様式 C-19

科学研究費補助金研究成果報告書

平成　21　年　6　月　25　日現在

研究種目：基盤研究（C）
研究期間：：2006～2008
課題番号：18560361
研究課題名（和文）　未知のウイルス及びスパイウェアの統一的な検知・駆除に関する研究

研究課題名（英文）　A Study on The Unified Detection and Extermination Method for Unknown Viruses or Spywares

研究代表者
厚井　裕司（KOI　YUJI）
岩手大学・工学部・教授
研究者番号：20333750

研究成果の概要：未知のウイルスとスパイウェアが非常に類似している点に注目して、シグネチャに過度に依存しない従来とは異なった方式を組み合わせることにより、未知のウイルスとスパイウェアを統一的に検出・駆除する方式を確立した。すなわち、スパムメール向けの学習アルゴリズムである Graham Bayes 理論をウイルスやスパイウェア等のマルウェア検知用に最適化したもので、実行ファイルにおけるバイナリ情報の文字列の特徴から未知のマルウェアを抽出する。実験では、95%の検知率(入力した実行ファイルからマルウェアを発見する率)と0.02%の誤検出率(マルウェア以外の実行ファイルを誤ってマルウェアと見なす率)を達成した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1,150,000	0	1,150,000
2007 年度	1,200,000	360,000	1,560,000
2008 年度	800,000	240,000	1,040,000
年度			
総　計	3,150,000	600,000	3,750,000

研究分野：工学
科研費の分科・細目：電気電子工学　通信・ネットワーク工学
キーワード：　ベイジアンウイルスフィルタ, ベイズの定理

１．研究開始当初の背景

(1) ウイルスやスパイウェアによる被害

研究開始時点において、未知のコンピュータウイルスや未知のスパイウェアが日々発生し、各所で被害をもたらしていた。被害が増加している原因としては、以下の２つの理由があった。

・特定の人物や企業を狙うウイルスやスパイウェアが増えて、検体の入手が難しくなっている。

・ウイルスのソースコードがオープンソースとして公開されており、亜種ウイルスや亜種スパイウェアが急増している。

(2) パターンマッチング方式の限界

コンピュータウイルスやスパイウェアに対して、防御および削除などの処置を行うためには、 まずこれらを検出する必要がある。検出方式としては、パターンマッチング方式と呼ばれるものである。 これはシグネチャあるいは定義ファイルと呼ばれるウイルスやスパイウェアに対する固有の情報を集めたファイルと対象ファイルを照らし合わせ、対象ファイル中にシグネチャと完全一致した部分が存在する場合にウイルスやスパイウェアとして検出するものである。このパターンマッチング方式は検体として入手した実際のウイルスやスパイウェアを詳細に解析してから シグネチャを生成するため、誤

検出が極めて少ないという特徴がある。しかし、シグネチャが生成可能となるのは、対象となるウイルスやスパイウェアが発見された後となるため、 常に対策が後手に周るという欠点も持っている。パターンマッチング方式を用いる限りは、未知ウイルスや未知スパイウェアの発生直後に検出をすることはできないという課題があった。

２．研究の目的

(1) スパイウェア対策の現状

スパイウェアとウイルスとの違いは、他のコンピュータやファイルへの感染活動を行わないこと、インストールされるまでの経路とコンピュータの利用者に及ぼす被害にある。長い間スパイウェアはウイルスとは区別され、ウイルス対策ソフトではスパイウェアを検出できなかったが、近年のスパイウェアの猛威に Norton AntiVirus 2006 やウイルスバスター2005 など、スパイウェア検出機能が強化されたウイルス対策ソフトが増えてきた。しかしながら、未だ検知率は低く未知のスパイウェアを検知することは非常に困難である。

(2) ボットウイルスの蔓延

最近では、ボットウイルスの増加が顕著であり、これは攻撃者が情報を盗み出すために攻撃対象のコンピュータを遠隔操作するウイルスで、感染後に深刻な被害をもたらす。政府のサイバークリーンセンターは 2007 年末に収集した検体の総数が 1 ヶ月前より約 13 万件増加して、ウイルス対策ソフトで検出できない割合が 17%で野放し状態であると発表した。このようにボットウイルスもスパイウェアも多くのコンピュータに感染するというよりも、情報を盗み出すために作成されている。したがって、ボットウイルスとスパイウェアは機能的に共通した特徴を持つ。さらに形成上も類似点が多く、亜種が多いことも注目すべき点である。

(3)未知ウイルスの検出

そこで本研究では、解凍したウイルスの特徴点を学習アルゴリズムで学習し、未来に発生する未知のウイルスを検出する方法を用いる。この方法は、入力を 2 つのカテゴリへ分類することに特化している PaulGraham ベイズ学習アルゴリズムを利用している。そのため現実的な処理時間で既知のウイルスの学習を行い、未来に発生する未知のウイルスの検出を可能としている。また，学習データは特別な技巧を必要とせずに、実行ファイルから得ることができる strings を使用しているため、実装が容易であるという特徴もある。この研究の最終的な目的は、過去のウイルスと類似性を持った未知ウイルスや未知スパイウェア等の未知マルウェアを検出可能にすることである。これによって、シグネチャ更新がなされるまでの間は未知ウイルスが検出できないという既存のアンチウイルスの弱点を補完することができる。

３．研究の方法

提案方法は、定期的に既知のウイルスやスパイウェアを学習させておくことで、共通点を持った未来のウイルスやスパイウェアを検出する手法である。提案方法と既存のアンチウイルスやアンチスパイウェアを組み合わせて用いることで、シグネチャ更新までこれらのマルウェアが検出できない既存のアンチウイルスやアンチスパイウェアの欠点を補完することができる。

提案方法は学習アルゴリズムである Paul Graham ベイズに、解凍したマルウェアから抽出できる表示可能な文字列（以下、strings）を学習させるものである。Paul Graham ベイズはウイルスやスパイウェアにだけ頻繁に現れる strings を危険なものと認識し、入力されたファイルにも同様の strings が多く含まれる場合には，それをウイルスやスパイウェアとして検出を行う。Paul Graham ベイズは、本来スパム向けに開発されたものであるため、パラメータの値を一部変更することでマルウェア向けへの改良を行う。またウイルスやスパイウェアを解凍するのは、そのままの状態で学習を行うことが、マルウェア検出率に悪影響を与えるからである。

(1)マルウェアとスパムの類似性

マルウェア検出に本来スパム向けの検出アルゴリズムである Paul Graham ベイズを選択した理由は、スパムとマルウェアは「ある目的を達成するための単語，あるいは命令の集合体」であり、類似性があると考えたためである。スパムは商品等の宣伝を目的として、望まない相手に対して強制的に送りつけられるメールである。そのため、どのスパムにも同様の単語が現れるという性質がある。オリジナルの Paul Graham ベイズではスパムに頻繁に現れる単語を学習することで、それらの単語を多く含んだメールをスパムとして検出している。このようなスパム検出フィルタは、ベイジアンフィルタと呼ばれ 90%以上の精度でスパムを検出している。

一方、マルウェアも多くのパソコンに感染または侵入していくという目的を持っている。それゆえ、マルウェアは目的を達成するために様々な行動をとる。しかし、亜種同士においては、似たような行動をとることが多い。これは、亜種マルウェアがオリジナルのものに多少の修正を加えたものだからである。そこで、一定の手続きでウイルスから自身の動作に関する命令を取り出せば、亜種同士にだけ頻繁に現れる命令が存在すると考えられる。このように亜種同士の命令が似ていると仮定すればスパムと同様に、既知の亜

種を学習することで，その後に発生する未来の亜種を検出できる可能性がある．

(2)学習データ：strings

提案方法では実行ファイルから自身の実行に関わるような命令を取り出し、それらを学習データとして用いる。ここでは学習データである strings について述べる。strings はバイナリに含まれる文字として表示可能なデータのことである。文字として表示が可能なデータとは ASCII コードの 0x20 から 0x7E までであり、日本語などの 2 バイト文字は含まない。実行ファイルは PE 形式で構成されており各種情報が strings として格納されている。よって、実行ファイル中から使用する共有ライブラリ名（以下、DLL）やシステムコール名（以下、API），動作時に参照する文字列、バイナリの一部などを抽出することが可能である。PE 形式である実行ファイル中には使用する DLL と API が strings として記述されている。実行ファイルが Microsoft Windows OS 上で動作するためには、OS に対して様々な処理を依頼する必要があり、その時に必要な API と DLL の情報が実行ファイル内部に含まれている。

(3)strings の長さ

バイナリから strings を抽出する際には、strings と見なす最短長を設定することが出来る。例えば、デフォルトの 4 で抽出を行った場合に、長さが 3 文字以下の strings は無効なものとして切り捨てられる。 最短長を 1 のように小さく設定する事で多くの情報を得ることが出来る。しかし、それに比例して意味を持たない余分な情報が混ざり、学習やカテゴリに分類するのに処理時間がかかるようになってしまう。 逆に最短長を長くすることで、必要な情報を切り捨ててしまう恐れがある。

(4)学習アルゴリズム

Paul Graham ベイズはベイズ理論による機械学習を行う分類器の一種であり、ある入力が与えられた場合にそれを適切なカテゴリへと分類する。本方式は入力を 2 つのカテゴリに分類することに特化しているために、高速に動作する特徴がある。一方，テキスト分類などで広く用いられている Naive ベイズは入力を複数のカテゴリに分類するアルゴリズムであり、Paul Graham ベイズより多くの計算量やメモリを必要とする。マルウェアの流行を防ぐためには、いかに素早く対応をしていくのかが重要であるため、Paul Graham ベイズのような高速なアルゴリズムを用いる必要がある。オリジナルの Paul Graham ベイズはベイズ定理をスパム向けに改良したもので、スパムと非スパムを学習し、新たな入力があった場合にそれをスパムカテゴリ、あるいは非スパムカテゴリに分類するものである．提案方法ではカテゴリをマルウェアと非マルウェアとし、入力を実行ファイルの特徴点である strings とすることで、スパム向けのフィルタをマルウェアフィルタとして応用できることを示す。提案方法の Pbase の値以外は、オリジナルの Paul Graham ベイズの定義に従う。

４．研究成果

(1) ウイルスとスパイウェアの類似性

未知のウイルスとスパイウェアが非常に類似している点に注目して、シグネチャに過度に依存しない従来とは異なった方式を組み合わせることにより、未知のウイルスとスパイウェアを統一的に検出・駆除する方式を確立した。

(2) ベイジアンフィルタの利用

既知マルウェアの特徴を機械学習することで、類似点を持った未知マルウェアを検出するベイジアンフィルタを実現した。これは、スパムメール向けの学習アルゴリズムである Graham Bayes 理論をマルウェア検知用に最適化したもので、実行ファイルにおけるバイナリ情報の文字列（最大 15 個で、過去の出現頻度より各々のマルウェア確率を算出）の特徴から未知マルウェアを抽出する。

(3) 検知率と誤検出率の向上

実験では、過去に発生した 2,000 種のマルウェアと 10,000 種の一般ファイルを無作為に混ぜてベイジアンフィルタに入力させた結果、95%の検知率(入力した実行ファイルからウイルスを発見する率)と 0.02%の誤検出率(マルウェア以外の実行ファイルを誤ってマルウェアと見なす率)を達成した。特に検体の入手が難しく、その大部分が亜種ウイルスで構成されるボットの検知に有効である。

（４）大学における評価

このフィルタを実際に大学の情報処理センターに 3 ヶ月間設置した評価でも、上記結果を実証できた。今後は大規模なハニーポットを構築して、スパイウェアやボットを含めた多くのマルウェアを収集して、実用的な検出性能を有することを実証する。更に下記の技術成果を組み合わせて、システムとして実用化する。

①未知のマルウェアフィルタ：ベイジアンフィルタによって実行可能圧縮（難読化）されたマルウェア検知率を 95%に向上

②メールの添付ファイルから未知マルウェアを検出・培養（特許第 3991074）

③マルウェアを特定する新しい固有情報抽出方法を発見（特許第 4025882）

現在、アンチウィルスメーカに技術成果を活用するように提案している。先方より検出性能については十分であり、さらに実験データを増やすよう求められている。

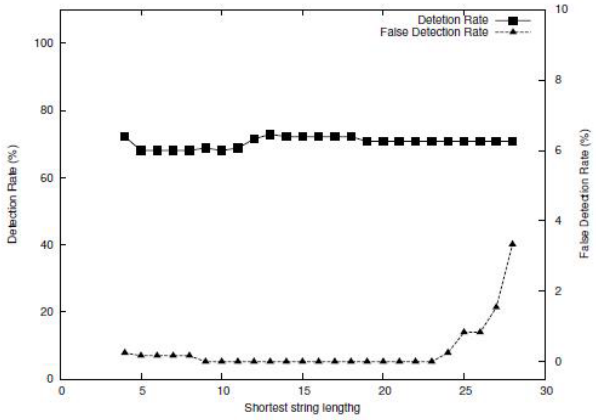

図 1　Netsky の検出率および誤検出率

表 1　Bagle の亜種検出率 (l=13)

	Input								
Signature	J	K	N	O	Y	Z	AB	AE	Nonvirus2
Bagle.J	○	○	○	○					○
Bagle.K	○	○	○	○					○
Bagle.N	○	○	○	○	○	○	○	○	○
Bagle.O	○	○	○	○	○	○	○	○	○
Bagle.Y	○	○	○	○	○	○	○	○	○
Bagle.Z	○	○	○	○	○	○	○	○	○
Bagle.AB	○	○	○	○	○	○	○	○	○
Bagle.AE	○	○	○	○	○	○	○	○	○
Mix	○	○	○	○	○	○	○	○	○

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 Huihuan Wang, Naoshi Nakaya, Ryuiti Koike, Yuji Koui, Dengfeng Zhang, Abdullah Mamun: A Performance Evaluation of Bayes Learning Algorithm For Spam Filter and Virus Filter, The First International Conference on Computer, Control and Communication (IC4 2007), Technical Session 2 no8, Nov. 2007.査読有
②Ryuiti Koike, Naoshi Nakaya, and Yuji Koui: Development of system for the automatic generation of unknown virus extermination software, In Proceedings of the International Symposium on Applications and the Internet 2007 (SAINT 2007), pp. 8-15. IEEE/IPSJ, Jan. 2007.査読有
③ Ryuiti Koike, Naoshi Nakaya, Yuji Koui :Filter for Detecting Unknown Computer Viruses Using Graham Bayes Learning Algorithm for Spam Detection, In Proceedings of the International Workshop on Data-Mining and Statistical Science (DMSS 2006), pp.93-103, Sep. 2006.査読有

〔学会発表〕（計３件）
①1Zhongda LIU, Naoshi NAKAYA, Yuuji KOUI: The Unknown Computer Viruses Detection Based on Similarity, IEICE TRANSACTIONS, Vol.E92-A, No.1, pp.190-196, Jan. 2009.
②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. 3125-3136, 2007.
③小池竜一, 中谷直司, 厚井裕司: 未知コンピュータウイルスを駆除するUSBフラッシュメ モリの開発, 情報処理学会論文誌, Vol. 48, No. 4, pp. 1595-1605, 2007.

〔産業財産権〕

○出願状況（計２件）

①名称：コンピュータウィルス検出装置、処理方法、及びプログラム
発明者：厚井裕司, 中谷直司, 山本博幸, 松浦千凡
権利者：国立大学法人岩手大学
種類：特許
番号：特願 2008-166875
出願年月日：平成 20 年 9 月 19 日
国内外の別：国内

②名称：コンピュータウイルス検出装置、コンピュータウイルス検出方法及びコンピュータウイルス検出プログラム
発明者：厚井裕司, 中谷直司, 水野節郎, 劉忠達
権利者：国立大学法人岩手大学
種類：特許
番号：特願 2008-166875
出願年月日：平成 20 年 6 月 26 日
国内外の別：国内

○取得状況（計２件）

①名称：コンピュータウィルス固有情報抽出装置、コンピュータウィルス固有情報抽出方法、コンピュータウィルス固有情報抽出プログラム
発明者：厚井裕司, 中谷直司, 小池竜一
権利者：国立大学法人岩手大学
種類：特許
番号：特許第 4025882
取得年月日：平成 19 年 10 月 19 日
国内外の別：国内

②名称：電子メール中継システム、方法及びプログラム並びにウィルス検知システム、方法及びプログラム
発明者：厚井裕司, 中谷直司
権利者：国立大学法人岩手大学
種類：特許
番号：特許第 3991074
取得年月日：平成 19 年 8 月 3 日
国内外の別：国内

〔その他〕
http://www.eng.iwate-u.ac.jp/jp/labo/cis.html

６．研究組織

(1)研究代表者
厚井　裕司(KOI　YUJI)
岩手大学・工学部・教授
研究者番号：20333750

(2)研究分担者（２００７年度まで）

中谷　直司（NAKAYA NAOSI）
岩手大学・工学部・助教
研究者番号：20322969

吉田　等明（YOSIDA HITOAKI）
岩手大学・情報処理センター・准教授
研究者番号：00220666

(3)連携研究者（２００８年度）

中谷　直司（NAKAYA NAOSI）
岩手大学・工学部・助教
研究者番号：20322969

吉田　等明（YOSIDA HITOAKI）
岩手大学・情報処理センター・准教授
研究者番号：00220666