DOI：10.3969/j.issn.1001-8972.2011.16.045

基于数据包分析的网页还原技术研究

陈淼
西南科技大学信息工程学院 621000

摘 要
本文结合应用环境和实际需求，研究了网络数据包信息还原涉及的数据包捕获技术、数据包分析技术、数据包重组技术、http信息识别技术。

关键词
网络数据包捕获；数据包重组；http信息提取

引言

在信息社会中，信息是维持生产活动、经济活动以及社会活动的重要资源，对政治、经济和文化都有着深远影响。不断探索网络信息监听与还原技术有助于建立可靠、高效的信息安全保障体系，对于维护社会政治稳定和国家信息安全具有重要的现实意义。因此，对因特网中一些重要数据信息进行还原和提取，是保证网络应用的健康发展和打击网络犯罪的一个重要手段。

一、系统设计

基于数据包分析的网页还原系统主要分以下几个模块：

1、网络数据包捕获模块

2、网络数据包协议分析与数据包重组模块

3、HTTP 协议信息还原模块

数据包捕获模块负责捕获流经用户所选网络设备的全部数据包，并将捕获得到的数据按照相应的规则以文件格式存储起来，以供后续模块进行分析、重组和还原。

数据重组模块首先读入数据包捕获得到的数据，分析每一个捕获到的数据包，将具有相同的源 IP 地址、源端口号、目的 IP 地址和目的端口号的数据包按照先后顺序存储在一起，重组成一个完整的数据包。

HTTP 协议信息还原模块通过一层层的数据包分解出来，读取出应用层http的负载信息，通过分析http请求头得到seq，并为此创建一个临时文件，保存数据包的简要信息，然后分析http响应头，得到content-encoding、conteng-type、content-length字段值，再存入临时文件，最后把http负载信息与具有相同seq、sport、dport的临时文件简要信息匹配，若匹配成功，则进行相应的数据插入，还原出相关网页信息。总流程图如图1 所示。

图1 总流程图

从降低系统设计复杂性和提高开发效率的角度考虑，结合实际需求出发，本系统没有采用数据库软件存储捕获到的网络数据包，而是直接以某种自定义文件格式将其存放在硬盘中。HTTP 协议信息还原模块既支持对于老版本 HTTP/1.0 协议网页信息的还原，也支持针对新版本 HTTP/1.1 协议的网页信息还原。

二、网络数据包捕获

原始数据包捕获是进行数据挖掘研究的基础，数据的可靠性决定了挖掘分析的准确度。在本次研究中我们利用 winpcap 提供的用户接口捕获校园网络拓扑中共享网络上主机的收/发数据包。经过测试，我们结合winpcap接口开发的数据包捕获软件捕包效率可以达到99.63%以上，见下表1[1]，可以忽略漏掉的少量数据包对实验结果的影响。

表1 数据包获取实验数据

N	Switchboard	Acquire ip packet	efficienc
1	326,719,129	325,831,088	99.72%
2	314,933,107	314,456,075	99.84%
3	693,752,426	691,505,725	99.67%
4	733,479,255	732,011,380	99.79%
5	1,466,177,102	1,460,754,475	99.63%
6	1,732,291,161	1,730,352,166	99.88%

三、网络数据包重组

由于 IP 不能保证可靠、有序的包传输，因此包有可能会被破坏或在到达时是无序的。另外，在捕获过程中会将重复的包视为TCP重发的结果，也可能会捕获那些不会到达到预定接收器的包。在这两种情况下，捕获过程都可能会接收重复的包。更复杂的是，TCP 不能保证重新传输的包会按照原始数据的同一方式再将这些数据分组。因此重构有序流就显得尤为重要。

如果一个流有起始、结束和两者间的所有东西，它就是完整的流。根据 SYN 可以确定包连接的 TCP 包所有字节流。

为了方便应用数据的恢复，需要将无序的数据片流有序化，使其表现为一个有序的数据片流(或者说是数据流)。在软件实现数据片有序化时，本设计采用了一个带头结点的双向链表队列，队列中的每个结点存储一个完整的TCP数据流的内容，另外还有两个元素，分别是指向前导和后续结点的结构体指针。

四、HTTP 协议信息还原

根据消息体与消息头之间有两个连续的回车换行符（ASCII 码为 \r\n）作为分隔的特征,利用匹配函数strstr()找到两个连续的回车换行符出现的位置，从此位置开始，到服务器端数据文件结尾的全部数据均为HTTP 消息体数据。

本模块首先使用find(ack,sport,dport)获得该数据包的信息,通过函数node_isempty()判断该数据包的data有效长度是否为0,如果不为0,则通过函数have_inserted_first_data()判断该数据包是否已经加入过http data内容,如果没有,则调用函数init_first_seq()初始化seq,然后使用函数insert_first_data插入数据包的第一条data内容；如果已经加入过data内容,则继续加入当前分片的data内容。然后通过函数get_complete_percent()判断当前数据包是否完整,如果不完整则完成data内容的处理；如果是完整的，则通过函数serialstream()重组当前数据包,最后把所有分片的内容整合在一起，写入文件。

结语

本文结合应用层http协议为例，进行了协议分析，通过编写程序实现了对网络中的http数据包进行捕获、分析和提取有用信息，得到了网络协议分析技术实际应用的结果。实现了一个融合解决 HTTP/1.0 和 HTTP/1.1 协议网页信息还原的通用HTTP协议信息还原系统。

参考文献
[1]Miao Chen, Shun-hua Tan, Guo-hai Yang , Yi-zhi Wang. Research on network business identification technology based on IP packets. IEEE ICACIA2010
[2]谭敏生，汤亮.基于HTTP的网络数据包还原技术研究[J]. 计算机技术与发展. 2007年，第17卷（6期）：176

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net