CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 36 卷第 11 期 2014 年 11 月
Vol. 36, No. 11, Nov. 2014

文章编号:1007-130X(2014)11-2128-04

基于 NP 模式的报文检测方法*

唐湘滟[1],程杰仁[1],殷建平[2],龚德良[3]

(1.海南大学信息科学技术学院,海南 海口 571101;
2.国防科学技术大学计算机学院,湖南 长沙 410073;3.湘南学院,湖南 郴州 423000)

摘　要:针对现有网络入侵检测方法中存在的不足,引入否定模式(NP)匹配的策略,提出了基于 NP 模式的报文检测方法。该方法先从待测报文内容模式集合中找出 NP 模式,根据 NP 模式将待测数据流分段;然后通过模式匹配引擎对分段内容进行模式匹配。实验结果表明,该方法能降低误报率,减少报文匹配次数,提高检测效率。

关键词:否定模式;模式匹配;入侵检测;网络安全

中图分类号:TP393.08　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2014.11.012

A packet detection method based on Negative Pattern

TANG Xiang-yan[1], CHENG Jie-ren[1], YIN Jian-ping[2], GONG De-liang[3]

(1. College of Information Science & Technology, Hainan University, Haikou 571101;
2. College of Computer Science, National University of Defense Technology, Changsha 410037;
3. Xiangnan University, Chenzhou 423000, China)

Abstract: To solve the problems in intrusion detection based on pattern matching, we present a packet detection method based on negative pattern, which uses the negative pattern matching strategy. Firstly, the method detects the NP pattern in the to-be-detected packets and divides the data into segments according to NP pattern. Secondly, the system detects the data segments by NP pattern. The experimental results show that the method can reduce the false alarm rate and improve the detection efficiency of the system.

Key words: negative Pattern; pattern match; intrusion detection; network security

1　引言

随着物联网、云计算、大数据等新兴技术的出现,对高速检测海量网络数据流的需求日益增加。模式匹配在入侵检测系统各个关键环节中起着非常重要的作用[1~3],其分为单模式匹配和多模式匹配两类。模式匹配由最初的 Boyer-Moore 算法[4]、KMP 算法[5]以及 Commentz-Walter 算法[6]等发展到复杂、效率更高的多模式匹配[7~9]。由于单纯的软件实现方法已经不能满足数以千计的模式匹配的实际要求,FPGA[10]、ASIC[11]、TCAM 等支持并行处理的专用硬件设备被大量应用,以提高模式匹配速度[12],当然这也增加了相应模块的设计复杂性。尽管多模式匹配技术得到了广泛的应用,其匹配速度比单模式匹配的速度快得多,但是仍然需要依赖硬件和一些辅助的软件策略方式实现报文检测[13,14],而且目前基于 FPGA 等器件和 Snort 规则集实现报文检测的方法,通常吞吐量不超过10 GBps[15]。为提高报文检测的吞吐量,满足用户的更高需求,本文基于否定匹配的思想提出了基于 NP 模式的报文检测方法。

* **收稿日期**:2014-05-20;**修回日期**:2014-07-20
基金项目:国家自然科学基金资助项目(61363071,61100194,61232016,60970034);海南省自然科学基金资助项目(614220);湖南省教育科学十二五规划课题资助项目(XJK011BXJ004);海南大学 D 类人才启动基金(kyqd1328);海南大学青年基金资助项目(qnjj1444)
通信地址:571101 海南省海口市学院路 4 号 1 栋 102
Address: Room 102, Building 1, 4 Xueyuan Rd, Haikou 571101, Hainan, P. R. China

2 否定模式匹配

传统的入侵检测模式匹配方法是根据给定的模式集合,在待检网络流中找出相同的字符串子链,这种检测方法消耗的计算资源和时间较多。因此,本文结合传统的入侵检测模式匹配方法,基于否定模式 NP(Negative Pattern)检测方法来实现报文检测,以减少匹配的次数,降低成本,提高效率。

定义 1 假定模式集合 $P=\{V_1, V_2, \cdots, V_n\}$,否定模式 NP 是 P 的一个 NP,当且仅当 NP 的任何 $K(K>1)$字节后缀不是 V_i 的子集,其中 $V_i \in P, i=1,2,\cdots,n$。

假定模式集合 $P=\{\text{virus},\text{worm}\}$,根据否定模式定义,字符串模式"free"是 P 一个否定模式。因为"free"的任何 $K(K>1)$字节后缀集合 $T=\{\text{ee},\text{ree},\text{free}\}$,$P$ 的 $K(K>1)$ 字节子集 $S=\{\text{vi},\text{ir},\text{ru},\text{us},\text{wo},\text{or},\text{rm},\text{vir},\text{iru},\text{rus},\text{wor},\text{orm},\text{viru},\text{irus},\text{worm},\text{virus}\}$,有 $\forall x \in T$,则 $x \notin S$。

定理 1 假定模式 $NP=\{P_1, P_2, \cdots, P_n\}$是某待检测的数据流 F 的一个否定模式,则根据 P_i($P_i \in NP, i=1,2,\cdots,n$)的最后两个字节将 F 分段后生成的字符串模式不会横跨两个相邻的数据段。

证明 假设 NP 为待检数据流 F 的一个否定模式,$P_i \in NP, i=1,2,\cdots,n$。(1)当 $n=1$ 时,定理 1 显然成立;(2)当 $n>1$ 时,假设存在一个模式 $V_i \in F$(其中 $i=1,2,\cdots,k$)横跨两个相邻的数据段,那么 V_i 至少有两个字节长,也就是 V_i 的最后两个字节,所以 V_i 一定包含一个 NP 的 $K(K>1)$字节后缀,而根据否定模式定义,V_i 不能够包含 NP 的后缀,所以这样的 V_i 是不存在的,定理 1 成立。综合(1)和(2)可知,定理 1 成立。 □

假定 $S=\{\text{virus},\text{worm}\}$,某待检测的数据流 F 为"…using computervirustorefertoaworm…",通过否定模式定义可以得知{…ngc,erv, ust, ert, tow…}是 S 的 NP 模式,所以根据 NP 模式{…ngc,erv, ust, ert, tow…}中各元素最后两个字节将 F 分段后生成的字符串模式为"… using","computer","virus","toa","refer","to","worm…"。该数据流 F 采用 NP 模式分段后没有丢失特征,即所有的模式都可以检测出来。因此,解决问题的关键是在报文内容中精确地找到 NP,以确保正确分段和降低错误率。

3 基于 NP 模式的报文检测方法

3.1 基于模式匹配的内容检测机制的体系结构

图 1 是基于模式匹配的内容检测机制的体系结构。从图 1 中可以看出系统检测的主要流程,系统捕获的数据包通过"网络协议分析"把报文的头部和内容分离开,根据一定的规则把可能含有入侵信号的报文内容送到"报文内容检测"中,利用"模式匹配引擎"对报文内容做检测,即根据 NP 模式把待测数据流分段,然后再对段内容进行模式匹配,检测出可疑的数据包,以有效地提高入侵检测的工作效率。

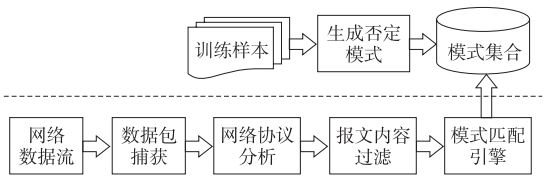


Figure 1 Content Detecting architecture based on pattern matching

图 1 基于模式匹配的内容检测机制的体系结构

3.2 报文内容检测中否定模式的数据结构

图 1 中系统送到"报文内容检测"中的是报文内容的字节流,通过"模式匹配引擎"基于 NP 模式对待测字节流进行分段和匹配,根据检测结果丢弃无入侵信号的数据包,将可疑的数据包传送到入侵检测的下一环节。

表 1 是关于报文内容检测机制的 NP 表数据结构。对于一个待匹配的模式集合,从相应的 NP 表可以确定是否组成 NP 模式。综合考虑匹配次数、时间、计算资源等各方面的因素,表 1 中否定模式的长度设为两个字节,实验结果也表明两个字节比较好。

Table 1 *NP* table of packet content filtering

表 1 报文内容过滤的 *NP* 表

字符串模式	NP 模式	字符串模式	NP 模式
…	…	…	…
er	否	us	否
es	是	…	…
…	…	yz	是
ev	是	…	…

3.3 基于 *NP* 匹配的报文内容检测的原理

"报文内容检测"主要是基于 NP 模式匹配结果来检测不含入侵信号的报文,以便排除正常流的干扰,提高报文检测效率。由于 NP 匹配是一个否定模式,而且找出所有的 NP 模式也许需要花

费很多的时间,所以不需要找出所有的 NP 模式。

查找 NP 可以从某一特定位置开始不停地查找直到找到 NP 位置,然后跳过检测窗口大小 w 字节再重复进行下一次查找。该查找方法需要解决两个主要问题:(1)循环次数不确定性和不可控性,尽管已经证明在实际例子中 NP 是经常出现的,但对一些特殊情况找到 NP 的概率仍然是不确定的;(2)关于内容检测的并行性问题。为了解决这两个问题,本文采用多层次 NP 匹配算法,利用递归方式查找 NP。即 $M(M<W$,M 由系统的并行度决定)等分数据包待测内容流,基于 NP 匹配结果,待测内容流将被分成 $M+1$ 段,如果有大于 $2w$ 字节的段则再分开,直到小于 $2w$ 字节,每次递归查找后向前跳一个字节,以避免重复检查同样的位置。多层次 NP 匹配算法描述如下所示:

算法 1 多层次 NP 匹配算法

```
NPSeek(byte *C, int iSL, int w, int iR, int N){
/*C: the buffer of content; C[i,k]: the k-bytes buffer starting from i; iSL: the length of the segment; w: the width of sliding windows; iR: the round of looking; N: the width of the NPs, in this prototype, N=2*/
if (iR>w-1 or iSL<2* w)
  EPM_Seek(C, iSL); //EPM the segment directly
for m from 1 to M parallel_do {
  if IsNP(C[iR+m*w-1,N])
    NPMatch[m]=TRUE;
}
lastNP=0;
for i from 1 to M do {
  if (NPMatch[i]==TRUE) { /* the following can be done in a new thread */
    NPSeek(C+lastNP*w, (i-lastNP)*w, w, iR+1, N);
    lastNP=i;
  }
}
Recursion_NP(byte *C, int Len, int w, int N){
/*C: buffer of content; C[i,k]: the k-bytes buffer starting from i; Len: length of content, i.e. Len=|C|; N: the width of the NPs; w: the width of the sliding window of the EPM engine*/
  NPSeek(C, Len, w, 0, N);
}
```

4 实验结果与分析

4.1 NP 查找匹配次数的评估

NP 匹配次数是由多层次 NP 匹配算法的递归深度决定的。设数据流的总长度为 L,检测窗口的大小为 w,递归深度为 D_r。假设每次查找都可以找到 NP,则需要 $\lceil L/w \rceil$ 次 NP 匹配。假设 NP 匹配的概率是 R_h,$0 \leqslant R_h < 1$,则 NP 查找次数的最大值为:

$$n_{NP} = L/w+(1-R_h)\times L/w+(1-R_h)^2\times L/w+\cdots+(1-R_h)^{D_r-1}\times L/w=\sum_{i=0}^{D_r-1}(1-R_h)^i\times L/w=(1-(1-R_h)^{D_r})/(1-(1-R_h))\times L/w=\{[1-(1-R_h)^{D_r}]\times L\}/(w\times R_h) \quad (1)$$

在最坏情况下 $R_h \to 0$,则有:

$$\lim_{R_h\to 0}\{[1-(1-R_h)^{D_r}]\times L\}/(w\times R_h)=L/w\times \lim_{R_h\to 0}[1-(1-R_h)^{D_r}]'/R_h'=L/w\times \lim_{R_h\to 0}((-1)\times(-1)\times D_r\times(1-R_h)^{D_r-1})/1=(L\times D_r)/w \quad (2)$$

4.2 检测次数分析

由 4.1 节可知,报文检测次数共计 $(n-1)(w-1)$,其中 n 为报文基于 NP 模式分割的段数,w 为窗口的大小,则有:

$$n = n_{NP}\times R_h = \{[1-(1-R_h)^{D_r}]\times L\}/(w\times R_h)\times R_h = [1-(1-R_h)^{D_r}]\times L/w \quad (3)$$

因此,降低的检查次数为:

$$n_{save} = \{[1-(1-R_h)^{D_r}]\times L/w-1\}(w-1) \quad (4)$$

当 $D_r = w-1(D_r < w)$,则次数的概率为:

$$R_{save} = n_{save}/(L-w+1) = \{[1-(1-R_h)^{w-1}]\times L/w-1\}(w-1)/(L-w+1) \quad (5)$$

若 $L \gg w$,$n \gg 1$,即有 $L-w+1 \approx L$,$[1-(1-R_h)^{w-1}]\times L/w-1 \approx [1-(1-R_h)^{w-1}]\times L/w$,则降低检测次数的概率为:

$$R_{save} \approx \{[1-(1-R_h)^{w-1}]\times L/w\}(w-1)/L = [1-(1-R_h)^{w-1}](w-1)/w \quad (6)$$

基于公式(6)有:

(1)明显有 $dR_{save}/dR_h \geqslant 0$,因此随着 NP 匹配概率(R_h)的增大,R_{save} 也增大。

(2)由 $d[1-(1-R_h)^{w-1}]/dw \geqslant 0$,$d[(w-1)/w]/dw \geqslant 0$,可得 $dR_{save}/dw \geqslant 0$,则 w 越大,R_{save} 也越大。

由图 2 可见,当检测窗口值越大,降低检测次数的概率的增长速度就越快,而且收敛速度越快;当 NP 匹配的概率小于 30%时,降低检测次数的概率的收敛速度由 R_h 决定。

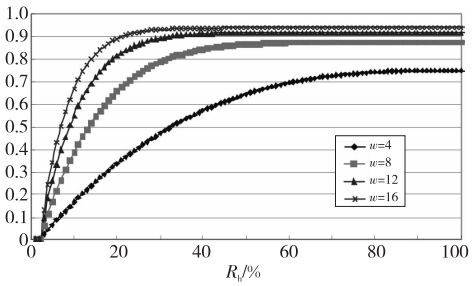


Figure 2 Linear relationship between the probability of *NP* matching and the size of the window

图 2 检测窗口的大小与 *NP* 匹配的概率关系

4.3 实验结果

本节实验所用的模式集合是来自 Snort 模式集合,该模式集合长度分布状态如图 4 所示,数据流集合是来自 MIT 的数据集[12]。本节将分析 4.2 节中 R_h、D_r、w 等各个参数的关系。经测试,Snort 模式集合有 62 647 个两字节 *NP*,因此 *NP* 匹配的概率为:$R_h = 62674/2^{16} = 95.63\%$。在处理的 1 376 598 个报文中,大约 52%报文含有内容(总长度是 242 576 387 B),所以平均报文流大小是 337 B。当 $w=4,8,12,16$,递归深度 D_r 为 $w-1$,则 R_{save} 分别约为 74.77%、87.23%、91.39%、93.46%。

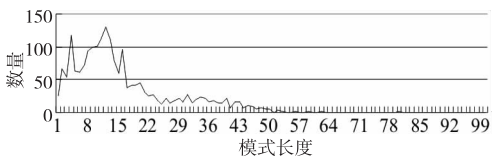


Figure 3 Length distribution state of Snort mode set

图 3 Snort 模式集合长度分布状态

5 结束语

大规模网络流量环境中检测报文内容需要消耗大量计算资源和时间。本文基于否定模式匹配的思想,提出了基于 NP 模式的报文检测方法。该方法根据 NP 模式把待测数据流分段,然后再对段内容进行模式匹配,根据匹配结果检测出可疑的数据报文。实验结果表明该方法能避免降低误报,减少检测次数,提高报文检测效率。

参考文献:

[1] Artan N S,Chao H J. TriBiCa:Trie bitmap content analyzer for high-speed network intrusion detection[C]// Proc of IEEE INFOCOM'07, 2007:125-133.

[2] Chang Y, Tsai M, Chung Y. Multi-character processor array for pattern matching in network intrusion detection system[C]// Proc of the 22nd International Conference on Advanced Information Networking and Applications, 2008:1.

[3] Artan N S, Chao H J. Design and analysis of a multipacket signature detection system[J]. International Journal of Security and Network, 2007,2(1):122-136.

[4] Boyer R S, Moore J S. A fast string searching algorithm[J]. Communications of the ACM, 1977,20(10):762-772.

[5] Dharmapurikar S, Lockwood J. Fast and scalable pattern matching for network intrusion detection systems[J]. IEEE Journal on Selected Areas in Communications, 2006, 24(10): 1781-1792.

[6] Ramaswamy R, Kencl L, Iannaccone G. Approximate fingerprinting to accelerate pattern matching[C]// Proc of the 6the ACM SIGCOMM Conference on Internet Measurement, 2006:1.

[7] Liu Yan-bing, Shao Yan, Wang Yong, et al. A multiple string matching algorithms for large-scale URL filtering[J]. Chinese Journal of Computer, 2014, 5: 1159-1169. (in Chinese)

[8] Chu Yan-jie, Li Yun-zhao, Wei Qiang. Swm: An improved multi-pattern matching algorithm and application[J]. Journal of Xidian University, 2014, 6: 197-203. (in Chinese)

[9] Song Tian, Li Dong-ni, Wang Dong-sheng, et al. Journal of software[J]. Memory Efficient Algorithm and Architecture for Multi-Pattern Matching, 2013, 7: 1650-1665. (in Chinese)

[10] Dharmapurikar S, Krishnamurthy P, Sproull T S, et al. Deep packet inspection using parallel bloom filters[J]. IEEE Micro, 2004, 24(1): 52-61.

[11] Lunteren J V. High-performance pattern-matching engine for intrusion detection[C]// Proc of IEEE INFOCOM'06, 2006:1.

[12] Yu F, Katz R H, Lakshman T V. Igabit rate packet pattern-matching using TCAM[C]// Proc of the 12th IEEE International Conference on Network Protocols, 2004:174-183.

[13] Juniper networks intrusion prevention solutions[EB/OL]. [2013-11-12]. http://www.juniper.net/products_and_services/intrusion_prevention_solutions/index.html.

[14] Fortinet-Muti-threat security systems for real time network protection[EB/OL]. [2013-12-25]. http://www.fortinet.com/.

[15] Snor intrusion detection system[EB/OL]. [2014-03-16]. http://www.snort.org.

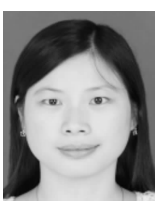
附中文参考文献:

[7] 刘燕兵,邵妍,王勇,等. 一种面向大规模 URL 过滤的多模式串匹配算法[J]. 计算机学报,2014,5:1159-1169.

[8] 褚衍杰,李云照,魏强. SWM:一种改进的多模式匹配算法及应用[J]. 西安电子科技大学学报,2014,6:197-203.

[9] 嵩天,李冬妮,汪东升,等. 存储有效的多模式匹配算法和体系结构[J]. 软件学报,2013,7:1650-1665.

作者简介:



唐湘滟(1981 -),女,湖南邵阳人,硕士,研究方向为网络安全和图书情报挖掘。**E-mail**:Tangxy36@163.com

TANG Xiang-yan, born in 1981, MS, her research interests include network security, library and information mining.