

一种综合多因素的网页浏览行为认证方法

陈冬祥 丁志军 闫春钢 王咪咪

(同济大学电子与信息工程学院 上海 208401)

(同济大学嵌入式系统与服务计算教育部重点实验室 上海 200092)

摘 要 在电子交易中,用户通过 PC 端的浏览器进行交易。由于钓鱼网站等盗号方式的威胁,传统的账号密码认证方式存在着失效的风险。现有的用户网页浏览行为认证方法主要针对用户的某一方面行为进行认证。若对大量用户仅进行单方面行为的认证,则难以区分特征相似用户,会造成认证失效。基于用户浏览网页的序列行为、超链接使用行为和操作浏览器行为的多因素浏览行为特征,采用机器学习方法构建了一种认证方法。实验结果表明,在一定的误报率情况下,该方法的侦测率达到了 90% 以上。

关键词 网页浏览,行为认证,多因素,机器学习

中图法分类号 TP391.9 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.02.032

Authentication Method Synthesizing Multi-factors for Web Browsing Behavior

CHEN Dong-xiang DING Zhi-jun YAN Chun-gang WANG Mi-mi

(Department of Electronics and Information Engineering, Tongji University, Shanghai 208401, China)

(The Key Laboratory of Embedded System and Service Computing(Tongji University), Ministry of Education, Shanghai 200092, China)

Abstract In the process of electronic trades, users trade through the PC browser. Due to the threats of phishing sites and other hacking way, there is a failure risk in the traditional account-password-authentication mode. The existing methods of user Web-browsing-authentication mainly aim at authenticating one aspect of the user's behaviors. For a large number of users, if the authentication is done only from the single aspect, it is difficult to distinguish the features among similar users, which will result in authentication invalidation. Based on the sequence behavior of user browsing Web, the behavior of hyperlink usage and browser manipulation, a authentication method synthesizing multi-factors was proposed by using the machine learning method. The experimental results show that this method achieves more than 90% detection rate under a certain false positive rate.

Keywords Web browsing, Behavior authentication, Multi-factors, Machine learning

1 引言

根据艾瑞网的统计数据,中国 2016 年的电商交易总额达到了 20 万亿,较 2015 年增加了 23.6%。电子商务交易是经济发展的重要成分^[1]。电子商务交易的基础是电子商务交易系统,电子商务交易系统是计算机技术的整合应用,其最重要的部分是支付安全。随着互联网技术的发展,电子支付面临的风险开始变得多种多样,如钓鱼网站、病毒等威胁着用户账号的安全。根据《360 2016 年互联网安全报告》显示,2016 年度新增的钓鱼网站高达 196.9 万个,大量的钓鱼网站会导致用户的账号被窃取^[2]。现有的认证技术^[3-5]主要利用账号密码进行一次性认证。一旦账号被盗用,非法用户的一切行为都会被认为是合法的,从而威胁用户的账户和资金安全。基于生物信息的认证方式是一种解决手段,但是需要额外的硬件支持^[6]。

用户的行为认证可以在传统的账号密码认证失效的情况下依旧为用户提供保障,同时又可以避免生物信息认证所需的额外的硬件支持。因此,行为认证比生物信息认证更适合作为传统的账号密码认证的弥补手段。行为认证分为一次性行为认证和持续性行为认证。一次性行为认证需要在应用程序上加上特定的认证环节,在该环节通过后使用期间的行为则不再被认证;而持续性行为认证可以确保账号的整个在线期间的安全性都可以得到保障,本文所述的浏览行为认证是一种持续性行为认证。通过分析用户持续的浏览行为,我们可以构建合法用户的行为模型,并基于此进行行为认证。在 PC 端,用户使用浏览器进行电子交易,其行为个性化反映在浏览行为中,而且不需要引入额外的软硬件采集浏览行为的数据。因此,本文通过挖掘用户的浏览行为来构建用户的行为模型。

由于本文提出的行为认证方法是一种仅需获取用户行为

到稿日期:2017-07-31 返修日期:2017-11-07 本文受上海市“科技创新行动计划”高新技术领域项目(16511100900),国家自然科学基金委重点项目(61332008)资助。

陈冬祥(1993—),男,硕士生,主要研究方向为可信计算, E-mail: 93Nicolas_chen@tongji.edu.cn; **丁志军**(1972—),男,教授,博士生导师,主要研究方向为服务计算; **闫春钢**(1963—),女,教授,博士生导师,主要研究方向为计算机协同与服务计算, E-mail: yanchungang@tongji.edu.cn (通信作者); **王咪咪**(1985—),女,博士生,主要研究方向为可信计算。

数据的隐式认证方式,因此其在实际应用中能够与口令卡、智能卡等双因子认证方案^[7]兼容。在账号注册初期,由于缺少足够的用户行为数据,因此可以利用口令卡、智能卡等为用户提供额外保障。但是口令卡等会带来额外的输入,影响用户体验,因此当用户的历史行为数据达到足够数量时可以停用口令卡、智能卡等认证,启用本文所述行为认证方法,从而在改善用户体验基础上维持甚至提高安全性。

现在已存在一些浏览行为认证成果。文献[8]提出了一种基于马尔科夫模型的个人访问模式,用于计算用户的可信度,并以此来进行行为认证,但是该方法的侦测率有限。文献[9]提出了一种基于关联规则挖掘(ARM)与 SimpleCart 分类器的行为认证方式,利用用户浏览过程中频繁访问站点的关联性建立行为模型并将其用于认证。但是,网络用户数量远多于主流网站数量,从而存在大量频繁访问站点高度重复的用户,因此该方法存在明显的风险,即无法侦测此群体内的非法用户。文献[10]提出了一种基于页面的浏览时间分布特征的方法,利用 SVM 分类器进行模型的建立与异常行为检测。但是页面内容的多少、环境干扰等常见的因素容易导致时间分布特征的不稳定,其侦测率同样有限。此外,上述成果并没有考虑用户在浏览网页时的操作行为的个性化差异。因此,

本文参考了部分已有的基于用户操作行为的成果。移动端基于用户触屏行为的行为认证^[11],PC 端基于用户键盘鼠标行为的行为认证^[12],这些成果反映出用户操作存在个性化的差异,但是并没有结合一个真实的应用场景,也没有考虑用户的操作行为受其所浏览的内容影响。由于浏览过程往往会持续一定的时间,同时,用户的行为又具有阵发性和长时休眠的特性^[13],在浏览的过程中往往会伴随着用户的其他行为,为了避免其他行为可能带来的干扰,因此将本文所提方案所用的特征用于刻画浏览行为除时间分布之外的个性化特征。

本文提出了一种改进的浏览行为认证方法,该方法不同于之前单纯地利用浏览日志数据进行特征提取以及行为认证。本文基于用户在使用浏览器进行浏览的过程中的页面访问记录和超链接使用记录和浏览器操作记录,不仅关注了用户对页面内容的偏好,而且关注了用户如何浏览站点,以及在浏览不同类型页面时对浏览器进行操作的行为习惯。使用多个子模型可以避免单一模型的易模仿以及在某些情况下失效的潜在风险,提高了侦测率。

本文提出的方法包括行为模型构建和非法检测两个阶段,如图 1 所示。

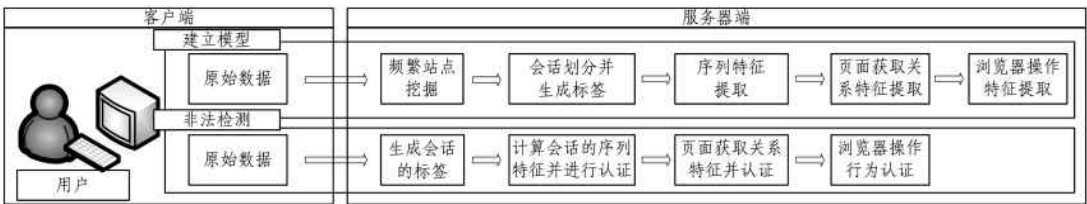


图 1 整体过程示意图

Fig. 1 Schematic diagram of whole process

模型建立阶段:1)依照会话的定义将用户浏览页面的序列进行划分,获取整个会话中所有的站点以构成的集合;2)依照所有历史记录的会话,提取出频繁访问的站点的集合,并提取出各个频繁访问站点的频繁访问板块(依据二级域名)的集合;3)将每个会话对应集合中所包含的所有频繁访问站点构成的集合视为该会话的标志;4)依据会话的标志,为标志相同的会话构建序列行为模型;5)建立各个频繁访问站点的页面获取关系(使用超链接跳转的行为)行为模型;6)建立各个频繁访问站点的操作行为模型。

非法检测阶段:1)根据历史数据划分会话;2)生成该会话的标志,检测标志的合法性;3)依据会话标志,进行序列行为模型认证;4)选择一个频繁访问站点,对其进行页面获取关系模型认证;5)利用上一步中相同的频繁访问站点进行操作行为模型认证;6)返回最终的认证结果。

2 数据采集

本文提出的浏览行为认证方法需要捕捉用户的以下几种数据:页面浏览序列、在页面上的超链接点击记录和在浏览页面时对浏览器进行操作的操作行为记录。我们基于开源库 QWebEngine 开发了一款浏览器,并布置了事件监听代码,用于捕捉用户的行为数据。数据存储格式如图 2 所示。

```
<pageview page_type="New">
  <url>https://123.sogou.com/</url>
  <classification>1473244572538</classification>
  <time>1473244572538</time>
  <operations>
    <pos type="press" time="1473244579294">224</pos>
    <textselect time="1473244578258">80</textselect>
    <pos type="release" time="1473244579390">290</pos>
    <pos type="press" time="1473244579594">149</pos>
    <pos type="release" time="1473244579650">149</pos>
    <link classification="INDEX">http://www.sina.com.cn</link>
    <wheel time="1473244584158">120</wheel>
    <wheel time="1473244584174">120</wheel>
  </operations>
</pageview>
```

图 2 数据存储格式

Fig. 2 Data storage format

<pageview>元素存储一次页面浏览所记录的数据。
<url>元素存储此页面的 URL。
<classification>元素存储对该页面的划分,INDEX 表示这个页面是一个导航页面,CONTENT 表示这个页面是一个内容提供页面。
<time>元素存储开始浏览这个页面时的毫秒级时间戳。

〈operations〉元素存储在浏览该页面时用户在浏览器上的所有的操作行为数据。

〈pos〉元素存储一次鼠标左键操作的数据。type 属性表示按压或释放,type 值为 press 表示按压事件,type 值为 release 表示释放;time 属性是该事件发生时的毫秒级时间戳;元素的内容是鼠标的纵坐标位置。

〈link〉元素存储在该页面上发生的超链接事件,classification 属性存储指向的页面的类型;元素的内容是超链接连接到的页面的 URL。

〈wheel〉元素表示浏览该页面时使用鼠标中轴滚动滑动条的事件,time 属性存储事件发生时的毫秒级时间戳;其存放的内容是距离,距离为负表示向上回滚,距离为正表示向下滚动。

利用上述数据存储格式,我们可以计算出用户的各种特

征,同时可以精确该特征所在页面的类型。本文所用的所有特征向量均可以从这种数据存储文件中计算得到。

3 行为模型

本文提出的综合多因素的浏览行为模型由序列行为模型、页面获取关系模型、操作行为模型 3 个子模型组成。在模型构建阶段,利用用户的历史数据来挖掘频繁访问站点集合和合法会话标志集合(历史上出现过的会话标志即为合法),并依次构建序列行为模型、页面获取关系模型和操作行为模型。在行为认证阶段,利用会话的标志、会话的序列特征、频繁访问站点的页面获取关系特征与操作性行为特征共同进行认证,综合各个子模型的结果,得到最终的认证结果。

3.1 基础概念

本文用到的相关术语如表 1 所列。

表 1 术语表
Table 1 Glossary

浏览序列元素	记录用户对一个页面进行一次浏览的数据元素。记浏览序列元素为 p ,图 2 中的〈pageview〉元素即为一个浏览序列元素
浏览序列	浏览序列元素 p 与时间戳 t 组成的二元组 (p,t) 依据 t 升序排序的序列,如: $((p_0,t_0),(p_1,t_1),\cdots,(p_n,t_n))$
会话	一个浏览序列片段若满足如下条件: 1) $((\cdots,(p_{k+1},t_{k+1}),(p_{k+2},t_{k+2}),\cdots,(p_{k+m},t_{k+m}),\cdots))$ 为一个浏览序列, $k<i<k+m$,且 $t_{i+1}-t_i<T$; 2) $t_{k+1}-t_k>T,t_{k+1+m}-t_{k+m}>T,T$ 为时间间隔阈值; 则 $((p_{k+1},t_{k+1}),(p_{k+2},t_{k+2}),\cdots,(p_{k+m},t_{k+m}))$ 为一个会话
板块	URL 所对应站点的二级域名记为一个板块,如:sports.sina.com.cn 的板块为 sports
频繁访问站点	若站点 h 满足: $\frac{\text{包含站点 } h \text{ 的会话数}}{\text{会话总数}} \geq \text{阈值}$,则 h 为一个频繁访问站点
Other 站点	将所有频繁访问站点集合以外的所有站点均视为 Other 站点
频繁访问板块	对于频繁访问站点 h ,若 h 的某个板块 sec 满足: $\frac{\text{包含站点 } h \text{ 的 sec 板块的会话数}}{\text{包含站点 } h \text{ 的会话数}} \geq \text{阈值}$,则将 sec 视为 h 的一个频繁访问板块
Other 板块	对于频繁访问站点 h ,将该站点的频繁访问板块集合以外的所有板块均视为 Other 板块
会话标志	会话中包含的所有频繁访问站点所构成的集合
段	会话序列 $((\cdots,(p_{k+1},t_{k+1}),(p_{k+2},t_{k+2}),\cdots,(p_{k+m},t_{k+m}),\cdots))$ 满足: $p_{k+1},p_{k+2},\cdots,p_{k+m}$ 属于同一个频繁访问站点,或者同属于 Other 站点,并且与 p_k,p_{k+m+1} 所属的站点不同,则 $((p_{k+1},t_{k+1}),(p_{k+2},t_{k+2}),\cdots,(p_{k+m},t_{k+m}))$ 为一个段
页面获取事件	标志着用户在浏览一个页面时,使用该页面上的一个超链接打开一个新的页面
页面获取关系图	在一种有权完全有向图中节点各自代表频繁访问站点中的一类页面,有向边表示在起始节点代表页面上通过超链接打开终点节点代表页面的所有页面获取事件。每个频繁访问站点对应一张页面获取关系图。页面获取关系图的数据格式如图 3 所示
板块导航节点	代表一个频繁访问板块的导航页面,或者 Other 板块的导航页面。如:sports 板块导航节点,代表页面 sports.qq.com
板块内容节点	代表某一个频繁访问板块的所有内容页面或者 Other 板块的所有内容页面。如:对于 sina.com.cn 站点而言,该节点对应 sports 板块下的 sports.sina.com.cn/nba/10247.html 等所有提供内容的页面
向下拖拽滑动条事件	代表用户使用鼠标左键按压浏览器窗口的滑动条向下拖拽并最终释放的事件
向下滚动滑动条事件	代表用户连续使用鼠标中轴滚轮向下滚动滑动条并且捕捉到的 UI 滚轮事件时间间隔小于阈值的事件
文本选取事件	代表用户在浏览页面时,使用鼠标进行一次文本选取的事件

	_Index	_Content	sports_ Index	sports_ Content	finance_ Index	finace_ Content	other_ Index	other_ Content
_Index	0.0	0.0	0.0	0.5625	0.0	0.3125	0.0	0.0
_Content	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
sports_ Index	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
sports_ Content	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.125
finance_ Index	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
finance_ Content	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
other_ Index	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
other_ Content	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

图 3 页面获取关系图的数据格式

Fig. 3 Data formats of page access relationship graph

3.2 模型特征

构建序列行为模型所需的特征如特征 1—特征 3 所示。

特征 1(段站比) 对于会话 session,计算段占比的表达式为:

$$\frac{SegNum(session|\text{会话标志} \cup \{other \text{ 站点}\})}{|\text{会话标志} \cup \{Other \text{ 站点}\}|}$$

其中,SegNum 函数计算 session 中各个站点段的总数之和。

特征 2(平均段长度) 对于会话 session,每个站点 $h(h \in \text{会话标志} \cup \{Other \text{ 站点}\})$ 的平均段长度的计算表达式为:

$$\frac{\sum_{seg \in Segs(session)} Len(seg)[Domain(seg) == h]}{|\{seg | Domain(seg) == h\}|}$$

其中,Segs 函数返回 session 中所有段构成的集合;Len 函数返回段中的元素个数;Domain 函数返回段的所属站点(一个频繁访问站点或者 Other 站点)。

特征 3(段数量比) 对于会话 *session*, 每个站点 $h(h \in \text{会话标志} \cup \{\text{Other 站点}\})$ 的段数量比的计算表达式为:

$$\frac{|\{seg | Domain(seg) == h, seg \in Segs(session)\}|}{|Segs(session)|}$$

其中, *Domain* 函数、*Segs* 函数与特征 2 所述相同。

序列行为模型利用特征 1—特征 3 构造特征向量的方法如下。

序列行为模型特征向量: 会话中各个频繁访问站点以及 Other 站点的特征 1—特征 3 按站点的字典序排列成一维向量, 即序列行为模型的特征向量。如: [段站比, 站点 1 的平均段长度, 站点 1 的段数量比, ..., 站点 *N* 的平均段长度, 站点 *N* 的段数量比, Other 站的平均段长度, Other 站的段数量比]。

构建页面获取关系模型所需的特征如特征 4 所示。

特征 4(获取权重) 一个页面获取关系图中有向边的权值表示该有向边代表的页面获取事件发生的次数占该频繁访问站点内的所有页面获取事件总数的比例; *A* 和 *B* 均为页面获取关系图中的节点, 对于有向边 $A \rightarrow B$, 获取权重的计算表达式为:

$$\frac{|\{link | link \in Links(session, h), link.from = A, link.to = B\}|}{Links(session, h)}$$

Links 函数返回所有 *Session* 会话中 *h* 站点下的超链接事件的集合; *link* 为超链接事件, *link.from* 表示该超链接所在的页面所属节点, *link.to* 表示该超链接指向的页面所属节点。

页面获取关系模型的原始数据结构为二维数组组成的有向完全图, 如图 3 所示, 各条有向边的权值即为相应的获取权重。

页面获取关系模型利用特征 4 构造特征向量的方法如下。

页面获取关系模型特征向量: 将页面获取关系图进行一维向量化, 即将二维数组自下而上地逐行添加到上一行的末尾, 最终迭代形成一维向量, 即为页面获取关系模型的特征向量。

构建操作行为模型所需的特征如特征 5—特征 10 所示。

特征 5(向下拖拽滑动条平均距离) 所有向下拖拽滑动条事件的向下距离的平均值。

特征 6(向下拖拽滑动条平均速度) 所有向下拖拽滑动条事件的向下速度的平均值。

特征 7(向下连续滚动滑动条平均距离) 所有的向下滚动滑动条事件距离的平均值。

特征 8(向下连续滚动滑动条平均速度) 所有的向下滚动滑动条事件速度的平均值。

特征 9(文本选取的平均发生次数) 文本选取的总次数除以浏览页面的总次数。

特征 10(文本选取平均长度) 所有的文本选取事件的选取字数的均值。

操作行为模型利用特征 5—特征 10 构造特征向量的方法如下。

操作行为模型特征向量: 对于某一频繁访问站点, 它的操

作行为模型特征向量为: [频繁访问板块 1 片段, 频繁访问板块 2 片段, ..., 频繁访问板块 *N* 片段]。其中频繁访问板块片段为: [板块特征 5, 板块特征 6, 板块特征 7, 板块特征 8, 板块特征 9, 板块特征 10]。

3.3 子模型的构建方法

3.3.1 序列行为模型的构建方法

采用带负样本训练的监督学习方法 C4.5 决策树^[14] 为用户的每个会话标志各训练一个序列行为模型。训练的正样本为用户会话集合中会话标志等于该标志的所有会话; 训练的负样本为其他用户会话集合中包含该标志的所有站点的会话。负样本从其他用户会话集合中随机均匀抽取, 正负样本的比例为 1:1。

将正负会话样本生成特征向量, 并给正样本打上标签 1, 给负样本打上标签 0, 生成样本集合 *D*。根据初始的样本集合 *D* 生成决策树根结点 *N*。利用熵计算公式 $Ent(D) = - \sum_k p_k \log_2 p_k$ 计算根结点对应样本集合 *D* 的熵。之后, 遍历 *D* 的所有属性以及该属性的属性值, 以寻求熵增益率最大化的分裂方法。 *X* 为属性, *a* 为属性值, 熵增益率 *Gain_ratio* 的计算公式为:

$$Gain_ratio(D, X, a) = \frac{Gain(D, X, a)}{IV(X, a)}$$

$$GAIN_{split(X,a)}(D) = Ent(D) - P_1 * Ent(D_1) - P_2 * Ent(D_2)$$

$$P_1 = \frac{|D_1|}{|D|}, P_2 = \frac{|D_2|}{|D|}$$

$$IV(X, a) = - \sum_{i=1}^2 \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|}$$

其中, *D*₁ 为 *D* 集合中 *X* 属性的属性值小于 *a* 的样本构成的子集, *D*₂ 为 *D* 集合中 *X* 属性值大于或等于 *a* 的样本构成的子集。将 *D* 分裂成 *D*₁ 和 *D*₂, 并以此分别生成子结点 *N*₁ 和 *N*₂。以 *N*₁ 和 *N*₂ 为根结点进行递归, 继续进行分裂直到结点对应的所有样本的标签一致, 令该结点为叶结点, 用该标签进行标记; 或者如果无其他属性可分, 则令该结点为叶结点, 将结点对应集合中多数样本的标签作为该结点的标签。结束后, 形成的树即为对应该标志的序列行为模型实例。

3.3.2 页面获取关系模型的构建方法

采用单类监督学习方法支持向量描述(SVDD)^[15] 为每个频繁访问站点构建一个页面获取关系模型。SVDD 进行训练时不需要负样本, 因此训练的样本均取决于用户的历史数据。对于频繁访问站点, 利用每个包含该站点的会话中在该站点内发生的页面获取事件数据生成一条特征向量, 将所有该站点的特征向量作为训练样本集。

由于直接一维化页面获取关系图形成的特征向量的维度过高(很容易达到 100 维及以上), 因此, 为了解决维度过高所带来的问题, 我们首先考虑对特征向量进行降维。我们选择主成分分析(PCA)^[16] 作为降维的方法, 因为 PCA 算法能够自动寻找主成分, 并且降维效果明显, 返回结果为一个矩阵。

利用样本中心化公式进行中心化:

$$x_i \leftarrow x_i - \bar{x}, \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

其中, x_i 为特征向量, m 为特征向量数。

计算协方差矩阵 XX^T (X 为特征向量), 并对其进行特征值分解, 选取最大的 d' 个特征值所对应的特征向量 $w_1, w_2, \dots, w_{d'}$, 则 $W=(w_1, w_2, \dots, w_{d'})$ 即为降维矩阵。利用降维矩阵对样本集进行矩阵变换。

利用变换后的样本集进行 SVDD 模型的训练。SVDD 训练的目标为: 在样本空间上, 寻求一个点 a 和直径 R , 使得目标函数 $F(R, a, \zeta_i) = R^2 + C \sum_i \zeta_i$ 最小化, 并且满足 $(x_i - a)^T (x_i - a) \leq R^2 + \zeta_i$, 其中 x_i 为样本, ζ_i 为松弛因子。求得最优解之后, (a, R) 即为模型训练结果。

3.3.3 操作行为模型的构建方法

每个频繁访问站点分别构造一个操作行为模型, 利用每个包含该频繁访问站点的会话中该站点的操作行为数据生成一条特征向量。由于操作行为模型特征向量的维度难以出现过高的情况, 因此我们无需对它进行降维, 直接采用 SVDD 进行训练, 训练方法与页面获取关系模型的 SVDD 训练部分相同。

3.4 模型认证算法

综合行为模型的认证算法的思想为: 对序列行为模型、页面获取关系模型、操作行为模型顺序依次进行认证, 若 3 个子模型中的某个模型返回非法结果, 则本次行为认证的最终结果是非法的, 跳过该子模型之后的子模型认证; 当 3 个子模型均返回合法结果时, 本次行为认证的最终结果为合法的。综合模型的认证算法如算法 1 所示。

算法 1 综合行为模型认证算法
输入: 用户的会话数据(连续的 $\langle \text{pageview} \rangle$ 元素序列) session, 用户的频繁访问站点集合 FH, 所有频繁访问站点的频繁访问板块集合 FS, 历史会话标志集合 MARKS

输出: 认证结果(合法/非法)

1. 生成会话 session 所访问过的站点集合 HS
2. 生成会话标志 mark, $\text{mark} \leftarrow \text{HS} \cap \text{FH}$
3. if mark $\in$ MARKS
4. 转步骤 7
5. else
6. 返回非法
7. 生成特征向量 OrderFeats, 初始化为[session 特征 1]
8. for h in HS//按照字典序遍历
9. 用[h 特征 2, h 特征 3]扩展 OrderFeats
10. end for
11. result1 $\leftarrow$ 算法 2 认证 OrderFeats 的结果
12. if result1 = 合法
13. 转步骤 16
14. else
15. 返回非法
16. 选择 $\text{HS} \cap \text{FH}$ 中在 session 内访问次数最多的站点为 ph
17. 依据 FS 中所有属于 ph 的板块, 生成页面获取关系图 map
18. 取出 session 中所有 ph 站点中的所有 $\langle \text{link} \rangle$ 元素, 生成页面获取关系事件集合 LINKS
19. for link in LINKS
20. from $\leftarrow$ link. from

21. to $\leftarrow$ link. to
22. 更新 map 中 from $\rightarrow$ to 边的权值
23. end for
24. 将 map 进行一维化, 生成特征向量 MapFeats
25. result2 $\leftarrow$ 算法 3 认证 MapFeats 的结果
26. if result = 2 合法
27. 转步骤 30
28. else
29. 返回非法
30. 依据 FS 中所有属于 ph 的板块, 生成操作行为模型特征向量 OprFeats
31. 取出 session 中所有 ph 站点中的所有的 $\langle \text{operations} \rangle$ 元素, 生成元素集合 OPRS
32. for opr in OPRS
33. sec $\leftarrow$ opr 所属板块
34. 依据 opr 更新 OprFeats 中 sec 板块的相应分量
35. end for
36. result3 $\leftarrow$ 算法 4 认证 OprFeats 的结果
37. if result3 = 合法
38. 返回合法
39. else
40. 返回非法

序列行为模型认证算法的思想为: 依据会话标志, 调取相应的序列模型(C4.5 决策树), 首先将待认证的序列模型特征向量匹配到决策树的根结点, 然后根据非叶节点的属性与属性值递归匹配该节点的相应子结点, 直到匹配到叶节点, 取出叶节点的类型作为认证结果并返回。序列行为模型的认证算法如算法 2 所示。

算法 2 序列行为模型认证算法
输入: 序列行为模型特征向量 OrderFeats, 站点 h, 站点 h 所对应的决策树 dt

输出: 认证结果(合法/非法)

1. node $\leftarrow$ dt 的根结点
2. While node 不是叶节点:
3. index $\leftarrow$ node 的属性下标
4. value $\leftarrow$ node 的属性值
5. if OrderFeats[index] $\leq$ value
6. node $\leftarrow$ node. 左子节点
7. else
8. node $\leftarrow$ node. 右子节点
9. 返回 node 的标签(合法/非法)

页面获取关系模型的认证算法的思想为: 调取受保护的频繁访问站点的降维矩阵以及 SVDD 分类器。对待认证的该站点页面获取的关系图进行一维化, 以形成一维向量, 并利用降维矩阵进行降维, 形成特征向量。之后利用 SVDD 分类器对该特征向量进行分类, 将分类结果作为认证结果进行返回。页面获取关系模型的认证算法如算法 3 所示。

算法 3 页面获取关系模型认证算法
输入: 页面获取关系模型特征向量 MapFeats, 站点 h 的中心向量 Cent, 站点 h 的降维矩阵 PCAMat, 站点 h 的 SVDD 分类器 svdd
输出: 认证结果(合法/非法)

1. 对 MapFeats 进行中心化, $\text{MapFeats} \leftarrow \text{MapFeats} - \text{Cent}$
2. 对 MapFeats 进行降维, $\text{MapFeats} \leftarrow \text{PCAMat} * \text{MapFeats}$
3. $\text{result} \leftarrow \text{svdd}$ 对 MapFeats 的分类结果
4. if $\text{result} = \text{正}$
5. 返回合法
6. else
7. 返回非法

操作行为模型的认证算法的思想为:按照板块排序将该站点的各个频繁访问板块的操作行为特征片段拼接成操作行为模型特征向量。调取该站点的操作行为模型 SVDD 分类器,利用该分类器对操作行为模型征向量进行分类,并将分类结果作为操作行为模型的认证结果并返回。操作行为模型的认证算法如算法 4 所示。

算法 4 操作行为模型认证算法

输入:操作行为模型特征向量 OprFeats, 站点 h 的操作行为模型 SVDD 分类器 svdd

输出:认证结果(合法/非法)

1. $\text{result} \leftarrow \text{svdd}$ 对 OprFeats 的分类结果
2. if $\text{result} = \text{正}$
3. 返回合法
4. else
5. 返回非法

4 实验

- 实验采用侦测率和误报率两个指标。
- 1) 侦测率:进行认证并判定为非法的非法样本占进行认证的所有非法样本的比例,即 $1 - \text{假正率}^{[17]}$, $\text{假正率} = \frac{\text{判正的负样本数}}{\text{判正的负样本数} + \text{判负的正样本数}}$,侦测率越高越好。
 - 2) 误报率:进行认证并判定为非法的合法样本占进行认证的所有合法样本的比例,即 $1 - \text{真正率}^{[17]}$, $\text{真正率} = \frac{\text{判正的正样本数}}{\text{判正的正样本数} + \text{判正的正样本数}}$,误报率越低越好。

实验环境: Intel(R) Core(TM) i5 CPU(3.2GHZ); 8GB RAM; Ubuntu Kylin 16.04; Python development language; Pycharm community。

实验所使用的数据集来源于 4 个偏好站点近似的用户在 10 天内的数据。对接受测试的用户在本文编写的采集器环境下捕捉所需的数据,整个数据采集过程、采集器的记录工作都在后台运行,不会对用户产生任何提示或者干扰。采用训练样本:测试样本 = 9:1 的比例进行验证,进行多次实验并对结果取均值。

对于任意用户,正样本为该用户本身的样本,负样本是从其他用户的样本中随机抽取的。在训练阶段,首先遍历训练样本集合中的会话,获取该用户的频繁访问站点集合和各类频繁访问站点的频繁访问板块集合,并生成用户的会话标志集合。重新遍历训练集合,依据用户会话标志集合计算每个会话的序列特征向量,该会话中每个频繁访问站点计算页面获取的关系图以及各个频繁访问站点的操作行为特征向量。利用各个子模型对应的训练集进行模型训练。在认证阶段,对于

一个待认证会话,计算该会话的序列行为模型特征向量、页面获取关系模型特征向量和操作行为模型特征向量,分别利用算法 2、算法 3、算法 4 进行认证。会话认证的流程如算法 1 所示。

利用文献[9]中的方法挖掘用户访问站点的偏好,并以此构建模型,用于进行行为认证并开展参照实验。4 个用户的侦测率以及本文提出的综合行为模型的侦测率参照如图 4 所示。

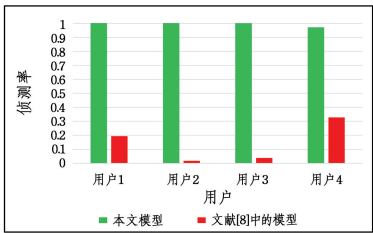


图 4 实验结果对比

Fig. 4 Comparison of experimental results

当面对的用户频繁访问的站点集合极为近似,甚至完全相同时,文献[9]提出的基于 ARM 关联规则挖掘浏览行为的模型的侦测率在 40% 以下,甚至大多数维持在 20% 以下,但是本文所提出的综合多因素的行为模型可以保持极高的侦测率,达到 100%。

序列行为模型的侦测率如图 5 所示。

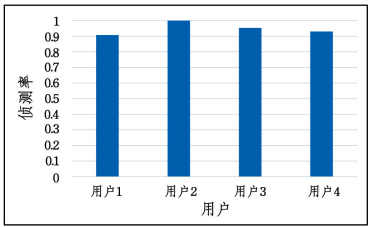


图 5 序列行为模型的侦测率

Fig. 5 Detection rate of browsing sequence behavior model

序列行为模型的实验结果主要由用户在浏览页面时对各频繁访问站点浏览的习惯决定。例如:用户 2 对站点 A 倾向于连续浏览,而对站点 B 倾向于切换浏览的个性化行为,使得用户 2 与其他用户在序列模型层面上显得极易被区分。

页面获取关系模型的侦测率如图 6 所示。

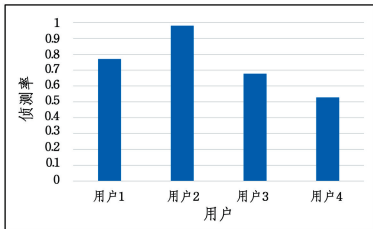


图 6 页面获取关系模型的侦测率

Fig. 6 Detection rate of page access relationship model

页面获取关系模型的实验结果由用户在一个站点下使用链接打开新页面的习惯决定。例如:如果一个用户在 qq.com 站点下大量使用 sports 和 news 板块内容页面的推荐系统,那么他们的两条有向边 sports_CONTENT → sports_CON-

TENT 和 news_CONTENT $\rightarrow$ news_CONTENT 的权重将会很大,当这一习惯极具个性化时,qq.com 站点下的页面获取关系模型将会很好地发现非法用户。

对于操作行为模型,特征向量的维度并不会过高,因此直接采用 SVDD 分类器即可。操作行为模型的侦测率如图 7 所示。

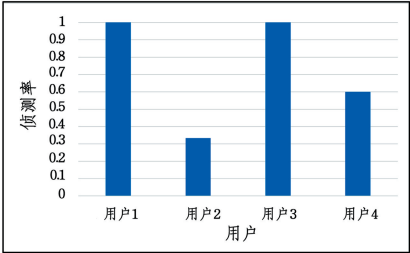


图 7 操作行为模型的侦测率

Fig. 7 Detection rate of operation behavior model

操作行为模型的训练结果与用户的操作行为的稳定性和个性化有重要的联系,例如用户 2 的操作层面并不稳定,将会导致侦测效果变差。

基于 3 个子模型共同进行侦测的综合模型的误报率在 10%~20%之间,如图 8 所示。

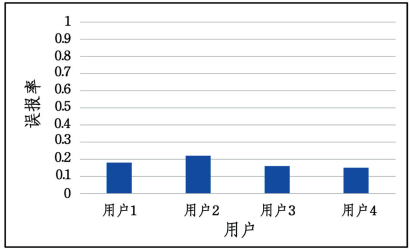


图 8 综合模型的误报率

Fig. 8 False alarm rate of comprehensive behavior model

产生误报的原因在于:用户的实际行为数据确实存在着极少量的特殊情况,但是为了确保模型的侦测效果,我们通常不会把这种极少数的情况学习到分类器的正样本情况中,例如在 SVDD 中通过松弛因子的设定会把这类正样本点舍去。

结束语 通过实验中的数据可以发现,若仅仅依靠会话中包括的站点对用户的行为进行建模,认证方法则会面临很高的失效风险。因此,应该提出一个模型,从多个方面去刻画一个用户,考虑到用户在不同方面的行为特征,尽可能细致地构建用户行为模型。

本文提及的用户浏览行为认证方法针对用户构建了一个综合的模型,该模型由上述 3 个子模型组合形成。3 个子模型分别对用户浏览行为中不同的因素建立模型。实验结果表明,每个模型都能够实现一定的认证效果,能够降低账号被盗用情况下造成损失的风险。3 个子模型组合起来共同进行认证时能达到极高的侦测率,可为用户的账号安全提供有力的保障,同时能把误报率维持在有限水平。综合考虑用户浏览行为的多种因素,可以规避仅仅考虑单一因素时检测失效的风险。相比于之前的成果,本文模型的误报率有所增加,但仍维持在较低水平。在实际应用中,在行为认证不被通过时,并不是直接进行账号冻结,而是进行强认证(如:短信、邮件;若

频繁进行,则会影响用户体验),以稍提高误报率为代价(这将略牺牲用户体验),来换取侦测率的进一步提高与健壮性的大大增强。

在未来的工作中,我们将考虑使用更准确的模型来刻画用户的各个子模型,同时尝试采用更新颖的模型集合手段来进一步提高侦测率,增强健壮性,降低误报率。

参 考 文 献

[1] 艾瑞网, 2017 中国电商导购行业研究报告[OL]. http://report.iresearch.cn/report_pdf.aspx?id=2957.

[2] 360 互联网安全中心,《2016 年中国互联网安全报告》[OL]. <http://zt.360.cn/1101061855.php?dtid=1101062514&did=490278985>.

[3] NENADIC A,ZHANG N,BARTON S. A security protocol for certified e-goods delivery[C]// International Conference on Information Technology,Coding and Computing. 2004:22-28.

[4] HUANG L,LIU G. On data security in e-commerce[C]// 2012 IEEE Symposium on Electrical and Electronics Engineering. 2012:94-97.

[5] KALAYE M R G,NIK M H,KORDESTANI H. Using template-based passwords for authentication in e-banking [C] // Ecommerce in Developing Countries; with Focus on E-security. 2013:1-9.

[6] HONG L,JAIN A. Integrating faces and fingerprints for personal identification[C]// Asian Conference on Computer Vision. 1998:16-23.

[7] WANG D,WANG P. Two Birds with One Stone:Two-Factor Authentication with Security Beyond Conventional Bound[J]. IEEE Transactions on Dependable & Secure Computing, 2016 (99):1.

[8] GE Y,YAN C,DING Z,et al. Web Access Patterns Mining for Individuals with Timing and Link Sequence[J]. Information Technology Journal, 2014, 13(4): 746-753.

[9] ZHONG J,YAN C,YU W,et al. A Kind of Identity Authentication Method Based on Browsing Behaviors [J]. International Symposium on Computational Intelligence & Design, 2015, 2: 279-284.

[10] ABRAMSON M, AHA D W. User Authentication from Web Browsing Behavior[C]// FLAIRS Conference. 2013.

[11] LIU Q,WANG M M,ZHAO P H,et al. A Behavioral Authentication Method for MobileGesture Against Resilient User Posture[C]// The 2016 3rd International Conference on Systems and Informatics (ICSAI 2016). 2016.

[12] ANIMA B A,JASIM M,RAHMAN K A,et al. User Authentication from Mouse Movement Data Using SVM Classifier[C]// International Conference on Cryptology & Network Security. 2016:692-700.

[13] BARABÁSI A L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435(7039): 207.

[14] QUINLAN J R. C4. 5: Programs for Machine Learning[M]. San Francisco: Morgan Kaufman, 1993.

[15] TAX D M J, DUIN R P W. Support vector domain description

[J]. Pattern Recognition Letters,1999,20(11-13):1191-1199.

[16] ABDI H,WILLIAMS L J. Principal component analysis[J]. Wiley Interdisciplinary Reviews Computational Statistics, 2010, 2(4):433-459.

[17] SPACKMAN K A. Signal detection theory: Valuable tools for evaluating inductive learning[C]//Proceedings of the 6th International Workshop on Machine Learning (IWML). Ithaca, NY, 1989:160-163.

(上接第 170 页)

机过程;通过采用近似最优的能量分配策略,推导出系统平均吞吐量以及系统的近似信道容量;同时得到当捕获能量小于超级电容存储容量时系统信道容量的上下界之间的常数差值为 1.77 bps/Hz,当捕获能量大于超级电容存储容量时系统信道容量的上下界之间的常数差值为 2.49 bps/Hz。实验结果表明,采用混合能量存储结构,随着捕获能量的增加和超级电容存储容量的增大,系统的性能提升得越大,即系统的能量利用率增加且信道容量增大。未来将对多节点多跳网络模型、衰减信道以及捕获的能量符合其他一般随机过程情形下系统的平均吞吐量和信道容量进行分析。

参 考 文 献

[1] HAN J H,DING X,SHI L,et al. Research on the time-varying charging and dynamic data routing strategy for rechargeable wireless sensor networks[J]. Journal on Communications,2012, 2012(12):1-10. (in Chinese)

韩江洪,丁煦,石雷,等. 无线传感器网络时变充电和动态数据路由算法研究[J]. 通信学报,2012,2012(12):1-10.

[2] PARADISO J A,STARNER T. Energy Scavenging for Mobile and Wireless Electronics [J]. IEEE Pervasive Computing,2005, 4(1):18-27

[3] SU B,LI Y Q,YU H Y,et al. The Wireless Sensor Node Harvesting Energy from Environment [J]. Chinese Journal of Sensors & Actuators,2008,21(9):1586-1589. (in Chinese)

苏波,李艳秋,于红云,等. 从环境中获取能量的无线传感器节点[J]. 传感技术学报,2008,21(9):1586-1589.

[4] JIANG X F,POLASTRE J,CULLER D. Perpetual environmentally powered sensor networks[C]// International Symposium on Information Processing in Sensor Networks. IEEE Press, 2005:463-468.

[5] SUDEVALAYAM S,KULKAMI P. Energy Harvesting Sensor Nodes;Survey and Implications[J]. IEEE Communications Surveys & Tutorials,2011,13(3):443-461.

[6] YAO X W,ZHENG X H,WANG W L,et al. A Bi-dimensional Wireless Energy Transfer Mechanism for Maximum Network Throughput [J]. Computer Science,2015,42(11):164-169. (in Chinese)

姚信威,郑星航,王万良,等. 吞吐量最大化的二维无限能量传输算法[J]. 计算机科学,2015,42(11):164-169.

[7] YANG J,ULUKUS S. Optimal Packet Scheduling in an Energy Harvesting Communication System[J]. IEEE Transactions on Communications,2010,60(1):220-230.

[8] OZEL O,TUTUNCUOGLU K,YANG J,et al. Transmission with Energy Harvesting Nodes in Fading Wireless Channels: Optimal Policies [J]. IEEE Journal on Selected Areas in Communications,2011,29(8):1732-1743.

[9] OZEL O,SHAHZAD K,ULUKUS S. Optimal Energy Allocation for Energy Harvesting Transmitters With Hybrid Energy Storage and Processing Cost[J]. IEEE Transactions on Signal Processing,2014,62(12):3232-3245.

[10] OZEL O,TUTUNCUOGLU K,ULUKUS S,et al. Fundamental limits of energy harvesting communications[J]. IEEE Communications Magazine,2015,53(4):126-132.

[11] OZEL O,ULUKUS S. Information-theoretic analysis of an energy harvesting communication system[C]//International Symposium on Personal, Indoor and Mobile Radio Communications Workshops. IEEE Xplore,2010:330-335.

[12] OZEL O,ULUKUS S. Achieving AWGN Capacity Under Stochastic Energy Harvesting [J]. IEEE Transactions on Information Theory,2012,58(10):6471-6483.

[13] RAJESH R,SHARMA V,VISWANATH P. Capacity of Gaussian Channels with Energy Harvesting and Processing Cost [J]. IEEE Transactions on Information Theory,2014,60(5):2563-2575.

[14] MAO W,HASSIBI B. On the capacity of a communication system with energy harvesting and a limited battery [C]// IEEE International Symposium on Information Theory Proceedings. 2013:1789-1793.

[15] JOG V,ANANTHARAM V. An energy harvesting AWGN channel with a finite battery[C]// IEEE International Symposium on Information Theory. 2014:806-810.

[16] TUTUNCUOGLU K,OZEL O,YENER A,et al. Binary energy harvesting channel with finite energy storage[C]// IEEE International Symposium on Information Theory Proceedings. 2013: 1591-1595.

[17] DONG Y,FARNIA F,OZGUR A. Near Optimal Energy Control and Approximate Capacity of Energy Harvesting Communication[J]. IEEE Journal on Selected Areas in Communications, 2014,33(3):540-557.

[18] DONG Y,OZGUR A. Approximate capacity of energy harvesting communication with finite battery[C]// IEEE International Symposium on Information Theory. 2014:801-805.

[19] SHAVIV D,NGUYEN P M,OZGUR A. Capacity of the Energy Harvesting Channel with a Finite Battery[C]// 2015 IEEE International Symposium on Information Theory (ISIT). IEEE, 2015:131-135.