目 录

[摘 要 3](#_Toc160392926)

[第一章 绪论 6](#_Toc160392927)

[1.1研究背景与意义 6](#_Toc160392928)

[1.2国内外研究现状 6](#_Toc160392929)

[1.3本文研究内容 7](#_Toc160392930)

[1.4论文结构 7](#_Toc160392931)

[第二章 相关技术介绍 8](#_Toc160392932)

[2.1 大数据与分布式架构理论介绍 8](#_Toc160392933)

[2.2 Hadoop与spark 8](#_Toc160392934)

[2.3 Spider爬虫技术 9](#_Toc160392935)

[2.4 Django框架 9](#_Toc160392936)

[2.5 Vue和Echart 9](#_Toc160392937)

[第三章 需求分析 11](#_Toc160392938)

[3.1系统功能分析 11](#_Toc160392939)

[3.1.1系统功能介绍 11](#_Toc160392940)

[3.1.2系统主要模块设计 11](#_Toc160392941)

[3.2系统流程分析 12](#_Toc160392942)

[3.3可行性分析 12](#_Toc160392943)

[3.4 非功能性需求分析 12](#_Toc160392944)

[3.3主要流程分析 13](#_Toc160392945)

[3.4本章小结 14](#_Toc160392946)

[第四章 系统设计 15](#_Toc160392947)

[4.1 系统总体架构设计 15](#_Toc160392948)

[4.2 前台页面设计 15](#_Toc160392949)

[4.3系统流程设计 15](#_Toc160392950)

[4.3.1 用户注册与登录流程设计 16](#_Toc160392951)

[4.3.2 学习资源推荐流程设计 16](#_Toc160392952)

[4.3.3收藏占比与类别统计流程设计 16](#_Toc160392953)

[4.4 数据库设计 16](#_Toc160392954)

[4.5 本章小结 17](#_Toc160392955)

[第五章 系统实现 18](#_Toc160392956)

[5.1 系统的配置和部署 18](#_Toc160392957)

[5.2 数据爬取与存储功能实现 19](#_Toc160392958)

[5.2.1菜鸟教程网爬取页面分析 19](#_Toc160392959)

[5.2.2菜鸟教程网数据爬取 19](#_Toc160392960)

[5.2.3数据上传到hadoop中 20](#_Toc160392961)

[5.3 首页功能实现 21](#_Toc160392962)

[5.4 系统用户注册与登录功能实现 22](#_Toc160392963)

[5.4.1用户注册界面 23](#_Toc160392964)

[5.4.2用户登录界面 23](#_Toc160392965)

[5.5学习资源推荐功能实现 24](#_Toc160392966)

[5.6数据管理功能实现 24](#_Toc160392967)

[5.7本章小结 25](#_Toc160392968)

[第六章 系统测试 26](#_Toc160392969)

[6.1 系统测试目的 26](#_Toc160392970)

[6.2 系统兼容性测试 26](#_Toc160392971)

[6.3 功能性测试 26](#_Toc160392972)

[6.4 本章小结 27](#_Toc160392973)

[结 论 28](#_Toc160392974)

# 摘 要

随着互联网的飞速发展，从网络获取学习学习资源已成为重要的一种途径。为了满足用户对个性化学习资源的需求，开发了一款基于大数据的学习资源推送系统。该系统充分利用Python编程语言、MySQL数据库以及Hadoop、Spark等大数据技术，对海量学习资源数据进行高效处理和分析。系统采用可视化Web界面，首先通过爬虫技术抓取学习资源内容、分类等数据。由于数据量庞大，使用Pandas进行初步的数据处理和存储，然后通过Hadoop分布式文件系统将数据传输至MySQL数据库管理系统。核心的分布式计算部分利用了Spark的强大处理能力，进行特征工程和机器学习算法的训练。最终，通过Django和Vue.js框架将推荐结果呈现给用户。

该系统的研究意义在于，通过爬虫技术成功抓取了菜鸟教程网的学习资源数据，经过清洗和特征工程处理后，采用协同过滤算法进行机器学习训练。这种学习资源推荐方法不仅提高了学习资源推荐的准确性和个性化程度，还为用户提供了更加便捷、高效的学习资源查看体验。此外，该系统的实现为大数据技术在学习资源推荐领域的应用提供了有益的探索和实践经验。

关键字：学习资源推荐；爬虫；python；Django；Spark；

作 者： 李浩喆

指导教师： 任艳

Abstract

With the rapid development of the Internet, obtaining learning resources from the network has become an important way. In order to meet users' needs for personalized learning resources, a learning resource push system based on big data has been developed. The system makes full use of Python programming language, MySQL database, Hadoop, Spark and other big data technologies to efficiently process and analyze massive learning resource data. The system adopts the visual Web interface, which first captures the learning resource content and classification data through crawler technology. Due to the huge amount of data, Pandas is used for preliminary data processing and storage, and then the data is transferred to the MySQL database management system through the Hadoop distributed file system. The distributed computing part of the core takes advantage of Spark's powerful processing power to train feature engineering and machine learning algorithms. Finally, the recommended results are presented to the user through the Django and Vue.js framework.

The research significance of this system lies in the learning resource data of Cainiao tutorial network through crawler technology. After cleaning and feature engineering processing, the collaborative filtering algorithm is used for machine learning training. This learning resource recommendation method not only improves the accuracy and personalization of learning resource recommendation, but also provides users with a more convenient and efficient learning resource viewing experience. In addition, the implementation of this system provides a useful exploration and practical experience for the application of big data technology in the field of learning resource recommendation.

Keywords:Learning Resource Recommendation; Crawler; Python; Django; Spark.

Written by Li Haozhe

Supervised by Ren Yan

绪论

1.1研究背景与意义

随着互联网的飞速发展，人们从网络上获取学习资源的途径变得越来越广泛。然而，如何在海量的学习资源中快速找到适合自己的内容成为了一个巨大的挑战。传统的资源推荐方法往往基于简单的内容过滤或用户行为统计，难以满足用户对个性化、精准推荐的需求。因此，开发一款基于大数据的学习资源推送系统显得尤为重要。

大数据技术为解决这一问题提供了有力支持。通过高效处理和分析海量数据，可以深入挖掘用户的学习习惯、兴趣偏好等信息，从而实现更加精准的资源推荐[1]。本研究旨在利用Python编程语言、MySQL数据库以及Hadoop、Spark等大数据技术，开发一款能够满足用户个性化学习资源需求的产品。系统除了可以让用户查看各类的菜鸟教程网各类学习资源的类别统计图，还能查看各类语言热度排行等图表数据。另外系统还可以提高学习资源推荐的准确性和个性化程度，为用户提供更加便捷、高效的学习资源查看体验。

1.2国内外研究现状

随着互联网的快速发展，学习资源的获取和学习方式的个性化需求逐渐凸显。国内外研究者针对如何高效、精准地为用户推荐学习资源进行了广泛研究。在国外，研究者们主要聚焦于推荐算法的创新和优化。协同过滤算法是最为经典的方法之一，它基于用户或物品的相似性进行推荐。近年来，深度学习算法在推荐系统中的应用也取得了显著成果，如基于神经网络的推荐系统可以更好地捕捉用户行为和资源特征的复杂关系[2]。

在国内，随着大数据技术的不断发展，越来越多的研究开始关注如何利用大数据进行学习资源的推荐。例如，有研究利用爬虫技术抓取网络学习资源，并通过数据挖掘和机器学习算法进行推荐。此外，国内研究者还对推荐系统的用户满意度、个性化程度等方面进行了深入探讨[3]。尽管如此，现有的研究仍存在一定局限性。一方面，大多数研究仅关注单一的推荐算法或技术，而缺乏对多种算法的融合和优化；另一方面，现有研究对学习资源的质量和权威性评估不足，可能导致推荐结果的质量参差不齐[4]。综上所述，基于大数据的学习资源推送系统在国内外均受到广泛关注，但仍存在诸多挑战和待解决的问题。本研究旨在突破现有研究的局限，通过整合多种技术和算法，提高学习资源推荐的准确性和个性化程度，为用户提供更加优质的学习资源推荐服务。

1.3本文的所作工作

基于大数据的学习资源推送系统的研究内容主要包括以下几个方面：

（1）爬虫程序: 开发高效、稳定的爬虫程序是学习资源推荐系统的第一步。此部分的目标是从菜鸟教程网上抓取学习资源数据，包括内容、分类、收藏数等。考虑到学习资源网站的更新频率和反爬策略，需要设计合理的抓取策略，确保数据的实时性和完整性。

（2）数据存储: 面对海量的学习资源数据，如何高效地存储和处理数据是关键。系统采用Pandas进行初步的数据处理和存储，然后通过Hadoop分布式文件系统（HDFS）进行大规模数据的存储。这种组合方式既保证了数据处理的速度，也确保了数据存储的稳定性[6]。

（3）机器学习: 为了实现个性化学习资源推荐，系统利用Spark进行大规模的分布式计算，进行特征工程和机器学习算法的训练。特别地系统采用协同过滤算法，基于用户的行为和学习资源相似度进行推荐。这种方法可以进一步提高学习资源推荐的准确性和个性化程度。

（4）Web可视化界面显示: 为了给用户提供直观、友好的界面，系统采用Django和Vue.js框架。通过这两种框架，用户不仅可以查看推荐学习资源，还可以对推荐结果进行反馈，形成一个互动的学习资源推荐体验。

本项目在开发和设计过程中涉及到原理和技术有: python等等；将按以下章节进行开发设计；

第一章绪论；剖析项目背景,说明研究的内容。

第二章开发技术。系统主要使用python等技术，并对此做了介绍。

第三章系统分析；包罗了系统总体结构、对系统的性能、功能、流程图进行了分析。

第四章系统设计；对软件功能模块和数据库进行详细设计。

第五章系统总体设计；对管理员、商家和用户的功能进行描述，

第六章对系统进行测试，

总结；在论文最后结束章节总结了开发这个系统和撰写论文时候自己的总结、感想,包括致谢。



相关技术介绍

系统是服务器端采用python编写的Web企业级项目，前台开发工具通过HBuilder开发，后台使用pycharm集成环境开发。根据web项目的一般性技术要求主要包括，前台技术，后台数据处理，爬虫数据获取技术等。前台技术主要是通过vue-cli结合Echart实现可视化界面，通过axios异步与后台沟通获取数据库数据，后台数据存储和计算使用hadoop框架，spark框架和mysql框架，数据的获取使用request库先爬取数据，并使用BeautifulSoup进行网页的数据的解析，通过pandas进行文本数据的存储[7]。以下介绍主要的技术。

2.1 大数据与分布式架构理论介绍

如今是一个大数据的年代，大数据（big data）有重要的5个特点（5V）包括：多样性（Variety）、体量（Volume）、速度（Velocity）、准确性（Veracity）和价值（Value）。多样性是指大数据包含多种类型和来源的数据，包括结构化、半结构化和非结构化数据。体量是指大数据具有大规模的数据集，通常在PB级别甚至更大。速度是指大数据需要在合理的时间内处理和分析数据，以产生有价值的洞察。准确性是指大数据的真实性和可信度，需要通过数据清洗和验证等手段来保证数据的准确性。价值是指大数据可以为企业和组织带来更高的商业价值，通过优化决策、提高效率和降低成本等方式实现[8]。

然而要处理这巨大的数据涉及到数据的存储和计算，这样就产生了分布式架构理论。分布式理论表面很简单，就是通过增加服务器的数量来处理庞大的数据。首先对数据或者计算进行有效的切割，然后再对这些数据或者计算进行多个服务器的存储和计算，通过这样的方式来增加数据存储和计算的效能，提升数据的处理能力。

2.2 Hadoop与spark

早期大数据技术主要集中在大企业中应用，随着Hadoop的开源，大数据技术得到了前所未有的发展。Hadoop和Spark是当前大数据领域中两种重要的技术框架，它们分别提供了不同的数据处理和分析能力，被广泛应用于数据仓库、数据挖掘、机器学习等众多领域。

Hadoop是一个分布式计算框架，它通过将大规模数据分散到大量计算节点上进行处理，实现了高效的数据处理能力。Hadoop具有简单易用的接口和灵活的数据存储方式，可以处理各种类型的大数据，如结构化数据、半结构化数据和非结构化数据。同时，Hadoop还提供了数据安全性和数据容错性方面的保障，保证了数据处理的高效性和可靠性。

Spark则是另一个重要的数据处理和分析框架，它采用了内存计算技术，通过将数据加载到内存中，实现了高速的数据处理能力。Spark具有高效的数据处理速度和灵活的数据分析能力，可以处理各种类型的大数据，如结构化数据、半结构化数据、日志文件等。同时，Spark还提供了丰富的数据分析工具和可视化工具，方便用户进行数据分析和数据展示[9]。

2.3 Spider爬虫技术

爬虫技术是一种自动化数据采集技术，通过编写程序模拟人类访问网站的行为，可以快速、准确地获取大量数据。在大数据时代，爬虫技术得到了广泛的应用和推广。爬虫技术是一种自动化数据采集技术，通过编写程序来模拟人类访问网站的行为，自动抓取网站上的信息。爬虫技术可以用于各种场景，例如数据挖掘、竞争情报、用户行为分析等。爬虫的工作原理通常是从一个或多个初始页面开始，根据一定的规则（如链接、关键字等）遍历网站上的所有页面，提取其中的数据并按照一定的格式进行存储。在提取数据的过程中，爬虫需要遵循网站的robots.txt协议和反爬虫策略，以确保采集数据的合法性和安全性。简单的信息的爬取可以使用python的第三方库比如request，urllib等，然而随着越来越多网站设置了反扒，数据的爬取也变得困难。本系统使用的爬虫技术主要是使用request库先爬取数据，并使用BeautifulSoup进行网页的数据的解析，最后通过pandas技术实现csv文本格式的存储[10]。

2.4 Django框架

Django简单来说就是python的应用的web框架，它拥有和springMVC类似的工作原理，提供MVT框架模式以提高系统的开发效率。M层就是模型层，在java应用中需要mybatis这样的ORM（对象模型映射）框架，但是Django内置了ORM框架的类库，这样就省去了额外的配置，减少了使用上的困难。另外，值得一提的是，框架对controller控制层进一步进行了封装，成为了更符合开发模式的路由，由路由来分发具体的操作请求，主要通过在相应的配置文件中进行配置，所以通过使用Django就能快速搭建一个web系统服务器[11]。

2.5 Vue和Echart

Vue和ECharts在可视化系统中的应用是一个非常有趣且实用的工具。Vue是一个流行的JavaScript框架，它提供了丰富的工具和组件，使得开发者可以轻松地构建用户界面。而ECharts则是一个使用JavaScript实现的，开源的可视化库，它提供了大量强大的图表类型和配置选项，可以帮助开发者创建出丰富而直观的数据可视化效果。在可视化系统中，Vue和ECharts的结合可以带来许多优势[11]。首先，Vue提供了强大的组件系统，可以轻松地创建复杂的用户界面，而ECharts则可以用来创建各种数据可视化图表。通过将这两个工具结合起来，开发者可以创建一个功能丰富、交互性强、用户体验良好的可视化系统[12]。

具体来说，Vue和Echart可以在以下几个方面帮助开发者实现可视化系统：

（1）创建复杂的用户界面：Vue提供了大量的组件和布局选项，可以帮助开发者创建各种复杂的用户界面。通过使用Vue，开发者可以轻松地实现数据可视化的各种功能，如数据展示、交互操作、动态更新等。

（2）高效的数据处理：Vue提供了强大的数据绑定和响应式系统，可以帮助开发者高效地处理数据。通过将ECharts图表与Vue的数据绑定起来，开发者可以轻松地实现数据的实时更新和动态交互。

（3）易于扩展和维护：Vue的组件化和模块化架构使得代码易于扩展和维护。开发者可以通过使用第三方插件和组件库来扩展ECharts的功能，同时也可以轻松地管理和维护整个系统的代码。

而ECharts在可视化系统中的作用则主要体现在以下几个方面：

（1）提供丰富的图表类型和配置选项：ECharts提供了大量的图表类型和配置选项，如折线图、柱状图、饼图、散点图等，可以帮助开发者根据不同的数据需求创建出各种直观的数据可视化效果。

（2）高度定制化：ECharts提供了丰富的配置选项，可以帮助开发者根据具体需求对图表进行高度定制化。例如，可以通过调整颜色、字体、动画效果等来增强图表的可视化效果。

总的来说，Vue和ECharts的结合可以为可视化系统带来许多优势，包括创建复杂的用户界面、高效的数据处理、易于扩展和维护以及提供丰富的图表类型和配置选项等。这种结合对于数据分析和数据可视化的应用场景非常有用，可以帮助开发者更好地理解和呈现数据，提高数据的使用效率和用户体验[14]。

# 需求分析

本章主要首先对系统开发的可行性进行分析，然后再对整体的系统开发流程和用户注册登录流程以及学习资源推荐功能流程进行分析。

3.1系统功能分析

系统的功能主要包括三个方面。首先是需要从菜鸟教程网站爬取到相应的数据，这些数据包括有学习资源分类信息，学习资源信息等。其次是将这些数据通过hadoop的HDFS组件存储到服务器的mysql中，通过pyspark对数据进行分布式计算处理。最后通过django搭建的web页面进行数据的可视化展示，在页面中也需要完成学习资源推荐功能，推荐的方式主要是通过用户点击的学习资源推荐相类似的学习资源，其次就是根据像是其他用户常看的学习资源推荐给相关用户。

根据以上的功能需求情况，整体的功能模块包括有前台vue项目模块，后台django后台项目模块和爬虫模块。前台vue的页面主要页面包括注册与登录页面，数据可视化展示页面，爬虫模块主要用来爬取菜鸟教程网的相关数据信息的，通过使用hadoop进行数据的存储，django后台用来提供前台所用的json数据以及给出推荐的相关的学习资源信息。其中学习资源推荐模块的实现是基于机器学习功能之后的应用阶段。系统功能模块图，如图3.1：



图3.1系统功能模块图

3.2 系统用例分析

3.2.1 用户管理的流程

用户注册、信息验证、分配权限，形成用户档案。后续可进行信息修改、权限调整或账号注销，确保用户信息的准确性与安全性。



图3.2 用户管理流程图

3.2.2 个人中心管理流程

 个人中心管理流程：用户登录后，进入个人中心，可查看个人信息、历史记录及操作日志，进行资料编辑或设置偏好，实现个性化管理。



图3.3 个人中心流程图

3.2.3 登录流程

登录流程：用户输入账号和密码，系统验证信息正确性，成功则进入主界面，失败则提示错误信息并要求重新输入。



图3.4 登录流程图

3.3可行性分析

基于大数据的学习资源推送系统具有可行性，具体分析如下：

（1）技术可行性：菜鸟教程网学习资源的内容、分类、收藏数等数据可以通过爬虫技术获取，并通过hadoop存储到数据库中，分布式计算使用spark技术。其次其它的技术包括有vue、django、python语言等都是学校学习过的，如果不懂也是可以通过网络或者查看文献的方式解决[15]。

（2）经济可行性：系统的开发使用的基本都是开源的软件，花费主要是个人的精力，不需要额外开支。另外如果实现商用，可以通过该技术提高学习资源平台的用户体验，吸引更多用户，增加平台收益。

（3）操作可行性：开发该系统的应具有丰富的计算机前后端技术，和大数据领域知识，能够实现系统的开发、测试和部署。但是对于用户来说，使用则非常的简单，只需要注册之后便能登录，查看对应的可视化界面，如果数据量不够，也是可以通过后台管理系统手工添加的，操作也非常简便。

综上所述，基于大数据的学习资源推送系统具有可行性，能够为用户提供有价值的学习资源推荐服务。但同时也要注意，系统涉及的功能较多，需要耐心开发。

3.4本章小结

本章首先对系统开发的可行性进行了分析，然后对系统的非功能性需求进行了分析，并对系统开发的整体流程以及主要流程操作进行了介绍。

系统设计

本章所介绍的是系统设计包括系统总体架构设计、前台页面设计、系统流程设计、数据库设计。

4.1 系统总体架构设计

本系统的是基于B/S架构的，前台通过浏览器进行访问，web服务器会对访问进行解析，然后通过HTTP协议于后台Django框架进行沟通。Django框架是基于MTV（模型、模板和视图层）的，视图层层会接收到路由分发的请求，进而调用业务逻辑层代码，业务逻辑层代码再进一步调用模型层代码对数据库进行操作了，操作结束后，数据库数据发生改变。其次，服务器再调取数据的操作时候，由于数据发生了变化，所以相应的发送到前台的客户端的数据也相应的变化。

4.2 前台页面设计

系统前台页面包括有注册于登录，考虑到本系统主要的功能设定是数据的爬取以及大数据可视化展示和情感分析，所以注册与登录页面设计相对简单，可视化主页面展示分为7块区域，顶部是系统的名称，下面划分为6个区域,，上部分3个区域主要是类别突击图展示区域，学习资源列表展示区域，语言热点排行区域，下部分3个区域分别收藏占比展示区域、学习资源详情介绍区域和学习资源推荐区域。

4.3 系统流程设计

4.3.1 用户注册与登录流程设计

用户注册与登录流程设计：用户填写注册信息，系统验证信息有效性并创建账号。登录时，用户输入账号和密码，系统验证后允许访问。流程简洁明了，确保用户信息安全。



图4.4用户登录流程图

4.3.2 学习资源推荐流程设计

学习资源推荐流程设计：根据用户学习历史、偏好和成绩，系统智能推荐相关学习资源。用户可浏览、筛选并获取资源，提升学习效率。

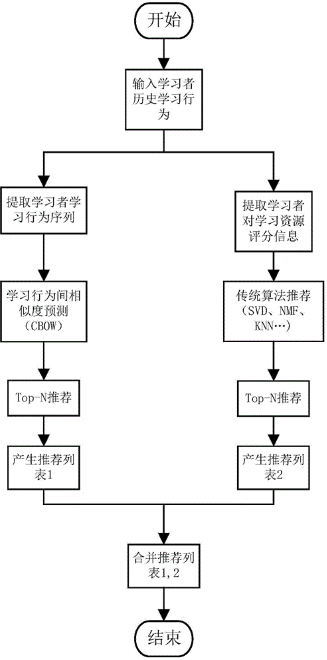


图4.5学习资源推荐流程图

4.3.3收藏占比与类别统计流程设计

收藏占比与类别统计流程设计：系统定期统计用户收藏资源的数量和类别，分析用户学习兴趣。通过可视化展示，帮助用户了解自己的学习倾向，优化学习计划。

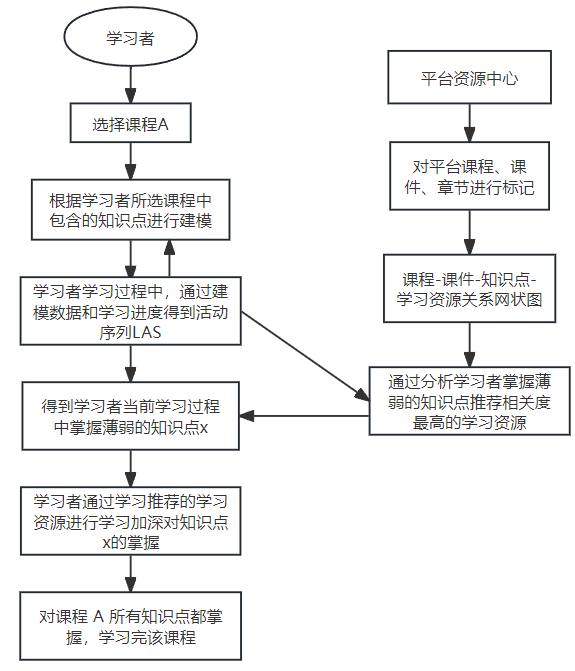


图4.6收藏占比与类别统计流程图

4.4 数据库设计

4.4.1 概念结构设计

本系统采用概念结构模型，首先需要创建数据库，数据库设计表是根据用户的需求抽象出来的，是一种线性表。根据项目中功能结构构建各个实体，并且确立实体之间的对应关系，整个项目的所有实体的各种字段和属性都存储在数据库表中，运用 ER 模型把问题转化成现实的问题。系统全局ER图如下4.7所示：

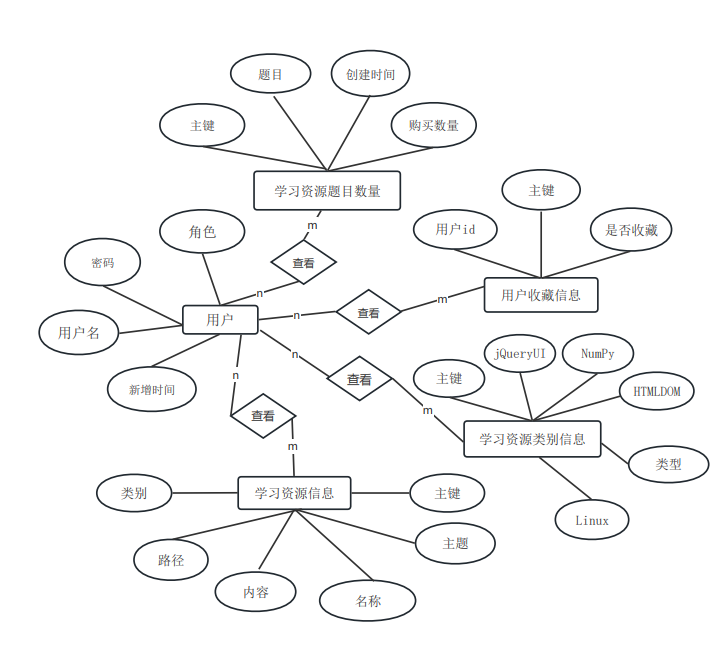


图4.7系统全局E-R图

4.4.2 逻辑结构设计

数据库是系统的数据源，根据需要展示可视化表格的要求，需要确定爬取数据的对象和属性，然后封装到数据库中，以下是主要需要爬取和储存的数据信息。

主要是对学习资源题目数量的记录。如表4-1所示：

表4-1学习资源题目数量信息表(title\_count)

|  |  |  |
| --- | --- | --- |
| 列名 | 注释 | 数据类型 |
| id | 主键 | Integer |
| title | 题目 | Text |
| count | 数量 | Integer |

主要是对用户收藏信息的记录。如表4-2所示：

表4-2用户收藏信息表(user\_like)

|  |  |  |
| --- | --- | --- |
| 列名 | 注释 | 数据类型 |
| id | 主键 | Integer |
| user\_id | 用户id | Integer |
| like | 是否收藏 | Integer |

主要是对学习资源类别信息的记录。如表4-3所示：

表4-3学习资源类别信息表(cainiao\_based)

|  |  |  |
| --- | --- | --- |
| 列名 | 注释 | 数据类型 |
| id | 主键 | Integer |
| jQueryUI | jQueryUI | Integer |
| NumPy | NumPy | Integer |
| HTMLDOM | HTMLDOM | Integer |
| … | … | Integer |
| Linux | Linux | Integer |

主要是对学习资源信息的记录。如表4-4所示：

表4-4学习资源信息表(new\_info)

|  |  |  |
| --- | --- | --- |
| 列名 | 注释 | 数据类型 |
| id | 主键 | Integer |
| title | 主题 | Text |
| name | 名称 | Text |
| content | 内容 | Text |
| url | 路径 | Text |
| design\_name | 类别 | Text |
| design\_url | 类别路径 | Integer |

4.5 本章小结

本章首先介绍了系统的整体框架，其次对系统的主要页面和模块进行了分析和设计；最后根据前台需要展示的数据列出来需要爬取的实体类和数据信息，这些数据是先通过request库进行爬取，使用pandas存储为csv文本格式的。

系统实现

5.1 系统的配置和部署

系统后台使用python书写，采用django框架搭建。以下是需要配置的一些插件，包括pyMysql,django,pip等，系统的虚拟环境采用python3.8，具体如下图所示。另外就是还需要配置JDK环境，主要是由于hadoop的配置需要依赖JDK环境。

电脑截图

描述已自动生成 图5.1系统的配置图片

后台程序需要连接数据库，在settings.py文件中进行相应的配置，以下是配置内容。

|  |
| --- |
| DATABASES = {  'default': {  'ENGINE': 'django.db.backends.mysql', # 数据库引擎  'NAME': hz\_cainiao\_project, # 数据库名称  'HOST': '127.0.0.1', # 数据库地址，本机 ip 地址 127.0.0.1  'PORT': 3306, # 端口  'USER': 'root', # 数据库用户名  'PASSWORD': 'root', # 数据库密码  }  } |

5.2 数据爬取与存储功能实现

5.2.1菜鸟教程网爬取页面分析

菜鸟教程网（www.cainiao.com）是国内知名的在线学习平台之一。该网站专注于提供各种编程语言、Web开发、数据库管理等领域的学习资源，旨在帮助用户快速掌握相关技能。菜鸟教程网的特色在于其丰富的学习资源。该网站提供了大量的文章教程、视频教程和在线实践等学习资料，内容覆盖广泛，既有基础知识讲解，也有高级技能传授。同时，菜鸟教程网注重学习资源的品质和权威性，邀请了众多行业内知名专家和经验丰富的开发者参与内容制作，确保了学习资料的专业性和实用性。因此从这个平台可以爬取到自己需要的数据，且该网站有大量关于计算机各类学习资源等数据，便于使用这些数据进行学习资源推荐。

5.2.2菜鸟教程网数据爬取

菜鸟教程网数据没有做反扒设置，只需要使用request库进行爬取即可，以爬取https://www.runoob.com/网页为例，爬取到指定的网友之后需要对网页需要的信息进行提取，这里使用BeautifulSoup进行提取，部分相关代码如下，然后将获取的数据保存到dataset/ list.csv中。

5.2.3数据上传到hadoop中

系统使用hadoop对数据进行保存和处理，获取的数据需要上传到hadoop中，上传的时候需要填写hadoop的ip等信息。

5.3 首页功能实现

通过以上步骤完成了数据的爬取和存储，接下来就是系统可视化页面展示阶段了。系统前台页面通过vue框架结合element-ui等插件实现，采用了Django web框架，后台使用python进行代码的书写。在用户输入网址后进入系统首页系统的首页需要展示的内容是最多的，如下图是首页信息展示流程图，首页在展示数据之前肯定是需要先从数据库调取相应的数据，经过web服务器的解析，然后进行展示。

用户进入本系统可查看系统主页信息，可视化主页面展示分为7块区域，顶部是系统的名称，下面划分为6个区域,，上部分3个区域主要是类别突击图展示区域，学习资源列表展示区域，语言热点排行区域，下部分3个区域分别收藏占比展示区域、学习资源详情介绍区域和学习资源推荐区域，具体如图。主页中的图标数据主要是通过Echart框架的功能，只需要创建对应的实体类变可以获取对应的展示信息。

 图5.4系统主界面图

5.4 系统用户注册与登录功能实现

5.4.1用户注册界面

用户注册流程图和实现界面展示如图5.5所示。注册的本质是后台拿到前台的数据，如图所示用户注册流程图，游客可以通过系统注册功能成功系统用户，成为用户后可以进行在线查看可视化数据等操作。注册需要用户填写相关的注册资料，其中用户名是全站唯一，在注册用户信息时候需要先对用户名进行一个校验，如果用户出现相同的则提示用户已存在,请勿重复注册。具体如图所示：

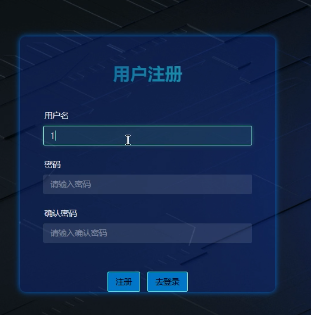


图5.5用户注册界面图

5.4.2用户登录界面

用户要想进行预测查看主页面等操作，必须登录系统。登录的操作是先获取到输入的用户名、密码，然后再和数据库中的用户表进行一一的比较，如果用户名和密码都存在与数据库，那么就可以登录成功，登录成功之后通过session把用户的信息进行保存，如果用户名或者密码错误，则登录失败并给出提示。用户登录实现界面展示如图5.6所示。

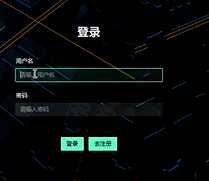


图5.6用户登录界面图

5.5学习资源推荐功能实现

系统利用大数据技术对海量学习资源数据进行高效处理。系统通过爬虫技术抓取学习资源数据，经Pandas初步处理后存入MySQL。核心的分布式计算部分利用协同过滤算法和机器学习算法的训练，然后将推荐结果通过Vue.js框架呈现。这样不仅提高了学习资源推荐的准确性和个性化程度，还为用户提供更便捷的阅读体验。



图5.7学习资源推荐界面图

**5.6数据管理功能实现**

考虑到有的时候数据爬取的数据不完整，或者需要修改或者添加数据，所以设置后台的数据管理功能，管理员可以根据需要增加或者修改爬取到的数据信息，然后再进行数据的保存操作。

图形用户界面

描述已自动生成图5.8 数据管理界面图

5.7本章小结

本章主要介绍了项目开发的MTV模式，然后着手代码的书写，介绍了系统的具体实现的一些功能，包括数据的爬取与储存，主页的实现，学习资源推荐功能等。

系统测试

6.1 系统测试目的

系统由于是个人开发的，开发过程中当然避免不了出现各类的问题，包括个人代码的问题以及兼容性等问题。正是在这样的背景下，需要进行测试，测试包括兼容性测试和典型测试用例的功能性测试两类。

6.2 系统兼容性测试

浏览器兼容性问题：随着ES6标准的制定，目前主流的浏览器都是符合ES6标准的，尤其是以谷歌为核心的内核，然后IE浏览器的早期版本使用的是微软自己的内容，对信息的兼容性产生影响。测试结果表明，目前的主要浏览器包括谷歌，IE,360,火狐浏览器最近的版本的运行都是没有问题的，状态良好，就是使用IE的早期版本有会有图片展示问题，通过对代码的修改，进行了改善，所以总的来说浏览器兼容性是没有问题的。其它兼容性问题：具体的比如Django框架版本的使用上需要使用2.0.13以上的版本，mysql需要安装5.5版本，python需要使用3.8版本，hadoop使用3.3版本等。

6.3 功能性测试

通过对系统的管理员和注册用户的具体操作进行典型的测试用例，测试主要的功能是否都能够正常使用。具体如下表

表6.1登录测试

本测试用例旨在验证用户登录功能的正确性和安全性。测试包括正常登录、错误密码登录、未注册用户登录等场景。正常登录测试将检查用户是否能成功进入系统；错误密码登录测试将验证系统是否能正确提示密码错误；未注册用户登录测试将确保系统能正确提示用户不存在。通过这些测试，确保登录功能的稳定性和安全性。

|  |  |
| --- | --- |
| 用例名 | 登录测试 |
| 目的 | 测试登录功能 |
| 前提 | 未登录的情况下 |
| 测试流程 | 1) 进入登录页面  2) 输入用户名和密码  3）提交 |
| 测试结果 | 1）当密码或者用户名错误的时候，提示用户名或者密码错误，页面不跳转；  2）当密码或者用户名都正确的时候，页面跳转到主页面； |
| 是否符合预期 | 是 |

表6.2添加数据管理测试用例

本测试用例关注数据管理功能的添加操作。测试将覆盖正常添加、数据格式错误、权限不足等多种情况。正常添加测试确保数据能正确录入系统；数据格式错误测试检查系统是否能识别并提示错误；权限不足测试将验证系统是否能阻止无权限用户进行添加操作。这些测试旨在确保数据管理功能的准确性和安全性。

|  |  |
| --- | --- |
| 用例名 | 添加数据管理测试用例 |
| 目的 | 测试管理员添加数据功能 |
| 前提 | 管理员用户登录系统 |
| 测试流程 | 1）点击对应的操作栏目  2）点击新增  3）填写内容后提交 |
| 测试结果 | 1）在相应的栏中展示新添加的数据。 |
| 是否符合预期 | 是 |

表6.3数据爬取测试用例

本测试用例旨在验证数据爬取功能的准确性和效率。测试将包括爬取正确数据、处理异常情况、爬取速度等多个方面。正确数据爬取测试确保系统能准确获取所需信息；异常情况处理测试检查系统是否能正确应对网络问题、数据变化等情况；爬取速度测试则关注系统的性能表现。这些测试有助于提升数据爬取功能的稳定性和效率。

|  |  |
| --- | --- |
| 用例名 | 数据爬取测试用例 |
| 目的 | 测试数据爬取信息功能 |
| 前提 | 爬虫文件书写完成 |
| 测试流程 | 运行爬虫程序 |
| 测试结果 | 爬取的文本信息，存储到dataset目录下对应的csv文件中 |
| 是否符合预期 | 是 |

6.4 本章小结

本章主要介绍了先对系统进行了兼容性的测试，然后在针对系统的主要功能进行了用例测试，测试结果表明，系统符合既定的功能需求目标。

# 结 论

在系统的开发中，前台主要使用的vue.js来开发页面。Vue如今非常的流行，这样更方便前台系统的维护和扩展。在后台的技术使用上，为伴随着人工智能的兴起，越来越多的人投入到python的学习当中，python语言结合Django框架，可以大幅度减少系统的配置，让开发者能够更多的把精力致力于系统逻辑上。另外再加上大数据技术的发展，hadoop的开源，让这些技术可以让小公司也能针对自己的用户进行有效的分析，这样便能够更好的推出适合用户使用的产品，立足于商业的竞争。

正是考虑到菜鸟教程网作为一款受欢迎的学习资源分享平台，拥有大量的学习资源数据信息。为了更好地利用这些信息，开发了基于大数据的学习资源推送系统。考虑到平台是没有设定了反扒机制，所以采用了request技术进行数据的访问，然后通过BeautifulSoup进行数据分析，最后通过pandas保存到本地csv文件中，最后通过hadoop进行数据分布式存储，通过spark进行数据的分布式计算。最后由于本系统是一个web网站，所以也设定了注册和登录功能，python的web框架采用Django，Django的使用可以极大的增加系统开发的效率，方便对数据的保存和使用。系统开发虽然结束，但是也是存在有一些问题的，比如在前台技术的使用上，用了vue脚手架和Echart联合开发，这样总体来说效率还是偏低的，主要是为了训练vue和echart的使用，因为现在也有BI大数据分析框架直接可以构建简单的数据分析图。

参考文献

[1] 尹志鹏, & 孔令外. 基于大数据的学习资源推荐系统研究[J]. 电子工程师, 2018. 6, 36-38.

[2] 范杲.. 基于大数据的学习资源智能推荐管理系统的设计与实现[J]. 智能计算机与应用, 2019.9(2), 158-161.

[3] 付毅, 赵展, & 王立平.基于大数据的学习资源管理系统研究[J]. 电子设计工程, 2019. 27(17), 87-91.

[4] 张立钧, & 高靖.基于大数据分析的医疗服务推荐系统研究[J]. 现代电子技术, 2019. 42(16), 101-104.

[5] 李丹, & 张宏峰.大数据分析在学习资源推荐系统中的应用研究[J]. 现代电子技术, 2019. 42(16), 97-100.

[6] 符笑良, & 魏润馨.基于大数据的健康饮食指南分享平台[J]. 电子设计工程, 2019. (22),78-80.

[7] 吴豪, 许森, & 杨麒麟.基于大数据和协同过滤算法的学习资源推荐系统研究[J]. 现代电子技术, 2018. 41(15), 108-110.

[8] 蔺龙, 刘喆, & 张晓玲.基于大数据的学习资源市场调研与推荐系统研究[J]. 计算机测量与控制, 2017. 25(6), 104-106.

[9] 李洪昌, 周磊. 基于PythonDjango框架的多媒体发布系统[J]. 物联网技术, 2018, 8(2):4.

[10] 李平, 刘小杰. 新时期高职院校毕业生就创业生态系统培育模式探析[J]. 湖北函授大学学报, 2020, 033(012):5-7.

[11] 肖航, 杨智, 吴东. 基于Django的网上书城系统的设计与实现[J]. 2021.

[12]李洪昌, 周磊. 基于PythonDjango框架的多媒体发布系统[J]. 物联网技术, 2018, 8(2):4.

[13] Zhang X . About the selection of graduate student enrollment source and system reform[J]. Francis Academic Press, 2020(11).

[14] Joseph M S , Dam F , Friestrom E D . A framework for graduate and postgraduate specialty pharmacy training[J]. American Journal of Health-System Pharmacy, 2021.

[15] Nsha B , Mwsa B , Lin Z , et al. Phenomapping a Novel Classification System for Patients With Destination Therapy Left Ventricular Assist Devices. 2021.

致 谢

首先，我要对所有支持和帮助我完成这项研究的人表示深深的感谢。第一要感谢我的导师任艳老师。在整个研究过程中，任艳老师给予了我细心的指导和无私的帮助。他严谨的治学态度、深厚的学术造诣和敏锐的洞察力，都深深地影响了我。他不仅在学术上给予我指导，还在生活中给予我关心和帮助，让我在研究过程中不断成长和进步。另外我也要感谢我的同学们。我们一起度过了无数个讨论和研究的夜晚，你们的支持和鼓励是我前进的动力。我们共同面对困难，共同解决问题，你们的友谊和支持是我完成这项研究的重要支柱。

同时我要感谢我的父母。他们一直以来的支持和鼓励是我前进的动力，他们的理解和支持让我能够全身心地投入到研究中。他们的爱和关怀是我人生中最宝贵的财富。最后我也要感谢那些在菜鸟教程网上发布学习资源的网站工作者。你们的学习资源等数据内容为我们的研究提供了宝贵的数据，你们的热情和反馈是我们研究的重要推动力。你们的每一句话、每一个表情符号，都成为我们分析的重要依据。在系统开发过程中，我也意识到自己的不足和需要改进的地方。我将继续努力，不断提高自己的研究能力和水平，为未来的研究打下坚实的基础。

总之，我要感谢的人太多太多，他们让我在研究的过程中不断成长和进步。未来，我将继续努力，为大数据分析领域做出更多的贡献！