基于数据报指纹关系的未知协议识别与发现*

宋 疆[1], 张春瑞[2], 张 楠[1], 李 芬[2], 吴艳梅[1]

(1. 电子科技大学 计算机科学与工程学院, 成都 611731; 2. 中国工程物理研究院 计算机应用研究所, 四川 绵阳 621900)

摘 要: 目前,关于窃密防范措施基本上只针对已知协议,为了保证网络的安全运行以及对攻击与危害行为的预警,迫切需要在当前结构复杂的网络环境下为决策者准确提供一种能高效地对未知协议进行识别的方法。为此,在整合已有的网络安全和数据挖掘技术的基础上,设计了基于数据报指纹关系的未知协议识别发现的解决方案。

关键词: 无线数据获取; 机器学习; 频繁集; 指纹特征

中图分类号: TP309 **文献标志码**: A **文章编号**: 1001-3695(2012)12-4604-03

doi:10.3969/j.issn.1001-3695.2012.12.052

Network traffic identification based on data finger-print

SONG Jiang[1], ZHANG Chun-rui[2], ZHANG Nan[1], LI Fen[2], WU Yan-mei[1]

(1. *School of Computer Science & Engineering, University of Electronic Science & Technology of China, Chengdu* 611731, *China*; 2. *Institute of Computer Application, China Academy of Engineering Physics, Mianyang Sichuan* 621900, *China*)

Abstract: As all of the current communication protocols are unconventional dedicated unknown ones while existing of prevention measures mainly aiming at the known protocols and based on port mapping or static features matching, they are useless for the monitoring and detection of the theft channel. To ensure the security of the network as well as the early warning of attacks and harmful behavior, policy-makers urgently need to provide an efficient way to identify an unknown protocol under the current structure of the complex network environment. To solve these problems, this project would integrate the existing network security and data mining technology to design solutions discovering unknown protocols based on the datagram fingerprint relations, to meet the national cyber security and many other needs. In addition to a significant meaning on the protection of network security and macro warning area, it promoted and improved the capability of independent innovation of China's network security and other aspects of the key technologies at the same time.

Key words: wireless data acquisition; machine learning; frequent set; fingerprint characteristics

网络的发展日趋复杂,保障信息网络的安全已成为国家信息化战略的核心内容。在特定的网络环境下,通过特殊手段进行窃密的威胁日趋严峻,此类窃密途径通常是通过无线通信的方式发送涉密信息,而这种通信采用的协议均为非常规的专用未知协议;同时,在电子对抗的情况下,从截获的通信流比特数据中对未知协议进行识别,现有的网络安全检测手段和协议识别方法在实际应用时存在诸多问题,如基于端口映射和静态特征匹配的方法不能有效识别未知协议等。

为了保证网络的安全运行以及对攻击与危害行为的预警,迫切需要在当前结构复杂网络环境下为决策者提供一种能高效地对未知协议进行识别的方法。

1 相关概念和算法介绍

1.1 模式匹配算法以及近似串匹配算法

1.1.1 模式匹配算法

模式匹配(pattern matching,或称为串匹配 string matching)[1]就是在字符集 Σ 上,给定模式集合 P,对于任意的一个字符串 T,找出 P 中模式 T 出现的所有位置。

模式匹配按照模式个数可以分为单模式匹配[2,3]和多模式匹配[4~6]。而未知协议的比特流特征寻找最大特点就是多候选模式,与多模式匹配算法所解决的问题是完全一致的。

1.1.2 近似串匹配算法

经典串匹配问题主要包括精确串匹配、正则表达式串匹配和近似串匹配。其应用领域包括信息检索、病毒检测等。

近似串匹配算法主要包括基于计算编辑距离矩阵的动态规划算法、基于自动机的算法等。编辑距离又称为 Levenshtein 距离[7],是指两个字串之间相互转换所需的最少编辑操作次数。

1.2 面向比特流数据的改进的 AC 算法

1.2.1 比特流数据的特点

比特是网络传输中信息的最基本的单位。比特流数据具有以下两个明显的特征:

a)单一性。比特流中包含的元素只有 0 和 1,不会出现其他元素。

b)顺序性。比特与比特之间的排列顺序是固定的,且一旦顺序改变,其代表的意义有可能截然相反。而数据库挖掘中强调的是集合的概念,对顺序没有要求。

收稿日期: 2012-04-20; **修回日期**: 2012-06-05 **基金项目**: 国家"242"信息安全计划资助项目(2010A14); 国家科技重大专项资助项目(2011ZX03002-002-03); 四川省科技支撑计划资助项目(2010FZ0101); 中国工程物理研究院科学发展技术基金资助项目(2012A0403021)

作者简介: 宋疆(1986-),男,四川宜宾人,硕士研究生,主要研究方向为网络与信息安全(songjiang_uestc@163.com).

由以上的特点可知,比特流数据挖掘虽然也是寻找频繁出现的模式,但是需要考虑顺序性和单一性的特点。

1.2.2　改进的 AC 算法

在实际的网络传输中,基于比特流数据的报文有可能不是以整个字节开头的,且其中的频繁序列也不一定是以字节为单位的,同时在传输过程中数据可能出现偏移或出错。为了避免这些实际可能出现的问题,以字节为单位的多模式匹配算法很难再满足实际需求,因此需要对传统的 AC 算法[8]进行改进,使其以比特为单位进行挖掘。改进的 AC 算法在算法的效率上有所改进,实现对所有的序列扫描一次,即可寻找到所有的特征序列并进行相关统计。

根据前述比特流的特点和 AC 算法的思路,以 1-3 位模式序列计数为例,其完整的统计过程的伪代码如下:

```
(a) create buffer[3] = '#'; count[8] = 0;
(b) char *str = read(file); n = getlength(file);
(c) for i = 1;n
        if(i-1<0||i-2<0){
        buffer[1] = str[i]; index = buffer[1];
        count[i]++;}
(d) if(i-2<0){
        buffer[1] = str[i-1]; buffer[2] = str[i];
        index = buffer[2]*2 + buffer[1];
        count[index]++;}
(e) else
        buffer[1] = str[i-2];buffer[2] = str[i-1];
        buffer[3] = str[i];
        index = buffer[3]*4 + buffer[2]*2 + buffer[1];
        count[index]++;
```

这样,要得到各序列的计数,整个统计过程仅需扫描一次源数据,即可得到所有长度模式序列的计数,理论上该算法可以统计任意两个长度之间的所有比特流出现的情况。

1.3　面向比特流数据的指纹特征提取

1.3.1　关联规则相关概念

1993 年 Agrawal 等人[9]提出了关联规则[10,11]的挖掘问题。关联规则挖掘是指在给定的数据库中寻找数据项之间的相互关系,从而发现未知规律,为决策提供依据的一项数据分析方法[11]。根据文献[12],关联规则的定义如下:

设 $I=\{I_1,I_2,I_3,\cdots,I_m\}$ 是项的集合,给定一个数据库 $D=\{t_1,t_2,t_3,\cdots,t_m\}$,其中每个事务 t 都是 I 的非空子集,即 $t\subseteq I$,每个事务 t 都与一个唯一的标志符 TID 对应。

定义 1　若 $t\subseteq I$,则称事务 t 支持项目集 I。

定义 2　若 D 的全部 n 个事务 t 中有 s 个支持项目集 I,则 I 的支持度为 $\frac{s}{n}$,记为 $\mathrm{supp}(I)$。

定义 3　若 $\mathrm{supp}(I)$ 不小于用户所定义的最小支持度,则称项目集 I 为频繁集;反之则为非频繁集。

定义 4　关联规则是形如 $X\Rightarrow Y$ 的蕴涵式,其中 $X,Y\subseteq I$ 且 $X\cap Y=\varnothing$,X 和 Y 分别称为关联规则的先导和后继。

定义 5　关联规则 $X\Rightarrow Y$ 的置信度 $\mathrm{conf}(X\Rightarrow Y)=\frac{\mathrm{supp}(X\cup Y)}{\mathrm{supp}(X)}$。

给定一个事务集 D,挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度和最小可信度的关联规则[13]。

1.3.2　面向比特流的频繁模式序列挖掘研究

为了便于叙述,对比特流作一些相关定义[14]:

设 S 是一个包含 n 位比特的随机比特流(即 S 的任意位置上出现 0 和 1 比特的概率是相同),P 是一个比特模式序列:

定义 6　若模式序列 P 包含 m 个比特位,则称 P 的长度为 m。

定义 7　若 S 中总共包含 r 个长度为 m 的模式序列,其中模式序列 P 出现了 k 次,则 P 的支持度为 $\frac{k}{r}$,记为 $\mathrm{supp}(P)$。对于确定的 m, r 值可以直接计算得到 $r=n-m+1$。

定义 8　若 $\mathrm{supp}(Q_0)=\mathrm{supp}(Q_1)=0.5\times\mathrm{supp}(P)$ 不小于用户定义的最小支持度,则模式序列 P 为频繁序列;反之,P 为非频繁序列。

定理　在长度为 n 的随机比特流 S 中,若模式序列 P 的长度为 m,其支持度为 $\mathrm{supp}(P)$,则其长度为 $m+1$ 的父序列 Q_0 与 Q_1 的支持度相同,且 $\mathrm{supp}(Q_0)=\mathrm{supp}(Q_1)=0.5\times\mathrm{supp}(P)$。

未知协议的比特流特征寻找最大特点就是多候选模式,与多模式匹配算法所解决的问题是完全一致的。利用成熟的多模式匹配算法[16],可以通过一遍扫描得到各候选序列出现的位置以及次数,以便进一步筛选。

2　基于指纹特征的数据帧定界

在比特流数据中,通过改进的数据挖掘算法提取出数据报指纹特征以后,需要对数据帧长度进行估计,即对数据帧进行定界。寻找相同指纹特征出现的最小位置差是进行帧定界的基本步骤:

a)假定比特流中提取出的频繁特征序列集合为 $A=\{X,Y,Z\cdots\}$。其中 $X,Y,Z\cdots$ 代表比特流中提取出的频繁特征序列;$X=\{X_0,X_1,X_2\cdots\}$,$\mathrm{pos}(X_i)$ 代表频繁特征序列 X 在比特流数据中出现的位置。对于 A 集合中的频繁序列 X 内的两个特征序列 X_i 和 X_{i+1},根据其在比特流中出现位置的先后次序,计算其位置差 $\mathrm{pos}(X_{i+1})-\mathrm{pos}(X_i)$(假定 $\mathrm{pos}(X_{i+1})>\mathrm{pos}(X_i)$),对所有不同位置的 X 进行两两验证,令 $\mathrm{repeate}_i=\min(\mathrm{pos}(X_{i+1})-\mathrm{pos}(X_i))$,并统计出现 $\mathrm{repeate}_i$ 的出现次数 count_i。

b)对 A 中所有元素作相同的处理,可以得到 $P_i=\{(x,y)\mid x=\mathrm{repeate}_i,y=\mathrm{count}_i\}$,其中 $i>0$。

c)对步骤 a)和 b)的结果 P_i 中的 y 值进行排序,对排序结果进行筛选,根据筛选结果给出数据帧定界的参考范围。

3　实验结果与分析

使用 Wireshark 抓包得到的数据具有以太网数据包的共同特征,采用 VC ++6.0 编写实验程序,在局域网环境中抓取数据报进行测试。

3.1　ARP 数据包实验

3.1.1　ARP 数据包频繁项集的提取

测试截取 ARP 包 400 个,选取阈值为 0.5、0.6、0.7 进行频繁项集的特征提取。运行测试程序后,统计结果如下:

a)阈值为 0.5。根据改进算法提取 4 位频繁项集,其中 0×0、0×1、0×2、0×4出现次数最多,分别是 44 112、9 470、9 306和 8 103 次。同时提取的 8 位频繁项集,可以找到的 8 位频繁项集有 62 个,筛选率为 75%。

b)阈值为 0.6。提取 8 位频繁项集,可以找到的 8 位频繁项集有 54 个,筛选率为 78.9%。

c)阈值为 0.7。提取 8 位频繁项集,可以找到的 8 位频繁项集有 34 个,筛选率为 86.7%。

图 1 为 ARP 在不同阈值下提取频繁集对比。从 ARP 数据包的频繁集提取曲线可以看出,阈值的选择直接影响到频繁集个数和筛选率。因此,在实验中选择合适的阈值是十分重要的。分析图 1 后,笔者选取 0.6 的阈值提取的频繁集个数和筛选率都是合适的,且筛选出的连续长序列对数据包关键部分的覆盖也是合适的。

3.1.2 ARP 数据帧切分

测试截取 ARP 包 400 个,提取 24~88 位的频繁序列。计算相邻频繁序列的位置差,并统计相同位置差的出现次数,筛选出现次数最多的值,作为数据包帧定界的参考集合。根据统计的 ARP 数据帧定界参考值出现次数绘制折线图(图 2),明显发现 ARP 数据包中以 336 作为相同频繁序列位置差的出现次数远远高于其他位置差,这与通过 Wireshark 解析到的 Ethernet帧长度一致。

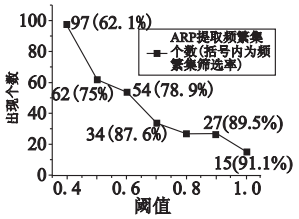

图1 ARP在不同阈值下提取频繁集对比

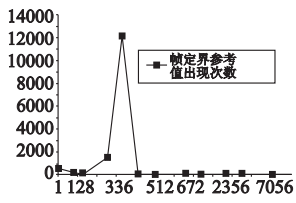

图2 ARP数据包帧定界

3.2 ICMP 数据包实验

3.2.1 ICMP 数据包频繁项集的提取

测试截取 ICMP 包 200 个,选取阈值为 0.5、0.6、0.7 进行频繁项集的特征提取。图 3 为 ICMP 在不同阈值下提取频繁集对比。运行测试程序后,统计结果如下:

a)阈值为 0.5。根据改进算法提取 4 位频繁项集,其中 0×0、0×1、0×8、0×2 出现次数最多,分别是 21 669、5 216、5 215和4 667 次。

同时提取的 8 位频繁项集,可以找到的 8 位频繁项集有 127 个,筛选率为 50.4%。其中出现次数最多的是 0×00、0×01、0×80、0×02,出现次数分别是 10 502、2 174、2 173 和 1 658 次。

b)阈值为 0.6。提取的 8 位频繁项集,可以找到的 8 位频繁项集有 104 个,筛选率为 59.4%。其中出现次数最多的是 0×00、0×01、0×80、0×02,出现次数分别是 10 502、2 174、2 173和 1 658 次。

c)阈值为 0.7。提取的 8 位频繁项集,可以找到的 8 位频繁项集有 56 个,筛选率为 78.1%。其中出现次数最多的是 0×00、0×01、0×80、0×02,出现次数分别是 10 502、2 174、2 173和 1 658 次。

与 ARP 数据包类似,ICMP 数据包的频繁集提取也与阈值选择有很大关系。分析图 3 后,笔者认为选取 0.7 的阈值提取的频繁集个数和筛选率都是合适的,这也在后面的实验中得到了验证。

3.2.2 ICMP 数据包切分

抓取 ICMP 包 200 个,通过改进算法对抓取的 ICMP 数据包进行处理,提取 24~88 位的频繁子串。计算相邻频繁子串的位置差,并统计相同位置差的出现次数,筛选出现次数最多的位置差,作为数据包帧定界的参考集合。

根据统计的 ICMP 数据包帧定界参考值出现次数明显发现,ICMP 数据包中位置差为 376 的出现次数远远高于其他位置差,这与 Wireshark 中解析的 Ethernet 帧长度一致。

3.3 实验结果分析

在频繁项集提取的实验中,通过对 AC 算法的改进,形成比特流快速统计算法,大大减小了 I/O 次数,提高了效率。同时,将实验结果中提取的频繁项集与真实数据包相比较,能够找出真正频繁项集,说明该算法对真实协议是有效的,其中 ARP 协议的结果最为明显,且筛选率较高,可以去除绝大多数冗余数据。

在数据帧进行定界实验中,通过对 ARP 包和 ICMP 包频繁集的提取,发现阈值的选择对于频繁集的提取的影响,因此,选择适当的阈值是十分必要的。图 4 为 APR 与 ICMP 提取频繁集的效果对比。从图 4 可知,阈值对 ICMP 包频繁集的影响更大,ARP 包和 ICMP 包在阈值为 0.7 时,从筛选率和提取的频繁集个数来看都是比较合适的。

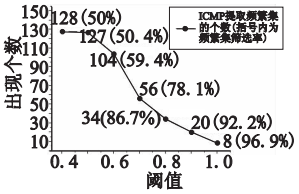

图3 ICMP在不同阈值下提取频繁集对比

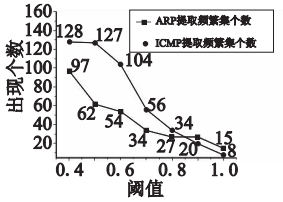

图4 APR与ICMP提取频繁集的效果对比

通过在实验室环境下进行大量的数据实验,验证了方法的有效性和正确性。

表 1 是 ARP 和 ICMP 在实验程序下的准确率。

表 1 ARP 和 ICMP 的准确率对比

协议类型	ARP	ICMP
准确率/%	62.5	96

通过对频繁序列的进一步处理和分析,能够得到对数据帧定界的参考范围,ARP 和 ICMP 的实验都说明了出现次数最多的位置差都与该协议的数据帧长度基本一致。该方法对数据帧进行定界是有效的。

4 结束语

本文根据比特流数据的实际特点,运用面向比特流数据的改进 AC 算法,提出了一个基于该改进算法的面向比特流数据的未知协议指纹特征提取和对未知协议的数据帧进行定界的方法,并采用真实数据证明将未知协议指纹特征提取和对比特流数据的未知协议数据帧定界的有效性和准确性。对比分析结果显示:指纹特征提取准确率达到 62.5% 和 96%,参考的帧长与实际帧长基本一致。本文方法在实际应用中还需要根据具体情况进行改进,例如变长数据帧的处理;如何利用帧长数据对比特流数据进行有效的数据帧切分;如何在帧长范围以内挖掘指纹特征之间的关联规则;如何根据前述的成果确定数据帧帧格式。这些都是未来需要进一步完成的工作。

参考文献:

[1] CORMEN T H, LEISERSON C E, RIVEST R L, *et al*. 算法导论[M].2 版.北京:高等教育出版社,2003.

(下转第 4614 页)

基于映射机制及带信任度的灵活委托授权模型 FDMTPM,该模型充分考虑角色成员关系的动态委托提出了模式度量角色的概念,模式度量角色对委托方式进行控制,既实现了灵活的委托需求又保证了安全需求。本文通过为角色/权限分配关联信任度阈值的方法较好地反映了角色中权限的不同敏感度,同时进行信任度的委托,实现了细粒度的委托,保证了模型在多步委托过程中委托能力的收敛性,提高了多步委托过程中权限传播的可控性。为保证模型的完整性,本文对模式约束机制以及委托撤销机制进行了研究。最后,给出了 FDMTPM 模型的一个具体应用实例,体现了模型的实用价值。

本文在完善和优化 RDBMPM 模型的基础上提出 FDMTPM 模型,该模型既继承了 RDBMPM 模型的优点,又具有 RDBMPM 模型等其他模型所不具有的委托灵活的和广泛的适应度。本文并没有对度量化算子构造方法进行很具体的研究,这将是笔者下一步的研究工作。此外,基于模式度量角色的可控委托模型的职责分离约束也是下一步考虑解决的问题,同时如何更合理地计算产生出各个角色的信任度阈值也是后续需要研究的问题。

参考文献:

[1] LI Ning-hui, WINSBOROUGH W H, MITCHELL J C. Distributed credential chain discovery in trust management (full version)[C]// Proc of the 8th ACM Conference on Computer and Communications Security. New York: ACM Press, 2001:156-165.

[2] LI Ning-hui, MITCHELL J C, WINSBOROUGH W H. Design of a role-based trust management framework[C]//Proc of IEEE Symposium on Security and Privacy. Washington DC: IEEE Computer Society,2002:114-130.

[3] 蔡伟鸿, 韦岗, 肖水. 基于映射机制的细粒度 RBAC 委托授权模型[J]. 电子学报,2010,38(8):1753-1758.

[4] BARKA E, SANDHU R. Framework for role-based delegation models [C]//Proc of the 16th Annual Computer Security Application Conference. Washington DC: IEEE Computer Society,2000:168-176.

[5] BARKA E, SANDHU R. A role-based delegation model and some extensions[C]//Proc of the 23rd National Information Systems Security Conference. 2000.

[6] BARKA E, SANDHU R. Role-based delegation model/ hierarchical roles (RBDM1)[C]//Proc of the 20th Annual Computer Security Applications Conference. Washington DC: IEEE Computer Society, 2004: 396-404.

[7] ZHANG Long-hua, AHN G-J, CHU B-T. A rule-based framework for role-based delegation and revocation[J]. ACM Trans on Information and System Security,2003,6(3): 404-441.

[8] ZHANG Xin-wen, OH S, SANDHU R S. PBDM: a flexible delegation model in RBAC[C]//Proc of the 8th ACM Symposium on Access Control Models and Technologies. New York: ACM Press,2003: 149-157.

[9] WAINER J, KUMAR A. A fine-grained, controllable user-to-user delegation method in RBAC[C]//Proc of the 10th ACM Symposium on Access Control Models and Technologies. New York: ACM Press, 2005: 59-66.

[10] CRAMPTON J, KHAMBHAMMETTU H. Delegation in role-based access control[J]. International Journal of Information Security, 2008,7(2): 123-136.

[11] 赵青松, 孙玉芳, 孙波. RPRDM: 基于重复和部分角色的转授权模型[J]. 计算机研究与发展, 2003,40(2): 221-227.

[12] 翟征德. 基于量化角色的可控委托模型[J]. 计算机学报,2006, 29(8): 1401-1407.

[13] BLAZE M, FEIGENBAUM J, LACY J. Decentralized trust management[C]//Proc of the '96 IEEE Symp. on Security and Privacy. Washington DC: IEEE Computer Society,1996: 164-173.

[14] BLAZE M, FEIGENBAUM J, IOANNIDIS J, *et al*. IETF RFC 2704, The KeyNote trust-management system version 2[S]. 1999.

[15] BECKER M Y, SEWELL P. Cassandra: distributed access control policies with tunable expressiveness[C]//Proc of the 5th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY 2004). Los Alamitos: IEEE Computer Society,2004: 159-168.

[16] LI Ning-hui, GROSOF B N, FEIGENBAUM J. Delegation logic: a logic-based approach to distributed authorization[J]. ACM Trans on Information and System Security,2003,6(1): 128-171.

[17] HONG Fan, ZHU Xian, WANG Shao-bin. Delegation depth control in trust-management system[C]//Proc of the 19th International Conference on Advanced Information Networking and Applications. Washington DC: IEEE Computer Society, 2005: 411-414.

(上接第 4606 页)

[2] FRANEK F, JENNINGS C G, SMYTH P W F. A simple fast hybrid pattern-matching algorithm[J]. Journal of Discrete Algorithms, 2007,4(5):682-695.

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

[4] 杨武,方滨兴,云晓春,等. 入侵检测系统中高效模式匹配算法的研究[J]. 计算机工程,2004,30(13):92-94.

[5] AHO A V, CORASICK M J. Efficient string matching: an aid to are independently specified[J]. TODS,1979,4(2):168-179.

[6] 冉占军,姚全珠,王晓峰,等. 模式匹配算法在入侵检测中的应用[J]. 现代电子技术,2009,32(2):63-67.

[7] LEVENSHETEIN V I. Binary codes capable of correcting spurious insertions and deletions of ones[J]. Problems of Information Transmission,1965,1(1):8-17.

[8] 卢汪节,鞠时光. 入侵检测系统中一种改进的 AC 算法[J]. 计算机工程与应用,2006,42(15):146-148.

[9] AGRAWAL R, IMIELINSKI T, SWAMI A, *et al*. Mining assocation rules between sets of items in large databases[C]//Proc ACM SIGMOD Conference on Management of Data. 1993:207-216.

[10] 刘步中. 基于频繁项集挖掘算法的改进与研究[J]. 计算机应用研究,2012,29(2):475-477.

[11] AGRAWAL R, IMIELINSKI T, SWAMI A, *et al*. Database mining: a performance perspective[J]. IEEE Trans on Knowledge and Data Engineering,1993,5(6): 914-925.

[12] HAN Jia-wei, PEI Jian, YIN Yi-wen, *et al*. Mining frequent patterns without candidate generation[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press,2000:1-12.

[13] 黄鹤. 关联规则算法综述[J]. 软件导刊, 2009,8(3):56-58.

[14] 金凌. 面向比特流的未知帧头识别技术研究[D]. 上海:上海交通大学,2010.

[15] FAN Jang-jong, SU K Y. An efficient algorithm for match multiple patterns[J]. IEEE Trans on Knowledge and Data Engineering, 1993,5(2):339-351.