基于互联网搜索的自动化设备标注引擎研究

山东大学网络空间安全实验室

陈平

议题目录

content

一、什么是自动化设备标注采集

二、自动化设备标注引擎的开发思路

三、自动化设备标注引擎的价值

四、自动化设备标注引擎未来的路

个人简介

陈平

ID： Murkfox

山东大学网络空间安全实验室研究员

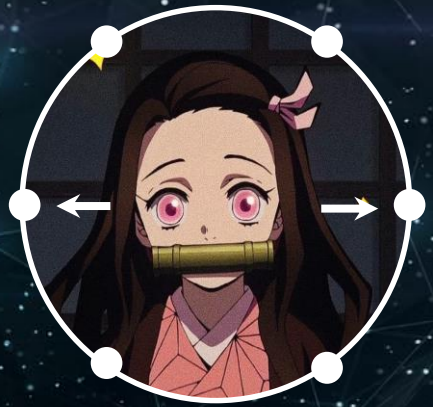
红日安全团队核心成员

主要研究领域：

工控安全

自动化设备注

红队攻击



代表文章：

ZMap扫描机制剖析

竞态攻击：Hyper-V安全问题分析

FACT：一款固件类比分析测试平台

第一部分

什么是自动化设备标注引擎

自动化设备标注引擎

- 自动化设备标注引擎是能够自动进行设备存活探测，设备信息标注，设备安全性检测的系统引擎

前言

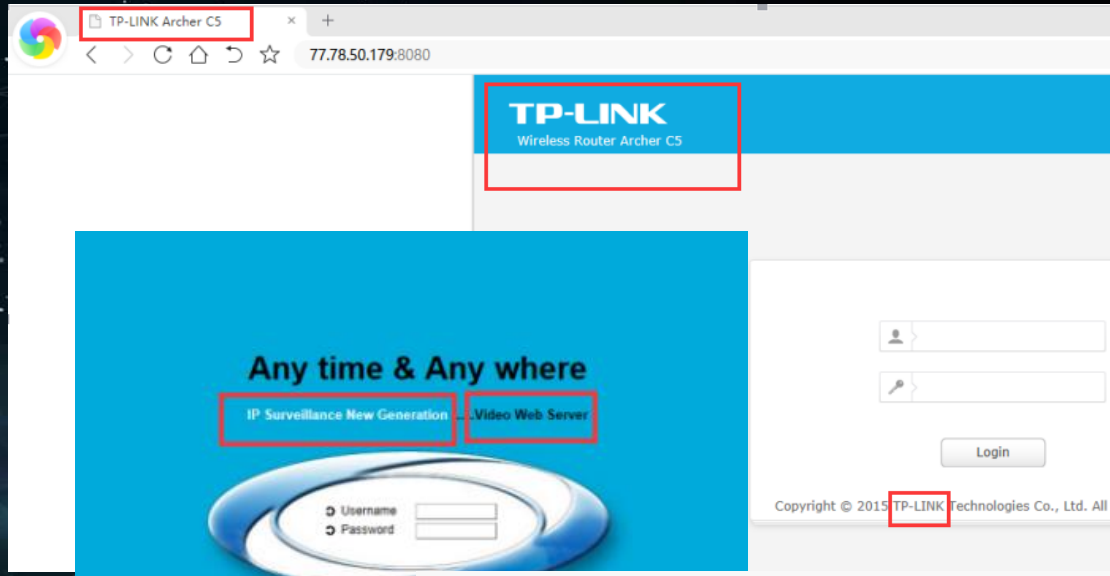
- 随着互联网时代的不断推进，工控设备、物联网设备等诸多终端机如雨后春笋般纷纷进场。由于设备型号的差异、生产厂商的不同导致设备通信数据的形式和内容变得不胜枚举。我们将很难再通过传统的人工编写设备识别脚本去进行设备识别。

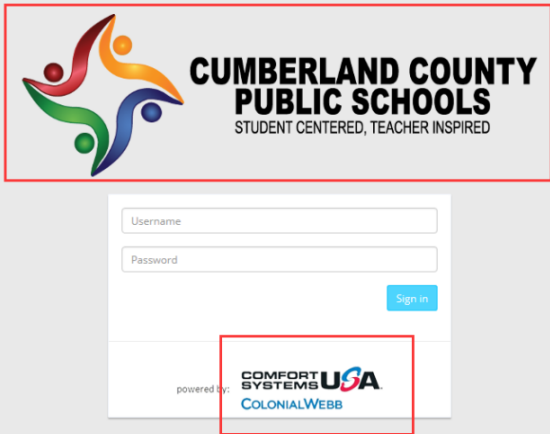
- 以此为契机开始了面对未知设备的自动化标注技术的研究

- 在随后的研究中，我们参考了 发表于 **USENIX Security·2018 《Acquisitional Rule-based Engine for Discovering Internet-of-Things Devices 》（基于采集规则，用于发现物联网设备的引擎）**论文中的原理，制作了名为 " 逆鱼 " 的自动化设备标注规则采集模块。并以此为核心制作了“破天”- 自动化设备标注引擎。

依旧是前言

- 在前期统计中发现，我们在采集的3040家厂商（路由器，工控，摄像头，交换机，打印机，NVR等）
- 发现有2899家厂商的设备提供网上管理或网络通信功能。
- 这些提供网上管理或网络通信功能的设备，有90%在数据通讯内容中会包含生产厂商或设备类型等相关信息





如何手动编写一条设备标注规则

- 访问设备并从通讯内容中获取相应关键字
- 通过某度某歌搜索该关键字，访问数个包含该关键字的网页，并搜索是否包含设备类型、生产厂商等相关设备信息
- 整理搜索到的相关信息，并根据相关信息编写规则
- 测试规则

```
{
    Keyword :  "TP-LINK Archer C5"
    {
        Vendor_Name : "TP-LINK"
        App_Type : "WI-FI"
        Device_Type : "Router"
    }
}
```

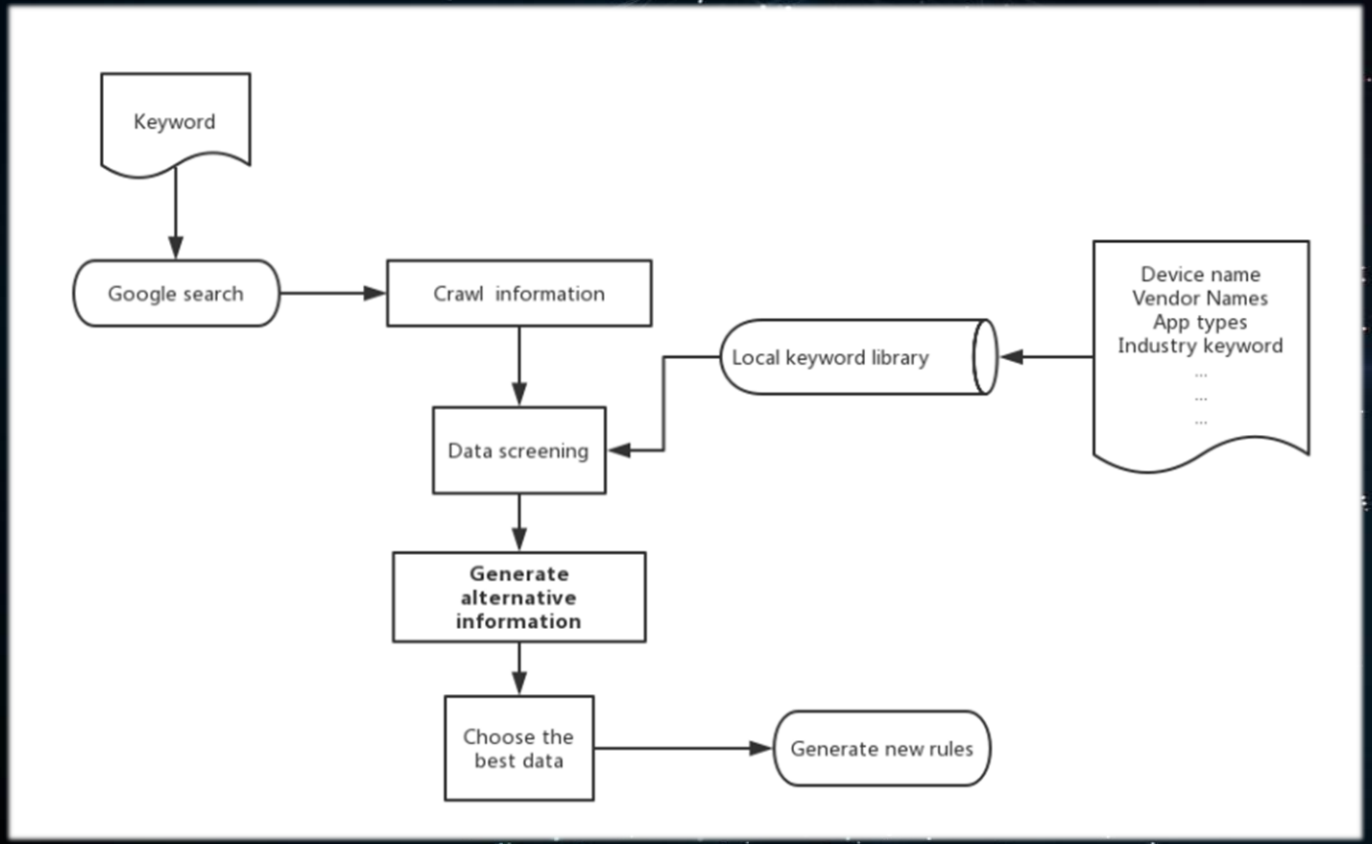
什么是自动化设备标注规则采集

- 自动化设备标注规则采集模块实际上就是自动化了上述步骤
- 我们模拟人工，通过正则匹配提取出关键字并通过搜索引擎API搜索包含该关键字并获取含有关键字的网页
- 我们取排名前10的网页，提取出这些网页的文本数据，根据本地依赖词库筛选出最能描述该设备相关信息的数据作为规则内容

```
{
    Vendor_name: "Huawei, D-LINK, TP-LINK, DaHua, Bacnet..."
    App_Type: "ICS, IOT-Cam, IP-Cam, WI-FI, Web, SSH, FTP, POP3..."
    Device_Type : "PLC, SCADA, RTOS, Router, UPS..."
    ...
    ...

}
```

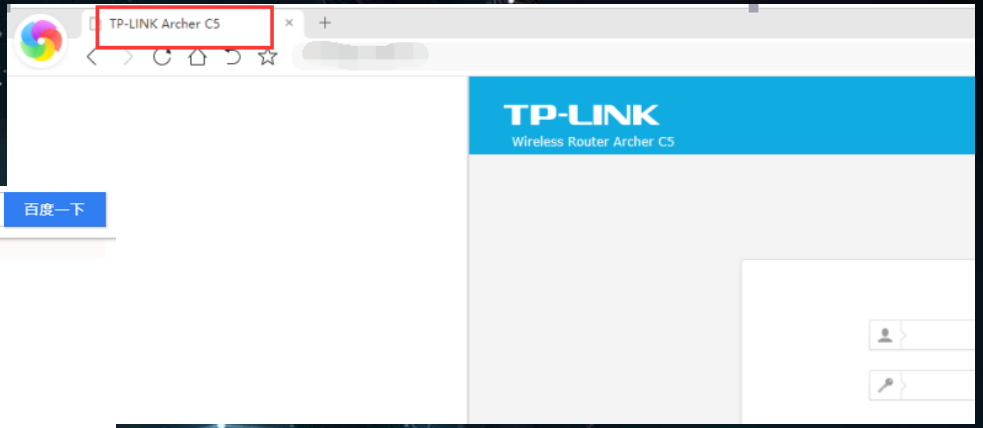
逆鱼 - 自动化设备标注规则采集模块



举个栗子

举个栗子

- 我们以一个TP-LINK的路由器为例，首先我们访问该设备，在网页的Title中我们获取到了关键字 "TP-LINK Archer C5"
- 将关键字通过百度搜索我们获得了许多网页



我们随机使用两个网页的文本内容通过本地依赖词库筛选有关设备的一些信息

TP-Link Archer C5 Login Instructions

This page shows you how to login to the TP-Link Archer C5 router.

Other TP-Link Archer C5 Guides

- Download Manuals for TP-Link Archer C5
- All TP-Link Archer C5 Screenshots.

Find Your TP-Link Archer C5 Router IP Address

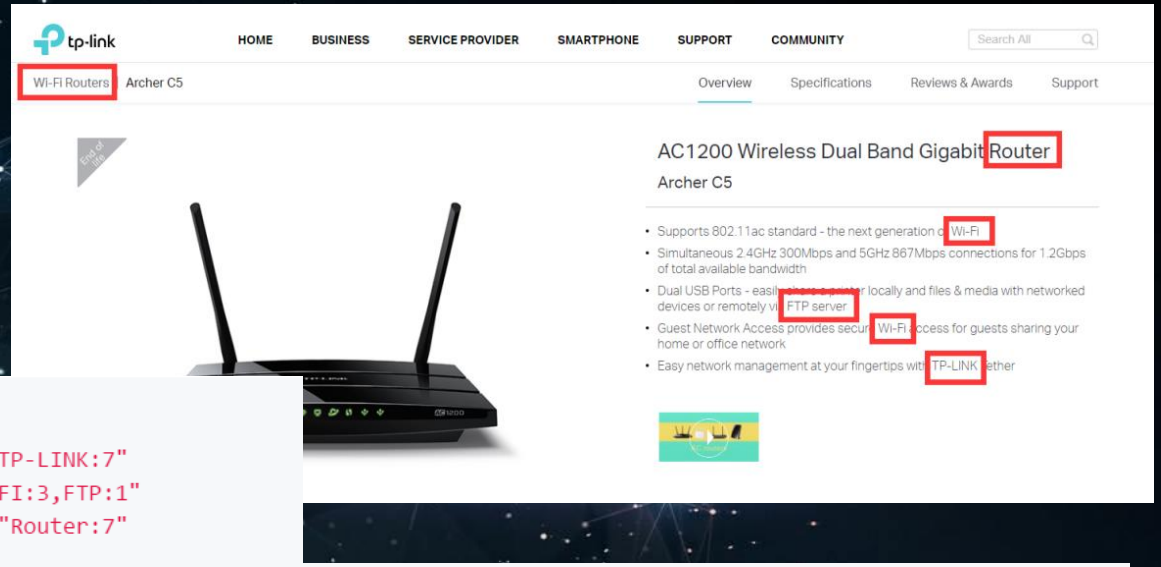
We need to know the Internal IP Address of your TP-Link Archer C5 router before we can login to it.

TP-Link Archer C5 IP Addresses
192.168.0.1

If you did not see your router's ip address in the list above. There are 2 additional ways that you can determine your router's IP address:

1. You can either follow our How To Find Your Routers IP Address guide.
2. Or you can use our free software called Router IP Address.

Now that you have your router's **Internal IP Address** we are ready to login to it.




```
{
    Vendor_name"TP-LINK:7"
    App_Type:"WIFI:3,FTP:1"
    Device_Type:"Router:7"

}
```

```
    Vendor_name: "Huawei, D-LINK, TP-LINK, DaHua, Bacnet..."
    App_Type: "ICS, IOT-Cam, IP-Cam, WI-FI, Web, SSH, FTP, POP3..."
    Device_Type : "PLC, SCADA, RTOS, Router, UPS..."
    ...
    ...

}
```

举个栗子

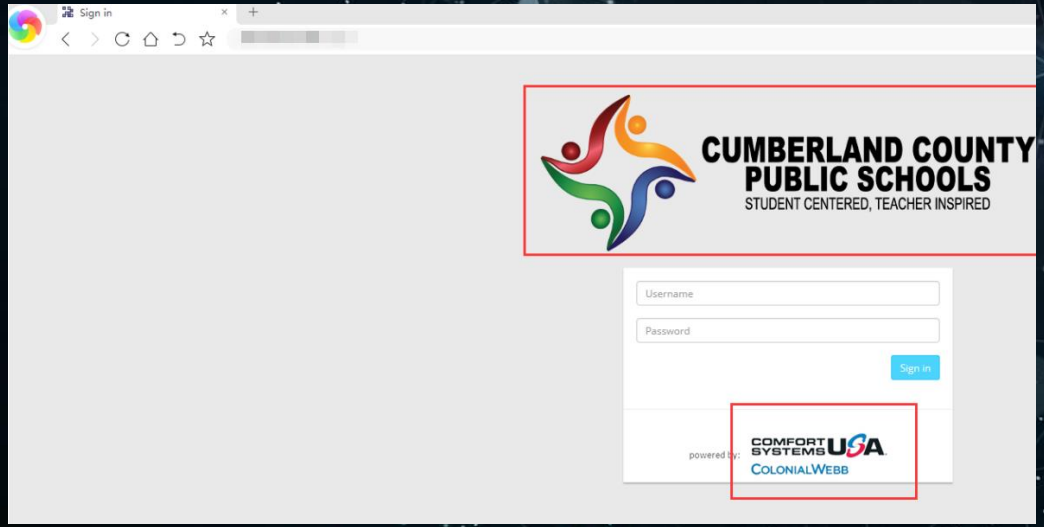
假设我们设定，出现次数最多的相关内容即为设备相关信息时，程序就会自动生成一条规则

这条规则规定了，当关键字中包含”TP-LINK Archer C5”时，将标注设备生产厂商为 “TP-LINK”,应用场景为”WIFI”，设备类型为”Router”。

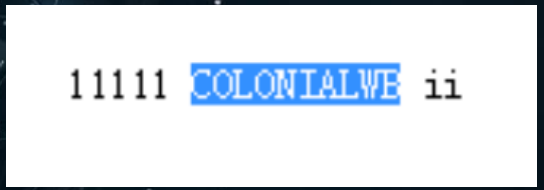
```
{
    Keyword :  "TP-LINK Archer C5"
    {
        Vendor_Name : "TP-LINK"
        App_Type : "WI-FI"
        Device_Type : "Router"
    }
}
```

测试结果

当我们进行实际测试的时候发现，并不是所有设备的通讯数据中包含字符串关键字，据不完全统计有30%的设备无法直接获取到相关信息，例如



我们需要进行图片文字提取在进行关键字查询



测试结果

面对未知设备，我们的设备识别率达到了88%

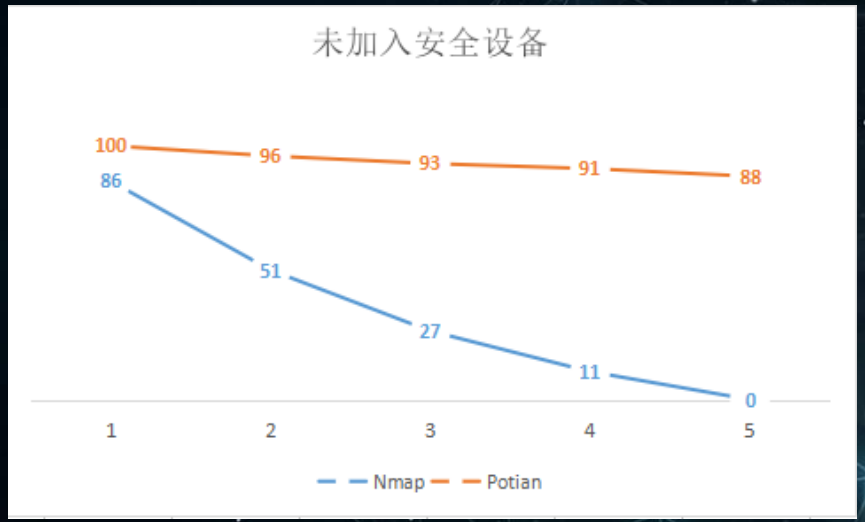
我们选择当无法获取到关键字或关键字搜索无效后，会搜索是通讯中是否包含图片信息，如果有我们将会提取图片中的文字信息再次进行规则匹配，这样操作后无法获取关键字的设备只占到了全部设备的7%

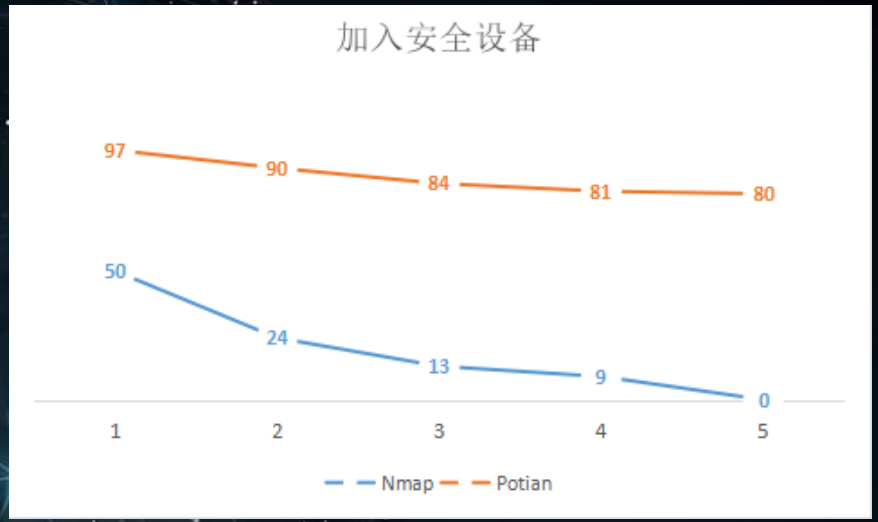
即设备识别准确度98%，设备漏报率为10%

面对已知设备我们的识别率达到了95%

即设备识别准确度99%，设备漏报率4%

测试结果


未加入安全设备
100
96
93
91
88
86
51
27
11
0
1
2
3
4
5
Nmap
Potian


加入安全设备
97
90
84
81
80
50
24
13
9
0
1
2
3
4
5
Nmap
Potian

总结

当然以上只是逆鱼所采集的规则带来的实验效果，与现一代扫描器诸如上文提到的 Nmap 、 Zmap 、masscan 等，它并不在速度上具有优越性，尤其是Masscan等采用了PF_RING框架去处理数据包的扫描器。

它真正的价值在于实践了上文提到的自动化设备标注规则采集，它可以自动化的去识别未知设备并自动编写相应的规则并存储到本地规则库。我们做过一项测试，一个经验丰富的工程师，在面对未知设备或服务时，从识别到编写出来可以用于实战的设备识别脚本，需要15-90分钟，而逆鱼只需要1分钟就可以完成相应工作。

第二部分

自动化设备标注引擎的开发思路

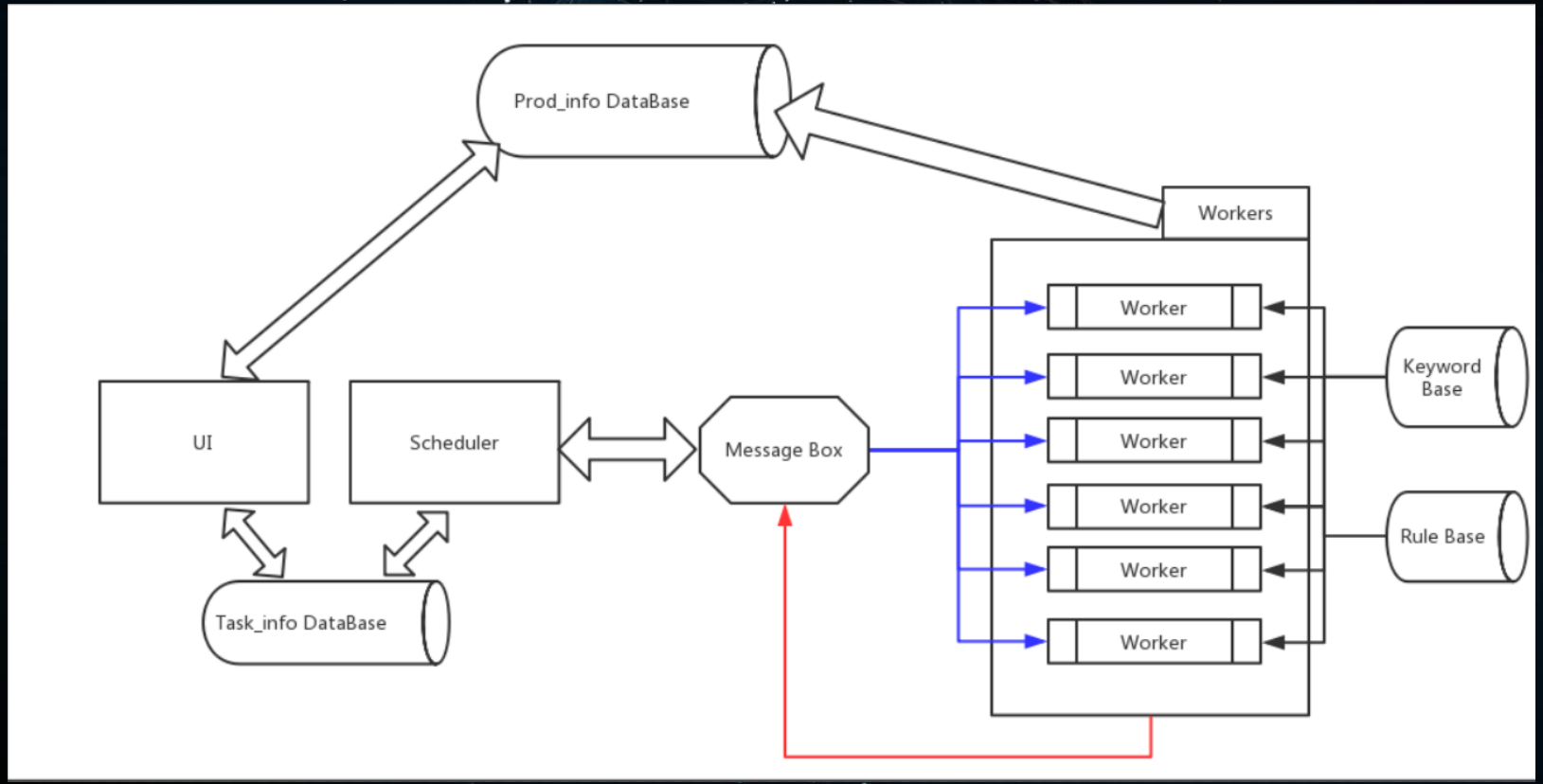
自动化设备标注引擎

自动化设备标注引擎不仅要具备自动识别未知设备的功能，同时它应具备自动的设备存活探测、设备信息标注、设备脆弱性探测以及漏洞检测功能。

我们采用 Celery 框架作为任务处理框架，使用 PF_RING 作为数据包处理框架，使其具备高效率的任务处理功能，使用逆鱼作为设备标注和规则采集模块，结合 Metaspliot 的漏洞检测功能实现了一个自动化设备标注引擎


1
Celery 分布式架构
2
数据库隔离
3
任务处理过程隔离

系统流程图


Prod_info DataBase
Workers
Worker
Worker
Worker
Worker
Worker
Worker
Keyword Base
Rule Base
UI
Scheduler
Message Box
Task_info DataBase

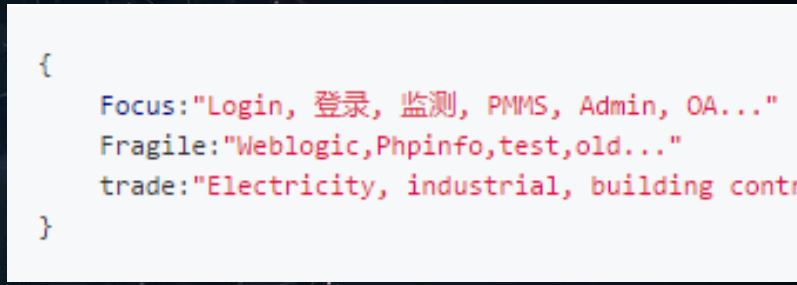
第三部分

自动化设备标注引擎的价值

用于自动生成基于设备的标注规则，面对未知设备无需人工对其进行识别和标注

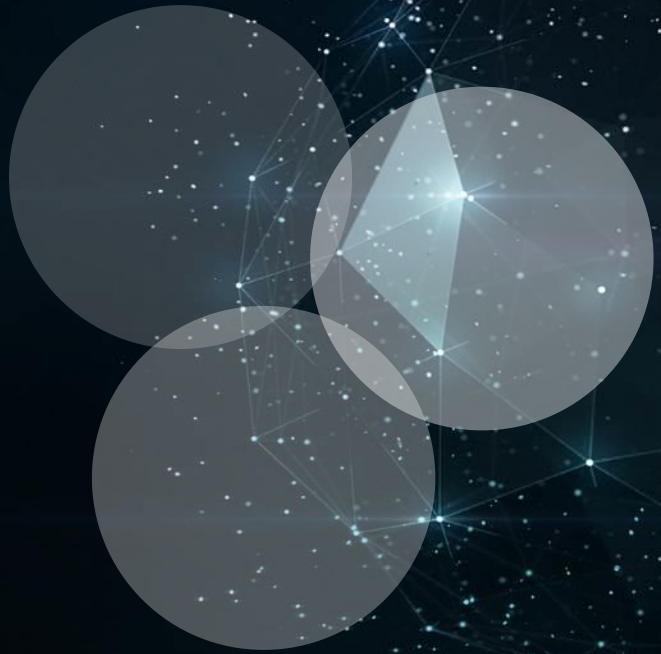
针对于网站，可应用在CMS识别以及业务分类，自动分析网站的重要程度以及脆弱程度

随着规则库的不断更新和积累会使其设备标注能力随着时间不断增强



第四部分

自动化设备标注引擎
未来的路



不仅仅局限于通信时所带有的关键字

与人工智能有更多的结合

制定统一设备信息标准，实现共享设备标注规则库


THANK YOU