同じ種類である必要はない。

【００８７】

図１８で示されたように、メモリ１３０は、可変、揮発、若しくは不揮発メモリ又は非可変若しくは固定したメモリの、任意の適切な組合せを用いて実行し得る。可変メモリは、揮発又は不揮発のいずれでも、任意の１つ以上の、スタティック又はダイナミックＲＡＭ、フロッピィディスク及びディスクドライブ、書込可能又は再書込可能な光ディスク及びディスクドライブ、ハードディスクドライブ、フラッシュメモリ等を用いて、実行され得る。同様に、非可変又は固定のメモリは、ＣＤ－ＲＯＭ若しくはＤＶＤ－ＲＯＭディスク及びディスクドライブ等、任意の１つ以上のＲＯＭ、ＰＲＯＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ及び光ＲＯＭディスクを用いて、実行され得る。

【００８８】

データ構成システム１００の種々の実施の形態が、プログラムされた汎用コンピュータ、専用コンピュータ、マイクロコンピュータ等の上で実行するソフトウェアとして実行され得る。これも当然のことながら、図１８で示された回路、ルーティン、及び／又はアプリケーションの各々は、適切にプログラムされた汎用データプロセッサの一部位として実行され得る。あるいは、図１８で示された回路、ルーティン、及び／又はアプリケーションの各々は、ＡＳＩＣ、デジタルシグナルプロセッサ（ＤＳＰ）、ＦＰＧＡ，ＰＬＤ、ＰＬＡ、及び／又はＰＡＬ、又は個別論理素子又は個別回路素子内の物理的に異なるハードウェア回路として、実行し得る。一般には、同様に図１及び２で示されたフローチャート

を実行し得る有限状態の機械を実行し得る任意の装置が、データ構成システム１００を実行するために用いられ得る。図１８で示された、特定の形式の回路、ルーティン、アプリケーション、対象、及び／又は管理者は、設計上の選択として解釈し、当業者には明らかでありかつ予期し得るものであろう。当然のことながら、図１８で示された、回路、ルーティン、アプリケーション、対象、及び／又は管理者は、同じ設計である必要はない。

【００８９】

メタデータ抽出回路、ルーティン、又はアプリケーション１４０は、データファイルと関連した少なくとも１つのメタデータ要素を抽出する。各データファイルのメタデータの少なくとも１つの要素は、構成されるべき複数のデータファイルから抽出される。典型的にはＪＰＥＧ画像ファイルフォーマットである、デジタル画像ファイル等のデータファイルは、典型的には標準的な交換可能な画像ファイルフォーマット（Ｅｘｉｆ）で保存された、デジタルファイル内の大量のメタデータを含む。このような抽出可能なメタデータは、写真が撮影された時又は次に再保存若しくは修正された時を示す。

【００９０】

メタデータ構成回路、ルーティン、又はアプリケーション１５０は、抽出されたメタデータ要素についての値に基づいて、抽出されたメタデータ要素を、要求された順序に構成する。抽出されたメタデータ要素は、時間の、アルファベット順の、数字の、及び／又は位置の特徴等の、任意の要求された構成する特徴を用いて構成され、割り当てられた又はインデックスが付された識別値に基づいて、抽出されたメタデータ要素を順序付けることができる。

【００９１】

類似性値判断回路、ルーティン、又はアプリケーション１６０は、少なくとも１つのパラメータ値の少なくとも１つについて、抽出されたメタデータ要素の少なくともいくつか及びそのパラメータ値を用いて、複数のデータファイルの少なくとも２つについての類似性値を判断する。それゆえ、類似性値判断回路、ルーティン、又はアプリケーション１６０は、データファイルの各々のこのような対の類似性値を得るために、パラメータ値を用いて、少なくとも１対のデータファイルについてのメタデータを比較する。

【００９２】

新規性値判断回路、ルーティン、又はアプリケーション１７０は、複数の類似性値に基づいて、データファイルについての少なくとも１つの新規性値を判断する。すなわち、新規性値判断回路、ルーティン、又はアプリケーション１７０は、要求されたデータファイルの数についての類似性値に基づいて、新規性値を判断する。

【００９３】

データ分割回路、ルーティン、又はアプリケーション１８０は、抽出されたメタデータ要素及び入力パラメータ値に基づいて、データファイルの少なくともいくつかを群に分割する。種々の例示的な実施の形態において、データ分割回路、ルーティン、又はアプリケーション１８０は、データファイルの少なくともいくつかについて判断された少なくとも１つの新規性値に基づいて複数のデータファイルの境界位置間の少なくとも１つの境界位置を判断することと、判断された境界位置の少なくともいくつかについて、信頼値を最大にする少なくとも１つのパラメータを判断することとによって、抽出されたメタデータ要素及び入力パラメータ値に基づいて、データファイルの少なくともいくつかを群に分割する。

【００９４】

信頼値判断回路、ルーティン、又はアプリケーション１９０は、判断された境界位置の少なくともいくつかについて、その境界位置についての信頼値を判断する。

【００９５】

動作中には、データ構成システム１００は、それぞれが対応するメタデータを有する複数のデータファイルを入力するか又はさもなければ得て、入力パラメータについての値をリンク２１０を通じてデータ送信装置２００から入力してもよく、及び／又はメモリ１３０から１つ以上のデータファイルを読み出してもよい。入力パラメータは、ユーザ入力装

置１０６を通じて入力されてもよい。データ送信装置２００から得られた場合には、入力／出力インタフェース１１０は、データファイル及び／又は入力パラメータを入力し、コントローラ１２０の制御下で、任意の適切なデータファイルを、メタデータ抽出回路、ルーティン、又はアプリケーション１４０に送る。

【００９６】

メタデータ抽出回路、ルーティン、又はアプリケーション１４０は、入力データファイルの少なくともいくつかと関連した少なくとも１つのメタデータ要素を抽出する。メタデータ抽出回路、ルーティン、又はアプリケーション１４０は、次に、コントローラ１２０の制御下で、抽出されたメタデータ要素をメモリ１３０に保存するか、又は抽出されたメタデータ要素をメタデータ構成回路、ルーティン、又はアプリケーション１５０に直接出力する。メタデータ構成回路、ルーティン、又はアプリケーション１５０は、コントローラ１２０の制御下で、抽出されたメタデータ要素を入力し、抽出されたメタデータ要素についての値に基づいて、抽出されたメタデータ要素を要求された順序に構成する。メタデータ構成回路、ルーティン、又はアプリケーション１５０は、次に、コントローラ１２０の制御下で、順序付けられた抽出されたメタデータをメモリ１３０に保存するか、又は順序付けられた抽出されたメタデータ要素を類似性値判断回路、ルーティン、又はアプリケーション１６０に直接出力する。

【００９７】

類似性値判断回路、ルーティン、又はアプリケーション１６０は、コントローラ１２０の制御下で、順序付けられたメタデータ要素及び／又は対応するデータファイルを入力し、少なくとも１つのパラメータ値の少なくとも１つについて、抽出されたメタデータ要素の少なくともいくつか及び又はそれらのデータファイルの内容及びそのパラメータ値を用いて、複数のデータファイルの２つで構成された対の少なくとも１対についての類似性値を判断する。類似性値判断回路、ルーティン、又はアプリケーション１６０は、次に、コントローラ１２０の制御下で、判断された類似性値をメモリ１３０に保存するか、又は、類似性値を、新規性値判断回路、ルーティン、又はアプリケーション１７０に直接出力する。

【００９８】

新規性値判断回路、ルーティン、又はアプリケーション１７０は、コントローラ１２０の制御下で、類似性値の少なくともいくつかを入力して、入力類似性値と関連した多数のデータファイルのそれぞれについて、そのデータファイルについての類似性値及び要求された数の周囲のデータファイルに基づいて、このようなデータファイルの各々についての少なくとも１つの新規性値を判断する。新規性値判断回路、ルーティン、又はアプリケーション１７０は、次に、コントローラ１２０の制御下で、判断された新規性値をメモリ１３０に保存するか、又は、判断された新規性値をデータ分割回路、ルーティン、又はアプリケーション１８０に直接出力する。

【００９９】

データ分割回路、ルーティン、又はアプリケーション１８０は、コントローラ１２０の制御下で、新規性値の少なくともいくつかを入力し、複数のデータファイルの種々の境界位置間の少なくとも１つの境界位置を判断することによって、データファイルの少なくともいくつかについて判断された少なくとも１つの新規性値に基づいて、対応するデータファイルを群に分割する。データ分割回路、ルーティン、又はアプリケーション１８０は、次に、コントローラ１２０の制御下で、判断された境界位置をメモリ１３０に保存するか、又は判断された境界位置を信頼値判断回路、ルーティン、又はアプリケーション１９０に出力する。

【０１００】

信頼値判断回路、ルーティン、又はアプリケーション１９０は、コントローラ１２０の制御下で、１つ以上の境界位置を入力し、判断された境界位置の少なくともいくつかについて、判断された境界位置の少なくともいくつかについての境界位置についての信頼値を判断する。信頼値判断回路、ルーティン、又はアプリケーション１９０は、次に、コント

ローラ１２０の制御下で、判断された信頼値をメモリに保存するか、又は判断された信頼値をデータ分割回路、ルーティン、又はアプリケーション１８０に出力する。データ分割回路、ルーティン、又はアプリケーション１８０は、次に、判断された境界位置の少なくともいくつかについての信頼値を最大にする少なくとも１つのパラメータ値を判断する。それゆえ、データ構成システム１００の動作中は、入力パラメータ値、抽出された順序付けられたメタデータ要素、及び／又は対応するデータファイルの内容は、順序付けられた抽出されたメタデータ要素、及び／又はデータファイルの対応する内容、及び入力パラメータ値に基づいて、読み出された／受け取られたデータファイルの少なくともいくつかを用いて、群に構成される。分割され、このようにして構成されたデータファイルは、次に、さらにメモリ１３０内に保存され、データ受信装置２２０に出力され、及び／又は表示装置１０２上に表示され得る。

【０１０１】

図１８は、データ構成ユニット１００を表示装置１０２から分離された装置として示しているが、ユーザ入力装置１０６、データ送信装置２００、及び／又はデータ受信装置２２０、及びデータ構成システム１００は、一体化された装置であってもよい。一体化された構成では、表示装置１０２、ユーザ入力装置１０６、データ送信装置２００、及び／又はデータ受信装置２２０からの２つ以上のデータ構成システム１００が、単一の装置に含められてもよい。

【０１０２】

あるいは、データ構成システム１００は、メタデータ抽出回路、ルーティン、又はアプリケーション１４０、メタデータ構成回路、ルーティン、又はアプリケーション１５０、類似性値判断回路、ルーティン、又はアプリケーション１６０、新規性値判断回路、ルーティン、又はアプリケーション１７０、データ分割回路、ルーティン、又はアプリケーション１８０、及び信頼値判断回路、ルーティン、又はアプリケーション１９０、コントローラ１２０、メモリ１３０、及び／又は入力／出力インタフェース１１０を含む、分離された装置であってもよい。さらに、分離された回路、ルーティン、及び／又はアプリケーションとして示されているが、メタデータ抽出回路、ルーティン、又はアプリケーション１４０、メタデータ構成回路、ルーティン、又はアプリケーション１５０、類似性値判断回路、ルーティン、又はアプリケーション１６０、新規性値判断回路、ルーティン、又はアプリケーション１７０、データ分割回路、ルーティン、又はアプリケーション１８０、及び信頼値判断回路、ルーティン、又はアプリケーション１９０は、それ自身が、種々の組み合わせで一体化されていてもよい。

【０１０３】

本発明を上述の概説された例示的な実施の形態に関連して述べてきたが、周知又は現在では予見できないかも知れないにしろ、種々の代替物、修正、変更、改良、及び／又は実質的な等価物が、少なくとも本技術分野における通常の知識を有する者には自明になるかも知れない。したがって、上述した本発明の例示的な実施の形態は、例示的であり、制限しないことを意図されている。本発明の精神及び範囲から逸脱せずに種々の変更が可能である。それゆえ、提出されて、それらが補正されるかも知れない特許請求の範囲は、すべての周知又は今後開発される代替物、修正、変更、改良、及び／又は実質的な等価物を含むように意図されている。

【図面の簡単な説明】

【０１０４】

【図１】本発明によるデータを構成するための方法の１つの例示的な実施の形態を概説するフローチャートである。

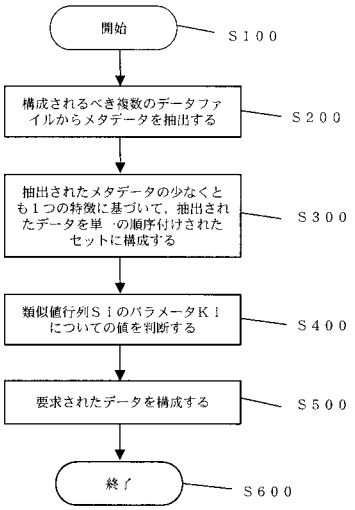
【図２】本発明による要求されたデータを構成するための方法の１つの例示的な実施の形態をより詳細に概説するフローチャートである。

【図３】類似性行列及び新規性スコアについて得られた結果の１つの例示的な実施の形態を図示する。

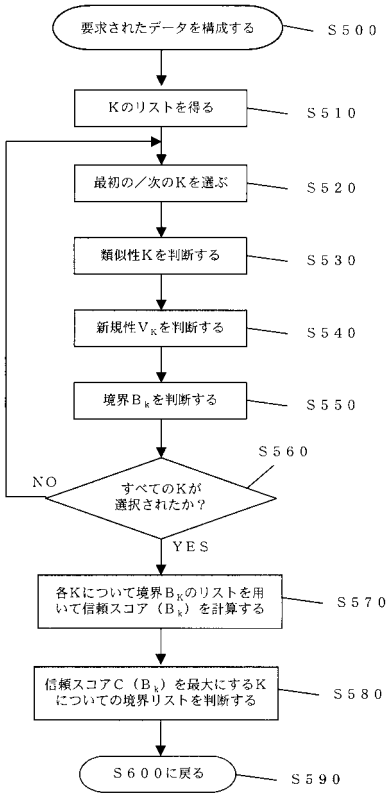
【図４】類似性行列及び新規性スコアについて得られた結果の１つの例示的な実施の形態

を図示する。
【図５】複数の類似性行列について得られた結果及びそれらの対応する新規性スコアの例示的な実施の形態を図示する。
【図６】複数の類似性行列について得られた結果及びそれらの対応する新規性スコアの例示的な実施の形態を図示する。
【図７】複数の類似性行列について得られた結果及びそれらの対応する新規性スコアの例示的な実施の形態を図示する。
【図８】複数の類似性行列について得られた結果及びそれらの対応する新規性スコアの例示的な実施の形態を図示する。
【図９】複数の類似性行列について得られた結果及びそれらの対応する新規性スコアの例示的な実施の形態を図示する。
【図１０】複数の類似性行列について得られた結果及びそれらの対応する新規性スコアの例示的な実施の形態を図示する。
【図１１】パラメータＫ値に応じて変化する境界について判断された新規性スコアの１つの例示的な実施の形態を図示する。
【図１２】２つの異なるパラメータＫ値について判断された類似性行列の例示的な実施の形態を図示する。
【図１３】２つの異なるパラメータＫ値について判断された類似性行列の例示的な実施の形態を図示する。
【図１４】信頼スコアの１つの例示的な実施の形態を示す。
【図１５】３つの異なるパラメータＫ値についての類似性行列の例示的な実施の形態を図示する。
【図１６】３つの異なるパラメータＫ値についての類似性行列の例示的な実施の形態を図示する。
【図１７】３つの異なるパラメータＫ値についての類似性行列の例示的な実施の形態を図示する。
【図１８】本発明によるデータ構成システムの１つの例示的な実施の構成図である。
【符号の説明】
【０１０５】
Ｓ２００：メタデータを抽出する
Ｓ３００：抽出されたデータを順序付けられたセットに構成する
Ｓ４００：パラメータＫについての値を判断する
Ｓ５００：要求されたデータを構成する
１００：データ構成システム
１４０：メタデータ抽出回路、ルーティン、又はアプリケーション
１５０：メタデータ構成回路、ルーティン、又はアプリケーション
１６０：類似性値判断回路、ルーティン、又はアプリケーション
１７０：新規性値判断回路、ルーティン、又はアプリケーション
１８０：データ分割回路、ルーティン、又はアプリケーション
１９０：信頼値判断回路、ルーティン、又はアプリケーション

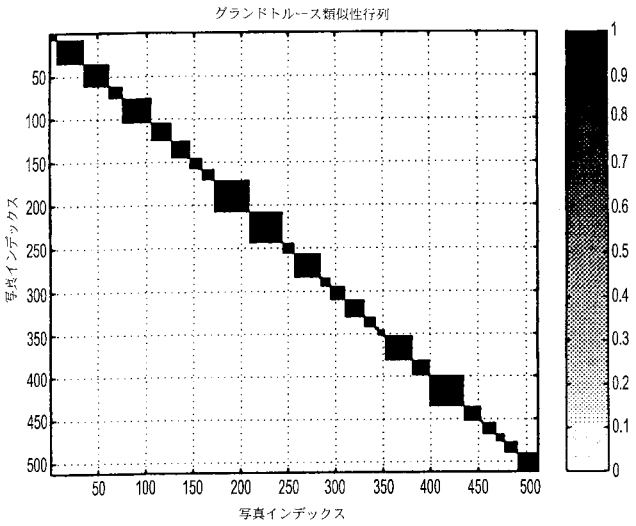
【図１】


開始
S100
構成されるべき複数のデータファイルからメタデータを抽出する
S200
抽出されたメタデータの少なくとも１つの特徴に基づいて，抽出されたデータを単一の順序付けされたセットに構成する
S300
類似値行列Ｓｉのパラメータ Ｋｉについての値を判断する
S400
要求されたデータを構成する
S500
終了
S600

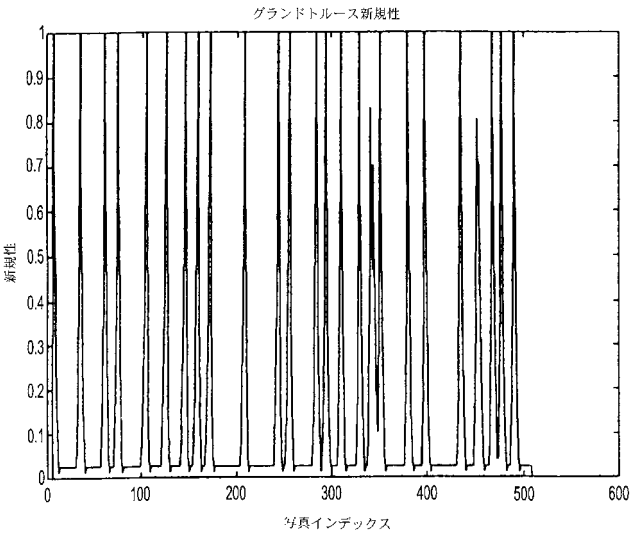
【図２】


要求されたデータを構成する
S500
Kのリストを得る
S510
最初の／次のKを選ぶ
S520
類似性Kを判断する
S530
新規性V_Kを判断する
S540
境界B_Kを判断する
S550
S560
NO
すべてのKが選択されたか？
YES
各Kについて境界B_Kのリストを用いて信頼スコア（B_k）を計算する
S570
信頼スコアC（B_k）を最大にするKについての境界リストを判断する
S580
S600に戻る
S590

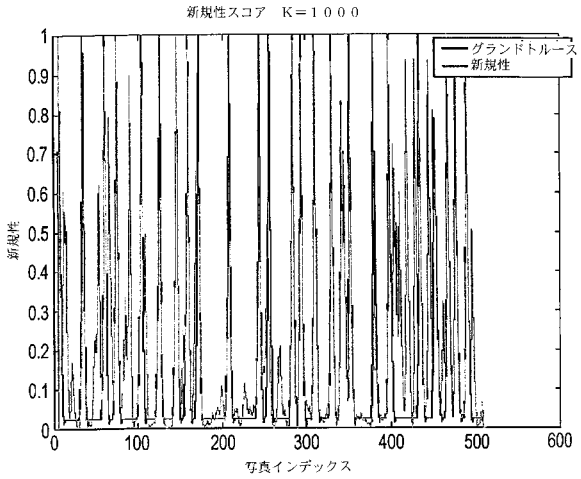
【図３】


グランドトルース類似性行列
写真インデックス
写真インデックス
50
100
150
200
250
300
350
400
450
500
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

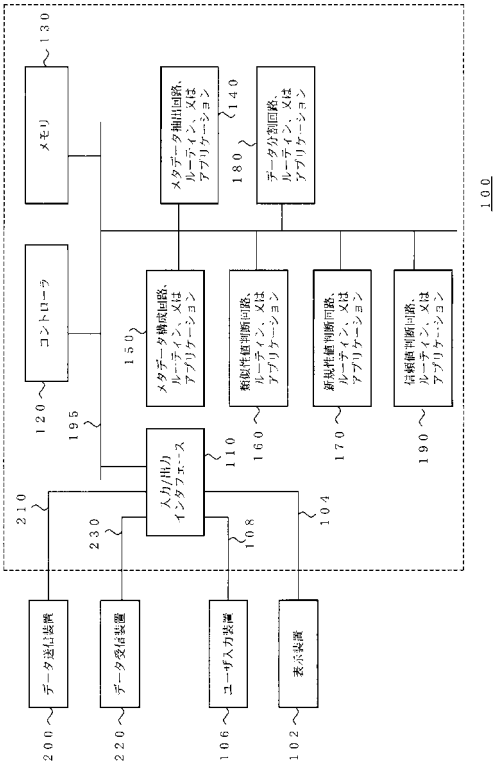
【図４】


グランドトルース新規性
新規性
写真インデックス
0
100
200
300
400
500
600
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

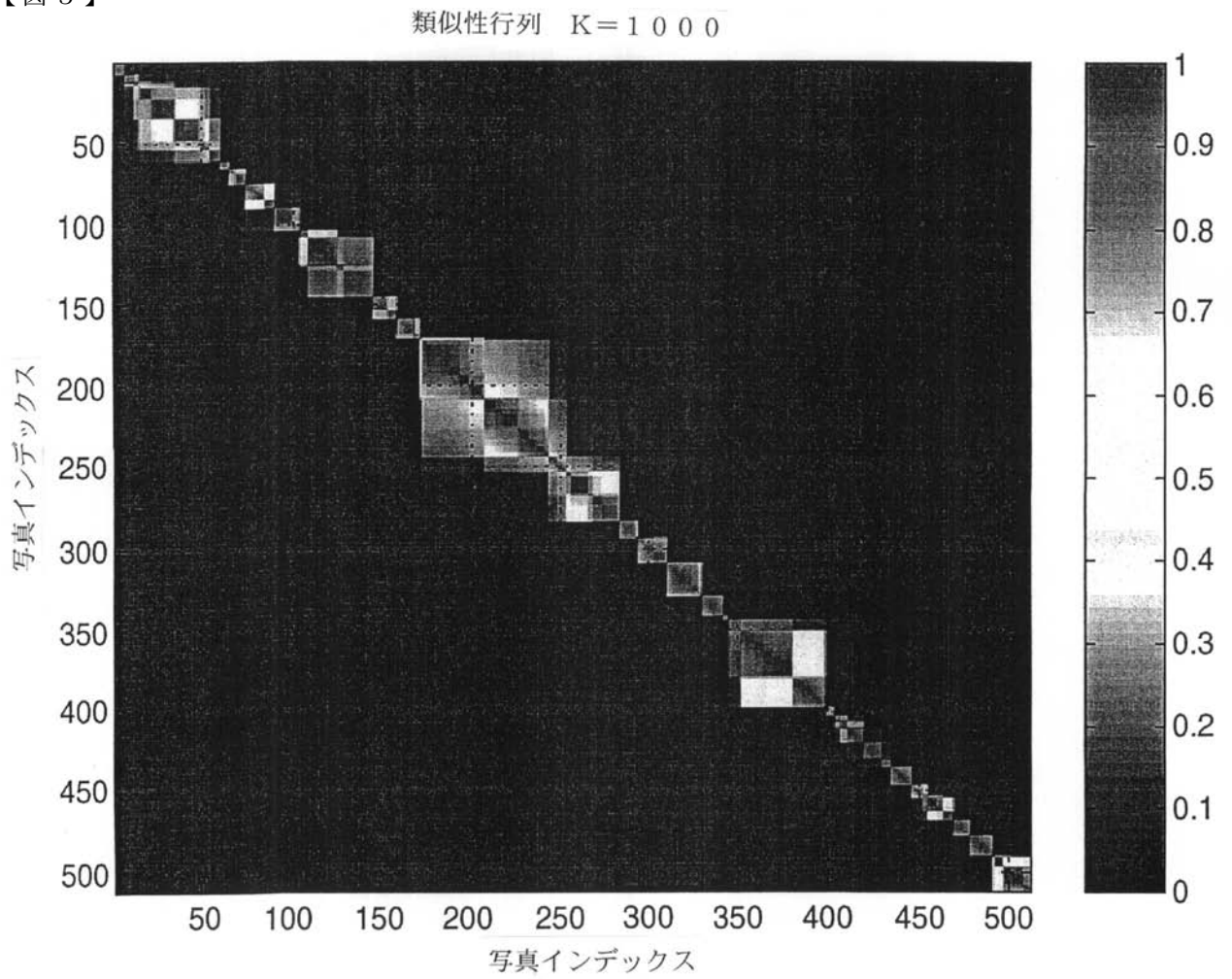
【図６】


新規性スコア　K＝１０００
グランドトルース
新規性
新規性
写真インデックス
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
100
200
300
400
500
600

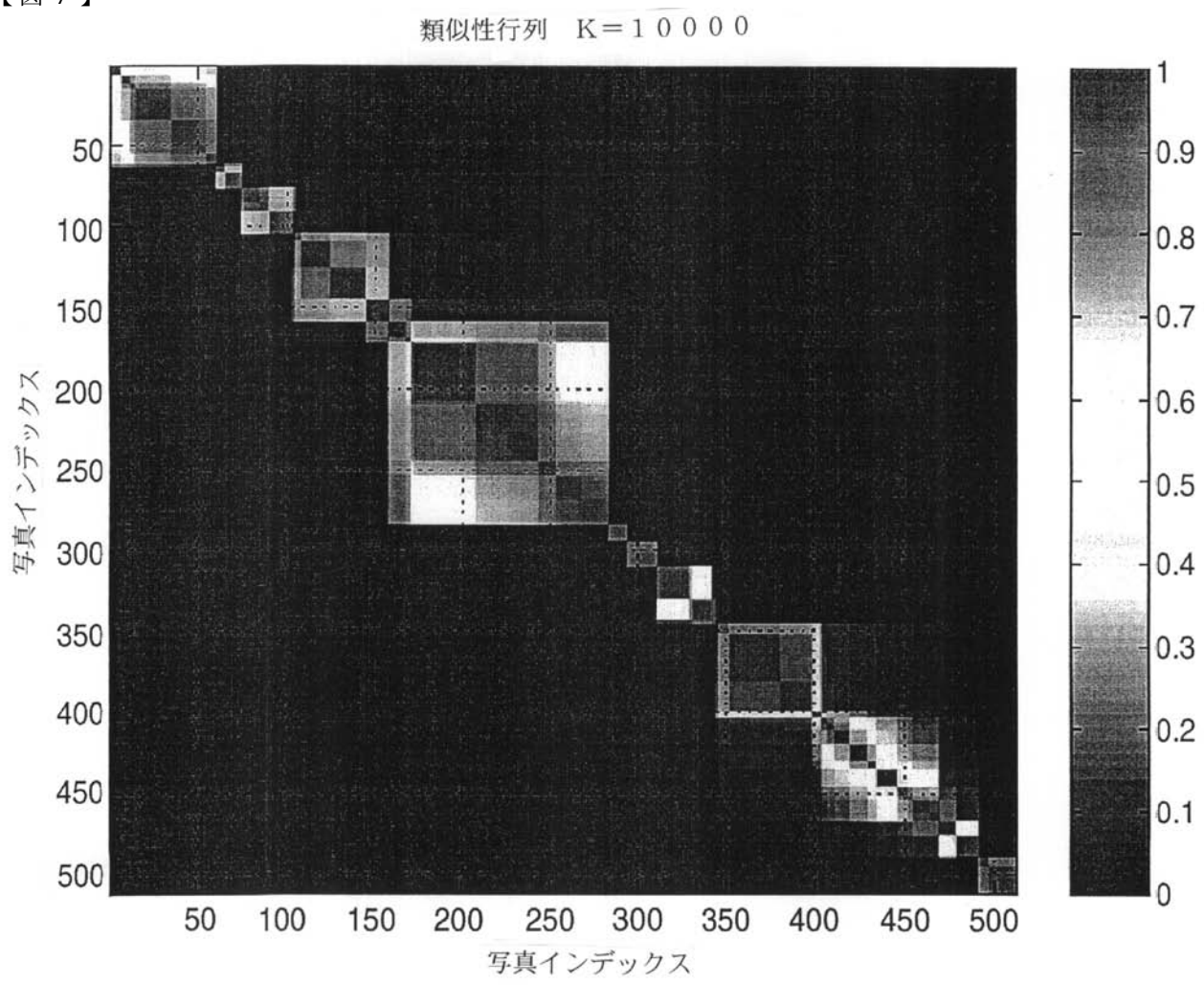
【図１８】


メモリ
130
メタデータ抽出回路、ルーティン、又はアプリケーション
140
データ分割回路、ルーティン、又はアプリケーション
180
コントローラ
120
195
メタデータ構成回路、ルーティン、又はアプリケーション
150
類似性値判断回路、ルーティン、又はアプリケーション
160
新規性値判断回路、ルーティン、又はアプリケーション
170
信頼値判断回路、ルーティン、又はアプリケーション
190
入力/出力インタフェース
110
210
230
108
104
データ送信装置
200
データ受信装置
220
ユーザ入力装置
106
表示装置
102
100

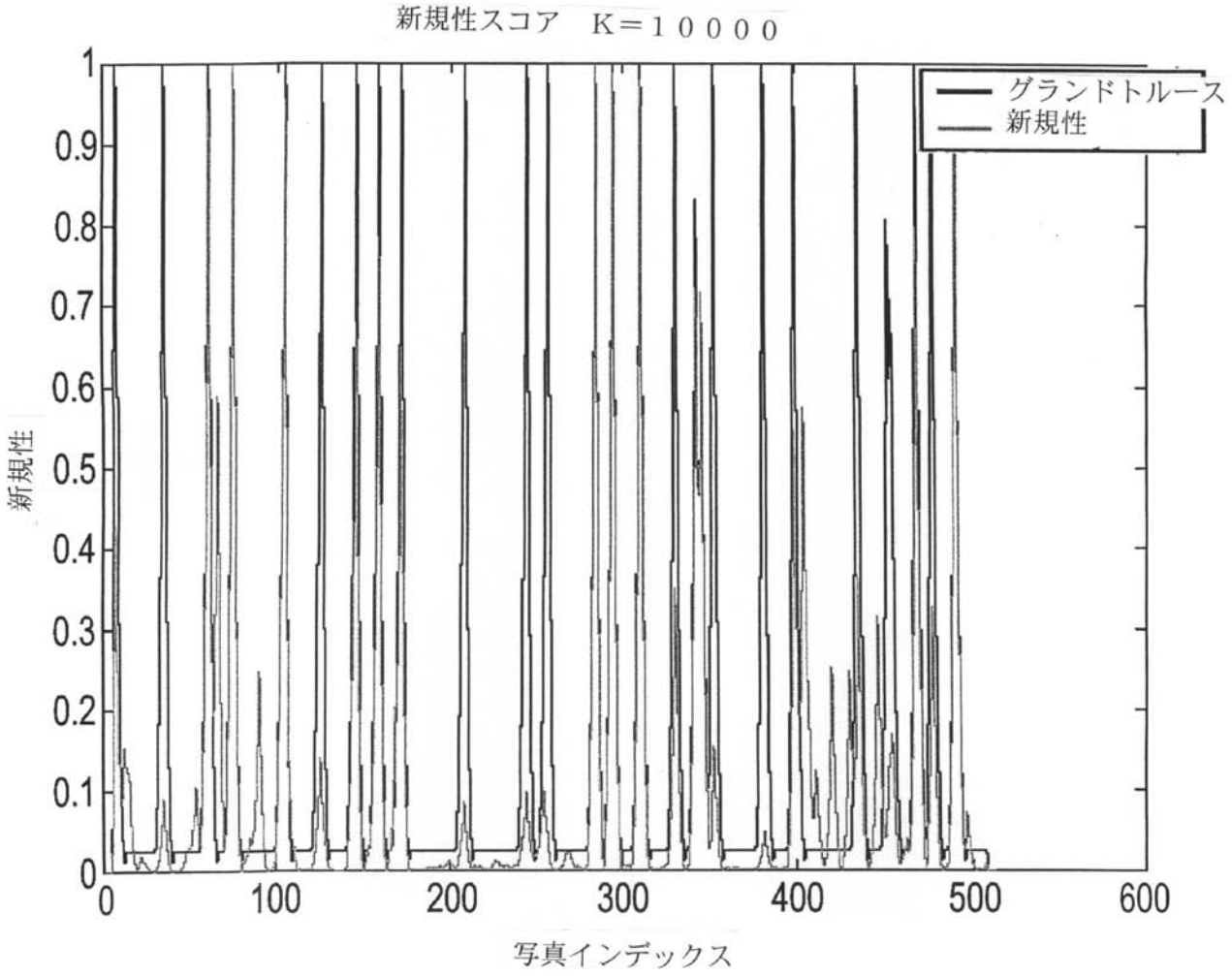
【図５】


類似性行列　Ｋ＝１０００
写真インデックス
50
100
150
200
250
300
350
400
450
500
写真インデックス
50
100
150
200
250
300
350
400
450
500
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

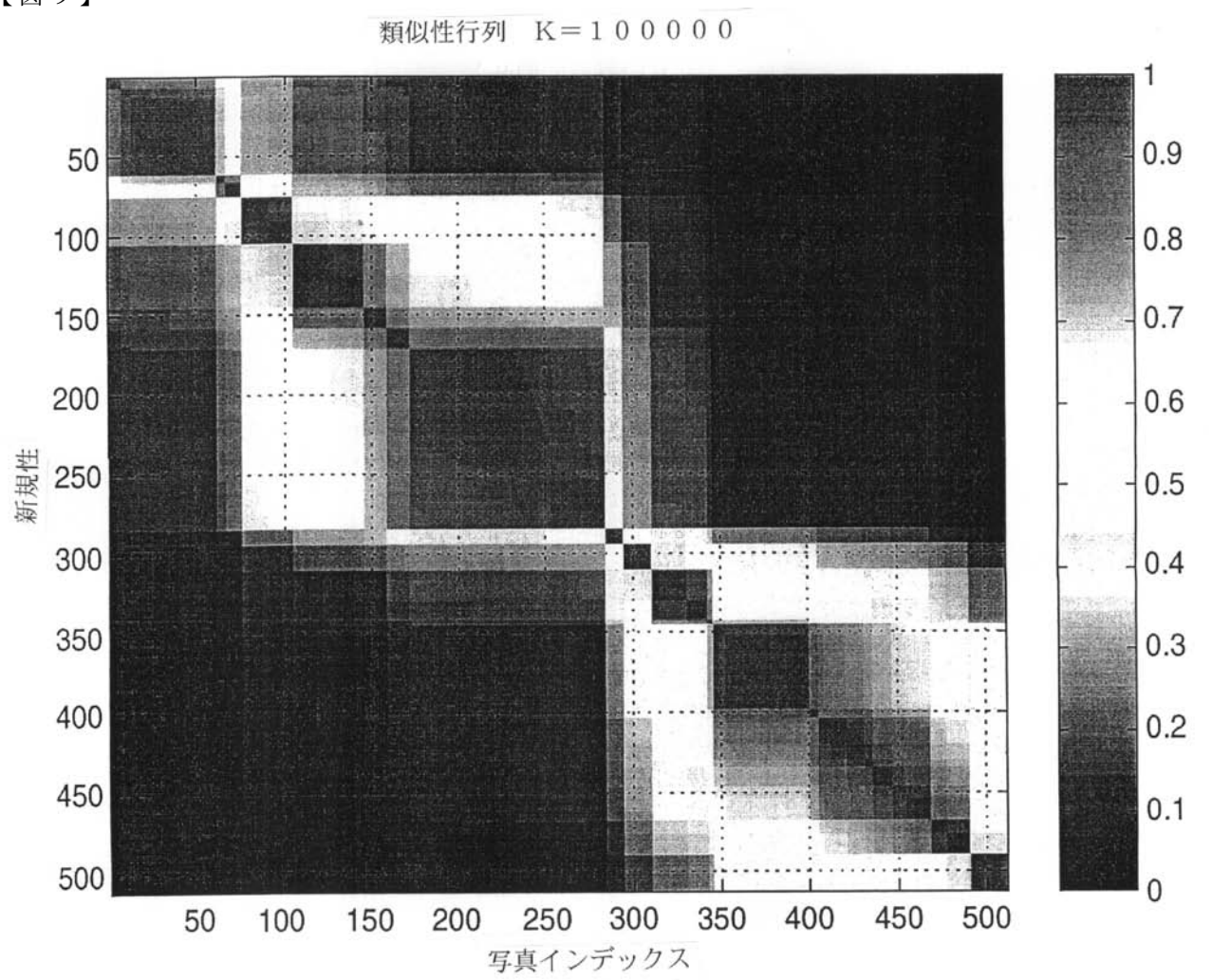
【図７】


類似性行列　Ｋ＝１００００
写真インデックス
50
100
150
200
250
300
350
400
450
500
写真インデックス
50
100
150
200
250
300
350
400
450
500
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

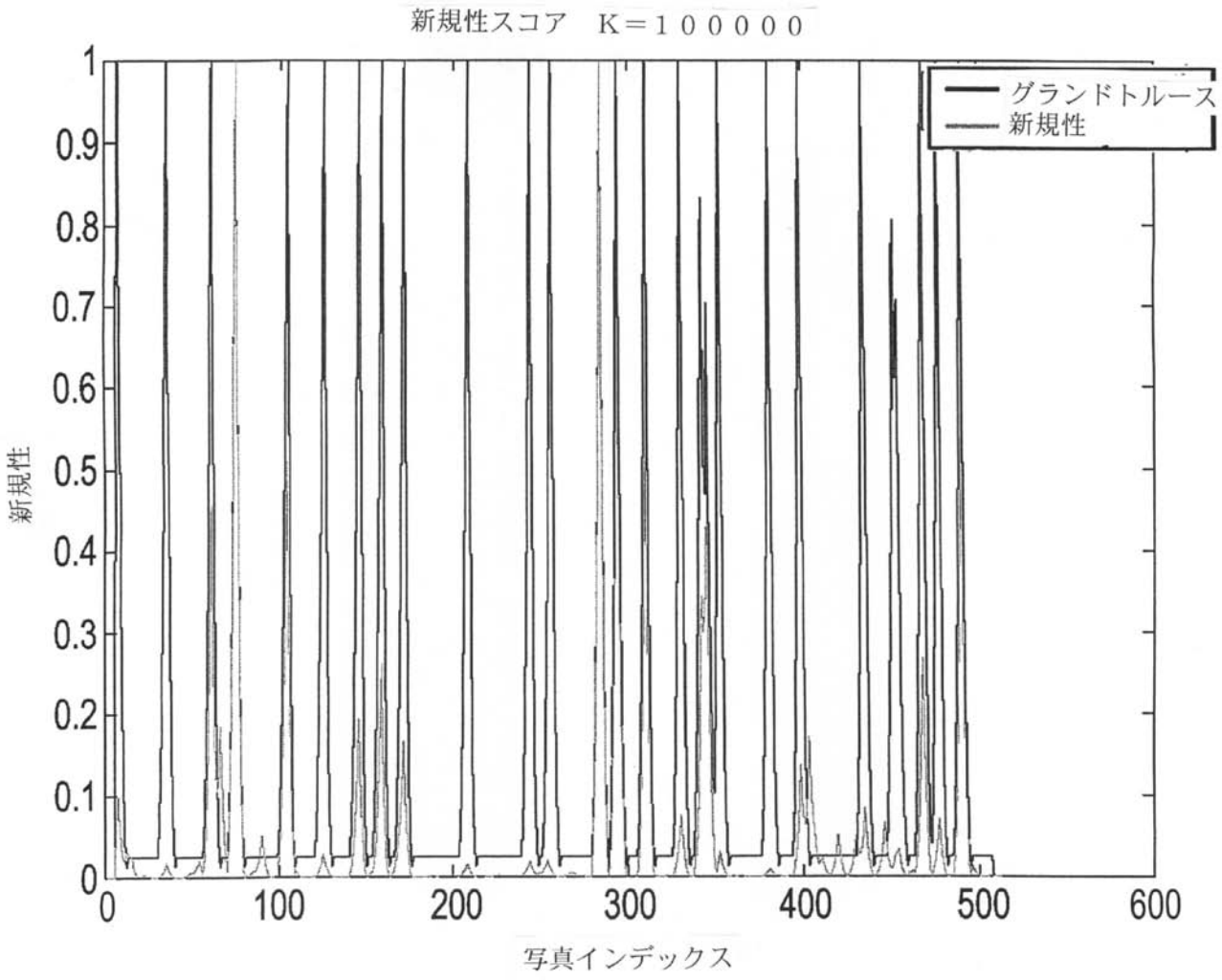
【図８】


新規性スコア　K＝１００００
グランドトルース
新規性
新規性
写真インデックス
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
100
200
300
400
500
600

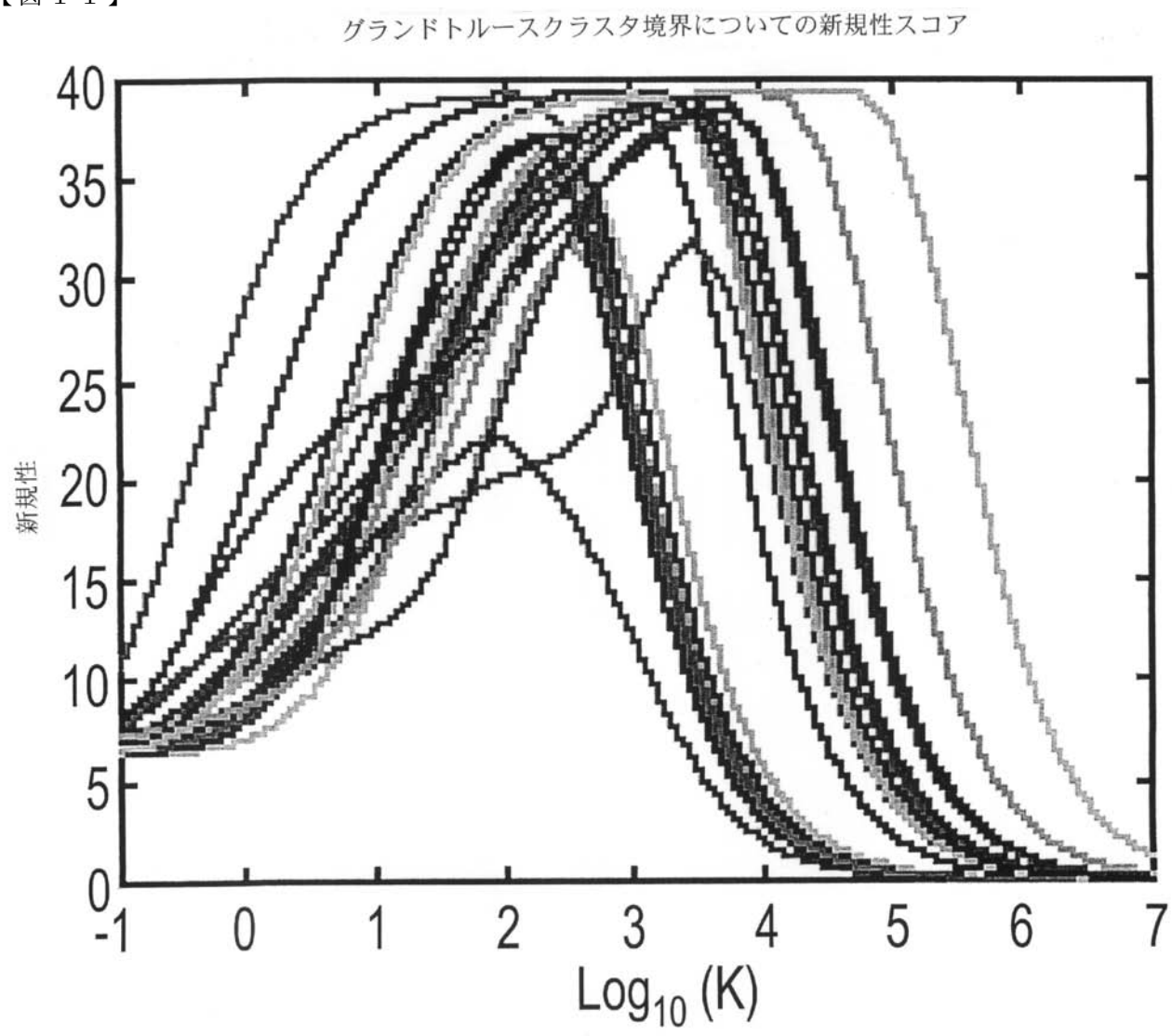
【図９】


類似性行列　Ｋ＝１００００
新規性
写真インデックス
50
100
150
200
250
300
350
400
450
500
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

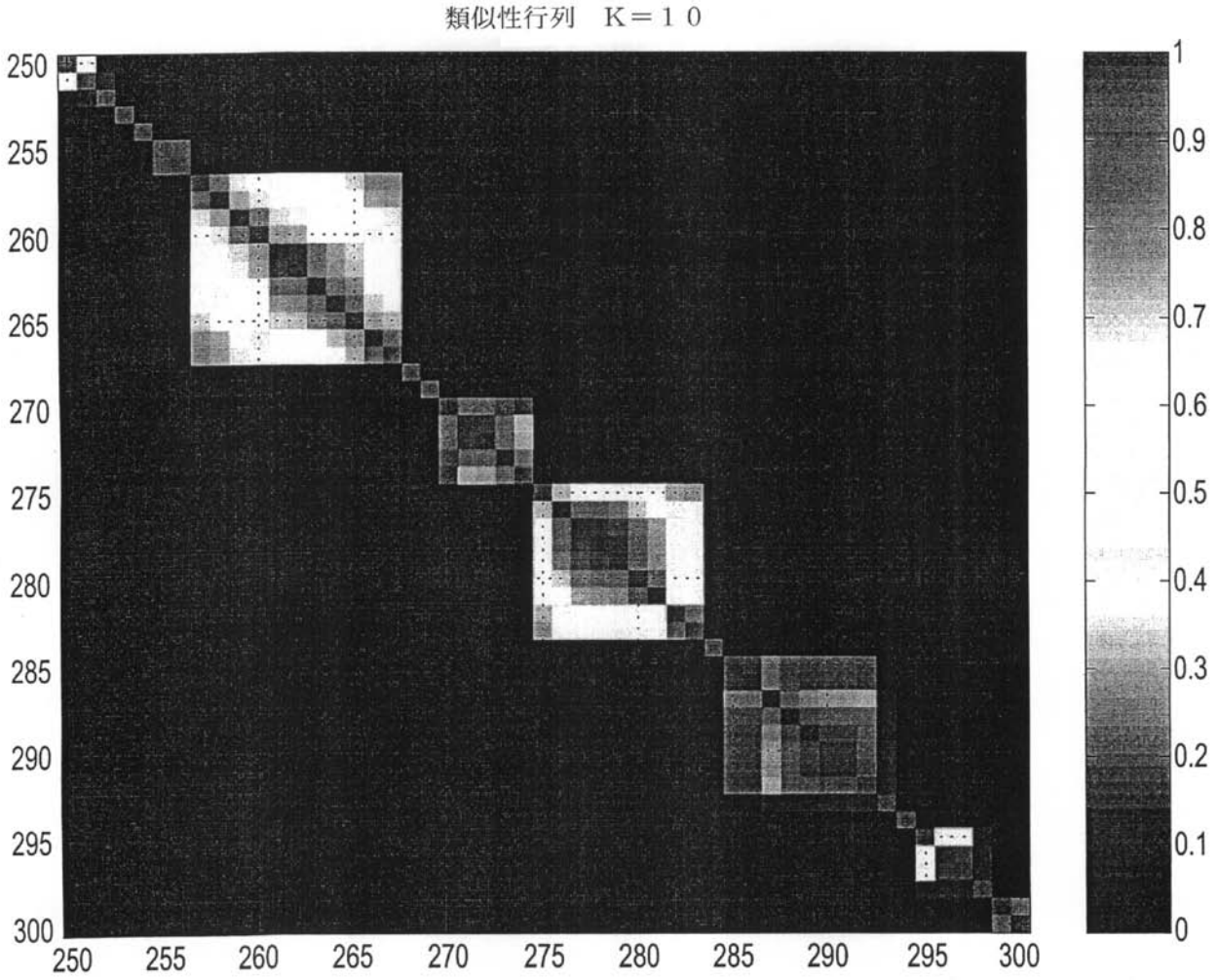
【図１０】


新規性スコア　K＝１００００００
グランドトルース
新規性
新規性
写真インデックス
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
100
200
300
400
500
600

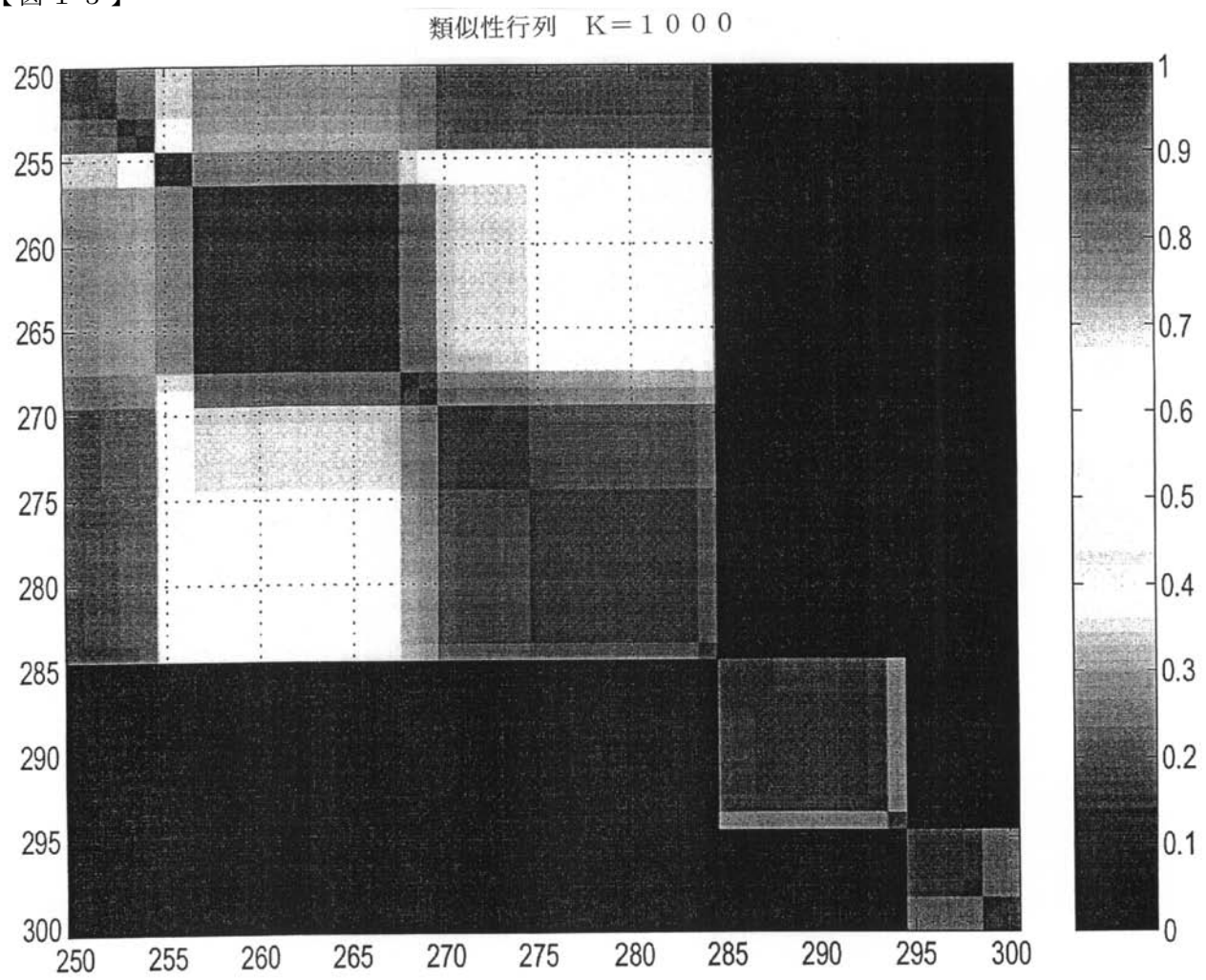
【図１１】


グランドトルースクラスタ境界についての新規性スコア
新規性
40
35
30
25
20
15
10
5
0
-1
0
1
2
3
4
5
6
7
Log10 (K)

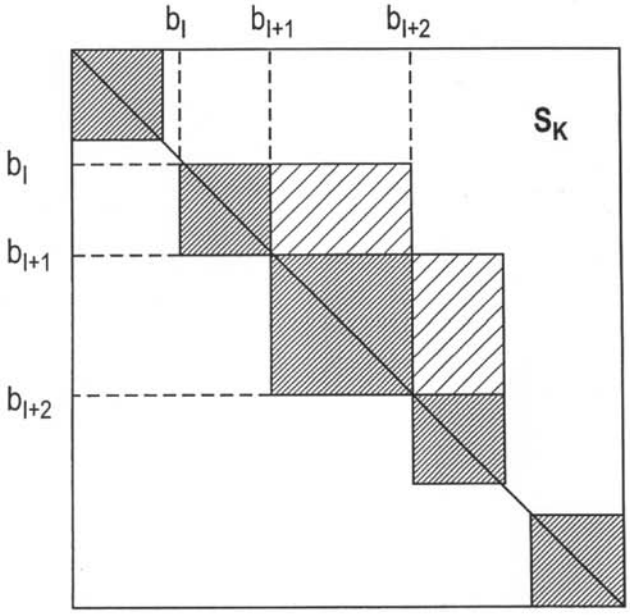
【図１２】


類似性行列　Ｋ＝１０
250
255
260
265
270
275
280
285
290
295
300
250
255
260
265
270
275
280
285
290
295
300
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

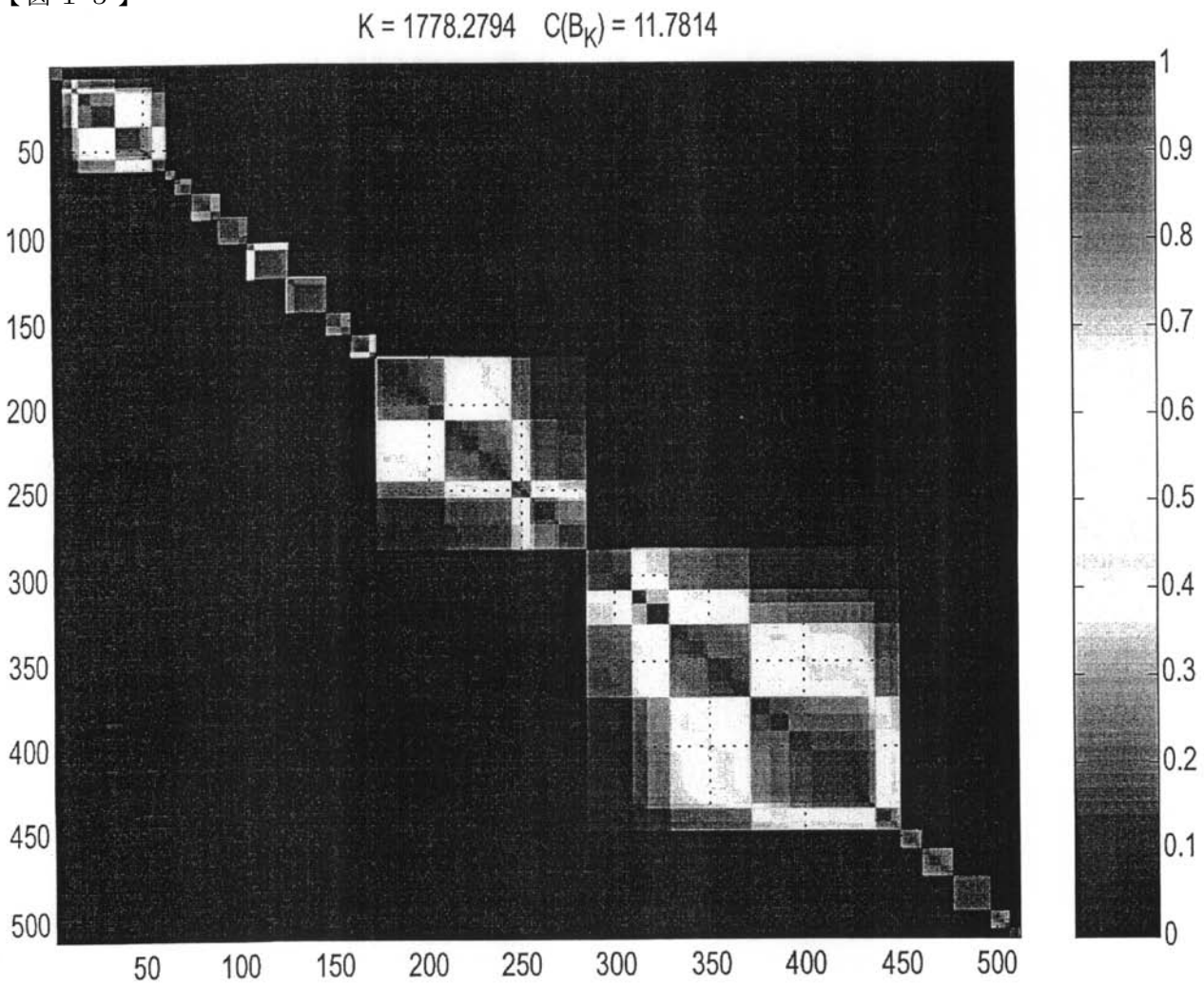
【図１３】


類似性行列　Ｋ＝１０００
250
255
260
265
270
275
280
285
290
295
300
250
255
260
265
270
275
280
285
290
295
300
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

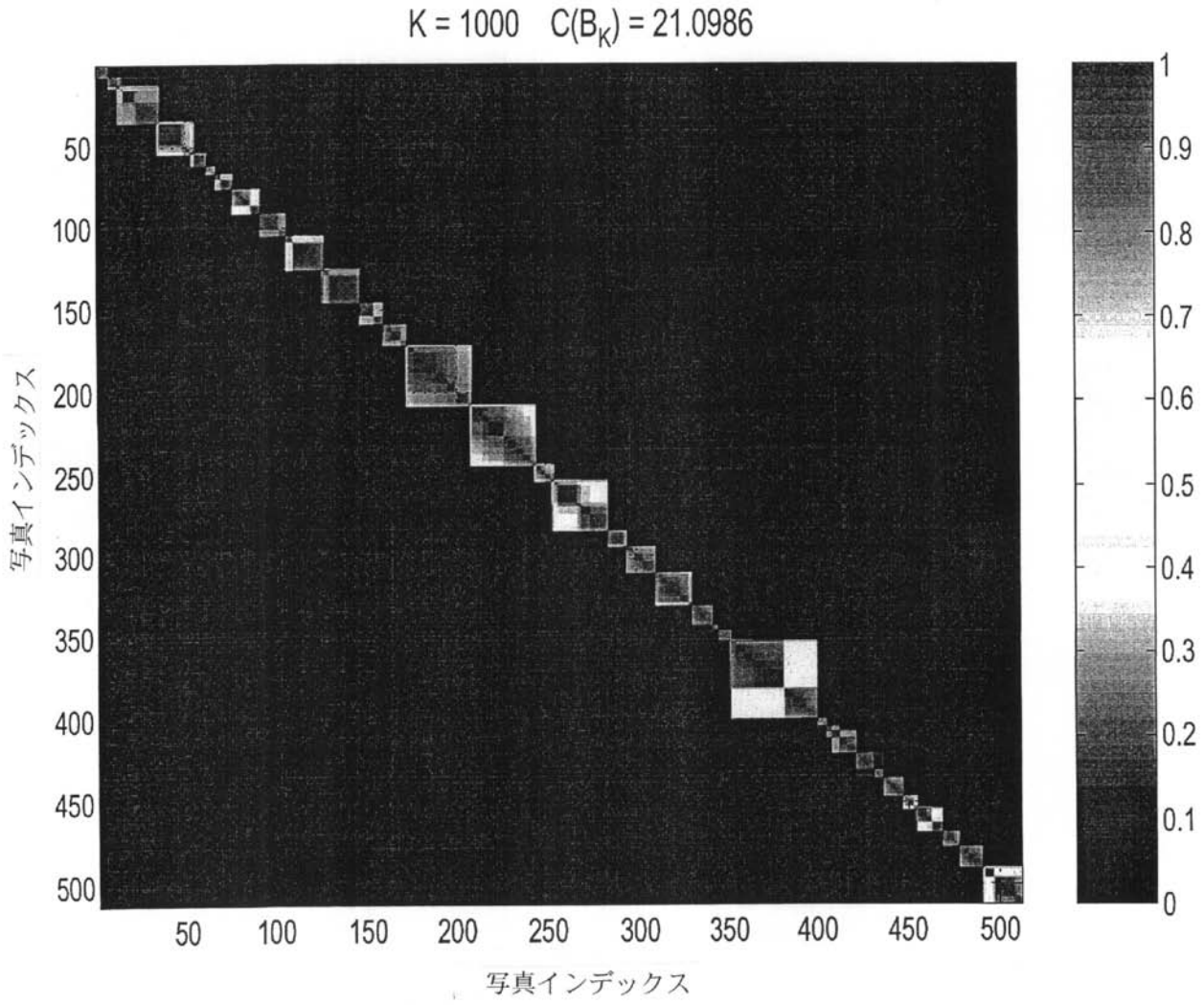
【図１４】


b_l
b_{l+1}
b_{l+2}
S_K
b_l
b_{l+1}
b_{l+2}

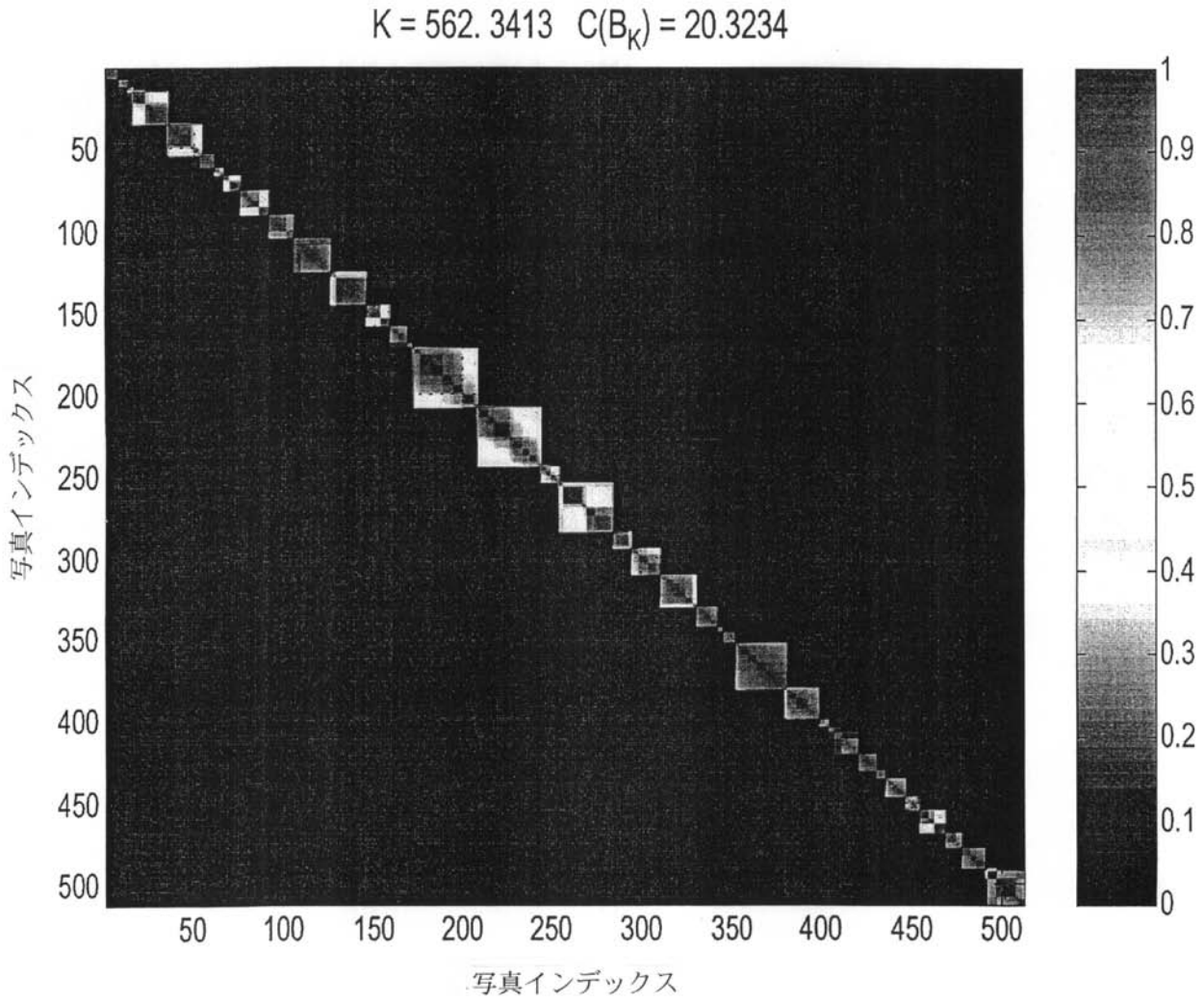
【図１５】


K = 1778.2794 C(B_K) = 11.7814
50
100
150
200
250
300
350
400
450
500
50
100
150
200
250
300
350
400
450
500
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

【図１６】


K = 1000 C(B_K) = 21.0986
写真インデックス
50
100
150
200
250
300
350
400
450
500
写真インデックス
50
100
150
200
250
300
350
400
450
500
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

【図１７】


K = 562. 3413 C(B_K) = 20.3234
写真インデックス
50
100
150
200
250
300
350
400
450
500
写真インデックス
50
100
150
200
250
300
350
400
450
500
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

フロントページの続き

(72)発明者　アンドレアス　ガーゲンソン
アメリカ合衆国　９４０２５　カリフォルニア州　メンロ　パーク　ウェイバリー　ストリート　２１０　ナンバー４

Ｆターム(参考) 5B075 KK35 ND08 NR12 NS10 UU40