(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2021年1月21日(21.01.2021)

WIPO | PCT

(10) 国際公開番号
WO 2021/009887 A1

(51) 国際特許分類:
H04L 12/70 (2013.01) *G06N 20/00* (2019.01)
G06F 21/55 (2013.01)

(21) 国際出願番号： PCT/JP2019/028178

(22) 国際出願日： 2019年7月17日(17.07.2019)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人: 日本電信電話株式会社 (**NIPPON TELEGRAPH AND TELEPHONE CORPORATION**) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目５番１号 Tokyo (JP).

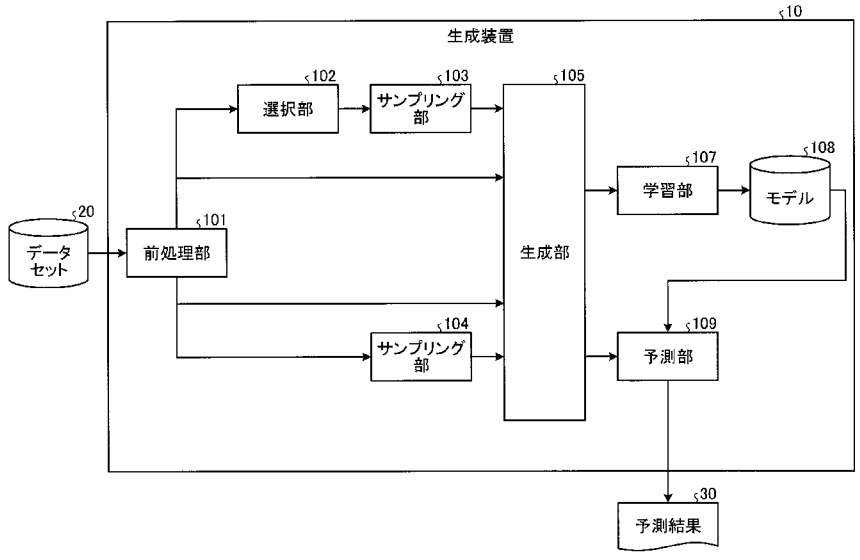
(72) 発明者: 胡 博(**HU, Bo**); 〒1808585 東京都武蔵野市緑町３丁目９－１１ ＮＴＴ 知的財産センタ内 Tokyo (JP). 神谷 和憲(**KAMIYA, Kazunori**); 〒1808585 東京都武蔵野市緑町３丁目９－１１ ＮＴＴ 知的財産センタ内 Tokyo (JP).

(74) 代理人: 特許業務法人酒井国際特許事務所 (**SAKAI INTERNATIONAL PATENT OFFICE**); 〒1000013 東京都千代田区霞が関３丁目８番１号 虎の門三井ビルディング Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH,

(54) **Title:** GENERATION DEVICE, GENERATION METHOD AND GENERATION PROGRAM

(54) 発明の名称： 生成装置、生成方法及び生成プログラム



10 Generation device
20 Data set
30 Prediction result
101 Pre-processing unit
102 Selection unit
103, 104 Sampling unit
105 Generation unit
107 Learning unit
108 Model
109 Prediction unit

(57) **Abstract:** A generation device (10) aggregates a plurality of pieces of traffic data for each prescribed target. Furthermore, the generation device (10) performs sampling of the traffic data of a target in which the number of pieces of aggregated traffic data exceeds a threshold value. Furthermore, the generation device (10) generates a feature vector that expresses a feature of the aggregated traffic data of the target that has not been subjected to sampling, and generates a feature vector that expresses a feature of the sampled traffic data of the target that has been subjected to sampling.

WO 2021/009887 A1

[続葉有]

KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

一 国際調査報告(条約第21条(3))

(57) 要約:生成装置(１０)は、複数のトラヒックデータを所定のターゲットごとに集約する。また、生成装置(１０)は、集約されたトラヒックデータの数が閾値を超えているターゲットのトラヒックデータをサンプリングする。また、生成装置(１０)は、サンプリングが行われなかったターゲットについては、集約されたトラヒックデータの特徴を表す特徴ベクトルを生成し、サンプリングが行われたターゲットについては、サンプリングされたトラヒックデータの特徴を表す特徴ベクトルを生成する。

明　細　書

発明の名称 ： 生成装置、生成方法及び生成プログラム

技術分野

[0001] 本発明は、生成装置、生成方法及び生成プログラムに関する。

背景技術

[0002] 従来、教師あり学習をNWフローに適用する技術が知られている。また、その際、データのサンプリングを行うことで、処理量を削減することが知られている。

先行技術文献

非特許文献

[0003] 非特許文献1：Walter de Donato, Antonio Pescape, and Alberto Dainotti, ”Traffic Identification Engine: An Open Platform for Traffic Classification”, IEEE Network March/April 2014

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、従来の技術には、モデルの精度が低下する場合があるという問題がある。例えば、NWフローから得られたトラヒックデータを基に、ホストごとに特徴量を生成する場合を考える。この場合、所定のサンプリングレートでサンプリングを行うと、もともとトラヒックデータが少ないホストについて、十分なトラヒックデータが得られず、モデルの精度が低下することが考えられる。

課題を解決するための手段

[0005] 上述した課題を解決し、目的を達成するために、生成装置は、複数のトラヒックデータを所定のターゲットごとに集約する前処理部と、前記前処理部によって集約されたトラヒックデータの数が閾値を超えているターゲットのトラヒックデータをサンプリングするサンプリング部と、前記サンプリング部によってサンプリングが行われなかったターゲットについては、集約され

たトラヒックデータの特徴を表す特徴ベクトルを生成し、前記サンプリング部によってサンプリングが行われたターゲットについては、サンプリングされたトラヒックデータの特徴を表す特徴ベクトルを生成する生成部と、を有することを特徴とする。

発明の効果

[0006] 本発明によれば、モデルの精度の低下を抑止することができる。

図面の簡単な説明

[0007] [図1]図１は、第１の実施形態に係る生成装置の構成例を示す図である。

[図2]図２は、トラヒックデータの一例を示す図である。

[図3]図３は、非選択的サンプリングについて説明するための図である。

[図4]図４は、選択的サンプリングについて説明するための図である。

[図5]図５は、第１の実施形態に係る生成装置の処理の流れを示すフローチャートである。

[図6]図６は、特徴量生成処理を並列に行う場合の構成例を示す図である。

[図7]図７は、前処理を並列に行う場合の構成例を示す図である。

[図8]図８は、生成プログラムを実行するコンピュータの一例を示す図である。

発明を実施するための形態

[0008] 以下に、本願に係る生成装置、生成方法及び生成プログラムの実施形態を図面に基づいて詳細に説明する。なお、本発明は、以下に説明する実施形態により限定されるものではない。

[0009] ［第１の実施形態の構成］

まず、図１を用いて、第１の実施形態に係る生成装置の構成について説明する。図１は、第１の実施形態に係る生成装置の構成の一例を示す図である。図１に示すように、生成装置１０は、データセット２０の入力を受け付け、予測結果３０を出力する。

[0010] 実施形態において、生成装置１０は、特徴ベクトル（特徴量）を生成し、生成した特徴ベクトルを使った学習及び予測を行うことができるものとして

説明する。一方で、生成装置１０は、少なくとも特徴ベクトルの生成ができればよい。例えば、他の装置が、生成装置１０によって生成された特徴ベクトルを受け取り、当該受け取った特徴ベクトルを使って学習及び予測を行うようにしてもよい。

[0011] データセット２０は、複数のトラヒックデータのセットである。例えば、トラヒックデータはNWフローから得られる情報である。ただし、トラヒックデータはNWフローから得られる情報に限られず、ネットワークに関する情報に基づくものであればどのようなデータであってもよい。また、トラヒックフローには、あらかじめクラスが設定されている場合がある。設定されたクラスは、モデルの学習時にはラベルとして用いられる。

[0012] 図２は、トラヒックデータの一例を示す図である。例えば、図２に示すように、トラヒックデータは、NWフローごとのタイムスタンプ（ts）、送信元IPアドレス（sip）、宛先IPアドレス（dip）、送信元ポート番号（sp）、宛先ポート番号（dp）、プロトコル（pr）、パケット数（pkt）、バイト数（byt）を含む。また、各トラヒックデータには、悪性（Malicias）又は良性（Benign）のいずれかクラスが設定されている場合がある。

[0013] なお、図２のトラヒックデータは、ターゲットr_iごとに集約されたNWフロー（Bag r_i of network flows）である。なお、集約する処理については後に説明する。また、この場合のターゲットは、送信元又は宛先のホストである。

[0014] 生成装置１０の各部について説明する。図１に示すように、生成装置１０は、前処理部１０１、選択部１０２、サンプリング部１０３、サンプリング部１０４、生成部１０５、学習部１０７、モデル１０８、予測部１０９を有する。

[0015] モデル１０８は、トラヒックデータから生成された特徴ベクトルを基に、トラヒックが悪性であるか良性であるかを予測するためのモデルである。また、生成装置１０は、クラスが既知のトラヒックデータを使い、モデル１０８の学習を行うことができる。

[0016] 前処理部１０１は、データセット２０を、学習用のデータと予測用のデータとに分けることができる。例えば、前処理部１０１は、データセット２０に含まれるトラヒックデータのうち、ラベルが付与されているものを学習用のデータとし、ラベルが未付与のものを予測用のデータとすることができる。なお、学習を行うか予測を行うかはあらかじめ決められていてもよく、その場合、前処理部１０１はデータを分ける必要はない。

[0017] また、前処理部１０１は、データセット２０のトラヒックデータを、所定のターゲットごとに集約する。ターゲットは、例えばホスト又はフローである。例えば、ターゲットがホストである場合、前処理部１０１は、送信元IPアドレス又は宛先IPアドレスが共通するトラヒックデータを同じ集合（Bag）として集約する。

[0018] 例えば、図２の例では、前処理部１０１は、送信元IPアドレス又は宛先IPアドレスがm1であるトラヒックデータを、Bag r_{m1}として集約する。また、前処理部１０１は、送信元IPアドレス又は宛先IPアドレスがb1であるトラヒックデータを、Bag r_{b1}として集約する。

[0019] 選択部１０２は、ターゲットを選択する。例えば、ターゲットがホストである場合、選択部１０２は、サーバm1、サーバm2、…、サーバb1、サーバb2、…のような順序でサーバを選択していく。また、ターゲットがフローである場合、選択部１０２は、5-tupleの値の組み合わせを順に選択していく。

[0020] サンプリング部１０３は、前処理部１０１によって集約されたトラヒックデータの数が閾値を超えているターゲットのトラヒックデータをサンプリングする。つまり、サンプリング部１０３は、選択部１０２によって選択されたターゲットのトラヒックデータが閾値を超えていればサンプリングを行う。逆に、サンプリング部１０３は、選択部１０２によって選択されたターゲットのトラヒックデータが閾値以下であれば、当該ターゲットについてはサンプリングを行わない。 なお、サンプリング部１０３は、サンプリングレートに従ってランダムにサンプリングを行ってもよいし、所定の基準でソートしたトラヒックデータの先頭から所定の数のトラヒックデータをサンプリ

ングしてもよい。

[0021] このように、本実施形態のように、トラヒックデータの数に応じて、サンプリングを行うか否かを選択的に決定する方法を選択的サンプリングと呼ぶ。一方、全てのターゲットに対してサンプリングを行う方法を非選択的サンプリングと呼ぶ。

[0022] まず、図３を用いて非選択的サンプリングについて説明する。図３は、非選択的サンプリングについて説明するための図である。図３に示すように、例えば、ホスト「Server b1」のトラヒックデータが2,000件であり、ホスト「Server b2」のトラヒックデータが200件であり、ホスト「Server m1」のトラヒックデータが20件であるものとする。

[0023] ここで、サンプリングレートを1.5%に設定しランダムサンプリングを行い、ホスト「Server b1」のトラヒックデータから30件、ホスト「Server b2」のトラヒックデータから3件、ホスト「Server m1」のトラヒックデータから0件がサンプリングされたものとする。

[0024] この場合、ホスト「Server b1」については、モデルの精度を向上させるために十分な情報を持つ特徴量が得られると考えられる。一方で、ホスト「Server b2」については、ホスト「Server b1」と比べてデータ数が非常に少ないため、モデルの精度を向上させることが難しくなることが考えられる。さらに、ホスト「Server m1」については、データが欠落し、特徴量が得られない。このように、非選択的サンプリングには、モデルの精度向上が困難になる場合があるという問題がある。選択的サンプリングによれば、このような問題が解決される。

[0025] 図４は、選択的サンプリングについて説明するための図である。図４の例では、閾値が20であるものとする。また、ホストごとの入力されるトラヒックデータの数は、図３の場合と同じであるものとする。ここでは、サンプリング部１０３は、閾値と同数のトラヒックデータをサンプリングするものとする。なお、サンプリング部１０３は、閾値以上かつホストに集約されたトラヒックデータの数未満の所定の数のトラヒックデータをサンプリングして

もよい。

[0026] ホスト「Server b1」のデータ件数は閾値を超えているため、サンプリング部１０３は、ホスト「Server b1」のトラヒックデータから20件をサンプリングする。また、ホスト「Server b2」のデータ件数は閾値を超えているため、サンプリング部１０３は、ホスト「Server b2」のトラヒックデータから20件をサンプリングする。ホスト「Server m1」のデータ件数は閾値を超えていないため、サンプリング部１０３は、ホスト「Server m1」のトラヒックデータからサンプリングを行わない。これは、ホスト「Server m1」のトラヒックデータの全件が特徴量生成の対象となることを意味する。

[0027] 生成部１０５は、サンプリング部１０３によってサンプリングが行われなかったホストについては、集約されたトラヒックデータの特徴を表す特徴ベクトルを生成する。また、生成部１０５は、サンプリング部１０３によってサンプリングが行われたホストについては、サンプリングされたトラヒックデータの特徴を表す特徴ベクトルを生成する。

[0028] 図４の例では、生成部１０５は、ホスト「Server b1」及び「Server b2」については、サンプリング部１０４によってサンプリングされたトラヒックデータから特徴ベクトルを生成する。一方で、生成部１０５は、ホスト「Server m1」については、サンプリングが行われないため、サンプリング前のトラヒックデータから特徴ベクトルを生成する。

[0029] なお、図１では、学習と予測を分けて説明するため、サンプリング部１０３とサンプリング部１０４を異なるブロックで表しているが、サンプリング部１０４はサンプリング部１０３と同じ処理を行う。

[0030] 学習部１０７は、特徴ベクトルを用いて、モデル１０８の学習を行う。この場合、図２に示すように、ホストごとの悪性又は良性を表すラベルは既知であるものとする。予測部１０９は、トラヒックデータの特徴ベクトルを学習済みのモデル１０８に入力し、当該トラヒックデータが悪性であるか良性であるかを示すラベルを予測する。

[0031] ［第１の実施形態の処理］

図５を用いて、第１の実施形態の生成装置１０の処理の流れを説明する。図５は、第１の実施形態に係る生成装置の処理の流れを示すフローチャートである。まず、図５に示すように、生成装置１０は、複数のトラヒックデータを含むデータセットの入力を受け付ける（ステップＳ１１）。次に、生成装置１０は、トラヒックデータをホストごとに集約する（ステップＳ１２）。

[0032] ここで、生成装置１０は、未選択のホストから1つを選択する（ステップＳ１３）。生成装置１０は、選択したホストについて、集約されたトラヒックデータの数が閾値を超えているか否かを判定する（ステップＳ１４）。

[0033] 生成装置１０は、トラヒックデータの数が閾値を超えていると判定した場合（ステップＳ１４、Yes）、当該ホストについてトラヒックデータのサンプリングを行う（ステップＳ１５）。一方、生成装置１０は、トラヒックデータの数が閾値を超えていないと判定した場合（ステップＳ１４、No）、当該ホストについてトラヒックデータのサンプリングを行わない。

[0034] 次に、生成装置１０は、特徴ベクトルを生成する（ステップＳ１６）。このとき、生成装置１０は、サンプリングが行われたホストについては、サンプリングされたトラヒックデータから特徴ベクトルを生成し、サンプリングが行われなかったホストについては、入力されたサンプリング前のトラヒックデータから特徴ベクトルを生成する。

[0035] その後、未選択のホストがある場合（ステップＳ１７、Yes）、生成装置１０は、ステップＳ１３に戻り処理を繰り返す。一方、未選択のホストがなくなった場合（ステップＳ１７、No）、生成装置１０は、生成装置１０は、各特徴ベクトルを使って学習又は予測を実行する（ステップＳ１８）。

[0036] ［第１の実施形態の効果］

これまで説明してきたように、生成装置１０は、複数のトラヒックデータを所定のターゲットごとに集約する。また、生成装置１０は、集約されたトラヒックデータの数が閾値を超えているターゲットのトラヒックデータをサンプリングする。また、生成装置１０は、サンプリングが行われなかったタ

ーゲットについては、集約されたトラヒックデータの特徴を表す特徴ベクトルを生成し、サンプリングが行われたターゲットについては、サンプリングされたトラヒックデータの特徴を表す特徴ベクトルを生成する。このように、生成装置１０は、ターゲットごとのトラヒックデータの件数に応じてサンプリングを行うか否かを決定することができる。このため、第１の実施形態によれば、特徴量を生成するためのデータが、サンプリングにより極端に少なくなることや、全くなくなってしまうことを防止できるため、モデルの精度の低下を抑止できる。

[0037] 生成装置１０は、閾値と同数、又は、閾値以上かつターゲットに集約されたトラヒックデータの数未満のトラヒックデータをサンプリングする。また、生成装置１０は、サンプリング部１０３は、所定の基準でソートしたトラヒックデータの先頭から所定の数のトラヒックデータをサンプリングする。このように、生成装置１０は、データの特性等に合わせて様々な方法でサンプリングを行うことができる。

[0038] ［その他の実施形態］

前処理、サンプリング及び特徴ベクトルの生成処理のうちの少なくとも一部は、並列処理により行われてもよい。図６及び７を用いて、生成装置１０が並列処理を行う場合の構成及び処理について説明する。

[0039] 図６は、前処理を並列に行う場合の構成例を示す図である。図６に示すように、生成装置１０は、前処理部１０１及びサンプリング部１０３を複数持ち、前処理及びサンプリングを並列して行うことができる。この場合、前処理部１０１がそれぞれ異なる閾値を定義し、サンプリング部１０３が自身に接続する前処理部１０１で定義された閾値を使ってサンプリングを行うようにしてもよい。これにより、集約後のターゲットごとのトラヒックデータの件数に応じて柔軟に閾値を定義することが可能になる。

[0040] 図７は、特徴量生成処理を並列に行う場合の構成例を示す図である。図７に示すように、生成装置１０は、生成部１０５を複数持ち、特徴ベクトルの生成処理を並列して行うことができる。この場合、特徴ベクトル生成に要す

る時間を短縮することが可能になる。

[0041] ［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ように構成されていることを要しない。すなわち、各装置の分散及び統合の具体的形態は図示のものに限られず、その全部又は一部を、各種の負荷や使用状況等に応じて、任意の単位で機能的又は物理的に分散又は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、その全部又は任意の一部が、CPU及び当該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。

[0042] また、本実施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又は一部を手動的に行うこともでき、あるいは、手動的に行われるものとして説明した処理の全部又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。

[0043] ［プログラム］

一実施形態として、生成装置１０は、パッケージソフトウェアやオンラインソフトウェアとして上記の抽出処理を実行する生成プログラムを所望のコンピュータにインストールさせることによって実装できる。例えば、上記の生成プログラムを情報処理装置に実行させることにより、情報処理装置を生成装置１０として機能させることができる。ここで言う情報処理装置には、デスクトップ型又はノート型のパーソナルコンピュータが含まれる。また、その他にも、情報処理装置にはスマートフォン、携帯電話機やPHS（Personal Handyphone System）等の移動体通信端末、さらには、PDA（Personal Digital Assistant）等のスレート端末等がその範疇に含まれる。

[0044] また、生成装置１０は、ユーザが使用する端末装置をクライアントとし、当該クライアントに上記の抽出処理に関するサービスを提供する抽出サーバ

装置として実装することもできる。例えば、抽出サーバ装置は、トラヒックデータを入力とし、第１の特徴量及び第２の特徴量を出力とする抽出サービスを提供するサーバ装置として実装される。この場合、抽出サーバ装置は、Webサーバとして実装することとしてもよいし、アウトソーシングによって上記の抽出処理に関するサービスを提供するクラウドとして実装することとしてもかまわない。

[0045] 図８は、生成プログラムを実行するコンピュータの一例を示す図である。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。

[0046] 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１１１０、キーボード１１２０に接続される。ビデオアダプタ１０６０は、例えばディスプレイ１１３０に接続される。

[0047] 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、生成装置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、生成装置１０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１０

９３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により代替されてもよい。

[0048] また、上述した実施形態の処理で用いられる設定データは、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は、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、上述した実施形態の処理を実行する。

[0049] 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（LAN（Local Area Network）、WAN（Wide Area Network）等）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

[0050] １０ 生成装置
２０ データセット
３０ 予測結果
１０１ 前処理部
１０２ 選択部
１０３、１０４ サンプリング部
１０５ 生成部
１０７ 学習部
１０８ モデル
１０９ 予測部

請求の範囲

[請求項1] 複数のトラヒックデータを所定のターゲットごとに集約する前処理部と、

前記前処理部によって集約されたトラヒックデータの数が閾値を超えているターゲットのトラヒックデータをサンプリングするサンプリング部と、

前記サンプリング部によってサンプリングが行われなかったターゲットについては、集約されたトラヒックデータの特徴を表す特徴ベクトルを生成し、前記サンプリング部によってサンプリングが行われたターゲットについては、サンプリングされたトラヒックデータの特徴を表す特徴ベクトルを生成する生成部と、

を有することを特徴とする生成装置。

[請求項2] 前記サンプリング部は、前記閾値と同数、又は、前記閾値以上かつ前記ターゲットに集約されたトラヒックデータの数未満のトラヒックデータをサンプリングすることを特徴とする請求項１に記載の生成装置。

[請求項3] 前記サンプリング部は、所定の基準でソートしたトラヒックデータの先頭から所定の数のトラヒックデータをサンプリングすることを特徴とする請求項１に記載の生成装置。

[請求項4] コンピュータによって実行される生成方法であって、

複数のトラヒックデータを所定のターゲットごとに集約する前処理工程と、

前記前処理工程によって集約されたトラヒックデータの数が閾値を超えているターゲットのトラヒックデータをサンプリングするサンプリング工程と、

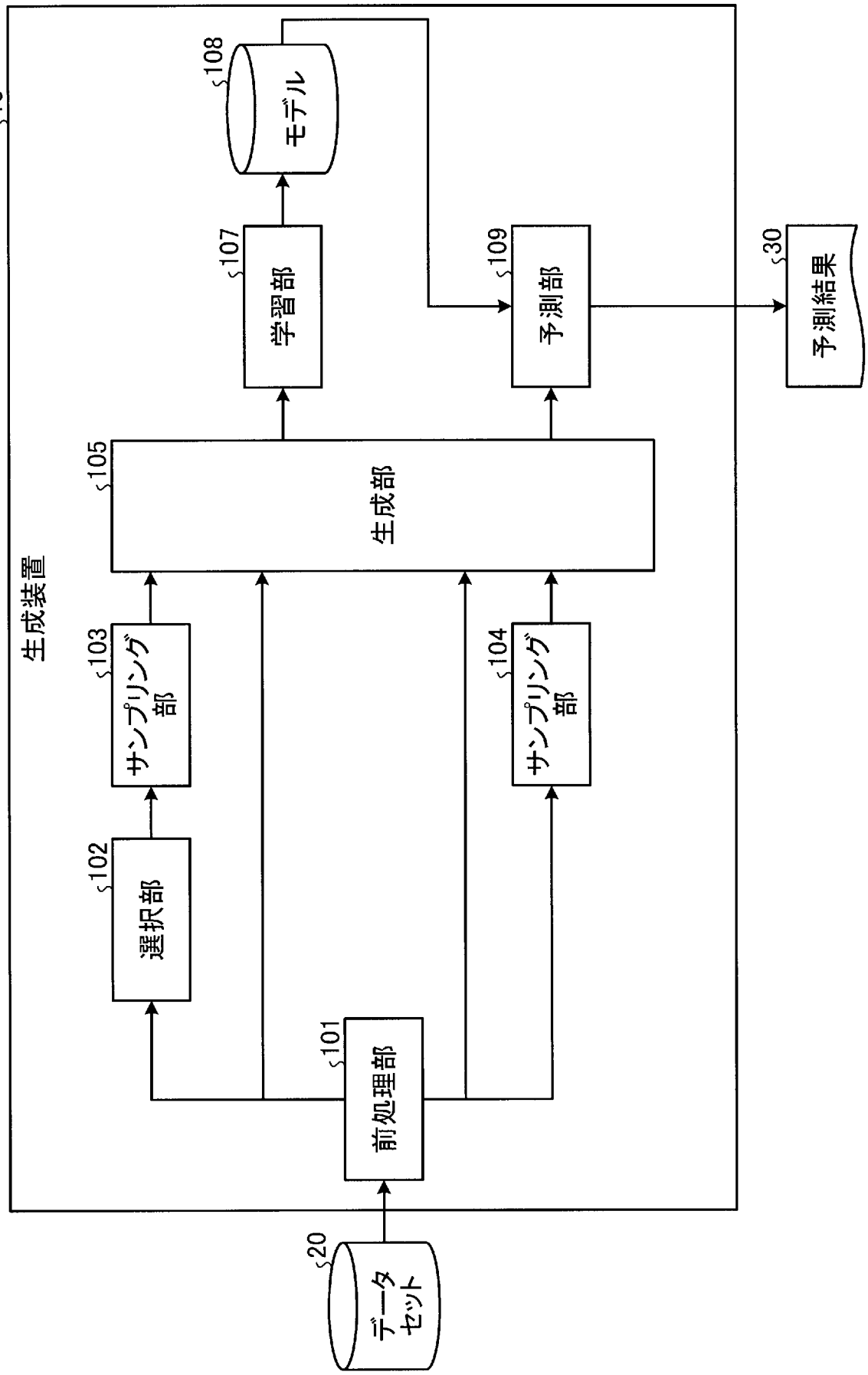
前記サンプリング工程によってサンプリングが行われなかったターゲットについては、集約されたトラヒックデータの特徴を表す特徴ベクトルを生成し、前記サンプリング工程によってサンプリングが行わ

れたターゲットについては、サンプリングされたトラヒックデータの特徴を表す特徴ベクトルを生成する生成工程と、

を有することを特徴とする生成方法。

[請求項5] コンピュータを、請求項１から３のいずれか１項に記載の生成装置として機能させるための生成プログラム。

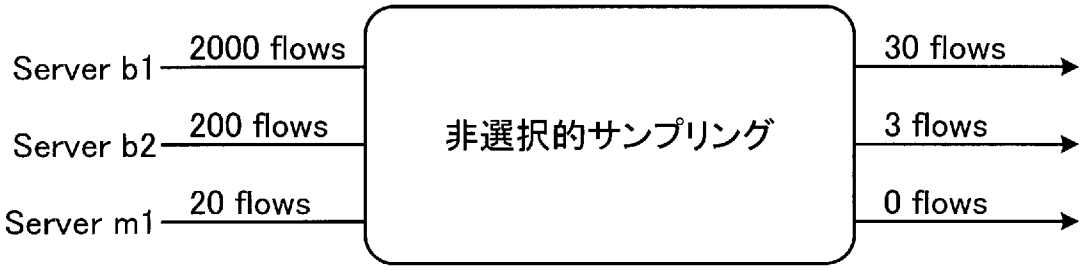
[図1]


10
生成装置
108
モデル
107
学習部
109
予測部
30
予測結果
105
生成部
103
サンプリング部
104
サンプリング部
102
選択部
101
前処理部
20
データセット

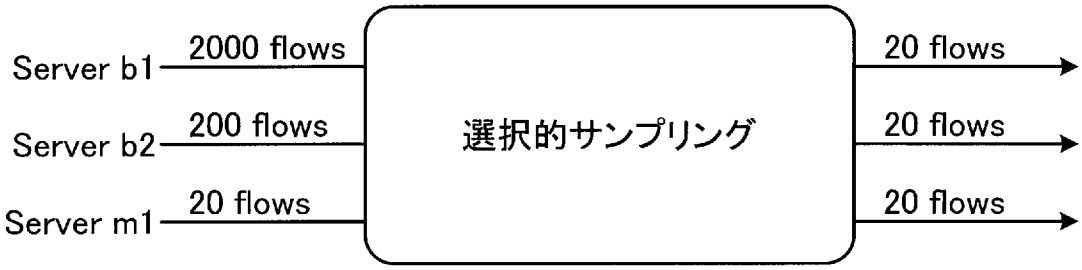
[図2]

Class	Target i	Bag r_i of network flows (ts, sip, dip, sp, dp, pr, pkt, byt)
Malicious	Server m1	8:00:00, m1, C_1, 80, 20000, tcp, 1, 100 8:00:00, C_2, m1, 30000, 80, udp, 2, 500 …
	Server m2	…
	…	…
Benign	Server b1	8:00:00, b1, C_1, 80, 20000, tcp, 1, 500 8:05:00, C_1, b1, 30000, 80, tcp, 1, 50 …
	Server b2	…
	…	…

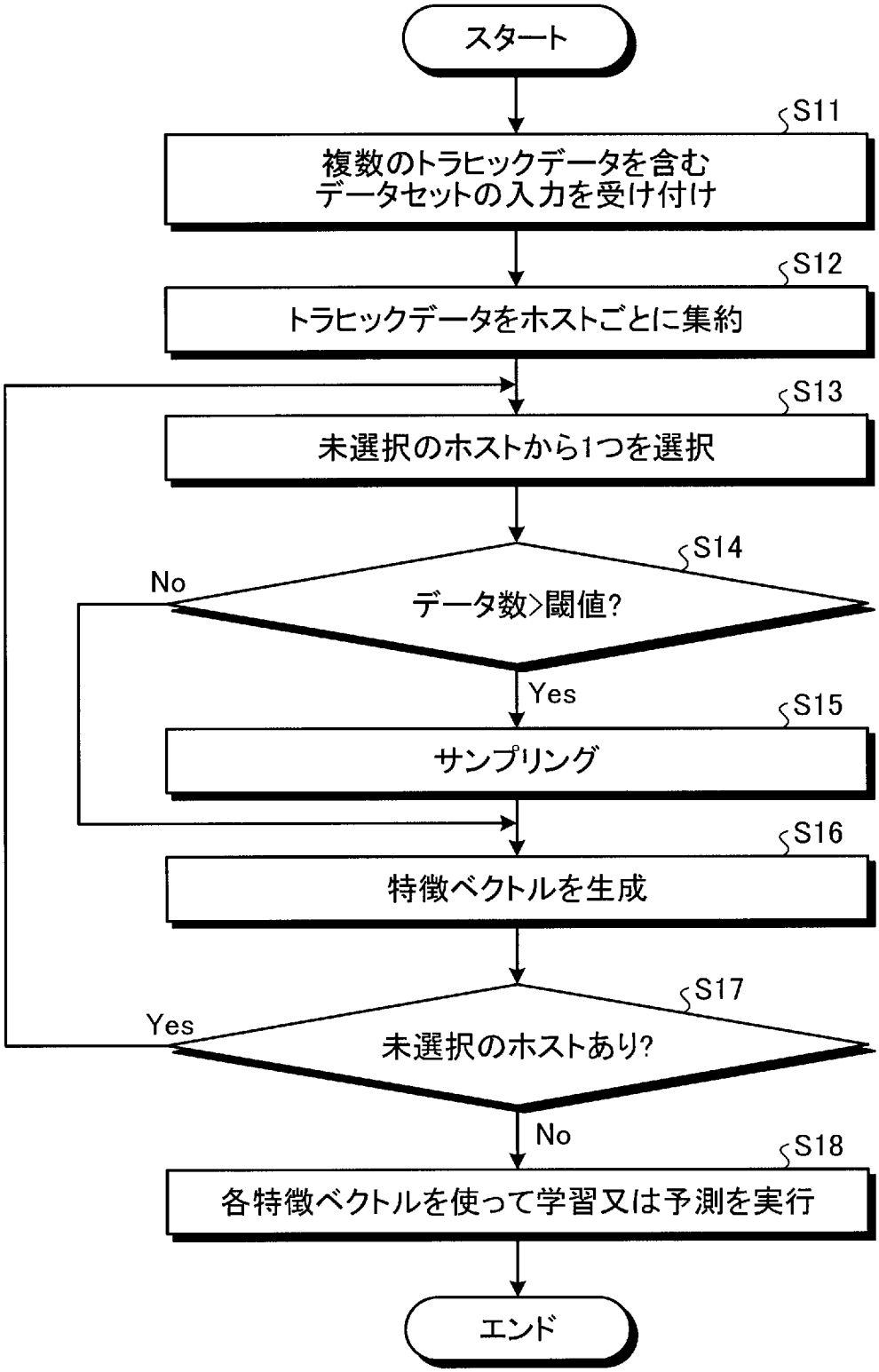
[図3]


Server b1
2000 flows
Server b2
200 flows
Server m1
20 flows
非選択的サンプリング
30 flows
3 flows
0 flows

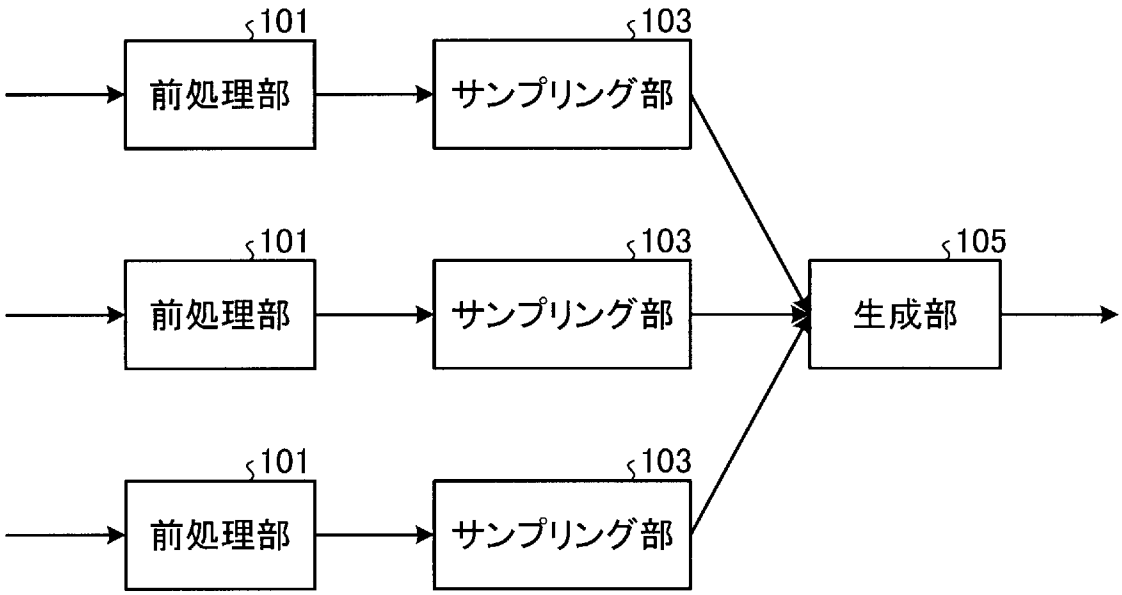
[図4]


Server b1
2000 flows
Server b2
200 flows
Server m1
20 flows
選択的サンプリング
20 flows
20 flows
20 flows

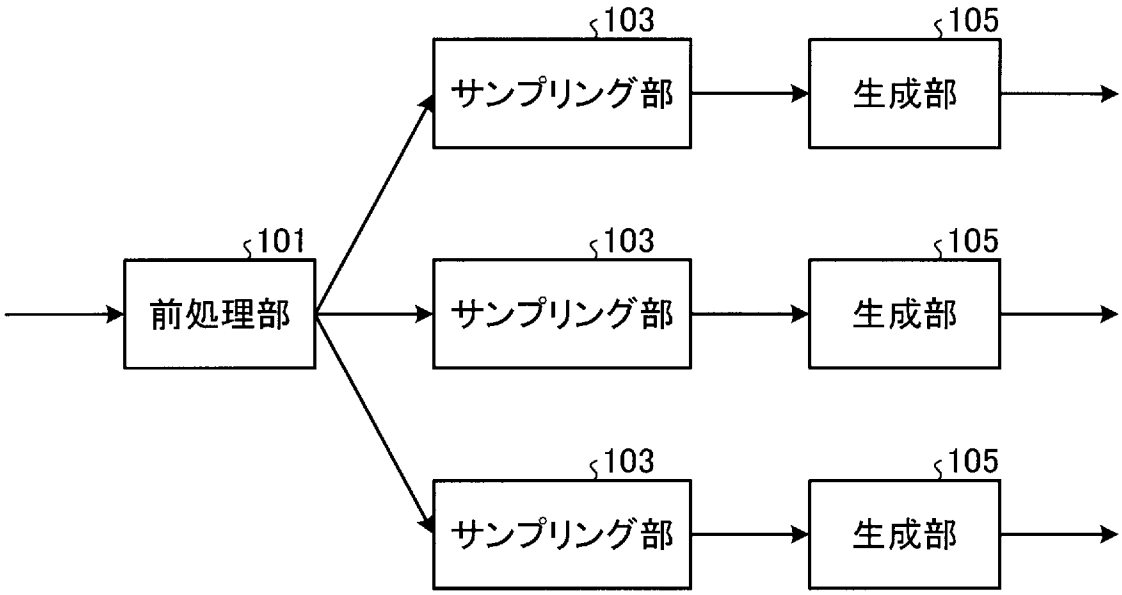
[図5]


スタート
S11
複数のトラヒックデータを含む
データセットの入力を受け付け
S12
トラヒックデータをホストごとに集約
S13
未選択のホストから1つを選択
S14
No
データ数>閾値?
Yes
S15
サンプリング
S16
特徴ベクトルを生成
S17
Yes
未選択のホストあり?
No
S18
各特徴ベクトルを使って学習又は予測を実行
エンド

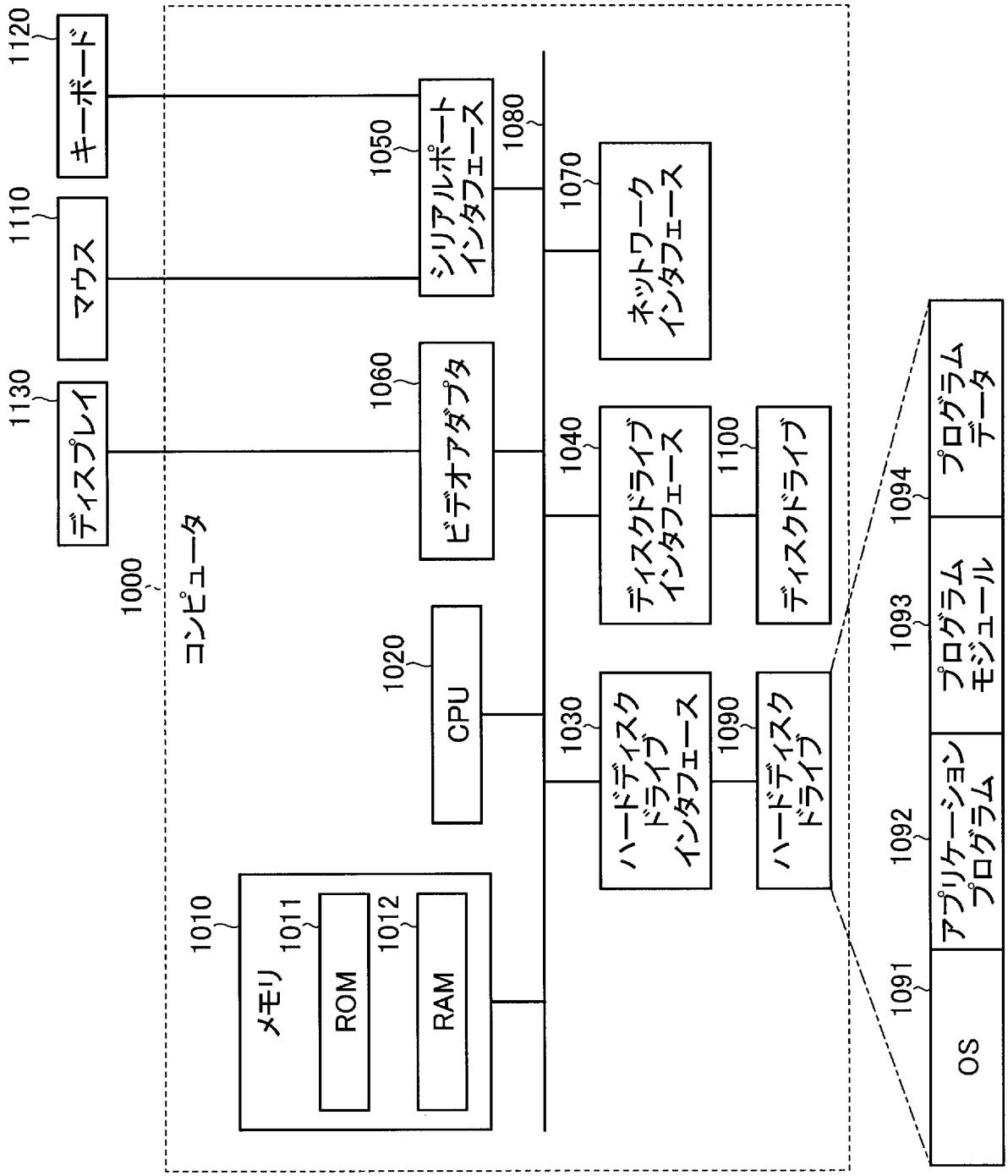
[図6]


101
前処理部
103
サンプリング部
101
前処理部
103
サンプリング部
105
生成部
101
前処理部
103
サンプリング部

[図7]


101
前処理部
103
サンプリング部
103
サンプリング部
103
サンプリング部
105
生成部
105
生成部
105
生成部

[図8]


1120
キーボード
1110
マウス
1130
ディスプレイ
1000
コンピュータ
1050
シリアルポート
インタフェース
1080
1070
ネットワーク
インタフェース
1060
ビデオアダプタ
1040
ディスクドライブ
インタフェース
1100
ディスクドライブ
1020
CPU
1030
ハードディスク
ドライブ
インタフェース
1090
ハードディスク
ドライブ
1010
メモリ
1011
ROM
1012
RAM
1091
OS
1092
アプリケーション
プログラム
1093
プログラム
モジュール
1094
プログラム
データ

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/028178

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl. H04L12/70(2013.01)i, G06F21/55(2013.01)i, G06N20/00(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl. H04L12/70, G06F21/55, G06N20/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan	1922-1996
Published unexamined utility model applications of Japan	1971-2019
Registered utility model specifications of Japan	1996-2019
Published registered utility model applications of Japan	1994-2019

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 2015/194604 A1 (NIPPON TELEGRAPH AND TELEPHONE CORP.) 23 December 2015, paragraphs [0031], [0046], [0124] & US 2017/0149808 A1, paragraphs [0058], [0074], [0156] & JP 2017-143583 A & JP 2018-38062 A & US 2017/0230396 A1 & EP 3145130 A1 & CN 106464577 A	1-5
Y	JP 2015-149695 A (KDDI CORPORATION) 20 August 2015, paragraph [0040] (Family: none)	1-5

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search 05 September 2019 (05.09.2019)	Date of mailing of the international search report 17 September 2019 (17.09.2019)
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/028178

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2019-075745 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 16 May 2019, paragraphs [0023]-[0024], fig. 2 (Family: none)	1-5

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１９／０２８１７８

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　H04L12/70(2013.01)i, G06F21/55(2013.01)i, G06N20/00(2019.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　H04L12/70, G06F21/55, G06N20/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１９年
日本国実用新案登録公報	１９９６－２０１９年
日本国登録実用新案公報	１９９４－２０１９年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Y	WO 2015/194604 A1 (日本電信電話株式会社) 2015.12.23, 段落[0031], [0046], [0124] & US 2017/0149808 A1, 段落[0058], [0074], [0156] & JP 2017-143583 A & JP 2018-38062 A & US 2017/0230396 A1 & EP 3145130 A1 & CN 106464577 A	1-5
Y	JP 2015-149695 A (ＫＤＤＩ株式会社) 2015.08.20, 段落[0040] (ファミリーなし)	1-5

☑　Ｃ欄の続きにも文献が列挙されている。　　☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
０５．０９．２０１９	１７．０９．２０１９

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｘ　９８４７
鈴木　肇
電話番号　０３－３５８１－１１０１　内線　３５９６

C（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2019-075745 A (日本電信電話株式会社) 2019.05.16, 段落[0023]-[0024], 図２ (ファミリーなし)	1-5