様式 C-19

科学研究費補助金研究成果報告書

平成２１年　５月２９日現在

研究種目：若手研究（B）
研究期間：2007～2008
課題番号：１９７００００５１
研究課題名（和文）　コンピュータウイルスの可視化と検出に関する研究

研究課題名（英文）　A study on computer viruses visualization and detection

研究代表者
中谷　直司（NAKAYA NAOSHI）
岩手大学・工学部・助教
研究者番号：２０３２２９６９

研究成果の概要：コンピュータウイルスによる被害を抑制するには、既存のアンチウイルスソフトウェアでは検出できない未知のウイルスを検出可能とすることが重要である。そこで、既知のウイルスの特徴とウイルスらしきファイルの特徴を比較することで、未知ウイルスを検出するいくつかの手法を提案した。提案するにあたり機械的に特徴データを処理した結果だけでなく、特徴を可視化することで特徴の抽出と比較を分離して考えた点は本研究の特色である。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	800,000	0	800,000
２００８年度	500,000	150,000	650,000
年度			
年度			
年度			
総　計	1,300,000	150,000	1,450,000

研究分野：総合領域
科研費の分科・細目：情報学　　計算機システム・ネットワーク
キーワード：コンピュータウイルス，未知コンピュータウイルス，可視化，機械学習，Paul Graham Bayes，ベクトル間距離

１．研究開始当初の背景

(1) 近年のインターネットをはじめとするネットワークの急速な発展にともない、コンピュータウイルス（以降、ウイルス）による被害は深刻なものとなってきている。特に近年のウイルスはメールを媒介としているために感染速度が速く、わずかな時間に世界規模の被害を発生させている。しかし現在のウイルス対策は、ウイルスの特徴点（以降、シグネチャ）をもとにしたパターンマッチによりウイルスの検出を行っているため、シグネチャの無い未知ウイルスは基本的に検出できず新種のウイルスによる被害がたびたび生じている。すなわち、未知ウイルスに対応した新しいウイルス検出方式が必要とされている。

(2) そこで未知ウイルス検出手法の確立を目的に、これまでに「過去のウイルスの機能をベイズ学習アルゴリズムにより学習し、その結果をもとにウイルスを検出する」という手法の研究を行い一定の成果を挙げることができた。

- 小池竜一，中谷直司，萩原由香里，厚井裕司，高倉弘喜，吉田等明，“ベイズ学習アルゴリズムを用いた未知のコンピュータウイルス検出手法，”情報処理学会論文誌, Vol.46, No.8, pp.1984-1996, 2005
- Koike, R., Nakaya, N., Koui, Y., “ Filter for Detecting Unknown Computer Viruses Using Graham Bayes Learning Algorithm for Spam Detection,” Proc. Of the Int. Workshop on Data-Mining and Statistical Science (DMSS2006), pp.93-103, 2006

しかし上記の研究では、ウイルスの亜種、すなわち良く似たウイルスならば高い確率で検出可能となったが、亜種ではない完全に新しいウイルスの検出精度は十分とは言えないものであった。これは学習アルゴリズムの問題ではなく、学習させた「過去のウイルスの機能」という情報が、十分にウイルスの機能を表現できていなかったためと考えられる。

(3) 以上の経緯を踏まえ本研究では、未知ウイルスの検出手法の確立を最終目的に、ウイルスを可視化するというアプローチを採ることとする。これは、バイナリファイルから何らかの情報を抽出しその情報をグラフィカルに表現したものが、通常のソフトウェア（以降、ノンウイルス）とウイルスとで明確に区別することができるのならば、それはウイルス特有の機能を抽出できたことに他ならないという発想に基づいている。なお、抽出した情報を可視化せずに何らかの方法で分類し、ウイルスを誤りなく分類できる最適な情報を探すといったアプローチを採ることももちろん可能である。しかし、このアプローチで失敗した場合、分類がうまくいかない原因が、情報の抽出手法と分類手法のどちらにあるのか切り分けることが難しくなる。一方、可視化を行えば、分類には人間というあいまいな情報処理に優れた高性能な分類機を使うことができ、抽出段階で生じた問題を解決することが容易になるものと思われる。そこで本研究では、あえてウイルスを可視化することとする。

２．研究の目的

(1) 本研究の最終目的は未知ウイルスの検出にあるが、これまでの経験からノンウイルスとウイルスを明確に区別できるほどの可視化が可能になったとすれば、その情報をベイズ学習アルゴリズムなどで学習し未知ウイルスを検出することは難しくはないと思われる。そこで本研究の研究期間内の主目標は、バイナリファイルから情報を抽出し、そのソフトウェアの持つ機能を的確に表したグラフィカルな出力を得る手法の開発とする。可視化が完了した後の未知ウイルスの検出アルゴリズムには、過去の研究で実績のあるベイズ学習アルゴリズムや、類似度算出手法などを用いる予定である。

(2) ウイルスも OS 上で動作するプログラムに過ぎず、その感染行動の実現には OS の機能を利用せざるを得ない。そこでウイルスが呼び出す OS の機能に着目し、ヒューリスティック法でウイルスを検出する手法は主流ではないが既に存在する。しかし、現状のそれはルールベースによるものであり、人手によるルールの作成が不可欠となる。また、ルールベースのヒューリスティック法を導入したウイルス対策ソフトウェアを解析することで、ウイルスを製作する側の人間も容易にルールを知ることができるため、ウイルスに対抗策を取られる可能性がある。そこで、ベイズ学習アルゴリズムなどを使って、確率的に未知ウイルスの検出を試みる点が本研究の特色の一つである。これらの学習アルゴリズムを用いればルールベースのヒューリスティック法と異なり、人手による操作なしに自動的にウイルス検出能力の向上を行うことが可能となる。また、ルールベースとは違い確率的に動作するため、ウイルスの製作者が対抗策を講じることは難しい。以上のような、学習アルゴリズムによる未知ウイルスの検出が可能になれば、現状では十分な対策が取れていない未知のウイルスによる被害を未然に防ぎ、ネットワーク全体のセキュリティの向上に貢献できるものと考える。

３．研究の方法

(1) 本研究では過去のコンピュータウイルスの機能をベイズ学習アルゴリズムなどで学習し、その学習結果に基づいたウイルス検出を行うことで、現状では未解決な未知ウイルス検出手法の確立を目指す。ただし、学習情報として与える「コンピュータウイルスの機能」の抽出には研究の余地が多分に存在するため、本研究では情報抽出手法と学習手法の問題を切り分けることを目的に、コンピュータウイルスの機能を的確に表現し、ノンウイルスとウイルスを明確に識別可能なバイナリファイルの可視化を目標とする。

(2) 初めに過去の研究である程度のウイルス検出結果が得られることが確認されている、抽出するウイルスの機能としてバイナリファイル中の文字列を利用した場合の可視化を試みた。バイナリファイル中にはプログラムの動作に必要な様々な情報が、文字列として表示可能な形で埋め込まれている。この文字列には、そのプログラムが動作する上で

欠かせない Windows API（Application Program Interface）関数名や DLL（Dynamic Link Library）名、ユーザーインターフェースに表示される各種文字列など、プログラムの機能と密接に関係しているものも含まれる。そこで、このバイナリファイル中の文字列を可視化することで、ウイルス間に類似性があるか、またノンウイルスとウイルスは識別可能かを検証する。

(3) 次にバイナリファイルの構造にも着目し、これまでの可視化手法では利用していなかったプログラムコードやリソースに関する情報を用いた可視化を行った。まず、バイナリファイルの構造からプログラムの動作そのものが書かれたコードセクションを特定し、そこから Windows API コールを順に取り出すことで API のシーケンスを得た。ここでいう Windows API とはプログラムが動作する上で必須となる OS 機能を呼び出すためのインターフェースのことであるため、今回取り出した API シーケンスはプログラムの動作そのものと密接に関係しており、亜種ウイルスをはじめとするウイルス同士では似通ったものとなる。そこで、既にウイルスとわかっているプログラムから取り出した API シーケンスと、未知のプログラムから取り出したそれとを n 次元ベクトル空間上に表現し、そのベクトル間距離をとることで未知プログラムがウイルスかどうかを判定する手法を試みた。

(4) また、亜種ウイルス同士ではプログラムアイコン等のリソース情報が似通っていることに着目し、リソース情報を格納しているリソースセクションの構造を可視化し、その違いを元にウイルスを検出する手法も試みた。この手法ではリソース情報の内容であるアイコンやメッセージの中身そのものではなく、あえてそれらの情報の格納構造を利用することで、亜種ウイルス間のわずかなリソースの差に影響を受けない検出が可能となっている。

４．研究成果

(1) 研究の方法(2)に基づき可視化した結果からウイルスの間にはある程度の類似性が見られ、特に亜種ウイルス同士ではその傾向が強いことがわかった（図 1）。しかし、バイナリファイル中の文字列をすべて可視化するとウイルスと一般の実行ファイルの間にもかなりの共通部分があり、ウイルスをウイルスとして正しく認識するには全体の傾向ではなく突出した少数の特徴に着目する必要があることが明らかになった。このことは、ウイルス検出に少数の特徴に着目し判定する Paul Graham Bayes を使った過去の研究が有効であったことを示すが、逆に亜種のように少数の特徴が共通するウイルス以外では未知ウイルスの検出が難しく、一般の実行ファイルをウイルスと誤検出する可能性も高くなることがわかった。そこで、過去のウイルスを学習する段階で一般の実行ファイルと共通する部分を学習データから除外する手法を試みたところ、検出率が上がり誤検出率が低下することが確認された。

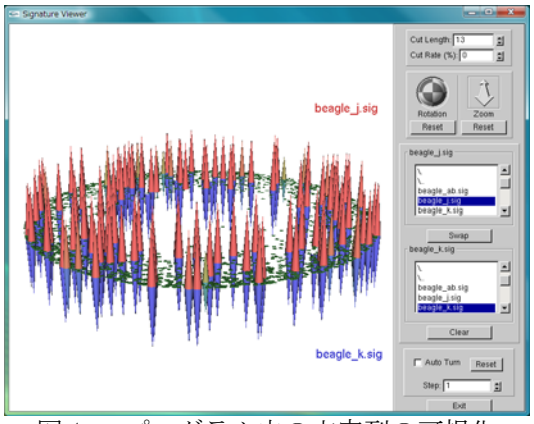

図 1.　プログラム中の文字列の可視化
上下で異なる亜種ウイルス（上：Beagle.J、下：Beagle.K）を表示しているが、差異はほとんど確認できない。

(2) 研究の方法(3)に基づきコードセクションから API シーケンスを取り出し、既知のウイルスやノンウイルスと比較することで未知ウイルスの検出を試みた。取り出した API シーケンスは亜種ウイルス間では類似性が見られることは確認できたが、シーケンスの長さが少しずつ異なるため n 次元ベクトル空間にうまく表現することができないことも分かった。これは、亜種ウイルス同士は機能が似通ってはいるが全く同一というわけではないため、機能に対応した API シーケンスもわずかに異なることに起因する。そこで、2 つのシーケンス間でグローバル・アライメントを取り適時ギャップを挿入することでベクトルの次元を合わせ、それら 2 つのベクトル間距離を取ることで類似度算出しウイルスかどうかを判定した。結果として、亜種ウイルス同士には高い類似度がみられ、事前に同じ種類の亜種ウイルスがあれば未知ウイルスを検出可能であることが確認された。しかし、亜種ウイルスの種類によっては亜種中でさらに 2 つのパターンに類似度がわかれるケースも存在した。なお、ノンウイルスとウイルス間の類似度は低い値に抑えられ、ノンウイルスをウイルスと誤検出することは実験した範囲では存在しなかった。

(3) 研究の方法(4)に基づきリソース情報の可視化を行った。方法でも述べたように亜種

ウイルス同士では、アイコンやメッセージなどのリソース情報にある程度の類似性がみられる。しかし、アイコンはともかくメッセージの内容は完全に同一ではないので、その内容を比較したのでは検出は難しい。そこで、リソースの内容ではなく情報の格納構造、すなわちリソースセクションの構造そのものを比較対象とする。最初にリソースセクションの構造を可視化（図 2）したところ、亜種ウイルス間では同一の構造を持っていることが確認された。また、このとき情報の内容そのものには差異があることも確認された。そこで、ウイルスやノンウイルスからリソースセクションの構造を取り出し比較することで、未知ウイルスを検出する実験を行ったところ、ある程度の誤検出は存在するが未知ウイルスを検出可能なことがわかった。

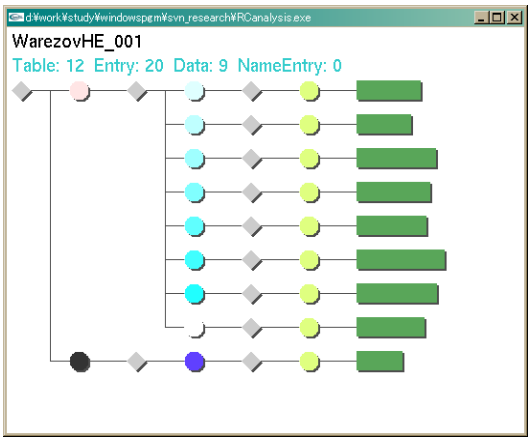

図 2.　リソースセクション構造の可視化
リソースセクションは木構造になっており、その木構造はプログラムごとに異なっている場合が多く、データサイズも合わせて考えるとほぼ固有なものとなる。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

① Dengfeng ZHANG, Naoshi NAKAYA, Yuuji KOUI, Hitoaki YOSHIDA, “An Automatic Unpacking Method for Computer Virus Effective in the Virus Filter Based on Paul Graham’s Bayesian Theorem, ” IEICE Transactions on Communications, Vol.E92-B, No.4, pp.1119-1127, 2009, 査読有

② Zhongda LIU, Naoshi NAKAYA, Yuuji KOUI, “ The Unknown Computer Viruses Detection Based on Similarity,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol.E92-A, No.1, pp.190-196, 2009, 査読有

③ 王卉歓, 中谷直司, 小池竜一, 厚井裕司, 朴美娘, “ベイズ学習アルゴリズムのスパムフィルタとウイルスフィルタへの適用の最適化,” 情報処理学会論文誌, Vol.48, No.9, pp.1325-3136, 2007, 査読有

〔学会発表〕（計 2 件）

① 青木智博, 中谷直司, 厚井裕司, “未解凍ウイルスに対するリソース情報を用いたシグネチャ生成法,” 平成 20 年度第 3 回情報処理学会東北支部研究会, 2009.1.6, 盛岡・岩手大学

② 三浦雄一郎, 中谷直司, 厚井裕司, “コンピュータウイルスの可視化,” 平成 19 年度第 3 回情報処理学会東北支部研究会, 2008.1.11, 盛岡・岩手大学

６．研究組織
(1)研究代表者
中谷　直司（NAKAYA　NAOSHI）
岩手大学・工学部・助教
研究者番号：２０３２２９６９