doi: 10.3969/j.issn.1671-1122.2009.10.013

基于元搜索引擎实现被篡改网站发现与攻击者调查剖析

诸葛建伟[1, 2]，袁春阳[3]

（1. 北京大学计算机科学技术研究所，北京 100871；2. 网络与软件安全保障教育部重点实验室（北京大学），北京 100871；
3. 国家计算机网络应急技术处理协调中心，北京 100029）

摘 要： 网站篡改是目前一类主要的网络攻击形式，对网站安全运营和形象造成了严重的威胁。网站篡改攻击发现与监测是安全研究和应急响应的一个重要任务，本文引入元搜索引擎技术对网站篡改攻击进行发现，并对攻击者的互联网踪迹进行进一步搜索和调查分析。概念验证性实验结果表明元搜索引擎技术可用于构建简单有效的网站篡改监测方案，并在攻击者调查剖析方面具有其他方法不可替代的优势。

关键词： 元搜索引擎；网页篡改；网站攻击；调查剖析

中图分类号： TP393.08 **文献标识码：** A

0 引言

随着上世纪九十年代 Web 的快速发展和普及，网站成为互联网上最重要的应用，据中国互联网络信息中心统计，截至 2008 年底中国网站数已达到 287.8 万个，较上年度增长 91.4%[1]。由于网站应用的普及，网站篡改 (Website Defacement) 也得到网络骇客们的青睐，成为其炫耀其攻击能力、表达观点诉求及相互间攻击的主要方式。网站篡改是一类较早出现且流行已久的网络攻击形式，一般是网络骇客 (Cracker) 所为，在利用特定攻击手段入侵网站后，将网站页面进行替换，从而宣示入侵成功或表达攻击者的某种观点诉求[2]。一般攻击者会在替换页面中留下他们的代号、所属团队等信息，作为声明他们“战绩”的依据，有时也会给未能保障网站安全的管理员留下讯息加以嘲讽。在由地缘政治、价值诉求等各种诱因所引发的网站篡改攻击中，攻击者还会以在替换页面中发表他们的言论和观点。本文引入元搜索引擎技术对网站篡改攻击进行发现，并对攻击者的互联网踪迹进行进一步搜索和调查分析。

1 基于元搜索引擎实现被篡改网站发现

1.1 搜索引擎与元搜索引擎

搜索引擎是网民在互联网中获取所需信息的基础应用，据中国互联网络信息中心统计报告，2008 年底搜索引擎的使用率为 68.0%，在各互联网应用中位列第四，全年搜索引擎用户增长 5100 万人，年增长率达到 33.6%[3]。该统计数据表明搜索引擎已经被普通网民所普遍接受和应用。

而网络黑客们则早已认识到搜索引擎在查找攻击目标并采集相关信息方面的作用，提出了Google Hacking技术方法[4]，通过 Google 等搜索引擎所提供的灵活接口精心构造复杂的搜索请求，利用搜索引擎强大的互联网内容爬取和索引能力从海量信息中过滤出与计算机安全相关信息内容，如存在特定安全漏洞的网站，个人敏感信息等。

而网络攻击者作为强烈依赖于互联网的一类特殊网民用户，其网络攻击、社区交流等各种行为在互联网上留下了大量踪迹。因此安全研究人员和应急响应组织也完全可以利用搜索引擎，针对各种不同攻击形式的特性，构造独特的搜索请求，对攻击者的网络踪迹进行发现和监测，以及更加深入的调查剖析。

为了提升搜索能力和覆盖范围，近年来研究者提出了元搜索引擎的概念[5]，元搜索引擎又称集合型搜索引擎，将多个单一搜索引擎集成在一起，提供统一的检索接口，将用户的检索提问同时提交给多个独立的搜索引擎，并根据多个独立搜索引擎的检索结果进行去重、排序等二次加工，并将综合后的结果输出给用户。

虽然目前的商业搜索引擎如 Google, Baidu, Yahoo! 等已具有足够的互联网信息采集和搜索能力，从而满足普通网民查找信息的需求。但对于如被篡改网站监测、攻击者网络踪迹调查等特殊安全研究需求，单一搜索引擎尚无法保证全面的监测范围，因此利用元搜索引擎技术综合多个搜索引擎的监测范围和搜索能力，将有助于提升对网络攻击的监测效果。

1.2 面向被篡改网站发现的元搜索引擎工具

本文在 The Honeynet Project 内部工具 Skynet 基础上，进一步集成 Baidu 搜索引擎查询接口，实现了一个面向被篡改网站发现和确认的元搜索引擎工具。该工具以 Perl 语言开发，设计框架如图 1 所示，其主体结构分为元搜索引擎模块

和结果确认综合模块。

元搜索引擎模块通过集成 Perl 程序档案库 (CPAN) 中已有的 Google AJAX Search API、Yahoo! Search Public API 及 Baidu 查询接口进行实现。但不同搜索引擎对复杂查询请求的语法要求存在较大差异，如对于在网页标题中查找“hacked by”关键字，搜索范围限制于“gov.cn”域名的查询，Google 所支持的查询输入为 allintitle: “hacked by” site:gov.cn，而 Baidu 所支持的查询输入为 site:（gov.cn）title:（“hacked by”）。因此元搜索引擎模块需针对数据库中用户指定的查询关键字、查询域名范围等信息构造各个引擎可接受的查询请求，然后通过查询接口向搜索引擎提交并获取返回结果信息。

被篡改网站页面存在较高的动态变化性，网站管理员在发现页面被篡改后将迅速移除替换页面并恢复至正常状态。而搜索引擎对某些网站的爬取更新周期可能较长，因此在搜索引擎返回结果中发现的被篡改网站可能已被修复。为了能够获取到在监测时间点确认的篡改网站攻击事件，工具中由结果确认综合模块对元搜索引擎模块获取的页面链接进行下载和内容确认，通过检查最新下载页面中是否仍含有用于识别篡改页面的查询关键字，进行该页面链接是否确认篡改页面的判断。经确认后的被篡改网站信息经过综合之后输出至 MySQL 数据库中提供给用户。确认机制的引入使得元搜索引擎工具能够输出确定性的篡改网站攻击事件信息，而非包含大量误报的疑似攻击事件信息，这对支持有效的安全应急响应具有重要意义。

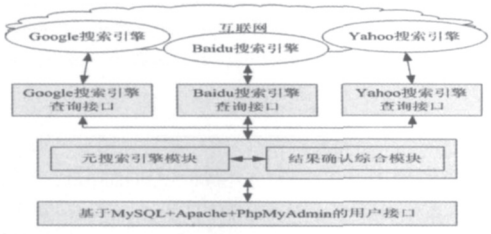

图1 面向被篡改网站发现的元搜索引擎工具设计框架

1.3 实验及结果分析

我们利用上述介绍的元搜索引擎工具进行了一个概念验证性的被篡改网站监测实验，为了保证检测准确率效果，目前所使用的查询策略只有一个简单的在 HTML 页面标题中搜索“hacked by”关键字，搜索站点范围为“.cn”域名，实验周期为 2009 年 3 月 25 日开始至 5 月 31 日，元搜索引擎工具设置为每天凌晨 1 点执行一次。

实验周期内共发现 423 个不同的篡改网站，每日新发现的篡改网站数量趋势如图 2 所示，在 3 月 25 日第一次运行时共发现 153 个确认的被篡改网站，平均每天新发现的篡改网站数量为 6.2 个，除第一次运行外最多一天能够新发现 43 个。我们对发现的被篡改网站进行了进一步统计分析，其域名分布如表 1 所示，其中所占比例最多的是“.cn”国家域名和“.com.cn”域名，均为 29.8%。而政府网站域名“.gov.cn”占 26.5%，共发现了 112 个被篡改网站，而“.gov.cn”域名网站仅占 .cn 域名总数的 1.1%[3]，这说明政府网站所面临的网站篡改攻击威胁更加严重。

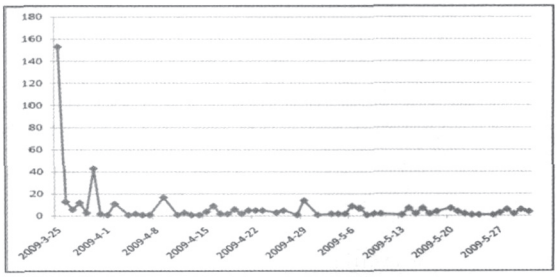

图2 利用元搜索引擎在网页标题中搜索”hacked by”新发现并确认的篡改网站数量趋势

表1 实验中发现的被篡改网站域名分布

域名	发现被黑网站数量	所占比例
.cn	126	29.8%
.com.cn	126	29.8%
.gov.cn	112	26.5%
.org.cn	29	6.9%
.edu.cn	17	4.0%
.net.cn	13	3.1%
总计	423	100%

2 利用元搜索引擎进行攻击者调查剖析

在发现被篡改网站的基础上，我们还进一步利用元搜索引擎技术进行了攻击者调查剖析实验，将从“黑页”中提取具有独特性的攻击者代号及联系方式等信息作为搜索关键字，利用元搜索引擎对攻击者网络踪迹进行全面搜索和监控，并通过安全研究人员的深入剖析给出攻击者的人物调查档案。

表2 被篡改网站的Top 10攻击者及进一步元搜索得到的页面数

Top 10 攻击者	篡改网站数	所属国家 / 地区	Google 搜索页面数	Baidu 搜索页面数	Yahoo! 搜索页面数
Iskorpitx	23	土耳其	260	227	99
Sinaritx	15	土耳其	51	46	72
"Iran Black Hats Team"	11	伊朗	30	97	25
MazHaR_FasHisT	10	伊朗	52	14	12
spo0feR	10	中国	22	44	39
GHoST61	8	土耳其	72	34	46
“黑客代号 1”	8	中国	50	10	10
RED	6	不详	N/A	N/A	N/A
“黑客代号 2”	6	中国	39	5	11
"Buzul Atay"	6	土耳其	13	6	14

从我们发现的 423 个被篡改网站中，我们对 Top 10 攻击者进行了进一步调查剖析，如表 2 所示，除中国之外，土耳其和伊朗的攻击者最为活跃，其中 Iskorpitx、Sinaritx 和 Iran Black Hats Team 均为持续性进行网站篡改攻击的黑帽子组织，在 CNCERT/CC 年报和 zone-h.org 统计信息中经常出现。Google/Baidu/Yahoo! 搜索引擎对 Top 10 攻击者独特的代号均能搜索到较多有效信息页面，其中包含一些曾被攻击者篡改的网站页面，以及一些反映攻击者基本轮廓的原始信息内容。从表中也可以看出，这三个主流搜索引擎对不同关键字所能

够搜索到的页面数比较各有优势，这也验证了元搜索引擎综合多个搜索引擎能力的必要性。

在实验中发现的 Top 10 攻击者中，我们基于元搜索引擎所提供的能力，对来自中国的

“黑客代号 1”和“黑客代号 2”(经分析，两个代号为同一攻击者) 进行了深入的调查剖析，如表 3 所示，元搜索引擎通过查询该攻击者独特的代号“黑客代号 1”和“黑客代号 2”，可以获取到大量关于该名攻击者网络攻击行为踪迹的有价值信息，从中可进一步分析得到攻击者其他代号（如“黑客代号 3”、“黑客代号 4”）和联系方式（如 QQ:8xxxxxx8、stxxxxx@xxxxxx.cn）等线索，并可利用这些线索扩大搜索范围。对表 3 中利用攻击者“黑客代号 1”/“黑客代号 2”的各个独特关键字搜索到的大量网页信息进行全面仔细的整理和分析之后，参考 The Honeynet Project 提出的黑客调查分析方法 [6]，我们总结出如表 4 所示的人物调查档案，可以清晰地展示”黑客代号 1”的基本情况和主要黑客历程。

表3 利用元搜索引擎对攻击者“黑客代号1”/“黑客代号2”的进一步搜索获取页面数

攻击者独特的搜索关键字	Google 搜索页面数	Baidu 搜索页面数	Yahoo! 搜索页面数
" 黑客代号 1"	38	9	15
" 黑客代号 2"	45	1	6
" 黑客代号 3"	41	42	44
" 黑客代号 4"	56	76	71
"QQ:8xxxxxx8"	57	3	26
stxxxxx@xxxxxx.cn	3	1	2
dxx@xxxxxxxxx.com	15	1	8

表4 “黑客代号1”/“黑客代号2”的元搜索调查剖析档案

基本信息	人物调查档案
黑客代号	" 黑客代号 1"
其他代号	" 黑客代号 2"/" 黑客代号 3"/" 黑客代号 4"
Email 地址	stxxxxx@xxxxxx.cn
即时通讯	QQ:8xxxxxx8, msn: dxx@xxxxxxxxx.com
博客	http://blog.xxxxxxx.cn/blog.asp?name= 黑客代号 3
	http://8xxxxxx8.qzone.qq.com/
所在团队	中国 xx 联盟 (2001-2003) -> x.x.x (2004-)
所属社群	灰帽子社群 / 计算机朋克
黑客动机	地缘政治诱因，进入社会团体，社区名声地位
擅长攻击方式	网站入侵、网站篡改
黑客历程	2001 年中国 xx 联盟核心成员，参与“中美黑客大战”；
	2003 年左右创建 XX 网站；
	2004 年遭中国 xx 联盟通缉，知名度提升；
	2004 年开始组建中国 x.x.x 黑客团队，2005 年对国外网站发动密集型攻击；
	2005 年 9 月在 zone-h.org 站点统计中高居榜首，2006 年正式招收成员；
	2008 年 7 月左右入侵中国 XX 部网站，至今仍活跃地进行网站篡改等攻击。
撰写文章	2003 年《XXXXXX》长篇连载及题记；
	2004 年《一次 XXXX 的入侵》、《XXXXXX 的高级手段与方法》等；
	2007 年《XXXXXXXX 破解器》等；
	2008 年《XXXXXX 的思考》等。

上述的攻击者调查剖析实验结果说明元搜索引擎是一种进行网络攻击者调查剖析简单有效的方法，所达到的调查剖析效果也能够满足安全研究和应急响应的基本需求。

3 总结

本文利用元搜索引擎技术对网站篡改攻击进行发现，并对攻击者的互联网踪迹进行进一步搜索。在 The Honeynet Project 内部工具 Skynet 基础上，集成 Google、Baidu 和 Yahoo! 搜索引擎构建了一个面向网站篡改主动发现和确认的元搜索引擎，并进行了国内万维网上被篡改网站的监测实验，在 2 个月时间内发现了超过 400 个确认的被篡改网站；进一步利用元搜索引擎技术对所发现的 Top 10 攻击者进行互联网踪迹搜索，根据搜索信息对其中的一名攻击者进行了全面的调查剖析，给出了较为完整的人物调查档案。实验结果表明元搜索引擎技术可用于构建简单有效的网站篡改监测方案，并在攻击者调查剖析方面具有其他方法不可替代的优势。（责编 岳逍远）

参考文献：

[1] 中国互联网络信息中心 (CNNIC), 第 23 次中国互联网络发展状况统计报告 [EB/OL], http://www.cnnic.cn/uploadfiles/pdf/2009/1/13/92458.pdf, 2009 年 1 月 .

[2] Wikipedia, Website defacement [EB/OL], http://en.wikipedia.org/wiki/Website_defacement, accessed June 2009.

[3] 国家计算机网络应急技术处理协调中心 (CNCERT/CC), CNCERT/CC 2004 年网络安全工作报告 [EB/OL], http://www.cert.org.cn/upload/2004CNCERTCCAnnualReport_Chinese.pdf, 2005 年 .

[4] 国家计算机网络应急技术处理协调中心 (CNCERT/CC), CNCERT/CC 2008 年网络安全工作报告 , 2009 年 .

[5] J. Long, Ed. Skoudis. Google Hacking for Penetration Testers [B], Syngress, 2005.

[6] M. Manoj, E. Jacob. Information retrieval on Internet using meta-search engines: A review, Journal of scientific & industrial research [J], 2008, 67(10): 739-746.

基金项目：国家 242 信息安全计划 (2007A16)，高等学校博士学科点专项科研基金资助课题 (200800011019)

作者简介：诸葛建伟 (1980-)，男，副研究员，博士，主要研究方向：网络与系统安全；袁春阳 (1979-)，男，工程师，博士，主要研究方向：网络安全监测与应急响应。

上接第 37 页

[5] Chiueh T, Sankaran H, Neogi A. Spout: A Transparent Distributed Execution Engine for Java Applets[C]. Proceedings of the 20th International Conference on Distributed Computing Systems (ICDCS'00), 2000. 2000: 394-401.

[6] Calder B, Chien A A, Wang J, et al. The Entropia Virtual Machine for Desktop Grids[C]. Proceedings of the 1st ACM/USENIX International Conference on Virtual Execution Environments, 2005. 2005: 186-196.

[7] Liang Z, Venkatakrishnan V N, Sekar R. Isolated Program Execution: An Application Transparent Approach for Executing Untrusted Programs[C]. Proceedings of Annual Computer Security Applications Conference (ACSAC'03), 2003. 2003: 182-191.

基金项目：江苏省自然科学基金 (BK2009485)；总装预研项目 (9140A06040107JB8101)；装备预研重点基金项目 (9140A06010408JB8101)；国家 863 项目 (编号：2007AA01Z126)

作者简介：缪嘉嘉 (1980 -)，男，讲师，博士，主要研究方向：信息安全、分布式计算；尹小虎 (1979 -)，男，技师，博士，主要研究方向：装备维修、信息安全；温研 (1981 -)，男，工程师，博士，主要研究方向：虚拟化、信息安全；冷健 (1975 -)，男，高级工程师，博士，主要研究方向：信息安全，密码学。