(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109063117 B
(45)授权公告日 2021.01.01

(21)申请号 201810855821.1

(22)申请日 2018.07.31

(65)同一申请的已公布的文献号
申请公布号 CN 109063117 A

(43)申请公布日 2018.12.21

(73)专利权人 中南大学
地址 410083 湖南省长沙市岳麓区麓山南路932号

(72)发明人 王建新 宁翔凯 李冬 王伟平 鲁鸣鸣

(74)专利代理机构 长沙市融智专利事务所(普通合伙) 43114
代理人 龚燕妮

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/36(2019.01)
G06F 16/951(2019.01)

(56)对比文件
CN 107341183 A,2017.11.10
CN 102663093 A,2012.09.12
CN 106845230 A,2017.06.13
US 8505094 B1,2013.08.06

审查员 刘芳

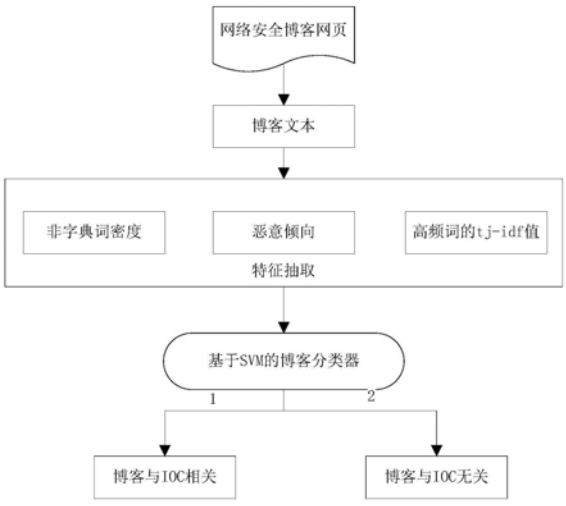
权利要求书3页 说明书8页 附图1页

(54)发明名称

一种基于特征抽取的网络安全博客分类方法及系统

(57)摘要

本发明公开了一种基于特征抽取的网络安全博客分类方法及系统,包括:爬取博客;计算每个博客的非字典词密度;计算每个博客的博客恶意倾向度;统计所有博客共同的高频词;计算每个博客中各个高频词的词频-逆文档频率;基于博客的非字典词密度、博客恶意倾向度以及每个博客中每个高频词的词频-逆文档频率,以及基于博客与IOC的相关或不相关进行编码来训练预设分类模型得到博客分类器;获取待分类博客的非字典词密度、博客恶意倾向度以及高频词的词频-逆文档频率并输入至训练后的博客分类器得到表示待分类博客与IOC的相关或不相关的分类器输出值。通过上述方法实现对网络安全技术博客中与IOC相关的博客和与IOC无关的博客精确分类。



CN 109063117 B

1.一种基于特征抽取的网络安全博客分类方法,其特征在于:包括如下步骤:

步骤1:利用网络爬虫技术,从网络安全博客网站爬取博客;

步骤2:对爬取的博客进行分词并利用预存的英文字典判断各个单词是否为字典词,再计算每个博客的非字典词密度,其中,若单词在所述英文字典内,所述单词为字典词;否则,所述单词为非字典词,第j个博客的非字典词密度计算公式如下:

$$density_j = \frac{|\{w_j\} - \{w_j\} \cap \{WordNet\}|}{|\{w_j\}|}$$

其中,{w_j}代表第j个博客的单词,{WordNet}代表在英文字典中的单词,|{w_j}|代表第j个博客单词集合中单词的个数;步骤3:分别计算每个博客的单词与预存的恶意词库的单词的平均相似度得到每个博客的博客恶意倾向度;

其中,将所述平均相似度作为对应博客的博客恶意倾向度,博客中的单词与恶意词库中的单词相似度的计算公式如下:

$$\mathrm{sim(w,m)} = \frac{\boldsymbol{W_1} \cdot \boldsymbol{W_2}}{||\boldsymbol{W_1}|| * ||\boldsymbol{W_2}||} = \frac{e_{11}e_{21} + \cdots + e_{1n}e_{2n}}{\sqrt{e_{11}^2 + \cdots + e_{1n}^2} \times \sqrt{e_{21}^2 + \cdots + e_{2n}^2}}$$

$W_1=(e_{11},\cdots,e_{1n})$,$W_2=(e_{21},\cdots,e_{2n})$

式中,sim(w,m)为博客中单词w和恶意词库中单词m的单词相似度,W_1、W_2分别为单词w和m对应的词向量,e_{11}、e_{1n}分别为词向量W_1的第1维和第n维元素,e_{21}、e_{2n}分别为词向量W_2的第1维和第n维元素;

步骤4:统计爬取的所有博客中各个单词出现的频次,并选取频次最高的N个单词作为高频词;

其中,N为正整数;

步骤5:计算各个高频词在每个博客中的词频-逆文档频率;

博客中各个高频词的词频-逆文档频率与所述高频词在每个博客中是否存在以及在对应博客中出现的频率相关;

步骤6:基于博客的非字典词密度、博客恶意倾向度以及每个博客中每个高频词的词频-逆文档频率构建输入向量,以及基于博客与网络威胁指标相关或不相关进行编码来构建输出向量,再利用构建的输入向量、输出向量训练预设分类模型得到博客分类器;

步骤7:获取待分类博客的非字典词密度、博客恶意倾向度以及各个高频词的词频-逆文档频率并输入至训练后的博客分类器得到分类器输出值;

所述分类器输出值表示待分类博客与网络威胁指标相关或不相关,所述网络威胁指标用于表示网络威胁的行为特征。

2.根据权利要求1所述的方法,其特征在于:步骤3中每个博客的博客恶意倾向度的获取过程如下:

首先,获取英文语料,并利用获取的英文语料训练出词向量模型;

然后,利用所述词向量模型计算出所述博客、所述恶意词库中每个单词的词向量;

再者,基于词向量计算所述博客中的每个单词与所述恶意词库中的每个单词的单词相似度;

最后,计算所述博客中所有单词与所述恶意词库中所有单词的单词相似度的平均相似

度。

3.根据权利要求2所述的方法，其特征在于：每个博客恶意倾向度的计算公式如下：

$$\mathrm{malic}_j = \frac{\sum_{k=1}^{|\{w_j\}|}\sum_{i=1}^{|\{m\}|} sim(w_{j,k}, m_i)}{|\{w_j\}|}$$

式中，malic_j为第j个博客的博客恶意倾向度，$sim(w_{j,k}, m_i)$为单词$w_{j,k}$与单词m_i的单词相似度，$|\{w_j\}|$为第j个博客的单词集合$\{w_j\}$中单词的个数，$|\{m\}|$为恶意词库$\{m\}$中单词的个数，$w_{j,k}$为第j个博客的第k个单词，m_i表示恶意词库中第i个单词。

4.根据权利要求2所述的方法，其特征在于：利用英文语料训练词向量模型的过程如下：

提取英文语料中的英文文本并输入Word2vec词向量模型进行训练得到词向量模型；

其中，迭代次数设置为50，词向量维度为300。

5.根据权利要求1所述的方法，其特征在于：步骤5中每个高频词的词频-逆文档频率的计算过程如下：

首先，计算每个高频词在博客中出现的频率；

其中，计算公式如下：

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

式中，$\mathrm{tf}_{i,j}$表示第i个高频词t_i在第j个博客中出现的频率，$n_{i,j}$表示第i个高频词在第j个博客中出现的次数，$\sum_k n_{k,j}$代表第j个博客中所有单词的出现次数之和；

然后，计算每个高频词在所有博客中的逆文档频率；

其中，逆文档频率的计算公式如下：

$$idf_i = log\frac{|D|}{|\{j: t_i \in d_j\}|}$$

式中，idf_i为第i个高频词t_i的逆文档频率，$|D|$表示博客总数，d_j为第j个博客，$|\{j:t_i \in d_j\}|$表示包含单词t_i的博客数目；

最后，计算每个高频词在每个博客中的词频-逆文档频率；

词频-逆文档频率的计算公式如下：

$\mathrm{tf\text{-}idf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$；

式中，$\mathrm{tf\text{-}idf}_{i,j}$为第i个高频词$t_i$在第j个博客中的词频-逆文档频率。

6.根据权利要求1所述的方法，其特征在于：所述预设分类模型为支持向量机或者逻辑回归模型或者随机森林模型。

7.根据权利要求1所述的方法，其特征在于：步骤1中从网络安全博客网站爬取博客时执行如下步骤：

首先，访问每个网络安全博客网站主页，分析博客链接特征，爬取所有博客的链接；

然后，依次访问每个博客链接，用Xpath锚定博客正文，标题，发布时间并爬取；

最后，将{链接，标题，发布时间，正文}作为一个博客条目存入数据库。

8.一种采用权利要求1-7任一项所述方法的分类系统，其特征在于：包括：

爬取模块，用于利用网络爬虫技术，从网络安全博客网站爬取博客；

特征获取模块，用于对爬取的博客进行分词并利用预存的英文字典判断各个单词是否为字典词，再计算每个博客的非字典词密度，其中，若单词在所述英文字典内，所述单词为字典词；否则，所述单词为非字典词；

特征获取模块，用于分别计算每个博客的单词与预存的恶意词库的单词的平均相似度得到每个博客的博客恶意倾向度；其中，将所述平均相似度作为对应博客的博客恶意倾向度；

特征获取模块，用于计算各个高频词在每个博客中的词频-逆文档频率，其中，高频词为统计爬取的所有博客中各个单词出现的频次，从中选取频次最高的N个单词得到；博客中各个高频词的词频-逆文档频率与所述高频词在每个博客中是否存在以及在对应博客中出现的频率相关；

分类器构建模块，用于基于每个博客的非字典词密度、博客恶意倾向度以及高频词的词频-逆文档频率构建输入向量，以及基于每个博客与网络威胁指标的相关或不相关进行编码来构建输出向量，再利用构建的输入向量、输出向量训练预设分类模型得到博客分类器；

所述博客分类器，用于基于待处理博客的非字典词密度、博客恶意倾向度以及高频词的词频-逆文档频率得到待处理博客的分类器输出值；

所述分类器输出值表示待处理博客与网络威胁指标的相关或不相关，所述网络威胁指标用于表示网络威胁的行为特征；

第j个博客的非字典词密度计算公式如下：

$$density_j = \frac{|\{w_j\} - \{w_j\} \cap \{WordNet\}|}{|\{w_j\}|}$$

其中，$\{w_j\}$代表第j个博客的单词，{WordNet}代表在英文字典中的单词，$|\{w_j\}|$代表第j个博客单词集合中单词的个数；

博客中的单词与恶意词库中的单词相似度的计算公式如下：

$$\mathrm{sim}(w,m) = \frac{\boldsymbol{W_1} \cdot \boldsymbol{W_2}}{||\boldsymbol{W_1}|| * ||\boldsymbol{W_2}||} = \frac{e_{11}e_{21} + \cdots + e_{1n}e_{2n}}{\sqrt{e_{11}^2 + \cdots + e_{1n}^2} \times \sqrt{e_{21}^2 + \cdots + e_{2n}^2}}$$

$W_1 = (e_{11}, \cdots, e_{1n})$，$W_2 = (e_{21}, \cdots, e_{2n})$

式中，sim(w,m)为博客中单词w和恶意词库中单词m的单词相似度，W_1、W_2分别为单词w和m对应的词向量，e_{11}、e_{1n}分别为词向量W_1的第1维和第n维元素，e_{21}、e_{2n}分别为词向量W_2的第1维和第n维元素。

一种基于特征抽取的网络安全博客分类方法及系统

技术领域

[0001] 本发明属于博客分类领域,具体涉及一种基于特征抽取的网络安全博客分类方法及系统。

背景技术

[0002] 近年来,网络威胁的攻击范围不断扩大,攻击频率也越来越高。许多公司因为网络攻击遭受了巨大损失,如何应对复杂多变的网络威胁成为了各个公司关注的焦点。许多网络安全专家在对网络威胁进行分析后,将获得的网络威胁情报发布在博客中。这类博客中包含了大量的网络威胁指标(Indicator of Compromise,简称IOC),如恶意网址,木马病毒名等。这些IOC代表了网络威胁的行为特征,对于检测和防御网络攻击具有重要作用。但是网络安全博客网站上还存在很多博客与新闻和安全产品推销相关,从所有博客中筛选出和IOC相关的博客能提前为对IOC博客有需求的公司或个人过滤无关内容,提升效率。

[0003] 从所有网络安全博客中筛选与IOC相关的博客对于网络安全具有重要意义。现有的文本分类方法大多采用深度学习技术,结合文章的标题进行分类。这种方法依赖于标题对文章主题的反映程度,大多用于将文本在话题上进行分类。而在我们对网络安全博客的分类中,我们最终的类别是博客与IOC相关和博客与IOC无关,但是在与IOC无关的博客中,有一部分博客也在描述网络威胁,只是没有对网络威胁的行为特征进行分析,这类博客和与IOC相关的博客在标题上的语义区分度不高,现有的方法无法准确地将其识别出来。

发明内容

[0004] 本发明的目的是针对上述问题,通过特征抽取的方法训练一个基于特征抽取的博客分类器,由于选取的是具有高区分度的特征,因此训练的博客分类器的可靠性高,达到对网络安全技术博客中与IOC相关的博客和与IOC无关的博客精确分类的效果。

[0005] 一种基于特征抽取的网络安全博客分类方法,包括如下步骤:

[0006] 步骤1:利用网络爬虫技术,从网络安全博客网站爬取博客;

[0007] 步骤2:对爬取的博客进行分词并利用预存的英文字典判断各个单词是否为字典词,再计算每个博客的非字典词密度;

[0008] 步骤3:分别计算每个博客的单词与预存的恶意词库的单词的平均相似度得到每个博客的博客恶意倾向度;

[0009] 其中,将所述平均相似度作为对应博客的博客恶意倾向度;

[0010] 步骤4:统计爬取的所有博客中各个单词出现的频次,选取频次最高的N个单词作为高频词;

[0011] 其中,N为正整数,;

[0012] 步骤5:计算各个高频词在每个博客中的词频-逆文档频率;

[0013] 博客中各个高频词的词频-逆文档频率与所述高频词在每个博客中是否存在以及在对应博客中出现的频率相关;

[0014] 步骤6:基于博客的非字典词密度、博客恶意倾向度以及每个博客中每个高频词的词频-逆文档频率构建输入向量,以及基于博客与IOC相关或不相关进行编码来构建输出向量,再利用构建的输入向量、输出向量训练预设分类模型得到博客分类器;

[0015] 步骤7:获取待分类博客的非字典词密度、博客恶意倾向度以及各个高频词的词频-逆文档频率并输入至训练后的博客分类器得到分类器输出值;

[0016] 所述分类器输出值表示待分类博客与IOC相关或不相关。

[0017] 本发明通过对博客与ICO相关或不相关分类过程中的经验总结,提取了3个特征(非字典词密度、博客恶意倾向度、高频词的tf-idf值)作为博客与IOC相关关系的识别特征,一方面,该3个特征均是涉及了博客中具体内容,即根据博客中单词情况提炼出来的,相较于依赖于标题对文本在话题上进行分类的情况,本发明提供的3个特征可以更准确地表示与IOC相关的博客和与IOC无关的博客之间的区别、差异,另一方面,本发明的3个特征也是依据大量的实验总结的,其效果也是通过了验证的。如非字典词密度越大,代表博客中非字典词越多,则博客更有可能与IOC相关。反之,则更有可能与IOC无关,如网络安全新闻和安全产品广告等。如博客恶意倾向度涉及博客单词与恶意词库中的单词的平均相似度,由于恶意词库中的单词是与网络威胁相关的单词,因此博客恶意倾向度可以有效地表述博客与IOC的关系程度。因此,本发明通过提取的此3个特征来训练得到的博客分类器的分类结果是可靠的。进而本发明通过训练的博客分类器可以处理任意一个博客,来识别其与IOC相关或无关。

[0018] 高频词是指统计所有出现过的单词在所有博客中出现的频次,并从中选择的频次排前N的单词。其中,N的取值范围优选为[300,1000]。应当说明,在待分类博客中计算出词频-逆文档频率的高频词的为步骤4选取的N个高频词。

[0019] 进一步优选,步骤2中每个博客的非字典词密度的计算公式如下:

[0020] $$density_j = \frac{|\{w_j\} - \{w_j\} \cap \{W\}|}{|\{w_j\}|}$$

[0021] 式中,$density_j$为第j个博客的非字典词密度,$\{w_j\}$为第j个博客的单词集合,$\{W\}$为字典中英文单词的集合,$|\{w_j\}|$为第j个博客的单词集合$\{w_j\}$中单词的个数。

[0022] 例如,本发明使用WordNet字典。本发明对博客分段后分句再分词,这样的目的是为了提升分词的准确性,然后判断每个单词是否在WordNet英文字典中,不在则判断为非字典词,最后统计博客中非字典词数和单词总数,用非字典词数除以单词总数作为该博客非字典词密度的值。

[0023] 进一步优选,步骤3中每个博客的博客恶意倾向度的获取过程如下:

[0024] 首先,获取英文语料,并利用获取的英文语料训练出词向量模型;

[0025] 然后,利用所述词向量模型计算出所述博客、所述恶意词库中每个单词的词向量;

[0026] 再者,基于词向量计算所述博客中的每个单词与所述恶意词库中的每个单词的单词相似度;

[0027] 最后,计算所述博客中所有单词与所述恶意词库中所有单词的单词相似度的平均相似度。

[0028] 本发明获取的英文语料是已经公知的,例如维基百科公开的英文语料。

[0029] 进一步优选,博客中的单词与恶意词库中的单词的单词相似度的计算公式如下:

[0030] $$\mathrm{sim(w,m)}=\frac{\boldsymbol{W_1}\cdot\boldsymbol{W_2}}{||\boldsymbol{W_1}||*||\boldsymbol{W_2}||}=\frac{e_{11}e_{21}+\cdots+e_{1n}e_{2n}}{\sqrt{e_{11}^2+\cdots+e_{1n}^2}\times\sqrt{e_{21}^2+\cdots+e_{2n}^2}}$$

[0031] $W_1=(e_{11},\cdots,e_{1n})$，$W_2=(e_{21},\cdots,e_{2n})$

[0032] 式中，sim(w,m)为博客中单词w和恶意词库中单词m的单词相似度，W_1、W_2分别为单词w和m对应的词向量，e_{11}、e_{1n}分别为词向量W_1的第1维和第n维元素，e_{21}、e_{2n}分别为词向量W_2的第1维和第n维元素=(,…,)。

[0033] 恶意词库中的单词是与网络威胁相关的单词，如“exploit”，“malicious”等，这些单词与恶意单词密切相关，共同构成恶意词库。

[0034] 进一步优选，每个博客恶意倾向度的计算公式如下：

[0035] $$\mathrm{malic}_j=\frac{\sum_{k=1}^{|\{w_j\}|}\sum_{i=1}^{|\{m\}|}sim(w_{j,k},m_i)}{|\{w_j\}|}$$

[0036] 式中，$malic_j$为第j个博客的博客恶意倾向度，$sim(w_{j,k},m_i)$为单词$w_{j,k}$与单词m_i的单词相似度，$|\{w_j\}|$为第j个博客的单词集合$\{w_j\}$中单词的个数，$|\{m\}|$为恶意词库$\{m\}$中单词的个数，$w_{j,k}$为第j个博客的第k个单词，m_i表示恶意词库中第i个单词。

[0037] 从上述公式可知，统计博客中所有单词与恶意词库中所有单词的单词相似度之和，再除以博客单词数作为博客恶意倾向的值。本发明基于对博客分词后的各个单词与恶意词库中的单词的单词相关性得到博客恶意倾向度，由此可知，得到的博客恶意倾向度可以有效地表示博客与IOC的相关程度。

[0038] 进一步优选，利用英文语料训练词向量模型的过程如下：

[0039] 提取英文语料中的英文文本并输入Word2vec词向量模型进行训练得到词向量模型；

[0040] 其中，迭代次数设置为50，词向量维度为300。

[0041] word2vec训练的词向量包含了单词的语义信息。在大规模的文本上训练的词向量质量更高，也就是说对单词的语义信息表达的更完善，例如博客大约67k篇，而维基开放下载的语料有14G之大，在其上训练的词向量能够更好的反映单词的语义信息。在训练后的词向量模型中，保存了单词-词向量的映射关系，也就是说每个单词对应了一个词向量。而且词向量包含了单词的语义信息。利用词向量，我们可以计算向量之间的距离，距离小则代表单词的语义相近，距离大则代表单词的语义差别大，因此本发明基于词向量来计算博客中的单词与恶意词库中单词的单词相似度。

[0042] 进一步优选，步骤5中每个高频词的词频-逆文档频率(tf-idf值)的计算过程如下：

[0043] 首先，计算每个高频词在每个博客中出现的频率tf；

[0044] 其中，频率tf计算公式如下：

[0045] $$\mathrm{tf}_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$

[0046] 式中，$tf_{i,j}$表示第i个高频词t_i在第j个博客中出现的频率，$n_{i,j}$表示第i个高频词在第j个博客中出现的次数，$\sum_k n_{k,j}$代表第j个博客中所有单词的出现次数之和；

[0047] 然后，计算每个高频词在所有博客中的逆文档频率idf；

[0048] 其中,逆文档频率idf的计算公式如下:

[0049] $$idf_i = log\frac{|D|}{|\{j: t_i \in d_j\}|}$$

[0050] 式中,idf_i为第i个高频词t_i的逆文档频率,|D|表示博客总数,d_j为第j个博客,|{j:t_i∈d_j}|表示包含单词t_i的博客数目;

[0051] 最后,计算高频词在每个博客中的tf-idf值;

[0052] 词频-逆文档频率的计算公式如下:

[0053] 的计算公式如下:

[0054] $tf\text{-}idf_{i,j} = tf_{i,j} \times idf_i$

[0055] 式中,$tf\text{-}idf_{i,j}$为第i个高频词t_i在第j个博客中的词频-逆文档频率。

[0056] 进一步优选,所述预设分类模型为支持向量机或者逻辑回归模型或者随机森林模型。

[0057] 例如,若采用支持向量机,则采用RBF函数作为其核函数。

[0058] 进一步优选,步骤1中从网络安全博客网站爬取博客时执行如下步骤:

[0059] 首先,访问每个网络安全博客网站主页,分析博客链接特征,爬取所有博客的链接;

[0060] 然后,依次访问每个博客链接,用Xpath锚定博客正文,标题,发布时间并爬取;

[0061] 最后,将{链接,标题,发布时间,正文}作为一个博客条目存入数据库。

[0062] 本发明还提供一种上述方法的分类系统,包括:

[0063] 爬取模块,用于爬取博客;

[0064] 特征获取模块,用于计算每个博客的非字典词密度;

[0065] 特征获取模块,用于分别计算每个博客的博客恶意倾向度;

[0066] 特征获取模块,用于计算每个博客各个高频词在每个博客中的词频-逆文档频率;

[0067] 分类器构建模块,用于基于每个博客的非字典词密度、博客恶意倾向度以及高频词的词频-逆文档频率构建输入向量,以及基于每个博客与IOC的相关或不相关进行编码来构建输出向量,再利用构建的输入向量、输出向量训练预设分类模型得到博客分类器;

[0068] 所述博客分类器,用于基于待处理博客的非字典词密度、博客恶意倾向度以及高频词的词频-逆文档频率得到待处理博客的分类器输出值;

[0069] 所述分类器输出值表示待处理博客与IOC的相关或不相关。

[0070] 有益效果

[0071] 1、本发明利用特征抽取结合常用的分类方法设计了一种网络安全博客分类方法。本发明提出了3个特征用于与IOC相关博客和与IOC无关博客的分类,将抽取的特征输入简单的分类器(如逻辑回归,支持向量机等)对博客分类。选取的特征是对博客分类过程中的经验总结,使两类文本的特征差异化,如非字典词密度越大,代表博客中非字典词越多,则博客更有可能与IOC相关。反之,则更有可能与IOC无关;博客恶意倾向度涉及博客单词与恶意词库中的单词的平均相似度,由于恶意词库中的单词是与网络威胁相关的单词,因此博客恶意倾向度可以有效地表述博客与IOC的关系程度;高频词的词频-逆文档频率(tf-idf值)本身就是针对博客中的高频词的一种特征表述,且其与高频词在每个博客中是否存在以及在对应博客中出现的频率相关,因此可以更准确的表征博客正文的内容特性,本发明

巧妙地将上述三个特征结合起来作为特征向量来训练得到博客分类器，从博客内容的多方面特征来获得与IOC相关的博客以及与IOC无关的博客之间的区别差异性，从而使得得到的博客分类器的可靠性更高，使得博客分类更加准确。

[0072] 2、与一般的采用深度学习进行文本分类的方法相比，特征抽取方法能够针对不同的文本类别针对性地提取特征。比如在描述网络威胁的博客中，有些博客中含有IOC，而有的博客不含有IOC。由于深度学习的方法是基于文章的语义对文本分类，对于上述的两种博客难以区分。而我们的方法提取了博客的非字典词密度，高频词的tf-idf值和文本的恶意倾向这3个特征，能够更好地区分与IOC相关的博客和与IOC无关的博客。

[0073] 3、为了表明本发明的优越性，本发明对此进行实验验证，在数据集上采用十折交叉验证测试其性能，并采用precision(精确率)，recall(召回率)和f1作为性能指标，得出本发明若采用逻辑回归或支持向量机的分类器，precision，recall和f1三类指标均高于95，因此也进一步验证了本发明分类结果的可靠性。

附图说明

[0074] 图1是本发明提出的一种基于特征抽取的网络安全博客分类方法的流程图。

具体实施方式

[0075] 下面将结合实施例对本发明做进一步的说明。

[0076] 如图1所示，本发明公开了一种基于特征抽取的网络安全博客分类方法，具体包括以下步骤：

[0077] 步骤1:针对网络安全博客，利用网络爬虫技术，从安全网站爬取网络安全博客。

[0078] 例如，本实施例中，以安全博客网站malwarebytes为例，其对应的博客列表页面的为https://blog.malwarebytes.com/page/1，其中最后1指明了第几页博客列表。我们从第一个博客列表页面开始遍历，直到页面为空停止，利用Xpath锚定每个列表页面的所有博客链接。然后访问每个博客链接，用Xpath锚定锚定该博客的标题，发布时间，正文。最后将{链接，标题，发布时间，正文}作为一个博客条目存入数据库。应当理解，爬虫技术是现有技术来实现的，例如利用Python的Scrapy框架爬取博客。

[0079] 步骤2:对博客分词后利用英文字典判断单词是否为字典词，进而计算每个博客的非字典词密度。

[0080] 本实施例中，为了保证分词的准确性，首先根据换行符“\n”对博客分段，分段后用NLTK自然语言处理工具包的分句器分句，最后用NLTK分词器分词。依次遍历每个单词，并判断单词是否在WordNet的英文字典中，不在则判断为非字典词。

[0081] 然后，统计非字典词数和单词总数，用非字典词数除以单词总数作为该博客的非字典词密度。如第j个博客非字典词密度的量化计算方法如下：

[0082] $$density_j = \frac{|\{w_j\} - \{w_j\} \cap \{WordNet\}|}{|\{w_j\}|}$$

[0083] 其中{w_j}代表第j个博客的单词，{WordNet}代表在WordNet字典中的单词，|{w_j}|代表博客单词集合中单词的个数。非字典词密度越大，代表博客中非字典词越多，则博客更有可能与IOC相关。反之，则更有可能与IOC无关，如安全类的新闻和安全产品广告等。

[0084] 步骤3:计算每个博客中单词与恶意词的平均相似度作为博客恶意倾向。

[0085] 本实施例中首先收集与网络威胁相关的单词,如“exploit”,“malicious”等,这些单词与网络威胁密切相关,共同构成恶意词库。对博客分词,计算博客单词与恶意词库中每个单词的相似度。

[0086] 其中,每个单词的词向量基于训练的词向量模型获取,词向量模型是从维基百科公开的下载地址获取英文语料。然后用加载英文语料,将其中的纯文本输入Word2vec词向量模型训练得到的。其中,单词相似度的量化计算方法如下:

[0087] $$\mathrm{sim(w,m)}=\frac{\boldsymbol{W_1}\cdot\boldsymbol{W_2}}{||\boldsymbol{W_1}||*||\boldsymbol{W_2}||}=\frac{e_{11}e_{21}+\cdots+e_{1n}e_{2n}}{\sqrt{e_{11}^2+\cdots+e_{1n}^2}\times\sqrt{e_{21}^2+\cdots+e_{2n}^2}}$$

[0088] 式中,$W_1=(e_{10},\cdots,e_{1n})$,$W_2=(e_{20},\cdots,e_{2n})$分别代表单词w和m对应的词向量,n代表了词向量的维度。两个单词的相似度在(-1,1)的范围内,其中-1表示两者的语义相反,1表示两者的语义很接近甚至相同。例如在我们训练的词向量模型中,单词”attachment”与”malicious”的相似度为0.2737,表明单词”attachment”与单词“malicious”的语义正相关。

[0089] 然后,统计博客单词与恶意词库单词的相似度之和,除以博客中单词数作为博客恶意倾向的值。对第j个博客而言,其恶意倾向的量化计算方法如下:

[0090] $$\mathrm{malic}_j=\frac{\sum_{k=1}^{|\{w_j\}|}\sum_{i=1}^{|\{m\}|}sim(w_{j,k},m_i)}{|\{w_j\}|}$$

[0091] 式中,malic_j为第j个博客的博客恶意倾向度,$\mathrm{sim}(w_{j,k},m_i)$为单词$w_{j,k}$与单词$m_i$的单词相似度,$\{w_j\}$代表第j个博客的单词集合,$\{m\}$代表恶意词库单词集合。$w_{j,k}$表示第j个博客的第k个单词,$m_i$表示第i个高频词。

[0092] 步骤4:统计所有博客中单词出现的频次,选取频次最高的1000个单词作为高频词。

[0093] 本实施例中从数据库中取出爬取的博客,对每个博客分段后分句再分词,并对每个单词做词型还原。例如“exploiting”,“exploit”,“exploits”等都还原为“exploit”。这样能够更准确地统计单词的词频,进而获得更可靠的高频词。

[0094] 然后统计在所有博客中出现的每个单词的出现次数,并根据单词的出现频次从高到低对单词排序,选取前1000个单词作为高频词。我们获得的前1000个高频词有“security”,“malware”,“sample”等。

[0095] 应当理解,本实施例中选择频次最高的1000个单词作为高频词,其他可行的实施例中,不限定为选择1000个,可以是其他数量。

[0096] 步骤5:计算高频词在每个博客的tf-idf值。

[0097] 本实施例中,首先,对每个博客分词并计算1000个高频词的tf值。第j个博客中第i个高频词的tf-idf值的量化计算方法如下:

[0098] $$\mathrm{tf}_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$

[0099] 式中,$\mathrm{tf}_{i,j}$表示第i个高频词t_i在第j个博客d_j中出现的频率,$n_{i,j}$表示第i个高频词在第j个博客中出现的次数。例如单词“malware”在高频词库中,它在博客B中出现了8次,博客B总共有2038个单词,则其在博客B中tf值为0.0039。

[0100] 然后，对每个高频词计算其在所有博客中的逆文档频率。第i个高频词t_i的逆文档频率idf_i的量化计算方法如下：

[0101] $$idf_i = log\frac{|D|}{|\{j: t_i \in d_j\}|}$$

[0102] 式中，D表示所有博客集合，$\{j: t_i \in d_j\}$表示包含单词t_i的博客集合。例如，单词“malware”出现在23128个博客中，则它的idf值为0.4663。

[0103] 最后，计算第i个高频词t_i在第j个博客的tf-idf值，其量化计算方法如下：

[0104] $tf\text{-}idf_{i,j} = tf_{i,j} \times idf_i$

[0105] 式中，$tf_{i,j}$表示第i个高频词在第j个博客中的频率，idf_i表示第i个高频词t_i的逆文档频率。则单词“malware”在博客B中的tf-idf值为0.0018。

[0106] 步骤6：构建输入向量和输出向量，然后训练基于SVM的博客分类器。

[0107] 本实施例中，首先从爬取的博客中选取一部分博客作为训练数据打标签。我们选取了2500个博客，邀请多为网络安全专家打标签，判断博客是否和IOC相关，1为相关，0为不相关。最终获得了1200篇与IOC相关的博客，2300篇与IOC无关的博客。例如，对于第j个博客而言，其对应的输入特征向量为$(density_j, malic_j, tf\text{-}idf_{0,j}, \cdots, tf\text{-}idf_{1000,j})$，共1002维；然后，构建第j个博客对应的输出向量y_j。y_j为1表示博客和IOC相关，为0表示博客和IOC无关。

[0108] 抽取每个博客的非字典词密度、恶意倾向和高频词的tf-idf值，并这些特征组合起来作为对应的特征向量。对于第j个博客而言，其对应的1002维输入特征向量为$(density_j, malic_j, tf\text{-}idf_{0,j}, \cdots, tf\text{-}idf_{1000,j})$。3500个博客训练数据的特征向量以及对应的标签构成训练集，训练一个采用RBF核函数的SVM博客分类器。其他可行的实施例中，可以选择其他类型的分类器。

[0109] 步骤7：对待分类的博客按照步骤2、3、4抽取特征，并用训练后的博客分类器进行分类。

[0110] 本实施例中，结合步骤2、3、4抽取待分类博客的特征向量$(density_{pred}, malic_{pred}, tf\text{-}idf_{0,pred}, \cdots, tf\text{-}idf_{1000,pred})$，输入上述训练后的SVM博客分类器中，获取分类器的输出。输出为1代表待分类博客和IOC相关，为0代表博客和IOC无关。

[0111] 基于上述方法，本发明还提供一种基于特征抽取的网络安全博客分类系统，包括：

[0112] 爬取模块，用于爬取博客；特征获取模块，用于计算每个博客的非字典词密度；特征获取模块，用于分别计算每个博客的博客恶意倾向度；特征获取模块，用于计算各个高频词在每个博客中的tf-idf值；分类器构建模块，用于基于每个博客的非字典词密度、博客恶意倾向度以及高频词的tf-idf值构建输入向量，以及基于每个博客与IOC的相关或不相关进行编码来构建输出向量，再利用构建的输入向量、输出向量训练预设分类模型得到博客分类器；所述博客分类器，用于基于待处理博客的非字典词密度、博客恶意倾向度以及高频词的tf-idf值得到待处理博客的分类器输出值；所述分类器输出值表示待处理博客与IOC的相关或不相关。

[0113] 应当理解，上述模块的具体实现过程与本发明提供的一种于特征抽取的网络安全博客分类方法相同，因此，再次不对其进行赘述，请参照上述方法步骤的执行过程，且需要说明的是上述模块的划分仅仅是功能性划分，其可以根据实际需求进行合并、划分和删减。

为了验证本发明的优越性，分别用逻辑回归(Logistic Regression)，支持向量机(SVM)，决策树(Decision Tree)和随机森林(Random Forest)作为分类器，在我们的数据集上采用十折交叉验证测试其性能。我们采用precision(精确率)，recall(召回率)和f1作为性能指标。其中，precision代表了分类为和IOC相关的博客中，真正和IOC相关的博客所占的比例，recall代表了正确分类的和IOC相关的博客在所有和IOC相关的博客中占的比例，f1值则是对precision和recall的一个加权调和平均，能够较好地反映分类的整体表现。其量化计算方法如下：

[0114] $$\text{precision} = \frac{TP}{TP + FP}$$

[0115] $$\text{recall} = \frac{TP}{TP + FN}$$

[0116] $$\text{f1} = \frac{2 \times precision \times recall}{precision + recall} = \frac{2 \times TP}{2 \times TP + FN + FP}$$

[0117] 其中，TP表示和IOC相关的博客中，被分类为和IOC相关的博客数目。FP表示和IOC无关的博客中，被分类为和IOC相关的博客数目。FN表示和IOC相关的博客中，被分类为和IOC无关的博客数目。综合评价如表1所示。

[0118] 表1:3个特征在不同分类器上的表现

[0119]

	Precision	Recall	F1
Logistic Regression	95.59	96.25	95.88
SVM	95.95	96.92	95.47
Decision Tree	86.04	86.75	85.90
Random Forest	95.59	89.50	94.45

[0120] 根据上表可以看出，我们提出的3个特征在Logistic Regression、SVM和Random Forest上均获得了较好的表现，在Decision Tree上获取的表现较差。当采用Logistic Regression时，3个特征对于博客分类的表现最好。综上，我们提出的3个特征能够有效地用于与IOC相关博客和与IOC无关博客的分类。

[0121] 需要强调的是，本发明所述的实例是说明性的，而不是限定性的，因此本发明不限于具体实施方式中所述的实例，凡是由本领域技术人员根据本发明的技术方案得出的其他实施方式，不脱离本发明宗旨和范围的，不论是修改还是替换，同样属于本发明的保护范围。

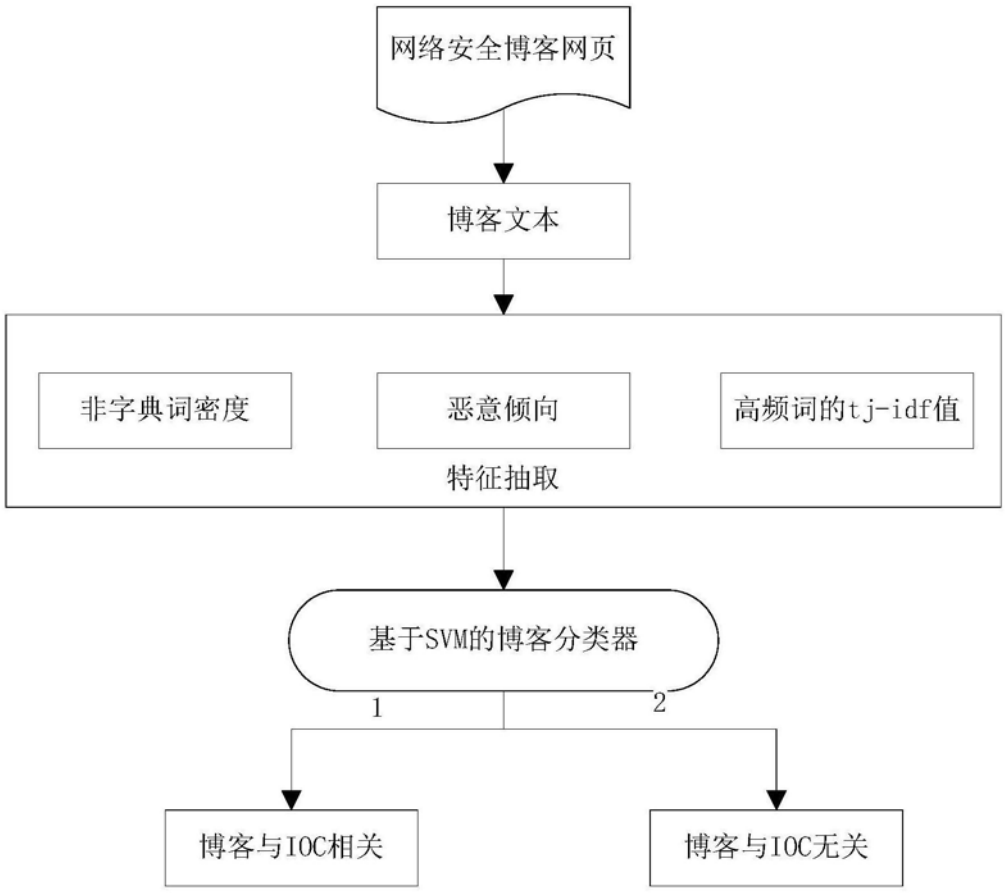

网络安全博客网页
博客文本
非字典词密度
恶意倾向
高频词的tj-idf值
特征抽取
基于SVM的博客分类器
1
2
博客与IOC相关
博客与IOC无关

图1