基于上下文语义的恶意域名语料提取模型研究

黄 诚[1,2],刘嘉勇[1],刘 亮[1],何 祥[1],汤殿华[2]
HUANG Cheng[1,2], LIU Jiayong[1], LIU Liang[1], HE Xiang[1], TANG Dianhua[2]

1.四川大学 电子信息学院,成都 610065
2.保密通信重点实验室,成都 610041
1.College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China
2.Science and Technology on Communication Security Laboratory, Chengdu 610041, China

HUANG Cheng, LIU Jiayong, LIU Liang, et al. Research on extraction model of malicious domain corpus based on context semantics. Computer Engineering and Applications, 2018, 54(9):101-108.

Abstract: To solve the problem of omitting and false positive in extracting malicious domains based on whitelist filtering technology in massive text, a contextual semantic-based model for extracting malicious domain corpus is presented. The proposed approach is based on the context words and phrases which describes malicious domains in a technical way, and natural language processing technology is used to automatically generate corpus from sentences which contain malicious domains. Malicious domain corpus is generated from many advanced persistent threat reports and articles with the proposed model. The malicious corpus extracted from documents is verified by random forest classifier.

Key words: malware detection; text mining; information extraction; malicious corpus

摘 要:针对目前基于白名单过滤技术在海量文本中恶意域名提取的漏报、误报等问题,提出了一种基于上下文语义的恶意域名语料提取模型。该模型分别从恶意域名所在语句的上下文单词、短语进行语义分析,并利用自然语言处理技术自动生成描述恶意域名的语料。通过该模型对公开的APT(Advanced Persistent Threat)分析文档数据提取了大量恶意域名语料数据。利用安全博客文章数据并结合基于随机森林算法的机器分类模型对论文提取的恶意语料的有效性进行了验证。

关键词:恶意域名;文本挖掘;提取模型;恶意语料
文献标志码:A **中图分类号:**TP393.08 doi:10.3778/j.issn.1002-8331.1612-0283

1 引言

近年来随着企业各种核心业务逐渐融合互联网,越来越多的组织或者公司都遭受到了各种黑客攻击,各种APT(Advanced Persistent Threat)攻击事件层出不穷。为了适应快速变化的网络犯罪技术,安全公司或相关机构也不断发现并溯源重大安全攻击事件,通过不同的渠道(博客、论坛、微博、专业报告等)来披露各种攻击技术细节及恶意域名等信息。这些已公开的攻击分析报告一般采用英文进行书写,其内容主要从攻击事件的目标、攻击者使用的恶意域名、IP地址、恶意工具等进行描述分析。内容中的恶意域名或者IP地址也有可能被黑客用于其他攻击中,为了检测并阻断这些潜在的黑客攻击行为,安全公司往往会将这些恶意域名进行整理并加入防火墙或者杀毒软件的黑名单列表。目前从文本中提取恶意域名的技术主要还是基于正则表达式和白名单技术,这种技术存在很大的误报率,即没有在白名单列表中的域名不一定就是恶意域名。因此,如何从海量技术文本中自动提取恶意域名在网络攻击检测与防御

基金项目:保密通信重点实验室基金(No.9140C110401140C11053)。

作者简介:黄诚(1987—),男,博士生,研究方向为信息系统安全;刘嘉勇(1962—),男,博士,教授,主要研究方向为信息安全理论与应用、网络通信与网络安全;刘亮(1982—),通讯作者,男,讲师,主要研究方向为信息安全,E-mail:liangzhai118@163.com;何祥(1988—),男,博士生,研究方向为信息系统安全;汤殿华(1986—),男,硕士,主要研究方向为信息安全理论与应用。

收稿日期:2016-12-20 **修回日期:**2017-02-16 **文章编号:**1002-8331(2018)09-0101-08
CNKI **网络出版:**2017-08-29, http://kns.cnki.net/kcms/detail/11.2127.tp.20170829.1420.004.html

方面具有重要的作用。

近年来,很多学者都在研究快速检测恶意域名并提取黑名单列表的方法。文献[1-2]提出的恶意域名检测技术主要是基于域名对应网页的内容进行分析和提取,其提取特征主要包含域名的书写规则、网页中的元素位置、网页静态和动态行为特征。文献[3-5]则是从域名自身的词法、句法等方面进行识别和检测。虽然以上研究实现了常规恶意域名的检测工作,但是目前针对文本中恶意数据挖掘的研究成果则微乎其微[6-7],而文本挖掘技术在计算机领域则相对成熟,文献[8-9]分析了目前文本挖掘的特征提取、语料生成算法和自动分类技术。

为解决文本中的恶意域名提取工作,论文试图将已有的文本挖掘和机器学习分类技术引入其中,并结合恶意域名文本描述的相关特征提出了基于上下文语义的恶意域名语料提取模型,并通过多个实验验证了模型的可行性和语料的有效性,扩展了目前在文本数据中恶意域名提取方法,实现了文本中的恶意域名的自动提取方法。

2 相关基础

2.1 语料库简介

语料库(Corpus)[10]是存放语言相关材料的仓库,其内容主要由同类型的语料组成。语料库的价值是通过语料的标注来体现的,标注的层次越高、规模越大、准确性越好,其价值就越高。标注后的语料一般称为熟语料,未标注的语料一般称为生语料[11]。目前,语料库在很多研究领域(尤其是文本处理、图像识别等)都获得了高度成功,已成为自然语言处理研究领域不可分割的一部分。

本文讨论的恶意域名语料是指从文本上描述域名安全性的单词或者短语,这些单词或者短语可以从文本语义上进行上下文描述。同时,通过这些语料可以训练有监督学习模型,从而对文本数据中域名的安全性进行标注或者预测。

2.2 上下文的定义

上下文目前还没有公认的标准定义,研究中采用较多的是Dey[12]提出的定义:“上下文是任何可用于刻画实体所处环境的信息。这里的实体可以是人,也可以是地点或者其他在用户和应用程序交互过程中所涉及的对象,甚至还包括用户和应用程序本身”。从这个定义可以看出,实体和上下文之间存在如下的关系:上下文是用来描述实体属性的信息,同时实体与上下文也是相对的概念。目前基于上下文的相关研究主要集中在推荐系统、信息检索、个性化推荐等领域中,在信息安全领域则相对较少。

本文讨论的是基于上下文语义的恶意域名语料提取模型,恶意域名是上下文定义中的实体对象,而恶意域名的属性信息则是通过文本中上下文的语义进行描述。安全人员在发布的博客文章或者正式报告中经常使用跟描述安全相关的单词或者短语来描述域名的安全性。例如,下面的句子是从实际APT攻击文档中提取的一段文字,通过这个例子可以对论文针对的上下文语义进行简单描述。

We also observed another HTTPS Gh0st variant connecting to a related command and control server at http://me.scieron.com.

此句中“http://me.scieron.com”即是论文研究的恶意域名对象,其上下文语义是指这个域名在句中前后的描述词语,这些描述用语既是基于安全人员的技术分析结果,又结合了安全人员的书写用语习惯。因此,可以通过这些词语可以对域名的安全属性进行描述。

2.3 向量空间模型

向量空间模型(Vector Space Model,VSM)由Gerard Salton等提出,是目前使用最广泛的文本表示模型[13]。在这种传统的向量空间模型中,文本集合 $D=\{d_1,d_2,\cdots,d_m\}$ 泛指机器可以读取的记录,可以是一个网页或是一篇文档;特征词集合 $T=\{t_1,t_2,\cdots,t_n\}$ 是用来体现文章主题内容的关键字。对于包含k个特征词的文本 $d_i=\{t_{i1},t_{i2},\cdots,t_{ik}\}(1\leqslant i\leqslant m,1\leqslant k\leqslant n)$,根据各个特征词 $t_{ij}(1\leqslant j\leqslant k)$ 在文本中的重要性计算其权值 $w_{ij}(1\leqslant i\leqslant m,1\leqslant j\leqslant k)$,则文本 d_i 可以用 k 维空间的一个向量 $\boldsymbol{w}_i=(w_{i1},w_{i2},\cdots,w_{ik})$,整个文本集合 D 就可以表示为 k 维空间的 m 个向量 $(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_m)$,如图1所示。

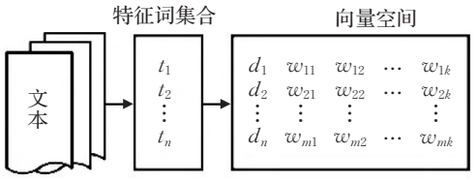

图1 向量空间模型表示流程

在实际用向量空间模型来表示文本信息时,有以下几方面内容需要详细研究:首先,从文本中生产特征词的过程中需要提出特征词提取算法,算法的优劣与否直接影响到文本的语义信息;其次,特征词集中词的数量一般比较庞大,需要使用排序算法对特征词进行重要性分析。因此,本文在基于文本数据的恶意域名语料提取研究中需要对以上问题进行详细分析。

3 基于上下文语义的恶意域名语料提取模型

3.1 设计思路

恶意域名语料提取模型主要研究如何从网络数据

中构建描述恶意域名的相关语料,并且这些语料可以从文本层面对域名的安全性进行区分。恶意域名语料提取模型生成的语料库可以结合机器分类算法实现从文本数据中自动提取恶意域名,进而生成恶意域名列表,实现威胁情报信息IOC(Indicator of Compromise)数据的自动提取,最终这些列表数据可供其他诸如防火墙、终端防护等安全设备使用。

作者的书写风格通常体现在其用字、遣词、造句等文法习惯中,因此,字、词、句等都可以作为特征来表示作者的书写风格。自书写风格识别开始引起研究者关注以来,如何选择和提取出更丰富、更有分辨力的书写风格特征就一直是研究的重点。论文结合文本的语言特点并依据向量空间模型相关理论,提出了基于文本上下文语义的恶意域名语料提取过程:提取模型的数据集可以用 $D=\{d_1,d_2,\cdots,d_m\}$ 进行表示,其中 d_m 可以是一个网页或是一篇文档。针对这些数据按照语料提取算法进行处理并生成语料,具体操作内容包含上下文单词和短语,通过算法进行处理后可以得到上下文语料 (C) 与2-gram语料 (G) 两种语料,从而得到恶意域名语料库。其中,恶意语料库中上下文语料 (C) 与2-gram语料 (G) 在一定程度上可以表明被描述域名的安全性,其选择的主要原因分析如下。

(1)上下文单词

通过对大量文本内容分析后发现,安全分析人员经常在技术博文中使用大量安全相关的单词或者短语来解读域名的安全性。例如,从2.2节中的描述文字中可以看出,其句子中包含很多专业用语来描述句中的域名,通过阅读这些文字可以给出该域名的安全性,同时句中包含了很多停用词(stop words,在句中没有实际含义的词语),如果去掉这些停用词可以得到如下的结果:

~~We~~ ~~also~~ observed ~~another~~ HTTPS Gh0st variant connecting ~~to~~ ~~a~~ related command ~~and~~ control server at http://me.scieron.com.

处理后的文字能很详细地描述域名"me.scieron.com"的属性和安全性:该域名被描述为恶意域名,已被攻击者用于恶意木马的通讯地址。经过处理后的句子其文字数量更少,单从每个词的语义很难对于目标域名的恶意性进行描述,但是经过处理后的单词组合成词包,通过这些词语可以直接推断域名的安全性,这个处理过程也在一定程度上减少语料库的大小。同时,如果有大量的数据进行训练,域名的安全性就可以通过上下文的文字内容进行判定。

为了获取恶意域名相关的文字,论文主要从采包含目标域名的句子进行分析。已有的研究模型中,往往是获取整个文本的所有词语,但是这个策略在实际中行不通,因为每个文档中都会包含正常域名和恶意域名,如果选择BOW(Bag-of-Words)模型[14]来提取所有的短语,那么机器学习分类模型将无法奏效。因此,只选择包含有目前域名的句子是一种更好的思路。

(2)上下文短语

虽然上下文单词可以在一定程度上表明句中域名的安全性,但是单个词语包含的信息量较少,无法表示更多的含义。在文本分类中,最具代表性的字符类特征是*N*-gram字符,*N*-gram字符又称为*N*元字符串,是指长度为*N*的字符序列,在文本分类任务中通常用于文本表示。假设有一大小固定为*N*的滑动窗口对文本内容进行滑动操作,每次滑动一个字符,形成长度为*N*的字符片段序列,则每个字符片段序列称为一个gram。借助向量空间模型VSM来表示文本,将所有的gram按频度进行统计和过滤后形成的列表,即可为该文本的特征向量空间,每一个gram表示一个特征向量维度。*N*-gram技术在20世纪80年代至20世纪90年代经常被用于拼写错误检查,输入字符预测,文献语种识别等。20世纪90年代以后在自然语言处理领域得到了新的发展,例如文本自动分类、自动分割等。但是在安全领域主要还是通过*N*-gram去提取片段代码或者数据,从而组合成不同的特征方便相似度比较或者分类处理[15]。

N-gram具有独立于语种,预处理简单,容错能力强,包含特征信息丰富等特征,从而弥补单纯基于上下文单词语料提取方式的不足。论文试图借助*N*-gram的短语提取方式和VSM空间向量模型理论,提出基于*N*-gram的上下文短语语料生成方法。在使用*N*-gram文本分类中,其中一个重要的问题是关于*N*值大小的确定,即*N*-gram字符串序列的长度。最佳*N*值的选择不能走过大或者过小两个极端,需要在两方面保持一定的平衡。通过查阅相关文献[9],同时鉴于论文实验数据为常规技术文本内容,因此,在实际提取过程中论文选择*N*=2作为此次上下文短语的生成长度。

通过以上对上下文单词、上下文短语提取思路的分析。论文恶意语料提取过程可以用图2进行详细描述。首先,从文本数据集 $D=\{d_1,d_2,\cdots,d_m\}$ 针对每个文本 d_m 进行处理,即从文本 d_m 提取包含有恶意域名集 $M=\{m_1,m_2,\cdots,m_i\}$ 中任一域名 m_i 的句子 $S=\{s_1,$

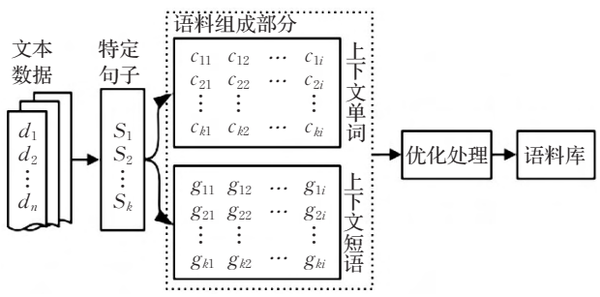

图2　恶意域名语料提取描述过程

$s_2,\cdots,s_k\}$;其次,针对每个包含域名的句子 s_k 从上下文本提取上下文语料 $c_k=\{c_{k1},c_{k2},\cdots,c_{ki}\}$ 和 $g_k=\{g_{k1},g_{k2},\cdots,g_{ki}\}$;然后,针对语料 c_k 和 g_k 进行进一步优化处理;最终得到恶意域名语料数据,从而形成语料库。

3.2　整体框架

根据以上对恶意语料提取思路的分析,论文提出恶意域名语料提取模型的总体结构设计图,如图3所示,模型总体上由数据输入层、业务逻辑层、数据输出层三部分构成。

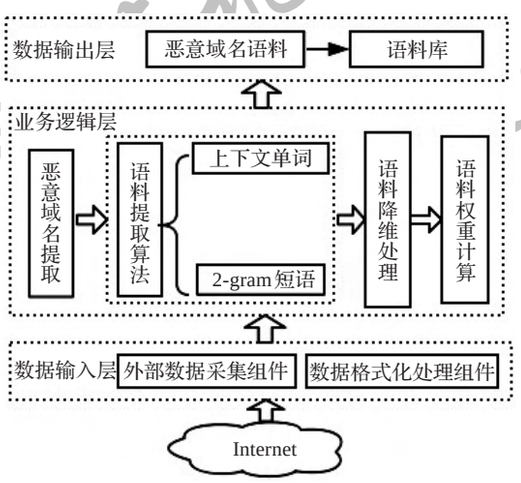

图3　恶意域名语料提取模型整体框架

各部分的主要功能设计如下:

(1)数据输入层

提供对外部数据的采集,并针对不同类别的数据进行格式化处理。外部数据采集的主要数据源是目前公开的恶意攻击APT攻击分析文章、文档或博客文章,由于包含不同的数据格式,因此需要采用数据格式化处理组件对其进行处理。

(2)业务逻辑层

属于恶意域名语料提取模型的核心技术层,实现了从格式化文本数据到最终生成恶意域名语料过程中的所有功能,包含恶意域名提取、语料提取算法、语料降维、权重计算等。

(3)数据输出层

提供带有权重的恶意域名语料数据,并可以通过此类数据构建语料库,供其他机器分类模型使用。

3.3　关键技术

3.3.1　语料提取算法

目前大部分语料库生成模型都是基于BOW模型[16],这些模型都是将完整的文本作为目标数据进行短语提取,从而生成对应的语料库,由于语料库包含了全文所有的短语,导致冗余信息太多,因而每个短语的信息量较低。同时,如果一份文本中同时包含恶意域名和正常域名,采用BOW模型提取的正常域名语料和恶意域名语料内容就会相同,因此,直接采用已有模型来提取描述恶意域名的语料行不通。通过对域名上下文的文字描述内容进行分析,本文提出了基于上下文语义的恶意语料提取算法,该算法仅仅从恶意域名所在句子的上下文语义进行分析,得到恶意语料的上下文单词和短语。

算法1　恶意域名语料提取算法

输入:包含恶意域名分析内容的文档集。

输出:生成可以描述恶意域名的语料(上下文单词、2-gram短语)。

步骤1　分别对每个文档进行格式化处理,只提取包含域名的句子。

步骤2　提取句子中所有域名,利用在线域名检测平台对域名进行安全性标注,并选择所有恶意域名。

步骤3　从所有句子中选择含有恶意域名的句子,并把这些句子进行下一步处理。

步骤4　通过2-gram生成算法对上一步得到的句子提取短语,从而生成恶意语料库中的2-gram短语。

步骤5　继续对步骤3的句子进行分词,移除停用词和时态还原等操作,然后将处理后的单词组合成词包,从而得到了上下文单词集合。

步骤6　将步骤4得到的2-gram和步骤5得到的上下文单词进行去重。

3.3.2　语料降维方法

降维操作的目标是将高维特征空间映射到一个低维的特征空间,在本文分类中,比较常见的降维方法主要包含特征词选择和特征词析取:特征词选择就是降维后的特征向量是降维前的特征空间的子集,所使用的手段有组合、转换、归纳等;特征词析取则主要通过特征词聚类、隐含语义索引、基于概念层次的降维方式进行处理的方法。通过对文本分类中现有降维方法的分析,同时结合恶意域名语料的实际内容和英文书写特征,本文提出了基于单词频度的选择方法与基于特征词主成份分析两种方法对恶意域名语料进行降维,如图4所示。

基于单词频度的选择方法主要考虑到很多英文停用词和标点符号会在文本中出现多次,同时大部分单词和符号对句子所表达的意思影响很小,其包含的信息熵很小。因此可以直接从文本中删除。停用词主要是用来连接各类词语,但在句子中没有任何含义的词语。通过分析NLTK的语料库发现[17],英文的常用停用词只有127个单词,但是其中一些词语还带有一定的感情色彩或者主观态度,可以影响到整个句子或者目标的含义,例如:no、not、too、very。虽然这些词语属于英文的停用词,但是实验中没有将这些有意义的停用词移除,而其他停用词在分词后进行了删除操作。同理,在句子中的

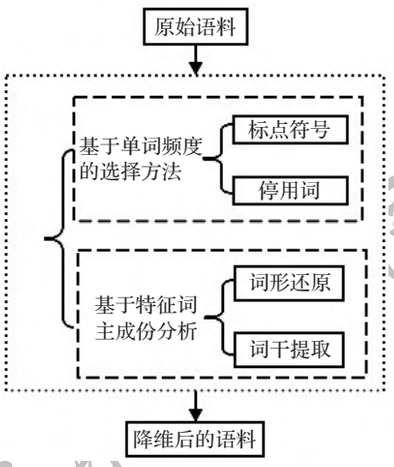

图4　恶意域名语料降维方法流程图

某些标点符号(如:!,?)可以在一定程度上影响被描述的域名,这些字符需要进行保留。另外的一些字符(如:",$)则对目标域名的描述没有任何帮助,这些字符同样需要删除。

基于特征词主成份分析方法主要考虑词的不同形态归并化处理,即词形规范化,用于降低整个语料的维度。其主要内容包含词形还原和词干提取,词形还原是把一个任何形式的语言词汇还原为一般形式,而词干提取是抽取此的词干或者词根形式。词形还原主要是针对动词在不同的语境和句子中不同的时态进行还原,比如第三人称单数、一般现在时、过去式等。目前这类操作主要有基于规则的方法、基于词典的方法、基于机器学习的方法和混合的方法,其中基于词典的词形还原方法也是最主流的方法。为了实现词的形态还原和词干提取操作,论文采用基于词典的方法对恶意域名语料数据进行处理,其主要思想是利用词典映射查询找到对应词形的原形,从而还原词的词根形式。其中,学术研究中也已经提出的Porter Stem Filter[18],Lancaster stemmer[19]等常规还原算法。论文在实现过程中主要利用NLTK和WordNet项目中的词典对语料进行还原操作,借助现有词典进行词形识别、词形和原形的映射,从而减少最终生成的恶意语料的维度。

3.3.3　语料权重计算

为了更准确地描述恶意域名语料库中每个语料的重要性,在对语料进行降维之后,需要计算每个语料在语料库中的权重。通过权重计算可以有效地筛选出对分类器比较有用的语料,常用的权重计算方法主要有:布尔权重,即通过二值(0或者1)去标注特征权重;频度权重,根据语料在本文中出现次数来计算其权重;TF-IDF(Term Frequency-Inverse Document Frequency)权重,语料在越多的文本中出现,越不重要;熵权重,通过计算语料的信息熵来表示权重;其他基于TF-IDF和信息熵改进的算法。

论文在结合实验数据并对比以上几类方法后,提出了基于TF-IDF算法的语料权重计算方法。恶意域名语料在经过前面的降维处理后,语料的维度有所减少,但是针对语料库中的每一个语料的权重则需要利用TF-IDF算法进行详细计算。基于TF-IDF算法的语料权重计算方法主要需要考虑以下两个因素:

(1)语料的频率TF(Term Frequency):该语料在所有语料去重操作前出现的频率。

(2)语料的逆文档频率IDF(Inverse Document Frequency):该语料在所有文本数据中分布情况的量化,常用方法是利用如下的公式进行计算:

$$IDF=\mathrm{lb}\left(\frac{N}{n_k}+0.01\right)$$

其中,N为文本集合中的文档数目,n_k为出现过该语料的文档数目。但是考虑到实验中总的数据量相对较少,根据上面公式计算得到的每个语料IDF值会比较接近。谷歌和微软公司都对互联网上的文本数据有深入研究,并把相关研究成果开放给研究人员使用[20]。其中,目前运用最广泛是文本*N*-gram短语相关的数据下载和接口查询功能,这些接口可以查询每个*N*-gram短语的IDF值,其代表了该短语在互联网中的实际分布情况,因此,这个值作为TF-IDF计算公式中的IDF值更佳。语料权重计算的详细算法步骤如下:

步骤1　计算恶意域名语料库(U)中每个语料(w)在语料去重操作前的TF频率值。

步骤2　通过微软在线API(Application Programming Interface)查询接口[21]计算每个语料(w)的IDF逆文档频率值。

步骤3　通过TF-IDF公式计算每个语料(w)的权重值。

步骤4　根据权重值对所有语料进行排序,并返回结果。

最后通过如上的处理得到了每个语料的权重值,其值代表了描述域名安全性的重要程度。

3.4　模型分析

论文提出的基于上下文语义的恶意语料模型解决了机器学习分类模型中的特征提取问题,利用机器学习相关理论和模型生成的语料可以直接用于提取富文本中的恶意域名。与传统标准的BOW模型不同,语料提取算法在传统BOW模型的基础上引入了上下文语义,从而提高语料的有效性;针对向量空间模型特征的稀疏性等问题,语料降维方法结合单词频度和特征词主成份分析方法去降低语料的维度;而语料权重计算方法虽然采用传统的TF-IDF算法,但是每个语料的IDF值则是基于海量数据的统计分析接口得到的,计算得到的TF-IDF值代表了语料描述恶意域名的相关性,在实际特征

选择时可以按照TF-IDF值选择特征的数量。因此,论文提出的恶意域名语料提取模型充分结合了文本处理技术和机器学习理论,为海量文本中恶意域名提取技术提供了一种新方向。

4 实验及结果分析

4.1 实验数据描述

为了验证论文提出的基于上下文语义的恶意语料模型,实验时需要包含有大量描述恶意域名或对恶意域名进行讨论的文本数据。通过对互联网上公开数据和技术报告进行分析后发现,从2008年到2015年间许多安全公司都发布了正式的APT攻击分析报告。这些报告主要针对各种黑客攻击行为进行剖析,也对攻击中所使用的各种资源(软件、域名、主机等)进行了描述和深度分析,尤其是针对域名和人员信息。实验中从互联网上共获取到277份APT攻击分析文档,通过对原始数据进行格式化处理,最终得到了4 073个顶级域名。

4.2 域名提取与标记

恶意域名语料提取框架的主要目的是自动提取描述恶意域名相关的语料,因此需要研究如何提取文档数据中的域名,并对这些域名的安全性进行判定。为了提取文档数据中的恶意域名,本文采用以下三个步骤对实验数据进行处理:URL提取器、黑名单标记和白名单过滤。首先,数据在经过预处理后,需要通过正则去匹配文本中所有的网址。其次,通过黑名单服务去标注恶意域名。最后,通过白名单过滤技术去降低黑名单标记服务的误报问题,从而让恶意域名提取结果更加准确。

(1)URL提取器

虽然可以利用Python中的正则表达式来提取文字中正常的网址,但是安全人员在引用恶意网址的时候为了防止用户点击恶意链接,往往会改用不同的书写方式。例如,下面所展示的几种书写方式都来源于真实的APT分析报告中。

A.hxxp://java-se.com/o.js

B.ct.datangcun[.]com

C.http://tempuri(dot)org/GetServerTime

通过与正常的链接对比发现,如果使用正常的URL正则表达式则无法提取这些恶意域名,因此,在提取URL时需要考虑这类书写方式的特殊性,将“[.]、hxxp、(dot)”这几种书写方法都进行考虑,最终,论文给出了如下的正则式来提取这类链接。

```
'''(?i)\b((?:(hxxps?|https?|ftps?)://|www\d{0,3}[.]|[a-z0-9.\-]+[.][az]{2,4}/)(?:[^\s()<>]+|\(([^\s()<>]+|(\([^\s()<>]+\)))*\))+(?:\(([^\s()<>]+|(\([^\s()<>]+\)))*\)|[^\s`!()\[\]{};:'".,<>?]))'''
```

(2)黑名单标记

通过前面的URL提取器从文本中提取了大量域名,但是这些域名没有标注是否为恶意域名,黑名单标记则是对这些网址进行安全性判定。鉴于很多互联网在线域名检测平台(如VirusTotal[22]、UrlQuery[23]等)都提供对域名安全性标记服务,同时VirusTotal作为目前最知名的在线恶意检测平台,已经包含了超过60款全球的安全杀毒软件。针对每个待检查对象,VirusTotal会详细给出有多少款杀毒软件已经将该对象列入了黑名单,其数字越大,说明其恶意性越肯定。所以论文在实际黑名单标记中也是采用了VirusTotal在线网站安全鉴定服务对域名的安全性进行标记。

(3)白名单过滤器

考虑到各种黑名单标记服务的误报率研究[24],论文又提出了白名单过滤器去修正黑名单标记服务的误报率问题。Alexa排名是亚马逊公司面向公众提供的评估网站流行度的一个服务。通常而言,Alexa通过长时间对各种网站流量、内容和链接等方面进行分析,最后再对域名进行排名。因此,如果一个域名在Alexa排名非常高,那么其安全性也相对较高。因此,本文选择Alexa排名来移除正常域名,所有在上一步被在线安全检测服务标记为黑名单的域名,都需要判定其是否在Alexa排名中,如果在排名中则从黑名单列表中移除。

最终,论文采用Alexa排名的前一百万网站作为白名单,通过黑名单标记和白名单过滤处理后,共有2 861个域名被VirusTotal标注为恶意域名并且没有出现在Alexa排名中,这些域名将作为恶意语料提取实验的域名数据。

4.3 语料提取实验结果

为了更准确地从APT攻击分析文档中提取恶意语料,论文分析了文档写作的语言风格,研究了自然语言处理技术和现有的黑名单标记技术。实验中从APT攻击分析文档中获得了2 861个恶意域名。基于论文提出的语料提取模型,总共获取了3 209个上下文单词和7 871个2-gram短语,并根据语料的权重进行了排序。表1列举了排名前15的上下文单词和短语,所有经过排名的语料库已经发布到恶意语料库开源项目[25]。

5 语料有效性评估实验

5.1 评估方法

经过上面实验得到了大量恶意域名语料,但是没有通过实验去验证这些语料的有效性。为了对比提取特征的有效性,本次实验通过真实数据来验证论文模型的可行性,实验主要采用机器分类算法来自动提取安全文章中的恶意域名。其中分类模型中的所有特征都来自

表1　排名靠前的恶意语料(单词与2-gram)

序号	单词	权重	2-gram	权重
1	domain	0.040 21	security crysys	0.011 61
2	malware	0.032 35	of the	0.011 30
3	server	0.026 62	the following	0.010 14
4	fidsecsys	0.022 37	crysys budapest	0.009 80
5	wa	0.022 25	the domain	0.009 07
6	cc	0.020 28	of cryptography	0.008 91
7	ip	0.019 54	to the	0.008 75
8	sample	0.019 54	cryptography and	0.008 66
9	file	0.018 76	the malware	0.008 26
10	used	0.018 38	by stteam	0.008 12
11	security	0.016 85	budapest university	0.007 92
12	md5	0.016 82	in the	0.007 74
13	crysys	0.016 56	system security	0.007 56
14	following	0.015 99	imagegif image	0.007 51
15	address	0.015 56	laboratory of	0.007 42

于上面实验生成的恶意域名语料,而训练数据和测试数据都是基于新的数据集。

恶意语料库有效性评估整体验证过程如图5所示,实验过程中具体流程如下:首先,对博客文章文件格式化分析,并提取出所有域名及域名所在句子内容;其次,通过在线域名安全检测平台对所有域名进行标注;然后,结合论文生成的恶意语料库和域名所在句子的文本内容对所有域名的特征进行赋值;最后通过基于随机森林分类的机器学习方法对测试数据进行训练,然后训练的机器分类模型对测试数据进行预测,通过其预测值与域名本来的标签进行对比分析。

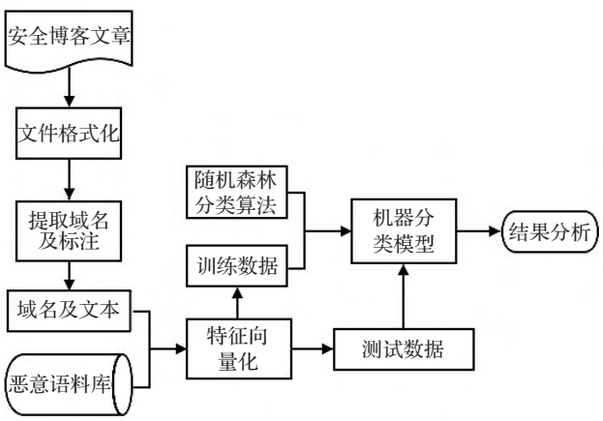

图5　恶意语料库有效性评估框架

5.2　评估数据

经过上面的实验得到了大量恶意域名语料,但是还没有实例去验证这些语料库的有效性。恶意语料提取实验主要是从公开的APT攻击分析报告中提取而来,目前许多公开的安全文章中都包含有大量恶意域名,例如:许多安全网站都发表过许多恶意分析文章。因此可以利用这些文章内容来评估论文模型提取的恶意语料的有效性。表2列举了实验中所采集的安全网址及采集的文章数量。实验时首先通过爬虫程序抓取了大量安全文章,其次通过程序对获取的网页文字进行了清理和格式化,最终从这些网站抓取了7 371篇安全文章,并提取了4 538个域名。为了对获取的域名进行机器自动分类实验,需要对这些域名的安全性进行标注。实验中继续采用VirusToal查询服务和Alexa排名列表来标注恶意域名和正常域名。

表2　安全网站网址及采集的文章数量

网址	数量及比例
http://blog.trendmicro.com	2 220(30.12%)
https://securelist.com	2 109(28.61%)
http://researchcenter.paloaltonetworks.com	1 275(17.30%)
http://blogs.cisco.com/security	1 012(13.73%)
https://blogs.mcafee.com/	551(7.48%)
https://www.alienvault.com	184(2.50%)
https://www.fireeye.com	20(0.27%)

实验中采用了两组特征集合进行了对比分析:第一个实验中采用上面得到的所有上下文单词和2-gram语料库,总共的特征总数达11 080个。每个域名的特征值则采用二值法进行标注,如果该域名包含这个特征则标注为1,否则为0。第二个实验中则采用部分上下文单词和2-gram语料库,为了提高模型的训练和分类速度,实验特征只选取了语料权重值最大的1 000个上下文单词和2-gram短语,总共合计2 000个特征。其特征的值也采用0或1进行标注。

为了进行机器分类计算,需要将所有的域名(包含恶意域名和正常域名)分类为训练数据和测试数据。实验中首先对两类域名进行随机排序,然后随机选择各自的70%数据作为训练样本,而剩下的30%的数据作为测试样本。

5.3　结果分析

为了更好地对比两组不同的特征值,实验中对这两组特征分类器的ROC(Receiver Operating Characteristics)曲线进行了描绘,图6显示了完整的ROC分类结果图,图中包含了使用完整特征和前2 000个特征的分类结果曲线,而图7则展示了特定区域(0~0.2)的ROC曲线图。结合两张图可以看到:基于论文生成的恶意语料的域名自动分类模型可以快速地把恶意域名从富文本中提取出来,从而验证了论文提出的基于上下文语义的恶意域名语料提取模型的有效性;同时基于部分恶意域名语料的分类模型也取得了83%以上的准确率。由于机器分类模型中经常会遇到过拟合现象,为了更好地验证论文提出的模型,实验中特别针对这种现象增加了十折交叉验证环节。通过十折交叉验证发现,该分类器的准确率可以达到0.87。

为更好地验证基于上下文语义的恶意语料提取模型的有效性,论文选择文献[3]中提出的基于词法的恶

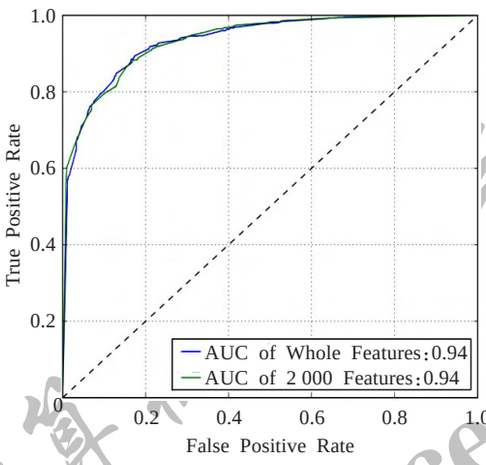

图6 域名分类模型的完整ROC曲线

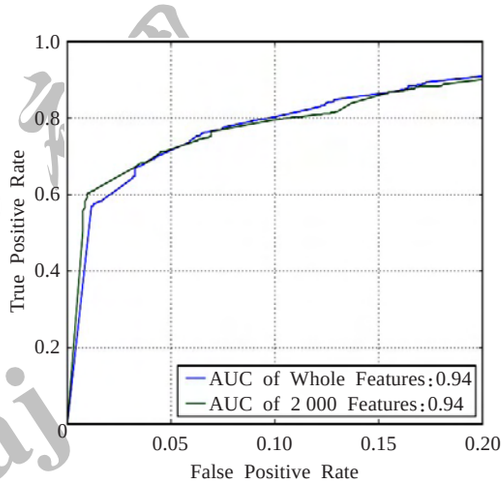

图7 域名分类模型的部分ROC曲线

意域名检测模型再次对评估数据进行训练和分类,同时也通过十折交叉试验进行验证,最终其分类准确率只能达到0.80。对比以上结果可以发现,论文提出的基于上下文语义的恶意域名提取模型扩充了特征关键字词集,并考虑了域名上下文语义,所以基于恶意域名语料库的机器分类模型的准确率提高了。

6 总结

本文通过对公开文本数据中域名上下文语义进行分析,提出了基于上下文语义的恶意域名语料提取模型,实现了恶意语料的提取、降维和权重计算方法。实验中利用该模型对公开APT分析文档进行分析,成功提取了恶意域名语料库(3 209个上下文单词和7 871个2-gram短语)。为验证该语料的准确性,本文又提出了基于安全博客文章的域名自动分类实验,其机器分类模型的所有特征都基于该恶意语料库。实验取得了87%的准确率,成功验证了语料提取模型的有效性。因此,本文的恶意域名语料提取模型为海量文本中恶意域名提取技术提供了一条新思路,并且生成的语料数据可用于各种威胁系统中的域名自动分类技术中。

参考文献:

[1] Cova M,Kruegel C,Vigna G.Detection and analysis of drive-by-download attacks and malicious JavaScript code[C]//Proceedings of the 19th International Conference on World Wide Web,2010:281-290.

[2] Zhang W,Wang W,Zhang X,et al.Research on privacy protection of WHOIS information in DNS[M]//Computer Science and its Applications.Berlin Heidelberg:Springer,2015:71-76.

[3] Wang W,Shirley K.Breaking bad:Detecting malicious do-mains using word segmentation[J].arXiv preprint arXiv:1506.04111,2015.

[4] Darling M,Heileman G,Gressel G,et al.A lexical approach for classifying malicious URLs[C]//2015 International Conference on High Performance Computing & Simulation(HPCS),2015:195-202.

[5] Park G,Taylor J M.Using syntactic features for phishing Detection[J].arXiv:1506.00037,2015.

[6] Joshi A,Lal R,Finin T.Extracting cybersecurity related linked data from text[C]//2013 IEEE Seventh International Conference on Semantic Computing(ICSC),2013:252-259.

[7] Bridges R A,Jones C L,Iannacone M D,et al.Automatic labeling for entity extraction in cyber security[J].arXiv pre-print arXiv:1308.4941,2013.

[8] 薛德军.中文文本自动分类中的关键问题研究[D].北京:清华大学,2004.

[9] 孙建文.基于集成特征选择的网络书写纹识别研究[D].武汉:华中师范大学,2011.

[10] 黄昌宁,李涓子.语料库语言学[M].北京:商务印书馆,2002.

[11] 郑家恒,张虎,谭红叶,等.智能信息处理—汉语语料库加工技术及应用[M].北京:科学出版社,2010.

[12] Dey A K.Understanding and using context[J].Personal and Ubiquitous Computing,2001,5(1):4-7.

[13] Salton G,Wong A,Yang C S.A vector space model for au-tomatic indexing[J].Communications of the ACM,1975,18(11):613-620.

[14] Wallach H M.Topic modeling:beyond bag-of-words[C]//Proceedings of the 23rd International Conference on Machine Learning,2006:977-984.

[15] Jiang W,Samanthula B K.N-gram based secure similar document detection[C]//Proceedings of the 25th Annual IFIP WG 11.3 Conference on Data and Applications Security and Privacy,2011:239-246.

[16] Wallach H M.Topic modeling:Beyond bag-of-words[C]//Proceedings of the 23rd International Conference on Machine Learning,2006:977-984.

(下转144页)