doi:10.11959/j.issn.1000-436x.2016064

基于深度学习的域名查询行为向量空间嵌入

周昌令[1,2]，栾兴龙[1,2]，肖建国[3]
(1. 北京大学计算中心，北京 100871；2. 北京大学信息科学技术学院，北京 100871；3. 北京大学计算机科学技术研究所，北京 100871)

摘 要：提出一种新的分析 DNS 查询行为的方法，用深度学习机制将被查询域名和请求查询的主机分别映射到向量空间，域名或主机的关联分析转化成向量的运算。通过对 2 组真实的校园网 DNS 日志数据集的处理，发现该方法很好地保持了关联特性，使用降维处理以及聚类分析，不仅可以让人直观地发现隐含的关联关系，还有助于发现网络中的异常问题如 botnet 等。

关键词：DNS；深度学习；上下文；降维；行为分析；层次聚类
中图分类号：TP393.07 **文献标识码：**A

Vector space embedding of DNS query behaviors by deep learning

ZHOU Chang-ling[1,2], LUAN Xing-long[1,2], XIAO Jian-guo[3]
(1. Computer Center, Peking University, Beijing 100871, China;
2. School of Electronics Engineering and Computer Scie ng University, Beijing 100871, China;
3. Institute of Computer Science & Technology, Peking University, Beijing 100871, China)

Abstract: A novel approach to analyze DNS query behaviors was introduced. This approach embeds queried domains or querying hosts to vector space by deep learning mechan then the relationship between querying of domains or hosts was mapped to vector space operations. By processing two real campus network DNS log datasets, it is found that this method maintains relationships very well. After doing mension reduction and clustering analysis, researchers can not only easily explore hidden relationships intuitively, but also discover abnormal network events like botnet.
Key words: DNS, deep learning, context, dimension reduction, behavior analysis, hierarchical clustering

1 引言

域名服务（DNS）是互联网最重要的基础应用之一，众多互联网中的业务都与它紧密关联，如 Web、邮件、内容分发网络（CDN）等。同时，一些恶意行为也利用或针对 DNS 的特性来达到攻击的目的。例如，僵尸网络（botnet）采用 FastFlux 手段来躲避打击，其基本思想就是不断变化域名与 IP 的对应关系。因此，主机的 DNS 查询行为与网络的运行状况紧密相关。

主机可能在多种情况下发起 DNS 查询的行为，根据发起方可以分为 2 大类别：一类是与用户活动相关的，包括用户主动发起的请求，如浏览 Web 网页等以及由用户触发的请求，网页中加载的图片、广告等；另一类是用户活动无关的，是由软件或系统自动产生的，如软件自动更新、证书检查、邮件黑名单查询以及受控僵尸节点请求指令等。

第一种类型的行为与用户的兴趣偏好相关。分析用户经常查询的网站域名的关联关系有助于理解用户需求，提升服务质量，改善用户体验。由于分析结果与实际的网络环境紧密相关，这方面的研究并不多。Moghaddam 等[1]用自组织映射（SOM）

收稿日期：2015-03-30；修回日期：2015-09-10
基金项目：国家 2012 年下一代互联网技术研发、产业化和规模商用专项基金资助项目（No.CNGI-12-03-001）；国家发展改革委 2011 年国家信息安全专项基金资助项目；国家高技术研究发展计划（“863 计划”）基金资助项目（No.2015AA011403）
Foundation Items: The Next-Generation Internet Technology Development, Industrialization and Large-scale Commercial Project, the National Development and Reform Commission 2012 (No.CNGI-12-03-001), National Information Security Special Project Funded by National Development and Reform Commission 2011, The National High Technology Research and Development Program of China(863 Program)(No.2015AA011403)

分析了无线用户访问的域名之间的关联性，他们发现一些逻辑上关联的网站域名生成的 SOM 图形状也非常相似，如“itunes”和“netflix”、“washingtonpost”和“cnet”等。他们的工作只限于无线，且只分析了人工标注的 100 个域名。这一类行为产生的 DNS 日志主要是 A 和 CNAME 记录，查询的大部分也是真实存在的域名。

第二种类型的行为反映了主机特性。相关研究主要集中在发现网络中的异常情况，如发送垃圾邮件[2]、恶意域名[3]，botnet[4,5]等。其中 botnet 受关注程度最高，因为它对网络的影响非常大，又采用了各种手段来躲避打击。其中域名生成算法（DGA, domain generation algorithm）是 botnet 应用得最多的一种手段，受控节点高频地查询不断变化的域名，主控节点在需要时把即将出现的域名与发布指令的 IP 对应关系注册上就可以控制该僵尸网络。这样由于所查询的变化目标域名绝大部分都是无效的，将产生大量返回失败的 DNS 查询记录（ServerFail 或 NxDomain）[6]。

此前对 DNS 域名查询行为的相关研究工作，主要集中在特征参数的选取[3,7]，以及关联信息的描述[5]等方面，通常再结合机器学习的手段来区分不同的行为。本文提出一种新的基于深度学习的方法来研究 DNS 域名查询行为：将被查询域名和发起查询请求的主机 IP 分别映射到 K 维的实向量空间，对域名或主机的分析转化成空间中的向量运算，通过降维还可以对域名或主机的关联特性进行直观的展示。

本文的主要贡献如下。

1) 提出了一种将 DNS 查询行为映射到向量空间的方法。通过构造被查询域名列表以及请求查询主机列表，用这 2 种列表作为深度学习的训练数据，获得域名和主机的向量表示，然后在向量空间中分析元素之间的关联性。

2) 借鉴了在自然语言处理（NLP）领域取得很好效果的深度学习优化算法[8]，实现了对海量 DNS 查询日志数据的高效处理。对一个典型的大中型校园网网络的核心 DNS 服务每天的查询日志进行单机分析，其中深度学习算法的运行时间仅需要 30~45 min。

3) 使用真实的校园网运行环境的 DNS 数据进行了验证。本文选取了 2 组来自不同校园网的数据集，在训练后得到的向量空间进行分析，发现映射后的向量很好地保留了域名或主机之间的关联特性，通过降维和交互式可视化处理后可以容易地发现隐蔽的关联关系。本文还通过计算向量之间的相似度对域名做层次聚类分析，结合域名信息熵发现与 DNS 相关的攻击行为如 botnet 等。

2 相关概念

为了便于后面描述，在此先定义一些相关的概念。

定义 1 派生邻近关系（derived proximity relationship）。按所描述的目标不同，可以分为被查询域名的派生邻近关系和请求查询主机的派生邻近关系。以被查询域名的派生邻近关系为例：在一段时间内，如果有一系列的主机都共同请求查询过 A、B 这 2 个域名，则认为 A 和 B 是邻近的。并且，发起共同请求的主机越多，则 A 与 B 的邻近程度越高。类似地，2 个主机查询的相同域名越多，它们也越邻近。

派生邻近关系通常反映了实际中存在的关联关系。以域名为例，假如多个域名所承载的业务存在逻辑上的联系，用户往往会顺序访问这些业务，如提供统一身份认证的系统与需要认证才能访问的业务系统，用户经常会先后查询它们的域名。又如用户点击网页的链接访问新的站点，以及网页中加载来自不同域名的图片时，起始站点的域名和关联站点的域名会先后被查询。这种先后查询的域名最终形成域名关联的上下文关系。本文通过保留这种先后顺序来研究派生邻近关系（或关联关系）。

定义 2 被域名查询列表（QDL, queried domains list）。在一段时间内，主机产生的 DNS 查询请求可以用序列 $QDL_k = \{d_1, d_2, \cdots, d_i, \cdots, d_n\}$ 表示，其中，每个主机 k 对应一条 QDL。同一个域名可以在列表内多次出现。

如图 1 所示，此列表可以从 DNS 查询日志信息中产生，图中每个方括号中的内容都是一个被查询域名列表。

图 1 被查询域名列表 QDL 示例

属于同一条 QDL 的域名保留了它们被查询的先后顺序，即保留了上下文的关联性（或派生邻近

关系)。由于域名与自身的邻近（关联）关系不需要考虑,因此同一域名在列表中连续出现时只保留一次,如果与其他域名交替重复出现,则全部保留。域名按被查询的先后顺序排列。如果主机在很长一段时间没有查询活动，或者列表的长度超过预设值，则产生一条 QDL 记录。不同主机的 QDL 长度一般不同，一个主机也可以有多条 QDL。在本文中仅将日志里返回 A 和 CNAME 的记录纳入 QDL。

类似地，研究主机的派生邻近关系时，要用到下面的定义。

定义 3　请求查询主机列表（QHL, querying hosts list)。在一段时间内，查询同一域名或子域的主机用序列 $QHL_k=\{h_1,h_2,\text{L},h_i,\text{L},h_n\}$ 表示。其中，相同的 h_i 可以多次出现在列表中。

实际中，不同主机查询 DNS 的域名是比较分散的，所以有时候需要关注查询相同域名后缀的主机。域名的多个字段由点分隔符“.”分开，例如, www.example.com。2 个或多个域名从右往左，它们所具有的公共字段为它们的公共域名后缀。一般地，需要关注的是最长的公共后缀。

本文中重点关注产生失败查询（返回 NxDomain 或 ServerFail 信息）的主机，这是因为它们通常与非正常的通信通道[6]以及恶意行为如 botnet[4]相关。由于这类失败查询的前缀大量变化，而后缀在一段时间是保持不变的，所以需要按最长公共后缀产生的 QHL 才能把有类似行为的主机放到同一个上下文环境中。本文在第 3.4 节具体描述用来提取这类查询的最长公共后缀的算法。

定义 4　向量空间嵌入（vector space embedding)。数学上嵌入是指一个数学结构经映射包含到另一个结构中[9]。如果存在一个保持结构的单射 $f:X\to Y$，其中目标结构 Y 为 k 维的向量空间，这个映射 f 就给出了一个向量空间嵌入。本文中向量空间嵌入特指对列表集合 L（由 QDL 组成或由 QHL 组成）中所有不同的元素 e（域名 D 或主机 H）所组成的集合 E，可以映射到 k 维的实向量空间。即对集合 E 来说，存在如下映射关系：$f_E:E\to R^k$。

将列表中的元素进行向量空间嵌入的基本思想最早由 Hinton 在 1986 年提出[10]，该文中称为分布式表示（distributed representation)。现在该方法主要用在自然语言处理（NLP）中，将单词语义研究转化成对应的 k 维实数向量的运算[8]，并取得了很好的效果[11]。

3　域名查询行为向量化方法

将 DNS 查询行为与自然语言处理进行类比，列表 QDL 或 QHL 对应文档，列表中的元素（域名或主机）对应单词，列表的集合 L 对应大量文档组成的语料。通过从 DNS 查询日志中构建 QDL 和 QHL，保留了域名查询行为的上下文关联关系，从而得到用于进行深度学习的训练数据。

3.1　深度学习

求解向量空间嵌入模型早期的方案是采用一个多层的神经网络进行训练。2013 年 Mikolov 在文献[8]中指出，可以通过一系列的优化措施，有效降低计算的复杂度。例如使用 3 层（输入—隐藏—输出）的神经网络，只对滑动窗口内的词计算联合概率，采用优化的 Huffman 编码让词频近似相等的单词其隐藏层激活的值基本一致，从而减少隐藏层数目等，此外还采用了一些其他的优化计算方法。实际上，一个优化的单机版本 word2vec[12]一天可训练上千亿单词。

本文借鉴文献[8]的方法，并采用了文献[13]中的 Skip-gram 模型。如图 2 所示，对于被查询域名列表 QDL（请求查询主机列表 QHL 可以类似地处理）中的元素 e_i，以及其上下文窗口中的各元素 e_{i+c}，它们所对应的向量空间表示分别为 $\vec{v}_i'$ 和 $\vec{v}_{i+c}'$。

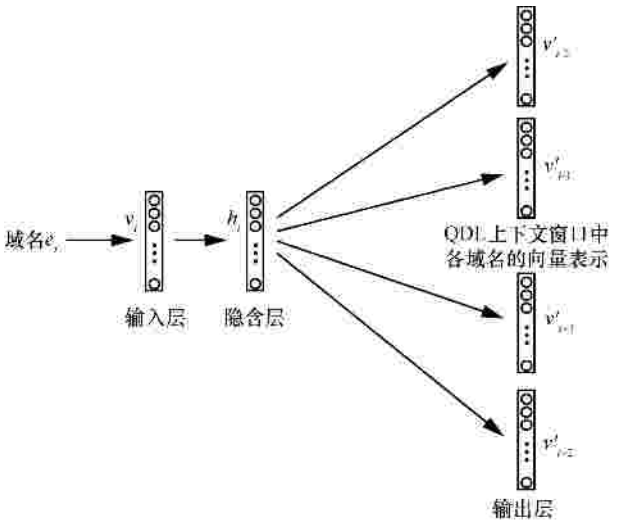

图 2　基于 Skip-gram 的域名向量空间表示

此模型由域名 e_i 来计算上下文窗口中各域名 e_{i+c} 出现的条件概率 $p(e_{i+c}\mid e_i;?)$，其中 $?$ 是待确定

的参数。对于给定的 QHL 列表集合，计算目标是所有的条件概率之积取最大值。为了便于计算，本文使用对数概率，这样可以将乘积转化为求和，从而有式(1)所定义的训练目标函数

$$\arg\max_{q}\sum_{i=1}^{E}\sum_{-C\le c\le C,c\ne 0}\log p(e_{i+c}\mid e_i;?) \tag{1}$$

其中，C 是上下文窗口的大小，训练数据是所有 QDL 组成的列表集合 L，其中不同的域名 e 的总数为 E。概率 p 的定义如下

$$p\left(e_j\left|e_i;?\right.\right)=\frac{\exp\left(\vec{v}_j\vec{v}_i\right)}{\sum_{l=1}^{L}\exp\left(\vec{v}_l\vec{v}_i\right)} \tag{2}$$

其中，$\vec{v}_i$ 和 $\vec{v}_i$ 分别是输入和输出层对元素 e_j 的向量表示，$?$ 参数是指所有这些待确定的向量。通常情况下，$\vec{v}_i$ 和 $\vec{v}_i$ 取值并不一样。同时参考文献[8]，本文中隐含层向量与输入层向量取相同值。

实际中，不同域名的总数 E 可能非常大（10^5~10^8），直接计算概率 p 不现实。本文采用 Negative Sampling 来提高计算速度，并用随机梯度下降（SGD）方法[13]来更新概率 p。这样，通过构建 QDL 或 QHL 产生上下文环境，类似于自然语言处理（NLP）中的词嵌入（word embedding），就可以将域名或主机 IP 转化成 k 维空间中的向量。

3.2 降维

降维（dimension reduction）是指采用映射关系 $f:R^k\to R^d, d<k$，将高维度空间中的点映射到低维度空间中。特别地，高维向量空间 R^k，当 $k>3$ 时人们无法直观理解其中的数据，故通常选择将其降维到 d =2 或 d = 3。

本文采用 t-SNE[14]来对得到的 k 维向量空间做降维处理，以方便可视化理解。一般地，对高维空间中的元素 x_1,x_2,L,x_n，t-SNE 按下式计算它们的联合概率 p_{ij}

$$p(j\mid i)=\frac{\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2s_i^2}\right)}{\sum_{k\ne i}\exp\left(-\frac{\left\|x_i-x_k\right\|^2}{2s_i^2}\right)}$$

$$p_{ij}=\frac{p(j\mid i)+p(i\mid j)}{2n} \tag{3}$$

其中，s 为高斯核，取值与 x_i 附近的点密度有关。

对目标空间 d 维的映射 y_1,y_2,L,y_n，定义两点间的联合概率为

$$q_{ij}=\frac{\left(1+\left\|y_i-y_j\right\|^2\right)^{-1}}{\sum_{k\ne l}\left(1+\left\|y_k-y_l\right\|^2\right)^{-1}} \tag{4}$$

最后，通过让高维和低维空间中的 KL 距离（Kullback–Leibler divergence）取极小值，得到向量在低维空间中的映射。此映射保持了节点映射前后的相似度，因此非常适合对向量空间嵌入后的可视化处理。

$$KL\left(P\parallel Q\right)=\sum_{i\ne j}p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{5}$$

3.3 层次聚类和相似度量

层次聚类(hierarchical clustering)[15]是一种可以根据给定的相似度阈值对节点聚类的方法，它需要计算节点之间的相似程度。本文中用 k 维向量的运算来度量节点的派生邻近程度或相似程度。令 e_i 是列表 QDL 或 QHL 中的元素，它可用 k 维向量 $\vec{v}_i\in R^k$ 来表示，则可以定义元素 e_i 和元素 e_j 之间的相似程度用单位长度向量的内积来计算

$$sim(e_i,e_j)=dot\left(\frac{\vec{v}_i}{\left\|\vec{v}_i\right\|},\frac{\vec{v}_j}{\left\|\vec{v}_j\right\|}\right) \tag{6}$$

本文只关心那些节点之间相似程度很高的簇，并且要求簇内的节点至少 2 个以上。因此选择 complete-linkage clustering 方法[16]，达到阈值后就停止迭代，这样可以大大提高计算效率。

3.4 域名最长公共后缀发现算法

本文在构建 QHL 时，重点关注返回的是查询失败（Server 或 NxDomain）的记录。由于查询的大多是一些并不存在的域名，它们几乎不重复，很多情况由大量变化的前缀和少数不变的后缀组合而成，因此按域名后缀构建 QHL 是合理的选择。由于 DNS 查询日志的数据量往往非常巨大，加上需要对从长到短的后缀分别组合，直接记录每个 IP 所有访问过的域名后缀效率不高且空间占用较多。本文中采用了 Counting Bloom Filter[17]来减少对存储空间的需求。具体算法描述如下，其中 *n*LD()函数将返回从长到短的各级域后缀的集合。

算法 1 域名最长公共后缀发现

输入： cbf counting Bloom filter

ipaddr: 发起查询的主机

domain: 域名
k 公共后缀最少出现次数
qhl: 查询主机列表
输出：更新后的查询主机列表 *qhl*

1) $key \leftarrow (ipaddr + domain)$
2) if $key \notin cbf$ then
3) for $j \leftarrow nLD(domain)$do
4) $cbf \leftarrow cbf \vee \{ipaddr + j\}$
5) if $(|cbf(ipaddr + j)| > k)$ then
6) $qhl \leftarrow qhl \vee \{ipaddr\}$
7) break
8) end
9) end
10) $cbf \leftarrow cbf \vee \{key\}$
11) end
12) return *qhl*

4　实验方法和数据分析

4.1　数据来源

本文使用了 2 组来自不同校园网环境的数据集。数据集 PKU_DNS 是在北京大学校园网的运行环境中对 5 台核心 DNS 服务器的流量进行采集得到的。此采集系统采用 Passive DNS 方案[18]，通过交换机端口镜像把校园网的几台核心 DNS 服务器的流量全部送到采集系统，从而记录下校园网用户详细的 DNS 查询日志。其数据规模如表 1 所示。

表 1　北京大学数据集 PKU DNS 的规模

DNS 日志产生速率	>4 000 条/秒
DNS 日志峰值速率	>20 000 条/秒
正常查询日志总量	约 350M 条/天
失败查询(NxDomain 等)	约 2M 条/天
DNS 原始查询流量	200~400 GB/天
占用空间（压缩率约 20%）	2~5 GB/天
被查询的不同域名数量	约 3×10^6/天
请求查询的不同主机数量	约 10^5/天
采集时间	20150301~20150319

数据集 BIT_DNS 是北京理工大学校园网中一台核心 DNS 服务器的 syslog 日志，其数据规模如表 2 所示。由于该服务器没有限制查询的来源 IP，而这几天恰好有来自校外的 DNS 放大攻击[19]，使该数据集 BIT_DNS 每天请求查询的不同主机数量偏大。

表 2　北京理工大学数据集 BIT DNS 的规模

DNS 日志产生速率	>1 000 条/秒
DNS 日志峰值速率	>7 000 条/秒
正常查询日志总量	约 100M 条/天
被查询的不同域名数量	约 7×10^5 /天
请求查询的不同主机数量	约 2×10^5 /天
采集时间	20150315~20150319

4.2　数据分析

数据分析使用的操作系统 ubuntu 14.04.2 LTS，16 GB 内存，4CPU。使用 python 代码，完成 3 部分工作：1)从 2 个不同格式的原始数据集分别生成相应的 QDL 和 QHL 列表；2)对各自得到的列表采用深度学习算法训练出向量空间嵌入表达，然后再调用 *t*-SNE[14]对得到 *k* 维向量空间结果进行降维处理后，生成 d3.js[20]需要的数据格式，方便通过浏览器交互地展示；3)对向量空间中的节点进行层次聚类[15]，输出高相似度的节点簇。在规模较大的 PKU_DNS 数据集上，单机上对每天的 DNS 日志文件进行列表提取过程大约需要 2~3 h，向量空间嵌入的过程大约需要 30~45 min，层次聚类过程大约需要 15~30 min。

本文所选用的参数情况是：被查询域名列表 QDL 的超时时间为 1 h，每个 QDL/QHL 的长度限制为不超过 1 000 条记录，当超时或长度超过限制时，输出对应的列表作为训练数据。由于 DNS 查询行为具有显著的按天重复的周期性[21]，因此训练数据以每天 24 h 为分隔。在计算最长公共后缀时，同一 IP 查询某个后缀失败次数不小于 10 会被记录。深度学习进行向量空间嵌入的参数[12]为：向量维度选取 $k = 100$，元素最少出现次数 $min\_count \geq 5$，上下文窗口 $window = 8$，随机梯度下降学习率 $a = 0.025$。

4.2.1　域名派生邻近关系分析

通过将域名向量空间嵌入后，可以预期关系越近的域名，在向量空间中的距离也越近。图 3 中展示了几组在向量空间中邻近的域名。第一组是与北京大学主页 www.pku.edu.cn 相邻的域名，可以看到，几乎全部是校园网的内部域名，其中 portal 是北大校内门户，pkunews 是新闻网，Web5 承载着主页的一些业务，www.bjmu.edu.cn 是医学部的主页，后面几个也是校园网用户经常访问的一些站点，这些域名之间有直接的逻辑关系。第 2 组是与

美国化学学会 pubs.acs.org 相邻的域名，可以发现除了化学学会的子域网站外，其他几个域名几乎都是与学术研究有关的网站，唯一的例外是其中第 5 条后缀为 rackcdn.com 的域名，分析发现此域名是由 pubs.acs.org 首页内加载了一个 Live Chat 的脚本所引起的。第 3 组域名展示的是美国物理学会刊物统计 counter.aps.org 相近的域名，可以看到，几乎全部是与学术及出版相关的，而且是和物理学科有关联的。最后一组域名则全部都是和人人网 www.renren.com 相关的域名。从这几组相近的域名关系可以看出，如果域名所承载的业务紧密关联，或者涉及的内容对用户具有相似性，在进行向量空间嵌入变换后，它们在向量空间也相邻。类似的方式分析在 BIT_DNS 数据集中的典型的邻近域名，也发现有类似的规律，如与北理工主页 www.bit.edu.cn 相似度较高的域名几乎都是北理工内部的一些网站。

与www.pku.edu.cn相近的域名	
域名	相似度
portal.pku.edu.cn	0.9113
pkunews.pku.edu.cn	0.8801
web5.pku.edu.cn	0.8697
www.bimu.edu.cn	0.8687
ipgw.pku.edu.cn	0.8674
bbs.pku.edu.cn	0.8431
tree.pku.edu.cn	0.8364
lib.pku.edu.cn	0.8236
green.pku.edu.cn	0.8065
old-lib.pku.edu.cn	0.7877

与pubs.acs.org相近的域名	
域名	相似度
www.opticsinfobase.org	0.8542
assets.acs.org	0.8521
sso.acs.org	0.8468
sciliamm.acs.org	0.8325
1131784da7184e34e737-f2bb466c05e9d6b8c5e18a9c4f84e1a0.ssl.cf1.rackcdn.com	0.8299
www.reaxys.com	0.8156
scholar.google.com	0.8115
www.ingentaconnect.com	0.7994
img.cn25.com	0.7987
scholar.googleusercontent.com	0.7979

与counter.aps.org相近的域名	
域名	相似度
www.physicscentral.com	0.8057
www.bssaonline.org	0.7790
journals.cambridge.org	0.7462
scholar-google.com	0.7438
www.cambridge.org	0.7427
jap.pearx-press.org	0.7412
s.addthis.com	0.7357
www.scopus.com	0.7322
mts-nn.nature.com	0.7295
www.cshlpress.com	0.7292

与www.renren.com相近的域名	
域名	相似度
www.physicscentral.com	0.8170
www.bssaonline.org	0.7597
journals.cambridge.org	0.7388
scholar-google.com	0.7256
www.cambridge.org	0.6987
jap.pearx-press.org	0.6839
s.addthis.com	0.6789
www.scopus.com	0.6736
mts-nn.nature.com	0.6734
www.cshlpress.com	0.6715

图 3　向量空间相邻的域名

图 4 是 PKU_DNS 数据集 2015 年 3 月 10 日的被访问域名嵌入后的向量空间进行了降维（使用 t-SNE），然后在二维空间中进行展示。为了更好地看到效果，做了如下处理。一是将部分域名按照它们的后缀进行了简单的标记，如 *.apple.com 的域名都标识为 apple，为了方便查看，本文在后期处理时在图中增加了标注信息。二是节点支持交互式查询，可以显示每个点对应的域名。可以发现，图中有许多明显的节点簇。它们大多数是由具有相同域名后缀的网站组成，或由属于同一个公司的不同域名后缀的域名组成（如 renren.com 和 xiaonei.com）。同一后缀的域名基本在同一个区域，但*.qq.com 是个例外，在图 4 中左下角的位置，形成了 2 个有一定距离的簇，这可能和它有差别很大的业务类型有关（即时通信和内容展示等）。另外一些簇是由截然不同的域名后缀组成的，例如标记为 scholar 学术的簇（位于左上角偏下的位置）。一些视频类网站如 56、土豆、爱奇艺等各自成簇，相互之间又靠得很近。除此之外，图 4 的中央偏左的位置由大量后缀各异的节点组成，且相互之间并不紧密。它们主要是由一些低频访问的站点组成。由此发现，不同数据集所得到的向量空间嵌入的结果并不一致，节点的绝对位置往往不一样，但节点之间的相对位置具有前述的规律，即属于同一公司的站点会被各自聚集在一起，如 apple（苹果）、taobao（淘宝）、sohu（搜狐）等。并且，不同数据集的簇的大小形状也略有差异。一方面，这和 t-SNE 算法有关，它的降维结果可能出现绝对位置的变化。另一方面也反映出这 2 个学校的用户访问网站的兴趣偏好有所差异。

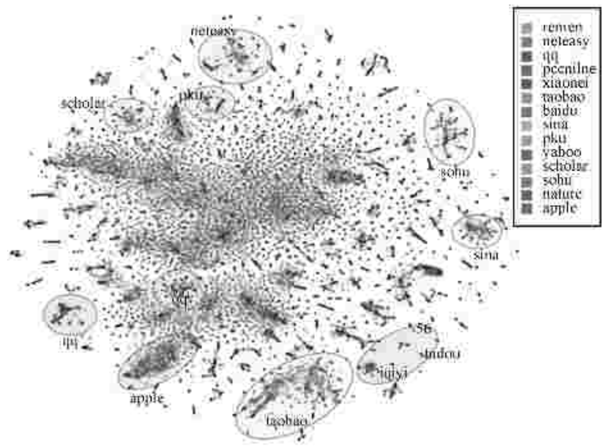

图 4　域名向量化后降维展示(t-SNE)

4.2.2　主机派生邻近关系分析

前面提到，对失败的 DNS 访问，采用的是 QHL 列表，即最终是对访问域名的主机 IP 进行向量化。如图 5 所示为 2015 年 3 月 11 日北京大学校园网内发起失败查询的 IP 经过向量化，并采用 t-SNE 降维后的数据（为了方便查看，本文在后期处理时在图中增加了标注信息）。其中明显独立成簇的节点主要有这几种情况：一是配置了错误的域名后缀，图

中标记为 pku 的簇是不少用户错误设置了 pku.edu.cn 后缀引起的，标记为 gsm 和 ccer 的 2 个簇是 2 个学院自己维护的机房的 IP，这些主机对每个待查域名都会添加错误的后缀再进行查询，结果就产生了大量的 NX 记录；另一种情况是自定义的软件通信通道，如邮件服务器访问 DNSBL 查询黑名单，防病毒软件（如 mcaffe 查询云病毒库）等；还有一些是用户端使用 BT 软件处理过期的 torrent 文件，大量重复查询一些不再提供服务 tracker 服务器引起的；最后就是由于 DNS 相关的攻击（如 botnet 等）形成的。

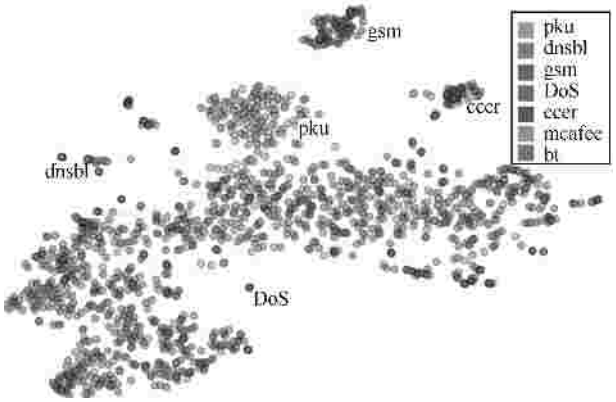

图 5　IP 向量化后降维展示(*t*-SNE)

这些标签数据是通过图中的聚簇信息，找到对应的 IP，然后根据前面归并时的最长后缀来确定。在处理过程中发现在图中有一个明显集中的簇由 6 个 IP 组成，它们的坐标非常接近，几乎重叠在一起。为了方便展示，本文在图中把它们的标识点放大显示，在图的中下位置。在这一天中，它们发起查询的部分域名后缀如图 6 所示。注意这只是后缀，实际查询时其前缀是高频地不断变化的，具有典型的 DGA 域名的特征。

a.fjhsxs.com	www.hsexpress.cn
b.fjhsxs.com	www.hxj8453.com
ggman.weiaojia.com	www.lundaddc.cn
mk.mhzjs.cn	www.weijie130.com
vip.mcgift.com.cn	www.weijie131.com
www.543ba.com	www.weijie132.com
www.543bk.com	www.weijie133.com
www.999ae.com	www.weijie666.com
www.999be.com	www.zszhanyi.cn

图 6　查询失败的域名后缀

进一步的分析发现，这几个 IP 基本都提供了匿名 DNS 服务，允许任意主机使用它们来进行递归查询。通过对图 6 这些域名的 Whois 查询以及利用当天的出口 NetFlow[22]数据对这几个 IP 的进出流量进行分析，确认这是一起针对目标域名的 DNS 放大攻击(DoS 攻击)[19]。

4.2.3　聚类分析

得到域名或主机的向量空间嵌入表示后，通过计算向量之间的内积可以得到任意 2 个节点之间的相似程度，继而可以使用层次聚类方法对节点进行成簇分析。本文只关心那些相似程度很高的节点簇，选择的聚类阈值是 0.9，并且要求簇内节点至少 2 个以上。由于 QHL 的解释和分析往往需要结合网络环境中的用户属性来进行，因此本文只对 QDL 得到的域名向量进行分析。前面提到业务关联的域名在向量空间中比较邻近，这些域名往往具有相同或类似的域名后缀，计算出这些域名的信息熵[23]就会比较低；而像 botnet 等利用 DGA 生成的一组域名其信息熵就会比较高。

通过层次聚类得到节点簇后，按它们的信息熵从高到低排列，发现在 BIT_DNS 数据集中，排在前面的 2 个域名簇，每簇都有上百个域名聚集在一起。它们的部分域名如图 7 和图 8 所示。可以看到，这 2 组域名在规律上类似，但又有所差别，并且不同组域名之间的相似度较低。进一步的分析发现，这些域名分别属于 conficker botnet 的 2 个变种。由于感染同种类别 botnet 的节点会以类似的规律定期查询相同的域名，使这些被查询域名之间产生派生邻近关系，从而在向量空间表示的节点之间非常相似。

riijlnimo.biz	bcdhaflh.net	iqqzmokgde.net
tuhfyfa.cc	qqwjbhqa.net	lspnzfc.net
mhgdmuic.info	xolxlnxho.net	lxiqeltz.net
hoynorbaf.org	ceisk.org	tyqqpui.net
mgkrxfu.biz	uqzfmfakkcw.biz	nvkzym.biz
mwkpwowj.cc	oqrksmcf.com	zatvxxiczl.cc
utnefbzyfy.com	helafi.info	izwcjusasmi.info

图 7　botnet1 查询的部分域名

有了被查询的 botnet 域名，可以利用 QDL 或原始的 DNS 数据源查到对应的感染 botnet 的主机，从而减少或消除对校园网安全的潜在威胁。由于聚类算法是无监督的方式，不需要预先知道 DGA 域

名生成算法的规律，因此这个方法可以用于发现未知类型 DGA 域名相关的攻击(如 botnet 或域名 DoS 攻击)。

ivoljjpg.cc	lelexqcs.info	hbdjdbu.com
aittfmp.cn	meyohwrpv.info	tshkd.com
hrviyokg.com	bhvof.org	jjjgcpmxvbg.info
xdnwa.org	kfzsm.org	pzqeytx.info
zhbwyda.ws	bltbxoirrvd.biz	gklvmeulwc.info
ildmidpdys.biz	xvamcooo.cc	olrhdrmf.biz
eyeiouzf.cn	uhyupqcdhg.cn	qtnukhqp.cc

图 8 botnet2 查询的部分域名

5 相关研究

在 DNS 行为分析方面，Dominik 等[24]评估了 1NN (1 最近邻)、多项式朴素贝叶斯分类器 (MNB, multinomial naive bayes classifier) 和模式挖掘 (PM, pattern mining) 这 3 种算法在利用 DNS 日志对用户行为模式进行挖掘方面的效果。袁春阳等[25]发现基于行为与域名查询关联在识别恶意域名时具有更好的效果，且可以监测到未知的病毒。Gao 等[21]提出了一种利用已知的恶意域名作为种子从 DNS 日志中发现未知的恶意域名的方法，其核心思想是利用域名解析请求在时间上的共现规律，将与恶意域名经常相伴出现的域名标记为可疑域名，再通过 TF-IDF 评分等方法将其可疑性量化。Choi 等[5]采用关联矩阵的方法来发现共同查询现象，他们对每个域名形成一个二值矩阵，列向量是时间窗口，行向量是主机查询情况，如果在时间窗口 y 主机 ip_x 查询了该域名，则矩阵对应位置(x, y)设置为 1，否则设置为 0。之后对不同的域名进行相似度聚类。Choi 的方法面临构造的矩阵维度非常大，运算复杂度很高的问题。

对 DNS 查询失败的记录分析方面，Krishnan 等[26]提出了利用阈值随机游走的方法来分析日志中的 NxDomain 记录，也考虑了共同查询现象。他们的方法能够用较少的失败查询数据就发现一些恶意域名，但不能区分不同类别的失败查询，故他们对如 spamhaus 之类的 DNSBL 需要放入白名单，也没有考虑主机错误配置 DNS 后缀对结果的影响。

Mikolov 提出的 word2vec[8]，首次发现了语义的相近关系可以直接用词向量的运算直接获得，极大地推动了词向量方法在自然语言领域的应用，在语义分析、词性分析、情感分析以及文档翻译[27]等方向取得了很好的进展。Levy 等[28]证明了关于采用类似于词嵌入 (word embedding) 的方式进行向量空间嵌入的方式可以保持元素之间的相关性，采用这种词嵌入方式表示与互信息 (PMI, pointwise mutual information) [29]表示在一定的前提下是等价的，而 PMI 就是用来处理元素之间关联度的。Perozzi 等[30]提出的 DeepWalk 方法将词向量思想推广到图的处理，基于深度学习方法，把对图中节点的随机游走 (random walk) 当成一个文档进行训练，在社会化网络多标签分类任务中取得了很好的效果。Tang 等[31]提出的 LINE 把 DeepWalk 的工作推广到一般的图处理。后两者的方法在应用到 DNS 查询行为时会把域名和主机同时向量空间嵌入，一是计算量大幅增加影响处理效率，二是影响了最终的相邻效果。

6 结束语

本文提出了一种将网络中的 DNS 查询行为在向量空间嵌入的方法。通过构造被查询域名列表和请求查询主机列表，将域名或主机的隐含关联关系用上下文共现机制来表示，然后利用深度学习的方法，将列表中的元素表示成 K 维实数向量，用随机梯度下降方法进行训练，最终得到元素在向量空间中的表示，从而将元素的关联分析转化成向量的运算。使用这种方法得到的向量很好地保持了域名或主机的关联信息。

本文使用真实校园网络的运行数据来验证向量空间嵌入方法。分别对北京大学和北京理工大学校园网核心 DNS 服务的查询日志数据集进行处理，作为深度学习的训练数据，并得到各自的域名或主机的向量表示。发现这 2 个数据集上，向量空间中相邻的节点往往具有显著的关联关系。为了直观地发现节点之间的关系，本文采用 t-SNE 方法进行降维，并采用了交互式的可视化来展示节点的信息，从而发现了在向量空间中邻近节点的一些规律。在域名向量空间表示中，属于同一公司如 apple、taobao 等的不同站点会因为其提供的业务具有关联性而各自独立成簇，不同组织或公司的域名，如果其站点提供的内容对用户具有相似性，向量空间嵌入后也可能邻近，如一些学术类的网站，这是用其他方法难于发现的。在主机向量空

间表示中，本文通过域名最长公共后缀的算法产生查询主机列表，得到的不同类别的查询失败记录所关联的节点各自独立成簇，可以很好地区分各种产生失败查询的情况，并可帮助发现与 DGA 相关的域名攻击情形。为了更好地发现成簇节点的特性，本文采用层次聚类方法对向量空间的节点进行分析。通过设置阈值只输出那些节点之间相似程度很高、节点数量较多的簇，再结合域名的信息熵，本文在北京理工大学数据集上发现 2 组属于不同 botnet 变种的查询域名集合。由于聚类方法属于无监督学习，因此可以用于发现未知类型的域名查询行为相关的攻击如域名放大攻击和 botnet 等。

在本文所使用的 2 个数据集上，数据规模不同，取得了类似的一些结果，因此本文的方法具有较好的适应性。但需要指出的是，本文所采用的深度学习机制依赖于训练数据的量。数据量越大，结果越趋于稳定，所展示的规律也更具有代表性。另外，本文的方法与自然语言处理所面临的情况也有所不同。在自然语言处理中，其训练语料是基本是不变的，而且单词的语义也是基本稳定的。而在域名查询行为中，被查询域名列表或请求查询主机列表与所处的网络环境相关，且是不断增长的，随着时间的增加域名查询行为也可能发现变化。因此本文的方法更适合在规模较大的网络环境中使用，且训练数据的列表的时间跨度不宜过长。

本文提出了一种新的思路来处理域名查询行为，在实际运行的数据中也取得了较好的效果。此方法有望在类似领域中得到应用，如网络流量分析、日志分析等。同时也有一些问题需要进一步解决：1) 如何自动对向量空间中的元素进行分类并产生标签；2) 如何更好地处理时间序列的数据，做到实时在线运算，及时发现问题；3) 此方法目前对训练数据量要求非常大，需要足够的数据结果才会稳定，但也意味着对当前数据的变化不敏感，如何发现当前新出现的异常，值得进一步探索。

参考文献：

[1] MOGHADDAM S, HELMY A. Spatio-temporal modeling of wireless users Internet access patterns using self-organizing maps[C]//2011 Proceedings IEEE INFOCOM. c2011: 496-500.

[2] CAGLAYAN A, TOOTHAKER M, DRAPAEAU D, et al. Behavioral analysis of fast flux service networks[C]//2010 43rd Hawaii International Conference on System Sciences. c2009: 1-9.

[3] BILGE L, KIRDA E, KRUEGEL C, et al. EXPOSURE: finding malicious domains using passive DNS analysis[C]//NDSS. c2011: 1-17.

[4] ANTONAKAKIS M, PERDISCI R. From throw-away traffic to bots: detecting the rise of DGA-based malware[C]// The 21st USENIX Security Symposium. c2012: 24.

[5] CHOI H, LEE H, LEE H, et al. Botnet detection by monitoring group activities in DNS traffic[C]//7th IEEE International Conference on Computer and Information Technology (CIT 2007). c2007: 715-720.

[6] CHEN Y, ANTONAKAKIS M. DNS noise: measuring the pervasiveness of disposable domains in modern DNS traffic[C]//Dependable Systems and Networks (DSN), 44th Annual IEEE/IFIP International Conference on. c2014: 598-609.

[7] CALLAHAN T, ALLMAN M, RABINOVICH M. On modern DNS behavior and properties[J]. ACM SIGCOMM Computer Commu ication Review, 2013,43 (3): 7.

[8] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv Preprint arXiv.1301.3781.20B.

[9] WIKIPEDIA. Embedding[EB/OL]. https://en.wikipedia.org/wiki/1301.3781.2013. Embedding, 2015.

[10] HINTON G E. Learning distributed representations of concepts[C]// The Eighth Annual Conference of the Cognitive Science Society. c1986: 1-12.

[11] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521 (7553): 436-444.

[12] REHUREK R. Word2vec in python, part two: optimizing. [EB/OL]. http://radimrehurek.com/2013/09/word2vec-in-python-part-two-ptimizing/,2015.

[13] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems. c2013:3111-3119.

[14] MAATEN L V D, HINTON G. Visualizing data using *t*-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.

[15] JAIN A, MURTY M, FLYNN P. Data clustering: a review[J]. ACM Computing Surveys (CSUR), 1999,31(3): 264-323.

[16] WIKIPEDIA. Complete-linkage clustering - wikipedia, the free encyclopedia[EB/OL]. https://en.wikipedia.org/w/index.php?title=Complete- linkage_clustering&oldid=625941679,2015.

[17] BRODER A, MITZENMACHER M. Network applications of bloom filters: a survey[J]. Internet Mathematics, 2004, 1 (4): 485-509.

[18] FJELLSKAL E B. Passive DNS tool[EB/OL]. https:// github.com/gamelinux/passivedns,2015.

[19] 马云龙, 姜彩萍, 张千里, 等. 基于IPFIX 的DNS异常行为检测方法[J]. 通信学报, 2014, 35(z1): 5-9.
MA Y L, JIANG C P, ZHANG Q L et al. DNS abnormal behavior detection based on IPFIX[J]. Journal on Communications. 2014, 35(z1): 5-9.

[20] BOSTOCK M. Data driven documents[EB/OL]. http: //d3js.org/.

[21] GAO H, YEGNESWARAN V, CHEN Y, et al. An empirical reexamination of global DNS behavior[J]. ACM SIGCOMM Computer Communication Review, 2013, 43 (4): 267-278.

[22] CISCO. Cisco IOS NetFlow[EB/OL]. http:// www.cisco.com/go/netflow.

[23] WIKIPEDIA. Entropy (information theory)-wikipedia, the free encyclopedia[EB/OL]. https://en.wikipedia.org/w/index.php?title= Entropy (information?theory)&oldid= 674556523.2015.

[24] HERRMANN D, BANSE C, FEDERRATH H. Behaviorbased tracking: exploiting characteristic patterns in DNS traffic[J]. Computers & Security, 2013, 39:17-33.

[25] **袁春阳, 李青山, 王永建. 基于行为与域名查询关联的僵尸网络聚类联动监测**[J]. **计算机应用研究**, 2012, 29(3):1084-1087.
YUAN C Y, LI Q S, WANG Y J. Linkage monitoring of clus for botnet based on relevance of behavior and domain inqui y[J]. Application Research of Computers, 2012, 29(3):1084-1087.

[26] KRISHNAN S, TAYLOR T, MONROSE F, et al. Crossing the threshold: detecting network malfeasance via sequential hypothesis testing[C]//2013 43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). c2013: 1-12.

[27] ZOU W Y, SOCHER R, CER D, et al. Bilingual word embeddings for phrase-based machine translation[C]//2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013).c2013: 1393-1398.

[28] LEVY O, GOLDBERG Y. Linguistic regularities in sparse and explicit word representations[C]//Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL 2014), c2014.

[29] WIKIPEDIA. Pointwise mutual information — Wikipedia, the free encyclopedia[EB/OL]. http://en.wikipedia.org/w/index.php? title= Pointwise?mutual?information&oldid= 650473510.

[30] PEROZZI B, SKIENA S. DeepWalk: online learning of social Representations[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. c2014:701-710.

[31] TANG J, QU M, WANG M, et al. LINE: Largescale Information Network Embedding[J]. arXiv preprint arXiv:1503.03578, 2015.

作者简介：

周昌令（1977-），男，重庆人，北京大学博士生，主要研究方向为网络与信息安全、无线网络、网络流量分析及网络管理等。

栾兴龙（1989-），男，山东烟台人，北京大学硕士生，主要研究方向为网络流量分析、自然语言主题模型等。

肖建国（1957-），男，辽宁鞍山人，北京大学教授，主要研究方向为图像处理、文本挖掘和网络信息处理。