(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6800825号
(P6800825)
(45)発行日　令和2年12月16日(2020.12.16)　(24)登録日　令和2年11月27日(2020.11.27)

(51)Int.Cl.　　F I
G06F 16/35 (2019.01)　　G06F 16/35
G06F 16/34 (2019.01)　　G06F 16/34

請求項の数 13　(全 26 頁)

(21)出願番号　特願2017-192750(P2017-192750)
(22)出願日　平成29年10月2日(2017.10.2)
(65)公開番号　特開2019-67191(P2019-67191A)
(43)公開日　平成31年4月25日(2019.4.25)
審査請求日　令和1年8月13日(2019.8.13)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(74)代理人　110002147
特許業務法人酒井国際特許事務所
(72)発明者　布目　光生
東京都港区芝浦一丁目１番１号　株式会社東芝内
(72)発明者　山崎　智弘
東京都港区芝浦一丁目１番１号　株式会社東芝内

審査官　鹿野　博嗣

最終頁に続く

(54)【発明の名称】情報処理装置、情報処理方法およびプログラム

(57)【特許請求の範囲】
【請求項１】
文書群に含まれる複数のキーフレーズを含むキーフレーズ群を階層クラスタリングするクラスタリング部と、
前記キーフレーズ群を複数の候補クラスタに分割する分割部と、
前記文書群を分類するための予め定められた複数の項目のうちの何れか１つ項目の選択操作を受け付ける項目選択部と、
前記複数の候補クラスタのそれぞれについて、選択された項目に対する有用性を表すスコアを算出するスコア算出部と、
前記複数の候補クラスタのうち、前記スコアが所定の順位の候補クラスタを参照クラスタとして決定する決定部と、
前記参照クラスタを複数のサブクラスタに分割するサブクラスタ生成部と、
選択された項目の下位層の予め定められた複数のサブ項目を抽出するサブ項目抽出部と、
前記複数のサブ項目のそれぞれ毎且つ前記複数のサブクラスタのそれぞれ毎の文書の情報量を表すように、展開画像の提示を制御する展開画像制御部と、
を備える情報処理装置。
【請求項２】
前記決定部は、前記複数の候補クラスタのうち前記スコアが最も高い１つの候補クラスタを前記参照クラスタとして決定する

請求項１に記載の情報処理装置。
【請求項３】
前記スコアは、対応する候補クラスタにおける選択された項目に分類される文書の情報量を表す
請求項１または２に記載の情報処理装置。
【請求項４】
前記スコアは、前記複数のサブ項目のそれぞれ毎且つ対応する候補クラスタを分割した複数のサブクラスタのそれぞれ毎の文書の情報量の分散を表す
請求項１または２に記載の情報処理装置。
【請求項５】
前記スコアは、対応する候補クラスタをユーザが過去に操作により選択した頻度または割合を表す
請求項１または２に記載の情報処理装置。
【請求項６】
前記スコア算出部は、第１パラメータと、第２パラメータと、第３パラメータとを合計して前記スコアを算出し、
前記第１パラメータは、対応する候補クラスタにおける選択された項目に分類される文書の情報量に応じた値を表し、
前記第２パラメータは、前記複数のサブ項目のそれぞれ毎且つ対応する候補クラスタを分割した複数のサブクラスタのそれぞれ毎の文書の情報量の分散に応じた値を表し、
前記第３パラメータは、対応する候補クラスタをユーザが過去に操作により選択した頻度または割合に応じた値を表す
請求項１または２に記載の情報処理装置。
【請求項７】
前記複数の項目のそれぞれ毎且つ前記複数の候補クラスタのそれぞれ毎の文書の情報量を表すように、初期画像の提示を制御する初期画像制御部をさらに備える
請求項１から６の何れか１項に記載の情報処理装置。
【請求項８】
前記複数のサブ項目のうち何れか１つのサブ項目が選択された場合、前記スコア算出部、前記決定部、前記サブ項目抽出部および前記展開画像制御部は、前記複数のサブ項目を前記複数の項目と置き換え、前記複数のサブクラスタを前記複数の候補クラスタと置き換えて、再度処理を実行して展開画像を提示する
請求項１から７の何れか１項に記載の情報処理装置。
【請求項９】
前記展開画像制御部は、前記展開画像とともに、前記複数の候補クラスタのうちの何れか１つを選択させるためのメニュー画像を提示し、
前記メニュー画像により何れか１つの新たな候補クラスタが選択された場合、
前記サブクラスタ生成部は、前記新たな候補クラスタを前記参照クラスタとして、新たな複数のサブクラスタに分割し、
前記展開画像制御部は、前記複数のサブ項目のそれぞれ毎且つ新たな複数のサブクラスタのそれぞれ毎の文書の情報量を表すように、新たな展開画像の提示を制御する
請求項１から８の何れか１項に記載の情報処理装置。
【請求項１０】
文書抽出部と、選択クラスタリング部と、選択分割部とをさらに備え、
前記項目選択部は、前記文書群を分類するための予め定められた複数の第１観点項目のうちの何れか１つの第１観点項目、および、前記文書群を分類するための予め定められた複数の第２観点項目のうちの何れか１つの第２観点項目の選択操作を受け付け、
前記文書抽出部は、前記文書群から、選択された第１観点項目および選択された第２観点項目の両者に分類される複数の文書を含む選択文書群を抽出し、
前記選択クラスタリング部は、前記選択文書群に含まれる複数のキーフレーズを含む選

択キーフレーズ群を階層クラスタリングし、
前記選択分割部は、前記選択キーフレーズ群を複数の候補クラスタに分割し、
前記スコア算出部は、前記複数の候補クラスタのそれぞれについて、選択された第１観点項目および第２観点項目に対する有用性を表す前記スコアを算出し、
前記決定部は、前記複数の候補クラスタのうち、前記スコアが所定の順位の２つの候補クラスタを第１参照クラスタおよび第２参照クラスタとして決定し、
前記サブクラスタ生成部は、前記第１参照クラスタを複数の第１サブクラスタに分割し、前記第２参照クラスタを複数の第２サブクラスタに分割し、
前記展開画像制御部は、前記複数の第１サブクラスタのそれぞれ毎且つ前記複数の第２サブクラスタのそれぞれ毎の文書の情報量を表すように、クラスタ展開画像の提示を制御する
請求項１に記載の情報処理装置。
【請求項１１】
前記複数の第１観点項目のそれぞれ毎且つ前記複数の第２観点項目のそれぞれ毎の文書の情報量を表すように、項目選択画像の提示を制御する初期画像制御部をさらに備える
請求項１０に記載の情報処理装置。
【請求項１２】
情報処理装置により実行される情報処理方法であって、
クラスタリング部が、文書群に含まれる複数のキーフレーズを含むキーフレーズ群を階層クラスタリングし、
分割部が、前記キーフレーズ群を複数の候補クラスタに分割し、
項目選択部が、前記文書群を分類するための予め定められた複数の項目のうちの何れか１つ項目の選択操作を受け付け、
スコア算出部が、前記複数の候補クラスタのそれぞれについて、選択された項目に対する有用性を表すスコアを算出し、
決定部が、前記複数の候補クラスタのうち、前記スコアが所定の順位の候補クラスタを参照クラスタとして決定し、
サブクラスタ生成部が、前記参照クラスタを複数のサブクラスタに分割し、
抽出部が、選択された項目の下位層の予め定められた複数のサブ項目を抽出し、
展開画像制御部が、前記複数のサブ項目のそれぞれ毎且つ前記複数のサブクラスタのそれぞれ毎の文書の情報量を表すように、展開画像の提示を制御する
情報処理方法。
【請求項１３】
情報処理装置で実行されるプログラムであって、
前記情報処理装置を、
文書群に含まれる複数のキーフレーズを含むキーフレーズ群を階層クラスタリングするクラスタリング部と、
前記キーフレーズ群を複数の候補クラスタに分割する分割部と、
前記文書群を分類するための予め定められた複数の項目のうちの何れか１つ項目の選択操作を受け付ける項目選択部と、
前記複数の候補クラスタのそれぞれについて、選択された項目に対する有用性を表すスコアを算出するスコア算出部と、
前記複数の候補クラスタのうち、前記スコアが所定の順位の候補クラスタを参照クラスタとして決定する決定部と、
前記参照クラスタを複数のサブクラスタに分割するサブクラスタ生成部と、
選択された項目の下位層の予め定められた複数のサブ項目を抽出するサブ項目抽出部と、
前記複数のサブ項目のそれぞれ毎且つ前記複数のサブクラスタのそれぞれ毎の文書の情報量を表すように、展開画像の提示を制御する展開画像制御部と
して機能させるプログラム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明の実施形態は、情報処理装置、情報処理方法およびプログラムに関する。
【背景技術】
【０００２】
業務において用いた文書をデータベースに記録し、その文書を他の業務において再利用させる文書管理システムが知られている。また、文書管理システムにより管理している文書群から、データマイニングおよびテキストマイニング等により知識を抽出し、抽出した知識を業務の分析および改善に役立てることも行われている。
【０００３】
また、文書管理システムにより管理している文書群から目的の文書を検索する手段として、キーフレーズ（単語および単語列等）検索およびファセット検索が知られている。ファセット検索は、文書を分類するための複数の項目および階層構造を予め定義しておき、ユーザに上位層から下位層に向かい順次に項目を選択させて、文書を絞り込む方法である。
【０００４】
文書管理システムにより管理している文書群の特徴をユーザに参照させる手段も、様々な方法が提案されている。例えば、文書群の特徴をユーザに参照させる手段として、ＯＬＡＰ（Ｏｎｌｉｎｅ　Ａｎａｌｙｔｉｃａｌ　Ｐｒｏｃｅｓｓｉｎｇ）機能が知られている。ＯＬＡＰ機能は、文書群の全体の特徴を俯瞰的に参照させたり、全体を表す情報から詳細を表す情報にドリルダウンさせながら文書群の特徴を参照させたりする方法である。また、文書群の特徴をユーザに参照させる手段として、ヒートマップも知られている。ヒートマップは、異なる２つの観点から分類した情報のそれぞれの特徴を２つの軸を有するマップ上に表す方法である。
【０００５】
ところで、ファセット検索では、予め項目の構造を定義しておかなければならない。例えば、しかし、項目の構造の設計、および、対応するデータベースの設計には、非常に多くのコストがかかる。また、文書管理システムの運用が進んだ段階で、新しい観点で文書を検索および文書群の特徴を参照したいというニーズが生じても、既に定義した項目の階層構造およびデータベースの構造を変えることは難しかった。
【０００６】
一方で、クラスタリングにより分類の項目を自動生成するといった手法も知られている。この方法であれば、事前に、項目の構造を設計する必要はない。しかし、クラスタリングにより分類の項目を自動生成する方法では、実際に利用できる項目には大きな限定がかかる。例えば、クラスタリングにより分類の項目を自動生成する方法では、数量表現、色および形等の離散属性、および、ソースコードのパッケージ名等の、階層情報および構造が文書内に記述されている項目にしか分類ができなかった。
【先行技術文献】
【特許文献】
【０００７】
【特許文献１】特開２０１７－０６８５３４号公報
【発明の概要】
【発明が解決しようとする課題】
【０００８】
発明が解決しようとする課題は、文書群を小さいコストで適切に分類した情報を提示することにある。
【課題を解決するための手段】
【０００９】
実施形態に係る情報処理装置は、クラスタリング部と、分割部と、項目選択部と、スコ

ア算出部と、決定部と、サブクラスタ生成部と、サブ項目抽出部と、展開画像制御部と、を備える。前記クラスタリング部は、文書群に含まれる複数のキーフレーズを含むキーフレーズ群を階層クラスタリングする。前記分割部は、前記キーフレーズ群を複数の候補クラスタに分割する。前記項目選択部は、前記文書群を分類するための予め定められた複数の項目のうちの何れか１つ項目の選択操作を受け付ける。前記スコア算出部は、前記複数の候補クラスタのそれぞれについて、選択された項目に対する有用性を表すスコアを算出する。前記決定部は、前記複数の候補クラスタのうち、前記スコアが所定の順位の候補クラスタを参照クラスタとして決定する。前記サブクラスタ生成部は、前記参照クラスタを複数のサブクラスタに分割する。前記サブ項目抽出部は、選択された項目の下位層の予め定められた複数のサブ項目を抽出する。前記展開画像制御部は、前記複数のサブ項目のそれぞれ毎且つ前記複数のサブクラスタのそれぞれ毎の文書の情報量を表すように、展開画像の提示を制御する。

【図面の簡単な説明】

【００１０】

【図１】第１実施形態に係る文書管理システムの構成図。

【図２】第１実施形態に係る事前処理部の構成図。

【図３】第１実施形態に係る第１画像制御部および第２画像制御部の構成図。

【図４】情報処理装置の処理の流れを示すフローチャート。

【図５】事前処理の詳細な処理の流れを示すフローチャート。

【図６】キーフレーズ群の構造を表すデンドログラムの一例を示す図。

【図７】デンドログラムおよび候補クラスタの一例を示す図。

【図８】複数の項目の階層構造を示す図。

【図９】初期画像の表示処理の詳細な処理の流れを示すフローチャート。

【図１０】初期画像の一例を示す図。

【図１１】展開画像の表示処理の詳細な処理の流れを示すフローチャート。

【図１２】初期画像および展開画像の一例を示す図。

【図１３】初期画像、展開画像および新たな展開画像の一例を示す図。

【図１４】メニュー画像が追加された展開画像を示す図。

【図１５】複数の候補クラスタのそれぞれのスコアを示す図。

【図１６】第２実施形態に係る第１画像制御部および第２画像制御部の構成図。

【図１７】複数の第１観点項目および複数の第２観点項目の一例を示す図。

【図１８】項目選択画像の一例を示す図。

【図１９】選択キーフレーズ群の構造を表すデンドログラムの一例を示す図。

【図２０】デンドログラムおよびクラスタ展開画像の一例を示す図。

【図２１】列選択および行選択をした場合の項目選択画像を示す図。

【発明を実施するための形態】

【００１１】

以下、図面を参照しながら本実施形態に係る文書管理システム１０について説明する。なお、以下の実施形態では、同一の参照符号を付した部分は略同一の構成および動作をするので、相違点を除き重複する説明を適宜省略する。

【００１２】

（第１実施形態）

図１は、第１実施形態に係る文書管理システム１０の構成を示す図である。文書管理システム１０は、業務等で作成された複数の文書を含む文書群を管理する。また、文書管理システム１０は、ユーザの操作に応じて、文書群を分類し、分類された複数の文書毎の情報量を表示する。

【００１３】

文書は、コンピュータにより情報内容を検索することが可能であれば、どのようなデータであってもよい。例えば、文書は、テキストを含むデータであってもよいし、プログラムコードを含むデータであってもよい。文書のファイル形式は、文書管理システム１０に

より取り扱いが可能であれば、どのようなものであってもよい。

【００１４】

また、複数の文書の情報量は、複数の文書の数であってもよいし、複数の文書に含まれる文字の数であってもよいし、複数の文書の合計のデータ量であってもよい。

【００１５】

また、文書管理システム１０は、１つの文書から１または複数のキーフレーズを抽出する。キーフレーズは、その文書に含まれる情報およびその文書に関連する情報等の、その文書の特徴を表す情報である。キーフレーズは、例えば、１つの単語であってもよいし、複数の単語が並んだ単語列であってもよい。また、キーフレーズは、プログラムコード中のコード列であってもよい。

【００１６】

文書管理システム１０は、表示装置１２と、入力装置１４と、記憶装置１６と、情報処理装置２０とを備える。

【００１７】

表示装置１２は、画像を表示することにより、画像をユーザに提示する。表示装置１２は、情報処理装置２０により生成された画像を受け取り、受け取った画像を表示する。表示装置１２は、例えば、液晶表示器等の表示デバイスである。

【００１８】

入力装置１４は、ユーザからの指示および操作を受け付ける。入力装置１４は、例えば、マウスまたはトラックボール等のポインティングデバイス、あるいはキーボード等の入力デバイスである。

【００１９】

記憶装置１６は、情報処理装置２０からデータを受け取り、受け取ったデータを記憶する。また、記憶装置１６は、記憶しているデータが情報処理装置２０により読み出される。記憶装置１６は、フラッシュメモリ等の半導体メモリ素子、ハードディスク、光ディスク等である。記憶装置１６は、ネットワークを介して情報処理装置２０と接続可能なサーバ装置であってもよい。

【００２０】

情報処理装置２０は、例えば、専用または汎用コンピュータである。情報処理装置２０は、ＰＣ、あるいは、情報を保存および管理するサーバに含まれるコンピュータであってもよい。情報処理装置２０は、一台の装置により実現されてもよいし、連携して動作する複数台の装置により実現されてもよい。また、情報処理装置２０は、ネットワーク上に実現される仮想的な装置（例えばクラウド）等であってもよい。

【００２１】

情報処理装置２０は、表示装置１２を制御して、表示装置１２に画像を表示させる。また、情報処理装置２０は、入力装置１４から情報を受け取って、ユーザから与えられた指示内容および操作内容を識別する。また、情報処理装置２０は、データを記憶装置１６に書き込み、記憶装置１６に記憶されたデータを読み出す。

【００２２】

情報処理装置２０は、通信部２２と、記憶回路２４と、処理回路３０とを有する。表示装置１２、入力装置１４、記憶装置１６、通信部２２、記憶回路２４および処理回路３０は、バスを介して接続される。

【００２３】

通信部２２は、有線または無線で接続された外部装置と情報の入出力を行うインターフェースである。通信部２２は、ネットワークに接続して通信を行ってもよい。

【００２４】

記憶回路２４は、ＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）およびＲＯＭ（Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）である。記憶回路２４は、起動用プログラムを読み出すスタートプログラムが記憶されている。また、記憶回路２４は、処理回路３０の作業領域として機能する。

【００２５】
処理回路３０は、１または複数のプロセッサを含む。処理回路３０は、情報処理を実行し、プログラムを読み出して記憶回路２４に展開して実行し、各部を制御してデータの入出力を行ったり、データの加工を行ったりする。プロセッサは、例えば、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）である。プロセッサは、ＣＰＵに限らず、プログラムを実行する他の種類のデータ処理デバイスまたは専用の処理デバイスであってもよい。
【００２６】
このようなハードウェア構成の文書管理システム１０は、記憶装置１６が、文書記憶部４２、クラスタ記憶部４４および項目記憶部４６として機能する。また、このようなハードウェア構成の文書管理システム１０は、処理回路３０が、プログラムを実行することにより、事前処理部３２、第１画像制御部３４および第２画像制御部３６として機能する。
【００２７】
図２は、第１実施形態に係る事前処理部３２の構成を文書記憶部４２、クラスタ記憶部４４および項目記憶部４６とともに示す図である。
【００２８】
事前処理部３２は、文書取得部５２と、キーフレーズ生成部５４と、文書登録部５６と、クラスタリング部５８と、分割部６０と、項目取得部６２と、項目登録部６４とを有する。
【００２９】
文書取得部５２は、他の装置から文書を取得する。キーフレーズ生成部５４は、文書取得部５２により取得された文書に対して形態素解析および複合語抽出処理等を行って、取得した文書に対する１または複数のキーフレーズを生成する。文書登録部５６は、文書取得部５２により取得された文書と、キーフレーズ生成部５４により生成された１または複数のキーフレーズとを対応付けて、文書記憶部４２に記憶させる。
【００３０】
文書取得部５２、キーフレーズ生成部５４および文書登録部５６は、複数の文書のそれぞれ毎に、これらの処理を実行する。これにより、文書記憶部４２は、複数の文書を含む文書群を記憶することができる。この文書群は、データベース化されており、任意のキーフレーズを指定することにより、指定されたキーフレーズに対応付けられた１または複数の文書を抽出することができる。
【００３１】
クラスタリング部５８は、文書記憶部４２から、文書群に含まれる複数のキーフレーズを含むキーフレーズ群を取得する。クラスタリング部５８は、取得したキーフレーズ群を階層クラスタリングする。例えば、クラスタリング部５８は、キーフレーズ群に含まれる複数のキーフレーズを複数のクラスタにクラスタリングする。さらに、クラスタリング部５８は、それぞれのキーフレーズをベクトル化する。そして、クラスタリング部５８は、クラスタ中心とのベクトル距離に応じて、キーフレーズの類似度を算出する。さらに、クラスタリング部５８は、それぞれのクラスタ内で同様の処理を繰り返して、階層化した複数のクラスタを生成する。
【００３２】
クラスタリング部５８は、階層化した複数のクラスタのそれぞれにラベルを付与してもよい。例えば、クラスタリング部５８は、クラスタ中心に近いキーフレーズをラベルとしてもよい。クラスタリング部５８は、階層クラスタリングしたキーフレーズ群をクラスタ記憶部４４に記憶させる。
【００３３】
分割部６０は、階層クラスタリングされたキーフレーズ群をクラスタ記憶部４４から読み出し、読み出したキーフレーズ群を複数の候補クラスタに分割する。例えば、分割部６０は、階層クラスタリングされたキーフレーズ群を表すデンドログラムを描き、描いたデンドログラムにおける、所定個（例えば、４個以上で最小）のクラスタに分割される高さ

を決定する。そして、分割部６０は、決定した高さでデンドログラムを切断した場合に生成される複数の階層クラスタを、複数の候補クラスタとする。分割部６０は、生成した複数の候補クラスタをクラスタ記憶部４４に記憶させる。

【００３４】

項目取得部６２は、文書群を分類するための予め定められた複数の項目を他の装置から取得する。項目取得部６２は、ユーザにより入力された複数の項目を取得してもよい。複数の項目は、木構造により階層化されている。項目登録部６４は、項目取得部６２により取得された複数の項目を項目記憶部４６に記憶させる。

【００３５】

図３は、第１実施形態に係る第１画像制御部３４および第２画像制御部３６の構成を文書記憶部４２、クラスタ記憶部４４および項目記憶部４６とともに示す図である。

【００３６】

第１画像制御部３４は、開始受付部７２と、第１算出部７４と、初期画像制御部７６とを有する。開始受付部７２は、ユーザによる開始操作を、入力装置１４から受け付ける。

【００３７】

第１算出部７４は、開始受付部７２が開始操作を受け付けると、複数の候補クラスタをクラスタ記憶部４４から取得する。また、第１算出部７４は、開始受付部７２が開始操作を受け付けると、項目記憶部４６から予め定められた複数の項目のうち最上位の複数の項目を取得する。

【００３８】

そして、第１算出部７４は、文書記憶部４２にアクセスして、予め定められた複数の項目のそれぞれ毎、且つ、複数の候補クラスタのそれぞれ毎の文書の情報量を算出する。すなわち、第１算出部７４は、文書群を最上位の複数の項目に従って複数の初期文書群に分類する。そして、第１算出部７４は、複数の初期文書群のそれぞれについて、複数の候補クラスタのそれぞれに分類される文書の情報量を算出する。例えば、最上位の複数の項目が４個、および、候補クラスタが５個の場合、第１算出部７４は、４×５＝２０個の文書の情報量を算出する。

【００３９】

初期画像制御部７６は、予め定められた複数の項目のそれぞれ毎、且つ、複数の候補クラスタのそれぞれ毎の文書の情報量を表すように、初期画像の提示を制御する。すなわち、初期画像制御部７６は、文書群を最上位の複数の項目に従って分類した複数の初期文書群のそれぞれについて、複数の候補クラスタのそれぞれに分類される文書の情報量を表す初期画像を生成する。そして、初期画像制御部７６は、生成した初期画像を表示装置１２に出力して、表示装置１２に初期画像を表示させる。

【００４０】

第２画像制御部３６は、項目選択部７８と、スコア算出部８０と、決定部８２と、サブクラスタ生成部８４と、サブ項目抽出部８６と、第２算出部８８と、展開画像制御部９０とを有する。

【００４１】

項目選択部７８は、ユーザによる、文書群を分類するための予め定められた複数の項目のうちの何れか１つ項目の選択操作を、入力装置１４から受け付ける。例えば、初期画像が表示された後、項目選択部７８は、初期画像に情報量が表示された最上位の複数の項目のうちの、何れか１つの項目の選択操作を受け付ける。

【００４２】

スコア算出部８０は、複数の候補クラスタのそれぞれについて、選択された項目に対する有用性を表すスコアを算出する。なお、スコアについては、詳細を後述する。

【００４３】

決定部８２は、複数の候補クラスタのうち、算出されたスコアが所定の順位の候補クラスタを、参照クラスタとして決定する。例えば、決定部８２は、複数の候補クラスタのうち、有用性が最も高いスコアの候補クラスタを、参照クラスタとして決定する。

【００４４】
サブクラスタ生成部８４は、参照クラスタを、複数のサブクラスタに分割する。例えば、サブクラスタ生成部８４は、参照クラスタを所定個（例えば、４個以上で最小）に分割して、複数のサブクラスタを生成する。
【００４５】
サブ項目抽出部８６は、項目記憶部４６にアクセスして、項目選択部７８により選択された項目の下位層の予め定められた複数のサブ項目を抽出する。
【００４６】
第２算出部８８は、サブクラスタ生成部８４から、複数のサブクラスタを取得する。また、第２算出部８８は、サブ項目抽出部８６から予め定められた複数のサブ項目を取得する。
【００４７】
そして、第２算出部８８は、文書記憶部４２にアクセスして、予め定められた複数のサブ項目のそれぞれ毎、且つ、複数のサブクラスタのそれぞれ毎の文書の情報量を算出する。すなわち、第２算出部８８は、選択された１つの最上位項目に分類される初期文書群を、さらに複数のサブ項目に従って複数のサブ文書群に分類する。そして、第２算出部８８は、複数のサブ文書群のそれぞれについて、複数のサブクラスタのそれぞれに分類される文書の情報量を算出する。例えば、複数のサブ項目が５個、および、複数のサブクラスタが６個の場合、第２算出部８８は、５×６＝３０個の文書の情報量を算出する。
【００４８】
展開画像制御部９０は、予め定められた複数のサブ項目のそれぞれ毎、且つ、複数のサブクラスタのそれぞれ毎の文書の情報量を表すように、展開画像の提示を制御する。すなわち、展開画像制御部９０は、複数のサブ文書群のそれぞれについて、複数のサブクラスタのそれぞれに分類される文書の情報量を表す展開画像を生成する。そして、展開画像制御部９０は、生成した展開画像を表示装置１２に出力して、表示装置１２に展開画像を表示させる。
【００４９】
なお、展開画像が表示された後、項目選択部７８は、展開画像に情報量が表示された複数のサブ項目のうちの、何れか１つのサブ項目の選択操作を受け付けてもよい。複数のサブ項目のうち何れか１つのサブ項目が選択された場合、スコア算出部８０、決定部８２、サブクラスタ生成部８４、サブ項目抽出部８６、第２算出部８８および展開画像制御部９０は、複数のサブ項目を複数の項目と置き換え、複数のサブクラスタを複数の候補クラスタと置き換えて、再度処理を実行して新たな展開画像の提示を制御する。
【００５０】
図４は、情報処理装置２０の処理の流れを示すフローチャートである。まず、Ｓ１１において、情報処理装置２０は、事前処理を実行する。続いて、Ｓ１２において、情報処理装置２０は、初期画像の表示処理を実行する。そして、Ｓ１３において、情報処理装置２０は、展開画像の表示処理を実行する。以下、Ｓ１１、Ｓ１２およびＳ１３の処理について詳細に説明する。
【００５１】
図５は、事前処理（Ｓ１１）の詳細な処理の流れを示すフローチャートである。情報処理装置２０は、Ｓ１１の事前処理において、以下のＳ２１からＳ２６の処理を実行する。
【００５２】
Ｓ２１において、情報処理装置２０は、他の装置から文書を取得する。続いて、Ｓ２２において、情報処理装置２０は、取得した文書に対して形態素解析および複合語抽出処理等を行って、取得した文書に対する１または複数のキーフレーズを生成する。続いて、Ｓ２３において、情報処理装置２０は、取得した文書と、生成した１または複数のキーフレーズとを対応付けて、文書記憶部４２に登録する。
【００５３】
情報処理装置２０は、Ｓ２１からＳ２３までの処理を、複数の文書のそれぞれに対して

実行する。これにより、文書記憶部４２は、複数の文書を含む文書群を記憶することができる。この文書群は、データベース化されている。情報処理装置２０は、文書群に対して任意のキーフレーズを指定することにより、指定されたキーフレーズに対応付けられた１または複数の文書を文書群から抽出することができる。

【００５４】

続いて、Ｓ２４において、情報処理装置２０は、文書記憶部４２から文書群に含まれる複数のキーフレーズを含むキーフレーズ群を取得する。そして、情報処理装置２０は、取得したキーフレーズ群を階層クラスタリングする。情報処理装置２０は、階層化した複数のクラスタのそれぞれにラベルを付与してもよい。情報処理装置２０は、階層クラスタリングしたキーフレーズ群をクラスタ記憶部４４に登録する。

【００５５】

続いて、Ｓ２５において、情報処理装置２０は、階層クラスタリングされたキーフレーズ群をクラスタ記憶部４４から読み出し、読み出したキーフレーズ群を複数の候補クラスタに分割する。そして、情報処理装置２０は、生成した複数の候補クラスタをクラスタ記憶部４４に登録する。

【００５６】

続いて、Ｓ２６において、情報処理装置２０は、文書群を分類するための予め定められた複数の項目を他の装置から取得する。情報処理装置２０は、ユーザにより入力された複数の項目を取得してもよい。そして、情報処理装置２０は、取得した複数の項目および複数の項目の階層構造を項目記憶部４６に記憶させる。

【００５７】

図６は、Ｓ２４において階層クラスタリングされたキーフレーズ群の構造を表すデンドログラムの一例を示す図である。キーフレーズ群を階層クラスタリングした場合、例えば、情報処理装置２０は、図６に示すようなデンドログラムにより表される階層構造の複数のクラスタを生成することができる。図６のデンドログラムは、末端ノードにキーフレーズが対応付けられる。また、高さ方向は、キーフレーズ間の類似度を表す。

【００５８】

図７は、キーフレーズ群の構造を表すデンドログラム、および、Ｓ２５において生成された候補クラスタの一例を示す図である。例えば、情報処理装置２０は、階層クラスタリングしたキーフレーズ群を分割して、複数の候補クラスタを生成する。例えば、情報処理装置２０は、図７に示すようなデンドログラムに基づき、所定個（例えば、４個以上で最小）の候補クラスタを生成する。情報処理装置２０は、複数の候補クラスタのそれぞれにラベルを付加してもよい。例えば、情報処理装置２０は、候補クラスタの中心位置近傍のキーフレーズを、その候補クラスタのラベルとしてもよい。

【００５９】

図７の例においては、情報処理装置２０は、類似度が１．００でクラスタを切断して、４個の候補クラスタを生成している。具体的には、情報処理装置２０は、ラベルが「アクション」の候補クラスタ、ラベルが「現象」の候補クラスタ、ラベルが「用語」の候補クラスタ、および、ラベルが「その他」の候補クラスタを生成している。

【００６０】

図８は、Ｓ２６において取得された複数の項目の階層構造を示す図である。例えば、情報処理装置２０は、図８に示すような、木構造で階層化された複数の項目を取得する。複数の項目の内容および階層構造は、例えばユーザにより予め定められている。情報処理装置２０は、このような複数の項目を他の装置から取得する。また、情報処理装置２０は、ユーザにより入力された複数の項目を取得してもよい。

【００６１】

これらの複数の項目のそれぞれは、文書群を分類するための情報である。情報処理装置２０は、記憶装置１６に記憶された文書群に対して、何れかの項目を指定することにより、その項目に関連付けられた文書を取得することができる。

【００６２】

図９は、初期画像の表示処理（Ｓ１２）の詳細な処理の流れを示すフローチャートである。情報処理装置２０は、Ｓ１２の初期画像の表示処理において、以下のＳ３１からＳ３４の処理を実行する。

【００６３】

Ｓ３１において、情報処理装置２０は、ユーザによる開始操作を、入力装置１４から受け付ける。続いて、Ｓ３２において、情報処理装置２０は、複数の候補クラスタをクラスタ記憶部４４から取得する。また、情報処理装置２０は、項目記憶部４６から予め定められた複数の項目のうち最上位の複数の項目を取得する。

【００６４】

続いて、Ｓ３３において、情報処理装置２０は、文書記憶部４２にアクセスして、予め定められた複数の項目のそれぞれ毎、且つ、複数の候補クラスタのそれぞれ毎の文書の情報量を算出する。すなわち、情報処理装置２０は、文書群を最上位の複数の項目に従って複数の初期文書群に分類する。そして、情報処理装置２０は、複数の初期文書群のそれぞれについて、複数の候補クラスタのそれぞれに分類される文書の情報量を算出する。

【００６５】

続いて、Ｓ３４において、情報処理装置２０は、複数の項目のそれぞれ毎、且つ、複数の候補クラスタのそれぞれ毎の文書の情報量に基づき、初期画像を生成する。そして、情報処理装置２０は、生成した初期画像を表示装置１２へと出力して、表示装置１２に初期画像を表示させる。

【００６６】

図１０は、Ｓ３４で表示される初期画像の一例を示す図である。情報処理装置２０は、Ｓ３４において、例えば、図１０に示すような、ヒートマップ状の初期画像を生成する。

【００６７】

初期画像は、一方の軸（項目軸）が項目を表し、他方の軸（クラスタ軸）が候補クラスタを表す２次元の格子状となっている。図１０の例では、項目軸が縦軸、クラスタ軸が横軸となっている。そして、初期画像は、複数の格子の内部のそれぞれの輝度または濃度が、対応する項目且つ対応する候補クラスタにより分類された文書の情報量を表す。

【００６８】

例えば、図１０の例の初期画像では、項目軸が、「機器」、「建屋」および「部品」の３個の項目を表す。また、この初期画像では、クラスタ軸が、「アクション」、「現象」、「用語」および「その他」の４個の候補クラスタを表す。そして、この初期画像では、「機器」および「アクション」の両者に対応する格子の内部の輝度または濃度が、項目が「機器」且つ候補クラスタが「アクション」に分類される文書の情報量を表す。他の格子の内部の輝度または濃度も同様である。

【００６９】

なお、初期画像は、図１０に示すようなヒートマップ状の画像に限られない。例えば、初期画像は、輝度または濃度に代えて、情報量を色によって表してもよい。また、初期画像は、情報量を数値または文字で表してもよい。また、初期画像は、オブジェクトまたはアイコンの種別の違いにより表してもよいし、ラベル文字の大小、フォント種別、描画線または線の傾き等の図形の変化により表してもよい。また、初期画像は、２次元のマップ状に限らず、情報量を短文のコメントまたはリストで表した文字情報であってもよいし、立体形状等の３次元以上の形状で情報量を表した画像であってもよい。

【００７０】

また、表示装置１２に表示された初期画像は、ユーザが、入力装置１４を用いて項目軸に表示された複数の項目のうちの何れか１つの項目を選択することが可能である。選択操作は、マウス等のポインティングデバイスにより行われてもよいし、音声等で行われてもよい。

【００７１】

図１１は、展開画像の表示処理（Ｓ１３）の詳細な処理の流れを示すフローチャートである。情報処理装置２０は、Ｓ１３の展開画像の表示処理において、以下のＳ４１からＳ

４８の処理を実行する。

【００７２】

Ｓ４１において、情報処理装置２０は、ユーザによる、初期画像に表示された複数の項目のうちの何れか１つ項目の選択操作を、入力装置１４から受け付ける。

【００７３】

続いて、Ｓ４２において、情報処理装置２０は、複数の候補クラスタのそれぞれについて、Ｓ４１で選択された項目に対する有用性を表すスコアを算出する。なお、スコアについては、図１５を参照して詳細を後述する。

【００７４】

続いて、Ｓ４３において、情報処理装置２０は、複数の候補クラスタのうち、算出されたスコアが所定の順位の候補クラスタを、参照クラスタとして決定する。例えば、決定部８２は、複数の候補クラスタのうち、有用性が最も高いスコアの候補クラスタを、参照クラスタとして決定する。

【００７５】

続いて、Ｓ４４において、情報処理装置２０は、参照クラスタを、複数のサブクラスタに分割する。例えば、情報処理装置２０は、参照クラスタを所定個（例えば、４個以上で最小）に分割して、複数のサブクラスタを生成する。また、この場合、情報処理装置２０は、複数のサブクラスタのそれぞれにラベルを付加してもよい。ラベルは、そのサブクラスタの中心近傍のサブフレーズ等であってもよい。

【００７６】

続いて、Ｓ４５において、情報処理装置２０は、項目記憶部４６から、選択された項目の下位層の予め定められた複数のサブ項目を抽出する。

【００７７】

続いて、Ｓ４６において、情報処理装置２０は、文書記憶部４２にアクセスして、複数のサブ項目のそれぞれ毎、且つ、複数のサブクラスタのそれぞれ毎の文書の情報量を算出する。すなわち、情報処理装置２０は、選択された項目の初期文書群をさらに複数のサブ項目に従って分割して複数のサブ文書群を生成する。そして、情報処理装置２０は、複数のサブ文書群のそれぞれについて、複数のサブクラスタのそれぞれに分類される文書の情報量を算出する。

【００７８】

続いて、Ｓ４７において、情報処理装置２０は、複数のサブ項目のそれぞれ毎、且つ、複数のサブクラスタのそれぞれ毎の文書の情報量に基づき、展開画像を生成する。そして、情報処理装置２０は、生成した展開画像を表示装置１２へと出力して、表示装置１２に展開画像を表示させる。

【００７９】

続いて、Ｓ４８において、情報処理装置２０は、展開画像を表示した後に、複数のサブ項目のうちの、何れか１つのサブ項目の選択操作を受け付けたか否かを判断する。何れか１つのサブ項目の選択操作を受け付けた場合（Ｓ４８のＹｅｓ）、情報処理装置２０は、処理をＳ４２に戻す。そして、情報処理装置２０は、複数のサブ項目を複数の項目と置き換え、複数のサブクラスタを複数の候補クラスタと置き換えて、再度、Ｓ４２～Ｓ４７の処理を実行して新たな展開画像を生成する。

【００８０】

展開画像を表示した後、何れのサブ項目も選択されずに終了操作がされた場合（Ｓ４８のＮｏ）、情報処理装置２０は、本フローの処理を終了する。

【００８１】

図１２は、初期画像および展開画像の一例を示す図である。情報処理装置２０は、初期画像を表示している状態において、何れか１つの項目が選択された場合、図１２の右側に示すようなヒートマップ状の展開画像を生成して表示装置１２に表示させる。

【００８２】

展開画像は、一方の軸（項目軸）がサブ項目を表し、他方の軸（クラスタ軸）がサブク

ラスタを表す２次元の格子状となっている。図１２の例では、項目軸が縦軸、クラスタ軸が横軸となっている。そして、展開画像は、初期画像と同様に、複数の格子の内部のそれぞれの輝度または濃度が、対応するサブ項目且つ対応するサブクラスタにより分類された文書の情報量を表す。

【００８３】

例えば、図１２の例では、初期画像の項目軸に表示された複数の項目のうち、「機器」の項目がユーザにより選択された。「機器」の項目が選択されたことに応じて、情報処理装置２０は、「機器」の下位層の「タービン」、「ノズル」、「ポンプ」、「配管」および「ロータ」の５個のサブ項目を抽出した。

【００８４】

また、図１２の例では、情報処理装置２０は、「機器」の項目が選択されたことに応じて、「アクション」、「現象」、「用語」および「その他」の４個の候補クラスタのそれぞれについて、選択された項目である「機器」に対する有用性を表すスコアを算出した。そして、情報処理装置２０は、スコアが最も高い「アクション」の候補クラスタを、参照クラスタとして決定した。さらに、情報処理装置２０は、「アクション」の参照クラスタを分割して、「検査」、「溶接」、「拡大」、「加工」および「位置」の５個のサブクラスタを生成した。

【００８５】

そして、情報処理装置２０は、図１２に示すような展開画像を生成した。図１２の例の展開画像では、項目軸が、「機器」の下位層のサブ項目である、「タービン」、「ノズル」、「ポンプ」、「配管」および「ロータ」の５個のサブ項目を表す。また、この展開画像では、クラスタ軸が、「検査」、「溶接」、「拡大」、「加工」および「位置」の５個のサブクラスタを表す。

【００８６】

なお、展開画像も、初期画像と同様に、ヒートマップ状の画像に限られない。また、表示装置１２に表示された展開画像は、ユーザが、入力装置１４を用いて項目軸に表示された複数のサブ項目のうちの何れか１つのサブ項目を選択することが可能である。

【００８７】

図１３は、初期画像、展開画像および新たな展開画像の一例を示す図である。情報処理装置２０は、展開画像を表示している状態において、何れか１つのサブ項目が選択された場合、新たな展開画像を生成して表示装置１２に表示させる。この場合、情報処理装置２０は、複数のサブ項目を複数の項目と置き換え、複数のサブクラスタを複数の候補クラスタと置き換えて、再度、展開画像を生成する処理を実行し、新たな展開画像を生成する。

【００８８】

図１４は、メニュー画像が追加された展開画像を示す図である。情報処理装置２０の展開画像制御部９０は、展開画像とともに、複数の候補クラスタのうちの何れか１つを選択させるためのメニュー画像の提示を制御してもよい。メニュー画像は、例えば、プルダウンメニュー９２のような、ユーザに操作により複数の候補クラスタのうちの何れか１つを選択させるためのユーザインターフェイス画像である。

【００８９】

情報処理装置２０の展開画像制御部９０は、メニュー画像により何れか１つの新たな候補クラスタが選択された場合、新たな候補クラスタを参照クラスタとして、再度、展開画像を提示する。具体的には、情報処理装置２０のサブクラスタ生成部８４は、新たな候補クラスタを参照クラスタとして、新たな複数のサブクラスタに分割する。

【００９０】

情報処理装置２０の第２算出部８８は、サブクラスタ生成部８４から、新たな複数のサブクラスタを取得する。第２算出部８８は、文書記憶部４２にアクセスして、予め定められた複数のサブ項目のそれぞれ毎、且つ、新たな複数のサブクラスタのそれぞれ毎の文書の情報量を算出する。そして、展開画像制御部９０は、複数のサブ項目のそれぞれ毎且つ新たな複数のサブクラスタのそれぞれ毎の文書の情報量を表す新たな展開画像を生成し、

表示装置１２に表示させる。

【００９１】

図１５は、複数の候補クラスタのそれぞれのスコアを示す図である。情報処理装置２０は、初期画像において、何れかの項目が選択された場合、クラスタ軸に表示する複数のサブクラスタを自動的に選択する。この場合、情報処理装置２０は、選択された項目に対して、最も有用なサブクラスタがクラスタ軸に表示されるように、選択された項目に対する有用性を表すスコアを、複数の候補クラスタのそれぞれに対して算出する。

【００９２】

ｎ番目（ｎは１以上の整数）の候補クラスタを“Ｃｎ”とした場合、情報処理装置２０は、下記の式（１）を演算して、選択された項目に対するｎ番目の候補クラスタのスコア（Ｖ（Ｃｎ））を算出する。

Ｖ（Ｃｎ）＝α＋β＋γ　…（１）

【００９３】

αは、スコアの第１パラメータである。αは、対応する候補クラスタにおける選択された項目に分類される文書の情報量を表す。αは、このような文書の情報量に係数等を乗じた値であってもよい。

【００９４】

例えば、項目として「機器」が選択され、「アクション」の候補クラスタのスコアを算出する場合、αは、アクションの候補クラスタに含まれる複数の文書の情報量のうちの、「機器」に分類される複数の文書の情報量に応じた値を表す。例えば、スコアは、対応する候補クラスタにおける選択された項目に分類される文書の情報量が大きい程、大きくなる。

【００９５】

βは、スコアの第２パラメータである。βは、複数のサブ項目のそれぞれ毎、且つ、対応する候補クラスタを分割した複数のサブクラスタのそれぞれ毎の文書の情報量の分散を表す。βは、このような分散に係数等を乗じた値であってもよい。

【００９６】

例えば、項目として「機器」が選択され、「アクション」の候補クラスタのスコアを算出する場合、選択された項目の下位層の複数のサブ項目は、「タービン」、「ノズル」、「ポンプ」、「配管」および「ロータ」となる。また、対応する「アクション」の候補クラスタを分割した複数のサブクラスタは、「検査」、「溶接」、「拡大」、「加工」および「位置」となる。情報処理装置２０は、「タービン」、「ノズル」、「ポンプ」、「配管」および「ロータ」の５個のサブ項目のそれぞれと、「検査」、「溶接」、「拡大」、「加工」および「位置」の５個のサブクラスタのそれぞれとの組み合わせ毎の文書の情報量を算出する。βは、これらの組み合わせ毎の情報量の分散を表す。例えば、スコアは、このような分散が大きい程、大きくなる。

【００９７】

γは、スコアの第３パラメータである。γは、対応する候補クラスタをユーザが過去に操作により選択した頻度または割合を表す。γは、このような頻度または割合に係数等を乗じた値であってもよい。

【００９８】

例えば、項目として「機器」が選択され、「アクション」の候補クラスタのスコアを算出する場合、γは、過去に、項目として「機器」が選択された後に、例えば、図１４に示すメニュー画面等により「アクション」の候補クラスタが選択された頻度または割合を表す。また、γは、時間的に近い選択操作が、時間的に遠い選択操作よりも大きな影響を与えるように、頻度または割合に重みが加えられてもよい。例えば、スコアは、対応する候補クラスタを過去にユーザが操作により選択した頻度または割合が多い程、大きくなる。

【００９９】

また、スコアは、α、βおよびγのうち、何れか１つのパラメータに基づく値、または、何れか２つのパラメータに基づく値であってもよい。また、スコアは、α、βおよびγ

に代えて、選択された項目に対する候補クラスタの有用性を表す他のパラメータにより表されてもよい。

【０１００】

（第１実施形態の効果）

以上のように、第１実施形態に係る情報処理装置２０は、文書群を、検索目的およびユーザの関心点に応じた２つの軸により分類して情報量を提供する。この場合において、情報処理装置２０は、一方の軸（項目軸）については予め定めた項目で分類するが、他方の軸（クラスタ軸）についてはクラスタリングにより分類する。そして、情報処理装置２０は、ユーザにより予め定められた複数の項目のうち何れか１つの項目が選択された場合、選択された項目に対して有用な複数のサブクラスタを自動的に抽出して、展開画像を生成する。

【０１０１】

例えば、情報処理装置２０は、選択された項目に分類される文書の情報量が多くなるように、複数のサブクラスタを自動的に抽出する。また、例えば、情報処理装置２０は、文書の情報量の分散が大きくなるように複数のサブクラスタを自動的に抽出する。また、例えば、情報処理装置２０は、過去にユーザが高い頻度で選択した複数のサブクラスタを自動的に抽出する。

【０１０２】

このように情報処理装置２０は、他方の軸（クラスタ軸）をクラスタリングにより分類するので、小さいコストで文書群を分類することができる。さらに、情報処理装置２０は、他方の軸（クラスタ軸）を、選択した項目に対して有用な複数のサブクラスタに分類するので、文書群を適切に分類することができる。以上のように、情報処理装置２０は、文書群を適切に分類した情報を小さいコストで提示することができる。

【０１０３】

（第２実施形態）

つぎに、第２実施形態に係る文書管理システム１０について説明をする。第２実施形態に係る文書管理システム１０は、第１実施形態に係る文書管理システム１０と略同一の構成および機能を有する。第２実施形態の説明では、第１実施形態で説明したユニットと、略同一の機能および構成のユニットには同一の符号を付けて、相違点を除き詳細な説明を省略する。

【０１０４】

図１６は、第２実施形態に係る第１画像制御部３４および第２画像制御部３６の構成を文書記憶部４２および項目記憶部４６とともに示す図である。

【０１０５】

第２実施形態において、項目記憶部４６は、文書群を第１観点により分類するための予め定められた複数の第１観点項目を記憶する。さらに、項目記憶部４６は、文書群を第１観点とは異なる第２観点により分類するための複数の第２観点項目を記憶する。

【０１０６】

第２実施形態に係る第１画像制御部３４は、開始受付部７２と、第１算出部７４と、初期画像制御部７６とを有する。

【０１０７】

第１算出部７４は、開始受付部７２が開始操作を受け付けると、項目記憶部４６から、複数の第１観点項目および複数の第２観点項目を取得する。そして、第１算出部７４は、文書記憶部４２にアクセスして、予め定められた複数の第１観点項目のそれぞれ毎、且つ、予め定められた複数の第２観点項目のそれぞれ毎の文書の情報量を算出する。すなわち、第１算出部７４は、文書群を複数の第１観点項目に従って複数の第１観点文書群に分類する。そして、第１算出部７４は、複数の第１観点文書群のそれぞれについて、複数の第２観点項目のそれぞれに分類される文書の情報量を算出する。

【０１０８】

初期画像制御部７６は、予め定められた複数の第１観点項目のそれぞれ毎、且つ、予め

定められた複数の第２観点項目のそれぞれ毎の文書の情報量を表すように、項目選択画像の提示を制御する。すなわち、初期画像制御部７６は、文書群を複数の第１観点項目に従って分類した複数の第１観点文書群のそれぞれについて、複数の第２観点項目のそれぞれに分類される文書の情報量を表す項目選択画像を生成する。そして、初期画像制御部７６は、生成した項目選択画像を表示装置１２に出力して、表示装置１２に項目選択画像を表示させる。

【０１０９】

第２実施形態に係る第２画像制御部３６は、項目選択部７８と、文書抽出部９４と、選択クラスタリング部９６と、選択分割部９８と、スコア算出部８０と、決定部８２と、サブクラスタ生成部８４と、第２算出部８８と、展開画像制御部９０とを有する。

【０１１０】

項目選択部７８は、ユーザによる、複数の第１観点項目のうちの何れか１つの第１観点項目、および、複数の第２観点項目のうちの何れか１つの第２観点項目の選択操作を、入力装置１４から受け付ける。

【０１１１】

文書抽出部９４は、文書記憶部４２にアクセスして、文書群から、選択された第１観点項目および選択された第２観点項目の両者に分類される複数の文書を含む選択文書群を抽出する。選択クラスタリング部９６は、文書抽出部９４により抽出された選択文書群に含まれる複数のキーフレーズを含む選択キーフレーズ群を取得する。そして、選択クラスタリング部９６は、取得した選択キーフレーズ群を階層クラスタリングする。選択分割部９８は、階層クラスタリングされた選択キーフレーズ群を複数の候補クラスタに分割する。

【０１１２】

スコア算出部８０は、選択分割部９８により分割された複数の候補クラスタのそれぞれについて、選択された第１観点項目および第２観点項目に対する有用性を表すスコアを算出する。決定部８２は、複数の候補クラスタのうち、算出されたスコアが所定の順位の２つの候補クラスタを、第１参照クラスタおよび第２参照クラスタとして決定する。例えば、決定部８２は、複数の候補クラスタのうち、有用性が最も高いスコアの候補クラスタを第１参照クラスタとして決定し、有用性が２番目に高いスコアの候補クラスタを第２参照クラスタとして決定する。

【０１１３】

サブクラスタ生成部８４は、第１参照クラスタを、複数の第１サブクラスタに分割する。例えば、サブクラスタ生成部８４は、第１参照クラスタを所定個（例えば、４個以上で最小）に分割して、複数の第１サブクラスタを生成する。また、サブクラスタ生成部８４は、第２参照クラスタを、複数の第２サブクラスタに分割する。例えば、サブクラスタ生成部８４は、第２参照クラスタを所定個（例えば、４個以上で最小）に分割して、複数の第２サブクラスタを生成する。

【０１１４】

第２算出部８８は、サブクラスタ生成部８４から、複数の第１サブクラスタおよび複数の第２サブクラスタを取得する。そして、第２算出部８８は、複数の第１サブクラスタのそれぞれ毎且つ複数の第２サブクラスタのそれぞれ毎の文書の情報量を算出する。すなわち、第２算出部８８は、複数の第１サブクラスタに従って複数の第１サブ文書群に分類する。そして、第２算出部８８は、複数の第１サブ文書群のそれぞれについて、複数の第２サブクラスタのそれぞれに分類される文書の情報量を算出する。

【０１１５】

展開画像制御部９０は、複数の第１サブクラスタのそれぞれ毎且つ複数の第２サブクラスタのそれぞれ毎の文書の情報量を表すように、クラスタ展開画像の提示を制御する。すなわち、展開画像制御部９０は、複数の第１サブ文書群のそれぞれについて、複数の第２サブクラスタのそれぞれに分類される文書の情報量を表すクラスタ展開画像を生成する。そして、展開画像制御部９０は、生成したクラスタ展開画像を表示装置１２に出力して、表示装置１２にクラスタ展開画像を表示させる。

【０１１６】
図１７は、項目記憶部４６に記憶された複数の第１観点項目および複数の第２観点項目の一例を示す図である。項目記憶部４６は、文書群を異なる観点で分類するための複数の第１観点項目および複数の第２観点項目を記憶する。複数の第１観点項目および複数の第２観点項目は、例えばユーザにより予め定められている。情報処理装置２０は、このような複数の第１観点項目および複数の第２観点項目を他の装置から取得する。また、情報処理装置２０は、ユーザにより入力された複数の第１観点項目および複数の第２観点項目を取得してもよい。
【０１１７】
複数の第１観点項目と複数の第２観点項目とは、文書群を異なる観点で分類するための情報である。従って、情報処理装置２０は、文書群を何れかの第１観点項目で絞り込んだ後、さらに何れかの第２観点項目で絞り込むことができる。
【０１１８】
図１８は、項目選択画像の一例を示す図である。情報処理装置２０は、例えば、図１８に示すような、ヒートマップ状の項目選択画像を生成する。
【０１１９】
項目選択画像は、一方の軸（第１観点軸）が第１観点項目を表し、他方の軸（第２観点軸）が第２観点項目を表す２次元の格子状となっている。図１８の例では、第１観点軸が縦軸、第２観点軸が横軸となっている。そして、項目選択画像は、複数の格子の内部のそれぞれの輝度または濃度が、対応する第１観点項目且つ対応する第２観点項目により分類された文書の情報量を表す。
【０１２０】
例えば、図１８の例の項目選択画像では、第１観点軸が、「機器」、「建屋」および「部品」の３個の第１観点項目を表す。また、この項目選択画像では、第２観点軸が、「品質部門」、「設計部門」および「製造部門」の３個の第２観点項目を表す。そして、この項目選択画像では、「機器」および「品質部門」の両者に対応する格子の内部の輝度または濃度が、第１観点項目が「機器」且つ第２観点項目が「品質部門」に分類される文書の情報量を表す。他の格子の内部の輝度または濃度も同様である。
【０１２１】
なお、項目選択画像は、第１実施形態で説明した初期画像と同様に、図１８に示すようなヒートマップ状の画像に限られない。また、表示装置１２に表示された項目選択画像は、ユーザが、入力装置１４を用いて、第１観点軸に表示された複数の第１観点項目のうちの何れか１つの第１観点項目と、第２観点軸に表示された複数の第２観点項目のうちの何れか１つの第２観点項目とを同時に選択することができる。
【０１２２】
図１９は、項目選択画像および選択キーフレーズ群の構造を表すデンドログラムの一例を示す図である。項目選択画像が表示された後、情報処理装置２０は、ユーザによる、複数の第１観点項目のうちの何れか１つの第１観点項目、および、複数の第２観点項目のうちの何れか１つの第２観点項目の選択操作を、入力装置１４から受け付ける。例えば、２次元の複数の格子が表示されている項目選択画像における何れか１つの格子（タイル）が選択された場合、情報処理装置２０は、そのタイルに対応する第１観点項目および第２観点項目の選択操作を受け付ける。
【０１２３】
第１観点項目および第２観点項目が選択された場合、情報処理装置２０は、文書記憶部４２にアクセスして、文書群から、選択された第１観点項目および選択された第２観点項目の両者に分類される複数の文書を含む選択文書群を抽出する。さらに、情報処理装置２０は、文書抽出部９４により抽出された選択文書群に含まれる複数のキーフレーズを含む選択キーフレーズ群を取得する。そして、情報処理装置２０は、取得した選択キーフレーズ群を階層クラスタリングする。例えば、情報処理装置２０は、選択キーフレーズ群に対して、図１９に示すデンドログラムに表されるような階層クラスタリングを行う。

【０１２４】
図２０は、選択キーフレーズ群の構造を表すデンドログラム、および、クラスタ展開画像の一例を示す図である。例えば、情報処理装置２０は、図２０に示すようなデンドログラムにより表される階層構造のクラスタを分割して、複数の候補クラスタを生成する。例えば、情報処理装置２０は、デンドログラムに基づき、階層クラスタリングされた選択キーフレーズ群を分割して、所定個（例えば、４個以上で最小）の候補クラスタを生成する。
【０１２５】
情報処理装置２０は、複数の候補クラスタのそれぞれにラベルを付加してもよい。例えば、情報処理装置２０は、候補クラスタの中心位置近傍のキーフレーズを、その候補クラスタのラベルとしてもよい。
【０１２６】
図２０の例においては、情報処理装置２０は、３個の候補クラスタを生成している。具体的には、情報処理装置２０は、ラベルが「Ａ」の候補クラスタ、ラベルが「Ｂ」の候補クラスタ、ラベルが「Ｃ」の候補クラスタ、および、ラベルが「Ｄ」の候補クラスタを生成している。
【０１２７】
続いて、情報処理装置２０は、分割した複数の候補クラスタのそれぞれについて、選択された第１観点項目および第２観点項目に対する有用性を表すスコアを算出する。続いて、情報処理装置２０は、複数の候補クラスタのうち、有用性が最も高いスコアの候補クラスタを第１参照クラスタとして決定し、有用性が２番目に高いスコアの候補クラスタを第２参照クラスタとして決定する。例えば、図２０の例では、ラベルが「Ｂ」の候補クラスタが第１参照クラスタとして決定され、ラベルが「Ａ」の候補クラスタが第２参照クラスタとして決定されている。
【０１２８】
続いて、情報処理装置２０は、第１参照クラスタを複数の第１サブクラスタに分割し、第２参照クラスタを複数の第２サブクラスタに分割する。そして、情報処理装置２０は、複数の第１サブクラスタのそれぞれ毎且つ複数の第２サブクラスタのそれぞれ毎の文書の情報量を表すクラスタ展開画像を生成する。
【０１２９】
クラスタ展開画像は、一方の軸が第１サブクラスタを表し、他方の軸が第２サブクラスタを表す２次元の格子状となっている。図２０の例では、縦軸が第１サブクラスタを表し、横軸が第２サブクラスタを表す。そして、クラスタ展開画像は、初期画像と同様に、複数の格子の内部のそれぞれの輝度または濃度が、対応する第１サブクラスタ且つ対応する第２サブクラスタにより分類された文書の情報量を表す。
【０１３０】
縦軸が、「コンプレッサ」、「フランジボルト」、「スパッタ」および「ステータ」の４個の第１サブクラスタを表す。また、このクラスタ展開画像では、横軸が、「破損」、「遅延」、「損傷」、「溶解」および「溶接」の５個の第２サブクラスタを表す。
【０１３１】
なお、クラスタ展開画像も、初期画像と同様に、ヒートマップ状の画像に限られない。また、クラスタ展開画像は、ユーザが、入力装置１４を用いて縦軸に表示された複数の第１サブクラスタのうちの何れか１つの第１サブクラスタを選択することが可能である。この場合、情報処理装置２０は、選択された第１サブクラスタのさらに下位層の複数のクラスタを縦軸に表示する。また、同様に、クラスタ展開画像は、ユーザが、入力装置１４を用いて横軸に表示された複数の第２サブクラスタのうちの何れか１つの第２サブクラスタを選択することが可能である。この場合、情報処理装置２０は、選択された第２サブクラスタのさらに下位層の複数のクラスタを横軸に表示する。
【０１３２】
図２１は、列選択および行選択をした場合の項目選択画像を示す図である。項目記憶部

４６に記憶された複数の第１観点項目および複数の第２観点項目は、木構造に階層構造化されていてもよい。

【０１３３】

この場合、情報処理装置２０は、ユーザによる、第１観点軸の複数の第１観点項目のうち何れか１つの第１観点項目の選択操作を、受け付けることができる。情報処理装置２０は、ユーザによる、第２観点軸の複数の第２観点項目のうち何れか１つの第２観点項目の選択操作を、受け付けることができる。

【０１３４】

例えば、ユーザにより項目選択画像の何れかの列が選択された場合、情報処理装置２０は、選択された列に対応する第２観点項目が選択されたと判断する。この場合、情報処理装置２０は、項目選択画像における横軸の項目を、選択された第２観点項目の下位層に展開される複数の項目に置き換える。

【０１３５】

また、例えば、ユーザにより項目選択画像の何れかの行が選択された場合、情報処理装置２０は、選択された行に対応する第１観点項目が選択されたと判断する。この場合、情報処理装置２０は、項目選択画像における縦軸の項目を、選択された第１観点項目の下位層に展開される複数の項目に置き換える。

【０１３６】

（第２実施形態の効果）

以上のように、第２実施形態に係る情報処理装置２０は、文書群を、検索目的およびユーザの関心点に応じた２つの軸により分類して情報量を提供する。この場合において、情報処理装置２０は、文書群を、予め定められた複数の第１観点項目および予め定められた複数の第２観点項目で分類して、項目選択画像を表示する。続いて、情報処理装置２０は、文書群から、ユーザにより選択された第１観点項目および選択された第２観点項目の両者に分類される複数の文書を含む選択文書群を抽出する。続いて、情報処理装置２０は、選択文書群をクラスタリングして複数の候補クラスタを生成する。続いて、情報処理装置２０は、選択された第１観点項目および第２観点項目に対して有用な２つの候補クラスタを自動的に決定する。そして、情報処理装置２０は、決定した２つの候補クラスタの一方を分類した複数の第１サブクラスタを一方の軸とし、他方を分類した複数の第２サブクラスタを他方の軸とした、クラスタ展開画像を生成する。

【０１３７】

このように情報処理装置２０は、２つの軸をクラスタリングにより分類するので、小さいコストで文書群を分類することができる。さらに、情報処理装置２０は、２つの軸を、抽出した選択文書群に対して有用な複数のサブクラスタに分類するので、文書群を適切に分類することができる。以上のように、情報処理装置２０は、文書群を適切に分類した情報を小さいコストで提示することができる。

【０１３８】

（プログラム）

情報処理装置２０で実行されるプログラムは、インストール可能な形式または実行可能な形式のファイルでＣＤ－ＲＯＭ、フレキシブルディスク（ＦＤ）、ＣＤ－Ｒ、ＤＶＤ等のコンピュータで読み取り可能な記録媒体に記録されて提供される。また、情報処理装置２０で実行されるプログラムを、インターネット等のネットワークに接続されたコンピュータ上に格納し、ネットワーク経由でダウンロードさせることにより提供するように構成してもよい。また、情報処理装置２０で実行されるプログラムをインターネット等のネットワーク経由で提供または配布するように構成してもよい。また、プログラムを、ＲＯＭ等に予め組み込んで提供するように構成してもよい。

【０１３９】

情報処理装置２０で実行されるプログラムは、事前処理モジュール（文書取得モジュール、キーフレーズ生成モジュール、文書登録モジュール、クラスタリングモジュール、分割モジュール、項目取得モジュールおよび項目登録モジュール）と、第１画像制御モジュ

ール（開始受付モジュール、第１算出モジュールおよび初期画像制御モジュール）と、第２画像制御モジュール（項目選択モジュール、スコア算出モジュール、決定モジュール、サブクラスタ生成モジュール、サブ項目抽出モジュール、第２算出モジュールおよび展開画像制御モジュール）とを有する。情報処理装置２０は、プロセッサ（処理回路３０）が記憶媒体（記憶装置１６等）からプログラムを読み出して実行することにより各モジュールが主記憶装置（記憶回路２４）上にロードされる。これにより、プロセッサ（処理回路３０）は、事前処理部３２（文書取得部５２、キーフレーズ生成部５４、文書登録部５６、クラスタリング部５８、分割部６０、項目取得部６２および項目登録部６４）、第１画像制御部３４（開始受付部７２、第１算出部７４および初期画像制御部７６）、第２画像制御部３６（項目選択部７８、スコア算出部８０、決定部８２、サブクラスタ生成部８４、サブ項目抽出部８６、第２算出部８８および展開画像制御部９０）として機能する。なお、これらの一部または全部がプロセッサ以外のハードウェアにより実現されてもよい。

【０１４０】

本発明のいくつかの実施形態を説明したが、これらの実施形態は、例として提示したものであり、発明の範囲を限定することは意図していない。これら新規な実施形態は、その他の様々な形態で実施されることが可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置き換え、変更を行うことができる。これら実施形態やその変形は、発明の範囲や要旨に含まれるとともに、請求の範囲に記載された発明とその均等の範囲に含まれる。

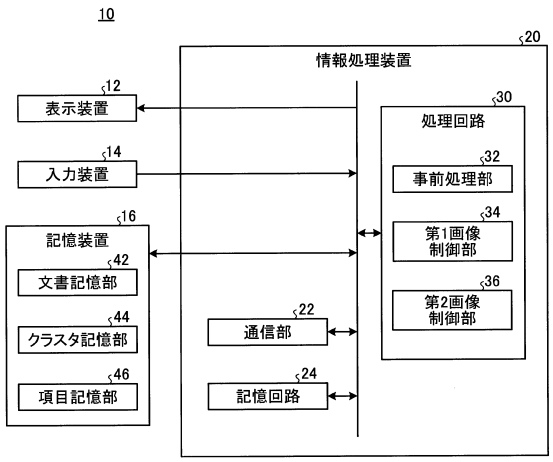
【符号の説明】

【０１４１】

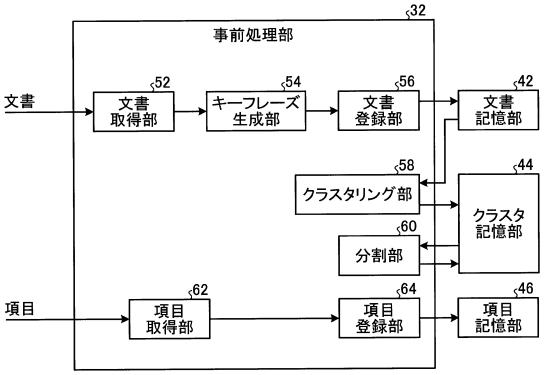
１０　文書管理システム
１２　表示装置
１４　入力装置
１６　記憶装置
２０　情報処理装置
２２　通信部
２４　記憶回路
３０　処理回路
３２　事前処理部
３４　第１画像制御部
３６　第２画像制御部
４２　文書記憶部
４４　クラスタ記憶部
４６　項目記憶部
５２　文書取得部
５４　キーフレーズ生成部
５６　文書登録部
５８　クラスタリング部
６０　分割部
６２　項目取得部
６４　項目登録部
７２　開始受付部
７４　第１算出部
７６　初期画像制御部
７８　項目選択部
８０　スコア算出部
８２　決定部
８４　サブクラスタ生成部
８６　サブ項目抽出部
８８　第２算出部

９０　展開画像制御部
９４　文書抽出部
９６　選択クラスタリング部
９８　選択分割部

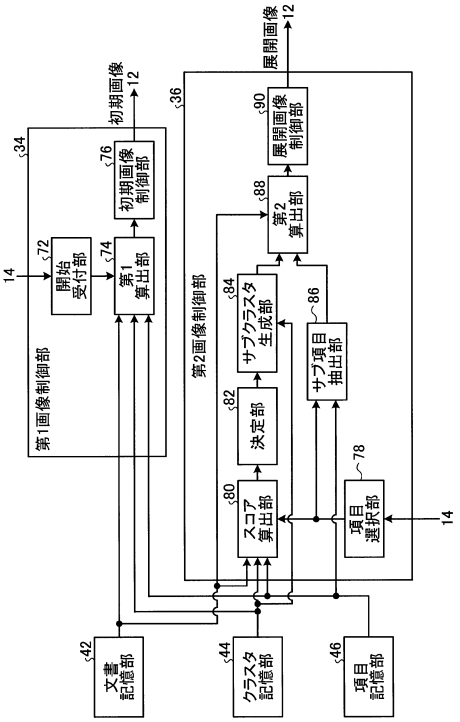
【図１】



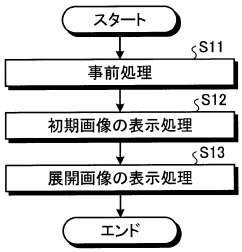
【図２】



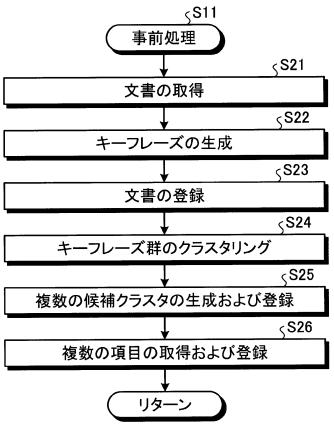
【図３】


展開画像
12
初期画像
12
34
76
初期画像制御部
72
開始受付部
74
第1算出部
14
第1画像制御部
36
第2画像制御部
90
展開画像制御部
88
第2算出部
84
サブクラスタ生成部
86
サブ項目抽出部
82
決定部
80
スコア算出部
78
項目選択部
14
42
文書記憶部
44
クラスタ記憶部
46
項目記憶部

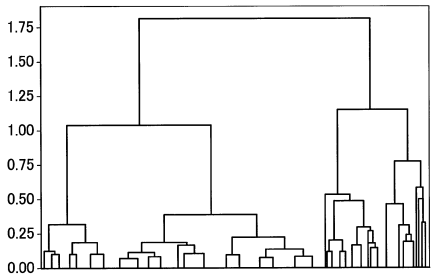
【図４】


スタート
S11
事前処理
S12
初期画像の表示処理
S13
展開画像の表示処理
エンド

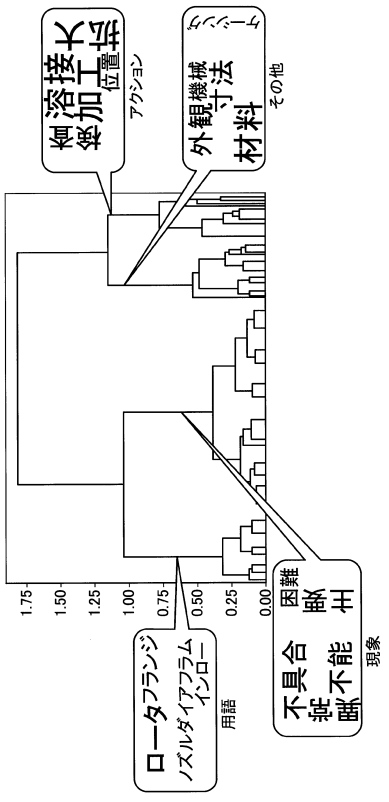
【図５】


S11
事前処理
S21
文書の取得
S22
キーフレーズの生成
S23
文書の登録
S24
キーフレーズ群のクラスタリング
S25
複数の候補クラスタの生成および登録
S26
複数の項目の取得および登録
リターン

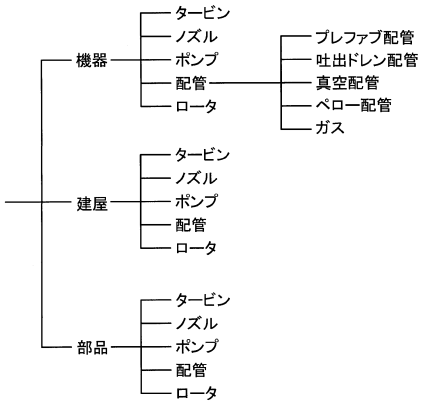
【図６】


1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

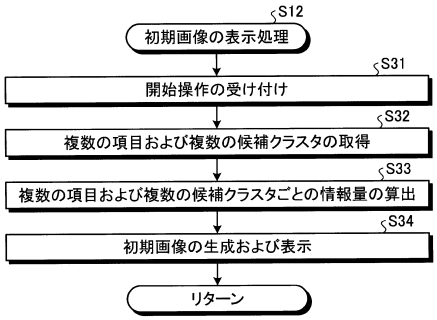
【図７】


溶接
接合
加工
位置
アクション
外観
機械
寸法
材料
その他
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
困難
不具合
故障
不能
現象
ロータ
フランジ
ノズル
ダイアフラム
インロー
用語

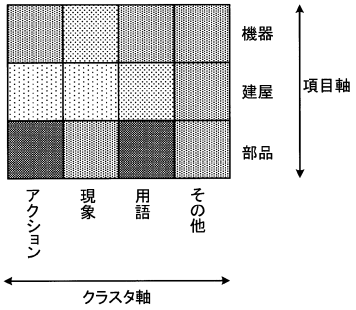
【図８】



【図９】

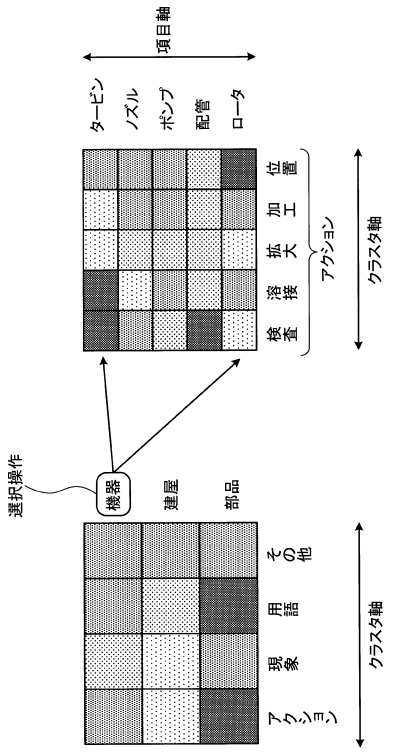


【図１０】

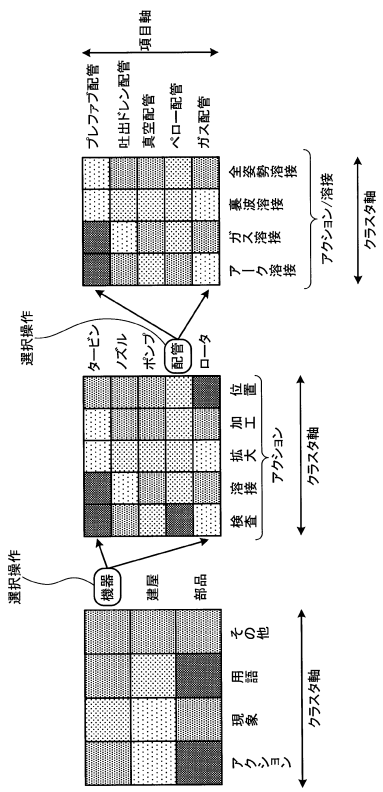


【図１１】

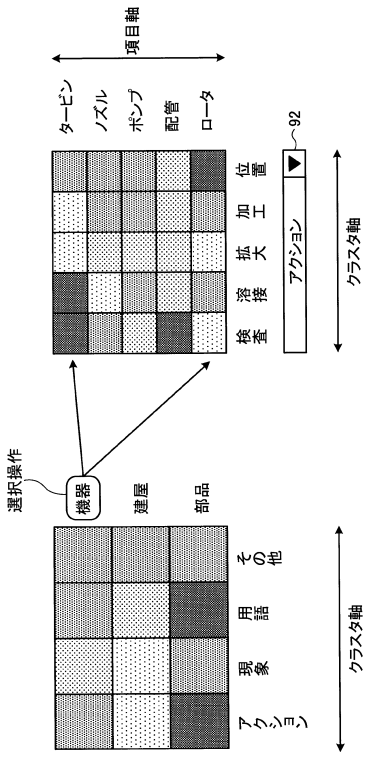
【図１２】



【図１３】



【図１４】

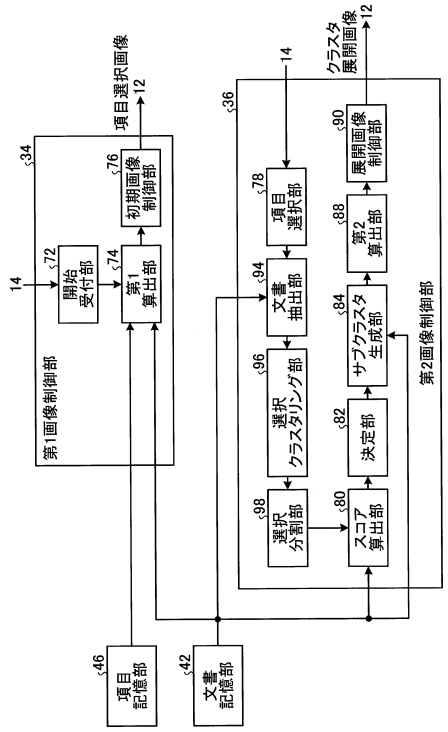


【図１５】

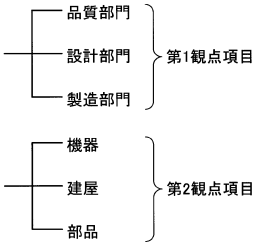
アクション　　現象　　用語　　その他

$$V(C_1) > V(C_2) > V(C_3) > V(C_4)$$

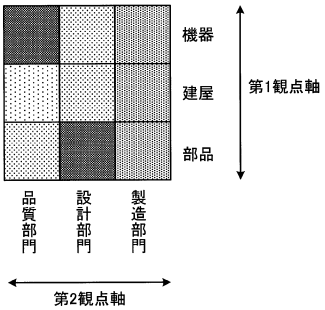
【図１６】



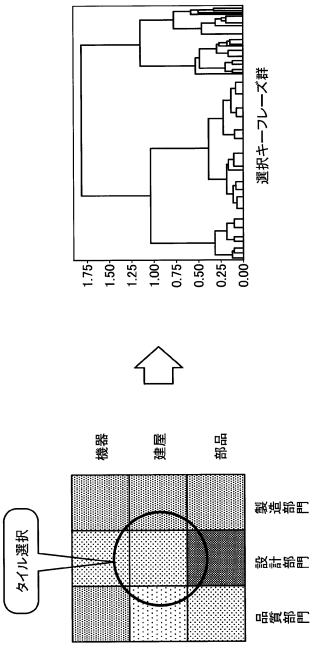
【図１７】



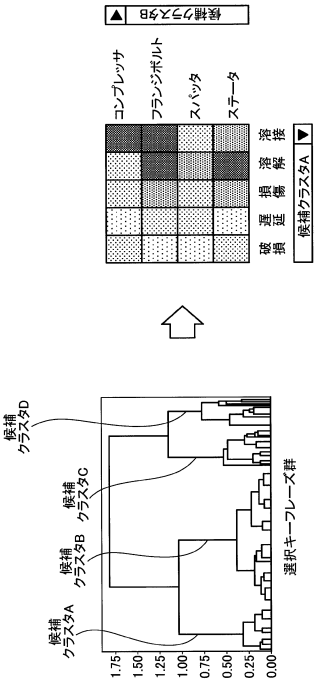
【図１８】



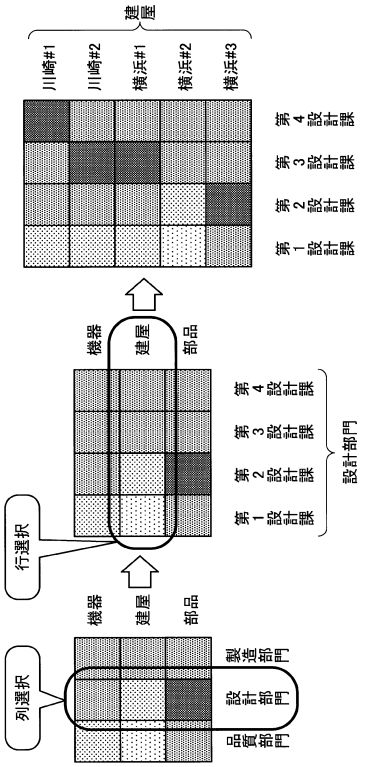
【図１９】



【図２０】



【図２１】



フロントページの続き

(56)参考文献　特開２０１１－１２８７０５（ＪＰ，Ａ）
国際公開第２００７／０６９６６３（WO，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１６／３５
Ｇ０６Ｆ　１６／３４