(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103428219 A
(43)申请公布日 2013.12.04

(21)申请号 201310374770.8

(22)申请日 2013.08.25

(71)申请人 金华比奇网络技术有限公司
地址 321017 浙江省金华市李渔路1118号创新大厦3楼

(72)发明人 池水明 周苏杭 陈勤 孙斌 张旻 方晓波

(74)专利代理机构 杭州求是专利事务所有限公司 33200
代理人 周烽

(51)Int.Cl.
H04L 29/06(2006.01)

权利要求书1页 说明书3页 附图1页

(54)发明名称

一种基于网页模板匹配的web漏洞扫描方法

(57)摘要

本发明公开了一种基于网页模板匹配的web漏洞扫描方法，该方法通过计算检测到页面与模板页面的结构相似度，判断该页面是否需要抓取并进行漏洞扫描；本发明对设备要求简单，漏洞检测准确率高，同时在应用中表现出良好的稳定性。

CN 103428219 A

1. 一种基于网页模板匹配的 web 漏洞扫描方法，其特征在于，包括如下步骤：

（1）维护一个扫描网站的目录树，目录树的每个节点均为一个目录；

（2）利用网络爬虫抓取网页，根据网页的 URL 地址将网页放到相应的目录节点；

（3）当从一个目录节点中随机抽取的网页数量达到指定值时，对该目录中的网页进行模板匹配分别记录该目录下网页的相似值和不相似值；

（4）模板匹配过程中维护两个队列，其中待分析队列保存目录中尚未处理的网页，模板队列保存匹配成功的网页；初始时，目录中一个网页保存到模板队列，其他网页都保存到待分析队列；

（5）待分析队列出队一个待分析网页，若待分析队列为空则转步骤 9；

（6）遍历模板队列，分别计算待分析网页与模板队列中网页的相似度；

（7）若相似度超过阈值则继续遍历模板队列，否则转步骤 5；

（8）若步骤 7 中模板队列遍历结束，且相似度均大于阈值则网页进入模板队列；

（9）计算目录中网页匹配成功的概率，即模板队列中网页数与目录下网页数的百分比；匹配成功的概率大于一定阈值，则不再继续爬行该目录下其他网页，否则需要继续爬行该目录下其他网页；

（10）对目录树中所爬取的所有网页进行 SQL 注入测试；

（11）扫描结束。

一种基于网页模板匹配的web漏洞扫描方法

技术领域

[0001] 本发明涉及信息安全和网页架构领域，尤其涉及一种基于网页模板匹配的web漏洞扫描方法。

背景技术

[0002] Web应用程序面向广大Web用户，一旦出现严重漏洞，其危害将非常大。Web应用程序存在许多种漏洞，导致易受到攻击，其中，SQL 注入攻击(SQL injection)是目前主流的Web攻击方法之一。SQL 注入攻击者利用Web 应用程序没有对用户输入数据的合法性进行判断，通过Web页面的输入区域(如URL、表单等)，用精心构造的 SQL 语句插入特殊字符和指令，从而对后端数据库进行攻击，以获得管理员权限。

[0003] SQL 注入漏洞检测的基本原理是采用模拟攻击方式，构造特殊的SQL 语句对目标Web 站点的URL 地址进行注入测试，然后根据返回的网页内容确定是否存在注入漏洞。例如，若在URL 地址后附加SQL 语句“and 1=1”测试语句执行后返回正常网页，而附加“”或“and 1=2”测试语句执行后返回包含数据库错误信息的网页或者其他与正常时相异的网页，则判定该 Web 页面存在 SQL 注入漏洞。所谓“注入点”就是可以实行注入的地方，通常是一个访问数据库的连接。SQL注入漏洞扫描的过程可描述为：1、利用网络爬虫抓取网站网页；2、分析网页页面结构，寻找可能的注入点；3、向注入点发送模拟攻击数据；4、通过分析返回数据判断被检测的网页是否存在SQL注入漏洞。在对Web漏洞扫描过程中需要抓取网站所有网页，再进行分析、测试，这种方式虽然可以获得较高的扫描准确率，但对大型网站，将导致过高的扫描时间。

[0004] 因此，在保证漏洞扫描的准确率的情况下，实现适当缩减扫描规模，以提高SQL注入漏洞扫描效率成为了当前亟需解决的问题。

发明内容

[0005] 为了提高SQL注入漏洞扫描效率，本发明提供了一种基于网页模板匹配的web漏洞扫描方法。

[0006] 本发明的目的是通过以下技术方案来实现的：一种基于网页模板匹配的web漏洞扫描方法，包括以下步骤：

(1)维护一个扫描网站的目录树，目录树的每个节点均为一个目录；

(2)利用网络爬虫抓取网页，根据网页的URL地址将网页放到相应的目录节点；

(3)当从一个目录节点中随机抽取的网页数量达到指定值时，对该目录中的网页进行模板匹配分别记录该目录下网页的相似值和不相似值；

(4)模板匹配过程中维护两个队列，其中待分析队列保存目录中尚未处理的网页，模板队列保存匹配成功的网页。初始时，目录中一个网页保存到模板队列，其他网页都保存到待分析队列；

(5)待分析队列出队一个待分析网页，若待分析队列为空则转步骤(9)；

（6）遍历模板队列，分别计算待分析网页与模板队列中网页的相似度；

（7）若相似度超过阈值则继续遍历模板队列，否则转步骤（5）；

（8）若步骤（7）中模板队列遍历结束，且相似度均大于阈值则网页进入模板队列；

（9）计算目录中网页匹配成功的概率，即模板队列中网页数与目录下网页数的百分比。匹配成功的概率大于一定阈值，则不再继续爬行该目录下其他网页，否则需要继续爬行该目录下其他网页。

[0007] （10）对目录树中所爬取的所有网页进行 SQL 注入测试；

（11）扫描结束。

[0008] 本发明的有益效果是，本发明利用网页模板匹配技术实现了一种高效的 web 漏洞扫描方法，而不再依赖对整个网站所有网页抓取和扫描。该方法通过计算检测到页面与模板页面的结构相似度，判断该页面是否需要抓取并进行漏洞扫描；本发明对设备要求简单，漏洞检测准确率高，同时在应用中表现出良好的稳定性。

附图说明

[0009] 图 1 是基于网页模板匹配的漏洞扫描的流程图。

具体实施方式

[0010] 本发明提供一种基于网络模板匹配的、高效的漏洞扫描解决方案。该方案在网页爬取过程中维护一个扫描网站的目录树，目录树的每一个节点均为一个目录，目录中可以包含子目录及该目录中的网页。漏洞扫描当爬取到一个网页时，根据网页的 URL 地址将网页存放到相应的目录节点中，当一个目录节点中的网页数量达到指定值时，对该目录中的网页进行模板匹配，计算出网页的相似度，如果相似度达到一定阈值，则可判断该目录中的网页由同一模板生成，该目录中的其他网页无需再爬取。最后，方案将对目录树中所提取的网页进行注入检测。

[0011] 下面结合附图详细描述本发明。

[0012] 如图 1 所示，基于网页模板匹配的漏洞扫描对象精简方法包括如下步骤：

（1）维护一个扫描网站的目录树，目录树的每个节点均为一个目录；

（2）利用网络爬虫抓取网页，根据网页的 URL 地址将网页放到相应的目录节点；

网络爬虫模块下载并解析了当前页面，分别保存了当前页面和页面中的 url 队列。由于爬取的网页最终用于 SQL 注入测试，因此在 url 抓取过程中需要过滤不存在注入点的静态 url，从而避免重复抓取又可以减少队列空间的开销。网盘模块最终返回当前页面下可能存在 SQL 注入漏洞的 url 队列。

[0013] （3）当从一个目录节点中随机抽取的网页数量达到指定值时，对该目录中的网页进行模板匹配分别记录该目录下网页的相似值和不相似值；

本发明依据超几何分布进行网页抽样，超几何分布是统计学上一种离散概率分布。它描述了由有限个物件中抽出 n 个物件，成功抽出指定种类的物件的次数（不归还）。

[0014] 在网页相似性计算的不放回抽检中，若 N 条 url 中有 M 条 url 为不相似网页链接，抽检 n 条时所得不相似数 X=k，则 $P(X=k)=C(M,k)\cdot C(N-M,n-k)/C(N,n)$，$C(a,b)$为古典概型的组合形式，a 为下限，b 为上限。此时我们称随机变量 X 服从超几何分布（hypergeometric

distribution)。本发明采用随机抓取限定数量 $\eta \in [Y, Z]$ 的网页数,Y 和 Z 为预先设定的值,计算得到这一数量内需要达到的相似度概率阈值 λ 。

[0015] (4)模板匹配过程中维护两个队列,其中待分析队列保存目录中尚未处理的网页,模板队列保存匹配成功的网页。初始时,目录中一个网页保存到模板队列,其他网页都保存到待分析队列;

为了解析目录中的网页,本发明设计了网页链接解析模块。解析的具体过程为:i) 获取一个网站上的链接;ii)设置变量用于过滤 <a> 标签和 <frame> 标签;iii)得到所有经过过滤的标签。解析的结果以网页的形式进行保存。

[0016] (5)待分析队列出队一个待分析网页,若待分析队列为空则转步骤(9);

(6)遍历模板队列,分别计算待分析网页与模板队列中网页的相似度;

本发明先将网页结构解析成标签的序列,通过对两个需要进行匹配的标签序列进行最长公子序列的计算得到标签序列间的相似度。

[0017] (7)若相似度超过阈值则继续遍历模板队列,否则转步骤(5);

(8)若步骤(7)中模板队列遍历结束,且相似度均大于阈值则网页进入模板队列;

(9)计算目录中网页匹配成功的概率,即模板队列中网页数与目录下网页数的百分比。匹配成功的概率大于一定阈值,则不再继续爬行该目录下其他网页,否则需要继续爬行该目录下其他网页。

[0018] 若抓取的网页中相似网页的概率大于 λ 则不再继续爬行该目录下其他网页,否则需要继续爬行该目录下其他网页。

[0019] (10)对目录树中所爬取的所有网页进行 SQL 注入测试;

SQL 盲注测试利用详细的出错结果获得数据是一种广泛的攻击技术。应用 SQL 盲注需要首先进行 SQL 盲注点的查找与确认工作,可以利用以下三点进行 SQL 盲注点的查找与确认工作:1)产生通用错误;2)确认盲注点;3)拆分注入。

[0020] 本发明的SQL注入测试主要采用SQL盲注测试,通过向服务器发送特意构造的SQL语句尝试获取数据库,并分析服务器反馈结果以确定是否存在SQL注入漏洞,若存在漏洞则返回漏洞类型。

[0021] (11)扫描结束。

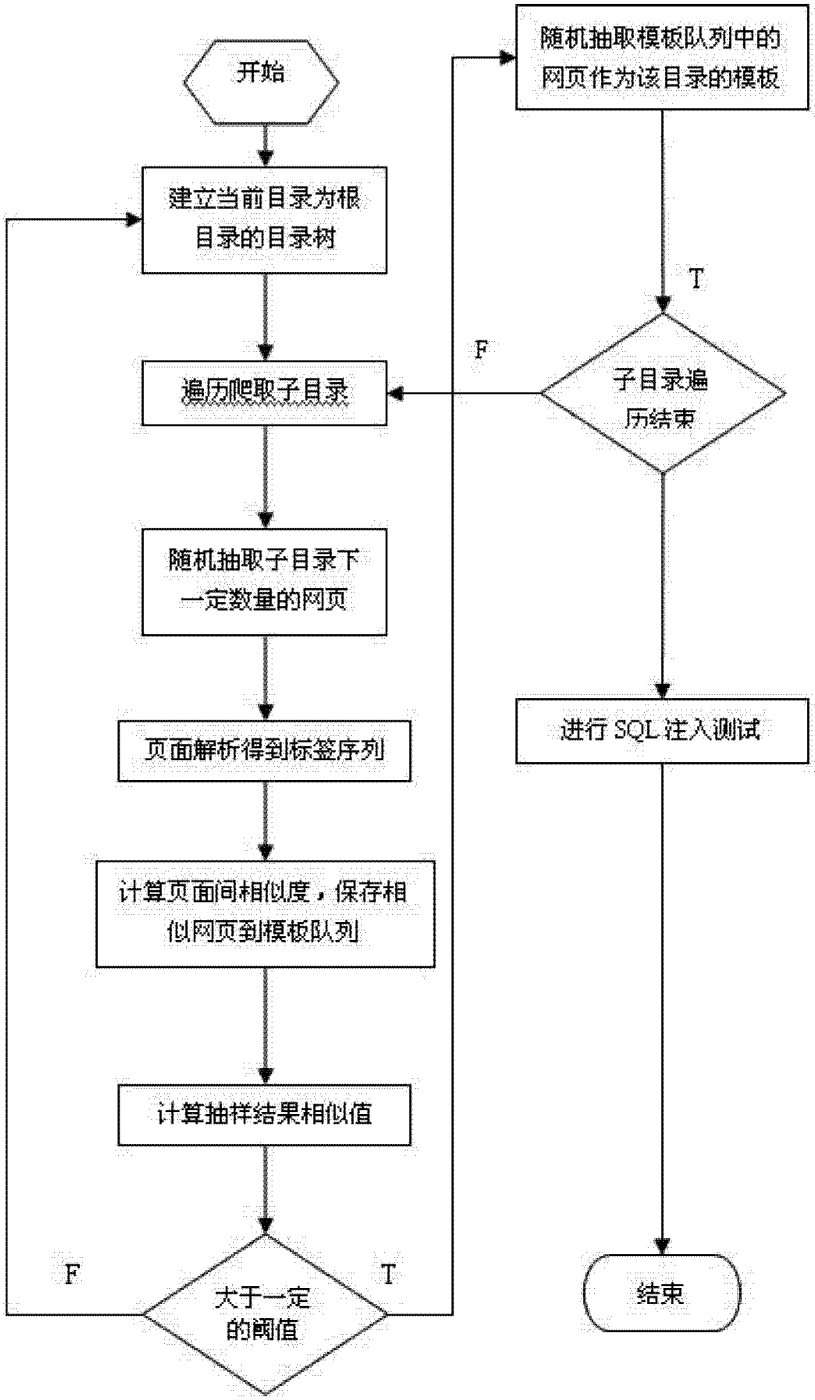
开始
建立当前目录为根目录的目录树
遍历爬取子目录
随机抽取子目录下一定数量的网页
页面解析得到标签序列
计算页面间相似度，保存相似网页到模板队列
计算抽样结果相似值
大于一定的阈值
F
T
随机抽取模板队列中的网页作为该目录的模板
T
子目录遍历结束
F
进行 SQL 注入测试
结束

图 1