A Linguistics-based Stacking Approach to

Disposable Domains Detection

一次性域名

Vogers 系统 分为三个模块

1 模块提取器，获取模块语言特镇

2 三个并行的可读性评分赋值器， 前缀 前缀中最长的标签。后缀 三部分语言特

3 二进制分类器

度量标准。长度和单词比率。 计算单词比率时应用了开源的的记号赋予器，根据预先加载的字典将字符串分割成记号

域名包扩 字母 数字 连字符。 Hello-12。被拆成 hello – 1 2。

S是个子域 k1 k2 k3是s的令牌串

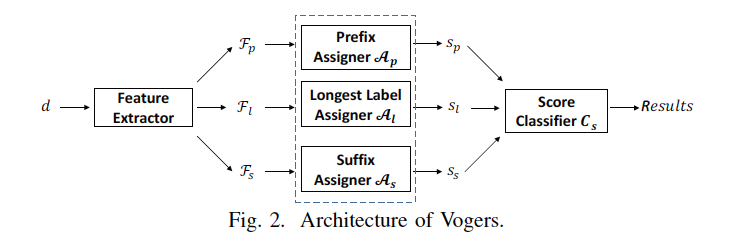
L（x）是序列x中的令牌数

 有太多令牌的字符串倾向于较低的比率

正常域名长度低于30。一次性域名只有60低于30

单词比例。 0.5为阈值则95%的一次性域名低于这个阈值，80%的正常域名高于

为了尽可能多地传输数据，可丢弃域使用某些压缩算法，这使得子域冗长且不可读。而普通的域名是用来提供直接的互联网服务的。它们总是简短易记，因此提示用户经常访问



然而，给定一个多级域名，很难区分哪个单独的点是前缀和后缀之间的分界点。因此，为了简化这个过程，我们保守地指定2LD作为后缀，其余的作为前缀。

因此，我们应该有意识地放大自动生成部分对整体可读性的影响。在这里，我们选择提取前缀中最长的标签作为“注意信号”，并将其与前缀放在同等的基础上。鉴于以上情况，我们从前缀、前缀中最长的标签和后缀三个方面来考虑域名的可读性。

然后，我们提取了每个方面的语言特征，将在第四节中详细阐述。

为了便于表示，我们分别将域名d的前缀、前缀中最长的标签和后缀标记为dp、dl和ds。最后，该模块输出三个特征向量，即dp的特征向量Fp、dl的特征向量Fl和ds的特征向量Fs。其中，Fp有12个维度，Fl和Fs分别有10个维度。然后这些特征向量被输入到下一个模块中。

我们构建了三个可读性分数赋值器来学习dp中不同域类型之间的可读性差异，dl和ds，分别表示为Ap、Al和as。每个赋值者以一个语言特征向量作为输入，并且输出介于0和1之间的可读性得分。越高分数越高，可读性越好。

更具体地说，每个赋值器实际上是一个二进制分类器。在训练阶段，我们根据不同的学习目标，采用不同的训练集或每个指派者。训练Ap时，数据集由可丢弃域和正常域的前缀组成。在训练Al时，我们直接提取每个前缀集中的样本构成此任务的训练集。在训练As时，我们采用了两个权威公共数据集的有效2ld。其中一个是Alexa榜单，它记录了最流行的良性域名。另一个是DGArchive[5]，它提供每日生成的DGA样本。在每个数据集中，我们分别构造了30000个正样本和30000个负样本。与自动生成的部分关联的特征向量标记为0（负），而与人造部分关联的特征向量标记为1（正）。

当训练达到收敛时，给定一个特征向量作为输入，相对赋值者将输出正负两个类别的概率。我们把正范畴的概率值作为最终的可读性得分。最后，该模块输出三个分数，即sp、sl和ss，分别对应于dp、dl和ds的可读性。

到目前为止，我们已经得到了域名三个代表部分的可读性得分。为了达到最终的检测目标，我们通过叠加集成将三个输出的可读性得分聚合成一个三维特征向量。接下来，我们构造了一个二值分类器，标记为Cs，它以聚集的特征向量作为输入，并输出最终的判别结果。为了训练这个分类器，我们仔细地选择10000个一次性区域作为负样本（标记为0），10000个正常区域作为正样本（标记为1）。为了避免错误地将恶意DGA域分类为可丢弃域，我们在正样本集中添加了5000个DGA域。所有这些样本都通过前两个模块转换成三维特征向量。重要的是，额外选择恶意DGA域作为阳性样本并不意味着恶意DGA域的可读性与普通域相似。我们真正要强调的是可丢弃域的前缀和后缀之间的可读性差异。

在这一部分中，我们设计了一系列的语言特征来量化所谓的可读性。受[6]、[7]的启发，我们构建了12个基本的语言特征：

（F1）长度。如图1（a）所示，为了携带尽可能多的信息，自动生成的部件自然比人工生成的部件长。

（F2）标签级别。由于dl和ds都是域d中的标签之一，因此我们只在dp中使用此功能。除此之外，我们还考虑了分隔点和连字符作为标签的分割点。如表三所示，与正常域相比，可弃域的子域往往具有更多的划分点，因此具有更高的标记级别。

（F3）平均标签长度。此功能等于标签的总长度除以标签级别。因为dl和ds都只有一个标签，所以这个特性专门用于前缀dp。

（F4-F6）元音比、辅音比和数字比。这三个特征分别反映了给定字符串中元音、辅音和数字的比率。元音和辅音在单词中的比例和位置决定了它的发音方式。只有一个元音-辅音比例合适的域才能更容易被记住并被不断地查询。

（F7-F8）：连续辅音比，连续数字比。如果一个长度大于2的字符串由纯辅音（数字）组成，我们认为该字符串是连续的辅音（数字）。这两个特征分别计算连续辅音和连续数字在给定输入中的比例。

F9：字符基数。此功能指字符串中不同字符的数目。字符串越长、越不可读，此功能就越高。

F10：内部数字编号。算法生成的字符串总是由字母和数字交替组成。然而，上述特性都不能揭示字母和数字之间的这种位置关系。我们引入这个特性来反映自动生成的字符串中的位置关系。为此，我们首先从给定字符串中提取所有连续的数字。接下来，我们判断与这些子字符串相邻的两个字符是字母还是连字符。

F11：字数比率。我们使用这个特性来显式地反映给定字符串中有意义单词的比例。第二节介绍了词比的具体计算方法。

F12：熵。这个特性反映了给定字符串的不确定性程度。该特征的具体计算方法可参考Shannon熵。

本节介绍了我们为评估Vogers的性能而进行的实验。首先，我们讨论了分类器的选择问题。接下来，我们将Vogers与现有技术进行比较。

A、 分类器选择

Vogers中有两个模块apply classifiers。一个是可读性分数赋值器，另一个是分数分类器。其中，前者并行使用三个分类器，后者则需要一个分类器。在本小节中，我们将阐述每个模块中最适合的分类模型。

1） 可读性评分分配人：如第三节B所述，我们为每位分配人准备了30000份阳性样本和30000份阴性样本。对于Ap和Al，我们挑选出100个一次性区域和100个正常区域，然后从每个区域中提取300个域。每区300个样本中，200个用于训练，100个用于测试。关于As，我们跳过了域区域的选择

直接从Alexa Top List和DGArchive中提取30000个样本，20000个用于培训，10000个用于测试。

我们通过实验为每个指派者选择性能最佳的分类模型。在这里我们选择了三种常用的分类模型作为候选，即随机森林（RF）、梯度提升决策树（GBDT）。这些指派者的任务是根据输入的特征向量为相关领域指派合理的可读性得分。因此，我们不必追求过分精确。我们评估了RF和GBDT的AUC值，以及它们的TPR

阈值为0.5时的值和FPR值。评价结果见表二。与RF相比，GBDT具有较高的AUC值和较低的FPR值。因此，我们决定采用GBDT作为三个可读性评分分配者的分类模型。

2） 分数分类器：该模块以三个可读性分数作为输入，输出最终的分类结果。与可读性评分相比，本模块更注重准确性。我们从Ap使用的训练集中提取14000个一次性样本和14000个正常样本。此外，我们在这个集合中添加了5000个恶意DGA样本。这里，一次性域被标记为1，并且

其他的标记为0。接下来，我们将这一组分为两部分，80%用于训练，20%用于测试。同样，我们使用RF和GBDT作为该模块的候选分类模型。表2中的指标表明，在相同AUC值为0.999的情况下，当阈值设置为0.5时，GBDT的FPR值仅为0.001，比RF的FPR值低0.002。因此，我们选择GBDT作为分数分类器。

与现有技术的比较

我们重建了文献[3]中提出的方法，并将其性能与V-A2中使用的Vogers数据集进行了比较。我们使用AUC、TPR和FPR来评估这些方法的性能。在计算TPR和FPR的值时，阈值仍设置为0.5。具体的评估结果如表三的起始栏所示，从中我们可以清楚地看到，Vogers在检测一次性域方面具有压倒性的优势。与先前的方法相比，Vogers将AUC值至少提高了0.0588，并将FPR值降低了0.19以上。

结合Zone-Miner的性能差和第二节的分析结果，我们可以得出一个初步的结论，即缓存命中率和子域尺度都不能从根本上刻画可丢弃域。主要原因是绝大多数DNS流量被几个热点域区域垄断。尽管一次性域只会被查询一次，但它显示的流量配置文件几乎与许多不受欢迎的域相同，尤其是前面提到的博客和电子商务域。真正支配Vogers的是它掌握了可弃域的本质，即可弃域的子域是自动生成的。通过某种算法将数据压缩成固定长度的字符串，不可避免地会使字符串无法满足常规的词法规则。为了在有限的域长度上传输尽可能多的信息，生产者必须对数据进行批量压缩，使相应的域区域充满混乱的子域。

六、 结论

本文讨论了可弃域的问题。这些自动生成的域被一些Internet服务用来从客户端收集临时信息。我们提出了一个基于语言学的叠加模型Vogers，来识别这些可丢弃的域。评价结果表明，Vogers检测一次性域的真阳性率为98.9%。此外，与现有技术相比，其假阳性率降低了19%以上。