(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6343986号
(P6343986)
(45)発行日　平成30年6月20日(2018.6.20)　(24)登録日　平成30年6月1日(2018.6.1)

(51)Int.Cl.　F I
G06F 21/55 (2013.01)　G06F 21/55 320

請求項の数 5　(全 29 頁)

(21)出願番号　特願2014-53371(P2014-53371)
(22)出願日　平成26年3月17日(2014.3.17)
(65)公開番号　特開2015-176434(P2015-176434A)
(43)公開日　平成27年10月5日(2015.10.5)
審査請求日　平成28年12月6日(2016.12.6)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100094525
弁理士　土井　健二
(74)代理人　100094514
弁理士　林　恒徳
(72)発明者　丸橋　弘治
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　湯上　伸弘
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

最終頁に続く

(54)【発明の名称】情報処理装置，プログラム，情報処理方法

(57)【特許請求の範囲】
【請求項１】
複数の変数それぞれの１つの変数値の組み合わせを含む変数値群をレコードとして複数含む第１のレコード群(T1)と，前記変数値群をレコードとして複数含み前記第１のレコード群と異なる第２のレコード群(T2)とを記憶する記憶部と，
前記第１のレコード群における前記複数の変数それぞれの１つの変数値の組み合わせを含む第１の変数値群が前記第２のレコード群において出現する第１の期待値を，複数の前記第１の変数値群の各変数値の前記第１のレコード群における出現率の各々と前記第２のレコード群に含まれる総レコード数とを乗じて算出し，
前記第１の変数値群から，前記第１の変数値群の前記第１の期待値と前記第２のレコード群における前記第１の変数値群の実際の出現数の差が閾値より大きい前記第１の変数値群を異常変数値群の候補として抽出し，
前記異常変数値群の候補として抽出された前記第１の変数値群と，前記第２のレコード群における前記複数の変数それぞれの１以上の変数値の組み合わせを含む第２の変数値群に含まれる前記複数の変数それぞれの１つの変数値の組み合わせを含む第３の変数値群との，重複する変数値群の数を，前記第３の変数値群の総数で除した値が，基準値より大きい場合，前記第２の変数値群及び前記異常変数値群の候補として抽出した前記第１の変数値群の少なくとも１つを出力する，処理部と，
を有する情報処理装置。
【請求項２】

請求項１において，
前記処理部は，さらに，前記第１の変数値群が前記第１のレコード群(T1)において出現する第２の期待値を，前記第１のレコード群に含まれる複数の変数値群の各変数値の前記第１のレコード群における出現率の各々と前記第１のレコード群に含まれる総レコード数とを乗じて求め，前記第２の期待値と，前記第１のレコード群(T1)における前記第１の変数値群の実際の出現数との差が大きい程，前記第１の変数値群に対応する前記閾値を大きい値に設定する情報処理装置。
【請求項３】
請求項２において，
前記第１のレコード群において，前記第１の変数値群の実際の出現数と前記第２の期待値との差の二乗値を前記第１の変数値群毎に算出し（T61），任意の変数値を含む第１の変数値群毎の二乗値の総和の平方根を前記任意の変数値の不安定度として算出し（T71-T73），前記第１の変数値群における各変数値の不安定度の各々と前記第２のレコード群のレコード総数とを乗算して，前記第１の変数値群に対応する閾値を算出する情報処理装置。
【請求項４】
複数の変数それぞれの１つの変数値の組み合わせを含む変数値群をレコードとして複数含む第１のレコード群(T1)と，前記変数値群をレコードとして複数含み前記第１のレコード群と異なる第２のレコード群(T2)とを記憶し，
前記第１のレコード群における前記複数の変数それぞれの１つの変数値の組み合わせを含む第１の変数値群が前記第２のレコード群において出現する第１の期待値を，複数の前記第１の変数値群の各変数値の前記第１のレコード群における出現率の各々と前記第２のレコード群に含まれる総レコード数とを乗じて算出し，
前記第１の変数値群から，前記第１の変数値群の前記第１の期待値と前記第２のレコード群における前記第１の変数値群の実際の出現数の差が閾値より大きい前記第１の変数値群を異常変数値群の候補として抽出し，
前記異常変数値群の候補として抽出された前記第１の変数値群と，前記第２のレコード群における前記複数の変数それぞれの１以上の変数値の組み合わせを含む第２の変数値群に含まれる前記複数の変数それぞれの１つの変数値の組み合わせを含む第３の変数値群との，重複する変数値群の数を，前記第３の変数値群の総数で除した値が，基準値より大きい場合，前記第２の変数値群及び前記異常変数値群の候補として抽出した前記第１の変数値群の少なくとも１つを出力する，
処理をコンピュータに実行させるプログラム。
【請求項５】
複数の変数それぞれの１つの変数値の組み合わせを含む変数値群をレコードとして複数含む第１のレコード群(T1)と，前記変数値群をレコードとして複数含み前記第１のレコード群と異なる第２のレコード群(T2)とを記憶する情報処理装置で実行される情報処理方法であって，
前記情報処理装置は，
前記第１のレコード群における前記複数の変数それぞれの１つの変数値の組み合わせを含む第１の変数値群が前記第２のレコード群において出現する第１の期待値を，複数の前記第１の変数値群の各変数値の前記第１のレコード群における出現率の各々と前記第２のレコード群に含まれる総レコード数とを乗じて算出し，
前記第１の変数値群から，前記第１の変数値群の前記第１の期待値と前記第２のレコード群における前記第１の変数値群の実際の出現数の差が閾値より大きい前記第１の変数値群を異常変数値群の候補として抽出し，
前記異常変数値群の候補として抽出された前記第１の変数値群と，前記第２のレコード群における前記複数の変数それぞれの１以上の変数値の組み合わせを含む第２の変数値群に含まれる前記複数の変数それぞれの１つの変数値の組み合わせを含む第３の変数値群との，重複する変数値群の数を，前記第３の変数値群の総数で除した値が，基準値より大き

い場合，前記第２の変数値群及び前記異常変数値群の候補として抽出した前記第１の変数値群の少なくとも１つを出力する，情報処理方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は，情報処理装置，プログラム，情報処理方法に関する。
【背景技術】
【０００２】
監視対象から取得した，複数の種類の変数を含むレコード(ログとも呼ばれる)の集合に基づき，所定の事象を検知する手法が提案されている。
【０００３】
監視対象は例えばWebサーバであり，レコードは例えばWebサーバのアクセスログである。アクセスログは，例えば，アクセス時の年月日時，アクセス元のIP(Internet Protocol)アドレス，アクセス先のURL(Uniform Resource Locator)を含む。所定の事象は，不正アクセスなどの異常事象である。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】特開２００６－４８２５３号公報
【特許文献２】特開２００６－１０７１７９号公報
【非特許文献】
【０００５】
【非特許文献１】Weng-Keen Wong, Andrew W. Moore, Gregory F. Cooper, Michael M. Wagner: Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks. AAAI/IAAI 2002: 217-223
【発明の概要】
【発明が解決しようとする課題】
【０００６】
前記した検知の手法としては，例えば，事象検知の対象となる，現在のレコードと過去のレコードとを比較し，比較結果に基づき所定の事象を検知する手法がある。しかし，このレコードには，検知対象となる所定の事象と関係のない様々なレコード(ノイズに起因するレコードと適宜記す)が含まれる。
【０００７】
そのため，前記した検知の手法において，このノイズに起因するレコードを原因とする誤検知が発生することがある。この比較対象のレコードは，現在のレコード，過去のレコードの何れか1つまたは両方である。
【０００８】
本実施の形態の一つの側面は，監視対象から取得したレコードに基づき所定の事象を検知する際の，誤検知を抑制することを目的とする。
【課題を解決するための手段】
【０００９】
本実施の形態の第１の側面は，複数の種類の変数を含む第1のレコード群と，前記複数の種類の変数を含む，前記第1のレコード群と異なる第2のレコード群とを記憶する記憶部と，前記第1のレコード群における前記各種類の1つの変数値の組み合わせを示す第1の変数群が，前記第2のレコード群において出現すると期待される期待値を算出し，前記第1の変数群の前記期待値と，前記第2のレコード群における前記第1の変数群を含むレコードの数とに基づき，前記第1の変数群を抽出し，抽出された前記第1の変数群と，前記第2のレコード群における前記各種類の1以上の変数値の組み合わせを示す第2の変数群とに基づき，前記第2の変数群及び抽出した前記第1の変数群の少なくとも1つを出力する，処理部と，を有することを特徴とする情報処理装置である。

【発明の効果】
【００１０】
第１の側面によれば，監視対象から取得したレコードに基づき所定の事象を検知する際の，誤検知の発生を抑制することができる。
【図面の簡単な説明】
【００１１】
【図１】図1は，異常事象検知の手法を説明する図である。
【図２】図2は，異常事象検知の手法の問題を説明する図である。
【図３】図3は，本実施の形態における情報処理システムSYSを説明する図の一例である。
【図４】図4は，図3の情報処理装置1のハードウェア構成を説明するブロック図の一例である。
【図５】図5は，図4の情報処理装置1のソフトウェア構成を説明するブロック図の一例である。
【図６】図6は，所定の事象(例えば，異常事象)の検出処理の流れを説明するフロー図である。
【図７】図7は，期待値の算出を説明する図である。
【図８】図8は，異常セットの出力を説明する図である。
【図９】図9は，閾値を算出する処理を説明するフロー図である。
【図１０】図10は，第1の変数群に対応する閾値を算出する処理を具体的に説明する図である。
【発明を実施するための形態】
【００１２】
［所定の事象の検知技術］
図1，図2を参照して，本実施の形態に関連する，所定の事象を検知する技術について説明する。ここで，所定の事象とは，例えば異常事象である。
【００１３】
図1は，異常事象検知の手法を説明する図である。過去のレコード群T1は，過去の複数のレコード(以下，過去のレコード群と適宜記す)を記憶するテーブルである。現在のレコード群T2は，現在の複数のレコード(以下，現在のレコード群と適宜記す)を記憶するテーブルである。このレコードは，監視装置が監視対象から取得したレコードであり，例えばWebサーバのアクセスログである。このWebサーバは，例えば電子商取引(Electronic Commerce)用のサーバや，ニュースサイト用のサーバである。なお，以下，"…"は省略を意味する。
【００１４】
以下に説明するレコードを記憶するテーブルは，1行に1つのレコードを記憶する。過去のレコード群を記憶するテーブルを示す過去のレコード群T1，現在のレコード群を記憶するテーブルを示す現在のレコード群T2は，PrimaryKey欄と，Date欄と，IP欄と，URL欄と，ユーザ欄とを有する。
【００１５】
PrimaryKey欄は，1つのレコードを識別する識別子を記憶する。Date欄は，このレコードが生成された年月日時を記憶する。なお，図示の都合上，年を省略し，時間については時だけを示している。Data欄では，MM/DD/HH形式で，月/日/時を示し，MMが月，DDが日，HHが時間を示す。
【００１６】
IP欄は，このレコードに含まれるアクセス元のIPアドレスを記憶する。以下，図において，値が異なるIPアドレスを，IP1～IPm(mは2以上の整数)で示す。
【００１７】
URL欄は，このレコードに含まれるアクセス先のURLを記憶する。以下，図において，値が異なるURLを，U1～Un(nは2以上の整数)で示す。
【００１８】

ユーザ欄は，Webサーバに対するアクセスを行ったユーザを一意に識別する情報を記憶する。以下，図において，異なるユーザをアルファベット(A，B，C，D…)で示す。

【００１９】

(異常事象検知の具体的手法)

ある手法は，以下の第1～第8の工程を行う。第1の工程は，過去のレコード群T1，現在のレコード群T2の総レコードを対象にして，各種類の変数の値(以下，変数の値を変数値と適宜記す)を組み合わせる。

【００２０】

図1の例では，変数値として，例えば，IPアドレス"IP2"，URL"U1"，ユーザ"C"を組み合わせる。この組み合わせ例を，図1の組み合わせテーブルT3，組み合わせテーブルT4における変数値の組み合わせ欄では"IP2＆U1＆C"と記す。

【００２１】

第2の工程は，図1で説明した過去のレコード群T1において，この組み合わせ"IP2＆U1＆C"を満たすレコードの数をカウントし，図1の組み合わせテーブルT3における出現数欄に記憶する。このカウントした数を例えば"45"とする。

【００２２】

第3の工程は，過去のレコード群T1における総レコード数(例えば，265)から，前記した組み合わせ"IP2＆U1＆C"を含まない，各種類の変数値の組み合わせの総数を減算し，図1の組み合わせテーブルT3における含まない数欄に記憶する。この記憶された総数を例えば"220"とする。

【００２３】

第4の工程は，この記憶された総数(例えば，220)に対する，この組み合わせ"IP2＆U1＆C"を満たす(に該当する)レコードの数の比率(以下，第1の比率と記す)を算出する。第1の比率は，"45/220"である。

【００２４】

この手法は，次に，図1で説明した現在のレコード群T2において，前記した第2～第4の工程を実行する。具体的には，第5～第7の工程を実行する。

【００２５】

第5の工程は，図1で説明した現在のレコード群T2において，第2の工程の説明において前記した組み合わせ"IP2＆U1＆C"を満たすレコードの数をカウントし，図1の組み合わせテーブルT4における出現数欄に記憶する。このカウントした数を例えば"48"とする。

【００２６】

第6の工程は，現在のレコード群T2における総レコード数(例えば，134)から，前記した組み合わせ"IP2＆U1＆C"を含まない，各種類の変数値の組み合わせの総数を減算し，図1の組み合わせテーブルT4における含まない数欄に記憶する。この記憶された総数を例えば"86"とする。

【００２７】

第7の工程は，この記憶された総数(例えば，86)に対する，この組み合わせ"IP2＆U1＆C"を満たす(該当する)レコードの数の比率(以下，第2の比率と記す)を算出する。第2の比率は，"48/86"である。

【００２８】

第8の工程は，第1の比率と，第2の比率とが有意に異なるか否かを判定(検定とも呼ぶ)し，第1の比率と，第2の比率とが有意に異なっていれば，前記した組み合わせを異常セットとして出力し，管理者に対応を促す。なお，第1の比率と，第2の比率とが有意に異なるか否かを判定するためには，例えばフィッシャーの正確確率検定を用いる。

【００２９】

この手法は，第1の工程で説明した，異なる種類の変数値の組み合わせを網羅的に行い，すなわち総組み合わせを作成し，各組み合わせに対して，第2～第8の工程を実行する。

【００３０】

この手法によれば，異常事象検知対象になる多数のレコードの中にノイズに起因するレ

コードが含まれる場合，ノイズに起因するレコードによる誤検知が頻発することがある。ノイズに起因するレコードは，異常事象と関係のない様々なレコードである。

【００３１】

ここでノイズに起因するレコードについて具体的に説明する。例えば，アクセス先のURL"U1"～URL"U5"(図示しない)が，電子商取引に関するURLであるとする。そして，電子商取引の主催者は，ユーザがこのURLにアクセスした数に応じて，商品の値引きポイント(クーポンとも呼ぶ)をユーザに付与するとする。ただし，商品の購入や，商品の閲覧(実際の店舗におけるウィンドショッピングに相当)を主目的とせずに，単に，値引きポイントを得る目的で，ユーザがこのURLにアクセスする場合，このユーザを悪意のユーザと見なし，ポイントを付与しない。電子商取引の主催者は，悪意のユーザによるアクセスを検知し，かかるユーザに対するポイントの付与を防止する必要がある。

【００３２】

ここで，過去のレコード群(例えば，図1のT1参照)，現在のレコード群(例えば，図2のT2参照)における，送信元IPアドレス"IP2"，アクセス先のURL"U1"，ユーザ"C"(組み合わせ"IP2＆U1＆C")のレコードを想定する。以下，このレコードを，想定レコードと適宜記す。

【００３３】

想定レコードの第1の比率は，図1の例では"45/220"(0.20)であり，想定レコードの第2の比率は"48/86"(0.56)である。第2の比率は，第1の比率に比べて，2倍以上であるので，この手法は，第1の比率と，第2の比率とが有意に異なると判定するとする。

【００３４】

そして，この手法は，この組み合わせを異常セットとして出力する。具体的には，この手法は，ユーザ"C"(送信元IPアドレス"IP2")によるURL"U1"に対するアクセスが，異常である旨を管理者に報知し対応を促す。この異常とは，ユーザ"C"が悪意のユーザであることを意味している。

【００３５】

ここで，ユーザ"C"が，現在のレコード群の月日時において，URL"U1"にアクセスする際，何らかの理由で，URL"U1"に適切にアクセスできない場合を想定する。例えば，ユーザ"C"がURL"U1"にアクセスしたが，何らかの理由で，URL"U1"のHTMLデータを完全に自装置にダウンロードできず，画像が完全に表示されない場合である。

【００３６】

この理由は，例えば，ユーザ"C"の自装置が実行しているブラウザの不具合や，自装置と，URL"U1"のHTMLデータを記憶するWebサーバとのネットワーク遅延や，Webサーバが過負荷になっていることである。

【００３７】

このように，ユーザ"C"が，URL"U1"に適切にアクセスできない場合，適切にアクセスができるまで，ユーザ"C"は，URL"U1"に対するアクセスを繰り返す。その結果，現在のレコード群において，この想定レコードの数が増える。このレコードは，悪意のあるユーザにより生じたレコードではなく，すなわち異常事象と関係のないレコード(換言すれば，ノイズに起因するレコード)である。

【００３８】

このように，検知対象となるレコード(すなわち，現在のレコード)にノイズに起因するレコードがあると，ノイズに起因するレコードによる誤検知が頻発する問題が生じることがある。以下，図2を参照して，この問題について説明する。

【００３９】

図2は，異常事象検知の手法の問題を説明する図である。図2の組み合わせテーブルT5，組み合わせテーブルT6は，図1の組み合わせテーブルT3，組み合わせテーブルT4と同様の構成を有する。図2の説明において，過去のレコード群における総レコード数は20000で，現在のレコード群における総レコード数は10000である。

【００４０】

この手法は，第1の工程を実行して，各種類の変数値を組み合わせる。図2の例では，変数値として，例えば，IPアドレス"IP1"，URL"U2"，ユーザ"A"を組み合わせる。この組み合わせ例を，図2の組み合わせテーブルT5，組み合わせテーブルT6における変数値の組み合わせ欄では"IP1＆U2＆A"と記す。

【００４１】

この手法は，過去のレコード群に対して，第2，第3の工程を実行する。そして，この手法は，第4の工程を実行し，ある組み合わせ(例えば"IP1＆U2＆A")に対応する第1の比率を算出する。第1の比率は，(例えば，3/19997)である。この工程の実行結果の一部を，図2の組み合わせテーブルT5に示す。以下，ある組み合わせを組み合わせXと記す。

【００４２】

さらに，この手法は，現在のレコード群に対して，第5，第6の工程を実行する。そして，この手法は，第7の工程を実行し，第4の工程で説明した前記組み合わせX(例えば"IP1＆U2＆A")に対応する第2の比率を算出する。第2の比率は，(例えば，7/9993)である。この工程の実行結果の一部を，図2の組み合わせテーブルT6に示す。

【００４３】

図2において，過去のレコード群における組み合わせと，現在のレコード群における，この組み合わせと同じ組み合わせとを矢印で示す。

【００４４】

ここで，図2において，各種類の変数値が多くなると，組み合わせXを満たすレコードの数が極端に少なくなる。例えば，組み合わせテーブルT5では，組み合わせX(例えば"IP1＆U2＆A")の場合，組み合わせXを満たすレコードの数は3である。組み合わせテーブルT6では，組み合わせXを満たすレコードの数は，7である。

【００４５】

組み合わせXを満たすレコードの数が極端に少ない場合において，この手法は，組み合わせXについて，第4の工程で説明した第1の比率と第7の工程で説明した第2の比率とが有意に異なるか否かを判定する。

【００４６】

過去のレコード群，現在のレコード群において，組み合わせXを満たす複数のレコードの何れかのレコードが，ノイズに起因するレコードなのか，また，ノイズに起因しないレコードなのかを事前に判別することは困難である。そのため，ノイズに起因するレコードも含めて第1の比率と，第2の比率とが有意に異なるか否かを判定することになる。

【００４７】

組み合わせXを満たすレコードの数が多い場合(例えば，過去のレコード群では"100"レコード，現在のレコード群では"200"レコード)を想定する。この多い場合，組み合わせXを満たす複数のレコードの中にノイズに起因する少数のレコード(例えば，"2"レコード)が含まれていても，この少数のレコードが，第1の比率，または，第2の比率に及ぼす影響は無視できるほど小さい。

【００４８】

ここで，組み合わせXを満たすレコードの数が少ない場合(例えば，過去のレコード群では"3"レコード，現在のレコード群では"7"レコード)を想定する。この少ない場合，組み合わせXを満たす複数のレコードの中にノイズに起因するレコードが含まれていると，このレコードが，第1の比率，または，第2の比率に及ぼす影響が大きくなる。その結果，ノイズに起因するレコードを原因として，組み合わせXにおいて，第1の比率と，第2の比率とが有意に異なると判定され，異常事象が発生したと検知されることがある。この検知は，実際にはノイズに起因するレコードに基づき検知されたもので，誤検知である。

【００４９】

以上で説明した所定の事象の検知技術に基づき，異常事象の検知を行う際の誤検知を抑制する処理を行う情報処理装置を下記の実施の形態で説明する。

【００５０】

[実施の形態]

(情報処理システム)
図3は，本実施の形態における情報処理システムSYSを説明する図の一例である。以下，同様の構成については，同じ符号を付し，適宜その説明を省略する。
【００５１】
情報処理システムSYSは，LAN(Local Area Network)，WAN(Wide Area Network)などのネットワークNを介して接続された，情報処理装置1と，クライアント端末2と，Webサーバ31…Webサーバ3p(小文字pは，2以上の整数)とを有する。
【００５２】
情報処理装置1は，監視対象(例えば，Webサーバ31…Webサーバ3p)から取得した，複数の種類の変数を含むレコードの集合に基づき，所定の事象を検知する装置である。情報処理装置1は，異常検知装置とも呼ばれる。クライアント端末2は，情報処理装置1を操作する端末である。
【００５３】
Webサーバ31…Webサーバ3p(pは2以上の整数)は，インターネットINを介して接続している外部端末(図示しない)にインストールされたブラウザとHTTP(Hypertext Transfer Protocol)による通信を行い，所定の情報処理を行うサーバである。Webサーバ31…Webサーバ3pの何れかのWebサーバ(例えば，Webサーバ31，32など)は，例えば，図1，図2で説明した電子商取引を実行するサーバである。Webサーバ31…Webサーバ3pの何れかのWebサーバは，電子商取引のユーザが使用する外部端末からアクセスリクエストを受信すると，ユーザの識別子(ユーザID)と，ユーザのパスワードを入力するようにユーザに促す。
【００５４】
Webサーバ31…Webサーバ3pの何れかのWebサーバは，入力されたユーザIDとユーザのパスワードとに基づき，このユーザを認証し，認証が成功した場合，このユーザによるサーバへのアクセスを許可する。
【００５５】
Webサーバ31…Webサーバ3pの何れかのWebサーバは，このアクセスを許可後，図1で説明したアクセスログ(レコード)を生成し，情報処理装置1に出力する。このレコードは，レコードの識別子，レコードが生成された年月日時，アクセス元(外部端末)のIPアドレス，アクセス先のURL，ユーザIDを含む。
【００５６】
(情報処理装置のハードウェアブロック図)
図4は，図3の情報処理装置1のハードウェア構成を説明するブロック図の一例である。
【００５７】
情報処理装置1は，バスBに接続された，CPU(処理部)11と，ストレージ(記憶部)12と，RAM13と，外部記憶媒体読み取り装置14と，ネットワークインターフェイス15とを有する。なお，CPUは"Central Processing Unit"の略語，RAMは"Random Access Memory"の略語である。以下，CPU(処理部)11をCPU11，ストレージ(記憶部)12をストレージ12と適宜記す。
【００５８】
CPU11は，情報処理装置1の全体を制御する中央演算処理装置(コンピュータ)である。ストレージ12は，例えばハードディスクドライブ(HDD：Hard Disk Drive)や，ソリッドステートドライブ(SSD:Solid State Drive)などの大容量記憶装置である。ストレージ12は，過去のレコード群T1と，現在のレコード群T2と，所定の事象(例えば，異常事象)を検出する検知用プログラムPGの実行ファイルを記憶する。
【００５９】
過去のレコード群T1は，複数の種類の変数を含む第1のレコード群であり，例えば図1の過去のレコード群T1である。現在のレコード群T2は，複数の種類の変数を含む，過去のレコード群T1と異なる第2のレコード群であり，例えば図1の現在のレコード群T2である。
【００６０】
RAM13は，CPU11が実行する処理や，検知用プログラムPGが実行する各ステップにおいて生成(算出)されたデータなどを一時的に記憶する。RAM13は，例えばDRAM(Dynamic Random

Access Memory)などの半導体メモリである。なお，RAM13が，過去のレコード群T1と，現在のレコード群T2とを記憶する記憶部でもよい。

【００６１】

CPU11は，情報処理装置1の起動時に，ストレージ12から検知用プログラムPGの実行ファイルを読み出し，RAM13に展開し，異常事象を検知する処理を実行する。なお，この実行ファイルを外部記憶媒体MDに記憶してもよい。

【００６２】

外部記憶媒体読み取り装置14は，外部記憶媒体MDに記憶されたデータを読み取る装置である。外部記憶媒体MDは，例えば，CD-ROM(Compact Disc Read Only Memory)，DVD(Digital Versatile Disc)などの可搬型記憶媒体や，USBメモリなどの可搬型の不揮発性メモリである。

【００６３】

ネットワークインターフェイス15は，例えばNIC(Network Interface Card)を有し，ネットワークNに対するインターフェイス機能を提供する。

【００６４】

(情報処理装置のソフトウェアブロック図)

図5は，図4の情報処理装置1のソフトウェア構成を説明するブロック図の一例である。検知用プログラムPGは，入力部101と，期待値算出部102と，閾値算出部103と，異常セット抽出部104と，出力部105とを有する。

【００６５】

入力部101は，Webサーバ31…Webサーバ3pが出力するレコード(アクセスログ)を取得し，取得したレコードに含まれる年月日時を参照し，取得したレコードを過去のレコード群と，現在のレコード群に分類する。例えば，入力部101は，異常事象検知処理を実行する日時(例えば，1月22日の10時)を基準にして所定の時間帯(7時から13時)に生成された複数のレコードを現在のレコード群に分類し，ストレージ12に記憶する(現在のレコード群T2参照)。

【００６６】

そして，入力部101は，異常事象検知処理を実行する日時(例えば，1月22日の10時)から所定の日(例えば，1日)前における前記した所定の時間帯(7時から13時)に生成された複数のレコードを過去のレコード群に分類し，ストレージ12に記憶する(過去のレコード群T1参照)。

【００６７】

入力部101は，過去のレコード群T1，現在のレコード群T2をストレージ12から読み出し，期待値算出部102と，閾値算出部103とに入力する。

【００６８】

期待値算出部102は，過去のレコード群T1における各種類の1つの変数値の組み合わせを示す第1の変数群が，現在のレコード群T2において出現すると期待される期待値を算出する。

【００６９】

閾値算出部103は，第1の変数群が，過去のレコード群T1において出現すると期待される期待値と，現在のレコード群T2の中で第1の変数群を含むレコードの数とに基づき，第1の変数群に対応する閾値を算出する。

【００７０】

異常セット抽出部104は，第1の変数群の期待値と，現在のレコード群T2における第1の変数群を含むレコードの数とに基づき，第1の変数群を抽出するか否かを判定し判定結果に基づき第1の変数群を抽出する。

【００７１】

出力部105は，抽出された第1の変数群と，現在のレコード群T2における各種類の1以上の変数値の組み合わせを示す第2の変数群とに基づき，第2の変数群及び抽出した第1の変数群の少なくとも1つを出力する。具体的には，出力部105は，抽出された第1の変数群と

，第2の変数群における各種類の1つの変数値の組み合わせを示す第3の変数群とに基づき，第2の変数群及び抽出した第1の変数群の少なくとも1つを出力する。出力部105は，前記した出力により，検知対象の事象が発生したことを管理者に報知する。

【００７２】

(異常事象の検出処理の流れ)

図6は，所定の事象(例えば，異常事象)の検出処理の流れを説明するフロー図である。以下のステップS1が実行される前に，情報処理装置1の入力部101は，過去のレコード群T1，現在のレコード群T2をストレージ12から読み出し，期待値算出部102，閾値算出部103に入力している。

【００７３】

ステップS1:閾値算出部103は，過去のレコード群T1において，各種類の1つの変数値の組み合わせを示す第1の変数群に対応する閾値を算出する。ステップS1の処理については，図9で詳細に説明する。

【００７４】

ステップS2:期待値算出部102は，過去のレコード群T1において各種類の各変数値が出現する割合を示す出現率(発生確率，出現確率とも呼ぶ)を算出する。なお，ステップS2～ステップS4の処理については，図7で具体的に説明する。

【００７５】

ステップS3:期待値算出部102は，現在のレコード群T2において，各種類の1つの変数値の組み合わせを示す，未選択の第1の変数群を選択する。

【００７６】

ステップS4:期待値算出部102は，選択した変数群が，現在のレコード群T2において出現すると期待される期待値を算出する。

【００７７】

ステップS5:異常セット抽出部104は，選択された変数群の期待値と，現在のレコード群T2における選択された変数群を含むレコードの数とに基づき，選択された変数群を抽出するか否かを判定し判定結果に基づき選択された変数群を異常セットとして抽出する。なお，ステップS5～ステップS8の処理については，図8で具体的に説明する。

【００７８】

ステップS6:異常セット抽出部104は，ステップS3において全ての変数群が選択されたか判定する。全ての変数群が選択されていない場合(ステップS6/NO)，ステップS3に戻る。全ての変数群が選択された場合(ステップS6/YES)，ステップS7に移る。

【００７９】

ステップS7:出力部105は，ステップS5において抽出された異常セット(変数群)の中から，出力対象となる異常セットを抽出する。

【００８０】

ステップS8:出力部105は，ステップS7において抽出した異常セットを出力する。出力方法としては，例えば，表示出力がある。表示出力の場合，出力部105は，図3のクライアント端末2の表示装置に，異常セットを出力する。

【００８１】

(異常事象の検出処理の具体例)

図1，図6～図8を参照して，異常事象の検出処理を具体的に説明する。図7は，期待値の算出を説明する図である。図8は，異常セットの出力を説明する図である。ここでは，既に，図6のステップS1で説明した閾値が算出されている。

【００８２】

(出現率の算出)

出現率の算出(図6のステップS2)について説明する。まず，期待値算出部102は，過去のレコード群T1の総レコード数を算出(例えば，20000)する。期待値算出部102は，この総レコードにおける各種類の各変数値の出現数(発生数とも呼ぶ)を算出する(ステップS2)。換言すれば，期待値算出部102は，この総レコードの中で，ある変数値を含むレコードの数

を算出する。期待値算出部102は，各変数値と，この各変数値の出現数とを例えばテーブル形式でRAM13に記憶する。

【００８３】

IP出現数テーブルT11は，変数の種類がIPアドレスの場合における，各変数値(例えば，IP1，IP2…)と，この各変数値の出現数とを記憶するテーブルである。IP出現数テーブルT11の例では，過去のレコード群T1においてIPアドレス"IP1"を含むレコードの数は1000である。

【００８４】

URL出現数テーブルT12は，変数の種類がURLの場合における，各変数値(例えば，U1，U2…)と，この各変数値の出現数とを記憶するテーブルである。URL出現数テーブルT12の例では，過去のレコード群T1においてURL"U1"を含むレコードの数は6000である。

【００８５】

ユーザ出現数テーブルT13は，変数の種類がユーザの場合における，各変数値(例えば，A，B…)と，この各変数値の出現数とを記憶するテーブルである。ユーザ出現数テーブルT13の例では，過去のレコード群T1においてユーザ"A"を含むレコードの数は1600である。

【００８６】

次いで，期待値算出部102は，各変数値の出現数を，過去のレコード群T1の総レコード数で除算した除算値を，各変数値の出現率として算出する(ステップS2)。期待値算出部102は，各変数値と，この各変数値の出現率とを例えばテーブル形式でRAM13に記憶する。

【００８７】

IP出現率テーブルT21は，変数の種類がIPアドレスの場合における，各変数値(例えば，IP1，IP2…)と，この各変数値の出現率とを記憶するテーブルである。期待値算出部102は，IP出現率テーブルT21の例では，過去のレコード群T1におけるIPアドレス"IP1"の出現数"1000"を，過去のレコード群T1の総レコード数(20000)で除算した除算値(0.05)を，IPアドレス"IP1"の出現率として算出する。

【００８８】

URL出現率テーブルT22は，変数の種類がURLの場合における，各変数値(例えば，U1，U2…)と，この各変数値の出現率とを記憶するテーブルである。期待値算出部102は，URL出現率テーブルT22の例では，過去のレコード群T1におけるURL"U1"の出現数"6000"を，過去のレコード群T1の総レコード数(20000)で除算した除算値(0.30)を，URL"U1"の出現率として算出する。

【００８９】

ユーザ出現率テーブルT23は，変数の種類がユーザの場合における，各変数値(例えば，A，B…)と，この各変数値の出現率とを記憶するテーブルである。期待値算出部102は，ユーザ出現率テーブルT23の例では，過去のレコード群T1におけるユーザ"A"の出現数"1600"を，過去のレコード群T1の総レコード数(20000)で除算した除算値(0.08)を，ユーザ"A"の出現率として算出する。

【００９０】

期待値算出部102は，現在のレコード群T2において，変数の各種類の1つの変数値の組み合わせを示す第1の変数群を選択し，RAM13に記憶する(ステップS3)。

【００９１】

例えば，変数の各種類は，IPアドレス，URL，ユーザである，期待値算出部102は，現在のレコード群T2において，変数の種類"IPアドレス"の変数値(例えば，IP1，IP2…)の何れか1つを選択する。期待値算出部102は，現在のレコード群T2において，変数の種類"URL"の変数値(例えば，U1，U2…)の何れか1つを選択する。期待値算出部102は，現在のレコード群T2において，変数の種類"ユーザ"の変数値(例えば，A，B…)の何れか1つを選択する。

【００９２】

そして，期待値算出部102は，選択した各種類の1つの変数値を組み合わせて，各種類の1つの変数値の組み合わせを示す第1の変数群とする。例えば，第1の変数群として，"IP1

, U2, A", "IP2, U2, B", " IP1, U1, B"などがある。以下，各変数値が組み合わされている状態を"&"(論理積)で示す。

【００９３】

期待値算出部102は，図7の現在の期待値テーブルT31に示すように，第1の変数群をRAM13に記憶する。

【００９４】

期待値算出部102は，現在のレコード群T2において，各種類の相違する変数値について前記した組み合わせを実行し，全ての組み合わせを示す全ての変数群を得る。例えば，現在のレコード群T2において，変数の種類"IPアドレス"の変数値がIP1，IP2の2つ，変数の種類"URL"の変数値がU1，U2，U3の3つ，変数の種類"ユーザ"の変数値がA，Bの2つの場合を想定する。この場合，期待値算出部102は，全ての組み合わせを示す全ての変数群は，12(2×3×2)個の変数群である。期待値算出部102は，全ての変数群の中から未選択の1つの変数群を選択する(ステップS3)。

【００９５】

(現在のレコード群T2における期待値の算出)

次に，現在のレコード群T2における期待値の算出(図6のステップS4)について説明する。

【００９６】

期待値算出部102は，現在のレコード群T2の総レコード数と，ステップS3で選択した変数群(第1の変数群)に含まれる各種類の1つの変数値の出現率に基づき，ステップS3で選択した変数群の期待値を算出する(ステップS4)。

【００９７】

具体的には，期待値算出部102は，ステップS3で選択した変数群(第1の変数群)に含まれる各種類の1つの変数値の出現率の各々を乗算し，この乗算値と，現在のレコード群T2の総レコード数とを乗算し，RAM13に記憶する(ステップS4)。期待値算出部102は，RAM13に記憶した乗算値を，現在のレコード群T2において，選択した変数群が出現すると期待される期待値，すなわち第1の変数群の期待値とする(ステップS4)。

【００９８】

前記の例では，現在のレコード群T2の総レコード数は"10000"である。選択された変数群は，例えば"IP1&U2&A"である。変数群"IP1&U2&A"における種類"IPアドレス"の1つの変数値"IP1"の出現率は"0.05"(図7のIP出現率テーブルT21参照)である。変数群"IP1&U2&A"における種類"URL"の1つの変数値"U2"の出現率は"0.10"(図7のURL出現率テーブルT22参照)である。変数群"IP1&U2&A"における種類"ユーザ"の1つの変数値"A"の出現率は"0.08"(図7のユーザ出現率テーブルT23参照)である。

【００９９】

期待値算出部102は，変数群"IP1&U2&A"における種類"IPアドレス", "URL", "ユーザ"の出現率"0.05", "0.10", "0.08"の各々を乗算する。期待値算出部102は，この乗算値(0.0004)と，現在のレコード群T2の総レコード数"10000"とを乗算し，乗算値"4.0"(10000×0.0004(0.05×0.10×0.08))を算出し，RAM13に記憶する。

【０１００】

期待値算出部102は，この乗算値"4.0"を，現在のレコード群T2において，変数群"IP1&U2&A"が出現すると期待される期待値"4.0"とする。

【０１０１】

期待値算出部102は，図7の現在の期待値テーブルT31に示すように，変数群と，この変数群の期待値とを対応付けてRAM13に記憶する。

【０１０２】

(異常セットの抽出)

次に，異常セットの抽出(図6のステップS5)について説明する。異常セット抽出部104は，現在のレコード群T2の中で，ステップS3において選択された変数群(第1の変数群)を含むレコードの総数と，この選択された変数群の期待値との差の絶対値を算出しRAM13に記

憶する。異常セット抽出部104は，この差の絶対値が，選択された変数群に対応して予め算出された閾値以上の場合，選択された変数群を現在のレコード群T2から抽出し，異常セットとしてRAM13に記憶する(ステップS5)。この閾値は，ステップS1において，予め算出された閾値である。

【０１０３】

選択された変数群が"IP1&U2&A"の場合，この変数群の期待値は，図8の現在の期待値テーブルT31によれば"4.0"である。また，現在のレコード群T2の中で，選択された変数群"IP1&U2&A"(IPアドレス"IP1"，URL"U2"，ユーザ"A")を含むレコードの数は"7"であるとする。この場合，異常セット抽出部104は，選択された変数群"IP1&U2&A"について，前記した差"+3.0"(7-4.0)を算出する。

【０１０４】

異常セット抽出部104は，図8の差算出テーブルT41に示すように，変数群と，現在のレコード群T2の中でこの変数群を含むレコードの数と，このレコードの数とこの変数群の期待値との差とを対応付けてRAM13に記憶する。差算出テーブルT41において，現在のレコード群T2の中で前記した変数群を含むレコードの数を出現数としている。

【０１０５】

異常セット抽出部104は，現在のレコード群T2の中で，選択された変数群を含むレコードの数と，このレコードの数とこの変数群の期待値との差の絶対値が，この変数群に対応する閾値以上の場合，この変数群を異常セットとして抽出し，RAM13に記憶する。

【０１０６】

例えば，変数群"IP1&U2&A"に対応する閾値が"3"，変数群"IP2&U2&B"に対応する閾値が"2"，変数群"IP1&U1&B"に対応する閾値が"4"とする。この場合，異常セット抽出部104は，差算出テーブルT41の例では，変数群"IP1&U2&A"と，変数群"IP1&U1&B"とを異常セットとして抽出する。

【０１０７】

異常セット抽出部104は，異常セットテーブルT42に示すように，異常セットとして抽出した変数群と，現在のレコード群T2の中でこの変数群を含むレコードの数(出現数)と，このレコードの数とこの変数群の期待値との差とを対応付けてRAM13に記憶する(ステップS5)。

【０１０８】

期待値算出部102，異常セット抽出部104は，現在のレコード群T2において，各種類の1つの変数値の全ての組み合わせを示す変数群についてステップS3～ステップS5の処理が終了するまで(ステップS6/NO)，ステップS3～ステップS5の処理を繰り返し，この処理が終了すると(ステップS6/YES)，ステップS7に移る。

【０１０９】

(抽出した異常セットの出力)

出力対象となる異常セットの抽出(図6のステップS7)，および出力(図6のステップS8)について説明する。

【０１１０】

出力部105は，各種類の1以上の変数値の組み合わせを示す第2の変数群を決定し，RAM13に記憶する(ステップS7)。各種類の1以上の変数値の組み合わせについて，例えば，各種類が，IPアドレス，URL，ユーザであり，各種類の変数値が，IPアドレスについて"IP1"～"IP3"，URLについて"U1"，"U2"，ユーザについて"A"，"B"であるとする。この例示を第1の場合と記す。

【０１１１】

第1の場合，各種類の1以上の変数値とは，IPアドレスについては，{IP1}，{IP2}，{IP3},{IP1，IP2},{IP1，IP3},{IP2，IP3},{IP1，IP2，IP3}である(7通り)。また，URLについては，{U1},{U2},{U1，U2}である(3通り)。ユーザについては，{A},{B},{A，B}である(3通り)。

【０１１２】

第1の場合，前記した各種類の1以上の変数値の組み合わせは，例えば，"{IP1}&{U1}&{A}"，"{IP1}&{U2}&{A}"，"{IP1}&{U1}&{B}"，"{IP1}&{U2}&{B}"，"{IP2}&{U1}&{A}"，"{IP2}&{U2}&{A}"，"{IP1}&{U1，U2}&{A}"，"{IP1,IP2}&{U1,U2}&{A,B}"，"{IP1，IP2}&{U1，U2}&{A}"，"{IP1,IP2,IP3}&{U1,U2}&{A,B}"などである。なお，この組み合わせの総数は63(7×3×3)である。

【０１１３】

出力部105は，決定した第2の変数群における，各種類の1つの変数値の組み合わせを示す第3の変数群の総数を算出する(ステップS7)。

【０１１４】

各種類の1つの変数値の組み合わせは，図8で説明した第1の変数群に対応し，例えば"IP1&U2&A"，"IP2&U2&B"，"IP1&U1&B"(図7の差算出テーブルT41参照)である。

【０１１５】

第1の場合において，例えば決定した第2の変数群"{IP1}&{U1}&{A}"における，各種類の1つの変数値の組み合わせを示す第3の変数群は"{IP1}&{U1}&{A}"であり，その総数は1である。

【０１１６】

第1の場合において，例えば決定した第2の変数群"{IP1，IP2}&{U1，U2}&{A，B}"における第3の変数群は，"{IP1}&{U1}&{A}"，"{IP1}&{U2}&{A}"，"{IP2}&{U1}&{A}"，"{IP2}&{U2}&{A}"，"{IP1}&{U1}&{B}"，"{IP1}&{U2}&{B}"，"{IP2}&{U1}&{B}"，"{IP2}&{U2}&{B}"である。この変数群の総数は，8(2×2×2)である。

【０１１７】

ここで，ステップS5において異常セットして抽出された，各種類の1つの変数値の組み合わせを示す変数群を，"{IP1}&{U1}&{A}"，"{IP1}&{U2}&{A}"，"{IP2}&{U1}&{A}"，"{IP2}&{U2}&{A}"，"{IP1}&{U1}&{B}"，"{IP1}&{U2}&{B}"，"{IP2}&{U1}&{B}"とする。この変数群の数は7である。

【０１１８】

出力部105は，第2の変数群に含まれる第3の変数群の総数と，第3の変数群の中で，抽出した第1の変数群に一致する変数群の数とに基づき，第2の変数群，抽出された異常セット(抽出された第1の変数群)の少なくとも1つを出力するか判定する(ステップS8)。そして，出力部105は，判定結果に基づき，第2の変数群，抽出された異常セットの少なくとも1つを出力する(ステップS8)。

【０１１９】

ここで，出力部105は，一致する変数群の数を第3の変数群の総数で除算した値が，予め定められた所定の値をよりも大きい場合，第2の変数群，抽出された異常セットの少なくとも1つを出力する(ステップS8)。なお，出力部105は，第2の変数群のみ，または，抽出された異常セットのみを出力してもよいし，第2の変数群および抽出された異常セットを出力してもよい。

【０１２０】

具体的には，出力部105は，第2の変数群に含まれる第3の変数群の中で，異常セットとして抽出された変数群に一致する変数群の一致数を算出し，RAM13に記憶する(ステップS8)。

【０１２１】

出力部105は，図8の出力対象異常セットテーブルT43に示すように，各種類の1以上の変数値の組み合わせ示す第2の変数群と，第2の変数群に含まれる第3の変数群の中で，異常セットとして抽出された変数群に一致する変数群の一致数とを対応付けてRAM13に記憶する。なお，各種類の1以上の変数値の組み合わせ示す第2の変数群は，図8の出力対象異常セットテーブルT43における"変数値の範囲"であり，一致数は，出力対象異常セットテーブルT43における"範囲に含まれる異常セット数"である。

【０１２２】

前記した第1の場合において，決定した第2の変数群"{IP1，IP2}&{U1，U2}&{A，B}"を例

示すると，異常セットとして抽出された変数群に一致する変数群は，前記した7つの変数群であり，一致数は7である。

【０１２３】

出力部105は，一致数を第3の変数群の総数で除算した値が，所定の値(例えば，0.8)をよりも大きい場合，前記異常セットとして抽出された変数群を出力する(ステップS8)。

【０１２４】

第1の場合における前記例示の場合，一致数"7"を第3の変数群の総数"8"で除算した値は"0.88"である。従って，出力部105は，異常セットして抽出された，各種類の1つの変数値の組み合わせを示す変数群である，"{IP1}&{U1}&{A}"，"{IP1}&{U2}&{A}"，"{IP2}&{U1}&{A}"，"{IP2}&{U2}&{A}"，"{IP1}&{U1}&{B}"，"{IP1}&{U2}&{B}"，"{IP2}&{U1}&{B}"を出力する。なお，出力部105は，決定した第2の変数群"{IP1，IP2}&{U1，U2}&{A，B}"を出力してもよい。

【０１２５】

第2の場合として，決定した第2の変数群が，"{IP1,IP3} & {U1,U2} & {A,B,C}"であるとする。異常セットして抽出された第1の変数群が，"{IP1}&{U1}&{A}"，"{IP1}&{U1}&{B}"，"{IP1}&{U1}&{C}"，"{IP1}&{U2}&{A}"，"{IP1}&{U2}&{B}"，"{IP1}&{U2}&{C}"，"{IP3}&{U1}&{A}"，"{IP3}&{U1}&{B}"，"{IP3}&{U1}&{C}"，"{IP3}&{U2}&{A}"，"{IP3}&{U2}&{B}"であるとする(11個)。

【０１２６】

この場合，決定した第2の変数群"{IP1,IP3} & {U1,U2} & {A,B,C}"において，各種類の1つの変数値の組み合わせを示す第3の変数群の総数は12(2×2×3)である。具体的には，この第3の変数群は，"{IP1}&{U1}&{A}"，"{IP1}&{U1}&{B}"，"{IP1}&{U1}&{C}"，"{IP1}&{U2}&{A}"，"{IP1}&{U2}&{B}"，"{IP1}&{U2}&{C}"，"{IP3}&{U1}&{A}"，"{IP3}&{U1}&{B}"，"{IP3}&{U1}&{C}"，"{IP3}&{U2}&{A}"，"{IP3}&{U2}&{B}"，"{IP3}&{U2}&{C}"である。

【０１２７】

そして，第3の変数群の中で，異常セットして抽出された第1の変数群に一致する変数群の一致数は11である。

【０１２８】

この一致数"11"を第3の変数群の総数"12"で除算した値"0.92"は，所定の割合(例えば，0.8)以上である。従って，出力部105は，決定した第2の変数群，異常セットの少なくとも1つを出力する。

【０１２９】

なお，出力部105は，ステップS8において，出力対象の第2の変数群が複数ある場合，例えば，第2の変数群の各々に対応する一致数が多い順に，第2の変数群の各々を出力してもよい。例えば，ステップS8において，第2の変数群X，第2の変数群Y，第2の変数群Zの3つの変数群が出力される場合を想定する。そして，第2の変数群X，第2の変数群Y，第2の変数群Zに対応する一致数が，それぞれ，30，20，10であるとする。この場合，出力部105は，第2の変数群X，第2の変数群Y，第2の変数群Zを，この順に出力する。

【０１３０】

出力対象となる異常セットの抽出(図6のステップS7)，および出力(図6のステップS8)を実行する理由について説明する。

【０１３１】

検知対象となる所定の事象(例えば，異常事象)に起因する異常セットは，互いに同じような変数値を含む集団を形成することが多い。例えば，図2で説明したアクセス数に応じてポイントをユーザに付与する電子商取引を行う場合を想定する。この電子商取引のサイトを特定するURLが，例えばU1，U2であるとする。この場合，前記した悪意のあるユーザは，これらのURLに頻繁にアクセスし，アクセス数に応じたポイントを取得しようと努める。また，かかる悪意のあるユーザは，通常，特定のユーザ(例えば，ユーザA，B，C)である。

【０１３２】
そのため，異常セットの変数群は，URL U1，U2の何れか1と，ユーザA，B，Cの何れか1と，ある送信元IPアドレスとを含む変数群を含む。この場合，例えば，ユーザA,B,Cが，ポイントを得るためだけに，URL U1，U2にアクセスする場合を想定する。この場合，例えば，異常セットは，｛IP1&U1&A｝，｛IP1&U2&A｝，｛IP1&U1&B｝，｛IP1&U2&B｝，｛IP1&U1&C｝，｛IP1&U2&C｝など，URL U1，U2，ユーザA，B，Cを含む同じような変数群を形成する。
【０１３３】
しかし，ノイズに起因しない異常セット(換言すれば，悪意あるユーザによるアクセスと関係する異常セット)に対して，ノイズに起因する異常セット(換言すれば，悪意あるユーザによるアクセスと関係しない異常セット)は，同じような変数群を形成することが少ない。換言すれば，ノイズに起因する異常セットは，通常，ランダムに出現する。
【０１３４】
すなわち，ノイズに起因しない各異常セットにおける各変数間の類似度は，ノイズに起因する各異常セットにおける各変数間の類似度に比べて高いと見なすことができる。そして，異常事象の検出において，ノイズに起因する各異常セットの出力を抑制すれば，異常事象の誤検知を抑制できる。
【０１３５】
そこで，発明者は，異常セットとして抽出された変数群に一致する変数群の数を第3の変数群の総数で除算した値が，予め定められた所定の値をよりも大きい場合，異常セットとして抽出された変数群を出力すれば，異常事象の誤検知を抑制できることを見出した。
【０１３６】
この除算値が予め定められた所定の値をよりも大きい場合とは，複数の異常セットにおける各変数間の類似度が，予め定められた値よりも高いことと実質的に同義である。すなわち，この複数の異常セットは，互いに同じような変数値を含む集団を形成しているとみなすことができるため，ノイズに起因しない異常セットと見なすことができる。情報処理装置1は，前記したステップS8の実行により，ノイズに起因する異常セットの出力を抑制する。
【０１３７】
以上説明したように，本実施の形態によれば，ノイズに起因する異常セットの出力を抑制して，ノイズに起因しない異常セットを出力することができる。その結果，ノイズに起因する異常セットによる異常事象の誤検知を抑制することができる。
【０１３８】
次に，図7で説明した期待値を算出する理由について説明する。以下，過去のレコード群において，大半のレコードが，検知対象となる所定の事象に関連しないレコードであり，僅かなレコードが，検知対象となる所定の事象に関連するレコードである状況を想定する。なお，以下，各種類の1つの変数値の組み合わせを単一セットと適宜記す。
【０１３９】
すなわち，過去のレコード群における単一セットの大半(例えば，99.99%)は，正常セットであり，過去のレコード群における単一セットの極僅か(例えば，0.01%)が，ノイズに起因しない異常セットである状況を想定する。正常セットとは，異常セットではない単一セットや，ノイズに起因する異常セットを含む。
【０１４０】
この想定に基づき，本実施の形態の情報処理装置1は，過去のレコード群T1のレコードと，現在のレコード群T2とを比較して，現在のレコード群における単一セットが異常セットか判定している。
【０１４１】
本実施の形態の情報処理装置1とは異なる処理により，所定の事象を検出する装置(以下，検出装置と記す)を想定する。検出装置は，例えば，過去のレコード群において単一セットが実際に出現した出現率に基づき，現在のレコード群においてこの単一セットが出現

するであろう期待値を算出する。
【０１４２】
そして，検出装置は，この期待値と，現在のレコード群におけるこの単一セットを含むレコードの数との差分の絶対値を算出し，この差分が予め定められた閾値以上の場合，この単一セットを異常セットとして出力する。
【０１４３】
ここで，過去のレコード群において単一セット｛IP1&U1&A｝を含むレコード(以下，レコードX1と記す)の数が30であるとする。この単一セットを含むレコードX1(レコード数30)は，例えば，前記したように，ポイントを得るためだけに，ユーザ"B"が，URL"U1"に30回アクセスすることにより得られたレコードである。そのため，このレコードX1に含まれる単一セットは，ノイズに起因しない異常セットである。
【０１４４】
そして，現在のレコード群において単一セット｛IP1&U1&A｝を含むレコード(以下，レコードX2と記す)の数が1であるとする。この単一セットを含むレコードX2(レコード数1)は，例えば，商品を購入するために，ユーザ"B"が，URL"U1"に1回アクセスすることにより得られたレコードである。そのため，このレコードX2に含まれる単一セットは，正常セットである。
【０１４５】
ここで，検出装置は，過去のレコード群の総レコード数と，現在のレコード群の総レコード数との比率に，過去のレコード群における単一セットの出現数を乗算して，現在のレコード群において単一セットが出現するであろう期待値を算出する。
【０１４６】
ここで，過去のレコード群の総レコード数が20000，現在のレコード群の総レコード数が10000とする。そして，過去のレコード群における単一セット｛IP1&U1&A｝の出現数を30とする。すると，現在のレコード群において単一セット｛IP1&U1&A｝が出現するであろう期待値は，例えば，15(30×(10000/20000))である。そして，単一セット｛IP1&U1&A｝に対応する閾値を"5"とする。
【０１４７】
すると，検出装置は，期待値"15"と，レコードX2の数"1"との差"14"を算出する。検出装置は，この差は前記した閾値以上であるので，この単一セットを異常セットとして抽出する。このレコードX2に含まれる単一セットは，正常セットであり，この抽出は誤抽出である。この誤抽出の結果，この単一セットが異常セットとして出力され，その結果，異常事象の誤検知が増える。
【０１４８】
すなわち，過去のレコード群T1に異常セットの集団があった場合に，異常セットの集団における単一セットが，現在のレコード群T2においては，正常セットであるにもかかわらず，ノイズに起因しない異常セットとして出力されることがある。
【０１４９】
このような誤抽出を防ぐためには，前記した期待値を適切に算出することが望ましい。しかし，前記したように，過去のレコード群の総レコード数と，現在のレコード群の総レコード数との比率に，過去のレコード群における単一セットの出現数を乗算して，この単一セットの期待値を算出するだけでは，期待値を適切に算出することは困難である。
【０１５０】
特に，前記した想定とは異なり，何らかの理由で，前記したように過去のレコード群における単一セットの極僅か(例えば，0.01%)ではなく，より多い(例えば，1%)単一セットが，異常セットの場合，かかる誤抽出が頻発する。
【０１５１】
かかる誤抽出を防ぐためには，過去のレコード群における異常セットの集団の影響を低減した期待値を算出すればよいことが発明者により見出された。
【０１５２】

発明者は，過去のレコード群に含まれる異常セットにおける各種類の変数値の数は，過去のレコード群における各種類の変数値の数に比べて非常に小さいことに着目した。この異常セットを｛IP2&U1&A｝と例示する。そして，この異常セットにおける各種類の変数値を，例えば"U2"とする。さらに，過去のレコード群における各種類の変数値を，例えば，"U1"～"U20"とする。このように，過去のレコード群に含まれる異常セットにおける各種類の変数値(例えば，"U2")の数(1)は，過去のレコード群における各種類の変数値(例えば，"U1"～"U20")の数(20)に比べて非常に小さい。

【０１５３】

そして，発明者は，過去のレコード群における異常セットの集団の影響を低減した期待値を算出するために，以下の処理を実行することが有効であると見出した。すなわち，図7で説明したように，図6のステップS3で選択した変数群(第1の変数群)に含まれる各種類の1つの変数値の出現率の各々を乗算し，乗算値と現在のレコード群T2の総レコード数とを乗算し，第1の変数群の期待値を算出する。この算出処理によれば，前記した期待値を適切に算出することができるので，その結果，異常セットの誤抽出を抑制することができる。

【０１５４】

（閾値の算出）

図9，図10で説明した，現在のレコード群T2において，各種類の1つの変数値の組み合わせを示す第1の変数群に対応して閾値を算出する理由について説明する。

【０１５５】

図6～図8で説明したように，異常セット抽出部104は，現在のレコード群T2の中で，第1の変数群を含むレコードの総数と，第1の変数群の期待値との差の絶対値が，第1の変数群に対応して予め算出された閾値以上の場合，選択された変数群を異常セットして抽出する。

【０１５６】

ここで，過去のレコード群T1における各種類の1つの変数値の組み合わせを示す第1の変数群の期待値と，現在のレコード群T2において第1の変数群が実際に出現する出現数との差は，変数群毎に異なる。

【０１５７】

また，過去のレコード群T1，現在のレコード群T2に含まれるレコードの時間帯が異なれば，同じ変数群でも，前記差は，変数群毎に異なる。

【０１５８】

例えば，アクセス先の　URL"U11"～URL"U15"(図示しない)が，ニュースサイトに関するURLであるとする。そして，あるオフィスから，このニュースサイトにアクセスした際のアクセスレコード(アクセスログ)に基づき，図6～図8で説明したように，異常事象を検知するとする。

【０１５９】

ここで，このオフィスの作業時間以外の非作業時間帯(例えば，昼休みの時間帯)には，不特定の従業員が，このニュースサイトにアクセスすると想定する。この場合，非作業時間帯に対応する，過去および現在のレコード群において，様々な変数群の組み合わせが出現する(変数群のばらつきが大きい)。変数群のばらつきが大きい場合，前記したように，異常事象に起因する異常セットは，互いに同じような変数値(変数群のばらつきが小さい)を含む集団を形成することが多いので，ノイズに起因する異常セットが多くなる確率が高くなると仮定する。

【０１６０】

この仮定では，過去のレコード群T1における第1の変数群の期待値から，現在のレコード群T2において第1の変数群が実際に出現する出現数が離れていても(前記差が大きい)，この変数群は，異常セットでないと見なしてもよい。

【０１６１】

一方，このオフィスの作業時間帯には，特定の従業員が，業務でこのニュースサイトに

アクセスすると仮定する。この場合，作業時間時間帯に対応する，過去および現在のレコード群において，変数群のばらつきは，非作業時間帯において出現する変数群のばらつき度合いに比べて，小さくなる。変数群のばらつきが小さい場合，前記したように，異常事象に起因する異常セットは，互いに同じような変数値(変数群のばらつきが小さい)を含む集団を形成することが多いので，ノイズに起因する異常セットが多くなる確率が低くなると仮定する。

【０１６２】

この仮定では，過去のレコード群T1における第1の変数群の期待値から，現在のレコード群T2において第1の変数群が実際に出現する出現数が離れていなくても(前記差が小さい)，この変数群は，異常セットと見なしてよい。

【０１６３】

以上説明したように，過去のレコード群T1における第1の変数群の期待値と，現在のレコード群T2において第1の変数群が実際に出現する出現数との差は，変数群毎に異なる。それにもかかわらず，変数群毎に同じ閾値を静的に定めると，適切な異常セットが出力されない。すなわち，異常事象の誤検知が多発する。

【０１６４】

そこで，本実施の形態では，過去のレコード群T2を参照して，変数群毎に，閾値を算出する。本実施の形態では，過去のレコード群T1における第1の変数群の期待値と，現在のレコード群T2において第1の変数群が実際に出現する出現数との差が大きい程，この変数群は，ノイズに起因する変数群である可能性が高いと見なす。そして，前記差が大きい程，この変数群に対応する閾値を大きくして，ノイズに起因する変数群が，異常セットとして抽出されるのを抑制する。

【０１６５】

図9，図10を参照して，所定の閾値を算出する処理(図6のステップS1)について説明する。図9は，閾値を算出する処理を説明するフロー図である。

【０１６６】

ステップS11：閾値算出部103は，過去のレコード群T1における各種類の各変数値の出現率を算出する。ステップS11は，図6のステップS2と同様の処理である。

【０１６７】

ステップS12：閾値算出部103は，過去のレコード群T1において，各種類の1つの変数値の組み合わせを示す，未選択の第4の変数群を選択する。第4の変数群は，第1の変数群と同義である。なお，ステップS12～ステップS19の処理については，図10で具体的に説明する。

【０１６８】

ステップS13：閾値算出部103は，選択した変数群が，過去のレコード群T1において出現すると期待される期待値を算出する。

【０１６９】

ステップS14：閾値算出部103は，過去のレコード群T1において，任意の変数群を含むレコードの総数とこの変数群の期待値との差の二乗値を変数群毎に算出する。

【０１７０】

ステップS15：閾値算出部103は，ステップS12において全ての変数群が選択されたか判定する。全ての変数群が選択されていない場合(ステップS15/NO)，ステップS12に戻る。全ての変数群が選択された場合(ステップS15/YES)，ステップS16に移る。

【０１７１】

ステップS16：閾値算出部103は，任意の変数値を含む変数群毎の二乗値の総和の平方を算出し，算出した平方値を前記任意の変数値の不安定度として算出する。この任意の変数値とは，複数の変数値の中の何れかの変数値である。

【０１７２】

ステップS17：閾値算出部103は，過去のレコード群T1において，各種類の1つの変数値の組み合わせを示す，未選択の第4の変数群を選択する。

【０１７３】
ステップS18：閾値算出部103は，選択した変数群に対応する閾値を算出する。
【０１７４】
ステップS19：閾値算出部103は，ステップS17において全ての変数群が選択されたか判定する。全ての変数群が選択されていない場合(ステップS19/NO)，ステップS17に戻る。全ての変数群が選択された場合(ステップS19/YES)，図6のステップS2に移る。
【０１７５】
(閾値を算出する処理の具体例)
図10は，第1の変数群に対応する閾値を算出する処理を具体的に説明する図である。
【０１７６】
閾値算出部103は，図7で説明したように，図6のステップS2で説明した各変数値の出現率を算出する(ステップS11)。なお，閾値算出部103は，期待値算出部102が算出した出現率をそのまま利用してもよい。
【０１７７】
閾値算出部103は，過去のレコード群T1において，変数の各種類の1つの変数値の組み合わせを示す第4の変数群を選択し，RAM13に記憶する(ステップS12)。
【０１７８】
例えば，変数の各種類は，IPアドレス，URL，ユーザである，閾値算出部103は，過去のレコード群T1において，変数の種類"IPアドレス"の変数値(例えば，IP1，IP2…)の何れか1つを選択する。閾値算出部103は，過去のレコード群T1において，変数の種類"URL"の変数値(例えば，U1，U2…)の何れか1つを選択する。閾値算出部103は，過去のレコード群T1において，変数の種類"ユーザ"の変数値(例えば，A，B…)の何れか1つを選択する。
【０１７９】
そして，閾値算出部103は，選択した各種類の1つの変数値を組み合わせて，各種類の1つの変数値の組み合わせを示す第4の変数群とする。例えば，第4の変数群として，"IP1&U2&A"，"IP2&U2&B"，" IP1&U1&B"などがある。
【０１８０】
閾値算出部103は，図10の過去の期待値テーブルT51に示すように，第4の変数群をRAM13に記憶する。
【０１８１】
閾値算出部103は，過去のレコード群T1において，各種類の相違する変数値について前記した組み合わせを実行し，全ての組み合わせを示す全ての変数群を得る。例えば，過去のレコード群T1において，変数の種類"IPアドレス"の変数値がIP1，IP2の2つ，変数の種類"URL"の変数値がU1，U2，U3の3つ，変数の種類"ユーザ"の変数値がA，Bの2つの場合を想定する。この場合，閾値算出部103は，全ての組み合わせを示す全ての変数群は，12(2×3×2)個の変数群である。閾値算出部103は，全ての変数群の中から未選択の1つの変数群を選択する(ステップS12)。
【０１８２】
(過去のレコード群T1における期待値の算出)
次に，過去のレコード群T1における期待値の算出(図9のステップS13)について説明する。
【０１８３】
閾値算出部103は，ステップS12で選択した変数群(第4の変数群)に含まれる各種類の1つの変数値の出現率の各々を乗算し，この乗算値と，過去のレコード群T1の総レコード数とを乗算し，RAM13に記憶する(ステップS13)。閾値算出部103は，RAM13に記憶した，前記した乗算値と過去のレコード群T1の総レコード数とを乗算した乗算値を，過去のレコード群T1において，選択した変数群が出現すると期待される期待値とする(ステップS13)。
【０１８４】
前記の例では，過去のレコード群T1の総レコード数は"20000"である。選択された変数群は，例えば"IP1&U2&A"である。変数群"IP1&U2&A"における種類"IPアドレス"の1つの変

数値"IP1"の出現率は"0.05"(図7のIP出現率テーブルT21参照)である。変数群"IP1&U2&A"における種類"URL"の1つの変数値"U2"の出現率は"0.10"(図7のURL出現率テーブルT22参照)である。変数群"IP1&U2&A"における種類"ユーザ"の1つの変数値"A"の出現率は"0.08"(図7のユーザ出現率テーブルT23参照)である。

【０１８５】

閾値算出部103は，変数群"IP1&U2&A"における種類"IPアドレス"，"URL"，"ユーザ"の出現率"0.05"，"0.10"，"0.08"の各々を乗算し，この乗算値(0.0004)と，過去のレコード群T1の総レコード数"20000"とを乗算し，乗算値"8.0"(20000×0.0004(0.05×0.10×0.08))を算出する。

【０１８６】

閾値算出部103は，この乗算値"8.0"を，過去のレコード群T1において，前記した変数群"IP1&U2&A"が出現すると期待される期待値"8.0"とする。

【０１８７】

閾値算出部103は，図10の過去の期待値テーブルT51に示すように，変数群と，この変数群の期待値とを対応付けてRAM13に記憶する。

【０１８８】

(二乗値の算出)

過去のレコード群T1において，任意の変数群を含むレコードの総数とこの変数群の期待値との差の二乗値を変数群毎に算出する処理(図6のステップS14)について説明する。任意の変数群を含むレコードの総数は，任意の変数群が過去のレコード群T1において出現する数である。

【０１８９】

閾値算出部103は，過去のレコード群T1の中で，ステップS12で選択した変数群を含むレコードの数と，選択した変数群の期待値との差を算出しRAM13に記憶する(ステップS14)。

【０１９０】

選択した変数群が"IP1&U2&A"の場合，この変数群の期待値は，"8.0"である。また，過去のレコード群T1の中で，変数群"IP1&U2&A"(IPアドレス"IP1"，URL"U2"，ユーザ"A")を含むレコードの数は"4"であるとする。この場合，閾値算出部103は，選択した変数群"IP1&U2&A"について，前記した差"-4.0"(4-8.0)を算出する。

【０１９１】

閾値算出部103は，図10の二乗値算出テーブルT61に示すように，変数群と，過去のレコード群T1の中でこの変数群を含むレコードの数(出現数)と，このレコードの数とこの変数群の期待値との差とを対応付けてRAM13に記憶する。二乗値算出テーブルT61において，過去のレコード群T1の中で前記した変数群を含むレコードの数を出現数としている。

【０１９２】

閾値算出部103は，過去のレコード群T1において，任意の変数群(選択した変数群)を含むレコードの数と前記変数群の期待値との差の二乗値を異なる変数群毎(選択した変数群について)に算出し，RAM13に記憶する(ステップS14)。

【０１９３】

図10の二乗値算出テーブルT61の例では，過去のレコード群T1において変数群"IP1&U2&A"を含むレコードの数は"4"であり，このレコード数とこの変数群の期待値(8.0)との差(-4.0(4-8))の二乗値は16($(-4.0)^2$)である。なお，過去のレコード群T1において変数群"IP2&U2&B
"を含むレコードの数は"6"であり，このレコード数とこの変数群の期待値(4.8)との差の二乗値は1.44($(1.2)^2$)である。

【０１９４】

閾値算出部103は，各種類の1つの変数値の全ての組み合わせを示す変数群についてステップS12～ステップS14の処理が終了するまで(ステップS15/NO)，ステップS12～ステップS14の処理を繰り返す。なお，過去のレコード群T1において変数群"IP1&U1&B"を含むレコードの数は"0"であり，このレコード数とこの変数群の期待値(9.0)との差の二乗値は81((0-

9)^2)である。過去のレコード群T1において変数群"IP1&U1&C"を含むレコードの数は"1"であり，このレコード数とこの変数群の期待値(例えば2.0)との差の二乗値は1((1-2.0)^2)である。

【０１９５】

閾値算出部103は，この処理が終了すると(ステップS15/YES)，ステップS16に移る。ステップS14で説明したように，閾値算出部103は，過去のレコード群T1における各種類の1つの変数値の組み合わせを示す第4の変数群が過去のレコード群T1において出現する数(出現数)と，第4の変数群が過去のレコード群T1において出現すると期待される期待値との差の二乗値を異なる第4の変数群毎に算出する。なお，第4の変数群毎における第4の変数群は，図10の二乗値算出テーブルT61に示した，例えば"IP1&U2&A"，"IP1&U1&B"，"IP1&U1&C"である。

【０１９６】

(不安定度の算出)

閾値算出部103は，任意の変数値を含む各変数群の前記二乗値の総和の平方を算出し，算出した平方値を前記任意の変数値の不安定度として，前記任意の変数値と対応付けてRAM13に記憶する(ステップS16)。

【０１９７】

例えば，任意の変数値をIPアドレスの"IP1"とする。そして，IPアドレス"IP1"を含む変数群は，図10の二乗値算出テーブルT61に示した"IP1&U2&A"，"IP1&U1&B"，"IP1&U1&C"であるとする。変数群"IP1&U2&A"の二乗値は"16"，変数群"IP1&U1&B"の二乗値は"81"，変数群"IP1&U1&C"の二乗値は"1"である。

【０１９８】

従って，閾値算出部103は，例えば，任意の変数値であるIPアドレス"IP1"を含む，異なる変数群毎の二乗値の総和"98"(16+81+1)を算出する。そして，閾値算出部103は，この総和"98"の平方値"9.9"((98)$^{1/2}$)を任意の変数値であるIPアドレス"IP1"の不安定度として，任意の変数値と対応付けてRAM13に記憶する(ステップS16)。

【０１９９】

図10のIPアドレス不安定度テーブルT71は，IPアドレスの各変数値と，この各変数値の不安定度とを対応付けて記憶したテーブルである。URL不安定度テーブルT72は，URLの各変数値と，この各変数値の不安定度とを対応付けて記憶したテーブルである。ユーザ不安定度テーブルT73は，ユーザの各変数値と，この各変数値の不安定度とを対応付けて記憶したテーブルである。

【０２００】

閾値算出部103は，過去のレコード群T1において各種類の1つの変数値の組み合わせを示す第4の変数群を選択し，RAM13に記憶する(ステップS17)。ステップS17は，ステップS12と同じ処理なので，その説明を省略する。

【０２０１】

(閾値の算出)

閾値算出部103は，ステップS17で選択した変数群における各種類の1つの変数値の不安定度の各々を乗算し，RAM13に記憶する(ステップS18)。閾値算出部103は，この乗算値と，現在のレコード群T2の総レコード数とを乗算し，この乗算値を定数倍した値を，前記変数群に対応する閾値とする(ステップS18)。

【０２０２】

前記の例では，現在のレコード群T2の総レコード数は"10000"である。前記した組み合わせの変数群は，例えば"IP1&U2&A"である。変数群"IP1&U2&A"における種類"IPアドレス"の1つの変数値"IP1"の不安定度は"9.9"(図10のIPアドレス不安定度テーブルT71参照)である。変数群"IP1&U2&A"における種類"URL"の1つの変数値"U2"の不安定度は"9.0"(図10のURL不安定度テーブルT72参照)である。変数群"IP1&U2&A"における種類"ユーザ"の1つの変数値"A"の不安定度は"3.1"(図10のユーザ不安定度テーブルT73参照)である。

【０２０３】

閾値算出部103は，変数群"IP1&U2&A"における種類IPアドレス"IP1"の不安定度"9.9"と，URL"U2"の不安定度"9.0"と，ユーザ"A"の不安定度"3.1"とを乗算する。そして，閾値算出部103は，この乗算値(276.21)と，現在のレコード群T2の総レコード数"10000"とを乗算した乗算値(276.21×10000)の定数倍(例えば，10^{-6})"2.8"を算出する。なお，定数倍は，情報処理装置1の管理者により予め定められた値である。

【０２０４】

閾値算出部103は，この乗算値"2.8"を，前記した変数群"IP1&U2&A"に対応する閾値"2.8"とする。

【０２０５】

閾値算出部103は，図10の閾値テーブルT81に示すように，前記した変数群と，この変数群に対応する閾値とを対応付けてRAM13に記憶する。

【０２０６】

閾値算出部103は，過去のレコード群T1において各種類の1つの変数値の全ての組み合わせを示す変数群について，ステップS18の処理が終了するまで(ステップS19/NO)，ステップS17，ステップS18の処理を繰り返す。

【０２０７】

そして，閾値算出部103は，この処理が終了すると(ステップS19/YES)，図6のステップS2に移る。以後，情報処理装置1は，算出した閾値を参照して，異常セットを抽出する(図6のステップS2～ステップS8)。

【０２０８】

ステップS18で説明したように，閾値算出部103は，第1の変数群における各種類の変数値毎に，この変数値を含む第4の変数群毎の二乗値の総和の平方値(不安定度)を算出する。第1の変数群における各種類の変数値毎に算出された平方値は，前記した変数値毎の不安定度である。そして，閾値算出部103は，第1の変数群における各種類の変数値の平方値の各々を乗算し，乗算値に基づき，第1の変数群に対応する閾値を算出する。ここで，閾値算出部103は，乗算値を定数倍(例えば，10^{-6})した値を，第1の変数群に対応する閾値として算出する。

【０２０９】

本実施の形態の情報処理装置によれば，過去のレコード群の内容に応じて，変数群毎に異なる閾値を動的に算出する。そして，この閾値を参照して，異常セットを抽出するので，適切な異常セットを出力できる。

【０２１０】

以上の実施の形態をまとめると，次の付記のとおりである。

【０２１１】

（付記１）

複数の種類の変数を含む第1のレコード群と，前記複数の種類の変数を含む，前記第1のレコード群と異なる第2のレコード群とを記憶する記憶部と，

前記第1のレコード群における前記各種類の1つの変数値の組み合わせを示す第1の変数群が，前記第2のレコード群において出現すると期待される期待値を算出し，前記第1の変数群の前記期待値と，前記第2のレコード群における前記第1の変数群を含むレコードの数とに基づき，前記第1の変数群を抽出し，抽出された前記第1の変数群と，前記第2のレコード群における前記各種類の1以上の変数値の組み合わせを示す第2の変数群とに基づき，前記第2の変数群及び抽出した前記第1の変数群の少なくとも1つを出力する，処理部と，

を有することを特徴とする情報処理装置。

【０２１２】

（付記２）

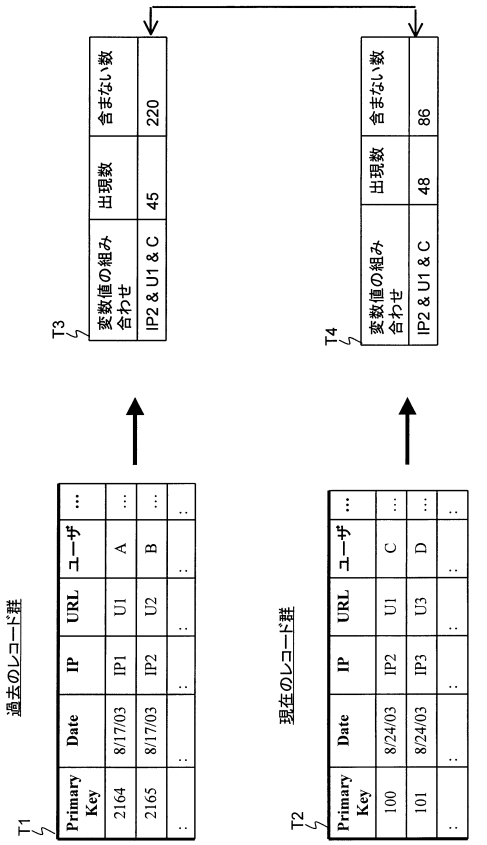
付記1において，

前記処理部は，抽出された前記第1の変数群と，前記第2の変数群における前記各種類の1つの変数値の組み合わせを示す第3の変数群とに基づき，前記第2の変数群及び抽出した前記第1の変数群の少なくとも1つを出力する情報処理装置。

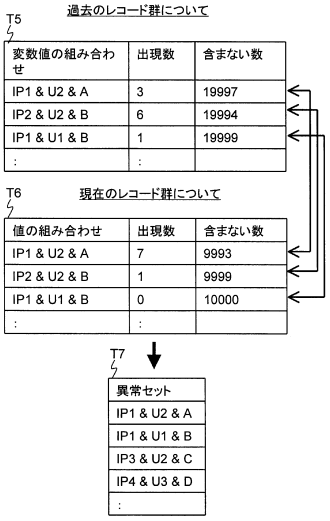
【０２１３】
（付記３）
付記2において，
前記処理部は，前記出力において，前記第2の変数群に含まれる前記第3の変数群の総数と，前記第3の変数群の中で，抽出した前記第1の変数群に一致する変数群の数とに基づき，前記第2の変数群，抽出した前記第1の変数群の少なくとも1つを出力するか判定する情報処理装置。
【０２１４】
（付記４）
付記3において，
前記処理部は，前記出力において，前記一致する変数群の数を前記第3の変数群の総数で除算した値が，所定の値よりも大きい場合，前記第2の変数群，抽出した前記第1の変数群の少なくとも1つを出力する情報処理装置。
【０２１５】
（付記５）
付記1において，
前記処理部は，前記期待値の算出において，前記第1のレコード群において各種類の変数値が出現する割合を示す出現率に基づき，前記第1の変数群の前記期待値を算出する情報処理装置。
【０２１６】
（付記６）
付記5において，
前記処理部は，前記期待値の算出において，前記第1の変数群に含まれる各種類の1つの変数値の前記出現率の各々を乗算し，乗算値と前記第2のレコード群の総レコード数とを乗算し，前記第1の変数群の前記期待値を算出する情報処理装置。
【０２１７】
（付記７）
付記1において，
前記処理部は，前記抽出において，前記第2のレコード群の中で前記第1の変数群を含むレコードの総数と，前記第1の変数群の前記期待値との差の絶対値が，前記第1の変数群に対応して予め算出された閾値以上の場合，前記第2のレコード群から前記第1の変数群を抽出する情報処理装置。
【０２１８】
（付記８）
付記7において，
前記処理部は，さらに，前記第1の変数群が，前記第1のレコード群において出現すると期待される期待値と，前記第2のレコード群の中で前記第1の変数群を含むレコードの数とに基づき，前記第1の変数群に対応する前記閾値を算出する情報処理装置。
【０２１９】
（付記９）
付記8において，
前記処理部は，前記閾値の算出において，前記第1のレコード群における前記各種類の1つの変数値の組み合わせを示す第4の変数群が前記第1のレコード群において出現する数と，前記第4の変数群が前記第1のレコード群において出現すると期待される期待値との差の二乗値を異なる第4の変数毎に算出し，前記第1の変数群における各種類の変数値毎に，当該変数値を含む前記第4の変数群毎の二乗値の総和の平方値を算出し，算出した平方値の各々を乗算し，乗算値に基づき，前記第1の変数群に対応する前記閾値を算出する情報処理装置。
【０２２０】
（付記１０）

付記9において，
前記処理部は，前記閾値の算出において，前記乗算値を定数倍した値を，前記第1の変数群に対応する前記閾値として算出する情報処理装置。
【０２２１】
（付記１１）
コンピュータに，
複数の種類の変数を含む第1のレコード群と，前記複数の種類の変数を含む，前記第1のレコード群と異なる第2のレコード群とを入力し，
前記各種類の1つの変数値の組み合わせを示す第1の変数群が，前記第2のレコード群において出現すると期待される期待値を算出し，
前記第1の変数群の前記期待値と，前記第2のレコード群における前記第1の変数群を含むレコードの数とに基づき，前記第1の変数群を抽出し，
抽出された前記第1の変数群と，前記第2のレコード群における前記各種類の1以上の変数値の組み合わせを示す第2の変数群とに基づき，前記第2の変数群及び抽出した前記第1の変数群の少なくとも1つを出力する，
処理を実行させることを特徴とするプログラム。
【０２２２】
（付記１２）
複数の種類の変数を含む第1のレコード群と，前記複数の種類の変数を含む，前記第1のレコード群と異なる第2のレコード群とを記憶する情報処理装置で実行される情報処理方法であって，
前記情報処理装置は，
前記各種類の1つの変数値の組み合わせを示す第1の変数群が，前記第2のレコード群において出現すると期待される期待値を算出し，
前記第1の変数群の前記期待値と，前記第2のレコード群における前記第1の変数群を含むレコードの数とに基づき，前記第1の変数群を抽出し，
抽出された前記第1の変数群と，前記第2のレコード群における前記各種類の1以上の変数値の組み合わせを示す第2の変数群とに基づき，前記第2の変数群及び抽出した前記第1の変数群の少なくとも1つを出力する，
ことを特徴とする情報処理方法。
【符号の説明】
【０２２３】
1…情報処理装置，2…クライアント端末，31～3p…Webサーバ，11…CPU，12…ストレージ，13…RAM，14…外部記憶媒体読み取り装置，15…ネットワークインターフェイス，101…入力部，102…期待値算出部，103…閾値算出部，104…異常セット抽出部，105…出力部。

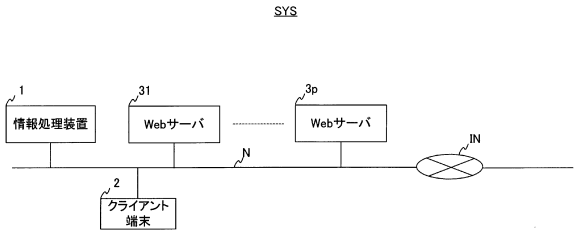
【図１】


過去のレコード群
現在のレコード群
T1
T2
T3
T4
Primary Key
Date
IP
URL
ユーザ
変数値の組み合わせ
出現数
含まない数
2164
2165
100
101
8/17/03
8/24/03
IP1
IP2
IP3
U1
U2
U3
A
B
C
D
IP1 & U1 & C
IP2 & U1 & C
220
45
86
48

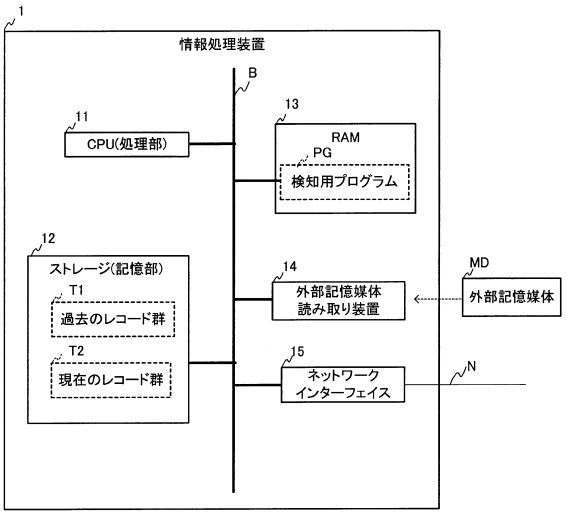
【図２】


過去のレコード群について
T5
変数値の組み合わせ
出現数
含まない数
IP1 & U2 & A
IP2 & U2 & B
IP1 & U1 & B
3
6
1
19997
19994
19999
現在のレコード群について
T6
値の組み合わせ
出現数
含まない数
IP1 & U2 & A
IP2 & U2 & B
IP1 & U1 & B
7
1
0
9993
9999
10000
T7
異常セット
IP1 & U2 & A
IP1 & U1 & B
IP3 & U2 & C
IP4 & U3 & D

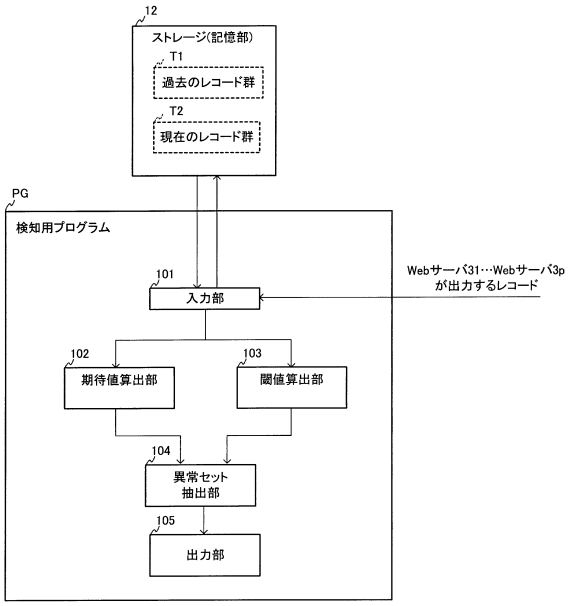
【図３】


SYS
1
情報処理装置
31
Webサーバ
3p
Webサーバ
N
IN
2
クライアント端末

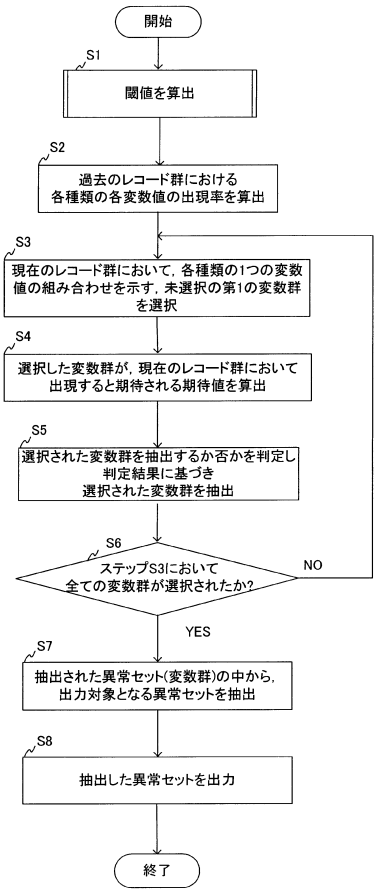
【図４】


1
情報処理装置
B
11
CPU(処理部)
13
RAM
PG
検知用プログラム
12
ストレージ(記憶部)
T1
過去のレコード群
T2
現在のレコード群
14
外部記憶媒体読み取り装置
MD
外部記憶媒体
15
ネットワークインターフェイス
N

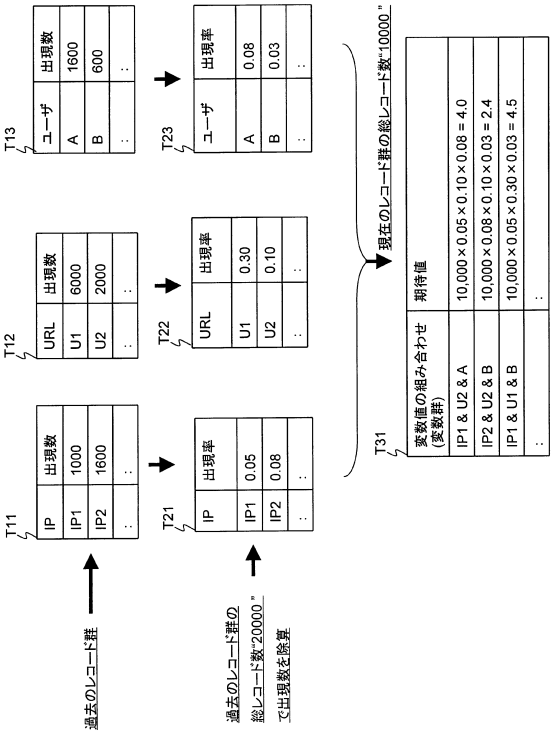
【図５】


12
ストレージ(記憶部)
T1
過去のレコード群
T2
現在のレコード群
PG
検知用プログラム
101
入力部
Webサーバ31…Webサーバ3pが出力するレコード
102
期待値算出部
103
閾値算出部
104
異常セット抽出部
105
出力部

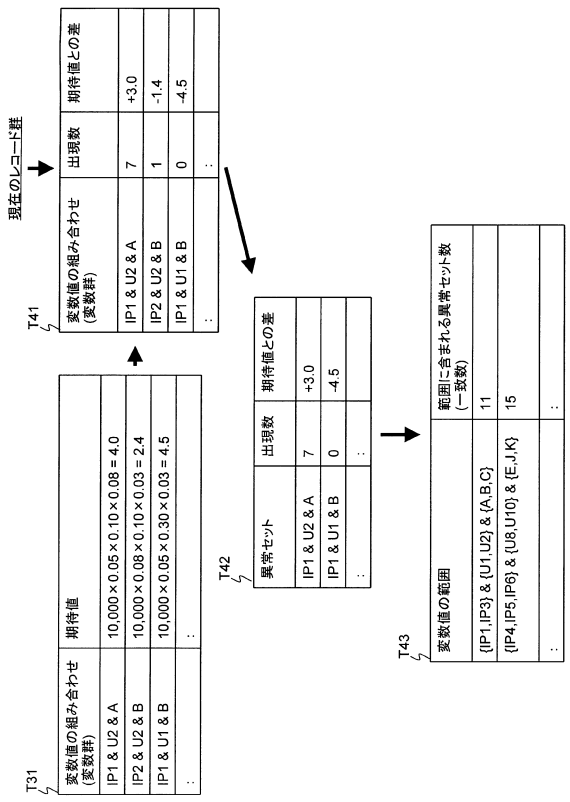
【図６】


開始
S1
閾値を算出
S2
過去のレコード群における
各種類の各変数値の出現率を算出
S3
現在のレコード群において，各種類の1つの変数
値の組み合わせを示す，未選択の第1の変数群
を選択
S4
選択した変数群が，現在のレコード群において
出現すると期待される期待値を算出
S5
選択された変数群を抽出するか否かを判定し
判定結果に基づき
選択された変数群を抽出
S6
ステップS3において
全ての変数群が選択されたか?
NO
YES
S7
抽出された異常セット(変数群)の中から，
出力対象となる異常セットを抽出
S8
抽出した異常セットを出力
終了

【図７】


T13
ユーザ
出現数
A 1600
B 600
T23
ユーザ
出現率
A 0.08
B 0.03
T12
URL
出現数
U1 6000
U2 2000
T22
URL
出現率
U1 0.30
U2 0.10
T11
IP
出現数
IP1 1000
IP2 1600
T21
IP
出現率
IP1 0.05
IP2 0.08
過去のレコード群
過去のレコード群の
総レコード数"20000"
で出現数を除算
現在のレコード群の総レコード数"10000"
T31
変数値の組み合わせ
(変数群)
期待値
IP1 & U2 & A 10,000×0.05×0.10×0.08 = 4.0
IP2 & U2 & B 10,000×0.08×0.10×0.03 = 2.4
IP1 & U1 & B 10,000×0.05×0.30×0.03 = 4.5

【図８】


現在のレコード群
T41
変数値の組み合わせ
(変数群)
出現数
期待値との差
IP1 & U2 & A 7 +3.0
IP2 & U2 & B 1 -1.4
IP1 & U1 & B 0 -4.5
T42
異常セット
出現数
期待値との差
IP1 & U2 & A 7 +3.0
IP1 & U1 & B 0 -4.5
T43
変数値の範囲
範囲に含まれる異常セット数
(一致数)
{IP1,IP3} & {U1,U2} & {A,B,C} 11
{IP4,IP5,IP6} & {U8,U10} & {E,J,K} 15
T31
変数値の組み合わせ
(変数群)
期待値
IP1 & U2 & A 10,000×0.05×0.10×0.08 = 4.0
IP2 & U2 & B 10,000×0.08×0.10×0.03 = 2.4
IP1 & U1 & B 10,000×0.05×0.30×0.03 = 4.5

【図９】

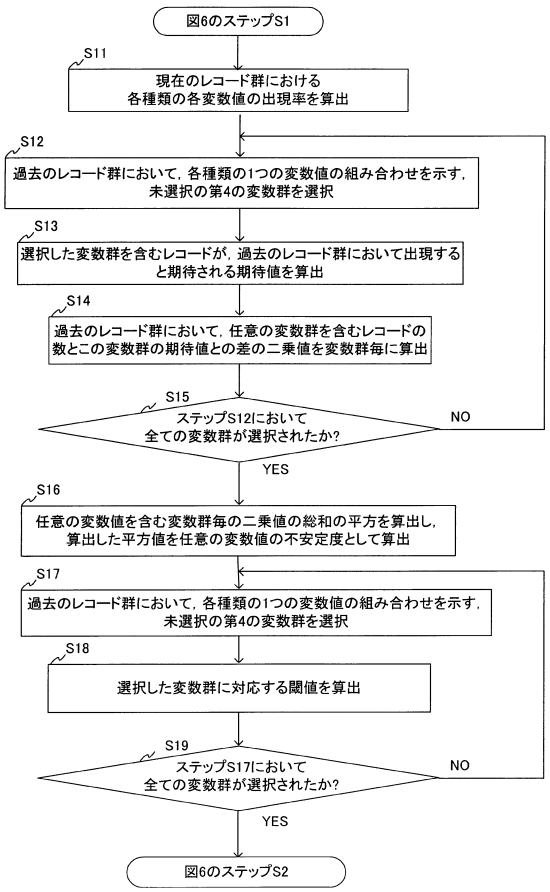
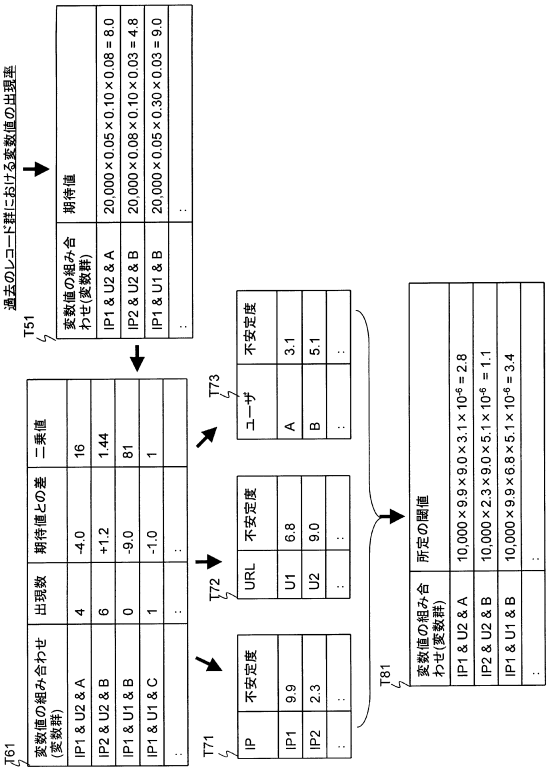


図6のステップS1
S11
現在のレコード群における
各種類の各変数値の出現率を算出
S12
過去のレコード群において，各種類の1つの変数値の組み合わせを示す，
未選択の第4の変数群を選択
S13
選択した変数群を含むレコードが，過去のレコード群において出現する
と期待される期待値を算出
S14
過去のレコード群において，任意の変数群を含むレコードの
数とこの変数群の期待値との差の二乗値を変数群毎に算出
S15
ステップS12において
全ての変数群が選択されたか?
NO
YES
S16
任意の変数値を含む変数群毎の二乗値の総和の平方を算出し，
算出した平方値を任意の変数値の不安定度として算出
S17
過去のレコード群において，各種類の1つの変数値の組み合わせを示す，
未選択の第4の変数群を選択
S18
選択した変数群に対応する閾値を算出
S19
ステップS17において
全ての変数群が選択されたか?
NO
YES
図6のステップS2

【図１０】


過去のレコード群における変数値の出現率
T61
変数値の組み合わせ(変数群)
出現数
期待値との差
二乗値
IP1 & U2 & A
4
-4.0
16
IP2 & U2 & B
6
+1.2
1.44
IP1 & U1 & B
0
-9.0
81
IP1 & U1 & C
1
-1.0
1
T51
変数値の組み合わせ(変数群)
期待値
IP1 & U2 & A
20,000×0.05×0.10×0.08 = 8.0
IP2 & U2 & B
20,000×0.08×0.10×0.03 = 4.8
IP1 & U1 & B
20,000×0.05×0.30×0.03 = 9.0
T71
IP
不安定度
IP1
9.9
IP2
2.3
T72
URL
不安定度
U1
6.8
U2
9.0
T73
ユーザ
不安定度
A
3.1
B
5.1
T81
変数値の組み合わせ(変数群)
所定の閾値
IP1 & U2 & A
10,000×9.9×9.0×3.1×10-6 = 2.8
IP2 & U2 & B
10,000×2.3×9.0×5.1×10-6 = 1.1
IP1 & U1 & B
10,000×9.9×6.8×5.1×10-6 = 3.4

フロントページの続き

審査官　圓道　浩史

(56)参考文献　特開２００２－１８９５９７（ＪＰ，Ａ）
米国特許出願公開第２００２／００８７５４０（ＵＳ，Ａ１）
特開２００６－０４８２５３（ＪＰ，Ａ）
米国特許出願公開第２００８／０１９８７５６（ＵＳ，Ａ１）
米国特許出願公開第２００８／０２６３６６３（ＵＳ，Ａ１）
特開２０１０－０９７３４２（ＪＰ，Ａ）
特開２０１１－１９６７３８（ＪＰ，Ａ）
米国特許出願公開第２０１１／０２２７７８６（ＵＳ，Ａ１）
再公表特許第２０１２／００１７９５（ＪＰ，Ａ１）
米国特許出願公開第２０１３／０１１７２９４（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ１２／１４
２１／００－２１／８８
Ｇ０９Ｃ　１／００－　５／００
Ｈ０４Ｋ　１／００－　３／００
Ｈ０４Ｌ　９／００－　９／３８