doi:10.3969/j.issn.1001-0505.2015.06.001

基于BB-BM算法的网络协议内容符合性测试方法

李 涛[1] 胡爱群[1] 高 尚[2]

([1]东南大学信息科学与工程学院,南京 210096)
([2]香港理工大学电子计算学系,香港 999077)

摘要: 为了检测网络通信协议的安全性,使用高效的模式识别方法对协议内容进行符合性测试.采用黑盒测试的方法,在检测端将协议服务器和检测模块分离,设计了协议安全性测试框架和测试流程;提出了以字节块为单位、分块计算摘要值再进行匹配的BB-BM算法.实验结果表明,使用该方法能够对网络协议按照内容种类划分值域空间,通过匹配算法进行符合性测试.在进行模式匹配时通过分块处理减少了模式串和目标串数量,从而导致跳跃距离增加,匹配次数减少,检测性能在最优和最差测试状态下较现有检测方法分别提高了20%和80%.在该测试框架下,以字节块为单位进行匹配有效提升了检测效率,适用于对字段格式固定的网络协议进行内容符合性测试.

关键词: 内容符合性;模式识别;BM算法;协议安全

中图分类号: TN918.91 **文献标志码:** A **文章编号:** 1001-0505(2015)06-1027-05

Context consistency test method for network protocol based on BB-BM algorithm

Li Tao[1] Hu Aiqun[1] Gao Shang[2]

([1]School of Information Science and Engineering, Southeast University, Nanjing 210096, China)
([2]Department of Computing, Hong Kong Polytechnic University, Hong Kong 999077, China)

Abstract: In order to check the security of network communication protocols, the efficient pattern recognition method is used to test protocols' context consistency. Testing framework and process for protocol security are designed based on the black testing method with the testing part being divided into the protocol server and the testing module. The BB-BM (block based Boyer Moore) algorithm is proposed, in which the words block is used as unit and matching blocks is carried out after the calculation of digests. The experimental results show that the proposed system can divide the value space of network protocol based on the context type. The consistency test is carried out by the recognition algorithm. The number of pattern strings and target strings decreases by the block division process during the pattern matching, and correspondingly the skip distance increases and the times of recognition decrease. Compared with the existing matching methods, the testing performance of the proposed method increases by 20% and 80% under the best and worst testing conditions, respectively. In this system, the testing efficiency is effectively improved by using words block, which is suitable for context consistency tests of network protocol with fixed field format.

Key words: context consistency; pattern recognition; BM (Boyer Moore) algorithm; protocol security

网络通信协议的实现依据标准协议规范,但由于不同实现者对协议有着不同的理解,设计的通信设备可能不尽相同.攻击者可以通过对安全协议进行部分修改以降低协议的安全性,窃取使用者数

收稿日期: 2015-04-23. **作者简介:** 李涛(1984—),男,博士,讲师,lit@seu.edu.cn.
基金项目: 国家发改委信息安全专项基金资助项目、国家重点基础研究发展计划(973计划)资助项目(2013CB338003).
引用本文: 李涛,胡爱群,高尚.基于BB-BM算法的网络协议内容符合性测试方法[J].东南大学学报:自然科学版,2015,45(6):1027-1031. [doi:10.3969/j.issn.1001-0505.2015.06.001]

据. 目前,主要利用一致性检测方法来研究实际通信协议与标准协议之间的差异[1]. 但协议一致性检测偏重于对协议运行状态的检测,针对协议内容字段的检测方法则较为缺乏,从而导致在检测结果不一致的情况下,无法判断协议实现者是对标准协议进行了更加安全的修改还是省略部分安全过程以达到非法目的[2]. 因此,在协议一致性检测之后,需要对协议的内容进行符合性检测.

内容符合性检测通过模式识别方法来判断协议实现是否与协议标准规定的字段内容相符. 对于模式识别方法,关键问题是如何增加跳跃距离来减少匹配次数,最后达到降低匹配时间、提高算法效率的目的[3].

在模式识别方法的研究中,暴力匹配(BF)算法最先被引入[4],该算法顺序比较字符串中每个字符,但由于对回溯的处理过于简单导致算法效率较低,匹配的时间复杂度为$O(mn)$,其中m,n分别为模式串和目标串长度. 针对BF算法的不足,Cook[5]从理论上论证了模式匹配问题可以在$O(m+n)$时间内解决. 基于此思想,Sunday[6]提出了KMP算法,Cho等提出了BM算法[7],匹配效率都得到了提高. 尤其是BM算法,其通过逆向匹配思想最大限度地增加了匹配失败时的跳跃距离,减少了匹配字符数目,在最优情况下算法的时间复杂度减少到$O(n/m)$;但由于匹配过程中规则较多,实现过程相对复杂. 基于此,学者们对BM算法进行了改进,出现了BMH算法[8]和BMHS算法[9],即利用BM算法中坏字符跳跃规则,将最大跳跃距离分别增大到m和$m+1$;但这些算法并没有解决BM算法中模式串与目标串之间块未对齐的问题,导致匹配效率较低,甚至在块匹配时会产生不可预知的错误.

本文设计了一种针对协议内容符合性的测试方法,并针对该方法下的特殊模式匹配环境,将BM算法进行了改进,提出了一种基于块的BM(BB-BM)算法,以提高检测效率.

1 系统设计

协议安全检测是根据相关标准中描述的语义、语法、时序,对具体实现的协议进行符合性检测[10]. 在测试过程中,通常将测试对象看作一个无法打开的黑盒,在不考虑黑盒内部结构与实现的前提下,通过黑盒提供的接口对其进行测试. 因此,对某个协议的检测,需要在与待测设备通信的另一端进行,通过相关接口与测试对象交互操作. 这就需要在设计通信协议安全检测方法时,将检测系统与协议服务器分离.

在进行协议安全检测时各实体之间的关系如图1所示. 图中,被测实现是指待检测协议的客户端软件实现;协议服务器是指被测协议的服务器端,其功能是在收到被测实现的反馈数据后,将数据交付至协议安全检测系统进行检测,在检测完成后,反馈协议服务器信息,根据反馈信息对被测实现进行响应;协议安全检测系统是实现协议内容符合性检测的核心,在收到协议服务器的报文输入后,通过分析正常检测序列与被测实现之间的完整通信数据,判断报文各字段的内容是否与标准协议中的规定相符,检测流程见图2.

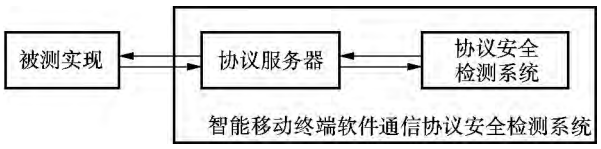

图1 协议安全检测系统检测框架

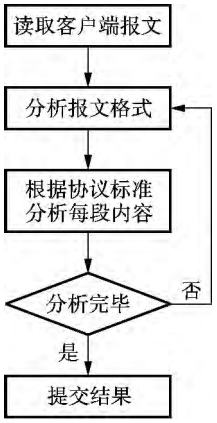

图2 内容符合性检测流程

2 基于块的BM算法

匹配算法是协议内容符合性检测的核心. 本文针对协议安全检测环境,对BM算法进行改进,提出了基于块的BM算法,以提高检测效率.

2.1 BM算法

BM算法的特点是从模式串的尾部开始匹配. 在模式串和目标串对齐后,从最右边的对齐位置开始向左逐个进行扫描匹配. 在一轮匹配结束后,BM算法采用坏字符规则和好后缀规则2种启发式规则将窗口右移.

假设$P=\{p_1, p_2, \cdots, p_m\}$为模式串,$T=\{t_1, t_2, \cdots, t_n\}$为目标串,则坏字符规则为

$$\text{skip}(t_j)=\begin{cases}m & t_j\notin P\\ m-\max\{\text{location}(t_j)\} & t_j\in P\end{cases}\quad(1)$$

式中，t_j 为匹配成功字符；$\text{skip}(t_j)$ 为 t_j 处失败时 P 右移的长度；$\text{location}(t_j)$ 为 t_j 在 P 中出现的位置；m 为模式串长度.

好后缀规则为

$$\text{shift}(p_{m-k})=\begin{cases}m-k-s+r & \exists s,r,p_{m-r}\neq p_{s-r}\\ \text{shift}(p_{s-r}) & \exists s,r,p_{m-r}=p_{s-r}\\ m-k & \text{其他}\end{cases}\quad(2)$$

式中，k 为已匹配成功的字符串长度；$\text{shift}(p_{m-k})$ 为 p_{m-k} 匹配失败时 P 右移的长度；r 为对于任意 s，满足 $\{p_{m-r+1},p_{m-r+2},\cdots,p_m\}=\{p_{s-r+1},p_{s-r+2},\cdots,p_s\}$ 条件的最大值；s 为在确定 r 之后，满足 $\{p_{m-r+1},p_{m-r+2},\cdots,p_m\}=\{p_{s-r+1},p_{s-r+2},\cdots,p_s\}$ 条件的最大值.

BM 算法是一种高效的模式匹配算法，但在协议安全检测环境中，匹配对象有其固有的特性：通常协议帧的格式由不同块组成，不同块之间由特定的分割符号来区分，或者是有固定的块长度. 因此，在进行匹配检测时以数据块为单位，而非以字符为单位. 例如，对于一个由{aa}，{bbbb}，{ccc}，{d}组成的目标串{aa，bbbb，ccc，d}以及由{ccc}，{d}组成的模式串{ccc，d}，直接应用 BM 算法匹配，经过第 1 次右移后的结果见图 3. 由图可知，模式串的块与目标串的块并未对齐，从而导致效率降低.

图 3　BM 算法的匹配结果

针对好后缀规则，令本轮模式串和目标串对应的位置如图 4 所示. 由图可知，$\{t_4,t_5\}=\{p_4,p_5\}=\{a,b\}$ 为好后缀，并且 $t_3\neq p_3$. 由于 $\{p_2,p_3\}=\{p_4,p_5\}$，$p_1\neq p_3$，则好后缀将使 p_1 和 t_3 对齐. 如果 $p_1\neq t_3$，对 p_2,p_3,p_4,p_5 进行匹配是无意义的.

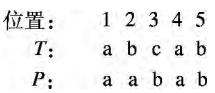

图 4　好后缀规则示例

2.2　BB-BM 算法

针对 2.1 节中的问题，在协议安全检测的应用场景下，本文对 BM 算法进行了改进，提出了 BB-BM 算法.

首先，在使用好后缀规则前，使用一次坏字符规则来进行对齐决策. 改进后的好后缀规则为

$$\text{shift}(p_{m-k})=\begin{cases}m-k-s+r & \exists s,r,t_{\text{bad}}=p_{s-r}\\ \text{shift}(p_{s-r}) & \exists s,r,t_{\text{bad}}\neq p_{s-r}\\ m-k & \text{其他}\end{cases}\quad(3)$$

式中，t_{bad} 为好后缀中的坏字符. 在此规则下，字符串将持续移动直到 $t_{\text{bad}}=p_{s-r}$.

其次，将以字符为单位进行匹配改成以块为单位进行匹配，并对每一个块进行 Hash 操作，将得到的数据重新组成一个新的序列进行匹配. 对于模式串的右移，BB-BM 算法继承了坏字符和好后缀规则，仅将距离单位由字符长度改为块长度.

BB-BM 算法的匹配步骤如下：

① 估算目标串中块的数目 b. 若估算困难，计算目标串中块的数目，保证 Hash 函数的值域足够使用.

② 根据估算的数目来决定 Hash 函数值域的空间大小，$S=\log(b+1)$.

③ 依次将模式串和目标串中的每一个块进行 Hash 操作. 对于较小的值域空间，可以依次赋值. 例如，对于一个 $S=4$ bit 的空间，可将第 1 块转化为 0001 B，第 2 块转化为 0010 B，以此类推. 在尾部需要定义结尾字符，如采用 0000 B 来标注模式串和目标串的结尾.

④ 转换完成后对模式串和目标串进行拼接，以块为单位进行模式匹配.

3　系统实现与性能分析

采用本文提出的测试方法对网络传输中常用的 IPSec 和 HTTP 协议进行内容符合性分析和测试.

3.1　协议内容符合性分析

协议内容符合性测试的流程如图 5 所示. IPSec 协议簇中起最主要作用的是 IKE，AH 协议和 ESP 协议，其中 IKE 是实现安全功能的核心，对其进行内容符合性检测尤为重要. 首先，对 IKE 协议头进行分段. 根据 RFC4306 中的 IPSec 协议规范，IKE 协议头包含 4 种内容类型，其值域空间大小 $S=\lceil\log(4+1)\rceil=3$ bit；然后，对模式串和目标串分块进行 Hash 操作，生成新的模式串及目标串；最后，对转换后的数据进行模式匹配.

对 HTTP 进行内容符合性检测，首先参考 RFC2616 标准规范对 HTTP 头进行分段处理；例如 request-line 中的 Method 字段，标准规范定义了

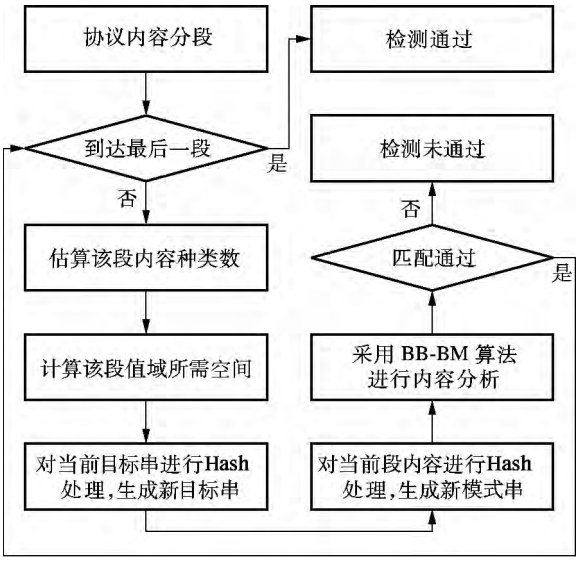

图5 内容符合性分析流程图

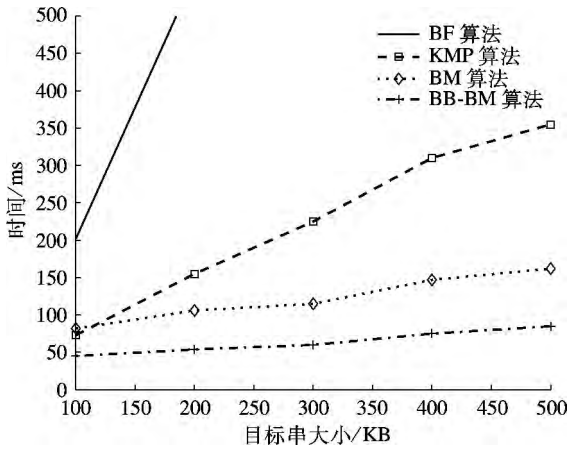

图6 4种算法的效率比较

OPTIONS, GET, HEAD, POST, PUT, DELETE, TRACE, CONNECT 共8种类型,因此其值域空间大小 $S=\lceil\log(8+1)\rceil=4$ bit. 然后,分别对模式串和目标串进行 Hash 操作,生成新的模式串和目标串. 最后,利用 BB-BM 算法进行模式匹配.

3.2 性能分析

采用本文所提方法,对由{aa},{bbbb},{ccc},{d}组成的目标串{aa, bbbb, ccc, d}以及由{ccc},{d}组成的模式串{ccc, d}进行测试. 首先,分析得到目标串中块的数量为4,计算得值域空间大小 $S=\lceil\log(4+1)\rceil=3$ bit. 然后,对每一块进行 Hash 操作. 为方便描述,将每一块的转换结果使用符号记录,转换关系如表1所示. 经过 Hash 操作转换后,模式串为{CD},目标串为{ABCD}. 最后,进行模式匹配,仅需右移1次,即可得到预期结果.

表1 Hash 转换对应关系

块	Hash 转化结果/B	符号
aa	001	A
bbbb	010	B
ccc	011	C
d	100	D
结束符	000	

下面对 BF 算法、KMP 算法、BM 算法以及 BB-BM 算法进行性能比较. 利用100~500 KB的目标串和5 KB的模式串进行测试. 在使用 BB-BM 算法进行块转换操作时,将模式串和目标串的块大小统一为8 B. 为了消除误差影响,采用对100次测量结果取均值的方法.

图6描述了4种算法的时间消耗. 由于模式串和目标串的内容对模式匹配算法的效率影响较大,因此对于不同目标串和模式串的选择,时间消耗可能存在差别.

由图6可知,由于 BF 算法消耗时间过长,当处理时间大于500 ms时不再显示. BM 算法的最优情况在匹配对象拥有特定格式时才会出现,故在小文件处理过程中,其效率可能低于 KMP 算法. 但 KMP 算法在处理大文件时,时间消耗增长较快;相反,BM 算法处理大文件的时间增长则不明显. BB-BM 算法继承了 BM 算法的优点,性能相对于 BM 算法有所提高,但由于需要进行 Hash 操作,其时间消耗并未达到理想的理论值.

4种算法的性能比较结果见表2. 由表可知,BB-BM 算法的匹配效率与模式串和目标串块的数量相关. 假设 Hash 操作后模式串的块数为 M,目标串的块数为 N,则最优时间复杂度为 $O(N/M)$,最差时间复杂度为 $O(MN)$. 而 BM 算法的最优时间复杂度为 $O(n/m)$,最差退化到 $O(mn)$;基于 BM 算法改进的 BMH 算法和 BMHS 算法在时间复杂度上相对于 BM 算法并未有较大改善. 对于本文中的检测示例,Hash 转换之前的目标串和模式串大小分别为10和4,使用 BM 算法的最优和最差时间复杂度分别为 $O(2.5)$ 和 $O(40)$. 使用 BB-BM 算法时,经过 Hash 转换后目标块串和模式块串大小分别为4和2,最好和最差时间复杂度分别为 $O(2)$ 和 $O(8)$,检测性能相对于使用 BM 算法分别提高了20%和80%. 虽然最好时间复杂度对于 BM 算法没有明显降低,但最坏情况的时间复杂度明显降低,其平均效率明显提升.

传统的 BF 算法、KMP 算法和 BM 算法适用于任何需要进行模式匹配的场景,适用范围更加广泛. BB-BM 算法是针对本文的网络协议内容符合性检测提出的,要求匹配对象拥有固定的字段格式

表 2　4 种算法的性能比较

比较类别	BF 算法	KMP 算法	BM 算法	BB-BM 算法
代码实现	简单	中等	复杂	复杂
适用范围	广泛	广泛	广泛	狭窄
算法效率	低	中	高	很高
时间复杂度	$O(mn)$	$O(m+n)$	$O(n/m)$	$O(N/M)$

以进行分块处理，因此其通用性较差. 但在网络协议内容符合性检测的场景下，BB-BM 算法具有较高的检测效率. 对于其他场景，如果模式串和目标串可以基于特定格式进行划分，并且可通过分块的方式进行处理，使用 BB-BM 算法也可对其进行模式匹配操作.

4　结语

本文针对网络协议内容的符合性测试提出了一种测试方法，分析了已有的模式匹配算法应用在协议内容符合性测试中的问题，对 BM 算法进行了改进，提出了 BB-BM 算法. 理论分析和性能测试的结果表明，BB-BM 算法在继承了 BM 算法的优点的同时，检测效率得到了提高.

本文的测试方法适用于构建协议安全测试系统，基于该方法进行协议内容符合性测试能显著提高测试效率. 但对于未知协议的检测，由于无法通过服务器端进行控制，故不建议采用本文方法，可以通过旁路监听客户端和服务器端的通信数据，采用模式识别和模糊测试结合的方法，将被测实现与协议标准进行匹配，达到检测的目的.

参考文献 (References)

[1] Heckel R, Mariani L. *Automatic conformance testing of web services* [J]. *FASE*, 2005, **3442**: 34-48.

[2] 冯登国，范红. 安全协议形式化分析理论与方法研究综述 [J]. 中国科学院研究生院学报，2003，20(4): 389-406.
Feng Dengguo, Fan Hong. Survey on theories and methods of formal analyses for security protocols [J]. *Journal of the Graduate School of the Chinese Academy of Sciences*, 2003, **20**(4): 389-406. (in Chinese)

[3] Chen L, Wang W. An improved NSSK protocol and its security analysis based on logic approach [C] //*International Conference on Communications, Circuits and Systems*. Xiamen, China, 2008: 772-775.

[4] Lecroq T. Fast exact string matching algorithms [J]. *Information Processing Letters*, 2007, **102**(6): 229-235.

[5] Cook S A. The complexity of theorem-proving procedures [C] //*Third Annual ACM Symposium on Theory of Computing*. New York, USA, 1971: 151-158.

[6] Sunday D M. A very fast substring search algorithm [J]. *Communications of the ACM*, 1990, **33**(8): 132-142.

[7] Cho S, Na J C, Park K, et al. Fast order-preserving pattern matching [J]. *Combinatorial Optimization and Applications*, 2013, **8287**: 295-305.

[8] Rytter W. On maximal suffixes and constant-space linear-time versions of KMP algorithm [J]. *Theoretical Computer Science*, 2003, **299**(1/2/3): 763-774.

[9] 毕智超. 字符串模式匹配算法的研究及改进 [J]. 电子测试，2013(20): 64-65.
Bi Zhichao. The research of improved matching algorithm of string pattern [J]. *Electronic Test*, 2013(20): 64-65. (in Chinese)

[10] Fu Y, Kone O. Security and robustness by protocol testing [J]. *IEEE Systems Journal*, 2014, **8**(3): 699-707.