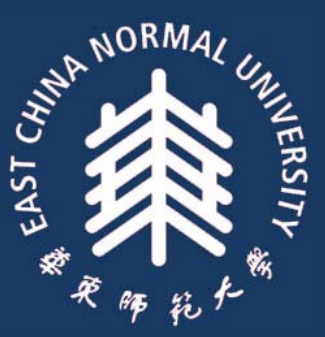


華東師範大學

基于CADAL平台的用户推荐系统开发实施方案

李欣

图书馆、数据科学与工程学院

大纲

- 项目需求调研
- 建设目标与服务
- 技术与实施方案
- 面临问题

项目需求调研

- 提升平台的可用性
  - CADAL平台运行多年，积累了大量用户数据，因此开展基于用户行为数据的分析与利用，并用于用户的精准推荐，对提升平台的可用性具有非常重要的现实意义。
- 技术成熟
  - 推荐系统技术成熟并不断发展
    - 经典算法成熟
    - 机器（深度/强化）学习、网络特征数据爬取
  - 数据科学作为一门新型交叉学科，近年来发展迅速
  - 用户画像（行为延伸）是基于用户行为数据实现标签化的过程，这些标签又可以被表示为用户的属性，包括个人资料、兴趣爱好、行为和情感特征等。
- 技术队伍优势
  - 图书馆、学院合作

建设目标与服务

➢总体设计方案

- 一期实现数据处理及基本软件开发
  - 基于CADAL平台数据（用户注册/系统记录/日志），设计多源数据的存储方案、基于资源/用户的协同过滤推荐系统方案
  - 系统开发为重点，从数据分析、清洗、结构化处理等，形成可分析数据，搭建实验环境。以注册用户访问及检索数据为主要研究对象，实现基于资源/用户日志数据的基本系统推荐功能
- 二期实现算法优化、丰富数据
  - 用户画像。重点在用户关联关系的数据采集/爬取，进一步精准刻画用户偏好
  - 推荐算法持续优化。针对数据稀疏等个性化原因，研究推荐算法改进方案；针对一些隐式或缺失的用户属性，研究利用机器学习方法从用户的关联关系非结构化、动态的数据中推断结果。如通过注册用户信息爬取其归属单位属性信息及其在互联网上偏好信息，以实现对稀疏数据的补充等。

通过分期建设，实现对用户日志的深度分析和关联关系数据的采集/爬取，实现CADAL平台的精准推荐功能，提升平台的实用性。

技术与实施方案

➢存储方案设计

1）数据库选择

- 基于项目一期结构化数据数据特点，使用MySQL作为数据库存储系统
- 二期将陆续加入其它非结构化数据，考虑使用MySQL+NoSQL结合方式处理，将Mysql+Neo4j结合，Mysql存储结构化数据，Neo4j存储图型结构
- 根据数据量极速扩大，可考虑使用Hive进行后台日志的存储

2）表设计（应该提供系统表数据接口或结构）

技术与实施方案

一级标签	二级标签	标签描述
注册信息	用户名	用户的唯一标识符。
	邮箱	
	专业	
	性别	
	生日	
	所在单位	
	常居地	
	兴趣	
借阅与归还信息	借阅书籍编号	
	借阅书籍时间	
	归还书籍时间	
检索信息	历史搜索内容	
	关键词	图书馆系统基于用户的搜索内容匹配的搜索关键词
评注信息	评注书籍名称	
	评注内容	
关注信息	关注用户名	
	被关注用户名	
浏览信息	浏览书籍编号	
	浏览书籍页数	

二级标签	标签描述
历史搜索内容	
关键词	图书馆系统基于用户的搜索内容匹配的搜索关键词
浏览书籍编号	

二级标签	标签描述
编号	书籍的唯一标识符
分类	
名称	
数目	
借阅时间	
借阅用户	
归还时间	
浏览时间	
浏览用户	
浏览书籍页数	
关注用户名	

数据要求：

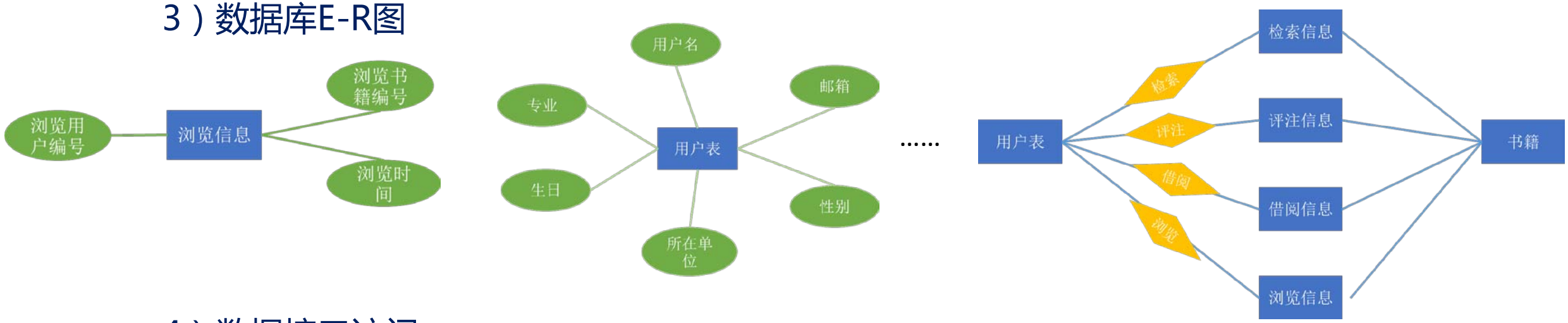
1. 上述三块标签给出推荐系统所要求的基本数据，若能提供其他方面的相关数据，越多越好。
2. 所有数据要求尽可能完整（每份数据都要有其唯一标识符和字段名），并且数据量越大越好。
3. 若暂无完整数据，可先提供一份数据样例，包括用户信息、书籍信息和用户行为信息。

注册用户数据 | 非注册用户数据 | 书籍数据 | 数据要求

技术与实施方案

➢存储方案的设计。

3）数据库E-R图



4）数据接口访问

- 推荐数据周期生成记录，提供接口URL方式访问

技术与实施方案

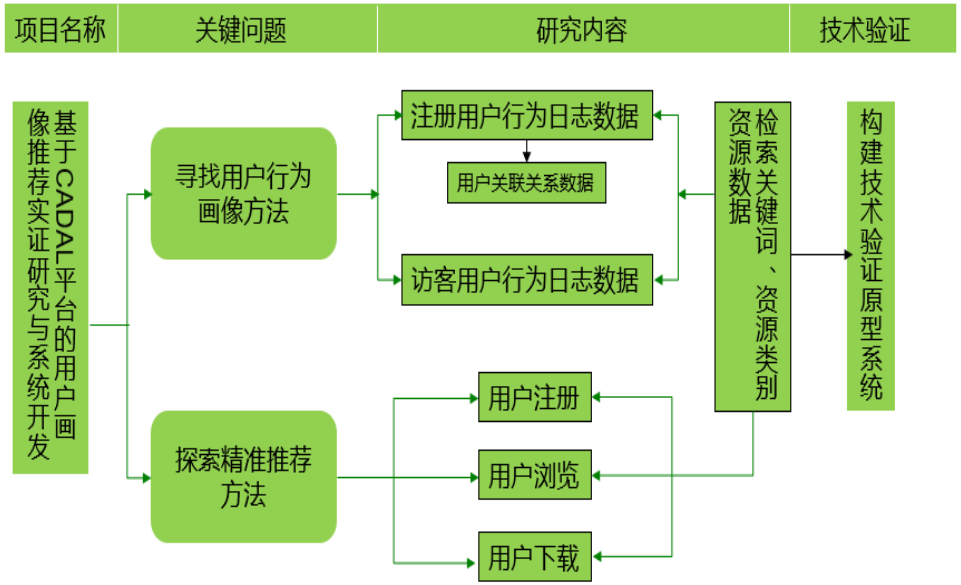
➢用户画像研究

基于已有数据信息，对用户行为，主要包括借还、检索、评注以及浏览信息进行画像数据设计

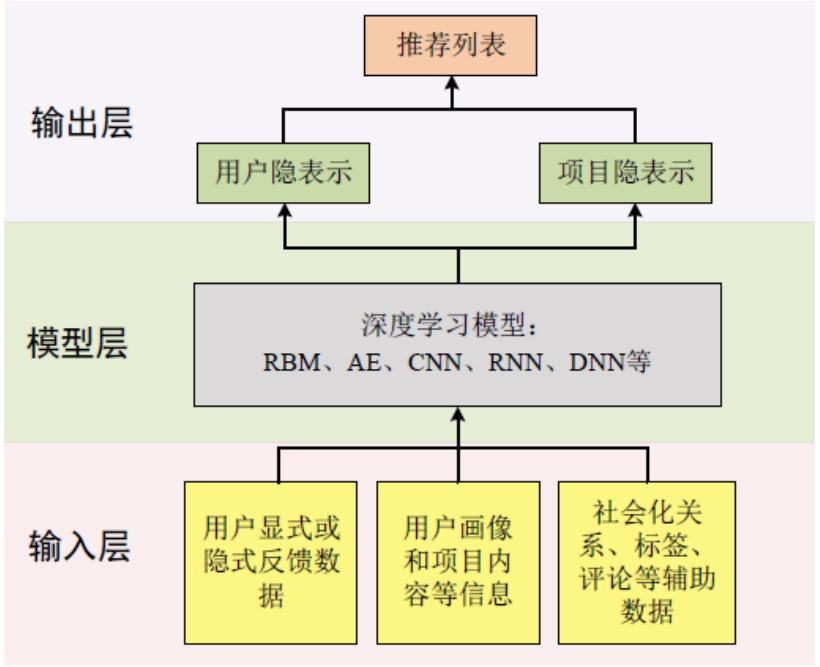
基本信息	用户名
	邮箱
	专业
	性别
	生日
	所在单位
借阅行为	借阅书籍基本信息
	借阅书籍主题
	借阅频率信息
检索行为	检索记录
	检索领域
	检索主题
评注行为	评注书籍信息
	评注情感倾向
浏览行为	浏览书籍信息
	浏览集中领域
	浏览书籍主题

技术与实施方案

➢一期设计方案



➢二期优化方案



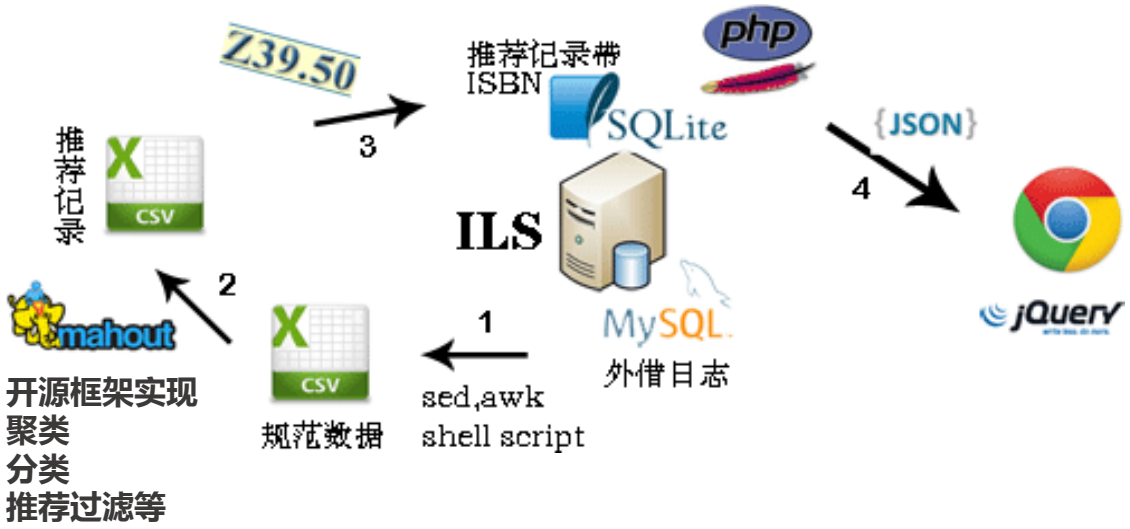
二期对系统数据积累要求较高

技术与实施方案

➢前期研究支撑

图书馆

Opac系统



学院

应用研究

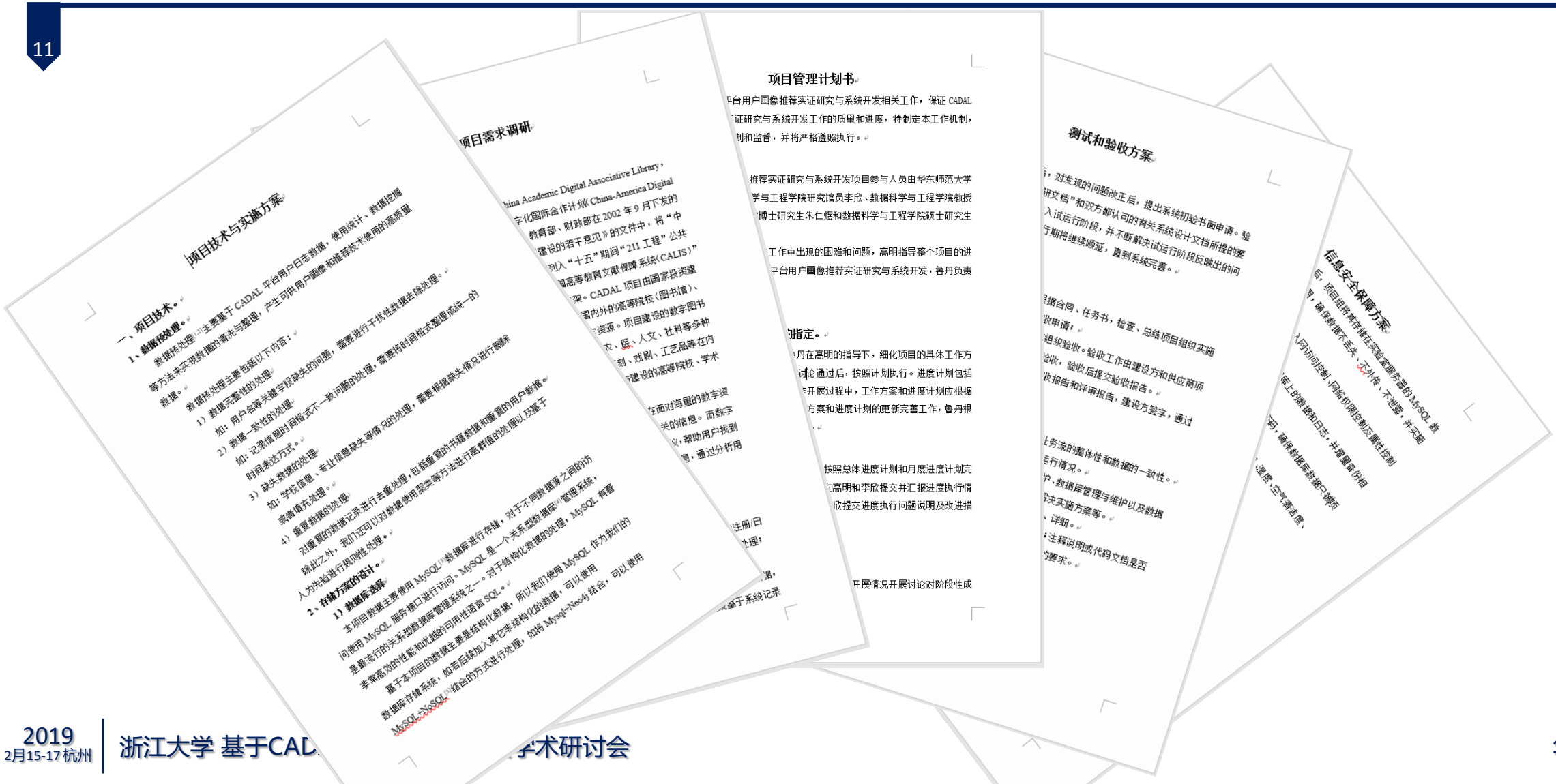
研究生教育质量评估用户画像项目, 教育部委托专项（华东师大/西交大）2018-2020年

学术研究

基于多个异构社交网络数据分析的用户建模及其应用，国家自科青年项目，2014

面向个性化课辅的学生学习行为画像及其应用研究，国家自科面上项目，2018

技术与实施方案



项目技术与实施方案

一、项目技术。

1、数据预处理。

数据预处理主要包括以下内容：

1）数据完整性的处理

如：用户名等关键字段缺失的问题，需要进行干扰性数据去除处理。

2）数据一致性的处理

如：记录信息时间格式不一致问题的处理，需要将时间格式整理成统一的时间表达方式。

3）缺失数据的处理

如：学校信息、专业信息缺失等情况的处理，需要根据缺失情况进行删除或者填充处理。

4）重复数据的处理

对重复的数据记录进行去重处理，包括重复的书籍数据和重复的用户数据。

除此之外，我们还可以对数据使用聚类等方法进行离群值的处理以及基于人为先验进行规则性处理。

2、存储方案的设计。

1）数据库选择

本项目数据主要使用 MySQL 数据库进行存储，对于不同数据源之间的访问使用 MySQL 服务接口进行访问。MySQL 是一个关系型数据库管理系统，是最流行的关系型数据库管理系统之一。对于结构化数据的处理，MySQL 有着非常高的性能和优越的可用性语言 SQL。

基于本项目的数据主要是结构化数据，所以我们使用 MySQL 作为我们的数据库存储系统，如若后续加入其它非结构化的数据，可以使用 MySQL+NoSQL 结合的方式进行处理，如将 Mysql+Neo4j 结合，可以使用

项目需求调研

hina Academic Digital Associative Library，

字化国际合作计划(China-America Digital

教育部、财政部在 2002 年 9 月下发的

建设的若干意见》的文件中，将“中

列入“十五”期间“211 工程”公共

智高等教育文献保障系统(CALIS)”

架。CADAL 项目由国家投资建

国内外的高等院校(图书馆)、

资源。项目建设的数字图书

农、医、人文、社科等多种

剧、戏剧、工艺品等在内

建设的高等院校、学术

面对海量的数字资

关的信息。而数字

义，帮助用户找到

息，通过分析用

注册/日

处理；

据，

基于系统记录

项目管理计划书

平台用户画像推荐实证研究与系统开发相关工作，保证 CADAL

证研究与系统开发工作的质量和进度，特制定本工作机制，

制和监督，并将严格遵照执行。

推荐实证研究与系统开发项目参与人员由华东师范大学

学与工程学院研究馆员李欣、数据科学与工程学院教授

博士研究生朱仁煜和数据科学与工程学院硕士研究生

工作中出现的困难和问题，高明指导整个项目的进

平台用户画像推荐实证研究与系统开发，鲁丹负责

指定。

丹在高明的指导下，细化项目的具体工作方

论通过后，按照计划执行。进度计划包括

开展过程中，工作方案和进度计划应根据

方案和进度计划的更新完善工作，鲁丹根

按照总体进度计划和月度进度计划完

高明和李欣提交并汇报进度执行情

欣提交进度执行问题说明及改进措

开展情况开展讨论对阶段性成

测试和验收方案

，对发现的问题改正后，提出系统初验书面申请。验

研文档”和双方都认可的有关系统设计文档所提的要

入试运行阶段，并不断解决试运行阶段反映出的问

行期将继续顺延，直到系统完善。

据合同、任务书，检查、总结项目组织实施

收申请；

组织验收。验收工作由建设方和供应商项

验收，验收后提交验收报告。

收报告和评审报告，建设方签字，通过

务流的整体性和数据的一致性。

运行情况。

户、数据库管理与维护以及数据

决实施方案等。

、详细。

：注释说明或代码文档是否

要求。

信息安全保障方案

后，项目组将其存储在实验室服务器的 MySQL 数

，确保数据不丢失、不外传、不泄露，并实施

入网访问控制、网络权限控制及属性控制

据上的数据和日志，并增量备份相

码，确保数据和数据日志的

温度、空气通过度，

面临问题

- 数据分析
  - 系统记录实验数据（应主要源于原来系统）？来自于系统记录/用户注册/日志等多源数据
  - 非注册用户数据
- 系统冷启动
  - 用户冷启动，即如何给新用户做个性化推荐
  - 资源冷启动，即如何将新的书籍推荐给可能它感兴趣的用户
  - 系统冷启动，即如何在刚上线的CADAL平台上设计个性化推荐系统
- 测试系统
  - 没有数据对测试带来问题

学院发展理念与使命

培养应用型、复合型、创新型的IT人才

- 践行“知行合一”理念，通过解决实际应用问题，进而培养创新和创造能力。

做真的研究，做有用的研究

- 立足解决现实应用问题，从而发现和抽象出研究问应用中去”，应用创新和学术创新相互促进，形成先，努力为实现“替代工程”和解决“卡脖子”问

传播和传授计算机科学和人工智能新理念和新知识

- 在我国推进“CS for ALL”和“AI for ALL”，前动能驱动的新工具的使用。

COOPERATION 合作单位

中国电信 CHINA TELECOM

交通银行 BANK OF COMMUNICATIONS

PingCAP

中国移动 China Mobile

批改网 www.pigai.org

DiDi

大学数字图书馆国际合作计划 CHINA ACADEMIC DIGITAL ASSOCIATIVE LIBRARY

Alibaba Group 阿里巴巴集团

华院数据 www.unidt.com

HUAWEI

感谢聆听！

xli@dase.ecnu.edu.cn