(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7586196号
(P7586196)

(45)発行日　令和6年11月19日(2024.11.19)

(24)登録日　令和6年11月11日(2024.11.11)

(51)国際特許分類
Ｇ０６Ｑ　10/063(2023.01)

ＦＩ
Ｇ０６Ｑ　10/063

請求項の数　10　（全24頁）

(21)出願番号　特願2022-571910(P2022-571910)
(86)(22)出願日　令和3年10月25日(2021.10.25)
(86)国際出願番号　PCT/JP2021/039367
(87)国際公開番号　WO2022/137778
(87)国際公開日　令和4年6月30日(2022.6.30)
審査請求日　令和5年6月14日(2023.6.14)
(31)優先権主張番号　特願2020-212788(P2020-212788)
(32)優先日　令和2年12月22日(2020.12.22)
(33)優先権主張国・地域又は機関
日本国(JP)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　110000338
弁理士法人 ＨＡＲＡＫＥＮＺＯ ＷＯＲＬＤ ＰＡＴＥＮＴ ＆ ＴＲＡＤＥＭＡＲＫ
(72)発明者　野澤 拓磨
東京都港区芝五丁目７番１号 日本電気株式会社内
(72)発明者　小山田 昌史
東京都港区芝五丁目７番１号 日本電気株式会社内
(72)発明者　董 于洋
東京都港区芝五丁目７番１号 日本電気
最終頁に続く

(54)【発明の名称】　情報処理装置、分析方法、および分析プログラム

(57)【特許請求の範囲】
【請求項１】
　複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化する分類手段と、
　グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する評価手段と、
複数の前記インサイトサブジェクトにおけるデータの粒度を統一する粒度統一手段と、を備え、
前記評価手段は、粒度が統一された複数の前記インサイトサブジェクトについて前記評価値を算出する、
情報処理装置。
【請求項２】
　複数の前記インサイトサブジェクトにおける表記を統一する表記統一手段を備え、
　前記分類手段は、表記が統一された前記インサイトサブジェクトをグループ化する、請求項１に記載の情報処理装置。
【請求項３】
前記評価手段は、動的時間伸縮法または関数データ解析により前記評価値を算出する、請求項１に記載の情報処理装置。
【請求項４】

前記評価手段は、グループ化された複数の前記インサイトサブジェクトを主成分分析することにより求めた、各主成分の寄与度の偏りの程度に基づいて前記評価値を算出する、請求項１から３の何れか１項に記載の情報処理装置。
【請求項５】
前記主成分分析により求められた主成分を用いて、グループ化された複数の前記インサイトサブジェクトに含まれるデータを表すことにより、当該データに含まれる外れ値を検出する外れ値検出手段を備える、請求項４に記載の情報処理装置。
【請求項６】
少なくとも１つのプロセッサが、
複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化すること、
グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出すること、および
複数の前記インサイトサブジェクトにおけるデータの粒度を統一すること、を含み、
前記評価値を算出する工程において、前記少なくとも１つのプロセッサが、粒度が統一された複数の前記インサイトサブジェクトについて前記評価値を算出する、
分析方法。
【請求項７】
コンピュータに、
複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化する処理と、
グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する処理と、
複数の前記インサイトサブジェクトにおけるデータの粒度を統一する処理と、を実行させる分析プログラムであって、
前記評価値を算出する処理において、前記コンピュータは、粒度が統一された複数の前記インサイトサブジェクトについて前記評価値を算出する、
分析プログラム。
【請求項８】
複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化する分類手段と、
グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する評価手段であって、グループ化された複数の前記インサイトサブジェクトを主成分分析することにより求めた、各主成分の寄与度の偏りの程度に基づいて前記評価値を算出する評価手段と、
前記主成分分析により求められた主成分を用いて、グループ化された複数の前記インサイトサブジェクトに含まれるデータを表すことにより、当該データに含まれる外れ値を検出する外れ値検出手段と、を備える情報処理装置。
【請求項９】
少なくとも１つのプロセッサが、
複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化すること、および
グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出すること、を含み、
前記評価値を算出する工程において、前記少なくとも１つのプロセッサが、グループ化された複数の前記インサイトサブジェクトを主成分分析することにより求めた、各主成分の

寄与度の偏りの程度に基づいて前記評価値を算出し、
前記少なくとも１つのプロセッサが、前記主成分分析により求められた主成分を用いて、グループ化された複数の前記インサイトサブジェクトに含まれるデータを表すことにより、当該データに含まれる外れ値を検出すること、
を含む分析方法。
【請求項１０】
コンピュータに、
複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化する処理と、
グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する処理であって、グループ化された複数の前記インサイトサブジェクトを主成分分析することにより求めた、各主成分の寄与度の偏りの程度に基づいて前記評価値を算出する処理と、
前記主成分分析により求められた主成分を用いて、グループ化された複数の前記インサイトサブジェクトに含まれるデータを表すことにより、当該データに含まれる外れ値を検出する処理と、を実行させる分析プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データセットの解析を行う情報処理装置等に関する。
【背景技術】
【０００２】
近年、様々な分野において、データを収集し、そのデータを分析することにより、人にとって意味のある知見を見出すことが行われている。このような知見はインサイトと呼ばれる。一般的なデータ分析作業では、分析者が、仮説を設定し、設定した仮説に基づいてデータ分析・可視化し、その仮説検証する、というサイクルを繰り返すことによってインサイトを見出している。
【０００３】
インサイトを見出すための上記のようなデータ分析作業は、非常に時間と労力を要するものであるため、これを自動化する技術の開発が進められている。例えば、下記の特許文献１には、データセットから自動でインサイトを提供するシステムが開示されている。分析者は、特許文献１に記載のシステムに、分析したい多次元データを入力すればよい。これにより、当該システムにより自動的にインサイトが決定され、決定されたインサイトがディスプレイに表示される。
【先行技術文献】
【特許文献】
【０００４】
【文献】米国特許第２０２０／０２５７６８２号明細書
【発明の概要】
【発明が解決しようとする課題】
【０００５】
特許文献１に記載の技術には、複数のデータセット間のインサイトを検出することができないという点で改善の余地があった。例えば、ある企業の製品販売データからなるデータセットと、他の企業についての製品販売データからなるデータセットの両方を解析することにより、一方のデータセットのみからは得られないインサイトが見つかる可能性がある。
【０００６】
しかしながら、特許文献１に記載の技術では、このような複数のデータセット間のインサイトを検出することは想定されていない。このため、当然のことながら、特許文献１に

記載の技術では、複数のデータセット間のインサイトを検出することはできない。

【０００７】

本発明の一態様は、上記の問題に鑑みてなされたものであり、その目的の一例は、複数のデータセット間におけるインサイトの検出を可能にする情報処理装置等を提供することである。

【課題を解決するための手段】

【０００８】

本発明の一態様に係る情報処理装置は、複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化する分類手段と、グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する評価手段とを備える。

【０００９】

本発明の一態様に係る分析方法は、少なくとも１つのプロセッサが、複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化することと、グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出すること、を含む。

【００１０】

本発明の一態様に係る分析プログラムは、複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化する処理と、グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する処理と、をコンピュータに実行させる。

【発明の効果】

【００１１】

本発明の一態様によれば、複数のデータセット間におけるインサイトの検出が可能になる。

【図面の簡単な説明】

【００１２】

【図１】本発明の例示的実施形態１に係る情報処理装置の構成を示すブロック図である。

【図２】本発明の例示的実施形態１に係る分析方法の流れを示すフロー図である。

【図３】本発明の例示的実施形態２に係る情報処理装置が実行する処理の概要を示す図である。

【図４】本発明の例示的実施形態２に係る情報処理装置の構成を示すブロック図である。

【図５】本発明の例示的実施形態２に係る分析方法の流れを示すフロー図である。

【図６】分析対象データと、当該分析対象データから生成されたインサイトサブジェクトの例を示す図である。

【図７】評価結果データと出力データの例を示す図である。

【図８】本発明の例示的実施形態３に係る情報処理装置の構成を示すブロック図である。

【図９】本発明の例示的実施形態３に係る分析方法の流れを示すフロー図である。

【図１０】インサイトスコアの算出方法と、外れ値の検出方法を説明する図である。

【図１１】上記情報処理装置の各機能を実現するソフトウェアであるプログラムの命令を実行するコンピュータの一例を示す図である。

【発明を実施するための形態】

【００１３】

〔例示的実施形態１〕

本発明の第１の例示的実施形態について、図面を参照して詳細に説明する。本例示的実施形態は、後述する例示的実施形態の基本となる形態である。

【００１４】

（情報処理装置１の構成）
本例示的実施形態に係る情報処理装置１の構成について、図１を参照して説明する。図１は、情報処理装置１の構成を示すブロック図である。図示のように、情報処理装置１は、分類部１１と評価部１２を備えている。
【００１５】
分類部１１は、複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化する。グループ化の際に、分類部１１は、評価部１２による評価値の算出が可能なインサイトサブジェクトをグループ化する。なお、以下では、検出対象のインサイトをインサイトタイプと呼ぶ。インサイトタイプは少なくとも１つ設定されていればよい。インサイトタイプの詳細は例示的実施形態２で説明する。
【００１６】
そして、評価部１２は、グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する。以下では、この評価値をインサイトスコアと呼ぶ。
【００１７】
例えば、ある店舗の月間の売上記録を示すデータセットが分析対象である場合、その店舗における日別の総売上を示すデータ（日付と総売上のデータ項目を関連付けたデータ）をインサイトサブジェクトとすることができる。同様に、その店舗におけるある商品の日別の売上を示すデータ（日付とある商品の売上のデータ項目を関連付けたデータ）をインサイトサブジェクトとすることができる。このようなインサイトサブジェクトは、例えばチャート等の形式で可視化することができるため、インサイトサブジェクトを可視化パターンと呼ぶこともできる。インサイトサブジェクトは、多次元データであるデータセットから得られる各可視化パターンを特徴づけるものであると言うこともできる。この場合、１つのインサイトサブジェクトにつき１つの可視化パターンが対応付けられる。
【００１８】
そして、検出対象のインサイト、すなわちインサイトタイプが、例えばインサイトサブジェクト間の相関であれば、分類部１１は、相関の有無を判定するためのインサイトスコア（例えば相関係数）の算出が可能なインサイトサブジェクトをグループ化する。例えば、分類部１１は、上記の例では、各店舗における日付と売上の関係を示すインサイトサブジェクトをグループ化してもよい。これにより、評価部１２は、各店舗における日付と売上についてインサイトスコアを算出することができる。インサイトスコアは、そのまま出力してもユーザがインサイトを発見する大きな助けとなる。また、インサイトスコアを用いることにより、インサイトスコアが高い、すなわちインサイトである可能性が高いインサイトサブジェクトの組み合わせを自動で検出することも可能になる。
【００１９】
以上のように、本例示的実施形態に係る情報処理装置１では、複数のデータセットのそれぞれから生成されたインサイトサブジェクトを、検出対象のインサイトごとにグループ化する分類部１１と、グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する評価部１２と、を備える、という構成が採用されている。
【００２０】
したがって、本例示的実施形態に係る情報処理装置１によれば、複数のデータセット間におけるインサイトの検出が可能になるという効果が得られる。言い換えれば、本例示的実施形態に係る情報処理装置１によれば、複数のデータセットを横断的に分析することで得られる複合インサイト（以下、横断的複合インサイトと呼ぶ）の発見に繋がる可能性のあるデータをユーザに提示することが可能になる。
【００２１】
なお、上述の情報処理装置１の機能は、プログラムによって実現することもできる。本例示的実施形態に係る分析プログラムは、コンピュータに、複数のデータセットのそれぞ

れから生成されたインサイトサブジェクトを、検出対象のインサイトごとにグループ化する処理と、グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する処理と、を実行させる。したがって、本例示的実施形態に係る分析プログラムによれば、複数のデータセット間におけるインサイト、すなわち横断的複合インサイトの検出が可能になるという効果が得られる。

【００２２】

（分析方法の流れ）

本例示的実施形態に係る分析方法の流れについて、図２を参照して説明する。図２は、本例示的実施形態に係る分析方法の流れを示すフロー図である。

【００２３】

Ｓ１１では、少なくとも１つのプロセッサが、複数のデータセットのそれぞれから生成されたインサイトサブジェクトを、インサイトタイプごとにグループ化する。そして、Ｓ１２では、少なくとも１つのプロセッサが、Ｓ１１でグループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値であるインサイトスコアを算出する。これにより、図２の分析方法は終了する。

【００２４】

なお、１つのプロセッサにＳ１１～Ｓ１２の処理を実行させてもよいし、Ｓ１１の処理とＳ１２の処理をそれぞれ別のプロセッサに実行させてもよい。後者の場合、各プロセッサは、１つの情報処理装置が備えているものであってもよいし、それぞれ異なる情報処理装置が備えているものであってもよい。また、Ｓ１１～Ｓ１２の処理を実行する少なくとも１つのプロセッサは、情報処理装置１が備えているものであってもよい。

【００２５】

以上のように、本例示的実施形態に係る分析方法においては、少なくとも１つのプロセッサが、複数のデータセットのそれぞれから生成されたインサイトサブジェクトをインサイトタイプごとにグループ化すること、およびグループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するためのインサイトスコアを算出すること、を含む、という構成が採用されている。このため、本例示的実施形態に係る分析方法によれば、複数のデータセット間におけるインサイト、すなわち横断的複合インサイトの検出が可能になるという効果が得られる。

【００２６】

〔例示的実施形態２〕

（概要）

本発明の第２の例示的実施形態について、図面を参照して詳細に説明する。本例示的実施形態では、複数のデータセットの入力を受け付けて、それらのデータセットについてのインサイトに関する情報を出力する情報処理装置２について説明する。図３は、情報処理装置２が実行する処理の概要を示す図である。

【００２７】

まず、情報処理装置２は、分析対象となる分析対象データ２１１ａと２１１ｂを取得する。分析対象データ２１１ａと２１１ｂは、何れも複数のレコードを含む多次元データのデータセットである。なお、分析対象データ２１１ａと２１１ｂを区別する必要がないときには単に分析対象データ２１１と記載する。図３に示す分析対象データ２１１ａと２１１ｂは何れもテーブル形式のデータである。

【００２８】

次に、情報処理装置２は、取得した分析対象データ２１１ａと２１１ｂのそれぞれからインサイトサブジェクトを生成する。図３の例では、分析対象データ２１１ａからI_1～I_3の３つのインサイトサブジェクトが生成され、分析対象データ２１１ｂからI_4、I_5の２つのインサイトサブジェクトが生成されている。

【００２９】

続いて、情報処理装置２は、生成したインサイトサブジェクトI_1～I_5をグループ化する。図３の例では、インサイトサブジェクトI_1とI_5がグループG^1に分類され、イン

サイトサブジェクトＩ$_3$とＩ$_4$がグループＧ^{2}に分類されている。グループＧ^{1}とＧ^{2}のインサイトタイプは同じであってもよいし、異なっていてもよい。ただし、グループＧ^{1}とＧ^{2}のインサイトタイプが同じである場合には、各グループにはそれぞれ異なるインサイトサブジェクトを分類する。

【００３０】

そして、情報処理装置２は、各グループに含まれるインサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値であるインサイトスコアを算出する。図３の例では、インサイトサブジェクトＩ$_1$とＩ$_5$のインサイトスコアが０．６、インサイトサブジェクトＩ$_3$とＩ$_4$のインサイトスコアが０．９と算出されている。インサイトスコアは、例えばインサイトサブジェクト間の相関の程度を０～１の数値（数値が大きいほど相関の程度が高い）で示すものであってもよい。この場合、インサイトサブジェクトＩ$_3$とＩ$_4$は、相関が高いことになる。

【００３１】

ここで、インサイトサブジェクトＩ$_3$は、分析対象データ２１１ａから生成されたものである。一方、インサイトサブジェクトＩ$_4$は、分析対象データ２１１ｂから生成されたものである。そして、インサイトサブジェクトＩ$_3$とＩ$_4$の相関が高いという知見は、人にとって有用なものである。つまり、情報処理装置２によれば、複数のデータセット間におけるインサイト、すなわち横断的複合インサイトの検出が可能になる。なお、詳細は以下説明するが、情報処理装置２は、相関以外にも様々なインサイトの検出を可能にする。

【００３２】

（情報処理装置２の構成）

図４は、情報処理装置２の構成を示すブロック図である。情報処理装置２は、情報処理装置２の各部を統括して制御する制御部２０と、情報処理装置２が使用する各種データを記憶する記憶部２１を備えている。また、情報処理装置２は、情報処理装置２が他の装置と通信するための通信部２２、情報処理装置２に対する入力を受け付ける入力部２３、および情報処理装置２がデータを出力するための出力部２４を備えている。以下では、出力部２４がデータを表示出力する表示装置である例を説明するが、出力部２４の出力態様は任意であり、例えば印字出力や音声出力等の態様でデータを出力するものであってもよい。また、入力部２３と出力部２４は、情報処理装置２に外付けされた、情報処理装置２の外部の機器であってもよい。

【００３３】

制御部２０には、データ取得部２０１、サブジェクト生成部２０２、表記統一部２０３、分類部２０４、粒度統一部２０５、評価部２０６、および出力データ生成部２０７が含まれている。また、記憶部２１には、分析対象データ２１１、評価結果データ２１２、および出力データ２１３が記憶されている。

【００３４】

分析対象データ２１１は、情報処理装置２による分析対象の対象となるデータである。分析対象データ２１１には、複数のデータセットが含まれている。各データセットは、複数のレコードを含む多次元データである。また、評価結果データ２１２は、評価部２０６による分析対象データ２１１の評価の結果を示すデータである。そして、出力データ２１３は、情報処理装置２による分析対象データ２１１の分析の結果をユーザに提示するためのデータ、すなわち分析対象データ２１１のインサイトに関するデータである。

【００３５】

データ取得部２０１は、情報処理装置２が分析する対象となる複数のデータセットを取得し、それらを分析対象データ２１１として記憶部２１に記憶させる。データ取得部２０１は、分析開始時までに分析対象データ２１１を取得して記憶部２１に記憶させればよい。分析対象データ２１１の取得方法は特に限定されない。例えば、データ取得部２０１は、情報処理装置２のユーザが入力部２３を介して入力したデータセットを取得してもよい。また、例えば、データ取得部２０１は、通信部２２を介した通信により、外部の装置から分析対象データ２１１を取得してもよい。

【００３６】
サブジェクト生成部２０２は、分析対象データ２１１に含まれる複数のデータセットのそれぞれからインサイトサブジェクトを生成する。より詳細には、サブジェクト生成部２０２は、複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることによりインサイトサブジェクトを生成する。例えば、あるデータセットが、日付、売上、および場所のデータ項目を含む多次元データである場合、サブジェクト生成部２０２は、日付と売上を関連付けたインサイトサブジェクトや、場所と売上を関連付けたインサイトサブジェクトを生成する。
【００３７】
表記統一部２０３は、各インサイトサブジェクトにおけるデータの表記を統一する。より詳細には、表記統一部２０３は、各インサイトサブジェクトに含まれる単語の中から類似した単語を抽出し、それらの単語を１つの単語に置き換えることにより、各インサイトサブジェクトにおける表記を統一する。なお、上記「類似」には、単語の文字列の類似の他、意味の類似も含まれる。
【００３８】
例えば、あるデータセットにおいて商品の販売地を表す「東京都」は、他のデータセットにおいて商品の販売地を表す「東京」と意味および文字列が類似した単語であり、これらは表記ゆれと呼ぶこともできる。また、例えば、あるデータセットにおいて商品の販売地を表す「都道府県」は、他のデータセットにおいて商品の販売地を表す「場所」と、意味が類似した単語である。
【００３９】
このような類似の単語を抽出する方法としては任意のものが適用可能である。表記統一部２０３は、「東京」と「東京都」のような表記ゆれの単語を抽出してもよい。この場合、表記統一部２０３は、例えば、単語間の編集距離が近い単語を抽出してもよい。編集距離は、レーベンシュタイン距離とも呼ばれ、２つの文字列がどの程度異なっているかを示す距離である。編集距離を求める際には、表記統一部２０３は、比較対象の一方の単語を構成する文字列に対して何回の変更処理（削除、挿入、置換）を行えば、比較対象の他方を構成する文字列に変換できるかを求める。この他にも、分析対象データ２１１は、例えば２つの文字列の長さと置換の要不要（部分的な一致）を測る距離であるジャロ・ウィンクラー距離に基づいて類似の単語を抽出してもよい。
【００４０】
また、意味が類似した単語を抽出する場合、分析対象データ２１１は、例えば、各データセットに含まれる各単語を分散表現で表し、分散表現の類似度が高い単語を抽出してもよい。分散表現の導出には、例えばword2vec等のプログラムを用いることができる。
【００４１】
表記統一部２０３は、類似した単語を抽出した後、それらの単語の表記を統一する。例えば、表記統一部２０３は、類似する２つの単語のうち一方の単語を他方の単語に全て置換することにより表記を統一してもよい。また、表記統一部２０３は、類似する２つの単語を、それらの単語を包括する上位概念的な単語に置換することにより表記を統一してもよい。
【００４２】
分類部２０４は、サブジェクト生成部２０２が生成したインサイトサブジェクトをグループ化する。より詳細には、分類部２０４は、インサイトの有無を判定するための評価値であるインサイトスコアを算出可能なインサイトサブジェクトをグループ化する。これにより、インサイトスコアに基づいてインサイトを検出することが可能になる。なお、１つのグループには任意の数のインサイトサブジェクトを含めることができる。そして、１つのグループには異なるデータセットから得られたインサイトサブジェクトを含めることができる。１つのグループには少なくとも１つのインサイトサブジェクトを含めることが好ましい。
【００４３】

なお、表記統一部２０３が複数のインサイトサブジェクトにおける表記を統一していた場合、評価部２０６は、表記が統一されたインサイトサブジェクトをグループ化する。異なるデータセット間では、表記が不統一であることも多く、表記が不統一であることが評価の支障となることも一般的には多いが、情報処理装置２によればそのような場合にも評価を行うことができる。つまり、情報処理装置２によれば、例示的実施形態１に係る情報処理装置１の奏する効果に加えて、表記が不統一なデータセットについても横断的複合インサイトを検出することが可能になるという効果が得られる。

【００４４】

例えば、年別の売上を示すインサイトサブジェクトが複数存在する場合、それらのインサイトサブジェクトの系列名は何れも「年」と「売上」となるから、分類部２０４は、それらを１つのグループに分類する。また、このようなインサイトサブジェクトの一部で、系列名が「売上」等の他の表記となっていた場合でも、表記統一部２０３が表記を統一するため、分類部２０４は、それらを１つのグループに分類することができる。

【００４５】

ここで、上記のとおり、グループ化はインサイトタイプごとに行われる。よって、各インサイトタイプについて、グループ化の基準を予め定めておけばよい。インサイトタイプとしては、例えば相関が挙げられる。インサイトタイプが相関であるインサイトサブジェクトをグループ化する場合、分類部２０４は、相関関係の強さを評価できる、言い換えれば相関係数を計算可能なインサイトサブジェクトをグループ化すればよい。また、インサイトタイプが外れ値であるインサイトサブジェクトをグループ化する場合、分類部２０４は、外れ値を検出できるインサイトサブジェクト、つまり対応するデータ間の距離を計算可能なインサイトサブジェクトをグループ化すればよい。具体的には、例えば、分類部２０４は、各系列名を示す単語が同一のインサイトサブジェクトを１つのグループに分類してもよい。

【００４６】

インサイトタイプとしては、相関以外にも任意のものを採用することができる。横断的複合インサイトを検出する場合、例えば、相互メジャー相関（Cross-measure correlation）、二次元クラスタリング、帰属（Attribution）等のインサイトタイプを設定してもよい。

【００４７】

また、例えば、分類部２０４は、シングルポイントインサイト（Single point insight）、すなわち１つのインサイトサブジェクトを入力とする横軸に順序が存在しない（non-ordinal dimension）インサイトサブジェクトをグループ化してもよい。このようなグループ化により、例えば、突出したＮｏ．１（Outstanding No.1）、突出した最下位（Outstanding No. Last）、突出した上位２つ（Outstanding Top 2）、または均一度（Evenness）等のインサイトを検出することが可能になる。

【００４８】

また、分類部２０４は、シングルシェープインサイト（Single shape insight）、すなわち１つのインサイトサブジェクトを入力とする横軸に順序が存在する（ordinal dimension）インサイトサブジェクトをグループ化してもよい。なお、横軸に順序が存在するデータとしては例えば時系列データが挙げられる。このようなグループ化により、変化点（Change point）、トレンド、季節性（Seasonality）、外れ値等のインサイトを検出することが可能になる。設定されるインサイトタイプには、横断的複合インサイトを検出可能なもの（例えば相関等）が少なくとも１つ含まれていればよく、横断的ではない複合インサイトを検出するためのもの（例えば変化点（Change point）等）が含まれていてもよい。

【００４９】

粒度統一部２０５は、各インサイトサブジェクトにおけるデータの粒度を統一する。この処理は、評価部２０６がインサイトサブジェクト間の関連性を評価できるようにするための処理であるから、粒度が揃っていないデータを対象として行われる。粒度の統一は、

データセットから生成されたインサイトサブジェクトに対して行ってもよいし、分析対象となる複数のデータセットに対して予め行っておいてもよい。なお、データの粒度は、一連のデータがどのような細かさ（単位）であるかを示す。

【００５０】

例えば、あるインサイトサブジェクトと他のインサイトサブジェクトが何れも月別の売上を示すものであるが、前者には毎月の売上が示されており、後者には隔月（奇数月）の売上が示されている場合、これらのデータの粒度は一致していない。この場合、両データ間の距離や類似度の評価ができないことがある。

【００５１】

粒度統一部２０５は、このようなデータに対して粒度を揃える処理を行う。例えば、粒度統一部２０５は、欠損値補完によりデータを補完して粒度を揃えてもよいし、ダウンサンプリングにより粒度を揃えてもよい。欠損値補完は、他のデータから欠損部を予測して補完する処理であり、具体例としては内挿等が挙げられる。ダウンサンプリングは、サンプリング粒度を粗い方に合わせる処理である。

【００５２】

上記の例において欠損値補完を行う場合、粒度統一部２０５は、他のインサイトサブジェクトにおける偶数月の売上を補完する。また、上記の例においてダウンサンプリングを行う場合、粒度統一部２０５は、あるインサイトサブジェクトにおける奇数月の売上のみが評価部２０６による評価に用いられるようにする。

【００５３】

評価部２０６は、分類部２０４により同じグループに分類された複数のインサイトサブジェクトの組み合わせについてインサイトスコアを算出し、その算出結果を示す評価結果データ２１２を生成して記憶部２１に記憶させる。例えば、評価部２０６は、同じグループに分類されたインサイトサブジェクトの組み合わせを入力としてインサイトスコアを返す関数f_Tを用いて上記の評価を行ってもよい。

【００５４】

f_Tは、インサイトタイプＴごとに予め定義される関数であり、検出したいインサイトを与えるインサイトサブジェクトが入力されると高い値になるように設計される。インサイトタイプＴに対応するインサイトグループをG_Tとすると、インサイトスコアは下記の式で表される。

【００５５】

$$（インサイトスコア）＝f_T（I_1，I_2，\cdots，I_n \mid I_i \in G_T）$$

評価部２０６は、同じグループに分類された複数のインサイトサブジェクトを組にして、各組のインサイトスコアを算出してもよい。この場合、２つのインサイトサブジェクトを入力とするf_Tを用いればよい。例えば、I_1～I_3の３つのインサイトサブジェクトがグループ化されている場合、評価部２０６は、I_1とI_2、I_1とI_3、およびI_2とI_3の各組をそれぞれf_Tに入力することにより、各組のインサイトスコアを算出する。

【００５６】

インサイトスコアの算出方法は、インサイトタイプに応じたものとすればよい。例えば、組にしたインサイトサブジェクト間の線形な相関の程度を評価する場合、評価部２０６は、ピアソン相関係数を算出するf_Tを用いてインサイトスコアを算出してもよい。この他にも、例えば、評価部２０６は、スピアマン順位相関係数やコサイン類似度、対応するデータ間のユークリッド距離やＥＭＤ（Earth Mover's distance）等をインサイトスコアとして算出してもよい。

【００５７】

なお、粒度統一部２０５がインサイトサブジェクトのデータの粒度を統一していた場合、評価部２０６は、粒度が統一された複数のインサイトサブジェクトの組み合わせについてインサイトスコアを算出する。異なるデータセット間では、データの粒度が不統一であることも多く、粒度が不統一であることが評価の支障となることも一般的には多いが、情報処理装置２によればそのような場合にも評価を行うことができる。すなわち、情報処理

装置２によれば、例示的実施形態１に係る情報処理装置１の奏する効果に加えて、粒度が不統一なデータを含むデータセットについても横断的複合インサイトを検出することが可能になるという効果が得られる。

【００５８】

出力データ生成部２０７は、評価結果データ２１２を用いて出力データ２１３を生成する。出力データ生成部２０７は、情報処理装置２の必須の構成要素ではないが、出力データ生成部２０７を設けることにより、情報処理装置２による分析の結果をより認識しやすい態様でユーザに提示することが可能になる。

【００５９】

（分析方法の流れ）

本例示的実施形態に係る分析方法の流れについて図５～図７を参照して説明する。図５は、分析方法の流れを示すフロー図である。また、図６は、分析対象データ２１１と、当該分析対象データ２１１から生成されたインサイトサブジェクトの例を示す図である。そして、図７は、評価結果データ２１２と出力データ２１３の例を示す図である。

【００６０】

Ｓ２１では、データ取得部２０１が、複数のデータセットの入力を受け付けて、分析対象データ２１１として記憶部２１に記憶させる。例えば、データ取得部２０１は、入力部２３を介して、図６に示す分析対象データ２１１の入力を受け付ける。分析対象データ２１１には、コンビニエンスストアにおける都道府県別の各月の売上を示すデータセット（D^S）と、スーパーマーケットにおける都道府県別の各月の売上を示すデータセット（D^T）が含まれる。

【００６１】

Ｓ２２では、サブジェクト生成部２０２が、分析対象データ２１１に含まれる各データセットからインサイトサブジェクトを生成する。例えば、図６に示すデータセットD^S、D^Tを用いる場合、サブジェクト生成部２０２は、データセットD^SからインサイトサブジェクトI^S_1とI^S_2を生成し、データセットD^TからインサイトサブジェクトI^T_1とI^T_2を生成することができる。

【００６２】

インサイトサブジェクトI^S_1は、コンビニエンスストアにおける都道府県別の売上を示すものであり、図６では、I^S_1を売上の棒グラフ（横軸が都道府県、縦軸が売上）として示している。また、インサイトサブジェクトI^S_2は、コンビニエンスストアにおける月毎の売上を示すものであり、図６では、I^S_2を売上の折れ線グラフ（横軸が日付、縦軸が売上）として示している。

【００６３】

同様に、インサイトサブジェクトI^T_1は、スーパーマーケットにおける都道府県別の売上を示すものであり、図６では、I^T_1を売上の棒グラフ（横軸が都道府県、縦軸が売上）として示している。また、インサイトサブジェクトI^T_2は、スーパーマーケットにおける月毎の売上を示すものであり、図６では、I^T_2を売上の折れ線グラフ（横軸が日付、縦軸が売上）として示している。

【００６４】

インサイトサブジェクトＩは、例えば下記のようなデータ形式とすることもできる。

$I = \{subspace, breakdown, measure, aggregation\}$

上記“subspace”（サブスペース）は、多次元データであるデータセットに含まれるレコードをどのようにフィルタしたかを示す。上記“subspace”は、各チャートの凡例に対応する。例えば、図６のI^S_2の折れ線グラフにおける“subspace”は「東京都」である。フィルタリングを行わないことは、“＊”等の記号で表せばよい。

【００６５】

上記“breakdown”（ブレークダウン）は、多次元データであるデータセットを集計するキーとして使用されるカラムを示す。上記“breakdown”は、各チャートの横軸に対応する。例えば、図６のI^S_2の折れ線グラフにおける“breakdown”は「日付」である。

【００６６】
上記“measure”（メジャー）は、多次元データであるデータセットにおいて数値データとして使用されるカラムを示す。上記“measure”は、各チャートの縦軸に対応する。例えば、図６の$I^{S}{}_{2}$の折れ線グラフにおける“measure”は「売上」の数値データである。
【００６７】
上記“aggregation”（アグリゲーション）は、“breakdown”ごとにデータを集計する際の方法（例えば関数）を示す。上記“aggregation”の例としては、合計、平均、最大値、最小値等が挙げられる。集計に用いられる関数が「合計」である場合、“aggregation”は省略してもよい。
【００６８】
例えば、図６に示す$I^{S}{}_{2}$であれば、$I^{S}{}_{2}$＝｛｛＊，東京都｝，日付，売上｝と表すことができる。Ｓ２２では、サブジェクト生成部２０２は、分析対象データ２１１に含まれる各データセットからこのようなデータ形式のインサイトサブジェクトを生成してもよい。
【００６９】
Ｓ２３では、表記統一部２０３が、Ｓ２２で生成された各インサイトサブジェクトにおけるデータの表記を統一する。例えば、図６に示す$I^{S}{}_{1}$、$I^{S}{}_{2}$、$I^{T}{}_{1}$、$I^{T}{}_{2}$の中では、$I^{S}{}_{1}$における横軸のラベル「都道府県」と、$I^{T}{}_{1}$における横軸のラベル「場所」の意味が類似している。また、$I^{S}{}_{1}$の系列名「東京都」、「大阪府」、「神奈川県」は、$I^{T}{}_{1}$の系列名「東京」、「大阪」、「神奈川」のそれぞれと意味および表記が類似している。表記統一部２０３は、このような単語を抽出し、それらの表記を統一する。例えば、表記統一部２０３は、$I^{S}{}_{1}$における横軸のラベルを「場所」に置換し、系列名「東京都」、「大阪府」、「神奈川県」を、それぞれ「東京」、「大阪」、「神奈川」に置換してもよい。
【００７０】
Ｓ２４では、分類部２０４が、Ｓ２２で生成されたインサイトサブジェクトであって、Ｓ２３で表記が統一されたインサイトサブジェクトをグループ化する。例えば、図６に示す$I^{S}{}_{1}$、$I^{S}{}_{2}$、$I^{T}{}_{1}$、$I^{T}{}_{2}$のうち、縦軸と横軸のラベルが共通するインサイトサブジェクトをグループ化するとする。この場合、分類部２０４は、縦軸のラベルが「売上」で横軸のラベルが「場所」である$I^{S}{}_{1}$と$I^{T}{}_{1}$をグループ化する。$I^{S}{}_{1}$の「都道府県」は表記統一部２０３により「場所」に置換済みであるからこのようなグループ化が可能になっている。また、分類部２０４は、縦軸のラベルが「売上」で横軸のラベルが「日付」である$I^{S}{}_{2}$と$I^{T}{}_{2}$をグループ化する。
【００７１】
$I^{S}{}_{1}$と$I^{T}{}_{1}$を含むグループをG^{1}、$I^{S}{}_{2}$と$I^{T}{}_{2}$を含むグループをG^{2}とすると、グループ化の結果は下記のように表される。
$I^{S}{}_{1}$，$I^{T}{}_{1}\in G^{1}$
$I^{S}{}_{2}$，$I^{T}{}_{2}\in G^{2}$
Ｓ２５では、粒度統一部２０５が、Ｓ２４でグループ化されたインサイトサブジェクトに含まれるデータの粒度を統一する。例えば、図６に示す$I^{S}{}_{2}$の「日付」は、奇数月の１日であるのに対し、$I^{T}{}_{2}$の「日付」は毎月の１日である。粒度統一部２０５は、このように粒度に差異があるデータを抽出し、それらのデータの粒度を揃える処理を行う。例えば、粒度統一部２０５は、$I^{T}{}_{2}$の「日付」のデータのうち、奇数月のデータを抽出（すなわちダウンサンプリング）することにより、「日付」データの粒度を揃えてもよい。また、粒度統一部２０５は、$I^{S}{}_{2}$の偶数月のデータを欠損値補完することにより、「日付」データの粒度を揃えてもよい。なお、欠損値補完は、データのサンプリング日付にずれがある場合にも有効である。例えば、粒度統一部２０５は、毎月１日のデータと、毎月１５日のデータの粒度を揃える場合、毎月１５日のデータを欠損値補完することにより、毎月１日のデータを生成してもよい。
【００７２】
Ｓ２６では、評価部２０６が、Ｓ２４でグループ化され、Ｓ２５でデータの粒度が統一

されたインサイトサブジェクトの組み合わせを評価し、評価結果を評価結果データ２１２として記憶部２１に記憶させる。より詳細には、評価部２０６は、同じグループに含まれるインサイトサブジェクトを組にして、その組についてのインサイトスコアを算出する、という処理を各グループについて行う。

【００７３】

例えば、評価部２０６は、$f_T(I_i, I_j)$の式で表されるスコア関数、すなわち評価対象とする２つのインサイトサブジェクトを入力とし、インサイトスコアを出力とする関数を用いてインサイトスコアを算出してもよい。このスコア関数を用いる場合、グループG^1のインサイトスコアは$f_T(I^S{}_1, I^T{}_1)$、グループG^2のインサイトスコアは$f_T(I^S{}_2, I^T{}_2)$と表される。

【００７４】

評価部２０６は、上述のような評価結果をリスト化することにより、例えば図７に示すような評価結果データ２１２を生成してもよい。図７に示す評価結果データ２１２は、インサイトサブジェクトの組み合わせと、その組み合わせについて算出されたインサイトスコアとを示すテーブル形式のデータである。また、図７に示す評価結果データ２１２には、インサイトスコアの順位を示す「ランク」と、「インサイトタイプ」についても示されている。このように、評価部２０６は、インサイトサブジェクトの組み合わせと、その組み合わせについて算出されたインサイトスコアに加えて、評価に関する各種情報を含む評価結果データ２１２を生成してもよい。

【００７５】

Ｓ２７では、出力データ生成部２０７が、Ｓ２６で生成された評価結果データ２１２を用いて出力データ２１３を生成し、出力部２４に出力させる。例えば、図７に示す評価結果データ２１２を用いる場合、出力データ生成部２０７は、インサイトスコア（ランク）が最も高いインサイトサブジェクトの組み合わせを示す出力データ２１３を生成し、出力部２４に出力させる。これにより、図５の処理は終了する。

【００７６】

出力データ２１３は、インサイトをユーザが認識しやすいように、当該インサイトを可視化したものであってもよい。可視化方法は、インサイトタイプに応じて決定すればよい。例えば、出力データ生成部２０７は、インサイトタイプが「相関」である場合、インサイトに関する情報として相関関係を表すのに適したチャート（例えば二次元の散布図）を出力データ２１３として生成してもよい。

【００７７】

図７の下側には、評価結果データ２１２に示されるインサイトサブジェクトの組み合わせのうち、最もインサイトスコアが高かった（つまり、ランクが１の）ものについてのインサイトに関する情報の例を示している。具体的には、図７に示されるインサイトに関する情報には、スーパーマーケットとコンビニエンスストアの売上の相関を示す散布図と、インサイトの詳細を示すインサイト情報とが含まれている。インサイト情報には、インサイトタイプとインサイトスコアの他、各インサイトサブジェクトの詳細とその元になったデータセットが示されている。このような情報を出力部２４に出力させることにより、情報処理装置２のユーザに、スーパーマーケットとコンビニエンスストアの売上の推移に強い相関がある、というインサイトを容易に認識させることができる。

【００７８】

無論、出力データ生成部２０７が生成する情報は、インサイトをユーザに認識させることができるようなものであればよく、図７の例に限られない。例えば、出力データ生成部２０７は、最もインサイトスコアが高かったインサイトサブジェクトの組み合わせについて、各インサイトサブジェクトのチャートを生成し、これを出力データ２１３としてもよい。

【００７９】

なお、分析結果をユーザに提示する際に、必ずしも新たな出力データ２１３を生成する必要はない。例えば、評価部２０６が、図７に示す評価結果データ２１２の全部または一

部を出力部２４に出力させることにより、分析結果をユーザに提示してもよい。また、評価部２０６は、ランクが１となった各インサイトサブジェクトや、インサイトスコアが所定の閾値以上となった各インサイトサブジェクトを構成するデータを出力させてもよい。このように、分析結果を出力させる態様は任意であり、図７のような例に限定されない。また、分析結果の可視化方法をユーザに選択させてもよい。この場合、出力データ生成部２０７は、ユーザが選択した方法で分析結果を可視化する。

【００８０】

このように、情報処理装置２は、複数のデータセットの分析結果として、インサイトの発見に繋がる可能性のあるチャートやデータ等を出力することができる。これにより、人手でチャートを比較する必要がなくなる。また、最終的にはインサイトをユーザが検討する場合であっても、分析に役立ちそうなデータセットを容易に絞り込むことができる。よって、分析・可視化に要する時間を大幅に短縮することができる。

【００８１】

また、情報処理装置２を用いることにより、全ての分析をユーザが行う場合に生じる判断基準のブレが発生する余地もない。さらに、分析をユーザが行う場合に生じる見逃しのリスク等も低減することができる。また、大規模なデータセットが分析対象である場合、ユーザによる複合インサイトの発見は困難であるが、情報処理装置２によれば、複合インサイト（横断的複合インサイトも含む）の発見が容易になる。

【００８２】

なお、図５のフローチャートにおいて、Ｓ２３の処理は、Ｓ２４の処理よりも先に行えばよく、例えばＳ２１とＳ２２の間に行ってもよい。また、Ｓ２５の処理は、Ｓ２６の処理よりも先に行えばよく、例えばＳ２１とＳ２２の間に行ってもよい。

【００８３】

（粒度の違いへの対応の変形例）

評価部２０６は、データの粒度が異なる複数のインサイトサブジェクトの組み合わせについてもインサイトスコアを算出可能な評価方法により、インサイトサブジェクトを評価してもよい。これにより、例示的実施形態１に係る情報処理装置１の奏する効果に加えて、粒度が不統一なデータを含むデータセットについても横断的複合インサイトを検出することが可能になるという効果が得られる。また、この場合、粒度統一部２０５を省略することができるという効果も得られる。

【００８４】

例えば、インサイトサブジェクトにおける横軸のデータに順序が存在する（ordinal dimensionである）場合には、評価部２０６は、ＤＴＷ（Dynamic Time Warping：動的時間伸縮法）や関数データ解析によりインサイトスコアを算出してもよい。なお、順序が存在するデータの例としては、例えば時系列データ等が挙げられる。ＤＴＷでは、$s=(s_1,\cdots,s_n)$と$t=(t_1,\cdots,t_m)$の要素間の距離を総当りで計算したコスト行列Ｗの端（１，１）から端（ｎ，ｎ）の最短経路を動的計画法で求める。ＤＴＷによれば、サンプルサイズが異なるデータ間の距離や類似度を計算可能であり、そのような距離や類似度をインサイトスコアの計算に用いることができる。また、関数データ解析を用いる場合、評価部２０６は、各インサイトサブジェクトのレコードを表現する連続的な関数を導出し、その関数を介してインサイトサブジェクト間の距離や類似度を計算し、それらをインサイトスコアの計算に用いることができる。

【００８５】

〔例示的実施形態３〕

本発明の第３の例示的実施形態について、図面を参照して詳細に説明する。上述の例示的実施形態において、インサイトサブジェクトをグループ化したときに、３つ以上のインサイトサブジェクトが１つのグループに分類されることがあり得る。このような場合、上述したスコア関数$f_T(I_i,I_j)$では、３つ以上のインサイトサブジェクトをまとめて評価することはできない。また、３つ以上のインサイトサブジェクトをまとめて評価する方法については、特許文献１にも記載も示唆もされていない。

【００８６】
本例示的実施形態では、３つ以上のインサイトサブジェクトをまとめて評価することが可能な評価方法について図８～図１０に基づいて説明する。図８は、本例示的実施形態に係る情報処理装置３の構成を示すブロック図である。図９は、本例示的実施形態に係る分析方法の流れを示すフロー図である。図１０は、インサイトスコアの算出方法と、外れ値の検出方法を説明する図である。
【００８７】
（情報処理装置３の構成）
図８に示すように、情報処理装置３は、評価部３１と外れ値検出部３２を備えている。なお、外れ値を検出する必要がない場合には外れ値検出部３２を省略してもよい。評価部３１は、図１に示した評価部１２および図４に示した評価部２０６と同様に、グループ化された複数のインサイトサブジェクトの組み合わせについてインサイトスコアを算出する。評価部３１は、３つ以上のインサイトサブジェクトをまとめて評価することができる点、言い換えれば３つ以上のインサイトサブジェクトにおけるインサイトの有無を示す１つのインサイトスコアを算出できる点で、評価部１２、２０６と相違している。
【００８８】
具体的には、評価部３１は、グループ化された複数のインサイトサブジェクトを主成分分析することにより求めた、各主成分の寄与度の偏りの程度に基づいて当該インサイトサブジェクトの組み合わせについてのインサイトスコアを算出する。主成分分析は、任意の数のインサイトサブジェクトを対象として行うことができる。このため、本例示的実施形態に係る情報処理装置３によれば、例示的実施形態１、２に係る情報処理装置１、２の奏する効果に加えて、３つ以上のインサイトサブジェクトをまとめて評価することが可能になるという効果が得られる。なお、評価方法の詳細およびこのような評価が可能である理由については、図９および図１０に基づいて後述する。
【００８９】
外れ値検出部３２は、評価部３１による主成分分析により求められた主成分を用いて、グループ化された複数のインサイトサブジェクトに含まれるデータを表すことにより、当該データに含まれる外れ値を検出する。このため、本例示的実施形態に係る情報処理装置３によれば、例示的実施形態１、２に係る情報処理装置１、２の奏する効果に加えて、評価のために行った主成分分析の結果を利用した効率のよい外れ値検出ができるという効果が得られる。なお、外れ値検出方法の詳細およびこのような方法で外れ値を検出することが可能である理由については、図９および図１０に基づいて後述する。
【００９０】
（情報処理装置３が実行する処理の流れ）
情報処理装置３が実行する処理の流れを図９に基づいて説明する。なお、図９の処理の前に、複数のインサイトサブジェクトがグループ化済であるとする。つまり、図８には示していないが、本例示的実施形態では、情報処理装置３が分類部１１（例示的実施形態１）または分類部２０４（例示的実施形態２）に相当する構成を備えていることを想定している。なお、情報処理装置３は、情報処理装置２が備える各種構成（例えば、データ取得部２０１やサブジェクト生成部２０２等）の一部または全部を備えていてもよい。
【００９１】
Ｓ３１では、評価部３１が、インサイトサブジェクトのグループを評価する。より詳細には、まず、評価部３１は、評価対象のグループに含まれる各インサイトサブジェクトにおける、主成分分析の対象とするデータを特定する。例えば、インサイトサブジェクトがＩ＝｛subspace, breakdown, measure, aggregation｝の形式で表されていた場合、評価部３１は、各インサイトサブジェクトにおける“measure”の項目のデータを主成分分析の対象とすればよい。
【００９２】
次に、評価部３１は、主成分分析の対象として特定したデータについて主成分分析を行う。例えば、評価部３１は、各インサイトサブジェクトにおける“measure”の項目のデー

タから多次元の相関行列を生成し、この相関行列を用いて主成分分析を行ってもよい。主成分分析により、固有値と固有ベクトルが算出される。

【００９３】

続いて、評価部３１は、算出された固有値を用いて、各主成分の寄与率を算出する。各主成分の寄与率はその軸方向（固有ベクトル）における情報量とみなすことができるから、各主成分の寄与率の偏り度合いを調べることで、インサイトサブジェクト間の相関の強さを定量的に評価することができる。

【００９４】

例えば、図１０には、相関がないインサイトサブジェクトを主成分分析して算出された各主成分の寄与率を示す棒グラフ１００１と、相関があるインサイトサブジェクトを主成分分析して算出された各主成分の寄与率を示す棒グラフ１００２を示している。なお、図１０において、ＰＣ１は第１主成分、ＰＣ２は第２主成分、ＰＣ３は第３主成分である。

【００９５】

棒グラフ１００１では、ＰＣ１～ＰＣ３の寄与率は概ね同程度であり、主成分間での偏り度合いは小さい。一方、棒グラフ１００２では、ＰＣ１の寄与率が最も高く、ＰＣ２の寄与率はその半分程度であり、ＰＣ３の寄与率はかなり小さく、全体として偏り度合いが大きい。このように、インサイトサブジェクト間の相関の有無は、各主成分の寄与率の偏り度合いに明瞭に反映される。

【００９６】

したがって、各主成分の寄与率の偏り度合いを定量的に評価すれば、その評価結果をインサイトスコアとすることができる。例えば、第１主成分の寄与率をインサイトスコアとしてもよい。これは、図１０に示されるように、各主成分の寄与率の偏り度合いが大きい場合（棒グラフ１００２）には、小さい場合（棒グラフ１００１）と比べて第１主成分ＰＣ１の寄与率が大きいためである。

【００９７】

また、図１０に示されるように、各主成分の寄与率の偏り度合いが大きい場合（棒グラフ１００２）には、ＰＣ１～ＰＣ３の中で寄与率が突出して高いもの（具体的にはＰＣ１）が存在する。一方、各主成分の寄与率の偏り度合いが小さい場合（棒グラフ１００１）には、寄与率が突出して高いものは存在しない。このため、例えば、各主成分の寄与率を入力とし、入力された寄与率の中に突出して高いものが含まれているほど高い値を出力するスコア関数を用いてインサイトスコアを算出することもできる。

【００９８】

なお、インサイトサブジェクト間の非線形な相関を検出したい場合には、評価部３１は、通常の主成分分析のかわりに、任意のカーネルを用いたカーネル主成分分析を実行してもよい。また、レコードのサンプリング粒度の違いなどで相関行列が計算できない場合には、評価部３１は、関数データ解析を用いた関数主成分分析を実行してもよい。

【００９９】

Ｓ３２では、外れ値検出部３２が、グループ化された各インサイトサブジェクトに含まれる外れ値の検出を行う。例えば、Ｓ３１で各インサイトサブジェクトにおける“measure”の項目のデータを用いた評価が行われていた場合、外れ値検出部３２も各インサイトサブジェクトにおける“measure”の項目のデータにおける外れ値を検出する。

【０１００】

外れ値の検出は、Ｓ３１における評価のために行われた主成分分析により求められた主成分を用いて、グループ化された複数のインサイトサブジェクトに含まれるデータを表すことにより行われる。

【０１０１】

図１０の１００３は、サンプルデータを主成分分析して求めた第１主成分ＰＣ１と第２主成分ＰＣ２により当該サンプルデータを表した点を、縦軸をＰＣ２、横軸をＰＣ１とする座標平面上にプロットしたものである。主成分分析後のプロットにおいて、他のデータと離れているデータは、元のサンプルデータにおいても他のデータと離れている。よって

、１００３において「外れ値」とされているプロットのように、他のデータから離れたデータを外れ値として検出すればよい。

【０１０２】

例えば、外れ値検出部３２は、主成分で表されたデータのHotellingのT^2統計量を算出し、算出したT^2統計量が顕著なデータを外れ値として検出してもよい。図１０の１００４は、同図の１００３に示すサンプルデータから算出したT^2統計量を、横軸がサンプル番号、縦軸がT^2統計量の座標平面にプロットしたものである。同図の１００３において「外れ値」とされていたプロットは、T^2統計量が他のプロットと比べて大きい値となっている。よって、外れ値検出部３２は、T^2統計量を用いて外れ値を検出することができる。

【０１０３】

また、T^2統計量はＦ分布やχ^2分布に従うことが知られている。このため、外れ値検出部３２は、統計的検定に基づいて得られたｐ値を用いてスコアを計算してもよい。この場合、外れ値検出部３２は、算出したスコアを用いて外れ値を検出すればよい。

【０１０４】

以上により、図９の処理は終了する。なお、Ｓ３１の評価結果とＳ３２で検出された外れ値は、評価結果データとして記憶しておけばよい。評価結果データは、そのまま出力してもよいし、例示的実施形態２と同様に、評価結果データから出力データを生成し、生成した出力データを出力してもよい。

【０１０５】

〔参考例〕

評価部３１による上述の評価方法は、横断的複合インサイトの検出に好適であると共に、横断的ではない、つまり１つのデータセットにおけるインサイトの検出にも好適である。このため、上述の情報処理装置３は、必ずしも分類部２０４（例示的実施形態２）や、分類部１１（例示的実施形態１）に相当する構成を備えている必要はない。

【０１０６】

本参考例に係る情報処理装置３は、評価対象となる複数のインサイトサブジェクトを取得する取得部と、上述の評価部３１を備えている。前記取得部が取得する複数のインサイトサブジェクトは、少なくとも１つのデータセットから生成されたものであればよい。つまり、複数のデータセットから生成された複数のインサイトサブジェクトを用いることが必須ではない点で、本参考例と上述の各例示的実施形態は相違している。

【０１０７】

本参考例の情報処理装置によれば、評価部３１は、取得部が取得した複数の前記インサイトサブジェクトを主成分分析することにより得られた、各主成分の寄与度の偏りの程度に基づいて、当該インサイトサブジェクトの組み合わせについてのインサイトスコアを算出する。よって、３つ以上のインサイトサブジェクトをまとめて評価することができなかったという従来の課題を解決することができる。

【０１０８】

また、本参考例に係る分析方法は、少なくとも１つのプロセッサが、評価対象となる複数のインサイトサブジェクトを取得すること、および、取得した複数の前記インサイトサブジェクトを主成分分析することにより得られた、各主成分の寄与度の偏りの程度に基づいて、当該インサイトサブジェクトの組み合わせについてのインサイトスコアを算出すること、を含む。そして、本参考例に係る分析プログラムは、コンピュータに、評価対象となる複数のインサイトサブジェクトを取得する処理と、取得した複数の前記インサイトサブジェクトを主成分分析することにより得られた、各主成分の寄与度の偏りの程度に基づいて、当該インサイトサブジェクトの組み合わせについてのインサイトスコアを算出する処理と、を実行させる。これらの分析方法および分析プログラムによっても、３つ以上のインサイトサブジェクトをまとめて評価することができなかったという従来の課題を解決することができる。

【０１０９】

〔変形例〕
上述の例示的実施形態１において、１つの情報処理装置１が行っていた処理は、複数の情報処理装置に分担させてもよい。言い換えれば、情報処理装置１が行う処理の一部を、少なくとも１つの他の情報処理装置に実行させてもよい。さらに言い換えれば、上述の各処理を少なくとも１つのプロセッサに行わせる場合、その少なくとも１つのプロセッサは、１つの情報処理装置１が備えているものであってもよいし、それぞれ異なる情報処理装置が備えているものであってもよい。これは、上述の例示的実施形態２における情報処理装置２、および例示的実施形態３における情報処理装置３についても同様である。
【０１１０】
〔ソフトウェアによる実現例〕
情報処理装置１～３の一部又は全部の機能は、集積回路（ＩＣチップ）等のハードウェアによって実現してもよいし、ソフトウェアによって実現してもよい。
【０１１１】
後者の場合、情報処理装置１～３は、例えば、各機能を実現するソフトウェアであるプログラムの命令を実行するコンピュータによって実現される。このようなコンピュータの一例（以下、コンピュータＣと記載する）を図１１に示す。コンピュータＣは、少なくとも１つのプロセッサＣ１と、少なくとも１つのメモリＣ２と、を備えている。メモリＣ２には、コンピュータＣを情報処理装置１～３として動作させるためのプログラムＰが記録されている。コンピュータＣにおいて、プロセッサＣ１は、プログラムＰをメモリＣ２から読み取って実行することにより、情報処理装置１～３の各機能が実現される。
【０１１２】
プロセッサＣ１としては、例えば、ＣＰＵ（Central Processing Unit）、ＧＰＵ（Graphic Processing Unit）、ＤＳＰ（Digital Signal Processor）、ＭＰＵ（Micro Processing Unit）、ＦＰＵ（Floating point number Processing Unit）、ＰＰＵ（Physics Processing Unit）、マイクロコントローラ、又は、これらの組み合わせなどを用いることができる。メモリＣ２としては、例えば、フラッシュメモリ、ＨＤＤ（Hard Disk Drive）、ＳＳＤ（Solid State Drive）、又は、これらの組み合わせなどを用いることができる。
【０１１３】
なお、コンピュータＣは、プログラムＰを実行時に展開したり、各種データを一時的に記憶したりするためのＲＡＭ（Random Access Memory）を更に備えていてもよい。また、コンピュータＣは、他の装置との間でデータを送受信するための通信インタフェースを更に備えていてもよい。また、コンピュータＣは、キーボードやマウス、ディスプレイやプリンタなどの入出力機器を接続するための入出力インタフェースを更に備えていてもよい。
【０１１４】
また、プログラムＰは、コンピュータＣが読み取り可能な、一時的でない有形の記録媒体Ｍに記録することができる。このような記録媒体Ｍとしては、例えば、テープ、ディスク、カード、半導体メモリ、又はプログラマブルな論理回路などを用いることができる。コンピュータＣは、このような記録媒体Ｍを介してプログラムＰを取得することができる。また、プログラムＰは、伝送媒体を介して伝送することができる。このような伝送媒体としては、例えば、通信ネットワーク、又は放送波などを用いることができる。コンピュータＣは、このような伝送媒体を介してプログラムＰを取得することもできる。
【０１１５】
〔付記事項１〕
本発明は、上述した実施形態に限定されるものでなく、請求項に示した範囲で種々の変更が可能である。例えば、上述した実施形態に開示された技術的手段を適宜組み合わせて得られる実施形態についても、本発明の技術的範囲に含まれる。
【０１１６】
〔付記事項２〕

上述した実施形態の一部又は全部は、以下のようにも記載され得る。ただし、本発明は、以下の記載する態様に限定されるものではない。

【０１１７】

（付記１）

複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化する分類手段と、グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する評価手段と、を備える情報処理装置。この構成によれば、複数のデータセット間におけるインサイトの検出を可能にすることができる。

【０１１８】

（付記２）

複数の前記インサイトサブジェクトにおける表記を統一する表記統一手段をさらに備え、前記分類手段は、表記が統一された前記インサイトサブジェクトをグループ化する、付記１に記載の情報処理装置。この構成によれば、表記が不統一なデータセットについても横断的複合インサイトを検出することが可能になる。

【０１１９】

（付記３）

複数の前記インサイトサブジェクトにおけるデータの粒度を統一する粒度統一手段をさらに備え、前記評価手段は、粒度が統一された複数の前記インサイトサブジェクトについて前記評価値を算出する、付記１または２に記載の情報処理装置。この構成によれば、粒度が不統一なデータを含むデータセットについても横断的複合インサイトを検出することが可能になる。

【０１２０】

（付記４）

前記評価手段は、動的時間伸縮法または関数データ解析により前記評価値を算出する、付記１または２に記載の情報処理装置。この構成によれば、粒度が不統一なデータを含むデータセットについても横断的複合インサイトを検出することが可能になる。

【０１２１】

（付記５）

前記評価手段は、グループ化された複数の前記インサイトサブジェクトを主成分分析することにより求めた、各主成分の寄与度の偏りの程度に基づいて前記評価値を算出する、付記１から４の何れかに記載の情報処理装置。この構成によれば、３つ以上のインサイトサブジェクトをまとめて評価することが可能になる。

【０１２２】

（付記６）

前記主成分分析により求められた主成分を用いて、グループ化された複数の前記インサイトサブジェクトに含まれるデータを表すことにより、当該データに含まれる外れ値を検出する外れ値検出手段をさらに備える、付記５に記載の情報処理装置。この構成によれば、評価のために行った主成分分析の結果を利用した効率のよい外れ値検出ができる。

【０１２３】

（付記７）

少なくとも１つのプロセッサが、複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化すること、およびグループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出すること、を含む分析方法。この構成によれば、複数のデータセット間におけるインサイトの検出を可能にすることができる。

【０１２４】

（付記８）

コンピュータに、複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化する処理と、グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する処理と、を実行させる分析プログラム。この構成によれば、複数のデータセット間におけるインサイトの検出を可能にすることができる。

【０１２５】

（付記９）

少なくとも１つのプロセッサを備え、前記プロセッサは、複数のデータセットのそれぞれから当該データセットに含まれる複数のデータ項目を関連付けることにより生成されたデータであるインサイトサブジェクトを、検出対象のインサイトごとにグループ化する処理と、グループ化された複数の前記インサイトサブジェクトの組み合わせについて、インサイトの有無を判定するための評価値を算出する処理とを実行する情報処理装置。

【０１２６】

なお、この情報処理装置は、更にメモリを備えていてもよく、このメモリには、前記をグループ化する処理と、前記評価する処理とを前記プロセッサに実行させるためのプログラムが記憶されていてもよい。また、このプログラムは、コンピュータ読み取り可能な一時的でない有形の記録媒体に記録されていてもよい。

【符号の説明】

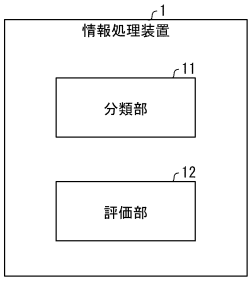
【０１２７】

１　情報処理装置
１１　分類部（分類手段）
１２　評価部（評価手段）
２　情報処理装置
２０３　表記統一部（表記統一手段）
２０４　分類部（分類手段）
２０５　粒度統一部（粒度統一手段）
２０６　評価部（評価手段）
３　情報処理装置
３１　評価部（評価手段）
３２　外れ値検出部（外れ値検出手段）

【図面】

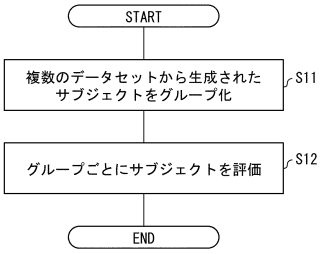
【図１】

図 1



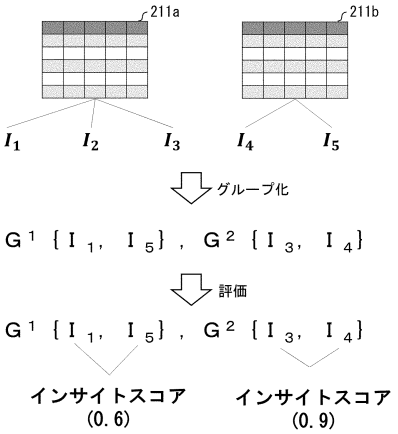
【図２】

図 2



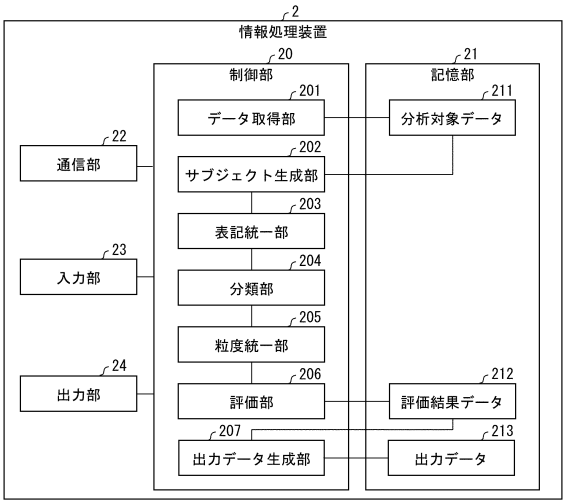
【図３】

図 3



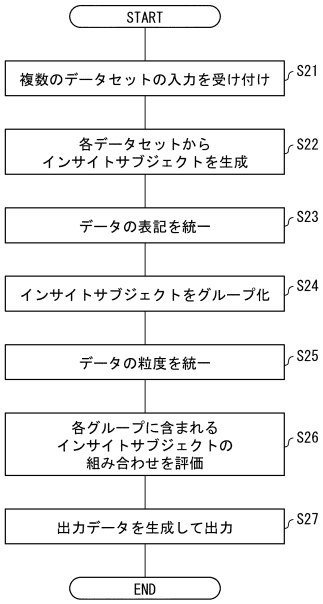
【図４】

図 4



【図５】

図 5



【図６】

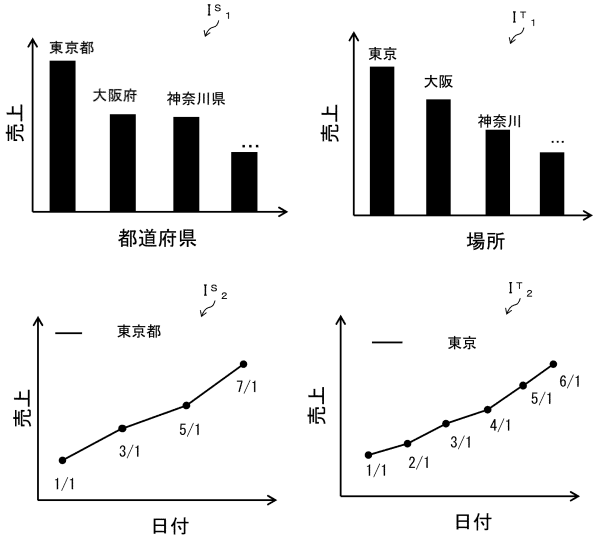
図 6

211（D^S）

日付	都道府県	売上
2020-01-01	東京都	100
2020-03-01	東京都	120
…	…	…
2020-01-01	大阪府	80
…	…	…

211（D^T）

日付	場所	売上
2020-01-01	東京	40
2020-02-01	東京	60
…	…	…
2020-01-01	大阪	30
…	…	…



【図７】

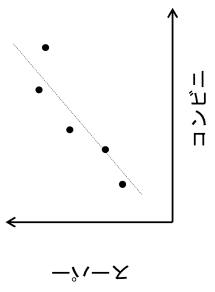
図 7

ランク	インサイトサブジェクト	インサイトスコア	インサイトタイプ
1	$I^S{}_2, I^T{}_2$	0.9	相関
2	$I^S{}_1, I^T{}_1$	0.3	相関
3	…	…	…

212

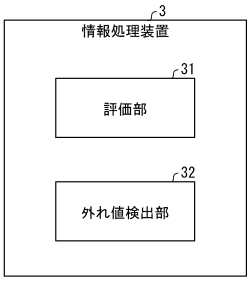
213

インサイト情報
・インサイトタイプ：相関
・インサイトスコア：0.9
データ 1
・データセット D^S,
・インサイトサブジェクト = {{*,*}, 日付, 売上}}
データ 2
・データセット D^T
・インサイトサブジェクト = {{*,*}, 日付, 売上}}



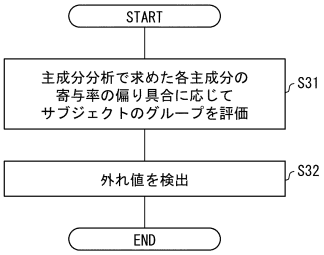
【図８】

図 8



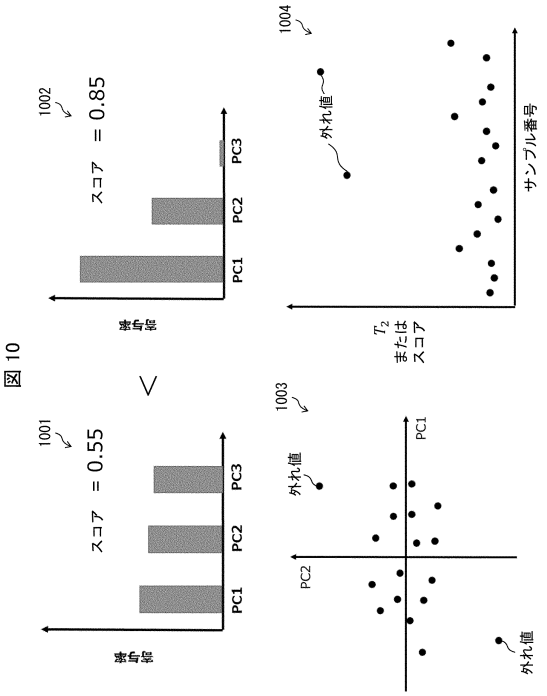
【図９】

図 9



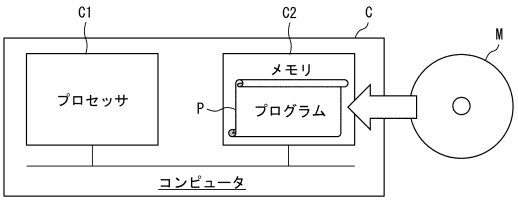
【図１０】

図 10



【図１１】

図 11



フロントページの続き

株式会社内
(72)発明者　草野 元紀
東京都港区芝五丁目７番１号 日本電気株式会社内
審査官　山崎 誠也
(56)参考文献　国際公開第２０２２／０２６３７８（WO，Ａ１）
特開２０２０－１８７５１１（ＪＰ，Ａ）
米国特許出願公開第２０２０／０２５７６８２（ＵＳ，Ａ１）
特開２０１９－１４８８９７（ＪＰ，Ａ）
特開２０２１－０４３８９９（ＪＰ，Ａ）
国際公開第２０１７／１６３２７７（WO，Ａ１）
塚越 雄登，次元間の関係に着目したドメインオントロジーに基づく異種データ間の関連性発見，情報処理学会 研究報告 知能システム（ＩＣＳ），日本，情報処理学会，2020年09月07日，p.1-8，ISSN:2188-885X
NEC、AIで予測分析した結果を可視化し、次の一手を提示するdotDataの新サービスを販売開始，[online]，2020年10月07日，p.1-3，[2022年1月6日検索], インターネット〈URL:https://web.archive.org/web/20201124042138/https://jpn.nec.com/press/202010/20201007_01.html〉
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｑ　１０／００－９９／００