(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3703617号
(P3703617)
(45)発行日　平成17年10月5日(2005.10.5)　(24)登録日　平成17年7月29日(2005.7.29)

(51)Int.Cl.[7]　F I
G06F 17/30　G06F 17/30 210D
G06F 17/30 170A

請求項の数 18　(全 26 頁)

(21)出願番号　特願平10-14314
(22)出願日　平成10年1月27日(1998.1.27)
(65)公開番号　特開平11-212978
(43)公開日　平成11年8月6日(1999.8.6)
審査請求日　平成13年11月1日(2001.11.1)

(73)特許権者　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号
(74)代理人　100066474
弁理士　田澤　博昭
(74)代理人　100088605
弁理士　加藤　公延
(72)発明者　小中　裕喜
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内

審査官　辻本　泰隆

(56)参考文献　特開平９－２５１４６７（ＪＰ，Ａ）
特開平５－２３３７０７（ＪＰ，Ａ）
最終頁に続く

(54)【発明の名称】文書情報解析方法、文書情報解析装置および記録媒体

(57)【特許請求の範囲】
【請求項１】
初期パターン集合生成手段と、第１の選択手段と、第２の選択手段と、出力手段と、パターン集合生成手段を備えた文書情報解析装置において、
上記初期パターン集合生成手段が、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成するステップと、
上記第１の選択手段が、所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択するステップと、
上記第２の選択手段が、所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、前記所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択するステップと、
上記出力手段が、前記支持率が所定の最小支持率以上であり、かつ、前記確信度が所定の最小確信度以上であるパターンを前記所定のクラスタに関連づけて出力するステップと

、
上記パターン集合生成手段が、前記支持率が前記最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するステップと
を備えた文書情報解析方法。
【請求項２】
第２の選択手段による所定のパターンの確信度の計算において、前記所定のパターンが出現した文書情報の文書ＩＤを記憶し、
所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の文書ＩＤを読み出し、その文書情報においてだけ、前記（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントする
ことを特徴とする請求項１記載の文書情報解析方法。
【請求項３】
第２の選択手段による所定のパターンの確信度の計算において、前記所定のパターンが出現した文書情報の属するクラスタの情報を記憶し、
所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の属するクラスタの情報を読み出し、そのクラスタにおいてだけ、前記（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントする
ことを特徴とする請求項１記載の文書情報解析方法。
【請求項４】
初期パターン集合生成手段において、初期のパターン集合を生成する前に、少なくとも１つのクラスタにおいて支持率が所定の最小支持率以上となるキーワード以外のキーワードを、すべての文書情報から除去する
ことを特徴とする請求項１から請求項３のうちのいずれか１項記載の文書情報解析方法。
【請求項５】
第１の選択手段において、各クラスタについて、所定のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を前記元の文書情報の代わりに使用する
ことを特徴とする請求項１から請求項４のうちのいずれか１項記載の文書情報解析方法。
【請求項６】
所定のパターン集合は、初期のパターン集合である
ことを特徴とする請求項５記載の文書情報解析方法。
【請求項７】
第２の選択手段において、新たなパターン集合を生成する前に、確信度が所定の最大確信度以上である多支持パターンを多支持パターン集合から除去する
ことを特徴とする請求項１から請求項６のうちのいずれか１項記載の文書情報解析方法。
【請求項８】
第２の選択手段において、初期状態では空集合である高確信度パターン集合を記憶し、
支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが前記高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットのいずれかに該当する前記高確信

度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを、そのクラスタに関連づけて出力するとともに、前記高確信度パターン集合に追加する
ことを特徴とする請求項１から請求項７のうちのいずれか１項記載の文書情報解析方法。

【請求項９】
第２の選択手段において、多支持パターン集合の各パターンについて、その確信度とともにそのパターンのすべてのサブセットの確信度を計算し、
前記確信度が所定の最小確信度以上でかつサブセットの確信度の最大値より大きい前記パターンだけを出力する
ことを特徴とする請求項１から請求項７のうちのいずれか１項記載の文書情報解析方法。

【請求項１０】
出力手段において、クラスタ毎に一括して、所定の順番でパターンを出力する
ことを特徴とする請求項８または請求項９記載の文書情報解析方法。

【請求項１１】
初期パターン集合生成手段と、選択手段と、パターン集合生成手段と、出力手段を備えた文書情報解析装置において、
上記初期パターン集合生成手段が、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成するステップと、
上記選択手段が、所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加するステップと、
上記パターン集合生成手段が、前記多支持パターン集合において、前記多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するステップと、
上記出力手段が、前記多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、前記累積多支持パターン集合を構成する各多支持パターンについて、前記所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、前記累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力するステップと
を備えた文書情報解析方法。

【請求項１２】
初期パターン集合生成手段において、初期の多支持パターン集合を生成する前に、少なくとも１つのクラスタにおいて支持率が所定の最小支持率以上となるキーワード以外のキーワードを、すべての文書情報から除去する
ことを特徴とする請求項１１記載の文書情報解析方法。

【請求項１３】
選択手段において、累積多支持パターン集合を構成する各多支持パターンについて確信度を計算する前に、元の文書情報から、当該クラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を前記元の文書情報の代わりに使用する

ことを特徴とする請求項１１または請求項１２記載の文書情報解析方法。
【請求項１４】
出力手段において、クラスタ毎に一括して、所定の順番で、確信度が所定の最小確信度以上である多支持パターンを出力する
ことを特徴とする請求項１１から請求項１３のうちのいずれか１項記載の文書情報解析方法。
【請求項１５】
複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、
所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択する第１の選択手段と、
所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、前記所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択する第２の選択手段と、
前記支持率が所定の最小支持率以上であり、かつ、前記確信度が所定の最小確信度以上であるパターンを前記所定のクラスタに関連づけて出力する出力手段と、
前記支持率が前記最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するパターン集合生成手段と
を備えた文書情報解析装置。
【請求項１６】
複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、
所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加する選択手段と、
前記多支持パターン集合において、前記多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するパターン集合生成手段と、
前記多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、前記累積多支持パターン集合を構成する各多支持パターンについて、前記所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、前記累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以

上である多支持パターンを選択し、出力する出力手段と
を備えた文書情報解析装置。
【請求項１７】
初期パターン集合生成手段と、第１の選択手段と、第２の選択手段と、出力手段と、パターン集合生成手段を備えた文書情報解析装置において、
上記初期パターン集合生成手段に、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成させるステップ、
上記第１の選択手段に、所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択させるステップ、
上記第２の選択手段に、所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、前記所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択させるステップ、
上記出力手段に、前記支持率が所定の最小支持率以上であり、かつ、前記確信度が所定の最小確信度以上であるパターンを前記所定のクラスタに関連づけて出力させるステップ、
上記パターン集合生成手段に、前記支持率が前記最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成させるステップ
を有するプログラムを記録した記録媒体。
【請求項１８】
初期パターン集合生成手段と、選択手段と、パターン集合生成手段と、出力手段を備えた文書情報解析装置において、
上記初期パターン集合生成手段に、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成させるステップ、
上記選択手段に、所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加させるステップ、
上記パターン集合生成手段に、前記多支持パターン集合において、前記多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが前記多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成させるステップ、
上記出力手段に、前記多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、前記累積多支持パターン集合を構成する各多支持パター

ンについて、前記所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、前記累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力させるステップ
を有するプログラムを記録した記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

この発明は、複数のクラスタに分類して記録されている、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報において、各クラスタについて、そのクラスタを特徴づけるキーワードの連言（パターン）を抽出するための文書情報解析方法、文書情報解析装置および記録媒体に関するものである。

【０００２】

【従来の技術】

近年の情報通信メディア、特にインターネットやＷＷＷ（World Wide Web）などの発展に伴い、容易に入手可能な情報の量が急激に増加している。例えば、ＷＷＷにおいて各種情報を検索するために、Ｙａｈｏｏ（商標）などの検索エンジンが提供されている。

【０００３】

従来、文書データベースにおいて検索エンジンなどを利用して所望の文書情報を検索する場合、１つまたは複数のキーワードを指定し、そのキーワードに該当する文書情報を文書データベースから読み出すようにしている。

【０００４】

しかしながら、大量の種々雑多な文書情報がそのまま記録されている場合、その文書情報の中から所望の文書情報を抽出するために、すべての文書情報を検査する必要があり実用的ではない。例えば、検索対象に対する利用者の知識が曖昧であり、利用者がその対象を的確にとらえるキーワードを想起することが困難である場合、利用者は抽象的なキーワードを使用して検索するため、関連性の低い長大な文書情報が抽出されることが多く、利用者にとって興味のある文書情報に到達するまでに長い時間を費やしてしまうことになる。

【０００５】

そこで、オペレータが予めそのような種々雑多な文書情報をその内容に基づいて分類し、内容に関連性があるもの同士をクラスタごとにまとめて記録することや、ＳＯＭ（Self-Organizing Maps）などの自己組織化アルゴリズムを使用して、種々雑多な文書情報をクラスタに分類することが考えられる。なお、自己組織化アルゴリズムには、例えば「Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration 」（Ｌａｇｕｓら著、「Proc. of the Second International Conference on Knowledge Discovery and Data Mining KDD'96」の第２３８頁～第２４３頁、１９９６年）に記載のものや、「Exploration of Document Collections with Self-Organizing Maps: A Novel Approach to Similarity Representation」（Ｍｅｒｋｌ著、「Proc. of the first European Symposium on Principles of Data Mining and Knowledge Discovery PKDD'97」の第１０１頁～第１１１頁、１９９７年）に記載のものがある。

【０００６】

このように、文書情報をクラスタに分類しておくことにより、所望の文書情報を検索する場合、まず、その文書情報に関連するクラスタを検索し、該当するクラスタだけを検査すればよいことになる。

【０００７】

しかしながら、文書情報をその詳細な内容にわたってクラスタリングする場合、オペレータによるクラスタリングでは、時間やコストが多く必要になることになる。一方、ＳＯＭなどの自己組織化アルゴリズムを使用したクラスタリングでは、各クラスタがどのような特徴に基づいて生成されたのか、また各クラスタの文書情報がどのような理由でそのクラスタに分類されたのかを判断することが困難な場合が多い。

【０００８】
近年、データベースから特徴的なパターンを抽出するいわゆるデータマイニング技術が提案されている。このデータマイニング技術により、所定の文書情報から、その文書情報において特徴的なキーワードのパターンを抽出することができる。したがって、特徴的なキーワードのパターンを抽出しておき、その文書情報に関連づけて記録しておくことにより、そのパターンに基づいて文書情報の検索を行えば、比較的容易に所望の文書情報に到達することになると考えられる。
【０００９】
そのようなデータマイニングの方法としては、例えば、「Fast Algorithms for Mining Association Rules」（Ａｇｒａｗａｌら著、「Proc. of VLDB'94」の第４８７頁～第４９９頁、１９９４年）（以下、文献１という）に記載のものや「An Effective Hash Based Algorithm for Mining Association Rules」（Ｐａｒｋら著、「Proc. of SIGMOD'95」の第１７５頁～第１８６頁、１９９５年）（以下、文献２という）に記載のものがある。さらに文書情報に適用したものとしては、「Pattern Based Browsing in Document Collections」（Ｆｅｌｄｍａｎら著、「Proc. of the First European Symposium on Principles of Data Mining and Knowledge Discovery PKDD'97」の第１１２頁～第１２２頁、１９９７年）（以下、文献３という）に記載のものがある。
【００１０】
【発明が解決しようとする課題】
文書情報から特徴的なキーワードのパターンを抽出する従来の文書解析方法は以上のように構成されているので、例えば文献３に記載の方法により、個々のクラスタ内において特徴的なパターンを抽出することはできるが、他のクラスタのパターンと比較して特徴的なそのクラスタ特有のパターンを抽出することが困難であるなどの課題があった。
【００１１】
すなわち、複数の類似するクラスタが存在する場合に、文書情報の検索を容易にするようなそのクラスタ特有のパターンを抽出することが困難であるという課題があった。
【００１２】
例えば、図９は、スポーツ関係の新聞記事を記録したデータベースの一例を示している。図９に示すデータベースにおいては、サッカー関連の文書情報と、高校総体関連の文書情報が、各クラスタに対応するファイルに記録されている。この場合、例えば「つり輪」という単一のキーワードにより構成されるパターンは、「高校総体総合」クラスタと「高校総体つり輪」クラスタにおいて多く出現するので、このパターンにより、「高校総体総合」クラスタまたは「高校総体つり輪」クラスタを特徴づけることは困難である。一方、例えば「つり輪」と「種目別」により構成されるパターンは、「高校総体つり輪」クラスタにおいて多く出現するので、このパターンにより、「高校総体つり輪」クラスタを特徴づけることが可能である。しかしながら、このような、クラスタ特有のパターン、例えば上述の「つり輪」と「種目別」とにより構成されるパターンを抽出することは、個々のクラスタ内だけにおいて特徴的なパターンを抽出する従来の方法では困難であった。
【００１３】
この発明は上記のような課題を解決するためになされたもので、複数のクラスタに分類して記録されている、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報において、各クラスタについて、そのクラスタを特徴づけるキーワードのパターンを抽出するための文書情報解析方法、文書情報解析装置および記録媒体を得ることを目的とする。
【００１４】
【課題を解決するための手段】
この発明に係る文書情報解析方法は、初期パターン集合生成手段と、第１の選択手段と、第２の選択手段と、出力手段と、パターン集合生成手段を備えた文書情報解析装置において、初期パターン集合生成手段が、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にす

べてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、第１の選択手段が、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、第２の選択手段が、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択し、出力手段が、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力し、パターン集合生成手段が、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するものである。

【００１５】

この発明に係る文書情報解析方法は、第２の選択手段による所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の文書ＩＤを記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の文書ＩＤを読み出し、その文書情報においてだけ、（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントするものである。

【００１６】

この発明に係る文書情報解析方法は、第２の選択手段による所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の属するクラスタの情報を記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の属するクラスタの情報を読み出し、そのクラスタにおいてだけ、（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントするものである。

【００１７】

この発明に係る文書情報解析方法は、初期パターン集合生成手段において、初期のパターン集合を生成する前に、少なくとも１つのクラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを、すべての文書情報から除去するものである。

【００１８】

この発明に係る文書情報解析方法は、第１の選択手段において、各クラスタについて、所定のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するものである。

【００１９】

この発明に係る文書情報解析方法は、第１の選択手段において、各クラスタについて、初期のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するものである。

【００２０】

この発明に係る文書情報解析方法は、第２の選択手段において、新たなパターン集合を

生成する前に、確信度が所定の最大確信度以上である多支持パターンを多支持パターン集合から除去するものである。

【００２１】

この発明に係る文書情報解析方法は、第２の選択手段において、初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットのいずれかに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを、そのクラスタに関連づけて出力するとともに、高確信度パターン集合に追加するものである。

【００２２】

この発明に係る文書情報解析方法は、第２の選択手段において、多支持パターン集合の各パターンについて、その確信度とともにそのパターンのすべてのサブセットの確信度を計算し、その確信度が所定の最小確信度以上でかつサブセットの確信度の最大値より大きい前記パターンだけを出力するものである。

【００２３】

この発明に係る文書情報解析方法は、出力手段において、クラスタ毎に一括して、所定の順番でパターンを出力するものである。

【００２４】

この発明に係る文書情報解析方法は、初期パターン集合生成手段と、選択手段と、パターン集合生成手段と、出力手段を備えた文書情報解析装置において、初期パターン集合生成手段が、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、選択手段が、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加し、パターン集合生成手段が、多支持パターン集合において、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成し、出力手段が、多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力するものである。

【００２５】

この発明に係る文書情報解析方法は、初期パターン集合生成手段において、初期のパターン集合を生成する前に、少なくとも１つのクラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードをすべての文書情報から除去するものである。

【００２６】

この発明に係る文書情報解析方法は、選択手段において、累積多支持パターン集合を構成する各多支持パターンについて確信度を計算する前に、元の文書情報から、当該クラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するものである。

【００２７】
この発明に係る文書情報解析方法は、出力手段において、クラスタ毎に一括して、所定の順番で、確信度が所定の最小確信度以上であるパターンを出力するものである。
【００３３】
この発明に係る文書情報解析装置は、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択する第１の選択手段と、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択する第２の選択手段と、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力する出力手段と、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するパターン集合生成手段とを備えたものである。
【００３４】
この発明に係る文書情報解析装置は、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加する選択手段と、多支持パターン集合において、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するパターン集合生成手段と、多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力する出力手段とを備えたものである。
【００３５】
この発明に係る記録媒体は、初期パターン集合生成手段と、第１の選択手段と、第２の選択手段と、出力手段と、パターン集合生成手段を備えた文書情報解析装置において、初期パターン集合生成手段に、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパ

ターンの集合を初期のパターン集合として生成させるステップ、第１の選択手段に、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択させるステップ、第２の選択手段に、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択させるステップ、出力手段に、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力させるステップ、パターン集合生成手段に、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成させるステップを有するプログラムを記録したものである。

【００３６】

この発明に係る記録媒体は、初期パターン集合生成手段と、選択手段と、パターン集合生成手段と、出力手段を備えた文書情報解析装置において、初期パターン集合生成手段に、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成させるステップ、選択手段に、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加させるステップ、パターン集合生成手段に、多支持パターン集合において、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成させるステップ、出力手段に、多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力させるステップを有するプログラムを記録したものである。

【００３７】

【発明の実施の形態】

以下、この発明の実施の一形態を説明する。

実施の形態１．

図１は、この発明の実施の形態１による文書情報解析装置の構成を示すブロック図である。図において、１は、例えば図９に示すような文書情報をクラスタに分類して、各クラスタに対応するファイルにそれぞれ対応する文書情報を記録する例えばハードディスク装置、フロッピーディスクを装着したフロッピーディスク駆動装置などの記録部である。なお、文書情報は、図９に示すように文書ＩＤと、１つまたは複数のキーワードで構成される。

【００３８】
２は、記録部１に記録された文書情報をクラスタ毎に読み出して記憶するとともに、制御部３により指定されたクラスタに対応する文書情報を候補パターン生成部６に出力する文書集合記憶部である。また文書集合記憶部２は、多支持パターン集合生成部４または高確信度パターン出力部５によりクラスタとパターンが指定されると、指定されたクラスタにおいて、指定されたパターンが出現する文書情報の数を計算し、多支持パターン集合生成部４または高確信度パターン出力部５に出力する。なお、文書集合記憶部２は、各文書情報において、指定されたパターンが出現するか否かを検査する場合、その文書情報を構成するキーワードの部分集合がそのパターンに一致するか否かを検査する。したがって、この検査を効率的に行うために、ハッシュテーブルやハッシュ木を用いて各文書のキーワード集合に対して部分集合となる可能性のあるパターンを絞り込んだり、各パターン、文書情報においてキーワード集合を整列したり、あるいはビットパターンで表すようにしてもよい。なお、文書集合記憶部２を所定のコンピュータネットワークに接続し、そのコンピュータネットワークを介して他の記録部から文書情報を読み出すようにしてもよい。
【００３９】
６は、文書集合記憶部２より供給されたあるクラスタの文書情報に用いられている各キーワードに対し、１つのキーワードにより構成される候補パターンを生成して、多支持パターン集合生成部４に出力するとともに、多支持パターン集合生成部４より供給され、その多支持パターン集合の多支持パターンを構成するキーワードの数をｎとしたとき、（ｎ－１）個のキーワードが共通する任意の２つの多支持パターンから（ｎ＋１）個のキーワードからなるパターンを生成し、そのパターンの任意のｎ個のキーワードが多支持パターン集合に含まれるものを候補パターンとして、新たな候補パターン集合を生成し、多支持パターン集合生成部４に出力する候補パターン集合生成部（初期パターン集合生成手段、パターン集合生成手段）である。
【００４０】
４は、候補パターン集合生成部６より供給される各候補パターンについて、制御部３により指定されたクラスタにおいてその候補パターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を、文書集合記憶部２を利用して計算し、その支持率が所定の最小支持率以上であるパターンを新たな多支持パターンとして選択し、それらの多支持パターンの集合を高確信度パターン出力部５と候補パターン集合生成部６に出力する多支持パターン集合生成部（第１の選択手段、選択手段）である。
【００４１】
５は、多支持パターン集合生成部４より供給される各パターンについて、制御部３により指定されたクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を、文書集合記憶部２を利用して計算し、多支持パターン集合生成部４より供給されたパターン集合から、確信度が所定の最小確信度以上であるパターンを選択し、そのパターンを図示せぬ装置に出力する高確信度パターン出力部（第２の選択手段、出力手段）である。なお、高確信度パターン出力部５は、図示せぬ装置に出力するかわりに、そのパターンを、指定されたクラスタに関連づけて記録部１に記録させるようにしてもよい。
【００４２】
３は、文書集合記憶部２、多支持パターン集合生成部４、高確信度パターン出力部５、および候補パターン集合生成部６を制御する制御部である。
【００４３】
なお、上述の文書集合記憶部２、制御部３、多支持パターン集合生成部４、高確信度パターン出力部５、および候補パターン集合生成部６は、各種電子回路でハードウェアとして構成してもよいが、ＣＰＵを文書集合記憶部２、制御部３、多支持パターン集合生成部４、高確信度パターン出力部５、および候補パターン集合生成部６として機能させるためのプログラムを記録した記録媒体（例えば、ハードディスク装置、フロッピーディスクなどの磁気ディスク、コンパクトディスクなどの光ディスク）から、そのプログラムを読み取

ることが可能なコンピュータで構成してもよい。

【００４４】

次に動作について説明する。

図２は、図１の文書情報解析装置の動作を説明するフローチャートである。

まず、ステップＳＴ１において、文書集合記憶部２は、記録部１に記録された文書情報を読み出し、候補パターン集合生成部６に出力する。

【００４５】

ステップＳＴ２において、制御部３は、文書集合記憶部２により文書情報を読み出されたクラスタのうち、後述のステップＳＴ３～ステップＳＴ１０の処理を実行されていないクラスタがあるか否かを判定し、すべてのクラスタについてその処理を実行したと判定した場合、パターン抽出の処理を終了し、その処理を実行されていないクラスタが少なくとも１つあると判定した場合、そのクラスタのうちの１つを指定する。

【００４６】

候補パターン集合生成部６は、ステップＳＴ３において、供給された文書情報に含まれる各キーワードを１つのキーワードからなる候補パターン（初期候補パターン）とし、クラスタ毎に候補パターン集合（初期候補パターン集合）を生成し、多支持パターン集合生成部４に出力する。そして、ステップＳＴ４において、多支持パターン集合生成部４は、制御部３により指定されたクラスタに属する各候補パターンについて、文書集合記憶部２を制御してそのクラスタにおいてそのパターンが出現する文書情報の数を計算させ、ステップＳＴ５において、そのクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率をそれぞれ計算し、その支持率が所定の最小支持率以上であるパターンを多支持パターンとして選択し、それらの多支持パターンの集合を多支持パターン集合として高確信度パターン出力部５と候補パターン生成部６に出力する。

【００４７】

次に、高確信度パターン出力部５は、ステップＳＴ６において、多支持パターン集合生成部４より供給された多支持パターンについて、文書集合記憶部２を制御して、指定されたクラスタ以外の残りのクラスタにおいてそのパターンが出現する文書情報の数を計算させ、ステップＳＴ７において、そのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、その確信度が所定の最小確信度以上であるパターンを選択し、図示せぬ装置に出力する。すなわち、指定されたクラスタにおいてそのパターンが出現する文書情報の数と、すべてのクラスタにおいてそのパターンが出現する文書情報の数とに基づいた確信度を利用することにより、そのクラスタ特有のパターンが選択される。

【００４８】

そして、ステップＳＴ８において、制御部３は、多支持パターン集合生成部４により選択された多支持パターンの数が所定の最小パターン数より小さいか否かを判定し、選択されたパターンの数が所定の最小パターン数より小さいと判定した場合、指定したクラスタに対する処理を終了し、ステップＳＴ２に戻る。一方、選択されたパターンの数が所定の最小パターン数以上であると判定した場合、制御部３は、ステップＳＴ９において、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したか否かを判定し、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したと判定した場合、指定したクラスタに対する処理を終了し、ステップＳＴ２に戻る。

【００４９】

一方、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達していないと判定した場合、制御部３は、候補パターン集合生成部６に制御信号を供給する。ステップＳＴ１０において、候補パターン集合生成部６は、その制御信号を受信すると、多支持パターン集合生成部４より供給された多支持パターン集合について、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターン（（ｎ－１）個のキーワードが共通である２つの多支持パターン）を構成するキーワードのうちの（ｎ＋１）

個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成し、そのパターン集合を候補パターン集合として多支持パターン集合生成部４に出力する。なお、このステップＳＴ１０の処理には、文献１に記載されている「apriori-gen 」関数を利用することができる。

【００５０】

そして、ステップＳＴ４に戻り、ステップＳＴ１０で生成された候補パターン集合について、ステップＳＴ４～ステップＳＴ１０の処理が実行される。

【００５１】

すなわち、ステップＳＴ５～ステップＳＴ１０の処理を反復する毎に、パターンを構成するキーワードの数を１つずつ増えていくとともに、それに伴い、支持率が所定の最小支持率以上であるパターンが次第に減少していく。

【００５２】

以上のように、この実施の形態１によれば、各クラスタについて、所定のパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率と、そのクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度とに基づいて、キーワードのパターンを抽出するので、そのクラスタ特有のパターンを抽出することができるという効果が得られる。

【００５３】

なお、ステップＳＴ４～ステップＳＴ１０の反復処理において、２回目以降の処理では、前回の処理において選択された多支持パターンを有していた文書情報またはクラスタだけについて、候補パターンが出現するか否かの検査を行うようにしてもよい。前回の処理において選択された多支持パターンを有しない文書情報は、そのパターンを含む今回の処理の候補パターンも同様に有しないので、このようにすることにより、処理する文書情報やクラスタの数が低減され、処理時間を短縮することができるという効果が得られる。

【００５４】

なお、オペレータが図示せぬ操作部を操作して選択したクラスタだけについて、パターンの抽出の処理を行うようにしてもよい。その場合、図示せぬ操作部は、制御部３に接続され、その操作に対応する信号に応じて制御部３は、各部を制御する。このようにすることにより、不要なクラスタに対して処理が行われないので、処理時間を短縮することができるという効果が得られる。

【００５５】

実施の形態２.

この発明の実施の形態２による文書情報解析装置は、実施の形態１による文書情報解析装置の文書集合記憶部２と多支持パターン集合生成部４に後述の機能を追加したものである。

【００５６】

すなわち、実施の形態２における文書集合記憶部２は、記録部１から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情報から除去する。また、実施の形態２における多支持パターン集合生成部４は、候補パターン集合生成部６より供給される候補パターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用する。

【００５７】

なお、その他の構成要素は、実施の形態１のものと同様であるので、その説明を省略する。

【００５８】

次に動作について説明する。

図３は、実施の形態２による文書情報解析装置の動作を説明するフローチャートである。なお、ステップＳＴ１～ステップＳＴ１０の処理は、実施の形態１のものと同様であるので、その説明を省略する。

【００５９】

実施の形態２においては、ステップＳＴ１の後のステップＳＴ２１において、文書集合記憶部２は、記録部１から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情報から除去する。

【００６０】

また、ステップＳＴ５の後のステップＳＴ２２において、多支持パターン集合生成部４は、候補パターン集合生成部６より供給された候補パターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用する。

【００６１】

なお、ステップＳＴ２２において更新された文書情報は、各クラスタについての処理が終了したときに消去され、次のクラスタについての処理は、ステップＳＴ２１において生成された文書情報に基づいて行われる。

【００６２】

以上のように、この実施の形態２によれば、パターンを構成する各キーワードについて支持率を計算し、その支持率が所定の最小支持率より低いキーワードを文書情報から適宜除去するようにしたので、パターンが出現するか否かを検査する対象になる文書情報の量を低減させることができ、処理時間を短縮することができるいう効果が得られる。

【００６３】

なお、ステップＳＴ２２の処理は、ステップＳＴ２２、ステップＳＴ４～ステップＳＴ１０の反復処理の最初の処理においてだけ行うようにしてもよい。２回目以降のステップＳＴ２２の処理においては、除去されるキーワードの数が少ないので、このようにすることにより、より処理時間を短縮することができる場合が多い。

【００６４】

実施の形態３．

この発明の実施の形態３による文書情報解析装置は、実施の形態２による文書情報解析装置の高確信度パターン出力部５に後述の機能を追加したものである。

【００６５】

すなわち、実施の形態３における高確信度パターン出力部５は、多支持パターン集合生成部４より供給されたパターンから、その確信度が所定の最小確信度以上であるパターンを選択し、そのパターンを図示せぬ装置に出力した後、確信度が所定の最大確信度以上であるパターンを除去し、残りのパターンを新たに多支持パターン集合として多支持パターン集合生成部４に戻す。なお、多支持パターン集合生成部４は、高確信度パターン出力部５から戻された多支持パターン集合を候補パターン集合生成部６に出力する。

【００６６】

なお、その他の構成要素は、実施の形態２のものと同様であるので、その説明を省略する。

【００６７】

次に動作について説明する。

図４は、実施の形態３による文書情報解析装置の動作を説明するフローチャートである。なお、ステップＳＴ１～ステップＳＴ１０、ステップＳＴ２１、およびステップＳＴ２２の処理は、実施の形態２のものと同様であるので、その説明を省略する。

【００６８】

実施の形態３においては、高確信度パターン出力部５は、ステップＳＴ７において、多支持パターン集合生成部４より供給されたパターンから、その確信度が所定の最小確信度以

上であるパターンを選択し、図示せぬ装置に出力した後、ステップＳＴ３１において、その確信度が所定の最大確信度以上であるパターンを除去し、残りのパターンを新たに多支持パターン集合として多支持パターン集合生成部４に戻す。なお、多支持パターン集合生成部４は、ステップＳＴ６で多支持パターン集合を候補パターン集合生成部６に出力する代わりに、ステップＳＴ３１で高確信度パターン出力部５より戻された多支持パターン集合を出力する。

【００６９】

以上のように、この実施の形態３によれば、確信度が所定の最大確信度以上であるパターンを多支持パターン集合から除去するようにしたので、そのようなパターンを含む、利用者にとって不必要に複雑なパターンの出力と、そのための処理を抑制することができるという効果が得られる。

【００７０】

実施の形態４.

この発明の実施の形態４による文書情報解析装置は、実施の形態３による文書情報解析装置の高確信度パターン出力部５に後述の機能を追加したものである。

【００７１】

実施の形態４においては、高確信度パターン出力部５は、初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないか、あるいは、確信度がそのパターンのサブセットに該当する高確信度パターン集合におけるパターンの確信度以上であるパターンだけを、そのクラスタに関連づけて出力するとともに、高確信度パターン集合に追加する。

【００７２】

なお、その他の構成要素は、実施の形態３のものと同様であるので、その説明を省略する。

【００７３】

次に動作について説明する。

図５は、実施の形態４による文書情報解析装置の動作を説明するフローチャートである。なお、ステップＳＴ１～ステップＳＴ１０、ステップＳＴ２１、ステップＳＴ２２、およびステップＳＴ３１の処理は、実施の形態３のものと同様であるので、その説明を省略する。

【００７４】

実施の形態４においては、高確信度パターン出力部５は、ステップＳＴ４１において、初期状態の空集合である高確信度パターン集合を記憶しておき、ステップＳＴ４２において、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを、そのクラスタに関連づけて出力し、その後、高確信度パターン集合に追加する。

【００７５】

以上のように、この実施の形態４によれば、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを選択するようにしたので、キーワード数が多くかつ確信度の低いパターンを除去することができ、キーワード数が少なく確信度の高いパターンが選択され、検索時の効率を向上させることができるという効果が得られる。

【００７６】

なお、ステップＳＴ１０において新たなパターン集合を生成するときに、各パターンのすべてのサブセットの確信度を同時に計算し、そのパターンのすべてのサブセットの確信度

のうちの最大値をそのパターンに対応して記憶して、ステップＳＴ４２においては、そのサブセットの確信度の最大値以上の確信度を有するパターンだけを選択し、そのパターンを出力するとともに、高確信度パターン集合に追加するようにしてもよい。このようにすることにより、ステップＳＴ４２における処理を簡略化することができ、処理時間を短縮することができるという効果が得られる。

【００７７】

実施の形態５．

この発明の実施の形態５による文書情報解析装置は、実施の形態４による文書情報解析装置の高確信度パターン出力部５に後述の機能を追加したものである。

【００７８】

すなわち、実施の形態５においては、高確信度パターン出力部５は、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないか、あるいは、そのパターンの確信度がそのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上であるパターンだけ高確信度パターン集合に追加していき、各クラスタに対する処理が終了したとき、クラスタ毎に一括して、所定の順番で高確信度パターン集合を出力する。

【００７９】

なお、その他の構成要素は、実施の形態４のものと同様であるので、その説明を省略する。

【００８０】

次に動作について説明する。

図６は、実施の形態５による文書情報解析装置の動作を説明するフローチャートである。なお、ステップＳＴ１～ステップＳＴ１０、ステップＳＴ２１、ステップＳＴ２２、ステップＳＴ３１、ステップＳＴ４１、およびステップＳＴ４２の処理は、実施の形態４のものと同様であるので、その説明を省略する。

【００８１】

実施の形態５においては、高確信度パターン出力部５は、ステップＳＴ４２において、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないか、あるいは、そのパターンの確信度がそのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上であるパターンだけ高確信度パターン集合に追加していく。なお、このときには、高確信度パターン出力部５は、そのパターンを図示せぬ所定の装置には出力しない。

【００８２】

そして、高確信度パターン出力部５は、ステップＳＴ８またはステップＳＴ９において、選択されたパターンの数が所定の最小パターン数より小さいと判定した場合、または、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したと判定した場合、ステップＳＴ５１において、そのクラスタに対応するパターンとして、所定の順番で高確信度パターン集合を出力する。

【００８３】

なお、このときの順番は、高確信度パターン集合を構成するパターンの支持率、確信度、パターンを構成するキーワードの数などに基づいて予め設定する。

【００８４】

また、Ｄを文書情報の集合、Ｄｐをパターンｐが出現する文書情報の集合としたときの、式（１）および式（２）により規定される各パターンｐに対する所定のクラスタc_iの条件付きエントロピーEnt_i（Ｄ｜ｐ）の値に基づいて順番を設定してもよい。

【数１】

$$Ent_i(D|P) = P(Dp) \times Ent_i(Dp) + P(\overline{Dp}) \times Ent_i(\overline{Dp}) \quad \cdots (1)$$

ただし、$\overline{Dp} = D - Dp$, $P(Dp) = |Dp| / |D|$

【００８５】
【数２】

$$Ent_i(D) = -(P_i^+ \log P_i^+ + P_i^- \log P_i^-) \quad \cdots (2)$$

ただし、$P_i^+ = |C_i \cap D| / |D|$, $P_i^- = 1 - P_i^+$

【００８６】
以上のように、この実施の形態５によれば、所定の順番でパターンを出力するようにしたので、利用者の要求に応じた順番でパターンを出力することができるという効果が得られる。
【００８７】
実施の形態６．
この発明の実施の形態６による文書情報解析装置は、実施の形態１による文書情報解析装置の制御部３による制御手順を後述のように変更したものである。
【００８８】
なお、その他の構成要素は、実施の形態１のものと同様であるので、その説明を省略する。
【００８９】
次に動作について説明する。
図７は、実施の形態６による文書情報解析装置の動作を説明するフローチャートである。まず、ステップＳＴ１において、文書集合記憶部２は、記録部１に記録された文書情報を読み出し、候補パターン集合生成部６に出力する。
【００９０】
ステップＳＴ２において、制御部３は、文書集合記憶部２により文書情報を読み出されたクラスタのうち、後述のステップＳＴ３～ステップＳＴ５、ステップＳＴ８～ステップＳＴ１０、ステップＳＴ６１およびステップＳＴ６３～ステップＳＴ６５の処理を実行されていないクラスタがあるか否かを判定し、すべてのクラスタについてその処理を実行したと判定した場合、パターン抽出の処理を終了し、その処理を実行されていないクラスタが少なくとも１つあると判定した場合、そのクラスタのうちの１つを指定する。
【００９１】
候補パターン集合生成部６は、ステップＳＴ３において、供給された文書情報に含まれる各キーワードを１つのキーワードからなる候補パターン（初期候補パターン）とし、クラスタ毎に候補パターン集合（初期候補パターン集合）を生成し、多支持パターン集合生成部４に出力する。そして、ステップＳＴ４において、多支持パターン集合生成部４は、制御部３により指定されたクラスタに属する各候補パターンについて、文書集合記憶部２を制御してそのクラスタにおいてそのパターンが出現する文書情報の数を計算させ、ステップＳＴ５において、そのクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率をそれぞれ計算し、その支持率が所定の最小支持率以上であるパターンを多支持パターンとして選択し、それらの多支持

パターンの集合を多支持パターン集合として候補パターン生成部６に出力する。

【００９２】

そして、ステップＳＴ６１において、多支持パターン集合生成部４は、多支持パターンを累積多支持パターン集合に追加する。なお、累積多支持パターン集合は、各クラスタについての処理を開始するとき（ステップＳＴ３）において、空集合に設定される。

【００９３】

そして、ステップＳＴ８において、制御部３は、多支持パターン集合生成部４により選択されたパターンの数が所定の最小パターン数より小さいか否かを判定し、選択されたパターンの数が所定の最小パターン数より小さくないと判定した場合、ステップＳＴ９に進み、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したか否かを判定する。

【００９４】

そして、ステップＳＴ８において選択されたパターンの数が所定の最小パターン数より小さいと判定した場合、または、ステップＳＴ９において選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したと判定した場合、ステップＳＴ６３に進み、制御部３は、多支持パターン集合生成部４に制御信号を供給し、そのときに選択されているパターンを高確信度パターン出力部５に出力させる。

【００９５】

一方、ステップＳＴ９において、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達していないと判定した場合、制御部３は特に何もしない。したがって、ステップＳＴ１０に進み、候補パターン集合生成部６は、多支持パターン集合生成部４より供給された多支持パターン集合について、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターンの集合を生成し、そのパターン集合を候補パターン集合として多支持パターン集合生成部４に出力する。そして、ステップＳＴ４に戻り、ステップＳＴ１０で生成された候補パターン集合について上述の処理を同様に実行する。

【００９６】

このようにして、ステップＳＴ８またはステップＳＴ９の条件が成立するまで、ステップＳＴ４、ステップＳＴ５、ステップＳＴ６１、ステップＳＴ８～ステップＳＴ１０の処理が反復的に実行される。

【００９７】

そして、反復処理が終了した後、ステップＳＴ６３においては、制御部３は、多支持パターン集合生成部４に制御信号を供給し、そのときに選択されているパターンを高確信度パターン出力部５に出力させる。

【００９８】

高確信度パターン出力部５は、ステップＳＴ６４において、多支持パターン集合生成部４より供給された各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、多支持パターン集合生成部４より供給されたパターン集合から、確信度が所定の最小確信度以上であるパターンを選択し、ステップＳＴ６５において、そのパターンを図示せぬ装置に出力する。そして、ステップＳＴ２に戻り、制御部３は、次のクラスタに対して同様の処理を行うように各部を制御する。

【００９９】

以上のように、この実施の形態６によれば、指定されたクラスタ以外のクラスタにおいてパターンが出現する文書情報を１回だけカウントすればよいので、処理時間を短縮することができるという効果が得られる。

【０１００】

なお、この実施の形態６においても、実施の形態５のように、所定の順番でパターンを出

力するようにしてもよい。
【０１０１】
実施の形態７.
この発明の実施の形態７による文書情報解析装置は、実施の形態６による文書情報解析装置の文書集合記憶部２と多支持パターン集合生成部４に後述の機能を追加したものである。
【０１０２】
すなわち、実施の形態７における文書集合記憶部２は、記録部１から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情報から除去する。また、多支持パターン集合生成部４は、候補パターン集合生成部６より供給される候補パターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用する。
【０１０３】
なお、その他の構成要素は、実施の形態６のものと同様であるので、その説明を省略する。
【０１０４】
次に動作について説明する。
図８は、実施の形態７による文書情報解析装置の動作を説明するフローチャートである。なお、ステップＳＴ１～ステップＳＴ５、ステップＳＴ８～ステップＳＴ１０、ステップＳＴ６１、およびステップＳＴ６３～ステップＳＴ６５の処理は、実施の形態６のものと同様であるので、その説明を省略する。
【０１０５】
実施の形態７においては、ステップＳＴ１の後のステップＳＴ７１において、文書集合記憶部２が、記録部１から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情報から除去する。
【０１０６】
また、ステップＳＴ６３の前のステップＳＴ７２において、高確信度パターン出力部５が、多支持パターン集合生成部４より供給されるパターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成する。そして、ステップＳＴ６３の処理では、その文書情報が元の文書情報の代わりに使用される。
【０１０７】
以上のように、この実施の形態７によれば、パターンを構成する各キーワードについて支持率を計算し、その支持率が所定の最小支持率より低いキーワードをすべての文書情報から除去するようにしたので、パターンが出現するか否かを検査する対象になる文書情報の数を低減させることができ、処理時間を短縮することができるいう効果が得られる。
【０１０８】
【発明の効果】
以上のように、この発明によれば、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラス

タにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択し、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力し、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数をｎとしたとき、複数のパターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するように構成したので、そのクラスタ特有のパターンを抽出することができる効果がある。

【０１０９】

この発明によれば、所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の文書ＩＤを記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の文書ＩＤを読み出し、その文書情報においてだけ、（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントするように構成したので、パターンの出現の検査の対象になる文書情報の数が低減され、処理時間を短縮することができる効果がある。

【０１１０】

この発明によれば、所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の属するクラスタの情報を記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を（ｍ＋１）個とすると、ｍ個のキーワードで構成されるパターンが出現した文書情報の属するクラスタの情報を読み出し、そのクラスタにおいてだけ、（ｍ＋１）個のキーワードで構成される所定のパターンが出現する文書情報をカウントするように構成したので、パターンの出現の検査の対象になるクラスタの数が低減され、処理時間を短縮することができる効果がある。

【０１１１】

この発明によれば、初期のパターン集合を生成する前に、少なくとも１つのクラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを、すべての文書情報から除去するように構成したので、パターンの出現の検査の対象になるキーワードの数が低減され、処理時間を短縮することができる効果がある。

【０１１２】

この発明によれば、各クラスタについて、所定のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するように構成したので、パターンの出現の検査の対象になるキーワードの数が低減され、処理時間を短縮することができる効果がある。

【０１１３】

この発明によれば、各クラスタについて、初期のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するように構成したので、処理の最初だけにおいて不要なキーワードが除去され、除去の処理にかかる時間を低減することができる効果がある。

【０１１４】

この発明によれば、新たなパターン集合を生成する前に、確信度が所定の最大確信度以上である多支持パターンを多支持パターン集合から除去するように構成したので、利用者の必要頻度が低く、かつそのクラスタにしか存在しないような複雑なパターンを除去することができる効果がある。

【０１１５】
この発明によれば、初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットのいずれかに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを、そのクラスタに関連づけて出力するとともに、高確信度パターン集合に追加するように構成したので、キーワード数が多くかつ確信度の低いパターンを除去することができ、キーワード数が少なく確信度の高いパターンが選択され、検索時の効率を向上させることができる効果がある。
【０１１６】
この発明によれば、多支持パターン集合の各パターンについて、その確信度とともにそのパターンのすべてのサブセットの確信度を計算し、その確信度が所定の最小確信度以上でかつサブセットの確信度の最大値より大きい前記パターンだけを出力するように構成したので、キーワード数が長くかつ確信度の低いパターンを選択するときの処理を簡略化することができ、処理時間を短縮することができる効果がある。
【０１１７】
この発明によれば、クラスタ毎に一括して、所定の順番でパターンを出力するように構成したので、利用者の要求に応じた順番でパターンを出力することができる効果がある。
【０１１８】
この発明によれば、複数のクラスタに分類された、文書ＩＤとその文書ＩＤに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ１つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加し、多支持パターン集合において、多支持パターンを構成するキーワードの数をｎとしたとき、複数の多支持パターンを構成するキーワードのうちの（ｎ＋１）個のキーワードで構成され、かつ、その（ｎ＋１）個のキーワードのうちの任意のｎ個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成し、多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力するように構成したので、指定されたクラスタ以外のクラスタにおいてパターンが出現する文書情報を１回だけカウントすればよいので、処理時間を短縮することができる効果がある。
【０１１９】
この発明によれば、累積多支持パターン集合を構成する各多支持パターンについて確信度を計算する前に、元の文書情報から、当該クラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するように構成したので、パターンの出現の検査の対象になるキーワードの数が低減され、処理時間を短縮することができる効果がある。
【０１２０】
この発明によれば、すべてのクラスタから選択されたクラスタだけについてパターンを選択し、出力するように構成したので、不要なクラスタに対して処理が行われないので、処理時間を短縮することができる効果がある。

【図面の簡単な説明】

【図１】　この発明の実施の形態１による文書情報解析装置の構成を示すブロック図である。

【図２】　図１の文書情報解析装置の動作を説明するフローチャートである。

【図３】　実施の形態２による文書情報解析装置の動作を説明するフローチャートである。

【図４】　実施の形態３による文書情報解析装置の動作を説明するフローチャートである。

【図５】　実施の形態４による文書情報解析装置の動作を説明するフローチャートである。

【図６】　実施の形態５による文書情報解析装置の動作を説明するフローチャートである。

【図７】　実施の形態６による文書情報解析装置の動作を説明するフローチャートである。

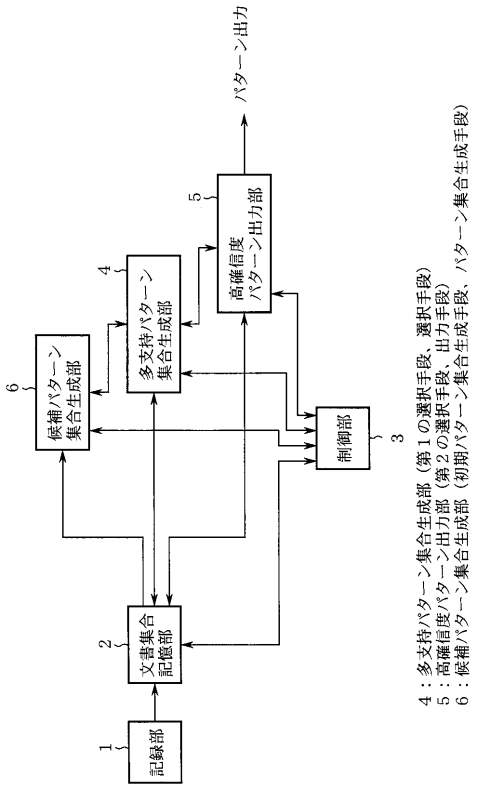
【図８】　実施の形態７による文書情報解析装置の動作を説明するフローチャートである。

【図９】　スポーツ関係の新聞記事を記録したデータベースの一例を示している。

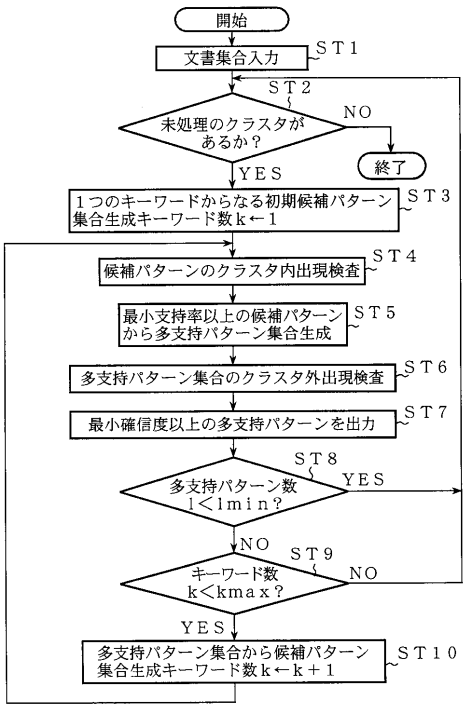
【符号の説明】

１　記録部、４　多支持パターン集合生成部（第１の選択手段、選択手段）、５　高確信度パターン出力部（第２の選択手段、出力手段）、６　候補パターン集合生成部（初期パターン集合生成手段、パターン集合生成手段）。

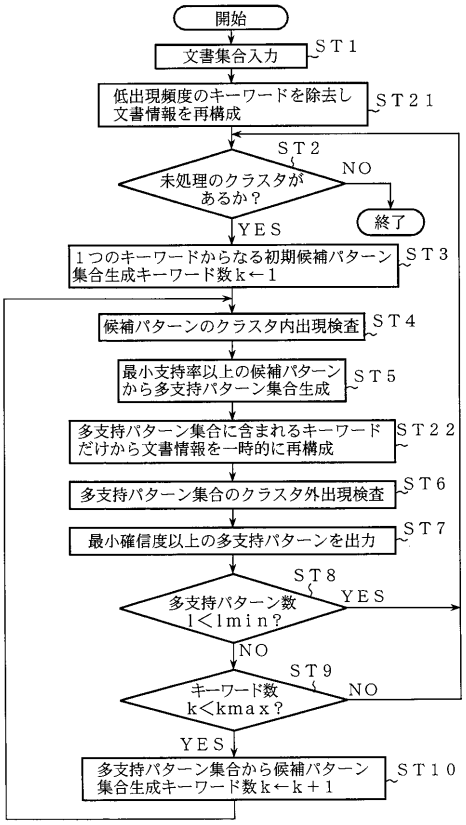
【図１】



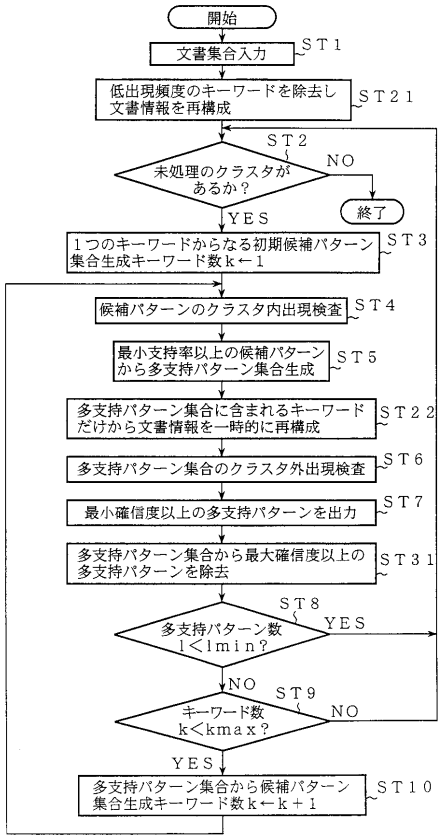
【図２】



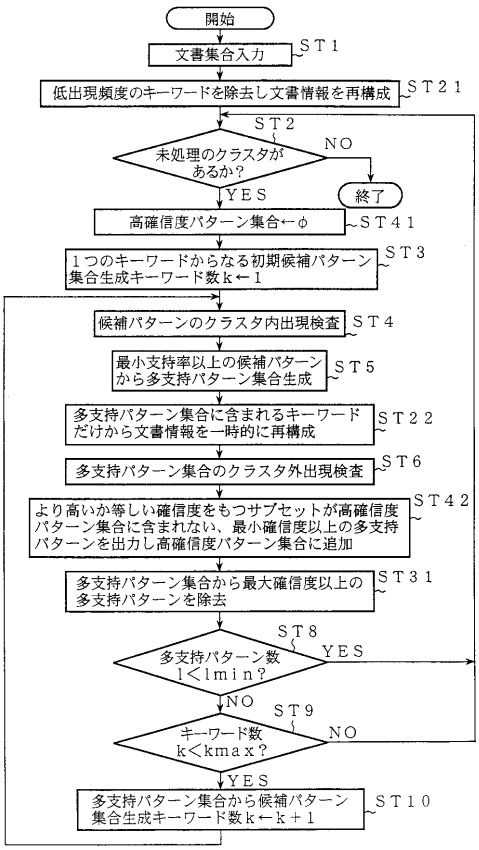
【図３】


開始
文書集合入力
ST1
低出現頻度のキーワードを除去し文書情報を再構成
ST21
ST2
未処理のクラスタがあるか？
NO
終了
YES
1つのキーワードからなる初期候補パターン集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターンから多支持パターン集合生成
ST5
多支持パターン集合に含まれるキーワードだけから文書情報を一時的に再構成
ST22
多支持パターン集合のクラスタ外出現検査
ST6
最小確信度以上の多支持パターンを出力
ST7
ST8
多支持パターン数 l<lmin？
YES
NO
ST9
キーワード数 k<kmax？
NO
YES
多支持パターン集合から候補パターン集合生成キーワード数k←k+1
ST10

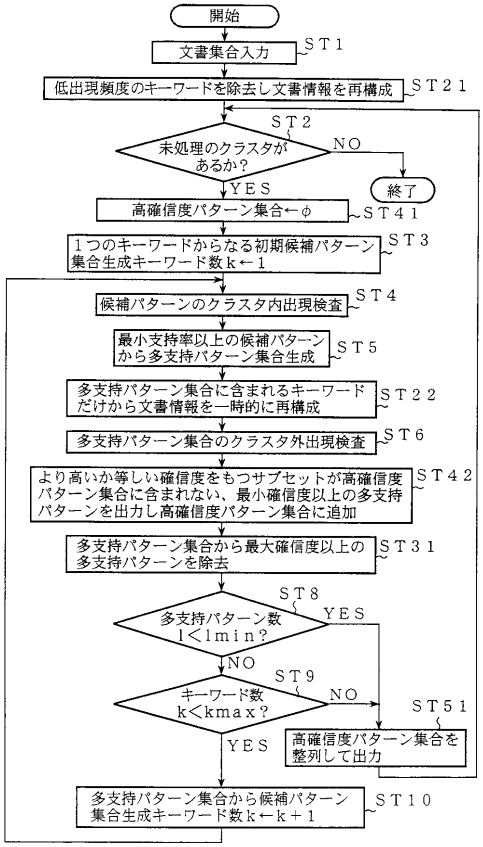
【図４】


開始
文書集合入力
ST1
低出現頻度のキーワードを除去し文書情報を再構成
ST21
ST2
未処理のクラスタがあるか？
NO
終了
YES
1つのキーワードからなる初期候補パターン集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターンから多支持パターン集合生成
ST5
多支持パターン集合に含まれるキーワードだけから文書情報を一時的に再構成
ST22
多支持パターン集合のクラスタ外出現検査
ST6
最小確信度以上の多支持パターンを出力
ST7
多支持パターン集合から最大確信度以上の多支持パターンを除去
ST31
ST8
多支持パターン数 l<lmin？
YES
NO
ST9
キーワード数 k<kmax？
NO
YES
多支持パターン集合から候補パターン集合生成キーワード数k←k+1
ST10

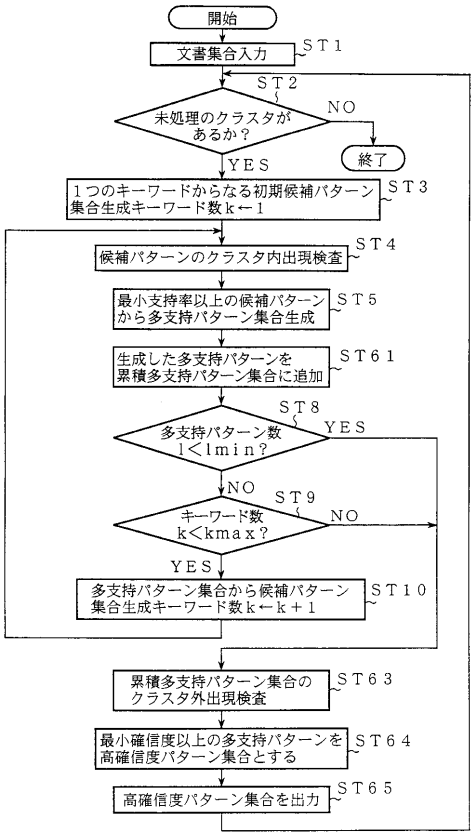
【図５】


開始
文書集合入力
ST1
低出現頻度のキーワードを除去し文書情報を再構成
ST21
ST2
未処理のクラスタがあるか？
NO
終了
YES
高確信度パターン集合←φ
ST41
1つのキーワードからなる初期候補パターン集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターンから多支持パターン集合生成
ST5
多支持パターン集合に含まれるキーワードだけから文書情報を一時的に再構成
ST22
多支持パターン集合のクラスタ外出現検査
ST6
より高いか等しい確信度をもつサブセットが高確信度パターン集合に含まれない、最小確信度以上の多支持パターンを出力し高確信度パターン集合に追加
ST42
多支持パターン集合から最大確信度以上の多支持パターンを除去
ST31
ST8
多支持パターン数 l<lmin？
YES
NO
ST9
キーワード数 k<kmax？
NO
YES
多支持パターン集合から候補パターン集合生成キーワード数k←k+1
ST10

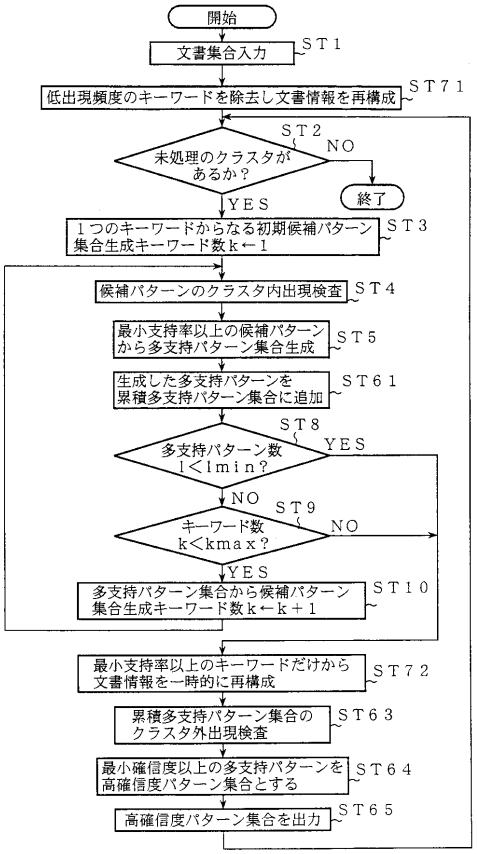
【図６】


開始
文書集合入力
ST1
低出現頻度のキーワードを除去し文書情報を再構成
ST21
ST2
未処理のクラスタがあるか？
NO
終了
YES
高確信度パターン集合←φ
ST41
1つのキーワードからなる初期候補パターン集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターンから多支持パターン集合生成
ST5
多支持パターン集合に含まれるキーワードだけから文書情報を一時的に再構成
ST22
多支持パターン集合のクラスタ外出現検査
ST6
より高いか等しい確信度をもつサブセットが高確信度パターン集合に含まれない、最小確信度以上の多支持パターンを出力し高確信度パターン集合に追加
ST42
多支持パターン集合から最大確信度以上の多支持パターンを除去
ST31
ST8
多支持パターン数 l<lmin？
YES
NO
ST9
キーワード数 k<kmax？
NO
YES
ST51
高確信度パターン集合を整列して出力
多支持パターン集合から候補パターン集合生成キーワード数k←k+1
ST10

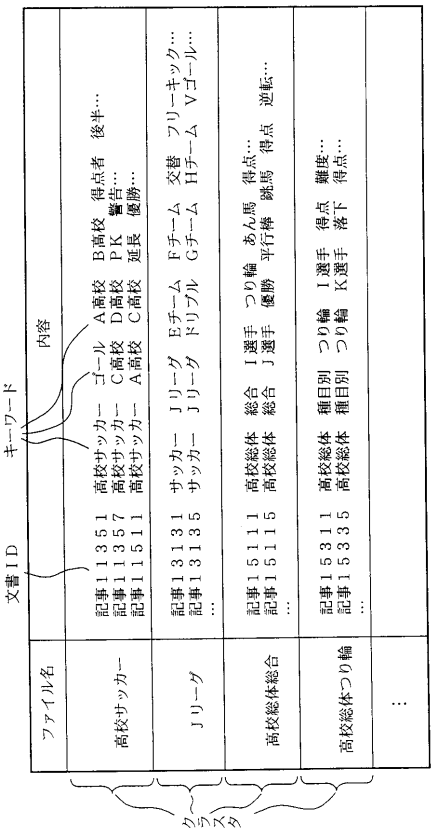
【図７】


開始
文書集合入力
ST1
ST2
未処理のクラスタが
あるか？
NO
終了
YES
１つのキーワードからなる初期候補パターン
集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターン
から多支持パターン集合生成
ST5
生成した多支持パターンを
累積多支持パターン集合に追加
ST61
ST8
多支持パターン数
l＜lmin？
YES
NO
ST9
キーワード数
k＜kmax？
NO
YES
多支持パターン集合から候補パターン
集合生成キーワード数k←k＋1
ST10
累積多支持パターン集合の
クラスタ外出現検査
ST63
最小確信度以上の多支持パターンを
高確信度パターン集合とする
ST64
高確信度パターン集合を出力
ST65

【図８】


開始
文書集合入力
ST1
低出現頻度のキーワードを除去し文書情報を再構成
ST71
ST2
未処理のクラスタが
あるか？
NO
終了
YES
１つのキーワードからなる初期候補パターン
集合生成キーワード数k←1
ST3
候補パターンのクラスタ内出現検査
ST4
最小支持率以上の候補パターン
から多支持パターン集合生成
ST5
生成した多支持パターンを
累積多支持パターン集合に追加
ST61
ST8
多支持パターン数
l＜lmin？
YES
NO
ST9
キーワード数
k＜kmax？
NO
YES
多支持パターン集合から候補パターン
集合生成キーワード数k←k＋1
ST10
最小支持率以上のキーワードだけから
文書情報を一時的に再構成
ST72
累積多支持パターン集合の
クラスタ外出現検査
ST63
最小確信度以上の多支持パターンを
高確信度パターン集合とする
ST64
高確信度パターン集合を出力
ST65

【図９】


ファイル名
内容
文書ID
キーワード
高校サッカー
記事11351 高校サッカー ゴール A高校 B高校 得点者 後半…
記事11357 高校サッカー C高校 D高校 PK 警告…
記事11511 高校サッカー A高校 C高校 延長 優勝…
Jリーグ
記事13131 サッカー Jリーグ Eチーム Fチーム 交替 フリーキック…
記事13135 サッカー Jリーグ ドリブル Gチーム Hチーム Vゴール…
…
高校総体総合
記事15111 高校総体 総合 I選手 つり輪 あん馬 得点…
記事15115 高校総体 総合 J選手 優勝 平行棒 跳馬 得点 逆転…
…
高校総体つり輪
記事15311 高校総体 種目別 つり輪 I選手 得点 難度…
記事15335 高校総体 種目別 つり輪 K選手 落下 得点…
…
…
クラスタ

フロントページの続き

(58)調査した分野(Int.Cl.[7], ＤＢ名)
G06F 17/30