(19)日本国特許庁（ＪＰ）　(12)特　許　公　報（Ｂ２）　(11)特許番号
特許第3333998号
(P3333998)
(45)発行日　平成14年10月15日(2002.10.15)　(24)登録日　平成14年８月２日(2002.8.2)

| (51)Int.Cl.[7] | 識別記号 | ＦＩ | |
|---|---|---|---|
| Ｇ０６Ｆ　17/30 | ２１０ | Ｇ０６Ｆ　17/30 | ２１０Ｄ |

請求項の数９(全 13 頁)

(21)出願番号　特願平4－250385
(22)出願日　平成４年８月27日(1992.8.27)
(65)公開番号　特開平6－75995
(43)公開日　平成６年３月18日(1994.3.18)
審査請求日　平成11年３月16日(1999.3.16)

(73)特許権者　000002945
オムロン株式会社
京都市下京区塩小路通堀川東入南不動堂町801番地
(72)発明者　岸大路　泰明
京都府京都市右京区花園土堂町10番地　オムロン株式会社内
(72)発明者　尾崎　時夫
京都府京都市右京区花園土堂町10番地　オムロン株式会社内
(72)発明者　久野　敦司
京都府京都市右京区花園土堂町10番地　オムロン株式会社内
(74)代理人　100080322
弁理士　牛久　健司

審査官　高瀬　勤

最終頁に続く

(54)【発明の名称】　自動分類付与装置および方法

(57)【特許請求の範囲】
【請求項１】　分類未付与文書に含まれる複数のキーワードを入力する手段，キーワードごとに，そのキーワードに関連の深い分類およびその分類の関連の深さを示す度合をあらかじめ記憶したキーワード／分類テーブルを参照して，入力されたキーワードに関連する分類の関連度合の合計値を分類ごとに算出し，この合計値の大きさの順序にしたがって付与すべき分類の候補を選択する手段，ならびに２つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照して，選択された複数の候補分類相互間の距離が妥当な範囲内にあるかどうかを検査し，妥当な範囲内にあれば候補分類を最終分類と決定する手段，を備えた自動分類付与装置。
【請求項２】　上記決定手段は，上記合計値が所定値以上である候補分類が１つである場合に，上記分類間距離テーブルを参照することなくその候補分類を最終分類と決定する，請求項１に記載の自動分類付与装置。
【請求項３】　上記合計値が所定値よりも大きい候補分類がない場合に，上記合計値の大きさの順序にしたがって複数の分類からなる分類パターンを作成し，同一分類パターンが所定回数出現したときに新たな分類を創設して付与する手段，をさらに備えた請求項１に記載の自動分類付与装置。
【請求項４】　一文書について複数の分類からなる分類の組があらかじめ付与された複数の分類付与済文書のそれぞれについて，それらの文書に付与された分類の組を入力するための手段，および入力された分類の組に２つの分類が同時に含まれる程度に基づいて，２つの分類間の距離を，すべての分類の中から選択されたすべての組合せの分類対について求め，分類間距離テーブルを作成

する手段，を備えた分類間距離テーブル作成装置。
【請求項５】　分類があらかじめ付与された複数の分類付与済文書のそれぞれについて，それらの文書に付与された分類とそれらの文書から抽出されたキーワードとを相互に関連させて入力するための手段，入力されたキーワードごとに，それらのキーワードに関連する分類の関連度合を求め，関連度合の大きさの順序にしたがって所定数の分類を選択する手段，キーワードごとに，それに関連する選択された分類の関連度合に基づいてキーワードを評価し，関連度合の低い分類のみが関連するキーワードを削除するキーワード評価手段，および削除されずに残ったキーワードのそれぞれについて，そのキーワードに関連の深い所定数の分類およびその分類の関連度合を対応させて記憶するキーワード／分類テーブルを作成する手段，を備えたキーワード／分類テーブル作成装置。
【請求項６】　上記キーワード評価手段は，２つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照してキーワードを評価し，分類間距離の大きい２つの分類が関連するキーワードを削除するものである，請求項５に記載のキーワード／分類テーブル作成装置。
【請求項７】　分類未付与文書に含まれる複数のキーワードを入力し，キーワードごとに，そのキーワードに関連の深い分類およびその分類の関連の深さを示す度合をあらかじめ記憶したキーワード／分類テーブルを参照して，入力されたキーワードに関連する分類の関連度合の合計値を分類ごとに算出し，この合計値の大きさの順序にしたがって付与すべき分類の候補を選択し，２つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照して，選択された複数の候補分類相互間の距離が妥当な範囲内にあるかどうかを検査し，妥当な範囲内にあれば候補分類を最終分類と決定する，自動分類付与方法。
【請求項８】　一文書について複数の分類からなる分類の組があらかじめ付与された複数の分類付与済文書のそれぞれについて，それらの文書に付与された分類の組を入力し，入力された分類の組に２つの分類が同時に含まれる程度に基づいて，２つの分類間の距離を，すべての分類の中から選択されたすべての組合せの分類対について求め，分類間距離テーブルを作成する，分類間距離テーブル作成方法。
【請求項９】　分類があらかじめ付与された複数の分類付与済文書のそれぞれについて，それらの文書に付与された分類とそれらの文書から抽出されたキーワードとを相互に関連させて入力し，入力されたキーワードごとに，それらのキーワードに関連する分類の関連度合を求め，関連度合の大きさの順序にしたがって所定数の分類を選択し，キーワードごとに，それに関連する選択された分類の関連度合に基づいてキーワードを評価し，関連度合の低い分類のみが関連するキーワードを削除するとともに，２つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照してキーワードを評価し，分類間距離の大きい２つの分類が関連するキーワードを削除し，削除されずに残ったキーワードのそれぞれについて，そのキーワードに関連の深い所定数の分類およびその分類の関連度合を対応させて記憶するキーワード／分類テーブルを作成する，キーワード／分類テーブル作成方法。
【発明の詳細な説明】
【０００１】
【技術分野】この発明は，学術論文，科学記事，特許公報とその抄録，その他の文書を複数のカテゴリーに分類する自動分類付与装置および方法，ならびにこの自動分類付与のために用いる分類間距離テーブルやキーワード／分類テーブルの作成装置および方法に関する。
【０００２】
【従来技術とその問題点】従来の自動分類付与装置には，特開平１－188934号公報や特開平２－98778 号公報に記載のものがある。これらの装置は電子化された文書からキーワードを抽出しキーワードの頻度だけで分類を決定したり，あらかじめ作成された生成規則を用いるものである。しかしながら頻度だけに基づいたのでは全く的はずれな分類を付与してしまう可能性があり，また生成規則を用いるものでもあらかじめ人間が生成規則辞書を作成しないといけないという煩わしさがあった。
【０００３】
【発明の開示】この発明は，既に分類が付与された文書に基づいて自動分類付与のためのデータ・ベースを作成し，このデータ・ベースに基づいて適切な分類を付与することのできる装置および方法を提供するものである。
【０００４】この発明による自動分類付与装置は，分類未付与文書に含まれる複数のキーワードを入力する手段，キーワードごとに，そのキーワードに関連の深い分類およびその分類の関連の深さを示す度合をあらかじめ記憶したキーワード／分類テーブルを参照して，入力されたキーワードに関連する分類の関連度合の合計値を分類ごとに算出し，この合計値の大きさの順序にしたがって付与すべき分類の候補を選択する手段，ならびに２つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照して，選択された複数の候補分類相互間の距離が妥当な範囲内にあるかどうかを検査し，妥当な範囲内にあれば候補分類を最終分類と決定する手段を備えている。
【０００５】この発明の実施態様においては，上記決定手段は，上記合計値が所定値以上である候補分類が１つである場合に，上記分類間距離テーブルを参照することなくその候補分類を最終分類と決定する。
【０００６】この発明の他の好ましい実施態様においては，上記合計値が所定値よりも大きい候補分類がない場

合に，上記合計値の大きさの順序にしたがって複数の分類からなる分類パターンを作成し，同一分類パターンが所定回数出現したときに新たな分類を創設して付与する手段がさらに設けられる。

【０００７】この発明による自動分類付与方法は，分類未付与文書に含まれる複数のキーワードを入力し，キーワードごとに，そのキーワードに関連の深い分類およびその分類の関連の深さを示す度合をあらかじめ記憶したキーワード／分類テーブルを参照して，入力されたキーワードに関連する分類の関連度合の合計値を分類ごとに算出し，この合計値の大きさの順序にしたがって付与すべき分類の候補を選択し，２つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照して，選択された複数の候補分類相互間の距離が妥当な範囲内にあるかどうかを検査し，妥当な範囲内にあれば候補分類を最終分類と決定するものである。

【０００８】この発明は上述した自動分類付与装置および方法で用いられる分類間距離テーブルを作成する装置および方法を提供している。

【０００９】この発明による分類間距離テーブル作成装置は，一文書について複数の分類からなる分類の組があらかじめ付与された複数の分類付与済文書のそれぞれについて，それらの文書に付与された分類の組を入力するための手段，および入力された分類の組に２つの分類が同時に含まれる程度に基づいて，２つの分類間の距離を，すべての分類の中から選択されたすべての組合せの分類対について求め，分類間距離テーブルを作成する手段を備えている。

【００１０】この発明による分類間距離テーブル作成方法は，一文書について複数の分類からなる分類の組があらかじめ付与された複数の分類付与済文書のそれぞれについて，それらの文書に付与された分類の組を入力し，入力された分類の組に２つの分類が同時に含まれる程度に基づいて，２つの分類間の距離を，すべての分類の中から選択されたすべての組合せの分類対について求め，分類間距離テーブルを作成するものである。

【００１１】この発明はさらに上記自動分類付与装置および方法で用いるキーワード／分類テーブル作成装置および方法を提供している。

【００１２】この発明によるキーワード／分類テーブル作成装置は，分類があらかじめ付与された複数の分類付与済文書のそれぞれについて，それらの文書に付与された分類とそれらの文書から抽出されたキーワードとを相互に関連させて入力するための手段，入力されたキーワードごとに，それらのキーワードに関連する分類の関連度合を求め，関連度合の大きさの順序にしたがって所定数の分類を選択する手段，キーワードごとに，それに関連する選択された分類の関連度合に基づいてキーワードを評価し，関連度合の低い分類のみが関連するキーワードを削除するキーワード評価手段，および削除されずに残ったキーワードのそれぞれについて，そのキーワードに関連の深い所定数の分類およびその分類の関連度合を対応させて記憶するキーワード／分類テーブルを作成する手段を備えている。

【００１３】好ましい実施態様においては，上記キーワード評価手段は，２つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照してキーワードを評価し，分類間距離の大きい２つの分類が関連するキーワードを削除するものである。

【００１４】この発明によるキーワード／分類テーブル作成方法は，分類があらかじめ付与された複数の分類付与済文書のそれぞれについて，それらの文書に付与された分類とそれらの文書から抽出されたキーワードとを相互に関連させて入力し，入力されたキーワードごとに，それらのキーワードに関連する分類の関連度合を求め，関連度合の大きさの順序にしたがって所定数の分類を選択し，キーワードごとに，それに関連する選択された分類の関連度合に基づいてキーワードを評価し，関連度合の低い分類のみが関連するキーワードを削除するとともに，２つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照してキーワードを評価し，分類間距離の大きい２つの分類が関連するキーワードを削除し，削除されずに残ったキーワードのそれぞれについて，そのキーワードに関連の深い所定数の分類およびその分類の関連度合を対応させて記憶するキーワード／分類テーブルを作成するものである。

【００１５】分類付与済文書への分類の付与は，一般に専門家によって行なわれるであろう。上記分類間距離は後述する実施例では分類間の技術距離として具体化されている。

【００１６】この発明によると，あらかじめ分類が付与された分類付与済文書における分類とキーワードのデータを用いて，自動分類付与のためのデータ・ベースとなる分類間距離テーブルおよびキーワード／分類テーブルが作成されている。既存の分類付与済文書に基づいてデータ・ベースが作成されるので，上述した従来例のように人間が生成規則辞書を作成する煩わしさがなくなる。また，データ・ベースの作成のためにより多くの情報（分類付与済文書のデータ）を与えれば与えるほどより正確な分類間距離テーブルおよびキーワード／分類テーブルが得られる。すなわち，この発明は学習機能をもっており，この学習により，より正確な分類の自動付与が達成される。さらに，この発明では分類間距離という概念を導入してこの分類間距離をキーワード／分類テーブルの作成および自動分類付与処理に利用しているので，妥当でない分類の付与を排除してより正しい分類の付与が可能となる。

【００１７】

【実施例の説明】図１は自動分類付与装置の電気的構成の概要を示している。

【００１８】自動分類付与装置は最も好ましい形態においてはコンピュータ・システム10を含み，このコンピュータ・システム10には入力装置11，出力装置12，内部メモリ13および外部メモリ14が接続される。入力装置11は後述する分類付与済文書の分類（コード），キーワード等を入力するとともに，分類未付与文書に記載された文章を入力するものであり，キーボード，マウス，イメージ・リーダ等を含む。分類未付与文書はキーボードから入力してもよいし，イメージ・リーダによって読込んだドット・データを文字コードに変換する文字認識処理により入力を達成することもできる。出力装置12は主に分類結果を出力するものであり，ＣＲＴ表示装置やプリンタを含む。分類結果は好ましくは文書の分類欄にプリンタによって印字される。内部メモリ13はコンピュータ・システム10のプログラムを格納するとともに各種処理のためのワーク・エリア（後述する各種テーブルの作成等）を含む。外部メモリ14は入力された文書データ，分類データ等を記憶する。プログラムを外部メモリ14に格納しておいてもよい。

【００１９】自動分類付与装置はあらかじめ分類が付与された複数の文書（分類付与済文書）における分類とキーワードに関するデータに基づいて分類のための基礎データを作成し，この基礎データを用いて分類未付与文書にその記載内容に適した分類を付与するものである。分類付与処理のための基礎データとしては，分類間の技術距離テーブル（図４）とキーワード／分類テーブル（図８）とがある。したがって，自動分類付与装置は，分類付与処理（図11および図14～図16）に先だって，分類間の技術距離テーブル作成処理（図３）およびキーワード／分類テーブルの作成処理（図７）を実行する。

【００２０】ここで文書とは文字で記載された内容が情報としての意味をもつすべての文書を含む。もちろん文書は人間が読むことができる形態で表わされていても，マシン・リーダブルな形態で表わされていてもよい。最も典型的な文書には技術文書があろう。中でも，特許公報，その抄録のような特許文献が最もなじみが深いものかも知れない。分類とはこのような文書を大系化して整理するためにその内容に応じて文書をグループ分けするのに用いる記号である。分類は大分類，中分類，小分類のようにヒエラルキー構造とすることもできよう。最も身近な分類には特許関係分書に付与されるＩＰＣ（国際特許分類），各企業で付与する社内分類等があろう。文書に記載された内容の輪郭を端的に表現する用語はキーワードと呼ばれている。キーワードは一般的には文書の中で用いられる用語の中から抽出される。特許文献や学術論文ではキーワードが特定の欄に羅列して表わされている。

【００２１】図２は分類付与済文書の一例を示している。

【００２２】分類付与済文書には，文書を識別するための文書番号が付与されている。また文書に付与された分類を記載する分類欄と文書から抽出されたキーワードを記載するキーワード欄が設けられている。この実施例では一つの文書に最大３種類の分類が付与されるものとする。この明細書では分類（コード）をＡ，Ｂ，Ｃ，Ｄ，Ｅ，Ｆ，ＧおよびＨと表現する。またキーワードをａ，ｂ，ｃ，ｄ，ｅ，ｆ，ｇ，…等の小文字のアルファベットで表わす。一般には専門家によって分類が付与された文書が分類付与済文書となろう。

【００２３】まず図３から図６を参照して分類間の技術距離テーブル作成処理について説明する。

【００２４】あらかじめ用意された分類付与済文書の分類欄に記載されている分類の組（一文書について最大３種類の分類からなる）が文書ごとに入力される（ステップ21）。一つの文書について分類の組が入力されるとＰ（Ｉ，Ｊ）テーブルおよびＱ（Ｉ，Ｊ）テーブルのデータが更新される（ステップ22）。

【００２５】Ｐ（Ｉ，Ｊ）テーブルは，図５に示すように，入力された分類の組の中で分類ＩまたはＪが含まれる分類の組の数を，分類ＩとＪのすべての組合せＰ（Ｉ，Ｊ）（Ｉ≠Ｊ，Ｉ，Ｊ＝Ａ～Ｈ）について記憶するものである。Ｑ（Ｉ，Ｊ）テーブルは，図６に示すように，入力された分類の組の中で分類ＩおよびＪがともに含まれる分類の組の数を，分類ＩとＪのすべての組合せＱ（Ｉ，Ｊ）（Ｉ≠Ｊ，Ｉ，Ｊ＝Ａ～Ｈ）について記憶するものである。

【００２６】分類付与済文書のすべてについて，その分類欄に記載されている分類の組の入力と，Ｐ（Ｉ，Ｊ）テーブルおよびＱ（Ｉ，Ｊ）テーブルのデータの更新が繰返して実行される（ステップ23）。これによりＰ（Ｉ，Ｊ）テーブルとＱ（Ｉ，Ｊ）テーブルとが完成する。

【００２７】分類Ｉと分類Ｊとの技術距離Ｌ（Ｉ，Ｊ）は，たとえば次式にしたがって算出される。

【００２８】

【数１】

$$L(I,J)=100-[Q(I,J)/P(I,J)]\times\alpha \quad \cdots\text{式}1$$

αは定数である。

【００２９】技術距離Ｌ（Ｉ，Ｊ）は０～100 の間の値をとる。

【００３０】分類Ｉと分類Ｊのすべての組合せについて式１にしたがって技術距離Ｌ（Ｉ，Ｊ）が算出され，図４に示すような分類間の技術距離テーブルが作成される。

【００３１】技術距離Ｌ（Ｉ，Ｊ）は，文書が技術文書である場合に，それらに付与される分類間の技術上の関連性の近さ，または遠さを表わしている。技術距離が大きければ２つの分類間の関連性が小さく，小さければ大きい。

【００３２】技術距離を一般文書についての分類間距離という概念に敷衍することができる。分類間距離は２つの分類間の関連性の近さまたは遠さを表わす。分類間距離または分類間の技術距離は式１のみならず他の演算式によっても定義することができよう。

【００３３】分類付与済文書が10枚あったとして，それらに付与された分類の組が次の10個であったと仮定する。

【００３４】（Ａ，Ｂ，Ｃ），（Ａ，Ｂ，Ｄ），（Ａ，Ｅ，Ｆ），（Ｂ，Ｆ，Ｇ），（Ｂ，Ｆ，Ｇ），（Ｃ，Ｄ，Ｅ），（Ｃ，Ｇ，Ｈ），（Ｃ，Ｇ，Ｈ），（Ｄ，Ｅ，Ｆ），（Ｄ，Ｇ，Ｈ）

【００３５】この場合に，Ｐ（Ａ，Ｂ）＝５，Ｑ（Ａ，Ｂ）＝２となる。α＝100 とすると，分類ＡとＢとの技術距離Ｌ（Ａ，Ｂ）は式１にしたがうと，

【数２】

Ｌ（Ａ，Ｂ）＝100 －（２／５）×100 ＝60　‥式２

となる。この値Ｌ（Ａ，Ｂ）＝60は単純化した一例であるから図４に示すものとは異なっている。

【００３６】続いて図７から図10を参照して，キーワード／分類テーブルの作成処理について説明する。

【００３７】あらかじめ用意されたすべての分類付与済文書に記載されている分類（最大３種類の分類）およびキーワードが，文書ごとに入力される（ステップ31）。後に示す自動分類付与処理と同じように，文書も入力して，入力された文書からキーワードを抽出するようにしてもよい。

【００３８】分類付与済文書についての分類とキーワードの入力ごとに図９に示すようなキーワード別分類頻度テーブルにおける度数（頻度）が加算される。たとえば，一文書について分類Ａ，ＢおよびＤとキーワードａ，ｂ，ｃ，ｅおよびｈが入力されたときには，キーワードａ，ｂ，ｃ，ｅおよびｈのそれぞれについて分類Ａ，ＢおよびＤの度数が＋１される。すべての分類付与済文書についての分類とキーワードの入力が終了すると，キーワード別分類頻度テーブルが完成し，このテーブルに基づいて図10に示すようなキーワード別分類ヒストグラムがキーワードごとに作成される（ステップ32）。

【００３９】このキーワード別分類頻度テーブルまたはキーワード別分類ヒストグラムは，キーワードごとに，そのキーワードと関連性がある分類についてその関連性（関係）の深さまたは強さを表わす度数から構成されている。度数はキーワードと分類との関係の深さまたは強さを表わしており，度数が大きいほど関係が強いといえる。たとえば，図10を参照して，キーワードａに最も関係が強い分類はＡであり，次に分類Ｂが関係が強く，第３番目は分類Ｄである。

【００４０】このようなキーワード別分類頻度テーブルまたはキーワード別分類ヒストグラムに基づいてキーワードの評価処理（その１）が行なわれる（ステップ33）。キーワードは特定の分類（できるだけ少数の分類）に強く関係している方が後に示す自動分類付与処理に役立つ。逆に言えば，強く関係している特定の分類が無く多くの分類に同程度に弱く関係しているキーワードは，分類付与処理のためのキーワードとして役に立たない。そこで，１または２，３程度の少数の特定の分類に関係しているとは言い切れない役に立ちそうもないキーワードを削除するのがこのキーワード評価処理（その１）である。

【００４１】一つのキーワードについて度数の大きいものから所定数（この実施例では３個）の分類を抽出し，その分類についての度数の和が求められ，これが所定数βよりも小さいかどうかがチェックされる。たとえば，度数の高いものからｎ番目の分類コードの度数をδ(n)とすると（ここでｎはキーワードを表わす符号とは異なり一般的な番号を表わす），

【数３】δ(1)＋δ(2)＋δ(3)＜β　‥式３

βはたとえば50

を満たすキーワードが削除される。

【００４２】上述したキーワードａについては，度数の高い３種類の分類Ａ，Ｂ，Ｄについての度数はそれぞれ80，70，10であり，これらの和は160 であるから，キーワードａは削除されない。

【００４３】続いて，既に作成された分類間の技術距離テーブルを参照したキーワードの評価処理（その２）が行なわれる（ステップ34）。

【００４４】キーワード評価処理（その１）において削除されなかったキーワードには度数の高い３種類の分類が対応しているが，これらの３種類の分類の中に相互の関連性がきわめて低い分類対が含まれている場合には，キーワードと３種類の分類との関連性に疑問があると考えられるので，このようなキーワードが削除される。

【００４５】このキーワード評価処理（その２）においては，あるキーワードについて度数の大きい３種類の分類をＩ，Ｊ，Ｋとすると，これらの３種類の分類から選択された１対の分類間の技術距離Ｌ（Ｉ，Ｊ），Ｌ（Ｉ，Ｋ），Ｌ（Ｊ，Ｋ）のうち１つでもしきい値γよりも大きいものがあれば，そのキーワードは削除される。すなわち，

【数４】｛Ｌ（Ｉ，Ｊ）＞γ｝or

｛Ｌ（Ｉ，Ｋ）＞γ｝or

｛Ｌ（Ｊ，Ｋ）＞γ｝＝真　‥式４

であればそのキーワードは削除される。

【００４６】たとえばキーワードａについては，図４の技術距離テーブルを参照すると，

Ｌ（Ａ，Ｂ）＝10

Ｌ（Ａ，Ｄ）＝14

Ｌ（Ｂ，Ｄ）＝30

であり，γ＝40とすると，式４の条件を満たさないので

削除されない。

【００４７】このようにして２種類のキーワード評価処理（その１）（その２）において削除されずに残ったキーワードのそれぞれについて，そのキーワードに関係する度数の最も高い分類から３番目に高い分類までの重要な３種類の分類とその度数とが対応づけられることにより，図８に示すようなキーワード／分類テーブルが作成される（ステップ35）。たとえば，キーワードａについては，分類Ａ（度数80）と分類Ｂ（度数70）と分類Ｄ（度数10）とが正しく関係するものとして対応づけられる。

【００４８】図11は自動分類付与処理の概要を示している。

【００４９】分類未付与文書に記載された文章が入力される（ステップ41）。上述したように，文書の文章はキーボードから入力されてもよいし，イメージ・リーダから入力されてもよい。または，あらかじめ外部メモリ14に格納しておいてこれを読出してもよい。いずれにしても入力された文章を構成する各文字を表わすコードの列がコンピュータ・システム10内に入力され，このコード列からキーワードを表わすコード列が抽出される（ステップ42）。入力された文章からキーワードを抽出する処理は公知であり，たとえば文章が分かち書きされ，助詞などの不要語が除かれることにより単語（主に名詞，動詞が含まれてもよい）が抽出される。この単語がここではキーワードとなる。したがって，先に説明したキーワード／分類テーブルに登録されていない単語（キーワード）が抽出されても問題は無い。キーワードの抽出処理の進行にともなって抽出されたキーワードは，図12に示すようなキーワード・リストに登録される（ステップ43）。

【００５０】このようにして，入力された文章からキーワードの抽出処理，抽出されたキーワードのリストの作成が終了すると，キーワード・リストに挙げられているキーワードのそれぞれについて，リストの順番に，キーワード／分類テーブルに登録されているかどうかが調べられ，登録されていればそのキーワードに対応する分類と度数が読取られ，キーワードごとに図13に示すような度数加算表に書加えられる。また，文類ごとに度数が加算される（ステップ44）。抽出されたキーワードがキーワード／分類テーブルに登録されていなければそのキーワードについては何らの処理も行なわれない。度数加算表はキーワードごとに，そのキーワードにキーワード／分類テーブルにおいて対応する分類についてその度数を記憶するとともに，分類ごとにその度数の合計を記憶するものである。

【００５１】このようにして作成された度数加算表を用いて，また必要に応じて先に作成された分類間の技術距離テーブルを参照して分類決定処理が行なわれる（ステップ45）。

【００５２】この分類決定処理において次の４種類の結論が得られる。

【００５３】(1) 文書への既存の分類（コード）の付与（最大３種類の分類）
(2) 新しい分類（コード）の付与
(3) 検討中であることを示すコードの付与
(4) 分類不可能であることを示すコードの付与

【００５４】図14から図16は分類決定処理（ステップ45）の詳細を示すものである。

【００５５】まず図14において，先にステップ44で作成された度数加算表における分類ごとの合計度数を用いてヒストグラムが作成され，このヒストグラムが正規化される（ステップ51）。

【００５６】図13に示す度数加算表に基づいて作成されたヒストグラムが図17に示されている。このようなヒストグラムの正規化は次式にしたがって行なわれる。

【００５７】分類Ｉの正規化された度数をＤ(Ｉ）とする。

【００５８】

【数５】

$$D(I) = (\text{分類Iの度数}) \Big/ \left(\sum_{J=A}^{H} \text{分類Jの度数}\right) \quad \cdots \text{式5}$$

【００５９】正規化されたヒストグラムが図18に示されている。この正規化されたヒストグラムに基づいて分類の付与か行なわれる。

【００６０】まず，正規化されたヒストグラムにおいて，度数の最も高い分類の度数が所定のしきい値ＴＨ１を越えているかどうかがチェックされる（ステップ52）。図18に示すヒストグラムにおいては分類Ｄの度数がしきい値ＴＨ１を越えている。

【００６１】このステップ52における判断がYES であれば次に，しきい値ＴＨ１を越えた度数をもつ分類が１つのみであるかどうかがチェックされる（ステップ53）。

【００６２】しきい値ＴＨ１を越えた度数をもつ分類が１つのみであればその分類が付与されることになる（ステップ54）。図18に示すヒストグラムではしきい値ＴＨ１を越える度数をもつ分類は分類Ｄのみであるから，このヒストグラムを生じさせた文書には１つの分類Ｄのみが付与される。

【００６３】分類の付与は上述したように文書の分類欄に，付与されるべき分類を表わす符号もしくは記号またはコードをプリンタによって印字することによって，または文書番号に対応して分類を表示，プリント・アウトもしくはメモリに記憶することによって行なわれよう。

【００６４】度数がしきい値ＴＨ１を越えた分類が１つだけでない場合には，度数がしきい値ＴＨ１を越えた分

類が２つかどうかがチェックされる（ステップ55）。

【００６５】度数がしきい値ＴＨ１を越えた分類が２つの場合には，これらの２つの分類間の技術距離が技術距離テーブル（図４）を参照して求められ（ステップ56），求められた技術距離が所定値よりも小さいかどうかが判定される（ステップ57）。

【００６６】２つの分類間の技術距離が所定値よりも小さければ，これらの２つの分類は技術的な観点からいって比較的近いから，これらの２つの分類は妥当とみなされ，その２つの分類が該当文書に付与されることになる（ステップ58）。

【００６７】２つの分類間の技術距離が所定値よりも大きい場合には，これらの２つの分類は比較的遠く，何らかの誤りを含んでいる可能性があるので分類不可能の旨が出力される（ステップ59）。この出力は，文書番号と分類不可能の旨を示す記号またはコードの表示，プリント・アウトもしくは記憶，または該当文書の分類欄への分類不可能の旨の印字によって達成される。

【００６８】正規化されたヒストグラムにおいて度数がしきい値ＴＨ１を越える分類が３つ以上ある場合には，図15を参照して，これらの３つ以上の分類の中から任意の３つの分類を選択して一つの組を構成する。そして，各組ごとにその組に含まれる分類の技術距離の合計を技術距離テーブルを参照して算出する（ステップ60）。

【００６９】たとえば，度数がしきい値ＴＨ１を越える分類がＡ，Ｃ，Ｄ，Ｆ，Ｇの５種類あったと仮定する。この５種類の分類の中から任意の３種類の分類が選ばれ組が構成される。生成される組は，（Ａ，Ｃ，Ｄ），（Ａ，Ｃ，Ｆ），（Ａ，Ｃ，Ｇ），（Ａ，Ｄ，Ｆ），（Ａ，Ｄ，Ｇ），（Ａ，Ｆ，Ｇ），（Ｃ，Ｄ，Ｆ），（Ｃ，Ｄ，Ｇ），（Ｃ，Ｆ，Ｇ），（Ｄ，Ｆ，Ｇ）の10組である。

【００７０】組（Ａ，Ｃ，Ｄ）の技術距離の合計Ｌ（Ａ，Ｃ，Ｄ）は次式で求められる。

【００７１】

【数６】

$$L(A, C, D) = L(A, C) + L(C, D) + L(D, A) \quad \cdots \text{式}6$$

【００７２】他のすべての組についても同じように技術距離の合計が算出される。

【００７３】続いて，このようにして算出された技術距離の合計がある所定値と比較され，その所定値よりも小さい組があるかどうかがチェックされる（ステップ61）。技術距離の合計があまりに大きいということは，その組に含まれる分類の中に関連性の薄いものが含まれている可能性があるので，そのような分類の組を排除するためである。

【００７４】技術距離の合計が所定値よりも小さい組が一つでもあれば，その中で技術距離の合計が最も小さい組が選択され，その組に含まれる３種類の分類が妥当なものとして該当文書に付与される（ステップ62）。

【００７５】度数がしきい値ＴＨ１を越える分類が３つの場合にはその３つの分類についての技術距離の合計が算出され，この合計が所定値よりも小さければその３つの分類が付与されることになるのはいうまでもない。

【００７６】技術距離の合計が所定値よりも小さい組がない場合には，分類付与不可能の旨が出力される（ステップ63）。

【００７７】正規化されたヒストグラムにおいて，度数がしきい値ＴＨ１を越える分類が存在しない場合には（ステップ52でNO），まだ定義されていない新しい分類に振分けられる文書である可能性がある。図19は，度数がしきい値ＴＨ１を越えるものが存在しない場合の正規化されたヒストグラムを示している。

【００７８】図16はこのような新分類の決定を含む処理を示すものである。

【００７９】図19に示すように第１のしきい値ＴＨ１よりも低い第２のしきい値ＴＨ２があらかじめ定められている。度数がこの第２のしきい値ＴＨ２を越える分類があるかどうかがチェックされる（ステップ64）。もし第２のしきい値ＴＨ２を越える度数をもつ分類が存在しなければ分類付与不可能ということになる（ステップ63）。

【００８０】しきい値ＴＨ２を越えた度数をもつ分類が一つでもあれば次にヒストグラム・パターン作成に移る（ステップ65）。図20に示すように，しきい値ＴＨ１とＴＨ２との間を等分し複数（この例では５個）のランクに分ける。度数の高い方からランク１，２，３，４，５となっている。しきい値ＴＨ２を越える度数をもつ分類のうち上位複数種類（この例では５種類）の分類が選ばれ，これらの分類がどのランクに属するかが判定され，この判定結果に基づいて図21に示すようなヒストグラム・パターンが作成される（ステップ65）。

【００８１】しきい値ＴＨ２を越える度数をもつ分類が５個以上無い場合にはしきい値ＴＨ２を越える度数をもつ分類のみでパターンを作成する。度数の高いものから合計５分類になるまで選択し，しきい値ＴＨ２以下のものにランク６を付与してまたはランクを付与せずにヒストグラム・パターンを作成してもよい。または分類不可能と判定してもよい。

【００８２】一方，図22に示すように新分類テーブルと未定分類テーブルとが設けられている。同一のヒストグラム・パターンをもつ文書の数が所定数に達したときにそのヒストグラム・パターンに新たな分類コードが付与され，この新たな分類コードが付与されたパターンが新分類コードとともに新分類テーブルに登録される。同一のヒストグラム・パターンをもつ文書の数が所定数に達しないヒストグラム・パターンがそのパターンをもつ文書の数（出現回数：カウント）とともに未定分類テーブルに登録されている。

【００８３】ステップ65で作成されたヒストグラム・パターンと同一のパターンが新分類テーブルにあるかどうかがチェックされ，もしあればそのパターンに与えられた新分類が付与されることになる（ステップ66，67，68）。

【００８４】新分類テーブルに同一パターンのものが存在しない場合には，作成されたヒストグラム・パターンは未定分類テーブルのパターンと比較される（ステップ69）。未定分類テーブルに同一のパターンがあればそのパターンのカウントが１つインクレメントされ（ステップ70，71），そのパターンのカウントが所定数に達したかどうかがチェックされる（ステップ72）。

【００８５】未定分類テーブルのあるパターンのカウントが所定数に達すると，そのパターンは新分類テーブルに移されかつそのパターンに新分類コードが割当てられ（ステップ73），そのパターンと同一のヒストグラム・パターンを生じさせた文書に新たに割当てられた新分類コードが付与される（ステップ74）。

【００８６】未定分類テーブルに同一パターンが存在しない場合には，作成されたパターンが未定分類テーブルに追加され，カウント１が与えられる（ステップ76）。この場合，およびステップ72において該当パターンのカウントが所定数に達しない場合には，その文書に検討中である旨のコードが付与される（ステップ75）。

【００８７】ヒストグラム・パターンを構成する分類の数は５個に限られず，ランクは必ずしも必要ではない。要するに，ヒストグラム・パターンが類似しているかどうかを判定できるものであればよい。

【図面の簡単な説明】

【図１】自動分類付与装置の構成を示すブロック図である。

【図２】分類付与済文書の例を示す。

【図３】分類間の技術距離テーブル作成処理を示すフロー・チャートである。

【図４】分類間の技術距離テーブルを示す。

【図５】Ｐ（Ｉ，Ｊ）テーブルを示す。

【図６】Ｑ（Ｉ，Ｊ）テーブルを示す。

【図７】キーワード／分類テーブルの作成処理を示す。

【図８】キーワード／分類テーブルを示す。

【図９】キーワード別分類頻度テーブルを示す。

【図１０】キーワード別分類ヒストグラムを示す。

【図１１】自動分類付与処理の概要を示すフロー・チャートである。

【図１２】キーワード・リストを示す。

【図１３】度数加算表を示す。

【図１４】分類決定処理を示すフロー・チャートである。

【図１５】分類決定処理を示すフロー・チャートである。

【図１６】分類決定処理を示すフロー・チャートである。

【図１７】度数加算表から作成されるヒストグラムを示す。

【図１８】正規化されたヒストグラムを示す。

【図１９】正規化されたヒストグラムを示す。

【図２０】ヒストグラム・パターンの作成の様子を示す。

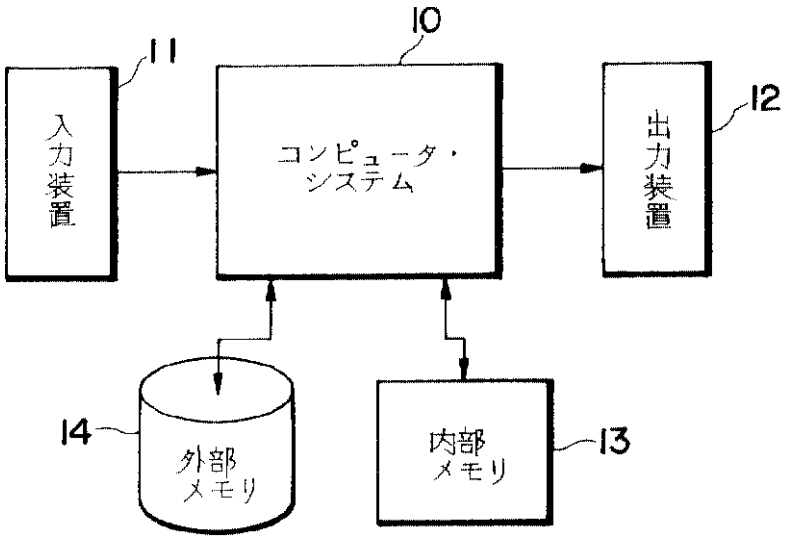
【図２１】ヒストグラム・パターンを示す。

【図２２】新分類テーブルと未定分類テーブルを示す。

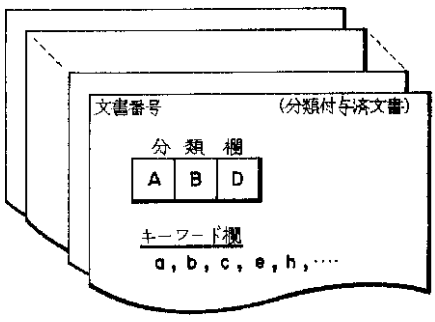
【符号の説明】

10　コンピュータ・システム
11　入力装置
12　出力装置
13　内部メモリ
14　外部メモリ

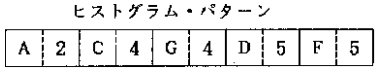
【図１】



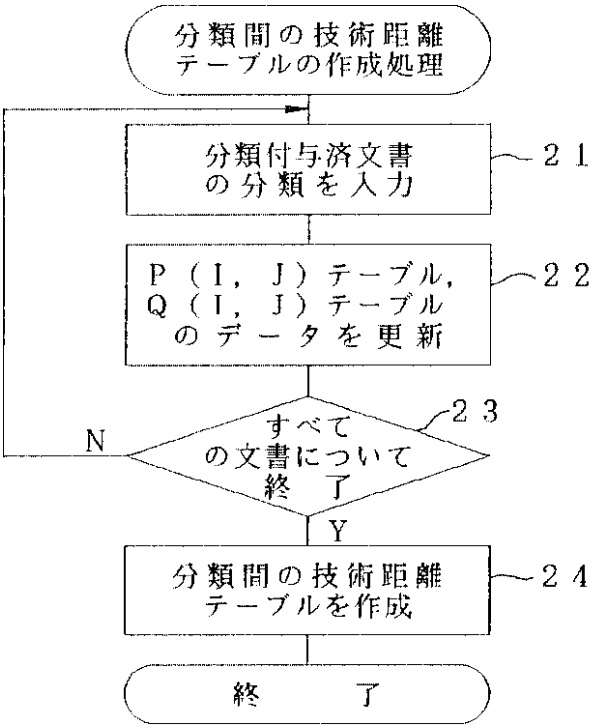

【図２】




【図２１】



【図３】




【図４】

分類間の技術距離テーブル

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| A | | 10 | 15 | 14 | ---------- | | | 80 |
| B | 10 | | 20 | 30 | | | | 90 |
| C | 15 | 20 | | 50 | | | | |
| D | 14 | 30 | 50 | | | | | |
| E | | | | | | | | |
| F | | | | | | | 10 | 20 |
| G | | | | | | 10 | | 30 |
| H | 80 | 90 | ---------- | | | 20 | 30 | |

【図５】

P（I，J）テーブル

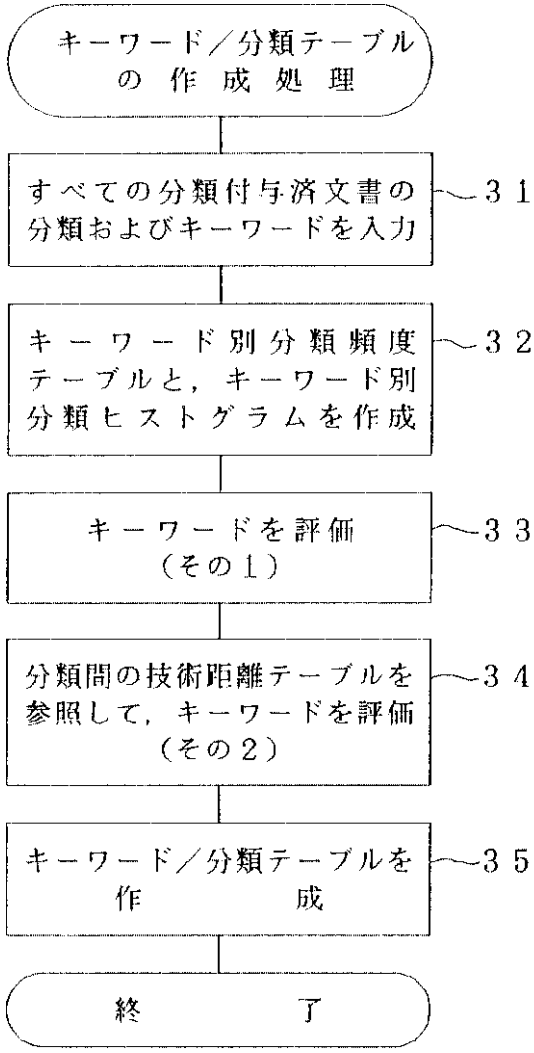
| |
|---|
| P（A，B）の数 |
| P（A，C）の数 |
| P（A，D）の数 |
| ⋮ |
| P（B，C）の数 |
| P（B，D）の数 |
| ⋮ |
| P（G，H）の数 |

【図６】

Q（I，J）テーブル

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | |
| B | | | | | | | | |
| C | | | | | | | | |
| D | | | | | Q（I，J）の数 | | | |
| E | | | | | | | | |
| F | | | | | | | | |
| G | | | | | | | | |
| H | | | | | | | | |

【図７】




【図８】

キーワード／分類テーブル

| | | 主要三分類とその度数 | | | | | |
|---|---|---|---|---|---|---|---|
| | | 分類 | 度数 | 分類 | 度数 | 分類 | 度数 |
| キーワード | a | A | 80 | B | 70 | D | 10 |
| | b | A | 60 | C | 30 | E | 50 |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | z | E | 40 | G | 70 | H | 60 |

【図１２】

キーワード・リスト

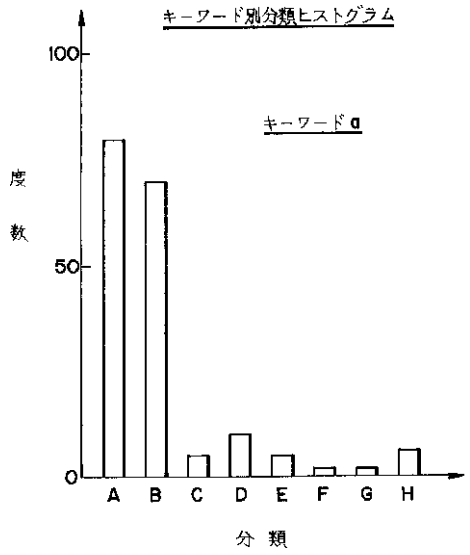
| |
|---|
| キーワードａ |
| キーワードｃ |
| キーワードｆ |
| ⋮ |

【図９】

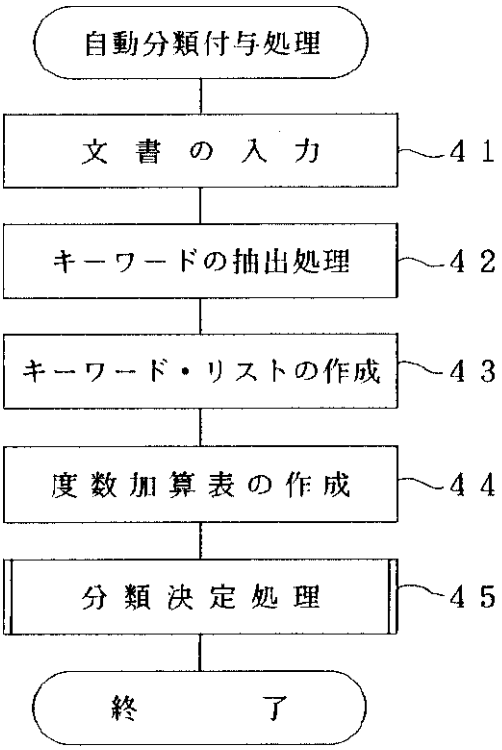
キーワード別分類頻度テーブル

| キーワード | 分類 A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| a | 80 | 70 | 5 | 10 | 5 | 2 | 2 | 6 |
| b | | | | | | | | |
| c | | | | | | | | |
| d | | | | | | | | |
| e | | | | | | | | |
| f | | | | | | | | |

【図１０】




【図１１】

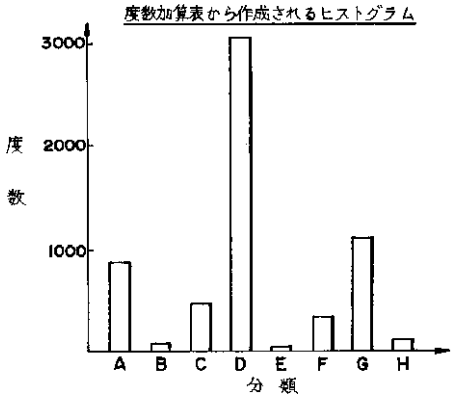



【図１３】

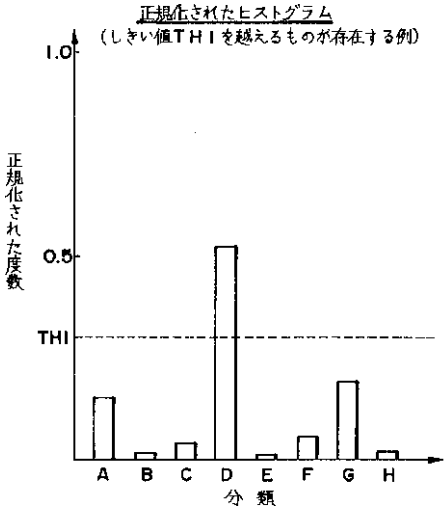
度数加算表

| キーワード＼分類 | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| キーワードa | 80 | 70 | | 10 | | | | |
| キーワードc | | | 10 | 50 | | | 50 | |
| キーワードf | 90 | | | 80 | | 10 | | |
| キーワードk | | | | | | | | |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 合計 | 880 | 70 | 250 | 3050 | 60 | 330 | 1090 | 100 |

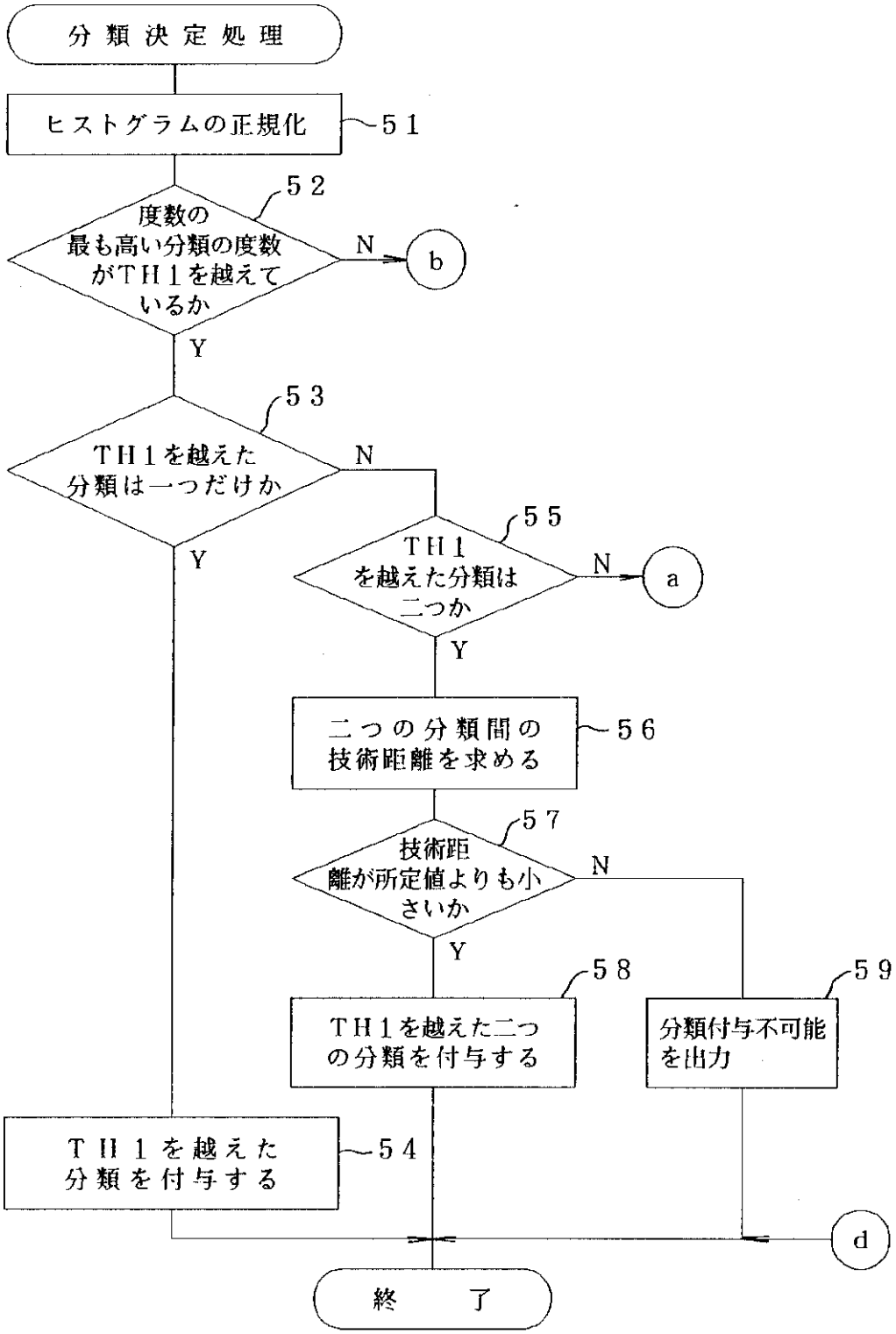
【図１７】



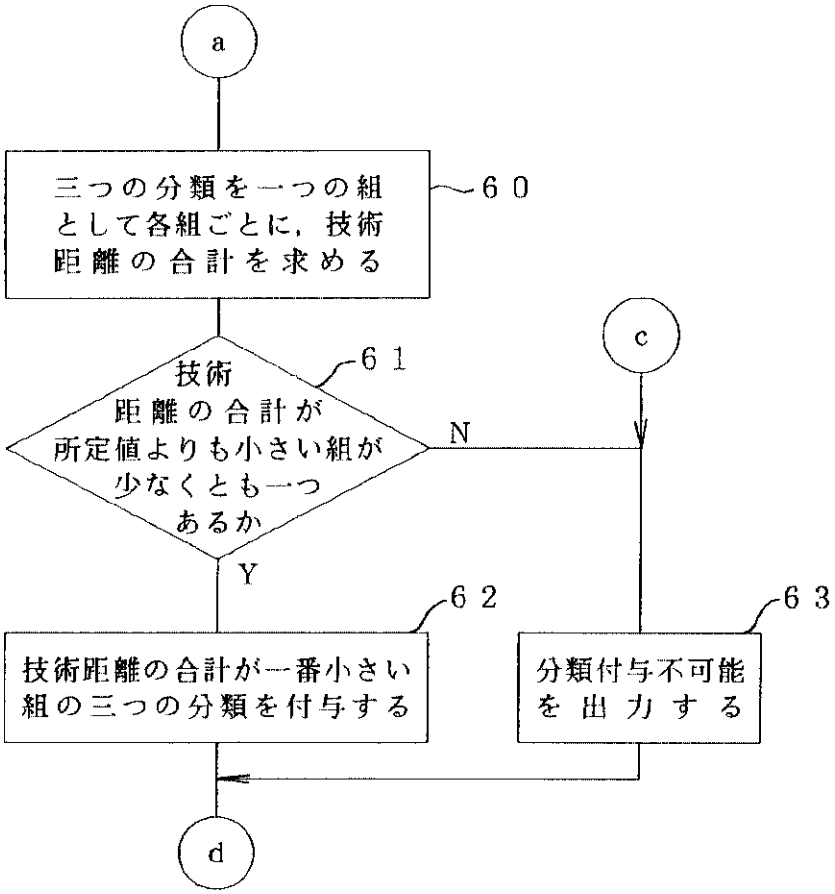

【図１８】



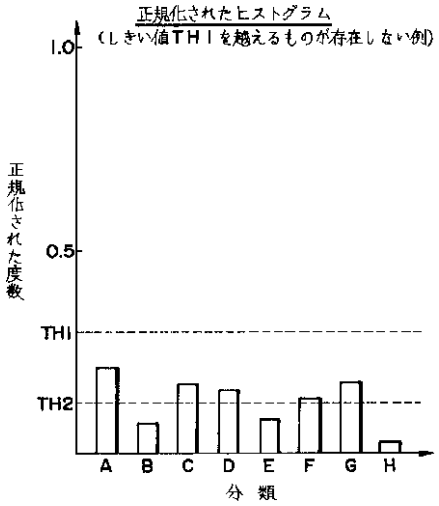
【図１４】


分類決定処理
ヒストグラムの正規化
51
52
度数の最も高い分類の度数がTH1を越えているか
N
b
Y
53
TH1を越えた分類は一つだけか
N
Y
55
TH1を越えた分類は二つか
N
a
Y
二つの分類間の技術距離を求める
56
57
技術距離が所定値よりも小さいか
N
Y
58
TH1を越えた二つの分類を付与する
59
分類付与不可能を出力
TH1を越えた分類を付与する
54
d
終　了

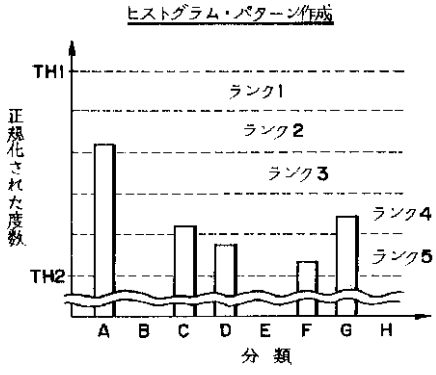
【図１５】




【図１９】



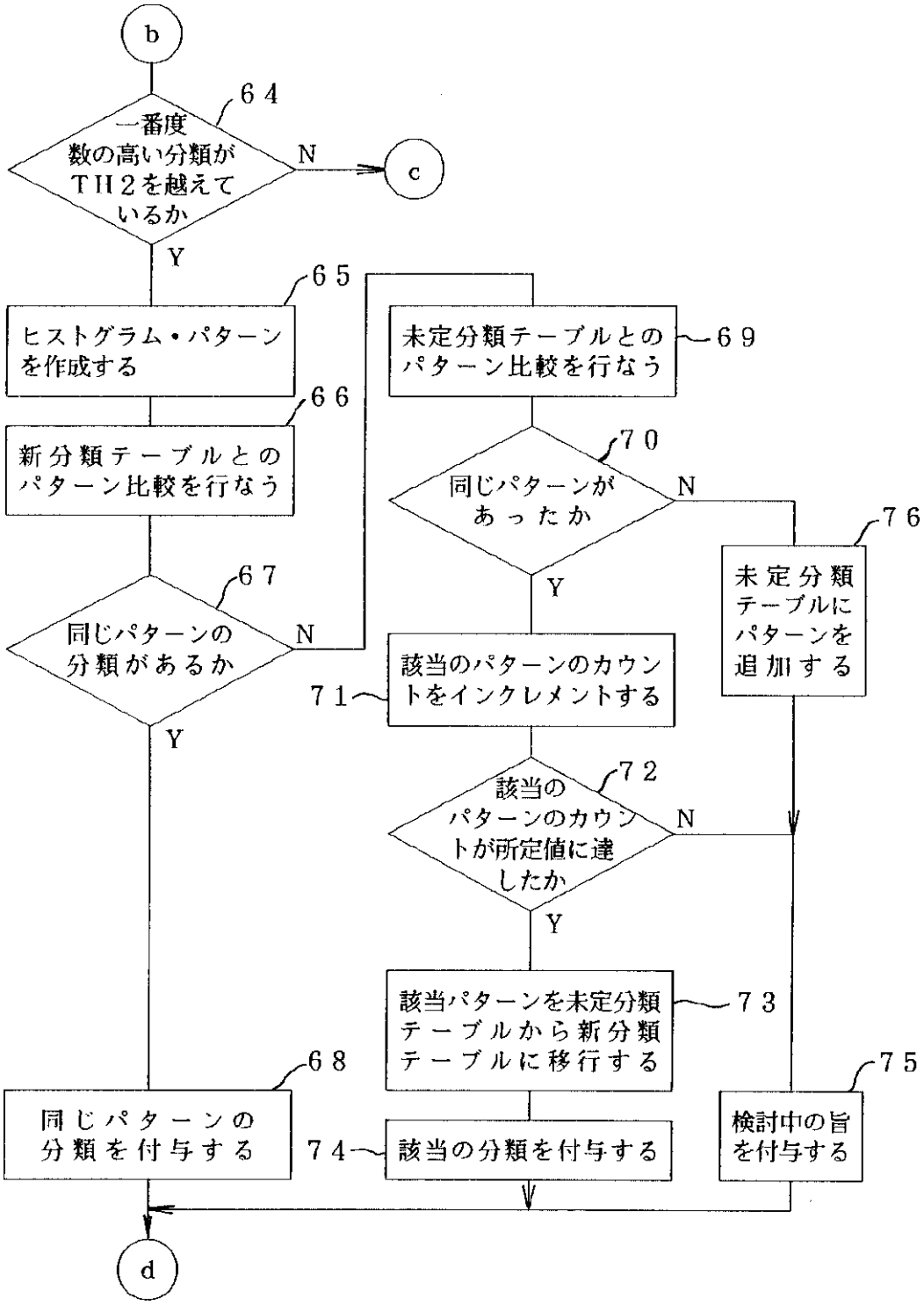

【図２０】




【図２２】

| NO. | ヒストグラム・パターン | | | | | | | | | | 新分類コード | 新分類テーブル |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | A | 1 | C | 3 | E | 3 | | | | | M | |
| 2 | B | 2 | D | 2 | F | 1 | G | 1 | | | N | |
| 3 | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| NO. | ヒストグラム・パターン | | | | | | | | | | カウント | 未定分類テーブル |
| 18 | A | 3 | C | 2 | D | 1 | F | 1 | G | 1 | | |
| 19 | C | 2 | D | 2 | E | 2 | F | 1 | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| | | | | | | | | | | | | |

【図１６】




フロントページの続き

(56)参考文献　特開　平２－105973（ＪＰ，Ａ）
特開　平３－78872（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.7，ＤＢ名)
G06F　17/30
ＪＩＣＳＴファイル（ＪＯＩＳ）