反病毒引擎及特征码自动提取算法的研究

金 庆[1,2]， 吴国新[1,2]， 李 丹[1,2]
(1. 东南大学 计算机科学与工程系，江苏 南京 210096;
2. 东南大学 计算机网络和信息集成教育部重点实验室，江苏 南京 210096)

摘 要：随着网络的广泛普及，计算机病毒带来的安全威胁日趋严重。提出了一种反病毒引擎的设计方案，该设计采用3种特征码格式(MD格式、两段检验和格式、字符串格式)。同时，又提出了针对VB应用程序的病毒特征码自动提取算法。最后通过实验1对这3种特征码格式进行了性能比较，通过实验2对自动提取算法的有效性和准确性进行了验证。

关键词：计算机病毒；反病毒引擎；病毒特征码；自动提取算法；计算机安全
中图法分类号：TP309 **文献标识码**：A **文章编号**：1000-7024 (2007) 24-5863-04

Research of anti-virus engine and automatic extraction of computer virus signatures

JIN Qing[1,2], WU Guo-xin[1,2], LI Dan[1,2]
(1. Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China; 2. Key Laboratory of Computer Network and Information Integration, Ministry of Education South University, Nanjing 210096, China)

Abstract: With the wide popularization of the Internet, security threats brought by computer virus become more and more serious. A design of anti-virus engine and an automatic extraction-arithmetic of computer virus signature are presented. This anti-virus engine adopts three formats for signatures. The first experiment compares the performance of the three formats and the second experiment demonstrates the accuracy and validity of the automatic extraction-arithmetic.

Key words: computer virus; anti-virus engine; virus signature; automatic extraction-arithmetic; computer security

0 引 言

多年来，计算机病毒一直是计算机用户的心头大患，特别是CIH病毒的出现，它不仅能够破坏计算机的软件系统，而且通过利用微软的VxD(虚拟设备驱动)技术，直接对硬盘的物理扇区进行写操作，从而破坏计算机系统的Flash BIOS芯片中的系统程序，导致主板损坏，给用户造成莫大的损失[1]。历年重大病毒影响情况如表1所示[1]。

在这个计算机病毒肆虐的时代，一个好的反病毒软件显的犹为重要。反病毒软件是由病毒扫描程序和反病毒引擎两部分构成的：病毒扫描程序处于软件前台，其主要功能就是为反病毒软件与用户提供交互接口，把扫描对象提交给反病毒引擎进行病毒扫描；反病毒引擎主要实现对前台传入的扫描对象进行文件格式分析和病毒扫描，将扫描的中间结果和最终结果返回给前台，并依据前台的返回结果进行相应的处理，同时引擎还肩负着病毒特征码库的加载、管理、升级等责任。

在整个反病毒软件的体系结构中，反病毒引擎是整个反病毒软件和各种反病毒应用的基础。本文着重介绍一种反病毒引擎的设计方案，并在此基础上提出了一种病毒特征码自动提取算法。

表1 历年重大病毒影响情况

病毒名称	持续时间	造成的经济损失
莫里斯蠕虫	1988年	6000多台计算机停机，直接经济损失达9600万美元
美丽杀手	1999年	政府部门和一些大公司紧急关闭了网络服务器，经济损失超过12亿美元
爱虫	2000年5月至今	众多用户电脑被感染，经济损失超过100亿美元
红色代码	2001年7月	网络瘫痪，直接经济损失超过20亿美元
求职信	2001年12月至今	大量病毒邮件堵塞服务器，经济损失达数百亿美元
Sql蠕虫王	2003年1月	网络瘫痪，银行自动提款机运行中断。经济损失超过20亿美元

1 系统设计

本引擎主要包括如下3个模块：特征码装载模块、病毒扫描模块和文件解析模块。三者关系如图1所示。

首先，反病毒引擎接收前台程序传入的扫描对象，并对其文件格式进行解析，这部分工作有文件解析模块完成；然后将

收稿日期：2006-12-15 **E-mail**：jq8205@163.com
作者简介：金庆 (1982－)，男，浙江绍兴人，硕士研究生，研究方向为计算机网络安全、网络应用；吴国新 (1956－)，男，安徽歙县人，教授，博士生导师，研究方向为计算机网络安全、网络管理；李丹 (1983－)，女，江苏张家港人，硕士研究生，研究方向为计算机网络安全、网络应用。

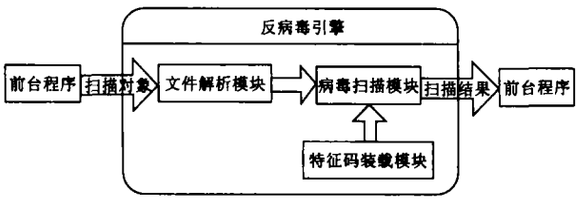

图 1 反病毒引擎的体系架构

解析结果传递给病毒扫描模块，该模块利用病毒特征码来扫描解析后的文件，如果文件与病毒特征码匹配，则断定该文件是病毒，给出病毒名，将结果返回给前台程序，反之认为不是病毒，继续扫描；特征码装载模块主要负责病毒特征码目标文件的装载和维护。

在反病毒引擎中至关重要的一块是病毒特征码的提取和维护。病毒特征码提取的准确性和及时性直接影响反病毒引擎的防毒效率，本系统采用 3 种病毒特征码的格式。分别是 MD5 格式，字符串格式和二段校验和(Two_checksum)格式。

1.1 MD5 格式

MD5 格式直接利用病毒样本的 MD5 值作为病毒特征码，用于检测特定病毒。该方法的优点是快速、简单，但一种特征码只能处理一种病毒，即使病毒微小的变动，都需要重新提取特征码，这样就造成特征码库过于庞大。该格式主要用于提取临时特征码，即当某种病毒突然爆发，而病毒特征码提取人员来不及对其进行准确处理，此时先用 MD5 格式得到一个临时特征码，用于防范该类病毒，然后在尽快的时间内利用字符串格式或二段校验和格式得到最终特征码。

1.2 字符串格式

字符串格式利用病毒文件特殊的字符串来表示一类病毒。该方式适用于所有的病毒。其优点是能够利用一段特殊的字符串来检查出一类病毒，而不是一种病毒。其缺点是需要耗费较多的扫描时间。

1.3 两段校验和格式

两段校验和格式是最普遍的病毒特征码格式，其包含两段病毒文件特殊位置的数据（通常是能代表该病毒特性）的 CRC 检验和。扫描文件时，先计算待查文件在该位置的校验和值，通过判断有无符合该值的特征码来断定文件是否是病毒。该方法准确率高，耗费时间少，能够利用一个特征码来检查出一类病毒等优点。

2 模块设计

2.1 特征码装载模块

特征码装载模块主要负责病毒特征码目标文件库的装入。病毒特征码目标文件库存放对病毒特征码源文件进行加密和压缩处理后得到特征码目标文件，处理的目的是为了保护病毒特征码的安全。病毒特征码目标文件主要包括两部分：文件头和数据体。文件头里包含了病毒特征码的版本、装载日期等信息，数据体包含了病毒的特征码值以及病毒名。

常见的特征码目标文件的组织格式是将病毒特征码和病毒名捆绑存放，即病毒特征码后面紧跟着其对应的病毒名，其好处是：简单，直接，处理速度快，但存在着资源浪费的问题。在实际中，存在多个不同的病毒特征码对应一个病毒名的情况。如果采用病毒特征码和病毒名捆绑存放，那么有多少个特征码就要存放多少个病毒名，而实际上这些特征码中许多是同一个病毒名的。这样的后果导致资源浪费，杀毒引擎过于庞大。本文对该组织格式进行的改进，将病毒特征码与病毒名分开存放，在每个病毒特征码后面存放一个偏移值 offset，该 offset 指向其病毒名，如图 2 所示。

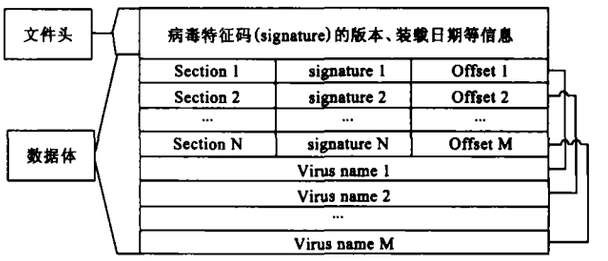

图 2 特征码目标文件的组织格式

这样，在特征码目标文件中，同一个病毒名只存放一次，该病毒名所有的病毒特征码都通过一个 offset 指向该病毒名。在 32 位机下，offset 所占空间最大不会超过 4 个字节，由于在实际情况中，病毒名数要远远少于病毒特征码数，所以该方法比病毒特征码和病毒名捆绑存放的方法要占更少的空间。

特征码装载流程如下：

(1) 解析特征码目标文件头，获取相关信息；

(2) 对特征码目标文件的数据体进行解密；

(3) 对解密后的特征码目标文件数据进行解压缩；

(4) 按照以下步骤逐一解析每一个节的数据，并装入到相应的病毒特征码容器中：①读取一行病毒特征码及其偏移值 offset；②利用 offset 恢复病毒名；③调用装载函数，解析并装载病毒特征码；④继续下一条病毒特征码。

2.2 文件解析模块

文件解析模块是反病毒引擎的重要组成部分，它主要有文件类型检测模块，解压缩模块，脱壳模块，脚本语法分析压缩模块，宏病毒预处理模块等组成。文件检测模块负责识别对象文件的类型，根据检测结果决定下一步操作。解压缩模块负责对打包文件进行解压缩。脱壳模块对加壳的文件进行脱壳，这主要是针对可执行文件(比如 PE、NE 等文件格式)。脚本语法分析压缩模块负责识别处理各种脚本文件，目前常见的有 vbs,js,php,perl 等脚本病毒，文件解析模块的实现流程如图 3 所示。首先文件类型检测模块对输入的文件进行类型检测，并根据检测出的结果决定操作，对于压缩文件，调用解压缩模块进行解压缩。对于脚本文件(包括 office 宏文件)，则调用脚本语法分析压缩器模块进行语法分析，并将结果输出，交由病毒扫描模块进行特征码匹配，如果是 office 宏文件还需要调用宏病毒预处理模块。对于二进制文件，需要区分是二进制可执行文件(如 PE、NE、COM、DOS 等)和其它二进制文件：对于二进制可执行文件需要判断是否加壳，如果有壳，需要调用脱壳模块进行脱壳。最终通过解析后输出，交由病毒扫描模块进行特征码匹配。

2.3 病毒扫描模块

病毒扫描模块主要负责对解析后的文件进行扫描，利用

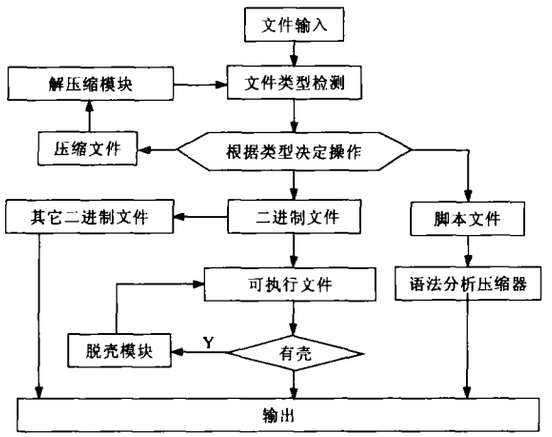

图 3　文件解析模块流程

特征码装载模块提供的病毒特征码去扫描文件，如果发现病毒，就通知前台程序，由前台决定下一步操作。

病毒扫描模块流程：①首先需要保证装载了最新的病毒特征码，这部分主要是调用病毒特征码装载模块完成；②病毒扫描模块接收由文件解析模块发送的解析后的扫描文件；③病毒扫描模块对解析后的文件类型进行判断；④根据解析后的文件类型调用相应的处理程序，进行扫描；⑤如果命中病毒特征码，引擎会发送相应的事件通知前台程序；⑥检查前台返回标志，决定下一步操作。

3　VB 应用程序病毒特征码自动提取算法

目前各家反病毒公司在特征码提取方面普遍使用手工提取的方法，这种方法虽然准确率较高，但是提取速度较慢。本文作者曾对几家反病毒公司的病毒特征码提取状况进行了调查，调查发现一个熟练的病毒分析师一天工作 8 小时计算，平均每天只能提取 12.8 个病毒的特征码。在病毒盛行的今天，一天就可能产生上百个病毒。12.8 个/天的处理速度显然是不够的，此时实现病毒特征码的自动提取显得犹为重要。本文在研究 VB 病毒的基础上，实现了一个 VB 编译程序病毒特征码自动提取算法：由于 VB 代码中项目信息和窗口信息位置比较固定，并且反映了程序实现的功能，所以提取的病毒特征码可以是程序入口点后的编译信息、项目信息和窗口信息(特别是启动窗口的信息)，算法思想是：先取程序入口点后第 19 个字节开始的 7 个字节，再取包括编译文件相关信息的 64 个字节，然后利用二段校验和的方法，得到病毒特征码。

本算法主要是针对 VB5.0 和 VB6.0 而设计的。由于 EXE 文件和 DLL 文件都属于 PE，且对于 VB 编写的 EXE 文件和 DLL 文件的判断过程是不同的，所以需要进行了区分处理。由于 EXE 和 DLL 可以通过 PE 文件头 (IMAGE_NT_HEADER 结构) 中的 Characteristics 字段进行区分，如果该字段值为 010fh，则为普通的可执行文件(EXE 文件)，若值为 210fh，则表示为 DLL 文件[5]。通过对 VB 编写的 EXE 文件结构和 DLL 文件结构深入分析后，发现如下规律：

规律 1　对于 EXE 文件，取其程序执行入口点后一个字节开始的 unsigned int 整型(可以确定其为一个 RVA，即相对虚拟地址)，将其转换为文件偏移，读取此位移处开始的 4 字节内容，判断是否是 VB5!，若是，则此 EXE 文件必为 VB 编译的。

规律 2　对与 DLL 文件，取导出表开始的 28 位移处的内容，再以此内容为位移读取数据，再以此数据为位移再加一后，读取此位置处内容，判断是否是 VB5!(在此过程中每次读取的内容都是 RVA，都需要先转换为文件中位移)。

(1)读取 DOS 文件头(IMAGE_DOS_HEADER)，判断其 DOS 标志位(e_magic)是否为“MZ”或“ZM”；

(2)读取 PE 文件头(IMAGE_NT_HEADER)，判断其 PE 标志位是否为“PE”；

(3)读取 PE OPTIONAL 文件头，并得到程序入口点地址和代码起始相对虚拟地址(relative virtual address，RVA)位置；

(4)根据 PE 文件头(IMAGE_NT_HEADER 结构)中的 Characteristics 字段进行区分 EXE 和 DLL；

(5)若是 EXE，则依据规律 1 判断是否是 VB 编译程序。若是 DLL，则依据规律 2 判断是否是 VB 编译程序，其中导出表的位置和大小可以从 PE 文件头中的数据目录中获取，与导出表对应的项目是数据目录中的首个 IMAGE_DATA_DIRECTORY 结构，从这个结构的 VirtualAddress 字段得到的就是导出表的 RVA 值；

(6)若判断出不是 VB 编译程序，则退出程序；否则找到程序入口点地址(文件偏移地址)：如果是 exe 文件，则跳过程序入口点代码后的 18 个字节，取后面的 7 个字节作为第 1 个特征码；如果是 dll，则需要跳过 36 个字节，取后面的 7 个字节作为第 1 个特征码；

(7)跳过第 1 个特征码 7 个字节后面的 25 个字节，以后面的 64 个字节作为第 2 个特征码；

(8)分别计算这 2 个特征码的校验和；

(9)得到该病毒特征码。

4　实验及结果描述

4.1　3 种病毒特征码的格式的性能比较

该实验进行了如下准备和实验。①分别准备了 MD5 方式(hdb)，字符串方式(ndb)，二段校验和方式 3 种类型的病毒特征码各 1 000、10 000、100 000、200 000 条；②分别用以上病毒特征码扫描同一个目录 (该目录包含各种类型的正常和病毒文件共 261 个，44 M)；③每种扫描各执行 10 次，去掉最长和最短时间之后求平均值；3 种格式花费时间的平均值如表 2 所示。

表 2　3 种格式花费时间的平均值

特征码个数	1 000	10 000	100 000	200 000
MD5 平均时间/S	5.026 625	4.077 625	6.179 752	4.317 375
二段校验和平均时间/S	5.550 753	4.315 251	8.536 125	13.412 211
字符串平均时间/S	5.924 752	15.634 371	52.626 251	110.446 932

3 种病毒特征码格式的性能特点：MD5 方式耗费时间最少，字符串方式耗费时间最多。

4.2　VB 应用病毒特征码自动提取算法的实验

在该实验中，我分别证明该算法的有效性和准确性。有效性证明实验如下准备：①准备 1 000 个病毒文件(其中 200 个

VB 编译文件),10 000 个病毒文件(其中 500 个 VB 编译文件),100 000 个病毒文件(其中 1 500 个 VB 编译文件),200 000 个病毒文件(其中 13 500 个 VB 编译文件);②利用 VB 病毒特征码自动提取算法自动提取特征码,统计个数。结果如表 3 所示。

表 3 VB 病毒特征码自动提取算法实验 1

病毒文件个数	1 000	10 000	100 000	200 000
VB 病毒文件个数	200	500	1 500	13 500
特征码提取个数	200	500	1 500	13 500

从表 3 可以得到,该算法能够有效的提取 VB 应用程序的病毒特征码,有效率 100%,从而证明了该算法的有效性。

准确性证明实验如下:①准备 1 000 个 VB 应用程序文件(其中 200 个病毒文件),10 000 个 VB 编译文件(其中 500 个病毒文件),100 000 个 VB 编译文件(其中 1 500 个病毒文件),200 000 个 VB 编译文件(其中 13 500 病毒文件);②针对所有的 VB 应用程序的病毒文件都利用 VB 病毒特征码自动提取算法提取特征码,保存在病毒特征码库中;③用病毒特征码分别扫描 1 000 个 VB 应用程序文件(其中 200 个病毒文件),10 000 个 VB 应用程序文件(其中 500 个病毒文件),100 000 个 VB 应用程序文件(其中 1 500 个病毒文件),200 000 个 VB 应用程序文件(其中 13 500 病毒文件);④统计误报率和漏报率。结果如表 4 所示。

表 4 VB 病毒特征码自动提取算法实验 2

VB 文件个数	1 000	10 000	100 000	200 000
病毒个数	200	500	1 500	13 500
检测到病毒数	200	500	1 500	13 500
误报文件数	0	2	4	32
漏报的文件数	0	0	0	0
误报率	0.00%	0.40%	0.26%	0.24%

从表 4 可以看出,该算法导致的病毒漏报率为 0,误报率要小于 0.5%,基本满足我们的需要。从而证明出该算法的准确性。对于误报的病毒,由于量比较小,可以采用人工分析的方法加以解决。

5 结束语

本文提出和设计了一种反病毒引擎的设计方案,并提出了针对 VB 应用程序的自动提取算法。在该反病毒引擎的设计方案中,采用 3 种特征码格式,并通过实验对这 3 种格式进行了性能上的比较。VB 编译程序的自动提取算法的提出主要是为了缓解病毒特征码提取人员的压力,提高工作效率。该算法充分理解 VB 应用程序文件结构的基础,利用全面、详细的实验来证明该算法的有效性和准确性。该算法导致的病毒漏报率为 0%,误报率要小于 0.5%,基本达到预期目标。

参考文献:

[1] 韩筱卿,王建锋,钟玮.计算机病毒分析与防范大全[M].北京:电子工业出版社,2006.

[2] 罗云彬.Windows 环境下 32 位汇编语言程序设计[M].2 版.北京:电子工业出版社,2006.

[3] Neal Hindocha,Eric Chien.Malicious threats and vulnerabilities in instant messaging[R].Technical Report,Symantec,2003.

[4] Tessa Lau,Oren Etzioni,Daniel S Weld.Privacy interfaces for information management [J]. Communications of the ACM, 1999,42(10):89-94.

[5] Roger Clarke.Internet privacy concerns confirm the case for intervention[J].Communications of the ACM,1999,42(2):60-67.

[6] 叶翔.主机安全防护系统研究与实现[D].武汉:华中科技大学,2004.

[7] Schleimer S, Wilkerson D, Aiken A. Winnowing: Local algorithms for document fingerprinting[C].Proceedings of the ACM SIGMOD International Conference on Management of Data, 2003:76-85.

[8] 张波云,殷建平,蒿敬波,等.基于多重朴素贝叶斯算法的未知病毒检测[J].计算机工程,2006,32(10):18-21.

[9] 禽敬波.殷建平,张波云.基于通用图灵机模型的病毒判定性定理证明[J].计算机科学,2005,32(8):243-245.

(上接第 5846 页)

4 结束语

入侵防御系统是网络安全技术发展到一定阶段的必然产物,它吸取、融合了防火墙和入侵检测技术,目的是为网络提供深层次的、有效的安全防护,IPS 的产生和发展也反应了安全产品的融合趋势。本文在分析了 IPS 的概念、特征、模型的基础上,提出了一种基于千兆网络数据控制卡的嵌入式 IPS 实现,在降低成本的同时保证了系统的性能,对于校园数据中心等中等规模的网络,有相当的实用价值。

参考文献:

[1] 李镇江,戴英侠,陈越. IDS 入侵检测系统研究[J].计算机工程,2001,27(4):12-14.

[2] CSI/FBI. Computer crime and security survey 2003 [EB/OL]. http://www.gocsi.com/forms/fbi/csi_fbi_survey.jhtml,2006.

[3] Neil Desai.Intrusion prevention systems: The next step in the evolution of IDS[EB/OL]. http://www.securityfocus.com/infocus/1670,2003.

[4] 李成华,周培源,张新访.基于主机内核的混合型入侵防御系统的设计与实现技术[J].计算机应用与软件,2006,23(7):119-122.

[5] Wickham T. Intrusion detection is dead. Long live intrusion prevention![EB/OL].http://www.sans.org/rr/papers/30/1028.pdf,2004.

[6] 卿昊,袁宏春.入侵防御系统(IPS)的技术研究及其实现[J].通信技术,2003(6):101-103.

[7] McAfee® IntruShield®.Network associates technology inc[EB/OL]. http://www.mcafee.com/us/enterprise/products/network_intrusion_prevention/index.html,2006.

[8] Intrusion prevention systems (IPS): Next generation firewalls [EB/OL].http://www.toplayer.com/content/cm/whitePaper_Registration.jsp,2007.