Journal of Computer Applications ISSN 1001-9081 2018-01-26
计算机应用 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(****)**-0000-00 doi:10.11772/j.issn.1001-9081.2017112846

基于最佳路径搜索的二进制协议格式关键词边界确定方法

闫小勇，李青

(信息工程大学 信息系统工程学院，郑州 450000)

(*通信作者电子邮箱 yanxiaoyong2016@163.com)

摘 要: 针对二进制协议报文格式逆向分析中字段切分问题，提出以格式关键词为逆向分析目标，通过改进的 n-gram 算法和最佳路径搜索算法，实现对二进制协议格式关键词的最优定界。将位置因素引入 n-gram 算法，提出基于迭代 n-gram-position 的格式关键词边界提取算法，有效解决了 n-gram 算法中 n 值不易确定和固定偏移位置格式关键词的边界提取问题；定义了频繁项边界命中率和左右分支信息熵为基础的分支度量，以关键词和非关键词的 n-gram-position 取值变化率存在差异为基础构造约束条件，提出基于最佳路径搜索的格式关键词边界选择算法，实现了对格式关键词的联合最优定界。通过在 AIS1、AIS18、ICMP00、ICMP03 和 NetBios 五种不同类型协议报文数据集上测试，本文提出算法的 F 值均在 83%以上。与 VDV 和 AutoReEngine 经典算法相比，F 值平均提升约 8%。

关键词: 二进制协议；格式关键词；边界确定；n-gram；最佳路径搜索

中图分类号:TP393 **文献标志码**：**A**

Method for Determining the Boundaries of Binary Protocol Format Keywords based on Best Path Search

YAN Xiaoyong, LI Qing

(School of Information System Engineering, Information Engineering University, Zhengzhou 450000, China)

Abstract: Aiming at the difficult problem that field segmentation in binary protocol reverse engineering, a novel algorithm with format keywords as the goal is proposed, which can optimally determine the boundaries of binary protocol format keywords by the improved n-gram algorithm and the best path search algorithm. By introducing position factor into the n-gram algorithm, the iterative n-gram-position algorithm is proposed to extract the candidate boundaries of format keywords, which can solve the problem that n is difficult to determine in the n-gram algorithm and the candidate boundaries extraction of format keywords with fixed offset position. The optimal path search algorithm is used to select the optimal boundaries from candidate boundaries. The branch metric of optimal path search algorithm is based on the hit ratio of frequent item boundaries and the left and right branch entropy. The constraint of the optimal path search algorithm is based on the difference of value change rate between keywords and non-keywords. By testing on AIS1, AIS18, ICMP00, ICMP03 and NetBios, the F values of our algorithm are all above 83%. Compared with VDV and AutoReEngine, the F value is increased averagely by about 8%.

Keywords: binary protocol; protocol format keyword; boundary determining; n-gram; optimal path search

0 引言

二进制协议结构紧凑，控制开销小，传输效率高，因而得到广泛应用。尤其在物联网中，二进制协议应用主导地位更为突出。二进制协议字段不受字节长度限制，不使用公开字符编码，使得二进制协议分析具有很高的难度。二进制协议逆向成为协议逆向工程的难题。在没有协议先验知识的条件下，二进制协议的字段切分十分困难，往往只能得到字段的组合；确切的字段边界分析很难做到，往往只能得到概率意义下的边界。为此本文提出从关键词的角度进行二进制协议逆向分析。

目前仅有少量研究是面向二进制协议，基于网络流量进行报文格式逆向分析。Tao[1]等改进多序列比对算法，使其适用于二进制协议，利用贝叶斯决策和最大似然准则实现二进制协议字段定界。多序列比对算法能够解决可变字段和不可变字段之间的定界问题，但对于可变字段和可变字段以及不可变字段和不可变字段之间的定界，因为缺少语义信息，很难实现。虽然作者最终利用贝叶斯决策和最大似然准则很大程度上提升了字段定界的准确性，但推断结果中仍存在字段组合问题，即多个字段被误判为一个字段。Li 等[2]设计了一种抗误码的未知二进制协议解析方法，提出模糊加权的多序列比对算法，在推断结果中同样会出现多个字段被误判为一个字段的情况。孟凡志等[3]提出基于概率比对的通信协议

收稿日期: 2017-12-05； 修回日期:2018-01-09； 录用日期:2018-01-15。基金项目:

作者简介: 闫小勇(1993—)，男（汉族），陕西陇县人，硕士研究生，主要研究方向：数据挖掘、协议逆向；李青(1976—)，女（汉族），河北正定人，副教授，博士，主要研究方向：协议逆向、可见光通信、无线自组织网、传感网。

本刊网络出版时间 2018-01-26。 http://www.joca.cn/CN/10.11772/j.issn.1001-9081.201711284;
知网网络出版时间 2018-01-26 14:26:35。 http://kns.cnki.net/kcms/detail/51.1307.TP.20180126.1426.004.html。

格式逆向分析方法。通过概率比对算法使字段准确对齐，再通过特征统计量的差异性进行字段分割，该方法同样会出现字段组合问题。上述方法均以获取二进制协议报文格式为最终目标，但因先验信息缺失，往往只能得到字段组合，同时字段边界也只是概率意义下的字段边界。

关键词与报文格式之间存在联系。对于文本类协议，关键词序列可以作为报文格式。Wang[4]等利用 n-gram 算法对同种应用协议数据做分解，提取协议关键词，用关键词序列代替协议报文做报文聚类，最后用多序列比对算法推断报文格式。Luo[5]等提出基于位置的协议逆向方法 AutoReEngine，利用 Apriori 算法提取频繁字符串，并对频繁字符串做基于位置的方差统计，方差小的作为协议关键词，最终协议报文格式为提取的关键词序列。黎敏等[6]建立隐半马尔科夫模型，描述协议报文字段之间的关系，通过最大似然概率分段方法实现报文字段的最佳划分。文中将报文格式简化为“关键词+变量字段”的分段形式，本质上是以关键词序列作为协议报文格式。以上研究均针对文本类协议，通过提取协议关键词序列近似协议报文格式。

对于二进制协议，字段组合灵活，很少出现关键词和变量字段交替出现的情形，因而关键词序列不等价于协议报文格式。如何从关键词角度进行二进制协议报文格式逆向分析成为亟待解决的问题。

本文面向二进制私有协议数据逆向分析，提出了一种基于最佳路径搜索的二进制协议格式关键词边界确定算法(Method for Determining the Boundaries of Binary Protocol Format Keywords based on Optimal Path Search, OBPFK)，实现了以关键词为核心的协议逆向分析。主要贡献有：1）将协议关键词做进一步划分，提出协议分类关键词和协议格式关键词定义；2）提出迭代 n-gram-position 算法，有效解决了 n-gram 算法中 n 值不易确定和固定偏移位置格式关键词的边界提取问题；3）利用最佳路径搜索算法实现了对格式关键词的联合最优定界。

本文剩余部分组织如下：第一节问题描述；第二节详细阐述算法细节；第三节对算法本身进行仿真与实验分析，并与经典算法对比；第四节结论。

1 问题描述

1.1 格式关键词

协议报文格式可以看作是字段序列，协议字段是具有特定语义的最小不可分割子序列[7]。同种类型协议报文的字段集合记为 $\boldsymbol{FD}=\{fd_1, fd_2, \mathbf{L}, fd_i, \mathbf{L}, fd_g\}$，其中 $fd_i\,(1\le i\le g)$ 为字段。一个协议报文可以唯一的划分为 g 个不相交字段。

绝大部分网络协议报文中存在协议键词，协议关键词是指满足一定条件（位置和频度）的字符串/比特串[8]。同种类型协议报文的关键词集合记为 $\boldsymbol{KW}=\{kw_1, kw_2, \mathbf{L}, kw_i, \mathbf{L}, kw_t\}$，其中 $kw_i\,(1\le i\le t)$ 为关键词。

协议关键词不同于协议字段，如图 1 所示。借用圆与圆之间的位置关系来描述：同一 $\boldsymbol{FD}$ 中，相邻字段位置仅存在相切关系；同一 $\boldsymbol{KW}$ 中，相邻关键词位置存在相交、相切和相离三种关系。

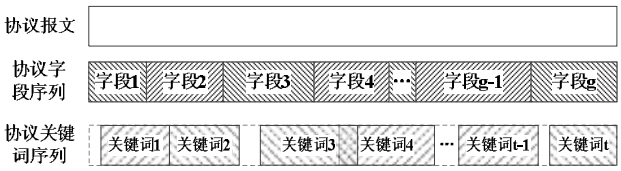

图 1 协议字段和协议关键词

Fig.1 Protocol fields and protocol keywords

协议关键词在协议数据分析中具有重要的作用，既可以用于区分不同协议报文，也可以作为协议字段，成为协议报文格式的一部分。根据其作用不同，本文将协议关键词分为分类关键词、格式关键词和其它关键词。分类关键词和格式关键词的定义如下：

定义 1：给定报文集合中不同类型协议报文的关键词集合为 $\boldsymbol{PKW}=\{\boldsymbol{KW_1}, \boldsymbol{KW_2}, \mathbf{L}, \boldsymbol{KW_i}, \mathbf{L}, \boldsymbol{KW_u}\}$，$\boldsymbol{KW_i}=\{kw_{i1}, kw_{i2}, \mathbf{L}, kw_{ij}, \mathbf{L}, kw_{it}\}$ 表示同种类型协议报文的关键词集合。$kw_{ij}\in\boldsymbol{KW_i}$，若 $kw_{ij}\notin\forall\boldsymbol{KW}_w\,(w\in(1,u), w\ne i)$，则 kw_{ij} 为 $\boldsymbol{KW_i}$ 对应协议的一个分类关键词。

定义 2：同种类型协议报文的字段集合为 $\boldsymbol{FD}=\{fd_1, fd_2, \mathbf{L}, fd_i, \mathbf{L}\ fd_g\}$，相应的关键词集合为 $\boldsymbol{KW}=\{kw_1, kw_2, \mathbf{L}, kw_j, \mathbf{L}, kw_t\}$。若 $kw_j=\sum_{i=i_1}^{i_2} fd_i$，则 kw_j 为该协议的一个格式关键词。

分类关键词与数据集中包含的协议有关，假设数据集中有两种不同类型协议报文，关键词 kw 只存在于一种类型的协议报文中，那么 kw 就是该种类型协议报文的一个分类关键词。格式关键词是协议字段，或者协议字段的组合。分类关键词集合和格式关键词集合可能存在交集，即协议关键词是分类关键词的同时，也可能是格式关键词。其它关键词既不是分类关键词，也不是格式关键词。格式关键词提取要求数据集纯度高，数据多样性好，覆盖时空分布。分类关键词提取对数据集样本没有特殊的要求。

协议报文格式可以看作协议字段序列，协议格式关键词序列不等价于协议报文格式，属于报文格式的一个子集。报文格式的形式会影响格式关键词边界的确定，以文本类协议和二进制协议为例。文本类协议的字段通常为特定的词，这些词由 ASCII 字符组成，易于理解。字段边界为预定义的分割符（例如空格或者回车换行），因此格式关键词的边界易于确定。二进制协议结构紧凑，字段之间没有分割符，面向私有协议时，因缺乏先验信息，格式关键词边界确定较为困难。

1.2 问题模型

二进制协议数据帧的最小单元为比特，值域 $\boldsymbol{D}=\{\backslash 0b0,\backslash 0b1\}$。第 k 个格式关键词 kf_k 表示为 $kf_k=b_{kfs_k}b_{kfs_k+1}\mathbf{L}\,b_l\,\mathbf{L}\,b_{kfe_k}, b_l\in\boldsymbol{D}$，$l$ 表示比特在协议数据帧中的偏移位置，kfs_k 和 kfe_k 分别为 kf_k 的起始边界点和终止边界点。二进制协议数据帧由格式关键词可以表示为 $fr_k=kf_1\,\|\,kf_2\,\|\,\mathbf{L}\,\|\,kf_k\,\|\,\mathbf{L}\,\|\,kf_m$，‘‖’是串行的意思。二进制协议数据帧集合可以表示为 $\boldsymbol{Fr}=\{fr_1,fr_2,\mathbf{L},fr_i,\mathbf{L},fr_n\}$。二进制协议格式关键词边界确定的最终目标是要解决一个后验概率问题 $p(i=kfs_k \text{ or } i=kfe_k\,|\,b_i)$，即在已知第 i 比特位观测值 b_i 的条件下，该比特位是格式关键词起始边界点或者终止边界点的概率。

格式关键词边界确定涉及两个关键问题：格式关键词边界提取和格式关键词边界选择。格式关键词边界提取结果为候选格式关键词边界，往往存在误警和虚警问题。关键词边界选择能够解决误警和虚警问题，筛选出与真实格式关键词边界最吻合的子集作为格式关键词边界确定的最终结果。

本文输入数据为同种类型二进制协议数据帧，第 k 个格式关键词边界在协议数据帧中的偏移位置为 kfs_k 和 kfe_k。二进制协议格式关键词边界确定问题可以描述为：在格式关键词数量 m 未知的条件下，对格式关键词边界 $\{kfs_1,kfe_1,kfs_2,kfe_2,\mathbf{L},kfs_m,kfe_m\}$ 的最佳估计。OBPFK 算法的系统框图如图 2 所示，具体步骤如下：

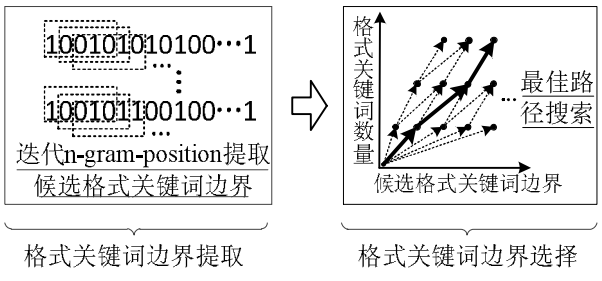

图 2 OBPFK 算法系统框图

Fig.2 System diagram of OBPFK

步骤 1：由迭代 n-gram-position 算法提取频繁项的边界 $\{fitem_1,fitem_2,\mathbf{L},fitem_p\}$ 作为候选格式关键词边界。

步骤 2：以 $\{fitem_1,fitem_2,\mathbf{L},fitem_p\}$ 中各比特位的频繁项边界命中率和左右分支信息熵之差为基础构造分支度量。以协议关键词和非协议关键词的 n-gram-position 取值变化率存在差异为基础构造约束条件。通过最佳路径搜索算法从候选格式关键词边界中筛选出与真实格式关键词边界最接近序列。

2 基于最佳路径搜索的二进制协议格式关键词边界确定方法

协议格式关键词边界确定受数据集样本多样性影响。因为时间和空间限制，获取的数据集样本可能存在多样性不足的问题，一些原本不是格式关键词的比特串被当作格式关键词提取出来，给结果带来误差。例如，AIS 协议[9]中经度和维度的高位在短时间内不会发生变化，如果仅用短时间内的数据样本做测试，提取的格式关键词中将包含经纬度高位，而经纬度高位不是真实的协议格式关键词。本文假设用于测试的数据集样本具有足够的多样性。

2.1 基于迭代 n-gram-positon 的格式关键词边界提取算法

由于二进制协议字段通常具有固定的长度和偏移[10]，使得二进制协议格式关键词一般也具有固定长度和偏移。本文在 n-gram 算法[11]基础上引入位置因素提出 n-gram-positon 算法，用于提取二进制协议数据帧的频繁项。同时，为了克服 n-gram-position 算法中 n 值不易确定的问题，采用不同 n 值的 n-gram-position 算法迭代提取频繁项，并将所有频繁项边界点作为候选格式关键词边界点。

n-gram 算法输入为协议报文 $fr=b_1b_2\,\mathbf{L}\,b_i\,\mathbf{L}\,b_s$，$b_i\in D$，输出 n-gram 为：$b_1b_2\,\mathbf{L}\,b_n, b_2b_3\,\mathbf{L}\,b_{n+1},\mathbf{L},b_{s-n+1}b_{s-n+2}\,\mathbf{L}\,b_s$。例如序列‘10001010’的所有 3-gram 为：‘100’，‘000’，‘001’，‘010’，‘101’。本文在原有算法基础上引入位置因素，提出 n-gram-position 算法。考虑位置因素后序列‘10001010’的所有 3-gram-position 为：‘100-1’，‘000-2’，‘001-3’，‘010-4’，‘101-5’，‘010-6’。n-gram-position 算法的输入为协议报文 $fr=b_1b_2\,\mathbf{L}\,b_i\,\mathbf{L}\,b_s$，$b_i\in D$，输出 n-gram-position 为：$b_1b_2\,\mathbf{L}\,b_n\text{-}1, b_2b_3\,\mathbf{L}\,b_{n+1}\text{-}2,\mathbf{L},b_{s-n+1}b_{s-n+2}\,\mathbf{L}\,b_s\text{-}(s-n+1)$。同一协议报文的 n-gram-position 不存在重复项。

n-gram-position 算法中最关键的是 n 值如何确定。文献[4]提出协议数据的 n-gram 近似服从齐普夫分布，并且 4-gram 最接近齐普夫分布，因此设置 n 为 4。本文为了防止不同协议数据下，单一 n 值的统计结果会出现偏差，做了不同 n 值的频繁项提取，并将所有频繁项的边界点作为候选格式关键词边界点。

本文设定最小支持度 minsupport，用于判定 n-gram-position 是否为频繁项，支持度的定义如下：

定义 3：给定报文集合 $\boldsymbol{Fr}$，$fr\in\boldsymbol{Fr}$ 是其中的一个报文，如果 fr 包含 n-gram，并且偏移位置为 position $\pm q$（q 为允许的误差范围），则 $fr(\text{n-gram-position})=1$，否则为 0。n-gram-position 在 $\boldsymbol{Fr}$ 上的支持度为：$\sup(\text{n-gram-position})=\sum_{fr\in\boldsymbol{Fr}} fr(\text{n-gram-position})/|\boldsymbol{Fr}|$，其中 $|\boldsymbol{Fr}|$ 为 $\boldsymbol{Fr}$ 中的报文总数。

构建候选格式关键词边界点的伪代码如下：

Begin
01. $fitem = \{\ \}$ //初始化
02. **for** $n = N_1$ *to* N_2 **do** //N_1, N_2为n取值的上下限
03. **for** i =1 *to* $s-n+1$ **do**
04. **if** sup $(n\text{-}gram-i) >$ minsupport **do**
05. $fitem = fitem \cup \{i, i+n-1\}$
06. **end**
07. **end**
08. **end**
End

2.2 基于最佳路径搜索的格式关键词边界选择算法

基于迭代 n-gram-position 提取的候选格式关键词边界点中往往存在误警和虚警情问题，需要通过格式关键词边界选择算法筛选出与真实边界最为接近的节点序列。利用候选格式关键词边界点可以构造有向图，如图 2 中最佳路径搜索部分。横轴表示候选格式关键词边界点，纵轴表示格式关键词数量，最终格式关键词边界可以表示为有向图中的一条路径（如有向图中实线所示）。通过最佳路径搜索算法可以找到这样一条路径。

2.2.1 分支度量与约束条件

1）分支度量

分支度量用于衡量有向图节点之间的转移权值。最佳路径搜索算法应用于格式关键词边界选择，最终目标是寻找一条从第一个候选边界点到最后一个候选边界点权值之和最小的路径。分支度量决定搜索路径的走向，是利用最佳路径搜索算法完成格式关键词边界选择的关键。本文在定义分支度量时考虑了两方面因素：频繁项边界命中率；左右分支信息熵。

频繁项边界命中率是指在迭代 n-gram-position 提取频繁项的结果中各比特位作为频繁项边界点出现的次数与迭代总次数的比值。用 $f_n^i = 1$ 表示协议数据帧第 i 比特位在 n-gram-position 提取频繁项过程中被确定为频繁项边界点，$f_n^i = 0$ 表示第 i 比特位没有被确定为频繁项边界点。式(1)表示第 i 比特位在迭代范围为 (N_1, N_2) 时被确定为频繁项边界点的总次数。第 i 比特位被确定为频繁项边界点的命中率如式(2)所示。

$$Z_i = \sum_{n=N_1}^{N_2} f_n^i \tag{1}$$

$$HR_i = Z_i / (N_2 - N_1 + 1) \tag{2}$$

HR 越大，其所对应的比特位越有可能成为格式关键词边界点。

左右分支信息熵反映了比特位左右两侧比特串的取值分布情况。基于协议关键词和非协议关键词的熵值存在差异，可以利用比特位左右分支信息熵的差值作为判定该比特位是否为格式关键词边界点的依据。左右分支信息熵包括左分支信息熵和右分支信息熵，计算公式如式(3)和式(4)。

$$H(L_r^i) = -\sum_{(b_{i-r}\,\mathbf{L}\,b_{i-1})\in Q_r} p(b_{i-r}\,\mathbf{L}\,b_{i-1}) \log p(b_{i-r}\,\mathbf{L}\,b_{i-1}) \tag{3}$$

$$H(R_r^i) = -\sum_{(b_{i+1}\,\mathbf{L}\,b_{i+r})\in Q_r} p(b_{i+1}\,\mathbf{L}\,b_{i+r}) \log p(b_{i+1}\,\mathbf{L}\,b_{i+r}) \tag{4}$$

$H(L_r^i)$ 表示第 i 比特位左侧长度为 r 的比特串信息熵，$H(R_r^i)$ 表示第 i 比特位右侧长度为 r 的比特串信息熵。Q_r 表示在协议数据帧集合中特定偏移位置处，长度为 r 的比特串的所有取值集合。因为长度 r 难以确定，为了尽可能避免 r 选择不当对结果造成的影响，本文计算了不同 r 取值的信息熵。为了便于比较，对不同 r 取值的信息熵做了归一化处理，最终的左右分支信息熵如式(5)和式(6)所示。

$$H(L^i) = \sum_{r=2}^{r_{th}} \frac{H(L_r^i)}{\log 2^r} = \sum_{r=2}^{r_{th}} \frac{H(L_r^i)}{r} \tag{5}$$

$$H(R^i) = \sum_{r=2}^{r_{th}} \frac{H(R_r^i)}{\log 2^r} = \sum_{r=2}^{r_{th}} \frac{H(R_r^i)}{r} \tag{6}$$

$$HD_i = \left| H(L^i) - H(R^i) \right| \tag{7}$$

比特位的左右分支信息熵之差越大，则该比特位成为格式关键词边界点的可能性就越大，即 HD_i 越大，第 i 比特位越有可能成为格式关键词边界点。

结合频繁项边界命中率以及左右分支信息熵，构造最佳路径搜索算法的分支度量，如式(8)。因为最佳路径搜索算法利用了最短路径搜索的思想，为了使最佳路径的权值之和最小，选择 HR_i 和 HD_i 加权之和的倒数作为分支度量。Bm_i 表示由其它节点到第 i 个节点的转移权值。Bm_i 越小，转移到第 i 个节点的可能性就越大。w_{HR} 和 w_{HD} 分别为 HR_i 和 HD_i 对应的权值，$w_{HR} + w_{HD} = 1$ 。

$$Bm_i = \frac{1}{w_{HR} \times HR_i + w_{HD} \times HD_i} \tag{8}$$

2）约束条件

约束条件控制节点之间是否可达，限制搜索路径的走向，同时约束条件能够压缩搜索空间，避免运算过程中遍历路径太多，导致运算量过大[12]。本文构造约束条件基于协议关键词和非协议关键词的 n-gram-position 具有不同的统计特性。以 n-gram-position 的取值变化率作为统计特性的衡量指标，如式(9)所示。

$$g_{(i,j,n)}^t = \frac{\left| v_{(i,j,n)}^t \right|}{G} \tag{9}$$

$\left| v_{(i,j,n)}^t \right|$ 表示数据集中所有报文在比特串 $b_i b_{i+1}\,\mathbf{L}\,b_j$ 的第 $t(1 \le t \le j-i-n+2)$ 个 n-gram-position 处的不同取值个数；

G 表示 n-gram-position 所有取值情况的个数，即 2^n；$g^t_{(i,j,n)}$ 为相应的取值变化率。

如图 3，协议数据中所有报文在比特串 $b_3b_4b_5b_6$ 的第 1 个 3-gram-position 处的不同取值个数 $\left|v^1_{(3,6,3)}\right|$ 为 2；G 为 3-gram-position 所有取值情况的个数，为 8；取值变化率 $g^1_{(3,6,3)}$ 为 0.25。

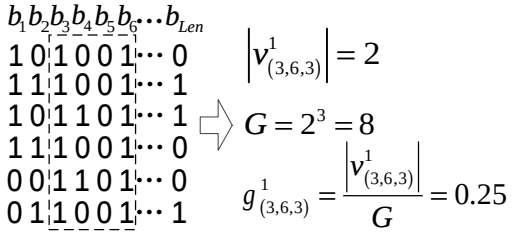

图 3 取值变化率

若两比特位之间比特串的 n-gram-position 取值变化率的方差大于设定阈值限 var_{th}，则认为 n-gram-position 取值变化率差异较大，进而判定该比特串既包含协议关键词部分，也包含非协议关键词部分，两节点之间不可达。取值变化率方差如式(10)，$\bar{g}_{(i,j,n)}$ 表示取值变化率均值。

$$var_{(i,j,n)} = \frac{\sum_{t=1}^{j-i-n+2}\left(g^t_{(i,j,n)} - \bar{g}_{(i,j,n)}\right)}{j-i-n+2} \tag{10}$$

2.2.2 最佳路径搜索算法

根据分支度量和约束条件，结合由候选格式关键词边界点生成的有向图。利用最佳路径搜索算法从有向图中找到与真实格式关键词边界最接近的一条路径作为最终格式关键词边界推断结果。最佳路径搜索算法的步骤如下：

1） 设置指针指向起始节点 $fitem_1$。

2） 判断搜索指针是否指向结束位置，若成立，转到步骤 5；否则指针移到下一个候选边界点。

3） 根据约束条件 $var_{(i,j,n)} < var_{th}$ 推测节点 $fitem_i$ 的可能上游节点 $fitem_j$。由已经得到的最短路径 $\boldsymbol{fitem}_s = (fitem_1, fitem_2, \mathbf{L}, fitem_i)$ 扩展为到节点 $fitem_j$ 的可行路径，并计算各可行路径的分支度量总和 $\sum_{i=fitem_2, i\in fitem_s}^{i=fitem_j} Bm_i$，直到完成对 $fitem_j$ 所有可行路径的遍历。

4） 根据最佳路径搜索原则，选择分支度量总和最小的一条路径作为到节点 $fitem_j$ 的最佳路径，转到步骤 2。

5） 根据搜索结果得到最佳路径，最终实现对协议格式关键词边界的确定。

3 仿真与实验分析

为了验证 OBPFK 算法在二进制协议格式关键词边界确定中的有效性，在 AIS1、AIS18、ICMP00、ICMP03 和 NetBIOS 五种类型协议报文的真实数据集上进行测试，测试数据集如表 1 所示。几种协议报文的格式规范可以参考文献[9][13][14]。由格式关键词定义，五种类型协议报文的格式关键词如表 2 所示。例如 AIS1 协议报文格式关键词列表中，1-6 表示该类协议数据帧的 1 到 6 比特为一个格式关键词。

表 1 测试数据集

Tab.1 Test data set

协议	报文类型	数量（帧）
AIS	AIS1	1000
	AIS18	1000
ICMP	ICMP00	1000
	ICMP03	1000
NetBIOS		1000

表 2 格式关键词

Tab.2 Format keywords

协议	报文类型	格式关键词（比特）
AIS	AIS1	1-6;144-148
	AIS18	1-6;39-46;144-147
ICMP	ICMP00	1-16
	ICMP03	1-16;33-72;137-144
NetBIOS		1-16;65-80

实验涉及的主要参数设置为：minsupport=0.95、N_1=3、N_2=6、r=4 和 var_{th}=0.05。minsupport 为最小支持度，用于判定 n-gram-position 是否为频繁项，本文针对同种类型协议报文做格式关键词边界确定，考虑到可能存在少量噪声数据，因此设置 minsupport=0.95。N_1 和 N_2 为 n-gram-position 算法中 n 取值范围的上下限，具体见 3.1.1 节。r 为统计左右分支信息熵时左右两侧比特串长度，经实验论证设置为 4。var_{th} 为 n-gram-position 取值变化率方差阈值，一般取较小值，经实验论证设置为 0.05。

本文采用协议逆向工程中常用的准确率、召回率和 F 值作为衡量指标。设 OBPFK 算法确定的格式关键词边界为 $\boldsymbol{W} = \{w_1, w_2, \mathbf{L}, w_q\}$，真实的格式关键词边界为 $\boldsymbol{Y} = \{y_1, y_2, \mathbf{L}, y_m\}$，准确率、召回率和 F 值的计算公式如式(11)-(13)所示。

$$precision = \frac{\left|\{w \,|\, |w-y| \le 2, w \in \boldsymbol{W}, y \in \boldsymbol{Y}\}\right|}{|\boldsymbol{W}|} \tag{11}$$

$$recall = \frac{\left|\{w \,|\, |w-y| \le 2, w \in \boldsymbol{W}, y \in \boldsymbol{Y}\}\right|}{|\boldsymbol{Y}|} \tag{12}$$

$$F = \frac{2 \times precision \times recall}{precision + recall} \tag{13}$$

3.1 算法参数及性能分析

3.1.1 算法参数分析

n-gram 算法中参数 n 影响频繁项提取。n 取值太小，提取的频繁项数量会偏多；n 取值太大，提取的频繁项数量会偏少。本文统计了不同 n 值时五种协议对应的 F 值，如图 4 所示。当 n 值逐渐增大时，F 值先增大后减小；不同协议，最佳 n 值可能不同。为了减小误差，本文 n 值为一个范围，而不是一个确定的值。由图 4 可以看出 $3 \le n \le 6$ 时，几种协议的 F 值相对较大，$n > 6$ 以后，部分协议 F 值下降较快。为了使 n 取值尽可满足不同协议，同时考虑到计算成本，本文设置 n 的取值范围为: $3 \le n \le 6$ 。

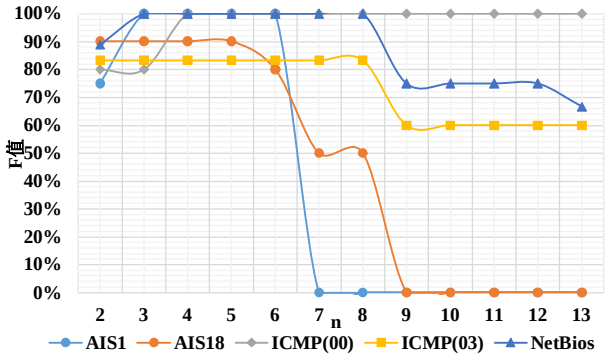

图 4 不同 n 下的 F 值统计

Fig.4 F under different n

3.1.2 算法鲁棒性分析

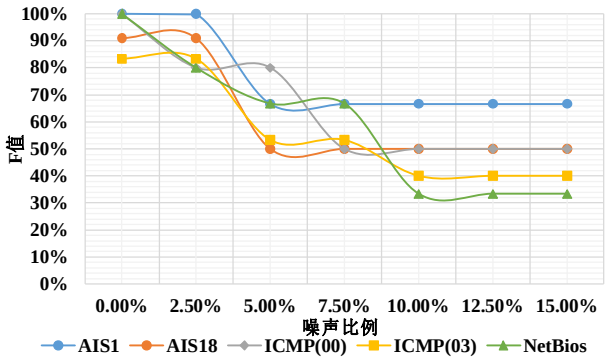

图 5 不同比例噪声下的 F 值统计

Fig.5 F under different proportional noise

为了测试 OBPFK 算法鲁棒性，本小节在同种类型协议报文集合中加入噪声数据，噪声数据为其它协议数据。统计不同比例噪声数据下的 F 值，如图 5 所示，随着噪声比例增加，算法性能逐渐下降，最终保持平稳。在有噪声的协议数据集中，算法性能与最小支持度阈值 minsupport 和噪声数据类型有关。当噪声数据所占比例大于1-minsupport 时，通过 minsupport 定义提取的频繁项数量将大幅下降，从而导致算法性能下降。若噪声数据的报文格式与目标协议数据的报文格式相似，则噪声数据对算法性能的影响很难消除。若噪声数据的报文格式与目标协议数据的报文格式差异较大，则适当调整 minsupport 取值，即可消除噪声对算法的影响。

3.1.3 算法整体性能分析

本小节对 OBPFK 算法的整体性能进行了测试。针对数据集中的五种不同类型协议报文，分别计算了准确率，召回率和 F 值。实验结果如图 6 所示，五种类型协议报文的 F 值均在 83%以上。

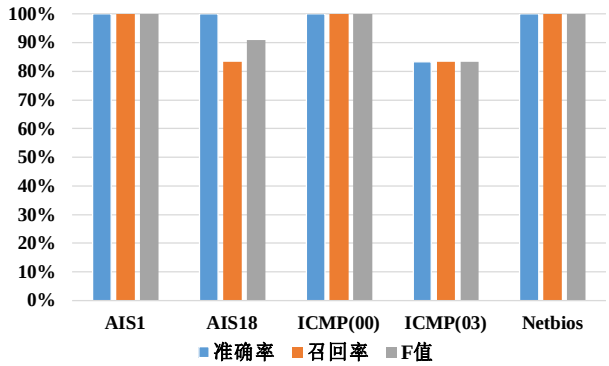

图 6 OBPFK 算法整体性能

Fig.6 The overall performance of OBPFK

3.2 与经典算法的性能对比

本小节对 OBPFK 算法与经典算法进行了性能对比。选取 VDV[15]和 AutoReEngine[5]作为对比算法。VDV 和 AutoReEngine 与 OBPFK 相似，均是基于关键词做协议报文格式逆向，但两者均以字节为单位，忽略了二进制协议字段不受字节长度限制的特性。OBPFK 以比特为单位做统计，相比于 VDV 和 AutoReEngine，具有更高的准确度。在不同协议数据集上，对三种算法的 F 值进行统计，结果如图 7 所示，可以发现 OBPFK 在三种算法中性能最好。相比于 VDV 和 AutoReEngine，OBPFK 算法 F 值平均提升约 8%。

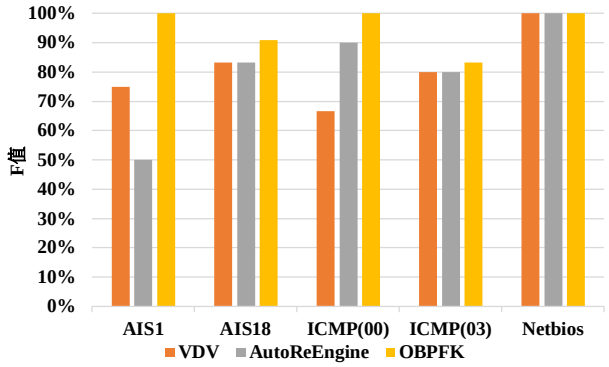

图 7 不同算法性能对比

Fig.7 Performance comparison of different algorithms

3.3 算法复杂度分析

假设数据集中协议数据帧个数为 n，数据帧长度为 l。VDV 算法计算构成协议数据帧各个字节的方差，计算复杂度

近似为$O(l)$。AutoReEngine 算法基于 Apriori 算法，计算复杂度近似为$O(nl^2)$。OBPFK 算法的复杂度主要由格式关键词边界提取和格式关键词边界选择两部分构成，为$O(nl+k^2)$，k 为候选格式关键词边界点个数。

4 结语

本文提出基于最佳路径搜索的二进制协议格式关键词边界确定方法 OBPFK，通过迭代 n-gram-position 算法提取候选格式关键词边界点，利用最佳路径搜索算法从候选边界点中筛选出与真实格式关键词边界点最接近的子集，作为协议格式关键词边界确定的结果。通过在 AIS1、AIS18、ICMP00、ICMP03 和 NetBios 五种不同类型协议报文数据集上测试，具有不错的效果。但 OBPFK 算法只能提取协议格式关键词，而协议格式关键词序列只是协议报文格式的一个子集。如何完整地确定二进制协议的字段边界，获取协议报文格式，是未来需要继续研究的课题。

参考文献

[1] TAO S, YU H, LI Q. Bit-oriented format extraction approach for automatic binary protocol reverse engineering [J]. Iet Communications, 2016, 10(6): 709-716.

[2] LI T, LIU Y, ZHANG C, et al. A noise-tolerant system for protocol formats extraction from binary data [C] // 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications. Ottawa: IEEE, 2014: 862- 865.

[3] 孟凡治, 李桐, 刘渊,等. 基于概率比对的通信协议格式逆向分析方法[J]. 计算机工程与设计, 2016, 37(9):2337-2341. (MENG F, LI T, LIU Y, et al. Format reverse method for communication protocol based on probability alignment [J]. Computer Engineering and Design, 2016, 37(9):2337-2341.)

[4] WANG Y, YUN X, Shafiq M Z, et al. A semantics aware approach to automated reverse engineering unknown protocols [C] // IEEE International Conference on Network Protocols. Austin: IEEE, 2012:1-10.

[5] LUO J, YU S. Position-based automatic reverse engineering of network protocols [J]. Journal of Network & Computer Applications, 2013, 36(3):1070-1077.

[6] 黎敏, 余顺争. 抗噪的未知应用层协议报文格式最佳分段方法*[J]. 软件学报, 2013(3):604-617. (LI M, YU S. Noise-Tolerant and Optimal Segmentation of Message Formats for Unknown Application-Layer Protocols [J]. Journal of Software, 2013(3):604-617.)

[7] 吴礼发, 洪征, 潘璠. 网络协议逆向分析及应用[M]. 北京：国防工业出版社，2016:63. (WU L, HONG Z, PAN F. Network protocol reverse analysis and application [M]. Beijing: National Defense Industry Press, 2016:63.)

[8] 王变琴, 余顺争. 自适应网络应用特征发现方法[J]. 通信学报, 2013(4):127-137. (WANG B, YU S. Adaptive extraction method of network application signature [J]. Journal on Communications, 2013(4):127-137.)

[9] G B T. 船载自动识别系统 (AIS) 技术要求 [S][D]. (G B T. Technical requirements of shipborne automatic identification system [S][D].)

[10] Ma J, Levchenko K, Kreibich C, et al. Unexpected means of protocol inference [C] // ACM SIGCOMM Conference on Internet Measurement. New York: ACM, 2006:313-326.

[11] Brown P F, Desouza P V, Mercer R L, et al. Class-based n-gram models of natural language [J]. Computational linguistics, 1992, 18(4): 467-479.

[12] 范亮, 王晓梅, 杨东煜. 一种利用最佳路径搜索的 PDU 容错定界算法[J]. 西安电子科技大学学报(自然科学版), 2016, 43(5):160-166. (FAN L, WANG X, YANG D. Algorithm for error-tolerant delimitation for the protocol data unit based on best path searching [J]. Journal of Xidian University (Natural Science), 2016, 43(5):160-166.)

[13] Blanchet M. Internet Control Message Protocol [J]. Alphascript Publishing, 1981, 11(4):348.

[14] McLaughlin L J. Standard for the transmission of IP datagrams over NetBIOS networks [R]. RFC1088, 1989:2.

[15] Burschka S, Biersack E. Traffic to protocol reverse engineering [C]// IEEE International Conference on Computational Intelligence for Security & Defense Applications. Ottawa: IEEE, 2009:1-8.

YAN Xiaoyong, born in 1993, M. S. candidate. His research interests include data mining, protocol reverse analysis.

LI Qing, born in 1976, Ph. D., associate professor. Her research interests include protocol reverse analysis, visible-light communication, wireless self-organizing network, sensor network.