基于信息熵的加密会话检测方法

陈 利 张 利 班晓芳 梁 杰
(中国信息安全测评中心 北京 100085)

摘 要 传统协议分析方法在检测网络加密会话时大都通过端口识别,在加密应用使用非常规端口或者在周知明文端口出现加密流量时无法进行有效的检测。为此,提出基于信息熵的加密会话检测方法。该方法先对数据流按端口进行会话重组,再计算会话数据包字符熵,进而统计出整个会话字符熵,判断熵值是否属于训练模型正态分布置信区间,通过信息分布均匀度来检测加密会话。实验表明,该方法无需特征指纹库,且检测准确率高,并能实现实时检测和处理。

关键词 信息熵,加密会话,协议识别,正态分布,入侵检测
中图法分类号 TP303.0 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.1.033

Encrypted Session Detection Approach Based on Information Entropy

CHEN Li ZHANG Li BAN Xiao-fang LIANG Jie
(China Information Technology Security Evaluation Center, Beijing 100085, China)

Abstract Traditional protocol analysis algorithms detect the network encrypted session through the port. It cannot work when encrypted session uses unknown port or encrypted traffic appeares at known plaintext port. To this end, we put forward a detection approach of encrypted session based on information entropy. Firstly it reorganizes net flow according to the port, then calculates the entropy of each packet and statistical entropy value of the entire session, at last determines whether the value belongs to the normal distribution confidence interval, and identifies the encrypted session through character distribution uniformity. Experiments show that the approach does not need fingerprint database, and can achieve higher correct detection rate, real-time detection and processing.

Keywords Information entropy, Encrypted session, Protocol identification, Normal distribution, Intrusion detection

网络通信加密技术是保障网络数据安全性的重要手段,保证了在网络上传输的数据不被第三方窃取。对安全性要求较高的应用可采用不同的加密技术,如链路加密、节点加密、端到端加密[1]。传统的网络数据分析技术(如DPI深度数据包解析)通过分析数据包有效载荷字段进行会话协议识别[2],该技术已经成为当前主流网络安全产品的主要网络分析手段。然而,对于加密通信的网络程序,由于嗅探到的数据经过加密处理变成了乱码,因此无法进行内容分析[3];另外,一些以窃取用户或企业私有信息为目的的木马程序,以传播非法信息为目的的翻墙程序[4],采用常规通信端口,利用加密技术进行数据通信,从而使传统协议分析手段无法进行有效检测。因此对加密会话的有效检测变得迫切。

熵理论目前被广泛应用于信息安全领域的数据分析和异常检测[5]。与传统熵相比,相对熵[6]对异常检测往往有更好的检测效果。本文提出的基于信息熵的加密会话检测方法,以加密会话网络通信特征为研究对象,研究加密数据中常见字符和非常见字符在加密前后的出现规律,发现其统计分布特性,通过计算大量加密通信程序样本的会话字符熵并按正态分布[7]特性训练出检测模型。检测过程中计算特定端口通信会话字符熵,以其熵值是否属于正态分布模型置信区间来判断其信息均匀度高低,从而进行加密会话检测。

1 信息熵概念和意义

信息是信息论中最重要、最基本的概念。信息论创始人香农给信息下的定义为:信息是事物运动状态或存在方式的不确定性的描述[8]。设一个概率系统中有 N 个事件 $X=X_i$ $(i=1,2,\cdots,N)$。X_i 发生的概率为 P_i,从概率的角度出发,事件 X_i 发生后,信息量定义为 $H_i=-\log_2 P_i$[9]。H_i 对由 N 个事件构成的概率系统而言,产生的平均信息量为:

$$H=-\sum_{i=1}^{N} P_i \log_2 P_i \tag{1}$$

平均信息量反映了事件的不确定性程度。"熵"是用来表示能量分布均匀程度的术语,能量分布越均匀,熵就越大[10]。香农提出的"信息熵"概念解决了对信息的量化度量问题,他指出信息熵描述的是信源的不确定性,能有效地表现出同一属性上对应数据的集中和分散情况。

2 基于信息熵的加密会话检测方法

2.1 加密会话字符分布特性

在大多数网络通信应用中,网络会话中出现的字符是有

到稿日期:2014-02-14 返修日期:2014-04-21

陈 利 助理研究员,主要研究领域为风险评估、入侵检测,E-mail:bhchenli@sina.com;张 利(1972-),研究员,主要研究领域为入侵检测、风险评估;班晓芳(1973-),副研究员,主要研究领域为入侵检测、风险评估;梁 杰(1983-),助理研究员,主要研究领域为入侵检测、风险评估。

统计规律的，常见字符出现的频率高，非常见字符出现的频率低，如大小写英文字符和阿拉伯数字出现频率高。如果对网络连接进行加密，常见字符和非常见字符都成了乱码，出现的频率就趋于相等。基于这种思想，我们可以计算特定端口的加密流量和非加密流量的字符熵，通过训练学习，找出稳定状态下的加密信息熵和明文信息熵的取值范围，检测阶段判断未知会话信息熵是否属于一定阈值范围，从而判定未知会话是否为加密会话。

2.2 数据流的字符熵

设数据流由一组会话$\{C_i\}$组成，其中$i=1,2,\cdots,N$，N为会话的数目。会话C_i由一组数据包$\{P_j\}$按顺序组成，其中$j=1,2,\cdots,M$，M为会话中数据包的个数。每个数据包的内容载荷都可以看成是256个ASCII码组成的字符集合，通过统计ASCII码在每个数据包内容载荷中出现的频率，可以对单个数据包、会话和数据流的信息量进行度量，从而将加密流量从正常流量中区分出来。对于单数据包P_{ij}（第i个会话中的第j个数据包），其字符熵H_{ij}采用如下公式进行计算，这里采用相对熵：

$$H_{ij}=[-\sum_{n=0}^{255}(\frac{c_n}{S})\log_2(\frac{c_n}{S})]/\log_2 S \tag{2}$$

其中，S为数据包中所有字符的总数，c_n为字符n在该数据包中出现的次数。

数据包P_{ij}所在会话的信息熵为该会话中所有数据包字符熵的平均值，计算如下：

$$H_i=\frac{1}{M}\sum_{j=1}^{M}H_{ij} \tag{3}$$

2.3 会话信息熵分布规律统计

通过计算加密通信会话和非加密通信会话的会话字符熵值，以会话时间为变量建立函数关系$h=f(t)$，绘制坐标曲线图，样本通信程序会话信息熵随会话时间变化的结果如图1所示，其中包括3个样本加密会话和非加密会话信息熵的一般训练值。

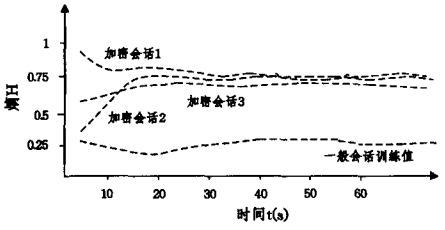

图1 样本通信程序会话信息熵随会话时间变化图

对大量加密会话和非加密会话的统计分析表明，当通信会话持续到一定时间后，会话信息熵值都会趋于稳定。对于非加密会话来说，随着会话密文数据增多，常见字符如大小写字母和阿拉伯数字出现次数多，非常见字符出现次数少，则会话字符分布均匀度低，其信息熵都会趋近于一个稳定的较低数值；而加密网络应用的数据结果则相反，随着会话数据增多，由于常见字符和非常见字符出现的概率趋同，因此会话字符分布均匀度较高，相应信息熵也会趋近于一个稳定的较高的数值，可以认为不同加密会话信息熵统计结果服从正态分布，则加密会话信息熵取值范围由均值μ和方差σ决定。

$$H\in[\mu-2\sigma,\mu+2\sigma] \tag{4}$$

通过计算网络会话每个数据包字符熵，进而计算整个会话字符熵，判断会话字符熵取值范围是否属于式(4)所示的正态分布置信区间，并进行加密会话的检测。实践中其对以窃取信息为目的的木马会话流检测有重要意义，如果某周知端口会话出现加密通信，或者本为明文传输的通信突然出现了加密信息（如80端口出现了加密会话），则有可能是木马伪装的正常应用，可进一步做内容分析。

3 实验及结果

本文实验是对大量加密会话流与一般网络数据流做加密会话检测。实验中搭建由若干主机构成的小型局域网，操作系统为Windows XP/7专业版，虚拟平台为Vmware Work Station7.0，检测程序开发平台为Python2.6.4。通过嗅探捕获大量常见的网络应用网络数据流（包括加密应用）作为训练和检测样本，另外还捕获若干加密通信的木马数据包作为加密会话流样本，样本会话列表如图2所示。

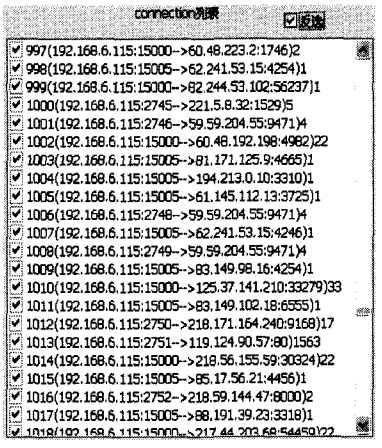

图2 实验样本会话列表

本文选取一个Web访问会话和TLSv1加密会话，由网络分析工具查看会话数据流内容，如图3、图4所示。

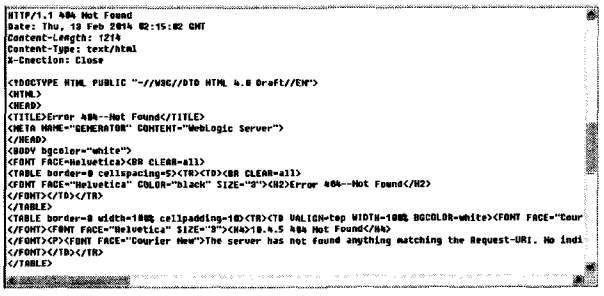

图3 Web访问会话数据流

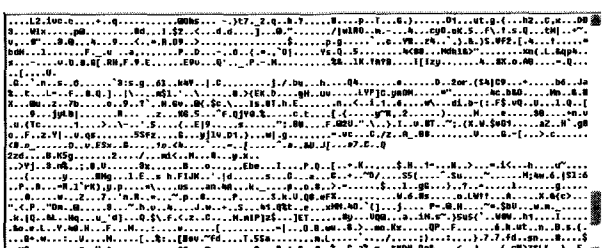

图4 TSLv1会话数据流

检测过程中，按照基于会话信息熵的检测模型，首先将数据流按会话重组，每个数据包分属于特定的会话，计算每个数据包的字符熵值，进而按均值计算整个会话的字符熵值。实

（下转第174页）

[3] Jøsang A, O'Hara S. Multiplication of Multinomial Subjective Opinions[C]// Proceedings of the International Conference on Information Proceeding and Management of Uncertainty. Dortmund, Germany, 2010: 248-257
[4] 王守信，张莉，李鹤松. 一种基于云模型的主观信任评估方法[J]. 软件学报，2010，21(6)：1341-1352
[5] Du Wei, Cui Guo-hua, Liu Wei. An uncertainty trust evolution strategy for e-Science[J]. Journal of Computer Science and Technology, 2010, 25(6): 1225-1236
[6] 张仕斌，许春香. 基于云模型的信任评估方法研究[J]. 计算机学报，2013，36(2)：422-431
[7] 陆玲玲，徐建，张宏. 人类心理认知习惯与云模型相结合的 P2P 信任模型[J]. 计算机科学，2012，39(8)：38-41
[8] Li De-yi, Liu Chang-yu, Gan Wen-yan. A new cognitive model: cloud model[J]. Int J of Inteligent Systems, 2009, 24(3): 357-375
[9] 王尚广，孙其博，张光卫. 基于云模型的不确定性 QoS 感知的 Skyline 服务选择[J]. 软件学报，2012，23(6)：1397-1412
[10] 马颖，田维坚，樊养余. 基于云模型的自适应量子粒子群算法[J]. 模式识别与人工智能，2013，26(8)：787-793
[11] 任剑. 基于云模型的语言随机多准则决策方法[J]. 计算机集成制造系统，2012，18(12)：2792-2797
[12] Huang Jing-wei, Fox M. An ontology of trust: formal semantics and transitivity[C]// Proceedings of the 8th international conference on Electronic commerce. ACM, New York, NY, USA, 2006: 259-270
[13] Williams C A, Heins R M. Risk Management and Insurance [M]. New York: MC Graw Hill, 1985
[14] Wang Shou-xin, Zhang Li, Wang Shuai, et al. A Cloud-Based Trust Model for Evaluating Quality of Web Services[J]. Journal of Computer Science and Technology, 2010, 25(6): 1130-1142
[15] 杨玉丽，彭新光，付东来. 基于云模型的主观信任评估机制[J]. 计算机工程与设计，2013，34(12)：4151-4155

(上接第 143 页)

验中 Web 访问会话共有 7 个网络连接，会话字符熵平均为 0.256，TLSv1 加密会话字符熵为 0.737。可以看出，TLSv1 会话字符熵属于置信区间$[\mu-2\sigma,\mu+2\sigma]$，从而被检测为加密会话，http 会话字符熵值不属于上述区间，则被识别为一般应用。大量实验证明，$\mu=0.755$ 和 $\sigma=0.045$ 的设置对于所有实验会话具有最佳检测效果。被检测应用的会话信息熵在正态分布曲线上的位置如图 5 所示。

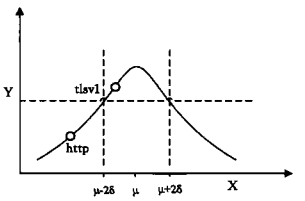

图 5 实验应用样本会话信息熵

本文选取了 48 个带加密功能的网络应用和 243 个一般网络应用，对加密应用和一般应用在真实网络环境中所产生的网络数据流进行加密会话检测。实验结果如表 1 所列。

表 1 检测结果统计

实验环境	网络状况	48 个加密网络应用			243 个一般网络应用		
		正确检测数(个)	错误检测数(个)	正确检测率(%)	正确检测数(个)	错误检测数(个)	错误检测率(%)
实验网	良好	46	2	95.8	238	5	2.05
公网	拥塞	44	4	91.7	234	9	3.7

经过大量样本实验，从正确检测率、错误检测率两个方面证明了该检测方法的可用性。在实验网环境下，由于网络状况良好、网络中丢包率低、重传包较少等原因，应用通信过程比较符合检测理想模型，因此正确检测率较高，错误检测率较低；在公网环境，拥塞的网络状况则会导致检测准确度有一定的降低。由表 1 中统计数据总体来看，该检测方法达到了比较高的正确检测率，将错误检测率降低在 5%以下，证明了该检测方法的可用性，对网络中加密会话的识别分析具有重要的意义，尤其对通过常规端口加密传输的泄密数据流检测，可从海量数据流中识别出可疑加密会话，有利于进一步针对性分析。

结束语 本文在对网络加密会话进行信息统计分析的基础上，提出基于信息熵的无指纹加密会话检测方法。该方法总结网络应用中常见字符和非常见字符在加密前后的出现规律，通过统计大量加密通信程序样本的会话信息熵并按正态分布特性训练出检测模型。检测中计算特定端口通信会话字符熵，以其熵值是否属于一定阈值范围来判断信息均匀度高低，从而进行加密会话检测。实验表明，该检测方法能够有效地检测出一般加密应用会话和加密木马会话，具有较高的正确检测率；同时，该方法高效地利用了处理器资源，在高速的网络环境中也能做到实时处理。此外，该方法已经被实际应用于网络入侵检测和木马行为监控等实时网络数据流分析当中，并取得了较好的效果，表明该方法具有很强的实用性。但是其仍存在一些不足，在时延较大的拥塞网络状况下，准确识别率略有降低，如何在拥塞网络环境下保证较高的准确识别率还有待提高。

参 考 文 献

[1] Lakhina A, Crovella M, Diot C. Characterization of Network-wide Anomalies in Traffic Flows[R]. Technical Report: BUCS-20040020. Boston University, 2004
[2] 高建明，龚亮亮，吕涛. 基于信息熵的目标平台识别方法[J]. 计算机应用与软件，2013，30(9)：171-184
[3] Kargupta H, Park B, Hershberger D, et al. Collective data mining: a new perspective toward distributed data mining[C]// Proceedings of Advances in Distributed and Parallel Knowledge Discovery. [S. l.]: AAAAI/ MIT Press, 2000: 128-175
[4] Sommer R, Paxson V. Outside the closed world: On using machine learing for network intrusion detection[C]// Proc. of 2010 IEEE Symposium on Secutiry and Privacy. 2010: 302-355
[5] 李文忠，左万利，赫枫龄. 一种基于信息熵的多维流数据噪声检测算法[J]. 计算机科学，2012，39(2)：123-144
[6] 王海龙，杨岳湘. 基于信息熵的大规模网络流量异常检测[J]. 计算机工程，2007，33 (18)：262-264
[7] Nehinbe J O. Automated technique for debugging network intrusion detection systems[C]// IEEE 2010 International Conference on Intelligent Systems, Modelling and Simulation (ISMS). Liverpool, 2010: 363-367
[8] 吴小叶，肖继民. 基于信息熵的网络异常流量的研究[J]. 广东通信技术，2008(4)：32-34
[9] Kim D S, Nguyen H N, Park J S. Genetic algorithm to improve SVM based network intrusion detection system[C]// Proc. of the 19th International Conference on Advanced Information Networking and Applications. 2005: 150-164
[10] 丁世飞，朱红，许新征，等. 基于熵的模糊信息测度研究[J]. 计算机学报，2012，30(8)：139-151