(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4月1日(01.04.2021)

WIPO | PCT

(10) 国際公開番号
WO 2021/059329 A1

(51) 国際特許分類:
G06F 16/9535 (2019.01) *G06F 16/955* (2019.01)

(21) 国際出願番号： PCT/JP2019/037283

(22) 国際出願日： 2019年9月24日(24.09.2019)

(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: 日本電気株式会社 (**NEC CORPORATION**) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 Tokyo (JP).

(72) 発明者: 川北 将 (**KAWAKITA, Masaru**); 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP).

(74) 代理人: 梶田 邦之, 外 (**KAJITA, Kuniyuki** et al.); 〒2110005 神奈川県川崎市中原区新丸子町９１５ 武蔵小杉フコク生命ビル４階 Kanagawa (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, 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,

(54) **Title:** INFORMATION COLLECTION DEVICE, INFORMATION COLLECTION METHOD, AND PROGRAM

(54) 発明の名称： 情報収集装置、情報収集方法、及びプログラム

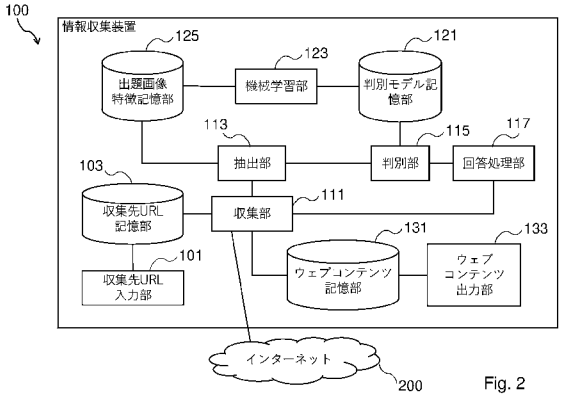

Fig. 2

100 Information collection device
101 Collection destination URL input unit
103 Collection destination URL storage unit
111 Collection unit
113 Extraction unit
115 Identification unit
117 Answer processing unit
121 Identification model storage unit
123 Machine learning unit
125 Question image feature storage unit
131 Web content storage unit
133 Web content output unit
200 Internet

(57) **Abstract:** [Problem] To efficiently collect web content accessible in accordance with an answer with a correct answer character string. [Solution] This information collection device is provided with: a collection unit 111 that collects first web content by using web address information; an extraction unit 113 that extracts, from the first web content, question image information obtained by applying an image effect to a correct answer character string for enabling access to second web content; and an identification unit 115 that identifies the correct answer character string from the question image information by using an identification model that is associated with the web address information and that is among two or more identification models for identifying character strings from an image.

WO 2021/059329 A1

［続葉有］

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：【課題】正解文字列の回答に応じてアクセス可能なウェブコンテンツの収集を効率よく行うこと。【解決手段】ウェブアドレス情報を用いて、第１のウェブコンテンツを収集する収集部１１１と、上記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出する抽出部１１３と、画像から文字列の判別を行うための２以上の判別モデルの中から、上記ウェブアドレス情報に関連付けられた判別モデルを用いて、上記出題画像情報から上記正解文字列を判別する判別部１１５と、を備える。

明　細　書

発明の名称 ： 情報収集装置、情報収集方法、及びプログラム

技術分野

[0001] 本発明は、ウェブコンテンツ情報の収集を行う情報収集装置、情報収集方法、及びプログラムに関する。

背景技術

[0002] 機械収集によるサーバ負荷の増大を抑止する等の目的で、ウェブサイトの閲覧者が人間であることを確認する認証方式が用いられている。このような認証方式の例として、反転チューリングテストの一種であるＣＡＰＴＣＨＡ（Completely Automated Public Turing test to tell Computers and Humans Apart）が知られている。例えば、特許文献１及び２には、このような反転チューリングテストを行うための装置が開示されている。

先行技術文献

特許文献

[0003] 特許文献1：特開２０１３－０６１９７１号公報

特許文献2：特開２０１４－１３０５９９号公報

発明の概要

発明が解決しようとする課題

[0004] 上述した特許文献１及び２などに開示されているＣＡＰＴＣＨＡを用いた出題画像は、例えばＯＣＲ（Optical Character Recognition/Reader）など、文字の視覚的な特徴に基づいた認識処理により、正解文字列を推定しうる。

[0005] しかしながら、例えば、アンダーグラウンドサイトなどにアクセスするのに課されている反転チューリングテストでは、より強力なアクセス制限を課すため、文字の機械的読み取りを妨げる画像効果が施される傾向にある。このような画像効果が施された文字列は、上述したような文字の視覚的な特徴に基づいた認識処理を用いて推定することが難しかった。このため、上述し

たアンダーグラウンドサイトなどの所定のウェブサイト内のコンテンツを効率よく収集することができなかった。

[0006] 本発明の目的は、正解文字列の回答に応じてアクセス可能なウェブコンテンツの収集を効率よく行うことが可能な情報収集装置、情報収集方法、及びプログラムを提供することにある。

課題を解決するための手段

[0007] 本発明の一つの態様によれば、情報収集装置は、ウェブアドレス情報を用いて、第１のウェブコンテンツを収集する収集部と、上記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出する抽出部と、画像から文字列の判別を行うための２以上の判別モデルの中から、上記ウェブアドレス情報に関連付けられた判別モデルを用いて、上記出題画像情報から上記正解文字列を判別する判別部と、を備え、上記２以上の判別モデルのそれぞれが、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである。

[0008] 本発明の一つの態様によれば、情報収集方法は、ウェブアドレス情報を用いて、第１のウェブコンテンツを収集することと、上記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出することと、画像から文字列の判別を行うための２以上の判別モデルの中から、上記ウェブアドレス情報に関連付けられた判別モデルを用いて、上記出題画像情報から上記正解文字列を判別することと、を備え、上記２以上の判別モデルのそれぞれが、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである。

[0009] 本発明の一つの態様によれば、プログラムは、ウェブアドレス情報を用いて、第１のウェブコンテンツを収集することと、上記第１のウェブコンテン

ツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出することと、画像から文字列の判別を行うための２以上の判別モデルの中から、上記ウェブアドレス情報に関連付けられた判別モデルを用いて、上記出題画像情報から上記正解文字列を判別することと、をコンピュータに実行させ、上記２以上の判別モデルのそれぞれが、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである。

発明の効果

[0010] 本発明の一つの態様によれば、１以上の搬送装置により対象物を適切に搬送することが可能になる。なお、本発明により、当該効果の代わりに、又は当該効果とともに、他の効果が奏されてもよい。

図面の簡単な説明

[0011] [図1]図１は、第１の実施形態に係る情報収集装置１００のハードウェア構成の例を示すブロック図である。

[図2]図２は、情報収集装置１００により実現される構成の例を示すブロック図である。

[図3]図３は、画像生成ルールのタイプの具体例を示す図である。

[図4]図４は、判別モデルを生成するための処理を概略的に示す図である。

[図5]図５は、判別モデル記憶部１２１により記憶される情報の具体例を示す図である。

[図6]図６は、第２の実施形態に係る情報収集装置１００の概略的な構成の例を示すブロック図である。

発明を実施するための形態

[0012] 以下、添付の図面を参照して本発明の実施形態を詳細に説明する。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。

[0013] 説明は、以下の順序で行われる。

[0014] ＜＜１．本発明の実施形態の概要＞＞

まず、本発明の実施形態の概要を説明する。

[0015] （１）技術的課題

機械収集によるサーバ負荷の増大を抑止する等の目的で、ウェブサイトの閲覧者が人間であることを確認する認証方式が用いられている。このような認証方式の例として、反転チューリングテストの一種であるＣＡＰＴＣＨＡ（Completely Automated Public Turing test to tell Computers and Humans Apart）が知られている。

[0016] 上述したＣＡＰＴＣＨＡを用いた出題画像は、例えばＯＣＲ（Optical Character Recognition/Reader）など、文字の視覚的な特徴に基づいた認識処理により、正解文字列を推定しうる。

[0017] しかしながら、例えば、アンダーグラウンドサイトなどにアクセスするのに課されている反転チューリングテストでは、より強力なアクセス制限を課すため、文字の機械的読み取りを妨げる画像効果が施される傾向にある。このような画像効果が施された文字列は、上述したような文字の視覚的な特徴に基づいた認識処理を用いて推定することが難しかった。このため、上述したアンダーグラウンドサイトなどの所定のウェブサイト内のコンテンツを効率よく収集することができなかった。

[0018] そこで、本実施形態では、正解文字列の回答に応じてアクセス可能なウェブコンテンツの収集を効率よく行うことを目的とする。

[0019] （２）技術的特徴

本発明の実施形態では、ウェブアドレス情報を用いて、第１のウェブコンテンツを収集し、上記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出し、背景画像を付加する処理を含む２以上の画像生成ルールの中から、上記ウェブアドレス情報に応じて、上記出題画像情報の生成に用いられる第１の画像生成ルールを推定し、上記第１の画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像に基づいた判別モデルを用いて、上記出題画像情報から上記正解文字列を判別する。

[0020] これにより、例えば、正解文字列の回答に応じてアクセス可能なウェブコンテンツの収集を効率よく行うことが可能になる。なお、上述した技術的特徴は本発明の実施形態の具体的な一例であり、当然ながら、本発明の実施形態は、上述した技術的特徴に限定されない。

[0021] ＜＜２．第１の実施形態＞＞

図１～図５を参照して、本発明が適用された第１の実施形態を説明する。

[0022] ＜２．１．情報収集装置１００の構成＞

図１は、第１の実施形態に係る情報収集装置１００のハードウェア構成の例を示すブロック図である。図１を参照すると、情報収集装置１００は、通信インタフェース２１、入出力部２２、演算処理部２３、メインメモリ２４、及び記憶部２５を備える。

[0023] 通信インタフェース２１は、外部の装置との間でデータを送受信する。例えば、通信インタフェース２１は、有線通信路を介して外部装置と通信する。

[0024] 演算処理部２３は、例えばＣＰＵ（Central Processing Unit）やＧＰＵ（Graphics Processing Unit）等である。メインメモリ２４は、例えばＲＡＭ（Random Access Memory）やＲＯＭ（Read Only Memory）等である。記憶部２５は、例えばＨＤＤ（Hard Disk Drive）、ＳＳＤ（Solid State Drive）、またはメモリカード等である。また、記憶部２５は、ＲＡＭやＲＯＭ等のメモ

リであってもよい。

[0025] 情報収集装置１００では、例えば記憶部２５に記憶された搬送制御用プログラムをメインメモリ２４に読み出して演算処理部２３により実行することにより、図２に示すような機能部が実現される。これらのプログラムをメインメモリ２４上に読み出してから実行してもよいし、メインメモリ２４上に読み出さずに実行してもよい。また、メインメモリ２４や記憶部２５は、情報収集装置１００が備える構成要素が保持する情報やデータを記憶する役割も果たす。

[0026] また、上述したプログラムは、様々なタイプの非一時的なコンピュータ可読媒体（non-transitory computer readable medium）を用いて格納され、コンピュータに供給することができる。非一時的なコンピュータ可読媒体は、様々なタイプの実体のある記録媒体（tangible storage medium）を含む。非一時的なコンピュータ可読媒体の例は、磁気記録媒体（例えば、フレキシブルディスク、磁気テープ、ハードディスクドライブ）、光磁気記録媒体（例えば、光磁気ディスク）、ＣＤ－ＲＯＭ（Compact Disc-ROM）、ＣＤ－Ｒ（CD-Recordable）、ＣＤ－Ｒ／Ｗ（CD-ReWritable）、半導体メモリ（例えば、マスクＲＯＭ、ＰＲＯＭ（Programmable ROM）、ＥＰＲＯＭ（Erasable PROM）、フラッシュＲＯＭ、ＲＡＭを含む。また、プログラムは、様々なタイプの一時的なコンピュータ可読媒体（transitory computer readable medium）によってコンピュータに供給されてもよい。一時的なコンピュータ可読媒体の例は、電気信号、光信号、及び電磁波を含む。一時的なコンピュータ可読媒体は、電線及び光ファイバ等の有線通信路、又は無線通信路を介して、プログラムをコンピュータに供給できる。

[0027] 表示装置２６は、ＬＣＤ（Liquid Crystal Display）、ＣＲＴ（Cathode Ray Tube）ディスプレイ、モニターのような、演算処理部２３により処理された描画データに対応する画面を表示する装置である。

[0028] 図２は、情報収集装置１００により実現される構成の例を示すブロック図である。

[0029] 図２を参照すると、情報収集装置１００は、収集先ＵＲＬ入力部１０１、及び収集先ＵＲＬ記憶部１０３を備える。また、情報収集装置１００は、収集部１１１、抽出部１１３、判別部１１５、及び回答処理部１１７を備える。さらに、情報収集装置１００は、判別モデル記憶部１２１、機械学習部１２３、及び出題画像特徴記憶部１２５を備える。これらの各々の機能部の具体的な動作ないし処理については後述する。

[0030] ＜２．２．技術的特徴＞

次に、第１の実施形態の技術的特徴を説明する。

[0031] 第１の実施形態によれば、情報収集装置１００（収集部１１１）は、ウェブアドレス情報を用いて、第１のウェブコンテンツを収集する。次に、情報収集装置１００（抽出部１１３）は、上記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出する。次に、情報収集装置１００（判別部１１５）は、画像から文字列の判別を行うための２以上の判別モデルの中から、上記ウェブアドレス情報に関連付けられた判別モデルを用いて、上記出題画像情報から上記正解文字列を判別する。ここで、上記２以上の判別モデルのそれぞれは、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである。

[0032] （１）第１のウェブコンテンツの収集

上記第１のウェブコンテンツの収集は、例えば次のようにして行われる。

[0033] まず、ユーザ又は管理システムは、収集先ＵＲＬ入力部１０１を用いて、収集したいコンテンツの場所を示すＵＲＬの集合を入力する。ＵＲＬの集合は、収集先ＵＲＬ記憶部１０３に格納される。なお、収集先ＵＲＬ入力部１０１は、情報収集装置１００に接続されたキーボード、外部記憶装置、外部ネットワークであってもよい。

[0034] 次に、収集部１１１は、収集先ＵＲＬ記憶部１０３に格納されているＵＲＬの集合のうち、１つのＵＲＬを上記ウェブアドレス情報として読み出す。

そして、収集部１１１は、インターネットにアクセスして、上記ウェブアドレス情報が示すウェブコンテンツ（上記第１のウェブコンテンツ）を取得した後、ＵＲＬ（上記ウェブアドレス情報）と上記第１のウェブコンテンツとのペアを、ウェブコンテンツ記憶部１３１に格納する。

[0035] さらに、収集部１１１は、上記第１のウェブコンテンツに含まれるＵＲＬを抽出して、当該抽出されたＵＲＬに再入力するよう構成されている。収集部１１１は、当該抽出されたＵＲＬが、例えばアンダーグラウンドサイトである場合を考慮して、隠匿オーバーレイネットワークにアクセスするために必要なプロキシ等のアクセス補助機能を用いてもよい。

[0036] （２）出題画像情報

上記出題画像情報の抽出は、例えば次のようにして行われる。

[0037] 例えば、抽出部１１３は、出題画像特徴記憶部１２５に記憶された情報を利用して、上記第１のウェブコンテンツから上記出題画像情報を抽出する。ここで、出題画像特徴記憶部１２５には、例えば収集先ＵＲＬ記憶部１０３に記憶された各々のＵＲＬによってアクセス可能なコンテンツから出題画像を抽出するための正規表現が記憶されている。

[0038] すなわち、抽出部１１３は、上記ウェブアドレス情報および収集部１１１により収集された上記第１のウェブコンテンツのペアと、出題画像特徴記憶部１２５が記憶するＵＲＬおよび出題画像を抽出するための正規表現とのペアを照合する。そして、抽出部１１３は、当該照合結果に従って、上記第１のウェブコンテンツから、上記出題画像情報を抽出することができる。

[0039] （３）画像生成ルール

図３は、画像生成ルールのタイプの具体例を示す図である。例えば、画像生成ルールは、図３に示すような４つのタイプに分けられる。

[0040] まず、第１のタイプの画像生成ルールに従って生成される出題画像３１は、例えば、背景および文字の色調が似ており文字が歪んでいる、という特徴を有する。また、第２のタイプの画像生成ルールに従って生成される出題画像３２は、例えば、背景である図形に含まれる文字を回答対象としており文

字が歪んでいる、という特徴を有する。また、第３のタイプの画像生成ルールに従って生成される出題画像３３ａ、３３ｂは、例えば、文字の配置が分散されており背景画像に文字が埋め込まれている、という特徴を有する。また、第４のタイプの画像生成ルールに従って生成される出題画像３４は、例えば、背景および文字の色調が似ており、文字の配置が分散されている、という特徴を有する。また、第５のタイプの画像生成ルールに従って生成される出題画像３５は、例えば、背景および文字の色調が似ており、文字の配置が分散されている、という特徴を有する。このような第１～第５のタイプの画像生成ルールは、例えばＣＡＰＴＣＨＡ（Completely Automated Public Turing test to tell Computers and Humans Apart）の出題ルールと捉えることができる。

[0041] また、第１～第５のタイプの画像生成ルールのそれぞれは、文字列に含まれる文字種を設定すること、文字列に含まれる文字数を設定すること、文字列を表示するための書体に関する情報を設定すること、及び上記背景画像に関する情報を設定することを含む。このような複数の設定により、正解文字列から上述した特徴を有する出題画像情報を生成することができる。

[0042] （４）判別モデル

例えば、上記２以上の判別モデルのそれぞれは、次に示すようにして、例えば機械学習部１２３によって生成される。図４は、判別モデルを生成するための処理を概略的に示す図である。

[0043] 図４を参照すると、ステップＳ４０１において、機械学習部１２３は、収集先ＵＲＬ記憶部１０３に記憶されている一つのウェブアドレス（以下、対象ウェブアドレスとも呼ぶ。）に関連付けれられた画像生成ルール及び画像生成ライブラリーコードのペアを取得してステップＳ４０３に進む。

[0044] 例えば、機械学習部１２３は、出題画像特徴記憶部１２５にアクセスすることにより、対象ウェブアドレスに関連付けられた画像生成ルール及び画像生成ライブラリーコードのペアを取得してもよい。この場合、出題画像特徴記憶部１２５は、収集先ＵＲＬ記憶部１０３に記憶されているウェブアドレ

スごとに、画像生成ルール及び画像生成ライブラリーコードのペアを関連付けて記憶する。このような関連付けは、例えばユーザ操作によって行われる。

[0045] ステップＳ４０３において、機械学習部１２３は、ステップＳ４０１により取得された画像生成ライブラリーコードを反復実行することにより、学習サンプルを生成する。

[0046] 具体的に、ステップＳ４０３において、機械学習部１２３は、対象ウェブアドレスに関連付けられた画像生成ルールに従って、候補正解文字列が取り得る文字種及び文字数を設定し、当該設定した条件に従って候補正解文字列をランダムに生成する。一例として、文字種として英数字が設定され、文字数として６～８が設定される。

[0047] また機械学習部１２３は、画像生成ルールに従って、文字列を表示するための書体に関する情報（フォント、文字の太さ、文字の色など）、及び背景画像に関する情報（模様、模様の太さ、模様の色など）を設定し、当該設定した条件に従って、各々の候補正解文字列に対応する候補出題画像を生成する。

[0048] ステップＳ４０５において、機械学習部１２３は、ステップＳ４０３により生成した学習サンプル（複数の正解文字列、及び複数の候補出題画像）を学習用データとして、判別モデルを生成して、ステップＳ４０７に進む。ここで、判別モデルは、任意の機械学習アルゴリズムにより得られる。例えば、機械学習のアルゴリズムは、サポートベクターマシン、又はディープラーニングであってもよい。判別モデルは、例えば任意の画素数から構成される画像情報（各画素の輝度情報および色差情報）と候補正解文字列との相関を評価するための評価関数などを含む。このような評価関数を用いた評価結果に基づいて、画像から正解文字列を判別することができる。

[0049] ステップＳ４０７において、機械学習部１２３は、判別モデルの判別精度が閾値以上であるか否かを判断し、閾値以上であれば（Ｓ４０７：Ｙｅｓ）ステップＳ４０９に進み、閾値未満であれば（Ｓ４０７：Ｎｏ）ステップＳ

４０３に戻ってステップＳ４０３およびステップＳ４０５を繰り返す。

[0050] ステップＳ４０９において、機械学習部１２３は、対象ウェブアドレスに関連付けて、ステップＳ４０５により生成した判別モデルを判別モデル記憶部１２１に記憶して、ステップＳ４１１に進む。

[0051] 図５は、判別モデル記憶部１２１により記憶される情報の具体例を示す図である。図５を参照すると、判別モデル記憶部１２１は、２以上の判別モデルのそれぞれにウェブアドレス情報が関連付けられたデータテーブル５００を有する。

[0052] ステップＳ４１１において、機械学習部１２３は、収集先ＵＲＬ記憶部１０３に記憶されている全てのウェブアドレスに対応する全ての判別モデルを生成したか否かを判断し、全ての判別モデルが生成された場合（Ｓ４１１：Ｙｅｓ）には図４に示す処理を終了し、全ての判別モデルが生成されていない場合（Ｓ４１１：Ｎｏ）にはステップＳ４０１に戻って、ステップＳ４０１～Ｓ４０９の処理を繰り返す。

[0053] 上記図４に示す処理に従って、機械学習部１２３は、判別モデルを生成することができる。

[0054] （５）判別モデルを用いた正解文字列の判別

判別部１１５は、判別モデル記憶部１２１を参照して、上記ウェブアドレス情報に関連付けられている判別モデルを特定し、特定した判別モデルを用いて、上記出題画像情報から上記正解文字列を判別する。例えば、図５に示すデータテーブル５００を参照すると、上記ウェブアドレス情報がウェブアドレスＵＲＬ１である場合、判別部１１５は、ウェブアドレスＵＲＬ１に関連付けられた判別モデル１を用いて、上記出題画像情報から上記正解文字列を判別することができる。

[0055] （６）回答処理

回答処理部１１７は、上述のようにして判別された正解文字列を用いて上記出題画像情報に対する回答を行う。この場合、収集部１１１は、上記回答に応じて上記第２のウェブコンテンツの収集を更に行う。

[0056] すなわち、収集部１１１は、上記回答情報を、インターネット２００を介して、上記ウェブアドレス情報が示すサーバ装置に送信する。そして、収集部１１１は、当該サーバ装置から応答として、ログイン成功の情報が送信される。ログイン成功の情報は、たとえばＳｅｔ－Ｃｏｏｋｉｅヘッダである。なお、ログイン成功の情報は、Ｓｅｔ－Ｃｏｏｋｉｅヘッダに限らず、例えばＳｅｔ－Ｃｏｏｋｉｅ２ヘッダなどの他の方式のＣｏｏｋｉｅヘッダであってもよい。その後、収集部１１１は、このログイン成功の情報を利用して、上記第２のウェブコンテンツを収集してウェブコンテンツ記憶部１３１に記憶する。

[0057] （７）閲覧処理

ウェブコンテンツ出力部１３３は、例えばユーザからの要求に応じて上記第２のコンテンツに関する情報を出力する。例えば上記第２のコンテンツに関する情報は、情報収集装置１００が有する表示装置２６上に表示される。これにより、ユーザは、例えば出題画像情報を解読して正解文字列の回答を行うこと無く、効率よく上記第２のコンテンツに関する情報を閲覧することができる。例えば、上記第２のコンテンツにアンダーグラウンドサイト上での交流情報が含まれている場合には、ユーザは、情報収集装置１００にアクセスするだけで効率よくこれらの交流情報を収集でき、セキュリティ対策や防犯対策などに利用することができる。

[0058] ＜＜３．第２の実施形態＞＞

続いて、図６を参照して、本発明の第２の実施形態を説明する。上述した第１の実施形態は、具体的な実施形態であるが、第２の実施形態は、より一般化された実施形態である。

[0059] ＜３．１．情報収集装置１００の構成＞

図６は、第２の実施形態に係る情報収集装置１００の概略的な構成の例を示すブロック図である。図６を参照すると、情報収集装置１００は、収集部１５０、抽出部１６０、及び判別部１７０を備える。

[0060] 収集部１５０、抽出部１６０、及び判別部１７０は、１つ以上のプロセッ

サと、メモリ（例えば、不揮発性メモリ及び／若しくは揮発性メモリ）並びに／又はハードディスクとにより実装されてもよい。収集部１５０、抽出部１６０、及び判別部１７０は、同一のプロセッサにより実装されてもよく、別々に異なるプロセッサにより実装されてもよい。上記メモリは、上記１つ以上のプロセッサ内に含まれていてもよく、又は、上記１つ以上のプロセッサ外にあってもよい。

[0061] ＜３．２．技術的特徴＞

第２の実施形態に係る技術的特徴を説明する。

[0062] 第２の実施形態によれば、情報収集装置１００（収集部１５０）は、ウェブアドレス情報を用いて、第１のウェブコンテンツを収集する。次に、情報収集装置１００（抽出部１６０）は、上記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出する。次に、情報収集装置１００（判別部１７０）は、画像から文字列の判別を行うための２以上の判別モデルの中から、上記ウェブアドレス情報に関連付けられた判別モデルを用いて、上記出題画像情報から上記正解文字列を判別する。ここで、上記２以上の判別モデルのそれぞれは、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである。

[0063] －第１の実施形態との関係

一例として、第２の実施形態の収集部１５０、抽出部１６０、及び判別部１７０は、それぞれ、第１の実施形態の収集部１１１、抽出部１１３、及び判別部１１５の動作を行ってもよい。この場合に、第１の実施形態についての説明は、第２の実施形態にも適用されうる。

[0064] なお、第２の実施形態は、この例に限定されない。

[0065] 以上、第２の実施形態を説明した。第２の実施形態によれば、例えば、正解文字列の回答に応じてアクセス可能なウェブコンテンツの収集を効率よく行うことが可能になる。

[0066] ＜＜４．他の実施形態＞＞

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。

[0067] 例えば、本明細書に記載されている処理におけるステップは、必ずしもシーケンス図に記載された順序に沿って時系列に実行されなくてよい。例えば、処理におけるステップは、シーケンス図として記載した順序と異なる順序で実行されても、並列的に実行されてもよい。また、処理におけるステップの一部が削除されてもよく、さらなるステップが処理に追加されてもよい。

[0068] また、本明細書において説明した情報収集装置の構成要素（例えば、収集部、抽出部、及び／又は判別部）を備える装置（例えば、情報収集装置を構成する複数の装置（又はユニット）のうちの１つ以上の装置（又はユニット）、又は上記複数の装置（又はユニット）のうちの１つのためのモジュール）が提供されてもよい。また、上記構成要素の処理を含む方法が提供されてもよく、上記構成要素の処理をプロセッサに実行させるためのプログラムが提供されてもよい。また、当該プログラムを記録したコンピュータに読み取り可能な非一時的記録媒体（Non-transitory computer readable medium）が提供されてもよい。当然ながら、このような装置、モジュール、方法、プログラム、及びコンピュータに読み取り可能な非一時的記録媒体も本発明に含まれる。

[0069] 上記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。

[0070]（付記１）

ウェブアドレス情報を用いて、第１のウェブコンテンツを収集する収集部と、

前記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出する抽出部と

、

画像から文字列の判別を行うための２以上の判別モデルの中から、前記ウェブアドレス情報に関連付けられた判別モデルを用いて、前記出題画像情報から前記正解文字列を判別する判別部と、

を備え、

前記２以上の判別モデルのそれぞれが、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである、情報収集装置。

[0071] （付記２）

前記画像生成ルールは、文字列に含まれる文字種を設定することを更に含む、付記１記載の情報収集装置。

[0072] （付記３）

前記画像生成ルールは、文字列に含まれる文字数を設定することを更に含む、付記１又は２記載の情報収集装置。

[0073] （付記４）

前記２以上の画像生成ルールのそれぞれは、文字列を表示するための書体に関する情報を設定することを更に含む、付記１乃至３のうち何れか１項記載の情報収集装置。

[0074] （付記５）

前記２以上の画像生成ルールのそれぞれは、前記背景画像に関する情報を設定することを更に含む、付記１乃至４のうち何れか１項記載の情報収集装置。

[0075] （付記６）

前記判別部は、前記２以上の判別モデルのそれぞれにウェブアドレス情報が関連付けられたデータテーブルを参照して、前記ウェブアドレス情報に関連付けられた判別モデルを特定する、付記１乃至５のうち何れか１項記載の情報収集装置。

[0076] （付記７）

前記判別された正解文字列を用いて前記出題画像情報に対する回答を行う回答処理部を更に備え、

前記収集部は、前記回答に応じて前記第２のウェブコンテンツの収集を更に行う、付記１乃至６のうち何れか１項記載の情報収集装置。

[0077] （付記８）

ユーザからの要求に応じて前記第２のウェブコンテンツに関する情報を出力するウェブコンテンツ出力部を更に備える、付記７記載の情報収集装置。

[0078] （付記９）

ウェブアドレス情報を用いて、第１のウェブコンテンツを収集することと、

前記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出することと、

画像から文字列の判別を行うための２以上の判別モデルの中から、前記ウェブアドレス情報に関連付けられた判別モデルを用いて、前記出題画像情報から前記正解文字列を判別することと、

を備え、

前記２以上の判別モデルのそれぞれが、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである、情報収集方法。

[0079] （付記１０）

ウェブアドレス情報を用いて、第１のウェブコンテンツを収集することと、

前記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出することと、

画像から文字列の判別を行うための２以上の判別モデルの中から、前記ウェブアドレス情報に関連付けられた判別モデルを用いて、前記出題画像情報

から前記正解文字列を判別することと、をコンピュータに実行させ、

前記２以上の判別モデルのそれぞれが、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである、プログラム。

産業上の利用可能性

[0080] ウェブサイトにアクセスしてウェブコンテンツを収集する情報収集装置において、正解文字列の回答に応じてアクセス可能なウェブコンテンツの収集を効率よく行うことが可能になる。

符号の説明

[0081] １００　情報収集装置

１０１　収集先ＵＲＬ入力部

１０３　収集先ＵＲＬ記憶部

１１１、１５０　収集部

１１３、１６０　抽出部

１１５、１７０　判別部

１１７　回答処理部

１２１　判別モデル記憶部

１２３　機械学習部

１２５　出題画像特徴記憶部

１３１　ウェブコンテンツ記憶部

１３３　ウェブコンテンツ出力部

２００　インターネット

請求の範囲

[請求項1] ウェブアドレス情報を用いて、第１のウェブコンテンツを収集する収集部と、

前記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出する抽出部と、

画像から文字列の判別を行うための２以上の判別モデルの中から、前記ウェブアドレス情報に関連付けられた判別モデルを用いて、前記出題画像情報から前記正解文字列を判別する判別部と、

を備え、

前記２以上の判別モデルのそれぞれが、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである、情報収集装置。

[請求項2] 前記画像生成ルールは、文字列に含まれる文字種を設定することを更に含む、請求項１記載の情報収集装置。

[請求項3] 前記画像生成ルールは、文字列に含まれる文字数を設定することを更に含む、請求項１又は２記載の情報収集装置。

[請求項4] 前記２以上の画像生成ルールのそれぞれは、文字列を表示するための書体に関する情報を設定することを更に含む、請求項１乃至３のうち何れか１項記載の情報収集装置。

[請求項5] 前記２以上の画像生成ルールのそれぞれは、前記背景画像に関する情報を設定することを更に含む、請求項１乃至４のうち何れか１項記載の情報収集装置。

[請求項6] 前記判別部は、前記２以上の判別モデルのそれぞれにウェブアドレス情報が関連付けられたデータテーブルを参照して、前記ウェブアドレス情報に関連付けられた判別モデルを特定する、請求項１乃至５のうち何れか１項記載の情報収集装置。

[請求項7] 前記判別された正解文字列を用いて前記出題画像情報に対する回答を行う回答処理部を更に備え、

前記収集部は、前記回答に応じて前記第２のウェブコンテンツの収集を更に行う、請求項１乃至６のうち何れか１項記載の情報収集装置。

[請求項8] ユーザからの要求に応じて前記第２のウェブコンテンツに関する情報を出力するウェブコンテンツ出力部を更に備える、請求項７記載の情報収集装置。

[請求項9] ウェブアドレス情報を用いて、第１のウェブコンテンツを収集することと、

前記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出することと、

画像から文字列の判別を行うための２以上の判別モデルの中から、前記ウェブアドレス情報に関連付けられた判別モデルを用いて、前記出題画像情報から前記正解文字列を判別することと、

を備え、

前記２以上の判別モデルのそれぞれが、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである、情報収集方法。

[請求項10] ウェブアドレス情報を用いて、第１のウェブコンテンツを収集することと、

前記第１のウェブコンテンツから、第２のウェブコンテンツへのアクセスのための正解文字列に画像効果が施された出題画像情報を抽出することと、

画像から文字列の判別を行うための２以上の判別モデルの中から、前記ウェブアドレス情報に関連付けられた判別モデルを用いて、前記

出題画像情報から前記正解文字列を判別することと、をコンピュータに実行させ、

前記２以上の判別モデルのそれぞれが、背景画像を付加する処理を含む画像生成ルールに従って複数の候補正解文字列から生成された複数の候補出題画像を学習用データとして機械学習された学習済みモデルである、プログラム。

[図1]

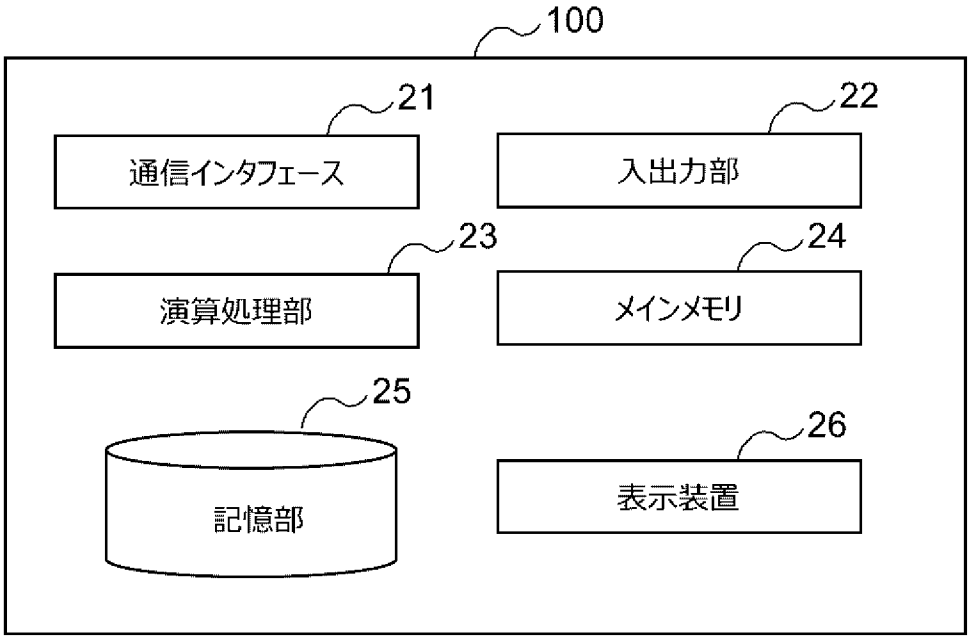

Fig.1

[図2]

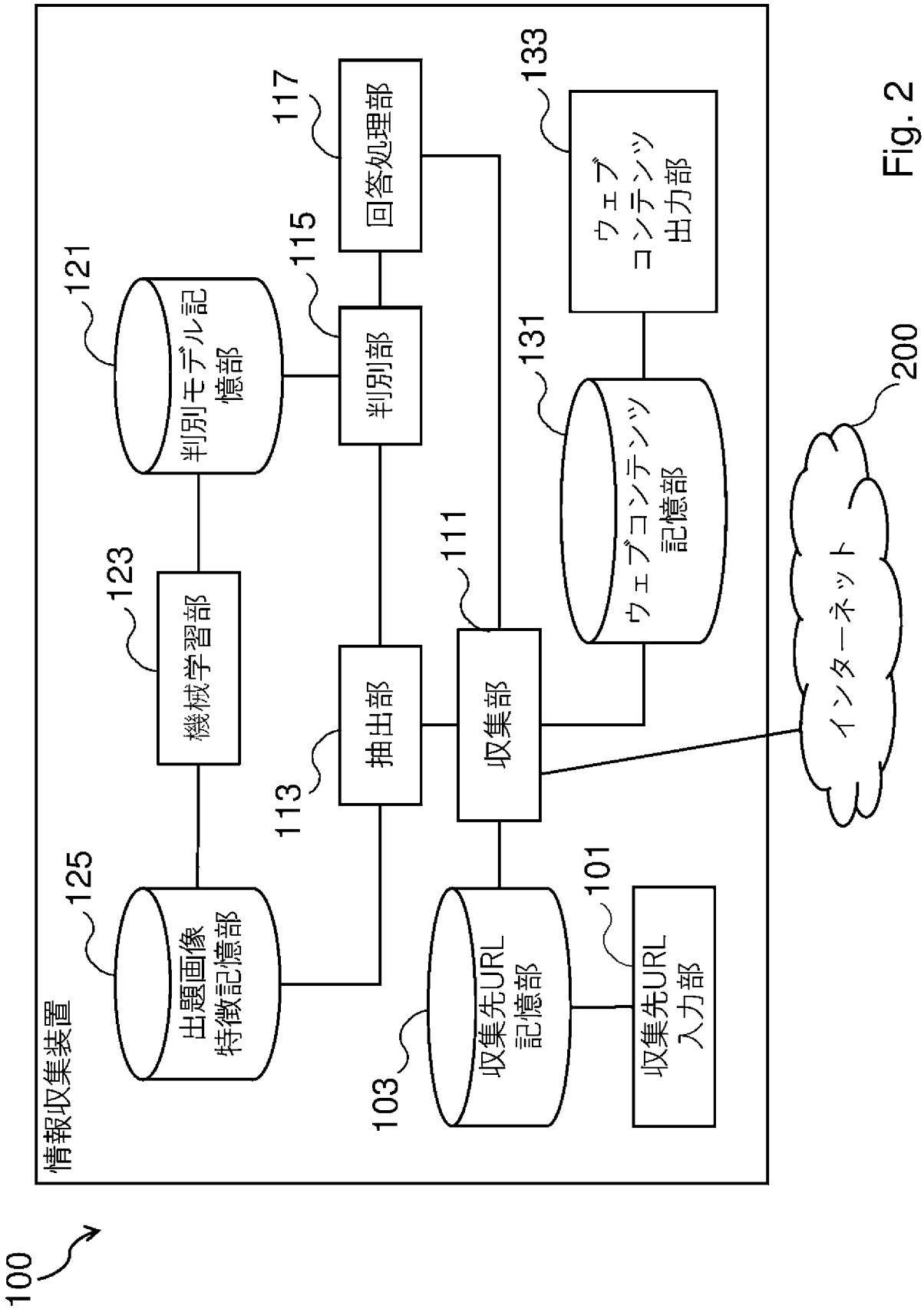
100
情報収集装置
125
出題画像
特徴記憶部
123
機械学習部
121
判別モデル記
憶部
113
抽出部
115
判別部
117
回答処理部
103
収集先URL
記憶部
101
収集先URL
入力部
111
収集部
131
ウェブコンテンツ
記憶部
133
ウェブ
コンテンツ
出力部
200
インターネット

Fig. 2

[図3]

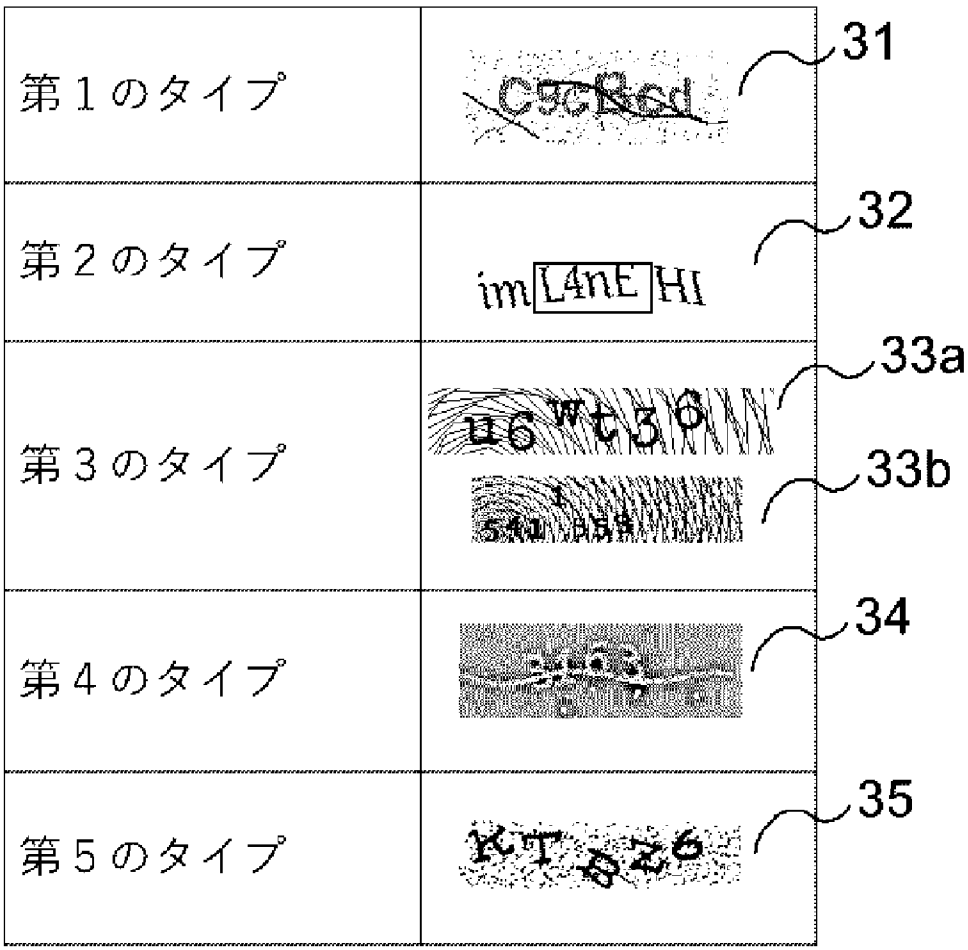

Fig. 3

[図4]

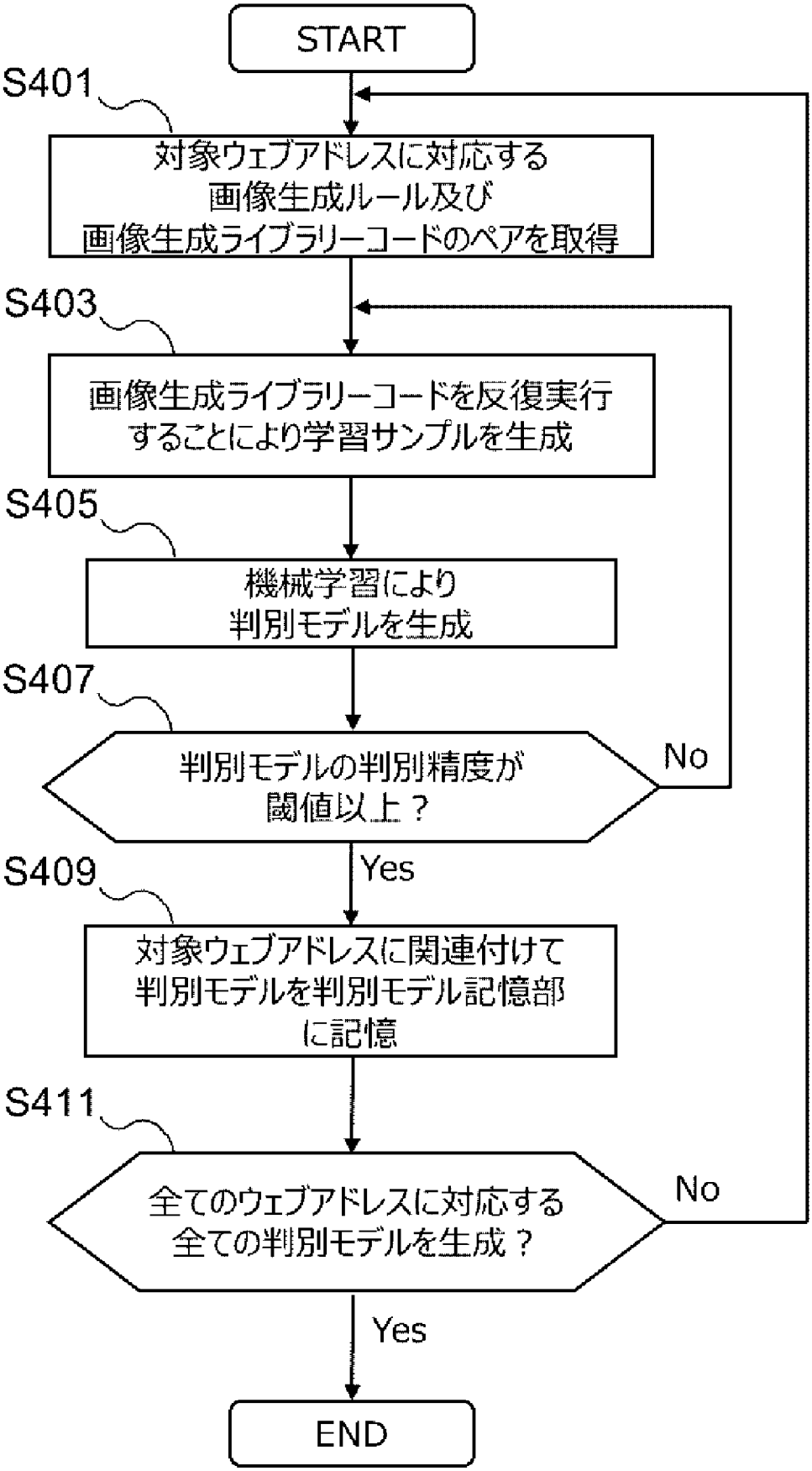
START
S401
対象ウェブアドレスに対応する
画像生成ルール及び
画像生成ライブラリーコードのペアを取得
S403
画像生成ライブラリーコードを反復実行
することにより学習サンプルを生成
S405
機械学習により
判別モデルを生成
S407
判別モデルの判別精度が
閾値以上？
No
Yes
S409
対象ウェブアドレスに関連付けて
判別モデルを判別モデル記憶部
に記憶
S411
全てのウェブアドレスに対応する
全ての判別モデルを生成？
No
Yes
END

Fig. 4

[図5]

500

ウェブアドレス情報	判別モデル
ウェブアドレスＵＲＬ１	判別モデル１
ウェブアドレスＵＲＬ２	判別モデル２
ウェブアドレスＵＲＬ３	判別モデル３
ウェブアドレスＵＲＬ４	判別モデル４
ウェブアドレスＵＲＬ５	判別モデル５
ウェブアドレスＵＲＬ６	判別モデル６
・ ・ ・	・ ・ ・

Fig. 5

[図6]

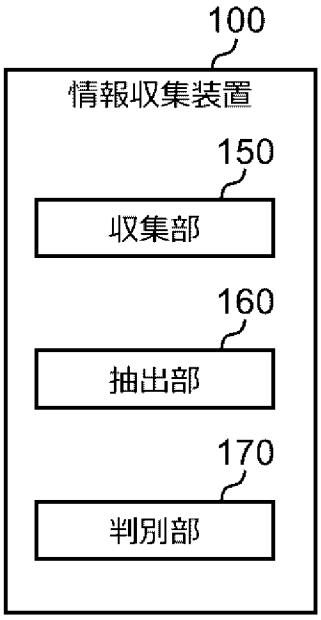

Fig.6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/037283

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl. G06F16/9535(2019.01)i, G06F16/955(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. G06F16/9535, G06F16/955

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	岩永 崇裕, 中尾 彰宏, ダークウェブマーケット統計分析・可視化のためのデータ自動収集とレポジトリ生成の提案, 信学技報, 04 July 2018, vol. 118, no. 124, pp. 25-30, [online], [retrieved on 16 August 2018], Internet: <URL: https://www.ieice.org/ken/user/index.php?cmd=download&p=X3RP&t=IEICE-NS&l=865f03745eb1304541dc28fd69a335a6c45686464a48b713c4c23bc78d079e3a&lang=>, ISSN 2432-6380, particularly, p. 27, right column, l. 1 to p. 30, right column, l. 2, fig. 2, (IWANAGA, Takahiro, NAKAO, Akihiro, Proposal of automatic data collection and repository generation for statistical analysis and visualization of dark web markets, IEICE Technical Report)	1-10

☒ 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	Date of mailing of the international search report
02.12.2019	10.12.2019

Name and mailing address of the ISA/	Authorized officer
Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/037283

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	GEORGE, D. et al., A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs, Science, THE AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE [online], 08 December 2017, [retrieved on 29 October 2019], Internet: <URL: https://science.sciencemag.org/content/358/6368/eaag2612.full>, particularly, pp. 1, 4-8	1-10

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１９／０３７２８３

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F16/9535(2019.01)i, G06F16/955(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.　G06F16/9535, G06F16/955

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	岩永　崇裕，　中尾　彰宏，ダークウェブマーケット統計分析・可視化のためのデータ自動収集とレポジトリ生成の提案，信学技報，2018.07.04，第118巻，第124号，p. 25-30，[オンライン]，[検索日 2018.8.16]，インターネット:<URL:https://www.ieice.org/ken/user/index.php?cmd=download&p=X3RP&t=IEICE-NS&l=865f03745eb1304541dc28fd69a335a6c45686464a48b713c4c23bc78d079e3a&lang=>，ISSN 2432-6380，特にp.27右欄第1行-p.30右欄第2行，図2	1-10

☑　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
０２．１２．２０１９	１０．１２．２０１９

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）　５Ｎ　６３０２
三橋　竜太郎
電話番号　０３－３５８１－１１０１　内線　３５８６

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
Y	GEORGE Dileep et al., A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs, Science, the American Association for the Advancement of Science [オンライン], 2017.12.08, [検索日 2019.10.29], インターネット:<URL: https://science.sciencemag.org/content/358/6368/eaag2612.full>, 特にp.1, 4-8ページ	1-10