文章编号:1007-1423(2018)07-0058-05 DOI:10.3969/j.issn.1007-1423.2018.07.012

基于网页浏览行为的反爬虫研究

刘洋

(四川大学计算机学院,成都 610065)

摘要:
在大数据的背景下,数据的潜在价值被不断地挖掘出来。能够有效识别或阻挡爬取行为的反爬虫方法对于商业服务网站来说尤为重要。基于网页浏览行为,提出一种新的反爬虫方法。该方法通过对真实用户和网络爬虫浏览网页的行为进行特征提取,然后构造并使用决策树对一个用户是否属于爬虫进行预测。该方法对网络爬虫的敏感性高,并具有较低的假阴率。

关键词:
网络爬虫;反爬虫方法;用户浏览行为;网站

0 引言

爬虫是一种按照特定规则在互联网上获取信息的程序。在大数据的背景下,每一家公司都希望在提供互联网服务的同时保护自己的数据,防止信息被网络爬虫的窃取。根据 Distil Networks 发布的"2018 Bad Bot Report"[1],在 2017 年,全部互联网流量中有 42.2%来自于爬虫程序,而不是真实的用户。在这些爬虫中有 21.8%属于不遵守爬虫协议的恶意爬虫,恶意爬虫的肆掠不仅会导致目标网站访问速度变慢甚至无法访问,而且会导致网站用户隐私信息、公司商业数据泄漏,造成持续的经济损失。所以,能够阻挡爬取行为的反爬虫方法对于商业服务网站来说尤为重要。

1 相关工作

现存的反爬虫方法主要可以分为两类,第一类会依据各种辅助信息主动识别网络爬虫,然后拒绝提供服务,或者返回无关的垃圾信息使其失效[2];第二类反爬虫方法不主动识别网络爬虫,而是利用各种复杂的前端数据渲染技术来增加网络爬虫的爬取难度。

用户通过 HTTP 向网站发出请求,网站可以通过 HTTP Request Header 获取发送请求的终端信息、发送请求的来源页面链接、用户的登录信息以及用户的 IP 地址[3]。通过埋点技术,网站还可以获取用户的鼠标的点击事件、窗口缩放事件、请求频率等信息。网站可以借助这些信息,判断一个请求的发起者是不是一个真实的用户,进而采取相应的应对措施。这种方法的优点是简单、对网站的改动小;但是也存在严重的缺陷,网络爬虫可以通过伪造 HTTP Request Header 内容,轻松地绕过网站的检测。另外,新浪新闻网站采用了异步数据加载的方法渲染前端数据来防止网络爬虫直接获取页面静态数据。还有一些网站,例如美团网、猫眼电影网会对传输数据进行加密,然后在前端利用自定义的字符映射关系表将数据解码后显示出来[4]。这种方法的缺陷是需要对网站进行较大的改动,且不易维护。并且随着 OCR(光学字符识别)技术的成熟,网络爬虫可以准确地识别网页截图中的文字内容,使这种反爬虫方法失效。

2 网页浏览行为特征提取

用户在浏览网页的过程中产生的行为数据信息统称为网页浏览行为,他们共同刻画了用户的行为特点、网页浏览习惯以及偏好[5]。所以,我们选择了部分浏览行为作为特征来帮助识别网络爬虫。表 1 展示了本文所使用的 9 种网页浏览行为特征。接下来我们需要将

具有连续值的特征进行标准化，以便于后续决策树的构造。

●特征是离散值，且取值划分大于 2，那么用该特征的每一个划分作为一个子分类；

●特征是离散值，且取值划分数为 2，那么用“符合此特征”和“不符合此特征”分为两个子分类；

●特征是连续值，那么取分裂点 *division_point*，将其划分为两个子分类：$feature > division\_point$ 和 $feature \leqslant division\_point$。

表 1 网页浏览行为特征

行为特征
A 站内访问活跃度
O 站外访问
C 访问内容丰富度
P 访问页面丰富度
T 访问页面轨迹
t 点击间隔时间
R 不同资源类型请求分布
s 页面滚动
TP 页面停留时间

（1）*A* 站内访问活跃度：该特征描述了某个用户在一段时间内平均发起的站内请求数量 A_1，站内请求的平均时间间隔 A_2。如果 $A_1 - A_1' > \alpha_1$（其中 A_1' 是相同时间段内所有真实用户的站内请求数量的均值），那么该用户就更有可能是网络爬虫；如果 $A_2' - A_2 > \alpha_2$（其中 A_2' 是相同时间段内所有真实用户发起的站内请求的平均时间间隔），那么该用户就更有可能是网络爬虫。

（2）*O* 站外访问：该特征描述了用户在网站中浏览时是否会通过点击站外链接。通常网络爬虫会爬取指定域名下的网页，所以如果用户没有点击站外链接，$O = 1$；否则 $O = -1$。

（3）*C* 访问内容丰富度：该特征描述了用户点击元素类型的数量 C_1，以及用户点击元素类型的分布 C_2。由于网络爬虫访问的页面元素类型比较集中，如果 $C_1' - C_1 > c_1$（其中 C_1' 是所有真实用户点击的页面元素类型的数量平均值），那么表示该用户只点击页面中固定的几种元素，可能是网络爬虫；如果点击元素类型的分布向量 C_2 与所有真实用户点击元素类型的分布向量 C_2' 的距离大于阈值 c_2（$C_2 \cdot C_2' > c_2$），那么表示该用户对某些网页元素的点击频率高于均值，可能被识别为爬虫。

（4）*P* 访问页面丰富度：该特征描述了用户在浏览的页面种类 P_1，以及访问的页面种类分布向量 P_2。网络爬虫具有明确的目标，所以访问的页面比较固定，相应地页面丰富度较低。

（5）*T* 访问页面轨迹：该特征描述了用户在浏览一个网站时的访问顺序，使用向量 *T* 表示（$T = \{t_1, t_2, t_3, \cdots t_n\}$），其中 t_i 表示第 i 个页面访问序列。假设网站的所有合法页面的访问轨迹集合为：$T' = \{t_1', t_2', t_3', \cdots t_n'\}$，那么当 $\exists t \in T, t \notin T'$，该用户就可能是网络爬虫。

（6）*t* 点击间隔时间：该特征描述了用户的两次点击的间隔时间。假设所有真实用户的点击时间间隔平均值为 t'，如果用户的点击间隔时间明显低于真实用户的平均点击时间间隔（$t' - t > \mu$），那么该用户就可能是网络爬虫。

（7）*R* 不同资源类型请求分布：网站的资源通常可以分为静态资源和动态资源两类，该特征描述了用户请求 HTML、文本、图片、CSS、JS 等不同资源类型的数量分布。如果请求资源类型的分布向量 *R* 与所有真实用户请求的资源类型分布向量 R' 的距离大于阈值 γ（$R \cdot R' > \gamma$），那么表示该用户对某些资源类型的访问频率高于均值，可能被识别为爬虫。

（8）*s* 页面滚动：该特征表示用户在浏览单个页面时鼠标向下滚动的次数 s_1，以及向上滚动的次数 s_2。用户在浏览网页的过程中，需要上下滚动鼠标查看网页的所有内容。而网络爬虫直接对网站代码进行解析不会产生滚动动作，虽然使用 Selenium 和 PhontomJs 技术的网络爬虫也会产生滚动行为，但是和真实用户相比，鼠标向下滚动的次数会大幅超过鼠标向上滚动的次数。

（10）*TP* 页面停留时间：用户在浏览网页的过程中，需要驻留一段时间浏览内容再决定访问下一个页面或者选择离开网站。该特征表示用户从打开单个页面到离开该页面的时间差。如果用户单个页面停留时间小于最小正常时间（$TP < \theta$），那么该用户可能被识别为爬虫。

3 决策树的构造

决策树（Decision Tree）是以类别为叶节点，属性为分支节点的一棵二叉树或多叉树[6]。在决策树中，每一个分支节点表示一个特征的划分，而每个分支对应这

个特征的值域分布，即表示对象所属分类的预测结果。从根节点到叶子节点的路径就是一条分类规则，分类的过程简单直观，方便人类专家进行检验[7]。

决策树的构造的关键是对特征进行筛选，选择合适的特征度量将对象划分成不同的类，而构造的目标是让决策树的每一个待分类分支尽可能地属于同一个类别。那么我们每次都应该选择信息增益最大的特征进行划分操作。假设 D 是样本数据集，样本数量为：n，其中包括两种类别的数据（真实用户类别样本数：n_1、网络爬虫类别样本数：n_2），那么划分前的信息熵计算公式如公式（1）所示：

$$entropy(D) = -\frac{1}{n}\left(n_1 \cdot \log_2 n_1 + n_2 \cdot \log_2 n_2\right) + \log_2 n \tag{1}$$

然后我们使用特征 *Feature* 对样本 D 进行划分，那么 *Feature* 作用后的信息熵的计算公式如公式（2）所示。其中，k 为样本 D 划分后的 k 个子类；D_i 表示第 i 个子分类下的样本数量。而信息增益的计算就是两者的差值，如公式（3）所示：

$$entropy_{Feature}(D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \cdot entropy(D_i) \tag{2}$$

$$Gain(Feature) = entropy(D) - entropy_{Feature}(D) \tag{3}$$

最后，我们依次选择信息增益最大的特征构造决策树。为了避免决策树对样本数据过拟合的现象，我们将样本数据集 D 分为三部分（训练集 D_1 占二分之一，验证集 D_2 占四分之一，验证集 D_3 占四分之一）。在划分的过程中，每次划分前，使用验证集 D_2 来验证 *Feature* 能否提高决策树的分类准确度。如果分类准确度提高，就把 *Feature* 加入决策树作为分支节点；否则此次划分失败。最后使用验证集 D_3 对决策树进行考察，自底向上依次删除非叶子节点。如果将该子树替换成叶节点能够提高分类准确度，那么执行该删除操作，否则保留该节点。

4 结语

本文提出一种基于网页浏览行为的反爬虫方法，它通过提取真实用户和网络爬虫浏览网页时产生的行为特征构造决策树，最后使用决策树对一个用户是否属于爬虫进行预测。该方法属于第一类的反爬虫方法，和同类的其他方法相比具有更高的真阳率，更低的假阴率；和第二类的反爬虫机制相比，该方法简便易实施，并且对网站的改动小，并且针对网络爬虫的敏感性高，假阴率较低。然而在少数情况下，使用该方法仍可能将网络爬虫判断为真实用户，主要是因为选择的网页浏览行为特征不够充足。接下来，我们希望能够提取更多有效的网页浏览行为特征，来杜绝假阴性。

参考文献：

[1]2018 Bad Bot Report [EB/OL]. https://resources.distilnetworks.com/whitepapers/2018-bad-bot-report，2018.

[2]刘庆杰，孙旭光，王小英. 通过 Filter 抵御网页爬虫[J]. 网络安全技术与应用，2010(1):70-71.

[3]陈利婷. 大数据时代的反爬虫技术[J]. 电脑与信息技术，2016(6):60-61.

[4]顾流，万仲保，石红芹. 基于 Web 页信息隐藏的研究与实现[J]. 微计算机信息，2006，22(24):186-187.

[5]张宁. 群体兴趣网的统计特性研究[J]. 上海理工大学学报，2008，30(3):243-248.

[6]Witten I H，Frank E. Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)[M]. 机械工业出版社，2005.

[7]史忠植. 知识发现.第 2 版[M]. 清华大学出版社，2011.

作者简介：

刘洋(1993-)，男，河南信阳人，硕士，研究方向为网络爬虫

收稿日期：2018-12-25　　修稿日期：2019-01-10

（下转第 70 页）

作者简介：
潘泽云(1992-)，男，江苏常州人，硕士，研究方向为图形图像处理、虚拟现实
丁利琼(1993-)，女，四川巴中人，硕士，研究方向为图形图像处理、计算机视觉
通信作者：程鹏(1985-)，四川成都人，讲师，博士，研究方向为图像处理、计算机视觉
收稿日期：2018-12-25　　修稿日期：2019-01-09

Real-Time Simulation of Rainy Weather Based on GPU

PAN Ze-yun[1], DING Li-qiong[1], CHENG Peng[2]

(1. College of Computer Science, Sichuan University, Chengdu 610065; 2. Sichuan Wisesoft Co., Ltd., Chengdu 610065)

Abstract：
Proposes a method of particle system based on GPU to simulate the rainfall weather in real time, introduces the force and motion of raindrops in the real world, and takes the light source and the position of viewpoint as controllable variables to ensure the authenticity and reusability of the scene. In order to ensure real-time performance, the data of raindrop pictures with different light sources and viewpoint positions are stored in two-dimensional texture array, and the powerful graphics processing ability of GPU is reasonably utilized. The experiment proves that the authenticity and reusability of rainfall weather are high and the real-time performance is good.
Keywords：
Particle System; Force Model; Motion Model; Real-Time Rendering; GPU

(上接第 60 页)

Research on Anti-Spider Method Based on Web Browsing Behavior

LIU Yang

(College of Computer Science, Sichuan University, Chengdu 610065)

Abstract：
In the era of the big data, the potential value of data is constantly being explored. Anti-spider methods that effectively identify or block web spider crawling behavior are especially important for websites who offer commercial services. Proposes a new anti-spider method based on the behavior of browsing the web page. The method extracts feature from web browsing behavior of real users and web spider, and then constructs and uses decision tree to predict whether a user is web spider. This method has high sensitivity to web spider and a low false-negative rate.
Keywords：
Web Crawler; Anti-Spider; User Browsing Behavior; Website