

面向存储的正则表达式匹配算法综述

姚 远,刘 鹏,单 征,田双鹏

(信息工程大学 信息工程学院, 郑州 450002)

(liupeng81@tom.com)

摘 要:正则表达式匹配是当前深度包检测领域中的关键性技术。介绍了面向存储的正则表达式匹配算法的基本思想和设计方法,给出了算法分类并比较了典型压缩算法间的差异,分析了正则表达式语法对算法设计的影响,最后论述了目前研究中面临的技术难点并对今后算法设计的发展趋势作了展望。

关键词:确定的有限自动机;面向存储;正则表达式;深度包检测;压缩算法

中图分类号: TP393 **文献标志码:** A

Survey on storage-oriented regular expressions matching algorithms

YAO Yuan, LIU Peng, SHAN Zheng, TIAN Shuang-peng

(Institute of Information Engineering, Information Engineering University, Zhengzhou Henan 450002, China)

Abstract: Regular expression matching is a key technology in current deep packet inspection. Basic ideas and methodology for storage-oriented regular expressions were introduced. The paper discussed the classification of algorithms, compared the main algorithms, and found out many factors of regular expression syntax that have influence on the algorithms. Finally, the key techniques and the difficulty were explored, and the study made suggestions on the algorithms design in future.

Key words: Deterministic Finite Automaton (DFA); storage-oriented; regular expression; deep packet inspection; compressed algorithm

0 引言

随着计算机网络应用的日益普遍,网络服务的多样化日益显著,IP 报文的内容检测对于网络安全和网络监控的应用变得至关重要。由于正则表达式具有较强的表达能力,能够描述更为广泛的负载特征,因此,正则表达式匹配替代精确字符串匹配成为深度包检测的主要手段。目前,正则表达式匹配的算法研究已成为当前深度包检测领域的一个研究热点。许多典型的网络安全应用如 Snort^[1]、Bro^[2]、Linux L7 Filter^[3]、Cisco 家族网络安全应用^[4]和 Citrix 应用防火墙^[5],其规则集普遍采用正则表达式。上述网络应用所采用的匹配算法还停留在传统的 DFA 和 NFA 上,算法设计缺乏可伸缩性,远远不能满足对于当前主干网带宽的线速处理能力。对正则表达式匹配算法的深入研究在网络监控、防火墙、P2P 流量识别、内容审计、应用层协议识别等方面都有很重要的实用价值。当前这一领域的算法研究还停留在实验室阶段,本文探讨近几年来研究人员在匹配算法方面所做的努力,主要包括各种算法的比较,语法特征对匹配算法设计的影响及面临的主要问题和前景的展望。

1 现有压缩算法分类

正则表达式是指通过某种模式表示一类字符串的公式,Friedl^[6]对正则表达式进行了深入的讲解。Becchi 等人指出当前的正则表达式匹配包括两种解决方案:FPGA 逻辑^[7-10]和面向存储的体系架构,其实现算法分别基于 NFA 和 DFA。

FPGA 逻辑的优点是并行性和可重构性,然而,FPGA 逻辑具有以下三点不足:其时钟频率低于通用处理器或 ASIC;缺乏规则更新的实时性;缺乏可伸缩性。面向存储的体系架构可以通过灵活的设计克服以上不足。由于 DFA 处理速度快,对于每个输入字符仅进行一次状态转换,大多数算法研究都在面向存储的体系架构下展开。这里的面向存储是指对于 DFA 状态数和 DFA 转换数,通过高速存储设备进行集中存储。其实现以内存为中心,包括通用处理器、网络处理器、FPGA/ASIC。

算法研究主要在 DFA 原型的基础上进行自动机设计。由于 DFA 占用过量的内存空间,主要算法都集中在如何压缩 DFA 空间。目前,国内外研究刚刚起步,并没有分类标准,根据对现有算法的研究,按照压缩方式可以分为以下三类:转换压缩、状态压缩和字母表压缩。转换压缩指对自动机中的转移函数 δ 边进行压缩,状态压缩是指对有穷状态集 Q 进行压缩,字母表压缩是指对有穷的输入字符表 Σ 进行压缩。一般而言,以上三类算法并没有内在联系,实际应用中可以同时采用多种压缩算法。表 1 列出了几种典型的压缩算法及其应用规模、处理速度、存储空间的比较。

在转换压缩中,主要思想是研究自动机边与边之间的冗余,通过消除冗余,以减少存储空间。通过 Kumar、Becchi、Ficara 等人的观察,这些冗余可以归纳为以下三点:

- 1) 许多状态接受同样的字符会转移到相同的目的状态;
- 2) 所有的正则表达式开始于一个单一的初始状态,大量状态转换会回到初始状态或与初始状态相邻的状态;

收稿日期:2009-06-17;修回日期:2009-08-04。 基金项目:国家 863 计划项目(2006AA01Z408)。

作者简介:姚远(1972-),男,湖北武汉人,副教授,博士研究生,主要研究方向:网络安全、并行计算; 刘鹏(1981-),男,河南开封人,硕士研究生,主要研究方向:网络安全; 单征(1977-),辽宁沈阳人,讲师,博士,CCF 高级会员,主要研究方向:网络安全; 田双鹏(1981-),男,山西忻州人,硕士研究生,主要研究方向:网络安全。

3) 大多数相邻的状态间具有大量相同的状态转换, 可以仅记录不同的状态转换而大量减少内存空间。

表 1 典型压缩算法比较

算法分类	典型算法	提出年份	应用规模	处理速度	存储空间
转换压缩	D ² FA	2006	大	中	中
	CD ² FA	2006	大	中	小
	Improved-D ² FA	2007	大	快	小
	δFA	2008	大	快	大
状态压缩	XFA	2008	中	中	中
	State-Merging	2007	中	中	中
	Multi-DFA	2006	大	快	小
	History-based FA	2007	中	中	中
	Hybrid-FA	2007	中	慢	中
字母表压缩	Counting-FA	2008	小	快	小
	Single-ACT	2007	大	中	中
	Multi-ACT	2008	中	慢	小

几种算法的提出都源于上述部分观察。Kumar 等人在 1) 的基础上提出了 D²FA^[11], 算法中引入了缺省转换, 其引入过程如下: 假设有两个状态 s_1 和 s_2 , 对于输入字符集 $\{C\}$ 会转换到相同的状态集 $\{S\}$ 。那么可以消除其中一种状态的状态转换, 如果消除 s_1 的状态转换, 那么就引入了一个从 s_1 到 s_2 的缺省转换。为了避免回路, DFA 向 D²FA 的转化也就是求最大生成树问题, 边的权值为相同的状态转换数。引入缺省转换后, 为了在存储空间和内存带宽上进行折中, 带来了如何限定缺省转换路径长度的问题。CD²FA^[12]、Improved-D²FA^[13]、δFA^[14] 三种算法在 D²FA 算法的基础上设计, 处理速度和存储空间都有不同程度的改进。相同测试环境下, D²FA 与 DFA 相比, 转换数所占内存空间不足 DFA 的 1/20; CD²FA 内存需求为 D²FA 的 1/10; Improved-D²FA 与 D²FA 相比, 转换数不足 D²FA 的 1/10, 与 CD²FA 相比, 内存需求为 CD²FA 的 1/10 到 1/2。

在状态压缩中, 算法设计的思想包括以下三种: 引入辅助变量、模式分组、自动机结合。由于存在 DFA 最小化算法, 对最小化的 DFA 进一步采取状态压缩不可能与原自动机等价, 因此, 需要引入辅助变量。XFA^[15-16] 通过引入辅助变量, 将 DFA 扩展为 7 元组, $M = \{Q, D, \Sigma, \delta, U_\delta, (q_0, d_0), F\}$, 其中 D 为有限的数值集合, U_δ 为转移函数 $Q \times \Sigma \times D \rightarrow Q \times D$, 自动机的其他元组根据 D 的引入作相应调整。State-Merging^[17]、History-based FA^[18]、Counting-FA^[19] 均不同程度地引入了辅助变量。几种引入辅助变量的算法在应用规模, 针对的模式

特征上不尽相同, 算法设计的差异也体现于此。

2009 年 2 月的 snortrules-snapshot-2.8 规则集包含近 2 万条正则表达式语法, 如果将这些规则合成单一的 DFA, 内存空间根本无法接受。Yu 等人提出的 Multi-DFA^[20] 采用了模式分组的思想, 避免了合成单一 DFA 引起的状态爆炸问题。模式分组不仅限于 DFA, 也同样适用于其他自动机。Hybrid-FA^[21] 结合了 NFA 和 DFA 的优点, 由一个 DFA 头和多个 NFA 尾表示, 内存带宽接近 DFA, 内存空间接近 NFA。然而, NFA 尾的存在使得 Hybrid-FA 的性能并不稳定, 无法应对带有恶意特征的 DDoS 攻击。

字母表压缩中, 2007 年 Becchi 等人提出了 Single-ACT^[13], 适用于从不同的源状态接收输入字符到达相同目的的状态的情况。2008 年 Kong 等人提出的 Multi-ACT^[22] 思想与单 ACT 类似, 匹配过程中, 不仅需要查找当前的状态和当前的输入字符, 也需要识别当前的 ACT。Single-ACT 所占用内存空间依据实际规则集变化较大, Multi-ACT 在 Single-ACT 的基础上进一步压缩了存储空间, 内存节省为 1/70 到 1/4, 同时运行时间增加 35% 到 85%。

2 语法特征对算法设计的影响

在实际的网络应用中, 新的规则特征不断出现, 用以表示新的规则特征的正则表达式语法越来越复杂。当前, Snort 等入侵检测系统规则集的正则表达式语法大多采用兼容 perl 的正则表达式 (Perl-Compatible Regular Expressions, PCRE)^[23]。PCRE 语法的复杂性对算法设计影响主要包括两个方面: DFA 状态爆炸使得内存需求无法满足, 传统 DFA 无法处理部分特殊的元字符。

DFA 状态爆炸是由正则表达式间的相关引起的, Yu 等人在 Multi-DFA 算法中给出了相关的概念, 其定义为: 如果两个正则表达式合成一个 DFA 后的状态数大于分别构造 DFA 的状态数之和, 就称这两个正则表达式相关。2006 年 Yu 等人首次提出了状态爆炸问题^[20], 并分析了计数字符组 (character-class) 与前缀交叠引起状态爆炸的两种情况, 重写规则 1 和重写规则 2 可以对两种状态爆炸情况分别处理^[20]。但是, 我们的前期研究发现重写规则 2 具有一定的局限性, 不能完全处理文献[20]中所描述的第二类特征, 对于从任意位置开始, 计数字符组长度为 j , 与前缀交迭, 存在后缀的模式无法处理。2008 年 Becchi 等人^[19]和 Smith 等人^[15-16]都对 * 合成引起的状态爆炸进行了分析。对文献[20]中描述的第二类特征修改后, 引起状态爆炸的模式特征可以概括为表 2。

表 2 引起状态爆炸的四类典型模式

模式特征	实例	空间复杂度	典型算法	备注
以“^”开始, 字符组长度 j , 与前缀交迭	$\wedge A + [A - Z] \{j\} D$	$O(k + j^2)$	重写规则 1	k 为模式长度
任意位置开始, 字符组长度 j , 与前缀交迭, 无后缀	$. * AB. \{j\}$	$O(k + 2^j)$	重写规则 2	k 为模式长度
任意位置开始, 字符组长度 j , 与前缀交迭, 有后缀	$. * AB. \{j\} CD$	$O(k + 2^j)$	无	k 为模式长度
* 合成引起的状态爆炸	$. * AB. * CD$ 与 $. * EF. * GH$	$O(nl2^n)$	XFA	n 为模式数, l 为被 * 分割的字符串长度

由于网络应用的多样化, 各种新的规则特征不断出现, 使得正则表达式匹配在实现上更加困难, 部分语法特征使得算法设计难度增大, 需要单独设计自动机进行处理。Counting-FA 和 XFA 都能处理计数字符组, 反向引用字符出现后,

Extend-FA^[19] 可以对其进行有效处理。2009 年 2 月的规则集出现了 468 个肯定顺序环视和 32 个否定顺序环视, 这些字符同属环视字符。Friedl 等人^[6]明确指出传统 DFA 不能处理环视字符, 而且当前也没有提出针对环视字符的算法。对这些

特殊的语法特征可以归纳为表 3。

表 3 三类型 PCRE 语法及处理方法

语法特征	实例	典型算法
计数字符组	<code>.*[A-Z].{n}</code>	XFA Counting-DFA
反向引用字符	<code>.*(abc bcd).\1y</code>	Extended-FA
环视字符	<code>.*Jeff(?=s\b)</code>	无

该研究方向的进一步工作,可以针对表 2 中的模式 3 和表 3 中的环视字符设计自动机,提出新的算法。对于从任意位置开始,字符组长度为 j ,与前缀交互,有后缀的模式可以借鉴位图的思想。而针对环视字符,则可以在回溯自动机^[24]和双向自动机^[25]的基础上进行设计。

3 技术难点及展望

算法设计的技术难点主要来自两个方面:不单要考虑正则语法的复杂性,同时要考虑网络环境的复杂性。存在以下一些难点问题。

1) 算法设计的一般性。当前虽然存在一些算法能够对部分复杂语法进行处理(例如:Counting-FA、extended-FA),然而,这些自动机针对性较强,只能对特定的语法特征进行处理,算法设计不具备一般性。

2) 适应各种规模大小的规则集。不同于 FPGA 逻辑下的算法设计,面向内存架构由于采用高速存储设备保存 DFA 状态和转换,算法设计应具备可伸缩性,适用于不同规模大小的规则集。

3) 实时规则更新。规则的添加、删除需要对自动机进行重新编译,如果规则集数量过大,那么重新编译的时间必然无法满足实时性的要求。规则更新时,尽可能地缩小自动机的重新编译时间能够有效保证接收数据的完整性。

4) 内存有效。由于状态爆炸的普遍存在,已有算法并没有根本解决这个问题。同时,对于适当规模的规则集,可以采用片上内存,提高内存读写效率。

5) 线速处理能力。正则表达式匹配算法真正要应用于实际网络环境,对网络中的数据流需要线速处理。面向存储的体系架构下实际测试的研究工作^[26-27]较少。在对自动机进行压缩的同时,需要充分考虑其处理时间,占用的内存带宽,要能满足实际网络需要,达到数吉比特的线速处理能力。

6) 并行处理能力。对规则进行分组,并在多核上进行任务映射,可以有效提高算法性能。

正则表达式匹配从 2006 年以来开始逐渐成为研究的一个热点,研究刚刚起步,具有广阔的发展前景,其未来研究的发展趋势包括以下几点:

1) 简单规则集到复杂规则集研究的转变。当前的算法研究主要停留在实验室阶段,进行实验评估的规则集往往采用大量简单的规则集或 10 到 20 条复杂的规则集,针对大量复杂规则集的算法设计和实验评估将是今后的研究方向。

2) 面对新的规则特征的算法设计。从当前的研究来看,大多数算法研究主要集中在减小 DFA 状态数方面,包括预编码策略^[28]、规则改写策略、有限自动机的改进和提出新的自动机算法解决状态爆炸等问题。然而,针对网络中复杂的恶意攻击流特征和病毒特征,新的规则集不断提出,实际的规则集中仍然存在已有方法无法处理的规则,对新的规则特征提出新的算法成为当前算法研究的主要发展趋势。

3) 专用算法到通用算法设计的转变。在已有算法中,所设计的自动机大多仅能处理很局限的一类模式,由于各种复

杂模式的存在,不可能针对每种模式设计一种单独的自动机,新的自动机设计将向着通用性发展。

4) 单包匹配到流匹配的转变。当前的自动机算法设计大多对单一数据包进行匹配,对数据流进行匹配的研究较少。将来的算法设计,需能实现对乱序数据流的匹配。

4 结语

面向存储的正则表达式匹配算法研究是一个具有重要理论研究价值和广泛应用背景的课题。近年来,这一课题的研究尚处于起步阶段,许多关键技术难点亟待解决。文中结合现有研究,对典型的面向存储的正则表达式匹配算法进行分类和比较,并对研究的难点进行了一些简要的阐述并展望了未来的研究方向,期望能推动国内对深度包检测领域这一新的研究热点的关注和研究。对该课题开展跟踪研究不仅能够满足用户高质量、全方位网络服务的要求,而且能够进一步丰富自动机理论,对网络应用的开发和设计人员具有很重要的参考价值,同时能够应用于 DNA 序列检测等其他研究领域。

参考文献:

- [1] Introduction to Snort [EB/OL]. [2008-02-27]. <http://www.snort.org/docs/>.
- [2] Bro intrusion detection system [EB/OL]. [2009-02-24]. <http://www.bro-ids.org/>.
- [3] Application layer packet classifier for Linux [EB/OL]. [2009-01-07]. <http://l7-filter.sourceforge.net/>.
- [4] Cisco systems. Cisco ASA 5505 adaptive security appliance [EB/OL]. [2008-09-16]. http://www.cisco.com/web/HK/solution_tc/smb/products/security/asa_5500_series_adaptive_security_appliances.html.
- [5] Citrix systems. Citrix application firewall [EB/OL]. [2008-07-09]. <http://www.citrix.com/English/ps2/products/product.asp?contentID=25636>.
- [6] FRIEDL J E F. Mastering regular expressions [M]. 3rd ed. [S. l.]: O'Reilly, 2006.
- [7] MITRA A, NAJJAR W, BHUYAN L. Compiling PCRE to FPGA for accelerating SNORT IDS [C]// Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and Communications Systems. New York: ACM Press, 2007: 127-136.
- [8] BRODIE B, CYTRON R K, TAYLOR D E. A scalable architecture for high-throughput regular-expression pattern matching [C]// ISCA '06: Proceedings of the 33rd Annual International Symposium on Computer Architecture. New York: ACM Press, 2006: 191-202.
- [9] LEE J, HWANG S H, PARK N. A high performance NIDS using FPGA-based regular expression matching [C]// Proceedings of the 2007 ACM Symposium on Applied Computing. New York: ACM Press, 2007: 1187-1191.
- [10] LO C-T D, TAI Y-G, PSARRIS K. Hardware implementation for network intrusion detection rules with regular expression support [C]// Proceedings of the 2008 ACM Symposium on Applied Computing. New York: ACM Press, 2008: 1535-1539.
- [11] KUMAR S, DHARMAPURIKAR S, YU F, et al. Algorithms to accelerate multiple regular expressions matching for deep packet inspection [C]// Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York: ACM Press, 2006: 339-350.
- [12] KUMAR S, TURNER J, WILLIAMS J. Advanced algorithms for fast and scalable deep packet inspection [C]// Proceedings of the 2006 ACM/IEEE Symposium on Architecture for Networking and Communications Systems. New York: ACM Press, 2006: 81-92.

(下转第 3177 页)

过先确定 16 个链接变量的值,再反过来确定相对应的 16 个明文字的值,是个逆推的过程,其特点是 16 个明文字一步修改到位,确定性强。

多消息修改技术主要应用于 MD5 算法第 16 步以后的运算,其中包含单消息修改技术。在应用中,首先要确保前面步骤已经满足的充分条件在之后的消息修改中不被间接修改;其次是确保逆推的过程和正推的过程在某一步骤中可以顺利衔接起来;最后,尽可能限制计算机搜索范围,有效降低计算复杂度。其特点是必须借助于计算机程序自动搜索功能,具有一定的随机性。

4 结语

本文介绍了 MD5 破译的整体思想和关键技术,并介绍了明文差分的引入原则,作为产生碰撞的第一步,对整个算法的破译起到关键性的作用;分析了 MD5 算法迭代中差分路径的控制方法,利用非线性函数的性质和差分的扩展来引导中间变量差分的继承、产生和消除,最后介绍了单消息修改技术和多消息修改技术在破译 MD5 中的实现方法。

参考文献:

- [1] RIVEST R. The MD4 message digest algorithm [C]// CRYPTO '90: Advances in Cryptology. Berlin: Springer-Verlag, 1991: 303–311.
 - [2] RIVEST R L. RFC 1320, the MD5 message-digest algorithm[S]. MIT Laboratory for Computer Science and RSA Data Security, 1992.
 - [3] ZHENG Y, PIEPRZYK J. HAVAL — A one-way hashing algorithm with variable length of output [C]// AUSCRYPT '92: Advances in Cryptology, LNCS 718. Berlin: Springer-Verlag, 1993: 83–104.
 - [4] PRENEEL B. Integrity primitives for secure information systems [M]. Berlin: Springer-Verlag, 1995: 116–125.
 - [5] FIPS 180-1, secure Hash standard[S]. NIST, US Department of Commerce, Federal Information Processing Standard PUB, 1996: 1–20.
 - [6] FIPS 180-2, secure Hash standard[S]. NIST, US Department of Commerce, Federal Information Processing Standard PUB, 1997: 1–10.
 - [7] WANG X Y, LAI X J, FENG D G. Cryptanalysis of the Hash functions MD4 and RIPEMD[C]// EUROCRYPT 2005, LNCS 3494. Berlin: Springer-Verlag, 2005: 1–18.
 - [8] WANG X Y, YU H B. How to break MD5 and other Hash functions [C]// EUROCRYPT 2005, LNCS 3494. Berlin: Springer-Verlag, 2005: 19–35.
 - [9] WANG X Y, YU H B. Efficient collision search attacks on SHA-0 [C]// CRYPTO 2005, LNCS 3621. Berlin: Springer-Verlag, 2005: 1–16.
 - [10] WANG X Y, YU H B. Finding collisions in the Full SHA-1[C]// CRYPTO 2005, LNCS 3621. Berlin: Springer-Verlag, 2005: 17–36.
 - [11] 王小云,冯登国,于秀源. HAVAL-128 的碰撞攻击[J]. 中国科学: E 辑, 2005, 35(3): 1–12.
 - [12] LIANG J, LAI X J. Improved collision attack on Hash function MD5 [J]. Journal of Computer Science & Technology, 2007, 22(1): 79–87.
 - [13] XIE T, FENG D G. A new differential for MD5 with its full differential path [EB/OL]. [2009–04–10]. <http://eprint.iacr.org/2008/230.pdf>.
 - [14] XIE T, LIU F B. Could 1-MSB input differential be the fastest collision attack for MD5 [EB/OL]. [2009–04–20]. <http://eprint.iacr.org/2008/391>.
-
- (上接第 3173 页)
- [13] BECCHI M, CROWLEY P. An improved algorithm to accelerate regular expression evaluation [C]// Proceedings of the 3rd ACM/IEEE Symposium on Architecture for Networking and Communications Systems. New York: ACM Press, 2007: 145–154.
 - [14] FICARA D, GIORDANO S, PROCISSI G, *et al.* An improved DFA for fast regular expression matching [J]. ACM SIGCOMM Computer Communication Review, 2008, 38(5): 29–40.
 - [15] SMITH R, ESTAN C, JHA S. Xfa: Faster signature matching with extended automata [C]// Proceedings of the 2008 IEEE Symposium on Security and Privacy. Washington, DC: IEEE, 2008: 187–201.
 - [16] SMITH R, ESTAN C, JHA S, *et al.* Deflating the big bang: Fast and scalable deep packet inspection with extended finite automata [C]// Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication. New York: ACM Press, 2008: 207–218.
 - [17] BECCHI M, CADAMBI S. Memory-efficient regular expression search using state merging [C]// INFOCOM 2007: 26th IEEE International Conference on Computer Communications. Washington, DC: IEEE, 2007: 1064–1072.
 - [18] KUMAR S, VARGHESE G. Curing regular expressions matching algorithms from insomnia, amnesia and acalculia [C]// Proceedings of the 3rd ACM/IEEE Symposium on Architecture for Networking and Communications Systems. Washington, DC: IEEE, 2007: 155–164.
 - [19] BECCHI M, CROWLEY P. Extending finite automata to efficiently match Perl-compatible regular expressions [C]// Proceedings of the 2008 ACM CoNEXT Conference. New York: ACM Press, 2008.
 - [20] YU F, CHEN Z, DIAO Y, *et al.* Fast and memory-efficient regular expression matching for deep packet inspection [C]// Proceedings of the 2006 ACM/IEEE Symposium on Architecture for Networking and Communications Systems. New York: ACM Press, 2006: 93–102.
 - [21] BECCHI M, CROWLEY P. A hybrid finite automaton for practical deep packet inspection [C]// Proceedings of the 2007 ACM CoNEXT Conference. New York: ACM Press, 2007: 1–12.
 - [22] KONG S, SMITH R, ESTAN C. Efficient signature matching with multiple alphabet compression tables [C]// Proceedings of the 4th international Conference on Security and Privacy In Communication Networks. New York: ACM Press, 2008.
 - [23] Perl compatible regular expressions [EB/OL]. [2009–04–11]. <http://www.pcre.org/>.
 - [24] 郝克刚,段振华,李新. 论回溯自动机[J]. 计算机学报, 1990, 13(5): 340–348.
 - [25] HOPCROFT J E, MOTWANI R, ULLMAN J D. Introduce to automata theory, languages, and computation [M]. 3rd ed. Upper Saddle River, New Jersey, USA: Addison-Wesley Professional, 2007.
 - [26] BECCHI M, CROWLEY P. Efficient regular expression evaluation: Theory to practice [C]// Proceedings of the 4th ACM/IEEE Symposium on Architectures for Networking and Communications Systems. New York: ACM Press, 2008: 50–59.
 - [27] BECCHI M, FRANKLIN M, CROWLEY P. A workload for evaluating deep packet inspection architectures [C]// IISWC 2008: IEEE International Symposium on Workload Characterization. Washington, DC: IEEE, 2008: 79–89.
 - [28] 陈曙晖,苏金树,范慧萍,等. 一种基于深度报文检测的 FSM 状态表压缩技术[J]. 计算机研究与发展, 2008, 45(8): 1299–1306.