第22卷第10期 常熟理工学院学报(自然科学) Vol. 22 No.10
2008年10月 Journal of Changshu Institute Technology(Natural Sciences) Oct.,2008

基于语义的启发式病毒检测引擎研究

崔 鹏

(辽东学院 信息技术学院,辽宁 丹东 118003)

摘 要:对形式化语义在启发式病毒检测引擎上的应用进行了研究,分析了基于虚拟机技术的反病毒检测引擎的效率问题及启发式病毒检测引擎,发现其规则的选取对病毒判断的准确率影响很大,提出基于形式化语义的启发式病毒检测引擎.进行了计算机病毒语义关系框架及数据结构的设计,通过深入剖析不同病毒程序传染部分的代码结构,总结出病毒程序的典型语义特征,形成描述其典型语义特征的语义关系框架;在检测时抽取蕴涵在待检程序中的语义,进而建立描述程序语义关系框架;计算二者的匹配程度来确定程序是否含有恶意程序,从而达到检测未知病毒的目的.

关键词:虚拟机;启发式病毒检测 ;语义框架

中图分类号:TP393.07 **文献标识码**:A **文章编号**:1008-2794(2008)10-0095-05

主动防御技术、启发式检测技术、虚拟机技术、基于免疫原理的病毒检测技术及人工智能技术是当前流行并且有效的病毒检测技术,虽然这些技术和早期的技术相比有了极大的改善,但无论哪种技术究其实质在病毒分析检测时还是需要对病毒代码进行分析,只不过分析的角度、方法不同,然后在分析的基础上建立病毒特征库供以后查杀病毒使用.

启发式病毒检测引擎是在多态病毒大面积流行的情况下出现的,多态病毒和早期的病毒区别在于采用了自动变形技术[1],在外观形态上没有固定的特征码.因而现在对病毒的检测不再是以检测特征码为主要手段.启发式病毒检测引擎是综合了虚拟机技术和传统的反病毒检测引擎技术发展来的[2],比现存其他技术有一定的优势,但也有明显的缺陷:病毒特征行为提取困难,病毒行为特征权值大小确定困难,本文研究的基于语义的启发式病毒检测引擎对以上问题进行了改善.

1 基于虚拟机技术的反病毒检测引擎

虚拟机技术是一种前沿的反病毒新技术,主要用来分析未知病毒和查、杀多态变形病毒.具体的思路是用程序代码虚拟CPU、各个寄存器甚至是硬件端口,将采集到的病毒样本送入该虚拟环境中执行,通过分析内存和寄存器以及端口的变化来了解程序的执行情况.当虚拟机技术加入病毒检测引擎中,由于该技术采用动态分析程序的变化,对于多态变形病毒和未知病毒的发现准确性很高.因为对于多态病毒,无论如何变化代码和加密代码,可是最终执行时刻还是会暴露出其破坏方式[3].基于虚拟机技术反病毒检测引擎的工作原理如图1所示[4].

首先,文件类型检测模块将检测的目标文件分为二进制可执行文件与文本类型两大类.目前基于文本格式的计算机脚本病毒已经成为目前的主流形式,文本类型的检测对象如果是OFFICE 文件,因为其中含有basic

收稿日期:2008-07-10

作者简介:崔 鹏(1976—),男,辽宁丹东人,辽东学院信息技术学院讲师,硕士,研究方向:计算机软件和网络应用.

宏代码,先经过一个预处理器提取其宏代码后交给下级语法分析器处理,分析完语法后再由脚本检测引擎从病毒特征码库中提取脚本检测码进行模式匹配,最后将结果送GUI 通知用户检测结果;对于普通的文本类的源程序文件就直接交语法分析器处理,省去了提取内含basic代码的过程.二进制执行文件先检测是否存在某种类型的加密外壳,如果是一个有外壳的执行文件,则转入一个递归的脱壳模块直到脱出真正的执行代码体;接着是检测变形病毒的虚拟机执行一段指令,对存在解密指令段的程序进行变形病毒的检测操作,如果不是则跳出虚拟机进入二进制代码的检测模块,然后将检测的结果通过GUI通知用户.根据计算机用户的指令,病毒检测的结果送入杀毒模块中,从杀毒规则库中提取相应的规则生成杀毒脚本,随后送入杀毒引擎中执行该脚本,整个检测、清除计算机文件型病毒的过程结束.

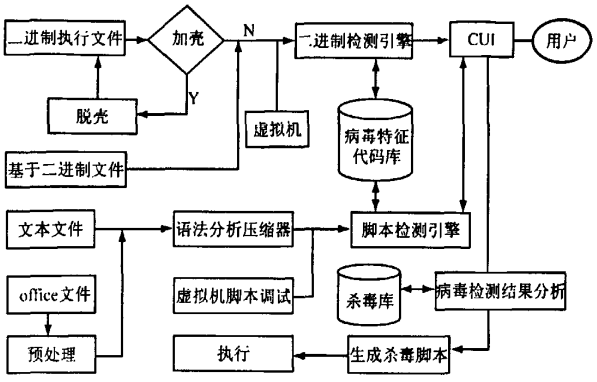

图1 基于虚拟机技术反病毒检测引擎结构图

但是该技术虚拟的CPU 执行速度比真正的CPU 慢10倍以上, 所以在查、杀效率上有待于提高[5].多数检测病毒情况发生在怀疑计算机系统被感染病毒时,此时只有少量的文件被病毒感染,多数文件处于未感染病毒的正常状态,并且由于多态病毒属于编写难度大的病毒,所以多态病毒的数量更少.反病毒软件在使用虚拟机时,是将每个执行文件放入虚拟机中运行一段时间,发现异常后在代码还原的状态下继续使用特征匹配来检测病毒.因此多数情况下,虚拟机是用机器效率来换取查准率[6].启发式病毒检测引擎的出现大大改善了虚拟机技术的效率问题.

2 启发式病毒检测引擎对虚拟机技术的改进

启发式[7]指的是具有自我发现的能力、运用某种方式和方法来判定事物的知识和技能.启发式分析就是利用计算机病毒的行为特征,结合以往的知识和经验,对未知的可疑病毒进行分析与识别.启发式代码扫描技术实际上就是把这种经验和知识移植到一个反病毒软件的具体程序中.它的基本原理是通过对一系列病毒代码的分析,提取一种广谱特征码,即代表病毒的某一种行为特征的特殊程序代码,当然仅仅是一段特征码还不能确定一定是某一种病毒,通过多种广谱特征码,也就是启发式规则的判断,综合考虑各种因素,确定到底是否是病毒,是哪一种病毒.

启发式扫描主要分为两类——静态启发式扫描和动态启发式扫描.静态启发式扫描程序有一个巨大的代码数据库,数据库把每一个代码串(称为特征码)与它所代表的行为特征联系在一起.并给每一个行为特征赋一个加权值.这些特征码可能会包含检查日期、文件大小或试图访问地址簿的指令和一些非正常的指令和行为.扫描程序扫描一个文件如果碰到符合自己代码数据库中的代码,就给这个文件做一个标记,并赋给一定的权值.有时候带标记的可能是一个无害的应用程序,但也要记录下来,当一个文件的标记的权值加起来超过一个底线,则这个文件可能是病毒.由此看来,静态试探式扫描技术并不精确,是特征码扫描技术的延伸.它是一种静态的行为检测技术,即把病毒可能执行的一些行为也作为特征码加入了病毒库中.动态启发式扫描技术主要增加了对CPU的模拟.对于加密病毒来说,直接使用静态启发式扫描是不能判定的.动态启发式扫描则解决了这个问题,它模拟了一个基本运行环境的计算机,对可能是病毒的文件先进行模拟运行,等加密病毒自解密后再进行静态启发式扫描.

图2是对基于虚拟机技术病毒检测引擎的完善:首先根据病毒特征库对病毒行为特征模糊量化处理,对每一个行为特征加一个模糊权值[8],根据各种行为特征权值判断是否异常.其次注意分析每个行为特征的关联,比

如C特征(修改计算机基本配置),如果只是单独出现此行为,则认为属于正常,因为这可能是某软件的安装程序;如果C与A特征同时出现,则该文件被病毒感染的概率就提高了,因此其检测权值相应地要提高.这种考虑行为特征关联而动态调整文件检测权值的方法有利于提高综合分析病毒程序行为的准确性.在改进的病毒检测引擎中,病毒行为分析器根据简单的规则,分析文件的检测权值,如果超出一定的阈值就送虚拟机处理,作为疑似多态病毒处理;反之,认为是健康文件,放弃该文件的检测工作.

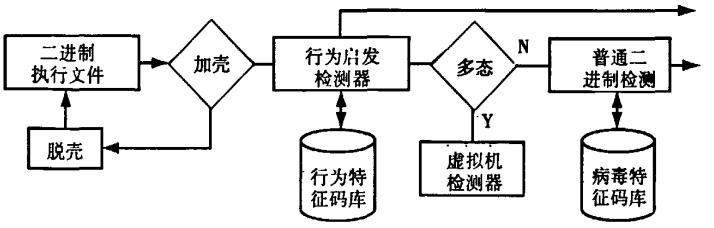

图2 启发式病毒检测引擎局部结构图

上述改进提高了虚拟机的效率.但是启发式诊断的正确率(包括检测率和误报率)和规则的选取有密切的关系.往往是某些规则对某种病毒很有效,但是又影响其它类型的病毒的检测,基于语义的启发式病毒检测对上述问题又进行了改进.

3　计算机病毒语义关系框架及数据结构的设计

病毒程序区别于正常程序的传染部分代码,包含着一些相对稳定的模块[9](程序段),因此,可以通过对相关病毒代码结构特点的分析研究,找到传染模块中相对稳定的病毒表达模式.根据对传统病毒、蠕虫病毒及木马程序典型行为分析总结出病毒程序关系框架Gy,如图3所示.

为了更合理地描述计算机病毒传染行为的语义特征,把不同病毒传染模式的语义特征完整、准确地描述出来,同时考虑到病毒技术不断变形、更新或变化,因此设计了面向对象的形式化描述.该方法是在人工智能框架知识表示的基础上,应用面向对象定义的形式化描述.如图4所示.

在这里恶意代码作为一个类,该类下的具体对象就是各种病毒程序包括传统病毒、蠕虫、木马等,每种病毒程序的典型传染行为语义特征作为事物的属性.知识库采用三层框架结构来存储.顶层框架描述计算机病毒的种类,即传统病毒、蠕虫和木马;每种具体的病毒程序则构成中间层框架.该层框架主要描述不同病毒程序传染行为的典型语义特征,它与顶层框架是隶属关系,反映该框架描述的语义特征是计算机病毒程序中的一种,同时继承了上层框架的所有特征;第三层框架描述的则是不同病毒程序中具体行为模块的语义特征.随着计算机及网络技术的发展演化出新的病毒技术,可以直接在此基础上扩展.

4　计算机程序语义框架建立

为了获得与程序对应的语义关系框架,必须实现原程序到语义关系框架的转换.可将转换系统视为映射函数T,并令:P=S1;S2;……;Si;……Sn;P表示原程序,其中Si(1≤i≤n)为P中的语句,n为原程序P的长度.令:Gp=(N,E)它是一个由节点集N和弧集E组成的框架,于是有T:P→Gp,映射函数T要解决的关键问题是如何在扫描原程序的过程中识别并提取程序的语义特征.本文借鉴了文献[10]中的框架抽取的思想和算法,并在此基础上做了改动使其满足病毒检测的需要.基本思想是:根据指令分类信息,首先对原程序进行子程序划分,确定子程序间的调用和被调用关系(也称为依赖关系),并用一棵结构树来表示这种依赖关系,继而抽取出原程序的整体框架结构.

对于一个代码量很大的程序要抽取出整个语义关系框架是很复杂的,本系统采用分治策略,将待检程序划分成多个子程序,然后再抽取每个子程序的框架结构.为描述每个子程序的框架结构信息、流程信息和子程序间的依赖关系,设计了子程序链结点、调用链结点、多次调用链结点、被调用链结点和程序块链结点五种数据结构.

其中每种数据结构的功能是:每个子程序结点描述了该子程序的内部结构特征;它的调用链描述了该子程序调用过的所有子程序的特征信息;调用链结点的多次调用链描述了它多次调用某个子程序的特征信息;它的

被调用链描述了该子程序被调用的特征信息;它的程序块链描述了该子程序的流程信息.所有的这些特征信息就组成了该子程序的框架结构信息以及与其它子程序之间的依赖关系,整个子程序链就描述整个程序的框架结构.

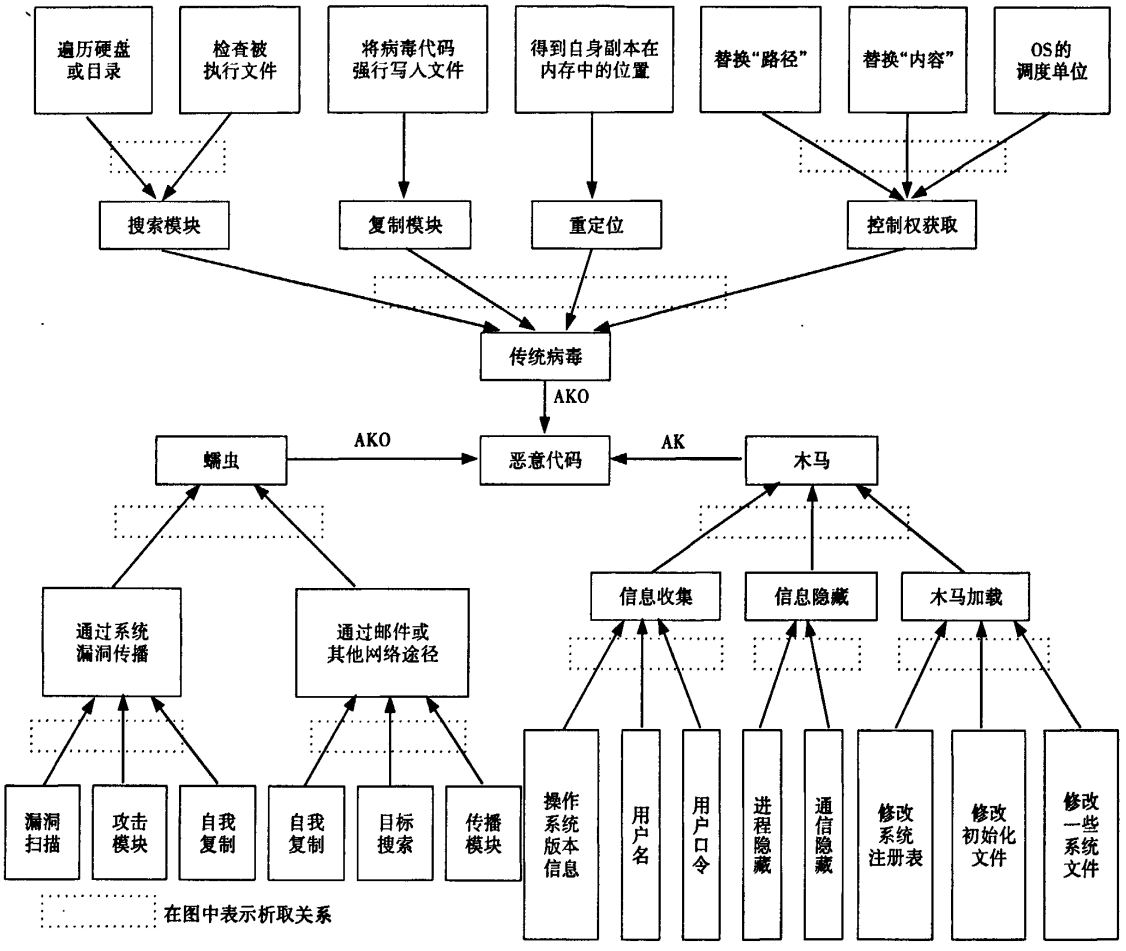

图3 计算机病毒传染行为模式的语义关系框架Cy

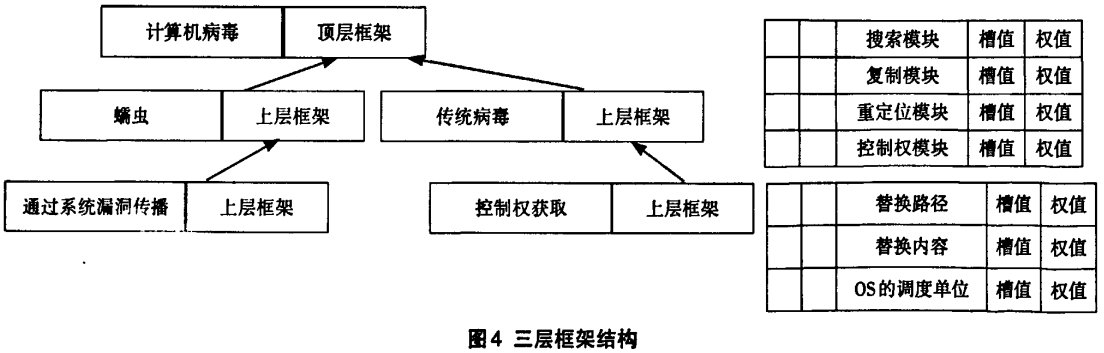

图4 三层框架结构

5 基于语义的启发式病毒扫描引擎

通过深入剖析不同病毒程序传染部分的代码结构,从而总结出病毒程序的典型语义特征,形成描述其典型语义特征的语义关系框架Cy;在检测时抽取蕴涵在待检程序中的语义,进而建立描述程序语义关系框架Gp;

计算二者的匹配程度来确定程序是否含有恶意程序,从而达到检测未知病毒的目的.

在检测前通过经验预先为阈值D确定一个固定的值,然后将语义关系框架抽取系统形成的程序语义关系框架相似度Y值清零.Gp中的槽值与Gy中的槽值进行匹配.若匹配,则相应的槽值标志位置1,即wi=1;若不匹配则标志位置0,即wi=0.经过匹配可以得到相匹配的槽的集合为{wl, w2, ……, wn},相应的权值为{xl, x2, ……, xn},按下面的函数计算Gp和Gv匹配程度.Similar()=a+wl*xl+wZ*xZ+……+wn*xn,其中a为整体误差,即由于各槽权值的设定不完善而引起的误差,在当前条件下可认为a=0,Y=Similar();将计算得到的相似度Y值与阈值D进行比较,若Y值大于等于阈值D,则认为Gp与Gy是匹配的,文件中含有恶意程序,反之则认为不匹配,文件中不含恶意程序,从而解决了启发式病毒检测误报率高的问题.

6 总　结

本文从虚拟机技术发展到启发式病毒检测技术,最后发展到基于语义的启发式病毒检测技术,解决了反病毒程序效率和准确率问题.该技术具有不需要病毒库,病毒库维护工作量小;不需要频繁升级;能防范未知病毒等特点.但是在阈值的选取和人工智能抽取程序语义方面还存在局限,有待研究改进.

参考文献:

[1] 慈庆玉,计算机变形病毒技术探讨[J]. 中国数据通信,2005,(1):37-40.

[2] 王振海,王海峰,基于多态病毒行为的启发式扫描检测引擎的研究[J]. 实验室研究与探索,2006,25(9):1089-1091.

[3] 曾宪伟,张智军,张志.基于虚拟机的启发式扫描反病毒技术[J]. 计算机应用与软件,2005,22(9):1252126.

[4] 王海峰,段友祥. 基于行为分析的病毒检测引擎的改良研究[J]. 计算机应用,2004,12:146.

[5] 王海峰,段友祥. 针对计算机黑客型病毒的网络防御体系研究[J]. 微型机与应用,2004,(6):426.

[6] 唐常杰,胡军. 计算机反病毒技术[M]. 北京:电子工业出版社,1990.

[7] 金晶,何昆,张世永. 基于智能扫描的病毒监视器研究[J]. 计算机工程,1999,(12):86-88.

[8] 段友祥,王海峰,满成城. 模糊逻辑在基于AIS的主机入侵检测中的应用[J]. 计算机工程与设计,2005,26(9):2450-2452.

[9] Singh P K, Lakhotin A. Static Verification of wormand Virus behavior in Binary Exeeutables using ModelCheeking [C]. IEEESystem,ManandCybemetiesSoeiety.information Assuranee Workshop,2003:298-300.

[10] 张有为,罗君宏,汪永红. 通用的汇编源程序框架分析技术研究[J]. 计算机工程与设计,2006,27(2).

Study on Semantic Anti-Virus Engine Based on Heuristic

CUI Peng

(Eastern Liaoning University, Dandong 118003, China)

Abstract: Because Virtual Machine technologe and Anti-Virus Engine Based on Heuristic have Obvious blemish in virus scanning, on at analysis Program Semantic and Virus Semantic Foundation are put forward based on semantic Anti-Virus Engine Based on Heuristic.By thoroughly analyzing different virus procedure to infect part of code structure, this paper has found typical semantic characteristic Program semantic; by thoroughly analyzing semantic in Program, the paper has found typical semantic characteristic Program semantic. Besides,two similar degrees are thoroughly compared with each other to judge whether a program is a virus .

Keywords: Virtual Machine; Anti-Virus Engine Based on Heuristic; Program Semantic

基于语义的启发式病毒检测引擎研究

作者： 崔鹏， CUI Peng
作者单位： 辽东学院,信息技术学院,辽宁,丹东,118003
刊名： 常熟理工学院学报
英文刊名： JOURNAL OF CHANGSHU INSTITUTE OF TECHNOLOGY
年，卷(期)： 2008,22(10)

参考文献(10条)

1. 慈庆玉 计算机变形病毒技术探讨[期刊论文]-中国数据通信 2005(01)

2. 王振海;王海峰 基于多态病毒行为的启发式扫描检测引擎的研究[期刊论文]-实验室研究与探索 2006(09)

3. 曾宪伟;张智军;张志 基于虚拟机的启发式扫描反病毒技术[期刊论文]-计算机应用与软件 2005(09)

4. 王海峰;段友祥 基于行为分析的病毒检测引擎的改良研究[期刊论文]-计算机应用 2004(12)

5. 王海峰;段友祥 针对计算机黑客型病毒的网络防御体系研究[期刊论文]-微型机与应用 2004(06)

6. 唐常杰;胡军 计算机反病毒技术 1990

7. 金晶;何昆;张世永 基于智能扫描的病毒监视器研究[期刊论文]-计算机工程 1999(12)

8. 段友祥;王海峰;满成城 模糊逻辑在基于AIS的主机入侵检测中的应用[期刊论文]-计算机工程与设计 2005(09)

9. Singh P K;Lakhotin A Static Verification of wormand Virus behavior in Binary Exeeutables using ModelCheeking 2003

10. 张有为;罗君宏;汪永红 通用的汇编源程序框架分析技术研究[期刊论文]-计算机工程与设计 2006(02)

本文读者也读过(10条)

1. 王振海.王海峰.WANG Zhen-hai.WANG Hai-feng 基于多态病毒行为的启发式扫描检测引擎的研究[期刊论文]-实验室研究与探索2006,25(9)

2. 崔鹏.CUI Peng 基于形式化语义的启发式病毒检测引擎研究[期刊论文]-辽东学院学报（自然科学版）2008,15(3)

3. 谭云松.Tan Yunsong 一种启发式反病毒技术的研究[期刊论文]-网络安全技术与应用2006(11)

4. 黄锦煜 浅谈基于主动防御的网络病毒防御技术[期刊论文]-科教导刊2009(4)

5. 王振海.王海峰.Wang,Zhenhai.Wang,Haifeng 针对多态病毒的反病毒检测引擎的研究[期刊论文]-微计算机信息2006,22(27)

6. 赵晓丽 基于语义的典型网络病毒过滤系统研究[期刊论文]-中国科技信息2009(19)

7. 张秋霞.孙秀英 计算机反病毒引擎策略[期刊论文]-黑龙江科技信息2007(16)

8. 李果.LI Guo 计算机病毒检测技术分析与对比[期刊论文]-工程建设与设计2007(10)

9. 李洪敏.凌荣辉 反病毒引擎技术初探[会议论文]-2003

10. 左志宏.周明天.ZUO Zhi-hong.ZHOU Ming-tian 计算机病毒描述语言VDL[期刊论文]-计算机应用研究2005,22(10)