Summary

Table 1
Universal extracted features of private dataset.

ID	Feature name	Description
1	IS_TRAP_FILE	Whether to access trap file such as robots. txt.
2	NIGHT_RATIO	Percentage of requests made between 12am and 7am.
3	IMAGE_RATIO	Percentage of image file requests.
4	HTML_RATIO	Percentage of html file requests.
5	REFERRER_RATIO	Percentage of requests with unassigned referrer.
6	HEAD_RATIO	Percentage of requests of type HEAD.
7	304_RATIO	Percentage of requests with status code 304.
8	ERROR_RATIO	Percentage of erroneous requests.
9	ERROR_UPSTREAM_RATIO	Percentage of requests with empty upstream status.
10	SESSION_TIME	Total time elapsed between the first and the last request.
11	AVERAGE_INTERVAL	Average time between two consecutive requests.
12	DEVIATION_INTERVAL	Standard deviation of the time between two consecutive requests.
13	REQUESTS_NUMBER	The total number of requests.
14	UNIQUE_TYPE	The total number of file type of requests.
15	MAX_BROWSER_FILE_RATE	Maximum number of embedded resources in a web page.
16	PENALTY	Penalty for each backward and forward navigation or loop.
17	SD_RPD	Standard deviation of the page depth across all requests.
18	CSR	Percentage of requests with continuous access to the page belong to the same directory.
19	RES	Average response time to requests.
20	SF-FILE_TYPE	Switching factor of file types for each session.
21	SF-REFERRER	Switching factor on unassigned referrer.
22	WIDTH	The number of leaf nodes generated in the graph of all requests.
23	DEPTH	The maximum depth of the tree within the graph of all requests.
24	TOTA_PAGE	The total number of pages requested.

H2 1.1 请求文件类型分布特征

对于不同的爬虫，请求的文件类型分布有所不同：

- 一般搜索引擎只进行文本匹配，因此只收集网站的文本数据，表现为文本请求占比极高
- 图站的图片爬虫只收集图片
- 部分爬虫不进行页面渲染以及动态内容的解析，因此不请求 js 或 css
- etc

H2 1.2 陷阱文件

爬虫可能会访问 robots. txt 文件和网站的 sitemap，这些都不是用户会进行的操作，同时恶意爬虫可能爬取禁止爬取的路径，这些目录/文件都可认为是陷阱，访问则为爬虫

H2 1.2 访问行为

- 访问速度
- HEAD 请求占比
- 夜晚占比
- 平均间隔
- 访问序列（DFS 或 BFS 等非随机访问序列）

H1 2. 数据集准备

来源：
Web Server Access Logs
处理方式参考：
基于访问序列和溯源图深度学习的网络爬虫检测技术研究_黄涵彬, p.32
Reinforcement learning based web crawler detection for diversity and dynamics, p.121

H1 3. 示例项目

mohammadhashemii/Web-Crawler-Detection: Web Crawler Detection using Unsupervised Algorithms

H1 4. 训练方案

- 普通 SVM
- 参考论文的模型架构
  - 访问序列
  - 溯源图
  - 强化学习
- 可附带特征选择模型