

新型内容过滤防火墙的研究

姚亚锋¹, 方贤进², 赛文莉³

(1. 南通职业大学 技师部, 江苏 南通 226007;

2. 安徽理工大学 计算机科学与工程学院, 安徽 淮南 232001;

3. 南通大学 医学研究中心, 江苏 南通 226001)

摘要:传统的防火墙不能充分快速地进行内容过滤,这直接影响了防火墙的功能。在模式匹配算法中,BM算法是一种针对单模式匹配的算法,AC算法是基于有限自动机的多模式匹配算法,人们在长期使用过程中感到这两种算法都存在弊端且效率较低。为此,结合两种经典的模式匹配算法,寻求一种高效率的模式匹配方法,并把它移植到防火墙的网络层的内容过滤模块中,优化后的AC-BM算法比以往的两种算法在复杂度和效率上都有明显的改进,而且在网络层进行这样的高效过滤,不仅增加了防火墙的内容过滤的充分性,而且提高了防火墙的效率。

关键词:新型防火墙;内容过滤;模式匹配;AC-BM算法

中图分类号:TP393.08

文献标识码:A

文章编号:1673-629X(2010)11-0158-04

Research of New Firewall for Content Filtering

YAO Ya-feng¹, FANG Xian-jin², SAI Wen-li³

(1. Technician Department of Nantong Vocational College, Nantong 226007, China;

2. Computer Department of Anhui Univ. of Science and Technology, Huainan 232001, China;

3. Medical Research Center of Nantong University, Nantong 226001, China)

Abstract: The traditional firewall cannot carry on content filtering fully and rapidly, which has reduced the function of firewall. In the pattern-matching algorithm, BM algorithm is for single pattern matching, and AC algorithm is for multi-pattern matching based on the finite automaton. But both of them have drawbacks because of their low efficiency. So combine two classical pattern-matching algorithms and transplant it to content filtering module of the new firewall in order to enhance its efficiency. The theoptimized AC-BM algorithm has the superiority compared to the former algorithms in complexity and efficiency. With this method the firewall not only increases function of content filtering, but also be more efficient.

Key words: new firewall; content filtering; pattern-matching; AC-BM algorithm

0 引言

随着网络技术的飞速发展和网络时代的到来,网络安全显得尤为重要。而作为抵御外来网络的威胁,防火墙技术仍为保障网络安全的一个重要手段,它作为一个门户,也作为一种屏障位于被保护的内部网络和外部非信任网络之间,选择性地允许和限制内网与外网之间的信息交互,既保障正常的因特网访问,也保护内部的重要数据不受破坏和越权访问^[1]。

目前,在防火墙的设计和研究过程中,不仅要注重防火墙的安全性,而且防火墙的效率显得尤为重要。

文中基于本防火墙的结构模型,提出了网络层的内容过滤防火墙概念,把模式匹配算法BM和AC算法运用到内容过滤防火墙过滤过程中,并通过分析把两种算法相结合推出AC-BM算法,用它来实现防火墙的内容过滤,进一步提高了防火墙的效率。

1 模式匹配算法在防火墙内容过滤的应用

防火墙过滤技术包括:包过滤防火墙、应用层网关防火墙、内容过滤防火墙。大多数包过滤防火墙在防火墙结构七层中的网络层工作,其基本原理是根据每个数据包头中的IP地址和TCP端口信息来检测这个数据包的内容,它只能对每个分包的包头信息内容进行检查,虽具有一定的安全性,但其缺点也是显而易见的,其始终无法过滤掉包内的不安全信息^[2]。而应用代理防火墙是工作在七层中的应用层,它的优势在于

收稿日期:2010-03-13;修回日期:2010-07-15

基金项目:安徽省自然科学研究项目(kj2007b242)

作者简介:姚亚锋(1978-),男,硕士研究生,助教,研究方向为算法分析与设计、计算机信息安全等;方贤进,博士研究生,副教授,研究方向为计算机信息安全、人工智能。

可以对整个数据包的内容进行简单过滤,但它必须根据不同的应用层协议来转化,使用效果和灵活性较差^[3]。

一个可靠的防火墙,在使用中安全性和高效性是两个必须考虑的参数。目前,对防火墙内容过滤的研究中模式匹配算法应用越来越广泛,其作用也越来越重要。整个模式匹配算法总体分为两大类:单模式串匹配算法和多模式串匹配算法^[4],常见的单模式串匹配算法主要有 BF 算法、BM 算法、KMP 算法等,多模式串匹配算法主要有 Wu - Manber 算法、AC(Aho - Corasick)算法和 DFSA 算法等。文中主要研究经典的单模式 BM 算法和多模式 AC 算法,并结合两种算法的优点,在此基础上推出一个新的快速的字符串多模式匹配算法——AC - BM 算法。目的在于把它运用到防火墙内容过滤模型中,提高防火墙的功能和效率。

2 经典的模式匹配算法及移植到防火墙中的弊端

2.1 单模式匹配 BM 算法

单模式匹配 BM 算法^[5]规则是把模式串 $P[0, 1, 2, \dots, m-1]$ 自左向右移动,而具体的模式串与主串之间字符的匹配从右向左进行,具体来说就是首先考察文本 $T[j+m-1]$ 字符与模式 $P[m-1]$ 字符是否一致,如一致则接着考察下一组字符: $T[j+m-2]$ 和 $P[m-2]$, 如果相等依次考察下去,直到 $P[0]$ 等于 $T[j]$, 这时就得到结论:模式串在文本中找到一次匹配;如果在比较的过程中,发现模式字符 $P[i]$ 和文本字符 $T[j+i]$ 不一致,显然比较工作不需要进行下去,此次匹配失败,但在此之前获得的成功匹配依然有效,也就是文本中 $T[j+i+1 \dots j+m-1] = P[i+1 \dots m-1]$, 充分利用这个等式关系可以从很大程度上避免再次匹配过程中重复劳动。

利用这个已知匹配,解决下一趟匹配将从哪个字符开始,也就是说模式串下一步应滑到什么位置,首先需要定义两个偏移量的偏移函数 Badchar 和 Goodsuffix。

(1) Badchar - 坏字符规则。

当 $P[i] \neq T[j]$ 时,右移模式串 P , 模式串右移的位置可以通过函数 delta1 表示。

$\text{delta1}(x) = -m; x < > P[j](1 \leq j \leq m)$, 即 x 在模式中未出现

$\text{delta1}(x) = -m - \max\{k \mid P[k] = x, 1 \leq k \leq m\}$; 其它情况

(2) Good Suffix - 好后缀规则。

好后缀规则是为了保证每个正确匹配不丢失而产生一个距离移动函数 delta2 , 它的大小与模式串有很大的关系。由前面的范例已知,在模式串 $P[j-s+1 \dots m-s] = P[j+1 \dots m]$ 的情况下,下一次匹配可让模式 P 右移 s 位。 S 在 delta2 中的定义如下:

$\text{delta2}(j) = \{\{s \mid P[j+1 \dots m] = P[j-s+1 \dots m-s]\} \& \& (P[j] \neq P[j-s])(j > s)\}$

不难发现在实际的匹配过程中,总是比较 delta1 函数的值和 delta2 函数的值,选取其中较大的一个。

2.2 多模式匹配 Aho - Corasick 算法

Aho - Corasick 算法^[6](简称 AC 算法)是多模式匹配算法中最典型的一种,它能同时对多个模式串进行检查匹配,故称为多模式匹配算法。具体做法是,把所有的模式串归并为一个集合(整体),根据有限自动机的理论同时匹配集合中的所有模式,构造有限自动机使得每个状态表示模式的一个前缀,当 T 的下一个字符不等于 P 期待接受的下一个字符时,该算法就会跳到该模式的最长前缀所代表的那个状态继续进行,由前面得知也是当前状态的适合的后缀。

2.3 效率分析与缺点

时间效率和空间效率:BM 算法时间复杂度为 $O(n * m)$, 空间复杂度为 $O(m + \sigma)$, σ 是与 P, T 相关的有限字符的长度。AC 算法的时间复杂度为 $O(M + n)$, M 为所有模式串的长度总和,空间复杂度为 $O(m)$ 。

缺点与不足:BM 算法是针对单个模式进行匹配的方法,由于防火墙需要处理大量的规则,把 BM 算法用于防火墙内容过滤过程中,显然效率很低。AC 算法虽是多模式匹配算法的经典,但它没有有效地利用坏字符和好后缀两个规则,把它应用到吞吐量大的防火墙中可见效率也不高。

3 改进的多模式匹配算法——AC - BM 算法的主要思想

为了提高效率,把 AC 算法和 BM 算法的基本思想有机地结合起来,提出了高效的多模式匹配算法,简称 AC - BM 算法^[7]。

AC - BM 算法是在有限自动机的多模式匹配算法^[8]的基础上,以连续跳跃的思想改进的新算法。算法将所有不同的规则放在一棵被称为模式树的树形结构上,根据树的数据结构,用单模式 BM 算法进行匹配检索。模式数从右边向左移动,只要模式树确定在适当的位置,字符之间的比较则从左向右进行。表面上看此算法与 BM 算法没有区别,但值得注意的是 BM 算法只能单个模式串与文本进行匹配,而 AC - BM 算法则可以实现模式树上多个串同时与文本匹配。另

外,可以充分利用 Badchar 和 Goodsuffix 这两个移动函数^[9]:

(1)首先进行初次匹配时模式树的移动方向是自右向左。而字符之间的比较是从左向右(从模式树的根结点到叶节点)。AC-BM 算法通过利用 Goodsuffix 和 Badchar 两种移动函数来进行匹配模式,不同的是 BM 算法只能单个模式串与文本进行匹配,而 AC-BM 算法则可以实现模式树上多个串同时与文本匹配,这样明显提高了效率。

(2)在一次匹配失效时,Goodsuffix 可以充分地利用部分成功匹配结果,反映在模式树的子树上或是子树的前缀,这时该把模式树滑向下一个匹配。假设有四个模式串 P: tiro, tired, tiade, tomato, tornadic。文本 T 为: automatone。过程如下,首先把四个模式串生成模式树,已知最短模式长度(tiro)是 4,就把模式树移动到文本最后倒数第四个字符位置(有下划线的 t),然后字符之间比较从左到右,直到比较到文本 T 的 n 处发现一次匹配失败,但此时发现字符串“to”在模式数上再次出现,由于“to”已经成功匹配过,所以接下来模式树应向左移到文本 T 下一个“to”的地方(见图 1)。

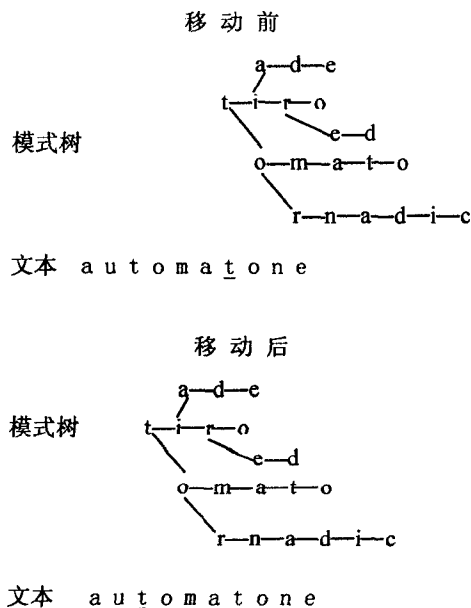


图 1 Goodsuffix 示意图

(3)如果发生匹配失效时,根据 Badchar 函数,模式树会滑到下一个与当前字母相同的位置来与文本继续匹配。倘若没有这样的字母,模式树把模式树中最短的子树长度当成本次移动的长度。假设有四个模式串 P: time, tired, tiring, tinted, tinsel。文本 T 为: timeisonmyside。具体步骤如下,同理在文本倒数四位对齐模式树(有下划线的 s),自左往右匹配模式数和文本的字符,显然在文本 T 一开始的“s”处匹配就失

败,但此时“s”这个字母在模式树的子树“tinsel”中存在,所以接下来模式树应左移使得模式树中的“s”与文本中的“s”对齐(见图 2)。

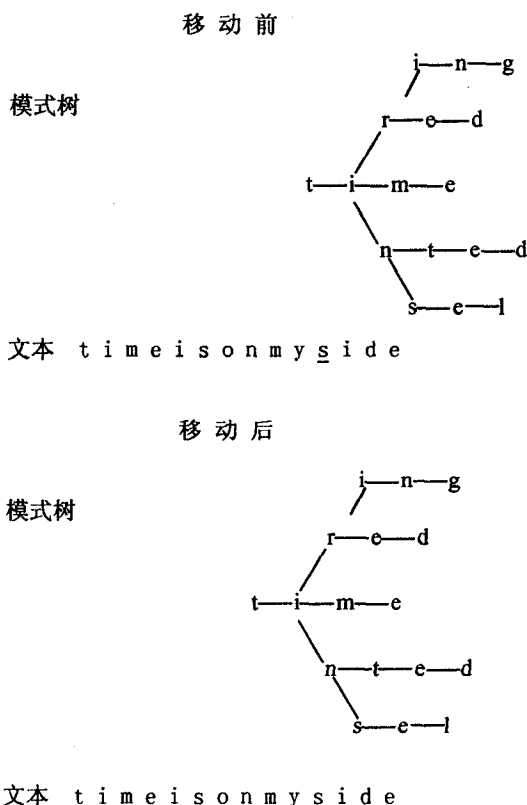


图 2 Badchar 示意图

(4)按照这样的方法反复进行上述过程,直至找到一次完全匹配成功或整个匹配结束,然后把最终的结果返回。

AC-BM 算法效率^[10]: 设模式串集合中,模式串最小长度为 minlen,最大长度为 maxlen,待匹配正文长度为 n , 则在最优情况下,时间复杂度为 $O(n/\text{minlen})$,在最坏情况下,时间复杂度为 $O(n * \text{maxlen})$ 。AC-BM 算法结合两种算法的优点,可以很好地利用到防火墙的内容过滤过程中。

4 算法对比与实验结果

比较 BM、AC 和 AC-BM 三种算法的模式匹配效率;测试环境为 CPU Intel P4, 2.0G, 内存 256M, 操作系统为 RedHat Linux9.0, 平台为 GNU C/C++。多个模式串采用有限自动机进行存储。数据采用 MIT Lincoln Lab 的 1998 DARPA Intrusion Detection Evaluation Data Set^[11], 数据测试见表 1。

通过使用检测模式串(假想数据包)对 AC 算法、BM 算法和 AC-BM 算法进行测试,基本以 200 条为增幅^[12],从表中可知,AC-BM 其速度是 BM 算法的 4

倍左右,是 AC 算法的 1.6 倍左右,而且 AC-BM 算法在模式串越多的情况下速度越快。这说明 AC-BM 算法完全适应较大的规则库和数据包,这一点正好满足防火墙的内容过滤功能。

表 1 三种模式匹配算法效率对比

模式个数 算法(耗时)	1	100	200	400	600	786	892	1000
BM(秒)	1.29	7.78	13.22	18.58	22.17	25.45	27.23	29.65
AC(秒)	×	10.31	10.46	10.50	11.01	11.12	11.18	11.20
AC-BM(秒)	×	6.38	6.89	7.03	7.14	7.18	7.21	7.22

5 AC-BM 在防火墙内容过滤中的应用

当今的网络安全十分重要,传统的防火墙还是以包过滤为主要手段^[13],但包过滤毕竟有它的弊端。而应用代理防火墙能够在应用层简单地检测数据包中的内容,但是使用 and 实现起来很麻烦,这就是此类防火墙的性能瓶颈。怎样使防火墙高效的工作,是研究的重点。数据包内容检测的关键技术就是前面提到的一些模式匹配算法^[14],因此希望以一种高效的模式匹配算法对数据包中的内容进行过滤检测,来达到有效内容过滤的目的。

而多模式匹配 AC-BM 算法正是一种可以满足上述条件的高效的算法。如果把它作为一种核心算法集成到防火墙内容过滤模块中,正好可以满足大量的防火墙规则有待于匹配的特点。以上所述新型防火墙结构框架如图 3 所示。首先在网络层把每个数据分包的包头里的内容提取出来,倘若包头里的 IP 是无效的,就把这个数据包直接过滤掉,这样就避免了重复过滤,从而提高了性能。

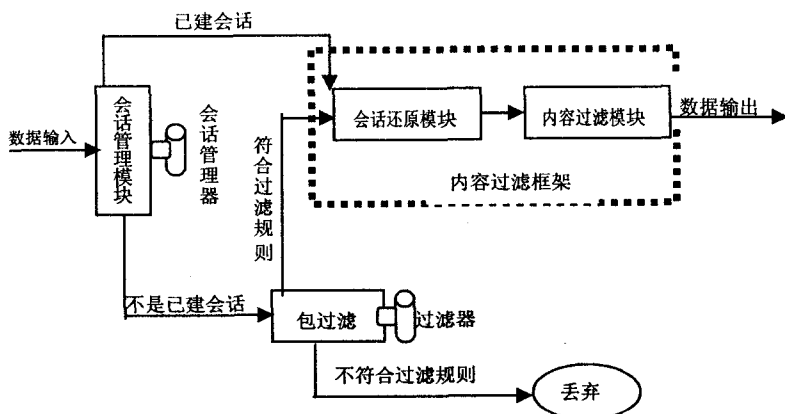


图 3 网络层内容过滤结构图

对于那些能够通过网络层过滤的数据包,如图先通过会话还原模块把整个会话进行还原,再用 AC-BM 算法对每个完整的会话进行内容扫描,若检测到非法内容就清除该数据包,否则就按照目的地的地址转发^[15],避免了不同协议之间的来回转换的弊端。与

传统防火墙对比,文中设计的防火墙体系具有内容过滤全面、速度快、设计合理和可靠性高等优点。

6 结束语

综上所述,通过对各种模式匹配算法的分析和研究,发现 AC-BM 算法结合了多种算法的优点具有高效率,为此把此算法集成到新型内容过滤模块中,作为网络层内容过滤的引擎,这样防火墙内容过滤将更加全面,而且过滤的效率也有很大提高。当然对于文中提出的基于内容过滤防火墙的研究才刚起步,细节方面亟待于进一步探索。

参考文献:

- [1] Stevens W R. TCP-IP 详解 I 协议[M]. 范建华等,译. 北京:机械工业出版社, 2004.
- [2] Goncalves M. 防火墙技术指南[M]. 宋书民,译. 北京:机械工业出版社, 2000.
- [3] Cormen T H, Ronald C L, Stein L R C. 算法导论(影印版)[M]. 第 2 版. 北京:高等教育出版社, 2003.
- [4] Aho A, Corasick M. Efficient String Matching: An Aid to Bibliographic Search[J]. Communications of the ACM, 1975, 18(6): 333-340.
- [5] 李 响, 李伟华. 面向入侵检测的模式匹配算法研究[J]. 计算机工程与应用, 2003(6): 1-2.
- [6] Fan J, Su K. An efficient algorithm for matching multiple patterns[J]. IEEE Transaction on Knowledge and Data Engineering, 1993, 5(2): 339-351.
- [7] Exact String Matching Algorithms[EB/OL]. 2002-07. <http://www-igm.univ-mtl.fr/~lecroq/string/index.html>.
- [8] Towards Faster String Matching for Intrusion Detection[EB/OL]. 2001-03-16. www.ilib.cn/A-jsjyyrj200103016.html.
- [9] 李长松. 具有入侵检测功能的防火墙系统的设计与实现[D]. 成都:电子科技大学, 2003.
- [10] 武舒凡, 胡建武. 防火墙包过滤技术发展研究[J]. 计算机应用研究, 2004(9): 144-146.
- [11] 席荣荣. 基于内容过滤的防火墙的关键技术的研究[D]. 太原:山西大学, 2004.
- [12] 张 娜. 内容过滤防火墙的设计与实现[D]. 合肥:合肥工业大学, 2006.
- [13] 方贤进, 李龙澍. 多模式匹配算法的优化研究[J]. 微计算机信息, 2007, 23(9): 211-213.
- [14] 方贤进, 李龙澍. 基于主观 Bayes 方法对 Web 使用挖掘的研究[J]. 计算机科学与技术, 2007(6): 56-59.
- [15] 方贤进, 李敬兆, 姚亚锋, 等. 一种校园网的网络安全策略[J]. 计算机科学与技术, 2006(5): 121-124.