网络化条件下漏洞信息的获取及处理方法研究

淮甲刚[1],黄曙光[2],唐和平[3]
(1.电子工程学院 研一队,安徽 合肥 230037;
2.电子工程学院 网络系,安徽 合肥 230037;
3.电子工程学院 博士生队,安徽 合肥 230037)

摘 要:针对目前众多计算机安全机构所使用的计算机漏洞信息的现状和存在问题,提出了开源漏洞库批量下载、权威漏洞库查询、信息搜索等漏洞信息自动获取方法,对获取的 XML、HTML 和文本结果文件进行信息抽取,实现了漏洞信息的多源融合。

关键词:漏洞;信息抽取;信息融合

中图分类号:TP393 **文献标识码**:A

Research on the collection and process methods for security vulnerability information via Internet

HUAI Jia Gang[1], HUANG Shu Guang[2], TANG He Ping[3]
(1.Team 1 of Master Electronic Engineering Institute, Hefei 230037, China;
2.Depatment of Networle, Electronic Engineering Institute, Hefei 230037, China;
3.Team of Doctor, Electronic Engineering Institute, Hefei 230037, China)

Abstract: This paper aims at the actuality and existing problems of security vulnerability used by many computer security departments, and proposes several strategies to automatically collect vulnerability information by methods of downloading the open source database, submitting a query to database and using a search engine then discusses information extracting technique from XML, HTML and text files and establish multi-source data fusing.

Key words: security vulnerability; information extraction; information integration

安全漏洞信息作为计算机安全研究的基础数据,是各种计算机安全事件处理的数据来源。目前,多个从事计算机安全研究的机构和公司都配备了相应的计算机漏洞数据库,但是由于采用的规范、标准等指标不同,这些漏洞数据库结构和内容差异较大,对漏洞属性的描述也存在很大差别(甚至互相矛盾),使得各个组织之间难以就漏洞信息进行交互和共享。此外,互联网上发布的漏洞信息多为原始描述信息,从网上手工查找下载最新公布的漏洞信息,要加以提取、整理、验证、入库,费时费力。因此,一般情况下自建漏洞库的数据更新速度跟不上新漏洞信息的发布速度,严重影响了其效能的发挥。

为了解决这些问题,信息工程大学的孙学涛等提出了通用脆弱点数据库的构建方法和标准[1];美国海军研究生院的 ARNOLD A D 等人指出,网上没有一个漏洞数据库的信息是完善的,应当从网上多途径获取漏洞信息建立关系数据库并进行挖掘[2];国家计算机网络入侵防范中心的王晓甜描述了安全漏洞自动收集系统的设计与实现[3];西安电子科技大学杨晓彦提出了一种漏洞信息收集和发布机制[4]。

本文在进一步分析广大用户对漏洞信息越来越高的要求和继承前人研究的基础上,开展网络化条件下通用漏洞信息获取、处理方法和流程的研究,以期建立更加统一、完整、规范、可直接为计算机安全研究服务的漏洞数据来源。

1 漏洞信息的获取策略

为了确保漏洞信息的权威性和准确性,必须慎重选

择信息来源。在本课题研究中,以国际漏洞统一 CVE 标号为标识,国际权威漏洞数据库中的信息为主体,其他途径获得的信息为补充,建立漏洞信息获取、处理、维护和使用的体系,如图 1 所示。

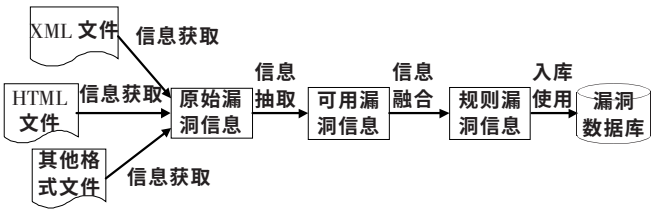

图 1 漏洞信息获取和处理流程

1.1 开源漏洞数据库批量下载

本课题对漏洞信息的获取,主要来源于网络批量下载。目前,由于各种原因,提供漏洞数据直接下载的组织只有美国国家标准与技术委员会的国家漏洞数据库 NVD 和美国安全组织创建的开源漏洞数据库 OSVDB。

(1)NVD(National Vulnerability Database)[5]

NVD 是美国国家标准与技术委员会 NIST(National Institute of Standards and Technology)的计算机安全资源中心 CSRC(Computer Security Resource Center)所创建的,提供的漏洞信息内容比较简练,目前收录有 CVE 编号的漏洞数量达到 37 301 条,包含 9 条漏洞属性,分别是:CVE 编号、易受攻击的系统号、影响的软件列表、发布时间、最后修改时间、CVSS 相关属性、CWE 编号、参考信息、漏洞摘要等。目前提供 XML 文件形式的漏洞信息下载,数据库的下载页面为:http://nvd.nist.gov/download.cfm。

(2)OSVDB(Open Source Vulnerability Database)[6]

OSVDB 目标是在安全漏洞方面为全世界免费提供准确、详细和公正的技术信息。该漏洞库收集了在操作系统、软件、协议和硬件设施以及信息技术基础组织部分的几乎所有漏洞。收录各种漏洞数量达到 54 686 条,包含漏洞描述、分类信息、解决方案、影响的系统、相关信息、来源、博客信息、注释信息等 8 条漏洞属性。目前提供 XML 格式漏洞信息下载,但需要注册。

1.2 权威漏洞数据库查询

国际上部分权威漏洞库出于商业目的,对普通用户只提供在线查询,这些数据库包括:美国计算机应急响应组 US-CERT,安全焦点 Security Focus;ISS 公司的 X-Force;Cerias 公司的漏洞库等。由于 NVD 和 OSVDB 中下载的漏洞属性较少,远远适应不了网络的需求,还需要从以上漏洞数据库里得到补充和完善,即通过对个别漏洞信息进行在线查询,将结果补充到相关漏洞信息。要实现漏洞信息的自动下载,必须解决两个问题:查询表单的自动填写和提交;查询结果的自动获取和下载。

(1)表单的自动填写和提交

目前,大多数漏洞网站提供的漏洞信息查询采用 input 表单,如图 2 所示。用户填写查询表单发送请求,浏览器在后台将数据 post 到目标网址,并获取响应数据。因此,只要获取浏览器中 post 的数据内容和目标网址,就可以用程序实现表单的自动填写和提交。获取 post 的内容和目标网址通过抓包实现。目前常用工具有 Ultra Network Snuffer 等,自动提交可以通过许多编程工具中集成的网络接口来实现,如.Net Framework 中的 HttpWebRequest()函数等。

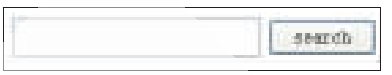

图 2 查询表单

```
<form name="Form1" method="post" id="Form1">
<input name="SearchInfo" type="text" id="SearchInfo"/>
<input type="submit" name="Submit" value="search" id="Submit"/>
</form>
```

(2)查询结果的获取

查询结果的获取是使用程序自动获得服务器对 post 信息处理以后的返回结果,同样可以通过编程工具中集成的网络接口来实现。为了完整展示从漏洞网站自动化的查询某一条漏洞信息并获取查询结果的过程。如,假定 postData 为 post 的信息内容,url 为 post 的目标地址。

```
byte[] data=encoding.GetBytes(postData);
HttpWebRequest request =(HttpWebRequest)WebRequest.Create(url);//准备请求…
request.Method="POST";//设置方法为 post;
Stream outstream=request.GetRequestStream();
outstream.Write(data,0,data.Length);
HttpWebResponse response=(HttpWebResponse)request.GetResponse();//发送请求;
Stream instream=response.GetResponseStream();
StreamReader sr=new StreamReader(instream);//返回结果
string content=sr.ReadToEnd();//结果存入 content;
```

1.3 其他获取途径

除了上述漏洞信息获取途径以外,还可以借助于其他途径来获取:

(1)大型软件厂商公司的门户网站(如 Microsoft、Cisco、Adobe 等)会及时公布其软件产品新发现的漏洞信息并提供解决方案,通过这些信息可以了解新漏洞部分属性。

(2)借助于对 Baidu、Google 等知名搜索引擎,重点在论坛、邮件列表、新闻组等网站,对所要了解的漏洞进行搜索。

(3)通过购买较完备的漏洞数据库(如国内应用最广泛的绿盟漏洞数据库),直接获取现有的结果,避免大量无意义的重复研究。

实践证明，根据实际情况，灵活应用多种漏洞信息获取手段和策略，可以取得更好的收集结果。

2 漏洞信息的处理技术

由于漏洞信息的获取来源不同，所获取的漏洞信息有 XML 文件、网页文件以及其他非结构化文本等，其结构、内容和所包含的信息量差别很大，对收集的原始漏洞信息的有效处理，并根据一定的属性筛选规则自动存入漏洞数据库中极为重要。漏洞信息处理可分为信息抽取和信息融合。

2.1 漏洞信息抽取技术

信息抽取是从大量结构化和非结构化的数据中，抽取出感兴趣的信息内容，形成结构化的记录。

2.1.1 XML 文件信息抽取

XML 是可扩展的(eXtensible)标记语言，它允许根据所提供的规则制定各种标记。该文档描述了漏洞 CVE-2009-0281 的 ID、影响软件、发布时间等属性。从漏洞库下载的 XML 文件片段如下。

```
<entry id="CVE-2009-0281">
<vuln:cve-id>CVE-2009-0281</vuln:cve-id>
-<vuln:vulnerable-software-list>
<vuln:product>cpe:warhound:walking_club</vuln:product>
</vuln:vulnerable-software-list>
<vuln:published-datetime>2009-01-27T13:30:00.360-05:00
</vuln:published-datetime>
</entry>
```

对属性信息的抽取主要用到了载入函数 xmldoc.load()、读取函数 SelectSingleNode()和条目内容属性 InnerText。

2.1.2 HTML 文件信息抽取

对所获取的 HTML 页面的抽取有 2 种方法：(1)首先进行去噪处理，然后将网页代码格式化成 XML 形式的文件，用 XPath 提取出感兴趣的漏洞信息，常用工具有 Chris Lovett 的 SgmlReader 和 Simon Mourier 的.NET HTML Agility Pack 等，该方法最终回归到 XML 抽取。(2)直接对网页源代码进行模式识别和字符串匹配。根据关键字和关键词匹配，找出属性名(如图中“发布时间”)；根据字体大小、颜色等不同(如图中的粗体)，对属性名和属性内容进一步区分和验证，如图 3 所示。

图 3 漏洞信息查询结果页面

此外，针对同一个网站所生成动态网页具有相同结构的特点，提前定义相对应的抽取模式，对该站所有页面按照抽取模式进行抽取，可以有效提高识别准确度。

2.1.3 文本信息抽取

上文提到的抽取策略仅适用于结构化和半结构化的信息，对于经由其他途径获取的信息，如黑客站点、软件厂商网站、或论坛上对部分漏洞信息的描述，归于非结构化信息，以普通文本为主要格式，要从中获取有价值的信息主要用到文本挖掘技术。文本挖掘属于数据挖掘的一部分，通过对文本信息的分类、聚类采用一定的挖掘算法，自动找出文本中所蕴含的漏洞属性，文本挖掘流程如图 4 所示。

图 4 文本挖掘流程图

2.2 漏洞信息融合技术

数据融合技术是利用计算机对不同途径获取的各种信息源，在一定准则下加以自动分析、综合，以完成所需的决策和评估任务而进行的信息处理技术。本文中使用数据融合技术来处理从网上获取和抽取后存储在本地的大量原始和多源的漏洞信息。主要流程包括结构分析、规范化、去重、补缺、冲突处理等。

2.2.1 结构分析

通常情况下，从原始漏洞信息中抽取的结果往往在总体结构、数据构成上很不统一。如 xfocus 漏洞库描述了 13 条漏洞属性，而绿盟漏洞库只有 8 条。即使描述漏洞的同一属性，采用的字段也可能不同，如 NVD 中对漏洞的特征描述属性名为 summary，而 OSVDB 中则为 Description。因此，必须对所抽取的漏洞信息进行结构分析，提炼出框架模型，自动地识别出漏洞信息描述的具体内容，将采取不同命名的同一漏洞属性尽可能对应起来，便于下一步的处理。

2.2.2 规范化

规范化是对采用不同标准描述的漏洞信息进行统一。规范化需要解决 3 个问题：(1)标准获取。识别出各个组织对同一漏洞属性描述上所采用的不同标准。(2)标准比较。根据各所采用标准的评定规则，比较各自的评定侧重点和优缺点。(3)标准选取和转换。根据对漏洞信息的要求，选择或者制定出最切合实际使用的标准，并对采用其他标准描述的信息进行转换。例如，在安全漏洞的危害级别评定中，目前主要的评定方式有 3 种：(1)以微软、FrSIRT 等为代表的“高、中、低”等常见的评定方法；(2)以 US-CERT 为代表，使用数值表示漏洞级别；(3) 目前正在普及的 CVSS(Common Vulnerability Severity System)漏洞评级标准[4]。3 种标准各有利弊，必

须经过判断比较,确定适合研究的评定方法。

2.2.3 冗余和补缺

冗余是规范化以后的漏洞信息中包含对同一属性的多个相同或相似描述:(1)不同漏洞库的描述造成的冗余,合理选取所有冗余中一项即可;(2)多个漏洞库对同一个漏洞属性的不同命名造成实质内容上的冗余,需要对属性命名的相似性做以判断,然后进行筛选。

补缺处理是对于所有漏洞数据库中都没有描述或没有确定描述的某个漏洞属性,通过一定的方法合理补充。补缺处理通常有如下方法:(1)根据同一软件中与该漏洞相似的漏洞的对应属性,推测出该属性最可能的值;(2)通过人工搜索或实验得到结果,进行补充;(3)暂时保留,等待别人的研究结果出来以后再进行补充。

2.2.4 冲突处理

数据冲突是对某一漏洞的同一属性来自不同数据库的多个描述互相矛盾。对于某些关键属性,应当做出合理取舍。比如对同一条漏洞 CVE-2006-1301,NVD 漏洞数据库给出的危害级别为高危级;而 FrSIRT 漏洞数据库给出的危害级别为低级,因此需要做以判断。在属性冲突处理中,可以采取如下方案:(1)按照多条冲突信息来源漏洞库的权威性、准确性等信息,制定各自权值,选择权值最大、即最可能、可靠的值;(2)取多条描述的均值,即使偏离正确值,也不会太大;(3)人工参与决策,通过实际验证或其他方法确定该属性。

本文从计算机安全漏洞信息的需求,介绍了漏洞信息的获取策略,从信息抽取和数据融合的角度探讨了对通过不同来源获取的原始和多源漏洞信息的处理流程。本课题的工程研究成果,可以直接应用于计算机安全研究上,为各种需求的用户提供标准统一的、最新的、全面的漏洞信息,提高其对安全情况处理的能力,并提高处理结果的时效性和可信度。

参考文献

[1] 孙学涛,李晓秋,谢余强.通用脆弱点数据库的构建[J].计算机应用,2002,22(9):42-44.

[2] ARNOLD A D, HYLA B M, ROWE N C. Automatically building an information-security vulnerability database [J]. US Naval Postgraduate Sch. Monterey, CA. Information Assurance Workshop[C], IEEE. 2006:376-377.

[3] 王晓甜,张玉清.安全漏洞自动收集系统的设计与实现[J].计算机工程,2006,32(20):177-179.

[4] 杨晓彦.网络安全信息系统的研究[D].西安:西安电子科技大学,2007.

[5] National Vulnerability Database[DB/OL].http://nvd.nist.gov,2008.

[6] Open Source Vulnerability Database[DB/OL].http://www.osvdb.org,2009.

(收稿日期:2009-09-14)

作者简介:

淮申刚,男,1984 年生,硕士研究生,主要研究方向:网络安全。

黄曙光,男,1963 年生,博士生导师,主要研究方向:网络安全。