A Novel Trie And Sequential Feature Analysis-Based Algorithm For Unknown Frame Structure Recognition

Chunfeng Luo, and Yuantao Gu
Department of Electronic Engineering, Tsinghua University, Beijing, China
Email: 13051353880@163.com, gyt@tsinghua.edu.cn

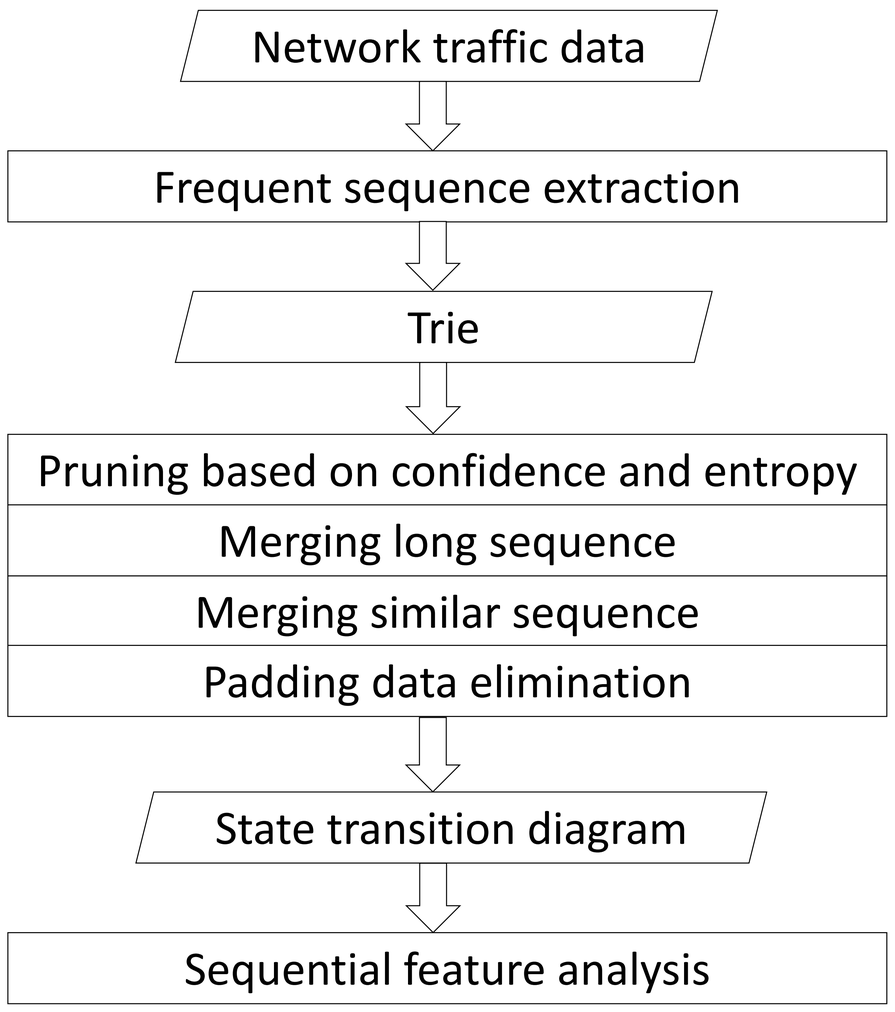
Introduction

- **Unknown frame structure recognition:**
  - **Dataset: Unknown data stream from network traffic data**
    - **Not structured in network protocol format**
    - **No prior knowledge about the frequent patterns in the stream**
  - **Main Goal: obtain sufficient knowledge about the transmission content and connectivity between the devices in the network**

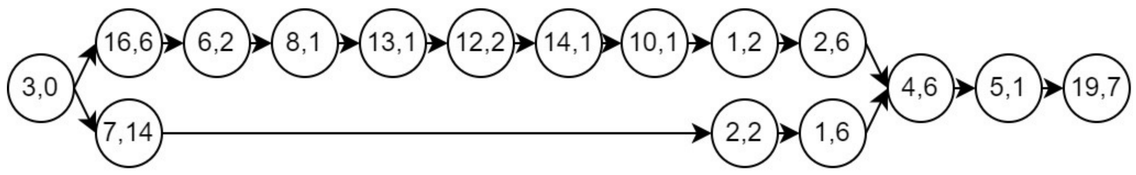
Motivation: How to Analyze frame structure in unknown network traffic data efficiently?

Proposed Approach

- **Frequent sequence extraction**
  - **Using sliding window to establish a Trie**
- **Pruning based on confidence and entropy**
  - $Confidence(B_1 \dots B_n) = \frac{count(B_1 \dots B_n)}{count(B_1 \dots B_{n-1})}$
  - $Entropy(T) = -\sum_{B\in[0\times00,0\times FF\}} conf(TB)\, ln\big(conf(TB)\big)$
- **Long sequence merge operation**
  - **For sequence AC and CB with common subsequence C:**
    - $count(ACB) = count(AC) - count(C) + count(CB)$
  - **Assuring that the error is small enough:**
    - $e = \frac{(1-\alpha_B)count(C)}{count(AC)} = \frac{(1-\alpha_B)}{\alpha_A}$
- **Similar sequence merge operation**
  - **Using DBSCAN to cluster sequences**
  - **Measuring distance by filtered hamming distance**
- **Padding data elimination**
  - **Recording occurrence location:** $L = \{l_1, l_2, \dots, l_n\}$ **and** $d_{i,j} = l_{i+j} - l_i$



Flowchart of analysis algorithm



Frequent pattern state transition diagram example

Result

- **Experimental results on the ethernet traffic data captured in Gulab:**
- **It shows that our algorithm:**
  - **can analyze the hierarchical relationship of long frequent sequences**
  - **has high efficiency and benefits following frequent pattern detection**

Data type	Ethernet
Data size	1075kB
Time	220s
Speed	4.88kB/s

Algorithm performance analysis

No.	Frequent pattern	Mask
1	[64 F6 9D 19 6A 52]	[FF FF FF FF FF FF]
2	[88 E0 F3 7A 66 F0]	[FF FF FF FF FF FF]
3	[D0 D3 D0 5D]	[FF FF FF FF]
4	[08 00 45 00]	[FF FF FF F7]
5	[00 45 00 00]	[FF FF FF FA]
6	[02 00 00 00]	[97 FF FF FF]

PCAP frequent pattern mining results

No.	Protocol layer	Description
1	Ethernet	MAC Address 1
2	Ethernet	MAC Address 2
3	PCAP	PCAP timestamp
4	IP	Ethernet protocol indicator and IP protocol indicator
5	IP	Ethernet protocol indicator and IP protocol indicator
6、7、8、10、12、13、16	Ethernet	PCAP length

Comparation of PCAP mining results and description