计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900035

私有二进制协议中变长域的格式挖掘方法

徐旭东　张志祥　张　献
海军工程大学电子工程学院　武汉 430033
(qddxxd@163.com)

摘　要　协议逆向工程是模糊测试领域的重要环节之一。针对目前私有二进制协议中对于变长域的格式挖掘工作没有很好的系统方法和变长域的关键词域边界挖掘不理想的问题,提出对变长域中的长度域和关键词域分别处理的方法。对于长度域,利用渐进多序列对比的结果,使用迭代窗的挖掘方式分别挖掘全局长度域和局部长度域,在SNMP协议构造的数据集上进行测试,具有较好的边界挖掘效果;对于关键词域,针对已有方法中无法挖掘关键词域前边界的问题,改进投票专家算法,增加反向查找树,能同时挖掘出关键词域的前边界和后边界,在ICMP和HTTP协议构造的数据集上测试,相对于传统的投票专家算法有较大改进。

关键词:二进制协议;协议格式挖掘;渐进多序列对比;投票专家算法;迭代窗
中图法分类号　TP393

Format Mining Method of Variable-length Domain in Private Binary Protocol

XU Xu-dong,ZHANG Zhi-xiang and ZHANG Xian
College of Electronic Engineering,Naval University of Engineering,Wuhan 430033,China

Abstract　Protocol reverse engineering is one of the important steps in fuzzy test field. Aiming at the problem that there is no good systematic method for the format mining of variable-length domain and the mining of keyword domain boundary of variable-length domain is not ideal in the private binary protocol,a method to deal with the length domain and keyword domain separately in variable-length domain is proposed. For the length domain,using the results of progressive multi-sequence alignment,the global length domain and the local length domain are respectively mined by using the iterative window mining method,and test on the data set constructed by SNMP protocol shows it has a good boundary mining effect. For the keyword domain,in view of the problem that the former boundary of the keyword domain cannot be mined with the existing methods,by improving the voting expert algorithm,and adding the reverse search tree,the front the back boundaries of the keyword domain can be mined at the same time. Test on the data set constructed by ICMP and HTTP protocol show that,there is greatl improvement compared with the traditional voting expert algorithm.

Keywords　Binary protocol,Protocol format mining,Progressive multiple sequence comparison,Voting expert algorithm,Iterative window

1　引言

在模糊测试[1]领域,对于内部软件的协议测试,需要应用协议逆向[2]技术对报文格式进行挖掘,将挖掘出的格式结合语义信息,构造测试用例,对软件进行测试。因此,协议格式挖掘是其中的重要步骤。

目前,针对协议格式挖掘的方法主要分为两类,一类为文本类协议[3],另一类为二进制协议[4]。针对文本类协议的方法已经有很多成熟的研究,罗建桢等[5]提出了非齐次左右型级联隐马尔可夫模型,其主要针对应用层网络协议,刻画报文中的字段跳转规律,使用最大似然概率准则确定协议关键词的长度,最终重构协议格式,但其应用场景受限于应用层的协议。Bossert等[6]利用上下文信息识别边界信息,同样利用此信息可反向增强聚类效果,但语义信息只针对文本类协议,二进制协议无法直接挖掘出语义信息。Li等[7]提出利用LDA和关联分析方法,从应用程序的网络轨迹自动逆向协议消息格式的方法,利用LDA模型提取关键词,利用语义信息推断协议格式,其弊端为同样不适用于无法获取语义信息的场景。

针对二进制协议,Tao等[8]利用改进的多序列对比的方法,对可能的字段边界进行特征构造,利用贝叶斯估计的判决方法选出最终的字段边界,但其方法只能够挖掘定长协议的字段边界;闫小勇等[9]利用n-grams的方法对报文序列分词,选出其中的频繁项作为关键词以构造边界特征,最后利用最佳路径搜索的办法得到最终的边界,其方法主要弊端在于会将数据部分的频繁项识别成关键词,造成误识别;侯方杰等[10]给关键词加上位置属性,根据关键词标记建立协议的状态转移模型,挖掘出关键词边界,特别对于长度较短的关键词拥有更好的效果,但其位置属性只适用于定长协议的格式挖掘;Zhang等[11]提出ProWord,利用改进的投票专家算法,构造查找树,对变长协议的关键词进行挖掘,但其只能够挖掘出关键词的后边界,而无法挖掘出关键词的前边界。

本文根据目前的研究内容总结出现有工作依然存在的问

通信作者:张志祥(hgzzx@sina.com)

题:1)对于变长二进制协议的变长部分(本文称之为变长域)格式挖掘没有系统的解决办法,某些工作只针对于长度域或关键词域;2)针对二进制协议中变长域的关键词域,没有很好的完全挖掘方法,如 ProWord 只能够挖掘关键词的后边界。针对这些问题,本文的主要贡献有:1)提出使用迭代窗的方式挖掘变长报文中的长度域;2)在投票专家算法的基础上加入反向查找树,用以挖掘变长报文中关键词域的前边界。

2 方法概述

变长二进制协议的长度变化主要由变长的数据部分引起,并且由于定长部分已有较为成熟的研究方法,因此本文主要研究协议的变长域。由于变长域的构成并不单一,主要由两类组成:长度域和关键词域。在由长度域组成的协议中,第一种组成形式为局部长度域和数据交叉出现,第二种组成形式为全局长度域和数据组合出现;在关键词域组成的协议中,组成形式为关键词域和数据交叉出现。

针对前面提到的问题,本文提出一种针对二进制协议中变长域的格式挖掘方法。1)二进制协议不同于文本类协议,其字段不仅有以字节为单位组成的,还有细分到比特一级的字段,如半字节(4 比特)字段等。因此,本文方法在格式划分时,将数据对象处理至半字节粒度,以便于更精确的格式划分。2)针对长度域的格式挖掘,在渐进多序列对比结果的基础上,使用迭代窗的方法,分别挖掘全局长度域和局部长度域,最后根据设置的判定规则,选出最终的长度域。3)针对关键词域的格式挖掘,由于投票专家算法[12-13]在二进制协议关键词挖掘中只能够挖掘出关键词的后边界,因此本文使用改进的投票专家算法,通过反向查找树的构造,使用两位投票专家在二进制报文上反向打分,最后根据投票阈值设置,选出最终的关键词域。

3 长度域挖掘方法

在变长协议产生的二进制报文中,长度域主要分为全局长度域和局部长度域。全局长度域表示所在的二进制报文总体长度,或长度域之后的所有字段长度;局部长度域表示其相连后方字段的长度。对于这两种长度域的挖掘都是建立在渐进多序列对比的结果上,而后根据规则判定最终真实的长度域。

3.1 渐进多序列对比

传统 Needleman Wunsch[14] 双序列对比算法构造了状态转换函数,用以计算两条序列的相似分值,构成一个相似度矩阵。根据得到的相似度矩阵,从矩阵右下角开始,按照最优比对的原则进行回溯,同时将空格插入两条序列,最终得到两条插入空格的序列比对结果。其中,状态转换函数定义如式(1)所示:

$$S_{i,j}=\max\begin{cases}S_{i-1,j-1}+f_{i,j}\\S_{i,j-1}+w\\S_{i-1,j}+w\end{cases}\tag{1}$$

其中,$S_{i,j}$ 表示当前位置得分,$f_{i,j}$ 表示匹配的得分或不匹配的罚分,w 是加入空格的罚分。

在 Needleman Wunsch 双序列对比算法的基础上进行拓展的渐进多序列对比[15]方法,主要分为两步:序列对比和空位回溯。

在序列对比阶段,主要思想如下:1)初始化 n 颗只有一个叶子节点的二叉树森林 $Forest=\{tree_1,tree_2,\cdots,tree_n\}$,在二进制报文样本集 $S=\{s_1,s_2,s_3,\cdots,s_n\}$ 的每条序列中加入森林中每棵树的序列属性,即 $tree_i.seq=s_i$;2)利用双序列对比算法,每次从 $Forest$ 中选择树根节点中两条长度最短的序列作为左右子树构造一颗新的二叉树,将左右对比之后的结果中添加空格较少的那条序列加入到这颗新的树根节点的序列属性中;3)同时删除原本的两条序列,将新建立的二叉树加入到森林中,如第一次迭代结束后,森林 $Forest=\{tree_3,tree_4,\cdots,tree_n,tree_{n+1}\}$;4)重复 2)和 3),当森林中只有一棵树时,结束迭代。

在二叉树的叶子节点中,保存原始的序列内容;在二叉树的中间节点中,保存每次对比产生的序列对比结果和空位列表。空位列表 $gapList=\{L_0:N_0,L_1:N_1,\cdots,L_i:N_i,\cdots,L_n:N_n\}$,其中 L_i 表示空位插入前,该字符在序列中的位置;N_i 表示此位置插入的空格数目。注意,在对比时,上次对比结果中保存的空位也被视为字符,因此每次对比产生的空位列表中的 L_i 只相对于本次对比前的序列。如此递归对比直至根节点。

在空位回溯阶段,从根节点开始,逐层向下更改每个节点的空位列表,直至将空格列表应用到叶子节点上。在回溯过程中,本文设置包含若干回溯规则的空位回溯算法,伪代码如算法 1 所列。

算法 1 空位回溯算法

```
Vba(node):
    ifnode.child==None:
到达叶子结点,apply(gapList,si)
    else:
    foreachchildinnode:
        for L,N in node:
    for Li,Niinchild.gapList:
    value_all=N0+N1+N2+…+Ni-1+Ni
        if L>Li+value_all:
          if i==n and L!=Li+value_all:
    child.gapList[L-value_all]=N
          if L==Li+value_all:
    child.gapList[Li]+=N
        else:
    value_all=N0+N1+N2+…+Ni+1
          if L-value_all>=Li and L-value_all<=Li+Ni:
    child.gapList[Li]+=Ni
          else:
    child.gapList[L-value_all]=N
Vba(child)
```

3.2 迭代窗挖掘方法

传统的滑动窗口是使用一个固定大小的窗口,在序列上进行滑动。由于本文长度域的值表示一段序列的大小,属于经常变化的字段,因此为了适应这种变化,本文使用迭代窗分别挖掘全局长度域和局部长度域。

在全局长度域的挖掘过程中,设计迭代窗大小为 $[L/2,L]$,其中 L 为二进制报文的长度,窗口起始点 T 为报文尾部,终止点 E 的变化区间为 $[2/L,K+1]$,即迭代窗口依次为 $[L/2,L]$,$[L/2-1,L]$,…,$[K+1,L]$,K 为窗口前部所查找长度域的长度。由于二进制报文中的长度域不仅为单纯的二进制表现形式,还有可能与字节、双字节等长度域混杂表示,因此在全局长度域挖掘过程中,查找迭代窗前部的半字节、字节和

双字节的序列部分 K，使得 $value(K)=Length(Window)$，每查找一次迭代一次，迭代过程中窗口起始点 T 固定，窗口终止点 E 向报文头部移动半字节，如图 1 所示。

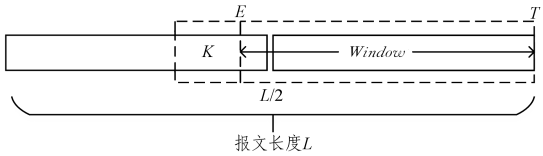

图 1 全局长度域挖掘

Fig. 1 Global length domain mining

在局部长度域的挖掘过程中，设计迭代窗大小为$[m,n]$。窗口起始点 T 在迭代过程中的变化区间为$[L,K+m]$，窗口终止点在每次迭代过程中的变化区间为$[T-m,T-n]$，即在每次迭代开始时，窗口起始点 T 向报文头部移动半字节，在每次迭代的过程中，窗口终止点 E 以半字节为移动单位，使窗口大小从 m 变化到 n。同样，在每次 E 移动后，查找迭代窗前部的半字节、字节和双字节的序列部分 K，使得 $value(K)=Length(Window)$，如图 2 所示。

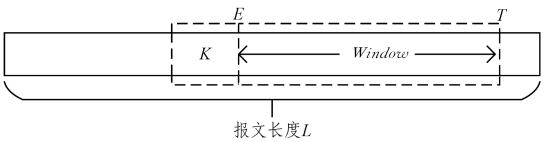

图 2 局部长度域挖掘

Fig. 2 Local length domain mining

挖掘结束后，存在以下两种可能，使得挖掘出的长度域并非真实的长度域：1)挖掘出的长度域为数据字段的内容部分；2)挖掘出的长度字段存在重复，如双字节 00 0e 与单字节 0e 同时被挖掘。

因此，需要借助渐进多序列对比结果，设置判定规则，以进行长度域的最终判别。

3.3 规则判定

借助渐进多序列对比结果，设置如下规则进行判定：1)针对单一序列，若挖掘到的长度域在对比结果中的偏移相同或相近(通常前后误差不超过一个字符)，则认为此长度域为候选长度域；2)针对单一序列，若存在同时挖掘出两种长度且重叠的长度域，优先选择长度更长的长度域作为候选长度域；3)在单一序列上，若选出的多个长度域在序列上连续，且它们的值大于序列长度，则从后向前只取小于序列长度的数值部分；4)由于前 3 条规则选择出的长度域并非全部或大部分出现在数据集中的所有序列，因此针对数据集 S 中的所有序列，根据前 3 条规则找出的候选长度域，设置阈值 $T1$，当某个长度域在所有序列中出现的频率大于阈值 $T1$ 时，则此长度域作为最终长度域。

4 关键词域挖掘方法

4.1 投票专家算法

投票专家算法设计了两个专家的投票作为输入。第一个投票专家指定为词的内部熵 H_I，它表示如果一个词在二进制报文中总是以一个整体出现，则它应该整体被保留，H_I 定义如式(2)所示：

$$H_I=-\log_2 P(w) \tag{2}$$

其中，$P(w)$表示二进制报文中子序列 w 的发生概率，H_I 越低表示 w 通常是整体出现，并且出现的频率越高。

另一个投票专家指定为词的边界熵 H_B，它表示如果一个词的后续内容有很多变化，那么这个词与后续内容之间应该被加上一个边界，H_B 的定义如式(3)所示：

$$H_B=-\sum_{c\in C}P(c|w)\log_2 P(c|w) \tag{3}$$

其中，c 表示子序列 w 后面所有可能出现的半字节的集合，$P(c|w)$表示半字节 c 在序列 w 后面发生的概率。H_B 越大表示 w 后面的内容变化越多，w 后的点越有可能是词边界。

为了实现词的有效划分，设计滑动窗口大小为 N，生成一颗深度为 $N+1$ 的词查找树，树中保存着数据集 S 中所有可能发生的字符组合。

例如，以字符序列“afee5d”构造查找树，设置窗口大小为 3，生成深度为 4 的树，如图 3 所示。其中每个节点表示序列中的一个字符，后面的数字表示此字符所在分支前的所有字符组成的序列所出现的次数。

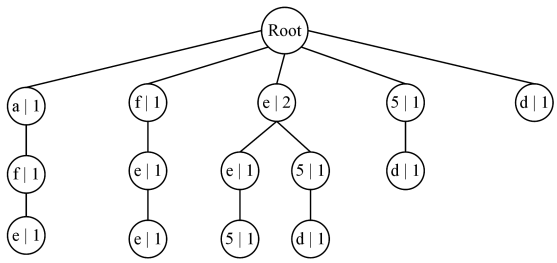

图 3 深度为 4 的查找树

Fig. 3 Search tree with depth 4

投票专家算法分为两个阶段，投票阶段和判决阶段。在投票阶段，使用大小为 n 的窗口在序列上滑动，根据查找树，两位专家依据式(4)和式(5)在窗口内进行投票。由于传统的投票专家算法使用词频率进行计算，在计算不同长度的子序列时，需要对表达式进行标准化，而本文直接使用词在所有数据集中出现的概率 $P(w)$及 $P(c|w)$，因此无需标准化。

$$x_i^I=\arg\min_{x_i^I=i+j}(H_I(w_{i,i+j})+H_I(w_{i+j+1,i+n})) \tag{4}$$

$$x_i^B=\arg\max_{x_i^B=i+j}H_B(w_{i,i+j}) \tag{5}$$

每个 x 点都存在一个投票分数 $V(x)$，如式(6)所示。

$$V(x)=\sum_i(1(x=x_i^I)+1(x=x_i^B)) \tag{6}$$

在判决阶段，设置两条规则以判定 x 处是否为边界：1)在 x 处的得票数大于左右两边相邻点的得票数；2)在 x 处的得票数大于系统所设置的票数阈值 $T2$。

通过投票阶段和判决阶段能够得到序列中大部分可能的词边界。

4.2 改进的前边界挖掘方法

本文设计的改进的前边界挖掘方法是在投票专家算法的基础上进行的改进。由于投票专家算法最初用于文本的词划分，而将其应用于变长二进制报文的关键词挖掘时，只能够有效地挖掘出关键词的后边界，而无法完成对前边界的有效挖掘。因此基于已挖掘出的后边界，本文设计多个迭代窗，进行前边界的挖掘，主要过程如下。

(1)传统投票算法的查找树中保存的是词频次，在投票时需进行标准化计算。本文提出在初始构造查找树时，在每个节点保存由词概率 $P(w)$和词条件概率 $P(c|w)$计算的 H_I 与 H_B，无需进行标准化，减少了计算量。

(2)传统投票专家算法进行二进制报文的关键词划分时只能够得到关键词的后边界，如图 4 所示，即每个边界之间包含着关键词和数据两部分，并不能将关键词完整分割开来。

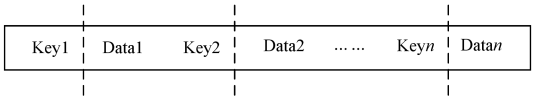

图 4 投票专家算法边界挖掘

Fig. 4 Voting expert algorithm boundary mining

因此，本文提出将二进制报文数据集 S 中的每一条序列作反向处理，根据反向序列，构建一颗反向查找树。例如，序列“afee5d”作反向处理后得到序列“d5eefa”，得到的深度为 4 的反向查找树如图 5 所示。

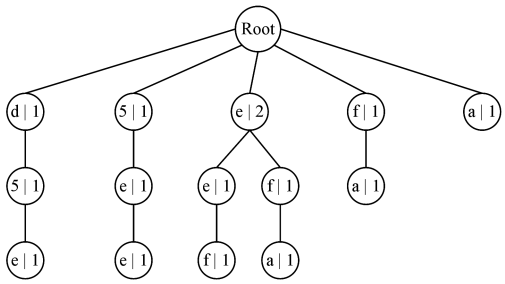

图 5 深度为 4 的反向查找树

Fig. 5 Reverse lookup tree with depth 4

根据反向查找树，两位投票专家依据式(4)和式(5)对序列再次进行投票，将两次结果叠加，最终得到的结果如图 6 所示。

图 6 改进投票专家算法边界挖掘

Fig. 6 Improved voting expert algorithm boundary mining

5 实验结果及分析

5.1 数据集及度量

在长度域挖掘实验中，本文选择 SNMP V1 协议的 Get-response 类型作为测试数据，如表 1 所列。在此协议中包含 5 个长度域，包括 1 个全局长度域，4 个局部长度域。

表 1 长度域数据集

Table 1 Length domain data set

协议	类型	数量/条
SNMPV1	Get-response	200

在关键词域挖掘实验中，本文选用 HTTP 协议中的 response 类型和 ICMP03 协议，如表 2 所列。在 HTTP 的 response 类型协议中，其二进制报文是变长的，共包含 8 个关键词；在 ICMP03 协议中，其二进制报文为定长，共包含 3 个关键词。

表 2 关键词域数据集

Table 2 Keyword domain data set

协议	类型	数量/条
HTTP	response	200
ICMP	03	200

在实验开始前，对数据进行预处理，将报文数据转换为以半字节(4bit)为单位的二进制报文，再进行实验。

本文选择 $F1$ 值作为度量标准，$F1$ 是准确率 P 和召回率 R 的函数，如式(7)所示：

$$F1=\frac{2\times R\times P}{R+P}$$
$$R=\frac{w}{W} \tag{7}$$
$$P=\frac{w}{Z}$$

其中，对于长度域挖掘，w 表示挖掘出的长度域中真实的长度域数量，W 为挖掘出的长度域数量，Z 表示真实的长度域总数；对于关键词域的挖掘，w 表示边界位置与真实位置差距小于 1 的数量，W 为挖掘出的边界数量，Z 表示真实的边界数量。

5.2 长度域挖掘结果及分析

设置长度域挖掘参数，全局长度域的窗口区间为$[L/2, L]$，L 为每条二进制报文的长度，局部长度域的窗口区间为$[6,90]$，设置频率阈值 $T1$，观察 $F1$ 值、召回率 R 与 $T1$ 的关系，如图 7 所示。

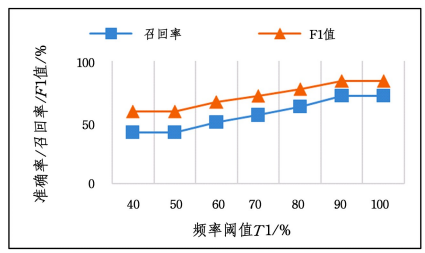

图 7 频率阈值 T1 对长度域挖掘的影响

Fig. 7 Effect of frequency threshold T1 on length domain mining

可以看出，设置不同的频率阈值会产生不同数量的假长度域，设置的频率阈值越高，召回率和 $F1$ 值越高，当频率阈值设为 100%时，所提方法的 $F1$ 值达到了 83.3%，证明其能够有效地挖掘出二进制报文的长度域。

5.3 关键词域挖掘结果及分析

为了观察参数票数阈值和窗口大小对 $F1$ 值的影响，首先设置固定参数，设置查找树和反向查找树的参数 $N=20$，内部熵专家投票票数为 3，边界熵专家投票票数为 2。

5.3.1 窗口大小参数分析

设置票数阈值 $T2=9$，观察不同窗口大小对关键词域边界划分效果的影响，如图 8 所示。

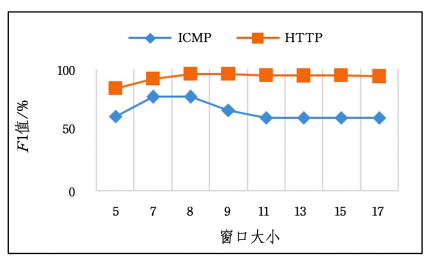

图 8 不同窗口大小对关键词域挖掘的影响

Fig. 8 Influence of different window sizes on keyword domain mining

可以看出，在窗口大小由小变大的过程中，$F1$ 值的变化是先上升后下降，窗口大小为 8 时，两种协议都达到了相对较高的水平，其中 HTTP 协议的关键词边界挖掘的 $F1$ 值最高达到了 95.62%，ICMP 协议的关键词边界挖掘的 $F1$ 值最高达到了 77.02%。通过实验数据展示发现，由于 ICMP 协议中存在一个由两个半字节组成的关键词，并没有将其边界挖

掘出，导致 ICMP 协议的 $F1$ 值普遍比 HTTP 协议的 $F1$ 值低，而 HTTP 协议的关键词普遍较长。因此，也说明本文的方法对关键词较长的协议拥有更好的挖掘效果。

5.3.2 票数阈值参数分析

设置窗口大小为 8，观察不同票数阈值对关键词域边界划分效果的影响，如图 9 所示。

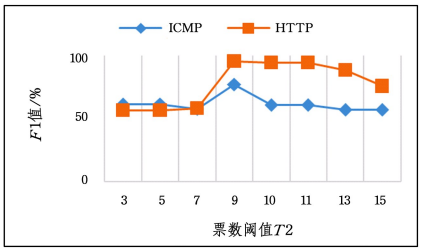

图 9 不同票数阈值 $T2$ 对关键词域挖掘的影响

Fig. 9 Influence of different vote threshold $T2$ on keyword domain mining

可以看出，随着票数阈值 $T2$ 逐渐变大，$F1$ 值先上升后下降，在票数阈值 $T2=9$ 时，两种协议的 $F1$ 值都达到较高值，其中 HTTP 协议的 $F1$ 值最高达到了 95.62%，ICMP 协议的 $F1$ 值达到了 77.02%。

5.3.3 对比分析

将本文方法与同样使用投票专家算法的 ProWord 作对比，所得出的结果如图 10 所示。

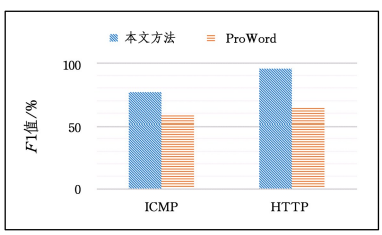

图 10 本文方法与 ProWord 对比

Fig. 10 Comparison of Proposed method with ProWord

两种方法都设置相同参数，票数阈值 $T2=9$，窗口大小为 8。可以看出，本文方法相较于 ProWord，$F1$ 有明显提高，对于 ICMP 协议，$F1$ 值提高了 17.9%，对于 HTTP 协议，$F1$ 值提高了 32.01%，说明本文方法在关键词域边界挖掘上具有较好的挖掘效果。

结束语 本文提出的对于变长域中的长度域和关键词域分别挖掘的工作思路，在长度域挖掘中，利用迭代窗的挖掘方式能够较好地挖掘出长度域，在关键词域挖掘中，通过与传统投票专家算法对比，本文的方法具有更好的挖掘效果。同时，通过实验也发现一些问题，例如方法依赖二进制报文中数据部分的多样性，过于单一的二进制报文不能够很好地将长度域和关键词域挖掘出，并且对于变长域的挖掘只是变长二进制报文格式挖掘的一部分，接下来工作需要结合固定域部分的格式挖掘，将协议格式进行完整解析。

参 考 文 献

[1] 黄影，邹颀伟，范科峰. 基于 Fuzzing 测试的工控网络协议漏洞挖掘技术[J]. 通信学报，2018，39(S2)：185-192.

[2] 张钊，温巧燕，唐文. 协议规范挖掘研究综述[J]. 计算机工程与应用，2013，49(9)：1-9.

[3] 钟晓欢. 基于文本类型的应用层协议逆向解析技术的研究[D]. 北京邮电大学，2014.

[4] 李美剑. 基于动态二进制分析的协议模型逆向提取及其应用研究[D]. 长沙：国防科学技术大学，2014.

[5] 罗建桢，余顺争，蔡君. 基于最大似然概率的协议关键词长度确定方法[J]. 通信学报，2016，37(6)：119-128.

[6] BOSSERT G，FRÉDÉRIC G，HIET G. Towards automated protocol reverse engineering using semantic information[C]// Acm Symposium on Information. ACM，2014.

[7] LI H，SHUAI B，WANG J，et al. IEEE 2015 11th International Conference on Computational Intelligence and Security (CIS)-Shenzhen，China (2015. 12. 19-2015. 12. 20)[C]// 2015 11th International Conference on Computational Intelligence and Security (CIS)-Protocol Reverse Engineering Using LDA and Association Analysis. 2015：312-316.

[8] 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.

[9] 闫小勇，李青. 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[J]. 计算机应用，2018，38(6)：206-211.

[10] 侯方杰，王雷，王嵩，等. 基于位置的自动化网络流协议逆向分析方法[J]. 计算机工程，2019，45(5)：84-87.

[11] ZHANG Z，ZHANG Z，LEE P P C，et al. ProWord：An unsupervised approach to protocol feature word extraction[C]// Infocom，IEEE. 2014.

[12] COHEN P，ADAMS N. An Algorithm for Segmenting Categorical Time Series into Meaningful Episodes[C]// International Conference on Advances in Intelligent Data Analysis. Springer-Verlag，2001.

[13] COHEN P，ADAMS N，HEERINGA B. Voting experts：An unsupervised algorithm for segmenting sequences[M]. IOS Press，2007.

[14] HERINGA J. Needleman-WunschAlgorithm[M]// Encyclopedic Dictionary of Genetics，Genomics，and Proteomics. 2004.

[15] HUNG C L，LIN Y S，LIN C Y，et al. CUDA ClustalW：An efficient parallel algorithm for progressive multiple sequence alignment on Multi-GPUs[J]. Computational Biology & Chemistry，2015，58：62-68.

XU Xu-dong，born in 1995，candidate. His main and research interests include software quality assurance，protocol reverse，etc.

ZHANG Zhi-xiang，born in 1967，Ph. D associate，professor. His main research interests include software quality assurance，artificial intelligence，etc.