面向比特流的未知短波协议识别技术①

牛 欢, 卢选民

(西北工业大学 电子信息学院, 西安 710129)

摘 要: 协议识别技术在信息对抗中发挥着极其重要的作用, 它是对信号进行解码的前提条件, 是通信对抗由信号层对抗, 转变为信号层与信息层对抗互相结合, 以信息层对抗为主的关键一步. 从海量比特流数据中识别未知协议的基本方法是对比特流数据进行挖掘, 寻找其中的特征序列, 在没有先验知识的情况下, 则需要对其中的频繁序列进行提取. 为适应比特流环境, 本文在 BNDM 算法的基础上做出改进, 进行前置编码, 极大地提高了二进制环境下搜寻频繁序列的效率. 实验表明, 上述方法能够实现海量比特流数据中对未知短波协议的识别以及对协议数据帧的定界和切分.

关键词: 比特流; 协议识别; 短波; 模式匹配; 数据挖掘

Identification Technology of Unknown Shortwave Protocol for Bit Stream

NIU Huan, LU Xuan-Min

(School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: As the foundation of further signal decoding, protocol identification technique plays a very important role in the information countermeasures. Furthermore, it is a key step for information countermeasures to evolve from signal layer to signal layer combined with information layer. The basic approach of unknown protocols identificated from massive bit-stream data is the bit stream data mining, and looking for information which can determine the type of protocol. In the case of lacking of prior knowledge, frequent pattern sequence appearing in the bit stream data needs to be extracted, and sequence that can identify the type of protocol should be screened out. In order to adapt to the environment of the bit stream, this paper makes an improvement based on the BNDM algorithm, and improves the efficiency of searching the frequent sequences in the binary environment. The experimental results show that unknown protocol identification, protocol data frame alignment and segmentation from massive bit-stream data are realized through the research results of this thesis.

Key words: bit stream; protocol identification; shortwave; pattern matching; data mining

按照国际无线电咨询委员会的划分, 短波是指其波长在 100m~10m, 频率为 3MHz~30MHz 的电磁波. 工作频段在此范围之内的无线电通信称之为短波通信. 相较于卫星通信、微波通信、光缆等通信方式, 短波通信在不建设中继站的条件下, 就能够实现远距离通信. 它有着许多较为显著的优点, 诸如使用时无需支付话费, 建设和维护费用低; 设备简单, 目标小, 容易隐蔽, 对自然灾害或战争的抗毁能力强, 且损毁后易修复; 造价低, 可大规模装备, 系统顽存性强; 电路调度较为容易, 灵活性较强, 既可固定通信, 又可由人背负或装入其他器械, 进行移动通信等[1]. 因而, 短波通信在现代战争中发挥着十分重要的作用.

如图 1 所示, 现代战场中的短波通信网络是由许多自主工作, 彼此之间经过通信链路相连接的通信节点构成. 在对短波通信信道进行侦察后, 可以捕获到相应的短波信号, 进而采用相应的技术手段对信号进

① 基金项目:2015 年西北工业大学本科毕业设计(论文)重点扶持项目(GDKY9005)
收稿时间:2015-06-02;收到修改稿时间:2015-09-08

行处理, 就可以得到一些比特信息[2-4]. 正常情况下, 通信双方所采用的短波通信协议是非公开的. 因此, 对于信道侦听所得到的比特流数据, 如何对其进行数据分析进而识别出其所使用的通信协议, 对通信侦察的顺利完成有着决定性的影响.

现有的协议识别技术, 其绝大部分只是在信号层面上进行检测, 分析所捕获信号的一些参数或者针对协议中的某一个特征来进行处理, 如基于二阶循环统计量的调制方式识别、基于端口扫描的协议识别方式以及基于观测统计特性的检测识别方法等[4]. 上述的协议识别技术一方面主要应用在有线通信中, 另一方面大都是针对已知的协议来实现, 并不能在比特流场景中对非常规的专用未知协议进行识别. 为解决上述问题, 本文从比特流层面的数据挖掘出发, 根据短波通信协议的特点, 在传统的单模式识别算法基础上做出改进, 从而对短波协议的识别提供一定的支持.

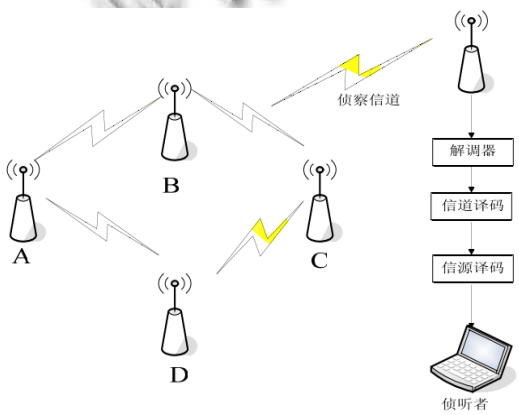

图 1 短波通信中数据的捕获过程

1 频繁序列的提取

在同种未知协议的原始数据大量累积的条件下, 面向比特流的未知协议识别的目标是从中寻找代表短波协议特征的特征序列[5]. 因此, 首先应通过面向比特流的频繁集挖掘算法, 提取其特征序列, 然后以该特征序列作为帧定位的标志对比特流数据进行切割. 上述的两个场景中, 前者主要体现为序列筛选问题, 可以采用改进的多模式匹配算法,将算法中匹配的过程替换为模式串的统计过程. 后者则是设法从比特流中正确的分离出每一个完整无误的帧, 进而确定其采用的协议格式. 帧提取问题也可以被转化为单模式串匹配问题: 协议同步码序列既定, 使用单模式匹配算法从比特流中找出同步码序列所有的位置, 则相邻两个特征序列之间就是一数据帧[6].

针对所捕获的比特流数据, 要对其所采用的协议进行识别, 首先需要在海量的比特流数据中提取此种协议的特征序列. 由前述内容可知, 可以利用同步码检测的方法来进行协议识别. 考虑到现有的帧同步检测技术大多是对已知的同步码进行检测, 因此, 本文主要研究如何利用频繁序列挖掘技术来进行未知同步码检测.

考虑到比特流数据具有单一性、顺序性两个明显的特征, 不能直接使用经典的数据挖掘方法处理纯比特流数据, 故需要进行一定的改进[7]. 进一步深入考虑, 在没有任何协议特征信息的情况下, 必须对所有的 2 进制序列进行匹配. 本质上来说, 是对所有的 2 进制模式序列的出现情况做全统计, 以备下一阶段的挖掘运用. 传统意义上的单模式匹配一次只能寻找一个特定的模式序列, 在匹配目标未知的前提条件下, 对源数据的扫描次数取决于候选模式的存在数目, 对于大量的源数据, 姑且不论匹配算法本身的效率如何, 磁盘 I/O 的代价已然相当可观, 因此, 此环境下使用单模式匹配算法是不明智的[8]. 而多模式匹配则能够以一次扫描寻找到多个模式, 这与比特流模式序列的统计的目标是一致的.

比较典型的短波协议有业余无线电封包通信的 AX.25 链接协议、D-STAR 数字通信协议以及 AIS 协议. 以 D-STAR 为例, 其数据包报头部分结构如图 2 所示.

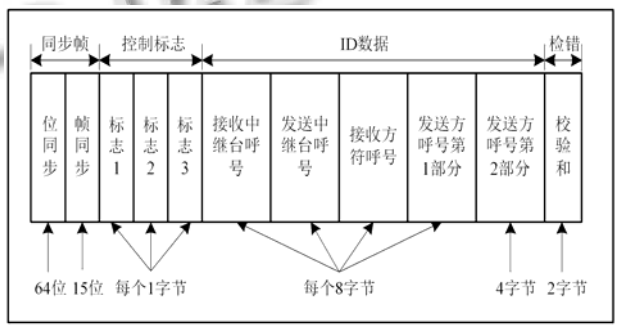

图 2 D-STAR 数据包报头部分的结构

表 1 列出了上述三种短波通信协议的同步码序列. 通过对多个短波协议进行分析研究, 能够发现, 充当帧同步段的特征序列, 其长度普遍不超过 128 比特. 因此, 可以做出这样的经验假设: 同步码序列长度范围为 2~128bits.

表 1　三种经典短波协议的同步码序列

协议类型	AX.25	AIS	D-STAR
同步码序列	01111110	010101010101 0101 010101011010 1010	1010 或 1001 111011001 010000
长度	8	32	79

基于以上分析, 在捕获的未知短波协议比特数据流中, 同步码序列出现的概率 P 有{P|O≤P≤1)}, 此同步序列长度 $m\in[2,128]$. 对于长度为 m 的比特序列, 对其进行频繁序列统计时的枚举空间为 2^m, 且空间复杂度与 m 存在指数关系. 如果对所有的比特序列进行一次性遍历, 则其空间复杂度为 $O(2^2\sim2^{128})$, 这是无法接受的. 继续分析, 如果首先对长度 m=4 的比特序列进行统计, 然后通过设定最小阈值过滤得到短频繁序列, 再通过一定的关联规则, 就可以将长度为 4 的短频繁序列拼接为长度 8 的频繁序列, 依此规则进行下去, 直到获得最终的同步码序列为止[9]. 相对于一次性遍历目标长频繁序列而言, 采用这种思路的好处是, 拼接获得的长频繁序列是建立在筛选过滤的短频繁序列基础之上的, 因此可以极大地提升工作效率.

为了验证算法的有效性及正确性, 编程利用改进的 AC 算法实现对短频繁序列进行统计, 然后通过拼接算法获得长频繁序列这一过程[10].

测试采用 AX.25 协议以及 D-STAR 协议分别解调后的数据文件, 其结果具有一致性, 这里以 AX.25 的数据文件为例进行说明. 表 2 为拼接得到的同步码候选列表, 这些长序列是一系列近似串. 显然, 标准同步序列应为编号为 1 的序列: 01111110(0x7E).

表 2　拼接后所得到的长频繁序列

编号	位数	拼接的长序列	出现概率(%)
1	8	01111110(0x7E)	100
2	8	01111001(0x79)	86.9
3	8	01111111(0x7F)	95.7
4	8	01111010(0x7A)	88.2
5	8	01111100(0x7C)	91.5
6	8	01110100(0x74)	83.6

图 3 是进行分频繁序列拼接后的结果, 有六个子串是标准同步码的近似串. 实验结果表明, 利用长字符串拼接技术对于同步码的识别率可以达到 94.1%, 能够比较准确地在比特流数据中对同步码进行识别.

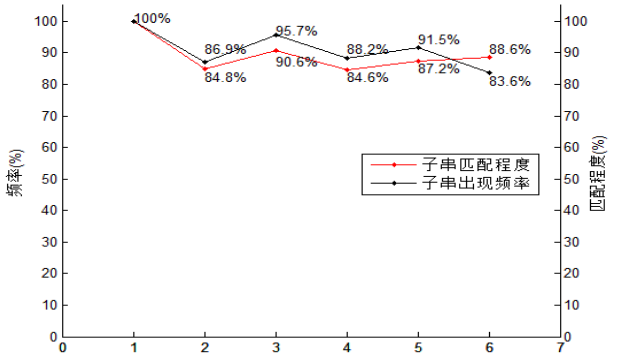

图 3　拼接后得到的频繁串匹配程度

2　数据帧定界与切分

2.1　单模式匹配算法效能分析与改进

在上述内容的基础上, 利用已经提取出的特征序列对比特流数据进行帧定界. 为此, 首先对应用较广的几种单模式匹配算法进行性能测试, 测试文本随机生成, 共 10M 字节, 模式串也使用同样的方法生成. 实验不断重复进行, 直到在 95%的置信度下相对误差低于 2%. 所有算法都采用了最优的实现, 但结果只有 Shift-And, Horspool, BNDM 和 BOM 算法在图上有相应的分布区域, 其他算法则因为太慢而没有在图上显示. 图 4 是实验结果图, 可见序列长度不超过 128 的前提下, BNDM 算法是比较合适的.

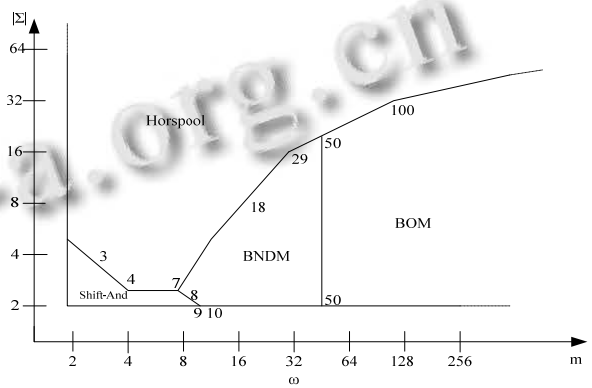

图 4　单模式匹配算法实验图

然而, 上述测试文本的字符集为{a, b, ……, z}, 而比特流中所包含的元素非 0 即 1, 并且模式串和目标串的字符集大小相同, 均为{0,1}, 远小于测试文本字符集. 在此环境下, 传统的单模式匹配算法效率十分低下. 依据比特流序列的特点, 需要对传统单模式匹配算法进行优化改进以提高在比特流环境下进行匹配的效率.

为此, 考虑从扩充字符集的方向对算法进行优化, 采用预编码的思想来进行实现. 编码的目标是在扩充字符集的同时能够对比特流文件进行压缩, 从而缩短目标串的长度, 提升匹配效率[11]. 由于目标串的长度远远大于模式串的长度, 因此, 编码后目标串的字符集必然大于模式串的字符集, 从而解决了模式匹配算法在比特流环境下效率低下的问题.

哈夫曼编码采用构造最优前缀码的贪心算法, 从而使得编码后的二进制序列能够达到最短. 考虑到此编码方法的前缀性质, 就可以使其译码方法变得非常简单. 这是由于任一字符的代码都不会是其他字符代码的前缀, 只要从编码文件中不断取出代表某一字符的前缀, 再转换为原字符, 即可逐个译出原文件中的所有字符. 这样, 一个任意长的哈夫曼编码序列可以被唯一地翻译为一个字符序列.

基于哈夫曼编码算法的优势, 本文采用逆向的思想, 将目标串和模式串看作是采用哈夫曼编码后的最短二进制序列, 通过对目标串和模式串的译码来达到编码的效果. 采用此种“编码”方法, 可以实现扩大字符集和使“编码”后的序列长度最短的目标, 避免编码对模式匹配算法效率的影响[12].

译码时采用线性链表存储字符序列, 具体实现如下:

```
template<class CharType,class weightType>

LinkList<CharType>HuffmanTree<CharType,WeightType>::DeCode(String strCode)  //对编码串 strCode 进行译码, 返回编码前的字符序列
{
LinkList<CharType>charList;
for(int pos=0;pos<strCode.Length();pos++)
{
                                    if(strCode[pos]=='0')
curPos=nodes[curPos].leftChild; //'0'表示左分支
        else
   curPos=nodes[curPos].rightChild; //'1'示右分支
if(nodes[curPos].leftChild==0&&nodes[curPOS].rightChild==0)
        {
charList.Insert(charList.Length()+1,LeafChars[curPos]);
        curPos=2*num-1;    //curPos 回归根结点
        }
}
return charList;        // 返回编码前的字符序列
}
```

2.2 前置编码 BNDM 算法效率分析

图 5 为目标比特流 01010000110110101 在编码前后的对比图, 为了验证在真实场景下, 编码 BNDM 算法同其他模式匹配算法的差异, 进行了多组实验测试. 实验数据以 AX.25 协议解调译码后比特流数据文件为例, 并分别在大小为 1Kb、10Kb、100Kb 三种情况下进行测试. 其中 AX.25 协议特征串为: 01111110,对照组取 KMP 算法、BM 算法和 BNDM 算法. 实验结果如表 3 所示.

表 3 AX.25 协议特征串检索测试(单位: ms)

文件大小	1Kb	10Kb	100Kb
算法	耗时	耗时	耗时
KMP	8036	80572	330312
BM	3266	12616	135943
BNDM	3984	16504	160336
前置编码	3058	12278	126878

由图可见, 前置编码 BNDM 算法相对 KMP、BM 和 BNDM 算法, 效率有了很大的提升, 并且数据文件越大, 其表现越明显. 相对于其它算法, 前置编码 BNDM 算法在编码阶段需要消耗一定的时间, 但是由于在匹配过程中节省了大量时间, 弥补了预处理阶段付出的时间代价, 因此总耗时仍然小于其它算法.

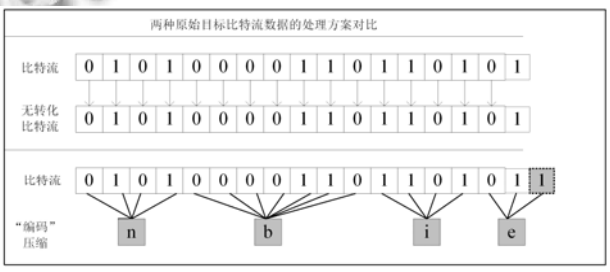

图 5 比特流编码过程示意图

可以看出, 采用哈夫曼译码的编码方法后, 解决了模式匹配算法在比特流环境下效能低下的问题. 基于此, 就可以采用前置编码 BNDM 算法进行帧定界切分. 图 6 给出了 AX.25 协议比特流数据的测试结果, 表明本文提出的方法是可行的.

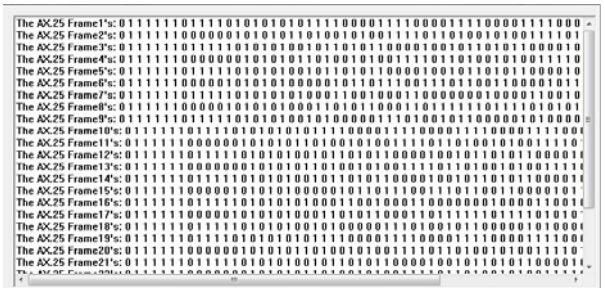

图 6 AX.25 协议比特流帧定位与切分

3 结语

本文所提出的未知短波协议识别技术, 主要是指在海量比特流数据中提取未知协议的特征序列, 并以此对其进行帧定界和切分. 考虑到单模式匹配算法在比特流环境下效能低下的问题, 本文将哈夫曼译码运用于二进制序列的编码, 弥补了单模式匹配算法在比特流场景中的缺陷. 实验结果表明, 本文提出的面向比特流的未知短波协议识别技术是可行的, 对目前相关研究较少的比特流数据分析有一定借鉴意义, 但对于数据帧切割之后的帧内结构分析, 有待于进一步研究.

参考文献

1 王坦.短波通信系统.北京:电子工业出版社,2008.
2 赵琦,刘荣科.编码理论.北京:北京航空航天大学出版社,2009.
3 何永君,舒辉,熊小兵.基于动态二进制分析的网络协议逆向解析.计算机工程,2010,36(9):268–270.
4 夏晓巍,方旭明,黄巍等.帧同步技术的研究与展望.信息安全与通信保密,2006,1(7):140–143.
5 聂东举,叶进,闫坤,等.基于算法的短波通信协议识别技术.系统工程与电子技术,2013,35(6):1307–1311.
6 孔东林,罗向阳,邓崎皓,等.基于 AC 自动机匹配算法的入侵检测系统研究.微电子学与计机,2005,22(3):89–95.
7 Zhou X, Wu Y. Signal modulation recognition based on KPCA and LAD. Systems Engineering and Electronics, 2011, 33(7): 1611–1616.
8 金凌.面向比特流的未知帧头识别技术研究[硕士学位论文].上海:上海交通大学,2011.
9 陈曙晖,苏金树.基于内容分析的协议识别研究.国防科技大学学报,2008,30(4):82–87.
10 林祎,彭华,赵振华.基于小波降噪的短波通信信号协议识别特征提取算法.信息工程大学学报,2012,13(4):438–442.
11 贺瑶,王文庆,薛飞.基于云计算的海量数据挖掘研究.计算机技术与发展,2013,23(2):69–72.
12 陆琳琳,田野.基于确定有限状态自动机的改进多模式匹配算法研究.计算机应用与软件,2013,30(7):321–323.