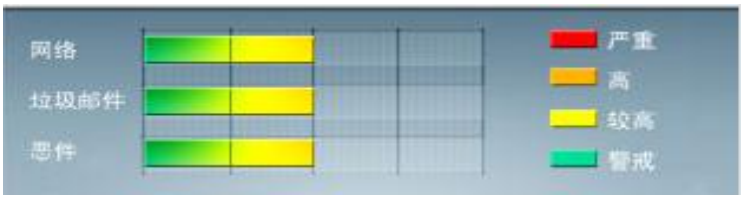
安全威胁每周警讯

2012/10/21～2012/10/27

本周威胁指数



***TrendMicro* 中国区网络安全监控中心**

前十大病毒警讯

排名	病毒名称	威胁类型	风险等级	趋势	病毒行为描述
1	X97M_OLEMAL.A	宏病毒	★★	→	宏病毒，主要通过用户浏览恶意网站感染，该病毒主要感染 Office Excel 文件，并且会在 Excel 中创建一个名为 StartUp 的宏脚本
2	TROJ_DOWNAD.INF	木马	★★★	→	DOWNAD 蠕虫关联木马
3	WORM_DOWNAD	蠕虫	★★★★	→	该病毒会攻击未安装微软 IE 漏洞 MS08-067 的电脑，并且会在受感染电脑产生五万个恶意程序网址并试图在同一时间内随机连结其中 500 个恶意网站下载病毒
4	WORM_DOWNAD.AD	蠕虫	★★★★	→	该病毒会攻击未安装微软 IE 漏洞 MS08-067 的电脑，并且会在受感染电脑产生五万个恶意程序网址并试图在同一时间内随机连结其中 500 个恶意网站下载病毒
5	TROJ_IFRAME.CP	木马	★★★	→	GIF、jpg 和 SWF 文件中被插入一个恶意的 iframe 标记时，趋势科技会将其判断为 TROJ_IFRAME.CP 病毒。当这些文件被执行时，会重定向到这些 URL，并下载恶意程序
6	CRCK_KEYGEN	破解程序	★★	↑	疑似木马病毒，通过访问恶意站点下载感染或由其他恶意程序下载感染
7	X97M_LAROUX.BK	宏病毒	★★	↓	宏病毒，主要通过用户浏览恶意网站感染，该病毒主要感染 Office Excel 文件，并且会在 Excel 中创建一个名为 StartUp 的宏脚本
8	TROJ_SPNR.15JC12	木马	★★★	↓	木马程序
9	WORM_ECODE.E-CN	蠕虫	★★★★	↓	它通过将受感染的移动驱动器与系统连接起来而入侵。 它可能是由远程站点的其他恶意软件/灰色软件/间谍软件下载而来。 它可能是用户在访问恶意网站时在无意中下载而来
10	X97M_LAROUX.CO	宏病毒	★★	↓	宏病毒，主要通过用户浏览恶意网站感染，该病毒主要感染 Office Excel 文件，并且会在 Excel 中创建一个名为 StartUp 的宏脚本

系统漏洞信息

MS12-060 ： Windows 常用控件中的漏洞可能允许远程执行代码 (2720573)

Microsoft Office 2003
Microsoft Office 2007
Microsoft Office 2010
Microsoft SQL Server 2000
Microsoft SQL Server 2005
Microsoft SQL Server 2008
Microsoft SQL Server 2008 R2
描述：http://technet.microsoft.com/zh-cn/security/bulletin/MS12-060

系统安全技巧

网络爬虫概述

网络爬虫(Web Crawler)，又称网络蜘蛛(Web Spider)或网络机器人(Web Robot)，是一种按照一定的规则自动抓取万维网资源的程序或者脚本，已被广泛应用于互联网领域。搜索引擎使用网络爬虫抓取 Web 网页、文档甚至图片、音频、视频等资源，通过相应的索引技术组织这些信息，提供给搜索用户进行查询。随着网络的迅速发展，万维网成为大量信息的载体，如何有效地提取并利用这些信息成为一个巨大的挑战。不断优化的网络爬虫技术正在有效地应对这种挑战，为高效搜索用户关注的特定领域与主题提供了有力支撑。网络爬虫也为中小站点的推广提供了有效的途径，网站针对搜索引擎爬虫的优化曾风靡一时。

传统网络爬虫从一个或若干个初始网页的 URL(Universal Resource Locator 统一资源定位符)开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列，直到满足系统的一定条件停止抓取。现阶段网络爬虫已发展为涵盖网页数据抽取、机器学习、数据挖掘、语义理解等多种方法综合应用的智能工具。

网络爬虫的安全性问题

由于网络爬虫的策略是尽可能多的“爬过”网站中的高价值信息，会根据特定策略尽可能多的访问页面，占用网络带宽并增加 Web 服务器的处理开销，不少小型站点的站长发现当网络爬虫光顾的时候，访问流量将会有明显的增长。恶意用户可以利用爬虫程序对 Web 站点发动 DoS 攻击，使 Web 服务在大量爬虫程序的暴力访问下，资源耗尽而不能提供正常服务。恶意用户还可能通过网络爬虫抓取各种敏感资料用于不正当用途，主要表现在以下几个方面：

1)搜索目录列表

互联网中的许多 Web 服务器在客户端请求该站点中某个没有默认页面的目录时，会返回一个目录列表。该目录列表通常包括可供用户点击的目录和文件链接，通过这些链接可以访问下一层目录及当前目录中的文件。因而通过抓取目录列表，恶意用户往往可获取大量有用的资料，包括站点的目录结构、敏感文件以及 Web 服务器设计架构及配置信息等等，比如程序使用的配置文件、日志文件、密码文件、数据库文件等，都有可能被网络爬虫抓取。这些信息可以作为挑选攻击目标或者直接入侵站点的重要资料。

2)搜索测试页面、手册文档、样本程序及可能存在的缺陷程序

大多数 Web 服务器软件附带了测试页面、帮助文档、样本程序及调试用后门程序等。这些文件往往会泄漏大量的系统信息甚至提供绕过认证直接访问 Web 服务数据的方法，成为恶意用户分析攻击 Web 服务器的有效情报来源。而且这些文件的存在本身也暗示网站中存在潜在的安全漏洞。

3)搜索管理员登录页面

许多网络产品提供了基于 Web 的管理接口，允许管理员在互联网中对其进行远程管理与控制。如果管理员疏于防范，没有修改网络产品默认的管理员名及密码，一旦其管理员登录页面被恶意用户搜索到，网络安全将面临极大的威胁。

4)搜索互联网用户的个人资料

互联网用户的个人资料包括姓名、身份证号、电话、Email 地址、QQ 号、通信地址等个人信息，恶意用户获取后容易利用社会工程学实施攻击或诈骗。

因此，采取适当的措施限制网络爬虫的访问权限，向网络爬虫开放网站希望推广的页面，屏蔽比较敏感的页面，对于保持网站的安全运行、保护用户的隐私是极其重要的。

基于网络爬虫技术的 Web 漏洞扫描

前面提到的网络爬虫对网站的间接安全威胁，是通过对网络站点的信息收集为不法份子的非法访问、攻击或诈骗作准备。随着安全技术的发展，利用网络爬虫技术对 Web 漏洞的直接探测已经出现，这会直接影响到 Web 服务器的安全。Web 服务器漏洞中，跨站脚本(Cross Site Script)漏洞与 SQL 注入(SQL Injection)漏洞所占比例很高，这两种漏洞均可以通过对网络爬虫的改进来进行探测。由于缺乏足够的安全知识，相当多的程序员在编写 Web 应用程序时对网页的请求内容缺乏足够的检查，使得不少 Web 应用程序存在安全隐患。用户可以通过提交一段精心构造的包含 SQL 语句或脚本的 URL 请求，根据程序的返回结果获得有关的敏感信息甚至直接修改后台数据。基于目前的安全现状，网络爬虫技术在 Web 漏洞扫描上的应用，大大提高了发现漏洞的效率。

基于网络爬虫技术的 Web 漏洞扫描大至分为如下过程：

1)页面过滤：通过自动化的程序抓取网站页面，对包含

等标签的 Web 页面进行 URL 提取处理，这些 HTML 标签中包含 URL 信息，便于恶意用户进行更深入的 Web 访问或提交操作。

2)URL 匹配：对 Web 页面中的 URL 进行自动匹配，提取由参数组合而成的动态查询 URL 或提交 URL，进行下一步的漏洞探测。如动态查询 URL “http://baike.xxxx.com/searchword/?word=frameset&pic=1”，其中 frameset 为 URL 中

动态的参数部分，可以进行参数变换。提交 URL 用于把 Web 用户的输入提交到服务器进行处理，其参数多为用户输入，同样可以进行参数变换。

3)漏洞试探：根据动态查询 URL 或提交 URL，自动在参数部分进行参数变换，插入引号、分号(SQL 注入对其敏感)及 script 标签(XSS 对 其敏感)等操作进行试探，并根据 Web 服务器返回的结果自动判断是否存在漏洞。如“URL 匹配”中的动态查询 URL 可以变换成 http://baike.xxxx.com/searchword/?word= &pic=1 进行跨站脚本漏洞探测。

如何应对爬虫的安全威胁

由于网络爬虫带来的安全威胁，不少网站的管理人员都在考虑对爬虫访问进行限制甚至拒绝爬虫访问。实际上，根据网站内容的安全性及敏感性，区别对待爬 虫是比较理想的措施。网站的 URL 组织应该根据是否为适合大范围公开，设置不同的 URL 路径，在同一 Web 页面中既有需要完全公开信息也有敏感信息时,应 通过链接、标签嵌入网页等方式显示敏感内容，另外尽可能把静态页面等经评估安全性较高的页面与安全性较差的动态页面从 URL 上分开。当限制爬虫时可以针对 URL 路径的安全性与敏感性对不同种类的爬虫与代理进行限制。

限制爬虫可以通过以下几种方法实现：

1) 设置 robots.txt 文件

限制爬虫最简单的方法是设置 robots.txt 文件。robots.txt 文件是搜索引擎爬虫访问网站的时候要查看的第一个文件，它告诉爬虫程序在服务器上什么文件是可以被查看的，如设置 Disallow: /，则表示所有的路径均不能查看。遗憾的是并不是所有的搜索引擎爬虫会遵守这个规则，因此仅仅设置 robots 文件是不够的。

2) User Agent 识别与限制

要对不理会 robots.txt 文件的爬虫访问进行限制，首先要把爬虫流量与普通用户的访问流量进行区分，即对其进行识别。一般的爬虫程序都可以通过其 HTTP 请求中的 User Agent 字段进行识别，该字段使服务器能够识别客户使用的操作系统及版本、CPU 类型、浏览器及版本、浏览器渲染引擎、浏览器语言、浏览器插件等。爬虫的 User Agent 字段一般与浏览器的有所不同，如 Google 搜索引擎爬虫 User Agent 字段中会有类似 Googlebot 的字符串，如 User-Agent: Googlebot/2.1 (http://www.google.com/bot.html)，百度搜索引擎爬虫则会有类似 Baiduspider 的字符串。不少 Web 服务器软件如 Apache，可以设置通过 User Agent 字段进行访问过滤，可以比较有效的限制大部分爬虫的访问。

3) 通过访问行为特征识别与限制

对于在 HTTP 请求的 User Agent 字段刻意伪装成浏览器的爬虫，可以通过其访问行为特征进行识别。爬虫程序的访问一般是有规律性的频率比较高，区别于真实用户浏览时的随意性与低 频率。对这类爬虫的限制原理与 DDoS 攻击的防御原理很相似，都基于统计数据。对于这类爬虫的限制只能通过应用识别设备、IPS 等能够做深度识别的网络设 备来实现。用网络设备限制网络爬虫，不仅比较全面，而且非常适合在多服务器情况下进行统一的管理，避免多服务器单独管理有可能造成的疏漏。

结束语

网络爬虫及其对应的技术为网站带来了可观访问量的同时，也带来了直接与间接的安全威胁，越来越多的网站开始关注对网络爬虫的限制问题。随着互联网的 高速发展，基于网络爬虫与搜索引擎技术的互联网应用将会越来越多，网站管理员及安全人员，有必要了解爬虫的原理及限制方法，准备好应对各种各样的网络爬虫。

来源：51CTO