·移动互联与通信技术· 文章编号:1000-3428(2019)05-0084-04 文献标志码:A 中图分类号:TP393.08

# 基于位置的自动化网络流协议逆向分析方法

侯方杰,王 雷,王 嵩,盛 捷

(中国科学技术大学 信息科学技术学院 自动化系,合肥 230001)

**摘 要**:现有自动化网络流协议逆向分析方法处理含有大量二进制报文数据的协议时难以准确推断报文格式。为此,提出一种改进的自动化网络流协议逆向分析方法(PoKE)。通过为关键词添加位置属性,提取出二进制报文数据中长度较短的关键词。利用关键词对报文进行标记,根据标记序列建立协议状态转移模型,同时采用基于报文分割和关键词提取的递归循环方式,实现更全面的关键词信息提取。实验结果表明,与 Biprominer 方法相比,PoKE 方法能提取出更多的关键词信息,从而建立更精确的二进制协议模型。

**关键词**:协议逆向;网络流;二进制协议;位置属性;关键词提取

**中文引用格式**: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.

**英文引用格式**:HOU Fangjie,WANG Lei,WANG Song,et al. Position-based automated protocol reverse analysis method on network flows[J]. Computer Engineering,2019,45(5):84-87.

# Position-based Automated Protocol Reverse Analysis Method on Network Flows

HOU Fangjie,WANG Lei,WANG Song,SHENG Jie

(Department of Automation,School of Information Science and Technology,
University of Science and Technology of China,Hefei 230001,China)

**【Abstract】** Existing automated network flows protocol reverse method is difficult to infer message format accurately when dealing with protocols with a large number of binary message data. This paper proposes an improved automated network flows protocol reverse analysis method called PoKE. By adding the position attribute to the keywords, PoKE extracts the short-length keywords from the binary message data, uses keywords to mark the messages, and establishes a state transition model of the protocol according to the marked message sequences. At the same time, PoKE extracts more detailed keywords information through recursive looping mode based on message segmentation and keywords extraction. Experimental results show that PoKE method can extract more keywords information than Biprominer method, thereby establishing a more accurate binary protocol model.

**【Key words】** protocol reverse; network flow; binary protocol; position attribute; keyword extraction

**DOI**:10.19678/j.issn.1000-3428.0050950

## 0 概述

协议逆向技术是指在没有协议先验知识的情况下,推断协议格式、语义、状态的技术。自动化协议逆向技术通常被划分为基于轨迹和基于网络流两大类[1]。基于轨迹的自动化协议逆向技术通过监控协议软件实体的执行来推断协议信息。该技术可以获取丰富的协议信息,但需要对协议的软件实体有访问权限。基于网络流的自动化协议逆向技术分析网络中传输的数据,通过聚类等方法获取协议的字段格式。

现有协议逆向技术可以有效提取报文头部格式信息,能较好地分析会话中不同类型报文的传输顺序,但是在分析协议负载时通常效果不好,尤其是当负载中包含大量二进制信息[2-3]。PIP[4]是较早的自动化协议逆向工具,其基本思想是使用 Needleman-Wunsch 算法对不同报文做字段比较,获取相似字段格式。Discoverer[5]是基于网络流的自动化协议逆向工具,通过对报文的每个字节做标记、循环聚类、融合 3 个步骤完成协议格式提取。文献[6]从报文分段的角度出发,使用专家投票方法对报文进行划分。该方法认为报文段内部的熵值较低,边界处的熵值较高,通过熵值的高低来判断报文边界。文献[7]也是从报文分段的角度出发,但使用的是语音

**基金项目**:国家科技重大专项(2017ZX03001019-004);中国科学院战略性先导科技专项(XDA06011203)。

**作者简介**:侯方杰(1991—),男,硕士研究生,主研方向为未来网络、网络协议解析;王 雷,副教授、博士;王 嵩、盛 捷,讲师、博士。

**收稿日期**:2018-03-27 **修回日期**:2018-05-09 **E-mail**:Devister@mail.ustc.edu.cn

识别领域的 *n*-gram 方法。文献[8]提出一种半自动协议逆向方法。该方法采用半自动化的处理方式,通过人工导入协议相关信息来提高协议推断的准确度。Biprominer、AutoReEngine 等方法[9-11]采用关键词提取的方式对协议格式进行推断和建模。但面对二进制报文尤其是短小的二进制关键词时,仍难以推断报文格式。针对其局限性,本文提出一种基于位置的自动化网络流协议逆向分析方法(PoKE)。该方法通过对关键词添加位置属性,实现从二进制协议负载部分提取更多样的关键词进行更精确的建模。

## 1　关键词提取相关协议逆向技术

通过提取关键词建立协议的状态转移模型是基于网络流的协议逆向技术的常见方法[12-13]。一般来说,关键词被定义为连续的具有特定值和长度的字节段。该方法一般包含 3 个步骤:关键词提取,报文标记,状态转移模型建立。在关键词提取步骤中,目的是提取出现频率高于阈值的关键词。在该步骤中,从长度为 1 的关键词开始统计,逐渐增加关键词的长度,并且大多数方法均基于 Apriori 性质[14],使用短的频繁关键词组合得到长的关键词候选集。在报文标记步骤中,报文将根据提取出的关键词进行标记。该步骤会把每个报文都标记为一串关键词的序列。在状态转移模型建立步骤中,根据得到的大量关键词序列,计算关键词之间的状态转移概率,并以此作为协议的模型描述。

Biprominer[9]是一个经典的基于关键词的协议逆向方法。ProDecoder[10]是其文本协议的改进版本。Biprominer 的关键词提取和报文标记步骤会递归循环调用。在报文标记步骤中,一条报文在被关键词标记后,可以划分为已标记部分和未标记部分。已标记部分将从报文中删除,未标记部分原封不动,再次进入关键词提取步骤,直至提取不出新的关键词为止,进入状态转移模型建立阶段。

AutoReEngine[11]是基于位置的协议逆向方法。AutoReEngine 在关键词提取步骤中会对关键词相对于报文头部和尾部位置做统计,获取该关键词相对于整个报文的位置方差,从而排除一些值相同但位置不同的关键词。但 AutoReEngine 基于位置的方法和本文方法不同。AutoReEngine 的位置使用方差进行衡量,用于排除一些位置不太固定的关键词,起到关键词过滤的效果。而本文方法的位置是每个关键词的固有属性,用于识别并提取特定位置的关键词。

后续有学者采用隐马尔科夫模型[15]和隐半马尔科夫模型[16]建立协议的状态转移模型[12-13],但在关键词提取方法上并没有太大改进。基于关键词提取的协议逆向技术在分析二进制协议负载部分时,主要问题在于难以提取报文中长度短的关键词。二进制协议的格式相对来说较多变,对格式产生影响的字段长度往往较短,有时只有 1 个或 2 个字节。对于这些长度短的关键词,现有方法很难根据频率进行有效提取。这是因为关键词频率一般定义为出现特定关键词的报文数(或会话数)与总报文数(或会话数)的比值。而由于负载的影响,单字节的关键词基本出现在所有报文中,其频率接近 100% 。双字节关键词的频率略低,但大多数仍高于阈值。因此,在实际应用中,往往会忽略长度短的关键词,导致对二进制协议负载分析效果不理想。

## 2　基于位置的自动化协议逆向分析方法

针对一般协议逆向技术在面对二进制协议负载时存在的问题,本文提出基于位置的自动化协议逆向分析方法。该方法通过对关键词添加位置属性提取出相同位置的关键词。对于长度短的关键词,虽然在整个报文组中出现频率相近,但在特定位置上的频率有差异。因此,PoKE 方法可以从二进制协议的负载部分提取出大量长度短的关键词。

### 2.1　总体描述

PoKE 协议逆向分析方法分为 4 个阶段:预处理,基于位置的关键词提取,报文标记与分割,状态转移模型建立,具体流程如图 1 所示。

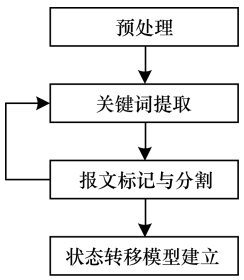


**图 1　PoKE 流程**

在预处理阶段,PoKE 根据通信双方 IP、端口号和协议类型对网络会话进行分组。对于 TCP 协议,还将根据 TCP 的 seq 和 ack 对报文进行重组。

在关键词提取阶段,对于一组给定的报文,会根据位置和出现频率信息提取出基于位置的关键词。每个被提取出来的关键词都满足 2 个条件:1)该关键词在该组报文中的位置是固定的;2)该关键词在该组报文中出现的频率高于某个特定的阈值。在关键词提取的过程中,基于 Apriori 特性[14],即频繁集的子集也是频繁的。本文使用 $k-1$ 个字节的频繁关键词生成 $k$ 个字节关键词候选集。

在报文标记与分割阶段,根据提取出的关键词对报文做标记,从而将报文划分为已标记部分和未标记部分。与 Biprominer 方法[9]不同的是,本文不会将已标记部分从报文中移除,而是在报文中提取出未标记部分,并按照报文标记的格式分组。对于所有新划分的分组,依次执行基于位置的关键词提

取。递归循环调用关键词提取和报文标记分割,直至提取不出关键词为止。

在状态转移模型建立阶段,将提取出的关键词作为协议状态,根据报文标记的关键词序列,建立状态之间的转移概率关系,并以此作为协议的模型。

本文的主要工作是基于位置的关键词提取,重点介绍关键词提取和报文格式标记与分割过程。预处理阶段可以使用多种网络工具完成,如 Wireshark、Dpkt 等;状态转移模型建立阶段使用马尔科夫模型建立状态转移矩阵[9]。

### 2.2 基于位置的关键词提取

基于位置的频繁关键词提取是指对于给定的报文组,从中提取出位置固定的频繁关键词。字符集的定义为:

$$D=\{\backslash 0x00,\backslash 0x01,\cdots,\backslash 0xFF\} \tag{1}$$

其中,每个元素都是一个单字节的字符。在字符集 $D$ 上定义的报文集为:

$$M=\{m_k\mid m_k=\overline{d_0d_1\cdots d_i\cdots d_{n_k}},d_i\in D,i=0,1,\cdots,n_k\} \tag{2}$$

包含 $n$ 个报文的报文组可以定义为:

$$G=\{m_i\mid m_i\in M,i=1,2,\cdots,n\} \tag{3}$$

报文组中的报文可以是网络会话报文,也可以是格式标记与分割阶段切割出来的新报文组。

基于位置的关键词是位置和值序列的二元组。位置指的是关键词首字符相对于报文的偏移量,值序列是 $D$ 中字符构成的字符序列。基于位置的关键词集合定义为:

$$W=\{w_k\mid w_k=(p,v_k),p\in\mathbb{N},v_k=\overline{d_0d_1\cdots d_i\cdots d_{n_k}},d_i\in D\} \tag{4}$$

一个基于位置的关键词 $w=(p,v)$ 在报文组 $G$ 中的频率定义为 $G$ 中包含 $w$ 的关键词的报文数量和 $G$ 中总报文数的比值,具体为:

$$P(w,G)=\frac{|\{m\mid w\in m,m\in G\}|}{|G|} \tag{5}$$

若 $w\in m$ 成立,则当且仅当关键词的值序列 $v$ 出现在报文 $m$ 的第 $p$ 个字节处,即:

$$w\in m\Leftrightarrow d_i^v=d_{i+p}^m,\forall i\in\{0,1,\cdots,len_v\} \tag{6}$$

其中,$d_i^v$ 表示字符序列 $v$ 的第 $i$ 个字符,$len_v$ 表示 $v$ 的长度。

基于位置的关键词提取算法基于 Apriori 性质,即频繁集的子集也是频繁的。具体来说,对于报文组 $G$ 和阈值 $T$,如果基于位置的关键词 $w_k$ 是频繁的,即满足 $P(w_k,G)\geqslant T$,则 $\forall w_{k-1}\in w_k,w_{k-1}$ 也是频繁的。基于该特性,每次在获取 $k$ Byte 关键词时,通过 $k$ Byte 关键词组合得到 $k+1$ Byte 关键词候选集,以减少计算量。

基于位置的关键词提取算法输入的是特定的报文组 $G$ 和用来判别是否频繁的阈值 $T$,输出的是基于位置的频繁关键词集合 $K=\cup K_i$,$K_i$ 表示 $i$ Byte 的关键词集合。如算法 1 所示,第 4 行始化单字节关键词候选集 $CG_i$,并限制候选集中关键词的位置不超过报文最大长度 $len_{max}$;第 7 行判断候选集中关键词频率是否大于阈值,若是则放入频繁集中;第 11 行根据 $i$ Byte频繁集生成 $i+1$ Byte 候选集,然后开始新一轮的关键词提取。如果 2 个 $i$ Byte 频繁关键词位置相差 1,对应位置的值也相同,则生成对应的 $i+1$ Byte 关键词放入候选集。

**算法 1** 基于位置的关键词提取算法

**输入** 报文组 $G$、阈值 $T$

**输出** 关键词集合 $K=\cup K_i$

1. BEGIN
2. $len_{max}=\max\limits_{m\in G}(len(m))$
3. $i=1;K_i=\varnothing$;
4. $CG_i=\{w\mid w=(p,v),w\in W,p<len_{max},len(v)=i\}$
5. while $CG_i\neq\varnothing$ do
6. for all $k\in CG_i$ do
7. if $P(k,G)\geqslant T$ then
8. 添加 k 到 $K_i$ 中;
9. end if
10. end for
11. 生成 $i+1$ 字节关键词的候选集 $CG_{i+1}$;
12. $i=i+1;K_i=\varnothing$;
13. end while
14. END

### 2.3 报文标记与分割

报文标记与分割阶段使用报文组 $G$ 所对应的基于位置的关键词频繁集 $K$,对报文组中每个报文做标记。被关键词标记的部分作为已标记部分,没有被关键词标记的部分作为未标记部分。未标记部分前后总有 2 个关键词,除非未标记部分出现在报文头部或报文尾部,如图 2 所示。


**图 2 报文标记与分割**

本文对未标记部分的报文段进行提取和分组。分组的依据是未标记部分的前后 2 个关键词。前后 2 个关键词相同的未标记部分会进入同一个分组。例如,在图 2 中,所有未标记部分 $U(1,2)$ 会被分到一个分组,$U(2,3)$ 会分到另外一个分组。分组完成后,对于每个新的报文组,分别进行基于位置的关键词提取。通过关键词提取和格式标记分割的递归循

环,能够有效应对关键词在报文样本中分布不均的情况,从而提取出更详细的关键词信息。

## 3　实验结果与分析

本文以 Oracle 数据库通信 TNS 协议比较本文方法 PoKE 和一般基于关键词提取的协议逆向分析方法 Biprominer。TNS 协议是一种二进制协议,且负载格式相对变化较多[17-19]。图 3 展示了 PoKE 方法在关键词提取方面与 Biprominer 方法的比较结果。虽然网络流协议逆向分析方法在协议状态建模方面性能较好[11-13],但在关键词提取方面却没有明显改进。本文采用 Biprominer 经典方法[9]作为对比方法。从图 3 可看出,PoKE 方法可以提取出大量长度短的关键词,并且由于位置的限制,随着关键词长度的增加,关键词数量急剧下降,长度超过 10 的关键词仅有 4 个。Biprominer 方法对于长度短的关键词的提取能力差,提取出的单字节和双字节关键词数量为 0,且随着关键词长度的增加,关键词数量下降速度较缓慢,存在较多 20 Byte 以上的关键词。对于一般的二进制协议,对格式有影响的字段通常较短,因此,PoKE 方法可以更有效地提取出协议中的关键词。

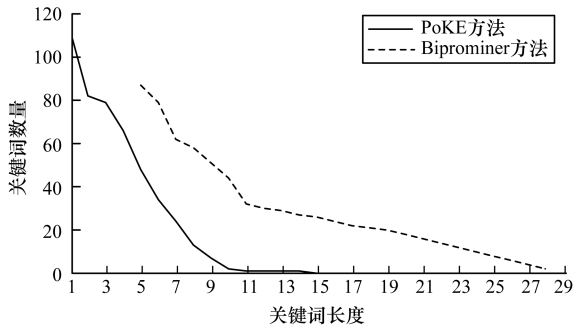


**图 3　协议分析逆向方法的关键词提取结果**

## 4　结束语

本文针对二进制协议负载部分格式难以确定的问题,提出一种基于位置的协议逆向分析方法 PoKE。该方法对关键词添加位置属性,在关键词提取时,使用位置属性对不同关键词进行区分,从而能够识别出长度短的关键词。实验结果表明,相对于一般协议逆向分析方法,PoKE 方法提取出的关键词更短,且能有效提取单字节和双字节的关键词。下一步将利用递归循环方式提取位置无关的比特级别关键词,获取更高精度的关键词信息。

### 参考文献

[1] LI Xiangdong, CHEN Li. A survey on methods of automatic protocol reverse engineering[C]//Proceedings of the 7th International Conference on Computational Intelligence and Security. Washington D. C., USA: IEEE Press,2011:685-689.

[2] NARAYAN J,SHUKLA S K,CLANCY T C. A survey of automatic protocol reverse engineering tools[J]. ACM Computing Surveys,2015,48(3):1-26.

[3] 潘璠,吴礼发,杜有翔,等. 协议逆向工程研究进展[J]. 计算机应用研究,2011,28(8):2801-2806.

[4] BEDDOE M. The protocol informatics project[EB/OL]. [2018-03-04]. http://www. 4tphi. net/ ~ awalters/PI/PI. html.

[5] CUI Weidong, KANNAN J. Discoverer: automatic protocol reverse engineering from network traces[C]// Proceedings of the 16th USENIX Security Symposium. Berkeley,USA:USENIX Association,2007:14-17.

[6] COHEN P,ADAMS N,HEERINGA B. Voting experts:an unsupervised algorithm for segmenting sequences[J]. Intelligent Data Analysis,2007,11(6):607-625.

[7] KRUEGER T, KRÄMER N, RIECK K. ASAP: automatic semantics-aware analysis of network payloads[C]// Proceedings of International Workshop on Privacy and Security Issues in Data Mining and Machine Learning. Berlin,Germany:Springer,2010:50-63.

[8] 杜有翔,吴礼发,潘璠,等. 一种基于报文序列分析的半自动协议逆向方法[J]. 计算机工程,2012,38(19):277-280.

[9] WANG Yipeng, LI Xingjian, MENG Jiao, et al. Biprominer: automatic mining of binary protocol features[C]//Proceedings of the 12th International Conference on Parallel and Distributed Computing, Applications and Technologies. Washington D. C., USA:IEEE Press,2011:179-184.

[10] WANG Yipeng, YUN Xiaochun, SHAFIQ M Z. A semantics aware approach to automated reverse engineering unknown protocols[C]//Proceedings of the 20th IEEE International Conference on Network Protocols. Washington D. C. ,USA:IEEE Press,2012:1-10.

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

[12] WHALEN S,BISHOP M,CRUTCHFIELD J P. Hidden Markov models for automated protocol learning[C]// Proceedings of SECURECOMM'10. Berlin, Germany: Springer,2010:415-428.

[13] CAI Jun, LUO Jianzhen, LEI Fangyuan. Analyzing network protocols of application layer using hidden semi-Markov model[EB/OL]. (2016-04-18)[2018-03-04]. http://dx. doi. org/10. 1155/2016/9161723.

[14] AGRAWAL R,SRIKANT R. Fast algorithms for mining association rules[C]//Proceedings of the 20th International Conference on VLDB. New York, USA: ACM Press,1994:487-499.

[15] RABINER L, JUANG B. An introduction to hidden Markov models[EB/OL]. [2018-03-04]. http://citeseerx. ist. psu. edu/viewdoc/download? doi = 10. 1. 1. 331. 8847&rep = rep1&type = pdf.

[16] YU Shunzheng. Hidden semi-Markov models[J]. Artificial Intelligence,2010,174(2):215-243.

[17] 王召. 基于数据库审计系统 TNS 协议解析的研究与实现[D]. 北京:华北电力大学,2015.

[18] 徐有为. 基于 TNS 协议的 Oracle 审计网关系统的设计与实现[D]. 北京:中国科学院大学,2014.

[19] Oracle TNS协议分析[EB/OL]. [2018-03-04]. https://wenku. baidu. com/view/0ba5df6925c52cc58bd6bedc. html.

编辑　陆燕菲