基于程序行为特征的病毒检测技术与应用

The Study of Anti – Virus Engine Base On the Analysis Of Virus Behaviors

王海峰 (临沂师范学院信息学院 山东临沂 276002)
夏洪雷 (临沂师范学院信息学院 山东临沂 276002)
(中国石油大学(华东)计算机与通讯工程学院 257061)
孙 冰 (中国石油大学(华东)计算机与通讯工程学院 257061)

摘要:分析计算机病毒行为特点,在此基础上提出基于病毒行为特征的检测方法,主要针对多态计算机病毒与变种病毒的检测。检测主体思路是分析病毒程序,提取其系统 API 函数调用序列,生成代表病毒行为特征的检测向量。然后通过计算向量之间相似性来决定待检测文件是否感染病毒。最后将此方法应用到反病毒引擎设计中。实验结果证明了行为特征向量检测有很好的应用前景。

关键词:病毒行为 反病毒引擎 多态病毒 特征向量

1 引言

近几年中计算机病毒正以惊人速度蔓延,对计算机系统安全构成严重威胁。早期计算机病毒没有使用自动变形技术,具有固定特征码。因此,反病毒软件利用病毒特征码匹配很容易检测出隐藏在系统中的病毒。病毒和反病毒技术这种"矛与盾"的较量中,反病毒专家尽管采用了各种各样的方法来检测计算机病毒,但是新病毒还是层出不穷,而且技术水平越来越高,隐蔽性越来越强。目前许多病毒采用自动变形技术来逃避特征码检测技术的检测,这就是所谓的多态病毒[1]。多态病毒是指每次传染产生的病毒副本在外观形态上都发生变化的病毒。因此,多态病毒在外观形态上没有固定的特征码。另一方面,病毒变种造成的危害远大于原病毒。如果计算机系统未能及时修复系统漏洞,病毒制造者修改源代码饶过反病毒软件的检测,变种病毒继续在网络中蔓延。

反病毒软件的检测策略是病毒特征码检测法。目前主要提取病毒特征码,由于病毒特征码是病毒机器码层的特征,因此具有很高的敏感度。某病毒的特征码不仅能区分病毒程序与非病毒程序,而且能区分该病毒与其他病毒。病毒特征码的高敏感性导致了无法完成对多态病毒与变种病毒的检测。机器码层属于低级特征层次,如果从更抽象的层次提取病毒特征就可以解决此问题。本文从程序行为特点方面考虑,由于计算机病毒属于行为比较特殊的程序,因此行为特征可以区别病毒与正常计算机程序,并且属于更抽象层次的特征,其敏感度较机器特征码会适度降低[2]。

2 病毒行为特征检测

2.1 病毒行为特点

病毒程序不同于普通计算机程序,带有破坏性与复制自身的特征。给每个病毒典型行为分类。以下给出一种具体分类方案:

(1) D:解密模板,多态病毒必然具有的行为特征。由于多态病毒感染文件时被随机加密,并且在其执行时解密模块是多变的代码。但是病毒设计者根据一个固定解密结构并利用相同功能程序演化手段进行随机变化,其解密模板一般是固定的,可以提取其特征码作为重点怀疑特征;

(2) G:解密库,因为多态病毒利用解密模板随机生成解密指令,所以必然携带等价指令变形库。该指令数据库有一定规模并且有很强的特征,可以作为行为识别特征;

(3) F:异常文件访问,病毒程序在感染时一般要遍历系统中所有执行文件,这是普通程序一般没有的

操作，可以作为重要行为怀疑特征；

（4）A：异常文件结构，比如 PE 文件头部出现异常标记，这可能是病毒判断感染的标志；PE 文件的最后一个节是可执行属性，这可能就是被病毒感染后添加的病毒体；PE 文件的入口点发生改变等；这些都可以作为行为怀疑特征；

（5）M：针对内存区域操作指令数量，病毒在感染和执行时会有大量内存区域的清除、移动、替换等操作，这类指令可以作为行为怀疑特征；

（6）C：修改计算机系统基本配置的指令，比如在注册表中添加启动项、注册服务进程、修改配置文件，由于普通的软件也有这类指令，所以只能作为行为怀疑特征；

（7）R：重定位，寄存型病毒在宿主程序中必须进行变量重定位，这是普通程序不具有的特点，因此可以作为行为识别特征；

（8）！：可疑指令，比如有的病毒运用抗虚拟机分析指令、为了引起结构异常故意使用错误指令、无效指令甚至 Intel 未公开指令，这些可作为行为怀疑特征[4]。

2.2 病毒行为特征码

上一小节讨论了病毒程序的行为特点，必须将这些行为特点量化处理为反病毒软件可以使用的特征码。本节主要设计病毒行为特征码。病毒基本上利用操作系统提供的 API 函数来实现各种功能。因为病毒要求短小、精悍，所以很多对计算机系统的操作必须调用操作系统提供的接口函数，而不是自己实现，其唯一目的就是尽量减少代码体积。系统 API 函数调用序列就能反映病毒的行为特点，比如对注册表的操作、对文件系统进行穷举搜索等，实现这些功能时 API 函数调用序列十分相似。变种病毒整体 API 调用序列十分相似。因此提取出原病毒 API 序列后，利用相似性判断可以检测出该病毒的各种变种。

首先，每个系统 API 函数调用表示为四字节的数字串，格式为 <模块序号，函数序号 >。将系统或主要编译系统中各种主要模块（各种动态连接库，如操作系统的 Kernel. dll，User32. dll；Visual C + + 6. 0 的 MFC42. dll 等）进行编号，占用两字节；再将各模块中函数进行编号，占两个字节。API 函数调用序列形式为 < 模块 1，函数 x，模块 2，函数 y，…，模块 v，函数 z >。利用反编译软件分析并提取已知病毒的 API 函数调用序列，建立病毒行为特征库。

2.3 病毒行为码检测

利用 API 函数调用序列作为病毒行为特征码，API 函数调用被量化处理后表示为向量形式。因此采用比较向量距离的相似性作为检测手段。提取可疑样本中 API 函数调用序列，量化处理为向量 Vu，计算 Vu 与行为特征库中已知病毒特征码 Vs 的向量距离。如果距离 Distance（Vs，Vu）< = ε（ε 表示病毒相似性阈值），则认为样本感染了该病毒。向量距离采用欧氏距离与向量加角余弦距离（如公式 1，2 所示），取两种距离的平均值作为向量相似性判断的依据。

$$D(Vs,Vu)=\left[\sum_{i=1}^{\min(|Vsl|,|Vul|)}(Vsi-Vui)^2\right]^{1/2}$$

公式 1 欧氏距离

$$\cos(Vs,Vu)=\frac{Vs^{T}Vu}{\|Vs\|_2\cdot\|Vu\|_2}$$

$$\|V\|_p=\left[\sum|Vi|^p\right]^{1/p}$$

公式 2 夹角余弦公式

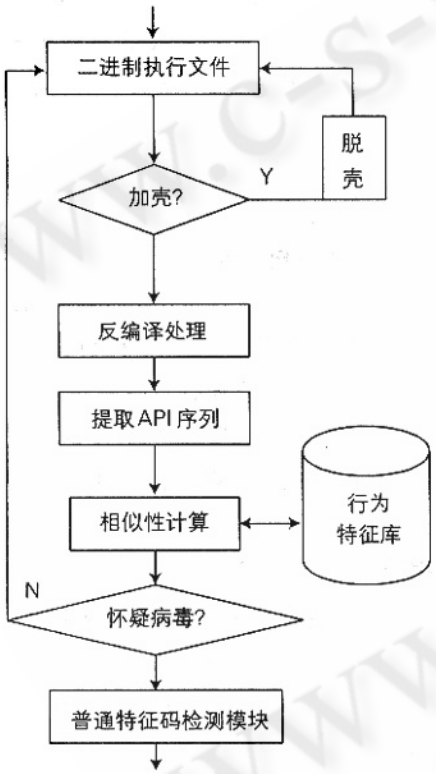
3 反病毒引擎原理与设计

早期反病毒软件只是针对个别计算机病毒，当时程序与检测所需数据紧密偶合[5]。随着病毒数量增多，反病毒软件逐渐演化为检测程序与独立特征码数据库的方式。依赖于特定数据集合上的一组用于计算机病毒检测和清除的程序序列被称为计算机病毒检测引擎。本文讨论的反病毒引擎以 Windows 系统中 Portable Executable（PE）文件格式为主要研究对象，以下给出文件型病毒检测引擎的结构图。

首先反病毒引擎检测是否加壳，如果执行文件存在保护壳，则脱壳后继续处理。接着利用反编译模块搜索执行文件中的系统 API 调用或者特定软件开发包中函数调用，经过分析整理出与病毒程序特点密切相关的 API 函数调用序列。将这种 API 函数调用序列转化为检测向量形式，再与行为特征库中病毒行为特征向量逐一计算相似性。如果相似性判断出可疑文件肯定是病毒或者肯定不是病毒，则系统继续处理后续文件；如果仍然判断为可疑病毒，则由普通特征码检测模块继续进行处理。

4 实验设计与结果分析

实验主要验证病毒行为特征向量检测的准确性与广谱性。实验设计思路如下:采用 IDAPRO 反编译软件仿真病毒检测引擎的反编译模块。IDAPRO 提供了强大的脚本功能,使用脚本编程可以将反编译程序的 API 函数调用输出到普通文本文件。然后利用 PERL 程序将文本文件中与病毒特点相关性强的 API 函数(网络传输函数、注册表函数、文件操作函数、内存操作函数等)调用序列提取出,并生成检测向量。最后进行相似性判断,输出报告结果。因为该仿真实验主要是利用脚本程序模拟,所以执行效率上有折扣。



首先,验证行为特征向量的广谱性。人工分析并提取生成了 W32. Mydoom. A 与 W32. Blaster. Worm 两种病毒的行为特征向量。相似性判断的阈值 ε 取 0. 9,然后将 W32. Mydoom. A 的 V1 – V7 与 W32. Blaster. Worm 的 V1 – V4 等 11 种变种病毒放入计算机[6],使用脚本程序生成各变种病毒的待检测向量,结果全部可以检测出。由此可见反映程序行为的特征向量具有很强的广谱性,可以准确检测出变种病毒。

其次,验证多态病毒检测准确性。首先设计一个普通具有蠕虫特点的实验病毒。该病毒自带 SMTP 引擎,可以使用电子邮件自动传播;利用 RPC DCOM 漏洞传播;修改注册表,实现自启动;将自己设为服务,实现后台运行。然后利用多态引擎对该病毒进行变形处理,主要变形原理是插入垃圾指令,等效指令替换,修改控制流。提取普通状态时的 API 函数调用序列作为行为特征向量,对每次变形后的病毒程序进行检测。实验中该病毒进行了 100 次变形,相似性低于 0. 9 的变形只有一次。由此可见行为特征向量检测多态病毒有很高准确性。

5 总结

有些商用反病毒软件提出了基于病毒行为进行启发式扫描,但是并没有简单有效的可行方案。在此提出程序 API 函数调用序列来反映病毒行为特点,并采用向量形式进行表示。该方法简单、可行,很容易进行相似性计算。检测多态病毒和变种病毒具有很高的准确性。虚拟机一直是检测多态病毒的主要手段,但是使用虚拟机技术降低了反病毒引擎的检测效率。这种行为特征向量检测方法可以克服虚拟机的缺点,提高反病毒引擎对多态病毒的检测效率。该方法建立在静态反编译的基础上,如果使用高效编程语言实现其中的主要模块,其检测效率会有更好的应用前景。同时基于行为特征检测方法可以作为目前反病毒引擎中一个重要的启发式模块,提高反病毒引擎的检测效率。

参考文献

1 祝恩、殷建平等,计算机病毒的本质特性分析及检测[J],计算机科学,2001,28(增刊):238 – 240。
2 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,微型机与应用,2004.6 (4 – 6)。
3 Understanding Virus Behavior in 32 – bit Operating Environments. Symantec,1997.
4 Virus Library http://www.viruslibrary.com/virusinfo
5 Fred Cohen. Computer Virus Theory and Experiments[J],Computer&Security,1987,6(1):22 – 35
6 Symantec Cooperation http://securityresponse.symantec.com/avcenter/.