

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7013940号

(P7013940)

(45)発行日 令和4年2月1日(2022.2.1)

(24)登録日 令和4年1月24日(2022.1.24)

(51)国際特許分類

F I

G 0 6 N 20/00 (2019.01)

G 0 6 N 20/00 1 6 0

H 0 4 L 43/04 (2022.01)

H 0 4 L 43/04

請求項の数 5 (全16頁)

(21)出願番号	特願2018-33888(P2018-33888)	(73)特許権者	000004226
(22)出願日	平成30年2月27日(2018.2.27)		日本電信電話株式会社
(65)公開番号	特開2019-149051(P2019-149051 A)	(74)代理人	東京都千代田区大手町一丁目5番1号 110002147 特許業務法人酒井国際特許事務所
(43)公開日	令和1年9月5日(2019.9.5)	(72)発明者	荒木 翔平 東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
審査請求日	令和1年8月1日(2019.8.1)	(72)発明者	胡 博 東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
		(72)発明者	神谷 和憲 東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
		(72)発明者	谷川 真樹

最終頁に続く

(54)【発明の名称】 分類装置および分類方法

(57)【特許請求の範囲】

【請求項1】

ダークネットにTCPヘッダのコントロールフラグがSYNであるパケットを送信していたIPアドレスであって、かつ、一定期間中に片方向通信を所定フロー数以上観測できたIPアドレスを送信元とするトラフィックデータを、処理対象の複数のデータとして取得する取得部と、

前記取得部によって取得された複数のデータから、それぞれ異なる種類の特徴量をそれぞれ抽出する抽出部と、

前記抽出部によってそれぞれ抽出された異なる種類の特徴量に基づいて、前記データを所定の単位でそれぞれ分類する分類部と、

前記分類部によって分類された複数の分類結果を用いて、各分類結果の組み合わせが同じパターンの出現回数を集計し、集計した出現回数が所定の閾値を超えたパターンを頻出パターンと特定し、当該頻出パターンに該当するIPアドレスを同じ種類のボットのIPアドレスであると判定する集計部と、

を有することを特徴とする分類装置。

【請求項2】

前記取得部は、複数のトラフィックデータを含む通信ログを取得し、

前記抽出部は、前記それぞれ異なる種類の特徴量として、前記トラフィックデータの送信バイト数、前記トラフィックデータの受信バイト数、前記トラフィックデータの送信フラグ、前記トラフィックデータの受信フラグ、前記トラフィックデータの宛先ポート番号お

よび前記トラフィックデータの送信元ポート番号に関する特徴量のうち、2種類以上の特徴量をそれぞれ抽出することを特徴とする請求項1に記載の分類装置。

【請求項3】

前記分類部は、教師なし機械学習手法を用いて、前記データをそれぞれ分類することを特徴とする請求項1に記載の分類装置。

【請求項4】

前記集計部は、頻出パターンマイニング手法を用いて、前記パターンの出現回数を集計することを特徴とする請求項1に記載の分類装置。

【請求項5】

分類装置によって実行される分類方法であって、

ダークネットにTCPヘッダのコントロールフラグがSYNであるパケットを送信していたIPアドレスであって、かつ、一定期間中に片方向通信を所定フロー数以上観測できたIPアドレスを送信元とするトラフィックデータを、処理対象の複数のデータとして取得する取得工程と、

前記取得工程によって取得された複数のデータから、それぞれ異なる種類の特徴量をそれぞれ抽出する抽出工程と、

前記抽出工程によってそれぞれ抽出された異なる種類の特徴量に基づいて、前記データを所定の単位でそれぞれ分類する分類工程と、

前記分類工程によって分類された複数の分類結果を用いて、各分類結果の組み合わせが同じパターンの出現回数を集計し、集計した出現回数が所定の閾値を超えたパターンを頻出パターンと特定し、当該頻出パターンに該当するIPアドレスを同じ種類のボットのIPアドレスであると判定する集計工程と、

を含んだことを特徴とする分類方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、分類装置および分類方法に関する。

【背景技術】

【0002】

従来、複数のデータから多次元の特徴量をそれぞれ抽出し、抽出した特徴量を用いて各データを分類し、分類結果を解析する技術が知られている。例えば、感染端末で構成されたボットネットを使った攻撃への対策として、通信ログに含まれる複数のトラフィックデータから多次元の特徴量を抽出し、抽出した特徴量を用いてIPアドレスをクラスタリングし、クラスタリングの結果を解析する。

【先行技術文献】

【非特許文献】

【0003】

【文献】Guofei Gu, Roberto Perdisci, Junjie Zhang, and Wenke Lee, “Bot Miner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection”, 17th USENIX Security Symposium p 139-p154, USENIX Association, 2008

【発明の概要】

【発明が解決しようとする課題】

【0004】

しかしながら、従来の技術では、分類結果からデータの特徴を把握することが容易でないという課題があった。例えば、通信ログに含まれるトラフィックデータから多次元の特徴量を用いてIPアドレスをクラスタリングする手法では、抽出する特徴量の次元数が増えるほど、どの特徴に着目してクラスタが形成されたのかを容易に把握することができない場合がある。また、抽出する特徴量の次元数が多いと、適切にクラスタリングされない場合がある。

10

20

30

40

50

【０００５】

本発明は、上記に鑑みてなされたものであって、分類結果からデータの特徴を把握することを容易にすることを目的とする。

【課題を解決するための手段】

【０００６】

上述した課題を解決し、目的を達成するために、本発明の分類装置は、複数のデータを取得する取得部と、前記取得部によって取得された複数のデータから、それぞれ異なる種類の特徴量をそれぞれ抽出する抽出部と、前記抽出部によってそれぞれ抽出された異なる種類の特徴量に基づいて、前記データを所定の単位でそれぞれ分類する分類部と、前記分類部によって分類された複数の分類結果を用いて、各分類結果の組み合わせが同じパターンの出現回数を集計する集計部と、を有することを特徴とする。

10

【０００７】

また、本発明の分類方法は、分類装置によって実行される分類方法であって、複数のデータを取得する取得工程と、前記取得工程によって取得された複数のデータから、それぞれ異なる種類の特徴量をそれぞれ抽出する抽出工程と、前記抽出工程によってそれぞれ抽出された異なる種類の特徴量に基づいて、前記データを所定の単位でそれぞれ分類する分類工程と、前記分類工程によって分類された複数の分類結果を用いて、各分類結果の組み合わせが同じパターンの出現回数を集計する集計工程と、を含んだことを特徴とする。

【発明の効果】

【０００８】

本発明によれば、分類結果からデータの特徴を把握することを容易にすることができるという効果を奏する。

20

【図面の簡単な説明】

【０００９】

【図１】図１は、第一の実施の形態に係る分類装置の構成例を示すブロック図である。

【図２】図２は、通信ログ記憶部に記憶される情報の一例を示す図である。

【図３】図３は、分類結果記憶部に記憶される情報の一例を示す図である。

【図４】図４は、第一の実施の形態に係る分類装置による処理の概要を説明する図である。

【図５】図５は、第一の実施の形態に係る分類装置による通信ログに対する前処理を説明する図である。

30

【図６】図６は、第一の実施の形態に係る分類装置による特徴量抽出処理を説明する図である。

【図７】図７は、第一の実施の形態に係る分類装置による特徴量抽出処理を説明する図である。

【図８】図８は、第一の実施の形態に係る分類装置による特徴量抽出処理を説明する図である。

【図９】図９は、特徴量セットの具体例を示す図である。

【図１０】図１０は、第一の実施の形態に係る分類装置による頻出パターンマイニング手法を用いて同じ種類のボットを分類する処理を説明する図である。

【図１１】図１１は、観測された特徴的なボットの一例を示す図である。

40

【図１２】図１２は、観測された特徴的なボットの一例を示す図である。

【図１３】図１３は、ボットと思われるＩＰアドレスを絞り込む処理を説明する図である。

【図１４】図１４は、第一の実施の形態に係る分類装置による処理を説明するフローチャートである。

【図１５】図１５は、分類プログラムを実行するコンピュータを示す図である。

【発明を実施するための形態】

【００１０】

以下に、本願に係る分類装置および分類方法の実施の形態を図面に基いて詳細に説明する。なお、この実施の形態により本願に係る分類装置および分類方法が限定されるものではない。例えば、以下の実施形態では、解析対象のデータとして通信データを例に説明す

50

るが、これに限定されるものではなく、どのようなデータを解析対象のデータとしてもよい。

【0011】

〔第一の実施の形態〕

以下の実施の形態では、第一の実施の形態に係る分類装置の構成、分類装置における処理の流れを順に説明し、最後に第一の実施の形態による効果を説明する。

【0012】

〔分類装置の構成〕

まず、図1を用いて、分類装置10の構成を説明する。図1は、第一の実施の形態に係る分類装置の構成例を示すブロック図である。図1に示すように、この分類装置10は、入力部11、出力部12、制御部13および記憶部14を有する。以下に分類装置10が有する各部の処理を説明する。

【0013】

入力部11は、キーボードやマウス等の入力デバイスを用いて実現され、操作者による入力操作に対応して、制御部13に対して各種指示情報を入力する。出力部12は、液晶ディスプレイなどの表示装置、プリンター等の印刷装置、情報通信装置、スピーカ等によって実現され、例えば、後述する頻出パターン等を操作者に対して出力する。

【0014】

また、記憶部14は、制御部13による各種処理に必要なデータおよびプログラムを格納するが、特に本発明に密接に関連するものとしては、通信ログ記憶部14aおよび分類結果記憶部14bを有する。例えば、記憶部14は、RAM(Random Access Memory)、フラッシュメモリ(Flash Memory)等の半導体メモリ素子、又は、ハードディスク、光ディスク等の記憶装置などである。

【0015】

通信ログ記憶部14aは、後述する取得部13aによって取得された通信ログを記憶する。例えば、通信ログとして、xFlow、Proxyログ、Firewallログ等の様々な形式の通信ログが用いられる。なお、以下の例では、主に、ネットワークのフロー情報であるxFlowを例として説明する。xFlowは、送信元IPアドレス、宛先IPアドレス、送信元ポート番号、宛先ポート番号、プロトコル、フラグ等を含む。

【0016】

ここで、図2を用いて、通信ログ記憶部14aに記憶される情報の一例を説明する。図2は、通信ログ記憶部に記憶される情報の一例を示す図である。通信ログ記憶部14aは、図2に例示するように、トラフィックデータについて、送信元のIPアドレスを示す「送信元IPアドレス」、宛先のIPアドレスを示す「宛先IPアドレス」、送信元のポート番号を示す「送信元ポート番号」、宛先のポート番号を示す「宛先ポート番号」、通信プロトコルを示す「プロトコル」およびTCPヘッダのコントロールフラグを示す「フラグ」を記憶する。

【0017】

分類結果記憶部14bは、後述する分類部13cによって分類された分類結果を記憶する。例えば、分類結果記憶部14bは、図3に例示するように、IPアドレスごとに、特徴量セットA、特徴量セットB、特徴量セットC・・・それぞれについての分類結果のラベルを記憶する。ここで、特徴量セットとは、予め設定された抽出する特徴量の組である。図3は、分類結果記憶部に記憶される情報の一例を示す図である。

【0018】

図3の例を挙げて説明すると、分類結果記憶部14bは、例えば、IPアドレス「A」について、特徴量セットAのラベル「1」、特徴量セットBのラベル「2」および特徴量セットCのラベル「1」を記憶する。また、分類結果記憶部14bは、IPアドレス「B」についても、特徴量セットAのラベル「1」、特徴量セットBのラベル「2」および特徴量セットCのラベル「1」を記憶する。また、分類結果記憶部14bは、IPアドレス「C」について、特徴量セットAのラベル「2」、特徴量セットBのラベル「1」および特

微量セットCのラベル「3」を記憶する。

【0019】

制御部13は、各種の処理手順などを規定したプログラムおよび所要データを格納するための内部メモリを有し、これらによって種々の処理を実行するが、特に本発明に密接に関連するものとしては、取得部13a、抽出部13b、分類部13cおよび集計部13dを有する。ここで、制御部13は、CPU (Central Processing Unit) やMPU (Micro Processing Unit) などの電子回路やASIC (Application Specific Integrated Circuit) やFPGA (Field Programmable Gate Array) などの集積回路である。

【0020】

取得部13aは、複数のトラフィックデータを含む通信ログを取得する。具体的には、取得部13aは、所定期間における複数のトラフィックデータを含む通信ログを取得すると、通信ログ記憶部14aに取得した通信ログに含まれるトラフィックデータを記憶する。なお、取得部13aは、リアルタイムに通信ログを取得してもよく、所定期間のトラフィックデータ群に区切るバッチ処理を行って通信ログ記憶部14aに格納するようにしてもよい。

10

【0021】

また、取得部13aは、取得した通信ログに対する前処理として、所定の条件を満たすIPアドレスのトラフィックデータのみを処理対象として通信ログ記憶部14aに格納するようにしてもよい。例えば、取得部13aは、所定の条件として、(1) ダークネットにTCPヘッダのコントロールフラグが「SYN」であるパケットを送信していたIPアドレスであって、かつ、(2) 一定期間中に片方向通信を20フロー以上観測できたIPアドレスを送信元とするトラフィックデータのみを通信ログ記憶部14aに格納するようにしてもよい。また、取得部13aは、上記した所定の条件(1)、(2)のうち、いずれか一方でも満たした場合であっても、該当するIPアドレスを送信元とするトラフィックデータを通信ログ記憶部14aに格納するようにしてもよい。

20

【0022】

抽出部13bは、取得部13aによって取得された複数のトラフィックデータから、それぞれ異なる種類の特徴量をそれぞれ抽出する。具体的には、抽出部13bは、予め設定された各特徴量セットで定められた各特徴量をそれぞれ抽出する。

【0023】

例えば、抽出部13bは、それぞれ異なる種類の特徴量として、トラフィックデータの送信バイト数、トラフィックデータの受信バイト数、トラフィックデータの送信フラグ、トラフィックデータの受信フラグ、トラフィックデータの宛先ポート番号およびトラフィックデータの送信元ポート番号に関する特徴量のうち、2種類以上の特徴量をそれぞれ抽出する。

30

【0024】

分類部13cは、抽出部13bによってそれぞれ抽出された異なる種類の特徴量に基づいて、トラフィックデータをIPアドレス単位でそれぞれ分類する。具体的には、分類部13cは、各特徴量セットについて抽出された特徴量の組に基づいて、各IPアドレスをそれぞれクラスタリングし、クラスタリング結果を基にそれぞれラベルを付与して分類結果記憶部14bに格納する。また、分類部13cは、例えば、教師なし機械学習手法を用いて、トラフィックデータをIPアドレス単位でそれぞれ分類するようにしてもよい。

40

【0025】

集計部13dは、分類部13cによって分類された複数の分類結果を用いて、各分類結果の組み合わせが同じパターンの出現回数を集計する。また、集計部13dは、各分類結果の組み合わせが同じパターンの出現回数を集計し、所定の条件を満たすパターンを出力する。具体的には、集計部13dは、分類部13cによって分類された複数の分類結果を用いて、各分類結果の組み合わせが同じパターンの出現回数を集計し、分類結果が同じ組み合わせのパターンが閾値以上出現した場合には、該パターンを出力部12を介して出力する。例えば、集計部13dは、頻出パターンマイニング手法を用いて、パターンの出現回

50

数を集計するようにしてもよい。

【0026】

ここで図4の例を用いて、第一の実施の形態に係る分類装置10による処理の概要を説明する。図4は、第一の実施の形態に係る分類装置による処理の概要を説明する図である。図4に例示するように、分類装置10の取得部13aは、xFlow等の通信ログを取得し、取得した通信ログに対する前処理として、所定の条件を満たすIPアドレスのトラフィックデータを処理対象のトラフィックデータとして選択する(図4の(1)参照)。

【0027】

ここで、図5を用いて、通信ログに対する前処理の一例を説明する。図5は、第一の実施の形態に係る分類装置による通信ログに対する前処理を説明する図である。図5に例示するように、例えば、取得部13aは、所定の条件として、ダークネットにTCPヘッダのコントロールフラグが「SYN」であるパケットを送信していたIPアドレスであって、かつ、一定期間中に片方向通信を20フロー以上観測できたIPアドレスを送信元とするフロー(トラフィックデータ)のみを使用する。

【0028】

図4の説明に戻って、分類装置10の抽出部13bは、予め設定された各特徴量セットA～Cで定められた各特徴量をそれぞれ抽出する。ここで、抽出部13bは、各特徴量セットA～Cで定められた各特徴量を抽出する処理を並行して行ってもよいし、順次行うようにしてもよい。そして、分類部13cは、各特徴量セットA～Cについて抽出された特徴量の組に基づいて、各IPアドレスをそれぞれクラスタリングし、クラスタリング結果を基にそれぞれラベルを付与する(図4の(2)参照)。

【0029】

ここで、図6～図8の例を用いて、分類装置10による特徴量抽出処理を説明する。図6～図8は、第一の実施の形態に係る分類装置による特徴量抽出処理を説明する図である。まず、図6を用いて、特徴量セットAで定められた各特徴量を抽出する処理を説明する。なお、以下では、特徴量セットAでは、抽出対象の特徴量として、通信量に関する特徴量が設定されているものとする。

【0030】

図6に例示するように、抽出部13bは、通信量に関する特徴量として、各送信元IPアドレスについて、1パケットあたりのバイト数の「最大」、「最少」、「平均」、「分散」を抽出する。例えば、抽出部13bは、送信元IP:Aの特徴量として、図6に例示するように、1パケットあたりのバイト数の「最大」として「59.0」バイト、「最少」として「40.0」バイト、「平均」として「41.5」バイト、「分散」として「4.04」バイトを抽出する。

【0031】

そして、分類部13cは、各送信元IPアドレスをそれぞれクラスタリングし、クラスタリング結果を基にそれぞれラベルを付与する。例えば、分類部13cは、各送信元IPアドレスを通信量が少ないものと多いものとでクラスタリングし、通信量が多い送信元IPアドレスにはラベル「2」を付与し、通信量が少ない送信元IPアドレスにはラベル「1」を付与する。

【0032】

次に、図7を用いて特徴量セットBで定められた各特徴量を抽出する処理を説明する。なお、以下では、特徴量セットBでは、抽出対象の特徴量として、TCPフラグに関する特徴量が設定されているものとする。

【0033】

図7に例示するように、抽出部13bは、TCPフラグに関する特徴量として、各送信元IPアドレスについて、送信パケットにおけるTCPフラグ「URG」、「SYN」、「PSH」、「RST」、「ACK」、「FIN」の割合をそれぞれ抽出する。例えば、抽出部13bは、送信元IP:Aの特徴量として、図7に例示するように、「URG」の割合として「0」、「ACK」の割合として「0.08」、「PSH」の割合として「0.

10

20

30

40

50

「03」、「RST」の割合として「0」、「SYN」の割合として「0.91」、「FIN」の割合として「0.01」を抽出する。

【0034】

そして、分類部13cは、各送信元IPアドレスをそれぞれクラスタリングし、クラスタリング結果を基にそれぞれラベルを付与する。例えば、分類部13cは、各送信元IPアドレスについて、TCPフラグの「ACK」が多いものとTCPフラグの「SYN」が多いものとでクラスタリングし、「ACK」が多い送信元IPアドレスにはラベル「1」を付与し、「SYN」が多い送信元IPアドレスにはラベル「2」を付与する。

【0035】

次に、図8を用いて特徴量セットCで定められた各特徴量を抽出する処理を説明する。なお、以下では、特徴量セットCでは、抽出対象の特徴量として、ポートに関する特徴量が設定されているものとする。図8に例示するように、抽出部13bは、ポートに関する特徴量として、クライアント側のIPアドレスについて、パケット送信時に使用された送信元ポート番号を昇順に並べたときの最少(0%)、25%、中間(50%)、75%、最大(100%)の送信元ポート番号をそれぞれ抽出する。

【0036】

例えば、抽出部13bは、クライアント側のIPアドレス：Aの特徴量として、図8に例示するように、「0%」のポート番号として「23」、「25%」のポート番号として「23」、「50%」のポート番号として「23」、「75%」のポート番号として「23」、「100%」のポート番号として「23」を抽出する。これは、パケット送信時に常に宛先ポート番号「23」が使用されていたことを意味する。

【0037】

そして、分類部13cは、各クライアント側のIPアドレスをそれぞれクラスタリングし、クラスタリング結果を基にそれぞれラベルを付与する。例えば、分類部13cは、各クライアント側のIPアドレスについて、使用ポート番号の傾向に応じてクラスタリングする。

【0038】

なお、上記の図6～図8の例では、特徴量セットA～Cの3つを例に挙げて説明したが、これに限定されるものではない。ここで、図9の例を用いて、特徴量セットの具体例を説明する。図9は、特徴量セットの具体例を示す図である。図9に例示するように、例えば、分類装置10は、特徴量セットとして、11つの特徴量セットが予め設定され、このうちの全て特徴量セットについて処理を実行しても、一部を実行してもよい。図9の例では、特徴量セットのセット名として、「送信フラグ」、「受信フラグ」、「送信量」、「受信量」、「送信量2」、「受信量2」、「送信元ポート」、「clientポート」、「宛先ポート」、「serverポート」および「OS」が設定されている。

【0039】

例えば、セット名「送信フラグ」では、抽出対象の特徴量として、TCPフラグに関する特徴量が設定されており、各送信元IPアドレスについて、送信パケットにおけるTCPフラグ「URG」、「SYN」、「PSH」、「RST」、「ACK」、「FIN」の割合をそれぞれ抽出する。また、セット名「OS」では、ダークネットへSYNパケットを送信した送信元のOSの種類を抽出する。

【0040】

また、集約単位として、送信元IPアドレスで集約することを示す「sa」、宛先IPアドレスで集約することを示す「da」、クライアント側のIPアドレスで集約することを示す「cli」、サーバ側のIPアドレスで集約することを示す「srv」のいずれかが設定されている。なお、「cli」の場合にはポート番号が大きい方のIPアドレスをクライアント側のIPアドレスとし、「srv」の場合にはポート番号が小さい方のIPアドレスをサーバ側のIPアドレスとする。また、抽出される特徴量の次元数についても設定されている。例えば、セット名「送信フラグ」では、次元数として「6」が設定されている。

10

20

30

40

50

【0041】

図4の説明に戻って、分類装置10の集計部13dは、頻出パターンマイニング手法を用いて、各分類結果の組み合わせが同じパターンの出現回数を集計し、ボットの分類を行う（図4の（3）参照）。

【0042】

ここで、図10を用いて、頻出パターンマイニング手法を用いて、同じ種類のボットを分類する処理を説明する。図10は、第一の実施の形態に係る分類装置による頻出パターンマイニング手法を用いて同じ種類のボットを分類する処理を説明する図である。図10に例示するように、集計部13dは、頻出パターンマイニング手法を用いて、大量に出現するパターンを検索する。図10の例では、特徴量セットAのラベルが「1」、特徴量セットBのラベルが「2」、特徴量セットCのラベルが「1」であるパターンを頻出パターンとして特定し、頻出パターンに該当するIPアドレス「A」、「B」を同じ種類のボットのIPアドレスであると判定する。

10

【0043】

このボットの特徴として、通信量が少なく、SYNパケットの送信が多く、通信プロトコルがtelnetであることが容易に把握することができる。つまり、分類装置10では、特徴量を意味のある側面で分割し、各側面のラベルの意味を解釈できる。また、分類装置10では、頻出する特徴の組み合わせを抽出し、複数ラベルで各ボットの全体像を把握することができる。

20

【0044】

ここで、図11および図12を用いて、特徴的なボットの例について説明する。図11および図12は、観測された特徴的なボットの一例を示す図である。図11に例示するように、各特徴量セットの頻出パターンから、SYNパケットが多く、送信パケットおよび受信パケットのバイト数が少なく、宛先ポートとして23、2323を使用し、OSがLinux（登録商標）2.4.xという特徴を有するmirai系と思われるボットを観測することができる。

【0045】

また、図12に例示するように、各特徴量セットの頻出パターンから、送信元ポートおよびclientポートに特徴があり、宛先ポートに特徴がないことから、固定の送信元ポート番号を使用するスキャナであって、様々なポートをスキャンする特徴があるボットを観測することができる。

30

【0046】

つまり、分類装置10では、複数の観点から特徴量を分割して抽出を行い、頻出する組み合わせを分類に用いることで、個々の分類結果の意味の解釈と、それらの組み合わせによって得られる最終的な分類結果の意味解釈を容易にした。このため、分類装置10では、何が存在しているかわからないデータのなからから、どのような種別のものが存在するのかを明らかにすることができる。

【0047】

また、第一の実施の形態に係る分類装置10では、ダークネットにTCPヘッダのコントロールフラグが「SYN」であるパケットを送信していたIPアドレスであるか等の条件を用いてボットと思われるIPアドレスを絞り込む。このため、図13に例示するように、第一の実施の形態に係る分類装置10では、未使用IPアドレスの空間であるダークネットの情報を活用し、ボットと思われるIPアドレスを絞り込むことで、処理対象のトラフィックデータを少なくし、処理負荷を軽減することができる。図13は、ボットと思われるIPアドレスを絞り込む処理を説明する図である。

40

【0048】

[分類装置の処理の流れ]

次に、図14を用いて、第一の実施の形態に係る分類装置10の処理の流れを説明する。図14は、第一の実施の形態に係る分類装置による処理を説明するフローチャートである。

【0049】

50

図 1 4 に示すように、分類装置 1 0 の取得部 1 3 a は、所定期間における複数のトラフィックデータを含む通信ログを取得すると（ステップ S 1 0 1 肯定）、取得した通信ログに対する前処理として、所定の条件を満たす IP アドレスのトラフィックデータのみを処理対象として選択する前処理を実行する（ステップ S 1 0 2）。

【 0 0 5 0 】

そして、抽出部 1 3 b は、取得部 1 3 a によって取得された複数のトラフィックデータから、それぞれ異なる種類の特徴量をそれぞれ抽出する（ステップ S 1 0 3）。具体的には、抽出部 1 3 b は、予め設定された各特徴量セットで定められた各特徴量をそれぞれ抽出する。

【 0 0 5 1 】

続いて、分類部 1 3 c は、各特徴量セットについて抽出された特徴量の組に基づいて、各 IP アドレスをそれぞれクラスタリングし（ステップ S 1 0 4）、クラスタリング結果を基にそれぞれラベルを付与する（ステップ S 1 0 5）。

【 0 0 5 2 】

そして、集計部 1 3 d は、分類部 1 3 c によって分類された複数の分類結果を用いて、同一のラベルの組み合わせの出現回数を集計し（ステップ S 1 0 6）、頻出するラベルの組み合わせを同じグループのデータとして出力部 1 2 を介して出力する（ステップ S 1 0 7）。

【 0 0 5 3 】

〔第一の実施の形態の効果〕

このように、第一の実施の形態に係る分類装置 1 0 は、複数のトラフィックデータを含む通信ログを取得し、複数のトラフィックデータから、それぞれ異なる種類の特徴量をそれぞれ抽出する。そして、分類装置 1 0 は、それぞれ抽出した異なる種類の特徴量に基づいて、トラフィックデータを IP アドレス単位でそれぞれ分類し、複数の分類結果を用いて、各分類結果の組み合わせが同じパターンの出現回数を集計する。このため、分類装置 1 0 は、分類結果からデータの特徴を把握することを容易にすることが可能である。

【 0 0 5 4 】

このように、第一の実施の形態に係る分類装置 1 0 では、少数の特徴量に分割して分類を行う。この分割は、例えばネットワークトラフィックデータであればパケットサイズの統計値、TCP フラグの割合のように、それぞれ異なる観点から特徴量の抽出を行い、それぞれ分類を行う。このように分類を行うことで、入力した特徴量に対応した分類結果を得ることができ、結果に対する意味の解釈性を向上させる。

【 0 0 5 5 】

例えば、分類装置 1 0 では、ネットワークトラフィックデータを入力情報として分類を行う際に、特徴量を TCP のフラグとパケットサイズの観点で分割して分類を行うとする。その場合、TCP フラグの分類では、SYN が多いや ACK が多いといった分類結果が得られ、パケットサイズの分類では送信量が多い、少ないといったような結果が得られると考えられる。このように、観点によって特徴量を分割し、少数の特徴量を入力情報として使用して分類を行うことによって、分類結果が簡単になり、なぜそうなったかの結果の解釈が容易となる。

【 0 0 5 6 】

また、分類装置 1 0 では、特徴量分割によって得られた分類結果を集計し、分類結果が同じ組み合わせのデータが閾値 θ 回以上出現したものを出力し、同じグループのデータとする。また、分類装置 1 0 では、それぞれの異なる分類結果を集計し、特定のパターンが多くのデータで出現した場合は、偶然同じ分類結果の組み合わせを得たのではなく、そのグループに共通した特徴であるとみなすことができる。この分類結果は解釈のしやすい結果であり、その結果を組み合わせることによって全体的な特徴を把握することが可能となる。

【 0 0 5 7 】

例えば、TCP フラグの分類の例で、SYN が少ない、送信量が少ないと言った組み合わせが抽出された場合には、この特徴を組み合わせるとこれはスキャンを行うボットである

10

20

30

40

50

と言った解釈をすることができる。また、分割数を増やし、特徴量抽出の観点を増やすことにより、より詳細に分類の特徴を把握することができる。

【0058】

〔システム構成等〕

また、図示した各装置の各構成要素は機能概念的なものであり、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、各装置の分散・統合の具体的形態は図示のものに限られず、その全部または一部を、各種の負荷や使用状況などに応じて、任意の単位で機能的または物理的に分散・統合して構成することができる。さらに、各装置にて行われる各処理機能は、その全部または任意の一部が、CPUおよび当該CPUにて解析実行されるプログラムにて実現され、あるいは、ワイヤードロジックによるハードウェアとして実現され得る。

10

【0059】

また、本実施の形態において説明した各処理のうち、自動的に行われるものとして説明した処理の全部または一部を手動的に行うこともでき、あるいは、手動的におこなわれるものとして説明した処理の全部または一部を公知の方法で自動的に行うこともできる。この他、上記文書中や図面中で示した処理手順、制御手順、具体的名称、各種のデータやパラメータを含む情報については、特記する場合を除いて任意に変更することができる。

【0060】

〔プログラム〕

図15は、分類プログラムを実行するコンピュータを示す図である。コンピュータ1000は、例えば、メモリ1010、CPU1020を有する。また、コンピュータ1000は、ハードディスクドライブインタフェース1030、ディスクドライブインタフェース1040、シリアルポートインタフェース1050、ビデオアダプタ1060、ネットワークインタフェース1070を有する。これらの各部は、バス1080によって接続される。

20

【0061】

メモリ1010は、ROM (Read Only Memory) 1011及びRAM1012を含む。ROM1011は、例えば、BIOS (Basic Input Output System) 等のブートプログラムを記憶する。ハードディスクドライブインタフェース1030は、ハードディスクドライブ1090に接続される。ディスクドライブインタフェース1040は、ディスクドライブ1100に接続される。例えば磁気ディスクや光ディスク等の着脱可能な記憶媒体が、ディスクドライブ1100に挿入される。シリアルポートインタフェース1050は、例えばマウス1051、キーボード1052に接続される。ビデオアダプタ1060は、例えばディスプレイ1061に接続される。

30

【0062】

ハードディスクドライブ1090は、例えば、OS1091、アプリケーションプログラム1092、プログラムモジュール1093、プログラムデータ1094を記憶する。すなわち、分類装置10の各処理を規定するプログラムは、コンピュータにより実行可能なコードが記述されたプログラムモジュール1093として実装される。プログラムモジュール1093は、例えばハードディスクドライブ1090に記憶される。例えば、装置における機能構成と同様の処理を実行するためのプログラムモジュール1093が、ハードディスクドライブ1090に記憶される。なお、ハードディスクドライブ1090は、SSD (Solid State Drive) により代替されてもよい。

40

【0063】

また、上述した実施の形態の処理で用いられるデータは、プログラムデータ1094として、例えばメモリ1010やハードディスクドライブ1090に記憶される。そして、CPU1020が、メモリ1010やハードディスクドライブ1090に記憶されたプログラムモジュール1093やプログラムデータ1094を必要に応じてRAM1012に読み出して実行する。

【0064】

50

なお、プログラムモジュール１０９３やプログラムデータ１０９４は、ハードディスクドライブ１０９０に記憶される場合に限らず、例えば着脱可能な記憶媒体に記憶され、ディスクドライブ１１００等を介してＣＰＵ１０２０によって読み出されてもよい。あるいは、プログラムモジュール１０９３及びプログラムデータ１０９４は、ネットワーク、ＷＡＮを介して接続された他のコンピュータに記憶されてもよい。そして、プログラムモジュール１０９３及びプログラムデータ１０９４は、他のコンピュータから、ネットワークインタフェース１０７０を介してＣＰＵ１０２０によって読み出されてもよい。

【符号の説明】

【００６５】

１０ 分類装置

10

１１ 入力部

１２ 出力部

１３ 制御部

１３ a 取得部

１３ b 抽出部

１３ c 分類部

１３ d 集計部

１４ 記憶部

１４ a 通信ログ記憶部

１４ b 分類結果記憶部

20

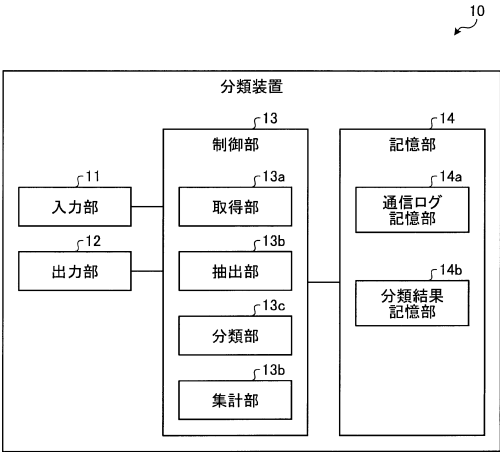
30

40

50

【図面】

【図 1】



【図 2】

送信元 IPアドレス	宛先 IPアドレス	送信元 ポート番号	宛先ポート 番号	プロトコル	フラグ	...
A	B	51001	80	TCP	SYN	...
A	B	51002	80	TCP	FIN	...
...	...	...	...	...	...	...

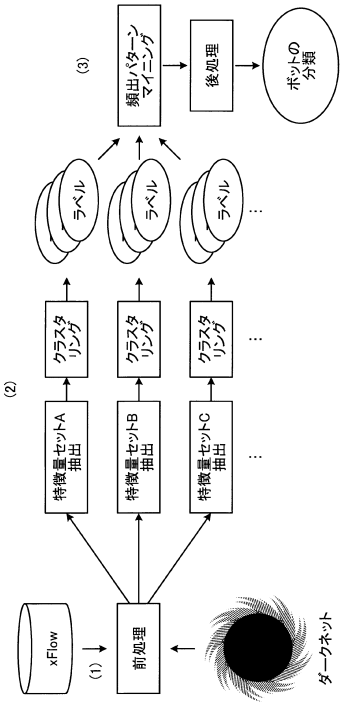
10

20

【図 3】

IP アドレス	特徴量 セットA	特徴量 セットB	特徴量 セットC	...
A	1	2	1	...
B	1	2	1	...
C	2	1	3	...
...	...	...	...	...

【図 4】



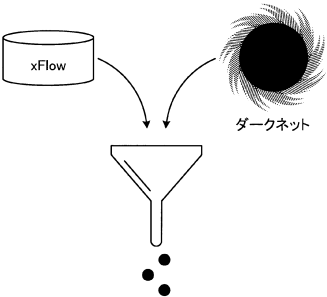
30

40

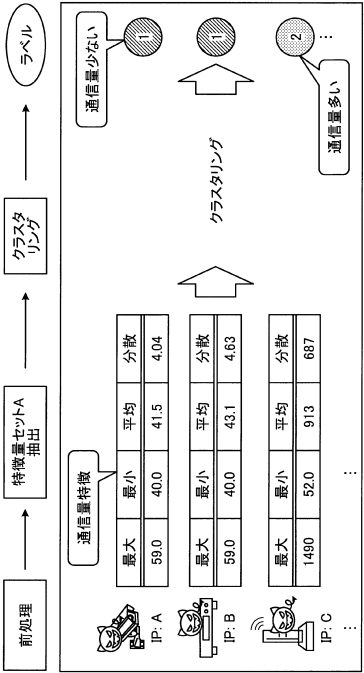
50

【図 5】

- ・次の2つの条件を満たすアドレスのフローを使用
- ・ダークネットにSYNを送信していたIPアドレス
- ・一定期間中に片方向20フロー以上観測できたIPアドレス



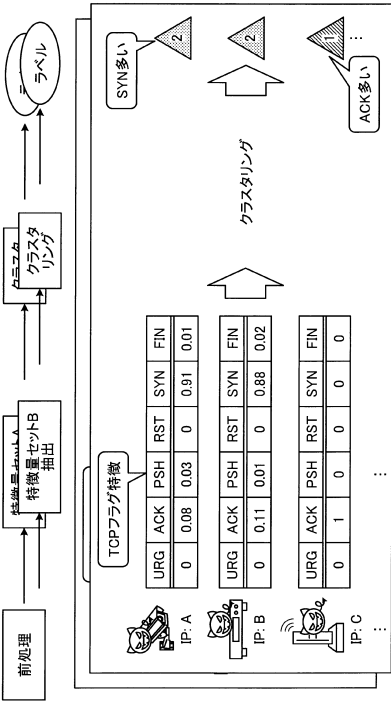
【図 6】



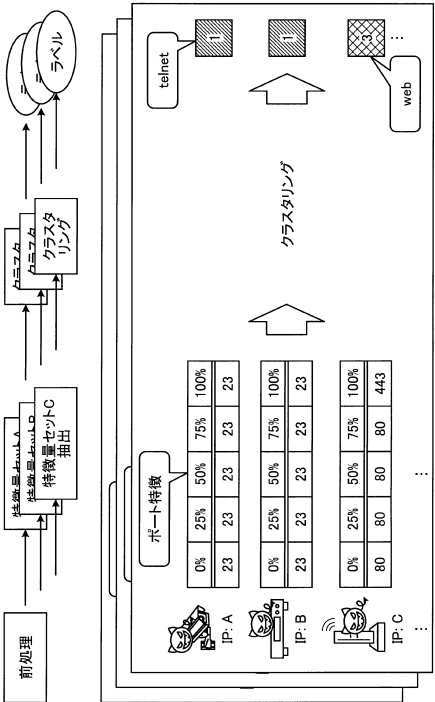
10

20

【図 7】



【図 8】



30

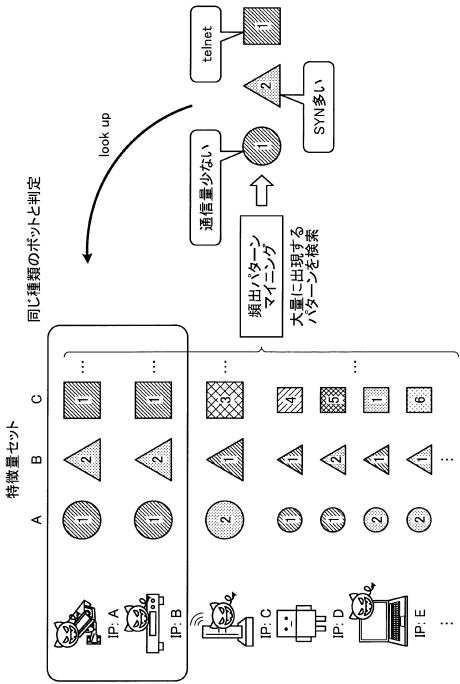
40

50

【図 9】

セット名	特徴量	集約単位	次元数
送信フラグ	TOPフラグ [URG, SYN, PSH, RST, ACK, FIN]の割合	sa	6
受信フラグ		da	6
送信量	1パケットあたりのバイト数の [最大, 最小, 平均, 分散]	sa	4
受信量		da	4
送信量2		sa	4
受信量2		da	4
送信元ポート	ACKがついた1パケットあたりのバイト数の [最大, 最小, 平均, 分散]	sa	4
宛先ポート	送信元ポート番号を昇順に並べた時の [min, 25%, mid, 75%, max]	cli	5
宛先ポート	宛先ポート番号を昇順に並べた時の [min, 25%, mid, 75%, max]	cli	5
serverポート		srv	5
OS	データネットワークでのOS分類	-	22

【図 10】



10

20

【図 11】

特徴量セット	頻出パターン	解釈結果
送信フラグ	○	SYNが多い
送信量	○	少ない
送信量2	○	少ない
受信フラグ	×	-
受信量	○	少ない
受信量2	○	少ない
送信元ポート	×	-
宛先ポート	○	23, 2323を使用
clientポート	×	-
serverポート	×	-
OS	○	Linux 2.4.x

【図 12】

特徴量セット	頻出パターン	解釈結果
送信フラグ	×	-
送信量	○	少ない、稀にデータを送信
送信量2	○	少ない、稀にデータを送信
受信フラグ	×	-
受信量	○	多い
受信量2	○	多い
送信元ポート	○	特徴あり
宛先ポート	×	-
clientポート	○	特徴あり
serverポート	×	-
OS	○	Windows 7or8

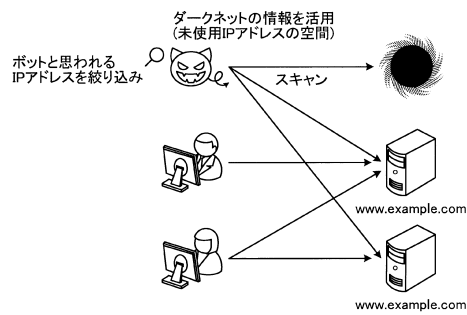
宛先ポートは特徴がない
=様々なポートをスキャン?

30

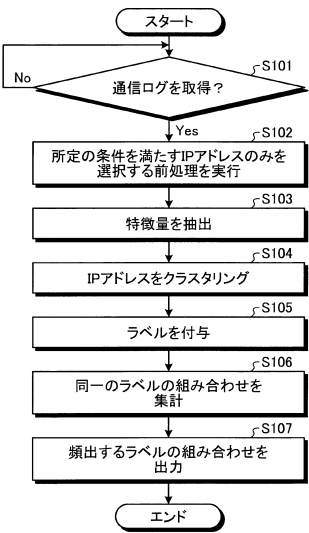
40

50

【図 1 3】



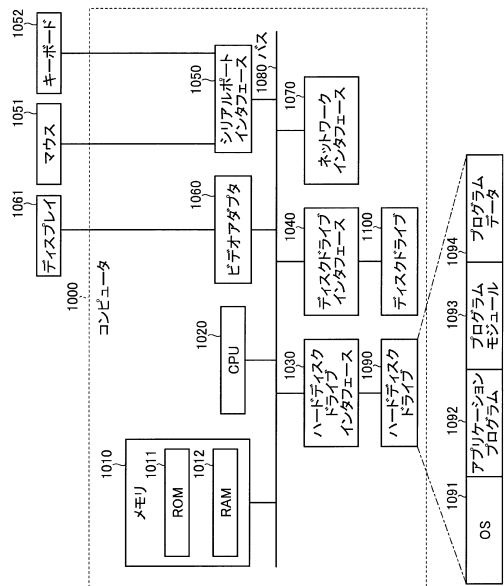
【図 1 4】



10

20

【図 1 5】



30

40

50

フロントページの続き

東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内

審査官 久保 光宏

(56)参考文献

Julan YI, "A Novel Research on Real-Time Intrusion Detection Technology Based on Data Mining", Proceedings of the 2015 2nd International Workshop on Materials Engineering and Computer Sciences (IWMECS 2015), 2015年10月, Pages 881-885, [online], [retrieved on 2019.04.10], Retrieved from the Internet: <URL: <https://www.atlantis-pess.com/proceedings/iwmeecs-15/25840718>>, ISBN: 978-94-6252-114-8, ISSN: 2352-538X, <DOI: <https://doi.org/10.2991/iwmeecs-15.2015.172>> .

Guofei GU, et al., "BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection", Proceedings of the 17th USENIX Security Symposium, 2008年07月31日, Pages 139-154, [online], [retrieved on 2019.04.10], Retrieved from the Internet: <URL: <https://www.usenix.org/conference/17th-usenix-security-symposium/presentation/botminer-clustering-analysis-network-traffic->> .

Tao BAN, "Data Mining Applied to Darknet Traffic Analysis", Journal of the National Institute of Information and Communications Technology, Vol.63, No.2, 2017年03月, Pages 45-54, [online], [retrieved on 2019.04.10], Retrieved from the Internet: <URL: <http://www.nict.go.jp/en/data/journal/index.html>> and <URL: <http://www.nict.go.jp/publication/shuppan/kihou-journal/journal-vol63no2.html>>, ISSN: 2433-6017.

村井 健祥 (外4名), 「PRACTICE ダークネットトラフィックへの時系列解析とスキャン特徴量によるスキャンツール等の分類」, 電子情報通信学会 2016年総合大会講演論文集, 分冊: 情報・システム2, 日本, 一般社団法人電子情報通信学会, 2016年03月01日, 第SS-14~SS-15頁, ISSN: 1349-1369.

デジタルアドバンテージ, 「基礎から学ぶWindowsネットワーク 第15回 信頼性のある通信を実現するTCPプロトコル (2)」, @IT (atmarkIT), [online], ITmedia Inc., 2004年01月29日, 2/3ページ, [令和3年5月18日検索], インターネット, <URL: https://www.atmarkit.co.jp/ait/articles/0401/29/news080_2.html> .

バン ヒョチャン (外1名), 「DDoS検出のためのトラフィック分析方法」, FIT2004 第3回情報科学技術フォーラム 一般講演論文集, 第4分冊, 日本, 社団法人電子情報通信学会, 2004年08月20日, 第417~418頁.

パロアルトネットワークス合同会社監修, 「Palo Alto Networks構築実践ガイド 次世代ファイアウォールの機能を徹底活用」, 初版, 日本, 株式会社技術評論社, 2015年08月25日, 第267~269頁, ISBN: 978-4-7741-7521-8.

三原 元 (外1名), 「数量化理論とCCCDATASET2009を利用したボットネットのC&Cサーバ特定手法の提案と評価」, コンピュータセキュリティシンポジウム2009論文集, 日本, 社団法人情報処理学会, 2009年10月19日, 第547~552頁.

(58)調査した分野 (Int.Cl., DB名)

G 0 6 N 3 / 0 0 - 9 9 / 0 0

H 0 4 L 4 3 / 0 0 - 4 3 / 5 5

G 0 6 F 2 1 / 0 0 - 2 1 / 8 8

G 0 6 F 1 6 / 0 0 - 1 6 / 9 5 8

C S D B (日本国特許庁)

I E E E X p l o r e (I E E E)