明　細　書

発明の名称 ： 注目評価対象抽出装置及びプログラム

技術分野

[0001] 本発明の実施形態は、注目評価対象抽出装置及びプログラムに関する。

背景技術

[0002] コンピュータ及びネットワーク環境では、特定の対象に対する数値的な時系列データが収集可能であるとともに、特定の対象に対するテキスト的な時系列データが入手可能となっている。このような環境においては、数値的な時系列データとテキスト的な時系列データに基づいて、数値的な時系列データの変動を、テキスト的な時系列データによって説明するパターンを発見するとともに、次期における注目すべき評価対象を予測する方法が研究されてきている。

[0003] 例えば、証券市場においては、株価を数値的な時系列データ、銘柄に関連するニュース記事をテキスト的な時系列データとみなすことにより、株価の変動を説明する、ニュース記事から抽出されるパターンを発見するとともに、次期における注目すべき銘柄を利用者に通知し、利用者の銘柄売買に関する意思決定を支援する方法が研究されてきている。

[0004] 評価対象を予測する方法としては、例えば２つの方法が提案されている。

[0005] 第１の方法は、明示的に与えられる属性値によって評価対象を事前に特徴付けておき、評価対象の発生に関するイベントの頻度に基づいた時系列データを生成し、評価対象の重要度あるいは評価対象の属性値の重要度を算出することにより、特定の問題領域において重要な評価対象を抽出する方式である。

[0006] 一方、第２の方法では、評価対象と特定の語の出現回数の経時的変化を対応させて、評価対象と特定の語との関係を視覚的に把握できるように利用者に表示する方式である。

先行技術文献

特許文献

[0007] 特許文献1：特開２００２－２０７７５５号公報

特許文献2：特開２００１－２１６３１１号公報

発明の概要

発明が解決しようとする課題

[0008] しかしながら以上のような２つの方法は、通常は特に問題ないが、本発明者の検討によれば、以下のような不都合がある。

[0009] 例えば、第１の方法は、明示的な属性値によって評価対象が事前に特徴付けられるため、特徴付けが事前に与えられないような評価対象を扱うことができないという不都合がある。また、第１の方法は、評価対象に関連する時系列データが、評価対象の発生に関するイベントの時系列データに限定されるという不都合もある。

[0010] 一方、第２の方法は、評価対象と特定の語との関係を視覚的に把握できるとしても、特定の関係が生じる場合を説明し得るようなパターンを自動的に発見できないという不都合がある。

[0011] 本発明が解決しようとする課題は、特徴付けが事前に与えられない評価対象を扱うことができ、時系列データが評価対象の発生に関するものに限定されず、且つパターンを自動的に発見し得る注目評価対象抽出装置及びプログラムを提供することである。

課題を解決するための手段

[0012] 実施形態の注目評価対象抽出装置は、時系列に沿って与えられる複数の評価対象に関連した複数のテキスト情報からパターンに基づいて注目すべき評価対象を抽出する前に、前記パターンを発見可能な装置である。

[0013] 前記注目評価対象抽出装置は、テキスト収集手段、数値収集手段、対象表現格納手段、アイテム割当て手段、クラス判別手段、トランザクション生成手段、パターン発見手段及びパターン格納手段を備えている。

[0014] 前記テキスト収集手段は、前記複数のテキスト情報を収集する。

[0015] 前記数値収集手段は、前記各評価対象に個別に関連した数値情報を収集する。

[0016] 前記対象表現格納手段は、前記各評価対象を個別に表す評価対象表現と当該各評価対象に個別に関連した対象を表す関連表現とを対応付けて格納する。

[0017] 前記アイテム割当て手段は、前記収集されたテキスト情報毎に、複数のアイテムを抽出すると共に、当該各アイテムのうち、前記評価対象表現又は前記関連表現に一致したアイテムがあるとき、当該評価対象表現又は当該関連表現に前記対象表現格納手段内で対応付けられた評価対象表現を評価対象アイテムとして当該テキスト情報に割り当てる。

[0018] 前記クラス判別手段は、前記評価対象アイテムとしての評価対象表現が表す評価対象に関連した前記数値情報に基づいて、当該数値情報の変化に関連したクラスを判別し、当該クラスを当該評価対象アイテムが割り当てられたテキスト情報に割り当てる。

[0019] 前記トランザクション生成手段は、前記収集されたテキスト情報毎に、前記割り当てられた評価対象アイテム、前記抽出された各アイテム及び前記割り当てられたクラスを含むクラス付きトランザクションを生成する。

[0020] 前記パターン発見手段は、前記トランザクションの集合から特徴的なアイテムの組合せを示す前記パターンを発見する。

[0021] 前記パターン格納手段は、前記発見されたパターンを格納する。

図面の簡単な説明

[0022] [図1]図１は、第１の実施形態に係る注目評価対象抽出装置の構成を模式的に示すブロック図である。

[図2]図２は、同実施形態における動作を説明するためのフローチャートである。

[図3]図３は、同実施形態における数値時系列データの一例を示す模式図である。

[図4]図４は、同実施形態におけるテキスト時系列データの一例を示す模式図

である。

[図5]図５は、同実施形態におけるアイテム集合の一例を示す模式図である。

[図6]図６は、同実施形態における評価対象知識の一例を示す模式図である。

[図7]図７は、同実施形態における数値時系列データの一例を示す模式図である。

[図8]図８は、同実施形態における数値時系列データの一例を示す模式図である。

[図9]図９は、同実施形態におけるクラスの割り当て結果の一例を示す模式図である。

[図10]図１０は、同実施形態におけるクラス付きのトランザクションの一例を示す模式図である。

[図11]図１１は、同実施形態におけるパターンの一例を示す模式図である。

[図12]図１２は、同実施形態におけるパターンの一例を示す模式図である。

[図13]図１３は、第２の実施形態に係る注目評価対象抽出装置の構成を模式的に示すブロック図である。

[図14]図１４は、同実施形態における動作を説明するためのフローチャートである。

[図15]図１５は、同実施形態におけるテキスト時系列データの一例を示す模式図である。

[図16]図１６は、同実施形態におけるアイテム集合の一例を示す模式図である。

[図17]図１７は、同実施形態におけるテキスト時系列データの評価結果の一例を示す模式図である。

発明を実施するための形態

[0023] 以下、各実施形態について図面を用いて説明する。なお、以下の注目評価対象抽出装置は、ハードウェア構成、又はハードウェア資源とソフトウェアとの組合せ構成のいずれでも実施可能となっている。組合せ構成のソフトウェアとしては、図１及び図１３に示す如き、予めネットワーク又は非一時的

なコンピュータ読取可能な記憶媒体（non-transitory computer-readable storage medium）Ｍからコンピュータにインストールされ、当該コンピュータのプロセッサに実行されることにより、当該コンピュータに注目評価対象抽出装置の機能を実現させるためのプログラムが用いられる。

[0024] ＜第１の実施形態＞

図１は第１の実施形態に係る注目評価対象抽出装置の構成を模式的に示すブロック図である。この注目評価対象抽出装置は、時系列に沿って与えられる複数の評価対象に関連した複数のテキスト情報からパターンに基づいて注目すべき評価対象を抽出する前に、当該パターンを発見可能な装置である。

[0025] 具体的には注目評価対象抽出装置は、テキスト時系列収集部Ｂ１、アイテム抽出部Ｂ２、評価対象知識格納部Ｂ３、数値時系列収集部Ｂ４、クラス判別部Ｂ５、トランザクション生成部Ｂ６、パターン発見部Ｂ７、パターン格納部Ｂ８及び評価対象抽出部Ｂ９を備えている。すなわち、注目評価対象抽出装置は、評価対象抽出部Ｂ９によって注目すべき評価対象を抽出する前に、各部Ｂ１～Ｂ８によってパターンを発見可能となっている。

[0026] ここで、テキスト時系列収集部（テキスト収集手段）Ｂ１は、時系列に沿って与えられる複数の評価対象に関連した複数のテキスト情報を収集する機能をもっている。ここで、「テキスト情報」は、「テキスト時系列データ」と呼んでもよい。

[0027] アイテム抽出部（アイテム割当て手段）Ｂ２は、テキスト時系列収集部Ｂ１により収集されたテキスト情報毎に、複数のアイテムを抽出すると共に、当該各アイテムのうち、評価対象知識格納部Ｂ３内の評価対象表現又は関連表現に一致したアイテムがあるとき、当該評価対象表現又は当該関連表現に評価対象知識格納部Ｂ３内で対応付けられた評価対象表現を評価対象アイテムとして当該テキスト情報に割り当てる機能をもっている。

[0028] 評価対象知識格納部（対象表現格納手段）Ｂ３は、各部Ｂ１，Ｂ２，Ｂ４～Ｂ７，Ｂ９から読出／書込可能な記憶装置であって、各評価対象を個別に表す評価対象表現と当該各評価対象に個別に関連した対象を表す関連表現と

を対応付けて格納している。例えば、評価対象知識格納部Ｂ３は、評価対象表現と関連表現とが対応付けて記述された評価対象知識が制御部（図示せず）により予め書込まれている。なお、評価対象知識格納部Ｂ３は、注目評価対象抽出装置のプログラムの一部に予め記述された評価対象知識が起動時に制御部から書き込まれる形態でもよく、ユーザの操作により記述された評価対象知識が書き込まれている形態でもよい。

[0029] 数値時系列収集部（数値収集手段）Ｂ４は、各評価対象に個別に関連した数値情報を収集する機能をもっている。ここで、「数値情報」は、「時系列に沿って与えられる各評価対象に個別に関連した数値情報」と呼んでもよく、「数値時系列データ」と呼んでもよい。また、「テキスト時系列データ」と「数値時系列データ」とを含む時系列データを「複合イベント時系列データ」と呼んでもよい。

[0030] クラス判別部Ｂ５は、アイテム抽出部Ｂ２により割り当てられた評価対象アイテムとしての評価対象表現が表す評価対象に関連した数値情報に基づいて、当該数値情報の変化に関連したクラスを判別し、当該クラスを当該評価対象アイテムが割り当てられたテキスト情報に割り当てる機能をもっている。

[0031] トランザクション生成部Ｂ６は、テキスト時系列収集部Ｂ１により収集されたテキスト情報毎に、アイテム抽出部Ｂ２により割り当てられた評価対象アイテム、アイテム抽出部Ｂ２により抽出された各アイテム及びクラス判別部Ｂ５により割り当てられたクラスを含むトランザクションを生成する機能をもっている。

[0032] パターン発見部Ｂ７は、トランザクション生成部Ｂ６により生成されたトランザクションの集合から特徴的なアイテムの組合せを示すパターンを発見する機能をもっている。なお、「特徴的なアイテムの組合せ」は、「頻度の高いアイテムの組合せ」又は「頻出するアイテムの組合せ」と呼んでもよい。また、発見されたパターンは、パターン発見部Ｂ７によりパターン格納部Ｂ８に書込まれる。

[0033] パターン格納部Ｂ８は、各部Ｂ１，Ｂ２，Ｂ４～Ｂ７，Ｂ９から読出／書込可能な記憶装置であって、パターン発見部Ｂ７により発見されたパターンを格納する。なお、評価対象知識格納部Ｂ３及びパターン格納部Ｂ８としては、同一の記憶装置内の別々の記憶領域を用いてもよく、別々の記憶装置を用いてもよい。

[0034] 評価対象抽出部Ｂ９は、パターン格納部Ｂ８内のパターンを新たに収集されたテキスト情報から生成されたトランザクションに適用することにより、注目すべき評価対象を抽出する機能をもっている。

[0035] 次に、以上のように構成された注目評価対象抽出装置の動作を図２のフローチャートを用いて説明する。以下の説明においては、テキスト時系列データ（テキスト情報）として、ニュース見出しを取り上げ、数値時系列データ（数値情報）として株価を取り上げる。また、評価対象としては、会社名（株価銘柄名）を取り上げることにする。

[0036] ステップＳａ１では、数値時系列収集部Ｂ４がインターネット回線などを通じて、株価情報を配信しているＷｅｂサイトから、評価対象毎の株価情報を示す数値時系列データを収集する。数値時系列データとしては、例えば、図３に示すように銘柄コード、日付、始値を含むデータが用いられる。ただし、数値時系列データは、株式市場が営業している日付のみのデータだけから構成されているとし、株式市場が営業していない日付に関するデータは含まれていないとする。

[0037] ステップＳａ２では、テキスト時系列収集部Ｂ１がインターネット回線などを通じて、ニュース見出しを配信しているＷｅｂサイトから、ニュース見出しを含むテキスト時系列データを収集する。テキスト時系列データとしては、例えば、図４に示すように、ニュースの配信日とニュース見出しを含むデータが用いられる。

[0038] ステップＳａ３では、収集したテキスト時系列データから、テキスト時系列収集部Ｂ１がひとつのテキストを取り出す。このとき、取り出すテキストが存在すれば、ステップＳａ４へと処理を進める一方、取り出すテキストが

存在しない場合には、ステップＳａ１１へと処理を進める。従って、図４に示すテキスト時系列データが収集されている場合、時系列データＴｎが取り出された直後の処理において、本ステップを実施する際に、ステップＳａ１１へと進むことになる。

[0039] ステップＳａ４では、アイテム抽出部Ｂ２が、取り出されたテキストを代表する各表現をアイテムとして抽出する。例えば、アイテム抽出部Ｂ２は、テキストに対して、形態素解析を実施することにより、各表現の品詞を特定し、その品詞として名詞が付与された表現を、アイテムとして抽出する。従って、テキスト番号Ｔ１のテキストに対して、形態素解析を実施して、名詞表現を抽出した場合には、「ａ社」、「会計」、「予約」、「ソフト」、「先行」、「販売」、「開始」といった表現がアイテムとして抽出されることになる。同様の処理を、図４に示す各テキストに実施することにより、図５に示すアイテム集合を各テキストから生成することができる。

[0040] ステップＳａ５では、アイテム抽出部Ｂ２が抽出したアイテムの集合の中からひとつのアイテムを取り出す。このとき、取り出すアイテムが存在する場合には、ステップＳａ６へと進む一方、取り出すアイテムが存在しない場合には、ステップＳａ７へと進む。従って、Ｔ１の場合においては、「開始」のアイテムが取り出された直後の処理において、本ステップを実施した場合に、ステップＳａ７へ移動することになる。

[0041] ステップＳａ６では、アイテム抽出部Ｂ２が、評価対象知識格納部Ｂ３に格納されている評価対象知識を参照することにより、当該アイテムが評価対象と関連するか否かを判定する。

[0042] アイテム抽出部Ｂ２は、評価対象であると判定した場合、当該テキストに、評価対象アイテムを割り当てる。例えば、評価対象知識格納部Ｂ３に、図６に示す評価対象表現と関連表現とから構成された評価対象知識が格納されているとする。このとき、テキスト番号Ｔ１のテキストからアイテム「ａ社」が取り出されるとする。以下、「テキスト番号Ｔ○のテキスト」は「テキストＴ○」ともいう（添字の○は任意の数字を表す。）。同様に、「アイテ

ム集合番号Ｉ○のアイテム集合」は「アイテム集合Ｉ○」ともいう。アイテム抽出部Ｂ２は、評価対象知識の関連表現を参照することにより、当該アイテムが関連表現として登録されていると判定することができる。そこで、アイテム抽出部Ｂ２は、テキストＴ１に対して、関連表現に対応する評価対象表現である「Ａ社」を評価対象アイテムに割り当てる。

[0043] 次に、テキストＴ７からアイテム「鶏肉」が取り出されるとする。アイテム抽出部Ｂ２は、評価対象知識の関連表現を参照することにより、当該アイテムが複数の箇所で関連表現として登録されていると識別することができる。そこで、アイテム抽出部Ｂ２は、テキスト番号Ｔ７のテキストに対して、それぞれの関連表現に対応する評価対象表現である「Ｄ１社」、「Ｄ２社」、「Ｄ３社」を、評価対象アイテムに割り当てる。

[0044] 一方、テキストＴ１のアイテム「会社」やテキストＴ７のアイテム「原産地」の場合には、対応する関連表現が存在しないので、それぞれのテキストに、新たな評価対象アイテムを割り当てることなく、ステップＳａ５に処理を戻すことになる。

[0045] ステップＳａ７では、クラス判別部Ｂ５が、テキストに割り当てられた評価対象アイテムをひとつ取り出す。このとき、取り出す評価対象アイテムが存在すれば、ステップＳａ８へと進む。一方、取り出す評価対象アイテムが存在しない場合は、ステップＳａ９へと進む。例えば、アイテム集合Ｉ１に対して、評価対象アイテム「Ａ社」だけが割り当てられているとすれば、「Ａ社」を取り出した直後の処理では、ステップＳａ９へと進むことになる。

[0046] ステップＳａ８では、クラス判別部Ｂ５が、取り出された評価対象アイテムに対応する数値時系列データを参照することにより、クラス判定の基準となる変動率を算出する。ここで、クラス判別部Ｂ５は、例えば式（１）に基づいて、当該評価対象アイテムに対応する変動率を算出することとする。式（１）において、itが評価対象アイテム、ｄが評価対象を含むテキストの配信日、dyを配信日ｄよりｙ日後の営業日、value（）を評価対象アイテムの指定日における株価の始値を返す関数とする。ただし、dが営業日でない場合は

、一日前の営業日とする。

[数1]

$$f(it,d)=\frac{value(it,dy)-value(it,d)}{value(it,d)} \quad \cdots(1)$$

[0047] ここで、ｙが２と与えられているとした場合について、変動率を算出してみることにする。アイテム集合Ｉ１から「Ａ社」が取り出されており、「Ａ社」に対応する数値時系列データが図３に示すように与えられているとする。このとき、アイテム集合Ｉ１に対応するテキストＴ１の配信日として、2010/4/26を得ることができるとともに、ｄｙとして2010/4/28を得ることができる。図３を参照することにより、value（Ａ社，2010/4/26）、value（Ａ社，2010/4/28）は、それぞれ９０２、８８６と与えられているため、ｆ（Ａ社，2010/4/26）＝（８８６－９０２）／９０２＝－０．０１７７４を得ることができる。次に、アイテム集合Ｉｎから「Ｋ社」が取り出されており、「Ｋ社」に対応する数値時系列データが図７に示すように与えられている場合を考えてみることにする。このとき、アイテム集合Ｉｎに対応するテキストＴｎの配信日として、2010/05/27を得ることができるとともに、ｙ日後の営業日ｄｙとして2010/5/31を得ることができる。図７を参照することにより、value（Ｋ社，2010/5/27）、value（Ｋ社，2010/5/31）は、８０２、８１６と与えられているため、ｆ（Ｋ社，2010/5/27）＝（８１６－８０２）／８０２＝０．０５９２３を得ることができる。同様に、アイテム集合Ｉｎから「Ｌ社」が取り出されており、「Ｌ社」に対応する数値時系列データが図８に示すように与えられている場合には、ｆ（Ｌ社，2010/5/27）＝（９２６－９１４）／９１４＝０．０１３１３と与えられる。

[0048] ステップＳａ９では、クラス判別部Ｂ５がテキストに対応する各アイテムに対して算出された変動率を統合した評価値を算出し、その評価値の値に従ってクラスを判別する。ここで、クラス判別部Ｂ５は、例えば式（２）に基づいて、変動率の平均値である平均変動率を算出し、当該平均変動率を、テ

キストに対応する評価値とする。式（２）において、Ｉはテキストに対応するアイテムの集合、｜　｜はアイテム集合を構成するアイテムの個数を算出する演算とする。ただし、アイテム集合が空集合の場合には、当該平均変動率は算出不能であるとする。

[数2]

$$\frac{1}{|I|}\cdot\sum_{it\in I} f(it,d) \qquad \cdots(2)$$

[0049] また、当該平均値が予め指定してあるクラス判別しきい値Ｔｈ（＞０）に－１掛けた値－Ｔｈ以下であるものを「下降」、Ｔｈ以上のものを「上昇」、－Ｔｈより大きくＴｈより小さいものを「平準」と判定する。ただし、平均変動率が算出不能の場合においては、クラスは割り当てられないとする。

[0050] 例えば、テキストＴ１の場合においては、ひとつのアイテム「Ａ社」のみが割り当てられているので、平均変動率は、－０．０１７７４と与えられる。また、テキストＴｎの場合においては、「Ｋ社」、「Ｌ社」のふたつのアイテムが割り当てられているので、各アイテムの平均値である（０．０５９２３＋０．０１３１３）／２＝０．０３６１８が平均変動率として与えられる。ここで、クラス判別しきい値が０．０１５と与えられているとすれば、テキストＴ１，Ｔｎに対して、クラス「下降」、「上昇」がそれぞれ割り当てられることになる。一方、テキストＴ２、Ｔ８にはアイテムが含まれていないため、平均変動率が算出不能となるため、クラスは割り当てられないことになる。

[0051] 以上のように、クラス判別部Ｂ５は、各テキストに対してクラスを判別し、当該クラスを割り当てることにより、図９に示すように、各テキストに対するクラスの割り当て結果を、テキストとクラスとの間の対応関係を表すデータとして得ることができる。ただし、クラスが割り当てられない場合には、斜線「／」が記載されている。

[0052] ステップＳａ１０では、トランザクション生成部Ｂ６が、クラス付きの各テキストから、当該テキストから抽出されるアイテム及び評価対象アイテム

と、割り当てられたクラスとを組み合わせることにより、トランザクションを生成する。従って、図４のテキストからは、図１０に示すようにクラス付きのトランザクションが生成されることになる。

[0053] ステップＳａ１１では、パターン発見部Ｂ７が、テキストに割り当てられるクラスの中からひとつのクラスを取り出す。このとき、取り出すクラスが存在する場合には、ステップＳａ１２へと処理を進める一方、取り出すクラスが存在しない場合には、本アルゴリズムを終了する。また、本実施形態においては、クラス「上昇」、「下降」のみからパターンの発見を試みることにする。

[0054] ステップＳａ１２では、パターン発見部Ｂ７が、指定されているクラスからなるトランザクションを抽出する。また、パターン発見部Ｂ７は、パターンの頻度に関するしきい値である最小支持度以上となる頻出するアイテムの組合せであるパターンを、参考文献R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, 20th VLDB Conference (1994)に記載のアルゴリズムによってすべて発見し、選択されているクラスに対応するパターンとして、パターン格納部Ｂ８に格納する。

[0055] パターン発見部Ｂ７は、例えば、本アルゴリズムを利用することにより、図１０に示したトランザクション内のクラス「上昇」に対して、「Ｉ社」と「サーバー」のように各アイテムの組み合わせをパターン候補として作成し、当該作成したパターン候補の中で頻度が基準値よりも高いパターン候補をパターンとして残す。これにより、パターン発見部Ｂ７は、図１１に示すように特徴的なアイテムの組合せを示すパターンを発見し、パターン格納部Ｂ８に格納することができる。また、クラス「下降」に対しても同様にしてパターン候補の頻度から、図１２に示すようにパターンを発見し、パターン格納部Ｂ８に格納することができる。

[0056] 評価対象抽出部Ｂ９は、パターン格納部Ｂ８内のパターンを新たに収集されたテキスト時系列データから生成されたトランザクションに適用することにより、注目すべき評価対象を抽出する。

[0057] 上述したように本実施形態によれば、収集されたテキスト情報毎に、複数のアイテムを抽出すると共に、当該各アイテムのうち、評価対象表現又は関連表現に一致したアイテムがあるとき、当該評価対象表現又は当該関連表現に対応付けられた評価対象表現を評価対象アイテムとして当該テキスト情報に割り当て、評価対象アイテムに関連した数値情報の変化に関連したクラスを判別し、テキスト情報毎に、評価対象アイテム、各アイテム及びクラスを含むクラス付きトランザクションを生成し、トランザクションの集合から特徴的なアイテムの組合せを示すパターンを発見する構成により、テキスト情報と数値情報の間に明確な対応関係が記述されていないとしても、テキスト情報と数値情報を適切に関連付けることができ、数値情報の変化を説明可能なパターンをテキスト情報の中から抽出することができる。

[0058] 補足すると、本実施形態によれば、特徴付けが事前に与えられない評価対象を扱うことができ、時系列データが評価対象の発生に関するものに限定されず、且つパターンを自動的に発見することができる。

[0059] また、従来の第２の方法では、発見されたパターンに基づいて、特定の関係が発生する注目すべき評価対象を予測できないという不都合もある。これに対し、本実施形態によれば、評価対象抽出部Ｂ９により、発見されたパターンに基づいて新たなテキスト情報を評価することにより、数値情報の変化が起きる可能性の高い評価対象として、注目すべき評価対象を予測（抽出）することができる。

[0060] ＜第２の実施形態＞

図１３は第２の実施形態に係る注目評価対象抽出装置の構成を模式的に示すブロック図であり、図１と同一部分についてはその詳しい説明を省略し、ここでは異なる部分について主に述べる。

[0061] すなわち、第２の実施形態は、第１の実施形態の具体例であり、評価対象抽出部Ｂ９が、パターン評価部Ｂ１０及び注目評価対象抽出部Ｂ１１を備えた形態となっている。これに伴い、トランザクション生成部Ｂ６は、パターン格納部Ｂ８にパターンが格納された後、テキスト時系列収集部Ｂ１によっ

て新たに収集されたテキスト情報毎に、アイテム抽出部Ｂ２によって割り当てられた評価対象アイテムと抽出された各アイテムとを含み、前述したクラスを含まないクラス無しトランザクションを生成する機能をもっている。なお、「クラス無しトランザクション」は、パターン発見に用いたクラス付きトランザクション（パターン発見用トランザクション）と区別がつけば任意の名称を用いてよい。このため、「クラス無しトランザクション」は、例えば、単に「トランザクション」と呼んでもよく、「クラス無しのトランザクション」、「注目評価対象抽出用トランザクション」又は「新規トランザクション」等と呼んでもよい。

[0062] ここで、パターン評価部（パターン評価手段）Ｂ１０は、パターン格納部Ｂ８内のパターンを当該クラス無しトランザクションに適用し、当該クラス無しトランザクションが当該パターンを含むか否かを評価する機能をもっている。ここで、「評価」は「判定」と呼んでもよい。

[0063] 注目評価対象抽出部（注目評価対象抽出手段）Ｂ１１は、評価の結果、パターンを含む場合には当該クラス無しトランザクション内の評価対象アイテムに関連付けた頻度を加算し、当該頻度の加算結果に基づいて当該頻度に関連付けられた評価対象アイテムを、注目すべき評価対象として抽出する機能をもっている。

[0064] 次に、以上のように構成された注目評価対象抽出装置の動作を図１４のフローチャートを用いて説明する。

[0065] ステップＳｂ１では、テキスト時系列収集部Ｂ１がインターネット回線などを通じて、ニュース見出しを配信しているＷｅｂサイトから、ニュース見出しを含むテキスト時系列データを収集する。テキスト時系列データとしては、例えば、図１５に示すように、ニュースの配信日とニュース見出しを含むデータが用いられる。但し、本実施形態のテキスト時系列データの配信日は、パターン発見部Ｂ７によるパターンの発見に用いたテキスト時系列データの収集日以降の日であるとする。

[0066] ステップＳｂ２では、テキスト時系列収集部Ｂ１が、収集したテキストの

中からひとつのテキストを取り出す。このとき、取り出すテキストが存在する場合には、ステップＳｂ３へと進む一方、取り出すテキストが存在しない場合には、ステップＳｂ９へと進む。例えば、図１５に示すテキストが収集されている場合には、テキストＥｍが取り出された直後の処理において、ステップＳｂ９へと進むことになる。

[0067] ステップＳｂ３では、アイテム抽出部Ｂ２が、取り出されたテキストに対して、図２のステップＳａ４と同様な処理を実施することにより、テキストを代表するアイテムを抽出する。従って、テキストＥ１からは、「ｉ社」、「ミドル」、「エンド」、「サーバー」、「音声」、「認識」、「機能」、「搭載」といったアイテムが抽出されることになる。

[0068] ステップＳｂ４では、アイテム抽出部Ｂ２が、抽出したアイテムの中からひとつのアイテムを順に取り出す。このとき、取り出すアイテムが存在する場合には、ステップＳｂ５へと進む一方、取り出すアイテムが存在しない場合には、ステップＳｂ６へと進む。例えば、テキストＥ１の場合には、アイテム「搭載」が抽出された直後の処理において、ステップＳｂ６へと進むことになる。

[0069] ステップＳｂ５では、アイテム抽出部Ｂ２が、取り出されたアイテムに対して、図２のステップＳａ６と同様な処理を実施することにより、取り出したアイテムに対応する評価対象アイテムが存在するかどうかを評価する。存在する場合には、アイテム抽出部Ｂ２は、当該評価対象アイテムを当該テキストに割り当てる。

[0070] ステップＳｂ６では、トランザクション生成部Ｂ６が、各テキストに対して、テキストから抽出されたアイテムと、テキストに対応する評価対象テキストからなるトランザクションを生成する。本処理は、図２のステップＳａ１０の処理に類似するが、図２のステップＳａ１０の処理の場合とは異なり、各トランザクションにはクラスが割り当てられていない。従って、図１５に示すテキストから、図１６に示すようにクラス無しのトランザクションを生成することができる。

[0071] ステップＳｂ７では、パターン評価部Ｂ１０が、パターン格納部Ｂ８に格納されているパターンの中から順にひとつのパターンを取り出す。このとき、取り出すパターンが存在する場合には、ステップＳａ８へと処理を進める一方、取り出すパターンが存在しない場合には、ステップＳｂ２へと処理を戻す。

[0072] ステップＳｂ８では、パターン評価部Ｂ１０が、取り出されているクラス無しのトランザクションとパターンを比較し、当該パターンを構成するアイテムを、当該クラス無しのトランザクションを構成するアイテムが含んでいるかどうかを評価する。このとき、当該パターンを含んでいるとすれば、当該パターンに対応するクラスを参照し、当該トランザクションに対応するクラス毎の評価対象アイテムの頻度を１加算する。

[0073] 例えば、トランザクションＦ１とパターンｕｐ４が選ばれている場合、パターンｕｐ４を構成するアイテム「Ｉ社」、「サーバー」、「搭載」は、トランザクションＦ１を構成するアイテムに含まれている。また、ｕｐ４のクラスは「上昇」と与えられているので、評価対象アイテム「Ｉ社」、クラス「上昇」の値を１加算する。

[0074] また、トランザクションＦ３とパターンｄｗ５が選択されている場合、パターンｄｗ４を構成する「原産地」、「偽装」は、トランザクションＦ３を構成するアイテムに含まれるため、評価対象アイテム「Ｄ２社」、クラス「下降」の頻度を１加算する。

[0075] 同様に、トランザクションＦｍとパターンｕｐ８が選ばれている場合には、評価対象アイテム「Ｄ３社」、クラス「上昇」及び評価対象アイテム「Ｊ社」、クラス「上昇」の頻度を１加算する。

[0076] すべてのトランザクションとパターンに対してこのようにパターンの評価を行なうことにより、図１７に示すように評価対象アイテムと頻度との関係を得ることができる。

[0077] ステップＳｂ９では、注目評価対象抽出部Ｂ１１が、各トランザクションを各パターンで評価することによって算出された頻度に基づいて、注目評価

対象を決定して出力する。

[0078] 例えば、抽出する注目対象の個数を設定し、頻度の高い評価対象アイテムを、当該個数分順に出力することができる。ここで、図１７に示す評価対象アイテムと頻度との関係に対して、上位３個のものを注目対象として抽出することにした場合、クラス毎に頻度を参照することにより、「Ｄ３社」、「Ｄ２社」、「Ｉ社」といった評価対象アイテムを注目評価対象として出力することができる。なお、注目評価対象抽出部Ｂ１１は、上位のものを抽出する場合に限らず、基準値より高い頻度のものを抽出してもよい。

[0079] 以上のようにすることにより、請求項２に記載の複合イベント時系列からのパターンの発見及び注目評価対象の抽出装置に基づいて、注目評価対象を決定することができる。

[0080] 第２の実施形態の注目評価対象抽出装置は、第２の実施形態に限られたものではなく、例えば以下の［１］～［６］に示す如き、変形例として用いることができる。

[0081] ［１］例えば、図２のステップＳａ１１及びステップＳａ１２において、クラスを選択し、クラスに基づいて分割したトランザクションの部分集合に基づいてパターンを発見する。これに限らず、図２のステップＳａ１１及びステップＳａ１２は、参考文献：櫻井茂明 著, “クラス付きアイテム集合からの頻出パターンの発見”, The 24th Annual Conference of the Japanese Society for Artificial Intelligence, 2010, 2A3-4, <https://kaigi.org/jsai/webprogram/2010/pdf/10.pdf> に記載のアルゴリズムを利用することにより、ＦＰ-tree及びＦＰ-growthに基づいた処理を実行して、トランザクションを分割すること無しに、クラス付きのトランザクションからパターンを発見することもできる。

[0082] 補足すると、パターン発見部Ｂ７は、本アルゴリズムにより、クラス付きのトランザクションに含まれるアイテム集合を木構造の形式に格納したＦＰ-treeを生成する。なお、クラス付きアイテム集合向けのＦＰ-treeにおいては、ＦＰ-treeのヘッダに、パターンに対応するパターンクラスを格納する領域

と、パターンが特徴パターンであるか可能性パターンであるかを識別するフラグと、パターンの特徴支持度が最小支持度以上であるかどうかを示すフラグとが追加されている。ＦＰ-treeの生成後、パターン発見部Ｂ７は、本アルゴリズムのＦＰ-growthを実行することにより、ＦＰ-treeから特定のアイテムで条件付けられたアイテム部分集合を生成し、アイテム部分集合から更にＦＰ-treeを生成する。パターン発見部Ｂ７は、本アルゴリズムにより、このＦＰ-treeの生成とアイテム部分集合の生成を再帰的に繰り返すことにより、パターンを発見する。

[0083] ［２］図２のステップＳａ９において、ひとつのクラス判別しきい値を利用して、３つのクラスに分割しているが、複数のクラス判別しきい値を利用することにより、４分割以上のクラスに分割することもできる。

[0084] ［３］図１４のステップＳｂ８において、パターンを含むトランザクションに対応するすべての評価対象アイテムに対して、その頻度を１加算しているが、複数の評価対象アイテムがひとつのトランザクションから抽出される場合には、評価対象アイテムの個数を重みとして正規化した値を加算することもできる。

[0085] ［４］パターンとトランザクションの両方に出現する評価対象アイテムに対しては、トランザクションのみに出現する評価対象アイテムよりも高い重みを与えて頻度を加算することもできる。

[0086] ［５］図１４のステップＳｂ９において、クラス毎の頻度に基づいて注目評価対象を決定しているが、個々のクラスの頻度を加算することにより、統合した評価値を算出することにより、当該評価値を用いて注目評価対象を決定することもできる。

[0087] ［６］図１４においては、すべてのトランザクションを処理した後で、注目評価対象を出力するようにしているが、トランザクション毎にパターンの評価を行い、パターンを含むトランザクションに対応する評価対象アイテムを注目対象として出力することもできる。

[0088] 上述したように本実施形態によれば、パターンを当該クラス無しトランザ

クションに適用し、当該クラス無しトランザクションが当該パターンを含むか否かを評価し、評価の結果、パターンを含む場合には当該クラス無しトランザクション内の評価対象アイテムに関連付けた頻度を加算し、当該頻度の加算結果に基づいて当該頻度に関連付けられた評価対象アイテムを、注目すべき評価対象として抽出する構成により、第１の実施形態と同様の効果を得ることができる。

[0089] また、パターン発見用のトランザクションとは異なり、クラスを含まないクラス無しトランザクションを生成するので、高速にトランザクションを生成することができる。

[0090] 以上説明した少なくとも一つの実施形態によれば、収集されたテキスト情報毎に、複数のアイテムを抽出すると共に、当該各アイテムのうち、評価対象表現又は関連表現に一致したアイテムがあるとき、当該評価対象表現又は当該関連表現に対応付けられた評価対象表現を評価対象アイテムとして当該テキスト情報に割り当て、評価対象アイテムに関連した数値情報の変化に関連したクラスを判別し、テキスト情報毎に、評価対象アイテム、各アイテム及びクラスを含むクラス付きトランザクションを生成し、トランザクションの集合から特徴的なアイテムの組合せを示すパターンを発見する構成により、特徴付けが事前に与えられない評価対象を扱うことができ、時系列データが評価対象の発生に関するものに限定されず、且つパターンを自動的に発見することができる。

[0091] なお、上記の各実施形態に記載した手法は、コンピュータに実行させることのできるプログラムとして、磁気ディスク（フロッピー（登録商標）ディスク、ハードディスクなど）、光ディスク（ＣＤ－ＲＯＭ、ＤＶＤなど）、光磁気ディスク（ＭＯ）、半導体メモリなどの記憶媒体に格納して頒布することもできる。

[0092] また、この記憶媒体としては、プログラムを記憶でき、かつコンピュータが読み取り可能な記憶媒体であれば、その記憶形式は何れの形態であっても良い。

[0093] また、記憶媒体からコンピュータにインストールされたプログラムの指示に基づきコンピュータ上で稼働しているＯＳ（オペレーティングシステム）や、データベース管理ソフト、ネットワークソフト等のＭＷ（ミドルウェア）等が上記実施形態を実現するための各処理の一部を実行しても良い。

[0094] さらに、各実施形態における記憶媒体は、コンピュータと独立した媒体に限らず、ＬＡＮやインターネット等により伝送されたプログラムをダウンロードして記憶または一時記憶した記憶媒体も含まれる。

[0095] また、記憶媒体は一つに限らず、複数の媒体から上記の各実施形態における処理が実行される場合も本発明における記憶媒体に含まれ、媒体構成は何れの構成であっても良い。

[0096] なお、各実施形態におけるコンピュータは、記憶媒体に記憶されたプログラムに基づき、上記の各実施形態における各処理を実行するものであって、パソコン等の一つからなる装置、複数の装置がネットワーク接続されたシステム等の何れの構成であっても良い。

[0097] また、各実施形態におけるコンピュータとは、パソコンに限らず、情報処理機器に含まれる演算処理装置、マイコン等も含み、プログラムによって本発明の機能を実現することが可能な機器、装置を総称している。

[0098] なお、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、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。

請求の範囲

［請求項1］ 時系列に沿って与えられる複数の評価対象に関連した複数のテキスト情報からパターンに基づいて注目すべき評価対象を抽出する前に、前記パターンを発見可能な注目評価対象抽出装置であって、

前記複数のテキスト情報を収集するテキスト収集手段（Ｂ１）と、

前記各評価対象に個別に関連した数値情報を収集する数値収集手段（Ｂ４）と、

前記各評価対象を個別に表す評価対象表現と当該各評価対象に個別に関連した対象を表す関連表現とを対応付けて格納する対象表現格納手段（Ｂ３）と、

前記収集されたテキスト情報毎に、複数のアイテムを抽出すると共に、当該各アイテムのうち、前記評価対象表現又は前記関連表現に一致したアイテムがあるとき、当該評価対象表現又は当該関連表現に前記対象表現格納手段内で対応付けられた評価対象表現を評価対象アイテムとして当該テキスト情報に割り当てるアイテム割当て手段（Ｂ２）と、

前記評価対象アイテムとしての評価対象表現が表す評価対象に関連した前記数値情報に基づいて、当該数値情報の変化に関連したクラスを判別し、当該クラスを当該評価対象アイテムが割り当てられたテキスト情報に割り当てるクラス判別手段（Ｂ５）と、

前記収集されたテキスト情報毎に、前記割り当てられた評価対象アイテム、前記抽出された各アイテム及び前記割り当てられたクラスを含むクラス付きトランザクションを生成するトランザクション生成手段（Ｂ６）と、

前記トランザクションの集合から特徴的なアイテムの組合せを示す前記パターンを発見するパターン発見手段（Ｂ７）と、

前記発見されたパターンを格納するパターン格納手段（Ｂ８）と

を備えたことを特徴とする注目評価対象抽出装置。

［請求項2］ 請求項１に記載の注目評価対象抽出装置において、

前記パターン格納手段にパターンが格納された後、前記テキスト収集手段によって新たに収集されたテキスト情報毎に、前記アイテム割当て手段によって割り当てられた評価対象アイテムと抽出された各アイテムとを含み、前記クラスを含まないクラス無しトランザクションを生成する手段（Ｂ６）と、

前記パターン格納手段内のパターンを前記クラス無しトランザクションに適用し、当該クラス無しトランザクションが当該パターンを含むか否かを評価するパターン評価手段（Ｂ１０）と、

前記評価の結果、前記パターンを含む場合には当該クラス無しトランザクション内の評価対象アイテムに関連付けた頻度を加算し、前記頻度の加算結果に基づいて当該頻度に関連付けられた評価対象アイテムを、前記注目すべき評価対象として抽出する注目評価対象抽出手段（Ｂ１１）と

を更に備えたことを特徴とする注目評価対象抽出装置。

［請求項3］ 対象表現格納手段（Ｂ３）及びパターン格納手段（Ｂ８）を備え、時系列に沿って与えられる複数の評価対象に関連した複数のテキスト情報からパターンに基づいて注目すべき評価対象を抽出する前に、前記パターンを発見可能な注目評価対象抽出装置のプロセッサにより実行され、非一時的なコンピュータ読取可能な記憶媒体（Ｍ）に記憶されたプログラムであって、

前記複数のテキスト情報を収集する処理（Ｂ１）を前記プロセッサに実行させる第１プログラムコード、

前記各評価対象に個別に関連した数値情報を収集する処理（Ｂ４）を前記プロセッサに実行させる第２プログラムコード、

前記各評価対象を個別に表す評価対象表現と当該各評価対象に個別に関連した対象を表す関連表現とを対応付けて前記対象表現格納手段（Ｂ３）に書込む処理を前記プロセッサに実行させる第３プログラム

コード、

前記収集されたテキスト情報毎に、複数のアイテムを抽出すると共に、当該各アイテムのうち、前記評価対象表現又は前記関連表現に一致したアイテムがあるとき、当該評価対象表現又は当該関連表現に前記対象表現格納手段内で対応付けられた評価対象表現を評価対象アイテムとして当該テキスト情報に割り当てる処理（Ｂ２）を前記プロセッサに実行させる第４プログラムコード、

前記評価対象アイテムとしての評価対象表現が表す評価対象に関連した前記数値情報に基づいて、当該数値情報の変化に関連したクラスを判別し、当該クラスを当該評価対象アイテムが割り当てられたテキスト情報に割り当てる処理（Ｂ５）を前記プロセッサに実行させる第５プログラムコード、

前記収集されたテキスト情報毎に、前記割り当てられた評価対象アイテム、前記抽出された各アイテム及び前記割り当てられたクラスを含むクラス付きトランザクションを生成する処理（Ｂ６）を前記プロセッサに実行させる第６プログラムコード、

前記トランザクションの集合から特徴的なアイテムの組合せを示す前記パターンを発見し、当該発見したパターンを前記パターン格納手段に書込む処理（Ｂ７）を前記プロセッサに実行させる第７プログラムコード、

を備えたことを特徴とするプログラム。