明　細　書

発明の名称 ： パターン抽出装置及び方法

技術分野

[0001] 本発明の実施形態は、複数のアイテムを含む対象情報からアイテムの組合せパターンを抽出する制御に関する。

背景技術

[0002] 従来から各種の活動や事象などを分析するために、複数のアイテムで構成された対象情報から、分析目的等に応じた特定のパターンすなわちアイテムの組合せを、効率良く抽出するための装置や方法が研究されている。

先行技術文献

特許文献

[0003] 特許文献1：特開２００３－７６９３７号公報

発明の概要

発明が解決しようとする課題

[0004] アイテムを含む複数の対象情報から、アイテム間の関連性を反映させて、特定のパターンを抽出させるパターン抽出装置及び方法を提供する。

課題を解決するための手段

[0005] 実施形態のパターン抽出装置は、複数のアイテムを含む対象情報内に存在する相互に異なる２以上のアイテムの組合せのパターンを抽出するパターン抽出装置であって、複数の対象情報を記憶する第１記憶部と、複数の対象情報それぞれに含まれる各アイテムに基づいて、相互に異なる２以上のアイテムで構成される候補パターンを生成する候補パターン生成部と、生成された候補パターンが複数の対象情報それぞれに出現する出現頻度に基づいて、候補パターンの抽出評価値を算出する候補評価値算出部と、算出された抽出評価値が所定の閾値を満たす候補パターンを判別し、閾値を満たす候補パターンを抽出するパターン抽出部と、を含むとともに、アイテム間の関連度を記憶する第２記憶部をさらに含む。候補評価値算出部は、候補パターンに含ま

れる各アイテム間の関連度を識別し、識別された関連度に基づく加重値及び出現頻度に基づいて、抽出評価値を算出する。

図面の簡単な説明

[0006] [図1]第１実施形態のパターン抽出装置の構成例を示す図であり、本装置を構成する各部間の関係を示すブロック図である。

[図2]データ格納部に格納される対象情報としてのトランザクション群の一例を示す図である。

[図3]アイテム間情報記憶部に記憶される関連度行列テーブルの一例を示す図である。

[図4]パターン抽出装置の動作を説明するためのフローチャートである。

[図5]アイテム抽出部が行う処理を説明するためのサブルーチンである。

[図6]長さが１となる候補パターンの頻度と支持度の一例を示す図である。

[図7]パターン格納部に格納される長さ１のパターンの一例を示す図である。

[図8]候補パターン生成部が行う処理を説明するためのサブルーチンである。

[図9]候補評価値算出部が行う処理を説明するためのサブルーチンである。

[図10]長さが２となる候補パターンの頻度と関連性支持度の一例を示す図である。

[図11]パターン格納部に格納される長さ２のパターンの一例を示す図である。

[図12]長さが３となる候補パターンの頻度と関連性支持度の一例を示す図である。

[図13]パターン格納部に格納される長さ３のパターンの一例を示す図である。

[図14]パターン抽出装置の他の実施形態の動作を説明するためのフローチャートである。

発明を実施するための形態

[0007] 以下、実施形態につき、図面を参照して説明する。

[0008] 実施形態に係るパターン抽出装置は、複数のアイテム（情報要素）を含む

対象情報が複数与えられた場合に、アイテム間に存在する関係を利用して、当該複数の対象情報の中からアイテムの特徴的な組合せ（パターン）を抽出する制御を遂行するものである。

[0009] なお、「パターン」の用語は、通常は２以上のアイテムの組合せを意味するが、以下の説明では、単一のアイテムに対しても「パターン」の用語が用いられる場合がある。一方、「パターン」の用語は、狭義には、上記「アイテムの特徴的な組合せ」を有するパターンを意味する。そして、かかる狭義のパターンを抽出するための候補となるパターンを「パターンの候補」または「候補パターン」と称する。

[0010] 本装置は、例えば、スーパーマーケットなどの日用品の販売における、購入商品の特徴的な組合せの発見、銀行業務における、店舗の特性と事務ミスの種類との間にある特徴的な因果関係の発見、番組推薦における、視聴者特性と視聴履歴との間にある視聴者の嗜好の発見などの分野に利用することができる。但し、これらは例示であって、これに限定されるものではない。

[0011] 以下の説明では、スーパーマーケットなどの日用品（食品）の販売業における商品をアイテムとして本装置で扱う場合について説明する。具体的には、あるスーパーマーケットでの食品を販売するフロアにおいて、購入者が購入した商品（食品）の特徴的な組合せ（すなわち「パターン」）を抽出、発見するために、商品購入者の一枚のレシートを１個の対象情報（トランザクション）とし、レシートに記載された商品（食品名等の品名）をアイテムとして扱う場合について説明する。

[0012] 図１から図１３は、第１実施形態を示す図である。図１は、本実施形態のパターン抽出装置１００のブロック構成図であり、図中の各ブロックを結線する矢印は、データの流れる方向を示している。本装置は、複数の対象情報としてトランザクションの集合を記憶する第１記憶部としてのデータ格納部１０と、アイテム間の関連性に関する情報を記憶する第２記憶部としてのアイテム間情報格納部２０と、トランザクションの集合からアイテムを抽出するアイテム抽出部３０と、抽出されたアイテムを用いて、相互に異なる２以

上のアイテムの組合せによるパターンの候補（候補パターン）を生成する候補パターン生成部４０と、候補パターンがトランザクションに出現する頻度を算出する候補頻度算出部５０と、候補パターンの頻度及び前記アイテム間の関連性に関する情報を用いて、候補パターンに対する抽出評価値を算出する候補評価値算出部６０と、算出された抽出評価値が基準値を満たす候補パターンをパターンとして抽出する処理を行う候補評価部７０およびパターン格納部８０と、を含む。

[0013] 本装置は、後述する各処理を実行するプログラムのデータを図示しないハードディスク装置などの外部記憶媒体に格納し、かかるプログラムをパーソナルコンピュータ（ＰＣ）に読み込ませることで実現することができる。この場合、例えば当該コンピュータのハードディスク装置やＲＡＭなどの記憶デバイスがデータ格納部１０、アイテム間情報格納部２０、パターン格納部８０として機能し、ＣＰＵなどの制御デバイスがアイテム抽出部３０、候補パターン生成部４０、候補頻度算出部５０、候補評価値算出部６０、及び候補評価部７０として機能することができる。

[0014] データ格納部１０は、後述する一連の処理に先立って、対象情報であるトランザクションの集合（以下、「トランザクション群」ともいう。）のデータを記憶するトランザクション記憶部として機能するとともに、後述する最小支持度に関するデータ、及び、アイテムの整列の優先度を示すデータを格納する。

[0015] データ格納部１０に記憶される対象情報としてのトランザクション群の一例を図２に示す。トランザクションは、複数のアイテム（この例では実際に購入された商品である「鶏肉」、「豚肉」、「牛肉」、「鮪」、「鯵」、「ビール」の６個すなわち６種のアイテム）の内の１種以上のアイテムで構成されたものであり、図２ではＡ０１、Ａ０２、Ａ０３、Ａ０４、Ａ０５の５個のトランザクションがトランザクション群としてデータ格納部１０の所定記憶領域に記憶（格納）された場合を例示している。

[0016] スーパーマーケットなどの日用品の販売業を対象とした場合、購入商品の

一覧が記載された１枚のレシートが１つのトランザクション（例えばＡ０１）に相当する。ただし、この例では、レシートに記載されている商品の金額や購入個数には着目しておらず、商品が購入されたかどうかに関する情報のみに着目している。このため、データ格納部１０には、図２に例示したように、商品名の情報のみをアイテムとし、購入個数が複数であっても１つの商品名が複数個存在しない形態のトランザクションのデータが格納されることになる。

[0017] 本実施形態では、データ格納部１０に格納されるトランザクション群のデータ構造として、各トランザクションを識別するためのトランザクション番号（Ａ０１～Ａ０５）と、当該トランザクションを構成するアイテムの一覧を示すアイテムリスト（この例では購入リスト）と、が対応付けられた構造となっている。Ａ０１乃至Ａ０５の各トランザクションは、購入リストに示す１又は複数種類のアイテムから構成される。すなわち、トランザクションＡ０１では４つのアイテム（換言すると４種類の商品、以下同様）、トランザクションＡ０２では３つのアイテム、トランザクションＡ０３では４つのアイテム、トランザクションＡ０４では２つのアイテム、トランザクションＡ０５では３つのアイテムで構成されていることが分かる。各トランザクションにおいて、各々のアイテムは、カンマ等の所定の記号で区切られることにより識別される。ここでは簡明のため、トランザクション群を構成する全てのトランザクションが複数のアイテムからなる場合を説明するが、１つ以上のアイテムがあればトランザクションとして成立し得る。

[0018] また、データ格納部１０に格納される最小支持度に関するデータは、分析者等により予め設定される数値データであり、本実施形態では、かかる数値は、後述する頻出アイテムを抽出するための基準値（閾値）であるとともに、複数のアイテムで構成された候補パターンから特徴的なパターンを抽出するための基準値（閾値）としても使用される。かかる最小支持度の数値データは、図示しないマウスやキーボードなどの操作入力部の操作により、処理に先立って、使用されるトランザクション群を構成するトランザクションの

数やアイテムの構成等に応じて任意に設定、変更等することができる。

[0019] 以下は最小支持度の数値を４０％とした場合について説明するが、この値に限定されるものではない。また、以下は最小支持度の数値を予め一律４０％に設定した場合について説明するが、パターンの次数ないし長さ（すなわち当該パターンが構成されるアイテムの個数）毎に異なる数値を予め設定することもできる。

[0020] さらに、データ格納部１０に格納されるアイテムの整列の優先度を示すデータは、後述する候補パターン生成の際に参照されるものであり、本例では、優先度の高い方から「鶏肉」、「豚肉」、「牛肉」、「鮪」、「鯵」、「ビール」の順となっている。このデータも、処理に先立って、図示しないマウスやキーボードなどの操作入力部の操作により、任意に設定、変更等することができる。本実施形態の優先度は、パターンを構成する複数のアイテムの整列順を規定する情報であり、例えば、品目のカテゴリ順、カテゴリ内の品目順、辞書や「あいうえお」順等の一定のルールに基づいて複数のアイテムを整列させるためのものである。当該優先度を用いることで、例えば、候補パターンの生成処理を円滑かつ迅速に行うことができる。

[0021] アイテム間情報格納部２０は、後述する一連の処理に先立って、使用されるトランザクション群を構成する各アイテム間の関連性に関する情報（以下「アイテム間知識」ともいう。）のデータを記憶する。アイテム間知識のデータは、各アイテム相互間（同一アイテム間をも含む）の関連度のデータ、より詳細には、関連度の高低を示す数値データであり、本実施形態では、アイテム間の関連度が高いほど大きな数値となる。

[0022] アイテム間情報格納部２０に記憶されるアイテム間知識の一例を図３に示す。本例では、アイテム間知識として、トランザクション群中に存在する各アイテムを行列形式で配置したデータテーブルを使用している。ここで、アイテム間知識は、アイテムとアイテムとの間にどの程度の関係があるかを、０から１までの範囲の数値によって示したものであり、以下は、当該数値を「関連度」と称する。ここで、関連度は、アイテム間の関係が強い（関連度

が高い）程その値が大きくなり、関係が弱い（関連度が低い）程その値が小さくなるように定義される。また、同一のアイテムに対する関連度は、最大値の１が与えられる。かかる関連度の具体的な数値は、図示しないマウスやキーボードなどの操作入力部の操作により、処理に先立って、処理対象となるトランザクションのアイテムの上位概念となるカテゴリの数やカテゴリの構造などに応じて、分析者らが任意に設定、変更等することができる。

[0023] 図３の例では、アイテム間情報格納部２０に記憶されるアイテム間知識として、トランザクション群中に存在するｎ個の各アイテムを行（ｉ）方向及び列（ｊ）方向に整列し、２つのアイテム間の関係を数値で表した行列（以下、この行列を「関連度行列」とも称する。）形式により各アイテム同士の関連度の値を登録する関連度行列テーブルが用いられる。この関連度行列テーブルでは、１を最大値として、アイテム同士の関連度が高くなるほど大きな値が登録される。具体的には、同一のアイテム間（「鶏肉」と「鶏肉」、「豚肉」と「豚肉」など）には最大値である１が登録され、異なるアイテム同士（「鶏肉」と「豚肉」、「鶏肉」と「ビール」など）には、アイテム間の関係の強さに応じた値が登録される。この例では、アイテムの属するカテゴリが同じである場合、すなわち肉類同士（「鶏肉」と「豚肉」、「鶏肉」と「牛肉」、「豚肉」と「牛肉」）の場合、魚類同士（「鮪」と「鯵」）の場合には中程度の関連度として０．５の値が登録される。一方、アイテムの属するカテゴリが異なる場合、すなわち上記肉類に属するアイテムと、魚類に属するアイテムと、飲料品のカテゴリに属するアイテムであるビールと、の間には関連がないものとして０の値が登録される。

[0024] なお、本実施形態において、「アイテム間知識」は、パターン（組合せ）として抽出したくないアイテム間の関連性の値と言い換えることもできる。すなわち、同じアイテム同士の組合せ（「鶏肉」と「鶏肉」、「豚肉」と「豚肉」など）は、分析時にパターンとして抽出しないように最大値（本例では１）が設定され、カテゴリが同じアイテムの組合せ（例えば「鶏肉」と「豚肉」）は、分析時にパターンとしてやや抽出しにくいような値（本例では

０．５）が設定され、さらに、カテゴリが大きく異なるアイテムの組合せ（例えば「鶏肉」と「ビール」）は、分析時にパターンとしてより抽出されやすいような値（本例では０）が設定されることになる。

[0025] アイテム抽出部３０は、データ格納部１０に格納されたトランザクション群のデータを読み出し、読み出されたデータから頻出するアイテムを抽出する処理を行う。具体的には、アイテム抽出部３０は、データ格納部１０から各トランザクション毎に構成するアイテムを抽出し、抽出されたアイテム毎に、その出現頻度すなわちアイテムが出現するトランザクションの個数（以下「アイテム頻度」ともいう）を算出する。算出されたアイテム頻度の情報は、アイテム抽出部３０から候補パターン生成部４０に渡される。さらに、アイテム抽出部３０は、算出されたアイテム頻度に基づいて、当該アイテムに対する支持度を算出し、該算出値が上述したデータ格納部１０に予め設定されている最小支持度（この例では４０％）以上となるアイテムのみを、頻出アイテムとしてパターン格納部８０に格納する。

[0026] ここで、任意の１つのアイテム（ｉｔ）に対する支持度の具体的な算出方法は、下記の数式１に従う。

[0027] [数1]

$$支持度(it)=\frac{アイテムitを含むトランザクションの個数}{トランザクションの総数}\times 100 \cdots (1)$$

[0028] 候補パターン生成部４０は、トランザクション群を参照しつつ、アイテムの集合で構成されたパターンの候補を生成する。具体的には、候補パターン生成部４０は、後述するパターン格納部８０に格納された長さｍのパターン（ｍ＝１の上述した頻出アイテム又は後述する高次（ｍは２以上）のパターン）を読み込み、データ格納部１０内のトランザクション群を参照しながら、頻出アイテムまたはパターンから、所定条件を満たす長さｍ＋１のパターンの候補（候補パターン）を生成する。候補パターン生成部４０が作成する候補パターンには、頻出アイテムを２個ずつ並べた２次パターン（図１０参照）、長さが２以上のパターン（図１１参照）に対して当該パターンに含ま

れるアイテムのうち所定条件を満たすアイテムを加えた３次以上のパターン（図１２参照）があり、かかる所定条件及び候補パターン生成部４０が行う処理の詳細については後述する。

[0029] また、パターン格納部８０に格納された高次のパターンは、トランザクション群中の複数のトランザクションに出現する長さが２以上（２次以上）のパターンの内、アイテムの特徴的な組合せを有するものとして抽出、格納されたものであり、かかる抽出、格納の処理の詳細については後述する。

[0030] 候補頻度算出部５０は、候補パターン生成部４０により生成された候補パターンのトランザクション群中に出現する出現頻度（トランザクションの数）を、候補パターン毎に算出し、該算出された候補パターン毎の頻度の値を候補評価値算出部６０に渡す。

[0031] 候補評価値算出部６０は、候補頻度算出部５０からの候補パターン毎の出現頻度の値と、上述したアイテム間知識（関連度行列テーブル）とを用いて、候補パターンに対する評価値として、当該パターンを構成するアイテム数の増大に対して単調に減少するように、アイテム間の関連性を反映させた評価値（抽出評価値）を算出する。以下、かかる評価値を「関連性支持度」と称する。候補評価値算出部６０は、候補パターン毎に関連性支持度を算出し、該算出された関連性支持度の値を候補評価部７０に渡す。

[0032] 候補評価部７０は、候補評価値算出部６０からの関連性支持度の値が所定の基準値を満たすか否かを候補パターン毎に判定し、かかる基準値を満たすと判定された候補パターンのデータをパターン格納部８０に格納する。本実施形態では、候補評価部７０は、データ格納部１０に設定された最小支持度を参照して、当該候補パターンの関連性支持度の値が最小支持度（本例では４０％）以上であるか否かを判定し、最小支持度以上の候補パターンのデータをパターン格納部８０に格納する処理を行う。かかる処理により、候補パターンの内の「パターン」すなわちアイテムの特徴的な組合せを有するものが抽出され、パターン格納部８０の記憶領域に保存される。

[0033] パターン格納部８０は、アイテム抽出部３０または候補評価部７０により

頻出アイテムまたはパターンのデータが格納されると、当該データが格納された旨を候補パターン生成部４０に通知し、格納されたパターンのデータを候補パターン生成部４０に提供する。また、パターン格納部８０に格納された頻出アイテムまたはパターンのデータは、格納に伴って自動で或いはユーザの操作入力部の操作により、適宜、図示しないＬＣＤなどの表示部に表示したり不図示のプリンタなどで印字出力することができる。

[0034] 以下、フローチャートを参照して、パターン抽出装置１００の詳細な処理内容について説明する。

[0035] 本実施形態のパターン抽出装置１００は、図４のフローチャートのステップＳ１から処理を開始する。ステップＳ１では、アイテム抽出部３０が、データ格納部１０に格納されているトランザクション群の読み込みを行なう。かかる処理により、上述のトランザクションを構成するアイテムのデータが、各トランザクション番号毎にＣＰＵの作業領域（ＲＡＭ）に読み込まれる。図２の例では、トランザクションＡ０１のアイテムとして「鶏肉」、「鮪」、「鯵」、「ビール」のデータが、トランザクションＡ０２のアイテムとして「鶏肉」、「豚肉」、「ビール」のデータが、トランザクションＡ０３のアイテムとして「鶏肉」、「豚肉」、「鮪」、「ビール」のデータが、トランザクションＡ０４のアイテムとして「牛肉」、「鯵」のデータが、トランザクションＡ０５のアイテムとして「鶏肉」、「鮪」、「鯵」のデータが、それぞれ読み込まれる。

[0036] ステップＳ２では、候補評価値算出部６０が、アイテム間情報格納部２０に格納されているアイテム間知識の読み込みを行なう。かかる処理によって、図３で上述した関連度行列テーブルに登録された各アイテム間の数値データがＣＰＵの作業領域（ＲＡＭ）に読み込まれる。読み込まれた数値データは、後述する抽出評価値算出の処理で使用されることになる。

[0037] 続くステップＳ３では、アイテム抽出部３０が図５のサブルーチンに従って、出現頻度の高いアイテム（頻出アイテム）の抽出及び格納の処理を行う。まず、アイテム抽出部３０は、ステップＳ２で読み込まれたトランザクシ

ョンの集合をサーチすることにより、トランザクションを構成するアイテムの種類をすべて抽出する（ステップＳ３１）。例えば、図２のトランザクションの集合の場合、アイテム抽出部３０は、「鶏肉」、「豚肉」、「牛肉」、「鮪」、「鯵」、「ビール」の６種類のアイテムを抽出する。本実施形態では、抽出されたアイテムは、候補アイテムとして扱われ、換言すると、特徴的なパターンの候補となり得る１次（長さ１）の候補パターンとして扱われる（図６参照）。

[0038] 次に、アイテム抽出部３０は、取り出された各アイテムに対して、アイテムごとに、以下のステップＳ３２乃至ステップＳ３５の処理を行う。まず、アイテム抽出部３０は、取り出された各アイテムの内、１つのアイテムについて、ステップＳ２で読み込まれたトランザクション群を参照して、当該アイテムが出現するトランザクションの個数をアイテム頻度として算出する（ステップＳ３２）。例えば、図２のトランザクション群の場合、アイテム「鶏肉」は、トランザクションＡ０１、Ａ０２、Ａ０３、Ａ０５に含まれているため、「鶏肉」の頻度として「４」が算出される。

[0039] 次に、アイテム抽出部３０は、算出された当該アイテムの頻度に基づいて、上述した数式１によりアイテムに対する支持度を算出し（ステップＳ３３）、その値がデータ格納部１０に予め設定されている最小支持度以上となるか否かを判定する（ステップＳ３４）。例えば、頻度が４と算出された上述のアイテム「鶏肉」の場合、トランザクションの総数が５であるため、その支持度は（４／５×１００＝）８０％となる。

[0040] このとき、アイテム抽出部３０は、算出されたアイテムの支持度がデータ格納部１０に格納されている最小支持度（本例では４０％）以上となる場合には、ステップＳ３５で当該アイテムを頻出アイテムとしてパターン格納部８０に格納してステップＳ３６に移行する。一方、アイテム抽出部３０は、算出されたアイテムの支持度が最小支持度（４０％）に満たない場合には、当該アイテムをパターン格納部８０に格納せずに、パターンの対象から除外（当該アイテムのデータを廃棄）してステップＳ３６に移行する。ステップ

Ｓ３６で、アイテム抽出部３０は、ステップＳ３１で抽出された全てのアイテムについての処理が完了したか否かを判定し、未だ処理が完了していない場合にはステップＳ３２に戻って上述したステップＳ３２乃至ステップＳ３５の処理を繰り返し、一方、全てのアイテムについての処理が完了した場合にはステップＳ３７に移行する。

[0041] 従って、支持度が８０％と算出された上述のアイテム「鶏肉」の場合、最小支持度（４０％）以上であるため（ステップＳ３４でＹＥＳ）、頻出アイテムとしてパターン格納部８０に格納される。同様にして、他のアイテム「豚肉」、「牛肉」、「鮪」、「鯵」、「ビール」についても、アイテム抽出部３０によって、頻度が各々「２」、「１」、「３」、「３」、「３」と算出され（ステップＳ３２）、支持度は各々「４０％」、「２０％」、「６０％」、「６０％」、「６０％」と算出され（ステップＳ３３）、図６に示すように、長さ１の候補パターンとして各アイテムの頻度と支持度が算出される。そして、本例では最小支持度が４０％に設定されているので、最小支持度に満たないアイテム「牛肉」だけがパターンの対象から除外され（ステップＳ３４でＮＯ）、図７に示すように、「牛肉」を除く、「鶏肉」、「豚肉」、「鮪」、「鯵」、「ビール」が、頻出アイテムとしてパターン格納部８０に格納されることになる。

[0042] ここで、後述する長さ２以上の候補パターン生成の処理のために、パターン格納部８０に格納される頻出アイテムは、アイテム抽出部３０がデータ格納部１０の上述したアイテムの整列の優先度を示すデータを参照することにより、予め定められた順序に従って整列される。本実施形態では、図６に示すように、「鶏肉」、「豚肉」、「鮪」、「鯵」、「ビール」の優先順で各パターンが整列される。以上のように、ステップＳ３（ステップＳ３１乃至ステップＳ３６）の一連の処理を行うことにより、最小支持度に満たないアイテム（本例では「牛肉」）がパターンの候補から除外され、高次の候補パターンの要素とならなくなるので、コンピュータの処理負担が大幅に減少し、高次のパターン抽出までの時間が短縮化される。なお、本実施形態では、

頻出アイテムは、パターンを構成するアイテムの数（パターンの長さ）が１となる「パターン」（すなわち特徴的な組合せ）とみなされる。

[0043] ステップＳ３７で、アイテム抽出部３０は、パターン格納部８０を参照して、頻出アイテムが存在するか否かについて判定し、存在すると判定された場合には頻出アイテムの抽出に成功したものとしてステップＳ４に移行し、一方、存在しないと判定された場合には頻出アイテムの抽出に失敗したものとして本装置の処理を終了する。すなわち、ステップＳ３の処理で頻出アイテムが１つも抽出されない場合には、本装置における処理が終了する。

[0044] ステップＳ４では、候補パターン生成部４０が図８のサブルーチンに従って、パターンの候補を生成する処理を行う。まず、候補パターン生成部４０は、ステップＳ４１で、パターン格納部８０から取り出すパターンの長さ（ｍの値）を設定する。具体的には、候補パターン生成部４０は、初めてステップＳ４１を実行する場合には、パターンの長さｍの値を１に設定し（ｍ＝１）、２度目以降はパターンの長さｍの値に１を加算する（ｍ＝ｍ＋１）。

[0045] 次に、候補パターン生成部４０は、パターン格納部８０に格納されているパターン（長さ１の頻出アイテムまたは長さ２以上のパターン）の内、前ステップで設定された長さｍのパターンが２個以上あるかを判定し（ステップＳ４２）、ＮＯすなわち０個又は１個しかないと判定された場合には、候補パターンを生成できないものとして処理を終了し、ＹＥＳすなわち２個以上あると判定された場合には、該当するパターンを全て取り出し（ステップＳ４３）、ステップＳ４４に移行する。候補パターン生成部４０は、ステップＳ４４で、取り出された全てのパターンの内、候補パターン生成条件に合致する２つのパターンがあるかを判定し、無いと判定された場合には、候補パターンを生成できないものとして処理を終了し、あると判定された場合にはステップＳ４５に移行する。

[0046] 本実施形態では、ステップＳ４４の候補パターン生成条件として、「先頭からｍ－１個までのパターンが同一のアイテムであり、最後の１つのアイテムが異なっていること」が設定されている。ただし、この前提として、各パ

ターンにおいては、アイテムは予め定められた順序に従って整列される必要がある。本例は、上述したように、「鶏肉」、「豚肉」、「牛肉」、「鮪」、「鯵」、「ビール」の順で優先度が付けられているため、この優先度に従って各パターンが整列される。

[0047] 候補パターン生成部４０は、ステップＳ４４の候補パターン生成条件を満たすパターンを２個取り出し（ステップＳ４５）、異なるアイテムを整列させて、ステップＳ４１で設定されたパターン長ｍよりも１だけ長いパターンの候補（候補パターン）を生成する処理を行う（ステップＳ４６）。

[0048] すなわち、ステップＳ４６で候補パターン生成部４０は、取り出された２個のパターンに共通するｍ－１個のアイテムに対して、相互に異なる最後の２つのアイテムを、アイテムの順序を守るようにして並べた、（ｍ－１＋２＝）ｍ＋１のパターンの長さを持つ候補パターンを１個生成する。続いて、候補パターン生成部４０は、生成された候補パターンを候補頻度算出部５０に提供する（ステップＳ４７）。さらに、候補パターン生成部４０は、長さｍ＋１の候補パターンを全て生成するまでステップＳ４５乃至ステップＳ４８の処理を繰り返し、全て生成したと判定されるとステップＳ４の処理を終了する。すなわち、このようなパターンの抽出とパターンの候補の生成を繰り返し実施することにより、パターン格納部８０に格納されている長さｍの頻出アイテムまたはパターンから、ｍ＋１のパターンの長さを持つすべてのパターンの候補が生成される。

[0049] 以下、候補パターンの生成処理について具体例を挙げて詳述する。

[0050] 例えば、ステップＳ４１でパターンの長さｍが１に設定され、パターン格納部８０には、図７に示す長さが１であるパターン（すなわち頻出アイテム）が格納されているものとする。ここで各アイテムは、上述のように「鶏肉」、「豚肉」、「牛肉」、「鮪」、「鯵」、「ビール」の順で優先度が与えられていることから、パターン格納部８０に格納される頻出アイテムは、「牛肉」を除いて「鶏肉」、「豚肉」、「鮪」、「鯵」、「ビール」の順で並べられている。

[0051] このとき、図７の各パターンは、いずれも長さｍが１であることから、ステップＳ４２を経てステップＳ４３で全て取り出される。一方、ステップＳ４４の候補パターン生成条件については、パターンの長さｍが１の場合には、共通するアイテムの個数は１個もない（すなわち０個である）ため、「前からｍ－１個までのパターンが同一のアイテム」であること及び「最後の１つのアイテムが異なっている」こととなり、図７の任意の２個のパターンの組合せが条件を満たすことになる。このため、候補パターン生成部４０は、上述したアイテムの優先度に従って、「鶏肉、豚肉」、「鶏肉、鮪」、「鶏肉、鯵」、「鶏肉、ビール」、「豚肉、鮪」、「豚肉、鯵」、「豚肉、ビール」、「鮪、鯵」、「鮪、ビール」、「鯵、ビール」の順で２個のパターンを取り出し（ステップＳ４５）、長さが２となる１０種類のパターンの候補（２次の候補パターン）をステップＳ４６で生成する（図１０参照）。当該生成された１０個の候補パターンの情報は、候補頻度算出部５０に提供され、各候補パターン毎に後述するステップＳ５乃至ステップＳ９の処理が繰り返し実行される。そして、かかる１０個全ての候補パターンについての処理が終了すると、ステップＳ９からステップＳ５に処理が戻される。

[0052] ステップＳ５では、候補頻度算出部５０が、候補パターン生成部４０から提供された候補パターンの中で頻度算出の処理が完了していないものがあるかについて判定し、未だ処理が完了していないものがある場合には１つの候補パターンを取り出してステップＳ６に進み、全ての候補パターンの頻度算出処理が完了している場合には、本装置の処理を候補パターン生成に関するステップＳ４へと戻す。

[0053] 以下のステップＳ６からステップＳ１０までは、取り出された１つの候補パターンに対する処理である。まず、ステップＳ６では、候補頻度算出部５０が、ステップＳ１で読み込まれているトランザクションの集合を参照することにより、取り出した１つのパターンの候補に対して、当該候補の出現頻度すなわち当該候補パターンを含むトランザクションの個数を算出する。

[0054] 例えば、図２に示すトランザクションの集合がステップＳ１で読み込まれ

、ステップＳ４で設定されたパターンの長さｍが２であり、ステップＳ６でパターンの候補として「鶏肉、豚肉」が取り出されている場合、当該パターンの候補はＡ０２及びＡ０３のトランザクションに含まれているため、その出現頻度として２が算出される（図１０参照）。同様に、他の候補パターンである「鶏肉、鮪」、「鶏肉、鯵」、「鶏肉、ビール」、「豚肉、鮪」、「豚肉、鯵」、「豚肉、ビール」、「鮪、鯵」、「鮪、ビール」、「鯵、ビール」に関しては、それぞれの出現頻度が、３、２、３、１、０、２、２、２、１と算出される。

[0055] ステップＳ７では、候補評価値算出部６０が、ステップＳ６で算出された候補パターンの出現頻度と、ステップＳ２で読み込まれたアイテム間知識（関連度行列テーブル）を用いて、当該パターンを構成するアイテム間の関連性を評価することにより、その頻度が低い程小さく、関連性の高いアイテムによってパターンが構成されている程小さくなる抽出評価値（以下、「関連性支持度」と称する。）を算出する。

[0056] 詳細には、図９のフローチャートに示すように、候補評価値算出部６０は、ステップ６で算出された出現頻度から当該候補パターンの支持度を算出する(ステップＳ７１)。ここで、候補パターンの支持度の算出は、上述した数式１と同様であり、かかる数式の「アイテムを含むトランザクションの個数」を「パターンを含むトランザクションの個数」と読み替えればよい。ステップＳ７２で、候補評価値算出部６０は、当該候補パターンに含まれる２個のアイテムの組合せを全て抽出する。ステップＳ７３で、候補評価値算出部６０は、ステップＳ２でアイテム間情報記憶部２０から読み込まれたアイテム間知識を参照（識別）して、抽出された組合せに対応する関連度を全て抽出する。続いて、候補評価値算出部６０は、抽出された関連度に基づく加重値を算出し（ステップＳ７４）、算出された加重値をステップＳ７１で算出された支持度に適用することで、当該候補パターンの抽出評価値（関連性支持度ｆ（ｐ））を算出する（ステップＳ７５）。

[0057] ここで、上述した加重値および関連性支持度ｆ（ｐ）は、パターンの長さ

ｍの増大に対して、単調に減少するように定義する必要がある。より詳細には、関連性支持度ｆ（ｐ）は、２つのパターンあるいはパターンの候補ｐ１及びｐ２に対して、ｐ１⊆ｐ２（ｐ１はｐ２の部分集合）の関係が成立する場合、ｆ（ｐ１）≧ｆ（ｐ２）といった関係が成立するように定義する必要がある。言い換えると、候補評価値算出部６０は、パターンの長さｍに対してトレードオフの関係が成り立つように加重値を算出する必要がある。

[0058] このような加重値および関連性支持度の定義や算出式には多様なものが考えられる。例えば、候補評価値算出部６０が算出する加重値の定義として、抽出された関連度を所定値（例えば１）から減じて得られる値を加重値とすることができる。あるいは、候補評価値算出部６０が算出する加重値の定義の例として、抽出された関連度と加重値との合計値を一定（例えば１）に保持したまま、当該合計値と抽出された関連度との差分値を、加重値とすることができる。

[0059] 本実施形態では、候補評価値算出部６０が算出する加重値及び関連度支持度として、下記数式２に示すように関連性支持度ｆ（ｐ）を定義する。

[0060] [数2]

$$f(p)=[1-\max_{it_i,it_j\in p,it_i\neq it_j}\{s(it_i,it_j)\}]\times\frac{\text{パターン}p\text{を含むトランザクションの個数}}{\text{トランザクションの総数}}\times 100\cdots\text{式(2)}$$

[0061] ただし、数式２の第１項において、ｓ（ｉｔ$_i$，ｉｔ$_j$）は、アイテムｉｔ$_i$とアイテムｉｔ$_j$との関連度を表す。また、ｍａｘ｛ｓ（ｉｔ$_i$，ｉｔ$_j$）｝は、パターンを構成する全てのアイテム（ｉｔ$_i$，ｉｔ$_j$）間の関連度の内の最大値である。

[0062] 数式２において、加重値である第１項は、パターンを構成する任意のアイテム間の関連度の最大値（ｍａｘ）を用いており、かかる関連度の最大値を定数１から減算している。このため、加重値である第１項は、パターンの長さｍが増大すると、アイテム間の関連度の最大値が単調に増加し、定数１からかかる最大値を減算することで、単調に減少する値が得られることになる。また、数式２の第２項は、分母（トランザクションの総数）の値が固定値

である一方、分子の値がパターンの長さｍの増大に伴って単調に減少する。このため、第１項に第２項を掛けて定数倍した関連性支持度ｆ（ｐ）は、パターンの長さｍの増大に対して、単調に減少するといえる。

[0063] 例えば、パターンの長さｍが２である「鶏肉、豚肉」の場合、「鶏肉」と「豚肉」の関連度は図３に示すように０．５と設定されている。このため、加重値である数式２の第１項の値は（１－０．５＝）０．５が算出される。また、上述のように、「鶏肉、豚肉」の頻度は２と算出される。このため、「鶏肉、豚肉」の関連性支持度ｆ（ｐ）は、（０．５×２／５×１００＝）２０％と算出される。一方、同じくパターンの長さｍが２である「鶏肉、鮪」の場合、「鶏肉」と「鮪」の関連度は０と設定されているので、数式２の第１項の値は（１－０＝）１が算出される。このため、ステップＳ７で、候補評価値算出部６０は、「鶏肉、鮪」の関連性支持度ｆ（ｐ）として、（１×３／５×１００＝）６０％を算出する（ステップＳ７５）。同様に、「鶏肉、鯵」、「鶏肉、ビール」、「豚肉、鮪」、「豚肉、鯵」、「豚肉、ビール」、「鮪、鯵」、「鮪、ビール」、「鯵、ビール」の関連性支持度は、図１０の関連性支持度の欄に示すように、各々、４０％、６０％、２０％、０％、４０％、２０％、４０％、２０％が算出される。

[0064] ステップＳ８では、候補評価部７０が、データ格納部１０に格納されている最小支持度の値と、算出された候補パターンの関連性支持度ｆ（ｐ）の値とを比較して、当該関連性支持度ｆ（ｐ）の値が閾値である最小支持度の値を満たすか否かを判定する。このとき、候補評価部７０は、当該候補の関連性支持度ｆ（ｐ）が最小支持度（本例では４０％）以上の場合に、当該候補パターンを「パターン」すなわちアイテムの特徴的な組合せを有するものとして登録するために、ステップＳ９に処理を移行する。一方、関連性支持度が最小支持度未満の場合には、当該候補をパターン格納部８０に登録することなくステップＳ５に処理を戻し、次の候補パターンについての処理を行なう。

[0065] ステップＳ９では、候補評価部７０が登録すると判定したパターンの候補

を、アイテムの特徴的な組合せを有するパターンとして、パターン格納部８０に格納する。例えば、図１０に示すように、パターンの長さが２となる候補パターンに対して関連性支持度が算出された場合、本例ではデータ格納部１０に格納されている最小支持度が４０％であるため、図１０に示す１０個の候補パターンの内、「鶏肉、鮪」、「鶏肉、鯵」、「鶏肉、ビール」、「豚肉、ビール」、「鮪、ビール」の候補パターンに対しては、ステップＳ８で基準値を満たすと判定され、当該５つのパターンが、パターンとして図１１に示すようにパターン格納部８０に登録される。一方、「鶏肉、豚肉」、「豚肉、鮪」、「豚肉、鯵」、「鮪、鯵」、「鯵、ビール」の候補パターンに対しては、ステップＳ８で基準値を満たさないと判定され、パターン格納部８０に格納されず、当該候補パターンのデータが廃棄される。このため、「鶏肉、豚肉」、「豚肉、鮪」、「豚肉、鯵」、「鮪、鯵」、「鯵、ビール」の候補パターンは、長さ３の候補パターン生成の対象から除外される。

[0066] そして、本例では、図１０に示す１０個の２次候補パターンの全てについてステップＳ６乃至ステップＳ９の処理が完了すると、ステップＳ５を経てステップＳ４に処理が戻され、長さ３の３次候補パターンの生成が開始される。

[0067] すなわち、パターン格納部８０には、図７に示す頻出アイテムの他に、図１１に示すように、長さ２のパターン（２次のパターン）が、上述したアイテムの優先度に従って、「鶏肉、鮪」、「鶏肉、鯵」、「鶏肉、ビール」、「豚肉、ビール」、「鮪、ビール」の順序で格納されている。この状態から、ステップＳ４１で、候補パターン生成部４０は、パターンの長さｍを２に設定する。

[0068] 続いて候補パターン生成部４０は、ステップＳ４２を経て、長さｍが２である図１１の各パターンを全て取り出し（ステップＳ４３）、長さ２のパターンを２つ組合せるために、候補パターン生成条件の適否を判定する（ステップＳ４４）。

[0069] ステップＳ４４の候補パターン生成条件について、長さｍが２の場合には

、共通するアイテムの個数が最大で１となるため、「鶏肉、鮪」と「鶏肉、鯵」は、「前方のｍ－１個のパターン」換言すると「最初のアイテム」である「鶏肉」が共通しており、かつ最後の１個のアイテムが相互に異なるため、候補パターン生成条件を満たしている。これに対して、「鶏肉、鮪」と「豚肉、ビール」は、「前方のｍ－１個のパターン」すなわち「最初のアイテム」が一致しておらず、このため候補パターン生成条件を満たしていない。

[0070] 候補パターン生成部４０は、このようにしてステップＳ４４で候補パターン生成条件の適否を判定し、「鶏肉、鮪」と「鶏肉、鯵」、「鶏肉、鮪」と「鶏肉、ビール」、「鶏肉、鯵」と「鶏肉、ビール」の３組を、候補パターン生成条件を満たすものと判定し、かかる３組をステップＳ４５で取り出す。さらに、候補パターン生成部４０は、ステップＳ４６で、長さが３となる３次の候補パターンとして、それぞれ、「鶏肉、鮪」と「鶏肉、鯵」から「鶏肉、鮪、鯵」を生成し、「鶏肉、鮪」と「鶏肉、ビール」から「鶏肉、鮪、ビール」を生成し、「鶏肉、鯵」と「鶏肉、ビール」から「鶏肉、鯵、ビール」を生成する（図１２参照）。当該生成された「鶏肉、鮪、鯵」、「鶏肉、鮪、ビール」、「鶏肉、鯵、ビール」の３個の候補パターンの情報は、候補頻度算出部５０に提供され、各候補パターン毎に上述したステップＳ５乃至ステップＳ９の処理が繰り返し実行される。そして、これら３個全ての候補パターンについての処理が終了すると、ステップＳ９からステップＳ５に処理が戻される。

[0071] 詳細には、ステップＳ５で、候補頻度算出部５０は、候補パターン生成部４０から提供された候補パターンから未処理の１つの候補パターン「鶏肉、鮪、鯵」を取り出してステップＳ６の頻度算出処理を行う。この場合、「鶏肉、鮪、鯵」はトランザクションＡ０１及びＡ０５に含まれているため、候補頻度算出部５０によって頻度「２」が算出され（ステップＳ６）、候補評価値算出部６０によって支持度４０（％）が算出される（ステップＳ７１）。

[0072] 続いて、候補評価値算出部６０は、候補パターン「鶏肉、鮪、鯵」から２

個のアイテムの全ての組合わせとして、「鶏肉、鮪」、「鶏肉、鯵」、「鮪、鯵」を抽出し（ステップＳ７２）、かかる組合わせに対応する関連度として、０、０、０．５を抽出し（ステップＳ７３）、抽出された関連度から、上述した数式２の第１項の加重値として、（１－ｍａｘ｛０，　０，　０．５｝）すなわち（１－０．５＝）０．５を算出する（ステップＳ７４）。また、「鶏肉、鮪、鯵」の支持度として「４０」が算出されているので、候補評価値算出部６０は、ステップＳ７５で、「鶏肉、鯵、鮪」の関連性支持度ｆ（ｐ）として、（０．５×４０＝）２０％を算出する（図１２参照）。この場合、関連性支持度が最小支持度（４０％）未満であり（ステップＳ８でＮＯ）、パターンとして登録することなくステップＳ５に処理が戻される。

[0073] 続いて、ステップＳ５で、候補頻度算出部５０は、候補パターン生成部４０から提供された長さ３の候補パターンからの中で未処理の１つの候補パターン「鶏肉、鮪、ビール」を取り出してステップＳ６の頻度算出処理を行う。この場合、「鶏肉、鮪、ビール」はトランザクションＡ０１及びＡ０３に含まれているため、候補頻度算出部５０によって頻度「２」が算出され（ステップＳ６）、候補評価値算出部６０によって支持度４０（％）が算出される（ステップＳ７１）。

[0074] 続いて、候補評価値算出部６０は、候補パターン「鶏肉、鮪、ビール」から２個のアイテムの全ての組合わせとして、「鶏肉、鮪」、「鶏肉、ビール」、「鮪、ビール」を抽出し（ステップＳ７２）、かかる組合わせに対応する関連度として、０、０、０を抽出し（ステップＳ７３）、抽出された関連度から、上述した数式２の第１項の加重値として、（１－ｍａｘ｛０，　０，　０｝）すなわち（１－０＝）１を算出する（ステップＳ７４）。また、「鶏肉、鮪、ビール」の支持度として「４０」が算出されているので、候補評価値算出部６０は、ステップＳ７５で、「鶏肉、鮪、ビール」の関連性支持度ｆ（ｐ）として、（１×４０＝）４０％を算出する（図１２参照）。この場合、関連性支持度が閾値の最小支持度（４０％）以上であるため、「鶏肉、鮪、ビール」の候補パターンは、「パターン」すなわち「アイテムの特

徴的な組合せを有するもの」として、ステップＳ９でパターン格納部８０に格納される（図１３参照）。

[0075] さらに、ステップＳ５で、候補頻度算出部５０は、候補パターン生成部４０から提供された候補パターンからの中で未処理の１つの候補パターン「鶏肉、鯵、ビール」を取り出してステップＳ６の頻度算出処理を行う。この場合、「鶏肉、鯵、ビール」はトランザクションＡ０１のみに含まれているため、候補頻度算出部５０によって頻度「１」が算出され（ステップＳ６）、候補評価値算出部６０によって支持度２０（％）が算出される（ステップＳ７１）。

[0076] 続いて、候補評価値算出部６０は、候補パターン「鶏肉、鯵、ビール」から２個のアイテムの全ての組合わせとして、「鶏肉、鯵」、「鶏肉、ビール」、「鯵、ビール」を抽出し（ステップＳ７２）、かかる組合わせに対応する関連度として、０、０、０を抽出し（ステップＳ７３）、抽出された関連度から、上述した数式２の第１項の加重値として、（１－ｍａｘ｛０，　０，　０｝）すなわち（１－０＝）１を算出する（ステップＳ７４）。また、「鶏肉、鯵、ビール」の支持度として「２０」が算出されているので、候補評価値算出部６０は、ステップＳ７５で、「鶏肉、鯵、ビール」の関連性支持度ｆ（ｐ）として、（１×２０＝）２０％を算出する（図１２参照）。この場合、関連性支持度が最小支持度（４０％）未満であり（ステップＳ８でＮＯ）、パターンとして登録することなくステップＳ５に処理が戻される。

[0077] そして、本例では、図１２に示す３個の３次候補パターンの全てについてステップＳ６乃至ステップＳ９の処理が完了すると、ステップＳ５を経てステップＳ４に処理が戻される。この場合、パターンの長さｍが３に設定されるが（ステップＳ４１）、パターン格納部８０には、図１３に示すように長さｍが３のパターンが１個しか格納されていないために、候補パターン生成条件を満たすパターンの組合せを取り出して長さ４の候補パターンを生成することができない（ステップＳ４２でＮＯ）。したがって、この場合は、候補パターン生成部４０によるステップＳ４５での候補パターンの生成が出来

ないものとして、本装置における処理を終了する。

[0078] 以上のように、パターンの長さが３の場合においては、図１３に示すように、３つの候補パターンの内の「鶏肉、鮪、ビール」のみがパターンとして抽出され、パターン格納部８０に登録されることになる。すなわち、頻度が同じ「２」であっても関連度の高いアイテムが含まれている「鶏肉、鯵、鮪」については、加重値ひいては抽出評価値（関連性支持度）が相対的に低く算出され、抽出対象とならない。また、関連度の低いアイテムのみで構成された「鶏肉、鯵、ビール」についても、頻度が低いために関連性支持度が低く算出され、やはり抽出対象とならない。

[0079] また、上述した例で示したように、長さ２の候補パターン「鶏肉、鯵」、「鶏肉、ビール」、「鯵、ビール」の関連性支持度ｆ（ｐ）は、それぞれ、４０％、６０％、２０％となり（図１０参照）、かかる３つの候補パターンを含む長さ３の候補パターン「鶏肉、鯵、ビール」の関連性支持度ｆ（ｐ）は２０％となる（図１２参照）。従って、数式２を用いた候補評価値算出部６０の演算結果によれば、パターンの長さの増大に伴って、抽出評価値である関連性支持度が単調に減少することを確認することができる。

[0080] 上述のように、本実施形態のパターン抽出装置１００は、候補パターンに対する抽出評価値を算出する際にアイテム間の関連性を考慮し、関連性が高いアイテムを含む候補パターンの加重値を相対的に小さい値となるように算出することにより、関連性が高いアイテムを含む候補パターンが相対的に抽出されにくくなり、分析者にとって自明と思われる、相互に関連性の高いアイテムで構成されたパターンが抽出されることを防止することができ、分析者の興味を惹くと考えられる、相互に関連性の低いアイテムで構成されたパターンを効率良く抽出することができる。

[0081] より具体的には、アイテム間の関連性を考慮せずに、単に最小支持度に基づいてパターンを抽出した場合には、図２のトランザクションから、出現頻度が２である候補パターン、すなわち肉類同士である「鶏肉、豚肉」や魚類同士である「鯵、鮪」もパターンとして抽出されることになる。このような

相互に関連性の高いアイテムで構成されたパターンは、分析者にとって、自明（当たり前）との印象が強いものであり、興味を惹くパターンとならない。これに対して、本実施形態のパターン抽出装置１００の抽出制御では、候補パターン中、ある程度出現頻度があり且つ関連性の低いアイテムで構成されたものが抽出対象となるので、上述の「鶏肉、豚肉」や「鯵、鮪」がパターンとして抽出されることを回避することができる。

[0082] さらに、本実施形態で説明した対象情報としてのトランザクションは、説明の簡明化のため、極めて小規模な構造のものを例示したが、実際には、より多くのアイテムの種類を扱うとともに、大量のトランザクションが対象とされ得る。このため、アイテム間の関連性を考慮せずに、単に最小支持度に基づいてパターンを抽出した場合には、相互に関連性の高いアイテムで構成されたパターンが多数抽出される可能性があり、「豚肉、鯵」といったようなカテゴリの異なる商品（アイテム）で構成されたパターンが、多くの同種の商品のパターンの中に埋もれてしまう虞がある。したがって、アイテム間の関連性を考慮せずに、最小支持度に基づいてパターンを抽出した場合には、分析者の興味を惹くパターンを効率的に発見することが著しく困難であると考えられる。

[0083] これに対し、本実施形態のパターン抽出制御では、上述のように、アイテム間の関連性を考慮し、候補パターンに含まれる各アイテム間の関連度をアイテム間情報記憶部２０から抽出し、抽出された関連度に基づいて加重値を算出し、かかる加重値を当該候補パターンのトランザクション中の出現頻度に基づく支持度に適用して抽出評価値を算出するために、関連性の高くないアイテムで構成されたパターンを効率良く抽出することが可能となる。従って、本実施形態のパターン抽出装置１００によれば、分析者の興味を惹く重要なパターンを、効率的に発見することが可能になる。

[0084] なお、パターン抽出装置の構成は、上述した実施形態に限定されるものではない。例えば、抽出評価値である関連性支持度の算出方法として、数式２を採用したが、単調性を満たすような関連性支持度の定義式としては、下記

の数式３、数式４に示すように、その定義を与えることもできる。

[0085] [数3]

$$f(p)=[1-\min\{1,\sum_{it_i,it_j\in p,it_i\neq it_j}s(it_i,it_j)\}]\times\frac{\text{パターン}p\text{を含むトランザクション の個数}}{\text{トランザクションの総 数}}\times 100\cdot\cdot\cdot\text{式(3)}$$

[0086] [数4]

$$f(p)=\prod_{it_i,it_j\in p,it_i\neq it_j}s(it_i,it_j)\times\frac{\text{パターン}p\text{を含むトランザクションの個数}}{\text{トランザクションの総数}}\times 100\cdot\cdot\cdot\text{式(4)}$$

[0087] ここで、数式３を用いた場合には、加重値となる第１項でアイテム同士の関連度を加算し、該加算値が１以上になると、第１項ひいては当該関連性支持度ｆ（ｐ）の値が０になる。したがって、数式３も、パターンの長さ（アイテム構成数の増大）に対して単調に減少する定義であることがわかる。

[0088] 一方、数式４を用いた場合には、加重値となる第１項はアイテム同士の関連度のかけ算をするので、実施形態の関連度行列の値をそのまま使うと、例えば「鶏肉」と「ビール」の場合に第１項ひいては当該関連性支持度ｆ（ｐ）の値が０になる。したがって、この場合には、アイテム間知識の他の実施形態として、同じアイテム同士の関連度を０とし、関連性が最も低いアイテム同士の関連度を１に設定すればよい。

[0089] また、上述した実施形態では、アイテム抽出部３０によるステップＳ３７の判定で、パターン格納部８０に頻出アイテムが存在しない場合には本装置の処理を終了することとしたが、これに限定されず、ステップＳ３７で頻出アイテムが存在しないと判定された場合に、アイテム抽出部３０が最小支持度の値（上記例では４０％）から所定の値（例えば２０％）だけ減算する処理を行い、該減算後の最小支持度以上の支持度のアイテムを頻出アイテムとして抽出するように、頻出アイテム抽出（ステップＳ３）の処理を再度やり直すこととしてもよい。この場合には、最小支持度の値を減少させて算出した旨及び算出に用いた最小支持度の値を表示部に適宜表示して分析者に知らせるようにすることが好ましい。

[0090] さらに、上述とは逆に、抽出されパターン格納部８０に格納された頻出ア

イテムが非常に多い場合、例えば、抽出対象となった頻出アイテムが、予め設定された所定数以上の場合、或いはステップＳ３１で抽出されたアイテムの内の予め設定された所定割合（％）以上のアイテムが頻出アイテムとして格納された場合に、アイテム抽出部３０が最小支持度の値を所定の値（例えば２０％）だけ増加させる処理を行って、該変更後の最小支持度以上の支持度のアイテムを頻出アイテムとして抽出するように、再度ステップＳ３の処理をやり直すこととしてもよい。この場合にも、最小支持度の値を増加させて算出した旨及び算出に用いた最小支持度の値を表示部に適宜表示して分析者に知らせるようにすることが好ましい。

[0091] さらにまた、長さ２以上の候補パターンに関する最小支持度についても同様の処理を行うことができる。図１４は、パターン抽出装置の他の実施形態の動作を説明するためのフローチャートであり、ステップＳ５で未処理の候補が無いと判定された場合に、ステップＳ５１でパターン格納部８０内のパターンの有無を判別し、パターンがある場合にはステップＳ４に戻り、一方、無い場合にはステップＳ５２で最小支持度の値（上記例では４０％）から所定の値（例えば２０％）だけ減算する処理を行い、該変更後の最小支持度でステップＳ８の判定処理を候補パターン毎に再度行う。

[0092] さらに、パターン格納部８０に格納されたパターンの個数が、予め定められた数未満（過少）の場合、または、予め定められた数を超える（過大の）場合に、最小支持度の値を減少させる処理、または、最小支持度の値を増加させる処理を行って、ステップＳ８の判定処理を候補パターン毎に再度行うこととしてもよい。

[0093] また、上述の実施形態ではアイテム間情報格納部２０に記憶するアイテム間知識として、２つのアイテム間の関係に対して関連度を定義していたが、これに限定されず、アイテム間知識は、アイテムの増大に対して単調性を保持するように与えることにより、３つ以上のアイテムに対して関連度を定義することもできる。

[0094] また、上述の実施形態では、日用品の販売における、購入商品の特徴的な

組合せの発見のためにパターン抽出装置１００を使用する場合について例示して説明したが、これに限定されず、他の多様な業務に使用することができる。例えば、銀行業務における、店舗の特性と事務ミスの種類との間にある特徴的な因果関係の発見に使用する場合の一例として、店舗毎に１個のトランザクションを使用し、当該店舗で発生したミスの種類をアイテムとして使用することができる。また、番組推薦における、視聴者特性と視聴履歴との間にある視聴者の嗜好の発見などの分野に利用する場合の一例として、視聴者毎に１個のトランザクションを使用し、当該視聴者が視聴した番組をアイテムとして使用することができる。

[0095] 上述した各処理は、コンピュータで実行可能なプログラムとして実現することが可能であり、当該プログラムがインストールされたコンピュータは、実施形態に係る各処理を遂行する情報処理装置として動作することが可能である。例えば、不図示の補助記憶装置に当該プログラムが格納され、ＣＰＵ等の制御部が補助記憶装置に格納されたプログラムを主記憶装置に読み出し、主記憶装置に読み出された該プログラムを制御部が実行し、コンピュータに実施形態に係る各処理を動作させることができる。

[0096] また、上記プログラムは、コンピュータ読取可能な記録媒体に記録された状態で、コンピュータに適用することも可能であり、インターネット等のネットワークを通じてコンピュータにダウンロードすることも可能である。コンピュータ読取可能な記録媒体としては、ＣＤ－ＲＯＭ等の光ディスク、ＤＶＤ－ＲＯＭ等の相変化型光ディスク、ＭＯ（Ｍａｇｎｅｔ　Ｏｐｔｉｃａｌ）やＭＤ（Ｍｉｎｉ　Ｄｉｓｋ）などの光磁気ディスク、フロッピー（登録商標）ディスクやリムーバブルハードディスクなどの磁気ディスク、コンパクトフラッシュ（登録商標）、スマートメディア、ＳＤメモリカード、メモリスティック等のメモリカードが挙げられる。また、特別に設計されて構成された集積回路（ＩＣチップ等）等のハードウェア装置も記録媒体として含まれる。

[0097] また、上述の実施形態では、図１に示した各部を１台のコンピュータで構

成したが、これに限定されず、図１に示した各部を適宜異なるサーバ装置等で実現し、ネットワーク等の通信回線を介して接続したコンピュータシステムとして構成することもできる。

[0098] なお、本発明の実施形態を説明したが、当該実施形態は、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。この新規な実施形態は、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、特許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

[0099] １００　パターン抽出装置

１０　データ格納部

２０　アイテム間情報記憶部

３０　アイテム抽出部

４０　候補パターン生成部

５０　候補頻度算出部

６０　候補評価値算出部

７０　候補評価部

８０　パターン格納部

請求の範囲

[請求項1] 複数のアイテムを含む対象情報内に存在する相互に異なる２以上のアイテムの組合せのパターンを抽出するパターン抽出装置であって、

複数の前記対象情報を記憶する第１記憶部と、

前記複数の対象情報それぞれに含まれる各アイテムに基づいて、相互に異なる２以上のアイテムで構成される候補パターンを生成する候補パターン生成部と、

前記生成された候補パターンが前記複数の対象情報それぞれに出現する出現頻度に基づいて、前記候補パターンの抽出評価値を算出する候補評価値算出部と、

前記算出された抽出評価値が所定の閾値を満たす候補パターンを判別し、前記閾値を満たす候補パターンを抽出するパターン抽出部と、を含むとともに、

前記アイテム間の関連度を記憶する第２記憶部をさらに含み、

前記候補評価値算出部は、前記候補パターンに含まれる各アイテム間の関連度を識別し、識別された関連度に基づく加重値及び前記出現頻度に基づいて、前記抽出評価値を算出することを特徴とするパターン抽出装置。

[請求項2] 前記候補評価値算出部は、任意のふたつの候補パターンｐ１、ｐ２において、候補パターンｐ１が候補パターンｐ２の部分集合である場合に、候補パターンｐ２の抽出評価値が候補パターンｐ１の抽出評価値以下になるような単調性が成立する定義に基づいて、前記抽出評価値を算出することを特徴とする請求項１記載のパターン抽出装置。

[請求項3] 前記候補評価値算出部は、前記候補パターンに含まれるアイテム間の関連度を前記第２記憶部から抽出し、前記抽出された関連度を所定値から減じて得られる値を前記加重値として算出することを特徴とする請求項１又は２に記載のパターン抽出装置。

[請求項4] 複数のアイテムを含む対象情報内に存在する相互に異なる２以上の

アイテムの組合せのパターンを抽出するパターン抽出方法であって、

複数の前記対象情報を記憶領域に記憶するステップと、

前記複数の対象情報それぞれに含まれる各アイテムに基づいて、相互に異なる２以上のアイテムで構成される候補パターンを生成するステップと、

前記アイテム間の関連度を記憶領域に記憶するステップと、

前記生成された候補パターンが前記複数の対象情報それぞれに出現する出現頻度に基づいて、前記候補パターンの抽出評価値を算出するステップと、

前記算出された抽出評価値が所定の閾値を満たす候補パターンを判別し、前記閾値を満たす候補パターンを抽出するステップと、

を含み、

前記候補パターンの抽出評価値を算出するステップは、前記候補パターンに含まれる各アイテム間の関連度を識別し、識別された関連度に基づく加重値及び前記出現頻度に基づいて、前記抽出評価値を算出することを特徴とするパターン抽出方法。

［請求項5］ 複数のアイテムを含む対象情報及び前記アイテム間の関連度を所定の記憶領域に記憶し、前記対象情報内に存在する相互に異なる２以上のアイテムの組合せのパターンを抽出するパターン抽出処理を遂行するコンピュータに、

複数の前記対象情報それぞれに含まれる各アイテムに基づいて、相互に異なる２以上のアイテムで構成される候補パターンを生成する機能と、

前記生成された候補パターンが複数の前記対象情報それぞれに出現する出現頻度に基づいて、前記候補パターンの抽出評価値を算出する機能と、

前記算出された抽出評価値が所定の閾値を満たす候補パターンを判別し、前記閾値を満たす候補パターンを抽出する機能と、を実現させ

、

前記候補パターンの抽出評価値を算出する機能は、前記候補パターンに含まれる各アイテム間の関連度を識別し、識別された関連度に基づく加重値及び前記出現頻度に基づいて、前記抽出評価値を算出することを特徴とするプログラム。