基于非负矩阵分解的域名查询行为分析

钱群 周昌令 栾兴龙 尚群 陈萍
北京大学

提纲

- 背景及相关研究
- DNS日志的向量化处理
- 基于降维矩阵的域名群组探测
- 异常域名组发现与受控主机行为特征
- 结论与展望

背景及相关研究

- DNS日志　IP(用户)←→域名(兴趣)
- 用户之间/域名之间/用户和域名之间的关联
- 僵尸网络→DGA域名→域名聚类→受控节点

DNS日志的向量化处理

词袋模型和文档特征矩阵

- 包含两篇文档的原始语料

Bob likes to play basketball, Jim likes too.
Bob also likes to play football games.

- 词袋模型

Bob likes to play basketball Jim likes too
Bob also likes to play football games

- 全部单词

Bob like to play basketball Jim too also football games

- 文档特征矩阵

$$\begin{Bmatrix} 1 & 2 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \end{Bmatrix}$$

DNS日志的向量化处理

- 单词：IP地址
- 文档：访问过某一域名的所有IP地址的组合
- 文档特征矩阵
  - 行向量：域名
  - 列向量：IP地址
  - 元素：X_{ij}代表IP地址j对域名i发起过X_{ij}次查询

	162.105.222.205	162.105.221.140	162.105.52.102	162.105.34.112	222.29.22.66	222.29.31.87	115.27.153.97	162.105.120.16
sensearch.baidu.com	1	0	0	0	2	1	0	0
www.zhihu.com	0	1	0	2	0	0	0	0
ucus.ucweb.com	0	0	1	0	0	0	0	0
apple.com	0	0	0	0	0	0	1	1

DNS日志的向量化处理

TF_IDF（词频_反向文件频率）

词频:给定单词在文档中出现的频率/归一化

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,k}}$$

反向文件频率：提高生僻词的权重

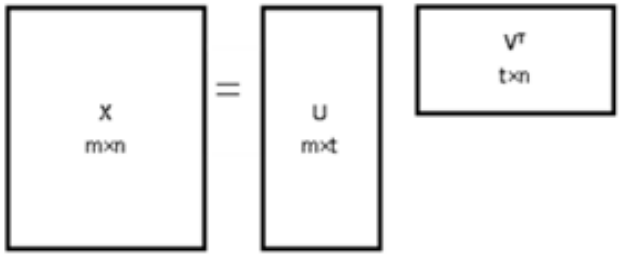
$$idf_j = \log \frac{|D|}{|\{i: j \in i\}|}$$

单词j对于文档i的重要性可衡量为

$$tfidf_{ij} = tf_{ij} \times idf_j$$

非负矩阵分解NMF

- 日志选取时间窗口：天级
- 域名归并：保留顶级域名及前面两段
- 矩阵维度大→降低维度→发现隐藏关联
- NMF：给定一个非负矩阵 $X=(x_{ij})_{m*n}$
  NMF为其找到两个非负矩阵 $U=(u_{ij})_{m*r}$ $V=(v_{ij})_{r*n}$
  满足 $X \approx UV$



域名群组与用户群组

- 降维后，特征向量的维度对应的不再是某一个特定的IP，而是各个IP地址的概率分布
- 通过筛选→相似规律的IP集合→兴趣用户组
- 兴趣用户组→访问兴趣→域名组

- 教育部官网、清华马克思主义学院、高校图书馆数字资源采购联盟、EBSCOhost文献数据、新东方在线；
- 去哪（qunar）、携程（ctrip），途牛（tuniu）、hotels.com（酒店预订）、四季酒店官网（fourseasons）；
- skype（网络电话）、facebook（脸书）、yahoo（雅虎），naver（韩国门户网站）、hotmail（微软提供的邮件服务）、soundcloud（德国的音乐分享社区）

异常域名组发现

- 异常域名（DGA域名）：可读性差，难于记忆，大量DGA域名其对应的控制端IP地址相对集中
- 异常行为：受控节点对异常域名的集中访问

- 结合Alexa全球排名前100万域名列表作为过滤，不在该列表中的域名即标记为异常域名

$$\text{异常比} = \frac{\text{某域名组中被标记为异常的域名数量}}{\text{该域名组相关域名的总数}} \times 100\%$$

- 异常比较大的域名组中有可能包含了DGA生成的非法域名组。

异常域名组发现

目标IP地址	域名			
221.8.69.25	oyiddwpj.cn	pemyy.cn	hyentconoz.cn	lcsboi.cn
	mmoogwe.cn	mwpadbcoatg.cn	mpqcapwgbd.cn	nkaykhos.cn
	icjurww.cn	xlbsoh.cn	zuxhzeuxvd.cn	tlfueywqi.cn
216.66.15.109	gfuwxkcvdf.ws	cimtpnom.biz	nyyrhcmclxf.biz	feyjizkboio.biz
	rwysfiat.ws	sfiddqe.ws	obkgps.biz	qlyszm.ws
	yzqueyjlacw.ws	vacxrfjm.ws	upmwtxpxsri.ws	awnczd.biz
38.102.150.27	oyybrjiucy.biz	ffqeixbm.biz	cfxbqnlt.biz	jvqqxaluugl.biz
	dzfygbon.biz	mwxho.biz	lcuxiwot.biz	yzmflltwi.biz
	otfkzbzowhg.biz	tihtq.biz	cubvu.biz	jsnxawbnu.biz

异常域名组发现

- 221.8.69.25的直接访问结果

Conficker Sinkhole By CNCERT/CC!

This domain is possibly used by Conficker Computer Worm. If you have any related problems, please contact CNCERT/CC.

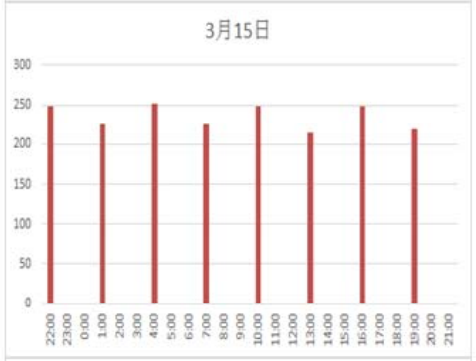
您所访问的域名可能正在被"飞客"蠕虫病毒使用。如有相关问题，请联系国家互联网应急中心（CNCERT/CC）。

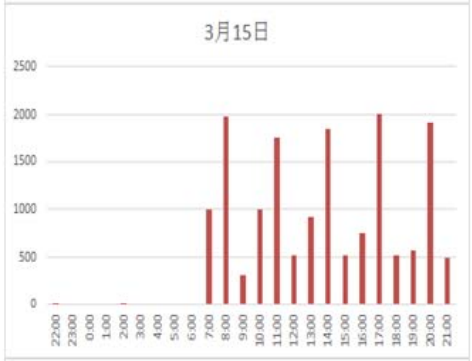
Email:cncert@cert.org.cn

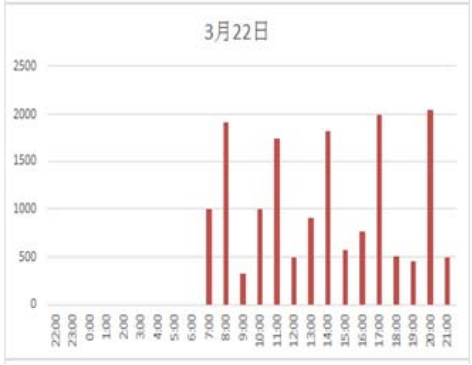
If the website you visited is a normal site , please click here

如果您在访问正常网站时遇到此问题，请点击这里

受控主机行为特征







结论与展望

- 自然语言处理领域常用的向量化、矩阵降维等方法可以用于文本格式的DNS日志的域名聚类分析
- 通过NMF降维后的特征矩阵对域名聚类分析、兴趣用户组发掘有较好的效果
- 用该方法找出了具有DGA特征的异常域名组以及对该域名组进行异常查询的受控节点
- 在今后的研究中，可以尝试使用不同的降维模型对文档特称矩阵进行处理

Q&A

- 谢谢！

北京大学 PEKING UNIVERSITY | 计算中心 COMPUTER CENTER