(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7259978号
(P7259978)

(45)発行日　令和5年4月18日(2023.4.18)

(24)登録日　令和5年4月10日(2023.4.10)

(51)国際特許分類

Ｈ０４Ｌ	*47/10*	*(2022.01)*	Ｈ０４Ｌ	47/10
Ｈ０４Ｌ	*41/16*	*(2022.01)*	Ｈ０４Ｌ	41/16
Ｇ０６Ｎ	*20/00*	*(2019.01)*	Ｇ０６Ｎ	20/00

ＦＩ

請求項の数　10　（全23頁）

(21)出願番号　特願2021-550731(P2021-550731)
(86)(22)出願日　令和1年9月30日(2019.9.30)
(86)国際出願番号　PCT/JP2019/038454
(87)国際公開番号　WO2021/064766
(87)国際公開日　令和3年4月8日(2021.4.8)
審査請求日　令和4年3月28日(2022.3.28)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100141519
弁理士　梶田 邦之
(72)発明者　沢辺 亜南
東京都港区芝五丁目７番１号 日本電気株式会社内
(72)発明者　岩井 孝法
東京都港区芝五丁目７番１号 日本電気株式会社内
審査官　鈴木 肇

最終頁に続く

(54)【発明の名称】　制御装置、方法及びシステム

(57)【特許請求の範囲】

【請求項１】

ネットワークを制御するための行動を学習する、学習手段と、

前記学習手段が生成した学習情報を記憶する、記憶手段と、を備え、

前記学習手段は、

前記ネットワークに対して行われた行動の報酬を、前記行動が行われた後のネットワークを介して提供されたアプリケーションの定常性に基づき定める、制御装置。

【請求項２】

ネットワークを制御するための行動を学習するステップと、

前記学習により生成された学習情報を記憶するステップと、

を含み、

前記学習するステップは、

前記ネットワークに対して行われた行動の報酬を、前記行動が行われた後のネットワークを介して提供されたアプリケーションの定常性に基づき定める、方法。

【請求項３】

前記学習するステップは、

前記行動が行われた後のネットワークを介して提供された前記アプリケーションが定常状態であれば、前記ネットワークに対して行われた行動に正の報酬を与え、

前記行動が行われた後のネットワークを介して提供された前記アプリケーションが非定常状態であれば、前記ネットワークに対して行われた行動に負の報酬を与える、請求項２

に記載の方法。
【請求項４】
前記学習するステップは、
前記ネットワークに対して行動を起こしたことにより変動するネットワークを介して提供された前記アプリケーションの状態に関する時系列データに基づいて前記ネットワークを介して提供された前記アプリケーションの定常性を判定する、請求項２又は３に記載の方法。
【請求項５】
前記学習するステップは、前記ネットワークを介して提供された前記アプリケーションの状態を、前記ネットワークに流れるトラヒックを特徴付ける特徴量、ユーザ体感品質及び制御品質のうち少なくとも１つから推定する、請求項４に記載の方法。
【請求項６】
前記学習するステップにより生成された学習モデルから得られる行動に基づき、前記ネットワークを制御するステップをさらに含む、請求項２乃至５のいずれか一項に記載の方法。
【請求項７】
ネットワークを制御するための行動を学習する、学習手段と、
前記学習手段が生成した学習情報を記憶する、記憶手段と、を含み、
前記学習手段は、
前記ネットワークに対して行われた行動の報酬を、前記行動が行われた後のネットワークを介して提供されたアプリケーションの定常性に基づき定める、システム。
【請求項８】
前記学習手段は、
前記行動が行われた後のネットワークを介して提供された前記アプリケーションが定常状態であれば、前記ネットワークに対して行われた行動に正の報酬を与え、
前記行動が行われた後のネットワークを介して提供された前記アプリケーションが非定常状態であれば、前記ネットワークに対して行われた行動に負の報酬を与える、請求項７に記載のシステム。
【請求項９】
前記学習手段は、
前記ネットワークに対して行動を起こしたことにより変動するネットワークを介して提供された前記アプリケーションの状態に関する時系列データに基づいて前記ネットワークを介して提供された前記アプリケーションの定常性を判定する、請求項７又は８に記載のシステム。
【請求項１０】
前記学習手段は、前記ネットワークを介して提供された前記アプリケーションの状態を、前記ネットワークに流れるトラヒックを特徴付ける特徴量、ユーザ体感品質及び制御品質のうち少なくとも１つから推定する、請求項９に記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、制御装置、方法及びシステムに関する。
【背景技術】
【０００２】
通信技術、情報処理技術の進展に伴い様々なサービスがネットワーク上にて提供される状況にある。例えば、ネットワーク上のサーバから動画データが配信され、端末にて当該動画データを再生することや、サーバから工場等に設置されたロボット等を遠隔制御することが行われている。
【０００３】
上記のようなネットワーク上で提供されるサービス、アプリケーションにおいて、エン

ドユーザが感じ取る品質（ＱｏＥ；Quality of Experience）や制御品質（ＱｏＣ；Quality of Control）を高める取り組みがなされている。

【０００４】

例えば、特許文献１には、個別のｗｅｂページの影響が除去された表示待ち時間の品質の推定を可能とする、と記載されている。特許文献１に記載された技術では、任意のエリア及び時間帯におけるトラフィック計測データに基づき当該エリア及び時間帯におけるｗｅｂページの表示待ち時間の品質を推定している。

【先行技術文献】

【特許文献】

【０００５】

【文献】特開２０１９－０７５０３０号公報

【発明の概要】

【発明が解決しようとする課題】

【０００６】

上記特許文献１に開示された技術では、ＳＶＭ（Support Vector Machine）と称される機械学習が用いられている。ここで、近年、深層学習（ディープラーニング）に代表される機械学習に関する技術が進展し、種々の分野への機械学習の適用が検討されている。

【０００７】

例えば、チェス等のゲームやロボット等の制御に機械学習を適用することが検討されている。ゲームの運用に機械学習を適用する場合には、ゲーム内のスコアの最大化が報酬に設定され、機械学習の性能が評価される。また、ロボットの制御では、目標動作の実現が報酬に設定され、機械学習の性能が評価される。通常、機械学習（強化学習）では、即時報酬及びエピソード単位の報酬の総和により学習の性能が議論される。

【０００８】

しかし、ネットワークの制御に機械学習を適用する場合には何を報酬に設定するのかが問題となる。例えば、ネットワークの制御では、ゲームに機械学習を適用する場合のように最大化するスコアの存在を観念することができない。例えば、ネットワークに含まれる通信機器におけるスループットを最大化することを報酬に設定したとしてもサービス、アプリケーションによっては適切な設定とはいえない。

【０００９】

本発明は、機械学習を用いた効率的なネットワークの制御を実現することに寄与する、制御装置、方法及びシステムを提供することを主たる目的とする。

【課題を解決するための手段】

【００１０】

本発明の第１の視点によれば、ネットワークを制御するための行動を学習する、学習部と、前記学習部が生成した学習情報を記憶する、記憶部と、を備え、前記学習部は、前記ネットワークに対して行われた行動の報酬を、前記行動が行われた後のネットワークの定常性に基づき定める、制御装置が提供される。

【００１１】

本発明の第２の視点によれば、ネットワークを制御するための行動を学習するステップと、前記学習により生成された学習情報を記憶するステップと、を含み、前記学習するステップは、前記ネットワークに対して行われた行動の報酬を、前記行動が行われた後のネットワークの定常性に基づき定める、方法が提供される。

【００１２】

本発明の第３の視点によれば、ネットワークを制御するための行動を学習する、学習手段と、前記学習手段が生成した学習情報を記憶する、記憶手段と、を含み、前記学習手段は、前記ネットワークに対して行われた行動の報酬を、前記行動が行われた後のネットワークの定常性に基づき定める、システムが提供される。

【発明の効果】

【００１３】

本発明の各視点によれば、機械学習を用いた効率的なネットワークの制御を実現することに寄与する、制御装置、方法及びシステムが提供される。なお、本発明により、当該効果の代わりに、又は当該効果と共に、他の効果が奏されてもよい。

【図面の簡単な説明】

【００１４】

【図１】一実施形態の概要を説明するための図である。

【図２】一実施形態に係る制御装置の動作の一例を示すフローチャートである。

【図３】第１の実施形態に係る通信ネットワークシステムの概略構成の一例を示す図である。

【図４】Ｑテーブルの一例を示す図である。

【図５】ニューラルネットワークの構成の一例を示す図である。

【図６】強化学習により得られる重みの一例を示す図である。

【図７】第１の実施形態に係る制御装置の処理構成の一例を示す図である。

【図８】特徴量とネットワークの状態を対応付ける情報の一例を示す図である。

【図９】行動と制御内容を対応付けたテーブル情報の一例を示す図である。

【図１０】特徴量の時系列データの一例を示す図である。

【図１１】第１の実施形態に係る制御装置の制御モード時の動作の一例を示すフローチャートである。

【図１２】第１の実施形態に係る制御装置の学習モード時の動作の一例を示すフローチャートである。

【図１３】強化学習実行部の動作を説明するための図である。

【図１４】スループットの時系列データの一例を示す図である。

【図１５】報酬の与え方を説明するための図である。

【図１６】制御装置のハードウェア構成の一例を示す図である。

【発明を実施するための形態】

【００１５】

はじめに、一実施形態の概要について説明する。なお、この概要に付記した図面参照符号は、理解を助けるための一例として各要素に便宜上付記したものであり、この概要の記載はなんらの限定を意図するものではない。なお、本明細書及び図面において、同様に説明されることが可能な要素については、同一の符号を付することにより重複説明が省略され得る。

【００１６】

一実施形態に係る制御装置１００は、学習部１０１と記憶部１０２を含む（図１参照）。学習部１０１は、ネットワークを制御するための行動を学習する。記憶部１０２は、学習部１０１が生成した学習情報を記憶する。学習部１０１は、ネットワークに対して行動をする（図２のステップＳ０１）。学習部１０１は、ネットワークに対して行われた行動の報酬を、行動が行われた後のネットワークの定常性に基づき定め、ネットワークを制御するための行動を学習する（図２のステップＳ０２）。

【００１７】

ネットワークにより提供されるサービスやアプリケーションでは、「ネットワークの安定性」が重要視される。制御装置１００は、ネットワークに対して行った行動（制御パラメータの変更）により得られる状態の定常性に基づき報酬を定める。即ち、制御装置１００は、機械学習（強化学習）の際にネットワークの状態が安定している収束状態に価値が高いものと捉え、そのような状況の場合に高い報酬を与えネットワークを制御するための学習を行う。その結果、機械学習を用いた効率的なネットワークの制御が実現される。

【００１８】

以下に具体的な実施形態について、図面を参照してさらに詳しく説明する。

【００１９】

[第１の実施形態]

第１の実施形態について、図面を用いてより詳細に説明する。

【００２０】
図３は、第１の実施形態に係る通信ネットワークシステムの概略構成の一例を示す図である。図３を参照すると、通信ネットワークシステムは、端末１０と、制御装置２０と、サーバ３０と、を含んで構成される。
【００２１】
端末１０は、通信機能を有する装置である。端末１０には、ＷＥＢ（ウェブ）カメラ、監視カメラ、ドローン、スマートフォン、ロボット等が例示される。但し、端末１０を上記ＷＥＢカメラ等に限定する趣旨ではない。端末１０は、通信機能を備える任意の装置とすることができる。
【００２２】
端末１０は、制御装置２０を介してサーバ３０と通信する。端末１０とサーバ３０により様々なアプリケーション、サービスが提供される。
【００２３】
例えば、端末１０がＷＥＢカメラの場合には、サーバ３０が当該ＷＥＢカメラからの画像データを解析し、工場等の資材管理が行われる。例えば、端末１０がドローンの場合には、サーバ３０からドローンに制御コマンドが送信され、ドローンが荷物等を搬送する。例えば、端末１０がスマートフォンの場合には、サーバ３０からスマートフォンに向けて動画が配信され、ユーザはスマートフォンを用いて動画を視聴する。
【００２４】
制御装置２０は、例えば、プロキシサーバやゲートウェイ等の通信機器であり、端末１０とサーバ３０からなるネットワークを制御する装置である。制御装置２０は、ＴＣＰ（Transmission Control Protocol）のパラメータ群やバッファ制御に関するパラメータ群の値を変更し、ネットワークを制御する。
【００２５】
例えば、ＴＣＰパラメータの制御としては、フローウィンドウサイズの変更が例示される。バッファ制御としては、複数バッファのキュー管理において、最低保証帯域、ＲＥＤ（Random Early Detection）のロス率、ロス開始キュー長、バッファ長に関するパラメータの変更が例示される。
【００２６】
なお、以降の説明において、上記ＴＣＰパラメータやバッファ制御に関するパラメータ等、端末１０とサーバ３０の間の通信（トラヒック）に影響を与えるパラメータを「制御パラメータ」と表記する。
【００２７】
制御装置２０は、制御パラメータを変更することで、ネットワークを制御する。制御装置２０によるネットワークの制御は、自装置（制御装置２０）のパケット転送時に行われてもよいし、端末１０やサーバ３０に制御パラメータの変更を指示することにより行われてもよい。
【００２８】
ＴＣＰセッションが制御装置２０により終端される場合には、例えば、制御装置２０は、端末１０との間で形成されるＴＣＰセッションのフローウィンドウサイズを変更することで、ネットワークを制御する。制御装置２０は、サーバ３０から受信したパケットを格納するバッファのサイズを変更したり、当該バッファからパケットを読み出す周期を変更したりしてネットワークを制御してもよい。
【００２９】
制御装置２０は、ネットワークの制御に「機械学習」を用いる。より具体的には、制御装置２０は、強化学習により得られる学習モデルに基づきネットワークを制御する。
【００３０】
強化学習には、種々のバリエーションが存在するが、例えば、制御装置２０は、Ｑ学習と称される強化学習の結果得られる学習情報（Ｑテーブル）に基づきネットワークを制御してもよい。

【００３１】
［Ｑ学習］
以下、Ｑ学習について概説する。
【００３２】
Ｑ学習では、与えられた「環境」における「価値」を最大化するように、「エージェント」を学習させる。当該Ｑ学習をネットワークシステムに適用すると、端末１０やサーバ３０を含むネットワークが「環境」であり、ネットワークの状態を最良にするように、制御装置２０を学習させる。
【００３３】
Ｑ学習では、状態（ステート）ｓ、行動（アクション）ａ、報酬（リワード）ｒの３要素が定義される。
【００３４】
状態ｓは、環境（ネットワーク）がどのような状態にあるかを示す。例えば、通信ネットワークシステムの場合には、トラヒック（例えば、スループット、平均パケット到着間隔等）が状態ｓに該当する。
【００３５】
行動ａは、エージェント（制御装置２０）が環境（ネットワーク）に対して取り得る行動を示す。例えば、通信ネットワークシステムの場合には、ＴＣＰパラメータ群の設定の変更や機能のオン／オフ等が行動ａとして例示される。
【００３６】
報酬ｒは、ある状態ｓにおいてエージェント（制御装置２０）が行動ａを実行した結果、どの程度の評価が得られるかを示す。例えば、通信ネットワークシステムの場合には、制御装置２０が、ＴＣＰパラメータ群の一部を変更した結果、スループットが上昇すれば正の報酬、スループットが下降すれば負の報酬の様に定められる。
【００３７】
Ｑ学習では、現在時点で得られる報酬（即時報酬）を最大化するのではなく、将来に亘る価値を最大化するように学習が進められる（Ｑテーブルが構築される）。Ｑ学習におけるエージェントの学習は、ある状態ｓにおける行動ａを採用した時の価値（Ｑ値、状態行動価値）を最大化するように行われる。
【００３８】
Ｑ値（状態行動価値）は、Ｑ（ｓ、ａ）と表記される。Ｑ学習では、エージェントが行動することによって価値の高い状態に遷移させる行動は、遷移先と同程度の価値を持つことを前提としている。このような前提により、現時点ｔにおけるＱ値は、次の時点ｔ＋１のＱ値により表現することができる（式（１）参照）。
【００３９】
【数１】

$$Q(s_t, a_t) = E_{s_{t+1}}\left(r_{t+1} + \gamma E_{a_{t+1}}\left(Q(s_{t+1}, a_{t+1})\right)\right) \quad \cdots (1)$$

【００４０】
なお、式（１）においてr_{t+1}は即時報酬、Es_{t+1}は状態S_{t+1}に関する期待値、Ea_{t+1}は行動a_{t+1}に関する期待値を示す。γは割引率である。
【００４１】
Ｑ学習では、ある状態ｓにおいて行動ａを採用した結果によりＱ値を更新する。具体的には、下記の式（２）に従いＱ値を更新する。
【００４２】

【数２】

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1})\right) \quad \cdots (2)$$

【００４３】
式（２）において、αは学習率と称されるパラメータであり、Ｑ値の更新を制御する。また、式（２）における「ｍａｘ」は状態Ｓｔ＋１の取り得る行動ａのうち最大値を出力する関数である。なお、エージェント（制御装置２０）が行動ａを選択する方式には、ε-ｇｒｅｅｄｙと称される方式を採用することができる。
【００４４】
ε-ｇｒｅｅｄｙ方式では、確率εでランダムに行動を選択し、確率１－εで最も価値の高い行動を選択する。Ｑ学習の実行により、図４に示すようなＱテーブルが生成される。
【００４５】
［ＤＱＮによる学習］
制御装置２０は、ＤＱＮ（Deep Q Network）と称される深層学習（ディープラーニング）を使った強化学習の結果得られる学習モデルに基づきネットワークを制御してもよい。Ｑ学習では、Ｑテーブルにより行動価値関数を表現しているが、ＤＱＮでは、ディープラーニングにより行動価値関数を表現する。ＤＱＮでは、最適行動価値関数を、ニューラルネットワークを使った近似関数により算出する。
【００４６】
なお、最適行動価値関数とは、ある状態ｓ時にある行動ａを行うことの価値を出力する関数である。
【００４７】
ニューラルネットワークは、入力層、中間層（隠れ層）、出力層を備える。入力層は、状態ｓを入力する。中間層の各ノードのリンクには、対応する重みが存在する。出力層は、行動ａの価値を出力する。
【００４８】
例えば、図５に示すようなニューラルネットワークの構成を考える。図５に示すニューラルネットワークを通信ネットワークシステムに適用すると、入力層のノードは、ネットワークの状態Ｓ１～Ｓ３に相当する。入力層に入力されたネットワークの状態は、中間層にて重み付けされ、出力層に出力される。
【００４９】
出力層のノードは、制御装置２０が取り得る行動Ａ１～Ａ３に相当する。出力層のノードは、行動Ａ１～Ａ３のそれぞれに対応する行動価値関数Ｑ（ｓｔ、ａｔ）の値を出力する。
【００５０】
ＤＱＮでは、上記行動価値関数を出力するノード間の結合パラメータ（重み）を学習する。具体的には、下記の式（３）に示す誤差関数Ｅ（ｓｔ、ａｔ）を設定しバックプロパゲーションにより学習を行う。
【００５１】
【数３】

$$E(s_t, a_t) = \left(r_{t+1} + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right)^2 \quad \cdots (3)$$

【００５２】
ＤＱＮによる強化学習の実行により、用意されたニューラルネットワークの中間層の構成に対応した学習情報（重み）が生成される（図６参照）。
【００５３】
ここで、制御装置２０の動作モードには、２つの動作モードが含まれる。
【００５４】
第１の動作モードは、学習モデルを算出する学習モードである。制御装置２０が「Ｑ学習」を実行することで、図４に示すようなＱテーブルが算出される。あるいは、制御装置２０が「ＤＱＮ」による強化学習を実行することで、図６に示すような重みが算出される。
【００５５】
第２の動作モードは、学習モードにて算出された学習モデルを用いてネットワークを制御する制御モードである。具体的には、制御モードの制御装置２０は、現在のネットワークの状態ｓを算出し、当該状態ｓの場合に取り得る行動ａのうち最も価値の高い行動ａを選択する。制御装置２０は、当該選択された行動ａに対応する動作（ネットワークの制御）を実行する。
【００５６】
図７は、第１の実施形態に係る制御装置２０の処理構成（処理モジュール）の一例を示す図である。図７を参照すると、制御装置２０は、パケット転送部２０１と、特徴量算出部２０２と、ネットワーク制御部２０３と、強化学習実行部２０４と、記憶部２０５と、を含んで構成される。
【００５７】
パケット転送部２０１は、端末１０やサーバ３０から送信されたパケットを受信し、当該受信したパケットを対向する装置に転送する手段である。パケット転送部２０１は、ネットワーク制御部２０３からの通知された制御パラメータに従い、パケット転送を行う。
【００５８】
例えば、ネットワーク制御部２０３からフローウィンドウサイズの設定値が通知されると、パケット転送部２０１は当該通知されたフローウィンドウサイズにてパケット転送を行う。
【００５９】
パケット転送部２０１は、受信したパケットの複製を特徴量算出部２０２に引き渡す。
【００６０】
特徴量算出部２０２は、端末１０とサーバ３０の間の通信トラヒックを特徴付ける特徴量を算出する手段である。特徴量算出部２０２は、取得したパケットからネットワーク制御の対象となるトラヒックフローを抽出する。なお、ネットワーク制御の対象となるトラヒックフローは、送信元ＩＰ（Internet Protocol）アドレス、宛先ＩＰアドレス、ポート番号等が同一のパケットからなるグループである。
【００６１】
特徴量算出部２０２は、抽出したトラヒックフローから上記特徴量を算出する。例えば、特徴量算出部２０２は、スループット、平均パケット到着間隔、パケットロス率、ジッター等を特徴量として算出する。特徴量算出部２０２は、算出した特徴量を算出時刻と共に記憶部２０５に格納する。なお、スループット等の算出については既存の技術を用いることができ、且つ、当業者にとって明らかであるのでその詳細な説明を省略する。
【００６２】

ネットワーク制御部２０３は、強化学習実行部２０４が生成した学習モデルから得られる行動に基づき、ネットワークを制御する手段である。ネットワーク制御部２０３は、強化学習の結果得られる学習モデルに基づきパケット転送部２０１に通知する制御パラメータを決定する。ネットワーク制御部２０３は、主に制御モード時に動作するモジュールである。

【００６３】

ネットワーク制御部２０３は、記憶部２０５から最新の（現在時刻の）特徴量を読み出す。ネットワーク制御部２０３は、当該読み出した特徴量から制御対象となっているネットワークの状態を推定（算出）する。

【００６４】

例えば、ネットワーク制御部２０３は、特徴量Ｆとネットワークの状態を対応付けたテーブル（図８参照）を参照し、現在の特徴量Ｆに対応するネットワークの状態を算出する。なお、トラヒックは端末１０とサーバ３０の間の通信により生じるものであるから、ネットワークの状態は「トラヒックの状態」と捉えることもできる。即ち、本願開示において、「トラヒックの状態」と「ネットワークの状態」は相互に読み替えが可能である。

【００６５】

Ｑ学習により学習モデルが構築された場合には、ネットワーク制御部２０３は、記憶部２０５に格納されたＱテーブルを参照し、現在のネットワーク状態に対応する各行動（アクション）のうち価値Ｑが最も高い行動を取得する。例えば、図４の例では、算出されたトラヒックの状態が「状態Ｓ１」であり、価値Ｑ（Ｓ１、Ａ１）、Ｑ（Ｓ１、Ａ２）、Ｑ（Ｓ１、Ａ３）のうち価値Ｑ（Ｓ１、Ａ１）が最大であれば、行動Ａ１が読み出される。

【００６６】

あるいは、ＤＮＱにより学習モデルが構築された場合には、ネットワーク制御部２０３は、図５に示すようなニューラルネットワークに現在のネットワーク状態を入力し、取り得る行動のうち最も価値の高い行動を取得する。

【００６７】

ネットワーク制御部２０３は、取得した行動に応じて制御パラメータを決定し、パケット転送部２０１に設定（通知）する。なお、記憶部２０５には、行動と制御内容を対応付けたテーブル（図９参照）が格納され、ネットワーク制御部２０３は、当該テーブルを参照してパケット転送部２０１に設定する制御パラメータを決定する。

【００６８】

例えば、図９に示すように、制御パラメータの変更内容（更新内容）が制御内容として記載されている場合には、ネットワーク制御部２０３は、当該変更内容に応じた制御パラメータをパケット転送部２０１に通知する。

【００６９】

強化学習実行部２０４は、ネットワークを制御するための行動（制御パラメータ）を学習する手段である。強化学習実行部２０４は、上記説明したＱ学習やＤＱＮによる強化学習を実行し、学習モデルを生成する。強化学習実行部２０４は、主に学習モード時に動作するモジュールである。

【００７０】

強化学習実行部２０４は、記憶部２０５に格納された特徴量から現在時刻ｔのネットワークの状態ｓを算出する。強化学習実行部２０４は、算出した状態ｓの取り得る行動ａのなかから上記ε-ｇｒｅｅｄｙ方式のような方法で行動ａを選択する。強化学習実行部２０４は、当該選択した行動に対応する制御内容（制御パラメータの更新値）をパケット転送部２０１に通知する。強化学習実行部２０４は、上記行動に応じたネットワークの変化に応じて報酬を定める。その際、強化学習実行部２０４は、ネットワークに対して行われた行動の報酬を、行動が行われた後のネットワークの定常性に基づき定める。

【００７１】

具体的には、強化学習実行部２０４は、行動ａを起こした結果、ネットワークが定常状態にあるか否かに基づき報酬を決定する。強化学習実行部２０４は、式（２）や式（３）

に記載された報酬ｒｔ＋１を定める際、ネットワークが定常状態であれば（ネットワークが安定していれば）、正の報酬を与える。対して、ネットワークの状態が非定常状態であれば（ネットワークが不安定であれば）、強化学習実行部２０４は、負の報酬を与える。

【００７２】

強化学習実行部２０４は、ネットワークに対して行動を起こしたことにより変動するネットワークの状態に関する時系列データに対して統計処理を実施することで、ネットワークの定常性を判定する。

【００７３】

具体的には、強化学習実行部２０４は、上記ε-ｇｒｅｅｄｙ方式のような方法で選択された行動ａに対応するネットワークの制御を実行後の次の時刻ｔ＋１から所定期間前までの特徴量（特徴量の時系列データ）を読み出す。強化学習実行部２０４は、当該読み出された特徴量の時系列データに対して統計処理を施すことで、ネットワークの状態が定常状態か否かを示す評価指標を算出する。

【００７４】

具体的には、強化学習実行部２０４は、上記時系列データを自己回帰（Autoregressive model；ＡＲ）モデルによりモデル化する。ＡＲモデルは、時系列データｘ１、ｘ２、・・・、ｘＮを下記の式（４）に示すように、現在時刻の値を、重みが付けられた過去の値の加算（線形和）により表現するものである。

【００７５】

【数４】

$$\mathrm{x(t)} = \mathrm{c} + \sum_{i=1}^{p} w_i x(t-i) + \epsilon(t)$$

・・・（４）

【００７６】

式（４）において、ｘ（ｔ）は特徴量、ε（ｔ）はノイズ（ホワイトノイズ）、ｃは時刻により変化しない定数、ｗｉは重みを示す。ｉは過去の時刻を指定するためのサフィックスであり、ｐは上記所定期間前を指定する整数である。

【００７７】

強化学習実行部２０４は、上記式（４）に示される重みｗｉを記憶部２０５から読み出した時系列データを用いて推定する。具体的には、強化学習実行部２０４は、最尤法、ユールウォーカー等のパラメータ推定手法により重みｗｉを推定する。なお、最尤法、ユールウォーカー等のパラメータ推定手法は公知の技術を用いることができるのでその詳細な説明を省略する。

【００７８】

次に、強化学習実行部２０４は、時系列データから得られたＡＲモデルに対して単位根検定を実施する。単位根検定を実施することで、強化学習実行部２０４は、時系列データの定常度（定常度合い）を得る。強化学習実行部２０４は、単位根検定の実行により、「非定常」に対する「定常」の割合を算出することができる。単位根検定は既存のアルゴリズムにより実現でき、且つ、当業者にとって明らかであるのでその詳細な説明を省略する。

【００７９】

強化学習実行部２０４は、単位根検定により得られた定常度に対して閾値処理（例えば、取得した値が閾値以上または未満であるかを判定する処理）を実行し、ネットワークの状態が定常状態にあるか否かを判定する。つまり、強化学習実行部２０４は、ネットワークの状態が、定常状態に向かう過渡的な「非定常状態」にあるのか、又は、特定の値を中心に収束している「定常状態」にあるのか判定する。

【００８０】
具体的には、強化学習実行部２０４は、定常度が閾値以上であればネットワークの状態は「定常」と判定する。強化学習実行部２０４は、定常度が閾値よりも小さければネットワークの状態は「非定常」と判定する。
【００８１】
図１０は、特徴量の時系列データの一例を示す図である。図１０Ａに示す時系列データに対して、強化学習実行部２０４が単位根検定を実施すると、ネットワークの状態は「非定常」と判定される。
【００８２】
この場合、強化学習実行部２０４は、式（２）や式（３）の報酬r_{t+1}に負の報酬（例えば、－１）を与え、Ｑテーブルや重みを更新する。対して、図１０Ｂに示す時系列データに対して、強化学習実行部２０４が単位根検定を実施すると、ネットワークの状態は「定常」と判定される。この場合、強化学習実行部２０４は、式（２）や式（３）の報酬r_{t+1}に正の報酬（例えば、＋１）を与え、Ｑテーブルや重みを更新する。
【００８３】
第１の実施形態に係る制御装置２０の制御モード時の動作をまとめると図１１に示すフローチャートのとおりとなる。
【００８４】
制御装置２０は、パケットを取得し、特徴量を算出する（ステップＳ１０１）。制御装置２０は、当該算出された特徴量に基づきネットワークの状態を特定する（ステップＳ１０２）。制御装置２０は、学習モデルを用いて、ネットワークの状態に応じた最も価値の高い行動によりネットワークを制御する（ステップＳ１０３）。
【００８５】
第１の実施形態に係る制御装置２０の学習モード時の動作をまとめると図１２に示すフローチャートのとおりとなる。
【００８６】
制御装置２０は、パケットを取得し、特徴量を算出する（ステップＳ２０１）。制御装置２０は、当該算出された特徴量に基づきネットワークの状態を特定する（ステップＳ２０２）。制御装置２０は、ε-ｇｒｅｅｄｙ方式等により現在のネットワーク状態にて取り得る行動を選択する（ステップＳ２０３）。制御装置２０は、当該選択された行動によりネットワークを制御する（ステップＳ２０４）。制御装置２０は、特徴量の時系列データを用いてネットワークの定常性を判定する（ステップＳ２０５）。制御装置２０は、判定結果により報酬を定め（ステップＳ２０６）、学習情報（Ｑテーブル、重み）を更新する（ステップＳ２０７）。
【００８７】
続いて、端末１０の種類ごとに制御装置２０の動作について具体的に説明する。
【００８８】
[端末がドローンの場合]
端末１０がドローンの場合、ネットワークの状態を示す指標（特徴量）として、例えば、ドローンからサーバ３０へ向けて送信されるパケットの平均パケット到着間隔が選択される。サーバ３０は、ドローンに対して制御パケット（制御コマンドを含むパケット）を送信する。当該制御パケットに対するドローンからの応答パケット（肯定応答、否定応答）の平均パケット到着間隔が特徴量として選択される。
【００８９】
制御装置２０は、サーバ３０とドローンの間のパケット送受信の間隔が安定するように、制御パラメータを決定しネットワークの制御を行う。端末１０がドローンの場合の取り得る行動（変更可能な制御パラメータ）としては、サーバ３０から取得した制御パケットを格納するバッファからのパケット読み出し間隔（パケット送信間隔）が考えられる。
【００９０】
強化学習実行部２０４は、ドローンからサーバ３０に送信される応答パケットの平均パ

ケット到着間隔が安定するように、バッファから制御パケットを読み出すパラメータを学習する。サーバ３０がドローン（制御対象）を遠隔制御するアプリケーションでは、ドローンとサーバ３０間で送受信されるパケット（制御パケット、応答パケット）が安定して相手側に届くことが重視される。

【００９１】

ここで、制御パケットや応答パケットのパケットサイズはあまり大きくない。そのため、サーバ３０からのスループットが高いが、パケットの送受信が安定しない状況（一度に多くの情報を送れるがパケットの到着にばらつきがある状況）よりも、スループットは低いがパケットの送受信が安定する状況の方が、ドローンの制御では価値が高い。

【００９２】

第１の実施形態に係る制御装置２０は、ネットワークの状態（トラヒックの状態）を特徴付ける特徴量を適切に選択（例えば、平均パケット到着間隔を選択）することで、ドローンの遠隔制御というアプリケーションに適したネットワーク制御を実現できる。

【００９３】

［端末がＷＥＢカメラの場合］

上記説明では、報酬r_{t+1}を決定する条件（基準）としてネットワークの定常性を用いる場合について説明したが、上記定常性に他の基準を加えて報酬r_{t+1}を決定してもよい。ここでは、端末１０がＷＥＢカメラである場合を例に取り、報酬r_{t+1}の決定に「ネットワークの定常性」以外の項目を考慮する場合について説明する。

【００９４】

端末１０がＷＥＢカメラの場合、ネットワークの状態を示す指標（特徴量）として、例えば、ＷＥＢカメラからサーバ３０に流れるトラヒックのスループットが選択される。強化学習実行部２０４は、ＷＥＢカメラからサーバ３０へのスループットが目標値の近傍で安定するように、学習モデルを算出する。

【００９５】

例えば、端末１０、サーバ３０との間で形成されるＴＣＰセッションのフローウィンドウサイズが制御パラメータに設定され、上記目標（スループットが目標値で安定）を実現するような行動が学習される。強化学習実行部２０４は、特徴量算出部２０２が算出した特徴量（スループット）の時系列データを用いてネットワークの定常性を判定する。

【００９６】

続いて、強化学習実行部２０４は、特徴量（スループット）の範囲に応じて報酬r_{t+1}を決定する。例えば、目標値が閾値ＴＨ２１以上、且つ、閾値ＴＨ２２以下とすれば、強化学習実行部２０４は、図１３に示すような方針（ポリシ）にて報酬r_{t+1}を決定する。このような報酬の与え方により得られた学習モデルを用いることで、ＷＥＢカメラからのスループットが目標とする値近傍で安定するようにネットワークは制御される。

【００９７】

具体的には、制御装置２０によるネットワーク制御により、図１４Ａに示すようなネットワークの状態（スループットが目標値近辺で安定）を実現できる。換言すれば、スループットの範囲を考慮して報酬r_{t+1}を決定することで、図１４Ｂに示すようなネットワークの状態に陥ることが回避される。図１４Ｂでは、最終的にネットワークの状態が安定しているが、定常時のスループットは目標値から大きく乖離している。

【００９８】

なお、図１３には、スループットが所定の範囲内であれば正の報酬を与える場合を記載したが、スループットが所定の値以上の場合に正の報酬を与えてもよい（図１５参照）。図１４Ｂの状況とは逆に、目標値から遠く離れた高い値でスループットが安定することが許容できる場合には、図１５に示すように報酬r_{t+1}が決定されてもよい。

【００９９】

スループットに設ける制限に関しては、制御装置２０のリソース（通信リソース）を考慮して決定すればよい。例えば、制御パラメータにフローウィンドウサイズを選択した場合、当該ウィンドウサイズを大きくすればスループットは高い値で安定すると考えられる。

しかしながら、大きなフローウィンドウサイズを用意するためにはメモリ（リソース）の消費が大きくなり、他の端末１０に割り当て可能なリソースが減少してしまう。制御装置２０は、上記のようなメリット、デメリットを考慮してテーブル更新ポリシを決定すればよい。

【０１００】

［端末がスマートフォンの場合］

上記では、１つの特徴量によりネットワークの定常性を判定したりする場合について説明したが、複数の特徴量によりネットワークの定常性の判定等が行われてもよい。以下、端末１０がスマートフォンである場合を例に取り、ネットワークの定常性が複数の特徴量により判定される場合について説明する。

【０１０１】

ここでは、サーバ３０から動画が配信され、スマートフォン（端末１０）にて当該動画が再生される場合を想定する。特徴量算出部２０２は、サーバ３０からスマートフォンに流れるトラヒックのスループットと平均パケット到着間隔を算出する。

【０１０２】

強化学習実行部２０４は、当該２つの特徴量からネットワークの定常性を判定する。具体的には、強化学習実行部２０４は、スループットの時系列データに基づきスループットが安定しているか否かを判定する。同様に、強化学習実行部２０４は、平均パケット到着間隔の時系列データに基づき平均パケット到着間隔が安定しているか否かを判定する。

【０１０３】

強化学習実行部２０４は、スループット及び平均パケット到着間隔が共に定常状態にある場合に、ネットワークが定常状態にあると判定し、報酬r_{t+1}に正の報酬を与え、他の場合には負の報酬を与える。

【０１０４】

以上のように、第１の実施形態に係る制御装置２０は、ネットワークの状態を、ネットワークに流れるトラヒックを特徴付ける特徴量を用いて推定する。制御装置２０は、ネットワークに対して行った行動（制御パラメータの変更）により得られる状態の時系列変化に応じて、当該行動に対する報酬を定める。そのため、ネットワークにて提供されるサービスやアプリケーションレベルで求められる、「ネットワークの安定性」に高い報酬が与えられ、アプリケーション等に適したネットワーク品質の向上が実現できる。即ち、本願開示では、強化学習の際にネットワークの状態が安定している収束状態に価値が高いものと捉え、そのような状況の場合に学習器が環境（ネットワーク）に適応できていると考える、報酬を決定している。

【０１０５】

［第２の実施形態］

続いて、第２の実施形態について図面を参照して詳細に説明する。

【０１０６】

第１の実施形態では、ネットワークに流されるトラヒックを特徴付ける特徴量（例えば、スループット）によりネットワークの状態を推定している。第２の実施形態では、端末１０におけるＱｏＥ（ユーザ体感品質）やＱｏＣ（制御品質）に基づきのネットワークの状態を決定する場合について説明する。

【０１０７】

例えば、端末１０がスマートフォンであって、動画再正アプリケーションが動作している場合を考える。この場合、端末１０は、再生動画の画質、ビットレート、途絶回数（バッファが空となった回数）、フレームレート等を制御装置２０に通知する。あるいは、端末１０は、ＩＴＵ（International Telecommunication Union）-T勧告Ｐ．１２０３に規定されたＭＯＳ（Mean Opinion Score）値を制御装置２０に送信してもよい。

【０１０８】

あるいは、スマートフォンにてＷＥＢページの閲覧（ブラウザが動作）が行われている場合には、端末１０は、ページ表示までの初期待機時間を制御装置２０に通知してもよい。

【０１０９】
例えば、端末１０がロボットである場合には、ロボットは、制御コマンドの受信間隔、作業完了時間、作業成功回数等を制御装置２０に通知してもよい。
【０１１０】
あるいは、端末１０が監視カメラである場合には、監視カメラは、監視対象（例えば、人の顔、物体等）の認証率、認証回数等を制御装置２０に通知してもよい。
【０１１１】
制御装置２０は、端末１０から当該端末１０におけるＱｏＥを示す値（例えば、上記初期待機時間等）を取得し、当該値に基づきネットワークの定常性を判定し、報酬r_{t+1}を決定してもよい。その際、制御装置２０は、第１の実施形態にて説明した方法と同様にして、端末１０から取得したＱｏＥの時系列データに対して単位根検定を実施し、ネットワークの定常性を評価すればよい。
【０１１２】
あるいは、制御装置２０は、端末１０とサーバ３０の間に流れるトラヒックから上記ＱｏＥを示す値を推定してもよい。例えば、制御装置２０は、スループットからビットレートを推定し、当該推定値に基づきネットワークの定常性を判定してもよい。なお、スループットからビットレートを推定する際には、以下の参考文献１に記載された方法を用いればよい。
［参考文献１］：国際公開第２０１９／０４４０６５号
【０１１３】
以上のように、第２の実施形態に係る制御装置２０は、ネットワークの状態を、ユーザ体感品質（ＱｏＥ）や制御品質（ＱｏＣ）から推定し、ユーザ体感品質等が安定している場合に高い報酬を与えても良い。例えば、ユーザが端末を使用して動画を視聴する場合を考える。この場合、本願開示では、フレームレートが頻繁に変わるネットワーク環境（フレームレートが安定しない環境）よりも、低いフレームレートであっても一定しているネットワーク環境の方が、ネットワーク品質が高いと判断している。換言すれば、制御装置２０は、このような高いネットワーク品質を実現する制御パラメータを強化学習により学習する。
【０１１４】
続いて、通信ネットワークシステムを構成する各装置のハードウェアについて説明する。図１６は、制御装置２０のハードウェア構成の一例を示す図である。
【０１１５】
制御装置２０は、情報処理装置（所謂、コンピュータ）により構成可能であり、図１６に例示する構成を備える。例えば、制御装置２０は、プロセッサ３１１、メモリ３１２、入出力インターフェイス３１３及び通信インターフェイス３１４等を備える。上記プロセッサ３１１等の構成要素は内部バス等により接続され、相互に通信可能に構成されている。
【０１１６】
但し、図１６に示す構成は、制御装置２０のハードウェア構成を限定する趣旨ではない。制御装置２０は、図示しないハードウェアを含んでもよいし、必要に応じて入出力インターフェイス３１３を備えていなくともよい。また、制御装置２０に含まれるプロセッサ３１１等の数も図１６の例示に限定する趣旨ではなく、例えば、複数のプロセッサ３１１が制御装置２０に含まれていてもよい。
【０１１７】
プロセッサ３１１は、例えば、ＣＰＵ（Central Processing Unit）、ＭＰＵ（Micro Processing Unit）、ＤＳＰ（Digital Signal Processor）等のプログラマブルなデバイスである。あるいは、プロセッサ３１１は、ＦＰＧＡ（Field Programmable Gate Array）、ＡＳＩＣ（Application Specific Integrated Circuit）等のデバイスであってもよい。プロセッサ３１１は、オペレーティングシステム（ＯＳ；Operating System）を含む各種プログラムを実行する。
【０１１８】

メモリ３１２は、ＲＡＭ（Random Access Memory）、ＲＯＭ（Read Only Memory）、ＨＤＤ（Hard Disk Drive）、ＳＳＤ（Solid State Drive）等である。メモリ３１２は、ＯＳプログラム、アプリケーションプログラム、各種データを格納する。

【０１１９】

入出力インターフェイス３１３は、図示しない表示装置や入力装置のインターフェイスである。表示装置は、例えば、液晶ディスプレイ等である。入力装置は、例えば、キーボードやマウス等のユーザ操作を受け付ける装置である。

【０１２０】

通信インターフェイス３１４は、他の装置と通信を行う回路、モジュール等である。例えば、通信インターフェイス３１４は、ＮＩＣ（Network Interface Card）等を備える。

【０１２１】

制御装置２０の機能は、各種処理モジュールにより実現される。当該処理モジュールは、例えば、メモリ３１２に格納されたプログラムをプロセッサ３１１が実行することで実現される。また、当該プログラムは、コンピュータが読み取り可能な記憶媒体に記録することができる。記憶媒体は、半導体メモリ、ハードディスク、磁気記録媒体、光記録媒体等の非トランジェント（non-transitory）なものとすることができる。即ち、本発明は、コンピュータプログラム製品として具現することも可能である。また、上記プログラムは、ネットワークを介してダウンロードするか、あるいは、プログラムを記憶した記憶媒体を用いて、更新することができる。さらに、上記処理モジュールは、半導体チップにより実現されてもよい。

【０１２２】

なお、端末１０、サーバ３０も制御装置２０と同様に情報処理装置により構成可能であり、その基本的なハードウェア構成は制御装置２０と相違する点はないので説明を省略する。

【０１２３】

［変形例］

なお、上記実施形態にて説明した通信ネットワークシステムの構成、動作等は例示であって、システムの構成等を限定する趣旨ではない。例えば、制御装置２０は、ネットワークを制御する装置と学習モデルを生成する装置に分離されていてもよい。あるいは、学習情報（学習モデル）を記憶する記憶部２０５は、外部のデータベースサーバ等により実現されてもよい。即ち、本願開示は、学習手段、制御手段、記憶手段等を含むシステムとして実施されてもよい。

【０１２４】

上記実施形態では、特徴量の時系列データに対して単位根検定を実施することとで、ネットワークの定常度を算出している。しかし、ネットワークの定常度は他の指標により算出されてもよい。例えば、強化学習実行部２０４は、データのばらつき度合いを示す標準偏差を計算し、「平均－標準偏差」が閾値以上の場合にネットワークは定常状態であると判定してもよい。

【０１２５】

上記実施形態では、１つの閾値を用いてネットワークの定常性（安定性）を判定しているが、複数の閾値を用いてより細かくネットワークの定常度合いが算出されてもよい。例えば、「極めて安定」、「安定」、「不安定」、「極めて不安定」のように４段階でネットワークの定常性が判定されてもよい。この場合、ネットワークの定常度合いに応じて報酬が決められていてもよい。

【０１２６】

なお、端末１０はセンサ装置である場合がある。センサ装置は、オン／オフモデルに従う通信パターン（通信トラヒック）を発生する。つまり、端末１０がセンサ装置等であれば、データ（パケット）がネットワークに流れる場合と流れない場合（無通信状態）が生じ得る。そのため、制御装置２０が、トラヒック（特徴量）の時系列データそのものを使って定常性判定（単位根検定）を実施するのではなく、変動パターンにより定常性が判定

されてもよい。制御装置２０は、特徴量が上下する時間間隔に関する時系列データを用いてネットワークの定常性を判定してもよい。あるいは、制御装置２０は、事前にオン／オフモデルに従うアプリケーションを把握している場合には、無通信状態は報酬に反映しない等の対応を行ってもよい。即ち、制御装置２０は、ネットワークの状態が「通信状態」にある場合に強化学習の報酬を与えるようにしてもよい。

【０１２７】

上記実施形態では、制御装置２０は、トラヒックフローを制御の対象（制御単位）とする場合について説明した。しかし、制御装置２０は、端末１０単位、又は、複数の端末１０をまとめたグループを制御の対象としてもよい。つまり、同じ端末１０であってもアプリケーションが異なればポート番号等が異なり、異なるフローとして扱われる。制御装置２０は、同じ端末１０から送信されるパケットには同じ制御（制御パラメータの変更）を適用してもよい。あるいは、制御装置２０は、例えば、同じ種類の端末１０を１つのグループとして扱い、同じグループに属する端末１０から送信されるパケットに対して同じ制御を適用してもよい。

【０１２８】

上述の説明で用いた複数のフローチャートでは、複数の工程（処理）が順番に記載されているが、各実施形態で実行される工程の実行順序は、その記載の順番に制限されない。各実施形態では、例えば各処理を並行して実行する等、図示される工程の順番を内容的に支障のない範囲で変更することができる。また、上述の各実施形態は、内容が相反しない範囲で組み合わせることができる。

【０１２９】

上記の実施形態の一部又は全部は、以下の付記のようにも記載され得るが、以下には限られない。

［付記１］

ネットワークを制御するための行動を学習する、学習部（１０１、２０４）と、

前記学習部（１０１、２０４）が生成した学習情報を記憶する、記憶部（１０２、２０５）と、を備え、

前記学習部（１０１、２０４）は、

前記ネットワークに対して行われた行動の報酬を、前記行動が行われた後のネットワークの定常性に基づき定める、制御装置（２０、１００）。

［付記２］

前記学習部（１０１、２０４）は、

前記行動が行われた後のネットワークが定常状態であれば、前記ネットワークに対して行われた行動に正の報酬を与え、

前記行動が行われた後のネットワークが非定常状態であれば、前記ネットワークに対して行われた行動に負の報酬を与える、付記１に記載の制御装置（２０、１００）。

［付記３］

前記学習部（１０１、２０４）は、

前記ネットワークに対して行動を起こしたことにより変動するネットワークの状態に関する時系列データに基づいて前記ネットワークの定常性を判定する、付記１又は２に記載の制御装置（２０、１００）。

［付記４］

前記学習部（１０１、２０４）は、前記ネットワークの状態を、前記ネットワークに流れるトラヒックを特徴付ける特徴量、ユーザ体感品質及び制御品質のうち少なくとも１つから推定する、付記３に記載の制御装置（２０、１００）。

［付記５］

前記学習部（１０１、２０４）が生成した学習モデルから得られる行動に基づき、前記ネットワークを制御する、制御部（２０３）をさらに備える、付記１乃至４のいずれか一つに記載の制御装置（２０、１００）。

［付記６］

ネットワークを制御するための行動を学習するステップと、
前記学習により生成された学習情報を記憶するステップと、
を含み、
前記学習するステップは、
前記ネットワークに対して行われた行動の報酬を、前記行動が行われた後のネットワークの定常性に基づき定める、方法。
［付記７］
前記学習するステップは、
前記行動が行われた後のネットワークが定常状態であれば、前記ネットワークに対して行われた行動に正の報酬を与え、
前記行動が行われた後のネットワークが非定常状態であれば、前記ネットワークに対して行われた行動に負の報酬を与える、付記６に記載の方法。
［付記８］
前記学習するステップは、
前記ネットワークに対して行動を起こしたことにより変動するネットワークの状態に関する時系列データに基づいて前記ネットワークの定常性を判定する、付記６又は７に記載の方法。
［付記９］
前記学習するステップは、前記ネットワークの状態を、前記ネットワークに流れるトラヒックを特徴付ける特徴量、ユーザ体感品質及び制御品質のうち少なくとも１つから推定する、付記８に記載の方法。
［付記１０］
前記学習するステップにより生成された学習モデルから得られる行動に基づき、前記ネットワークを制御するステップをさらに含む、付記６乃至９のいずれか一つに記載の方法。
［付記１１］
ネットワークを制御するための行動を学習する、学習手段（１０１、２０４）と、
前記学習手段が生成した学習情報を記憶する、記憶手段（１０２、２０５）と、を含み、
前記学習手段（１０１、２０４）は、
前記ネットワークに対して行われた行動の報酬を、前記行動が行われた後のネットワークの定常性に基づき定める、システム。
［付記１２］
前記学習手段（１０１、２０４）は、
前記行動が行われた後のネットワークが定常状態であれば、前記ネットワークに対して行われた行動に正の報酬を与え、
前記行動が行われた後のネットワークが非定常状態であれば、前記ネットワークに対して行われた行動に負の報酬を与える、付記１１に記載のシステム。
［付記１３］
前記学習手段（１０１、２０４）は、
前記ネットワークに対して行動を起こしたことにより変動するネットワークの状態に関する時系列データに基づいて前記ネットワークの定常性を判定する、付記１１又は１２に記載のシステム。
［付記１４］
前記学習手段（１０１、２０４）は、前記ネットワークの状態を、前記ネットワークに流れるトラヒックを特徴付ける特徴量、ユーザ体感品質及び制御品質のうち少なくとも１つから推定する、付記１３に記載のシステム。
［付記１５］
前記学習手段（１０１、２０４）が生成した学習モデルから得られる行動に基づき、前記ネットワークを制御する、制御手段（２０３）をさらに備える、付記１１乃至１４のいずれか一つに記載のシステム。
［付記１６］

コンピュータ（３１１）に、
ネットワークを制御するための行動を学習する処理と、
前記学習により生成された学習情報を記憶する処理と、
を実行させ、
前記学習する処理は、
前記ネットワークに対して行われた行動の報酬を、前記行動が行われた後のネットワークの定常性に基づき定める、プログラム。

【０１３０】

なお、引用した上記の先行技術文献の各開示は、本書に引用をもって繰り込むものとする。以上、本発明の実施形態を説明したが、本発明はこれらの実施形態に限定されるものではない。これらの実施形態は例示にすぎないということ、及び、本発明のスコープ及び精神から逸脱することなく様々な変形が可能であるということは、当業者に理解されるであろう。

【符号の説明】

【０１３１】

１０　端末
２０、１００　制御装置
３０　サーバ
１０１　学習部
１０２、２０５　記憶部
２０１　パケット転送装置
２０２　特徴量算出部
２０３　ネットワーク制御部
２０４　強化学習実行部
３１１　プロセッサ
３１２　メモリ
３１３　入出力インターフェイス
３１４　通信インターフェイス

【図面】

【図１】

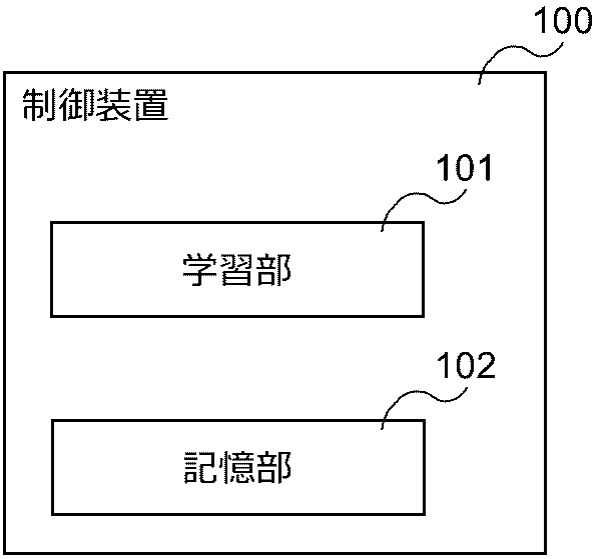

Fig.1

【図２】

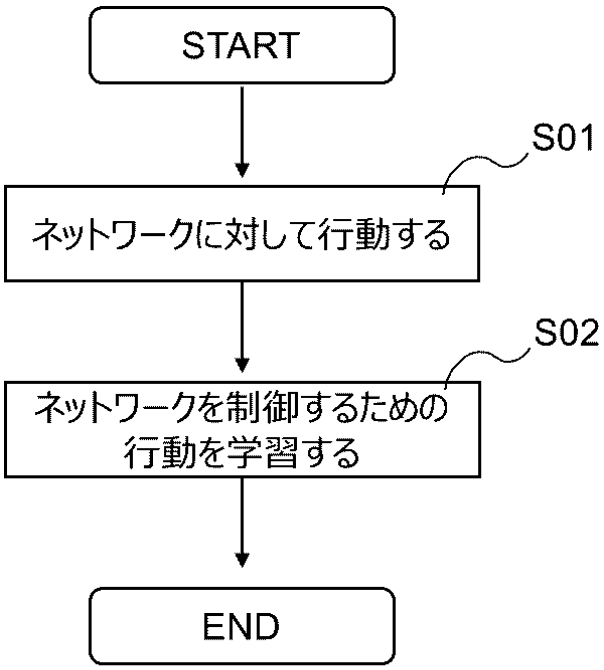

Fig.2

【図３】

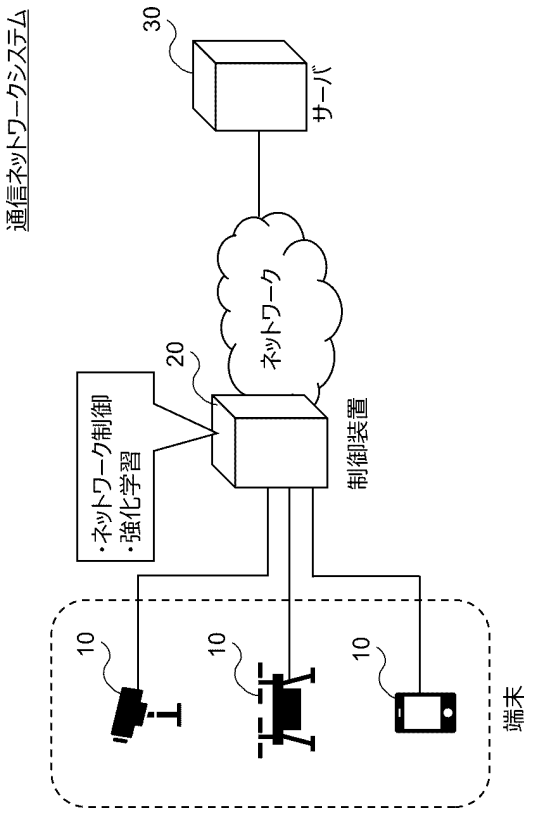

Fig.3

【図４】

Qテーブル

	行動A1	行動A2	行動A3	…
状態S1	Q(S1,A1)	Q(S1,A2)	Q(S1,A3)	…
状態S2	Q(S2,A1)	Q(S2,A2)	Q(S2,A3)	…
状態S3	Q(S3,A1)	Q(S3,A2)	Q(S3,A3)	…
…	…	…	…	…

Fig.4

【図５】

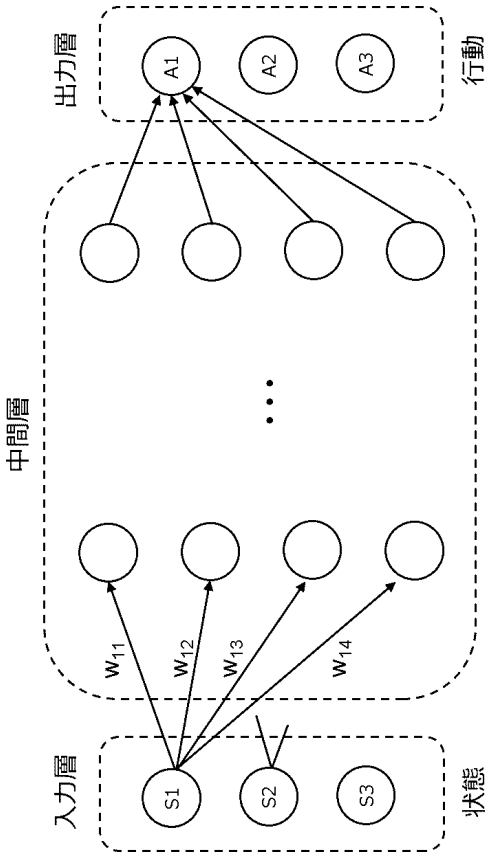

Fig.5

【図６】

重み

W11	W12	W13	・・・
W21	W22	W23	・・・
W31	W32	W33	・・・
⋮	⋮	⋮	⋮

Fig.6

【図７】

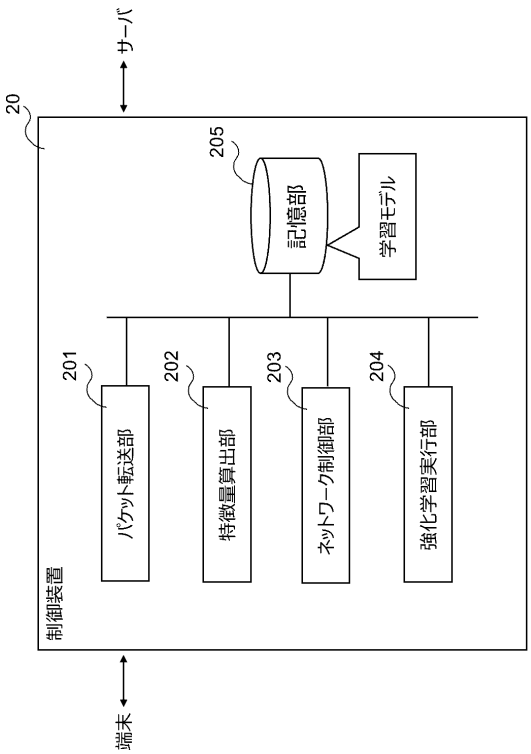

Fig.7

【図８】

特徴量	ネットワークの状態
F＜TH11	状態S1
TH11≦F＜TH12	状態S2
TH12≦F＜TH13	状態S2
⋮	⋮

Fig.8

【図９】

行動	制御内容
行動A1	ウィンドウサイズをAバイト増加
行動A2	ウィンドウサイズをBバイト増加
行動A3	ウィンドウサイズをCバイト増加
⋮	⋮

Fig.9

【図１０】

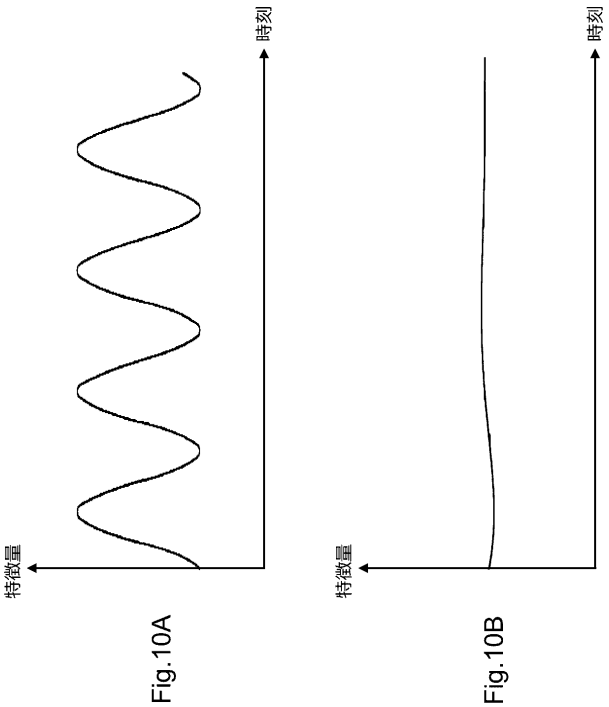

Fig.10A　Fig.10B

【図１１】

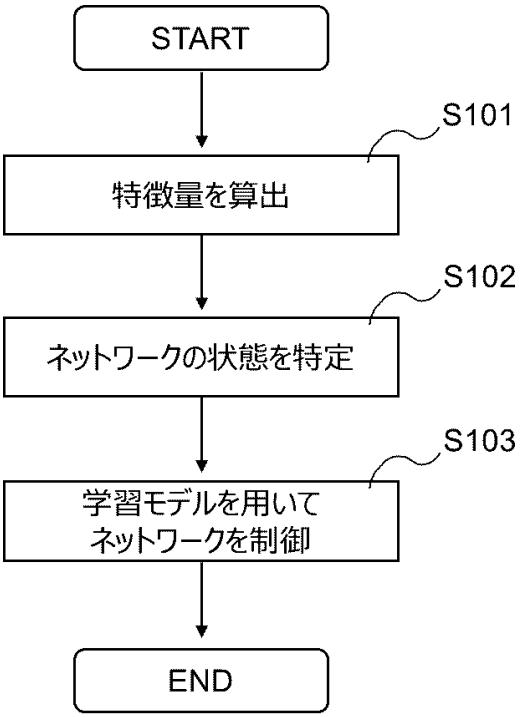

Fig.11

【図１２】

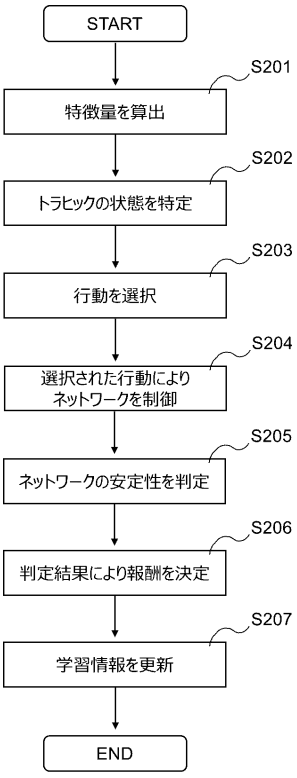

Fig.12

【図１３】

	スループット (T＜TH21)	スループット (TH21≦T≦TH22)	スループット (TH22＜T)
ネットワークは安定	負の報酬	正の報酬	負の報酬
ネットワークは不安定	負の報酬	負の報酬	負の報酬

Fig.13

【図１４】

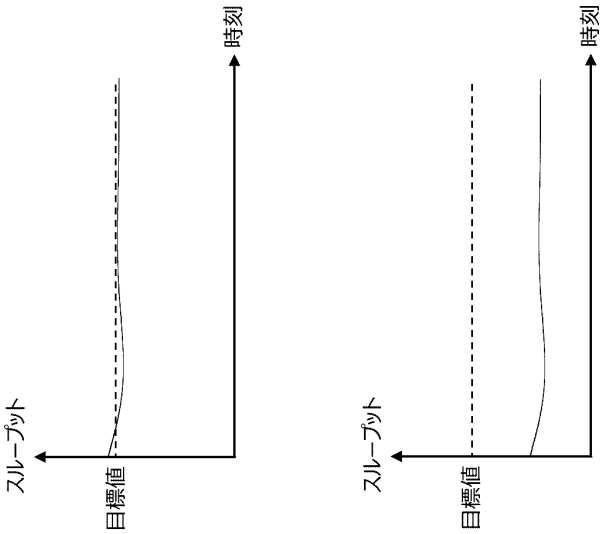

Fig.14A

Fig.14B

【図１５】

	スループット (T＜TH31)	スループット (TH31≦T)
ネットワークは安定	負の報酬	正の報酬
ネットワークは不安定	負の報酬	負の報酬

Fig.15

【図１６】

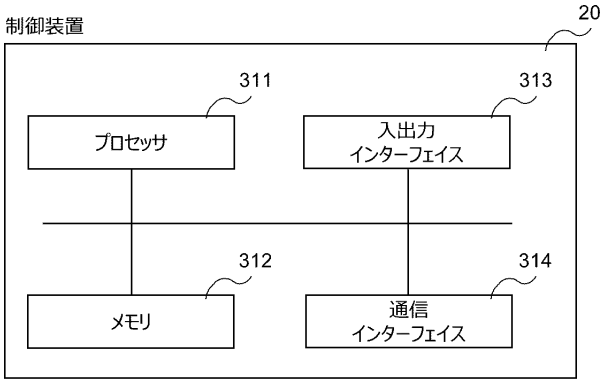

Fig.16

フロントページの続き

(56)参考文献 特開２０１３－１０６２０２（ＪＰ，Ａ）
国際公開第２０１９／１７６９９７（WO，Ａ１）
特開２０１９－０４１３３８（ＪＰ，Ａ）
米国特許出願公開第２０１９／０１４１１１３（ＵＳ，Ａ１）
特開２００９－０２７３０３（ＪＰ，Ａ）

(58)調査した分野 (Int.Cl.，ＤＢ名)

Ｈ０４Ｌ １２／００－１３／１８
Ｈ０４Ｌ ４１／００－４９／９０５７
Ｈ０４Ｌ ６１／００－６５／８０
Ｈ０４Ｌ ６９／００－６９／４０
Ｇ０６Ｎ ３／００－ ３／１２
Ｇ０６Ｎ ７／０８－９９／００