(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5571035号
(P5571035)
(45)発行日　平成26年8月13日(2014.8.13)　(24)登録日　平成26年7月4日(2014.7.4)

(51)Int.Cl.　F I
G06F 21/56 (2013.01)　G06F 21/00 156A
G06F 21/00 156D

請求項の数 5　(全 20 頁)

(21)出願番号　特願2011-120293 (P2011-120293)
(22)出願日　平成23年5月30日 (2011.5.30)
(65)公開番号　特開2012-248058 (P2012-248058A)
(43)公開日　平成24年12月13日 (2012.12.13)
審査請求日　平成25年7月17日 (2013.7.17)

(73)特許権者　000004226
日本電信電話株式会社
東京都千代田区大手町一丁目５番１号
(74)代理人　100089118
弁理士　酒井　宏明
(74)代理人　100112656
弁理士　宮田　英毅
(72)発明者　岩村　誠
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　伊藤　光恭
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

審査官　戸島　弘詩

最終頁に続く

(54)【発明の名称】特定装置、特定方法及び特定プログラム

(57)【特許請求の範囲】
【請求項１】
パックされたプログラムコードをアンパックすることで得られるオリジナルコードから、少なくとも間接分岐命令と解釈可能な箇所を示すアドレスである間接分岐箇所と、直接分岐命令と解釈可能な箇所である直接分岐箇所を抽出する抽出部と、
前記抽出部により抽出された箇所間の分岐関係を解析することで、前記間接分岐箇所が直接的又は間接的に分岐先となる前記直接分岐箇所を識別する識別部と、
前記抽出部により抽出された箇所各々について分岐命令となる第１の確率を算出し、該箇所各々について算出した該第１の確率に基づいて、前記間接分岐箇所の先頭と、該間接分岐箇所が分岐先となる前記識別部により識別された前記直接分岐箇所の先頭とのうち、少なくとも一つの箇所の先頭が機械語命令の先頭となる第２の確率を算出する算出部と、
前記算出部により算出された第２の確率が閾値以上である場合に、前記間接分岐箇所からの分岐先を示す分岐先アドレスが格納される分岐先アドレス格納場所を特定する特定部と
を備えたことを特徴とする特定装置。
【請求項２】
前記分岐先アドレス格納場所が同一となる前記間接分岐箇所が複数抽出された場合に、前記算出部は、前記抽出部により抽出された箇所各々について前記第１の確率を算出し、該箇所各々について算出した該分岐命令となる該第１の確率に基づいて、前記分岐先アドレス格納場所が同一となる前記間接分岐箇所各々の先頭と、該間接分岐箇所各々のうちい

ずれかが分岐先となる前記識別部により識別された前記直接分岐箇所の先頭とのうち、少なくとも一つの箇所の先頭が機械語命令の先頭となる前記第２の確率を算出することを特徴とする請求項１に記載の特定装置。

【請求項３】

前記算出部は、各箇所について、前記抽出部により抽出された前記箇所各々に該当する前記第１の確率を用いて、該当箇所を含む、他の命令との分岐元又は分岐先の分岐関係を示す分岐ツリー内に存在するすべての箇所が分岐命令とならない第３の確率を１から減算することで、前記第２の確率をそれぞれ算出することを特徴とする請求項１または２に記載の特定装置。

【請求項４】

特定装置で実行される特定方法であって、

前記特定装置の抽出部が、パックされたプログラムコードをアンパックすることで得られるオリジナルコードから、少なくとも間接分岐命令と解釈可能な箇所を示すアドレスである間接分岐箇所と、直接分岐命令と解釈可能な箇所である直接分岐箇所を抽出して記憶部に格納する抽出工程と、

前記特定装置の識別部が、前記記憶部に記憶された箇所であって、前記抽出工程により抽出された箇所間の分岐関係を解析することで、前記間接分岐箇所が直接的又は間接的に分岐先となる前記直接分岐箇所を識別する識別工程と、

前記特定装置の算出部が、前記記憶部に記憶された箇所であって、前記抽出工程により抽出された箇所各々について分岐命令となる第１の確率を算出し、該箇所各々について算出した該第１の確率に基づいて、前記間接分岐箇所の先頭と、該間接分岐箇所が分岐先となる前記識別工程により識別された前記直接分岐箇所の先頭とのうち、少なくとも一つの箇所の先頭が機械語命令の先頭となる第２の確率を算出する算出工程と、

前記特定装置の特定部が、前記算出工程により算出された第２の確率が閾値以上である場合に、前記間接分岐箇所からの分岐先を示す分岐先アドレスが格納される分岐先アドレス格納場所を特定する特定工程と

を含んだことを特徴とする特定方法。

【請求項５】

コンピュータを請求項１～３のいずれか一つに記載の特定装置として機能させるための特定プログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、特定装置、特定方法及び特定プログラムに関する。

【背景技術】

【０００２】

近時、コンピュータウィルス等の悪意あるソフトウェアに対する対策が不可欠となっている。コンピュータウィルス等の悪意あるソフトウェアとは、例えば、マルウェアが該当する。マルウェアには、パッカーと呼ばれるツールにより、機械語のオリジナルコードに対して解読を困難にするため隠蔽処理であるパッキングが施される。オリジナルコードにパッキングが施されることで、アンチウィルスソフトのパターンマッチング機構を回避され、マルウェアの解析が困難となる。また、プログラムの解析を困難にする機構を持つパッカーもある。なお、オリジナルコードを復元する処理を「アンパッキング」とも称する。プログラムの解析を困難にする機構とは、例えば、Ａｎｔｉ－ＤｅｂｕｇやＡｎｔｉ－ＶＭなどがある。

【０００３】

また、実行可能形式のファイルを受け付け、オリジナルコードを隠蔽しつつも実行可能形式を保ったファイルを出力するパッカーであるランタイムパッカーがある。ランタイムパッカーによりパッキングされたプログラムは、オリジナルコードを復元し、通常はローダが行う動的ライブラリのリンク処理等を実施した後に、オリジナルコードのエントリポ

イントへ処理を渡す。
【０００４】
ここで、マルウェアの脅威を把握する際には、オリジナルコードを抽出し、抽出したオリジナルコードから外部関数を呼び出すときに使用される外部関数のアドレスが格納される格納場所であるインポートアドレス格納場所を特定する必要がある。なお、外部関数とは、例えば、Ｗｉｎ３２ ＡＰＩ（Ａｐｐｌｉｃａｔｉｏｎ Ｐｒｏｇｒａｍ Ｉｎｔｅｒｆａｃｅ）、ＤＬＬ（Ｄｙｎａｍｉｃ Ｌｉｎｋ Ｌｉｂｒａｒｙ）などがある。
【０００５】
なお、インポートアドレス格納場所を特定する手法として、逆アセンブル結果に基づき、オリジナルコード領域から間接ｃａｌｌ命令を抽出し、抽出した間接ｃａｌｌ命令により示されるメモリ領域をインポートアドレス格納場所として特定する特定手法がある。
【先行技術文献】
【特許文献】
【０００６】
【特許文献１】特開２０１０－９２１７９号公報
【非特許文献】
【０００７】
【非特許文献１】Ｄａｎｉｅｌ Ｑｕｉｓｔ, Ｌｏｒｉｅ Ｌｉｅｂｒｏｃｋ, Ｊｏｓｈｕａ Ｎｅｉｌ, Ｉｍｐｒｏｖｉｎｇ ａｎｔｉｖｉｒｕｓ ａｃｃｕｒａｃｙ ｗｉｔｈ ｈｙｐｅｒｖｉｓｏｒ ａｓｓｉｓｔｅｄ ａｎａｌｙｓｉｓ, Ｊｏｕｒｎａｌ ｉｎ Ｃｏｍｐｕｔｅｒ Ｖｉｒｏｌｏｇｙ （Ｐｕｂｌｉｓｈｅｄ ｏｎｌｉｎｅ： ６ Ａｐｒｉｌ ２０１０）［ｏｎｌｉｎｅ］、［２０１１年5月１６日検索］、［インターネット］（ＵＲＬ：ｈｔｔｐ：／／ｃｓｒ．ｌａｎｌ．ｇｏｖ／ｖｅｒａ／ｊｃｖ－ｑｕｉｓｔ－ｌｉｅｂｒｏｃｋ－ｎｅｉｌ．ｐｄｆ）
【発明の概要】
【発明が解決しようとする課題】
【０００８】
しかしながら、上述の従来の特定手法では、逆アセンブル結果が不正確である場合に、誤ったインポートアドレス格納場所が特定されるという課題がある。例えば、データ部分を誤って間接ｃａｌｌ命令として抽出されると、誤ったインポートアドレス格納場所が特定される。
【０００９】
すなわち、オリジナルコードから間接ｃａｌｌ命令を抽出し、抽出した間接ｃａｌｌ命令により示されるメモリ領域をインポートアドレス格納場所とする上述の従来の特定手法では、オリジナルコードに対応する正確な逆アセンブル結果が必要となる。ここで、マルウェア作者は、アンチウィルスソフトベンダ等による解析や対策から逃れることを目的として、ソースコードや逆アセンブルに要するシンボル情報などを非公開とすることがある。この場合、正確な逆アセンブル結果が得られず、誤ったインポートアドレス格納場所を特定されることがある。
【００１０】
開示の実施形態は、上述に鑑みてなされたものであって、適切なインポートアドレス格納場所を特定可能となる特定装置、特定方法及び特定プログラムを提供することを目的とする。
【課題を解決するための手段】
【００１１】
開示する特定装置は、一つの態様において、抽出部と、算出部と、特定部とを有する。抽出部は、パックされたプログラムコードをアンパックすることで得られるオリジナルコードから、少なくとも間接分岐命令と解釈可能な箇所を示すアドレスである間接分岐箇所を抽出する。算出部は、前記抽出部により抽出された箇所が分岐命令となる第１の確率を算出し、算出した第１の確率に基づいて、抽出された該箇所の先頭が機械語命令の先頭と

なる第２の確率を算出する。特定部は、前記算出部により算出された第２の確率が閾値以上である場合に、前記間接分岐箇所からの分岐先を示す分岐先アドレスが格納される分岐先アドレス格納場所を特定する。

【発明の効果】

【００１２】

開示する特定装置の一つの態様によれば、適切なインポートアドレス格納場所を特定可能となるという効果を奏する。

【図面の簡単な説明】

【００１３】

【図１】図１は、実施例１における特定装置の構成の一例を示すブロック図である。

【図２】図２は、実施例１における特定装置に入力されるオリジナルコードの一部を示す図である。

【図３】図３は、実施例１における暫定分岐情報格納テーブルに記憶された情報の一例を示す図である。

【図４】図４は、実施例１における暫定分岐ツリー情報テーブルに記憶された情報の一例を示す図である。

【図５】図５は、隠れマルコフモデルの一例を示す図である。

【図６】図６は、隠れマルコフモデルの一例を示す図である。

【図７】図７は、実施例１における分岐命令解析部による処理の一例について示す図である。

【図８－１】図８－１は、実施例１における分岐命令解析部による処理の一例について示す図である。

【図８－２】図８－２は、実施例１における分岐命令解析部による処理の一例について示す図である。

【図９】図９は、実施例１における分岐命令解析部による処理の一例について示す図である。

【図１０】図１０は、実施例１における特定装置による処理の流れの一例を示すフローチャートである。

【図１１】図１１は、特定装置による一連の処理を実行するための特定プログラムによる情報処理がコンピュータを用いて具体的に実現されることを示す図である。

【発明を実施するための形態】

【００１４】

以下に、開示する特定装置、特定方法及び特定プログラムの実施例について、図面に基づいて詳細に説明する。なお、本実施例により開示する発明が限定されるものではない。各実施例は、処理内容を矛盾させない範囲で適宜組み合わせることが可能である。

【実施例１】

【００１５】

図１は、実施例１における特定装置の構成の一例を示すブロック図である。図１に示す例では、特定装置１００は、記憶部１１０と、制御部１２０とを有する。また、特定装置１００は、以下に詳細に説明するように、オリジナルコード１０の入力を受け付け、インポートアドレス格納場所を出力する。

【００１６】

なお、以下では、インポートアドレス格納場所２０を「分岐先アドレス格納場所」とも記載する。言い換えると、以下では、特定装置１００が、パッキングされたプログラムコードをアンパックすることで得られるオリジナルコード１０の入力を受け付ける場合を例に説明する。言い換えると、パッキング前のプログラムコードの入力を受け付ける場合を用いて説明する。

【００１７】

ただし、これに限定されるものではなく、例えば、特定装置１００は、パッキングされたファイルを受け付け、受け付けたファイルからオリジナルコード１０を抽出した上で、

後述する一連の処理を実行することでインポートアドレス格納場所情報を出力しても良い。

【００１８】

図２は、実施例１における特定装置に入力されるオリジナルコードの一部を示す図である。図２に示す例では、説明の便宜上、オリジナルコード１０に記載された一連の機械語のうちの機械語命令と解釈される部分の先頭を示す「アドレス」と対応付けて、「アドレス」から始まる機械語が機械語命令であると解釈した場合に得られる「ニーモニック」を示した。

【００１９】

図２に示す例では、アドレス「０ｘ０００５」とニーモニック「ｃａｌｌ　０ｘ０００Ｆ」とを対応付けて記憶する。すなわち、オリジナルコード１０のうち、「０ｘ０００５」が機械語命令であると解釈した場合には、ニーモニック「ｃａｌｌ　０ｘ０００Ｆ」が得られることを示す。なお、図２に示す例では、ニーモニックが「ｃａｌｌ」で始まる場合には、ｃａｌｌ命令を示す。また、ニーモニックが「ｊｕｍｐ」で始まる場合には、ｊｕｍｐ命令を示す。また、図２において、［］がある場合には、間接分岐命令を示し、［］がない場合には、直接分岐命令を示す。例えば、ニーモニック「ｃａｌｌ　０ｘ０００Ｆ」は、「０ｘ０００Ｆ」に分岐する直接ｃａｌｌ命令を示す。また、ニーモニック「ｃａｌｌ　［０ｘ２００４］」は、アドレス「０ｘ２００４」に格納されている値を分岐先アドレスとして分岐する間接ｃａｌｌ命令を示す。

【００２０】

記憶部１１０は、制御部１２０と接続される。記憶部１１０は、制御部１２０による各種処理に用いるデータを記憶する。記憶部１１０は、例えば、ＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）やＲＯＭ（Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）、フラッシュメモリ（Ｆｌａｓｈ　Ｍｅｍｏｒｙ）などの半導体メモリ素子、又は、ハードディスクや光ディスクなどである。図１に示す例では、記憶部１１０は、暫定分岐情報格納テーブル１１１と、暫定分岐ツリー情報テーブル１１２とを有する。

【００２１】

暫定分岐情報格納テーブル１１１は、後述する制御部１２０の分岐命令解析部１２１による処理結果を記憶する。図３は、実施例１における暫定分岐情報格納テーブルに記憶された情報の一例を示す図である。

【００２２】

図３に示す例では、暫定分岐情報格納テーブル１１１は、「アドレス」と、「種別」と、「命令確率」と、「暫定分岐先アドレス」と、「暫定分岐先アドレス格納場所」とを対応付けて記憶する。ここで、「アドレス」は、オリジナルコード１０に記載された一連の機械語のうち機械語命令と解釈される部分の先頭を示す。「種別」は、分岐命令の種別を示す。例えば、「種別」は、「アドレス」から始まる機械語が機械語命令であると解釈した場合に、「直接ｃａｌｌ命令」か「間接ｃａｌｌ命令」か「直接ｊｕｍｐ命令」か「間接ｊｕｍｐ命令」かを示す。また、「命令確率」は、「アドレス」から始まる機械語が機械語命令である確率を示す。「命令確率」は「第１の確率」とも記載する。「暫定分岐先アドレス」と「暫定分岐先アドレス格納場所」とは、それぞれ、「アドレス」から始まる機械語が機械語命令についての分岐先アドレス、又は、分岐先アドレスが格納された格納場所を示す。

【００２３】

例えば、暫定分岐情報格納テーブル１１１は、アドレス「０ｘ０００５」と、種別「直接ｃａｌｌ」と、命令確率「０．２」と、暫定分岐先アドレス「０ｘ０００Ｆ」とを含むレコードを記憶する。すなわち、暫定分岐情報格納テーブル１１１は、アドレス「０ｘ０００５」から始まる機械語が機械語命令であると解釈した場合に、分岐先アドレスが「０ｘ０００Ｆ」となる直接ｃａｌｌ命令である確率が「０．２」であることを記憶する。なお、暫定分岐情報格納テーブル１１１は、直接分岐命令についてのレコードでは、「暫定分岐先アドレス」を記憶する。また、暫定分岐情報格納テーブル１１１は、間接分岐命令

についてのレコードでは、「暫定分岐先アドレス格納場所」を記憶する。

【００２４】

なお、暫定分岐情報格納テーブル１１１に記憶された情報は、分岐命令解析部１２１により格納され、分岐関係解析部１２２により用いられる。

【００２５】

暫定分岐ツリー情報テーブル１１２は、後述する制御部１２０の分岐関係解析部１２２による処理結果を記憶する。具体的には、暫定分岐ツリー情報テーブル１１２は、分岐命令解析部１２１により抽出された分岐命令と解釈可能な部分各々について、他の命令との分岐元又は分岐先を示す分岐関係を示す暫定分岐関係ツリーを記憶する。

【００２６】

図４は、実施例１における暫定分岐ツリー情報テーブルに記憶された情報の一例を示す図である。図４に示す例では、図３における「アドレス」を用いて、分岐命令解析部１２１により抽出された部分各々を示した。図３に示すように、暫定分岐ツリー情報テーブル１１２は、例えば、「０ｘ０００５」が分岐元となり「０ｘ０００Ｆ」が分岐先となる関係を記憶し、「０ｘ０００Ｆ」が分岐元となり「０ｘ１００４」が分岐先となる関係を記憶する。

【００２７】

なお、暫定分岐ツリー情報テーブル１１２に記憶された情報は、分岐関係解析部１２２により格納され、インポートアドレス格納場所特定部１２３により用いられる。

【００２８】

制御部１２０は、記憶部１１０と接続される。制御部１２０は、各種の処理手順などを規定したプログラムを記憶する内部メモリを有し、種々の処理を制御する。制御部１２０は、例えば、ＡＳＩＣ（Ａｐｐｌｉｃａｔｉｏｎ Ｓｐｅｃｉｆｉｃ Ｉｎｔｅｇｒａｔｅｄ Ｃｉｒｃｕｉｔ）、ＦＰＧＡ（Ｆｉｅｌｄ Ｐｒｏｇｒａｍｍａｂｌｅ Ｇａｔｅ Ａｒｒａｙ）、ＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）、ＭＰＵ（Ｍｉｃｒｏ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）などの電子回路である。図１に示す例では、制御部１２０は、分岐命令解析部１２１と、分岐関係解析部１２２と、インポートアドレス格納場所特定部１２３とを有する。

【００２９】

分岐命令解析部１２１は、パックされたプログラムコードをアンパックすることで得られるオリジナルコード１０から、少なくとも間接分岐命令と解釈可能な箇所を示すアドレスである間接分岐箇所を抽出する。また、分岐命令解析部１２１は、間接分岐箇所に加えて、直接分岐命令と解釈可能な箇所である直接分岐箇所を抽出する。すなわち、分岐命令解析部１２１は、例えば、オリジナルコード１０の入力を受け付けると、分岐命令と解釈可能な部分をすべて抽出する。

【００３０】

より詳細な一例をあげて説明すると、分岐命令解析部１２１は、オリジナルコード１０から、直接ｃａｌｌ命令と間接ｃａｌｌ命令と直接ｊｕｍｐ命令と間接ｊｕｍｐ命令とのうち、いずれかであると解釈可能な部分を抽出する。

【００３１】

なお、分岐命令解析部１２１により抽出された箇所は、分岐命令と解釈可能である箇所であり、実際には分岐命令ではない箇所も含まれることがあり得る。

【００３２】

ここで、分岐命令について簡単に説明した上で、直接分岐命令と間接分岐命令との違いについて簡単に説明する。分岐命令とは、次に実行される命令を切り替える命令である。なお、一連の命令は、分岐命令でない場合には、逐次順番に実行される。分岐命令には、例えば、ｊｕｍｐ命令やｃａｌｌ命令などがある。ｊｕｍｐ命令やｃａｌｌ命令は、次に実行する命令を変更する。言い換えると、ｊｕｍｐ命令やｃａｌｌ命令の次に実行される命令は、ｊｕｍｐ命令やｃａｌｌ命令の次に記載された命令ではなく、ｊｕｍｐ命令やｃａｌｌ命令のオペランドとして指定された命令となる。例えば、ｊｕｍｐ命令やｃａｌｌ

命令は、オペランドとしてメモリアドレスを指定しており、オペランドとして指定されたメモリアドレスに格納されている値が分岐先アドレスとなる。つまり、ｊｕｍｐ命令やｃａｌｌ命令の次に、オペランドとして指定されたメモリアドレスに格納されている値が取得されたり処理されたりする。

【００３３】

直接分岐命令と間接分岐命令との違いについて簡単に説明する。直接分岐命令には、例えば、直接ｃａｌｌ命令と直接ｊｕｍｐ命令とがある。また、間接分岐命令には、例えば、間接ｃａｌｌ命令と間接ｊｕｍｐ命令とがある。直接分岐命令では、命令内に分岐先となる分岐先アドレスが明示されており、命令解読時に分岐先が決定される。間接分岐命令では、命令内に分岐先アドレスが明示されておらず、実際に間接分岐命令を実行する際に初めて分岐先アドレスが決定される。例えば、間接分岐命令では、命令内には分岐先アドレスが明示されておらず、分岐先アドレス格納場所から分岐先アドレスを読み出すことで、分岐先アドレスが決定される。例えば、間接分岐命令は、実行時にロードされたＤＬＬが分岐先となり、ＤＬＬがロードされるメモリアドレスが動的に決定される場合に利用される。

【００３４】

また、分岐先アドレス格納場所と、間接分岐箇所と、直接分岐箇所との関係について簡単に説明する。分岐先アドレス格納場所を参照する分岐命令は、間接分岐箇所のみとなる。また、間接分岐箇所を分岐先とする分岐命令は、直接分岐箇所のみとなる。間接分岐箇所を分岐先とする直接分岐箇所は、ない場合もある。直接分岐箇所を分岐先とする分岐命令は、直接分岐箇所のみとなる。直接分岐箇所を分岐先とする直接分岐箇所は、ない場合もある。

【００３５】

分岐命令解析部１２１の説明に戻る。分岐命令解析部１２１は、公知の手法を用いて、オリジナルコード１０から分岐命令と解釈可能な箇所を抽出する。例えば、分岐命令解析部１２１は、機械語にて記載された命令である機械語命令として解釈可能なバイト列を識別することで、直接分岐箇所や間接分岐箇所を抽出する。詳細な一例をあげて説明すると、ＩＡ－３２命令セットの場合、分岐命令解析部１２１は、１バイト目が「０ｘＥＢ」であれば、直接ｊｕｍｐ命令であると解釈可能な部分として抽出する。

【００３６】

また、分岐命令解析部１２１は、抽出した間接分岐箇所が分岐命令となる第１の確率を算出する。また、同様に、分岐命令解析部１２１は、抽出した間接分岐箇所が分岐命令となる第１の確率に加えて、抽出した直接分岐箇所が分岐命令となる第１の確率を算出する。なお、分岐命令解析部１２１による第１の確率を算出する算出手法の一例については、後述する。

【００３７】

また、分岐命令解析部１２１は、抽出した間接分岐箇所と直接分岐箇所とについて、算出した第１の確率と対応付けて暫定分岐情報格納テーブル１１１に格納する。また、分岐命令解析部１２１は、抽出した直接分岐箇所について格納する際には、抽出した直接分岐箇所が実際に直接分岐命令である場合に分岐先と解釈される分岐先アドレスを「暫定分岐先アドレス」として併せて格納する。また、分岐命令解析部１２１は、抽出した間接分岐箇所を格納する際には、抽出した間接分岐箇所が実際に間接分岐命令である場合に分岐先を示す分岐先アドレスが格納される分岐先アドレス格納場所と解釈される位置を「暫定分岐先アドレス格納場所」として併せて格納する。

【００３８】

例えば、分岐命令解析部１２１は、先頭が「０ｘ０００５」から始まる箇所が直接ｃａｌｌ命令である第１の確率が「０．２」であると算出した場合を用いて説明する。また、先頭が「０ｘ０００５」から始まる箇所が直接ｃａｌｌ命令であり、分岐先の先頭のアドレスが「０ｘ０００Ｆ」である場合を用いて説明する。この場合、分岐命令解析部１２１は、アドレス「０ｘ０００５」と、種別「直接ｃａｌｌ」と、命令確率「０．２」と、暫

定分岐先アドレス「０ｘ０００Ｆ」とを含むレコードを暫定分岐情報格納テーブル１１１に格納する。

【００３９】

分岐関係解析部１２２は、分岐命令解析部１２１により抽出された箇所間の分岐関係を解析することで、間接分岐箇所が直接的又は間接的に分岐先となる直接分岐箇所を識別する。そして、分岐関係解析部１２２は、分岐命令解析部１２１により抽出された箇所間の分岐関係を示すツリーである暫定分岐関係ツリーを生成する。なお、分岐関係解析部１２２は、「識別部」とも称する。

【００４０】

例えば、分岐関係解析部１２２は、分岐命令解析部１２１により生成された暫定分岐情報格納テーブル１１１内に記憶されたレコードのうち、種別が「直接ｃａｌｌ」又は「直接ｊｕｍｐ」となるレコードを識別する。そして、分岐関係解析部１２２は、種別が「直接ｃａｌｌ」又は「直接ｊｕｍｐ」となるレコードの「先頭」から抽出された箇所が、他のレコードの「暫定分岐先アドレス」と一致するか否かを判定する。すなわち、直接分岐命令として抽出された箇所の分岐元が、直接分岐命令として抽出された箇所と一致するか否かを判定する。なお、レコードの「先頭」から抽出された箇所とは、例えば、図３の「アドレス」に示す箇所が該当する。ここで、分岐関係解析部１２２は、一致すると判定した場合には、一致すると判定したレコードに対応する箇所各々が分岐元と分岐先とが関係あると判定する。また、分岐関係解析部１２２は、種別が「直接ｃａｌｌ」又は「直接ｊｕｍｐ」となるレコードについての判定結果に加えて、種別が「間接ｃａｌｌ」又は「間接ｊｕｍｐ」となるレコードの「先頭」と「暫定分岐先アドレス格納場所」とを加えることで、暫定分岐関係ツリーを生成する。

【００４１】

例えば、分岐関係解析部１２２は、図３に示す暫定分岐情報格納テーブル１１１に基づいて処理を実行する場合には、図４に示すような暫定分岐ツリーを生成し、暫定分岐ツリー情報テーブル１１２に格納する。図４は、実施例１における分岐関係解析部により生成される暫定分岐関係ツリーの一例を示す。図４に示す例では、説明の便宜上、直接分岐箇所については４角形で表し、間接分岐箇所については３角形で表し、暫定分岐先アドレス格納場所については丸で表した。例えば、分岐関係解析部１２２は、直接分岐箇所についての分岐関係の判定結果からツリーを生成する。その後、分岐関係解析部１２２は、直接分岐箇所の分岐先となる間接分岐箇所と、間接分岐箇所の暫定分岐先アドレス格納場所とを加えることで、暫定分岐ツリーを生成する。なお、図４に示す例では、「０ｘ０００５」が分岐元となり「０ｘ０００Ｆ」が分岐先となり、「０ｘ０００Ｆ」が分岐元となり「０ｘ１００４」が分岐先となる。

【００４２】

インポートアドレス格納場所特定部１２３は、間接分岐箇所の先頭と、間接分岐箇所が分岐先となる直接分岐箇所の先頭とのうち、少なくとも一つの箇所の先頭が機械語命令の先頭となる第２の確率を算出する。なお、間接分岐箇所の「先頭」や直接分岐箇所の「先頭」とは、間接分岐箇所や直接分岐箇所としてレコードから抽出された箇所のうち、先頭部分にある箇所を示す。すなわち、図３に示す例では、「アドレス」に示された箇所が該当する。

【００４３】

また、インポートアドレス格納場所特定部１２３は、暫定分岐先アドレス格納場所が同一となる間接分岐箇所が複数抽出された場合に、暫定分岐先アドレス格納場所が同一となる間接分岐箇所各々の先頭と、暫定分岐先アドレス格納場所が同一となる間接分岐箇所各々のうちいずれかが分岐先となる直接分岐箇所の先頭とのうち、少なくとも一つの箇所の先頭が機械語命令の先頭となる第２の確率を算出する。つまり、インポートアドレス格納場所特定部１２３は、抽出された間接分岐箇所の分岐先アドレスが格納されている暫定分岐先アドレス格納場所が実際に分岐先アドレス格納場所となる確率を算出する。

【００４４】

ここで、計算の単純化を目的として、分岐命令解析部１２１により抽出された箇所各々が機械語命令として解釈される事象がお互いに独立であると仮定して説明する。また、暫定分岐先アドレス格納場所「Ｘ」が実際に分岐先アドレス格納場所となる第２の確率を「ＰＸ」として説明する。また、「Ｘ」と同一の暫定分岐ツリー内に存在する分岐命令「Ｙｉ（１≦ｉ≦Ｎ，ＮはＸと同一ツリー内に存在する分岐命令の数を示し、ｉはツリー内の分岐命令各々を表すインデックスを示す。）」として説明する。また、分岐命令「Ｙｉ」が実際に分岐命令となる第１の確率を「ＰＹｉ」とした上で説明する。この場合、「ＰＸ」は、下記の式（１）にて算出可能となる。

【００４５】

【数１】

$$PX = 1 - \prod(1 - PYi) \qquad \cdots (1)$$

【００４６】

すなわち、式（１）に示すように、「Ｘ」と同一の暫定分岐ツリー内に存在する分岐命令のうち、少なくとも一つの箇所の先頭が機械語命令の先頭となる第２の確率は、「１」から、「Ｘ」と同一の暫定分岐ツリー内に存在する分岐命令のすべてが機械語命令の先頭とはならない確率を減算することで得られる。なお、式（１）の「Π」は、総積を示す。言い換えると、式（１）に示す例では、「Π（１－ＰＹｉ）」は、「Ｘ」と同一の暫定分岐ツリー内に存在するすべての分岐命令「Ｙｉ」について「１－ＰＹｉ」を乗算した結果を示す。

【００４７】

すなわち、インポートアドレス格納場所特定部１２３は、抽出された箇所各々に該当する第１の確率を用いて、抽出された同一ツリー内にある箇所すべてが分岐命令とならない第３の確率を１から減算することで、第２の確率を算出する。ここで、第３の確率とは、例えば、式（１）における「Π（１－ＰＹｉ）」を示す。

【００４８】

より詳細な一例をあげて説明すると、インポートアドレス格納場所特定部１２３は、処理対象となる箇所について算出された第１の確率を１から減算した値を処理対象となる箇所ごとに算出し、処理対象となる箇所ごとに算出された値各々を積算し、積算することで得られた値を１から減算することで、第２の確率を算出する。処理対象となる箇所とは、同一の暫定分岐ツリー内に存在する分岐命令となり、例えば、間接分岐箇所の先頭一つであったり、間接分岐箇所の先頭、及び間接分岐箇所が直接的又は間接的に分岐先となる直接分岐箇所の先頭であったり、分岐先アドレス格納場所が同一となる間接分岐箇所各々の先頭及び間接分岐箇所各々のうちいずれかが分岐先となる直接分岐箇所の先頭であったりする。

【００４９】

例えば、図３における暫定分岐先アドレス格納場所「０ｘ２００４」がインポートアドレス格納場所２０となる第２の確率は下記の式（２）にて算出される。

【００５０】

【数２】

$$P(0x2004) = 1 - (1 - 0.2) \times (1 - 0.4) \times (1 - 0.1) \times (1 - 0.6) \times (1 - 0.8) = 0.96544 \qquad \cdots (2)$$

【００５１】

同様に、図３における暫定分岐先アドレス格納場所「０ｘ２０ＡＡ」がインポートアドレス格納場所２０である第２の確率は下記の式（３）にて算出される。

【００５２】

【数３】

$$P(0x20AA) = 1 - (1 - 0.4) = 0.4 \qquad \cdots (3)$$

【００５３】

また、インポートアドレス格納場所特定部１２３は、算出された第２の確率が閾値以上である場合に、間接分岐箇所が間接分岐命令であるとした場合における間接分岐箇所からの分岐先を示す分岐先アドレスが格納される分岐先アドレス格納場所を特定する。例えば、閾値として「０．５」を用いる場合には、インポートアドレス格納場所特定部１２３は、アドレス「０ｘ２００４」をインポートアドレス格納場所２０として特定し、アドレス「０ｘ２０ＡＡ」をインポートアドレス格納場所２０として特定しない。

【００５４】

なお、インポートアドレス格納場所特定部１２３は、間接分岐箇所について算出された第１の確率が閾値以上となる場合に分岐先アドレス格納場所を特定しても良い。言い換えると、同一の暫定分岐ツリー内に存在する分岐命令のうち、少なくとも一つの箇所の先頭が機械語命令の先頭となる第２の確率を用いることなく、間接分岐箇所について算出された第１の確率のみを用いても良い。

【００５５】

ここで、同一の暫定分岐ツリー内に存在する分岐命令のうち、少なくとも一つの箇所の先頭が機械語命令の先頭となる第２の確率が閾値以上となる場合に、分岐先アドレス格納場所を特定する点について補足する。暫定分岐ツリー内に分岐命令が存在する場合、その分岐先も機械語命令と考えられることを踏まえ、インポートアドレス格納場所特定部１２３は、少なくとも一つの箇所の先頭が機械語命令の先頭となる第２の確率を算出して用いても良い。

【００５６】

ここで、インポートアドレス格納場所特定部１２３により分岐先アドレス格納場所として特定されるのは、間接分岐箇所として抽出された箇所のうち、算出された第２の確率が閾値以上となった間接分岐箇所についての暫定分岐先アドレス格納場所となる。言い換えると、インポートアドレス格納場所特定部１２３により分岐先アドレス格納場所として特定されるのは、直接分岐箇所の暫定分岐先アドレスではない。この点について補足する。上述したように、例えば、ランタイムパッカーによりパッキングされたプログラムは、通常はローダが行う動的ライブラリのリンク処理等を実施した後に、オリジナルコードのエントリポイントへ処理が渡される。ここで、Ｗｉｎ３２　ＡＰＩやＤＬＬに含まれる外部関数は、通常、実行時にメモリ上に一度ロードされた上で、実行時に決定されたメモリアドレスが参照されて実行される。このことを踏まえ、インポートアドレス格納場所特定部１２３は、直接分岐箇所の分岐先アドレスではなく、間接分岐箇所の分岐先アドレス格納場所を特定する。

【００５７】

ここで、機械語命令として解釈可能な部分が実際に機械語命令である第１の確率を算出する分岐命令解析部１２１による処理の一例について説明する。以下では，隠れマルコフモデルに基づき箇所各々について命令確率を算出する場合を用いて説明するが，これに限定されるものではない。

【００５８】

分岐命令解析部１２１は、例えば、隠れマルコフモデルによりオリジナルコード１０をモデル化し、Ｆｏｒｗａｒｄ／Ｂａｃｋｗａｒｄアルゴリズムを用いることで、オリジナルコード１０の各バイトが機械語命令の先頭である確率を算出する。なお、以下に説明する処理は一例であり、これに限定されるものではない。

【００５９】

隠れマルコフモデルによりオリジナルコード１０をモデル化する点について簡単に説明する。なお、以下では、オリジナルコード１０に対して逆アセンブルが行われることで、オリジナルコード１０を構成する複数のバイナリ値が複数の単語に分割され、分割された複数の単語それぞれに「命令部」又は「データ部」のいずれかの状態であることを示す「タグ」が割り当てられ、「命令部」としての「タグ」が割り当てられた単語の命令長に基づいて、ニーモニック（アセンブルコード）を当てはめられた場合を用いて説明する。すなわち、オリジナルコード１０を構成する複数のバイナリ値が複数の単語に分割されてお

り、複数の単語それぞれにタグが付されている場合を用いて説明する。

【００６０】

また、以下では、「入力バイナリ列：Ｘ」とは、「逆アセンブル」の対象となる「オリジナルコード１０」のバイナリ列を示し、「Ｎ」個のバイナリ値であるとする。式（４）に示すように、「逆アセンブル」の対象となる「オリジナルコード」を構成するＮ個のバイナリ値は、「ｘ１～ｘＮ」として表される。

【００６１】

【数４】

$$\text{入力バイナリ列}\quad X = x_1^N = x_1, x_2, \ldots, x_N \qquad \cdots (4)$$

【００６２】

また、「単語列：ｗ」とは、「入力バイナリ列：Ｘ」を１命令の「命令部」もしくは１データの「データ部」としての単語として分割したものである。式（５）に示すように、「入力バイナリ列：Ｘ」を分割したＭ個の単語それぞれは、「ｗ１～ｗＭ」として表される。「ｗｉ」は１命令もしくは１データを表す。なお、「命令部」は、複数のバイナリ値から構成される場合もあるため、「単語数：Ｍ」≦「入力バイナリ数：Ｎ」となる。

【００６３】

【数５】

$$\text{単語列}\quad W = w_1^M = w_1, w_2, \ldots, w_M \qquad \cdots (5)$$

【００６４】

また、「タグ列：Ｔ」とは、単語「ｗ１～ｗＭ」それぞれに対して、「命令部」か「データ部」であるかの「タグ」が割り当てられたものである。式（６）に示すように、単語「ｗ１～ｗＭ」に対応付けてタグ「ｔ１～ｔＭ」として表される。

【００６５】

【数６】

$$\text{タグ列}\quad T = t_1^M = t_1, t_2, \ldots, t_M \qquad \cdots (6)$$

【００６６】

また、「命令タグ集合：Ｉ」は、「命令部」としての状態を表す「タグ」の集合であり、「データタグ集合：Ｄ」は、「データ部」としての状態を表す「タグ」の集合である。ここで、タグ「ｔｉ（１≦ｉ≦Ｍ）」は、命令部かデータ部かのいずれかとなる。この結果、式（７）に示すように、「ｔｉ」は、「命令タグ集合：Ｉ」あるいは「データタグ集合：Ｄ」のいずれかに属する。

【００６７】

【数７】

$$\text{命令タグ集合 } I,\ \text{データタグ集合 } D \qquad t_i \in I \cup D \qquad \cdots (7)$$

【００６８】

図５及び図６は、隠れマルコフモデルの一例を示す図である。すなわち、図５に示すように、「命令タグ集合：Ｉ」に属するタグを「継続命令状態：Ｓ」及び「データ直前命令状態：Ｔ」の２種類に更に分割し、「データタグ集合：Ｄ」に属する「データ状態：Ｕ」と合わせて３種類の状態から構成される隠れマルコフモデルを前提とする。

【００６９】

「継続命令状態：Ｓ」は、１命令を出力したのち、引き続き「継続命令状態：Ｓ」に留まる場合と、「データ直前命令状態：Ｔ」に遷移する場合とがある。

【００７０】

「データ直前命令状態：Ｔ」は、「継続命令状態：Ｓ」と同様に、１命令を出力するが、その遷移先は、「データ状態：Ｕ」のみとなる。一般的に、後方にデータが続く命令は、無条件分岐であることが多い。この結果、命令状態を、継続命令状態と、データ直前命

令状態に分割することで、逆アセンブルの精度を向上することが期待できる。

【００７１】

ここで、「継続命令状態：Ｓ」、「データ直前命令状態：Ｔ」、又は、「データ状態：Ｕ」のいずれかの「状態ｉ」から始まる確率（初期確率）を「πｉ」とし、「状態ｉ」から「状態ｊ」へ遷移する確率（遷移確率）を「ａｉｊ」とし、「状態ｉ」におけるシンボルとしての「単語ｗ」が出力される確率（シンボル出力確率）を「ｂｉ（ｗ）」とする。

【００７２】

このような隠れマルコフモデルの一例において、「データ状態：Ｕ」で出力されるシンボルをデータ１バイトとすると、これにより、「データ状態：Ｕ」におけるシンボル出力確率「ｂＵ（ｗ）」において、「ｗ」は、「０以上２５５以下の範囲にある整数」とすることができる。

【００７３】

これに対して、「命令タグ集合：Ｉ」に属する「状態ｉ」において出力されるシンボルの長さ（シンボル長）は、１命令の長さとなる。ここで、複合命令セットコンピュータ（ＣＩＳＣ：Ｃｏｍｐｌｅｘ　Ｉｎｓｔｒｕｃｔｉｏｎ　Ｓｅｔ　Ｃｏｍｐｕｔｅｒ）アーキテクチャの代表的な例であるＩｎｔｅｌ社の「ｘ８６命令」の場合、１命令の長さは最大で１６バイトにも及ぶため、そのままで統計的に信頼できるシンボル出力確率「ｂｉ（ｗ）」を学習することは容易ではない。こうした状況に対応するため、図６を用いて近似的にシンボル出力確率「ｂｉ（ｗ）」（ｉはＩに属する）を算出する方法について述べる。「ｘ８６命令」は、「ＰＲＥＦＩＸ（命令長：０～４バイト）」、「ＯＰＣＯＤＥ（命令長：１～２バイト）」、「ＭｏｄＲＭ（命令長：０～１バイト）」、「ＳＩＢ（命令長：０～１バイト）」、「ＤＩＳＰＬＡＣＥＭＥＮＴ（命令長：０～４バイト）」、「ＩＭＭＥＤＩＡＴＥ（命令長：０～４バイト）」といった命令部から構成される。また、これらの命令部間の遷移パターンは、図６に示すパターンとなる。

【００７４】

ここで、図６に示す遷移パターンによって遷移する各命令部を「状態」とし、「命令開始状態」と「命令終了状態」とを除いた各状態（ＰＲＥＦＩＸ，ＯＰＣＯＤＥ，ＭｏｄＲＭ，ＳＩＢ，ＤＩＳＰＬＡＣＥＭＥＮＴ，ＩＭＭＥＤＩＡＴＥ）では、１バイトの命令部を出力するとする。

【００７５】

また、「単語ｗ」を１バイトごとに分解した結果を、式（８）によって表し（「ｘａ～ｘｂ」）、対応する命令部の種別を、式（９）によって表すとする。

【００７６】

【数８】

$$W\text{を}1\text{バイトごとに分解した結果}: x_a^b \quad \cdots (8)$$

【数９】

$$\text{対応する命令部種別}: v_a^b \in \{\mathrm{PREFIX}, \mathrm{OPCODE}, \mathrm{ModRM}, \mathrm{SIB}, \mathrm{DISPLACEMENT}, \mathrm{IMMEDIATE}\} \quad \cdots (9)$$

【００７７】

このとき、「命令部１バイトを出力する確率は、その時点での命令部の状態によってのみ決まる」と仮定し、更に、「命令部の状態へ遷移する確率は、一つ前の命令部の状態によって決まる」と仮定すると、「命令タグ集合：Ｉ」に属する「状態ｉ」におけるシンボルとしての「単語ｗ」のシンボル出力確率「ｂｉ（ｗ）」は、式（１０）に示すように、近似することができる。なお、ここで示す「命令部の状態」とは、図６のＯＰＣＯＤＥやＭｏｄＲＭなどを示す。

【００７８】

【数１０】

$$b_i(w) \approx \prod_{i=a}^{b} P(x_i \mid v_i) P(v_i \mid v_{i-1}) \quad \cdots (10)$$

【００７９】

これにより、隠れマルコフモデルにおける最尤状態系列算出の問題として、オリジナルコード１０から命令部とデータ部とを識別するために用いるモデルパラメータは、命令部間の状態遷移確率と、命令部ごとの１バイトの出現確率のみとすることができる。この命令部に関するモデルパラメータは、「継続命令状態：Ｓ」と「データ直前命令状態：Ｔ」とで個別に持たせる。

【００８０】

次に、分岐命令解析部１２１によるモデルパラメータの学習について説明する。分岐命令解析部１２１は、『命令部間の遷移確率「Ｐ（ｖｉ｜ｖｉ－１）」及び各命令部における１バイト値のシンボル出力確率「Ｐ（ｘｉ｜ｖｉ）」』を、各状態（タグ）間での遷移回数及び各状態（タグ）におけるシンボル出現回数をカウントすることで算出する。ここで、『命令部間の遷移確率「Ｐ（ｖｉ｜ｖｉ－１）」及び各命令部における１バイト値のシンボル出力確率「Ｐ（ｘｉ｜ｖｉ）」』は、式（１１）～式（１４）を算出するためのモデルパラメータである。

【００８１】

【数１１】

$$\pi_i(i \in I \cup D) \quad \cdots (11)$$

【数１２】

$$a_{ij}(i, j \in I \cup D) \quad \cdots (12)$$

【数１３】

$$b_i(w)(i \in D, 0 \le w \le 255) \quad \cdots (13)$$

【数１４】

$$b_i(w)(i \in I) \quad \cdots (14)$$

【００８２】

ここで、式（１１）は、「命令タグ集合：Ｉ」又は「データタグ集合：Ｄ」のいずれかに属する「状態ｉ」の初期確率「πｉ」を示す。式（１２）は、「命令タグ集合：Ｉ」又は「データタグ集合：Ｄ」のいずれかに属する「状態ｉ」から「命令タグ集合：Ｉ」又は「データタグ集合：Ｄ」のいずれかに属する「状態ｊ」への遷移確率「ａｉｊ」を示す。式（１３）は、「状態ｉ」が「データタグ集合：Ｄ」に属する場合のシンボル出力確率「ｂｉ（ｗ）」を示す。式（１４）は、「状態ｉ」が「命令タグ集合：Ｉ」に属する場合のシンボル出力確率「ｂｉ（ｗ）」を示す。

【００８３】

例えば、分岐命令解析部１２１は、「初期状態」、「継続命令状態：Ｓ」、「データ直前命令状態：Ｔ」及び「データ状態：Ｕ」の間での遷移確率を、図７に示すように、算出する。なお、図７は、実施例１における分岐命令解析部による処理の一例について示す図である。

【００８４】

なお、逆アセンブルされていないタグなしのオリジナルコード１０について処理を実行する場合の一例について簡単に説明する。分岐命令解析部１２１は、タグ付きのオリジナルコード１０に基づいて決定されたモデルパラメータと、タグなしのオリジナルコード１０とを用いて、バウム・ウェルチアルゴリズムによって新たなモデルパラメータを決定して用いる。

【００８５】
次に、Ｆｏｒｗａｒｄ／Ｂａｃｋｗａｒｄアルゴリズムについて簡単に説明する。分岐命令解析部１２１は、Ｆｏｒｗａｒｄアルゴリズムに基づき、モデルパラメータθが与えられたときのオリジナルコード１０候補Ｘの出力確率Ｐ（Ｘ｜θ）を算出する。図８－１及び図８－２、図９は、実施例１における分岐命令解析部による処理の一例について示す図である。

【００８６】
図８－１に示すような１６進数表記の「入力バイナリ列」がオリジナルコード１０として特定装置１００に入力された場合を用いて説明する。また、「入力バイナリ列」を先頭から１バイトずつずらしながら、命令として解釈した場合の命令長を取得することで、図８－２に示すデータが得られた場合を用いて説明する。例えば、図８－２に示すように、「入力バイナリ列」が「５５」である場合に、「命令長：１」が取得され、これに対応するニーモニックが「ＰＵＳＨ　ＥＢＰ」となる場合を用いて説明する。

【００８７】
図９に示す行列は、横軸に「入力バイナリ列」が配置され、縦軸に「継続命令状態：Ｓ」、「データ直前命令状態：Ｔ」及び「データ状態：Ｕ」が配置される。ｊ行目ｉ列目の要素には、「ｘ１，．．．，ｘｉ－１」を出力し且つ「状態ｊ」で「ｘｉ（状態ｊが命令状態の場合は、ｘｉを命令の先頭としたときの命令全体）」を出力する「累積最大確率値」が格納される。また、各要素には、「累積最大確率値」以外にも、「遷移元要素リスト」と「累積最大確率値算出の元になった遷移元要素」が格納される。

【００８８】
ここで、各要素における「遷移元要素リスト」は、図８－２に示す命令長と、式（１１）～式（１４）に示される遷移状態相関関係とを利用することで算出可能である。具体的には、図９に示す行列における１行目１列目（継続命令状態：Ｓ）の場合、「５５」は、１バイト命令であり、遷移先は、１行目２列目（継続命令状態：Ｓ）と、２行目２列目（データ直前命令状態：Ｔ）となる。つまり、１行目２列目と、２行目２列目の「遷移元要素リスト」へ、１行目１列目を追加する。分岐命令解析部１２１は、全要素について同様の処理を繰り返すことで、各要素における「遷移元要素リスト」が算出する。

【００８９】
なお、分岐命令解析部１２１は、すべての入力バイナリ列を出力し終える場合には、図９に示す行列における終了状態（出力確率は「１」）の列に遷移するとする。なお、例外として、１列目の要素の遷移元は、図９に示す行列における初期状態（累積最大確率値は「１」）としておく。

【００９０】
なお、累積最大確率値の算出手法についても簡単に補足する。例えば、分岐命令解析部１２１は、ｊ行目ｉ列目の遷移元要素が、ｎ行目ｍ列目であり、ｎ行目ｍ列目の累積最大確率値を「Ｐｎｍ」、「遷移元状態：ｎ」から「現状態：ｊ」に遷移する確率（図７のモデルパラメータを参照）を「ａｎｊ」とすると、「最大確率値算出の元となった遷移元要素」は、式（１５）に示すように、「Ｐｎｍ×ａｎｊ」が最大となる「ｍ」及び「ｎ」を探すことで算出される。そして、「Ｐｎｍ×ａｎｊ」の最大値に、「ｘｉ」（現状態が命令状態の場合は、ｘｉを命令の先頭としたときの命令全体）のシンボル出力確率を乗算した値を、ｊ行目ｉ列目の累積最大確率値として算出し、対応する要素に格納する。

【００９１】
【数１５】

$$\arg\max_{nm}\left(P_{nm}\times a_{nj}\right) \qquad \cdots（１５）$$

【００９２】
図９に示すように、分岐命令解析部１２１は、図９に示す行列の要素間の遷移において、すべてのタグ系列の総和を計算する。例えば、分岐命令解析部１２１は、ｎ行目ｍ列目の要素からｊ行目ｉ列目の要素への遷移（ｍ＜ｉとする。）に対して、式（１６）に示す

ように、ｎ行目ｍ列目の確率値Ｐｎｍに、状態「ｎ」から状態「ｊ」に遷移する確率「ａｎｊ」（例えば、図７のモデルパラメータを参照）を乗算し、遷移元となるすべてのｍ，ｎについて和をとる。そして、式（１６）の値に「ｘｉ」のシンボル出力確率を乗算した値をｊ行目ｉ列目の確率値として算出し、このような計算を初期状態から終了状態まで算出して確率Ｐ（Ｘ｜θ）を得る。

【００９３】

【数１６】

$$\sum_{nm} P_{nm} \times a_{nj} \qquad \cdots（16）$$

【００９４】

分岐命令解析部１２１は、上述したようなモデルを利用し、Ｆｏｒｗａｒｄ／Ｂａｃｋｗａｒｄアルゴリズムを用いることで、各バイトが機械語命令である第１の確率を算出する。

【００９５】

なお、特定装置１００は、既知のパーソナルコンピュータ、ワークステーション、携帯電話、ＰＨＳ（Ｐｅｒｓｏｎａｌ Ｈａｎｄｙｐｈｏｎｅ Ｓｙｓｔｅｍ）端末、移動体通信端末又はＰＤＡ（Ｐｅｒｓｏｎａｌ Ｄｉｇｉｔａｌ Ａｓｓｉｓｔａｎｔ）などの情報処理装置を利用して実現しても良い。例えば、ＰＤＡなどの情報処理装置に、図１に示した記憶部１１０や制御部１２０の各機能を搭載することによって実現しても良い。

【００９６】

［特定装置による処理］

図１０は、実施例１における特定装置による処理の流れの一例を示すフローチャートである。

【００９７】

図１０に示すように、特定装置１００では、オリジナルコード１０の入力があると（ステップＳ１０１肯定）、分岐命令解析部１２１が、分岐命令と解釈可能な箇所を抽出する（ステップＳ１０２）。例えば、分岐命令解析部１２１は、直接分岐命令又は間接分岐命令と解釈可能な箇所を抽出する。また、分岐命令解析部１２１は、抽出した箇所の先頭が機械語命令の先頭となる確率を算出する（ステップＳ１０３）。

【００９８】

そして、分岐関係解析部１２２は、分岐命令解析部１２１により抽出された箇所間の分岐関係を解析することで、間接分岐箇所が直接的又は間接的に分岐先となる直接分岐箇所を識別し（ステップＳ１０４）、分岐命令解析部１２１により抽出された箇所間の分岐関係を示すツリーである暫定分岐関係ツリーを生成する（ステップＳ１０５）。

【００９９】

そして、インポートアドレス格納場所特定部１２３は、間接分岐箇所の先頭と、間接分岐箇所が分岐先となる直接分岐箇所の先頭とのうち、少なくとも一つの箇所の先頭が機械語命令の先頭となる確率を算出する（ステップＳ１０６）。図４における暫定分岐先アドレス格納場所「０ｘ２００４」を例に説明すると、インポートアドレス格納場所特定部１２３は、「１－（１－０．２）×（１－０．４）　×（１－０．１）×（１－０．６）×（１－０．８）＝０．９６５４４」であると算出する。

【０１００】

そして、インポートアドレス格納場所特定部１２３は、閾値以上の確率が算出されると（ステップＳ１０７肯定）、間接分岐箇所が間接分岐命令である場合に間接分岐箇所からの分岐先を示す分岐先アドレスが格納される分岐先アドレス格納場所を特定する（ステップＳ１０８）。例えば、所定の閾値として「０．５」を用いる場合には、インポートアドレス格納場所特定部１２３は、アドレス「０ｘ２００４」をインポートアドレス格納場所２０として特定する。一方、インポートアドレス格納場所特定部１２３は、閾値以上の確率が算出されないと（ステップＳ１０７否定）、そのまま処理を終了する。

【０１０１】

なお、上記の処理手順は、上記の順番に限定されるものではなく、処理内容を矛盾させない範囲で適宜変更しても良い。例えば、上記のステップＳ１０３をＳ１０５の後に実行しても良い。

【０１０２】

［実施例１の効果］

上述したように、実施例１によれば、パックされたプログラムコードをアンパックすることで得られるオリジナルコード１０から、少なくとも間接分岐命令と解釈可能な箇所を示すアドレスである間接分岐箇所を抽出し、抽出された箇所が分岐命令となる第１の確率を算出し、算出した第１の確率に基づいて、抽出された該箇所の先頭が機械語命令の先頭となる第２の確率を算出する。そして、閾値以上の第２の確率が算出されると、間接分岐箇所が間接分岐命令である場合に間接分岐箇所からの分岐先を示す分岐先アドレスが格納される分岐先アドレス格納場所を特定する。この結果、適切な分岐先アドレス格納場所を特定可能となる。すなわち、逆アセンブル結果が不正確であったとしても、機械語命令の先頭となる第１の確率に基づいて処理を実行することで、誤ったインポートアドレス格納場所２０を特定することを減らすことが可能となる。

【０１０３】

すなわち、実施例１によれば、プログラムコードにおいて、外部関数を呼び出すときに使用される分岐先アドレス格納場所を適切に特定可能となる。従来の手法では、プログラムコードから間接ｃａｌｌ命令を探し出し、探し出した間接ｃａｌｌ命令が示すメモリ領域を分岐先アドレス格納場所として特定していた。この従来の手法では、オリジナルコード１０がマルウェアである場合など、逆アセンブルが困難で正確な逆アセンブル結果が得られない場合には、誤ったインポートアドレス格納場所２０を特定することになる。これに対して、実施例１によれば、プログラムコード内の間接分岐命令と解釈できるすべての箇所について、機械語命令と解釈できる第１の確率に基づき、分岐先アドレス格納場所である確率が高い箇所を特定可能となる。この結果、従来の手法よりも高精度に分岐先アドレス格納場所を特定可能となる。また、分岐先アドレス格納場所を特定できれば、例えばマルウェアが利用しようとしている外部ＡＰＩを特定でき、マルウェアがどんな機能を持つか等の脅威把握や、脅威に基づくマルウェアの分類作業が可能となる。

【実施例２】

【０１０４】

さて、これまで本発明の実施例について説明したが、本発明は上述した実施例以外にも、その他の実施例にて実施されても良い。そこで、以下では、その他の実施例を示す。

【０１０５】

［システム構成］

また、本実施例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、上述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（図１～図１０）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。

【０１０６】

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、その全部又は一部を、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又は物理的に分散・統合して構成することができる。例えば、図１に示す例では、記憶部１１０を特定装置１００の外部装置としてネットワーク経由で接続するようにしても良い。

【０１０７】

［プログラム］

図１１は、特定装置による一連の処理を実行するための特定プログラムによる情報処理

がコンピュータを用いて具体的に実現されることを示す図である。図１１に例示するように、コンピュータ３０００は、例えば、メモリ３０１０と、ＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）３０２０と、ネットワークインタフェース３０７０とを有する。コンピュータ３０００の各部はバス３１００によって接続される。

【０１０８】

メモリ３０１０は、図１１に例示するように、ＲＯＭ３０１１及びＲＡＭ３０１２を含む。ＲＯＭ３０１１は、例えば、ＢＩＯＳ（Ｂａｓｉｃ Ｉｎｐｕｔ Ｏｕｔｐｕｔ Ｓｙｓｔｅｍ）等のブートプログラムを記憶する。

【０１０９】

ここで、図１１に例示するように、ハードディスクドライブ３０８０は、例えば、ＯＳ３０８１、アプリケーションプログラム３０８２、プログラムモジュール３０８３、プログラムデータ３０８４を記憶する。すなわち、開示の技術に係る特定プログラムは、コンピュータによって実行される指令が記述されたプログラムモジュール３０８３として、例えばハードディスクドライブ３０８０に記憶される。具体的には、上記実施例で説明した記憶部１１０と同様の情報処理を実行する手順各々が記述されたプログラムモジュールが、ハードディスクドライブ３０８０に記憶される。

【０１１０】

また、上記実施例で説明した記憶部１１０に記憶されるデータのように、特定プログラムによる情報処理に用いられるデータは、プログラムデータ３０８４として、例えばハードディスクドライブ３０８０に記憶される。そして、ＣＰＵ３０２０が、ハードディスクドライブ３０８０に記憶されたプログラムモジュール３０８３やプログラムデータ３０８４を必要に応じてＲＡＭ３０１２に読み出し、各種の手順を実行する。

【０１１１】

なお、特定プログラムに係るプログラムモジュール３０８３やプログラムデータ３０８４は、ハードディスクドライブ３０８０に記憶される場合に限られない。例えば、プログラムモジュール３０８３やプログラムデータ３０８４は、着脱可能な記憶媒体に記憶されても良い。この場合、ＣＰＵ３０２０は、ディスクドライブなどの着脱可能な記憶媒体を介してデータを読み出す。また、同様に、更新プログラムに係るプログラムモジュール３０８３やプログラムデータ３０８４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良い。この場合、ＣＰＵ３０２０は、ネットワークインタフェースを介して他のコンピュータにアクセスすることで各種データを読み出す。

【０１１２】

［その他］

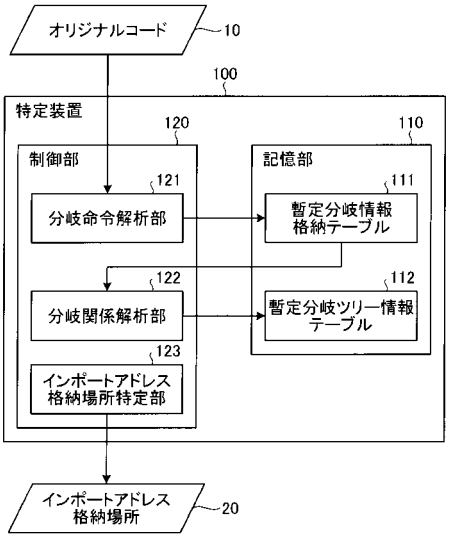
なお、本実施例で説明した特定プログラムは、インターネットなどのネットワークを介して配布することができる。また、特定プログラムは、ハードディスク、フレキシブルディスク（ＦＤ）、ＣＤ－ＲＯＭ、ＭＯ、ＤＶＤなどのコンピュータで読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行することもできる。

【符号の説明】

【０１１３】

１００　特定装置
１１０　記憶部
１１１　暫定分岐情報格納テーブル
１１２　暫定分岐ツリー情報テーブル
１２０　制御部
１２１　分岐命令解析部
１２２　分岐関係解析部
１２３　インポートアドレス格納場所特定部

【図１】



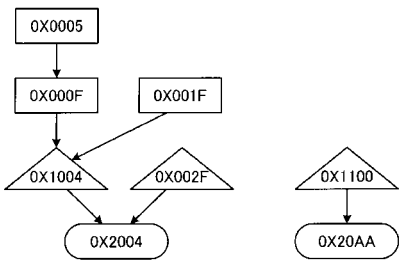
【図２】

アドレス	機械語命令として解釈したときのニーモニック
⋮	
0x0005	call 0x000F
⋮	
0x000F	call 0x1004
⋮	
0x001F	jmp 0x1004
⋮	
0x002F	call [0x2004]
⋮	
0x1004	jmp [0x2004]
⋮	
0x1100	call [0x20AA]
⋮	

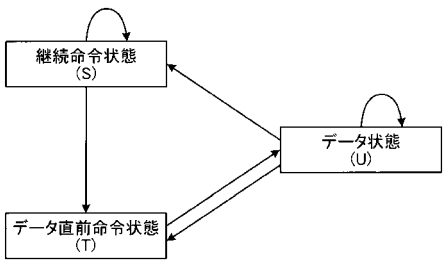
【図３】

アドレス	種別	命令確率	暫定分岐先アドレス	暫定分岐先アドレス格納場所
0x0005	直接call	0.2	0x000F	—
0x000F	直接call	0.4	0x1004	—
0x001F	直接jmp	0.6	0x1004	—
0x002F	間接call	0.8	—	0x2004
0x1004	間接jmp	0.1	—	0x2004
0x1100	間接call	0.4	—	0x20AA

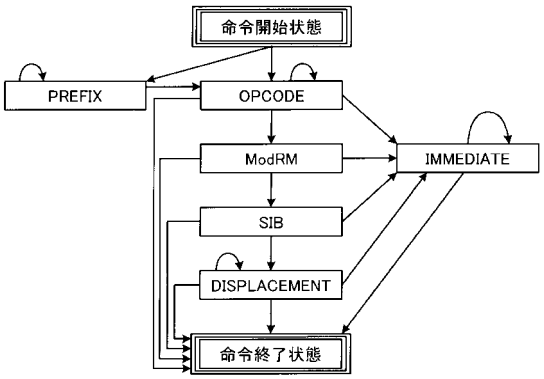
【図４】



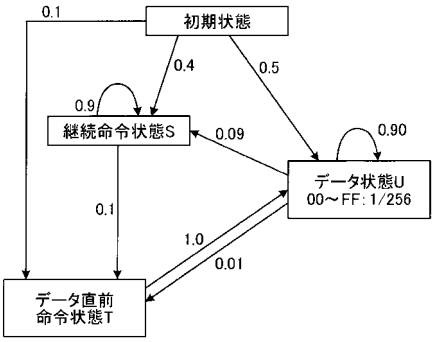
【図５】



【図６】



【図７】



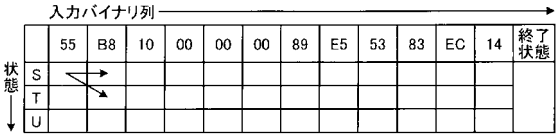
【図８－１】

55	B8	10	00	00	00	89	E5	53	83	EC	14

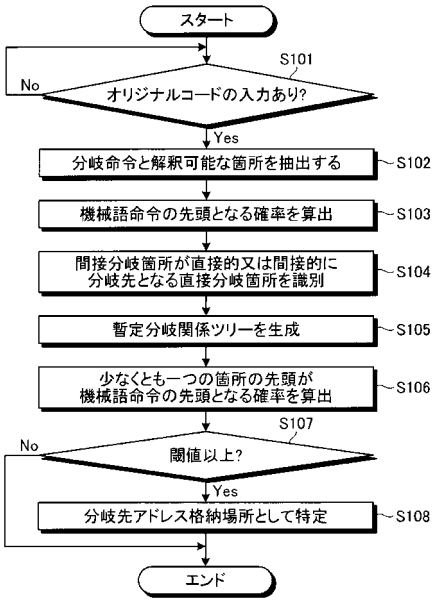
【図８－２】

入力バイナリ列	命令長	（参考）ニーモニック
55	1	PUSH EBP
B8	5	MOV EAX,10
10	2	ADC BYTE PTR DS:[EAX],AL
00	2	ADD BYTE PTR DS:[EAX],AL
00	2	ADD BYTE PTR DS:[EAX],AL
00	6	ADD BYTE PTR DS:[ECX+EC8353E5],CL
89	2	MOV EBP,ESP
E5	2	IN EAX,53
53	1	PUSH EBX
83	3	SUB ESP,14
EC	1	IN AL,DX
14	-	（命令として解釈できない）

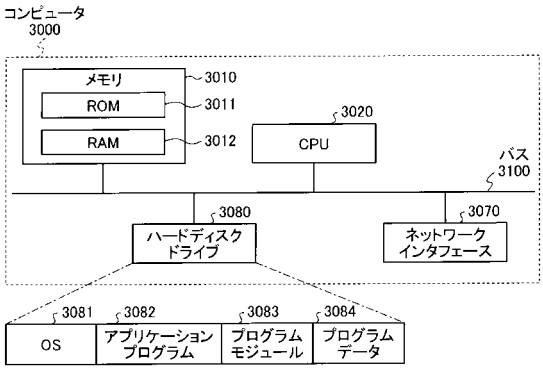
【図９】



【図１０】



【図１１】



フロントページの続き

(56)参考文献 特開２０１０－０９２１７９（ＪＰ，Ａ）
特開２００９－１９３１６１（ＪＰ，Ａ）
米国特許出願公開第２０１０／０２８１５４０（ＵＳ，Ａ1）
岩村 誠，マルウェアのエントリポイント検出後におけるコード領域識別手法，電子情報通信学会技術研究報告，日本，社団法人電子情報通信学会，２０１０年 6月１０日，Ｖｏｌ．１１０，Ｎｏ．７８，第１９－２４頁
泉田 大宗，動的エミュレーションと静的解析を併用したバイナリコードの解析手法，情報処理学会研究報告 平成２１年度▲6▼ ［ＤＶＤ－ＲＯＭ］，日本，社団法人情報処理学会，２０１０年 4月１５日，第１－８頁

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ２１／００－２１／８８
Ｇ０９Ｃ１／００－５／００
Ｈ０４Ｋ１／００
Ｈ０４Ｌ９／００
Ｇ０６Ｆ９／０６，９／４４，－９／４４５，９／４８－９／５０