(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2008-176753
(P2008-176753A)
(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
G06F 13/00	(2006.01)	G06F	13/00	540E	5K030
H04L 12/56	(2006.01)	H04L	12/56	400Z	

審査請求 未請求　請求項の数 26　ＯＬ　（全 24 頁）

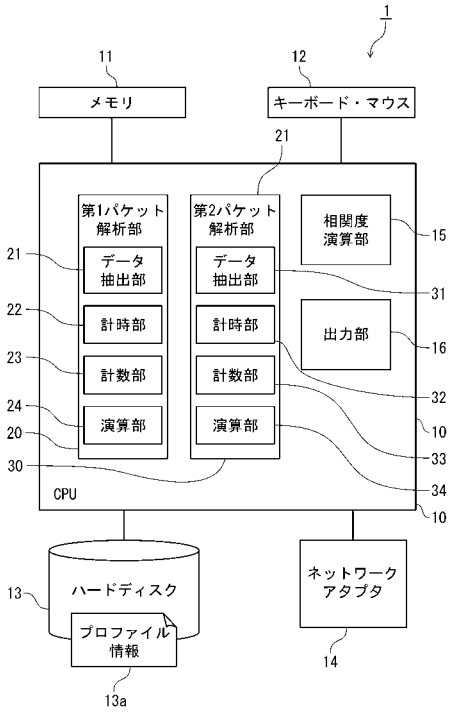
(21)出願番号　特願2007-12071 (P2007-12071)
(22)出願日　平成19年1月22日 (2007.1.22)

(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　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
(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
(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

最終頁に続く

(54)【発明の名称】データ類似性検査方法及び装置

(57)【要約】
【課題】　他のコンピュータに対して不正な処理を行うマルウェア等のソフトウェアと、検査するソフトウェアとの相関関係を高速、的確に自動算出する技術を提供すること。
【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パケットと、検査対象ソフトウェアが送信するパケットを、所定のパラメータに着目してその相関度を算出することによりデータの類似性を検査する。パラメータの種類をプロファイル情報１３ａとして備えておき、第１パケット解析手段２０が、それを参照してパラメータを取得する。これを第１のプロファイルデータとする。同様に第２パケット解析手段３０が、第２のプロファイルデータを作成する。そして、相関度演算手段１５が、両プロファイルデータから相関関係式を用いて相関度を算出する。
【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のソフトウェアが送信する第１の送信データと、検査対象の第２のソフトウェアが送信する第２の送信データとを比較してその類似性を検査するコンピュータのデータ類似性検査方法であって、
該送信データに関して、少なくとも送信されるパケットに含まれる情報のいずれかについて予め着目する単数又は複数のパラメータの種類をプロファイル情報として記憶手段に格納しておき、
第１パケット解析手段が、該プロファイル情報を参照し、第１のソフトウェアから送信される第１パケットを受信して、該パラメータの種類に該当するパラメータの一部又は全部を取得して第１のプロファイルデータとして記憶手段に記録する第１パケット解析ステップ、
第２パケット解析手段が、該プロファイル情報を参照し、第２のソフトウェアから送信される第２パケットを受信して、該パラメータの種類に該当するパラメータの一部又は全部を取得して第２のプロファイルデータとして記憶手段に記録する第２パケット解析ステップ、
コンピュータの相関度演算手段が、該第１のプロファイルデータと、該第２のプロファイルデータとから、予め記憶手段に格納されるパラメータの種類に応じた相関関係式を用いて相関度を算出する相関度演算ステップ、
該相関度を出力する出力ステップ
を有することを特徴とするデータ類似性検査方法。
【請求項２】
前記プロファイル情報に、
パケットの送信先ポートのポート番号の一部又は全部をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、送信先ポート番号をパケットから抽出する
ことを特徴とする請求項１に記載のデータ類似性検査方法。
【請求項３】
前記プロファイル情報に、
パケットの送信先ポートのポート番号と各ポート番号に対する送信回数をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、送信先ポート番号をパケットから抽出すると共に、所定時間における若しくはソフトウェアの所定回数の実行処理における各ポート番号に対するパケットの送信回数を計数する
ことを特徴とする請求項１又は２に記載のデータ類似性検査方法。
【請求項４】
前記プロファイル情報に、
パケットの送信先ポートのポート番号間の遷移情報をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、送信先ポート番号の遷移情報をパケットから抽出する
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データ類似性検査方法。
【請求項５】
前記プロファイル情報に、
パケットの送信先アドレスが複数ある場合における該送信先アドレス間の差分値又は差分値に基づく所定の統計値を導く計算をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、各送信先アドレスをパケットから抽出すると共に、該送信先アドレス間の差分値又は差分値に基づく所定の統計値を算出する
ことを特徴とする請求項１ないし４のいずれかに記載のデータ類似性検査方法。

【請求項６】
前記プロファイル情報に、
パケットの送信先アドレスと送信する送信先アドレスの数をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、所定時間において若しくはソフトウェアの所定回数の実行処理においてパケットを送信する送信先アドレスの数を計数する
ことを特徴とする請求項１ないし５のいずれかに記載のデータ類似性検査方法。
【請求項７】
前記プロファイル情報に、
パケットの送信元ポートのポート番号の数をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、所定時間において若しくはソフトウェアの所定回数の実行処理においてパケットを送信する送信元ポート番号の数を計数する
ことを特徴とする請求項１ないし６のいずれかに記載のデータ類似性検査方法。
【請求項８】
前記プロファイル情報に、
パケットの送信元ポートのポート番号が複数ある場合における該送信元ポート番号間の差分値又は差分値に基づく所定の統計値を導く計算をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、各送信元ポート番号をパケットから抽出すると共に、該送信元ポート番号間の差分値又は差分値に基づく所定の統計値を算出する
ことを特徴とする請求項１ないし７のいずれかに記載のデータ類似性検査方法。
【請求項９】
前記プロファイル情報に、
パケットが準拠するプロトコル、又は、ＴＣＰプロトコルにおけるフラグの少なくともいずれかの種類又はその数をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、該プロトコル又はフラグを該パケットから取得すると共に、数がパラメータの種類である場合にはその数を計数する
ことを特徴とする請求項１ないし８のいずれかに記載のデータ類似性検査方法。
【請求項１０】
前記プロファイル情報に、
パケットに含まれるペイロード(Payload)成分のダイジェスト(Digest)値又はその数又はそのいずれかに基づく所定の統計値を導く計算をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、各ペイロードのダイジェスト値をパケットから抽出すると共に、数がパラメータの種類である場合にはダイジェスト値の数を計数し、該ダイジェスト値若しくはその数に基づく所定の統計値がパラメータの種類である場合にはそれを算出する
ことを特徴とする請求項１ないし９のいずれかに記載のデータ類似性検査方法。
【請求項１１】
前記プロファイル情報に、
パケットに含まれるペイロード(Payload)成分のデータ量又は該データ量に基づく所定の統計値を導く計算をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、該ペイロード成分をパケットから抽出すると共に、該データ量を測定する、又は測定した該データ量に基づく所定の統計値を算出する
ことを特徴とする請求項１ないし１０のいずれかに記載のデータ類似性検査方法。
【請求項１２】
前記コンピュータのデータ類似性検査方法において、
該送信データに関して、少なくとも送信されるパケットが所定時間に送信先ポートに到

達する数をパラメータの種類に含めてプロファイル情報として記憶手段に格納しておき、
前記相関度演算ステップの前のいずれかの時点で
前記第１パケット解析手段が、該プロファイル情報を参照し、第１のソフトウェアから送信される第１パケットを所定時間に受信する数を計数して第１のプロファイルデータとして記憶手段に記録する第１パケット受信回数計数ステップ、
前記第２パケット解析手段が、該プロファイル情報を参照し、第２のソフトウェアから送信される第２パケットを所定時間に受信する数を計数して第２のプロファイルデータとして記憶手段に記録する第２パケット受信回数計数ステップ
の各ステップを行う
ことを特徴とする請求項１ないし１１のいずれかに記載のデータ類似性検査方法。
【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既知のマルウェアであり、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る検査対象のソフトウェアであって、
前記請求項１ないし１２に記載のデータ類似性検査方法における前記第１パケット解析ステップを単数又は複数の第１のソフトウェアについて実行処理した後、
コンピュータの検査開始指示信号送出手段が、該第２のソフトウェアからの送信データ内容又は送信データのパケットに基づいて検査開始指示信号を送出する検査開始指示ステップ、
該検査開始指示信号を契機として実行処理される前記第２パケット解析ステップ、
コンピュータの相関度演算手段が、該単数又は複数の第１のプロファイルデータと、該第２のプロファイルデータとから、予め記憶手段に格納されるパラメータの種類に応じた相関関係式を用いて相関度を算出する相関度演算ステップ、
該第１のソフトウェアとの各相関度の少なくとも一部を出力するマルウェア相関度出力ステップ
を有することを特徴とするマルウェアの検査方法。
【請求項１４】
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のソフトウェアが送信する第１の送信データと、検査対象の第２のソフトウェアが送信する第２の送信データとを比較してその類似性を検査するデータ類似性検査装置であって、
該送信データに関して、少なくとも送信されるパケットに含まれる情報のいずれかについて予め着目する単数又は複数のパラメータの種類をプロファイル情報として格納する記憶手段と、
該プロファイル情報を参照し、第１のソフトウェアから送信される第１パケットを受信して、該パラメータの種類に該当するパラメータの一部又は全部を取得して第１のプロファイルデータとして記憶手段に記録する第１パケット解析手段と、
該プロファイル情報を参照し、第２のソフトウェアから送信される第２パケットを受信して、該パラメータの種類に該当するパラメータの一部又は全部を取得して第２のプロファイルデータとして記憶手段に記録する第２パケット解析手段と、
該第１のプロファイルデータと、該第２のプロファイルデータとから、予め記憶手段に格納されるパラメータの種類に応じた相関関係式を用いて相関度を算出するコンピュータの相関度演算手段と、
該相関度を出力する出力手段と
を備えたことを特徴とするデータ類似性検査装置。
【請求項１５】
前記プロファイル情報に、
パケットの送信先ポートのポート番号の一部又は全部をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、送信先ポート番号をパケットから抽出する

ことを特徴とする請求項１４に記載のデータ類似性検査装置。
【請求項１６】
前記プロファイル情報に、
パケットの送信先ポートのポート番号と各ポート番号に対する送信回数をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、送信先ポート番号をパケットから抽出すると共に、所定時間における若しくはソフトウェアの所定回数の実行処理における各ポート番号に対するパケットの送信回数を計数する
ことを特徴とする請求項１４又は１５に記載のデータ類似性検査装置。
【請求項１７】
前記プロファイル情報に、
パケットの送信先ポートのポート番号間の遷移情報をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、送信先ポート番号の遷移情報をパケットから抽出する
ことを特徴とする請求項１４ないし１６のいずれかに記載のデータ類似性検査装置。
【請求項１８】
前記プロファイル情報に、
パケットの送信先アドレスが複数ある場合における該送信先アドレス間の差分値又は差分値に基づく所定の統計値を導く計算をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、各送信先アドレスをパケットから抽出すると共に、該送信先アドレス間の差分値又は差分値に基づく所定の統計値を算出する
ことを特徴とする請求項１４ないし１７のいずれかに記載のデータ類似性検査装置。
【請求項１９】
前記プロファイル情報に、
パケットの送信先アドレスと送信する送信先アドレスの数をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、所定時間において若しくはソフトウェアの所定回数の実行処理においてパケットを送信する送信先アドレスの数を計数する
ことを特徴とする請求項１４ないし１８のいずれかに記載のデータ類似性検査装置。
【請求項２０】
前記プロファイル情報に、
パケットの送信元ポートのポート番号の数をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、所定時間において若しくはソフトウェアの所定回数の実行処理においてパケットを送信する送信元ポート番号の数を計数する
ことを特徴とする請求項１４ないし１９のいずれかに記載のデータ類似性検査装置。
【請求項２１】
前記プロファイル情報に、
パケットの送信元ポートのポート番号が複数ある場合における該送信元ポート番号間の差分値又は差分値に基づく所定の統計値を導く計算をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、各送信元ポート番号をパケットから抽出すると共に、該送信元ポート番号間の差分値又は差分値に基づく所定の統計値を算出する
ことを特徴とする請求項１４ないし２０のいずれかに記載のデータ類似性検査装置。
【請求項２２】
前記プロファイル情報に、
パケットが準拠するプロトコル、又は、ＴＣＰプロトコルにおけるフラグの少なくともいずれかの種類又はその数をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、該プロトコル又はフラグを該パケット

から取得すると共に、数がパラメータの種類である場合にはその数を計数する
ことを特徴とする請求項１４ないし２１のいずれかに記載のデータ類似性検査装置。
【請求項２３】
前記プロファイル情報に、
パケットに含まれるペイロード(Payload)成分のダイジェスト(Digest)値又はその数又はそのいずれかに基づく所定の統計値を導く計算をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、各ペイロードのダイジェスト値をパケットから抽出すると共に、数がパラメータの種類である場合にはダイジェスト値の数を計数し、該ダイジェスト値若しくはその数に基づく所定の統計値がパラメータの種類である場合にはそれを算出する
ことを特徴とする請求項１４ないし２２のいずれかに記載のデータ類似性検査装置。
【請求項２４】
前記プロファイル情報に、
パケットに含まれるペイロード(Payload)成分のデータ量又は該データ量に基づく所定の統計値を導く計算をパラメータの種類として格納しておき、
前記第１及び第２パケット解析手段がそれぞれ、該ペイロード成分をパケットから抽出すると共に、該データ量を測定する、又は測定した該データ量に基づく所定の統計値を算出する
ことを特徴とする請求項１４ないし２３のいずれかに記載のデータ類似性検査装置。
【請求項２５】
前記コンピュータのデータ類似性検査装置において、
記憶手段が、該送信データに関して、少なくとも送信されるパケットが所定時間に送信先ポートに到達する数をパラメータの種類に含めてプロファイル情報として備えると共に、
前記第１パケット解析手段が、該プロファイル情報を参照し、第１のソフトウェアから送信される第１パケットを所定時間に受信する数を計数して第１のプロファイルデータとして記憶手段に記録し、
前記第２パケット解析手段が、該プロファイル情報を参照し、第２のソフトウェアから送信される第２パケットを所定時間に受信する数を計数して第２のプロファイルデータとして記憶手段に記録する
ことを特徴とする請求項１４ないし２４のいずれかに記載のデータ類似性検査装置。
【請求項２６】
マルウェアを検査するマルウェア検査システムであって、
前記請求項１４ないし２５のいずれかに記載のデータ類似性検査装置を備え、
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既知のマルウェアであり、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る検査対象のソフトウェアである構成において、
前記第１パケット解析手段が、単数又は複数の第１のソフトウェアについて各第１のプロファイルデータを記録すると共に、
該第２のソフトウェアからの送信データ内容又は送信データのパケットに基づいて検査開始指示信号を送出するコンピュータの検査開始指示信号送出手段を備え、
前記第２パケット解析手段が、該検査開始指示信号を受信すると作動し、
前記相関度演算手段が、該単数又は複数の第１のプロファイルデータと、該第２のプロファイルデータとから、予め記憶手段に格納されるパラメータの種類に応じた相関関係式を用いて相関度を算出し、
前記出力手段が、該第１のソフトウェアとの各相関度の少なくとも一部を出力する
を有することを特徴とするマルウェア検査システム。
【発明の詳細な説明】
【技術分野】

【０００１】
本発明はソフトウェアが出力するデータの類似性を検査する方法と装置に関し、特にネットワーク上で不正な処理を行うコンピュータから送信されるパケットに基づいてそのデータの類似性あるいは、そのソフトウェアの類似性を検査する方法と装置に係る技術である。
【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、ウィルス(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】
マルウェアの自動解析方法としては、特許文献１のような技術が知られている。
【０００７】
【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
【特許文献１】特表2006-522395号公報
【発明の開示】
【発明が解決しようとする課題】
【０００８】
このように、従来の技術においてはミクロ的なインシデントの解析、マクロ的なマルウェアの解析が別個に行われており、ネットワーク上で発生しているインシデントが、解析済みのどのマルウェアの作用によるものなのか、特定は人手に頼らざるを得なかった。
自動的に特定する場合にも、挙動が極めて類似していれば正解が得られやすいが、実際のマルウェアは、検出されにくいように様々な挙動をするように作成されているため、受信するコンピュータによってその態様は多様に変化する。
【０００９】

本発明は、このように他のコンピュータに対して不正な処理を行うマルウェア等のソフトウェアと、検査するソフトウェアとの相関関係を高速、的確に自動算出する技術を提供するものである。

【課題を解決するための手段】

【００１０】

本発明は、上記の課題を解決するために、次のようなデータ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のソフトウェアが送信する第１の送信データと、検査対象の第２のソフトウェアが送信する第２の送信データとを比較してその類似性を検査するコンピュータのデータ類似性検査方法を提供するものである。

本方法においては、送信データに関して、少なくとも送信されるパケットに含まれる情報のいずれかについて予め着目する単数又は複数のパラメータの種類をプロファイル情報として記憶手段に格納しておく。

【００１１】

そして、第１パケット解析手段が、プロファイル情報を参照し、第１のソフトウェアから送信される第１パケットを受信して、そこに格納されているパラメータの種類に該当するパラメータの一部又は全部を取得する。これを第１のプロファイルデータとして記憶手段に記録する。（第１パケット解析ステップ）

【００１２】

その上で、第２パケット解析手段が、プロファイル情報を参照し、第２のソフトウェアから送信される第２パケットを受信して、パラメータの種類に該当するパラメータの一部又は全部を取得して第２のプロファイルデータとして記憶手段に記録する。（第２パケット解析ステップ）

【００１３】

続く相関度演算ステップにおいて、コンピュータの相関度演算手段が、該第１のプロファイルデータと、該第２のプロファイルデータとから、予め記憶手段に格納されるパラメータの種類に応じた相関関係式を用いて相関度を算出し、出力ステップにおいて相関度を出力する。

【００１４】

また、本発明は次のようなデータ類似性検査装置を提供することもできる。

すなわち、請求項１４に記載のように、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のソフトウェアが送信する第１の送信データと、検査対象の第２のソフトウェアが送信する第２の送信データとを比較してその類似性を検査するデータ類似性検査装置を提供する。

【００１５】

本装置において、送信データに関して、少なくとも送信されるパケットに含まれる情報のいずれかについて予め着目する単数又は複数のパラメータの種類をプロファイル情報として格納する記憶手段を備える。

また、プロファイル情報を参照し、第１のソフトウェアから送信される第１パケットを受信して、該パラメータの種類に該当するパラメータの一部又は全部を取得して第１のプロファイルデータとして記憶手段に記録する第１パケット解析手段と、プロファイル情報を参照し、第２のソフトウェアから送信される第２パケットを受信して、該パラメータの種類に該当するパラメータの一部又は全部を取得して第２のプロファイルデータとして記憶手段に記録する第２パケット解析手段を備える。

【００１６】

以上の各パケット解析手段で記録された各プロファイルデータを利用し、予め記憶手段に格納されるパラメータの種類に応じた相関関係式を用いて相関度を算出するコンピュータの相関度演算手段、相関度を出力する出力手段を備える。

【００１７】

本発明において、上記プロファイル情報に、パケットの送信先ポートのポート番号の一部又は全部をパラメータの種類として格納しておき、第１及び第２パケット解析手段がそれぞれ、送信先ポート番号をパケットから抽出する構成でもよい。

【００１８】

上記プロファイル情報に、パケットの送信先ポートのポート番号と各ポート番号に対する送信回数をパラメータの種類として格納しておき、第１及び第２パケット解析手段がそれぞれ、送信先ポート番号をパケットから抽出すると共に、所定時間における若しくはソフトウェアの所定回数の実行処理における各ポート番号に対するパケットの送信回数を計数する構成でもよい。

【００１９】

上記プロファイル情報に、パケットの送信先ポートのポート番号間の遷移情報をパラメータの種類として格納しておき、第１及び第２パケット解析手段がそれぞれ、送信先ポート番号の遷移情報をパケットから抽出する構成でもよい。

【００２０】

上記プロファイル情報に、パケットの送信先アドレスが複数ある場合における該送信先アドレス間の差分値又は差分値に基づく所定の統計値を導く計算をパラメータの種類として格納しておき、第１及び第２パケット解析手段がそれぞれ、各送信先アドレスをパケットから抽出すると共に、該送信先アドレス間の差分値又は差分値に基づく所定の統計値を算出する構成でもよい。

【００２１】

上記プロファイル情報に、パケットの送信先アドレスと送信する送信先アドレスの数をパラメータの種類として格納しておき、第１及び第２パケット解析手段がそれぞれ、所定時間において若しくはソフトウェアの所定回数の実行処理においてパケットを送信する送信先アドレスの数を計数する構成でもよい。

【００２２】

上記プロファイル情報に、パケットの送信元ポートのポート番号の数をパラメータの種類として格納しておき、第１及び第２パケット解析手段がそれぞれ、所定時間において若しくはソフトウェアの所定回数の実行処理においてパケットを送信する送信元ポート番号の数を計数する構成でもよい。

【００２３】

上記プロファイル情報に、パケットの送信元ポートのポート番号が複数ある場合における該送信元ポート番号間の差分値又は差分値に基づく所定の統計値を導く計算をパラメータの種類として格納しておき、第１及び第２パケット解析手段がそれぞれ、各送信元ポート番号をパケットから抽出すると共に、該送信元ポート番号間の差分値又は差分値に基づく所定の統計値を算出する構成でもよい。

【００２４】

上記プロファイル情報に、パケットが準拠するプロトコル、又は、ＴＣＰプロトコルにおけるフラグの少なくともいずれかの種類又はその数をパラメータの種類として格納しておき、第１及び第２パケット解析手段がそれぞれ、該プロトコル又はフラグを該パケットから取得すると共に、数がパラメータの種類である場合にはその数を計数する構成でもよい。

【００２５】

上記プロファイル情報に、パケットに含まれるペイロード(Payload)成分のダイジェスト(Digest)値又はその数又はそのいずれかに基づく所定の統計値を導く計算をパラメータの種類として格納しておき、第１及び第２パケット解析手段がそれぞれ、各ペイロードのダイジェスト値をパケットから抽出すると共に、数がパラメータの種類である場合にはダイジェスト値の数を計数し、該ダイジェスト値若しくはその数に基づく所定の統計値がパラメータの種類である場合にはそれを算出する構成でもよい。

【００２６】

上記プロファイル情報に、パケットに含まれるペイロード(Payload)成分のデータ量又

は該データ量に基づく所定の統計値を導く計算をパラメータの種類として格納しておき、第１及び第２パケット解析手段がそれぞれ、該ペイロード成分をパケットから抽出すると共に、該データ量を測定する、又は測定した該データ量に基づく所定の統計値を算出する構成でもよい。

【００２７】

また、本発明は上記のデータ類似性検査方法において、該送信データに関して、少なくとも送信されるパケットが所定時間に送信先ポートに到達する数をパラメータの種類に含めてプロファイル情報として記憶手段に格納しておくこともできる。

本構成において、上記の相関度演算ステップの前のいずれかの時点で第１パケット解析手段が、該プロファイル情報を参照し、第１のソフトウェアから送信される第１パケットを所定時間に受信する数を計数して第１のプロファイルデータとして記憶手段に記録する第１パケット受信回数計数ステップ、第２パケット解析手段が、該プロファイル情報を参照し、第２のソフトウェアから送信される第２パケットを所定時間に受信する数を計数して第２のプロファイルデータとして記憶手段に記録する第２パケット受信回数計数ステップの各ステップを行う。

本技術を実装したデータ類似性検査装置を提供することもできる。

【００２８】

本発明は、上記データ類似性検査方法の技術により、次のようなマルウェアの検査方法を提供してもよい。

すなわち、請求項１３に記載の発明によれば、上記第１のソフトウェアとして、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既知のマルウェアを、第２のソフトウェアとして、広域ネットワークにおいて実際に実行される検査対象のソフトウェアをそれぞれ用い、上記の第１パケット解析ステップを単数又は複数の第１のソフトウェアについて実行処理する。

【００２９】

その後、コンピュータの検査開始指示信号送出手段が、該第２のソフトウェアからの送信データ内容又は送信データのパケットに基づいて検査開始指示信号を送出する検査開始指示ステップ、この検査開始指示信号を契機として実行処理される前記第２パケット解析ステップ、コンピュータの相関度演算手段が、該単数又は複数の第１のプロファイルデータと、該第２のプロファイルデータとから、予め記憶手段に格納されるパラメータの種類に応じた相関関係式を用いて相関度を算出する相関度演算ステップ、該第１のソフトウェアとの各相関度の少なくとも一部を出力するマルウェア相関度出力ステップを順に実行する。

【００３０】

さらに、マルウェアを検査するマルウェア検査システムとして提供することもできる。

すなわち、上記のデータ類似性検査装置を備え、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既知のマルウェアであり、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る検査対象のソフトウェアである構成において、第１パケット解析手段が、単数又は複数の第１のソフトウェアについて各第１のプロファイルデータを記録すると共に、第２のソフトウェアからの送信データ内容又は送信データのパケットに基づいて検査開始指示信号を送出するコンピュータの検査開始指示信号送出手段を備える。

【００３１】

そして、第２パケット解析手段が、該検査開始指示信号を受信すると作動し、相関度演算手段が、該単数又は複数の第１のプロファイルデータと、該第２のプロファイルデータとから、予め記憶手段に格納されるパラメータの種類に応じた相関関係式を用いて相関度を算出し、出力手段が、該第１のソフトウェアとの各相関度の少なくとも一部を出力する。

【発明の効果】

【００３２】

本発明は、以上の構成を備えることにより、次の効果を奏する。
まず、２つのソフトウェアが送信するデータの類似性を検査する際に、両ソフトウェアから出力されるパケットから、好適なパラメータの種類に着目し、これをプロファイル情報として格納することができる。
【００３３】
そしてプロファイル情報に基づいてパラメータを抽出すると共に、パラメータ毎に設定してある相関関係式に基づいて両者の比較を行う。本方法によれば、パケットの送信態様が複雑で、しかも頻繁に変化するマルウェアについても複数のパラメータを融合して相関度を求めることが容易であり、その結果として高精度に相関度を算出することができる。
【００３４】
またプロファイル情報を書き換えるだけで新しいマルウェアの検出にも迅速簡便に対応することが可能であり、急速に、かつ多岐にわたって進化するマルウェアの検査方法としても最適である。
【００３５】
特に本発明では、パラメータの種類として特に好適なものを提供するので、マルウェアの一般的な不正処理のパケットに対し、効率よく相関度を求めることができる。
【発明を実施するための最良の形態】
【００３６】
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は本発明に係るデータ類似性検査装置（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。
【００３７】
本装置（１）には、演算処理等を司るＣＰＵ（１０）を中心として、ＣＰＵ（１）と協働するメモリ（１１）、ユーザが入力等を行うキーボード及びマウス（１２）、データを読み書き自在に格納するハードディスク（１３）、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（１４）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。
【００３８】
本発明は、上記ＣＰＵ（１０）に、第１パケット解析部（２０）と、第２パケット解析部（３０）と、相関度演算部（１５）と、出力部（１６）とを備える。第１及び第２パケット解析部は、以下説述するさまざまなパラメータを取得するために、データ抽出部（２１）（３１）、計時部（２２）（３２）、計数部（２３）（３３）、演算部（２４）（３４）を備える。
【００３９】
はじめに、本発明の最も基本的な処理を図２を用いて説明する。本実施例では不正処理を行う第１のソフトウェアとしてマルウェアを、第２のソフトウェアとして検査対象となるソフトウェアを用いる。
まず、プロファイル情報（１３ａ）を読み出す。（プロファイル読出処理：Ｓ１０）
ここでプロファイル情報（１３ａ）とは、ソフトウェアから送信されるデータの類似性を検査するために着目するパラメータの種類とその抽出方法を予め定義したものである。パラメータの種類として本発明で用いるのは、例えばパケットに含まれる送信先ホストにおけるポート番号や、送信元ポートのポート番号などであり、それぞれについて、ＴＣＰ（Transmission Control Protocol）（RFC793により規定されている）のヘッダ部分における最初の０～１５ビット（送信元ポート番号）、１６ビット～３１ビット（送信先ポート番号）に配置されていることが記録される。
本発明ではこのパラメータの種類としていかなるものを用いるのかについても重要な意

義がある。

【００４０】

パラメータの種類をハードディスクに格納する場合に、具体的にはコンピュータで用いられる変数の名称として記録されている。例えば、送信先ホストにおけるポート番号という種類を表す変数は「DstPort」としており、送信元ホストのポート番号は「SrcPort」としている。そして、このパラメータとしては「139」「445」などのデータである。以下、パラメータの種類とパラメータとはこのように区別して使用する。

【００４１】

最初のプロファイル情報読出処理（Ｓ１０）では、まずこのようにパラメータの種類（変数及びその抽出方法に関するデータ）を読み出し処理する。具体的な処理方法は、ＣＰＵ（１０）とハードディスク（１３）の協働によるものであり公知である。

そして、読み出したプロファイル情報（１３ａ）に基づいてマルウェアからのパケットを解析する。（第１パケット解析処理：Ｓ１１）

【００４２】

ここでの解析方法は、プロファイル情報によって様々であるから、具体的なパラメータの種類と共に解析方法を後述する。

そして、得られたパラメータのデータを、第１プロファイルデータ（１３ｂ）としてハードディスク（１３）やメモリ（１１）などの記憶手段に記録する。（プロファイル記録処理：Ｓ１２）

上記処理（Ｓ１０～S１２）は第１パケット解析部（２０）による処理である。

【００４３】

本発明では以上の各処理を経て、マルウェアに関する第１のプロファイルデータを生成する。１つのマルウェアに関してのプロファイルデータを作成するだけでもよいが、多数のマルウェアのプロファイルデータを生成しておくことが好ましく、これによって次の検査対象ソフトウェアとの比較の際に、多数の候補から最も相関が認められるマルウェアを特定することができる。

【００４４】

検査対象ソフトウェアについてもマルウェアと同様の処理を行う。すなわち、プロファイル情報（１３ａ）を読み出し（プロファイル情報読出処理：Ｓ２０）、検査対象ソフトウェアが送信するパケットをプロファイル情報に基づいて解析処理する。（第２パケット解析処理：Ｓ２１）

そして、取得されたパラメータを第２プロファイルデータ（１３ｃ）として記録する。（プロファイル記録処理：Ｓ２２）

以上の各処理（Ｓ２０～S２２）は第２パケット解析部（３０）による処理である。

【００４５】

以上の第２プロファイルデータ（１３ｃ）と、予め記録されている第１プロファイルデータ（１３ｂ）とを比較するために、相関度演算部（１５）が、予め定義された相関関係式に基づいて相関度を算出する。（相関度算出処理：Ｓ３０）

相関関係式は、プロファイル情報のパラメータの種類ごとに定義しておくことが好ましく、この場合にはプロファイル情報にパラメータの種類に対応して格納される。もっとも、本発明では相関関係式は１種類として、ハードディスク（１３）に別に格納しておいてもよい。相関関係式についても後述する。

【００４６】

そして、出力部（１６）からマルウェアごとの相関度を出力する。（出力処理：Ｓ３１）

出力の方法は、一覧形式にしてモニタからの画面表示、プリンタからの印刷出力、ネットワークアダプタ（１４）を介して他のコンピュータへのデータ送信など、いかなる方法でもよい。ＣＰＵ（１０）の出力部（１６）は公知のハードウェアと協働して同処理を司る。

【００４７】

本発明では、さらにマルウェアの特定方法又はマルウェアの特定システムとして提供することもできる。基本的な技術としては上記データ類似性検査装置と同様であるが、本件出願人らが提案しているnicter（非特許文献１参照）に組み合わせてマルウェアの特定を行うことを提案する。

【００４８】

図３において、まず広域ネットワーク（４０）に複数設けたセンサー（４１）で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（４２）に入力する。マクロ解析の結果はデータベース（４３）に格納される。

【００４９】

一方、ネットワーク（４４）上で、キャプチャ（４５）によって多数のマルウェア検体を採集し、ミクロ解析器（４６）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（４７）に格納する。

【００５０】

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、それぞれのデータベースから相関分析器（４８）で相関分析を行うことが考えられている。

【００５１】

相関分析の結果はデータベース（４９）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（５０）を介してユーザ（５１）に通知されたり、レポート（５２）として出力されたりする。

【００５２】

このシステムに対して、本発明を適用し、マクロ解析器（４２）とミクロ解析器（４６）とともに、相関分析器（４８）に本発明のデータ類似性検査装置（１）を実装する。

もっとも、本装置（１）を分割してマクロ解析器（４２）に第２パケット解析部（３０）を、ミクロ解析器（４６）に第１パケット解析部（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、相関度演算部（１５）を備えた相関分析器（４８）で相関分析してもよい。

【００５３】

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。

【００５４】

ミクロ解析器（４６）における動作として、図４に示すようなマルウェア動的解析環境を用いた解析が好適である。

すなわち、マルウェア動的解析環境とは仮想的にローカルアドレス空間、グローバルアドレス空間を設けて、感染したホストが各空間においてどのような挙動を示すかを測定する箱庭環境である。

【００５５】

そして、感染したホスト（７０）から仮想グローバル空間（６０）の各ＩＰアドレスのホスト（６１）に対してどのようにグローバルアドレスをスキャンするか、あるいはダミーのＩＲＣサーバ（６２）、ＨＴＴＰサーバ（６３）、ＦＴＰサーバ（６４）、ＴＦＴＰサーバ（６５）に対してどのようにアクセスするかなどを測定する。

このような仮想グローバル空間（６０）を用いることで、グローバルアドレス空間への挙動を測定することができる。

【００５６】

また、仮想ローカル空間（６６）においては、感染したホスト（７０）から複数のローカルＩＰアドレスのホスト（６７）に対してどのようにローカルアドレスをスキャンするか、あるいはダミーのＳＭＴＰサーバ（６８）、ＤＮＳサーバ（６９）に対してどのようにアクセスするのかなどを測定する。

このような仮想ローカル空間（６６）を用いることで、ローカルアドレス空間への挙動

を測定することができる。

【００５７】

一方、ダークネットと呼んでいる、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広域ネットワーク（４０）上のセンサー（４４）で検知する。

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。

【００５８】

ここでセンサー（４４）はダークネットにあたるＩＰアドレスが付与されたネットワーク端末で構成する。

さらに、本発明ではセンサー（４４）又はマクロ解析器（４２）に公知の不正な挙動を検出する機能を付与する。例えば非特許文献２に該方法のアルゴリズムが記載されているが、マクロ解析器の構成は非特許文献３などにも開示されるようにさまざまな技術を適用することができる。以下ではセンサー（４４）に同機能を付与したものとして説明する。

【００５９】

【非特許文献２】Suzuki, K., Baba, S., Takakura, H.: Analyzing traffic directed to unused IPaddress blocks, IEICE Technical Report, vol.105, no.530, IA2005-23, pp.25-302006年１月

【非特許文献３】Takeuchi, J., Sato, Y., Rikitake K., Nakao K.:Development of Incident Detection System Based on Change Point Detection, SCIS2006, The 2006 Symposium on Cryptography and Information Security, Japan, 2006年１月

【００６０】

相関分析器（４８）に本装置（１）を実装する場合には、検査開始指示信号として受信したパケットの生データを用いるのが簡便であり、該信号に基づいて図２のプロファイル情報読出処理（Ｓ２０）から相関度算出処理（Ｓ３０）、出力処理（Ｓ３１）までを行ってもよい。

【００６１】

このとき、図５に示すような装置構成をとってもよい。

すなわち、図３の相関分析器（４８）において、プロファイルデータ生成部（８０）と相関度演算部（８２）を設け、プロファイルデータ生成部（８０）には第１パケット解析部（２０）と第２パケット解析部（３０）の共通する処理手段を設け、相関度演算部（８２）は本装置（１）の相関度演算部（１５）と同様の手段とする。

【００６２】

本構成において、プロファイルデータ生成部（８０）が予め相関分析結果データベース（４９）にミクロ解析から得たマルウェアにつき複数のプロファイルデータ（８３）～（８７）・・を作成する。

また、ダークネットからの生データが検査開始指示信号として入力されると、同じプロファイルデータ生成部（８０）が該生データに関するプロファイルデータ（８１）を作成し、相関度演算部（８２）で相関度を演算する。

【００６３】

以上、説明したように、本発明における第１パケット解析部、第２パケット解析部などは１台の装置に実装せずに複数の装置に分割して備えてもよいし、また各パケット解析部における機能に応じて複数の装置に分割して配置してもよい。さらに、各パケット解析部は同様の処理内容をなすものであるから、これを上記のように１個の処理部で構成してもよい。

【００６４】

次に、本発明に係るパケット解析部の詳細な処理内容を説述する。ここではプロファイル情報（１３ａ）におけるパラメータの種類と重要な関連があるため、好適なパラメータの種類とあわせて説明する。

まず、本実施例で用いるパラメータの種類としては次の表１に掲げる項目がある。

【００６５】
【表１】

データの内容	パラメータの種類（変数）
送信先ポート	DstPort_Count
送信先ポート	DstPort_Trans
送信元ポート	SrcPort_Unique
送信元ポート	SrcPortDif_Stats
送信先ＩＰアドレス	DstIPDif_Stats
送信先ＩＰアドレス	DstIP_Unique
プロトコル	Protocol_Count
フラグ	Flag_Count
時間	NumPacketRate
ペイロード	PayloadSig_Count
ペイロード	Payload_Stats
TTL	TTL_Stats
ID番号	Id_Stats
シーケンス番号	SeqNum_Stats

【００６６】
(DstPort_Count)
不正処理における送信データのパケットをどの送信先ポート番号に送るかという情報は、マルウェアを特徴づけるもっとも基本的な情報である。
そこで、最初のパラメータとしてDstPort_Countは、パケットの送信先ポートのポート番号と各ポート番号に対する送信回数を表す。そして、第１パケット解析部（２０）におけるデータ抽出部（２１）が、送信されたパケットのヘッダ部分から送信先ポート番号の部分を抽出する。
本実施例では、送信先ポート番号だけを用いてもよいが、送信先ポート番号に何回パケットを送ったかもパラメータとして利用する。
【００６７】
ここで、パケットを送った回数は、単位時間あたりの回数を計数する。すなわち、データ抽出部（２１）で送信先ポート番号を抽出すると共に、計時部（２２）で単位時間を計時し、その間の送信回数を計数部（２３）で計数する。単位時間とは例えば１秒間など、適宜定められる時間である。
以下、他のパラメータにおいても送信回数はこれと同様にして求める。
【００６８】
（DstPort_Trans）
次に、DstPort_Transはパケットの送信先ポートのポート番号間の遷移情報である。遷移情報とは、例えば単位時間あるいは一連の当該ホストへのパケット送信において、送信先のポートがどのような順でスキャンされるかに関する情報である。
例えば、ポート番号139、445、3127、6659の順でスキャンした場合に、139の次が445であるという情報をパラメータとする。
より好適には、遷移確率を用いて、139の次が445である確率をパラメータとしてもよい。具体的には遷移順序を記憶しておき、各ポート番号について139の次にパケットが送られた割合を遷移確率とする。全てのポート番号について格納せず、例えば上位１０位の遷移確率を格納するようにしてもよい。
【００６９】
(SrcPort_Unique)
送信元ポート番号は、送信先ポート番号に比較してマルウェアが自由に設定できる特徴がある。マルウェアによっては固定した送信元ポート番号を用いるため、どのポートをいくつ用いているかをパラメータとして利用すると好適である。また、送信元ポート番号が変動する場合でも、どのようにポート番号を変化させているかがパラメータとして利用できる。
そこで、データ抽出部（２２）が送信先ポート番号を抽出し、その番号をパラメータと

してプロファイルデータに格納する。

【００７０】

(SrcPortDif_Stats)

また、送信元ポート番号の変動については、送信元ポート番号を順にデータ抽出部（２１）で抽出して、前後のポート番号の差分を演算部（２４）で求めると共に、その平均値を算出する。なお、平均値以外に任意の統計計算式を用いて統計値を求めてもよい。

これは送信元ポート番号をある値ごとに変化させて送信する場合などに好適な指標となる。

【００７１】

(DstIPDif_Stats)

グローバルアドレス空間及びローカルアドレス空間において、どのようなＩＰアドレスに対してパケットを送ってくるかも利用することができる。ＩＰアドレスをどのようにスキャンするのかはマルウェアが標的を探す上で重要であり、その検索順序もさまざまな方法がある。サブネットマスクが１６ビットか、２４ビットか、近いＩＰアドレスから検索するのかランダムに検索するのか、などマルウェアの特定に役立つ情報である。

該情報は仮想ホストに割り当てたＩＰアドレスから取得することができる。

【００７２】

送信するＩＰアドレスが特異な特徴を持つ場合など、ＩＰアドレス自体をパラメータとすることもできるし、ＩＰアドレスを順に検出して、規定の方法で数値に変換し、演算部で差分を求め、その平均値を用いてもよい。平均値を他の統計値としてもよい。

【００７３】

(DstIP_Unique)

DstIP_Unique変数はいくつの送信先ＩＰアドレスに送信するかに関するパラメータである。そこで、単位時間における送信先ＩＰアドレスの数を計数する。マルウェアは単位時間に大量のパケットを送信するものと、一定の時間をおいて送信するものとがあり、その間隔を指標として用いることができる。単位時間やその他所定の時間における回数でもよいし、一連のパケットの送信における送信頻度でもよい。

【００７４】

（Protocol_Count）

Protocol_Count変数は、マルウェアがどのプロトコルによっていくつのパケットを送信しているか、そのプロトコル毎のパケット数をパラメータとするものである。マルウェアは様々なプロトコルを利用するため、ＴＣＰによるのかＵＤＰによるのか、どのプロトコルでスキャンされているか、データ抽出部（２１）により特定する。

【００７５】

（Flag_Count）

同様に、ＴＣＰをプロトコルとして用いている場合、ＴＣＰのどのフラグにどのようなパケットがくるかの情報を用いることもできる。

フラグの種類としては、URG（urgent）フラグ、ACK（acknowledge）フラグ、PSH（push）フラグ、RST（reset）フラグ、SYN（synchronize）フラグ、FIN（finis、終了）フラグなどがある。

そして、さまざまなスキャン方法が知られており、TCP SYNスキャン、TCP FINスキャン、TCPNullスキャン、TCP Xmasスキャン、TCP Maimon スキャン、TCP ACKスキャン、UDPスキャンなど、スキャンの種類をパラメータとしてもよい。

【００７６】

(NumPacketRate)

本発明では、単位時間あたりに送信先ホストにどれだけのパケットを送信してくるかの情報を用いることもできる。マルウェアによって連続的に送ってくる場合や、時間をおきながらパケットを送ってくる場合もある。

本発明では計時部（２２）及び計数部（２３）を用いてこれを計数してパラメータとする。

【００７７】
(PayloadSig_Count)
UDPによる探索を行うマルウェアなど、パケットのペイロードフィールドが攻撃用のコードを含む場合がある。そこで、ペイロードのダイジェスト値を比較することも有効である。まず実際のデータ（ダイジェスト値）を抽出して、その値自体や、ダイジェスト値の数（種類数）をパラメータとして用いることができる。
【００７８】
(Payload_Stats)
Payload_Statsはペイロードのサイズの統計値である。平均値など任意の統計値を用いることができる。
【００７９】
(TTL_Stats,Id_Stats,SeqNum_Stats)
その他、本実施例では、ＴＣＰにおけるＴＴＬフィールド、ＩＤフィールド、シーケンス番号フィールドなどの値や、その統計値を用いることもできる。
【００８０】
本発明は、以上説述した通りのパラメータの種類を用いることを提案し、これをプロファイルデータとして第１パケット解析部（２０）が、データ抽出部（２１）、計時部（２２）、計数部（２３）、演算部（２４）によって解析する。
また、これと全く同様に、第２パケット解析部（３０）が、データ抽出部（３１）、計時部（３２）、計数部（３３）、演算部（３４）によって第２のソフトウェアを解析する。
【００８１】
さらに、相関度演算部においては第１のプロファイルデータ（１３ｂ）と第２のプロファイルデータ（１３ｃ）から、任意の相関算出式により相関度を算出する。
次に、本実施例で用いた相関関係式について説明する。
【００８２】
本発明で用いたプロファイルデータは、ポート番号ごとに値をもつ場合など、多次元データとなる場合も多く、これらの多次元データ間の相関度を演算する必要がある。
例えば、第１のソフトウェアの使用したポート番号群P_Aと、第２のソフトウェアが使用したポート番号群P_Bに対して、$p \in P_A$、$q \in P_B$とする。このとき各ポート番号における第１のソフトウェアの送信回数$C_A(p)$と第２のソフトウェアの送信回数$C_B(q)$とあらわすことができる。
【００８３】
$C_A(p)$とC_B（p）を２つの確率変数と考える。このとき$p \in P_A \cup P_B = \Omega$（標本空間）である。そして、$C_A(p)$と$C_B$（p）の相関度は次式（数１）によって表される。なお、m_Aとm_Bは、それぞれ$C_A(p)$とC_B（p）の平均値である。
【００８４】
【数１】

$$\frac{\sum_{p\in\Omega}\{C_A(p)-m_A\}\{C_B(p)-m_B\}}{\sqrt{\sum_{p\in\Omega}\{C_A(p)-m_A\}^2}\sqrt{\sum_{p\in\Omega}\{C_B(p)-m_B\}^2}}$$

【００８５】
相関度演算部（１５）において、上記式に従って、相関度を算出することにより、多次元データの相関度が得られる。該演算方法は公知のプログラム手法によることができる。なお、同様に多次元データとなるのは、上記においてDstPort_Count、Protocol_Count 、Flag_Count、PayloadSig_Countなどである。
【００８６】
一方、１次元のデータとしては、SrcPort_Unique、DstIP_Unique、NumPacketRateなど

があり、これらは単純に次式（数２）によって類似度を求めることができる。
（数２）
（ａ／ｂ）^{k}

なお、a<b、kは任意に定義可能であり、ここでは１を用いている。
【００８７】
相関度演算部（１５）で用いる相関関係式としては、プロファイルデータによって任意であるが、パラメータがデータの系列である場合、次のような相関関係式を用いることができる。
【００８８】
(1) ピアソンの積率相関係数
２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nに対して、ピアソンの積率相関係数は次式（数３）によって求められる。
【００８９】
【数３】

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

ここで,
$\bar{x},\bar{y}$
はそれぞれデータ系列x,yの平均値をあらわす.
【００９０】
(2) スピアマンの順位相関係数
２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nに対して、スピアマンの順位相関係数は次式（数４）によって求められる。
【００９１】
【数４】

$$1-\frac{6\sum_{i=1}^{n}D_i^2}{n(n^2-1)}$$

ここで、D_iはx_iの系列x内での順位とy_iの系列y内での順位の差を表す。
【００９２】
(3)ケンドールの順位相関係数
２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nに対して、ケンドールの順位相関係数は次式（数５）によって求められる。
【００９３】
【数５】

$$\frac{K-L}{\frac{1}{2}n(n-1)}$$

ここで、Kは、i=1，2，…，nについてx_iの系列x内での順位がy_iの系列y内での順位より大きい（または順位が等しい）場合の数、Lは、y_iの系列y内での順位がx_iの系列x内での順位より大きい(または順位が等しい)場合の数とする。
【００９４】

(4) 特定条件を満たす割合

２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nについて、x_i,y_iが共に特定の条件を満たす場合の数を$g_{condition}$回とする。このとき、xとyの類似度を以下のように定義する。（数６）

（数６）

$g_{condition}/n$

例えば、x_i,y_iが実数値である場合、$x_i>0$、$y_i>0$が条件として考えられる。

【００９５】

この他、類似度と逆の概念として距離(非類似度)が考えられるが、２つのデータ系列の間の距離の定義として以下がある。距離を相関分析に用いる場合は、２つのプロファイルデータ間で距離が小さいほど類似度が高いと考える。

【００９６】

(5) ユークリッド距離

２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nに対して、ユークリッド距離は次式（数７）によって求められる。

【００９７】

【数７】

$$\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

【００９８】

(6) ユークリッド平方距離

２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nに対して、ユークリッド平方距離は次式（数８）によって求められる。

【００９９】

【数８】

$$\sum_{i=1}^{n}(x_i-y_i)^2$$

【０１００】

(7) チェビシェフ距離

２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nに対して、チェビシェフ距離は次式（数９）によって求められる。

（数９）

$\max(|x_i-y_i|)$

ここで|x|はxの絶対値をあらわす。またmax Xはi = 1, 2, …, nについてXの最大値を表す。

【０１０１】

(8) マンハッタン距離

２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nに対して、マンハッタン距離は次式（数１０）によって求められる。

【０１０２】

【数１０】

$$\sum_{i=1}^{n}|x_i - y_i|$$

【０１０３】

(9) マハラノビス距離

２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nに対して、マハラノビス距離は次式（数１１）によって求められる。

【０１０４】

【数１１】

$$\sum_{i=1}^{n}\sum_{j=1}^{n}s^{ij}(x_i - y_i)(x_j - y_j)$$

但し、s^{ij}は、分散共分散行列s^{ij}の逆行列のi,j要素とする。

【０１０５】

(10) ミンコフスキー距離

２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nに対して、ミンコフスキー距離は次式（数１２）によって求められる。

【０１０６】

【数１２】

$$\left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/r}$$

ここでrとpは重みを調整するためのパラメータであり、r≠0の範囲で任意の値を設定可能である。pは個々のデータの差分に与える重みを調節するために用い、rは全体としての距離を調節するために用いる。

【０１０７】

(11) 不一致割合

２つのデータ列 x={x_i},y={y_i}, i=1,2,…,nについて、データx_i,y_iが不一致の場合がc回ある場合、xとyの距離を以下のように定義することが出来る。（数１３）

（数１３）

c/n

また、xとyの類似度を以下のように定義できる。（数１４）

（数１４）

(n-c) / n

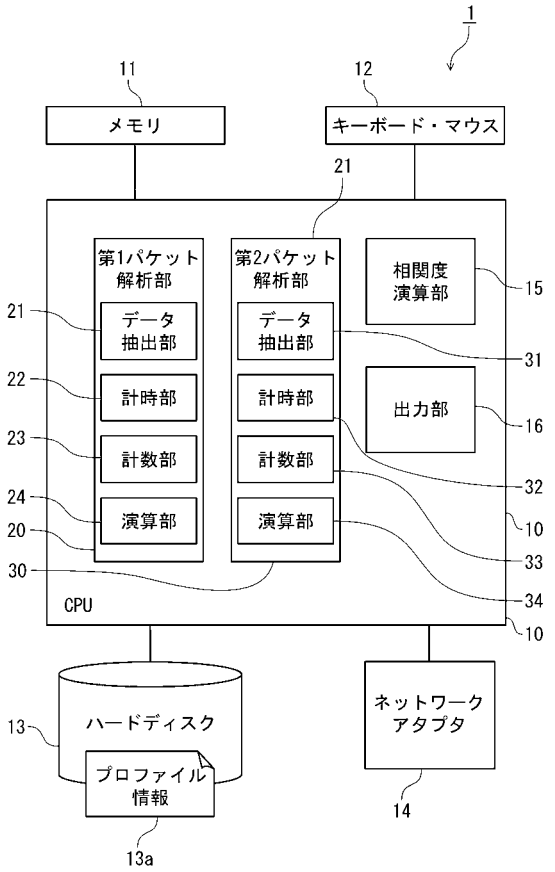
【０１０８】

本発明は以上説述したように、プロファイルデータを用いて、２つのソフトウェアの送信データを比較することに最大の特徴があり、これらの比較には上記に示した様々な相関関係式の他、公知の任意の相関関係式を使用することができる。

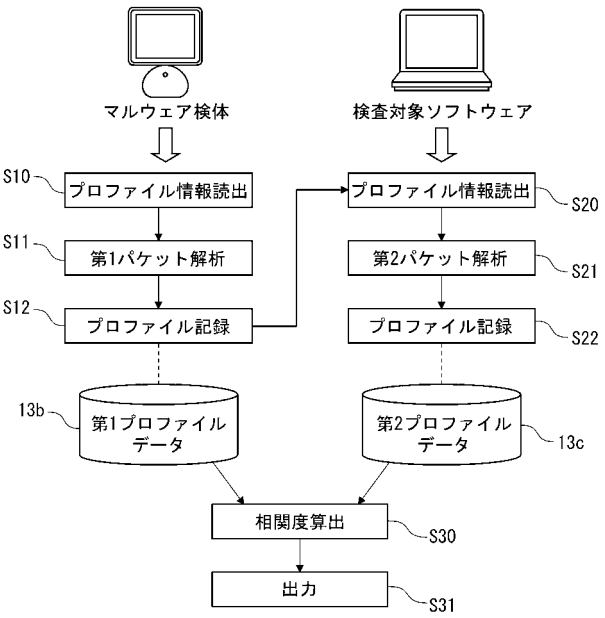
【図面の簡単な説明】

【０１０９】
【図１】本発明に係るデータ類似性検査装置の構成図である。
【図２】本発明に係るデータ類似性検査方法の処理フローチャートである。
【図３】本発明に係るマルウェアの特定システムの構成図である。
【図４】本発明におけるマルウェア動的解析環境を用いたマルウェアの解析方法の説明図である。
【図５】本発明における相関解析部の別構成を示す構成図である。
【符号の説明】
【０１１０】
１　データ類似性検査装置
１０　ＣＰＵ
１１　メモリ
１２　キーボード・マウス
１３　ハードディスク
１３ａ　プロファイル情報
１４　ネットワークアダプタ
１５　相関度演算部
１６　出力部
２０　第１パケット解析部
２１　データ抽出部
２２　計時部
２３　計数部
２４　演算部
３０　第２パケット解析部
３１　データ抽出部
３２　計時部
３３　計数部
３４　演算部

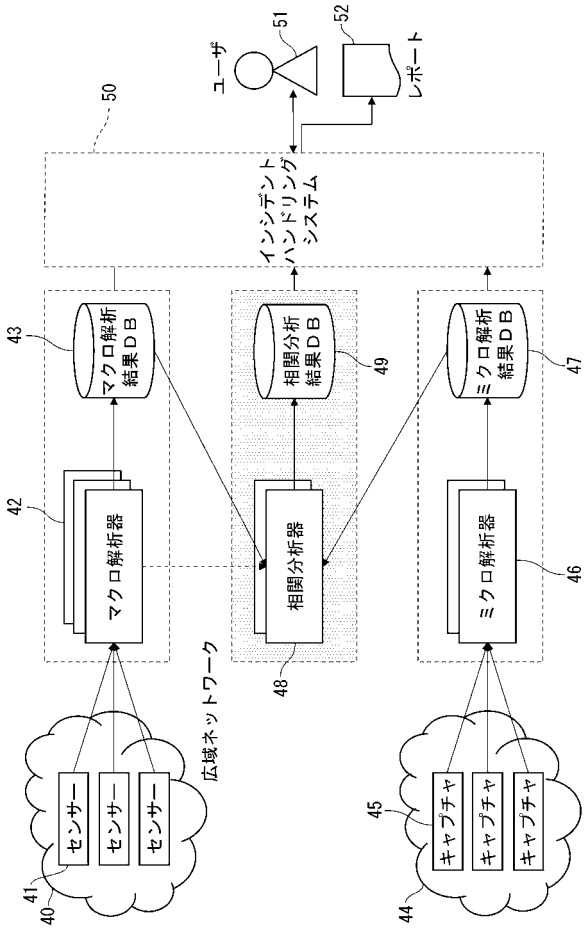
【図１】


1
11
メモリ
12
キーボード・マウス
21
第1パケット解析部
第2パケット解析部
相関度演算部
15
データ抽出部
計時部
計数部
演算部
出力部
16
31
32
10
33
34
CPU
ハードディスク
13
プロファイル情報
13a
ネットワークアダプタ
14

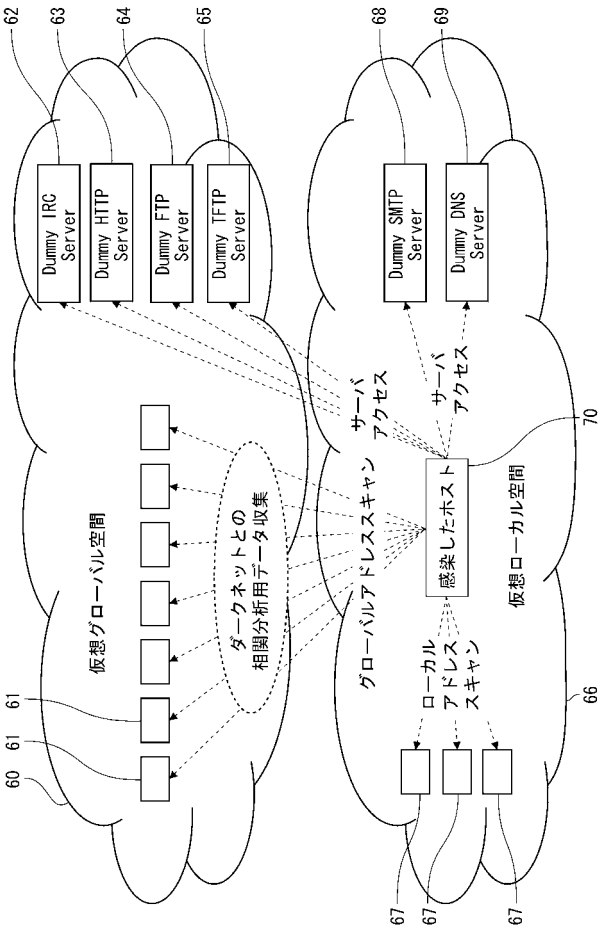
【図２】


マルウェア検体
検査対象ソフトウェア
S10
プロファイル情報読出
S20
S11
第1パケット解析
第2パケット解析
S21
S12
プロファイル記録
S22
13b
第1プロファイルデータ
第2プロファイルデータ
13c
相関度算出
S30
出力
S31

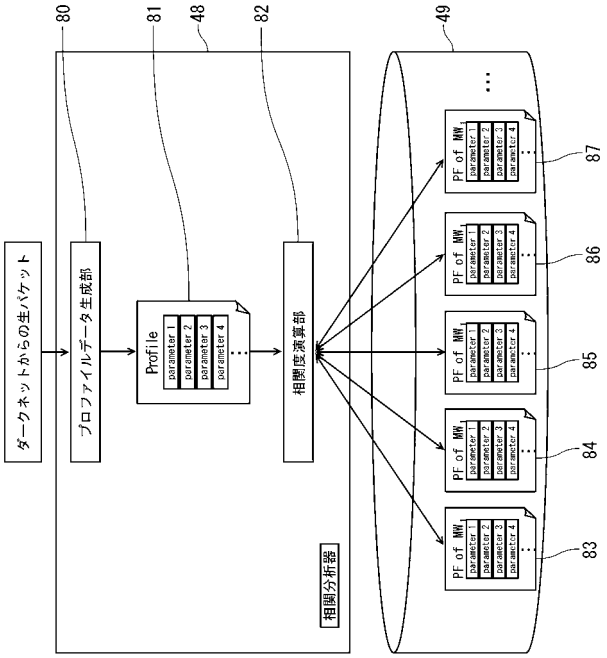
【図３】


ユーザ
51
52
レポート
50
インシデントハンドリングシステム
43
マクロ解析結果ＤＢ
49
相関分析結果ＤＢ
47
ミクロ解析結果ＤＢ
42
マクロ解析器
相関分析器
ミクロ解析器
46
48
広域ネットワーク
センサー
センサー
センサー
41
40
キャプチャ
キャプチャ
キャプチャ
45
44

【図４】


62
63
64
65
68
69
Dummy IRC Server
Dummy HTTP Server
Dummy FTP Server
Dummy TFTP Server
Dummy SMTP Server
Dummy DNS Server
サーバアクセス
サーバアクセス
70
仮想グローバル空間
ダークネットとの相関分析用データ収集
グローバルアドレススキャン
感染したホスト
仮想ローカル空間
ローカルアドレススキャン
61
61
60
66
67
67
67

【図５】


80
81
48
82
49
ダークネットからの生パケット
プロファイルデータ生成部
Profile
parameter 1
parameter 2
parameter 3
parameter 4
相関度演算部
相関分析器
PF of MW1
PF of MW2
PF of MW3
PF of MW4
PF of MWn
parameter 1
parameter 2
parameter 3
parameter 4
83
84
85
86
87

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

Ｆターム(参考) 5B276 FD08
5B285 AA05 AA06 BA01 CA32 CA36 DA04
5K030 GA15 HA08 JA10 KA04 KA07 MA04 MB01 MC08 MD08