|  |
| --- |
| 基于聚类分析的HTTP协议攻击数据包检测工具 |
| **方案设计书** |

学　 院： 软件学院

指导教师： 胡海波

项目组成员： 方芳 万剑 任立峰

联系电话： 13983697592

2012-4-23

|  |
| --- |
|  |

# 摘要

Web应用已成为目前Internet上增长最快的信息服务,作为Web应用载体的HTTP协议常常被攻击者用于实施网络攻击。由于HTTP攻击方式的改进和规避方法的提高，在过去的几年里HTTP攻击频繁爆发并且危害越来越大，对于IDS固定特征值的匹对和单独依赖防火墙的防护变得越来越不可靠。

为了应对这种问题，自动提取特征、合理的将攻击归类并做出响应的防护，在近几年应运而生，并逐渐成为研究的焦点。本系统根据Nemean系统的核心思想，首先根据Nemean聚类算法进行基准向量的求解，求解的过程中也使用了数据包的抓取，提取Nemean特征数据和量化处理，最后根据聚类的结果再确定基准向量。然而本系统的最主要的模块是对数据包的分析模块，来确定攻击数据包的分类。在这个过程中通过对请求数据包的抓取过滤，解析提取HTTP数据包中Nemean特征数据，然后经过量化处理计算，比对基准向量进而分类的方式对HTTP数据包进行分类，最终判断其是否含有攻击性行为，并对攻击性行为归类做出警报。

通过本系统，可以很好的提高未知HTTP攻击的防范处理，加强HTTP应用服务的安全性。使用本系统的方法可以对其他协议进行分类，同时由于Nemean聚类算法的准确度较高，所以我们可以根据聚类的结果来提取特征值，为IDS特征值库提供特征值从而提高IDS的准确率。

**关键词**： **HTTP协议 Nemean聚类 基准向量 攻击分类 报警**

目录

[摘要 1](#_Toc323026463)

[1. 选题背景 3](#_Toc323026464)

[2. 创新与特色 3](#_Toc323026465)

[3. 系统方案 4](#_Toc323026466)

[3.1 总体方案 4](#_Toc323026467)

[3.2 模块原理介绍 5](#_Toc323026468)

[3.2.1 基准向量生成模块 5](#_Toc323026469)

[3.2.2 数据包分析模块 6](#_Toc323026470)

[3.2.3 报警和显示模块 7](#_Toc323026471)

[4. 可行性分析 7](#_Toc323026472)

[5. 进度规划 8](#_Toc323026473)

# 选题背景

随着互联网的飞速发展，Web应用已成为目前internet上增长最快的信息服务，所谓的web应用是指以浏览器作为用户界面，以HTTP作为传输协议，以网络服务器作为信息库的一种“浏览器-Internet-服务器”的通信交互方式。越来越多的政府机构、金融机构、企事业单位逐渐将自己的业务与工作平台建立在Web之上，Web技术已经不再是一种单纯的通信技术，在基于Web的通信网络中，更多的是承载着政治、经济、文化与利益的相关内容，对Web服务器的攻击可以产生更加深远的影响。目前基于HTTP协议的Web通信架构在一定程度上促进了Internet的蓬勃发展，但也为网络攻击开辟了一个全新的战场。

但是，作为目前应用最为广泛HTTP服务器，在提供Web服务时一般采用“盲方式”，也就是不限制访问者的来源、不对访问/浏览时间做约束、不对访问者的下载行为做约束。同时HTTP协议在注重功能实现与简单性之余，并没有太多的考虑自身的安全性。HTTP协议的“请求”与“响应”均采用明文结构，这种方式使攻击者很容易构造自己的HTTP“请求”或者“响应”用于各种攻击目的。与此同时，防火墙也只是验证HTTP协议本身的合法性，完全不能理解HTTP协议所承载的数据，也无从判断对HTTP服务器的访问行为是否合法。攻击者知道正面攻破网络防火墙十分困难，于是从简单的端口扫描攻击转向通过应用层协议进入企业内部，而这些攻击过程中产生的网络层数据，和正常数据没有区别。

综上所述，我们应该采取更加安全可靠的防护措施，针对诸如此类问题我们设计数据包的攻击检测。通过对HTTP请求数据包提取的Nemean数据特征，与已建立好的基准向量进行比对，从而判断该数据包是否含有攻击性行为，并对攻击方式进行分类。

# 创新与特色

1. 系统使用Nemean聚类算法确定基准向量准确率高。根据资料显示可得Nemean在基于HTTP的攻击探测中正确率达90%以上，由此可知我们基于HTTP的攻击检测分类的准确率较高。
2. 系统使用的Nemean数据特征简单易得。因为Nemean数据特征都是可以在数据包中直接提取的三个参数，因此我们需要分析的数据特征可以在获得数据包的同时简单快速的得到。
3. 系统可以判断未知的行为并对其进行分类。当一种未知的攻击行为攻击时，虽然并没有该攻击的具体信息，但是我们可以根据Nemean数据特征对其进行分类，判断它属于哪种攻击类型。
4. 系统有很好的扩展性。因为现在很多的入侵检测系统都是基于数据包特征值对攻击的分类，因此我们可以为入侵检测系统提供更高准确率的分类效果，以便提高对特征值提取的准确率。同时可以用相同的方法对其他协议的攻击进行分类。
5. 系统可以在判定数据包的攻击性后进行报警提示，并显示对其进行的攻击分类的结果。

# 系统方案

## 总体方案

由于Web应用已成为目前Internet上增长最快的信息服务，而HTTP协议则是Web上最重要的协议，Web服务攻击中最多的也是基于HTTP协议。因此根据现在计算机信息安全保护的需要，我们开发了这样一个系统，而这个系统主要是对于数据包基于对HTTP协议的攻击进行检测，当检测到它为非攻击的，则允许他进入网络；当检测它为攻击的，则经过计算对比进行分类，并报警和显示。系统的部署图如下图3.1所示

图3.1 系统部署图

在系统中，我们首先对数据包进行抓取，对抓取的数据包进行解析取得Nemean数据特征后，我们对得到的Nemean数据特征再进行量化；为了判断该数据包是否有攻击性，我们将量化后的Nemean数据特征进行计算；最后根据计算的结果与基准向量进行比对分类，并根据比对分类结果进行报警和显示。系统功能架构图如图3.2所示

图 3.2 系统功能架构图

## 模块原理介绍

### 基准向量生成模块

依据Nemean聚类算法，若有足够多已知的针对HTTP的网络攻击，我们首先提取数据包中的Nemean数据特征，Nemean数据特征可以看作一个三维的向量D（x，y，z），其中x，y，z分别指接收数据包的字节长度，接收数据包的请求类型和响应数据包中的响应代码，为了使量纲一致，我们在得到Nemean数据特征后先对三个值进行量化处理。通过Nemean聚类算法能够生成基准向量，归纳出各类行为特征，从而用于判断未知数据包类别。如图3.3所示



图 3.3 基准向量

其中，基准向量x0为接收数据包的字节长度，y0为接收数据包的请求类型，z0为响应数据包中的响应代码。每次提取x，y，z后对其进行量化，通过对足够多的已知攻击类型的数据包进行解析量化处理，聚类划分产生基准向量以用作对未知类型的数据包检测。

Nemean聚类算法的基本思想是计算两个Nemean数据特征A、B的来判断两个向量之间的相似度。其中的计算如公式3.1所示

 （3.1）

其中、为A、B特征向量，向量中每个值均为正数；为两个向量的内积；、分别为A、B的向量范数；为A、B的相似度，。

Nemean聚类算法则是根据来进行进行聚类，同时我们再根据聚类的结果得到基准向量，以备下一个模块使用。

### 数据包分析模块

数据包分析模块是该系统的总体模块，它负责整个系统数据包的处理流程，与基准向量生成模块类似，它包括对HTTP数据包的筛选抓取，解析提取HTTP数据包Nemean数据特征，对数据特征进行量化，对量化后的数据特征进行计算并与基准向量生成模块中生成的基准向量进行比对，比对的过程则是根据Nemean聚类的算法来求出该数据包的Nemean数据特征与基准向量的相似度，并与上一个模块得到的聚类结果进行比对，判断它的攻击类比，并将对比结果传递给下一模块。

主要原理：设计数据包的接受机制，Collector完成对网络上流经的数据报进行快速抓取，Filter对抓取到的数据报进行进一步的过滤，对不符合HTTP条件规则的数据报直接抛弃，对符合条件的数据进行下一步处理。Analyzer对过滤后系统关系的数据包进行分析和解析，从中提取有价值的信息（Nemean数据特征）。Quantizer对提取出来的信息进行量化处理，得出有计算价值的数字数据。Calculator对量化后的数据进行计算，得到最终一个可对比的值交给Compareter进行对比，从而得出对比结果，并交由下一模块处理。原理如图3.4所示

图 3.4 数据包分析原理

### 报警和显示模块

接收来自数据报分析模块中所得到的分析结果数据，根据分析结果数据判定其是否属于具有恶意行为的非法数据包，如果为非法数据包，则判定并显示其非法数据包类型，同时做出警报通知。

# 可行性分析

1. Nemean聚类算法，它的基本思想是通过Nemean特征向量来进行聚类并建立一个基准向量，与基准向量来计算角度并根据角度来进行分类。这个方法准确率高，由于我们使用过其他的聚类方法，因此有一定聚类算法的知识，对于实现此方法有一定的知识准备。
2. 本系统对数据包的分析依据为Nemean特征向量，而对于它的三个参数可以通过数据包直接提取，简单易得。
3. 本系统只是针对基于当前流行的web服务中最主要的HTTP来进行分析，针对性较强，也降低了对数据包的分析难度。但由于本系统的扩展能力很强，因此虽然只是根据一个协议的攻击但是却可以根据本系统方法来扩展到对其他一个或多个协议进行检测分类。
4. 本系统只是使用Nemean数据特征，而不使用特征值，降低了提取特征值的难度。但也正是因为这样，我们可以对未知的攻击进行分类。

# 进度规划

项目分为三个阶段：

1. 2012.5—2012.6
2. 对Nemean聚类算法的研究和改进
3. 对改进后的算法进行编程和验证
4. 2012.6—2012.7
5. 根据聚类算法确定Nemean数据特征的量化方式
6. 使用大量的不同类别的HTTP协议攻击的数据包来确定基准向量
7. 开发出系统的1.0版本，并对版本进行测试
8. 2012.7—2012.8
9. 对系统进行调试和优化，确定最终版本
10. 完成整个项目的相关文档并提交最后作品