
# 基于CADAL平台的用户推荐系统设计

李欣

华东师范大学
图书馆
数据科学与工程学院

# 目录

# 目录

# 推荐系统及项目简介



- 推荐系统由来
  - 随着信息技术和互联网的发展，人类从信息匮乏时代走向了信息过载时代
  - 推荐系统是解决信息超载问题一个非常有潜力的办法
  - 推荐系统现已广泛应用于很多领域，其中最典型是电子商务领域。同时学术界对推荐系统的研究热度一直很高，逐步形成了一门独立的学科。

# 推荐系统及项目简介



## 什么是推荐系统

- 推荐系统是根据用户的历史行为、社交关系、兴趣点、所处上下文环境等信息去判断用户当前需要或感兴趣的物品/服务的一类应用
- 信息过滤
- 推荐系统的核心任务是联系用户和内容提供商
  - 对用户而言，推荐系统能帮助用户找到喜欢的物品/服务，帮忙进行决策，发现用户可能喜欢的新事物
  - 对内容提供商而言，推荐系统可以给用户提供个性化的服务，提高用户信任度和粘性，增加营收

# 推荐系统及项目简介



- 推荐系统的发展
  - 1994年国外学界提出
  - 目前已广泛集成到很多商业应用系统中
  - 尤其是网络购物平台中


  - Amazon：网络书城的推荐算法每年贡献30个百分点的创收
  - Netflix：2/3 被观看的电影来自推荐
  - Google新闻：38%的点击量来自推荐
  - ……

入口：用户、物品、评价。出口：推荐列表

# 推荐系统及项目简介 CADAL平台



- China Academic Digital Associative Library （大学数字图书馆国际合作计划）。
- “十五”期间“211工程”公共服务体系建设的重要组成部分。
- 一期建设100万册(件)数字资源，提供一站式的个性化知识服务。

# 推荐系统及项目简介 项目背景



CADAL数字图书馆有海量的数字资源，用户难以找到与用户需求相关的信息

推荐系统

收集和统计用户行为信息，向用户推荐有关的文献信息或有用的建议

## 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。

# 推荐系统及项目简介 项目背景



- 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

# 推荐系统及项目简介 项目背景



## 相似性推荐

### 面向问题的思路

- 借阅量逐年下滑的触动，考虑寻求提升读者借阅量的方法
- 互联网思维影响，以用户为中心的思考，丰富“用户空间”内容
- 让推荐更有针对性、主动性、智能化

# 推荐系统及项目简介 项目背景



- 做法
  - 改变传统推荐：把一类图书推荐给一类用户
  - 互联网做法影响……
    - 亚马逊网站上有35%的销售额是来自于个性化推荐[1]
    - 60%的销售额间接受到推荐影响[1]
  - 利用读者历史数据

[1]：亚马逊前科学家Greg Linden博客

# 推荐系统及项目简介 项目愿景



- 一期实现数据处理及软件开发

  - 基于CADAL平台提供的用户访问数据，设计来自于系统记录/用户注册/日志的多源数据处理方案
  - 对系统记录/用户注册/日志数据进行全面的预处理
  - 对处理好的数据设计合理的存储方案
  - 基于上述设计逐步实现画像系统

# 推荐系统及项目简介 项目愿景



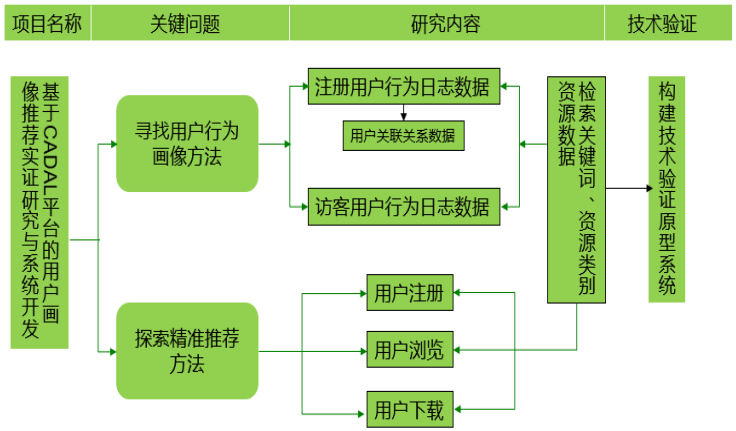
- 二期实现研究环境建设

  - 爬取外部数据丰富用户数据，精准刻画用户偏好
  - 精准推荐算法持续优化，并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，研究利用机器深度学习和强化学习等方法从用户的关联关系非结构化、动态的数据中推断结果
  - 通过爬取方式获取用户更多信息

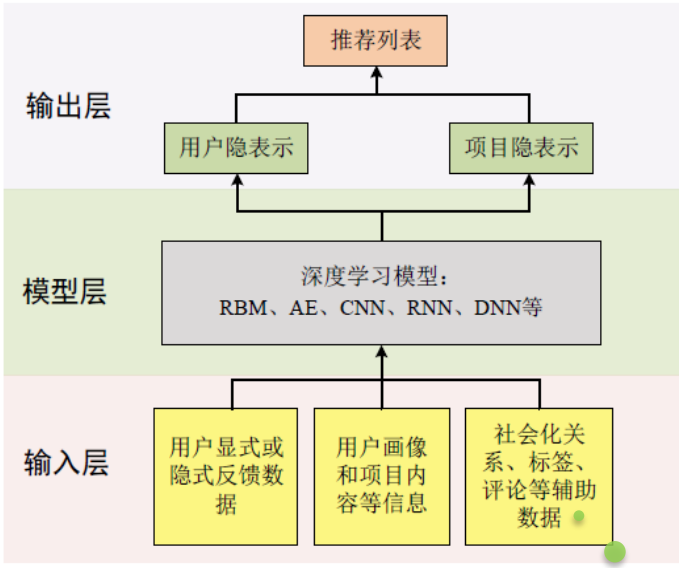
# 推荐系统及项目简介 项目实施方案



## 一期实现




## 二期实现




二期对系统数据积累要求较高

通过爬取方式获取用户更多信息

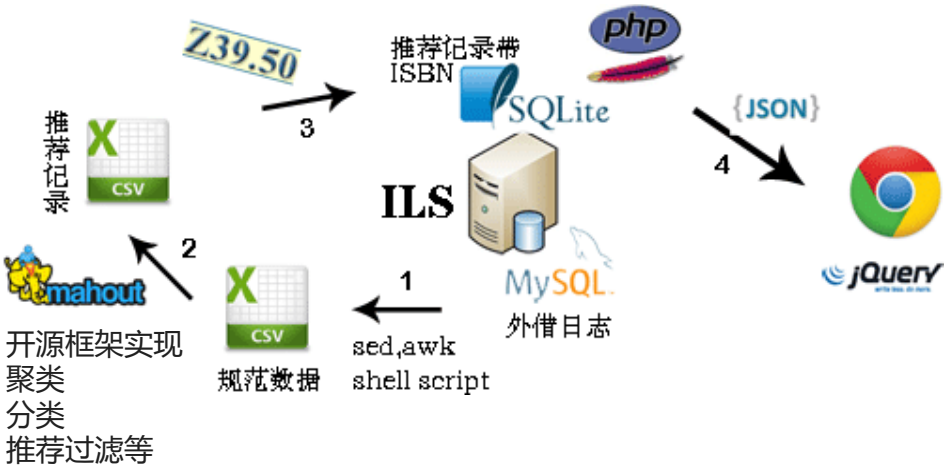
# 推荐系统及项目简介 项目参与人员



- 项目负责人
- 项目指导老师
  - 高明（数据科学与工程学院教授）
- 项目参与者
  - 鲁丹、汪志莉（图书馆）
  - 朱仁煜（数据科学与工程学院2017级博士生）
  - 李娜（数据科学与工程学院2017级硕士生）

# 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

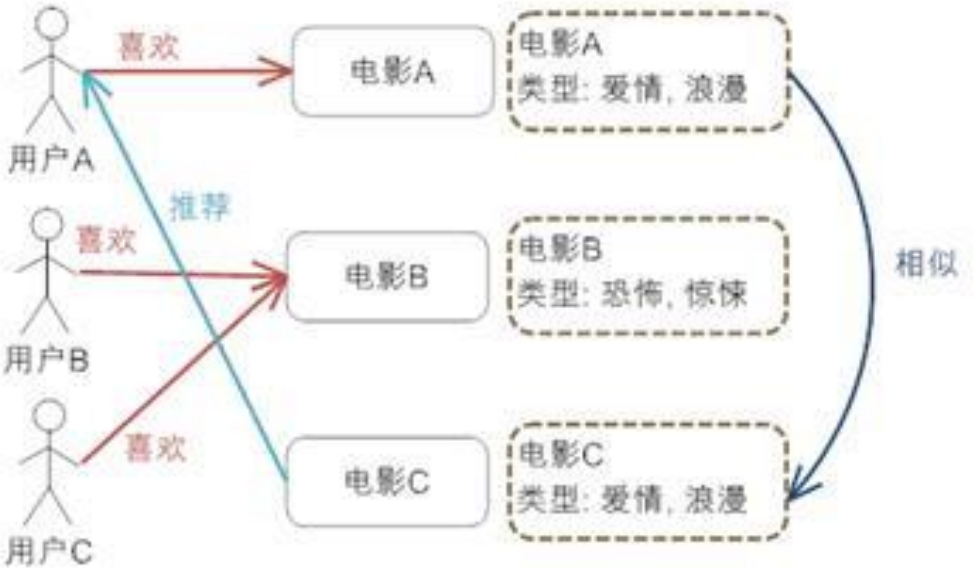
# 目录

# 2.1 推荐算法调研



**传统的推荐算法可分为两种: 一是基于内容的推荐;**

根据用户过去喜欢的**物品**（item），为用户推荐和他过去喜欢的物品相似的产品。




仅仅依赖物品特征衡量其相似度，忽略用户对物品的兴趣，缺乏个性化

# 2.1 推荐算法调研



**传统的推荐算法可分为两种: 二是协同过滤推荐。**

- 基于启发式的协同过滤推荐。可分为基于用户的协同过滤推荐和基于项目的协同过滤推荐。
- 基于模型的协同过滤推荐。如分类算法，设置一评分阈值，评分高于阈值--推荐，评分低于阈值--不推荐，问题变成了一个二分类问题。
- 基于图的协同过滤推荐。

# 2.1 推荐算法调研



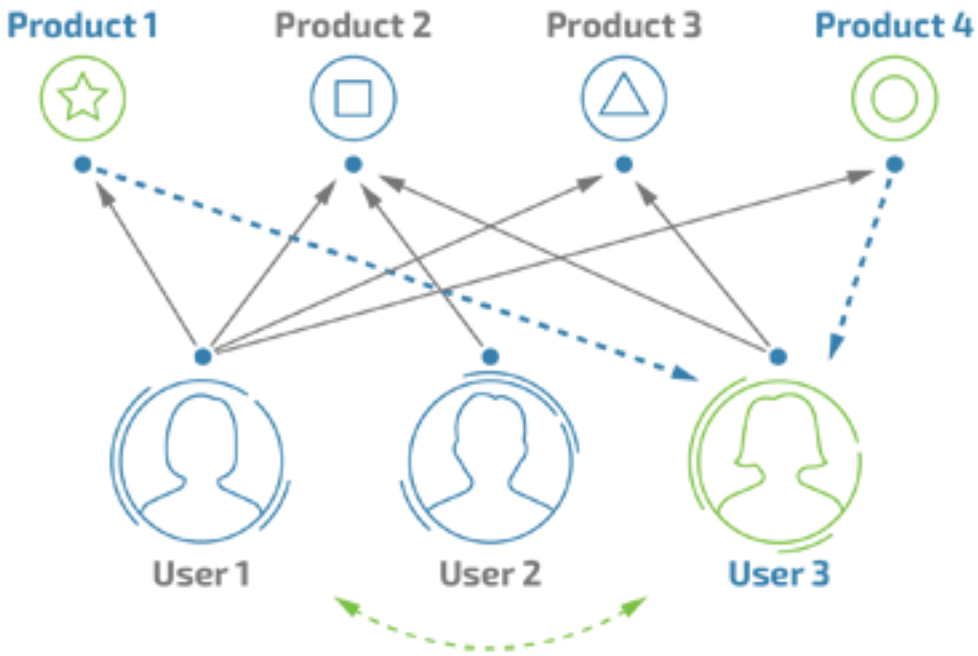
用户协同过滤算法

通过用户历史行为数据发现用户对内容的喜欢(如商品购买，收藏，内容评论或分享)，并对这些喜好进行度量和打分。根据不同用户对相同内容的态度和偏好程度计算用户之间的关系。

在有相同喜好的用户间进行商品推荐。

如果user1，user3用户都购买了p2，p3图书，并且给出了5星的好评。那么user1和user3就属于同一类用户。可以将user1看过的图书p1和p4也推荐给用户user3

前提：
需要知道用户相似性
故：适合用户数较少场景




基于历史数据，容易受到数据稀疏和冷启动问题的影响

# 2.1 推荐算法调研

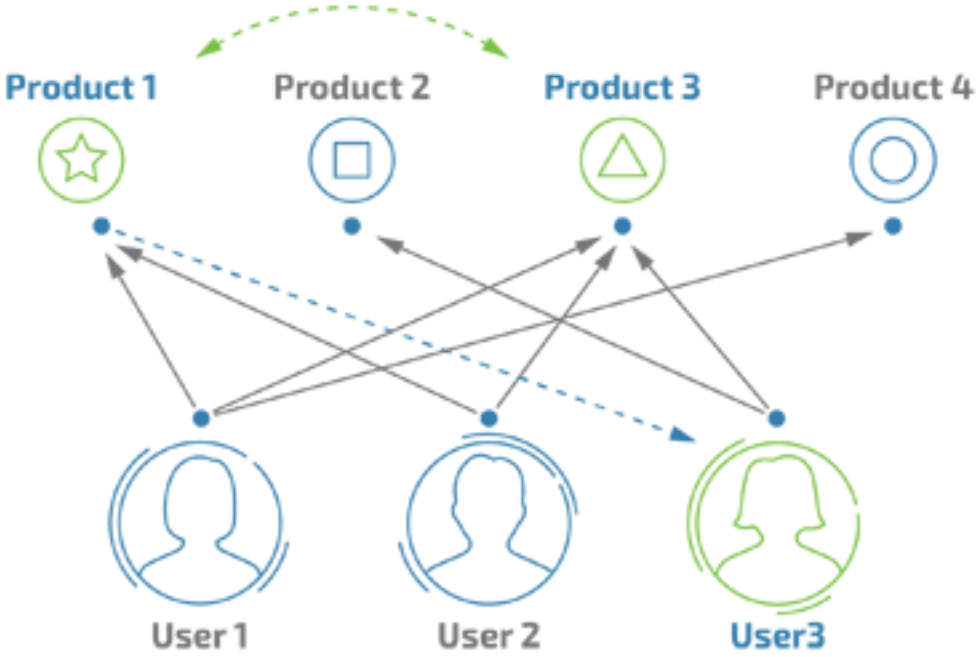


项目协同过滤算法

与基于用户的协同过滤算法类似，将商品和用户互换。
通过计算不同用户对不同物品的评分获得物品间的关系。
基于物品间的关系对用户进行相似物品的推荐。这里的评分代表用户对商品的态度和偏好。

在有相同特征的商品间进行商品推荐。

如果user1同时购买了p1和p3，那么说明p1和p3相关度较高。当user3也购买了p3时，可以推断他也有购买p1的需求。




使用较多：
根据用户喜好决定物品相似度
故：适合物品数明显小于用户数的场景

基于历史数据，容易受到数据稀疏和冷启动问题的影响

# 2.2 推荐算法设计



## 逻辑回归的推荐算法优势

基于CADAL海量数据的需求，采用基于逻辑回归的推荐算法。方法通过历史数据训练得到人工提取和构造的特征（用户特征、物品特征和用户-物品交叉特征）的权重，并以此去计算用户对物品的偏好程度。

- **形式简单，训练速度快** 是工业界中在海量数据推荐场景下应用最广的算法之一
- **特征的设计灵活可控** 除了从历史数据中提取特征之外，还可以人为地自由设计与组合出新特征
- **较好的可解释性** 从特征的权重可以看到不同的特征对最后结果的影响，某个特征的权重值比较高，那么这个特征最后对结果的影响会比较大。
- **可缓解冷启动问题** 对于没有任何历史记录的新用户，可根据提取得到的用户专业等特征进行书籍的推荐

# 2.2 推荐算法设计



## 基于逻辑回归的推荐算法步骤：

- 获取原始数据----用户行为、用户画像和资源属性
- 特征提取----用户特征和资源特征
- 特征工程----清洗、标准化、平滑化、离散化、特征组合、特征筛选等
- 训练算法模型并预测
- 输出结果展示

# 2.3 推荐算法验证



- **算法验证指标确定**
- **算法的验证**
  - 线下，线下测试数据可以从借阅历史数据中抽取测试数据。比如，将某个用户前几次借阅的图书当做训练集，将最后一次借阅的数据当做测试集对算法性能进行验证。
  - 线上，一个时间较长的过程，需要有真实用户进行借阅才可以进行性能的衡量。

算法选择：不同系统、用户各异（大学图书馆、cadal…）

# 2.4 数据结构分析



## 用户数据表

| 字段名 | 字段类型 | 说明 |
|---|---|---|
| id | Integer | 用户id，主键，自增长 |
| username | String | 用户名 |
| password | String | 密码 |
| status | Integer | 账户状态 |
| role_id_list | String | 角色id |
| real_name | String | 真实姓名 |
| balance | BigDecimal | 账户余额 |
| head_portrait_path | String | 头像地址 |
| booklist_notification | String | 我的书单被点赞或者收藏，是否通知 |
| comment_notification | String | 我的评论被点赞或者回复，是否通知 |
| booklist_update_notification | String | 书单更新，是否通知 |
| hidden_browsing_history | String | 是否隐藏个人浏览历史 |
| affiliation_library | String | 所属图书馆 |
| email | String | 邮箱 |
| sex | String | 性别 |
| birthday | Date | 生日 |
| hometown | String | 故乡 |
| habitual_residence | String | 常居地 |
| school | String | 学校 |
| major | String | 专业 |
| telephone | String | 手机号 |
| brief_introduction | String | 简介 |
| registration_date | Date | 注册时间 |
| account_type | String | 账号类型，独立账号D 子级账号Z 父级账号F |
| parent_id | int | 父级账号id |
| integral | int | 用户积分 |
| background_url_path | String | 人中心主页背景图片URL |
| recently_login | Date | 最近登录时间 |
| seetime | String | 阅读时间 |

## 用户借阅资源表

| 字段名 | 字段类型 | 说明 |
|---|---|---|
| userid | String | 借阅者id |
| username | String | 借阅者 |
| resourceid | String | 资源id |
| ssno | String | 资源ssno |
| resourcename | String | 资源名 |
| chapter | String | 借阅章节 |
| resource_image_url | String | 资源封面 |
| cover_image | String | 资源封面 |
| author | String | 作者 |
| borrow_start_time | String | 借阅开始时间 |
| borrow_end_time | String | 借阅结束时间 |

目前得到CADAL提供两个数据表结构：

1）用户数据表

2）用户借阅资源表（借阅历史）

*）缺少资源表

# 2.4 数据结构分析



## 书目资源数据表（图书馆）

| | | |
|---|---|---|
| 语种 | char | 10 |
| 馆藏地 | char | 10 |
| 编目日期 | datetime | 0 |
| 目录级别 | char | 10 |
| 文献形态 | char | 10 |
| 记录状态 | char | 10 |
| 国家 | char | 10 |
| MARC类型 | char | 10 |
| 记录号码 | char | 10 |
| 建档 | datetime | 0 |
| 更新 | datetime | 0 |
| 索书号 | varchar | 50 |
| 主要责任者 | varchar | 255 |
| 题名 | varchar | 255 |
| 版本 | varchar | 50 |
| 出版发行 | varchar | 50 |
| 载体形态 | varchar | 50 |
| 其它责任者 | varchar | 50 |
| 书目机构号 | varchar | 50 |
| ISBN/ISSN | varchar | 50 |
| 分类号 | varchar | 50 |
| 馆有 | varchar | 50 |


需要CADAL提供相关数据表真实数据：

1）了解数据组织方式

2）确定特征提取方式

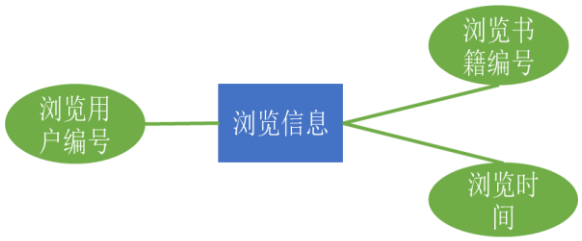
3）避免重复劳动

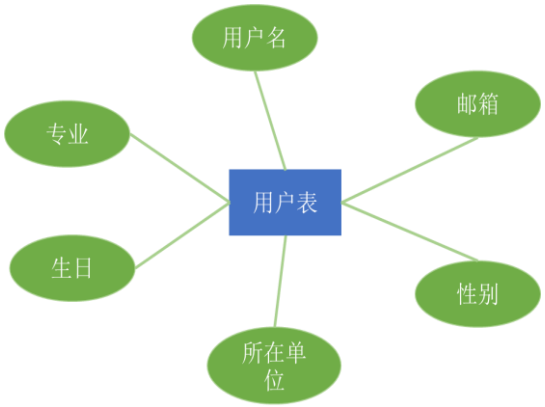
# 目录

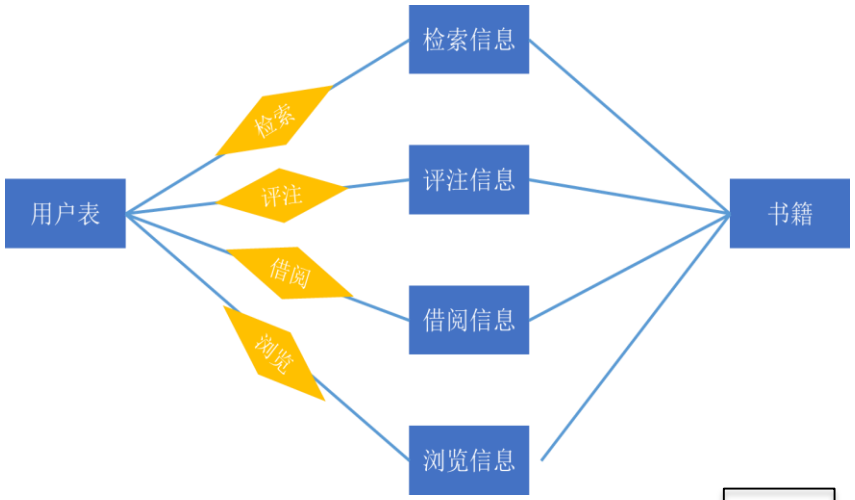
# 3.1 系统E-R图设计

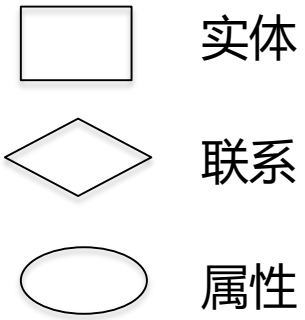


## 数据表E-R图



# 3.2 Demo系统



- 华师大图书馆数据（用户数据、资源和借阅历史数据）。
- 数据清洗，入SQL Server/MySQL数据库。
- 实现逻辑回归算法。
- 实现部分用户特征的提取。

# 3.3 特征提取



- 用户特征
  - 用户借阅资源的数目
- 图书特征
  - 资源被借阅的次数、资源语种
- 交互特征
  - 当前资源作者在当前用户已借阅图书资源作者中的数目
  - 当前资源的作者在当前用户已借阅资源作者中的占比
  - 当前资源被当前用户借阅次数
  - 当前资源在当前用户的借阅资源中占比
  - 当前资源的语种在当前用户已借阅资源语种中的数目
  - 当前资源的语种在当前用户已借阅资源语种中的占比

# 目录

# 4.1 下一步工作计划



- 数据预处理与特征提取

保证特征提取高质量

- 需要对基础数据进行预处理，处理成算法可接受的数据类型等
- 涉及到去重、去空格等特殊字符操作也是需要处理的过程
- 基于处理过的数据，进行多维度的特征提取。

# 4.1 下一步工作计划



- 算法代码开发

按照算法的设计方案，进行算法的开发

- 在算法开发的同时，注重算法性能和效果的兼顾
- 算法的开发将主要基于Python语言、Mysql关系型数据库。

# 4.1 下一步工作计划



- 算法性能验证

针对开发的算法模型，进行算法性能的验证

线下和线上两部分

- 线下部分
  - 在算法开发完毕时，从已有数据中提取部分数据当做测试数据，使用衡量指标对算法性能进行衡量
- 线上部分
  - 将针对线上借阅行为对算法性能进行测试

# 4.1 下一步工作计划



- 算法嵌入

算法结果-->存数据库，提供CADAL系统响应的API接口文件

- 用户端展示
  - CADAL系统可直接读取数据并在前台展示
- 数据更新
  - 采用离线更新方式，更高效的保证算法的运行。
  - 周或旬更新

# 感谢聆听

xli@dase.ecnu.edu.cn