 单位代码 **10006**

学号 **17373281**

分类号 **TP311.5**

****

毕业设计(论文)

网站访客唯一性识别和智能流控方案的设计与实现

|  |  |
| --- | --- |
| 学院名称 | 软件学院 |
| 专业名称 | 软件工程 |
| 学生姓名 | 江一帆 |
| 指导教师 | 吕云翔 |

**2021**年**6**月

网站访客唯一性识别和智能流控方案的设计与实现

学生：

江一帆

指导教师：

吕云翔

摘要

随着互联网的发展与移动设备的普及，一位用户可能使用多台不同的设备或经由不同的浏览器访问相同的网站。同时，网络爬虫也被频繁地用于获取网络数据，可能导致网站核心内容被复制等问题。清华大学出版社作为国内领先的综合性教育与专业出版机构，其网站在用户的唯一性识别以及反恶意爬虫等方面都有较高的需求。本课题的目标是根据清华大学出版社的实际需求，设计并实现一个识别用户身份与爬虫，并进行网站流量控制的方案。通过IP地址、浏览器Fingerprint等信息对访客进行标记，通过接口请求频率以及身份标识一致性对爬虫进行识别，将用户、爬虫、用户表示信息等存储在前后端数据库中。通过对访问频率过高的用户设置请求限制的方式，达到网站流量控制的目的。

关键词：计算机

网站，反爬虫，访客识别，流量控制

The Design and Realization of Website Visitor Recognition and Smart Website Traffic Control

Author :

Jiang Yi-fan

Tutor :

Lu Yun-xiang

Abstract

With the development of the Internet and the popularization of mobile devices, a user may visit a same website through several different devices or different browsers. At the same time, web crawlers are also frequently used to obtain data from websites, which may cause problems such as leakage of the core content of the website. As a leading comprehensive and professional publishing institution in China, Tsinghua University Press has a high demand in unique user identification and anti-malicious-crawler. The purpose of this project is to design and implement a scheme to identify user and crawler, as well as controlling website traffic according to the actual needs of Tsinghua University Press. Users are marked by IP address, browser Fingerprint and other information, crawlers are identified by interface request frequency and identity consistency. User, crawler and user identity information are stored in front-end and back-end database. Website traffic control is achieved by setting request limitation for users with high request frequency.

Key words:

Computer Website, Anti-Crawler, Website Visitor Recognition, Website Traffic Control

目录

[1 绪论 1](#_Toc69416463)

[1.1 课题背景与意义 1](#_Toc69416464)

[1.2 国内外研究现状 1](#_Toc69416465)

[2 系统设计与技术选型 3](#_Toc69416466)

[2.1 系统设计 3](#_Toc69416467)

[2.1.1 需求分析 3](#_Toc69416468)

[2.1.2 系统整体设计 6](#_Toc69416469)

[2.2 技术选型 8](#_Toc69416470)

[2.3 项目任务 11](#_Toc69416471)

[3 项目具体实现 12](#_Toc69416472)

[3.1 用户识别部分 12](#_Toc69416473)

[3.2 爬虫识别部分 14](#_Toc69416474)

[3.3 请求限制部分 15](#_Toc69416475)

[参考文献 16](#_Toc69416476)

# 绪论

## 课题背景与意义

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或经由不同的浏览器访问相同的网站。

除此之外，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。根据统计，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为带来的[1]。对于网站而言，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。

论文课题由北京航空航天大学软件学院吕云翔老师提出，来源于吕老师于清华大学出版社联系到的对于识别用户唯一性、识别爬虫并进行接口请求限制的实际需求。清华大学出版社成立于1980年6月，是教育部主管、清华大学主办的综合性大学出版社，现年出版图书、音像制品、电子出版物等近3000种。作为国内领先的综合性教育与专业出版机构，清华大学出版社的网站在用户的唯一性识别以及反恶意爬虫等方面都有较高的需求。

本课题的目标是根据清华大学出版社提出的保障网站安全性与性能的实际需求，结合国内外研究现状和发展状态，设计并实现一个识别用户身份与爬虫，并进行网站流量控制的方案，希望通过对访客进行标记以及对恶意爬虫进行识别，对恶意用户与爬虫设置访问限制，达到智能流量控制的目的。

## 国内外研究现状

本课题所涉及到的需求主要可以分为三部分，分别是访客唯一性识别、恶意爬虫识别以及网站流量调控。

在访客唯一性识别方面，服务器产生的Web日志可以成为良好的用户识别数据源，因为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。其中，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。由于MAC地址是网卡物理地址，与网络无关，通过在Web日志中加入MAC地址，还可以实现用户的唯一性识别，增加用户识别的准确性本课题所涉及到的需求主要可以分为三部分，分别是访客唯一性识别、恶意爬虫识别以及网站流量调控。

在访客唯一性识别方面，服务器产生的Web日志可以成为良好的用户识别数据源，因为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。其中，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。由于MAC地址是网卡物理地址，与网络无关，通过在Web日志中加入MAC地址，还可以实现用户的唯一性识别，增加用户识别的准确性[2]。

在爬虫识别方面，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[3]。其他比较常用的主动型方法有设置单位阈值，倘若某一IP、User Agent标识超出了阈值，对其实施监控；对HTTP请求头的每个属性进行“是否常规访问”的判断[4]；设置cookie检测等。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。此外，还有基于机器学习的爬虫识别，通过提取会话的身份信息、日志的统计信息、日志中的异常信息进行模型训练，并使用训练后的模型通过会话信息识别爬虫[6]。

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[7]，当遭受来自无法产生大量流量的国家的访问（通常外国访问者使用一个网站是没有意义的），可以限制来自该国家的流量。重定向到验证页面也是十分有效的方法，常见的验证方式有验证码、短信验证、邮箱验证。使用Honeypot陷阱可以跟踪许多种类的爬虫，并且通过一个简单的规则列表就可以将请求分类为恶意或非恶意[8]。

(undone,待扩展

# 系统概述及需求分析

## 项目概述

本课题的内容是设计并实现一个识别用户身份与爬虫，并进行网站流量控制的系统，以接口的形式接入清华大学出版社的网站。本系统通过收集客户端设备、浏览器的特征值、IP地址等信息对用户进行标识，通过请求频率对爬虫进行标识，对爬虫程序设置请求频率、总次数的限制以达到网站流量控制的目的，对网站在安全性、负载等方面提供一定的保护。

本系统的全部设计与开发工作均由本文作者完成，包括需求分析、系统设计、系统编码、测试等。

## 功能需求

本项目的功能需求主要集中在数据的收取与处理方面，可分为如下几部分：记录用户标识数据、标记用户、记录用户请求数据、标记爬虫以及设置接口访问限制。项目的用例图如图2.1所示。

（undone，具体说明用例）



图2.1 项目用例图

### 记录用户标识数据

记录用户标识数据的目的是为了在标识用户身份时可以获得更精确的结果，具有高信息熵和高跨浏览器稳定性的用户标识数据越多，就越能唯一确定一名用户。更多有效的用户标识数据也能减少数据之间的碰撞。所记录的用户标识数据将用于计算用户Fingerprint。

数据的收集工作是在用户进入网站时完成的，也就是页面第一次加载时，无法获取到的用户标识数据将会以undefined替代。

记录用户标识数据的用例说明见表2.1。

表2.1 记录用户标识数据

|  |  |
| --- | --- |
| 用例名称 | 记录用户标识数据 |
| 简要描述 | 该用例描述了系统记录用户标识数据的过程 |
| 参与者 | 网站管理员 |
| 涉众 | 用户 |
| 前置条件 | 用户进入到网站 |
| 后置条件 | 记录成功后进行用户标记 |
| 基本事件流 | 1. 用户进入网站 2. 系统收集用户标识数据，包括useragent、ip地址、fingerprint等。（A-2） |
| 备选事件流 | A-2 存在无法获取的数据   1. 以undefined代替 |
| 补充约束 | 无 |
| 待解决问题 | 无 |

### 标记用户

标记用户的目的是基于上一步收集到的用户标识数据，将用户标识数据进行加权计算，得到一个唯一的用户Fingerprint值用作区分用户的依据之一。此外，IP地址也将作为次要依据用于区分用户。

在生成Fingerprint后，将会完成前后端数据库的数据存储，并对已经存在用户进行数据的比对与更新。

标记用户的用例说明见表2.2。

表2.2 标记用户

|  |  |
| --- | --- |
| 用例名称 | 标记用户 |
| 简要描述 | 该用例描述了系统根据获取的数据标记用户的过程 |
| 参与者 | 网站管理员 |
| 涉众 | 无 |
| 前置条件 | 已经获得用户标识数据 |
| 后置条件 | 标记成功后保存到数据库中 |
| 基本事件流 | 1. 传入收集到的用户标识信息 2. 根据数据区分用户后，存入IndexedDB中，并同步到MongoDB。（B-2） |
| 备选事件流 | B-2 同一用户已存在   1. 检查数据是否重复 2. 添加不重复的新数据 |
| 补充约束 | 无 |
| 待解决问题 | 无 |

### 记录用户请求数据

记录用户请求数据的目的是为了后续判断客户端是否为爬虫程序。需要记录的内容包括请求数据，例如header，并记录请求的时间等，用于计算单位时间内请求的频率。

记录用户请求数据的用例说明见表2.3。

表2.3 记录用户请求数据

|  |  |
| --- | --- |
| 用例名称 | 记录用户请求数据 |
| 简要描述 | 该用例描述了系统记录用户请求数据的过程 |
| 参与者 | 网站管理员 |
| 涉众 | 用户 |
| 前置条件 | 用户进入到网站 |
| 后置条件 | 记录成功后进行爬虫标记 |
| 基本事件流 | 1. 用户进入网站 2. 系统收集用户请求数据，包括请求的header、请求的时间等。（C-2） 3. 计算请求的频率 |
| 备选事件流 | 无 |
| 补充约束 | 无 |
| 待解决问题 | 无 |

### 标记爬虫

标记爬虫的目的是基于获取的用户请求数据，对比数据库中记录的用户标识数据、判断请求频率是否超过上限等，将请求过高的客户端标记为爬虫程序。对于爬虫程序，还需要保存在数据库中，并设置访问限制信息。对于已经被标记的爬虫程序的客户端，则需要更新限制信息。

标记爬虫的用例说明见表2.4。

表2.4 标记爬虫

|  |  |
| --- | --- |
| 用例名称 | 标记爬虫 |
| 简要描述 | 该用例描述了系统根据获取的数据标记爬虫的过程 |
| 参与者 | 网站管理员 |
| 涉众 | 无 |
| 前置条件 | 已经获得用户请求数据 |
| 后置条件 | 标记成功后保存到数据库中 |
| 基本事件流 | 1. 传入收集到的用户请求信息 2. 根据信息进行频率、合法性等检查以区分爬虫 3. 区分爬虫后，存入MongoDB中。（D-2） |
| 备选事件流 | D-2 同一用户已存在   1. 检查数据是否重复 2. 更新请求频率、访问限制等信息 |
| 补充约束 | 无 |
| 待解决问题 | 无 |

### 设置访问限制

设置访问限制的过程是根据上一步标识的爬虫信息，对请求频率过高的客户端设置访问限制，达到流量控制的目的。

设置访问限制的用例说明见表2.5。

表2.5 设置访问限制

|  |  |
| --- | --- |
| 用例名称 | 设置访问限制 |
| 简要描述 | 该用例描述了系统根据爬虫标记限制访问的过程 |
| 参与者 | 网站管理员 |
| 涉众 | 无 |
| 前置条件 | 已经完成爬虫标记 |
| 后置条件 | 无 |
| 基本事件流 | 1. 传入需要限制的用户数据。 2. 根据数据对请求进行限制。（E-2） |
| 备选事件流 | E-2 同一用户已存在   1. 更新限制信息 |
| 补充约束 | 无 |
| 待解决问题 | 无 |

## 非功能性需求

本项目的非功能性需求主要包含一下几方面：

在正确性需求方面，能对跨浏览器的用户做到识别或近似识别，及时更新用户标识数据。能对高请求频率的爬虫程序进行识别，及时更新爬虫标识数据和限制数据，对爬虫程序设置访问限制。

在安全性需求方面，需要保证除前端数据库外的数据不被用户随意更改。

系统由于采用接入的形式，因此对界面没有特殊需求，可以制作简单的网站用于功能展示。

在稳定性方面，该系统接入网站后，如果系统运行的环境没有发生变化，系统需要保持长期稳定地运行。

## 项目任务

本项目主要包括如下几个任务：

* + - 1. 分析清华大学出版社提出的网站访客识别和智能流控具体需求。
      2. 进行需求建模，设计方案的架构与数据库
      3. 对网站获取到的访客数据进行提取，整理系统需要的信息。
      4. 在提取的信息中进行访客身份的识别，并在数据库中标记。
      5. 设置合适的规律判断方法，对爬虫进行识别。
      6. 根据流量控制要求，对频繁访问的接口的请求次数与频率进行限制。
      7. 交付清华大学出版社运行测试。

# 相关原理与技术

## 依赖管理CDN与NPM

在项目过程中使用了一些第三方库，包括ClientJS、Dexie等，需要引入外部依赖。在开发过程中，主要通过CDN与NPM的方式引入、管理依赖。

CDN又被称为内容分发网络，全称是Content Delivery Network。CDN的原理是当用户发送请求时，从离用户最近的服务器返回用户请求的内容，以此提高项目的加载速度。它的一个好处是可以通过缓存来加载jQuery，减少加载时间。

NPM的全称是Node Package Manager，它是Node.js包含的依赖包管理组件，在安装Node.js时会一起安装。虽然它最初是用于Node.js的包管理组件，能自动处理Node.js版本或部署等很多问题，但是随着NPM的发展，现在还可以用来自动安装与管理JavaScript依赖包，多个依赖包之间的版本关系等也可以实现自动处理。

本项目使用到了request-ip库、ClientJS库、Dexie库等。

request-ip库用于获取request的IP地址，它会在request中寻找特定的headers，并在内容不存在时设置为默认值。通过12个值依次判断ip地址的值，包括X-Client-IP、X-Forwarded-For、CF-Connecting-IP、Fastly-Client-Ip、True-Client-Ip、X-Real-IP、X-Cluster-Client-IP、X-Forwarded, Forwarded-For and Forwarded、req.connection.remoteAddress、req.socket.remoteAddress、req.connection.socket.remoteAddress、req.info.remoteAddress。

ClientJS库用于查找客户端的设备信息并生成数字浏览器指纹。通过本机的JavaScript函数、方法以及一些开源的资源来查找客户端的信息。由于从不同的浏览器和设备中收集用户设备特征值信息会使处理多平台兼容性变得困难，通过使用ClientJS可以简单地创建一个客户端对象，并生成浏览器或设备的指纹ID。

Dexie库是简单易学的IndexedDB封装API，解决了原生IndexedDB的API过于复杂的问题。Dexie具有近乎原生的表现，也支持了批量操作。

## 浏览器指纹Fingerprint

在选择用于区分用户的设备信息时，用到了不同的浏览器指纹作为区分依据。浏览器指纹可以被分为普通指纹和高级指纹，普通指纹一般比较容易获取，但也往往容易被修改，例如HTTP的Header。高级指纹几乎可以直接确定一个唯一的浏览器标识，例如Canvas、AudioContex等。但是这里的普通指纹和高级指纹都是从同一浏览器上获取的，很多特征值在面对不同浏览器的情况下都是不稳定的，例如User Agent、Canvas在相同设备的不同浏览器中会出现不同的值。

跨浏览器指纹Fingerprint的提出就是为了解决跨浏览器识别的问题，理论上同一设备在不同浏览器上也会生成相同或接近的Fingerprint值。Fingerprint获取浏览器能提供的高信息熵、高稳定性的特征值，计算得到一个哈希值。这里的特征值可以是UA、设备时区、定位信息或者使用的语言等。这些特征值具有不同的信息熵，信息熵大的特征值对于唯一确定一位用户更有作用。而将指纹信息综合起来，可以极大地降低特征值之间的重合率，提高客户端唯一标识的准确性。因此往往会在计算Fingerprint时，赋予信息熵较大的特征值更大的权重。

## 前端数据库IndexedDB

在选择前端数据存储时综合对比了LocalStorage、WebSQL与IndexedDB。

虽然LocalStorage使用key-value键值对的形式存储数据，但数据的存储格式都是字符串，因此并不能支持数据的检索，也不能自定义创建索引。另外，LocalStorage的容量根据浏览器的不同在2.5MB 到 10MB之间浮动，标准并不统一而且容量也不大。

另一方面WebSQL已经停止维护，只有Chrome还支持WebSQL，在兼容性上并不好。

综上，采用了更为兼容、容量更大且检索更方便的IndexedDB。IndexedDB为NoSQL型数据库，支持数据的查找、索引等。

NoSQL的全称是Not Only SQL，也被称为非关系型数据库，是相对于传统关系型数据库的另一种数据库的统称。NoSQL对于存储大规模的数据有比较好的支持，因为NoSQL的数据存储没有固定模式，可以十分方便地进行横向扩展。

由于IndexedDB的API十分复杂，并且异常处理等机制并不完善，因此使用Dexie封装的API来简化与完善对IndexedDB的操作。

## 后端Node.js与Express

用户请求数据的获取、整理以及用户请求限制的部分主要在服务端完成。后端采用Node.js，依据CommonJS标准并使用Express框架帮助Node.js开发。

Node.js是一个开源的JavaScript跨平台运行时环境。得益于Node.js，使用JavaScript编写网站前端的开发者除了客户端业务，还可以设计服务器端业务，并且不需要学习其他的服务端语言。Node.js可以自动处理并发连接的问题，而不需要开发者花费精力管理多线程并发。在 Node.js中还可以使用新的ECMAScript标准。

Express是当下最流行的，基于Node.js的Web应用开发框架，它有着灵活、小规模的特点。在提供核心的Web应用功能的同时，Express也保留了Node.js的功能。

CommonJS是一种开发规范，对应Node.js的模块化特性。在Node.js中一个文件就是一个模块，其中定义的函数、变量等都是私有的，对其他模块是不可见的。CommonJS规定模块中的module对象代表当前模块，module的exports方法是会将模块暴露给外部。CommonJS的代码是在模块的作用域内运行的，并且模块只会在第一次加载时运行一次，然后缓存运行结果，以后加载时直接读取缓存。模块还可以根据它在代码中的出现顺序进行多次加载。但是想让模块再次运行，就需要清除缓存。

## 服务器Nginx

服务器端使用了Nginx。Nginx是由俄罗斯开发者设计的一款高性能、轻量级的Web和反向代理服务器，也是一款IMAP/POP3/SMTP邮件代理服务器。Nginx可以在大多数 Unix Linux操作系统上运行，它的特点是内存占用少、并发能力强。Nginx在同类型的网页服务器中有比较好的并发能力，它的稳定性、低系统资源消耗等优点也让它变得十分受欢迎。相比于Apache服务器，Nginx在有处理高连接并发的需要时，将会是一个不错的选择。

## 后端数据库MongoDB

在后端的数据库选择上，为了与前端数据库IndexedDB匹配，采用了同为Nosql型数据库的MongoDB。MongoDB的性质介于关系型数据库和非关系型数据库之间，相比于其他非关系数据库，MongoDB功能十分丰富，也十分接近关系型数据库。MongoDB的数据存储形式采用了分布式文件存储，因此对于Web应用可以方便地进行扩展并保持高性能。MongoDB数据存储的文档类似于JSON对象，数据结构由键值key-value对组成，其中value值可以包含其他文档、数组及文档数组等。

## 其他

使用GitHub完成项目代码、说明文档和版本的管理。

# 项目总体设计

## 系统工作流程

系统的工作流程整体上为功能的顺序执行，在用户开始访问网站时，记录用户的标识信息，并基于用户的身份标识信息，通过计算完成用户身份的识别并完成数据的前后端存储。当用户发送请求时，收集用户的请求信息，包括header、时间、请求的接口等，将请求信息中的身份信息与所记录的进行对比，并根据请求的时间计算单位时间内的请求频率，标记请求频率超过限制的客户端并更新数据库标识、限制信息。根据限制信息对客户端设置访问限制。

系统整体的流程图如图4.1所示。

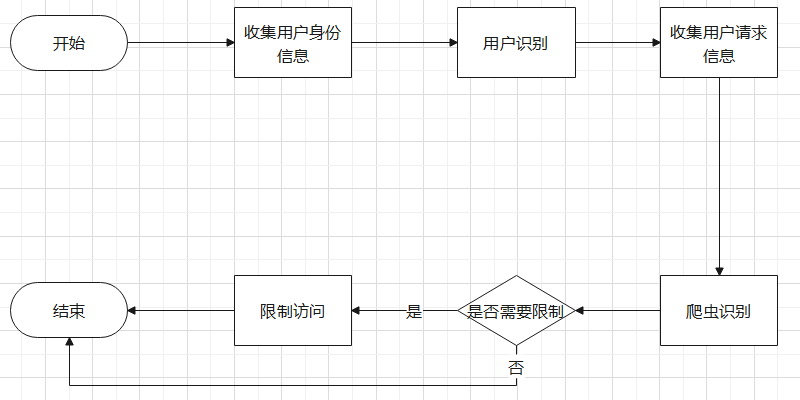


图4.1 系统流程图

## 系统架构设计

本系统在架构设计上采用了三层结构，包括接入层、业务层和数据层。系统架构图如图4.2所示。

由于本系统不需要前端展示层，因此这里用接入层表示。接入层中为系统接入网站的部分。

业务层主要包括记录用户标识数据、标识用户、记录请求数据、标识爬虫、设置访问限制这五项功能的实现。其中，可以进一步区分为在前端通过JavaScript实现的记录用户标识数据和标识用户部分，以及在后端通过Express框架下的Node.js实现的记录请求数据、标识爬虫、设置访问限制部分，其中设置访问限制也用到了部分服务器Nginx的功能。前后端的数据通信则使用了Axios实现，它是一个可以在浏览器和Node.js中使用的HTTP库，基于promise并且可以自动转换JSON格式的数据。主要使用了其中的post与get方法。

数据层主要包括对前端数据库IndexedDB与后端数据库MongoDB封装的增删改查等操作的类，数据的格式也统一为JSON。对MongoDB的操作通过后端的Node.js完成，对IndexedDB的操作在前端通过JavaScript完成。

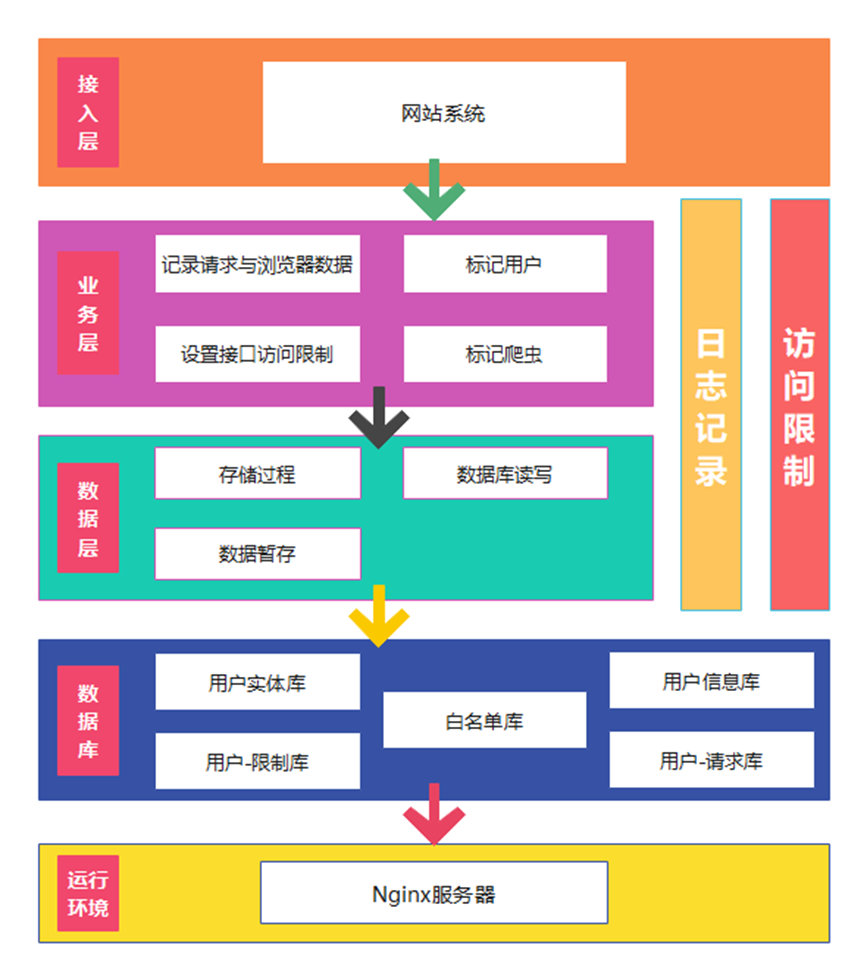


图4.2 系统架构图

## 数据库设计

### 数据库逻辑设计

根据需求提取出了数据库实体包括用户标识、用户信息、用户请求、用户限制，进一步挖掘实体中的关系，可得数据库ER图，如图4.3所示。

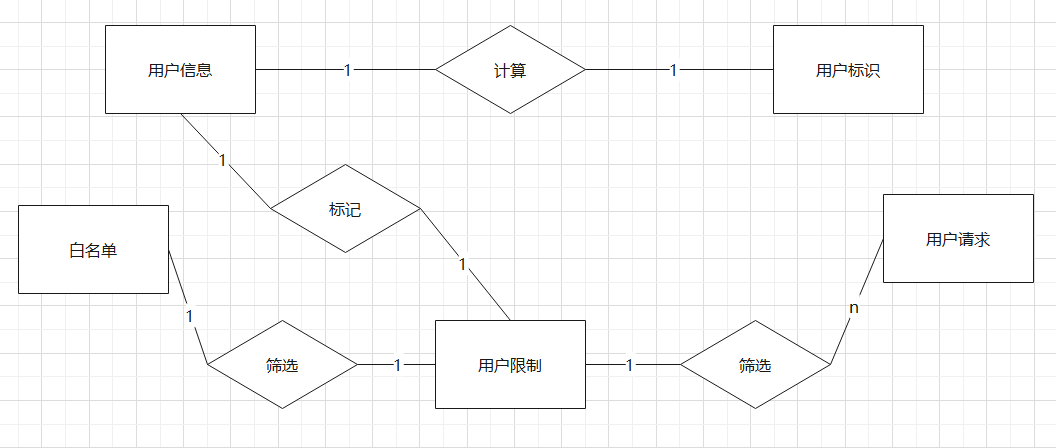


图4.3 数据库ER图

虽然用户限制表和用户信息存在实体联系，但是在实际使用过程中用户信息对用户限制表并非必需，且在使用时仅是辅助检测，因此在具体实现中不必体现两者的联系。经过转化后的数据库类图如图4.4所示，其中，省略了数据库属性内容，将会在后续的数据库表中详细说明。

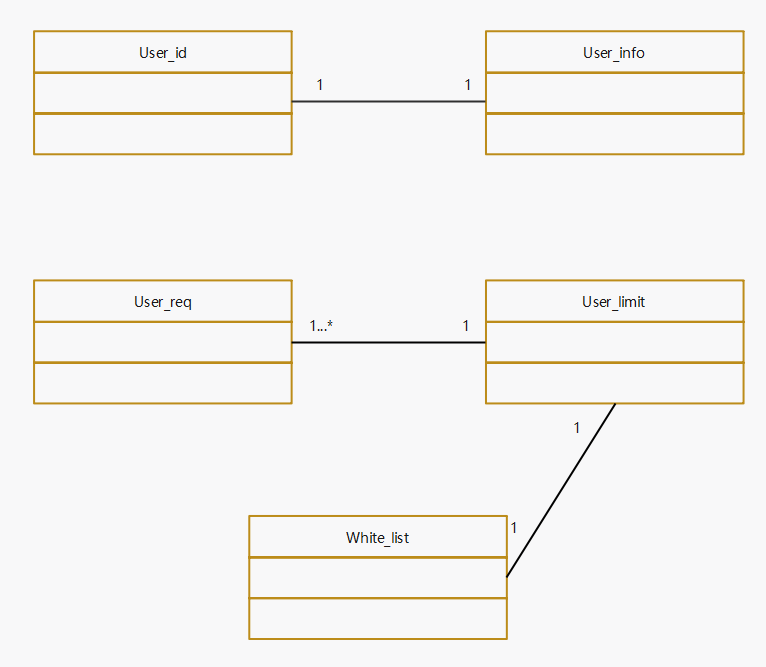


图4.4 数据库类图

### 数据库表设计

根据数据库ER图与数据库类图，设计系统的数据库表，如至表所示。

表4.1 用户标识表

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 意义 |
| Fingerprint | Number | 唯一表示用户客户端的至多16位数字 |
| IP | Array | 客户端IP地址池数组，每个IP地址的类型为String |

用户标识表记录了根据客户端信息计算而得出的唯一Fingerprint值，以及该Fingerprint值下所使用到的所有IP地址值，便于在需要时解决使用IP池的爬虫的问题。

表4.2 用户信息表

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 意义 |
| Fingerprint | Number | 唯一表示客户端身份的至多16位数字 |
| User Agent | String | 客户端浏览器User Agent字段的值 |
| CPU | String | 客户端CPU类型 |
| screenPrint | String | 客户端屏幕信息，包括分辨率、色深等 |
| colorDepth | Number | 客户端屏幕色深 |
| availableResolution | String | 客户端屏幕可用分辨率 |
| mimeTypes | String | 客户端多功能网络邮件扩展类型 |
| fonts | String | 客户端安装的字体 |
| timeZone | String | 客户端所在时区 |
| language | String | 客户端用户语言 |
| core | Number | 客户端CPU核心数 |
| screenOrientation, | String | 客户端屏幕方向 |
| screenAngle | Number | 客户端屏幕倾斜角度 |
| screenHeight | Number | 客户端屏幕竖向分辨率 |
| screenWidth | Number | 客户端屏幕横向分辨率 |

用户信息表记录了被唯一的Fingerprint所标识的客户端详细信息，在必要时可以结合IP池、Fingerprint数值接近程度等进一步判断是否为同一用户。

表4.3 用户请求表

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 意义 |
| IP | String | 发出请求的客户端IP地址 |
| API | String | 客户端请求的接口名称 |
| Time1 | Date | 接收到客户端请求的时间，仅记录一分钟内的最多10次时间 |
| Time2 | Date | 接收到客户端请求的时间，仅记录一分钟内的最多10次时间 |
| … | … | … |

用户请求表记录了发送请求的用户的IP地址以及所请求的接口，同时记录了请求的时间，用于计算每分钟用户请求的频率。

表4.4 用户限制表

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 意义 |
| IP | String | 需要设限的客户端IP地址 |
| Time | Date | 客户端访问限制的解除时间 |

用户限制表记录了被设限的客户端IP地址，当用户请求表中计算得到的每分钟请求频率高于10次时将会加入用户限制表，并将Time字段设置为一分钟后的时间，限制来自该IP地址的请求频率为每分钟1次。

表4.5 白名单表

|  |  |  |
| --- | --- | --- |
| 字段 | 类型 | 意义 |
| IP | String | 不受访问频率限制的客户端IP地址 |

白名单表记录了不受访问频率限制的客户端IP地址，在设限前需要过滤掉白名单中的IP地址。

# 项目具体实现

## 项目实现概述

用户识别部分采用IP地址与浏览器Fingerprint相结合的方式。将同Fingerprint不同IP地址的字段进行整合，看作是同一用户，用以解决IP池的问题。通过timezone、fonts等其他信息可以做更精确的用户区分。

数据以JSON的形式保存在NoSQL型数据库中，前端为IndexedDB，后端为MongoDB。

在后端采用Node.js，对MongoDB的增删改查操作进行了封装，便于数据的整理。

对于爬虫检测的部分，采用设置阈值、HTTP请求检测的方式。在Node.js端对容易被频繁起请求的接口处接入爬虫检测。通过记录时间段内的请求次数以及最近的访问时间，将访问过于频繁的用户进行标记。对于请求中User Agent、platform等信息与记录不匹配的用户也进行标记。将受标记的用户与白名单用户进行匹配，去除需要放行的用户后，对于其他受标记的用户进行频率与流量速率限制。

为了加快请求响应速度与访问限制的灵活性，将影响严重的用户加入Nginx的blockip配置文件中，并适时进行更新。

在访问限制部分，过时的用户代理、过时的浏览器版本、不合理的海外IP、不合理的Referer等会被特别标记，在需要时可以在这些方面设置更为严格的限制策略。

（undone，细化）

## 用户识别部分

为了实现跨浏览器识别，最为直接的方法是获取到用户设备的唯一标识，这样就可以解决用户在同一台设备的不同浏览器上访问网站的识别问题。MAC地址就是这样一种设备的唯一标识，但是由于MAC地址只能在IE浏览器中的ActiveX控件中获得，IE浏览器的普及性与使用率十分低，并且一般情况下网站并没有方法获取到准确的用户设备的MAC地址，便放弃MAC地址识别的方法，转而使用一些可以通过JavaScript等方式获得的用户特征值作为识别用户的依据，也就是浏览器追踪技术。

浏览器追踪技术到目前已经进入3代。第一代主要根据用户状态区分用户，包括对比用户的cookie、evercookie等，这些信息往往需要用户登录后才能被有效地获取。第二代追踪技术出现了浏览器指纹的概念，用户的区分主要是通过使用不同的浏览器特征值，例如UA、浏览器插件信息等。第三代追踪技术主要通过收集用户的行为或习惯特征，并建立用户的特征模型来实现不受设备、浏览器限制的用户人追踪。目前第三代追踪技术的实现比较复杂，还在进一步探索中。

由于第三代追踪技术需要与网站业务有深度的结合，因此选择2.5代追踪技术。2代与3代之间的2.5代追踪技术浏览器Fingerprint专注于解决跨浏览器识别用户的问题。将浏览器特征值与权重结合计算后可以生成一个Fingerprint哈希值，用于标记不同的用户，这里采用了ClientJS提供的Fingerprint计算方法，通过clientjs.getFingerprint()方法获取到Fingerprint的哈希值。ClientJS的计算方法中采用的特征值包括：user agent、screen print、color depth、current resolution、available resolution、device XDPI、device YDPI、plugin list、font list、local storage、session storage、timezone、language、system language、cookies、canvas print，共16种。

根据Crossing Browser Tracking的研究，各个特征值在单浏览器下的信息熵和跨浏览器下的信息熵与稳定性如表3.1所示。

表5.1 特征值信息熵与跨浏览器稳定性

|  |  |  |  |
| --- | --- | --- | --- |
| 特征值 | 单浏览器 | 跨浏览器 | |
| 信息熵 | 信息熵 | 稳定性 |
| User Agent | 6.71 | 0.00 | 1.39％ |
| Accept | 1.29 | 0.01 | 1.25％ |
| Content encoding | 0.33 | 0.03 | 87.83％ |
| Content language | 4.28 | 1.39 | 10.96％ |
| List of plugins | 5.77 | 0.25 | 1.65％ |
| Cookies enabled | 0.00 | 0.00 | 100.00％ |
| Use of local/session storage | 0.03 | 0.00 | 99.57％ |
| Timezone | 3.72 | 3.51 | 100.00％ |
| Screen resolution and color depth | 7.41 | 3.24 | 9.13％ |
| List of fonts (Flash) | 2.40 | 0.05 | 68.00％ |
| List of HTTP headers | 3.17 | 0.64 | 9.13％ |
| Platform | 2.22 | 1.25 | 97.91％ |
| Do Not Track | 0.47 | 0.18 | 82.00％ |
| Canvas | 5.71 | 2.73 | 8.17％ |
| WebGL Vendor | 2.22 | 0.70 | 16.09％ |
| WebGL Renderer | 5.70 | 3.92 | 15.39％ |
| Use of an Ad blocker | 0.67 | 0.28 | 70.78％ |
| Screen Ratio | 1.40 | 0.98 | 97.57％ |
| List of fonts (JavaScript) | 10.40 | 6.58 | 96.52％ |
| AudioContext | 1.87 | 1.02 | 97.48％ |
| CPU Virtual cores | 1.92 | 0.59 | 100.00％ |

用户识别部分的详细流程图如图3.1所示。

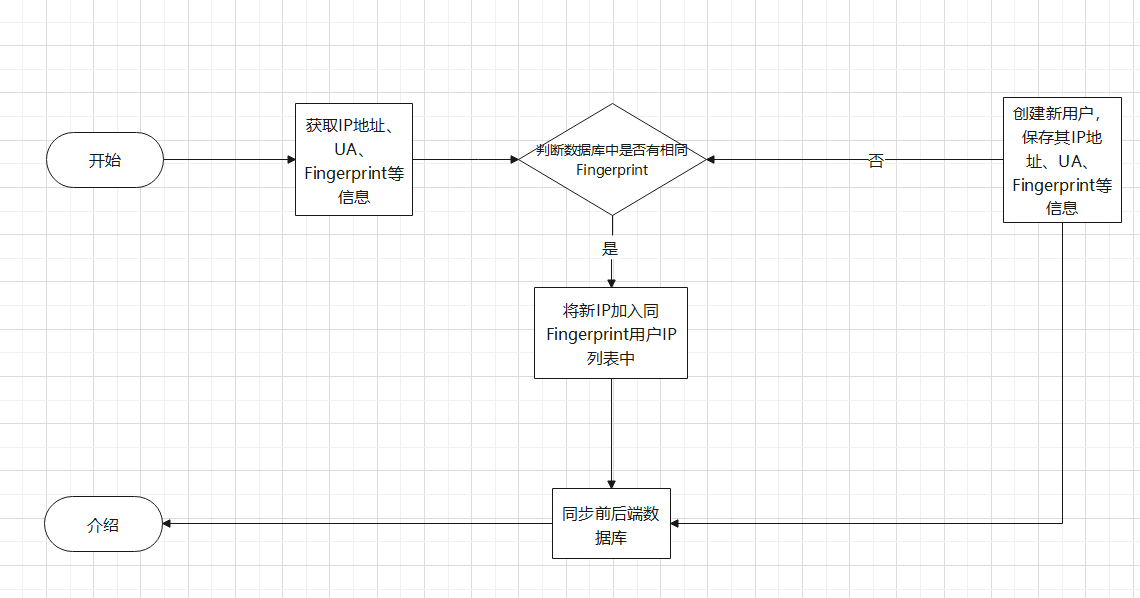


图5.1 用户识别流程图

## 爬虫识别部分

本项目的爬虫识别方式主要为特征识别以及频率控制。特征识别反爬虫是指通过客户端的特征、属性来区分正常用户和爬虫程序。客户端的属性值可以作为特征，并不等同于服务端可以通过某一特定的属性值来唯一确定用户身份，属性值只能作为判断用户身份的依据之一。频率控制则是根据同一ip在单位时间内访问单个接口的次数或一段时间内访问的总次数对请求进行限制。

Navigator对象的cookieEnable、 platform、 plugins等属性和Screen对象的orientation等属性可以作为判断依据。通过将浏览器请求headers中的User Agent值与JavaScript获取到的navigator.userAgent 属性值进行对比，可以判断客户端是否随机切换User Agent。如果User Agent值不同，可以对当前的客户端进行标记。User Agent中还包括了操作系统信息，结合navigator. platform提供的操作系统信息，如果两者的属性值不同，也可以标记当前的客户端。

类似地，屏幕分辨率、CPU核心数等也可以作为客户端的特征值。除非客户端计算机运行在虚拟机中或是年代久远，否则JavaScript获取到的hardwareConcurrency，即计算机核心数量一般为2个以上。

在使用不同的渲染工具时，浏览器的插件数量往往也不相同，因此这个属性值也可以作为一种客户端的特征信息。

由于通过JavaScript可以更改上述的部分属性值，所以这种特征识别方式得到的结果也并不是完全可靠的。因此在编码时仅做标记用于参考。主要限制访问频率过高的爬虫程序。

访问频率指的是单位时间内客户端向服务器端发出网络请求的次数，用于表示网络请求的频繁程度。正常用户浏览网页时发出请求的频率不会像爬虫那么高，可以将访问频率过高的用户标记为爬虫。

限制爬虫请求频率的第一步是要确定客户端的身份标识，这里使用IP地址作为主要依据。然后根据客户端的身份标识记录此客户端在单位时间内的请求次数，拒绝请求次数过多的客户端发出的请求。

在项目中通过在Node.js端记录用户请求次数以及请求的接口，计算出时间段内请求的频率，对请求过多的用户进行限制。

## 请求限制部分

根据沟通，初步采用的流量控制方案为：某一个接口请求太频繁或者总请求次数过多则限制访问一段时间。每分钟请求接口超过10次则限制其访问1分钟、每2个时总请求数大于1000次则限制其访问频率为每10秒一次等。

（undone，更新最终版）

在Node.js端完成这部分工作，连接了MongoDB数据库用于统计每个接口的请求次数、总请求次数以及当前的限制情况。

将IP地址信息与此前记录的用户信息进行匹配，通过这种方式对使用IP池的爬虫进行标记，可以用于更好地区分爬虫与普通用户。

爬虫识别与请求限制部分的详细流程图如图3.2所示。

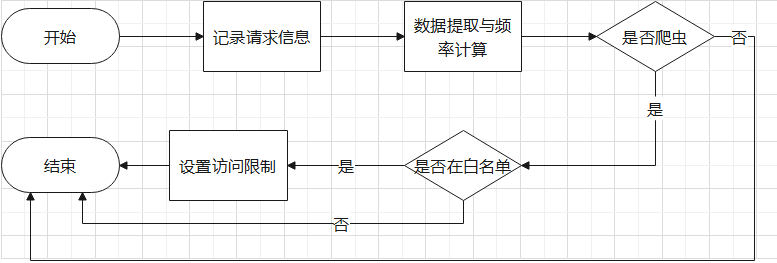


图5.2 爬虫识别与请求限制流程图

参考文献

1. Edward Roberts. Bad Bot Report 2020: Bad Bots Strike Back.[R]. San Mateo: Imperva, 2020.
2. 林中明. 基于Hadoop的Web用户识别与新闻智能推荐算法研究[D]. 郑州：战略支援部队信息工程大学，2018.
3. 伏康，杜振鹏. 网站反爬虫策略的分析与研究[J]. 电脑知识与技术，2019(28)：28-30.
4. 张晔，孙光光，徐洪云，庞婷，曲潇洋. 国外科技网站反爬虫研究及数据获取对策研究[J]. 竞争情报，2020(01)：24-28.
5. Y. Liu, Z. Yang, J. Xiu, C. Liu. Research on an anti-crawling mechanism and key algorithm based on sliding time window[A]. In: 2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS)[C]. Beijing: IEEE, 2016: 220-223
6. W. Zhu, J. Qin, R. Kong, H. Lin, Z. He. A System Framework for Efficiently Recognizing Web Crawlers[A]. In: IEEE SmartWorld 2018 Organizing and Program Committees. 2018 IEEE SmartWorld [C]. Guangzhou: IEEE, 2018: 1130-1133
7. H. Wang, C. Li, L. Zhang, M. Shi.Anti-Crawler strategy and distributed crawler based on Hadoop[A]. In: 2018 IEEE 3rd International Conference on Big Data Analysis(ICBDA)[C]. Shanghai: IEEE, 2018: 227-231
8. P. Lewandowski, M. Janiszewski, A. Felkner. SpiderTrap—An Innovative Approach to Analyze Activity of Internet Bots on a Website[J]. IEEE Access, 2020, 8: 141292-141309
9. G. Neelima, S. Rodda. Predicting user behavior through sessions using the web log mining[A]. In: 2016 International Conference on Advances in Human Machine Interaction (HMI) [C]. Doddaballapur: IEEE, 2016:1-5
10. 刘洋. 基于网页浏览行为的反爬虫研究[J]. 现代计算机(专业版)，2019(07)：50-60+70.
11. Cao, Y., Li, S., & Wijmans, E. (Cross-)Browser Fingerprinting via OS and Hardware Level Features[R]. San Diego: NDSS, 2017.