网络用户角色辨识及其恶意访问行为的发现方法

王 建 张仰森 陈若愚 蒋玉茹 尤建清
(北京信息科技大学智能信息处理研究所 北京 100101)

摘 要 随着互联网络技术的快速发展,各种恶意访问行为危及到网络的信息安全,因此辨识访问用户的角色并识别用户的恶意访问行为对于网络安全具有十分重要的理论意义和实用价值。首先,以网络日志数据为基础,通过建立IP辅助数据库,构建IP用户的日角色模型,在此基础上,引入滑动时间窗技术,将时间的变化动态地融入用户角色辨识,建立了基于滑动时间窗的用户角色动态辨识模型。然后,在分析用户恶意访问流量特征的基础上,将用户访问流量特征和用户信息熵特征进行加权,构建基于多特征的用户恶意访问行为的辨识模型。该模型能够对爆发性和高持续性的恶意访问行为以及少量但大规模分散访问的恶意行为进行识别。最后,采用大数据存储和Spark内存计算技术,对所建立的模型进行实现。实验结果表明,在网络流量产生异常时,所提出的模型能够发现具有恶意访问行为的用户,并准确且高效地辨别出该用户的角色,从而验证了其有效性。

关键词 网络用户,数据挖掘,角色辨识,恶意访问行为,滑动时间窗
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.10.030

Identification of User's Role and Discovery Method of Its Malicious Access Behavior in Web Logs

WANG Jian ZHANG Yang-sen CHEN Ruo-yu JIANG Yu-ru YOU Jian-qing
(Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing 100101, China)

Abstract With the rapid development of Internet technology, a variety of malicious access behavios endanger the information security of network. There is theoretical significance and practical value for network security to identify user's role and discover malicious access behaviors. Based on Web logs, an IP assisted database was constructed to build IP user's daily role model. On this basis, the sliding time window technique was introduced, and the dynamic change of time was integrated into user's role identification. A dynamic identification model of user's role based on sliding time window was established. Then, analyzing the characteristics of user's malicious access traffic, the user access traffic and the characteristics of user's information entropy were weighted to construct an identification model based on multi-characteristics of the user's malicious access behavior. The model can not only identify explosive and highly persistent malicious access behaviors, but also identify the malicious access behaviors which are small but widely distributed. Finally, the model was implemented by using big data storage and Spark memory computing technology. The experimental results show that the user of malicious access behavior can be found by using the proposed model when the network traffic is abnormal, and the user's role can be identified accurately and efficiently, thus verifying its validity.

Keywords Web users, Data mining, Identification of use's role, Malicious access behavior, Sliding time window

1 引言

随着互联网在规模和复杂度等多方面的爆炸式增长,Web资源也在飞速增长,其中包括了用户访问的日志信息、内容信息以及丰富的链接信息,为数据挖掘提供了丰富的资源。网络日志挖掘旨在通过对网络日志进行有效的数据分析,来发掘隐藏在日志数据背后的Web用户角色信息和访问行为模式,以便对一些具有特殊角色和恶意访问行为的网络用户进行重点监测与跟踪。

互联网用户访问角色分为两大类:人类访问与非人类访问。通常情况下,非人类访问指的是"网络机器人"模仿正常人类进行网页浏览、社区互动、文件下载等访问动作。网络用户角色的辨识指在全部流量中识别出用户的非人类访问行为,并对用户的访问角色进行标注。网络用户的角色标注在甄别网络爬虫、机器人程序以及识别各种网络欺诈和安全威胁方面具有重要的辅助作用。

到稿日期:2017-09-09 返修日期:2017-11-26 本文受国家自然科学基金(61370139,61602044)资助。

王 建(1993-),男,硕士生,CCF会员,主要研究方向为大数据处理、自然语言处理,E-mail:455858538@qq.com;**张仰森**(1962-),男,博士,教授,CCF会员,主要研究方向为自然语言处理、人工智能,E-mail:zhangyangsen@163.com(通信作者);**陈若愚**(1982-),男,博士,讲师,主要研究方向为自然语言处理;**蒋玉茹**(1978-),女,博士,副教授,主要研究方向为自然语言处理;**尤建清**(1980-),男,硕士,讲师,主要研究方向为自然语言处理。

从流量角度来看,非人类访问行为产生的大部分流量是由善意的爬虫机器人或程序在搜索引擎上搜集数据时贡献的。但仍有部分流量是由具有恶意的用户或程序产生的,例如:2016 年 10 月 21 日上午,美国最主要的 DNS 服务商 Dyn 遭受到拒绝服务(DDoS)攻击,造成了托管 DNS 网络的中断。导致 Twitter、Spotify、Netflix、AirBnb、CNN、华尔街日报等数百家网站无法访问。此外,广告点击流量、垃圾邮件发送等恶意流量对广告公司以及用户也会造成相应的损失。由此可见,开展网络用户角色的辨识与恶意访问行为的识别,对于国家网络安全具有十分重要的理论意义和实用价值。

2　相关工作

近年来,与 Web 日志挖掘相关的研究已经成为热点[1],但与用户角色辨识相关的工作仍处于起步阶段[2]。目前,与角色辨识相近的研究工作多以用户访问模式[3]和兴趣挖掘为主[4]。

Chen 等[5]于 1996 年借鉴数据挖掘算法,构建了 Web 挖掘领域的基本理论,定义将图片、音视频等文件过滤之后剩余的 Web 服务器日志记录作为网络用户在网站中的访问记录。这一定义的提出为日后 Web 日志挖掘相关理论的研究开创了先河,同时也为各类用户访问行为及兴趣的挖掘提供了丰富的日志资源。Xu 等[6]分析网络犯罪的用户在网络行为中的恐怖主义角色,通过可视化的方法形成网络用户角色之间的通联关系网。郭岩等[7]围绕网络日志中是否蕴含用户访问 Web 的规律性特性,研究并分析了日志规模与用户数、Web 文档数以及单位用户访问的 Web 文档数之间的关系,提出了一个基于用户行为的相关文档检索模型。通过用户对 Web 访问动机的分析,认为一定时间段的 Web 访问日志中蕴含了用户的稳定兴趣。邢东山等[8]在分析用户浏览模式挖掘算法的基础上,利用支持-偏爱度的概念,提出了基于网站访问矩阵的用户浏览偏爱路径挖掘算法,利用 Web 日志建立以引用网页 URL 为行、浏览网页 URL 为列、路径访问频度为元素值的网站访问矩阵,通过对该矩阵进行支持-偏爱度计算生成用户浏览偏爱路径,反映了用户的浏览兴趣与意图。

在基于网络日志的恶意访问行为发现的相关领域,现有研究已经取得了一定的成果。目前,恶意访问行为检测方面的研究可以分为以统计网络流量特征为中心的方法[9]和以分析网络报文属性为中心的方法[10]。

Alan 等[11]根据不同网络协议之间的差别,分别利用其协议流量特征来训练人工神经网络(ANN)以发现并减轻已知的和实时网络环境中未知的恶意流量。Leung 等[12]利用聚类算法将网络数据分组划分为各个类别,再根据这些类别识别出流量分组是否异常。Rubinstein 等[13]针对聚类时数据属性较多的问题对属性进行重要程度的度量,挑选重要的属性来达到降维的目的,从而提高了异常检测的效率问题。Li 等[14]提出了一种解析数据包转发模型(APFM),可简化分组转发的计算过程,降低了恶意流量检测过程中内存的使用并缩短了仿真时间。孙知信等[15]提出了一种基于 IP 地址对数据库的防范分布式拒绝服务攻击策略。该策略建立正常流量的 IP 地址对数据库(SDIAD),并采用改进的滑动窗口无参数算法,对新的 IP 地址对进行累积分析,以快速、准确地检测出恶意性攻击行为,孙知信等[15]认为该方法主要用于边缘路由器,无论是靠近攻击源端还是靠近受害者端,都能够有效地检测出恶意性攻击行为,并且有很好的检测准确率。桂兵祥等[16]建立了一种恶意性攻击源回溯跟踪机制,通过监控正常通信流和恶意通信流之间的网络报文属性变化来判断网络是否有恶意访问行为。

综上所述,目前在网络角色辨识方面的研究成果较少,但可借鉴用户行为及兴趣方面的研究方法,提出网络用户角色辨识模型。而在恶意访问行为识别方法上,现有的研究在正确率和误报率上并不理想。以上研究内容均是基于少量数据进行挖掘的,处理延时较大且无法处理海量日志数据。

因网络日志的规模具有了大数据的特征,故大数据存储、计算等相关技术是本文在处理网络日志数据时需要使用的关键技术。李清等[17]针对大规模日志数据的聚类问题,使用 Hadoop 对原始日志数据进行预处理,并结合 k-means 和 DBSCAN 聚类算法各自的优势,提出了 DBk-means 算法,取得了更好的聚类效果。赵龙[18]提出了基于 Hive 的 Web 海量搜索日志分析机制,利用 HQL 语言以及 Hadoop 分布式文件系统(HDFS)和 MapReduce 编程模式对海量搜索日志进行分析处理,对用户搜索行为进行分析研究。

因此,本文通过大数据并行处理技术对网络访问日志进行分析挖掘,能够近乎实时地辨别用户的角色并识别恶意访问行为,因此能够对重点的用户进行监测。利用用户角色辨别模型和恶意流量的识别模型,通过对真实网络环境所产生的日志数据进行实验,能够分析得到相应的判定参数,从而提高用户角色的辨识率和恶意流量的识别准确率并降低漏报率,提升角色辨识和恶意流量识别的计算效率。

3　基于滑动时间窗的用户角色动态辨识模型

本文对网络日志的用户流量进行分析,构建了用户角色识别模型,用于在全部流量中识别出用户的非人类访问行为,并标注出用户的访问角色。将用户访问角色分为两大类:人类和非人类。对于具有两种访问角色的用户,引入非人类访问行为(Non-Human Traffic,NHT)概率作为该用户的访问角色。用户访问角色分类集如图 1 所示。

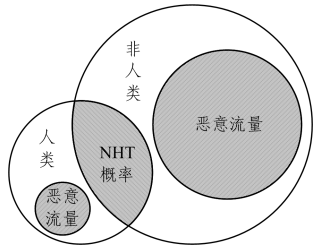


图 1　用户访问角色分类集

Fig. 1　Classification of user's access role

3.1　面向用户角色标注的 IP 辅助数据库的构建

为了更直观快速地对用户角色进行辨识,仅通过网络日志的数据来分析挖掘是远远不够的,因此需构建辅助数据库来对用户的基本信息加以表述。本文采用了一种借助 GeoLite2 IP 地址数据库、纯真 IP 地址数据库、中国互联网络信

息中心 CNNIC IP 地址数据库、淘宝 IP 地址数据库、中国行政区域经纬度信息表等多个网络数据源的方法，进行 IP 基本信息的综合推理，构建了用户信息数据库。该信息数据库包括 IP 地址、所属国家、所属省市、经度、纬度、服务(包括数据中心、CDN、云服务、教育单位、互联网公司等)。此外，本文还搜集了各大互联网公司的爬虫 IP，构建了爬虫网段数据集，用于辅助网络用户角色的识别。表 1 列出了辅助数据库的部分数据。

表 1 辅助数据库的部分数据

Table 1 Partial data of auxiliary database

	属性	取值
用户信息数据集	IP 地址	121.69.50.22
	所属国家	中国
	所属省市	北京
	经度	116.40
	纬度	39.90
	服务	无
爬虫网段数据集	起始地址	123.125.71.1
	终止地址	123.125.71.255
	服务	Google 爬虫

3.2 基于滑动时间窗的用户角色动态辨识模型

利用 3.1 节构建的 IP 地址辅助数据库得到用户的基础信息，结合用户行为特征，以天为时间单位，计算网络用户的日角色值。所谓日角色值是指在某天 IP 用户为 1(机器人)或 0(人类)。统计分析一段时间内用户的日角色值情况，引入滑动时间窗口，最终得到非人类访问行为概率，以辨识网络用户的访问角色。

3.2.1 用户 IP 地址日角色评估模型

设用户的 IP 地址为 x，若 x 在第 t 天发生访问行为，则可得到用户 IP 地址 x 在第 t 天的日角色值，用 $DC_t(x)$ 表示，其取值范围为 0(人类)或 1(机器人)，如式(1)所示：

$$DC_t(x)=\begin{cases}0, & \text{其他}\\ 1, & x\in A\,|\,x\in B\,|\max\limits_{\text{hour}}(x)\geqslant\alpha\,|\,act_t(x)\geqslant\tau\end{cases}\tag{1}$$

其中，集合 A 代表辅助数据集中的爬虫网段库，集合 B 代表辅助数据集中的用户信息数据集的服务类字段，$\max\limits_{\text{hour}}(x)$ 表示 IP 地址 x 的最大小时访问量，$act_t(x)$ 表示 IP 地址 x 在第 t 天的活跃小时数。参数 α 表示用户访问的次数，τ 表示用户活跃的时长。经过对 20 人点击鼠标访问网页速度的实验调查发现，每个人在一个小时内点击鼠标访问网页的次数不会超过 3600 次，而其坚持坐在电脑前操作电脑的持续时间也不可能超过 20 h，而机器访问却具有高频性和持久性的特点。因此，如果访问的次数或时长超过这个数值，那么访问者一定是机器。为此，将参数 α 设置为 3600(次)，τ 设置为 20(小时)，如果访问频次超过 α 或访问时长超过 τ，则判定该用户的日角色值为机器，这里 α 和 τ 的值可根据具体项目的需要进行调整。用户 IP 地址的日角色评估的计算过程如算法 1 所示。

算法 1 用户 IP 地址日角色值算法

输入：IP 地址 x，时间 t，日志数据集，辅助数据集

输出：用户 IP 地址 x 在第 t 天的日角色值

Step 1 对 IP 地址 x 与辅助数据集中的爬虫网段库进行对比，若 x 在爬虫网段库中，则 $DC_t(x)$ 置 1；否则 $DC_t(x)$ 置 0，结束。

Step 2 对 IP 地址 x 与辅助数据集中的用户信息数据集的服务类字段(此类 IP 发出的网页浏览行为在大多数情况下属于 NHT 行为)进行对比，若在服务类字段不为空，则 $DC_t(x)$ 置 1；否则 $DC_t(x)$ 置 0，结束。

Step 3 对于 IP 地址 x 在第 t 天的日志数据集，计算其每小时的访问量，若最大小时访问量超过 α 次(机器用户角色的高频性)，则 $DC_t(x)$ 置 1；否则 $DC_t(x)$ 置 0，结束。

Step 4 对于 IP 地址 x 在第 t 天的日志数据集，若其当日活跃小时数超过 τ 个小时(机器用户角色的持久性)，则 $DC_t(x)$ 置 1；否则 $DC_t(x)$ 置 0，结束。

3.2.2 引入滑动时间窗的用户角色动态辨识模型

通过一段时间对 IP 地址日角色值的监测统计，可判断出该 IP 用户的角色值。由于考虑到 IP 用户访问的动态性，不能仅靠一天或几天的数据来判断它的角色，为此引入滑动时间窗口，该窗口是一个窗口大小可变的、动态滑动的时间段，IP 用户的角色值可依赖于该用户近 T 个时间窗内的日角色值，越接近当前窗口其日角色值的权重越高，随着时间的推移，用户角色值也会发生动态变化，因此其能更精准地识别该用户的角色。若以 $UC_t(x)$ 表示某用户 x 在第 t 天的角色值，则 $UC_t(x)$ 如式(2)所示：

$$UC_t(x)=\frac{1}{2}DC_t(x)+\frac{1}{2^2}DC_{t-1}(x)+\cdots+\frac{1}{2^T}DC_{t-T}(x)\tag{2}$$

其中，t 为当前日期；T 为滑动时间窗的大小；$DC_i(x)$ 为 x 在第 i 天的日角色值，其值域为[0,1]。

由于只考虑人类和机器两类用户，因此用户角色值只有两种，根据概率统计经验，将 $UC_t(x)$ 的值域分为两段，以 0.5 为界，当 $UC_t(x)$ 大于或等于 0.5 时，可认定此 IP 地址的访问行为由人类主动行为产生，分值愈高越真实；当 $UC_t(x)$ 小于 0.5 时，则认为此 IP 地址的网页访问行为有较高可能性是由机器人主导。因此构建用户角色 $Role_t(x)$ 的辨识模型，如式(3)所示：

$$Role_t(x)=\begin{cases}\text{人类}, & UC_t(x)\geqslant 0.5\\ \text{机器人}, & UC_t(x)<0.5\end{cases}\tag{3}$$

4 基于多特征的用户恶意访问行为的识别模型

首先根据用户访问角色分类集对网络用户角色进行识别，在此基础上分别对真人和机器角色产生的恶意流量进行识别。基于用户产生的网络流量特征和信息熵的统计特性，构建用户恶意访问行为识别模型，用于识别由恶意的用户或程序产生的流量。

4.1 网络用户恶意流量的特征分析

网络用户产生恶意流量的目的大多是通过消耗主机资源导致系统拒绝正常用户服务，或者占用大量宽带资源使得网络堵塞。从网络流量的角度来分析，恶意流量的产生会使网络流量在短时间内发生突增现象，并且在短时间内不会削减，趋于一种高峰稳定状态。通过用户访问流量特征能够对爆发性和高持续性的恶意访问行为进行识别，而通过用户信息熵特征能够对少量但大规模分散访问的恶意行为进行辨识。因此，对用户访问流量特征和用户信息熵特征进行加权，构建基于多特征的用户恶意访问行为的辨识模型。

4.1.1　网络用户流量特征

设用户 IP 地址为 x,t 为当前日期,引入滑动时间窗口 T,得到用户 IP 地址流量特征值 $f_t(x)$。$f_t(x)$的计算式如式(4)所示:

$$f_t(x)=\max\{df_t(x,m),cf_t(x,m,T)\} \tag{4}$$

其中,$f_t(x)$的取值为 0 或 1,若 $f_t(x)$的取值为 1,则认定该用户 IP 地址 x 产生恶意流量,否则反之;$df_t(x,m)$为用户 IP 在当前日期 t 下的流量函数,用于衡量用户 IP 地址在当日的流量是否异常。$df_t(x,m)$的计算式如式(5)所示:

$$df_t(x,m)=\begin{cases}0, & C_t(x)<C_{\text{avg}}(t)+m\times C_{sd}(t)\\ 1, & C_t(x)\geqslant C_{\text{avg}}(t)+m\times C_{sd}(t)\end{cases} \tag{5}$$

其中,$C_t(x)$为用户 IP 地址 x 在当前日期 t 下的流量数,$C_{\text{avg}}(t)$为当前日期 t 的全部用户的流量平均值,$C_{sd}(t)$为当前日期 t 的流量标准差,m 为高爆发性流量异常参数。流量函数可以识别出日爆发性的恶意流量,但对于高量持续性的非爆发流量则无法识别,$cf_t(x,n)$为用户 IP 地址在一段时间 T 内的持续流量函数,用于检测持续高频访问的用户 IP 地址,$cf_t(x,n)$的计算式如式(6)所示:

$$cf_t(x,n,T)=\begin{cases}0, & \sum_{i=t}^{t-T}df_t(x,n)<T\\ 1, & \sum_{i=t}^{t-T}df_t(x,n)\geqslant T\end{cases} \tag{6}$$

其中,n 为高持续性流量异常参数,T 为时间窗口大小。参数 m 和 n 的选取决定着用户 IP 地址的流量特征值何时为异常。6.2 节会通过真实的网络流量进行参数选取,以更好地适应网络恶意流量行为的判定。

4.1.2　网络用户的信息熵特征

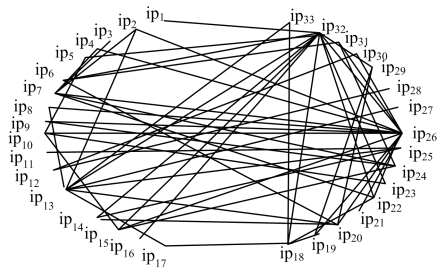
流量特征值以用户 IP 地址为分析对象而得到,识别出"单"用户的多恶意流量。若仅从用户 IP 的角度出发无法识别出"多对一"的恶意流量,即少量但大规模分散访问的恶意行为,因此引入网络信息熵的概念来针对群体恶意攻击行为进行识别。信息熵的物理意义是体系混乱程度的度量。通过计算单位时间内网络报文目的 IP 的熵值,来度量被访问地址的离散程度,以达到检测恶意流量的目的。

设用户 IP 地址为 x,t 为日期,集合 $D=\{mdip_1,mdip_2,\cdots,mdip_n\}$为第 t 天用户 IP 地址 x 访问的所有目的 IP 地址集合。集合 $P=\{p_1,p_2,\cdots,p_n\}$为集合 D 中目的 IP 地址在第 t 天所处网络的概率分布,p_i 表示目标 IP 地址 i 在当日流量中的占比,则可得到用户 IP 地址对应的所有目的 IP 地址的熵特征值,以 $e_t(D)$表示,如式(7)所示:

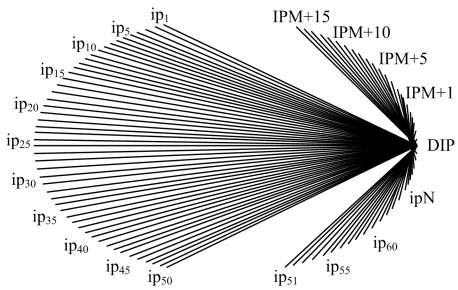
$$e_t(D)=-\sum_{i=1}^{n}p_i\ln p_i \tag{7}$$

恶意流量的产生会导致网络日志中被访问的目的地址比较集中,此时网络关系中目的地址 IP 的信息熵与正常网络流量的熵相比会相对减少,如图 2 所示。图 2(a)是正常网络流量的关系图,图 2(b)是发生"多对一"这种大规模分散访问的网络关系图。根据式(5)可知,恶意网络流量的熵远远小于正常网络流量的熵,因此将信息熵作为判断 IP 恶意流量的标准之一。同样地,利用类似流量函数的判断方法构建熵函数 $ef_t(x,m')$(见式(6)),并将该函数用于衡量网络流量的熵是否异常。式(6)中的参数 m'为信息熵异常参数,其选取方法将在 6.2 节中讨论。

$$ef_t(x,m')=\begin{cases}0, & e_t(D)\geqslant e_{\text{avg}}(D)-m'\times e_{sd}(D)\\ 1, & e_t(D)<e_{\text{avg}}(D)-m'\times e_{sd}(D)\end{cases} \tag{6}$$



(a)正常网络流量关系图



(b)发生"多对一"大规模分散访问的网络关系图

图 2　正常网络关系与恶意流量网络关系图

Fig. 2　Normal network relationship and malicious traffic network relationship

4.2　基于多特征的网络用户恶意访问行为辨识模型

综合用户产生的网络流量特征和用户信息熵特性,对爆发性和高持续性的恶意访问行为进行识别。将 IP 用户的流量特征与用户的信息熵特征进行加权,构建网络用户恶意访问行为的判定模型如下。

设用户 IP 地址为 x,t 为当前日期,计算用户 IP 地址 x 的恶意值,以 $MA_t(x)$表示,则可判断用户 IP 地址 x 是否产生恶意流量。鉴于 IP 用户流量可以反映其访问的高爆发性和高持续性,但对少量但大规模分散访问的恶意行为缺乏识别能力,而用户网络信息熵对大规模分散访问具有较好的描述性。因此,两个特征对于恶意访问的描述具有同样的效力,视其在评估恶意流量时具有同样的权重。构建用户恶意访问评价函数 $MA_t(x)$,如式(8)所示:

$$MA_t(x)=\frac{1}{2}f_t(x)+\frac{1}{2}ef_t(x,m') \tag{8}$$

其中,为了将用户恶意值归一化,将各特征的权重设为 0.5。$MA_t(x)$的取值范围为 0～1,表示用户 IP 地址的恶意程度,取值大于 0 时表示用户发生恶意访问,需要重点监测和追踪。式(8)综合了网络用户行为特性和 IP 所在网络熵的特性,既避免了单独分析用户行为特性造成的误报,同时又考虑了单独分析源 IP 可能造成的"多对一"大流量访问异常情况的漏报。

5　上述模型实现的大数据处理存储及计算技术

由于本文原始数据量(日均 3.7 亿条,共计 116 亿条)较为庞大,因此给以周计或者更长时间跨度的挖掘带来了一些挑战。针对海量数据,对大数据进行分析处理的前提是以适当的方式对大数据进行存储,利用传统的关系型数据库进行

管理和挖掘几乎不可能。目前,大数据的存储主要有分布式文件系统、非关系型数据库、分布式数据仓库等多种形态。由于本文所处理的数据源为网络日志,且不存在基于行级的数据更新操作,因此采用 Hive 数据仓库对数据进行存储。其实质是将数据存储在分布式文件系统中,将数据库模式(Schema)存储在关系型数据库中,使用类似于 SQL 语言的查询语言(如 HQL)对数据进行简单的检索和处理。

接着对存储在数据仓库中的静态数据进行集中计算,由于本文研究内容对实时性的要求不高,但对数据的准确性、全面性更为看重,因此采用批量计算技术进行分析和挖掘。并采用伯克利大学开源的 Hadoop MapReduce 通用的并行计算框架 Spark 内存计算技术进行处理和计算,Spark 由于能够将计算的中间结果存储到内存中,减少了磁盘的 I/O 开销,因此能够更好地应用于需要多次迭代的算法中。

6 实验分析与结果

6.1 实验环境与数据源

实验环境是由 46 台实验主机构成的集群,使用的网络日志数据包括服务器上有关 Web 访问的访问日志、引用日志、代理日志、错误日志等文件。数据源来自于国内某科研单位的网络日志,日志按照时间正序记录,记为 4 个数据集 D1—D4,分别为 2017 年 01 月 01 日起一天、一周、半个月、一个月的日志数据。

本文对 Web 日志进行一般预处理,预处理后的日志中包括用户 IP 地址、服务器 IP 地址、应用协议(HTTP,HTTPS,SSL 等)、所访问的 URL、传输协议、访问日期和时间、访问方法(GET 或 POST)。为了能够更直观地了解用户访问域名所承载的内容,本文参考 wiki 的近 400 个域名分类建立了域名分类体系,共计 4 个大类 43 小类,并按照所构建的分类分级体系对网络日志中的重点域名进行了手工标注,构建了域名信息数据集。构建域名信息数据集实质上是给域名打上分类的属性标签,以便于了解域名背后所承载的功能,其中包括域名、信息传播者中承担的角色、媒体类型、网站对应的组织机构、承载服务。表 2 与表 3 分别列出了经过数据预处理后的网络日志格式及域名信息数据集。

表 2 数据预处理后的网络日志

Table 2 Web log after data preprocessing

用户 IP	203.100.175.180
服务器 IP	211.82.96.1
应用协议	DNS
URL	cpsc.gov
传输协议	UDP
事件	2017-01-25 09:42:52
方法	GET

表 3 数据预处理后的域名信息

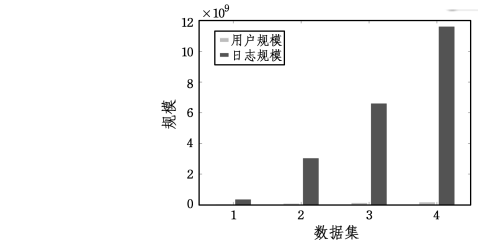
Table 3 Domain information after data preprocessing

域名	google.com	youtube.com	fc2.com
信息传播中承担的角色	信息中转平台	信息交互平台	信息发布平台
媒体类型	文本	视频	综合
网站对应的组织机构	企业	企业	企业
承载服务	搜索引擎	非实时社交平台	综合门户、音视频

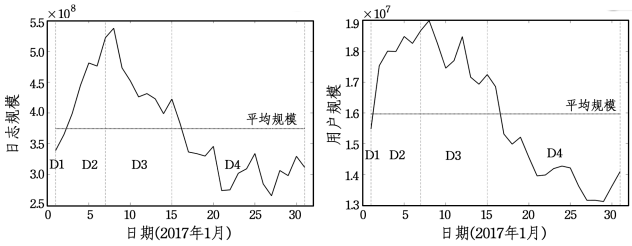
6.2 实验结果

6.2.1 用户访问角色判定及其分布实验

对日志规模和用户规模进行统计,图 3 给出了各数据集日志及用户规模图。



(a)各数据集中的用户规模



(b)D1—D4 中时间与日志规模的关系 (c)D1—D4 中时间与用户规模的关系

图 3 各数据集日志及用户规模图

Fig. 3 Number of dataset logs and users

由图 3(a)可知,数据集 D1—D4 的日志规模增长实际上隐含了随着时间的推移日志规模的不断累计,而用户规模趋于稳定的一种特征。总用户规模约 1480 万,总日志规模约 116 亿条。图 3(b)描述了每日的日志规模情况,日均 3.75 亿条。图 3(c)描述了每日的用户规模情况,日均活跃用户 1600 万。可以发现,无论是用户还是日志,在规模上后半个月相比前半个月都有明显的减少。这可能是因为数据集的选取时间是 2017 年 1 月,临近过年时期,用户较少导致,对本文的研究内容并无影响。

对数据集 D1—D4 分别绘制了用户角色分类情况图,4个数据集的滑动时间窗口取值为其各自日志时间的跨度,如图 4—图 6 所示。图 4 分别给出了 D1—D4 中用户二分类角色所占的比例关系;图 5 分别给出了 D1—D4 中具有 NHT 概率角色的用户分布情况;图 6 分别给出了 D1—D4 中各类角色访问产生的日志数量。

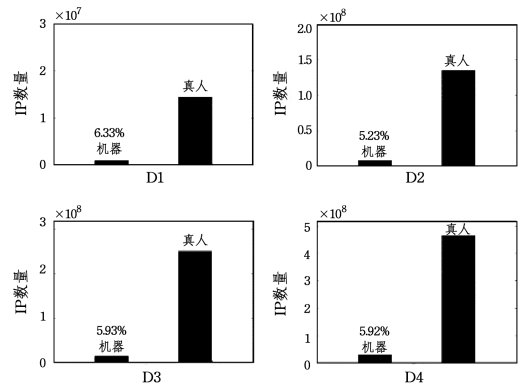


图 4 各数据集中用户二分类角色分布图

Fig. 4 Binary role distribution map of users of each data set

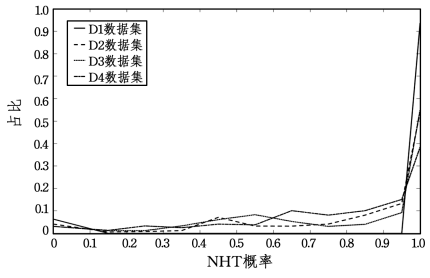


图 5　各数据集中 NHT 概率角色的用户分布情况图

Fig. 5　User's role distribution of NHT probability in different data sets

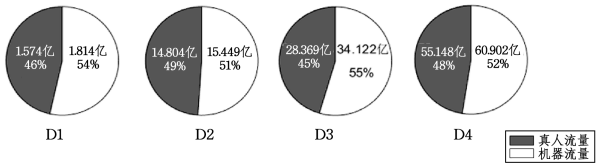


图 6　各数据集中不同角色的访问流量

Fig. 6　Access traffic of different role in different data sets

在引入时间窗口后,随着窗口的扩大,用户非人类访问行为的概率从两极化逐渐趋于平稳化。将用户角色值概率化后,每一个数据集中人类主导的流量与机器主导的流量比例约为 1∶1,也就是说网络日志流量中有一半的流量是由机器人主导产生的。

6.2.2　用户恶意访问流量的识别及参数确定

在对 IP 流量异常进行判断时,流量异常参数 m 和 n 的选取决定着用户 IP 地址的流量特征值何时为异常。取值过大时会导致异常流量用户的漏报率增高,取值过小时误报率会增高。将全量数据集(720 h)作为实验数据进行流量统计,以确定参数 m 和参数 n。首先将数据集以小时为单位进行切割,对流量进行随机采样,并统计其平均流量和标准差。流量变化情况及其平均流量如图 7 所示。

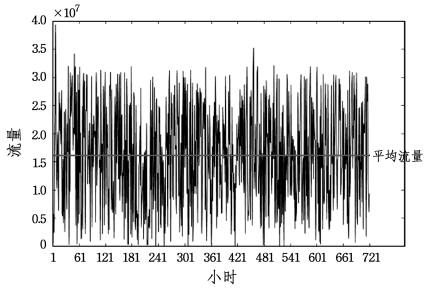


图 7　流量情况

Fig. 7　Flow distribution

采用 $C(h)$表示全量数据集中第 h 小时的流量。将小时平均流量表示为 C_{avg},小时流量标准差表示为 C_{sd},然后计算在每一个小时内的流量是平均值与多少倍的标准差之和,将其倍数函数定义为 $f(h)$,其分布情况如图 8(a)所示,计算方法如式(9)所示。从 $f(h)$的分布情况来看,曲线在点(0.967,3.1)处的斜率发生突变,意味着在 720 h 的全量数据集上,有 96.7%的时间,其瞬时流量都小于平均值加 3.1 倍的标准差,除此之外,3.3%的时间瞬时流量都可判定为异常流量。因此,设参数 m 为 3.1,认为当 IP 的流量超过用户平均值加上 3.1 倍的标准差时,其访问行为是异常的;同样地,对参数 n 的选取采用相似的方法,最终将参数 n 选取为 1.3。本文认为对于那些产生高量持续性的非爆发流量 IP,其持续流量超过用户平均值加上 1.3 倍的标准差时,其访问行为是异常的。

$$f(h)=\frac{C(h)-C_{avg}}{C_{sd}} \tag{9}$$

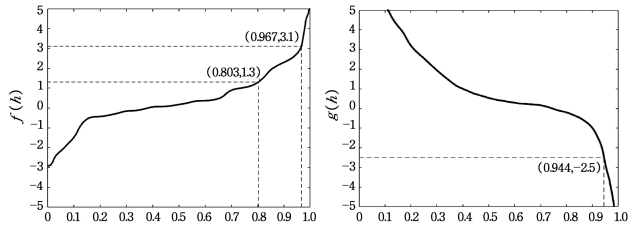


图 8　倍数函数 $f(h)$与 $g(h)$

Fig. 8　Multiplier function $f(h)$ and $g(h)$

网络信息熵异常参数 m'的选取决定着网络的熵何时低于正常值,采用与流量异常参数 m 相同的选取方式,构建熵的分布函数,如图 8(b)所示,分布函数 $g(h)$的结果显示有 96.2%的用户所在网络流量高于平均值与 2.5 倍的标准差之和,除此之外,4.8%的网络流量判定为恶意流量。

为了衡量检测效果,采用 MITLincoln 实验室开发的 DAPAR1999 第二周的入侵检测数据集来验证本文识别模型的有效性。采取上述确定的用户恶意访问流量的识别参数进行检测仿真,共计识别出 38 次恶意访问行为。DAPAR1999 第二周的数据集是在第一周的正常数据集中插入了属于 18 种类型的 43 次攻击实例。其中正确识别 37 次,误报 1 次,漏报 6 次,由此可见本文模型在检测恶意流量时具有较高的正确率,但由于数据集中可能存在各种未知的攻击类型,会导致漏报率仍然不是很理想,因此对未知的攻击类别的识别需要更深入的研究和优化。

结束语　本文对网络日志数据进行挖掘,首先从用户角度出发,对其角色进行识别,将分析得到的非人类访问行为概率作为该用户的访问角色。然后从流量角度出发,对恶意流量进行识别,包括爆发性恶意流量识别和高持续性恶意流量识别。综合了网络用户行为特性和 IP 所在网络信息熵的特性,既避免了单独分析用户行为特性造成的误报,同时又考虑了单独分析源 IP 可能造成的"多对一"大流量访问异常情况的漏报。在计算技术上采用了类似 MapReduce 的通用并行框架 Spark,使计算效率得到大幅提高。从实验结果来看,在网络流量产生异常时,所提模型能够发现导致异常的用户,并准确且高效地辨别出该用户的角色。

参 考 文 献

[1]　KEMMAR A,LEBBAH Y,LOUDNI S. A Constraint Programming Approach for Web Log Mining[J]. International Journal of Information Technology and Web Engineering (IJITWE),2016,11(4):24-42.

[2]　SISODIA D S,VERMA S,VYAS O P. Agglomerative Approach for Identification and Elimination of Web Robots from Web Server Logs to Extract Knowledge about Actual Visitors[J]. Journal of Data Analysis and Information Processing,2015,3(1):1-10.

(下转第 201 页)

[13] WANG X，HUA G，HAN T. Discriminative tracking by metric learning[C]// Computer Vision-ECCV 2010. 2010：200-214.

[14] KHOSHNESHIN M，STREET W N. Collaborative filtering via euclidean embedding[C]// Proceedings of the Fourth ACM Conference on Recommender Systems. ACM，2010：87-94.

[15] GOLDBERGER J，HINTON G E，ROWEIS S T，et al. Neighbourhood components analysis[C]// International Confernece on Neural Information Processing Systems. MIT Press，2005：513-520.

[16] WEINBERGER K Q，SAUL L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research，2009，10(2)：207-244.

[17] HSIEH C K，YANG L，CUI Y，et al. Collaborative metric learning[C]// Proceedings of the 26th International Conference on World Wide Web. 2017：193-201.

[18] KOREN Y，BELL R，VOLINSKY C. Matrix Factorization Techniques for Recommender Systems[J]. Computer，2009，42(8)：30-37.

[19] KOREN Y. Collaborative filtering with temporal dynamics[J]. Communications of the ACM，2010，53(4)：89-97.

[20] SHEN Y Y，YAN Y，WANG H Z. Recent Advances on Supervised Distance Metric Learning Algorithms [J]. Acta Automatica Sinica，2014，40(12)：2673-2686. (in Chinese)
沈媛媛，严严，王菡子. 有监督的距离度量学习算法研究进展[J]. 自动化学报，2014，40(12)：2673-2686.

[21] ZHAO G，QIAN X，XIE X. User-service rating prediction by exploring social users' rating behaviors[J]. IEEE Transactions on Multimedia，2016，18(3)：496-506.

[22] MASSA P，AVESANI P. Trust-aware recommender systems [C]// Proceedings of the 2007 ACM Conference on Recommender Systems. ACM，2007：17-24.

[23] SALAKHUTDINOV R，MNIH A. Probabilistic Matrix Factorization [C] // International Conference on Neural Information Processing Systems. Curran Associates Inc. ，2007：1257-1264.

[24] MA H，KING I，LYU M R. Learning to recommend with social trust ensemble[C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM，2009：203-210.

(上接第 165 页)

[3] JOSHILA GRACE L K，MAHESWARI V，NAGAMALAI D. Analysis of Web Logs And Web User In Web Mining[J]. International Journal of Network Security & Its Applications，2011，3(1)：99-110.

[4] XU X F，YANG L，WANG W. Novel role analysis method for network domain users[J]. Chinese Journal of Network and Information Security，2017，3(3)：22-27. (in Chinese)
许小丰，杨力，王巍. 新颖的网络域名用户关键角色识别方法[J]. 网络与信息安全学报，2017，3(3)：22-27.

[5] CHEN M S，PARK J S，YU P S. Efficient data mining for path traversal patterns[J]. IEEE Transactions on Knowledge and Data Engineering，1998，10(2)：209-221.

[6] XU J J，CHEN H. CrimeNet explorer：a framework for criminal network knowledge discovery[J]. ACM Transactions on Information Systems (TOIS)，2005，23(2)：201-226.

[7] GUO Y，BAI S，YANG Z F，et al. Analyzing Scale of Web Logs and Mining Users' Interests [J]. Chinese Journal of Computers，2005，28(9)：1483-1496. (in Chinese)
郭岩，白硕，杨志峰，等. 网络日志规模分析和用户兴趣挖掘[J]. 计算机学报，2005，28(9)：1483-1496.

[8] XING D S，SHEN J Y，SONG Q B. Discovering Preferred Browsing Paths from Web Logs [J]. Chinese Journal of Computers，2003，26(11)：1518-1523. (in Chinese)
邢东山，沈钧毅，宋擒豹. 从 Web 日志中挖掘用户浏览偏爱路径[J]. 计算机学报，2003，26(11)：1518-1523.

[9] JIN X. Web Log Mining Based-on Improved Double-Points Crossover Genetic Algorithm[J]. Journal of Multimedia，2014，9(6)：804-809. (in Chinese)

[10] YANG J G，WANG X T，LIU G Q. DDoS attack detection method based on network traffic and IP entropy[J]. Application Research of Computers，2016，33(4)：1145-1149. (in Chinese)
杨君刚，王新桐，刘故箐. 基于流量和 IP 熵特性的 DDoS 攻击检测方法[J]. 计算机应用研究，2016，33(4)：1145-1149.

[11] SAIED A，OVERILL R E，RADZIK T. Detection of known and unknown DDoS attacks using Artificial Neural Networks[J]. Neurocomputing，2016，172(C)：385-393.

[12] LEUNG K，LECKIE C. Unsupervised anomaly detection in network intrusion detection using clusters [C] // Proceedings of Australasian Computer Science Conference. Australia，2005. 333-342.

[13] RUBINSTEIN B，NELSON B，HUANG L，et al. Stealthy poisoning attacks on PCA-based anomaly detectors[J]. Acm Sigmetrics Performance Evaluation Review，2009，37(2)：73-74.

[14] LI Q，CHI L J，ZHANG Z X. A Novel Approach to Simulate DDoS Attack[J]. International Journal of Wireless and Microwave Technologies(IJWMT)，2011，1(2)：33-40.

[15] SUN Z X，LI Q D. Defending DDos Attacks Based on the Source and Destination IP Address Database [J]. Journal of Software，2007，18(10)：2613-2623. (in Chinese)
孙知信，李清东. 基于源目的 IP 地址对数据库的防范 DDos 攻击策略[J]. 软件学报，2007，18(10)：2613-2623.

[16] GUI B X，ZHOU K，ZHOU W L. An IP Traceback Model Based Traffic Entropy Variations for DDoS Attacks [J]. Journal of Chinese Computer Systems，2013，34(7)：1607-1609. (in Chinese)
桂兵祥，周康，周万雷. 通信流熵变量 DDoS 攻击 IP 回溯跟踪模型[J]. 小型微型计算机系统，2013，34(7)：1607-1609.

[17] LI Q，SHEN T，GUAN Y. Research on Clustering Algorithm for Large Data Sets[J]. Intelligent Computer and Applications，2012，2(5)：42-45. (in Chinese)
李清，沈彤，关毅. 面向大规模日志数据的聚类算法研究[J]. 智能计算机与应用，2012，2(5)：42-45.

[18] ZHAO L. The Design and Implementation of Massive Search Logs Analysis Platform Based on Hadoop[D]. Dalian：Dalian University of Technology，2013. (in Chinese)
赵龙. 基于 Hadoop 的海量搜索日志分析平台的设计和实现[D]. 大连：大连理工大学，2013.