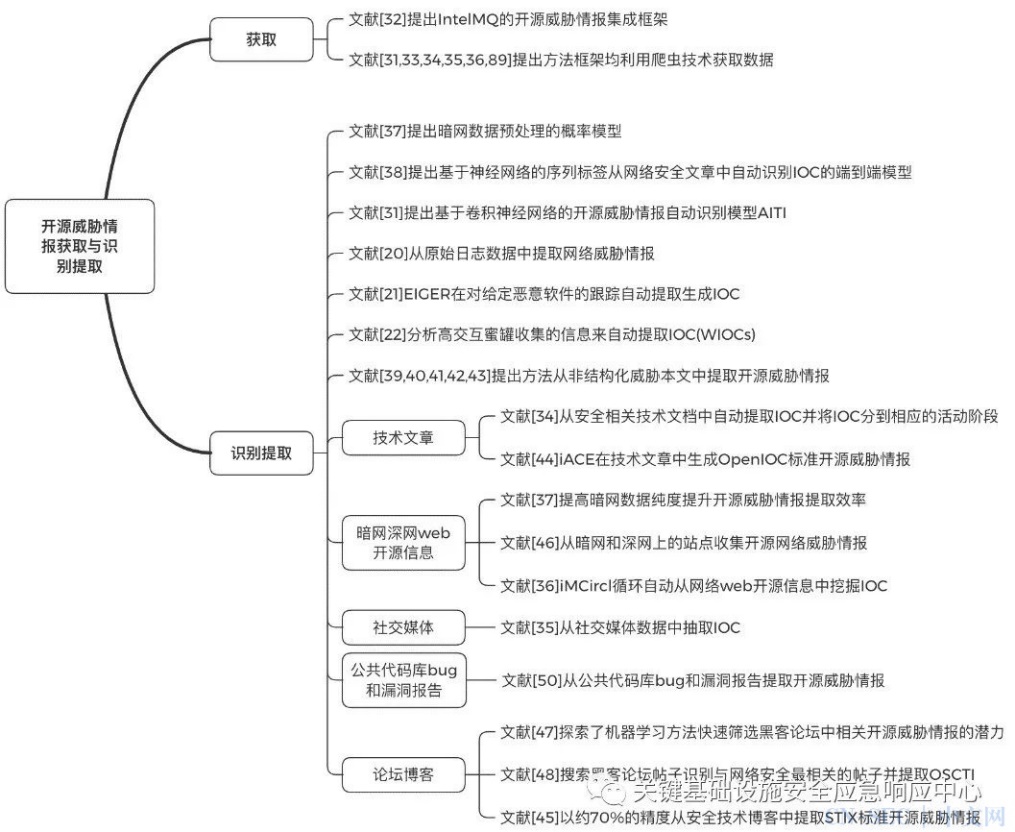
参考维度：1.数据集 2.nlp相关性 3.论文数目 4.专业性低一些

* 开源威胁情报挖掘

在当前研究工作中，主要用过命名实体识别结合SVM、逻辑回归、随机森林、朴素贝叶斯等机器学习分类方法进行话题分类

开源威胁情报另外一个重要流程是威胁情报的识别和提取。这部分工作包括信息预处理、IOC提取与威胁情报生成等子模块。其中信息预处理主要运用文本处理方法删除掉下载信息中的非用户生成信息；IOC提取一般是通过应用NLP中NER技术或其他人工智能技术，如正则表达式匹配，BiLSTM+CRF等，针对预处理过的非结构文本信息进行遍历定位出IOC，并应用机器挖掘技术获取目标实体关系，最终根据实际需要进行标准化威胁情报格式输出。





威胁情报也有其固定的栺式，目前主流的

有三种：CybOX、STIX、TAXII。

CybOX（Cyber Observable eXpression）主要提供一套标准且支持扩展的语法，用来描述判断威胁的指标；STIX（Structured Threat Information eXpression）提供了描述威胁情报内容的斱法，比如利用手段、攻击目标、攻击目的、身份等多项特征，详细地描述威胁攻击事件的细节；TAXII（Trusted Automated eXchange of Indicator Information）旨在提供安全的传输和威胁情报信息的自动化交换。当前的应用是通过 TAXII 来传递共享数据，用 STIX

来对威胁情报进行描述 ，用 CybOX 来提供威胁情报词汇。

* Web管理后台

前期经过长时间的数据采集、去脏、打标，我们积累了充足的网页源码样本，但是页面的源码需要经过预处理才能转化为模型可理解、可计算的形式。具体过程主要是去除页面中无用的代码、符号等干扰因素，之后将页面进行分词，具体规则简略描述如下

机器学习方案由于不依赖关键字，具有良好的泛化能力，能识别传统基于关键字方案漏报的部分；同时，模型可通过不断迭代自进化，灵活度高；在识别能力上，机器学习模型是通过综合学习多维特征，建立各维度关联关系，从而指导决策，具备更缜密的判断逻辑。新的识别方案上线之后，也确实有很多意外收获，比如识别出了Django的调试页面等等。

由于管理后台一般情况下相较普通用户登录页面而言更加简洁，所以这里引入了信息熵来去除相对复杂的页面，比如部分新闻页面中存在登录框导致的误报等，其他步骤则主要是去除干扰的文本，最终获得一系列的英文单词和中文词语。

紧接着，我们选择经典的word2vec对数据做进一步处理。word2vec通过计算向量之间的距离来体现词与词之间的关系，就是简单化的神经网络，我们通过训练word2vec模型将文本内容转化为固定长度的数字构成的向量，这些带有文本特征的向量作为机器学习模型输入，优势体现在：

加速模型收敛；

降低输入的维度，节省运算开销；

增加语义信息，有利于后续模型的特征学习。

* 恶意软件检测

Tobiyama提出一种基于进程行为进行恶意进程检测到方法，研究人员利用LSTM（Long Short-Term Memory，长短期记忆网络）进行特征提取，用CNN（Convolutional Neural Networks，卷积神经网络）进行特征分类，而进程行为是一系列的API调用序列。特征是从转化为含有本地特征的图片的进程行为日志文件中提取的，这些本地特征表示进程活动。

Rhode调查了基于行为数据对可执行文件是否恶意文件进行预测。模型使用Cuckoo沙箱来手机PE样本的活动数据。收集的特征有：系统CPU利用率、用户CPU使用、发送的包、收到的包、发送的字节、收到的字节、内存使用、交换区使用、当前运行的进程总数和分配的最大进程ID等。考虑到训练速度的问题，作者选择了GRU（Gated Recurrent Unit ，门控循环单元）来替代LSTM

Hardy提出一种基于Stacked AutoEncoders（SAE）的恶意软件检测模型。模型会使用收集的PE文件产生的Windows API调用。如图所示，使用PE解析器从每个文件中提取Windows API调用。API查询数据库会把API调用转变成对应API函数的32位表示。因此，SAE也被用于特征学习、参数调优和恶意软件检测。该模型的准确率大概是96.85%。

**Hou提出一种基于Linux kernel系统调用和SAE的安卓恶意软件检测框架。使用了一种名为component travelsal的动态分析方法来对每个给定的安卓应用的代码路径进行自动执行。为了获取系统调用之间的关系，作者构建了一个带权有向图，其中节点代表系统调用，节点的大小表示频率，边表示系统调用的序列流。**

**roidDetector是一款基于安卓恶意软件检测引擎的在线深度学习应用。作者Yuan等通过动态分析和静态分析从每个APP中提取特征，提取的特征分为三类，分别是必须的权限、敏感API和动态行为。DroidDetector的检测准确率大约为96.76%，由于传统的机器学习算法。在静态分析阶段，会对Android-Manifest.xml和classes.dex进行语义分析来得到APP请求的120个权限。动态分析包含在DroidBox中运行每个APP，进行动态标记分类并监控13个APP行为的分析。**

**恶意外连流量**

