基于统计学习的挂马网页实时检测

王 涛[1] 余顺争[2]
(广东工业大学自动化学院 广州 510006)[1]
(中山大学信息科学与技术学院电子与通信工程系 广州 510006)[2]

摘 要 近年来挂马网页对Web安全造成严重威胁,客户端的主要防御手段包括反病毒软件与恶意站点黑名单。反病毒软件采用特征码匹配方法,无法有效检测经过加密与混淆变形的网页脚本代码;黑名单无法防御最新出现的恶意站点。提出一种新型的、与网页内容代码无关的挂马网页实时检测方法。该方法主要提取访问网页时HTTP会话过程的各种统计特征,利用决策树机器学习方法构建挂马网页分类模型并用于在线实时检测。实验证明,该方法能够达到89.7%的挂马网页检测率与0.3%的误检率。

关键词 挂马网页,HTTP会话,决策树,机器学习

Real-time Detection of Malicious Web Pages Based on Statistical Learning

WANG Tao[1] YU Shun-zheng[2]
(Faculty of Automation, Guangdong University of Technology, Guangzhou 510006, China)[1]
(Department of Electronics and Communication Engineering, Sun Yat-Sen University, Guangzhou 510006, China)[2]

Abstract Malicious Web pages impose increasing threats on Web security in recent years. Currently, there are mainly two client-side protection approaches including anti-virus software packages and blacklists of malicious sites. Anti-virus techniques commonly use signature-based approaches which might not be able to efficiently identify malicious HTML codes with encryption and obfuscation. Furthermore, blacklisting techniques are difficult to keep up-to-date. This paper presented a novel classification method for real-time detecting malicious Web pages which is independent with the contents of Web pages. Our approach characterizes malicious Web pages using HTTP session information. With representative statistical features and decision tree algorithm in machine learning, we built an effective classification model for online real-time detecting malicious Web pages. Experiment results demonstrate that we are able to successfully detect 89.7% of the malicious Web pages with a low false positive rate of 0.3%.

Keywords Malicious Web pages, HTTP session, Decision tree, Machine learning

1 引言

丰富的Web服务在为信息共享带来便利的同时,也成为攻击者入侵用户系统的主要平台。挂马网页指被攻击者植入了恶意HTML脚本代码的网页,主要利用浏览器与Web应用程序漏洞把各种恶意程序传播到用户系统。一旦用户浏览挂马网页,浏览器就会加载运行恶意脚本代码并自动下载执行恶意程序。通过安装的恶意程序,攻击者可以控制用户主机,盗取用户的隐私信息,帮助某些流氓厂商提高安装量或点击率,对某个网站服务器发动DDoS攻击等等。据瑞星公司2009年抽样统计[1]显示,每天约有30%的网民上网时会遇到挂马网站。

攻击者利用挂马网页传播恶意程序的行为称为网页挂马攻击。目前,网页挂马攻击已取代传统的扫描攻击方式,并成为传播病毒木马的主要手段[2]。病毒蠕虫主要通过大量扫描发现有系统漏洞的主机(如某个开放的网络服务端口),并通过推送模式把恶意程序传播到漏洞主机,但此方式不能穿越NAT以及网络边界防火墙。网页挂马攻击采用取回模式的感染方式,在用户浏览被俘获网站时自动将恶意程序植入到用户系统,整个过程在后台进行并且用户无法察觉。因此,一旦攻击者俘获具有较大访问量的正常网站并用于实施挂马攻击,将会造成大面积感染。图1是一个典型的网页挂马攻击交互过程。其中,恶意程序分发站点(malware distribution site)是提供恶意木马病毒下载的站点。通常,攻击者为逃避跟踪监测,会利用多次重定向链接将Web用户引导至恶意程序分发站点,自动下载恶意程序到本地并执行。

现阶段检测挂马网页的主要方法包括网页恶意代码特征匹配与基于高交互虚拟蜜罐系统的动态行为监测。网页恶意代码特征匹配[3,4]是将恶意脚本代码视为脚本病毒,通过检查脚本代码是否与已知特征码匹配进行判定。此方法具有固有缺陷:需要将加密脚本解释成为明文脚本再来检测,但目前正常网页为保护知识产权也普遍使用加密技术;浏览器插件

到稿日期:2010-02-05 返修日期:2010-05-07 本文受国家高技术研究发展计划(863计划)专题课题(2007AA01Z449),国家自然科学基金-广东联合基金重点项目(U0735002),国家自然科学基金面上项目(60970146),教育部博士点专项基金(20090171120001)资助。

王 涛(1983—),男,博士生,主要研究方向为计算机网络安全,E-mail:wangtaosea@msn.com;余顺争(1958—),男,教授,博士生导师,主要研究方向为网络安全、网络行为分析、网络测量等。

普及带来的漏洞导致零日攻击增多，使得基于代码特征匹配的检测方法失效；多重定向、混淆变形技术都可以用于躲避基于代码特征的检测。基于高交互虚拟蜜罐系统检测挂马网页主要是对访问网页后的系统行为与状态进行监测，因为恶意程序下载执行后通常会有可疑的系统调用或修改系统进程以及注册表等等。虚拟蜜罐系统有两种部署方式：第一，直接部署在客户端[5]作为浏览器代理，将用户请求的网页先载入虚拟系统并监控系统行为与状态，将检测出的恶意网页拦截。此方法虽然不会造成误报并能发现零日攻击，但需要耗费一定的系统资源并带来浏览延迟。第二，大型机构部署虚拟蜜罐网对互联网海量网页进行主动扫描检测[6,7]，将发现的挂马网页列入黑名单并以插件形式提供给用户端浏览器，显然，由于互联网网页数量过大，基于蜜罐技术扫描检测挂马网页存在时间差，因此导致黑名单无法及时防御新出现的挂马网页或恶意站点。

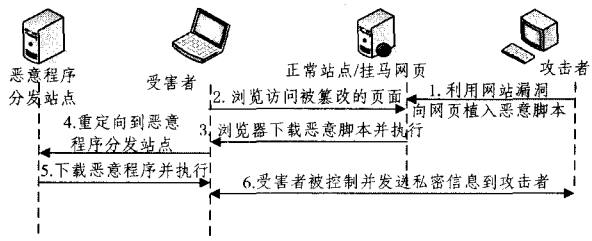


图 1　典型的网页挂马攻击过程

为此，本文提出一种新型的、与网页内容代码无关的挂马网页客户端实时检测方法。该方法主要提取访问网页时HTTP会话过程的各种统计特征，利用有监督机器学习方法构建挂马网页分类模型，并基于网页分类模型在线实时检测挂马网页。

2　挂马网页检测模型

本文中，一个网页会话（即网页的 HTTP 请求-响应会话过程）是指浏览器访问某个网页，执行此页面内脚本代码，然后发出对网页内嵌的广告、墙纸图片、装饰条块、背景音乐等内嵌对象的请求。在请求内嵌对象时，可能依多重链接访问多个其它站点。最后，浏览器将页面呈现出来。

图 2 给出了一个典型的挂马网页会话实例。用户在访问初始页面（www. 5axs. com）后，经过 7 次重定向被诱导至恶意程序分发站点，自动下载并执行恶意程序。在此实例中，www. 22na. com，cxcce332. 7766. org 为外部媒介站点，rewdsds333. 8866. org 为最终的恶意程序分发站点。通常，外部媒介站点是由攻击者控制的服务器，或者是被攻击者俘获的正常网站（一般是在正常网站的网页或脚本文件中植入重定向代码）。

```
Level 0: http:// www.5axs.com
Level 1: <script> http:// www.5axs.com /gg/baidu.js
Level 2: <script> http://www.22na.com/templets/img/toppic .jpg
Level 3: <iframe> http://cxcce332.7766.org/03/03.htm?2
Level 4: <iframe> http://cxcce332.7766.org/ganiniang360.html
Level 5: <iframe> http://cxcce332.7766.org/go2.jpg
Level 6: <script> http://cxcce332.7766.org/03/logo.gif
Level 7: <exe> http://rewdsds333.8866.org/w/x3.exe
```

图 2　挂马网页多次重定向的实例

基于以上分析，本文提出的挂马网页检测模型结构如图 3 所示。首先，利用用户浏览时间间隔分离各个网页会话，并根据各个请求包头部域中 Referer 的信息对会话进行重组。其次，根据已采集的数据集建立白名单（可信服务器域名列表）与黑名单（恶意站点域名列表）。统计每个网页 HTTP 会话过程的各种特征，特征向量被送到机器学习分类检测模块并判断是否可疑，检测并确定后的恶意站点域名添加到黑名单中。

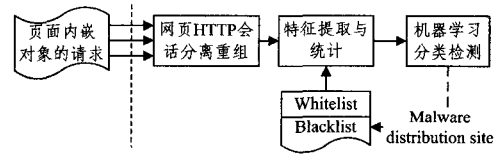


图 3　挂马网页检测模型结构图

3　挂马网页特征分析

为了分析挂马网页的特征，本文采集了访问正常网页和挂马网页的 HTTP 请求响应数据集。

正常网页（WebClean）会话数据的采集：本文选取 Alexa[8] 提供的网站访问流量排名列表的前 50,000 个网站主页作为正常网页集。这些网站安全措施比较严密，可认为这些网站主页基本上都是可信的，而且能够代表大多数页面的特征。利用 IE 自动地按顺序访问这些页面，每个页面请求等待足够时间间隔，同时截获访问各个网页的 HTTP 流量。由于有些页面无法访问，我们共采集正常网页会话实例为 41,800 个。

挂马网页（WebMalware）会话数据的采集：利用高交互虚拟蜜罐系统采集[9]。截至目前共采集了 816 个挂马网页会话实例。表 1 是所采集数据的概要说明，包括实例数量以及采集日期。

表 1　数据集

Corpus abbreviation	Number of Instances	Crawl Date
WebClean	41,800	Sept. 2009
WebMalware	816	Sept. -Oct. 2009

3.1　站点 IP 地址分布

正常站点与恶意程序分发站点 IP 地址前缀的累计概率分布如图 4 所示。可见，恶意程序分发站点的 IP 地址集中分布于几个区域，并与正常站点的分布有较大差异。如在 IP 段 90. * —96. *，约有 19.8%的恶意站点 IP 在此区间，而正常站点只有约 5%；约 7.5%的恶意站点 IP 在 120. * —125. * 区间内，而正常站点约只有 2.2%；约 10.3%的恶意站点 IP 在 192. * —195. * 区间内，而正常站点却只有 5.5%。总体来说，正常站点的 IP 地址分布比较分散，而恶意站点的 IP 分布较为集中。因此，我们将网页会话过程中 IP 地址在上述几个可疑区间的站点数量作为一个特征。

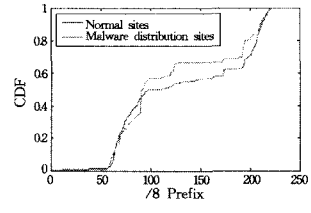
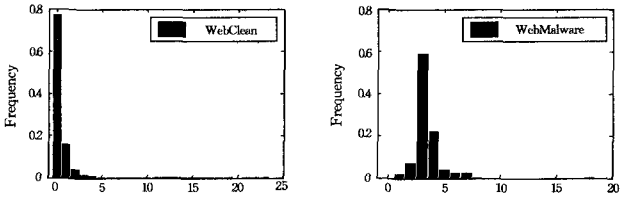


图 4　恶意程序分发站点 IP 地址前缀累计概率分布

3.2　可疑外部域名数量

正常网页所引用的外部对象大多由知名站点提供，我们对正常网页 HTTP 会话过程中引用的所有外部链接所在网

站的域名做统计，将常被引用的外部域名看作是可信的网站并列入白名单，白名单之外的网站称为可疑网站。经过统计，少部分外部域名被频繁引用，如 www. google-analytics. com 在本文采集的数据中被约 6 万个网页引用。同时，恶意程序分发站点一般都是由黑客直接管理并不对外提供正常的 Web 服务，因此不会在白名单内。即使攻击者利用白名单中的正常网站作为中间媒介站点来实现挂马，但用户最终也需要被连接到恶意程序分发服务器才能下载恶意程序。根据对挂马网页会话过程的观测可以发现，多数包含不止一个可疑外部域名。图 5 是各个网页会话中可疑外部域名数量的统计。对正常网页集(见图 5(a))，约 76. 8%的网页不会引用可疑的外部域；对挂马网页集(见图 5(b))，网页通常引用 2～5 个可疑的外部域名，甚至更多。在实际应用中，使用动态白名单，实时增添一些知名度高且可信的域名，由此筛选出一些可疑的域名。所以，可以把网页 HTTP 会话过程中引用外部可疑域名的数量作为检测挂马网页的一个特征。

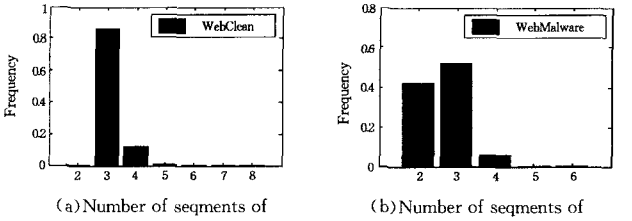


(a)Number of suspicious external domains　(b)Number of suspicious external domains

图 5　网页引用可疑外部域名的数量分布

3.3　域名段数

基于分隔符“.”，可称全域名 www. sohu. com 的段数为 3 段，二级域名 sohu. com 为 2 段。图 6 统计了两类数据集引用的所有外部域名段数：对正常网页集(见图 6(a))，网页所引用的外部域名一般为三段或更多(多级子域名)，只有约0. 2%的二段外部域名；对挂马网页集(见图 6(b))，两段的外部域名约占 41. 9%。可见，正常网站一般使用多个子域名来区分不同的服务器并对外提供服务，如 ad. doubleclick. net，g. doubleclick. net，因此其域名段数基本都在 3 段以上；而恶意站点一般直接使用注册的二级域名对外提供服务，如 ccndk822. cn，ewrewr34. cn。因此，将网页是否有引用过二段的外部域名作为一个特征。



(a)Number of seqments of external domains　(b)Number of seqments of external domains

图 6　网页所引用外部域名的段数分布

3.4　不完整头部信息域的请求数量

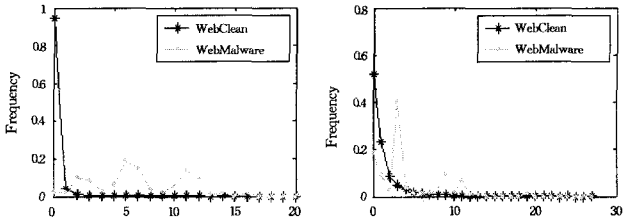
大部分正常的 HTTP 请求头默认情况下会有以下几个信息域：Accept，Accept-Language，Accept-Encoding，User-Agent，Referer，Host，Connection。一些挂马网页会话过程中会出现缺少多个常见的头部信息域的请求。图 7 是访问挂马网页后自动发送到恶意程序分发站点的请求头部信息，缺少 Accept-Language，Accept-Encoding 与 Referer 3 个信息域。我们将网页会话中缺少这 3 个信息域的请求数量作为特征。数据集中约 55. 3%的挂马网页会话包含此类具有不完整头部信息域的请求，而只有 1. 24%的正常网页会话出现过此类请求。

```
GET /ooo/147.exe HTTP/1.0
Accept: */*
User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)
Host: 121.12.170.84
Connection: Keep-Alive
```

图 7　不完整头部信息域的请求实例

3.5　各种常见类型文件的请求数量

我们将在网页会话过程中到可疑外部域(白名单外)的各种常见类型文件的请求数量作为特征值。图 8 是 html，js 两种类型文件的请求数量的分布情况：约 94. 3%的正常网页没有引用可疑外部域的 html 文件，而约 96%的挂马网页引用多于 2 个的 html 文件；约 52. 2%的正常网页没有引用可疑外部域的 js 文件，而挂马网页中只有 20%。同时，正常页面所引用可疑外部域的 html，js 文件一般不多于 3 个。因此，如果一个网页会话过程中出现过多到可疑外部域的 html，js 文件的请求，则此网页可能是挂马网页。



(a)Number of requested html files　(b)Number of requested js files

图 8　不同类型文件的请求数量分布

3.6　重定向层数

对于一个网页会话，利用各个请求包头部域中 Referer 的信息对会话进行重组，构建一个链接树。此链接树以用户请求的原始页面(landing Webpage)作为根节点，每个请求以其 Referer 域内的对象作为父节点。图 9 是一个网页 HTTP 会话过程的重定向链接树，其中 M，N 是不同的外部站点。由根节点开始，最长的链接路径(包含外部站点)长度称为页面重定向链接层数(page redirection steps)。

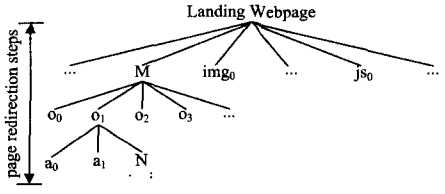


图 9　网页会话的重定向链接树

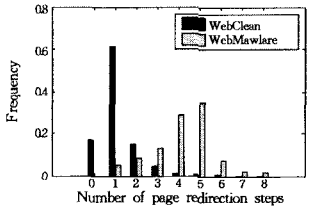


图 10　网页会话重定向链接层数分布

图 10 统计比较了正常网页集与挂马网页集中所有实例的重定向链接层数。对正常网页，约 16. 8%的网页不会引用外部对象，约 61. 6%的网页是直接引用外部对象，只有约 6. 6%的网页会话包含 3 次以上重定向链接；对挂马网页，约

87.1%的网页会话包含3次以上重定向链接，最终将用户引导到恶意程序分发站点。所以，由统计结果可知，网页编辑人员引用外部对象时，基本上都会直接引用，其对象的请求大多由初始页面直接产生，不会经过多层的链接才访问到远程对象；而攻击者经常利用多次重定向来躲避检测。

3.7 典型的可疑特征

一些挂马页面HTTP会话过程具有典型的可疑特征，如所引用的外部域名采用了代码混淆方法，或URL中含有重定向地址，或直接采用IP地址，或使用一些特殊的服务端口，这在正常的网页会话过程中很少出现，实例如图11所示。

```
Type1: http://6b%6b%36%2e%75%73/1.js
Type2: http://okm4.org/\"http://23weg.bnr56.cn/aa/a3a.htm\"
Tyep3: http://121.12.170.223:9090/a/3.css
Type4: http://wesdrt.cn:338/a0036159/a03.htm
```

图11 4种典型的可疑特征

3.8 特征表

经过汇总，共提取特征24个，如表2所列。

表2 特征集

Feature	Count
Number of suspicious external IPs	1
Number of suspicious external domains	1
Whether external domains with 2 segments exist	1
Number of requests with 13 common types to suspicious external sites	13
Number of page redirection steps	1
Number of different Sever headers	1
Number of different User-Agent headers	1
Number of requests with incomplete headers	1
Typical suspicious features	4
All	24

4 分类模型训练

4.1 C4.5决策树

本文采用C4.5决策树方法[10]训练挂马网页分类模型。决策树是用于分类和预测的一种树结构，是以实例为基础的归纳学习算法。它着眼于从一组无次序、无规则的实例中推理出决策树表示形式的分类规则。利用决策树方法处理分类问题分为两个步骤：第一步利用训练集建立并精化一棵决策树，建立决策树模型；第二步利用生成完毕的决策树对输入样本进行分类。对输入的待测样本，从根节点依次测试待测样本的特征值，直到到达某个叶节点，从而确定该待测样本所在的类。

设训练集 $S=\{X_1,X_2,\cdots,X_N\}$，其中包含 M 个不同的类 $w_i(i=1,2,\cdots,M)$。设 N_i 是属于类 w_i 的样本的个数。由此可以得到训练集 S 对分类的平均信息量

$$I(N_1,N_1,\cdots,N_M)=-\sum_{i=1}^{M}p_i\log_2(p_i) \tag{1}$$

式中，$p_i=N_i/N$ 是样本属于类 w_i 的概率。设每个样本可由包含 d 个特征的特征向量$(A_1,A_2,\cdots,A_d)$表示。对任一离散特征 $A_i(1\leqslant i\leqslant d)$，假设 A_i 存在 k 个不同取值$\{a_1,a_2,\cdots,a_j,\cdots,a_k\}$，那么根据 A_i 的取值，可以将训练集 S 划分为 k 个子集 $S_1,S_2,\cdots,S_k$，其中 $S_j=\{X\mid X\in S,S.A=a_j\}$。如果选 A_i 为测试属性，那么这些子集表示从代表集合 S 出发的所有树枝。设 N_{ij} 表示 S_j 中类为 w_i 的样本的个数。由特征 A_i 进一步划分训练集后，训练集 S 对分类的平均信息量为

$$E(A_i)=\sum_{j=1}^{k}[(\frac{N_{1j}+N_{2j}+\cdots+N_{Mj}}{N})\cdot I(N_{1j},N_{2j},\cdots,N_{Mj})] \tag{2}$$

式中，对于给定子集 S_j，

$$I(N_{1j},N_{2j},\cdots,N_{Mj})=-\sum_{i=1}^{M}p_{ij}\log_2(p_{ij}) \tag{3}$$

式中，$p_{ij}=N_{ij}/|S_j|$表示 S_j 中的样本属于类 w_i 的概率；$|S_j|$表示 S_j 中的样本个数。因此在属性 A_i 上分支获得的信息增益表示为

$$Gain(A_i)=I(N_1,N_2,\cdots,N_M)-E(A_i) \tag{4}$$

$Gain(A_i)$指由于知道特征 A_i 的值而导致的平均信息量的减小，即分类不确定性的降低。因此，选择信息增益最大的特征创建决策树节点，根据特征的不同取值创建各个分支。再对各分支的子集递归调用该方法，建立决策树节点的分支，直到所有子集仅包含同一类别的数据为止。

对于非离散的特征，C4.5决策树算法采用离散化其取值空间的策略，将其转化成为离散特征进行计算。C4.5决策树方法处理分类问题有以下优势：C4.5决策树方法在模型构建和样本预测过程中都不依赖于样本的分布，因此该方法能够有效避免样本分布变化所带来的影响，具有良好的分类稳定性；C4.5决策树处理分类问题具有更高的效率。

我们将挂马网页作为正例子(positive class)，正常网页作为负例子(negative class)，并采用评价分类模型的4个主要指标：检测率(True Positive Rate)，即挂马网页被正确检测出来的比率；误检率(False Positive Rate)，即正常网页被误检为挂马网页的比率；精确率(Precision)，即被判为正例子的集合中真实挂马网页的比率；准确率(Accuracy)，即被正确检测出的样本占训练集所有样本的比率。

4.2 模型检测性能

在训练分类模型时，采用十折交叉验证来测试模型性能。训练数据集被随机地分为10份，轮流将其中9份做训练，1份做测试，10次结果的均值作为对算法性能的估计。模型的ROC曲线如图12所示。由于实际环境下正常网页数量要远远大于挂马网页，因此在保证一定检测率的条件下，模型误检率越低越好。C4.5决策树检测模型达到了较高的检测率(89.7%)以及低误检率(0.3%)，模型的精确率为85.7%，准确率为99.5%。

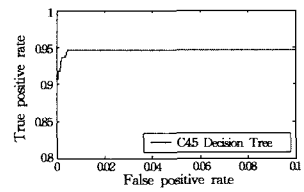


图12 C4.5分类模型ROC曲线

表3列出了C4.5决策树模型中信息增益排名前五的特征。可见，可疑外部域名的数量具有最大的特征信息增益，根据此特征可以过滤出很多正常网页，其次是重定向链接层数。

为了衡量检测模型的分类稳定性，我们改变训练集中正常网页样本与挂马网页样本的比例。表4是模型检测性能随样本分布的变化情况。可以看出，在训练集样本数量与分布不同的情况下，分类模型的性能基本保持稳定，其准确率保持上升。另外，随着正常样本数量的增加，分类模型的检测率有所降低，这是因为新增的一些模糊样本(与挂马网页特征相近)影响了模型的分类规则，但这类模糊样本数量较小，因此分类模型依然保持了较高的检测率与较低的误检率。

(下转第129页)

言的事实标准。本文在 XACML 基础上，通过引入时态约束来弥补 XACML 在描述异构策略组合时不能有效描述时态约束的不足。XACML 在描述策略组合时，没有考虑策略之间的安全属性，也无法描述策略组合后的安全属性，如何在 XACML 中引入策略的安全等级将是下一步需要解决的问题。

参 考 文 献

[1] 邓集波，洪帆. 基于任务的访问控制模型[J]. 软件学报，2003，14(1)：76-82

[2] Gong L，Qian X. Computational Issues in Secure Interoperation [J]. IEEE Transactions on Software Engineering，1996，22(1)：43-52

[3] Xacml T C. OASIS eXtensible Access Control Markup Language (XACML) [DB/OL]. http://www. oasis-open. org/committees/xacml/

[4] Hada S，Kudo M. XML access control language：Provisional authorization for XML documents [DB/OL]. http:// www. trl. ibm. com /projects /xml / xacl /xacl-spec. html

[5] Ashley P，Hada S，Karjoth G，et al. The enterprise privacy authorization language(EPAL) [DB/OL]. http://www. w3. org/2003/p3p-ws/pp/ibm3. html

[6] Ribeiro C，Z' lšquete A，Ferreira P，et al. SPL：An access control language for security policies with complex constraints [C]// NDSS '01：Network and Distributed System Security Symposium. 2001

[7] Bharadwaj V G，Baras J S. Towards automated negotiation of access control policies [C]// Proceedings of IEEE 4th International Workshop on Policies for Distributed Systems and Networks. Washington DC，USA：IEEE Computer Society Press，2003：111-119

[8] Wainer J，Kumar A，Barthelmess P. DW-RBAC：A Formal Security Model of Delegation and Revocation in Workflow Systems [J]. Information Systems，2007，22(3)：365-384

[9] James B D，Bertino E，Latif U，et al. A Generalized Temporal Role-Based Access Control Model [J]. IEEE Transaction on Knowledge and Data Engineering，2005：4-22

[10] 唐卓，赵林，李肯立，等. 一种基于风险的多域互操作动态访问控制模型[J]. 计算机研究与发展，2009，43(6)：948-955

[11] Li Ninghui，Wang Qihua，Qardaji W，et al. Access Control Policy Combining：Theory Meets Practice[C]// Proceedings of the 14th ACM symposium on Access control models and technologies. June 2009

[12] Cheng chen，Rohatgi P，Wagner G M，et al. Fuzzy Multi-Level Security：An Experiment on Quantified Risk-Adaptive Access Control[C]// IEEE Symposium on Security and Privacy. 2007：222-230

[13] 许峰，赖海光，等. 面向服务的角色访问控制技术研究[J]. 计算机学报，2005，28(4)：686-693

[14] 黄建，卿斯汉. 带时间特性的角色访问控制[J]. 软件学报，2003，14(11)：1944-1954

[15] Dewri R，Poolsappasit N，Ray P，et al. Optimal Security Hardening Using Multi-Objective Optimization on Attack Tree Models of Networks[C]// Proceedings of the 14th ACM Conference on Computer and Communications Security(CCS'07). New York，USA：ACM Press，2007：204-213

(上接第 90 页)

表 3　信息增益排名前五的特征

Rank	Feature
1	Number of suspicious external domains
2	Number of page redirection steps
3	Number of requests with incomplete headers
4	Whether external domains with 2 segments exist
5	Number of requested html files

表 4　不同样本分布下 C4.5 分类模型的性能

Malicious Webpage percentage	TP	FP	Precision	Accuracy
50%	96.6%	0.6%	99.4%	98%
20%	95.7%	0.4%	98.4%	98.8%
10%	92.2%	0.3%	97.2%	98.9%
All Sources	89.7%	0.3%	85.7%	99.5%

结束语　当前检测挂马网页的主要手段有网页代码特征匹配与高交互虚拟蜜罐技术。前者难以对抗代码加密与混淆变形技术，后者资源消耗较大，难以在客户端直接部署。针对这些不足，本文提出一种轻量级的、基于访问网页的 HTTP 会话统计特征的挂马网页检测方法，它无需对网页 HTML 代码、数据载荷进行特征匹配。基于低维特征与有监督的 C4.5 决策树学习，训练了能有效检测挂马网页的分类模型。实验证明，我们能达到 89.7%的检测率与 0.3%的误检率。下一步工作是进一步发掘更多挂马网页的特征，研究在线学习算法，以适应不断更新的挂马网页特征。

参 考 文 献

[1] 2009 年上半年中国大陆地区互联网安全报告[EB/OL]. See http://it. rising. com. cn/new2008/News/NewsInfo/2009-07-21/1248160663d53890. shtml

[2] Provos N，McNamee D，Mavrommatis P，et al. The ghost in the browser analysis of Web-based malware[C]// Proceedings of the First Workshop on Hot Topics in Understanding Botnets. Cambridge，MA，2007

[3] Hou Yung-Tsung，Chang Yimeng，Chen Tsuhan，et al. Malicious Web content detection by machine learning[J]. Expert Systems with Applications，2010，37(1)：55-60

[4] Seifert C，Komisarczuk P，Welch I. Identification of Malicious Web Pages with Static Heuristics[C]// IEEE Australasian Telecommunication Networks and Applications Conference. Adelaide，2008：91-96

[5] Moshchuk A，Bragin T，Deville D，et al. SpyProxy：Execution-based Detection of Malicious Web Content[C]// Proc. of the USENIX Security Symposium. Boston，MA，Aug. 2007：27-42

[6] Provos N，Mavrommatis P，Rajab M A，et al. All Your iFRAMEs Point to Us[C]// Proc. of the USENIX Security Symposium. San Jose，CA，July 2008：1-15

[7] Zhuge Jianwei，Thorsten H，Song Chengyu，et al. Studying Malicious Websites and the Underground Economy on the Chinese Web[C]// Proceedings of 2008 Workshop on the Economics of Information Security(WEIS'08). June 2008

[8] Top 1，000，000 Sites[EB/OL]. http://www. alexa. com/topsites，September 2009

[9] Seifert C，Steenson R. Capture-honeypot client [EB/OL]. https://www. client-honeynet. org/capture. html，2006

[10] Witten I H，Frank E. Data Mining：Practical Machine Learning Tools and Techniques (2nd ed) [M]. San Francisco：Elsevier Inc.，2005