

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2008-176752

(P2008-176752A)

(43) 公開日 平成20年7月31日(2008.7.31)

(51) Int.Cl.	F I	テーマコード (参考)
G06F 21/22 (2006.01)	G06F 9/06 660N	5B276
G06F 21/20 (2006.01)	G06F 15/00 330A	5B285

審査請求 未請求 請求項の数 16 O L (全 20 頁)

(21) 出願番号 特願2007-12070 (P2007-12070)
 (22) 出願日 平成19年1月22日 (2007.1.22)

(71) 出願人 301022471
 独立行政法人情報通信研究機構
 東京都小金井市貫井北町4-2-1
 (74) 代理人 100130111
 弁理士 新保 斉
 (72) 発明者 衛藤 将史
 東京都小金井市貫井北町4-2-1 独立
 行政法人情報通信研究機構内
 (72) 発明者 園田 光太郎
 東京都小金井市貫井北町4-2-1 独立
 行政法人情報通信研究機構内
 (72) 発明者 吉岡 克成
 東京都小金井市貫井北町4-2-1 独立
 行政法人情報通信研究機構内

最終頁に続く

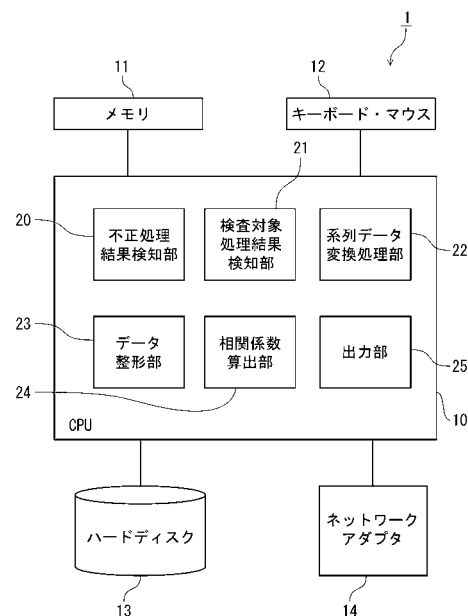
(54) 【発明の名称】 系列データ間の類似性検査方法及び装置

(57) 【要約】

【課題】 系列データ間の類似性を高精度に検査する手法を創出し、それによって広域ネットワークにおけるインシデントの解析結果と、各マルウェアの特性とを効率よく比較し、両者の相関を得ること。

【解決手段】 ネットワーク上で他のコンピュータに対して不正処理を行うマルウェアが送信する系列データと、検査対象のソフトウェアが送信する系列データとを比較してその類似性を検査する類似性検査方法を提供する。不正処理結果検知手段20が、マルウェアの系列データを得るとともに、検査対象処理結果検知手段21が、検査対象の系列データを得る。系列データ変換処理手段22が、両系列データをフーリエ変換して正規化した後に、相関係数算出手段24が両者の相関係数を算出する。

【選択図】 図1



【特許請求の範囲】

【請求項 1】

ネットワーク上で他のコンピュータに対して不正処理を行う第 1 のソフトウェアの処理結果から得られる第 1 の系列データと、検査対象の第 2 のソフトウェアの処理結果から得られる第 2 の系列データとを比較してその類似性を検査する類似性検査方法であって、

不正処理結果検知手段が、該第 1 のソフトウェアの不正処理の結果を検出しその結果を第 1 の系列データとして得る不正処理結果検知ステップ、

検査対象処理結果検知手段が、該第 2 のソフトウェアの処理結果を検出しその結果を第 2 の系列データとして得る検査対象処理結果検知ステップ、

コンピュータの系列データ変換処理手段が、該第 1 の系列データ及び該第 2 の系列データを次の各工程：

コンピュータの演算手段が、入力された系列データについて、離散フーリエ変換して横軸に周波数、縦軸に周波数成分の強度を表すスペクトラムを得る離散フーリエ変換処理工程、

コンピュータのデータ抽出手段が、該スペクトラムについて、該縦軸において所定の閾値を超える周波数強度を持つ要素を抽出し、その出現位置の値の系列を得る出現位置系列取得工程、

コンピュータの出現位置値正規化処理手段が、該スペクトラムにおける最も強度の強いスペクトルの出現位置の値で、該出現位置の値の系列の値を全て除算し、正規化された出現位置の値の系列を得る出現位置値正規化処理工程、

コンピュータの調波構造正規化処理手段が、該正規化された出現位置値の系列について、標準偏差を用いた正規化処理を行う調波構造正規化処理工程

により変換する系列データ変換処理ステップ、

コンピュータの相関係数算出手段が、変換後の第 1 の系列データと変換後の第 2 の系列データとから所定の相関関係式を用いて相関係数を算出する相関係数算出ステップ

を有する

ことを特徴とする系列データ間の類似性検査方法。

【請求項 2】

前記系列データ変換処理ステップにおいて、

前記離散フーリエ変換処理工程の後に、

コンピュータのデータ抽出手段が、所定の閾値以上の高周波数成分を除去する高周波数成分除去処理工程を含む

ことを特徴とする請求項 1 に記載の系列データ間の類似性検査方法。

【請求項 3】

前記系列データ変換処理ステップの後に、

コンピュータのデータ整形手段が、変換後の第 1 及び第 2 の系列データについて、基本周波数における出現位置において両系列データを同期すると共に、該基本周波数間で出現位置値がない場合には所定の値を補完するデータ整形ステップを有する

ことを特徴とする請求項 1 又は 2 に記載の系列データ間の類似性検査方法。

【請求項 4】

前記不正処理結果検知手段及び検査対象処理結果検知手段が、それぞれ第 1 及び第 2 のソフトウェアによる、他のコンピュータのネットワークアドレスに対する連続的なスキャンを検知する構成であって、前記第 1 及び第 2 の系列データとして、スキャンしたネットワークアドレスの値の列を用いる

ことを特徴とする請求項 1 ないし 3 のいずれかに記載の系列データ間の類似性検査方法。

【請求項 5】

前記第 1 のソフトウェアが、閉じられたネットワークにおいて検査のために実行されるマルウェアであり、前記第 2 のソフトウェアが、広域ネットワークにおいて実際に実行され、マルウェアと疑われる挙動を示すソフトウェアであり、

10

20

30

40

50

前記請求項 1 ないし 4 のいずれかに記載の系列データ間の類似性検査方法を用いて、該第 2 のソフトウェアの種類を、該第 1 のソフトウェアとの類似性を検査することにより特定する

ことを特徴とするマルウェアの検査方法。

【請求項 6】

2 つ以上の系列データを比較して系列データ間の類似性を検査する類似性検査方法であって、

コンピュータの系列データ変換処理手段が、該各系列データを次の各工程：

コンピュータの演算手段が、入力された系列データについて、離散フーリエ変換して横軸に周波数、縦軸に周波数成分の強度を表すスペクトラムを得る離散フーリエ変換処理工程、

コンピュータのデータ抽出手段が、該スペクトラムについて、該縦軸において所定の閾値を超える周波数強度を持つ要素を抽出し、その出現位置の値の系列を得る出現位置系列取得工程、

コンピュータの出現位置値正規化処理手段が、該スペクトラムにおける最も強度の強いスペクトルの出現位置の値で、該出現位置の値の系列の値を全て除算し、正規化された出現位置の値の系列を得る出現位置値正規化処理工程、

コンピュータの調波構造正規化処理手段が、該正規化された出現位置値の系列について、標準偏差を用いた正規化処理を行う調波構造正規化処理工程

により変換する系列データ変換処理ステップ、

コンピュータの相関係数算出手段が、変換後の各系列データから所定の相関関係式を用いて相関係数を算出する相関係数算出ステップ

を有する

ことを特徴とする系列データ間の類似性検査方法。

【請求項 7】

前記系列データ変換処理ステップにおいて、

前記離散フーリエ変換処理工程の後に、

コンピュータのデータ抽出手段が、所定の閾値以上の高周波数成分を除去する高周波数成分除去処理工程を含む

ことを特徴とする請求項 6 に記載の系列データ間の類似性検査方法。

【請求項 8】

前記系列データ変換処理ステップの後に、

コンピュータのデータ整形手段が、変換後の各系列データについて、基本周波数における出現位置において各系列データを同期すると共に、該基本周波数間で出現位置値がない場合には所定の値を補完するデータ整形ステップを有する

ことを特徴とする請求項 6 又は 7 に記載の系列データ間の類似性検査方法。

【請求項 9】

ネットワーク上で他のコンピュータに対して不正処理を行う第 1 のソフトウェアの処理結果から得られる第 1 の系列データと、検査対象の第 2 のソフトウェアの処理結果から得られる第 2 の系列データとを比較してその類似性を検査する類似性検査装置であって、

該第 1 のソフトウェアの不正処理の結果を検出しその結果を第 1 の系列データとして得る不正処理結果検知手段と、

該第 2 のソフトウェアの処理結果を検出しその結果を第 2 の系列データとして得る検査対象処理結果検知手段と、

該第 1 の系列データ及び該第 2 の系列データを変換処理する系列データ変換処理手段であって、

入力された系列データについて、離散フーリエ変換して横軸に周波数、縦軸に周波数成分の強度を表すスペクトラムを得る離散フーリエ変換処理部と、

該スペクトラムについて、該縦軸において所定の閾値を超える周波数強度を持つ要素を抽出し、その出現位置の値の系列を得る出現位置系列取得部と、

10

20

30

40

50

該スペクトラムにおける最も強度の強いスペクトルの出現位置の値で、該出現位置の値の系列の値を全て除算し、正規化された出現位置の値の系列を得る出現位置値正規化処理部と

該正規化された出現位置値の系列について、標準偏差を用いた正規化処理を行う調波構造正規化処理部と

を少なくとも含むコンピュータの系列データ変換処理手段と、
変換後の第1の系列データと変換後の第2の系列データとから所定の相関関係式を用いて相関係数を算出するコンピュータの相関係数算出手段と
を少なくとも備える
ことを特徴とする系列データ間の類似性検査装置。

10

【請求項10】

前記系列データ変換処理手段が、
離散フーリエ変換処理部から出力されたスペクトラムにおいて、所定の閾値以上の高周波数成分を除去する高周波数成分除去処理部を含む
ことを特徴とする請求項9に記載の系列データ間の類似性検査装置。

【請求項11】

前記系列データ間の類似性検査装置が、
調波構造正規化処理部において正規化された第1及び第2の系列データについて、基本周波数における出現位置において両系列データを同期すると共に、該基本周波数間で出現位置値がない場合には所定の値を補完するデータ整形手段を備えた
ことを特徴とする請求項9又は10に記載の系列データ間の類似性検査装置。

20

【請求項12】

前記不正処理結果検知手段及び検査対象処理結果検知手段が、それぞれ第1及び第2のソフトウェアによる、他のコンピュータのネットワークアドレスに対する連続的なスキャンを検知する構成であって、前記第1及び第2の系列データとして、スキャンしたネットワークアドレスの値の列を用いる
ことを特徴とする請求項9ないし11のいずれかに記載の系列データ間の類似性検査装置。

【請求項13】

前記第1のソフトウェアが、閉じられたネットワークにおいて検査のために実行されるマルウェアであり、前記第2のソフトウェアが、広域ネットワークにおいて実際に実行され、マルウェアと疑われる挙動を示すソフトウェアであり、
前記請求項1ないし4に記載の系列データ間の類似性検査方法を用いて、該第2のソフトウェアの種類を、該第1のソフトウェアとの類似性を検査することにより特定する
ことを特徴とするマルウェアの検査装置。

30

【請求項14】

2つ以上の系列データを比較して系列データ間の類似性を検査する類似性検査装置であって、

該各系列データを変換処理する系列データ変換処理手段であって、
入力された系列データについて、離散フーリエ変換して横軸に周波数、縦軸に周波数成分の強度を表すスペクトラムを得る離散フーリエ変換処理部と、
該スペクトラムについて、該縦軸において所定の閾値を超える周波数強度を持つ要素を抽出し、その出現位置の値の系列を得る出現位置系列取得部と、

40

該スペクトラムにおける最も強度の強いスペクトルの出現位置の値で、該出現位置の値の系列の値を全て除算し、正規化された出現位置の値の系列を得る出現位置値正規化処理部と

該正規化された出現位置値の系列について、標準偏差を用いた正規化処理を行う調波構造正規化処理部と

を少なくとも含むコンピュータの系列データ変換処理手段と、
変換後の各系列データから所定の相関関係式を用いて相関係数を算出する相関係数算出

50

手段

とを備える

ことを特徴とする系列データ間の類似性検査装置。

【請求項 15】

前記系列データ変換処理手段が、

離散フーリエ変換処理部から出力されたスペクトラムにおいて、所定の閾値以上の高周波数成分を除去する高周波数成分除去処理部を含む

ことを特徴とする請求項 14 に記載の系列データ間の類似性検査装置。

【請求項 16】

前記系列データ間の類似性検査装置が、

調波構造正規化処理部において正規化された第 1 及び第 2 の系列データについて、基本周波数における出現位置において両系列データを同期すると共に、該基本周波数間で出現位置値がない場合には所定の値を補完するデータ整形手段を備えた

ことを特徴とする請求項 14 又は 15 に記載の系列データ間の類似性検査装置。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は複数の数値の列からなる系列データについて、2 つ以上の系列データ間の類似性を検査する方法とその装置に関し、特に該方法によりネットワーク上のスキャン特性の類似性を検査する技術に係るものである。

【背景技術】

【0002】

インターネットにおけるインシデント対策の研究分野では、広域ネットワークでのパッシブモニタリングを行い、観測されたトラフィックを分析することで、インシデント検知を行うための研究が盛んに行われている。

また、本件発明者らが推進するインシデント対策のためのプロジェクトnicter（非特許文献 1 を参照。）では、広域観測網において観測されたトラフィックから、実時間でインシデントを検知する技術が研究されている。

広域ネットワークにおいて実際のインシデントを解析する技術をここではマクロ解析と呼ぶこととする。

【0003】

その一方で、ウイルス(virus)、ワーム(worm)、ボット(bot)といったマルウェア(malware)検体を収集・分析し、個々のマルウェアの特徴を抽出する技術も研究が進められている。このように閉じられたネットワーク空間において、マルウェア検体の分析を行うことを、上記のマクロ解析に対して、ミクロ解析と呼ぶこととする。

【0004】

マルウェアに起因するインシデントに迅速かつ的確に対処するためには、広域観測網において検出された事象(結果)に対し、その原因となったマルウェアを特定し、提示することが重要である。

このようなインシデント(結果)とマルウェア(原因)との相関関係を得るためには、それぞれの特徴を効果的に抽出した上で相関分析を行う必要がある。

【0005】

ミクロ解析においてスキャン攻撃の特徴抽出手法としていくつかの先行研究が提案されているが、広域ネットワークでのインシデントとマルウェアとの相関分析を行うことを前提とする、個々のホストのネットワーク的挙動を分析する研究はいまだ少ない。すなわち、マクロ解析結果とミクロ解析結果との相関関係を検査して、マクロ解析において得られた特定のホストについてマルウェアの特定を行う技術はほとんど提供されていない。

【0006】

ところで、ネットワークインシデントの研究分野では、スペクトラム解析アルゴリズムや時系列解析アルゴリズムといったアルゴリズムを用いた、さまざまなトラフィック解析

10

20

30

40

50

手法が提案されている。

非特許文献 2 に開示される研究では定点観測網から得られるパケット数の変動に着目した解析を行っている。これは、送信元および送信先の IP アドレスとポート番号といったパラメータ毎のパケット数の変動データに対してウェーブレット解析を施し、そこで得られる時間周波数成分の変化に基づいて脅威を検知する手法である。

【 0 0 0 7 】

また、非特許文献 3 に開示される研究では、非特許文献 2 の技術と同様、パケット数の変動に着目した解析を行っている。ここでは系列データ(単位時間あたりのパケット数)に対してSDAR (Sequential Discounting AR estimating) と呼ばれる時系列解析アルゴリズムを用いることで軽快な処理を実現し、リアルタイムでの異常検知を可能としている。

10

【 0 0 0 8 】

以上の 2 つの提案手法はその目的がインシデント検知であるため、上述したようなマルウェアの特徴抽出に適しているとは言えない。

これらに対して、非特許文献 4 に開示される研究はフーリエ変換を用いたマルウェアの特徴抽出である。該文献では、フーリエ変換によって得られたスペクトラムの調波構造に着目し、マルウェアの識別を行っている。

しかし、解析対象となるデータは、上の二例と同じくパケット数の変動データを前提としているため、宛先 IP アドレス等のパラメータの遷移情報を検査対象とすることができない。

【 0 0 0 9 】

20

【非特許文献 1】中尾康二、吉岡克成、衛藤将史、井上大介、力武健次著「nicter: An Incident Analysis System using Correlation between Network Monitoring and Malware Analysis」Proceedings of The 1st Joint Workshop on Information Security, JWIS2006, Page363-377, 2006年9月

【非特許文献 2】石黒正揮、鈴木裕信、村瀬一郎著「ウェーブレット解析を用いた周波数成分変化に基づくインターネット脅威検出法」電子情報通信学会(2006年暗号と情報セキュリティシンポジウム(SCIS2006)) 2006年1月

【非特許文献 3】竹内純一、佐藤靖士、力武健次、中尾康二著「変化点検出エンジンを利用したインシデント検知システムの構築」電子情報通信学会(2006年暗号と情報セキュリティシンポジウム(SCIS2006)) 2006年1月

30

【非特許文献 4】John Heidemann,Urbashi,Mitra,Antonio Ortega,Christos Papadopoulos 著「Detecting and identifying malware: A new signal processing goal」IEEE Signal Processing Magazine, Volume 23, Issue 5, pp.107-111 2006年9月

【発明の開示】

【発明が解決しようとする課題】

【 0 0 1 0 】

上記従来技術では、インシデントのマクロ解析結果と、マルウェアのミクロ解析結果とを効果的に融合させて当該インシデントの詳細な情報を特定することができない。また、非特許文献 4 の技術によっても、パケット数の変動データのみを検査の対象としており、これはネットワークの混雑状況によるパケット数の変動の影響を大きく受けやすく、実際のマルウェアの挙動を正確に把握するためには極めて不十分である。

40

【 0 0 1 1 】

本発明はこのような従来技術の有する問題点に鑑みて創出されたものであり、系列データ間の類似性を高精度に検査する手法を創出し、それによって広域ネットワークにおけるインシデントの解析結果と、各マルウェアの特性とを効率よく比較し、両者の相関を得ることを可能にすることを目的とするものである。

同時に、同様の特徴を有する系列データの汎用的な類似性検査方法を提供することも目的とする。

【課題を解決するための手段】

【 0 0 1 2 】

50

本発明は、上記の課題を解決するために、次のような系列データ間の類似性検査方法を提供する。

すなわち、請求項 1 に記載の発明は、ネットワーク上で他のコンピュータに対して不正処理を行う第 1 のソフトウェアの処理結果から得られる第 1 の系列データと、検査対象の第 2 のソフトウェアの処理結果から得られる第 2 の系列データとを比較してその類似性を検査する類似性検査方法であって、次の各ステップを有する。

【 0 0 1 3 】

(1)不正処理結果検知手段が、該第 1 のソフトウェアの不正処理の結果を検出しその結果を第 1 の系列データとして得る不正処理結果検知ステップ、

(2)検査対象処理結果検知手段が、該第 2 のソフトウェアの処理結果を検出しその結果を第 2 の系列データとして得る検査対象処理結果検知ステップ、

(3)コンピュータの系列データ変換処理手段が、該第 1 の系列データ及び該第 2 の系列データを次の各工程：

(3-1)コンピュータの演算手段が、入力された系列データについて、離散フーリエ変換して横軸に周波数、縦軸に周波数成分の強度を表すスペクトラムを得る離散フーリエ変換処理工程、

(3-2)コンピュータのデータ抽出手段が、該スペクトラムについて、該縦軸において所定の閾値を超える周波数強度を持つ要素を抽出し、その出現位置の値の系列を得る出現位置系列取得工程、

(3-3)コンピュータの出現位置値正規化処理手段が、該スペクトラムにおける最も強度の強いスペクトルの出現位置の値で、該系列の全ての出現位置の値を除算し、正規化された出現位置の値の系列を得る出現位置値正規化処理工程、

(3-4)コンピュータの調波構造正規化処理手段が、該正規化された出現位置値の系列について、標準偏差を用いた正規化処理を行う調波構造正規化処理工程により変換する系列データ変換処理ステップ、

(4)コンピュータの相関係数算出手段が、変換後の第 1 の系列データと変換後の第 2 の系列データとから所定の相関係数式を用いて相関係数を算出する相関係数算出ステップを有することを特徴とする。

【 0 0 1 4 】

請求項 3 に記載の発明は、上記の (3) 系列データ変換処理ステップの後に、

(3')コンピュータのデータ整形手段が、変換後の第 1 及び第 2 の系列データについて、基本周波数における出現位置において両系列データを同期すると共に、該基本周波数間で出現位置値がない場合には所定の値を補完するデータ整形ステップを有することを特徴とする。

【 0 0 1 5 】

請求項 4 に記載の発明は、上記の不正処理結果検知手段及び検査対象処理結果検知手段が、それぞれ第 1 及び第 2 のソフトウェアによる、他のコンピュータのネットワークアドレスに対する連続的なスキャンを検知する構成であって、前記第 1 及び第 2 の系列データとして、スキャンしたネットワークアドレスの値の列を用いることを特徴とする。

【 0 0 1 6 】

また本発明は、次のようなマルウェアの検査方法として提供することもできる。

すなわち、請求項 5 に記載の発明は、上記の第 1 のソフトウェアが、閉じられたネットワークにおいて検査のために実行されるマルウェアであり、第 2 のソフトウェアが、広域ネットワークにおいて実際に実行され、マルウェアと疑われる挙動を示すソフトウェアであり、請求項 1 ないし 4 のいずれかに記載の系列データ間の類似性検査方法を用いて、該第 2 のソフトウェアの種類を、該第 1 のソフトウェアとの類似性を検査することにより特定することを特徴とする。

【 0 0 1 7 】

さらに本発明は、用途を限定されない 2 つ以上の系列データを比較して系列データ間の類似性を検査する類似性検査方法として提供することもできる。

この場合において、

(A)コンピュータの系列データ変換処理手段が、該各系列データを次の各工程：

(A-1)コンピュータの演算手段が、入力された系列データについて、離散フーリエ変換して横軸に周波数、縦軸に周波数成分の強度を表すスペクトラムを得る離散フーリエ変換処理工程、

(A-2)コンピュータのデータ抽出手段が、該スペクトラムについて、該縦軸において所定の閾値を超える周波数強度を持つ要素を抽出し、その出現位置の値の系列を得る出現位置系列取得工程、

(A-3)コンピュータの出現位置値正規化処理手段が、該スペクトラムにおける最も強度の強いスペクトルの出現位置の値で、該出現位置の値の系列の値を全て除算し、正規化された出現位置の値の系列を得る出現位置値正規化処理工程、

(A-4)コンピュータの調波構造正規化処理手段が、該正規化された出現位置値の系列について、標準偏差を用いた正規化処理を行う調波構造正規化処理工程により変換する系列データ変換処理ステップ、

(B)コンピュータの相関係数算出手段が、変換後の各系列データから所定の相関関係式を用いて相関係数を算出する相関係数算出ステップを有することを特徴とする。

【0018】

請求項7に記載の発明は、上記の(A)系列データ変換処理ステップにおいて、(A-1)離散フーリエ変換処理工程の後に、

(A-1')コンピュータのデータ抽出手段が、所定の閾値以上の高周波数成分を除去する高周波数成分除去処理工程

を含むことを特徴とする。

【0019】

請求項8に記載の発明は、上記の(A)系列データ変換処理ステップの後に、

(A')コンピュータのデータ整形手段が、変換後の各系列データについて、基本周波数における出現位置において各系列データを同期すると共に、該基本周波数間で出現位置値がない場合には所定の値を補完するデータ整形ステップを有する

ことを特徴とする。

【0020】

本発明は、上記請求項1ないし4の各処理を実行する系列データ間の類似性検査装置として提供してもよい。

【0021】

また、請求項5の各処理を実行するマルウェアの検査装置として提供してもよい。

【0022】

さらに、上記請求項6ないし8のいずれかの各処理を実行するより汎用的な系列データ間の類似性検査装置として提供してもよい。

【発明の効果】

【0023】

本発明は、上記構成を備えることにより次のような効果を奏する。

すなわち、本発明によれば、複数の数値の列からなる系列データにおいて、その類似性を高精度に検査する検査方法及び装置を提供することができる。

特に、本発明は、系列データの遷移に着目するものであり、各系列データ間の値が異なる値域にいたり、系列データの要素の数が異なったり、系列中で多少の入れ違いが生じていても、正規化処理、整形処理によって良好に類似性を検査することができる。

【0024】

マルウェアの挙動のうち、IPアドレスなどのネットワークアドレスを連続してスキャンする構成が知られているが、本発明における系列データとしてスキャンされたIPアドレスを用いることで、2つのマルウェアの類似性検査にも用いることができる。

特に、

10

20

30

40

50

【 0 0 2 5 】

また、ハニーポッドなど閉じられたネットワーク空間におけるマルウェアの検体における挙動と、広域ネットワークで生じているインシデントにおける挙動とを比較することで、マルウェアの特定のための検査方法として用いることもできる。

【 発明を実施するための最良の形態 】

【 0 0 2 6 】

以下、本発明の実施形態を、図面に示す実施例を基に説明する。なお、実施形態は下記に限定されるものではない。

図 1 は本発明に係る系列データの類似性検査装置（以下、本装置と呼ぶ。）（ 1 ）の全体構成図である。本装置（ 1 ）は、公知のパーソナルコンピュータやネットワークサーバによって構成するのが簡便である。

10

【 0 0 2 7 】

本装置（ 1 ）には、演算処理等を司る CPU（ 1 0 ）を中心として、CPU（ 1 ）と協働するメモリ（ 1 1 ）、ユーザが入力等を行うキーボード及びマウス（ 1 2 ）、データを読み書き自在に格納するハードディスク（ 1 3 ）、インターネット等のネットワーク接続を行うネットワークアダプタ（ 1 4 ）などが備えられている。また、図示しないモニタを接続して画面表示を行ったり、スピーカを接続して音声出力を行うことも可能である。

これらの構成はいずれも周知の事項であって、その構造や作用については説明を省略する。

【 0 0 2 8 】

20

本発明はこのようなコンピュータを用いて、2つ以上の系列データが類似しているか否か、あるいは類似度を検査する処理方法と、該方法を実装した装置を提供するものである。以下では、2つ以上の系列データとして、ネットワークにおいて2種のソフトウェアがそれぞれ複数のIPアドレスを順にスキャンしていく際の該アドレスを時系列で並べた数値列を用いて説明する。

【 0 0 2 9 】

このようなスキャンは、大規模なネットワーク障害を起こすために大量のパケットを大量のIPアドレスに向けて送出するマルウェアや、脆弱なサーバを探索する際にみられる挙動であり、本実施例ではそのようなマルウェアの挙動を比較することを目的としている。このような処理の意義については後記で詳述する。

30

【 0 0 3 0 】

本装置（ 1 ）の CPU（ 1 0 ）には、順に不正処理結果検知部（ 2 0 ）、検査対象処理結果検知部（ 2 1 ）、系列データ変換処理部（ 2 2 ）、データ整形部（ 2 3 ）、相関関数算出部（ 2 4 ）、出力部（ 2 5 ）を備えている。

このうち、データ整形部（ 2 3 ）については、入力される系列データにより、必ずしも備えなくてもよいが、本実施例のようにIPアドレスなど、異なる値域の系列データを入力する際には必要である。

【 0 0 3 1 】

本発明の中核となるのは系列データ変換処理部（ 2 2 ）と、相関係数算出部（ 2 3 ）である。まず系列データ変換処理部（ 2 2 ）において入力された系列データを離散フーリエ変換することに特徴がある。そして、単にフーリエ変換するのみならず、これを最適な方法によって正規化処理し、相関係数算出部（ 2 3 ）で相関関数を得ることを可能にしている。

40

【 0 0 3 2 】

このために、図 2 に示すように、系列データ変換処理部（ 2 2 ）にはさらに、離散フーリエ変換処理部（ 2 2 0 ）、高周波数成分除去処理部（ 2 2 1 ）、出現位置系列取得部（ 2 2 2 ）、出現位置値正規化処理部（ 2 2 3 ）、調波構造正規化処理部（ 2 2 4 ）を備えている。

このうち、高周波数成分除去処理部（ 2 2 1 ）については、同処理を行うことが好ましいが、入力される系列データによっては必ずしも備えなくてもよい。

50

【 0 0 3 3 】

以上の構成を備えた本装置（ 1 ）によって、図 3 に示す処理フローチャートによって系列データ間の類似性を検査する。

（不正処理結果検知処理： S 1 0 ）

まず、不正処理結果検知部（ 2 0 ）が、第 1 のソフトウェアによるネットワーク上での IP アドレスのスキャンを検知する。該不正処理結果検知部（ 2 0 ）の動作としては、例えば実験用に閉じられたネットワーク空間において、仮想的に複数のコンピュータからなるネットワークを設け、検体として収集してあるマルウェアを実験的に実行処理させてみる。そして、その際のマルウェアの挙動のうち、ネットワーク内でパケットを送信する宛先 IP アドレスの遷移を抽出する。

本処理により、既知のマルウェアがパケットを送信する際の宛先 IP アドレスの系列データを得て、ハードディスク（ 1 3 ）に格納する。

【 0 0 3 4 】

（離散フーリエ変換処理： S 1 1 ）

このようにして得られた宛先 IP アドレスを時系列でグラフに表すと、図 4 の（ A ）のようになる。グラフに示されるように、周期的に小さなアドレスから大きなアドレスまで順にスキャンしていく様子が分かる。同グラフにおいて Y 軸は IP アドレスの値を表し、入力される系列データからは時間成分を取り除いているため、X 軸は時間ではなく単にパケットの到着順を表している。

【 0 0 3 5 】

一般にマルウェアがスキャンを行う際には標的とするネットワークに対して、一定の方法で宛先 IP アドレスを変動させながらパケットを送信する。その変動パターンはマルウェアが持つスキャンエンジン毎に大きく異なり、アドレス値を 1 つずつ単純増加させるものや、任意のタイミングでアドレス値を大きくずらすもの、あるいはランダムにアドレス値を決定するものなどがある。

【 0 0 3 6 】

このような特徴を捉えるためのアルゴリズムとして本発明ではスペクトラム解析を用いることを提案し、宛先 IP アドレスの遷移を信号波形として捉えてフーリエ変換を施すこととした。

抽出された周波数成分を用いて、他のスキャンとの類似性を評価する。

【 0 0 3 7 】

ここで、離散フーリエ変換とは離散群上のフーリエ変換であり、コンピュータによって高速に計算できることが周知である。離散フーリエ変換をコンピュータ上で行う方法は、高速フーリエ変換（ FFT ）としてさまざまなアルゴリズムが提案されているが、最も基本的なものは、Cooley-Tukey 型 FFT アルゴリズムと呼ばれ、非特許文献 5 に開示されるものが知られている。

【 0 0 3 8 】

【非特許文献 5】 J.W.Cooley and J.W.Tukey: Math. of Comput. 19 (1965) 297.

【 0 0 3 9 】

離散フーリエ変換処理部（ 2 2 0 ）ではこのような周知のアルゴリズムを任意に用いて、図 4（ A ）のような入力された系列データを周波数成分に分解する。これによって得られたスペクトラムが、図 4 の（ B ）に示されるグラフである。

該スペクトラムでは、X 軸が周波数を、Y 軸が周波数成分の強度を表していることになる。（なお入力する時系列が時間ではなく到着順であるため、厳密な意味での周波数とは異なるが、本発明においては影響しないため、以下でもこの表現により説明する。）

【 0 0 4 0 】

本方法は次のような利点がある。

まず、フーリエ変換は直流成分を無視することで一連の系列データの中での相対的なアドレス値の変動を捉えることができる。すなわち、スキャン対象となるアドレス帯の大小にかかわらず、元の信号波形同士に類似性が見られるならば、それを検出することが可能

10

20

30

40

50

である。

【 0 0 4 1 】

また、一般にフーリエ変換によって得られたスペクトラムから強度の高い成分のみを抽出し、それらの成分に対して逆フーリエ変換を施した場合に、元の信号を高い水準で復元できることが知られている。(図4の(C)を参照)。

この特性を利用して宛先IPアドレスの系列データから、アドレス遷移を特徴づける支配的な要素を一定の数だけ抜き出して使用することができる。これにより、攻撃元ホストから到達したパケット数の大小に関わらず、一定の要素数を用いた類似性の検証を行うことが可能となる。

【 0 0 4 2 】

さらに、フーリエ変換によって得られるスペクトラムでは、パケット到達順序の入れ違いといった軽微な特徴は高周波数帯域に表れる。よってフーリエ変換を行った後に一定の高周波数帯域の要素を除去することで、ネットワーク状態の悪化によるパケット到達順序の入れ違いやパケットロスの影響を吸収することが可能となる。

【 0 0 4 3 】

(高周波数成分除去処理：S 1 2)

このようなフーリエ変換の利点を利用して、図5に示すように、高周波数成分除去処理部(2 2 1)では所定の閾値Aにより、それより高い周波数成分を除去する。すなわち図5のグラフにおける右側の信号は利用しない。

上記した通り、パケット到達順序の入れ違いやパケットロスといった軽微な特徴は高周波数帯に表れる。よって本実施例では、ネットワーク状況によってもたらされるスキャンパターンへの影響を抑えるため、スペクトラム中の高周波数帯域の除去を行っている。

【 0 0 4 4 】

(出現位置系列取得処理：S 1 3)

次に、出現位置系列取得部(2 2 2)において、高レベルスペクトルの閾値B(図5)により、所定の閾値を超える周波数強度を持つ要素のみを抽出する。これにより比較対象とする要素数を削減することができる。

【 0 0 4 5 】

そして、これより先の処理においては、周波数強度(Y軸)ではなく、高周波数成分除去処理(S 1 2)と出現位置系列取得処理(S 1 3)で選択されたスペクトルの出現位置(X軸)(以下、この値をインデックス値と呼ぶ。)の系列(I)を用いて相関係数の導出処理を行う。

【 0 0 4 6 】

本処理(S 1 3)により、図4に示すBのスペクトラムから、支配的なインデックス値を取得することができる。例えば、図示するように

{1,2,4,9,10,13,15,18,・・・}

のようなインデックス値の系列が得られる。

【 0 0 4 7 】

(出現位置値正規化処理：S 1 4)

調波構造の抽出本来は同一のスキャンパターンであっても、観測点のネットワーク条件の違いにより採取されるパケット数が大きく異なる場合がある。例えばホスト(A)からのスキャンが3周期分の変動をしたのに対して、ホスト(B)からのスキャンは1周期分しか採取されなかった場合が考えられる。

また、ホスト(A)からのパケットの全てが観測地点に到達するのに対し、ホスト(B)からのパケットは2つに1つしか到達しなかった場合にはホスト(B)の周期はホスト(A)の2分の1となる。

【 0 0 4 8 】

このような条件の違いを補うため、調波構造を維持したまま基本周波数を取り除く必要がある。これは言い換えれば、スペクトラムにおけるX軸のスケールをそれぞれのサンプル数に合わせて正規化する処理であると言える。

10

20

30

40

50

この処理は、上記処理で得られたスペクトラムのうちもっとも強度の高いスペクトルのインデックス値(l_p)で全てのスペクトルのインデックス値(l_i)を除算し、正規化された個々のインデックス値(N_i)を得ることで実現する。

【0049】

すなわち、出現位置値正規化処理部(223)では、次式(数1)によりインデックス値を正規化する。

(数1)

$$N_i = l_i / l_p$$

10

以降の処理では、この正規化されたインデックス値の系列Nを用いる。

【0050】

(調波構造正規化処理：S15)

インデックス値の系列Nでは、最初の段階でモニタリングされたスキャンパッケージの数によって、インデックス値の取り得る値が大きく異なっている。これにより一つのインデックス値が持つ重みも異なってしまうため、そのまま相関係数を求めた場合には不正確な結果が算出される可能性がある。

そこで調波構造正規化処理部(224)では以下のように、ホスト毎に異なるインデックス値の重みを標準偏差を用いて正規化する。

【0051】

20

まず、 n 個の要素を持つ系列Nが与えられたとき、その平均値を M とすると、標準偏差 SD_N は次式(数2)によって得られる。

【0052】

【数2】

$$SD_N = \sqrt{\frac{\sum_{i=1}^n (N_i - M)^2}{n}}$$

30

【0053】

そしてこの標準偏差 SD_N をもとに、各要素の基準値 S_i は以下の式(数3)によって求められる。

【0054】

【数3】

$$S_i = \frac{N_i - M}{SD_N}$$

40

以降の処理は、この正規化されたインデックス値の系列Sを用いて行う。

【0055】

以上、離散フーリエ変換処理(S11)ないし、調波構造正規化処理(S15)までが本発明に係る系列データ変換処理の詳細な内容である。

次に、本発明では同様の処理を広域ネットワーク上におけるインシデントの解析結果に対して用いる。

【0056】

50

(検査対象処理結果検知処理 : S 2 0)

すなわち、CPU (1 0) の検査対象処理結果検知部 (2 1) が、例えばダークネット (darknet) と呼ばれる、実際には使用されていない IP アドレス領域に対して送信されるパケットをネットワーク上で検知し、その宛先 IP アドレスの遷移を抽出する。

【 0 0 5 7 】

このような IP アドレスに向けたパケットは規則に準じたホストに向けたものではないから、設定ミスか、ワームによるスキャン、探索、後方散乱メールなどの悪意による処理と考えられる。このような不正処理は、送信元 IP アドレスが偽られている場合も多い。

抽出された宛先 IP アドレスの系列データはハードディスク (1 3) に格納される。

【 0 0 5 8 】

そして、この系列データに対して、離散フーリエ変換処理 (S 2 1) 、高周波数成分除去処理 (S 2 2) 、出現位置系列取得処理 (S 2 3) 、出現位置値正規化処理 (S 2 4) 、調波構造正規化処理 (S 2 5) を順次行う。該処理内容は、上記と全く同様であるので、説明を省略する。

【 0 0 5 9 】

(データ整形処理 : S 1 6 , S 2 6)

これまでの一連の手続きにより、個々の系列に対して要素数の削減やスケール合わせのための正規化処理が済んだ。これにより初めて他のデータとの比較を行えるようになったが、実際に相関分析を行う前に、比較対象である 2 つの系列の同期と系列長を整える必要がある。データ整形部 (2 3) では以下の処理を行う。

【 0 0 6 0 】

ここではまず、2 つの系列の同期を基本周波数 (もっとも強度の大きい周波数成分) のインデックスを軸として揃える。さらに、系列長の違いを埋めるため、ずれた要素に対して Zero-Padding 処理、すなわち各要素に数値 0 を代入する処理を行う。

以上の手続きによって、2 つの系列の同期と長さが整い、適正な相関処理が行えるようになる。図 3 では各系列データに対してデータ整形処理を行っている場合を図示しているが、本処理はどちらか一方を他方の系列データに揃える処理でもよい。

【 0 0 6 1 】

(相関関数算出処理 : S 3 0)

最後に、相関係数算出部 (2 4) の演算処理によって、正規化された 2 つの系列 S と S の相関係数 C を以下の式 (数 4) を用いて求める。

【 0 0 6 2 】

【 数 4 】

$$C_{\alpha\beta} = \frac{1}{n} \sum_{i=1}^n S_{\alpha i} \times S_{\beta i}$$

【 0 0 6 3 】

最終的に導出される相関係数は、-1 から 1 の間の値をとり、相関性の高い 2 つの系列ほど相関係数は 1 に近づき、相関性の低い系列の相関係数は -1 に近づくという特徴を持つ。

なお、ここで用いている相関関数は周知の相関関数を任意に用いることができ、上記はその一例である。

【 0 0 6 4 】

(出力処理 : S 3 1)

本装置 (1) は出力部 (2 5) から、該相関係数を出力することにより、最初に入力した 2 つの系列データ間の類似度を出力することができる。出力の態様としては、ネットワークアダプタ (1 4) から他のコンピュータに結果を送信してもよいし、モニタから出力したり、ハードディスク (1 3) に格納してもよい。

10

20

30

40

50

また、複数のマルウェアとの類似度を検査して、その一覧表をレポートとして出力してもよい。

相関係数のように実数で出力せず、所定の閾値を用いて、「相関がある」「相関がない」の2値で出力してもよい。

【0065】

本実施例の構成は以上の通りであるが、マルウェアの特徴はスキャンパケットの宛先IPアドレスだけでなく、攻撃元および攻撃先のポート番号やパケット送出タイミングなどにも表れると考えられる。

よって不正処理結果検知部(20)や検査対象処理結果検知部(21)でこれらの系列データを抽出して適用することで、より多面的なマルウェアの識別が行うこともできる。これらの抽出方法は、公知の技術を適宜用いることができる。

【0066】

本発明では、相関分析を行うことを前提として個々のホストのネットワーク的挙動を分析する技術を提案した。この方法によって従来技術の問題であった次の諸点につき解決した。

【0067】

(a) 宛先IPアドレス帯の位置に依存しない

観測地点に割り当てられるIPアドレス帯は適度に散らばっている。複数のセンサにおいて同一ホストからのスキャンパケットが観測されることが保証されないため、宛先IPアドレス帯の位置に依存しない手法を実現した。

【0068】

(b) サンプル数が異なるデータ同士を比較できる

観測地点に割り当てられたIPアドレスの個数は一定ではなく、サブネット長が/24のものから/16や/8のものまでさまざまである。観測アドレス数が異なると、単一のホストから採取できるパケット数も大きく変動する。本発明ではパケット数が異なっても比較を可能にした。

【0069】

(c) パケットロス・パケット到達順序の入れ違いを吸収できる

攻撃元ホストとの間のネットワーク状態の悪化により、パケットロスが発生したり、パケットの到達順序が頻繁に入れ替わることが知られている。本発明は、これらの軽微な特徴を吸収した上で、相関分析を可能にした。

【0070】

(別実施例)

本発明は、上記ネットワークのインシデントに係る系列データにとどまらず、任意の系列データに対して適用することが可能であり、特に、系列データの値域が異なるもの、系列データの要素数が異なるもの、系列の要素に多少の入れ替わりが生じるもの、などの系列データに適用すると好適である。

【0071】

(マルウェア特定システムへの適用)

本件出願人らにより、図6に示すシステムが提案されている。

同図において、まず広域ネットワーク(60)に複数設けたセンサー(61)で上記したダークネットに対するパケットなどを検知し、マクロ解析器(62)に入力する。マクロ解析の結果はデータベース(63)に格納される。

【0072】

一方、ネットワーク(64)上で、キャプチャ(65)によって多数のマルウェア検体を採集し、ミクロ解析器(66)によりその静的、動的な性質を解析する。その解析結果もデータベース(67)に格納する。

【0073】

このように、実際にインシデントを発生させているマルウェアをマクロ解析器によってマクロ的に解析すると共に、検体を解析してマルウェアのミクロ的な解析を行い、それぞ

10

20

30

40

50

れのデータベースから相関分析器(68)で相関分析を行うことが考えられている。

【0074】

相関分析の結果はデータベース(69)に格納されて、さまざまな出力方法によるインシデントハンドリングシステム(70)を介してユーザ(71)に通知されたり、レポート(72)として出力されたりする。

【0075】

このシステムに対して、本発明を適用し、マクロ解析器(62)に検査対象処理結果検知部(21)を、ミクロ解析器(66)に不正処理結果検知部(20)を備えて、それぞれの挙動を検出すると共に、その結果を系列データ変換処理部(22)、データ整形部(23)、相関係数算出部(24)を備えた相関分析器(68)において相関分析してもよい。

10

【0076】

従来、マクロ解析とミクロ解析の結果を融合することが技術的に困難であったが、本発明の方法を適用することによって、これが実現され、広域ネットワークで生じているインシデントの原因を高速、的確に特定することができる。

【0077】

(実験例)

本発明方法の評価実験を示す。ここでは、(1)同一の系列同士を比較した際に最大の相関係数が得られること。(2)外形が近いスキャンパターンを持つ系列同士を比較した場合にも高い相関係数が得られること。(3)サンプル数が異なる場合でも相関係数を導出することができること。(4)対象とするアドレス帯が異なる場合でも相関係数を導出することができること。(5)全く異なるスキャンパターンの場合の5項目について検証を行う。

20

【0078】

(1) 同一の系列同士の比較(図7)

まず始めに、あるホストからのスキャンパターンと全く同一のデータを用意し、これら2つの系列を本装置(1)に入力することで相関係数の導出を行った。結果は図7に示すとおり、相関係数が1.00となり、期待通りに最大の値を得ることが出来た。なお、図中では2つのホストからのスキャンパターンを表しているが、重なっているため1本の線に見える。

また、図は上から(A)IPアドレスの遷移、(B)スペクトラム、(C)相関係数を示している。以下も同様である。

30

【0079】

(2) 外形が近いスキャンパターンを持つ系列同士の比較(図8)

次に、スキャンパターンが外形的に似ていると判断できる2つの系列を用意し、これらを本装置(1)に入力し、相関係数を求めた。結果は図8のとおり、相関係数は0.98となり、外形が近いスキャンパターン同士の類似性の高さを確認することが出来た。

【0080】

(3) サンプル数が異なる系列同士の比較(図9)

ケース(1)で用いた2つの系列データの一つのスキャンパケットを1/4周期にした上で、これらのデータに対して相関分析を行った。このような系列同士の比較でも高い相関性が得られることが期待される。結果は図9に示すとおり、相関係数は0.87となり、このケースにおいても十分な効果を得ることが出来た。

40

【0081】

(4) アドレス帯が異なるスキャン同士の比較(図10)

スキャン対象となるアドレス帯が異なりながらも、アドレス値の遷移が類似している2つの系列データを用意し、これらを用いて相関分析を行った。この場合においても、高い相関係数が得られることが期待される。結果は図10に示すとおり、相関係数が0.96となり、一定の相関性の高さを示すことが出来た。

【0082】

(5) 外形が全く異なるスキャンパターンを持つ系列同士の比較(図11)

50

最後に、異なる２種類のマルウェアによるスキャンパターンをもつ系列データを入力したときの結果を示す。この場合には当然に低い値が出力されなければならない。

実験の結果、このときの相関係数は 0.08 となり、顕著に低い値を得ることができ、本発明の効果が確認された。

【図面の簡単な説明】

【0083】

【図1】本発明に係る類似性検査装置の構成図である。

【図2】本発明に係る系列データ変換処理部の構成図である。

【図3】本発明に係る類似性検査方法の処理フローチャートである。

【図4】本発明に係る離散フーリエ変換の説明図である。

10

【図5】スペクトラムから閾値を用いてデータを抽出する処理の説明図である。

【図6】マルウェア特定システムの構成図である。

【図7】本発明に係る実験例(1)における実験データである。

【図8】本発明に係る実験例(2)における実験データである。

【図9】本発明に係る実験例(3)における実験データである。

【図10】本発明に係る実験例(4)における実験データである。

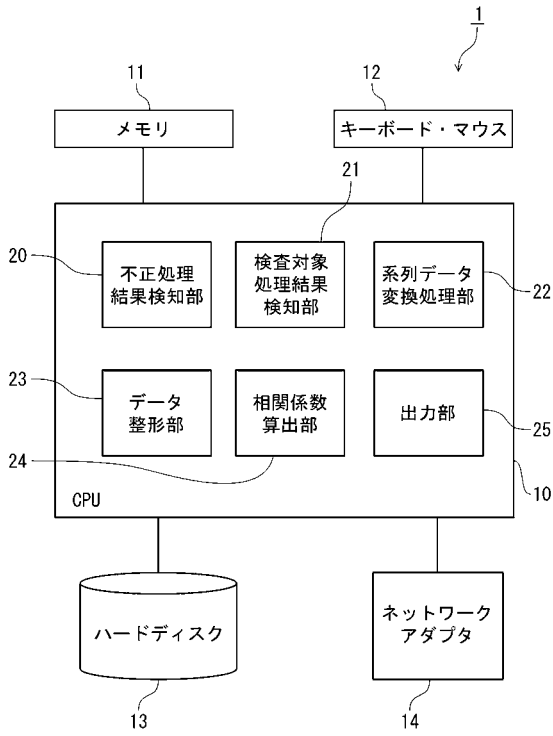
【図11】本発明に係る実験例(5)における実験データである。

【符号の説明】

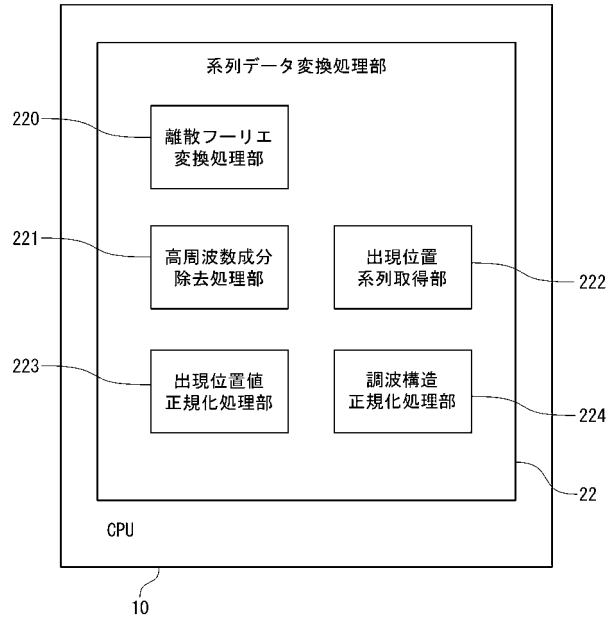
【0084】

1	類似性検査装置	20
10	CPU	
11	メモリ	
12	キーボード・マウス	
13	ハードディスク	
14	ネットワークアダプタ	
20	不正処理結果検知部	
21	検査対象処理結果検知部	
22	系列データ変換処理部	
23	データ整形部	
24	相関係数算出部	30
25	出力部	

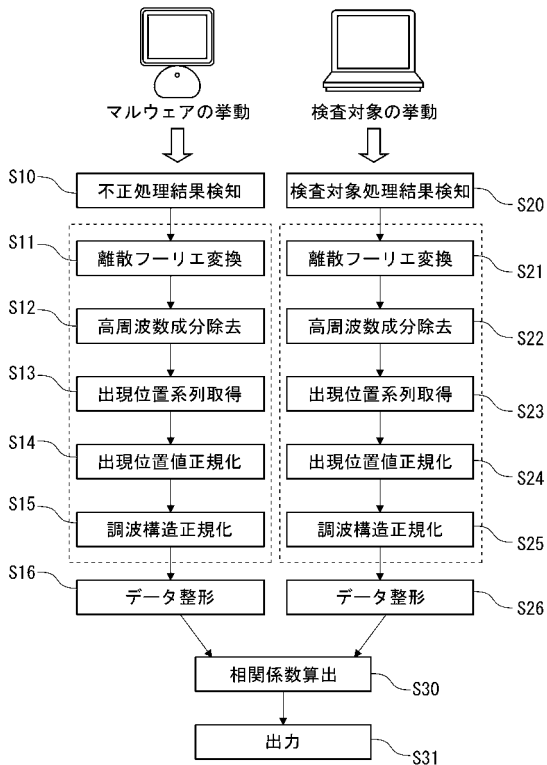
【 図 1 】



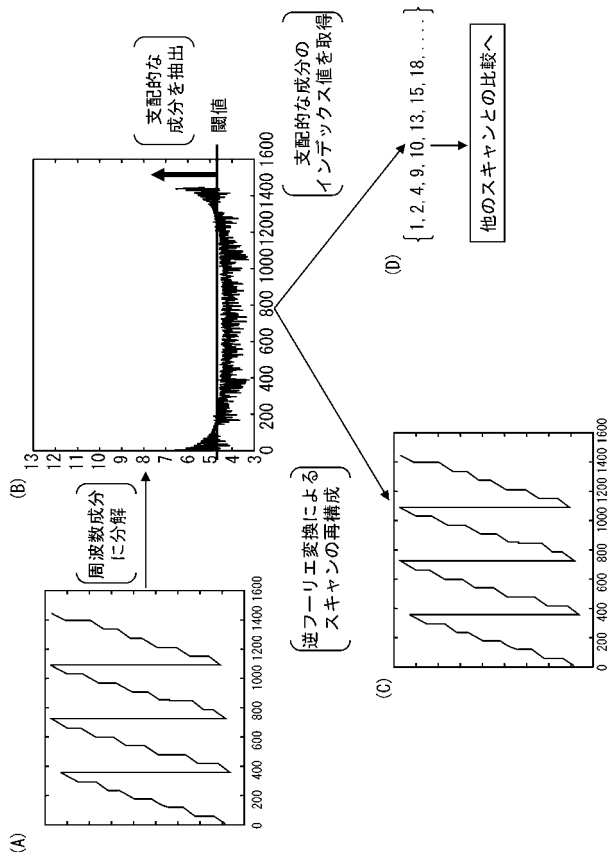
【 図 2 】



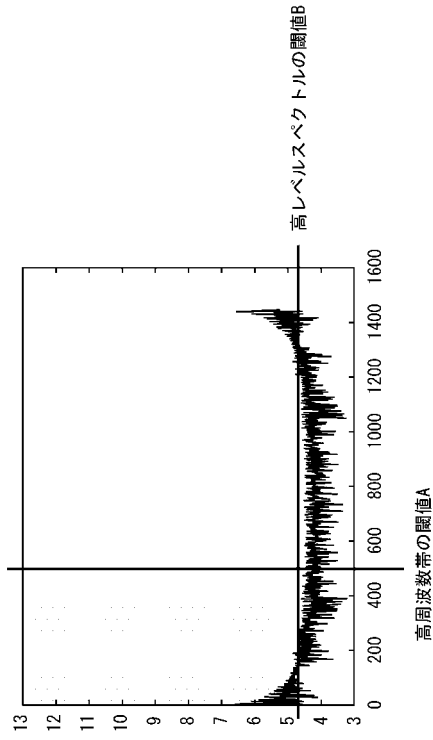
【 図 3 】



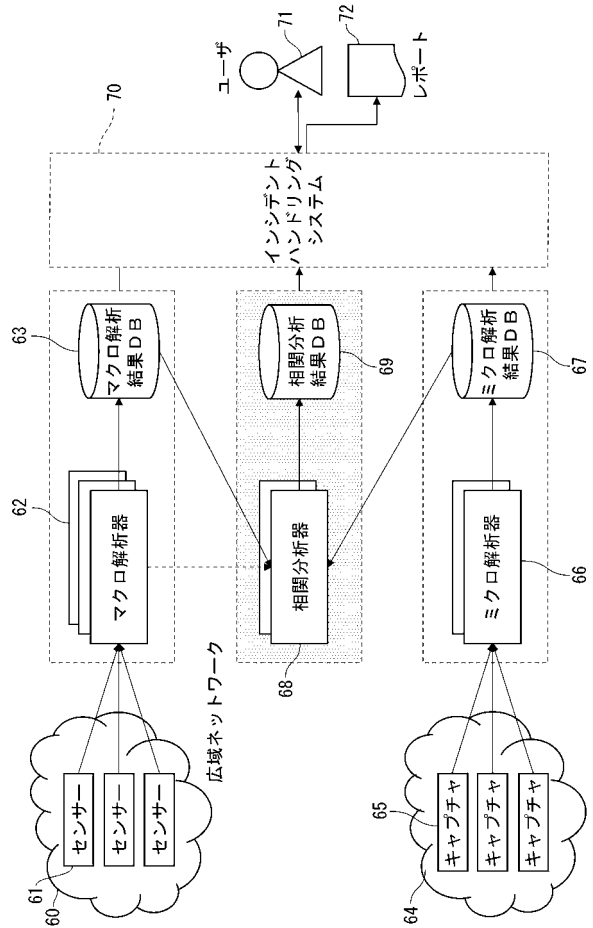
【 図 4 】



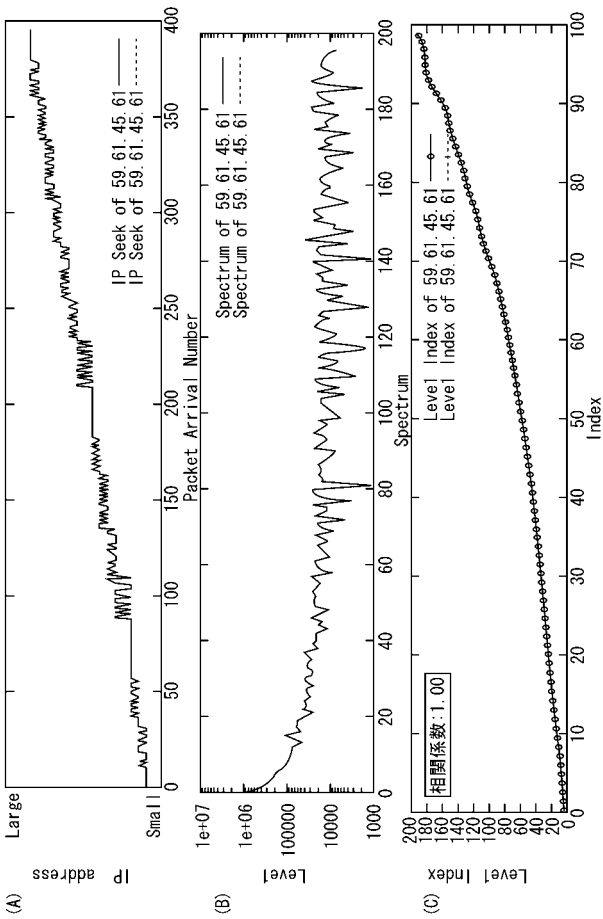
【 図 5 】



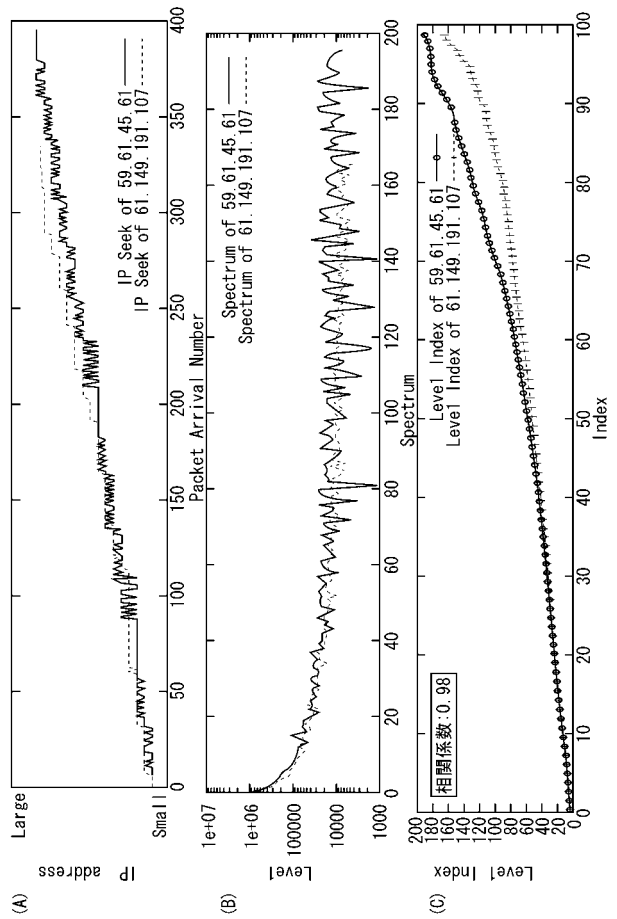
【 図 6 】



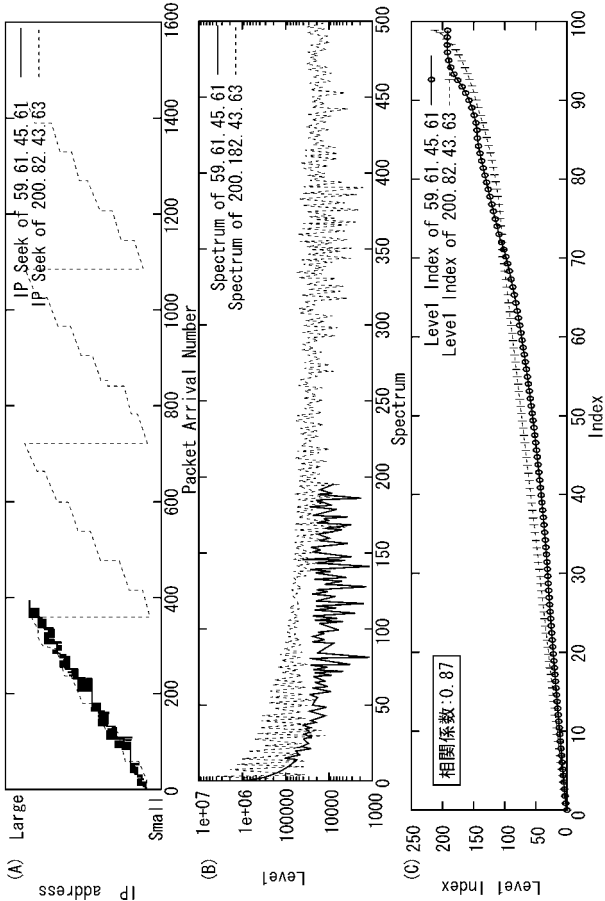
【 図 7 】



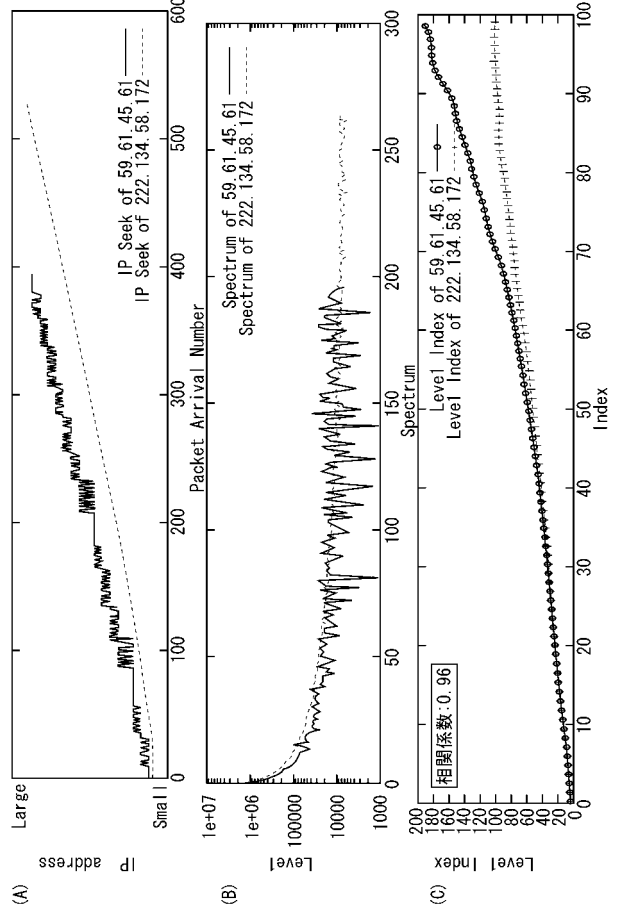
【 図 8 】



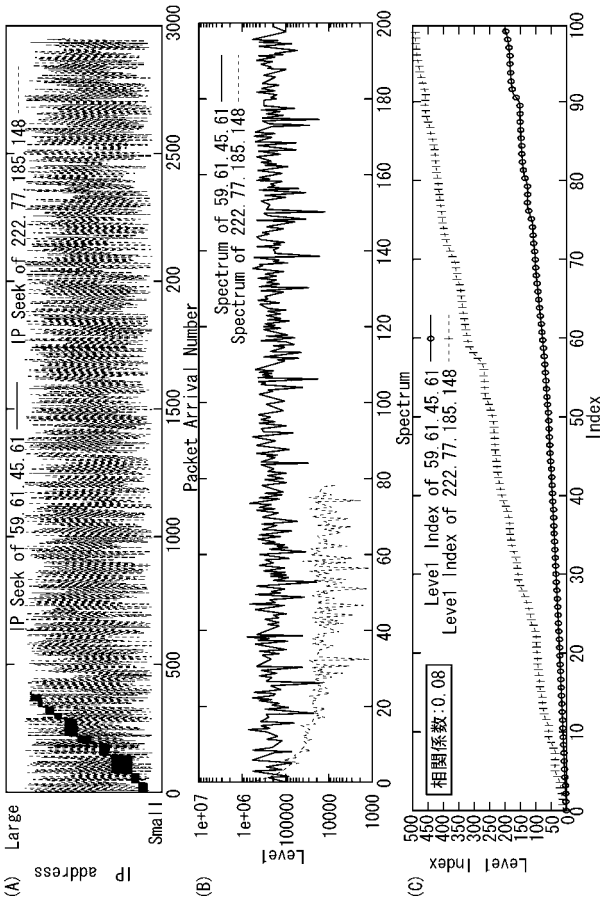
【 図 9 】



【 図 10 】



【 図 11 】



フロントページの続き

(72)発明者 井上 大介

東京都小金井市貫井北町4-2-1 独立行政法人情報通信研究機構内

(72)発明者 中尾 康二

東京都小金井市貫井北町4-2-1 独立行政法人情報通信研究機構内

Fターム(参考) 5B276 FD08

5B285 AA06 BA01 CA31 CA36 CA37