·通信保密·

一种基于代码特征的网页木马改良模型研究

胡　明，刘嘉勇，刘　亮
（四川大学 信息安全研究所，四川 成都 610065）

【摘　要】鉴于传统的检测方案无法准确地检测复杂多变的网页木马行为，简要介绍了网页木马的原理，结合了传统检测方案的优点，总结了检测网页木马的分析流程，归纳了网页木马的典型特征，以空间向量模型为理论依据，提出了一种改进的基于代码特征的网页木马检测模型，并依据此网页木马模型对网页木马样本进行分析与总结，为有效的检测网页木马提供理论依据、为防止其恶意行为提供新思路。

【关键词】网页木马；代码特征；检测模型
【中图分类号】TP393.08　　【文献标识码】A　　【文章编号】1002-0802(2010)08-0155-03

An Improved Web Trojan Model Based on Code Characteristics

HU Ming，LIU Jia-yong，LIU Liang
(Institute of Information Security, Sichuan University, Chengdu Sichuan 610065, China)

【Abstract】For the traditional detection could not deal with complicated and variable web trojan behaviors effectively, the article briefly describes the principle of web trojan, and in combination with the advantages of traditional scheme, summarizes the typical characteristics of the Web Trojan, gives the procedure of how to detect its existence. With space vector model as the theoretical basis, this article puts forwards a detection model of web Trojan based on code characteristics, with which the samples of Web Trojan are analyzed, thus providing a new idea for effective detection of web Trojan and prevention of its malicious behaviors.

【Key words】web Trojan; code characteristic; detection model

0 引言

互联网的飞速发展与广泛普及，给人们生活带来极大的便利的同时也带来了许多潜在的网络安全威胁。而在这些安全威胁中，网页木马是危害面最广泛、传播效果最佳的病毒形式之一。所谓的网页木马就是利用网站，浏览器和第三方软件存在安全薄弱点和漏洞，在网站上植入一些恶意代码，利用漏洞绕过杀毒软件，执行木马程序，从而达到破坏、窃取计算机信息的目的。如何有效的检测和分析网页木马的特征是防范其危害的关键，传统的网页木马检测方法包括基于行为特征的病毒库引擎检测[1]和基于代码特征的远程网站网页木马检测[2]。文献[1]以网页木马的行为作为研究重点，建立病毒库检测引擎，结合病毒特征码数据库，来实现对网页木马的检测，侧重于分析网页木马的内部行为特征，而文献[2]则侧重于检测以代码特征为特点的网页木马的外部行为。

收稿日期：2009-11-26。
作者简介：胡　明（1986-），男，硕士研究生，主要研究方向为网络系统与信息安全；刘嘉勇（1962-），男，博士，教授，主要研究方向为网络系统与信息安全；刘　亮（1982-），男，硕士，主要研究方向为网络系统与信息安全。

现结合上述两者的特点，综合分析网页木马的特征，并提出一种基于特征向量的网页木马静态分析模型，能更准确的检测网页木马。分析流程如图 1 所示。

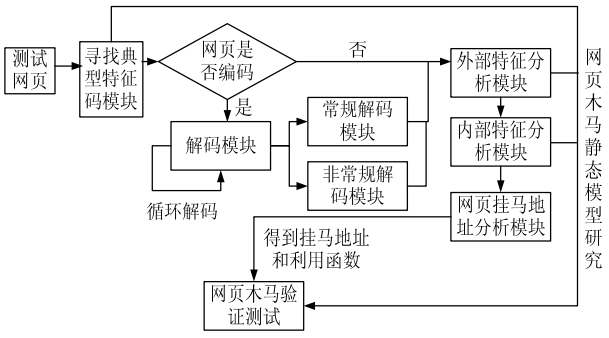

图 1 网页木马分析流程

①首先将测试网页放入寻找典型特征码模块进行处理，获取初步典型特征码；

②判断网页是否编码，如果未编码则直接进入外部特征分析模块，否则就进入解码模块分析。通过常规解码模块和非常规解码模块两种方式进行分析解码；

③接着通过外部和内部特征分析模块对未编码或解码后的网页进行分析，提取与其相符合的特征集合；

④再利用网页木马静态模型进行分析，即将典型特征模块、外部特征模块和内部特征模块总结出来的特征集合利用特征向量的方法进行判别，检测其是否为网页木马；

⑤然后进入网页挂马地址分析模块，判断是否存在网页木马特征，接着分析挂马地址和利用函数；

⑥最后与网页木马静态模型的测试结果进行对比，验证网页木马静态模型的准确性。

1 网页木马静态分析原理及其模型

1.1 静态分析原理

所谓的静态分析就是从挂马网页中使用正则表达式[3-4]分离出 html 内容和脚本内容，剥离出一些特定特征的 JavaScript 脚本，通过对某些网页木马所使用的关键函数进行解码处理，就容易取出像类标识符(CLSID)对象和 shellcode 等关键内容。现通过对大量的网页木马样本进行实验研究，对网页木马使用的关键函数进行归纳总结并组成一个函数词库，为了更全面更准确的检测网页木马行为，提出一种基于代码特征的网页木马静态分析模型，并将这些关键函数作为提出模型的一组关键词条。这些关键的特征属性函数及其意义如下：

①document.write/document.writeln()函数是将字符串转换成 Html 代码，函数外面必须含有 Html 标签，脚本才能够执行，否则将会被当做字符串输出在网页上；

②eval()函数是检查 Jscript 代码并执行，可以解析转义字符；

③unescape()函数解码所有使用%xx 十六进制形式编码的字符，并用 ASCII 字符集中等价的字符代替；

④fromCharCode 方法实现转换字符串的 unicode 字符串值的序列，返回带有恶意脚本的字符串；

⑤setTimeOut()方法通过设定延时值来执行某个特定的函数。如：setTimeOut('function()'，0)则表示立即执行函数 function()；

⑥字符串拆分再组合即将某些函数或者关键代码拆分成字符串，并用"+"号连接，对于网页木马中出现的 CLSID，通常都采用拆分再组合的方式绕过部分软件的扫描。

1.1.2 网页木马静态解析模型的研究

利用向量空间模型[5]对所研究的未知网页进行量化分析。

首先将网页木马的特征属性函数设置为关键词，即将 document.Write、eval、unescape、fromCharCode、setTimeOut 等这类函数设置为关键词，为了以后扩充方便可以把这类关键词总结成一词库，把这类关键词组成一组词条 $T(T_1, T_2, \cdots T_n)$，并依据词条 T_i 在网页木马中的危险度赋予一个危险度值 K_i，组成一组与词条 T 相对应的另一组词条 $K(K_1, K_2, \cdots K_n)$；另外网络字符编码也是判别网页木马的主要参考依据，如果存在过长的编码字符串，则说明有可能是经过编码后或者加密后的 shellcode 及其相关脚本。于是定义了阈值 A_0 和超过阈值 A_0 的过长字符串，A 表示这些字符串包括以"%xx"，"%uxxxx"，"\x00"，"&xx"，"&#x00"编码开头的字符串及不包含转义字符的原始字符串等，把这些过长字符串组成一组词条即：$A(A_1, A_2\cdots\cdots A_m)$。下面使用词频分析法(TF 法) 对关键词的权重进行计算，所得权重值如式（1）所示：

$$W_i = \frac{ft_i}{\sqrt{\sum_{k=1}^{n} ft_k}} 。 \quad (1)$$

由上述总结的网页木马特征规律，得出网页木马可疑度公式 S_{xi}，综合反映了网页木马的内部和外部特征。如式（2）所示：

$$S_{xi} = Sus(i,j) = \sum_{i=1}^{n} ft_i W_i K_i F_i + \sum_{j=1}^{m} A_j L_j 。 \quad (2)$$

参数说明：ft_i 表示 T_i 在网页木马样本出现的次数，W_i 表示特征值 T_i 的权重值，K_i 表示每个词条的危险程度值， F_i、L_j 表示是否存在关键字，F_i、L_j=1 表示存在，F_i、L_j=0 时表示不存在，A_j 表示附加参数的危险程度值，S_{xi} 表示对样本 x 第 i 个特征词的木马可疑度。j 为固定值，对于不同的特征词条，均添加相同的附加参数 A_j。当 i=1 时，S_{x1} 如式（3）所示：

$$S_{x1} = \sum_{i=1}^{n} ft_1 W_1 K_1 F_1 + \sum_{j=1}^{m} A_j L_j 。 \quad (3)$$

最后将(T_i，S_{Xi})映射成为向量空间中的一个点，即把(T_i，S_{Xi})作为一组特征向量。

要判别一个网页木马是否可疑，首先求解出被测试网页木马样本可疑度，以网页木马参考模型向量作为基础向量，然后求解出两者向量之间的夹角余弦，通过这个夹角余弦来衡量两个样本的相似度，最后通过设置相似阈值来判别是否进行下一步处理。

相似度 $R_{x,y}$[6]如式（4）所示：

$$R_{x,y} = \cos(S_{xi}, S_{yi}) = \frac{\sum_{i=1}^{n} S_{xi} \times S_{yi}}{\sqrt{(\sum_{i=1}^{n} S_{xi}^2)(\sum_{i=1}^{n} S_{yi}^2)}} 。 \quad (4)$$

2.2 网页木马解析步骤

2.2.1 寻找典型特征码模块

查找目标样本中的典型特征函数和判断编码方式。网页木马通常不会直接给出溢出利用函数和 shellcode，而是通过网页字符转换或者加密来绕过杀毒软件的查杀。从目标样本中查找典型特征函数有助于判断该样本的编码方式，为网页解码提供解码依据。

2.2.2 解码模块

解码模块包括成常规解码模块和非常规解码模块，常规解码模块适用利用脚本提供的编码方法，字符转义和利用工具编码等编码方式的解码。非常规解码模块包括自写算法函数解码，带密钥的加密函数解码等，要解码这种类型的编码

则需要从典型特征的方面着手研究。对于常规解码只需要借助部分解码工具如 CAL9000 即可以实现，对于非常规的编码可以通过修改或者删除关键字来实现，移除函数关键词 unescape 或者将 document.write(unescape(s))替换成 alert('s')函数进而提取出关键代码。

2.2.3 外部特征提取模块

网页木马外部特征体现在未编码的和经过解码后的特征属性函数，网页木马间接利用这些特征函数来进行地址空间的分配和 shellcode 的执行。如果需要批量获取特征码，则需要建立正则表达式来匹配特征函数。

2.2.4 内部特征提取模块

网页木马都有其实现功能的内部特征，包含以下方面：

①漏洞利用系统的类标示符(CLSID)，它并不是网页木马的存在的必备条件，但经常出现在网页木马中，通常以拆分再组合的形式出现在网页木马中，其作用是在本地系统文件中寻找并且调用所对应的 ActiveX 组件；

②解析执行的一段 shellcode，利用缓冲区溢出的执行一段恶意的病毒程序；

③调用缓冲区溢出利用函数，通过这个函数去执行黑客精心设计的 shellcode 程序。

2.2.5 网页挂马地址分析模块

通过对利用漏洞溢出后跳转执行的 shellcode 进行分析。通常网页木马会给溢出后的跳转地址分配大量的空间，并且添加“%u9090”等的填充字符头，同时给 shellcode 添加自写密钥从而避免被“0x00”字符截断。通过对“%uxxxx”字符的逆向解析，可以得出溢出调用函数和挂马地址。如果存在大量地址空间的分配和过长编码字符串或者解析出挂马地址和函数，则判定为网页木马。

2.3 实验结果

利用提出的网页木马静态解析模型对某个网页木马样本进行分析，首先建立参考样本模型，将特征词条 *Ti* 出现频率均赋值为 1，表示每个特征属性函数均在网页木马中出现一次；设置 L_1=1，$L_{j\neq1}$=0，A_0=120，表示网页木马中包含一串超过阈值 A_0 的过长字符串；取 K_i=0.1，A_j=0.3,表示特征属性函数的权重值为 0.1，过长字符串的权重值为 0.3，其相对权重比为 1:3。通过对目标样本与参考样本计算得出的相似度值来判别是否为网页木马。通过对十个测试样本的统计，绘出样本余弦统计图，横坐标表示所测试的特征样本 x，纵坐标表示测试特征样本与参考网页木马样本的相似度 R_{xy}，如图 2 所示。

实验表明：从图 2 可以看出相似度分布均匀，求得临界阈值为 $\sqrt{\frac{1}{5}}$；当相似度大于等于 $\sqrt{\frac{2}{5}}$ 时，该测试样本为网页木马的概率为 90%；当相似度等于 $\sqrt{\frac{1}{5}}$ 时，无法确定是否为网页木马，需要进一步测试，通过上述网页挂马地址分析模块的验证测试，但如果存在过长字符串 A 和大量地址空间的分配，判定为网页木马；当相似度小于 $\sqrt{\frac{1}{5}}$，则判定为为正常网页。

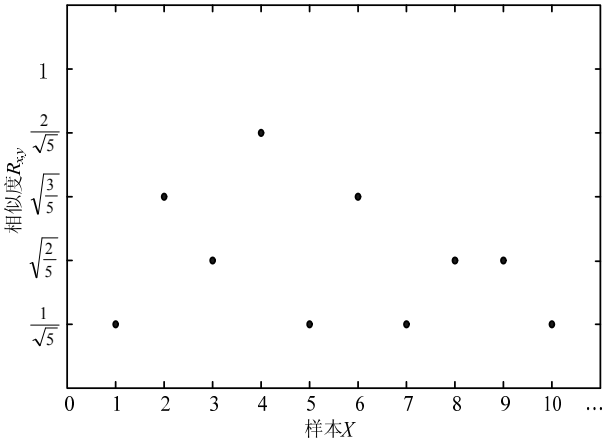

图 2 样本余弦统计

3 结语

网页木马的泛滥已经严重地威胁到信息系统的安全，影响了互联网正常有序的发展。为了更准确的检测和分析网页挂马的行为，防范网页木马的传播，从网页木马的基本原理入手，给出网页木马的流程图，并提出一种基于向量空间的网页木马静态解析模型，通过对测试样本的统计，以阈值区间来进行网页木马判别。实验表明：在阈值区间内的样本为网页木马概率相当大。在网页挂马地址分析中给出更详尽的木马行为特征，将是下一步研究的方向。

参考文献

[1] 王海峰，段友祥，刘仁宁. 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(24):109-110.

[2] 吴润浦，方勇，吴少华. 基于统计与代码特征分析的网页木马检测模型[J]. 信息与电子工程,2009,7(01):71-75.

[3] 董敏,毕盛,齐得昱. 基于正则表达式的测试数据自动生成技术[J]. 计算机工程, 2009,35(16):29-31.

[4] 石倩，陈荣，鲁明羽. 基于规则归纳的信息抽取系统实现[J]. 计算机工程与应用,2008,44(21):166-168.

[5] 段丽娟. Web 挖掘的敏感信息过滤模型[J]. 信息安全与通信保密,2007(01):69-71.

[6] 刘滨，王世华. Intranet 非法站点及不良信息检测系统的设计与实现[J]. 信息安全与通信保密,2005(12):103-106.

一种基于代码特征的网页木马改良模型研究

作者： 胡明， 刘嘉勇， 刘亮， HU Ming， LIU Jia-yong， LIU Liang
作者单位： 四川大学,信息安全研究所,四川,成都,610066
刊名： 通信技术
英文刊名： COMMUNICATIONS TECHNOLOGY
年，卷(期)： 2010,43(8)
被引用次数： 1次

参考文献(6条)

1. 王海峰;段友祥;刘仁宁 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(24)

2. 吴润浦;方勇;吴少华 基于统计与代码特征分析的网页木马检测模型[期刊论文]-信息与电子工程 2009(01)

3. 董敏;毕盛;齐得昱 基于正则表达式的测试数据自动生成技术[期刊论文]-计算机工程 2009(16)

4. 石倩;陈荣;鲁明羽 基于规则归纳的信息抽取系统实现[期刊论文]-计算机工程与应用 2008(21)

5. 段丽娟 Web挖掘的敏感信息过滤模型[期刊论文]-信息安全与通信保密 2007(01)

6. 刘滨;王世华 Intranet非法站点及不良信息检测系统的设计与实现[期刊论文]-信息安全与通信保密 2005(12)

本文读者也读过(10条)

1. 吴润浦.方勇.吴少华.WU Run-pu.FANG Yong.WU Shao-hua 基于统计与代码特征分析的网页木马检测模型[期刊论文]-信息与电子工程 2009,7(1)

2. 方刚.FANG Gang 对网站被挂马的分析与防范[期刊论文]-实验室研究与探索2010,29(7)

3. 梁玲.Liang Ling 基于变形方法的网页木马免杀技术研究[期刊论文]-电脑开发与应用2010,23(4)

4. 罗川.辛茗庭.凌志祥.Luo Chuan.Xin Ming-ting.Ling Zhi-xiang 网页木马剖析与实现[期刊论文]-计算机安全2007(12)

5. 葛先军.李志勇.宋巍巍 基于网页恶意脚本链接分析的木马检测技术[会议论文]-2008

6. 江璜.JIANG Huang 网页木马与跨域漏洞[期刊论文]-电脑知识与技术（学术交流） 2006(1)

7. 杜振华.张健.马勇.张鑫.苏圣魁 一种恶意网页检测系统的研究与设计[会议论文]-2008

8. 韩法旺 Web网页木马研究初探[期刊论文]-科技信息2008(19)

9. 李新.徐阳.李成友 支招防御网页木马[期刊论文]-中国教育网络2010(4)

10. 张慧琳.诸葛建伟.宋程昱.韩心慧.邹维 基于网页动态视图的网页木马检测方法研究[会议论文]-2009

引证文献(1条)

1. 孙飞帆.施勇.薛质 基于权重分析的网页木马检测模型[期刊论文]-信息安全与通信保密 2012(12)