(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 111212055 A
(43)申请公布日 2020.05.29

(21)申请号 201911394776.5

(22)申请日 2019.12.30

(71)申请人 上海安洵信息技术有限公司
地址 201100 上海市闵行区万源路2158号泓毅大厦A1007室

(72)发明人 郑华东 陈权 吴海波

(51)Int.Cl.
H04L 29/06(2006.01)
H04L 12/24(2006.01)
G06F 16/951(2019.01)
G06F 21/56(2013.01)

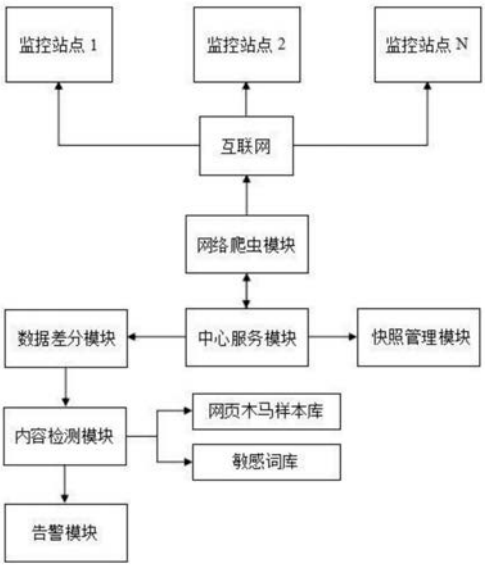
权利要求书1页 说明书3页 附图1页

(54)发明名称

非侵入式网站远程检测系统及检测方法

(57)摘要

本发明公开了一种非侵入式网站远程检测系统及检测方法，包括用于通过互联网获取被监控网站的网页源码和资源的网络爬虫模块、用于调度各功能模块的中心服务模块、用于对网站的正常状态进行备份的快照管理模块、用于将网络快照与当前源码和资源进行对比的数据差分模块以及用于快速检查出风险项的内容检测模块，中心服务模块的输出端分别连接网络爬虫模块、快照管理模块、数据差分模块的输入端连接，网络爬虫模块的输出端通过互联网连接各个监控站点，数据差分模块的输出端连接内容检测模块的输入端。本发明采用数据差分算法，检测网站页面源代码的变化，在不对被监管网站进行任何更改的前提下，对目标网站进行检测，不影响网站的运营管理。



CN 111212055 A

1.非侵入式网站远程检测系统及检测方法,其特征在于:包括用于通过互联网获取被监控网站的网页源码和资源的网络爬虫模块、用于调度各功能模块的中心服务模块、用于对网站的正常状态进行备份的快照管理模块、用于将网络快照与当前源码和资源进行对比的数据差分模块以及用于快速检查出风险项的内容检测模块,中心服务模块的输出端分别连接网络爬虫模块、快照管理模块、数据差分模块的输入端连接,网络爬虫模块的输出端通过互联网连接各个监控站点,数据差分模块的输出端连接内容检测模块的输入端。

2.根据权利要求1所述的非侵入式网站远程检测系统及检测方法,其特征在于:所述检测系统还包括用于对网站站长发出风险告警的告警模块,告警模块的输入端连接内容检测模块的输出端。

3.根据权利要求1所述的非侵入式网站远程检测系统及检测方法,其特征在于:所述内容检查模块包括用于对网页木马样本库进行检测的风险代码识别单元和用于对敏感词库进行检测的敏感词识别单元。

4.根据权利要求1至3任一项所述的非侵入式网站远程检测方法,其特征在于:该检测方法包括以下几个步骤:

A.先为目标网站创建网站快照,对网站的正常状态进行备份;

B.通过网络爬虫获取被监控网站的网页源代码和资源,将网站当前源码和资源与网站快照进行对比后,提取发生变化的差异项;

C.将提取的差异项通过网页木马样本库和敏感词库进行对比,检查出网页源代码和资源中存在风险项;

D.当检查出网页存在风险项时,通过短信和邮件对网站站长发出告警。

非侵入式网站远程检测系统及检测方法

技术领域

[0001] 本发明涉及网络安全技术领域，特别是一种非侵入式网站远程检测系统及 检测方法。

背景技术

[0002] 随着网络黑客活动猖獗，经常会发生企业的网站被挂上影响用户安全的木 马链接而造成严重的后果，由于网络防护技术能力不足，导致网站的防护措施 不到位，往往是网站被篡改后很难及时发现，引发严重的后果。

[0003] 为了保障这些网站的安全，上级部门专门成立了监管部门，对这些网站进 行监管，为这些站点提供一定的防护。传统采用的监管方式是网站检测，需要 在网站服务器上部署检测程序，实施起来非常不方便，同时因为需要对原始服 务的运行环境和安全策略做一定的修改，对服务本身的运行管理也会造成一定 影响。另外，在监管部门在对下属站点进行日常的监管中，被监管的站点往往 因为监管措施实施困难或者是因为其他方面的考虑而拒绝监管部门通过侵入式 (即在网站服务器上安装额外的监管软件或硬件)对自己的网站进行监管，而 监管部门在面对运行环境各异、部署环境千差万别的各类网站时，也很难提供 一种通用的实施方便的侵入式检测技术方案。

[0004] 由于网站大多是动态网页，内容是实时变化的，如何从这些变化的内容中 区分出合法的变更和非法的变更，例如政府部门首页，每天都有新的新闻推送， 任意时刻都可能会有新的留言或回复等变化的信息，如何将这些内容与黑客入 侵后插入的内容进行区分识别，是监管部门需要解决的问题。

发明内容

[0005] 本发明需要解决的技术问题是提供一种非侵入式网站远程检测系统及检测 方法，在不对被监管网站进行更改的前提下，对目标网站进行有效检测，不影 响网站的运营管理。

[0006] 为解决上述技术问题，本发明所采取的技术方案如下。

[0007] 非侵入式网站远程检测系统及检测方法，包括用于通过互联网获取被监控 网站的网页源码和资源的网络爬虫模块、用于调度各功能模块的中心服务模块、 用于对网站的正常状态进行备份的快照管理模块、用于将网络快照与当前源码 和资源进行对比的数据差分模块以及用于快速检查出风险项的内容检测模块， 中心服务模块的输出端分别连接网络爬虫模块、快照管理模块、数据差分模块 的输入端连接，网络爬虫模块的输出端通过互联网连接各个监控站点，数据差 分模块的输出端连接内容检测模块的输入端。

[0008] 上述非侵入式网站远程检测系统及检测方法，所述检测系统还包括用于对 网站站长发出风险告警的告警模块，告警模块的输入端连接内容检测模块的输 出端。

[0009] 上述非侵入式网站远程检测系统及检测方法，所述内容检查模块包括用于 对网页木马样本库进行检测的风险代码识别单元和用于对敏感词库进行检测的 敏感词识别单

元。

[0010] 上述非侵入式网站远程检测方法，该检测方法包括以下几个步骤：

[0011] A.先为目标网站创建网站快照，对网站的正常状态进行备份；

[0012] B.通过网络爬虫获取被监控网站的网页源代码和资源，将网站当前源码和 资源与网站快照进行对比后，提取发生变化的差异项；

[0013] C.将提取的差异项通过网页木马样本库和敏感词库进行对比，检查出网页 源代码和资源中存在的风险项；

[0014] D.当检查出网页存在风险项时，通过短信和邮件对网站站长发出告警。

[0015] 由于采用了以上技术方案，本发明所取得技术进步如下。

[0016] 本发明采用高效的数据差分算法，来检测网站页面源代码的变化，在不对 被监管网站进行任何更改的前提下，对目标网站进行有效的检测，不会影响网 站的运营管理。

附图说明

[0017] 图1为本发明的结构框图。

具体实施方式

[0018] 下面将结合附图和具体实施例对本发明进行进一步详细说明。

[0019] 非侵入式网站远程检测系统及检测方法，其结构框图如图1所示，包括网 络爬虫模块、中心服务模块、快照管理模块、数据差分模块、内容检测模块和 告警模块。网络爬虫模块用来通过互联网获取被监控网站的网页源码和资源， 中心服务模块用来调度各功能模块，快照管理模块用来对网站的正常状态进行 备份，数据差分模块用来将网络快照与当前源码和资源进行对比，内容检测模 块用来快速检查出风险项，告警模块用来对网站站长发出告警。中心服务模块 的输出端分别连接网络爬虫模块、快照管理模块、数据差分模块的输入端连接， 网络爬虫模块的输出端通过互联网连接各个监控站点，数据差分模块的输出端 连接内容检测模块的输入端，内容检测模块的输出端连接告警模块的输入端。

[0020] 网络爬虫模块是采用网络爬虫技术，模拟一个正常网民的信息去访问被监 管网站，抓取网站的网页源代码和资源作为网站快照存放于本地磁盘，后续由 中心服务模块的监控任务定时发送请求，获取当前的网站源码和资源进行数据 差分对比。

[0021] 中心服务模块用于调度各功能模块，同时提供对监控网站的管理以及资源 配置的工作，每间隔一段时间自动通过网络爬虫模块抓取目标网站。

[0022] 快照管理模块用于对网站的正常状态进行备份，作为后续对网站进行内容 检测时的对照依据。快照管理模块主要提供，快照创建，快照读取，快照更新， 快照删除的功能。

[0023] 数据差分模块采用数据差分算法，通过网站快照与网站当前源码和资源进 行对比，快速提取大声变化的差异项，并将差异项传送给内容检测模块进行检 查。

[0024] 因为网页源码中存在大量的javascript逻辑代码及网站自己的文字内容，如 果对网站网页进行全文内容检查，会产生大量误报，本发明中先通过数据差分 提取差异内容，只对差异内容进行检查则能避免有效误报情况，同时也能节约 服务器的性能开销。

[0025] 内容检测模块包括风险代码识别单元和敏感词识别单元，风险代码识别单 元用来对网页木马样本库进行检测，敏感词识别单元用来对敏感词库进行检测。内容检测模块

通过网页木马样本库和敏感词库能够从数据差分模块的产生的结 果中，快速检查出风险项。

[0026] 告警模块是在内容检测模块检查出网页存在风险项时，告警模块采用短信 和邮件的方式对网站站长发出风险告警，通知站长及时处理，避免引起不必要 的损失。

[0027] 当网站更新升级时，升级完成后由站长通知监管部门，监管部门对网站快 照进行及时更新，避免检测漏洞，对网站进行全面防护，提高了网站检测的效 果，保证了网站的安全。

[0028] 为了避免误报的情况发生，检测时会先提取出变化的内容，并对内容进行 精准的识别，识别依据采用风险代码样本库和敏感词库，样本库和敏感词库都 能进行升级更新，提供更高的可用性。

[0029] 本发明中的侵入式网站远程检测方法包括以下几个步骤：

[0030] A.先通过网络爬虫技术，获取目标网站的网页源码和资源，为目标网站创 建网站快照，由监管人员对目标站点进行审计，确定网站状态正常后进行快照 抓取，并将网络快照存放于本地磁盘，对网站的正常状态进行备份，为后续对 网站进行内容检测提供对照依据。

[0031] B.每隔一段时间自动通过网络爬虫技术，模拟一个正常网民的信息去访问被 监管的网站，获取被监控网站的网页源代码和资源，并将获取的被监控网站的 当前源码和资源与备份的网站快照进行差分对比，快速提取发生变化的差异项， 然后传给内容检测模块进行检查；

[0032] C.提取的差异项交由内容检测模块进行分析，将差异项通过网页木马样本库 和敏感词库进行对比，快速发现网页中存在的风险代码或敏感词，提取出网页 源代码和资源中存在的风险项；

[0033] D.当检查出网页存在风险代码或敏感词时，告警模块第一时间通过短信和 邮件形式对网站站长发出告警，通知网站站长及时进行处理，维护网站安全， 避免不必要的损失。

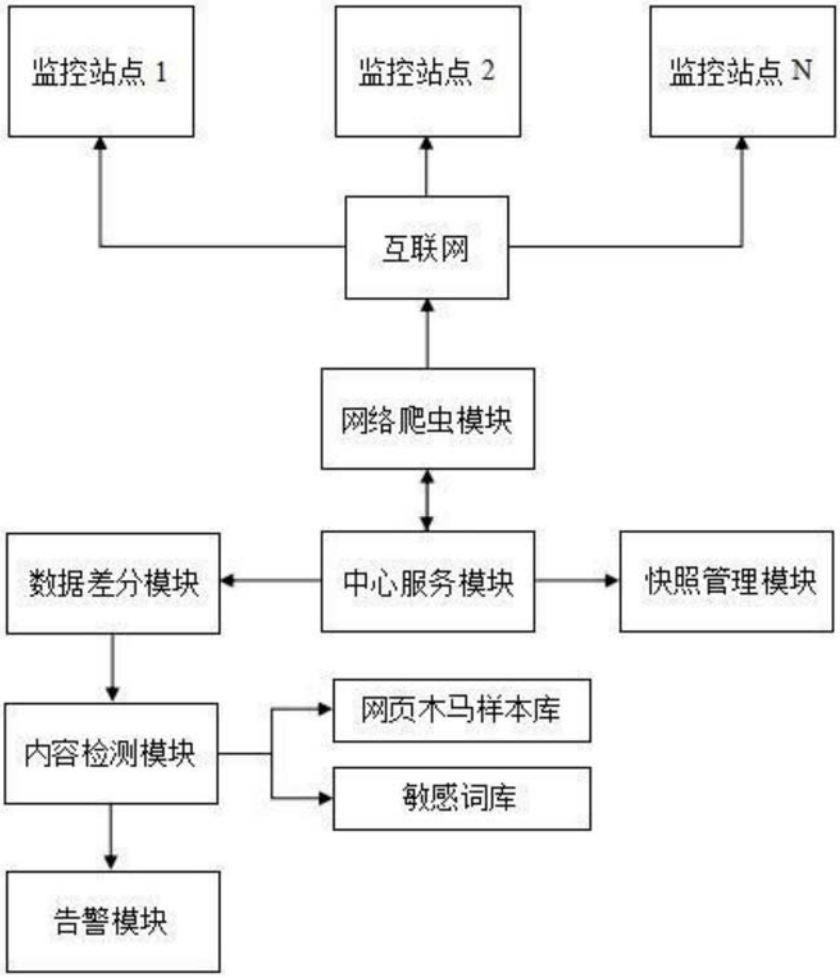

监控站点 1
监控站点 2
监控站点 N
互联网
网络爬虫模块
数据差分模块
中心服务模块
快照管理模块
内容检测模块
网页木马样本库
敏感词库
告警模块

图1