怀化学院

课程项目
开 题 报 告 书

题　　目　基于Oracle的豆瓣图书数据分析的设计与实现

学生姓名　　粟泳璋

学　　号　　2400131241

院　　别　　计算机与人工智能学院（软件学院）

专　　业　　软件工程

指导教师　　杨攀　助教

2025 年 4 月 11 日

设计题目	基于Oracle的豆瓣图书数据分析的设计与实现

一、选题的目的、意义及相关研究动态和自己的见解：

目的：

豆瓣网站作为中国最具影响力的文化社交网站之一，汇聚了大量用户生成的书籍

论、评分数据。这些数据蕴含着读者的阅读兴趣、喜好以及书籍的质量、影响力等信息，对于出版机构、图书馆、书店以及广大读者都具有重要意义。豆瓣读书作为豆瓣网站一个重要组成部分，拥有庞大的书籍数据库和用户群体。用户可以在豆瓣读书上查找书信息、发布读书笔记和评论等。这些书籍信息和用户评论数据为书籍推荐、用户行分析、市场趋势预测等提供了宝贵的数据支持。通过数据分析可以得到一些信息，如同分类书籍的数量、评分的分布情况、出版社的排名、作者的排名等。这些信息助

出版机构了解市场趋势、优化出版策略；图书馆和书店可以根据数据分析结果调整藏书结构、提高服务质量；读者可以根据数据分析结果选择符合书籍。

意义：

现代社会，随着经济的发展和人们生活水平的提高，人们对于生活品质的要求也在不断攀升。通过对豆瓣读书数据的分析，我们可以深入了解读者的阅读偏好、兴趣点以及阅读趋势。这些数据不仅可以帮助出版机构精准定位目标读者群体，制定更符合市场需求的出版策略，还能为图书馆、书店等提供选书参考，为读者提供更加精准的书目推荐。豆瓣读书数据分析可以揭示图书市场的现状和发展趋势，如畅销书排行、不同分类书籍的销售情况等。这些信息有助于出版机构、书店等了解市场动态，制定更加合理的销售策略，促进图书市场的健康发展。同时，数据分析还可以发现潜在的市场机会，为新的出版项目提供决策支持。通过参与豆瓣读书数据分析项目，个人可以学习到数据爬取、清洗、预处理、分析以及可视化等技能，提升自己在数据科学领域的能力。

研究动态：

随着信息技术的飞速发展，大数据时代已然来临，各领域对于数据分析的重视程度与日俱增。在图书领域，豆瓣作为国内极具影响力的图书社区，积累了海量的图书数据，涵盖图书基本信息、用户评价、评分、评论等多方面内容，为深入的数据分析提供了丰富素材，吸引了众多学者和研究人员的关注。与此同时，数据库管理系统在数据存储与处理方面发挥着关键作用，Oracle 凭借其强大的功能、高度的可靠性以及出色的性能，成为大数据分析项目中常用的数据库管理系统之一，为基于豆瓣图书数据的分析提供了坚实的技术支撑。国外在数据可视化技术与图书数据分析结合方面开展了大量研究，并取得了显著成果。例如，全球知名的图书社交平台 Goodreads，充分运用数据挖掘和可视化技术，深入分析用户的阅读记录、评价以及互动信息，为读者提供个性化的图书推荐服务，极大地提升了用户的阅读体验 。此外，不少国外学者运用先进的数据分析算法，对大规模图书数据进行挖掘，在预测图书销售趋势、分析读者阅读偏好演变等方面取得了重要突破，为图书出版行业的决策提供了科学依据 。国内对于豆瓣图书数据的研究也在持续升温。部分研究聚焦于运用网络爬虫技术获取豆瓣图书数据，并借助数据清洗和预处理手段，提高数据质量，为后续分析奠定基础 。在数据分析方法上，学者们综合运用多种技术，如通过文本挖掘技术分析用户评论中的情感倾向，了解读者对图书的喜好与不满；利用聚类分析算法对图书进行分类，挖掘不同类别图书的特征 。同时，一些研究尝试构建基于豆瓣图书数据的推荐系统，旨在为用户提供精准的图书推荐，提高图书与用户需求的匹配度 。在 Oracle 数据库应用于数据分析的研究中，国内外均有众多成果。学者们深入研究 Oracle 数据库的性能优化，通过调整数据库参数、优化查询语句等方式，提升数据处理效率 。此外，针对大数据环境下的数据存储与管理难题，利用 Oracle 的分布式架构和数据分区技术，实现海量数据的高效存储与快速检索 。在实际应用中，诸多企业借助 Oracle 数据库构建数据分析平台，整合多源数据，进行深度分析，为企业决策提供有力支持 。

个人见解：

在数据获取与整合方面，现有的研究多集中于通过常规网络爬虫获取数据，可能面临数据获取不全面、网站反爬机制阻碍等问题。未来可尝试探索更高效、稳定的数据采集策略，例如与豆瓣平台建立合作获取官方数据接口，或者利用模拟浏览器行为、分布式爬虫等技术手段，突破反爬限制，获取更为全面、准确的图书数据，同时涵盖不同语言版本、小众分类等易被忽视的数据类别，以丰富数据维度。在数据整合时，不仅要整合图书基本信息、评分评论，还可关联其他相关平台数据，如电商平台图书销量数据、学术数据库中相关图书研究文献引用量等，形成多源融合的综合性数据集，为更深入分析提供数据基础。数据分析方法上，目前多采用较为基础的文本挖掘、聚类分析等技术。后续研究可引入深度学习模型，如基于 Transformer 架构的自然语言处理模型，对用户评论进行更精准的情感分析，挖掘隐含在文字中的复杂情感与潜在观点；运用图神经网络，将图书、作者、读者、标

签等构建成知识图谱，深入分析它们之间的关联关系，发现图书领域潜在的知识脉络与传播路径，从而为个性化推荐、图书主题拓展研究提供更强大的技术支撑。在基于 Oracle 数据库的实现层面，虽然已有不少关于性能优化的研究，但在应对豆瓣图书数据不断增长、数据类型愈发复杂的情况下，仍需进一步探索创新。一方面，可以充分利用 Oracle 的 In-Memory 数据库特性，将频繁访问的热点数据驻留在内存中，大幅提升数据读取与分析速度，以满足实时性数据分析需求，例如在动态更新图书推荐榜单时能快速响应。另一方面，针对海量非结构化数据（如大量的用户评论），可结合 Oracle 对 JSON 等非结构化数据格式的支持，设计更合理的数据存储结构与查询方式，实现结构化与非结构化数据的协同处理，提高数据处理的灵活性与效率。

从应用价值角度出发，现有研究成果在实际应用场景的拓展上稍显不足。后续研究可以紧密结合图书出版行业、图书馆管理以及读者服务等实际需求，开发具有针对性的应用系统。例如，为图书出版机构构建基于数据分析的选题策划辅助系统，通过分析豆瓣图书数据中的热门主题趋势、读者需求痛点等信息，为出版机构规划选题方向、预测市场需求提供科学依据；为图书馆提供智能馆藏优化方案，依据读者在豆瓣上对不同图书的关注度、借阅行为数据等，合理调整馆藏结构，提高图书馆资源利用率，为读者提供更贴合需求的图书资源服务。

二、课题的主要内容：

主要完成以下几个方面的内容：

(1)展示爬取到的所有电影数据，包括电影名，电影评分等基本信息，点击电影名能跳转到详情页使用户更方便的观察。

(2)根据评分，统计数据中每个分段的书籍数量，然后用柱状图展示出来，能直观展示每个评分段的书籍数量分布。

(3)根据书籍简介，然后用词云图展示出来，方便其中的热门文学观点的观察。

(4)通过扇形图展示各国的书籍占比数。

(5)根据点赞人数统计出热度排行前十的书籍。

三、研究方法、设计方案或论文撰写提纲：

1、 研究方法

数据收集法

网络爬虫技术：使用 Python 的 Scrapy 框架，编写网络爬虫程序，从豆瓣图书页面获取数据。通过精心设计爬虫规则，能够精准定位并采集图书名称、作者、出版社、出版年份、评分、评论数量、用户评论内容等关键信息。针对豆瓣的反爬

虫机制，采用设置合理的请求间隔时间、随机更换 User - Agent、使用代理 IP 等策略，确保数据采集的稳定性与持续性，以获取大规模、高质量的原始图书数据。

数据补充与整合：除了豆瓣平台自身数据，为丰富分析维度，从其他公开数据源收集相关数据。例如，从电商平台获取图书销量数据，从图书馆馆藏系统获取部分图书借阅频次数据。利用数据清洗与预处理技术，将多源数据按照统一的标准进行格式转换、去重、缺失值处理等操作，然后依据图书唯一标识符（如 ISBN 号）进行数据整合，构建综合性的图书数据集，为后续深入分析奠定坚实基础。

数据存储与管理 - Oracle 数据库应用

数据库设计：依据图书数据的特点与分析需求，运用数据库设计的规范化理论，在 Oracle 数据库中精心设计数据库表结构。创建 “图书基本信息表” 存储图书名称、作者、出版社等静态信息；“评分与评论量表” 记录图书评分、评论数量及用户评论详情；“关联数据表” 用于存储与其他数据源整合后的补充信息等。合理设置表间关系，通过主键、外键约束确保数据的一致性与完整性，为高效的数据存储与查询提供保障。

数据加载与优化：利用 Oracle 提供的数据加载工具，如 SQL*Loader，将经过清洗和预处理的数据快速、准确地加载到数据库中。针对大数据量存储与查询性能问题，运用 Oracle 的分区表技术，按照出版年份、图书类别等维度对数据进行分区存储，提高数据检索效率。同时，通过优化数据库参数配置，如调整内存分配参数、设置合理的缓存策略等，进一步提升数据库整体性能，确保能够高效管理海量的豆瓣图书数据。

数据分析方法

描述性统计分析：运用统计分析工具，对收集到的图书数据进行描述性统计。计算各类图书的平均评分、评分标准差，以了解图书整体质量水平及评分的离散程度；统计不同年份、出版社出版图书的数量，分析图书出版趋势与出版社活跃度；计算评论数量的均值、中位数等，评估图书受关注程度的集中趋势与分布情况，从宏观层面初步把握豆瓣图书数据的整体特征。

文本挖掘与情感分析：针对用户评论这一非结构化文本数据，采用自然语言处理技术进行文本挖掘。使用 Python 的 NLTK、SpaCy 等库对评论进行分词、词性标注、去除停用词等预处理操作，然后运用词袋模型、TF - IDF 算法将文本转化为可用于分析的数值向量形式。在此基础上，构建基于机器学习的情感分析模型，如朴素贝叶斯分类器、支持向量机等，对用户评论进行情感倾向判断，分为正面、负

面、中性三类，深入了解读者对图书的情感态度与意见反馈。

聚类分析与关联规则挖掘：运用聚类分析算法，如 K - Means 聚类，对图书数据进行分类。根据图书的评分、评论数量、主题标签等特征，将具有相似特征的图书聚为一类，挖掘不同类别图书的潜在特征与共性，为图书推荐、市场细分等应用提供依据。同时，采用关联规则挖掘算法，如 Apriori 算法，分析图书之间的关联关系，例如发现同时被大量用户关注或评论的图书组合，以及用户购买行为中不同图书之间的关联模式，为图书营销策略制定提供参考。

2、 设计方案

（1）查阅相关资料，熟悉开发工具软件的使用；

（2）根据项目的具体需求，进行一定的界面实现逻辑规划；

（3）针对特定的任务在熟练掌握原理的前提之下，灵活运用相关知识点；

（4）以项目驱动的方式完成整个设计流程。

四、完成期限和预期进度：

（1）接受任务、查阅资料、撰写总体设计方案、撰写开题报告：2025 年 4 月—5 月

（2）程序的开发、运行测试：2025 年 5 月—2025 年 6 月

（3）毕业论文的撰写：2025 年 6 月—7 月

（4）毕业论文的修改：2025 年 6 月—7 月。

五、主要参考文献~~（不少于10篇）~~：

[1] 钱雪忠。数据库原理及应用 [M]. 北京：北京邮电大学出版社，2024：15-30.

[2] Bain T.SQL server 2000 数据仓库与 Analysis Services [M]. 北京：中国电力出版社，2023：45-60.

[3] 王珊。数据库技术与联机分析处理 [M]. 北京：科学出版社，2022：22-44.

[4] 佚名。豆瓣书籍数据可视化分析工具 [J].CSDN 博客，2025 (1)：1-10.

[5] 佚名.Oracle 数据库技术 [J].Oracle 中国，2024 (9)：1-15.

[6] 佚名。大数据毕业设计 —— 基于 Hadoop 实现的豆瓣电子图书数据分析和推荐系统的设计与实现 [J].CSDN 博客，2025 (3)：1-12.

[7] 佚名.Data crawling and analysis based on Douban books [J].Kaggle，2025：1-8.

[8] 佚名。最权威的行业数据 [J].OpenBookScan，2025：1-5.

[9] 佚名.Python 毕业设计 —— 基于大数据 +Hadoop 的豆瓣电子图书推荐系统设计和实现 [J].CSDN 博客，2025 (3)：1-14.

[10] 佚名.Oracle Developer Tools for Visual Studio [J].Oracle 中国，2025：1-11.

六、指导教师意见：

年　　月　　日