《软件开发实践》课程报告

课题名称：eMousica 在线教育平台

课题负责人名（学号）：方涛（2015141463052）

同组成员名单（角色）：胡恒昌（前端开发）

姜美羡（测试）

指导教师：

评阅成绩：

评阅意见：

提交报告时间：200　　年　　月　　日

基于 OpenSource Learning 的在线教育平台

软件工程 专业

学生　方涛 胡恒昌 姜美羡　　指导老师　洪玫

[摘要]

在当前资源海量化，新兴技术日益更新的时代背景下。教师的查找资料，再学习的任务愈加艰巨，备课压力陡增。同时，学生缺少一款组织有序且与时俱进的学习工具。

本项目旨在开发一款基于包含爬虫、数据挖掘、个性教育的在线教育系统。用以将老师从繁重的查询资料、备课任务中解放。通过爬虫获取优质的互联网知识，进行文本挖掘的处理后，可以通过个性化推荐模块在既定的教学大纲内向学生推送最合适的内容，供其挑选与学习、以达到因材施教的效果。

预期将大大减轻教师备课授课的压力，也能极大程度的激发学生的学习兴趣，提高学习效率。

关键词：教育　网络　数据库　爬虫

1. 引言（项目背景）

1.1 中国在线教育行业概述

在线教育指的是通过应用信息科技和互联网技术进行内容传播和快速学习的方法。与传统教育机构的教育方式相比，在线教育具有效率高、方便（打破了时空限制，可碎片化学习）、低门槛、教学资源丰富的特点。基于上述特点，再加上“互联网+”推动，在线教育平台兴起，市场需求也与日俱增。

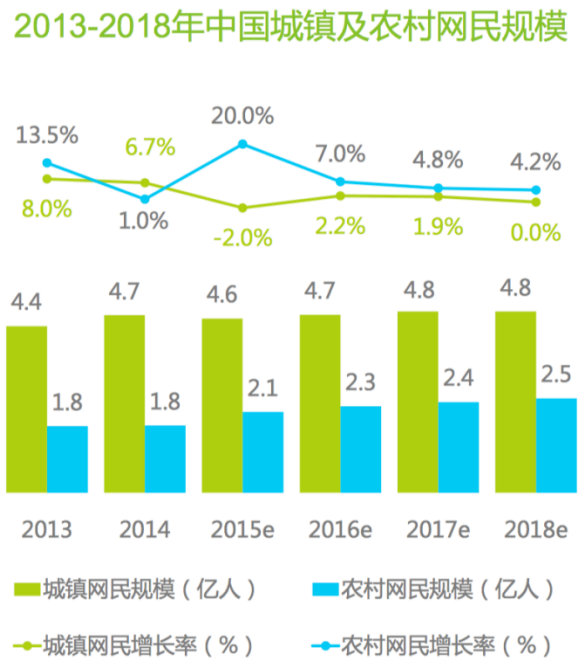
1.2 社会环境分析

1.2.1 移动端增速高于 PC 端网民，学生期待多种类型的学习终端

随着移动互联网的发展，移动端的网民增长速度明显高于 PC 端的速度，但网民数量在一定时期趋向饱和，增长速度也逐步放缓。但移动端流量数据的优势，让移动端成为下一个互联网的主要战场。可以预见，提早布局移动端市场，成为了当下互联网企业的最有发展潜力的战略。

1.2.2 农村地区网民开始快速增长，教育亟待普及

从我们咨询所获数据显示，农村网民的增长速度预计从15年开始



高于城镇网民，对在线教育而言，目前的主要战场在一线城市，但二三线城市有好的经济基础和好的网络环境，同时，对于优质的教育资源有较强的需求，是发展的蓝海。农村对于优质教育资源的需求也很强烈，但网络基础设施相对薄弱，在这种环境下发展移动互联网教育更符合用户选择，更具优势。

图1 中国城镇及农村网民规模图

在eMousika未来远大的预想中，这类基于互联网开源知识的学习工具也非常适合云贵川偏远地区的孩子们去使用。

2. 相关工作（对现有的系统进行分析）

2.1 行业生产上下游企业分析

图 2. 行业参与企业图

根据在线教育行业的运作方式，大致可以将行业参与企业分为四个方面：内容生产方，技术支持方，平台提供方，平台资源整合方。

内容生产作为在线教育行业的上游，其资源来源广、种类多样。包括新东方教育机构、华圆教育、北大青鸟、弘成教育以及好未来等传统教育培训机构，北京大学、清华大学、中国人大等知名学校机构，沪江、学而思网校、101远程教育网等互联网培训平台机构。

在产品加工方面，有科大讯飞、方直科技、启明科技、诺亚教育平台等企业提供技术设备，有道、知乎、猿题库、百度文库、印象笔记等提供教育工具，以及科大讯飞、天喻信息、拓维教育等提供教育信息化。

当前在线教育平台依据运作方式分为三类：B2C，B2B2C，O2O。B2C企业：沪江网校，学而思网校，101远程教育网，华圆网校、正保远程教育。B2B2C企业：百度教育、淘宝同学、CC课堂、YY教育、第九课堂等。O2O企业：决胜网、跟谁学、疯狂老师。

平台资源整合方面，主要由政府监管机构、行业监管机构、第三方教学评测机构、NPO&NGO等进行。

我们的产品—eMousika，意图在在线教育行业的流水生产线的上游，添加一条新的内容来源。把知乎、豆瓣、微信公众账号的文章利用起来，亦作为教育的材料推送给大家。结合当下新生一代与生俱来的『互联网产品亲切有好感』和当前市场竞争产品的空缺。eMousika一定会具有广阔的市场前景。

2.2 现有主要线上教育平台及分析

产品名称	简介	特色/优势分析
沪 江	沪江，是专业的互联网学习平台，致力于为用户提供便捷、优质的全方位网络学习产品和服务。	（1）较早布局了移动端,利用众多移动端产品抓住移动流量入口； （2）建立了学习资讯、学习社区、学习工具及学习平台四大业务体系,涵盖中小幼、语言、留学、职场、兴趣等丰富内容。
淘宝教育	旨在利用自身的用户、流量优势，为线下教育机构和在线教育机构搭建一个承载虚拟教育服务的平台。	（1）精准的分析和定位用户,根据用户购买的情况,推荐与购买实物相关的课程 （2）深入线下教育体系,为试点学校 提供相关的直播、录播、互动技术支持 （3）加强了对教育信息化硬件产品的开发,将在线教育嵌入更 多的互动元素
腾讯课堂	腾讯推出的专业在线教育平台,聚合大量优质教育机构和名师,下设职业培训、公务员考试、托福雅思、考证考级、英语口语、中小学教育等众多在线学习精品课程	（1）依托于QQ海量用户群,借助 QQ 语音技术、文本传输技术、 支付体系切入在线教育领域 （2）为机构和用户提供交互性学习渠 道,解决在线教育缺少交互性的壁垒
跟谁学	是一个 O2O 找好老师学习服务电商平台，在北京 2014 年 6 月由陈向东带领创建。	（1）教育 O2O 平台,为入驻机构、教师、学生同时创造价值。 （2）互联网技术+教育+平台+电商 （3）产品的自我驱动性增长+口碑传播 （4）多线条、多板块、多维度、 多地域的 O2O 运营 （5）个性化服务
一起作业网	中国最大的中小学在线学习平台，为老师、学生和 家长三方提供有价值的个性化互动教学服务。	（1）运用大数据对于教育领域的重塑,辅助学生不断的进行优化练习。 （2）引入出版商、培训机构、教育游戏、金融机构等等教育

（i）沪江

沪江作为互联网学习平台，为用户提供便捷、优质的网络学习产品和服务。自成立以来，沪江建立了学习资讯、学习社区、学习工具及学习平台四大业务体系，涵盖中小幼、语言、留学、职场、兴趣等丰富内容。近年来，沪江积极扶持互联网教育创业团队，努力打造在线教育生态圈，实现产业共赢。

2015年4月沪江进行品牌战略升级，“沪江网”正式更名为“沪江”，同时启用新Logo，旗下包含四大业务体系：“沪江网”专注于为用户提供学习资讯；“沪江社团”是用户的专业学习社区；学习工具如CCTalk、开心词场、小D词 典、听力酷等；录播平台“沪江网校”和直播平台“CC课堂”两大学习平台。品牌升级之后，“沪江”成为公司的母品牌，平台化和移动化成为公司当前的主要方向，构建为用户服务的完整生态平台。其优化升级后不再局限于录播模式，以用户为中心，围绕用户以免费+直播+录播的学习模式运营，并专注内容研发和优质内容的引入，为用户提供极致的内容体验。

沪江当下的学习方式可以总结为“系统化学习+碎片化学习”，系统化学习由沪江网校承载，碎片化学习由移动端承载。沪江较早布局了移动端，利用众多移动端产品抓住移动流量入口，既作为学习工具，又作为用户获取渠道。先后发布了直播课堂CCTalk、大沪江App，升级了开心词场，更好地实现移动端整合，抢占移动入口。

（ii）淘宝教育

淘宝教育坐拥海量流量数据，但要从淘宝引向淘宝教育，还需要精准的分析和定位用户，根据用户购买的情况，推荐与购买实物相关的课程，如设定关键词，然后匹配的方式等。现阶段，流量的转化和技术创新一样重要。

目前淘宝教育的课程内容课程覆盖外语学习、职业培训、学历考试、IT技能、营销管理、中小学辅导、学前教育、生活百科、文体艺术、美容养生等。淘宝开设和生活、美容等相关的课 程，以便吸引没有学习目的的用户和参与技能培训的用户，提升平台的用户留存率。

其挑战战略：

1.淘宝同学，真是更名“淘宝教 育”强化“淘宝的教育平台价值”；

2.转化淘宝流量，为入驻机构导流；

3.淘宝教育正加强对教育信息化硬件产品的开发，将在线教育嵌入更多的互动元素；

4.与淘宝实体物品交易结合，购买 实体物品，可获赠相关的使用培训课程，是引流的重要步骤；

5.深入线下教育体系，为试点学校提供相关的直播、录播、互动技术支持。

而淘宝教育未来的主要战略是完善淘宝教育生态链，实现“支付+内容+工具+平台+服务+互动”全产业链的贯通。近期内更多的通过平台优势，助力线下机构的互联网转型。

1.简化教育网店的开设流程；

2.提供直播、录播、互动交流的技术支持；

3.开发移动端应用，帮助企业打通移动端市场；

4.为PC端企业导流，实现在线教育O2O。

（iii）腾讯课堂

腾讯课堂依托QQ客户端及腾讯视频，做到即时课程互动。并最大限度模拟线下课堂；通过QQ群的天然群聚效应，老师用QQ群进行学生管理、互动，做到线上线下教育无缝连接。

其优势在于

1.腾讯课堂依托于QQ海量用户群，借助QQ语音技术、文本传输技术、 支付体系切入在线教育领域；

2.借助流量优势，成立在线教育平台，吸引商家入驻；

3.为机构和用户提供交互性学习渠道，解决在线教育缺少交互性的壁垒；

4.入驻机构在腾讯课堂达到一定评分之后将提供“万元广点通基金” 并安排专业人员进行推广指导，让教育机构在短时间内以“零成本”获得第一批学生。

综上所述，在当前在线教育行业大背景。创新性的基于细分的自动化挖掘open source的爬虫技术以及结合机器学习文本聚类的教育系统仍属市场空白阶段。市场潜力巨大，应用场景细分。属于强需求产品。有较好的发展前景。

3. 问题描述（需求分析）

3.1 产品描述

本项目全称为基于协同过滤的在线教育平台,是使用协同过滤算法实现的教 育系统。目标用户在于在校大学生群体,提供互联网的在线教育信息支持,同时 也满足自学者群体的需求。

随着网络技术的进步,互联网的普及,网络阅读成本降低。在校大学生能够 熟练使用计算机,是网络阅读的主要群体。调查显示,在线阅读、手机阅读、手 持式阅读器阅读等数字阅读方式开始普及。网络在线阅读排第一位,手机阅读排 第二位。数字媒介阅读代替书面阅读。廉价的电子书成为他们主要选择的阅读对 象。惯以纸质为载体的报纸、杂志和图书所占比例较小。深受大学生依赖的网络 阅读,呈现一种时效性、阶段性、冗杂性的本质,从阅读的形式上分析,这种网 络阅读就是一种浅阅读。浅阅读已经成为一种流行,作为网络时代新的阅读方式, 浅阅读除了与传统阅读一样获取信息外,更注重追求阅读过程中的视觉快感和心理愉悦,而难以获得实质上的深刻学习。而我们的项目希望将用户引向网络深阅读。

本教育平台主要面对在校大学生,同时可适用于在职人员和其他贫困山区求学学生。本软件核心功能在于对网络教育资源进行分类以及根据用户喜好进行智能推送,从而实现对用户的个性化教育。

3.2 产品功能

本项目旨在开发一款基于包含爬虫、数据挖掘、个性教育的在线教育系统。 用以将老师从繁重的查询资料、备课任务中解放。通过爬虫获取优质的互联网知识,进行文本挖掘的处理后,可以通过个性化推荐模块在既定的教学大纲内向学 生推送最合适的内容,供其挑选与学习、以达到因材施教的效果。产品主要功能包括：

(a)数据挖掘功能

数据挖掘一项从大量数据或者数据库中提取有用信息的技术,一般是指从大 量的数据中自动搜索隐藏于其中的有着特殊关系性(的信息的过程。数据挖掘通 常与计算机科学有关,并通过统计、在线分析处理、情报检索、机器学习、专家 系统(依靠过去的经验法则)和模式识别等诸多方法来实现上述目标。

本项目组主要利用数据挖掘技术分析标签内容和抓取的页面的内容。网络爬 虫是一种按照一定的规则,自动地抓取万维网信息的程序或者脚本。另外一些不 常使用的名字还有蚂蚁、自动索引、模拟程序或者蠕虫。爬虫是一个自动下载网 页的程序,它根据既定的抓取

目标,有选择的访问万维网上的网页与相关的链接，获取所需要的信息。

本项目采用深度优先策略,其基本方法是对于每个目标网站,按照深度由低 到高的顺序,依次访问下一级网页链接,直到不能再深入为止(即与该目标信息 网页相关信息全部爬取完毕)。 爬虫在完成一个爬行分支后返回到上一链接节 点进一步搜索其它链接。 当所有链接遍历完后,爬行任务结束。 这种策略比较 适合垂直搜索或站内搜索。

(b)数据分析功能

系统需要对储存的数据进行分析归类,按基础的属性分级及标签归类的方法,使数据可以供推送模块精确调用。本项目拟使用朴素贝叶斯算法实现对文本 的分类。

(c)智能推送功能

本平台能够根据用户历史的评分记录,结合所有用户的历史评分记录,对个 人推荐可能感兴趣的书名目录。

(d)登录、登出及注册功能

提供用户使用邮箱号进行注册的功能以及数据库表的支持。

3.3 系统处理数据及其描述

a. 网站登录

输入数据描述：

用户名(user_name),char 类型，允许输入 1-20 个字符；用户密码（password),char 类型，输入密码必须为英文字母加数字组合形式，且长度大于 8 个字符,并且加密

处理流程（算法）：

匹配数据库用户数据：数据库按用户名称搜索，之后匹配用户密码是否正确

输出结果描述：

用户信息正确：跳转网页至主页；用户名称不存在：提示不存在该用户，并提示重新输入信息；用户密码输入错误：网页显示密码输入错误，并提示用户重新输入或找回密码

b. 用户注册

输入数据描述：

用户名(user_name),char 类型，允许输入 1-20 个字符，且不能与用户密码（password),char 类型，输入密码必须为英文字母加数字组合形式，且长度大于 8 个字符,并且加密

处理流程（算法）：

向数据库中写入用户名字段以及对应的用户密码

输出结果描述：

用户名与用户密码格式均正确：用户点击提交并系统反馈用户注册成功。用户名格式不正确：提示用户名格式错误；用户密码格式不正确：提示用户密码格式错误；

c. 书本搜索

输入数据描述：

输入书名(search_item),类型为 char，长度不超过 50 个字符。

处理流程（算法）：

获取输入的数据，如果有书名号则用正则表达式去除，然后在数据库中匹配每条记录的 name 字段，匹配成功则返回，否则返回缺失。

输出结果描述：

在网页中由上到下显示匹配到的合适的书，若数据库中无记录则返回缺失条目信息

d. 用户评分

输入数据描述：

用户选择打分分数，分数处于 1-5 之间，只能为整数

处理流程（算法）：

保存用户打分、用户名、和书本 ID 在数据库中

输出结果描述：

如果数据库写入成功，则反馈用户打分成功信息；数据库写入失败，则返回失败信息并允许用户重新打分。

e. 收藏书本

输入数据描述：

每本书详细介绍界面提供给用户收藏按钮，用户点击即可收藏书本。

处理流程（算法）：

用户确认收藏书本后，系统保存用户名，书本 ID 在数据库中 Reference 表中用户确认收藏书本后，系统保存用户名，书本 ID 在数据库中 Reference 表中。

输出结果描述：

如果书本收藏成功（数据库写入成功）则返回用户收藏成功提示；如果书本收藏失败（网络原因等）则返回用户收藏失败提示并允许再次收藏。

f. 浏览记录

输入数据描述：

用户选择浏览记录页面。

处理流程（算法）：

服务器从数据库选择 History 表并根据用户名查找浏览记录，返回书本名与浏览时间。

输出结果描述：

页面返回书本名与浏览时间，仅显示 3 日内浏览记录，每一行显示书名与浏览时间；若用户近几日未访问本网站或因网络问题无法显示，则报出页面为空信息。

g. 书籍推荐

输入数据描述：

用户第一次注册时选择自己喜好的个人方向（书籍类型）

处理流程（算法）：

服务器存储用户名称与用户选择的个人标签，并在数据库中匹配该标签相符的书籍，将该类书籍定期推送；此外，还将使用协同过滤算法，根据用户长期的阅读记录自动推送可能喜欢的书籍

输出结果描述：

用户推荐列表生成正确则按列显示用户所喜欢的书籍名称、书籍标签；用户推荐列表生成失败则显示列表查询错误提示。

h. 标签系统

输入数据描述：

系统算法接收书籍 ID（int 类型）、书籍简介文本信息（char 类型，总长度不超过 3000 个字符）

处理流程（算法）：

服务器接收书籍简介后，对文本进行分词并建立单词向量，输入贝叶斯分类算法中计算与各个类别的相似度，生成标签矩阵，并根据设定的阈值划分可能的标签,最后将该标签集合和书本 ID 存入数据库

输出结果描述：

输出书籍分类列表（即近含有 0 和 1 的标签向量）并存入数据库中可由网页端查看;列表生成错误则返回空值并不予写入数据库。

i. 个人主页

输入数据描述：

用户成功登陆状态下选择个人主页选项，系统自动向服务器发送用户名字段与个人主页请求

处理流程（算法）：

系统向数据库中在用户信息表中依据用户名查找相应列，并返回列中各字段

输出结果描述：

输出用户名，用户年龄，所在城市，个人爱好等信息

j. 用户评论

输入数据描述：

用户输入评论文章（char 类型，总长度不超过 1000 个字符）

处理流程（算法）：

网页向服务端提交用户名，提交时间和评论内容，服务器将改信息按列存入数据库中评论表中

输出结果描述：

网页输出用户名，评论内容，评论时间，并且评论按时间顺序显示

3.4 系统运行环境

客户端:操作系统 Windows/Mac OS/Android/IOS

服务端:操作系统 Windows Azure Cent OS

数据库 SQL Azure 操作系统:Windows 10

开发语言:Python / html / CSS / javascript

开发工具:Visual Studio 2013 Professional / IE / Pycharm / Chrome

4. 解决方案（概要设计）

4.1 开发过程模型

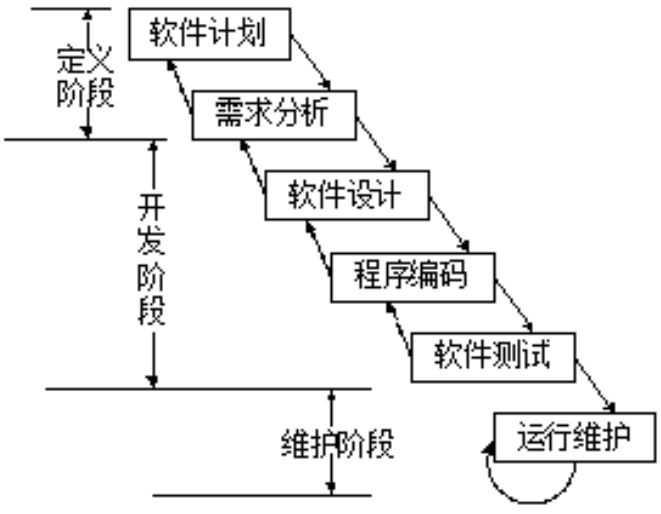


图. 瀑布模型

我们的项目开发采用瀑布模型的开发方式，将软件生命周期划分为制订计划、需求分析、软件设计、程序编写、软件测试和运行维护等六个基本活动，并且规定了他们自上而下、相互衔接的固定次序，如同瀑布流水，逐级下落。有利于开发过程中人员的组织、管理，从而提高了大型软件项目开发的质量和效率。

4.2 开发标准规范

4.2.1 接口设计

4.2.1.1 用户接口

(a) 用户注册接口

在 emousica.com/signup 网址下用户输入相关注册信息(包括注册邮箱、密码、 用户昵称以及喜好标签),提交表单完成注册功能。注册成功,系统则返回注册成功提示并跳转至主页面。

(b) 用户登陆接口

在 emousica.com 网址下用户输入相关登陆信息; 登陆成功,系统自动跳转至主页面并显示用户在线。

(c) 用户图书搜索接口

用户在 emousica.com/index 网址下,在搜索框中输入图书相关信息,点选搜索 后系统返回数据库中所存储的相关图书索引列表。

(d) 用户评论接口

用户在 emousica.com/index/content 网址下,在评论页面输入相关评论,点选发 表评论后系统刷新页面并返回提示评论成功,显示用户评论信息。

(e) 用户打分接口

用户在 emousica.com/index/content 网址下,在书本内容页面输入评分分数,点 选发表评分后系统刷新页面并返回提示评分成功,显示用户评分信息。

4.2.1.2 外部接口

其他支持软件:

(a) 结巴分词: 输入文本内容,该软件处理后返回分词后的列表结果。

(b) Azure 文本关键词提取: 输入文本内容,该软件处理后返回提取关键词的列表结果。

4.2.1.3 内部接口

说明本系统之内的各个系统元素之间的接口的安排。

(a) 爬虫模块:

Public void getBookSources(int start , int end ,String keyWords)

调用该模块时输入开始爬取的网页数,爬取结束时的网页数,以及关键词,该 模块将从目标网页按照关键词搜索后从开始到结束页数爬取书本关键信息,包 括书名、作者、图片(URL)以及书本内容等。爬取成功之后自动存入服务器 上的数据库中。

(b) 文本分类模块

Public String[] classification(String text)

调用该模块时输入需要分类文本的内容,该模块调用结巴分词自动将文本内容 分词成数组并使用贝叶斯分类算法进行分类,最后返回可能的标签的数组。

(c) 用户推荐模块:

Public String[] recommender(int userID)

调用该模块时输入目标用户 ID,该模块将从数据库中自动调用所有用户的历史 打分记录,并结合该用户的历史打分记录对该用户进行喜好推荐,返回可能喜 好书目的列表。

(d) 用户注册模块:

Public boolean sign_up(int password , String mailbox , String name,String[] labels)

调用该模块时输入用户注册时输入的邮箱地址、用户昵称、密码和偏好标签, 该模块自动匹配用户是否在数据库中已经存在,若不存在则写入数据库中。以 布尔形式返回是否注册成功。

(e) 用户登陆模块:

Public boolean login(int password , String mailbox)

调用该模块时输入用户注册时输入的邮箱地址、密码,该模块自动匹配用户是否在数据库中已经存在,并匹配用户名和密码是否匹配。以布尔形式返回是否登陆成功。

4.2.2 运行设计

4.2.2.1 运行模块组合

(a) 获取数据库书籍资源:

先调用爬虫模块,该模块将从目标网页按照关键词搜索后从开始到结束页数爬 取书本关键信息,包括书名、作者、图片(URL)以及书本内容等,按表存入数据库中,之后调用文本分类模块,向该模块中传入爬取的文本的内容,并使用结巴分词模块对文本内容进行分词,之后使用贝叶斯算法进行分类,最终将生成的文本分类标签存入数据库中。

(b) 生成用户推荐列表:

平时用户使用打分评论模块,该模块存储用户 ID,书本 ID 和用户对该书的评分在数据库中,之后生成用户推荐列表的时候调用推荐模块,输入用户 ID,该模块从数据库表中调取用户评分记录,使用协同过滤算法生成用户推荐列表并返回。

(c) 用户完成注册:

用户填写表单后发送提交表单请求,系统先调用注册模块对用户信息是否符合格式进行检查,之后调用查重模块遍历数据库检查用户数据是否已经存在,最终返回注册结果。

4.2.2.2 运行控制

(a) 运行爬虫与分类并存储:

管理人员在服务端运行 bookSources.py 脚本,输入开始结束页面数以及关键词, 等待程序运行结束提示。

(b) 运行用户推荐模块:

用户发送请求后根据用户传入的用户 ID,服务端运行 recommend.py 脚本,运行完毕后返回用户推荐列表。

(d) 注册功能:

用户填写表单后发送提交表单请求,系统运行注册模块检查输入信息是否合理后存储用户信息。

4.3 运行时间

说明每种运行模块组合将占用各种资源的时间。

(a) 运行爬虫与分类并存储:

运行期间占用带宽与内存; 结束后写入数据库,运行时间与爬取资源数量有关; 平均爬取一页运行时间:10min。

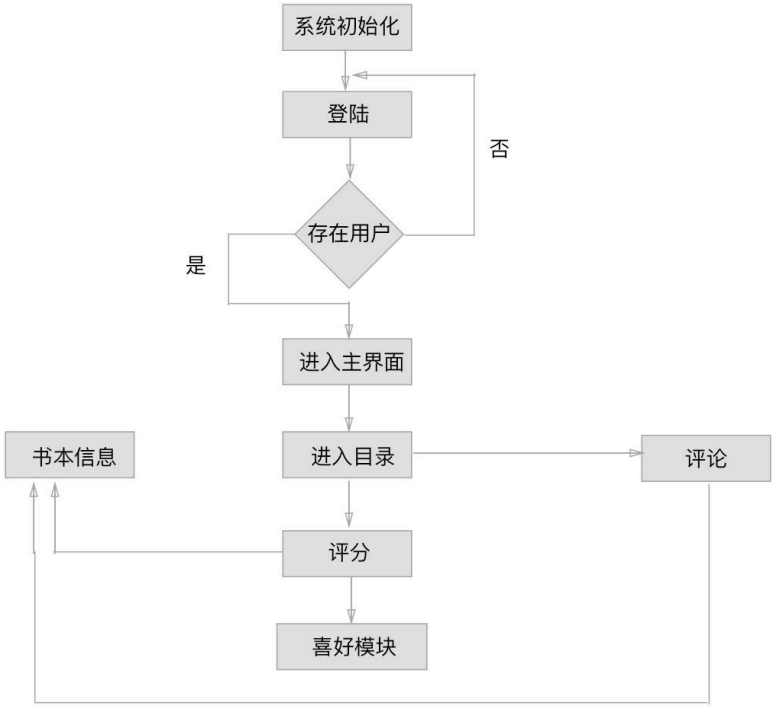
(b) 运行用户推荐模块:

占用内存,使用带宽,平均每个用户生成推荐列表处理时间不超过 10s。

(c) 注册功能:

写入数据库, 平均完成一个用户信息注册时间不超过 1S。

4.3 系统整体框架结构



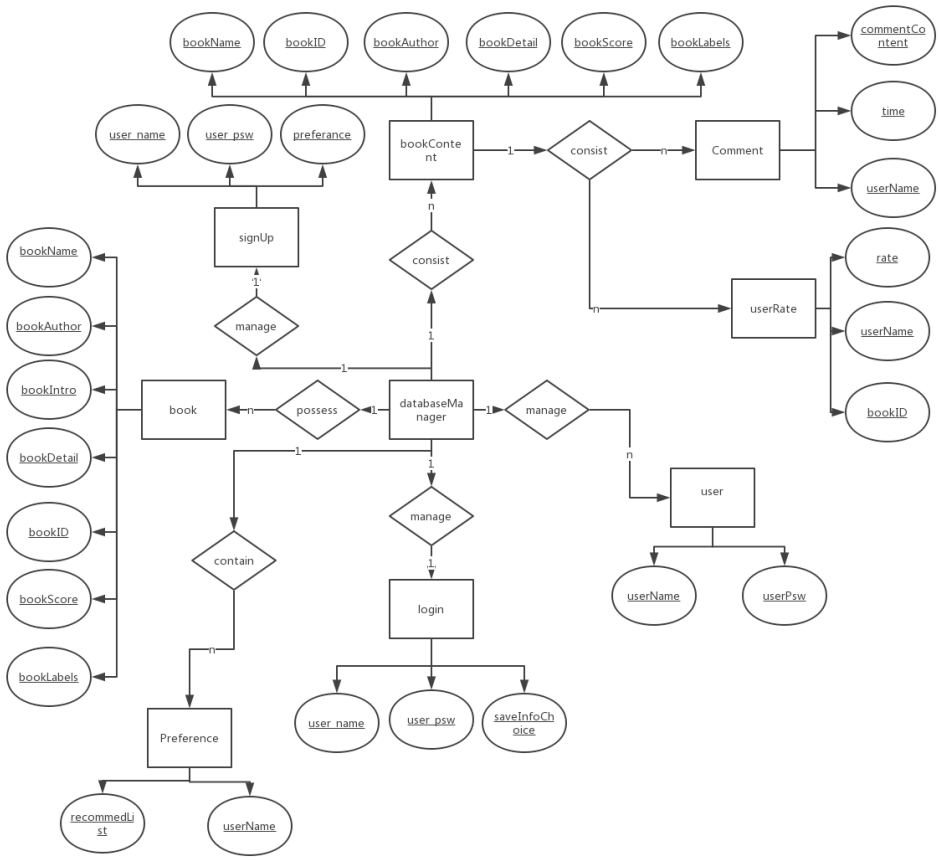
我们程序采用集中式系统,一个主机带多个终端。终端没有数据处理能力,运算全部在主机上进行。是将项目集中存放在中央服务器中,在工作的时候,大家只在自己电脑上操作, 从同一个地方下载最新版本,然后开始工作,做完的工作再提交给中央服务器保存。

系统元素名称	元素功能
推荐系统	系统需要根据用户自身的偏好信息定向提供 相符的资源。先收集用户偏好,即找到相似 用户和物品,计算用户间以及物品间的相似 度,最终找到相似的用户或物品。
文本分类系统	本项目采用朴素贝叶斯算法实现。服务器接
	收书籍简介后,对文本进行分词并建立单词 向量,输入贝叶斯分类算法中计算与各个类 别的相似度,生成标签矩阵,并根据设定的 阈值划分可能的标签,最后将该标签集合和 书本 ID 存入数据库。
数据管理模块	管理用户数据的注册、提取、校验、登陆， 以及用户浏览

	历史记录的存储调取,用户喜 好标签的存储、修改、调取。
用户界面模块	提供与用户交互接口,允许用户在系统进行 登陆、注册、浏览、信息修改等操作。
爬虫模块	采用开源爬虫框架 scrapy 进行分布式爬虫集 群操作。采用超文本分类和随机森林聚类算 法进行分类。根据网页链接网页的相关类型 对网页进行分类,依靠相关联的网页推测该 网页的类型。一个自动下载网页的程序,它 根据既定的抓取目标,有选择的访问万维网 上的网页与相关的链接,获取所需要的信息。 用深度优先策略,其基本方法是对于每个目 标网站,按照深度由低到高的顺序,依次访 问下一级网页链接,直到不能再深入为止(即 与该目标信息网页相关信息全部爬取完毕)。
搜索模块	获取输入的数据,如果有书名号则用正则表 达式去除,然后在数据库中匹配每条记录的 name 字段,匹配成功则返回,否则返回缺失。 最终生成搜索结果列表返回。

4.4 系统数据结构设计

4.4.1 逻辑结构设计要点



本程序逻辑结构设计要点在于数据结构模块化划分的清晰，一级逻辑数据模块包括：login、preference、book、signup、bookContent、Comment、UserRate、user，这些模块彼此相对功能独立，但是也有千丝万缕的联系。

Login 是本平台的登入的第一个模块，其包括 user_name(用户名)、user_psw（用户密码）、saveInfoChoice（保存个人化信息选项），这些二层数据对接下来饿模块也具有影响，比如对用户的 user 数据

库操作的时候，需要将 user_name 保存下来然后对服务器进行相关信息的提交，做个人化信息、书籍的推出的时候，也需要用到 saveInfoChoice。

Book 包括 bookName（书籍名字）、bookAuthor（书籍作者）、bookIntro（书籍简介）、bookDetail（书籍内容）、bookID（书籍编码）、bookScore（书籍评分）、bookLabel（书籍标签），这些书籍也与 bookContent、userRate 里的而成逻辑数据有着紧密的联系。

4.4.2 物理结构设计要点

存取方法：调用外部依赖库 mysqlclient，使用 Django 的 model 模版创建表结构，并且在 view 里进行相关操作。

存取单元结构：根据逻辑结构的指标以及 DBMS 支持的数据类型，所确定的数据项的存储类型和长度以及元组的存储结构等，即：数据文件及其数据项在介质上的具体存储结构。

存放位置；指根数据库文件和索引文件等在介质上的具体存储位置。

存储介质： 用于存储文件的物理存储设备包括磁盘、磁带、光盘、磁盘阵列、磁带库、光盘阵列，具体包括：介质容量的大小、存取速度与费用

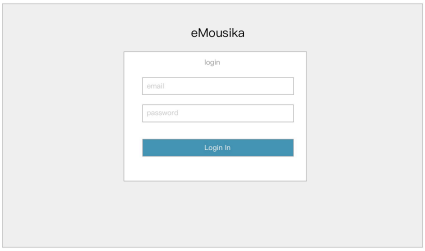
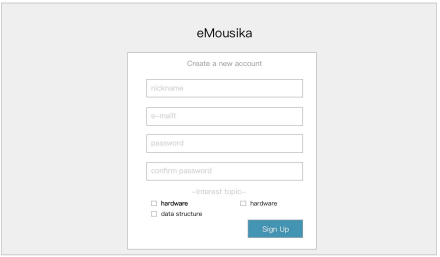
4.4.3 数据结构与程序的关系

服务器程序在对用户进行增删改、书籍进行增删改、评分进行增删改、喜好推荐表进行修改的过程中需要对数据库数据结构也就是数据表进行查询和修改，在查询用户、书籍等操作中都需要对数据库中的所有表，进行联合查询、修改。

物理数据结构主要用于各模版之间函数的信息传递，接口传递的信息将是以数据结构封装了的数据，以参数传递或返回值的形式在各模版间传输。出错信息将送入显示模版中，一些数据的测试模块则送入准备模块中准备打印格式。

我们程序的数据结构在程序上的操作时线性的操作，从表的一段开始，向另外一端逐个按给定值与关键码进行比较，若找到，则是查找成功，并给出数据元素在表中俄位置，若整个表监测完，未找到相同的关键码，则查找失败，给出失败信息。从数据结构的逻辑关系层面考虑，顺序查找的方向是可以从左到右，也可以是从右到左。但是如果进一步考虑存储结构，该结论就不一定正确，比如单链表只能从左到右，如果决定使用链表，又要考虑从右到左的查找，显然必须启用双向链表，为了操作方便性而付出空间代价。

4.5 系统的界面设计

<table>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>登录界面</td><td>注册界面</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>主界面</td><td>欢迎界面</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>阅读界面</td><td>评论节目</td></tr>
</table>

5. 实现细节（详细设计、实现）

5.1 模块描述

图16 系统架构图

我们的平台包括以下功能模块：

1.爬虫模块：主要实现平台基本的社交的功能的划分和实现；

2.分析归类模块：主要借助文本分析的手段，对用户正在做的项目进行相关资料的自动搜集，以及用户感兴趣的话题的自动搜索；

3.智能推荐模块：主要根据用户的标签和想要做的创意对用户希望找到的朋友进行便签分析以及历史数据，比如日志的文本信息的分析；

4.用户行为分析模块：主要实现用户对于组群功能需求的满足，我们使用灵活的数据库结构来满足这一设计；

5.2 关键技术

5.2.1 数据挖掘技术

数据挖掘一项从大量数据或者数据库中提取有用信息的技术，一般是指从大量的数据中自动搜索隐藏于其中的有着特殊关系性（属于Association rule learning）的信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

本项目组主要利用数据挖掘技术分析标签内容和抓取的页面的内容。

5.2.2 文本信息提取技术

- **词干提取、去停用词**

在英语等欧洲语言中，语素是语言中包含有意义的最小单元，所有的词均由语素构成，语素包括词干和词缀两部分，词干承担词的主要涵义，词缀承担词的附加涵义。在文本分类中，为了减少文本处理时的特征维数，提高文本分类的效率和精确度，需要对被分类文本中的所有英语等欧洲语言的词汇中的词缀部分进行去除，只保留词干部分，这部分工作称为词干提取。从词干提取使用的具体技术手段划分，词干提取的算法可划分为四个类别：基于规则的词干提取算法、词典查找法、基于统计的提干提取算法、混合词干提取算法。

● **文本结构化**

经过文本预处理后，一个完整的文本已经转化为适合文本处理的中间形式文本。但是，文本中的信息是无结构的数据，这不仅不便于计算机的读取，也不便于对文本内容进行各种统计，导致文本分类难以高效、顺利地进行。因此，为了克服上述不便之处，在文本分类前，需要把所有被分类的文本进行结构化表示。

文本结构化表示需要使用一定形式的数据结构，这种数据结构称为文本表示模型。文本表示模型有很多种，例如布尔模型、概率模型、基于图结构的文本表示模型等。现在使用最为广泛的文本表示模型是向量空间模型（Vector Space Model，VSM）。

6. 实验验证（测试、分析、评价）

此功能测试用例对测试对象的功能测试应侧重于所有可直接追踪到用例或业务功能和业务规则的测试需求。这种测试的目标是核实数据的接受、处理和检索是否正确,以及业务规则的实施是否恰当。主要测试技术方法为用户通过 GU (I 图形用户界面) 与应用程序交互 , 对交互的输出或接受进行分析,以此来核实需求功能与实现功能是否一致。

参考文献

[1]刘露,彭涛,左万利,戴耀康. 一种基于聚类的 PU 主动文本分类方法[J]. 软件学报,2013,11:2571-2583.
[2]平源. 基于支持向量机的聚类及文本分类研究[D].北京邮电大学,2012.
[3]杨杰明. 文本分类中文本表示模型和特征选择算法研究[D].吉林大学,2013.
[4]李荣陆. 文本分类及其相关技术研究[D].复旦大学,2005.
[5]王煜. 基于决策树和 K 最近邻算法的文本分类研究[D].天津大学,2006.
[6]苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,09:1848-1859.
[7]周平红. 我国高等教育信息化水平测评与发展预测研究[D].华中师范大学，2012.
[8]范坤. 推进我国教育信息化建设进程的对策研究[D].华中师范大学，2004.
[9]牛龙飞. 基于 Web 的我国教育信息化公共服务平台的设计与实现[D].华中师范大学，2013.
[10] 艾瑞咨询 2015 中国在线教育行业发展报告

附录：

1. 项目开发计划；
2. 软件需求规格说明书；
3. 软件设计文档；
4. 软件源代码；
5. 软件测试文档；
6. 用户手册；