福昕高级PDF编辑器

高效·安全·专业

立即下载　　点击购买

- OFFICE格式互转
- OCR文字识别
- 文本图像编辑
- 加密和签署
- 交互式动态表单
- 互联PDF文档

中国科学院大学
University of Chinese Academy of Sciences

Detection of Parking Domains

Domain Security Homework

Huang ZH Cui SY Wang ZX

2018/12/10

Outline

1. **引言概述**

   1. 域名停靠定义　　2. 特点及安全问题　　3.方案设计

2. **数据观察统计**

   1. 高频二级域名统计　　2. 可疑域名统计

3. **可获取源码域名判别**

   1. 数据准备　　2. 特征提取　　3. 模型训练

4. **不可获取源码域名判别**

   1. 获取DNS和Whois　　2. 数据统计　　3. 域名过滤

5. **总结**

中国科学院大学
University of Chinese Academy of Sciences

1. 引言概述

Domain Security Homework

相关背景知识

域名停靠:

如果域名所有者有一个理想的域名或者带有流量的域名，但又不想立即使用域名建立网站，这时就有可能会选择域名停放服务自动生成包含了广告商链接的广告网页，域名所有者将从域名停放服务中获得收益。

引言概述

University of Chinese Academy of Sciences

实例

xiaomw.com

相关背景知识

- 特点

1. 空域名，没有任何服务
2. 停放广告或不良信息
3. 域名与现有的大流量的域名相似（抢注域名）

Domain Security Homework

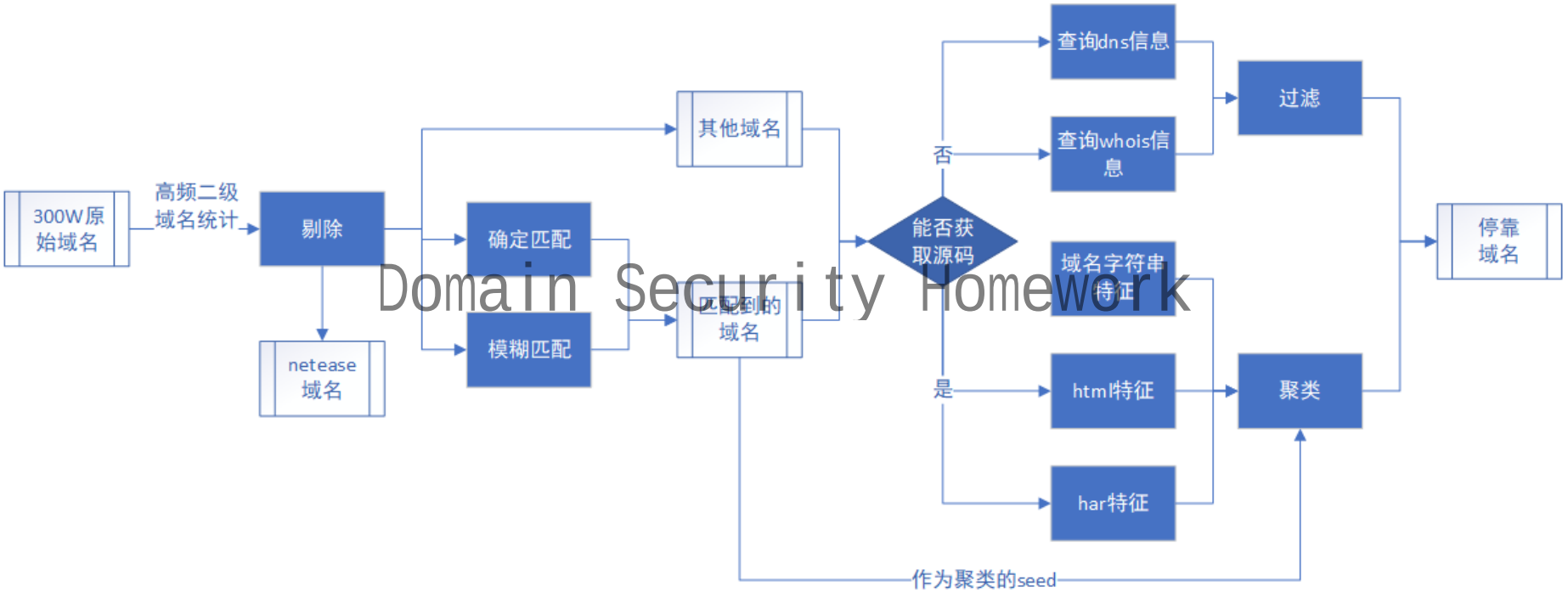
- 安全问题

1. 通过注册域名和域名停靠服务等环节，形成的黑色产业链。
2. 被用来做搜索引擎优化（SEO）

引言概述

University of Chinese Academy of Sciences

方案设计



中国科学院大学
University of Chinese Academy of Sciences

2. 数据观察统计

Domain Security Homework

高频二级域名统计

对原始300万数据中的FQDN的二级域名进行统计。探索了解数据集的分布情况。

通过截取每个域名的二级域名部分，且统计出现频率最高的二级域名如下：

二级域名	
netease	588946
irs01	274097
imtmp	233467
ksyunacc	28581
dnssina	25161
baiduyundns	20517
taobao	18988
baiyangzs	15083

高频二级域名统计

- 从表中可以看到二级域名为netease的在总体样本中占到20%的比例。加上irs01，imtmp，这三个二级域名占到总体比例的大约34%。
- 针对irs01，imtmp查询whois信息：

Domain Name: irs01.com
Registry Domain ID: 1679194804_DOMAIN_COM-VRSN
Registrar WHOIS Server: grs-whois.hichina.com
Registrar URL: http://whois.aliyun.com
Updated Date: 2018-09-29T01:54:23Z
Creation Date: 2011-09-27T09:13:43Z
Registrar Registration Expiration Date: 2019-09-27T09:13:43Z
Registrar: **Alibaba Cloud Computing (Beijing) Co., Ltd.**
Registrar IANA ID: 420
Reseller:

Domain Status: ok https://icann.org/epp#ok
Registrant City:
Registrant State/Province: shang hai
Registry Registrant ID: Not Available From Registry
Name Server: NS1.DNSV2.COM
Name Server: NS2.DNSV2.COM
DNSSEC: unsigned
Registrar Abuse Contact Email: DomainAbuse@service.aliyun.com
Registrar Abuse Contact Phone: +86.95187
URL of the ICANN WHOIS Data Problem Reporting System: http://wdprs.internic.net/
>>>Last update of WHOIS database: 2018-11-16T11:56:29Z <<<

高频二级域名统计

imtmp whois：

Domain Name: imtmp.net

Registry Domain ID: 1578231260_DOMAIN_NET-VRSN

Registrar WHOIS Server: whois.markmonitor.com

Registrar URL: http://www.markmonitor.com

Updated Date: 2018-01-09T04:00:44-0800

Creation Date: 2009-12-07T17:46:48-0800

Registrar Registration Expiration Date: 2020-12-07T17:46:48-0800

Registrar: MarkMonitor, Inc.

Registrar IANA ID: 292

Registrar Abuse Contact Email: abusecomplaints@markmonitor.com

Registrar Abuse Contact Phone: +1.2083895740

Domain Status: clientUpdateProhibited (https://www.icann.org/epp#clientUpdateProhibited)

Domain Status: clientTransferProhibited (https://www.icann.org/epp#clientTransferProhibited)

Domain Status: clientDeleteProhibited (https://www.icann.org/epp#clientDeleteProhibited)

Registrant Organization: **Shenzhen Tencent Computer Systems CO.,Ltd**

Registrant State/Province: Guang Dong

Registrant Country: CN

Admin Organization: Shenzhen Tencent Computer Systems CO.,Ltd

Admin State/Province: Guang Dong

Admin Country: CN

Tech Organization: Shenzhen Tencent Computer Systems CO.,Ltd

Tech State/Province: Guang Dong

Tech Country: CN

Name Server: ns1.imtmp.net

Name Server: ns2.imtmp.net

Name Server: ns3.imtmp.net

DNSSEC: unsigned

URL of the ICANN WHOIS Data Problem Reporting System: http://wdprs.internic.net/

>>> Last update of WHOIS database: 2018-11-16T06:59:30-0800

高频二级域名统计

在通过上面比较简单的对数据进行一个统计，并查询一些相关的whois信息后，决定剔除二级域名为netease的样本，暂时认为这些是正常域名，从而减少后续要处理的样本总量。

可疑域名统计

通过搜索引擎和相关报告搜集Parking-service的服务器记录与原域名的NS记录和CNAME记录进行匹配

Parking Service	Parking Service	Parking Service
sedoparking.com	dotzup.com	internettraffic.com
namedrive.com	voodoo.com	1and1.com
parked.com	parkingcrew.com	namesilo.com
whypark.com	rookmedia.net	above.com
astoriacompany.com	bodis.com	dnsexit.com
fabulous.com	smartname.com	ztomy.com
domainsponsor.com	parklogic.com	pql.net
trafficz.com	domainapps.com	airportparkingtucson.com
domainhop.com	trafficz.com	
parkingdots.com	dopa.com	

ww1.909900tk.com

NS: sedoparking.com

方法

① 确定匹配： 从论文和资料查询获取到常见Parking-Service提供者服务器的记录50条，对CNAME与NS记录中包含这些特殊资源记录的域名进行标注；

② 模糊匹配：对NS记录和CNAME记录中包含“park“ 敏感词的域名进行标注得到1030条记录；再人工过滤掉一定不是parking domain的域名。

③ 直接匹配： 对域名字符串中包含“park“ 敏感词的域名进行标注(但人工筛选后发现这种匹配方法常匹配到停车场、公园的域名，因此放弃这种匹配的结果)

3. 可获取源码域名判别

Domain Security Homework

可获取源码域名判别

根据[1] Parking Domains具有如下特征

- 页面具有第三方广告链接较多；
- 存在通过重定向跳转和通过iframe内嵌广告页面的情况；
- 在域名字符串角度，可能是某些流行域名的typo-squatting

[1]NDSS2015-Parking Sensors Analyzing and Detecting Parked Domains

面临问题

已标注数据少，无法通过有监督算法进行有效学习；

方法

以有限的已标注的park domains作为seed，通过聚类发现新的可疑域名

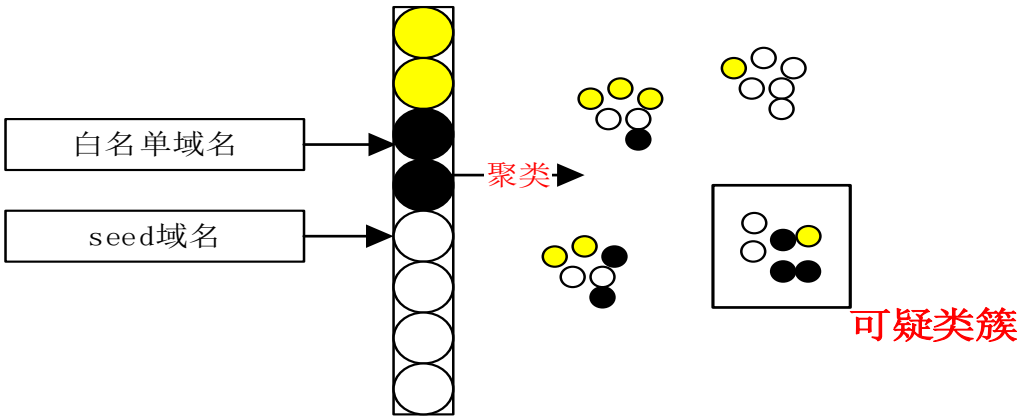
可获取源码域名判别

域名数据

- 白名单域名：Alexa TOP 3000 (额外从ALEXA TOP网站获取,2300个)
- Seed域名：① “确定匹配”和“模糊匹配”所得域名；

  ② 通过WHOIS信息反查手动获取

  对①②所得域名进行人工筛选，得到1128个域名作为seed
- 待检测域名：不属于其他以上两类的域名且可获取源码的域名560162个



数据获取

HTML数据获取

借助scrapy框架，访问指定的域名，直接保存原始页面的所有源代码，以供后面多次处理。

HAR数据获取

通过selenium模拟浏览器，选择直接提取har特征，而不保存har文件。对每个har的['log'][‘entries‘]中的每个请求响应的参数进行记录，计算出第三方的请求比例、第三方的数据比例、第三方的html内容比例、初始响应大小和比例。

字符特征信息获取

① 对主域名字符串中的数字比例进行统计；

② 通过对主域名中字符的增、减、替换、顺序调整以及顶级域替换等检验方式，判断改域名是否可能是某个流行域名的typo.

可获取源码域名判别

特征项

特征项	含义	提取来源
size	源文件大小	HTML 源码
a_nums	<a>标签数量	HTML 源码
href_max_length	href 链接最长长度	HTML 源码
href_avg_length	href 链接平均长度	HTML 源码
avg_a_href	href 链接数量 / <a>数量	HTML 源码
external_link_ratio	外链比例	HTML 源码
frame_nums	<frame>与<iframe>数量之和	HTML 源码
location_flag	标识源码中是否包含 window.location	HTML 源码
digit_ratio	域名字符串中数字比例	域名字符串
digit_flag	标识域名是否是纯数字域名	域名字符串
typo_flag	标识域名是否是某域名的 typo	域名字符串
third_party_request	第三方请求数量	HAR 文件
total_request_requests	总请求数量	HAR 文件
thirdParty_bodySize	第三方响应的大小	HAR 文件
thirdyparty_contentsize	第三方内容的大小	HAR 文件
total_contentsize	总内容大小	HAR 文件
first_bodysize	首个响应大小	HAR 文件
total_bodysize	总响应大小	HAR 文件

可获取源码域名判别

特征项

对seed域名和白域名通过决策树进行特征评价

RANK	特征	score
1	avg_a_herf	0.2198
2	a_nums	0.2131
3	external_link_ratio	0.1679
4	size	0.1044
5	fame_nums	0.101
6	external_link_ratio	0.0365
7	href_max_length	0.0354
8	total_contentSize	0.0215
9	href_avg_length	0.0203
10	total_body_size	0.0181
11	External_link_ratio	0.0156
12	digit_ratio	0.0136
13	typo_flag	0.0079
14	third_party_ratio	0.0058
15	location_flag	0.0057
16	digit_flag	0.0054
17	First_body_size	0.0042
18	Total_body_size	0.0016

可获取源码域名判别

特征选择

由于HAR特征含义与部分HTML特征含义有重合，同时其特征重要程度不高而且获取较为繁琐，故实际判别时舍去了HAR特征

特征项	含义	提取来源
size	源文件大小	HTML 源码
a_nums	<a>标签数量	HTML 源码
href_max_length	href 链接最长长度	HTML 源码
href_avg_length	href 链接平均长度	HTML 源码
avg_a_href	href 链接数量 /<a>数量	HTML 源码
external_link_ratio	外链比例	HTML 源码
frame_nums	<frame>与<iframe>数量之和	HTML 源码
location_flag	标识源码中是否包含 window.location	HTML 源码
digit_ratio	域名字符串中数字比例	域名字符串
digit_flag	标识域名是否是纯数字域名	域名字符串
typo_flag	标识域名是否是某域名的 typo	域名字符串

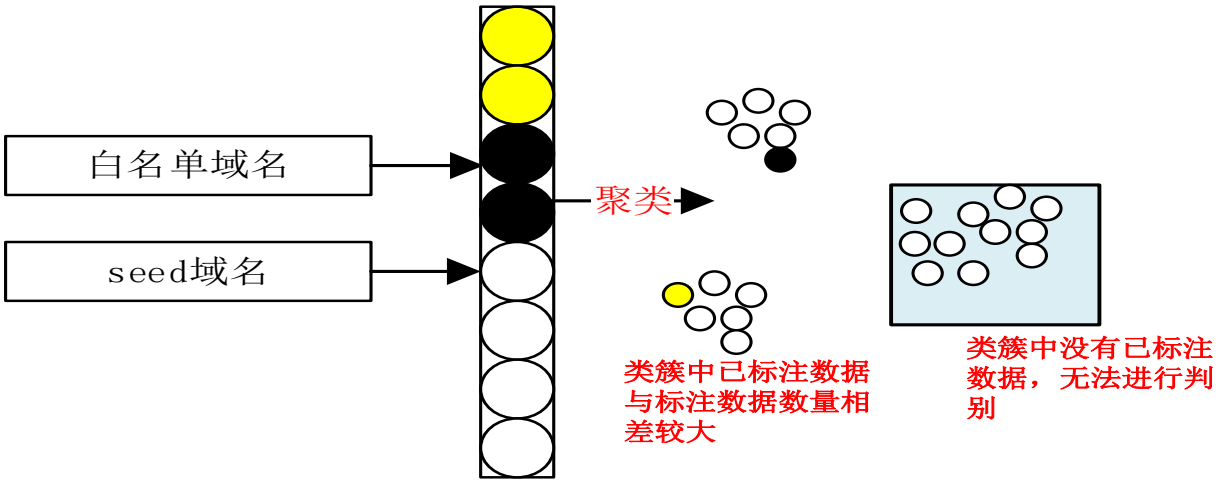
可获取源码域名判别

模型-Keans

尝试一：有标注数据+待判别数据，令k=3直接进行聚类

- 结果：标注数据与待判别数据数量悬殊，几乎无法得到有意义结果
- 改进：① 数据分组对待判别数据进行分组，使标注和待判别数据数量大致持平

②非平衡数据处理：由于已标注的可疑域名数量只有白域名数量1/2，因此对可疑域名数量进行复制来平衡数据集

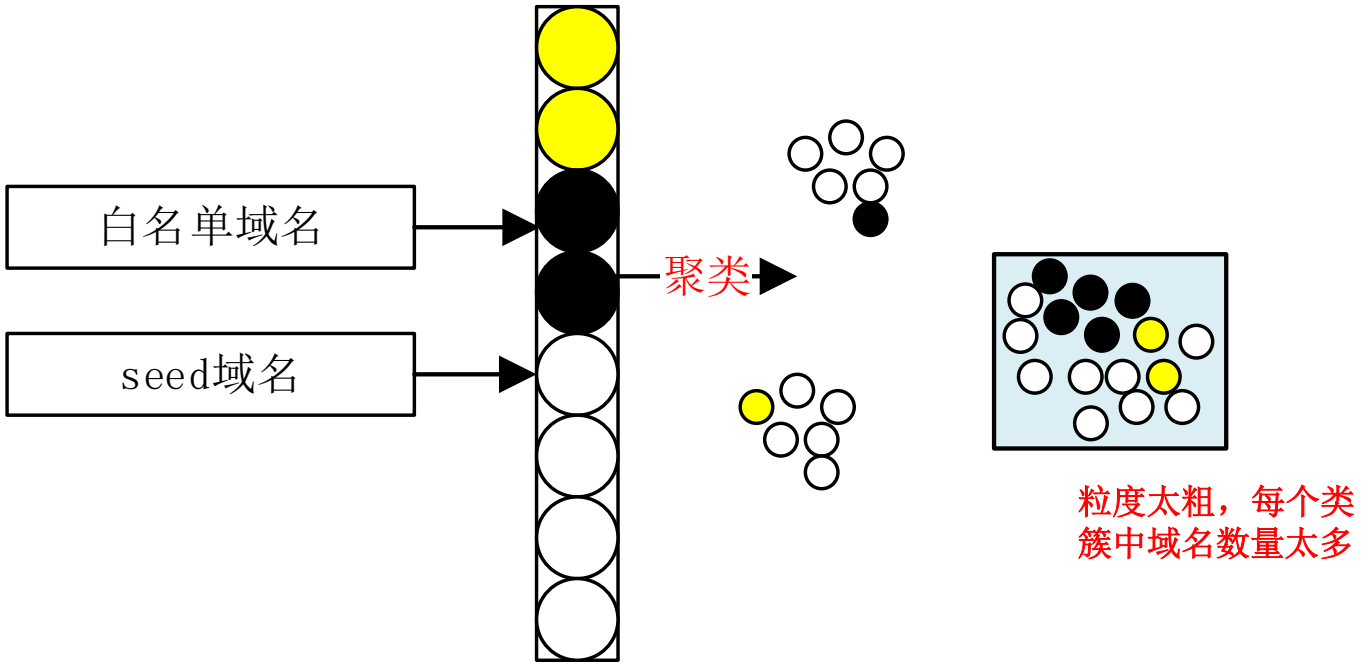


可获取源码域名判别

模型-Keans

尝试二：每4000个待判别域名与标注标注域名为一组,令k=3进行聚类

- 判别： 如果每个类簇中seed域名数量大于白域名数量，则认为该类簇中待判别域名是parking domains
- 结果：每个类簇中域名数量太多，判别结果粒度太粗
- 改进：增大k值，细化聚类结果的粒度



可获取源码域名判别

尝试三：在方法二基础上增大k值，分别令k=10,15,20,…50进行聚类

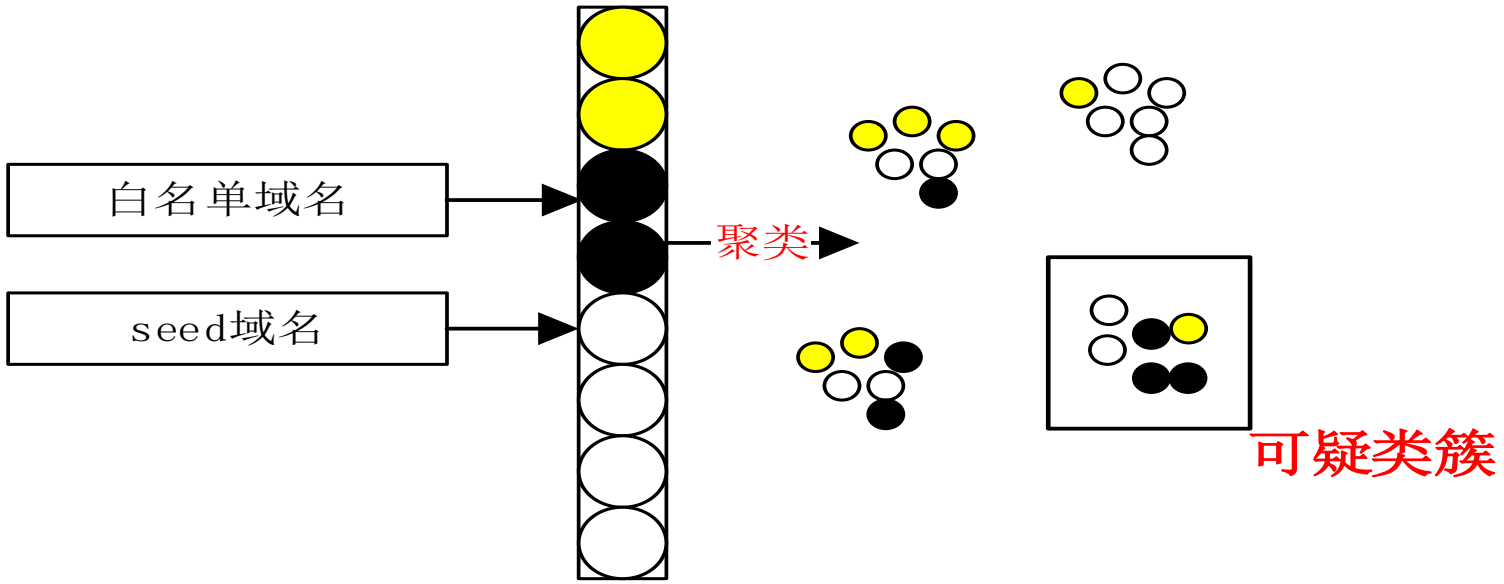
判别：①如果每个类簇中seed域名数量大于白域名数量，则认为该类簇中待判别域名是parking domains

②人工对判别为parking的域名进行抽样验证

结果：k=50时效果最佳

聚类结果

判定368271个域名为可疑的parking domains



可获取源码域名判别

中国科学院大学 University of Chinese Academy of Sciences

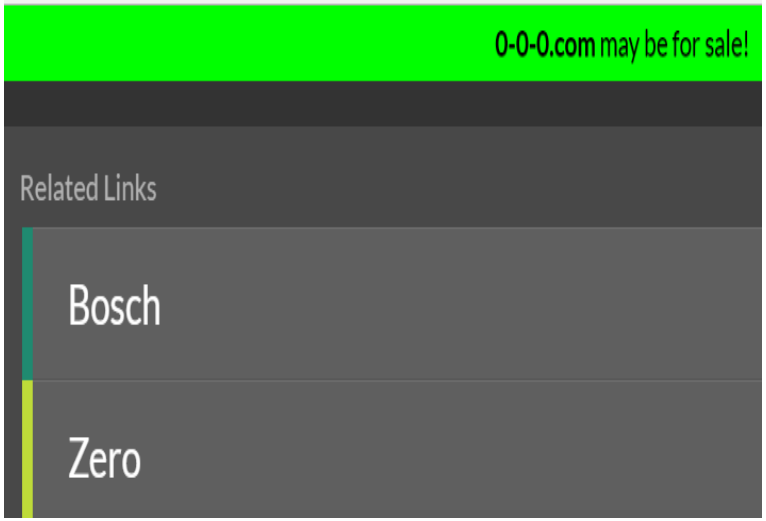
Kmeans准确性：

使用Kmeans模型对Alexa白域名和seed域名进行聚类，然后对判断结果与实际标签比对，得到准确率为54%，作为参考

最终结果

对于聚类得到的368271个可疑的parking domains，经过人工随机抽样发现并不全是parking domains,但却包含了赌博、色情类的域名。因此，我们认为聚类所得的“parking domains”能够为异常域名的发现提供一定的基础数据与参考。

聚类得到的parking domains



聚类得到的赌博类域名

4. 不可获取源码域名判别

Domain Security Homework

不可获取源码域名判别

爬取域名的DNS和Whois信息

将剩下的无法获取源码的大约170万域名，重新查询其DNS信息和Whois信息，期望从中能够发现可用于判别的信息。

爬取的DNS 字段：

Name Servers、CNAME

爬取的Whois字段：

domain name：查询的域名（二级域名）

registrar、creation date、expiration date、last updated、name servers

查询速度较慢，最终实际爬取的DNS和Whois仅约37万条。

不可获取源码域名判别

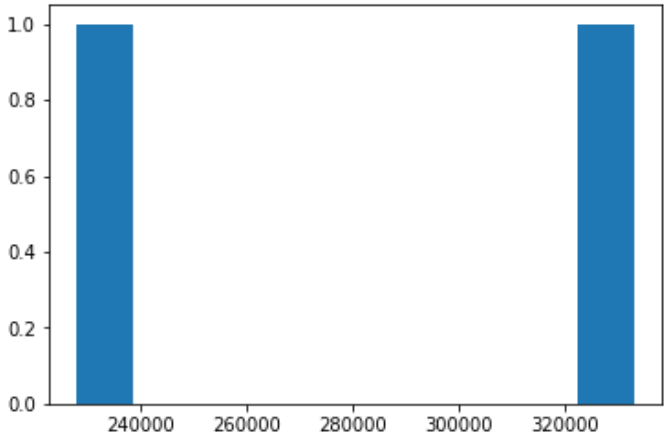
University of Chinese Academy of Sciences

统计数据缺失情况

DNS总记录：374211

name server字段缺失数量：338162

cname 字段缺失数量：233429



Whois总记录：374211

数据缺失情况严重

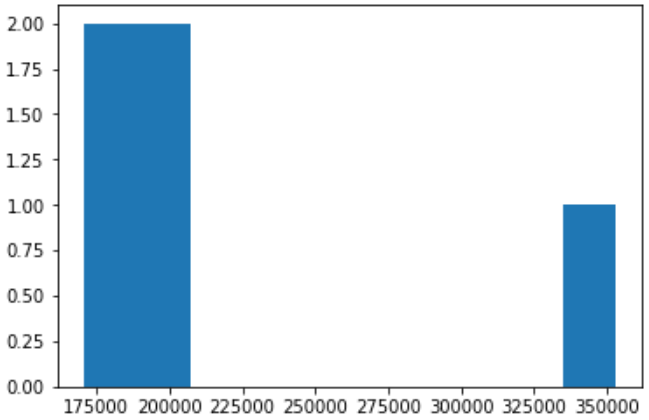
registrar字段缺失数量：179666

creation_date字段缺失数量：198314

expiration_date字段缺失数量：198165

last_updated字段缺失数量：362203

name servers字段缺失数量：182202



不可获取源码域名判别

基于DNS信息对域名过滤

用已知的停靠服务商的name servers服务器，来匹配待检测域名DNS的name servers字段。

如果域名的name server是已知的停靠服务商的服务器，则判定其为停靠域名。

[ ]:

	domain	ns	cname
0	www.gzrail.com.cn	ns1.parkingcrew.net.,ns2.parkingcrew.net.	NaN
1	www.lifanku.co	ns3.parklogic.com.,ns1.parklogic.com.,ns5.park...	lifanku.co.
2	javzoo.org	ns2.parklogic.com.,ns3.parklogic.com.,ns1.park...	NaN
3	sh.k6p.co	ns3.above.com.,ns4.above.com.	NaN
4	petite-une-deux-trois.com	ns2.parklogic.com.,ns4.parklogic.com.,ns3.park...	NaN
5	popunderjs.club	ns1.parkingcrew.net.,ns2.parkingcrew.net.	NaN
6	888.zznzyy.com	sk.s5.ans1.ns148.ztomy.com.,sk.s5.ans2.ns148.z...	NaN
7	api.tranzz.com	ns5.parklogic.com.,ns1.parklogic.com.,ns3.park...	tranzz.com.
8	chap.ga	ns2.parkingcrew.net.,ns1.parkingcrew.net.	NaN
9	cjmaniacs.com	ns1.above.com.,ns2.above.com.	NaN
10	host.gc18.info	ns4.parklogic.com.,ns5.parklogic.com.,ns1.park...	gc18.info.

找出824个停靠域名

不可获取源码域名判别

基于Whois信息对域名过滤

将上一步中找出的停靠域名的name server和cname字段的二级域名都添加到已知的停靠服务商相关的域名列表中，并去重。

再用已知的停靠服务商相关域名列表，来匹配待检测域名whois的name servers字段和domain name字段。

想找到的新的停靠域名和上一步中的停靠域名合并去重后，**新增了24个停靠域名。**

818	NaN	2g.wap.sg	ns1.parkingcrew.net,ns2.parkingcrew.net
819	NaN	www.imagechunk.com	ns1.above.com,ns2.above.com
820	nvc-elc.com	www.nvc-elc.com	ns5.parklogic.com,ns4.parklogic.com,ns1.parklo...
821	freeeduhost.com	ns2.freeeduhost.com	ns3.parklogic.com,ns1.parklogic.com,ns5.parklo...
822	NaN	www.iquicksearch.com	ns2626.ztomy.com,ns1626.ztomy.com
823	NaN	www.116www.com	ns1626.ztomy.com,ns2626.ztomy.com
17	NaN	www.javbus2.com	NS1.PARKINGCREW.NET,NS2.PARKINGCREW.NET,ns1.pa...
23	NaN	www.tp22.net	150.NS1.ABOVE.COM,150.NS2.ABOVE.COM,150.ns1.ab...
26	NaN	www.hdkan.co	ns2.bodis.com,ns1.bodis.com
28	NaN	NLB_100.str.strlab.com	NS1.SEDOPARKING.COM,NS2.SEDOPARKING.COM,ns1.se...

5. 总结

Domain Security Homework

Outline

中国科学院大学
University of Chinese Academy of Sciences

确认判别parking domains: 2472

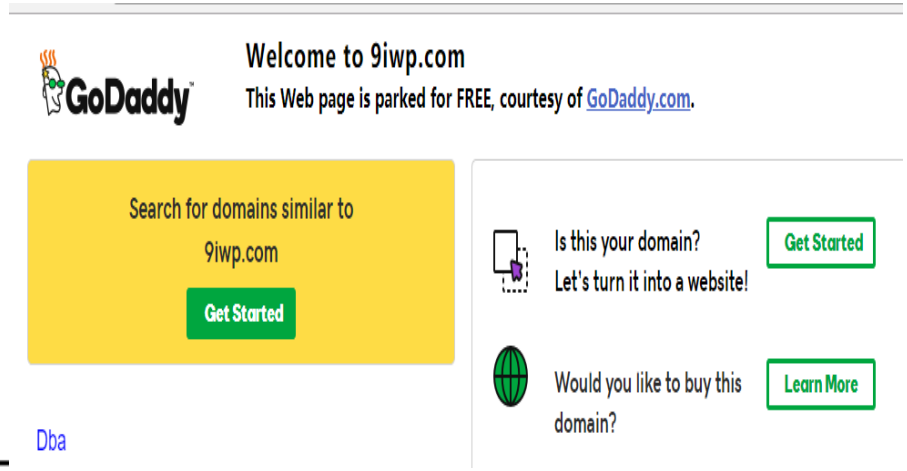
Domain Security Homework

可疑parking domains: 368271

Outline

University of Chinese Academy of Sciences

Parking Domains	Parking Domains
ww1.motioninjoy.com	www.gzrail.com.cn
ww1.xiaomw.com	www.lifanku.co
ww1.hdzone.info	javzoo.org
ww1.zakume.net	sh.k6p.co
ww1.windswow.com	petite-une-deux-
sh.k6p.cn	trois.com
ww1.llysvip.com	popunderjs.club
ww1.electracode.com	888.zznzyy.com
ww1.909900tk.com	chap.ga
gjthub.com	cjmaniacs.com
www.baidu368.com	host.gc18.info
getphoto.net	hhaazz.com
ww17.1004j.com	tcqa.com
ww1.iphonerm.com	www.xiaoniao.today
btscene.net	518ziyuan.com
16e7.com	ja16b.com
33483.cc	www.shuketxt.com
sanya30.com	bus222.com
9iwp.com	y1.xsdd.org
btscene.net	aikandy.org
baixuefeng.com	www.521gav.com



总结

优点：

1. 适当剔除一些可信度较高的域名，使得后面的模型训练和计算复杂度降低。
2. 用较少的已标注域名进行可疑域名的发现。

不足与改进的地方：

1. 默认将包含netease二级域名的所有FQDN都认为是正常的域名。
2. 在聚类结果的判别中，将正常域名或停靠域名误判的概率较高。

新发现：

1. 可以通过WHOIS信息反查去获取可疑的新的parking domains

谢谢！

Domain Security Homework

Q&A