清华大学出版社是教育部主管、清华大学主办的综合性大学出版社，是国内领先的综合性教育与专业出版机构。

随着互联网的发展与移动设备的普及，互联网用户及终端设备数量呈现爆发式的增长。一位用户可能使用多台不同的设备访问相同的网站。

出于对获取各类网站上大量有价值信息的需要，网络爬虫也被大量地用于获取网络数据。

根据Imperva发布的统计报告，2019年，恶意爬虫的流量占所有互联网流量的24.1%，上升到有史以来最高的百分比，而37.2%的互联网流量是非人的。

恶意爬虫可能会导致网站核心内容被复制、注册用户被扫描、网站带宽负担加重等问题。

Web日志中包含了访问主机IP、访问时的服务器时间、访问页面、请求方式等信息。HTTP协议中包含的User Agent字段能够识别访客使用的操作系统及版本、CPU类型、浏览器及版本等。MAC地址，还可以实现用户的唯一性识别。

妥协型的方法有设置Robots.txt协议。设置Robots.txt 文件，来告诉一些规范的爬虫使用者页面的爬取规则。主动型方法有分析不同访问行为的特征，然后采用阈值、决策树、偏见网络等方法对网络爬虫进行检查。但是，该方法的一个缺点是它的滞后和耗时，它不能实时地发现和处理爬行请求；对HTTP请求头的每个属性进行判断；设置cookie检测等。第二种方法是基于陷阱技术。你可以在网页上设置一个隐藏的链接，用户不能点击它，但是可以被网络爬虫访问。使用honeypot陷阱可以跟踪许多种类的爬虫。能准确地捕捉爬虫，而当爬虫设计师意识到这一陷阱时，它便会变得毫无用处。基于滑动时间窗的爬虫实时检测方法提高了对不符合爬虫规则的检测的准确性和效率。

用户和网络爬虫的行为特征的差异主要体现在两个方面:一是访问时间的不同。用户在网页阅读有时快，有时慢，二是访问内容不同。原因是用户与爬虫程序相比，目的感更低。因此从这两个方面进行细分和设置权重。

确定特征空间的维数

2: Cookie可以用作一个维度。cookie有三种状态。cookie不存在，cookie存在但检查失败，cookie存在且检查成功。可以使用滑动窗口中不重复的cookie的数量进行判断。

3:会话可以用作一个维度。滑动窗口中不重复的会话数，单个会话是否对应不同的IP，最大会话窗口访问数

4:可将访问时间作为一个维度。最大裕度的时间，访问之间的中间时间，时间间隔的方差。

5: Referer可以用作一个维度。我主要是从三个角度来判断:referer不存在，referer存在但不是我们网站的，referer存在并是我们网站的。

6:用户代理可以作为一个维度。没有声明、声明为未知爬虫、声明为已知爬虫但IP不正确、声明为已知爬虫且IP正确。

基于滑动时间窗的爬虫检测算法。该系统的优点是实时、提高了准确率，降低了系统资源消耗成本。

在智能流控方面，许多爬虫可能会通过使用代理的方式来避开IP识别，当遭受来自无法产生大量流量的国家的访问，可以限制来自该国家的流量。重定向到验证码界面

当网站项目需要添加本系统时，不需要在原有系统的基础上做很大的改动。使原系统与本系统具有高内聚性和低耦合性。