面向HTTPS流量的目标网页识别技术研究

* 研究背景：

随着HTTPS协议的普及应用，Web访问数据越来越多地被加密传输。

网络服务提供商希望对流量进行监控。

用户访问网页的流量被HTTPS的流量加密。

https流量是被tls协议加密的http流量，https具有多种协议号。

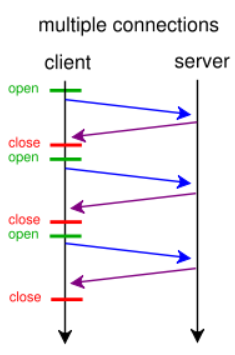
* 研究目标：

由于网页中的每个“元素”是特殊的，访问流量时传输这些“元素”会生成不同的流量特征，这些流量特征在一次访问中经由若干个“链接”传输。

我们希望经由“划分流量”，“总结特征”，“使用分类器识别网页”三步完成我们的目标：

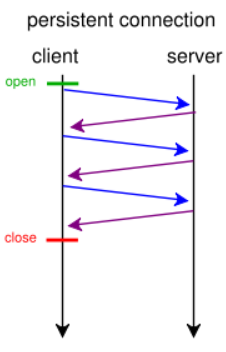
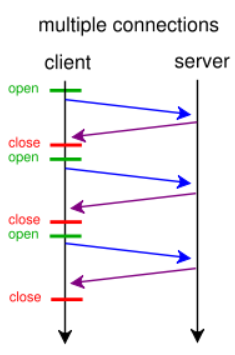
1. 划分流量：

在http1.0时代

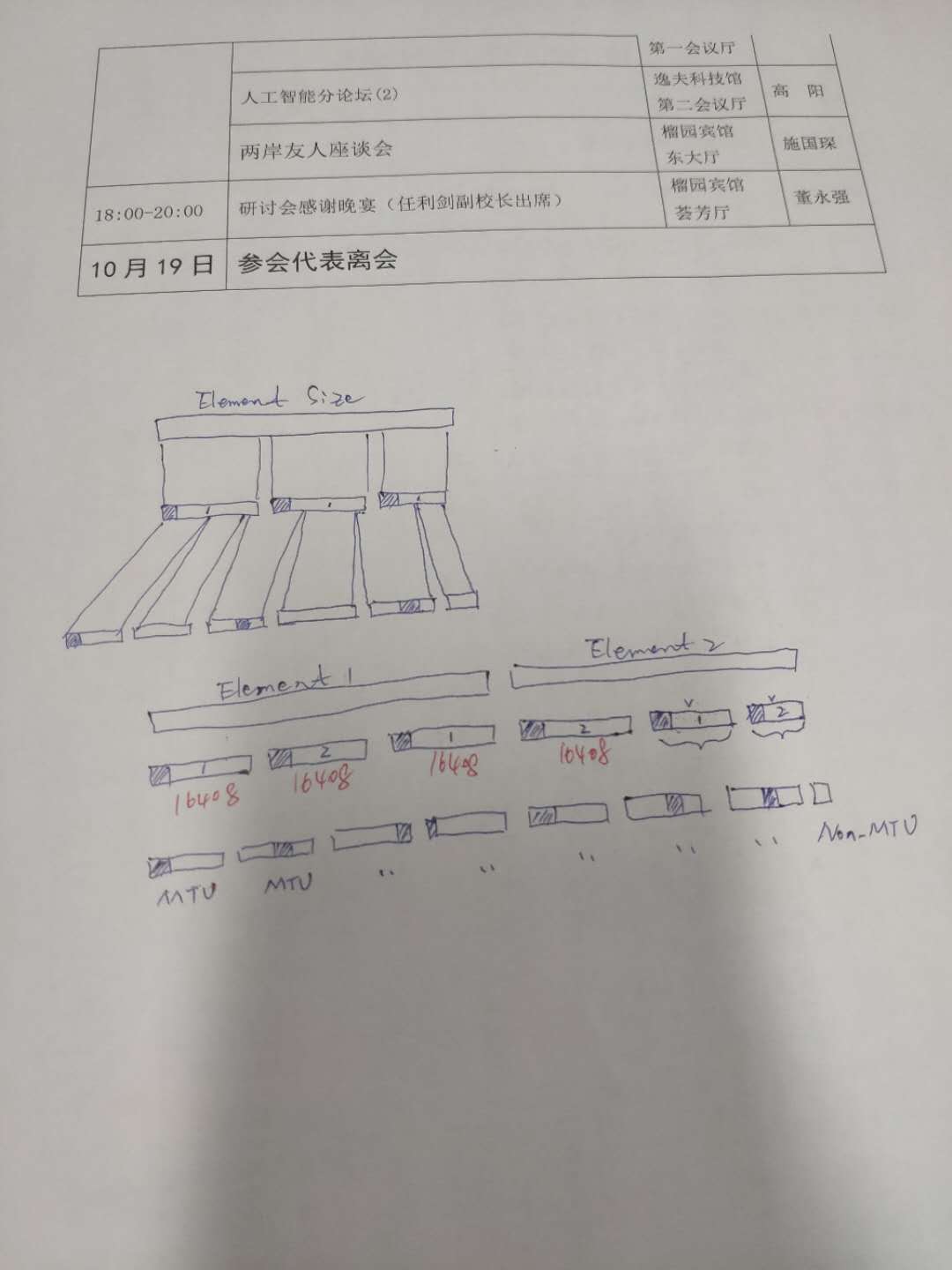


我们可以通过流量的上下行来划分每个流量。

在http1.1时代，引入了持久化和管道化，作为应对手段，我们引出了新的划分流量的方法。



现在被广泛使用的http2.0，我们需要找到新的识别流量的方法。



在http2.0中，网页中的一个或多个元素备份在一个连接（流）中传输，这个连接由一个特定的五元祖信息确定。

有些过大的元素在传输时会划分成若干个片段，在wireshark生成的json文件中我们可以获得每个片段的大小。由于http2.0中不同元素的片段是穿插着传输的，因此我们没有办法准确的将每个元素片段和元素对应起来，但是每个网页中所拥有的元素是固定的，因此每个连接中的片段数量和大小应该是有规律的。我们需要先整理出这些片段，观察其中的特征。