浏览器识别调研报告

指导老师：杨嵘

郑延钦

2021.12

目录

1. 主流浏览器

1.1. IE 浏览器

IE 是微软公司旗下浏览器，是目国内用户量最多的浏览器。IE 诞生于 1994 年，当时微软为了对抗市场份额占据将近百分之九十的网景 Netscape Navigator，于是在 Windows 中开发了自己的浏览器 Internet Explorer。1996 年，微软从 Spyglass 手里拿到 Spyglass Mosaic 的源代码和授权，开始开发自己的浏览器 IE。后来，微软以 IE 和 Windows 捆绑的模式不断向市场扩展份额，使 IE 成为市场的绝对主流。现在装了 Windows 系统的电脑基本无法卸载 IE。

1.2. Opera 浏览器

Opera 是挪威 Opera Software ASA 公司旗下的浏览器。1995 年，opera 公司发布第一版 Opera 浏览器，使用自己研发的 Presto 内核。当时 opera 公司的开发团队不断完善 Presto 内核，使 Opera 浏览器一度成为顶级浏览器。直到 2016 年奇虎 360 和昆仑万维收购了 Oprea 浏览器，从此也丢弃了强大的 Presto 内核，改用当时 Google 开源的 webkit 内核。后来 Opera 浏览器跟随 Google 将浏览器内核改为 Blink 内核。

1.3. Safari 浏览器

2003 年，苹果公司在苹果手机上开发 Safari 浏览器，利用自己得天独厚的手机市场份额使 Safari 浏览器迅速成为世界主流浏览器。Safari 是最早使用 webkit 内核的浏览器也是现在苹果默认的浏览器。

1.4. Firefox 浏览器

Firefox 浏览器是 Mozilla 公司旗下浏览器。Firefox 采用 Gecko 作为内核。Gecko 是一个开源的项目，代码完全公开，因此受到很多人的青睐。Firefox 的问世打破了 IE 浏览器从 98 年后独步浏览器市场的局面。

1.5. Chrome 浏览器

Chrome 浏览器是 google 旗下的浏览器。Chrome 浏览器至发布以来一直讲究简洁、快速、安全，所以 Chrome 浏览器到现在一直受人追捧。最开始 Chrome 采用 webkit 作为浏览器内核，直到 2013 年，google 宣布不再使用苹果的 webkit 内核，开始使用 webkit 的分支内核 Blink。

1.6. 浏览器国内市场占有率排行

第一名，Google Chrome，谷歌浏览器依据其简约的设计风格、较快的页面访问速度和丰富的拓展功能，占据了市场 58.52%的份额，稳居国内第一。

第二名，Microsoft Edge，Edge 浏览器市场占比 15.84%。自从 2020 年 1 月 Edge 稳定版推出以来，得到一致好评。一方面使用了与谷歌浏览器相同的 Chromium 内核，大大增强了系统的稳定性和功能性;另一方面良好的兼容性，全面覆盖 Windows 7、Windows 8 和 Windows 10 系统，方便用户使用。

第三名，QQ 浏览器，腾讯重力打造的一款浏览器，市场占比 7.08%。采用 Chromium 内核+IE 双内核，完美支持 HTML5 及各种新的 Web 标准，超小的安装包和超强的稳定性成为了好评度最高的国货浏览器。

第四名，搜狗浏览器，市场占比 4.37%。作为引领“新上网革命”的先驱，搜狗浏览器注重国内用户的使用习惯, 独创的预取引擎技术和简约而富有创意的页面设计，在实现超快网速的同时，完美地增强了国内用户的上网体验。

第五名，火狐浏览器（Firefox），市场占比 3.89%。

第六名，IE 浏览器，国内市场占比 3.48%。

第七名，苹果 Safari 浏览器，在国内市场仅占比 3.30%。

第八名，用友云浏览器，作为一款专门为企业用户打造的专业浏览器，其主要功能是企业应用和工作协同，市场占比 0.88%。

第九名，猎豹浏览器，国内市场占比 0.84%，由原金山团队倾力打造。

第十名，遨游五浏览器。

以上内容选自《2019-2025 年浏览器产业深度调研及未来发展现状趋势预测报告》。

2. 浏览器内核

浏览器内核（Rendering Engine），是指浏览器最核心的部分，负责对网页语法的解释并渲染网页。

所以，通常所谓的浏览器内核也就是浏览器所采用的渲染引擎，渲染引擎决定了浏览器如何显示网页的内容以及页面的格式信息。不同的浏览器内核对网页编写语法的解释也有不同, 因此同一网页在不同的内核的浏览器里的渲染效果也可能不同, 这也是网页编写者需要在不同内核的浏览器中测试网页显示效果的原因。**由于浏览器内核在渲染页面时需要请求外部资源如图片等，不同的浏览器内核渲染页面时请求资源的网络流量会产生不同的特征，因此可以利用该特征识别浏览器类别。**

常见的浏览器内核可以分这四种：Trident、Gecko、Blink、Webkit。

目前主流的浏览器基本使用这四种内核中的一种或多种：

1、IE 浏览器内核：Trident 内核，也是俗称的 IE 内核；

2、Chrome 浏览器内核：统称为 Chromium 内核或 Chrome 内核，以前是 Webkit 内核，现在是 Blink 内核；

3、Firefox 浏览器内核：Gecko 内核，俗称 Firefox 内核；

4、Safari 浏览器内核：Webkit 内核；

5、Opera 浏览器内核：最初是自己的 Presto 内核，后来是 Webkit，现在是 Blink 内核；

6、360 浏览器、猎豹浏览器内核：IE+Chrome 双内核；

7、搜狗、遨游、QQ 浏览器内核：Trident（兼容模式）+Webkit（高速模式）；

8、百度浏览器、世界之窗内核：IE 内核；

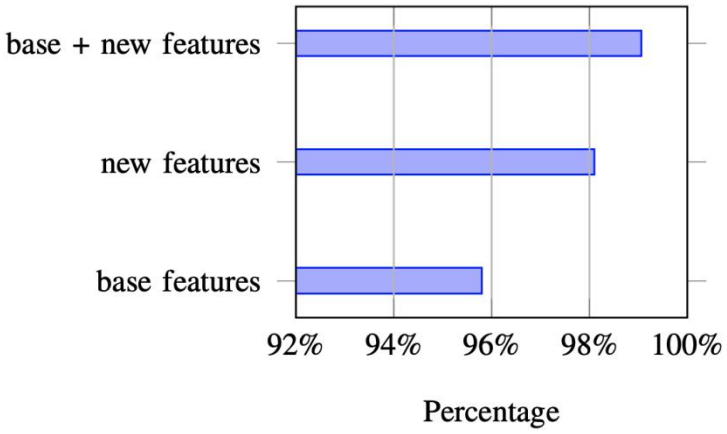
9、2345 浏览器内核：以前是 IE 内核，现在也是 IE+Chrome 双内核；

3. 基于流量识别浏览器类别

以下实验均是在本地主机上直接抓取相关流量研究, 没有考虑到实际应用中网关环境下的流量复杂性与多变性。

3.1. 论文 1

《Analyzing HTTPS Encrypted Traffic to Identify User’ s Operating System, Browser and Application》：在 2017 年 CCNC 会议中，Jonathan Muehlstein 等人提出通过识别流量特征可以确定客户端设备的浏览器类型。论文中主要工作为选取了合适的流量特征，通过简单的 SVM 模型就可以实现对浏览器的准确分类，准确率高达 99%。



(c) Browser Accuracy Results

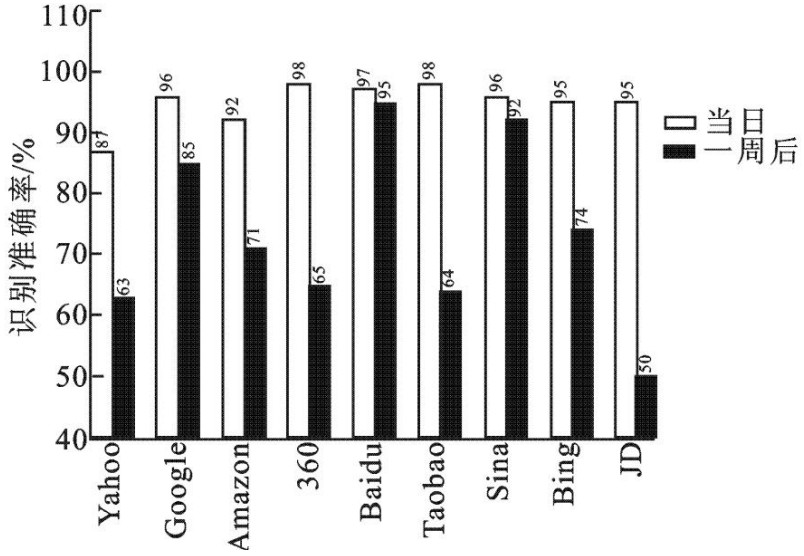
base 特征：前向数据包数、前向总字节数、最小前向到达间时差、最大前向到达间时差、平均前向到达间时差、STD 前向到达间时差、前向数据包数均值、STD 前向数据包数、反向包数、反向总字节数、最小后向到达间隔时间差、最大反向到达间隔时间差、平均反向到达间隔时间差、STD 反向到达间隔时间差、反向数据包数均值、STD 反向数据包数、平均前向 TTL 值、最小前向数据包、最小向后包、最大向前包、最大向后包、总包数、最小数据包大小、最大数据包大小、平均数据包大小、数据包大小方差。

new 特征：TCP 初始窗口大小、TCP 窗口缩放因子、SSL 压缩方法、SSL 扩展计数、SSL chipper 方法、SSL 会话 ID 长度、前向峰值最大吞吐量、后向峰值的平均吞吐量、最大后向峰值吞吐量、最小后向峰值吞吐量、后向 STD 峰值吞

吐量、前向爆发数、后向爆发数、转发最小峰值吞吐量、前向峰值平均吞吐量、前向 STD 峰值吞吐量、平均后峰值到达间时差、到达间的最小后峰值时间差、到达间最大后峰值时差、STD 后向峰值到达间隔时间差、平均前峰值到达间隔时间差、到达间最小前峰值时间差、到达间最大前峰值时间差、STD 前峰到达间隔时间差、保活数据包数、TCP 最大段尺寸、前向 SSL 版本。

3.2. 论文 2

《利用包长特征的浏览器被动识别方法》：该文发表于 2017 年西安电子科技大学学报，刘长江等人通过数据包长度特征来识别加密流量中的浏览器类别。论文主要检测访问不同网站时对浏览器识别的准确率，对比了朴素贝叶斯、随机森林和逻辑回归方法，并且验证了随时间变化包长特征对浏览器识别准确率的变化。实验表明：访问不同网站对浏览器识别的准确率不同，用同一天的数据做训练集和测试集，对多个浏览器识别的准确率平均在 95%以上，用间隔七天后的数据做测试集，对多个浏览器识别的准确率大幅下降，准确率平均为 65%。



4. 总结

基于网络流量特征检测浏览器类别的相关工作不多，从目前少有的论文中也可得知检测浏览器类别难度不高，简单的机器学习模型就可以有 90%以上的准确率。**以上实验的检测前提是浏览器访问指定的网站（比如百度或者微博）时可以分析流量特征来识别浏览器，如果浏览器没有访问指定的网站则无法识别，有一定的局限性。**以上实验均是在本地主机抓取流量进行实验，在网关实际使用的鲁棒性还有待考证。

暂时没有相关工作直接证明网络流量特征可以识别浏览器内核类别，但是考虑到 Jonathan Muehlstein 等人的工作中识别的四个浏览器类别分别使用了四个不同的主流浏览器内核，且浏览器产生的流量与内核渲染逻辑相关，或许可以间接证明网络流量特征可以识别浏览器内核类别，但是对于双核可以切换内核的浏览

器，该类浏览器类别或许较难检测，检测到两种内核特征有可能是客户端使用多个浏览器所导致。以上结论仅为推测，还需实验验证。

通过被动流量进行浏览器识别，从技术的角度来说可以进行工程实现，涉及技术为爬虫流量抓取、抓包和机器学习算法，难点在于流量的特征会随访问的指定网站变化而变化，需要随时更新模型，最终工程中的识别效果暂时无法确定。

参考资料

https://ofirpele.droppages.com/CCNC2017.pdf
https://zhuanlan.zhihu.com/p/250573121
https://www.doc88.com/p-8921319535529.html