基于协同过滤的在线教育平台

需求规格说明书

版本<1.0>

目录

1 引言----------3

1.1 目的----------3

1.2 范围----------3

1.3 定义、简略----------3

1.4 引用文件----------4

2 总体描述----------4

2.1 产品描述----------4

2.2 产品功能----------5

2.3 用户特点----------6

2.4 约束----------6

2.5 假设和依赖关系----------6

2.6 需求分配----------7

3.具体需求----------7

3.1 外部接口需求----------7

3.2 功能需求----------8

3.3 性能需求----------10

3.4 设计约束----------11

3.5 软件系统属性----------11

3.6 其他需求----------12

4 附件----------12

1、引言

1.1 目的

本需求规划书的目的在于通过本文档定义基于协同过滤在线教育平台产品的需求，以求在项目组员与相关成员之间达成一致的需求描述。

预期读者：项目小组成员，利益相关者。

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。

1.2 背景

a. 待开发的软件系统的名称:基于协同过滤的在线教育平台；

b. 本项目任务由方涛提出，开发者主要由胡恒昌、姜美羡担任，目标用户是贫困山区儿童、普通在校大学生和在职人员。

c. 该软件与其他教育平台、知识科普网络存在数据上的交互，且依赖大量网络教育资源。

1.3 定义

（1）在线教育：

或称远程教育、在线学习，现行概念中一般指的是指一种基于网络的学习行为，与网络培训概念相似。

（2）文本分类：

文本分类用电脑对文本集(或其他实体或物件)按照一定的分类体系或标准进行自动分类标记。

（3）网络爬虫（Reptilia）：

是一种自动获取网页内容的程序。是搜索引擎的重要组成部分，因此搜索引擎优化很大程度上就是针对爬虫而做出的优化。

（4）协同过滤推荐（Collaborative Filtering recommendation）：

协同过滤分析用户兴趣，在用户群中找到指定用户的相似（兴趣）用户，综合这些相似用户对某一信息的评价，形成系统对该指定用户对此信息的喜好程度预测。

（5）朴素贝叶斯算法（Naive Bayesian Model）：

朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法。

1.4 参考资料

[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

2、总体描述

2.1 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。

2.2 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。

(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。

2.3 用户特点

本教育平台针对对于知识渴求，但是又苦于线上互联网信息的冗杂、线下信息更新的缓慢的苦恼的学习者，他们中包括苦于备课的老师、缺乏自主学习的在校学生，渴望教育平台和二次学习资源的救治人员，互联网线上自学者，缺乏教育人力物力的偏远山区的孩子等等。同时我们的产品主要针对互联网使用者推出，应用于 PC Web 端。

2.4 约束

a.文本分类、网络爬虫、智能推荐算法功能的实现不能配置于客户端，应在服务端配置；

b.开发语言使用 Python2.7

c.开发环境选择 Linux

d.文本分类正确率应大于 70%

2.5 假设和依赖关系

a.使用的组件：服务器搭建使用 django 服务端框架；前端 WEB 网页实现使用 BootStrap 商业组件；

b.界面设定：使用绿色作为基调颜色，且网页基准色调提供接口随时改变；

c.用户设定：具备互联网基本常识；

d.工期：项目核心功能实现工期在 30 个工作日以内；

e.经费：无；

f.人员：本项目核心开发人员为三位：方涛、胡恒昌、姜美羡；

g.设备:服务端使用 linux 操作系统开发;客户端支持普通 PC 使用谷歌 Chrom 等浏览器　　登录本网页；

2.6 其他需求

a.提供数据库管理和查询的接口；

b.提供用户依据标签对书目进行查询的功能

3、具体需求

3.1 外部接口需求

3.1.1 用户界面

- 拥有登录、注册界面让用户以个性化账号进入平台，从而获得个性化服务
- 主界面能够显示不同的知识范畴的领域提供给用户选择
- 次界面侧边栏能够将知识领域已层级结构显示爬的所有书籍
- 顶栏可以查看弹出个人信息框，可以选择收缩侧边栏
- 设置框可从右侧弹出并且进行基本设置
- 书本介绍界面可以对书本进行打分和查看、发表评论

3.1.2 软件接口

- 通过微软 Azure 云服务 Cognitive Services 套件中自然语言文本关键词提取模块的对媒体内容进行分级与分类
- 访问微软 Azure 函数库 API 以调用 MS 功能
- 第三方哑巴分词 Python 函数库针对中文切词

3.1.3 通讯接口

- HTTP 协议：前端与服务器通过 Post 和 Get 方法获得和发送信息，实现数据的交互传输。
- Socket：就是服务端不断监听双方约定好的端口号，客户端通过服务器 IP 去请求链接，连接成功后，得到 Socket 的输入输出流，保证登录操作的必须性。

3.1.4 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

3.2 功能需求

3.2.1 功能性需求分类

一级模块	二级模块	功能点编号及名称	优先级
M1_登入系统	M1_C1 登录	F_1 提交登录表单	P1
		F_2 忘记并找回密码	
	M1_C2 注册	F_3 提交注册信息表单	P2
M2_查阅书籍	M2_C1 目录结构找寻书籍	F_4 主界面显示大领域方向	P1
		F_5 次界面显示领域方向下的知识分级	P1
	M2_C2 搜索栏寻找	F_6 通过直接检索数据库返回	
M3_反馈信息	M3_C1 举报功能	F_7 举报反馈错误信息的按钮	P2
	M3_C2 打星反馈	F_8 记录个人对特定书籍的喜好	
		F_9 载入书本热度信息	
	M4_C3 评论	F_10 录入书本的评论数据库	
M4_推荐系统	M4_C1 推荐个性化书籍	F_11 根据个人浏览记录和打星状况给个人进行个性化推荐书籍	
	M5_C2 学习热门推荐	F_12 按照当下点击量和访问量来给所有人推荐当前学习热门	

3.2.2 登入系统模块

- 有效输入核查

i. 登录、注册表单内 email 的 input 栏需要包含‘@’字符

ii. 登录、注册表单内 nickname、password 由字符串组成

iii. 注册表单内的 checkbox 可以选择大于等于 0 项

- 操作的准确顺序

i. 登录表单，输入有效值，点击‘登录’按钮 submit 表单值

ii. 注册表单，输入有效值，点击‘注册’按钮 submit 表单值

● 异常状况

i. 网络状况异常，信息通讯交换不稳定

ii. 输入值非法，如长度溢出、不符合检测条件等

3.2.3 查阅书籍模块

● 有效输入核查

i.目录结构查阅书籍的有效输入即为 mouse 左键单击带有 href 的已注册链接的按钮

ii. 搜索书籍的有效输入为在 input 框内输入合法字符串后输入‘回车’

● 操作的准确顺序

i. 目录结构查阅书籍：先在主界面选择特定的知识领域，进入次界面后按照目录分层的结构引导找到自己想要阅览的书籍。

ii. 搜索书籍：在输入框内输入自己想要查询的标签对应的书籍，即在右侧可以列出书本。

● 异常状况

i. 带有 href 属性的元素的值未在注册表单里登记

ii. 查询的字符串非法

iii. 网络状况异常，信息通讯交换不稳定

3.2.4 反馈模块

● 有效输入核查

i. 举报功能：单击“举报”按钮即可对书本内页的内容进行质询和举报

ii. 打星反馈模块选择：mouse 左键单击“star”的 div 元素可以对星星进行调整

iii. 打星反馈模块提交：mouse 左键单击“submit”可以将实时监测的 star_number 通过 Post 方法传递给服务器。

vi. 评论输入:在 Commend 框内输入合法字符串

v. 评论提交：mouse 左键单击“Send”讲评论的内容插入进评论内容数据库

● 操作的准确顺序

i. 举报功能：发现内容错误，即可以点击“举报”进行内容检举。

ii. 评星功能：先选择星星的数量（1～5）来代表自己对此书的喜好程度，再点击“Submit”保存到个人信息数据库和书本数据库。

iii. 评论功能：在书本页面选择“Commend”栏目，在评论区的最下方的输入行表单内输入自己的评论，点击“Send”将其保存到数据库，并且可以在网页动态渲染时查看到。

● 异常状况

i. 网络状况异常，信息通讯交换不稳定

ii. 输入值非法

3.2.5 推荐模块

● 有效输入核查

i. 进入推荐模块：在平台主界面的上方导航栏选择“Hot topic”和“Recommend”即可看到系统推荐的学习图书资源。

● 异常状况

i. 网络状况异常，信息通讯交换不稳定

3.3 性能需求

3.3.1 处理能力

目前系统处理能力主要与服务器的访问承载量有关，目前服务器性能较差，大约承载量的最大用户数为 40 人左右。

3.3.2 响应时间

Web 使用页面响应时间按照时段来分级：

时间段	操作	响应时间（秒）
平时	登录 / 注册	0.3～0.8
	选择目录操作	2～5
	查询书籍	4～10
	评论、打星评分	2～3
访问高峰	登录 / 注册	1～3
	选择目录操作	4～8
	查询书籍	10～15

	评论、打星评分	3～6

3.4 设计约束

- 平台约束为 Web 端，因其拓展性高，以浏览器为支撑能够兼容多个平台运行。
- 使用编程语言为 Python、JavaScript
- 产品语言暂时约束为英文，因目前 Azure 文本聚类只支持英文。
- 数据命名首字母大写

3.5 软件系统属性

3.5.1 可靠性

软件有用，而且好用，但是如果经常突然宕机或无法使用，用户肯定不会满意这种提心吊胆的感觉。因此软件还必须可靠，让人放心使用。

软件的可靠性质量属性，主要体现了系统持续不间断地满足客户相关应用目标的能力。它们属于软件外在的与其使用价值可获取度有关的质量属性。其决定性因素包括其成熟度，成熟性—与由软件中的缺陷而造成故障的频度相关的软件属性。例如软件的系统缺陷率、平均无故障时间 MTBF 等。

3.5.2 可用性

与软件持续正常运行，而可供用户使用相关的软件属性。例如平均可用时间等。如果在用户急需软件来完成某项任务时，系统却不能使用，那么将有可能严重损坏客户的利益。

3.5.3 安全保密性

只有授权用户才可以对数据库、用户进行管理和增删修改。登录账户的权限不同，而且必须防止信息的非法、非授权的泄漏。

识别检测对象的系统资源，分析这一资源被攻击的可能指数，了解支撑系统本身的脆弱性，评估所有存在的安全风险

设备备份机制、容错机制，防止在系统出现单点失败时，系统的备份机制保证系统的正常运行。

3.5.4 可维护性

一旦遇到故障，软件重新恢复工作性能所需的时间越短，受损数据的修复程度越高，则因故障而造成的损坏就越轻。

软件在运行时，有可能会遇到故障（例如硬件失效），或者其某些特定接口部分遭到侵害;在此情形下，软件应当仍然保持一定工作能力，从而避免彻底中断服务而造成的更大损失。

3.5.5 可移植性

软件被交付之后，除了普通的维护需求之外，很多时候还会遇到客户的其它要求。客户有可能升级了新的硬件设备，而原有软件并不能直接在新的平台环境下运行;客户的业务不断发展，业务量急剧增长，原有系统的性能无法支撑;客户面临很大的竞争压力，需要开展新的业务，而原有软件不具备相应的功能。这些问题最终都依靠开发者来加以解决。下列软件的内部相关质量属性，将直接影响到解决前述这些问题的成本。

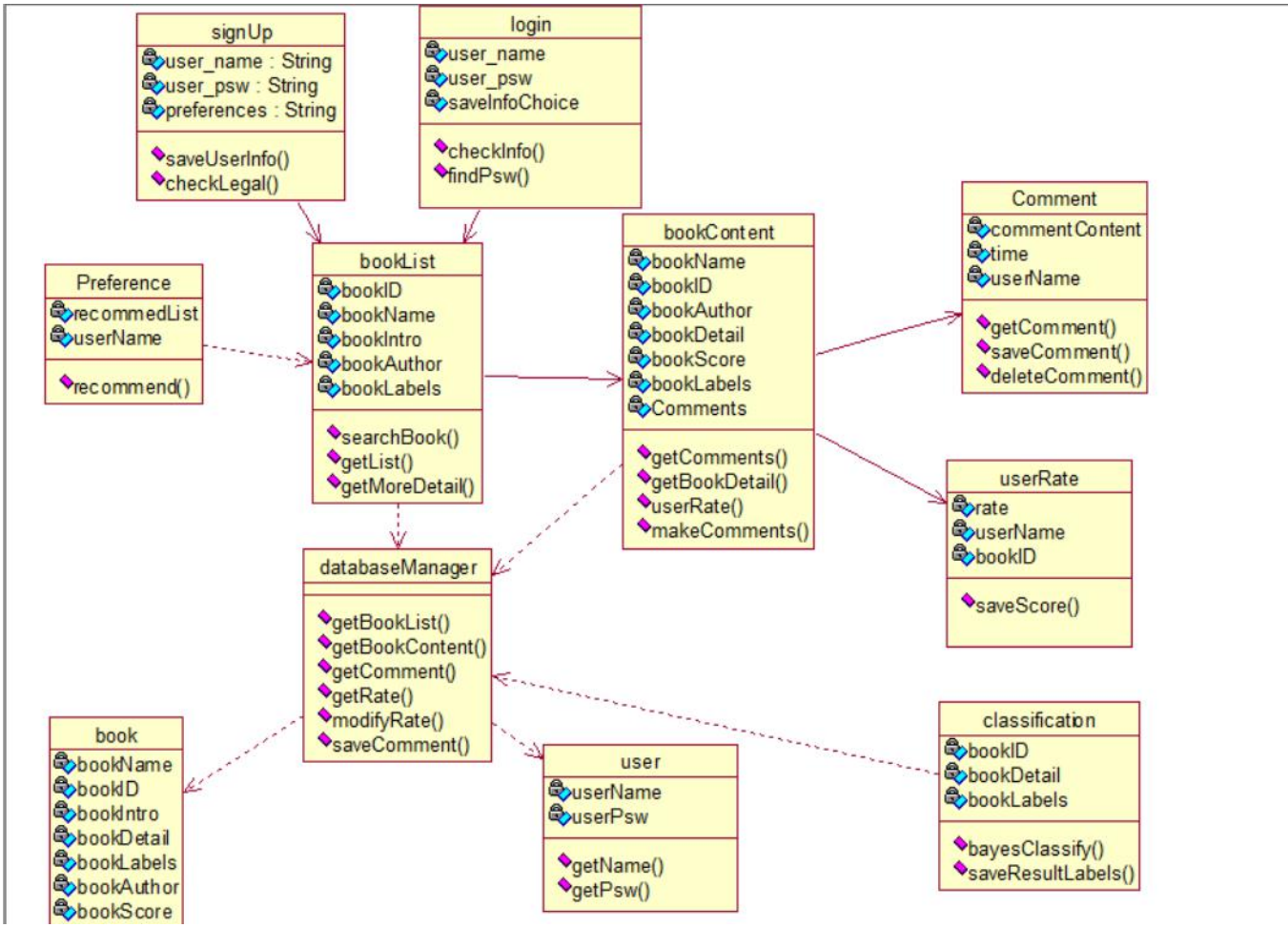
我们的平台支持通过重新编译或改造，从而能够在新的环境下被运行使用，为此所需要付出努力相关的软件属性。例如平台移植、本地化移植等。

3.6 其他需求

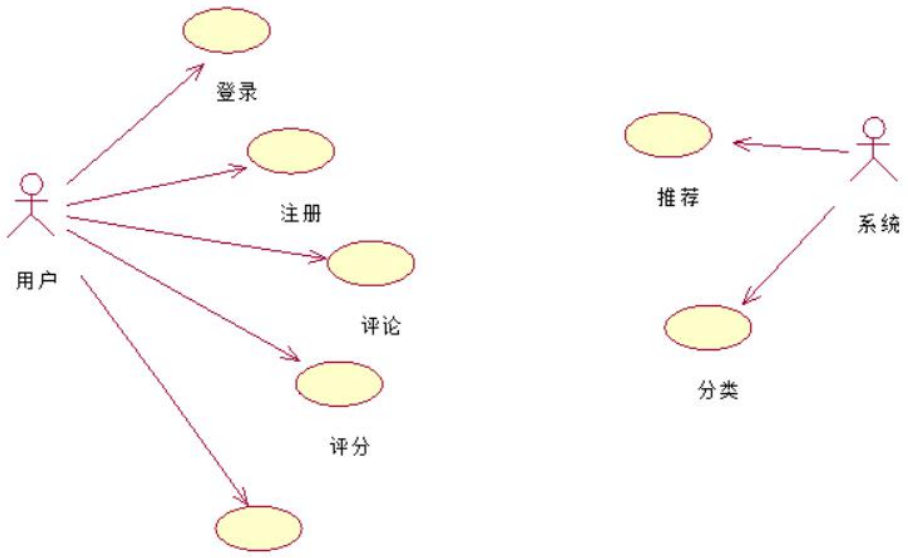
- 我们平台爬取的网站不受版权问题限制。
- 政府政策支持和鼓励网络平台的教育方针政策。
- 国民素质不断提高，能够主观意识到自主学习的益处。

4、附件

5.1 类图

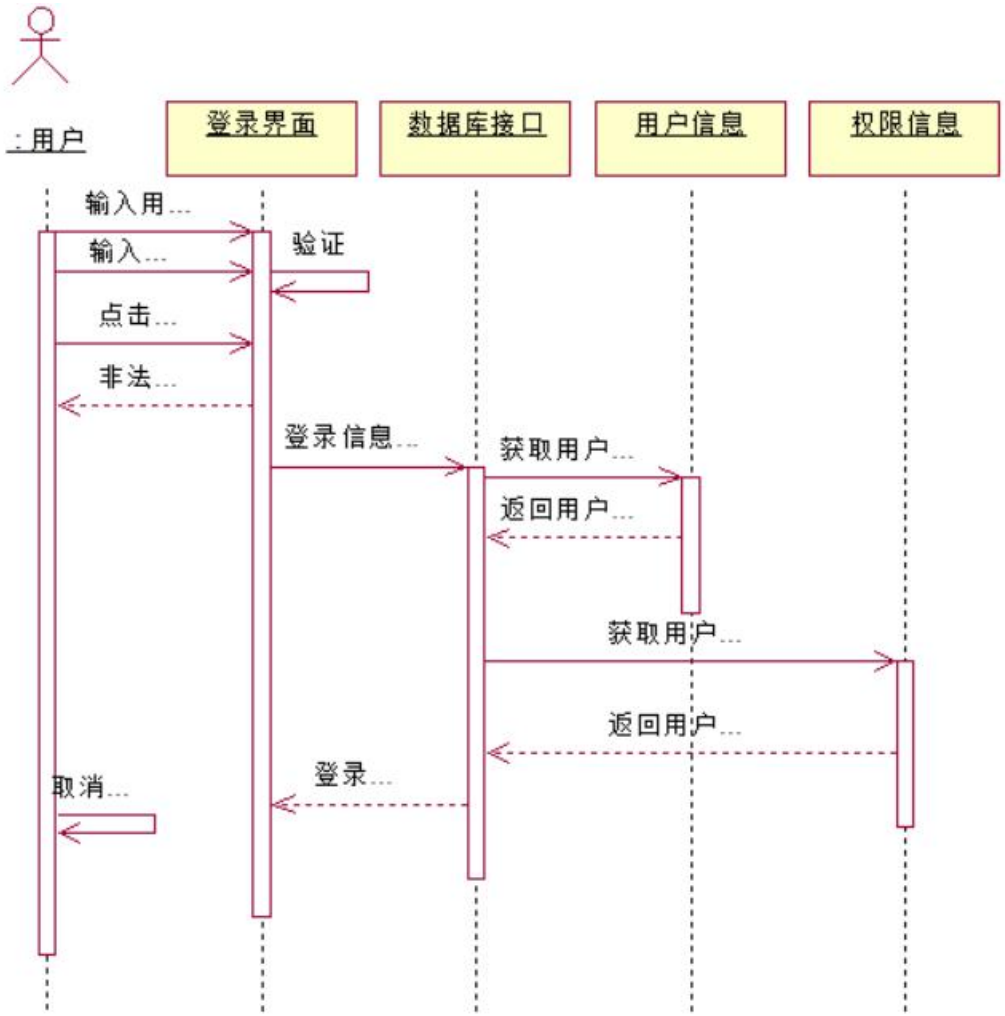


5.2 用例图

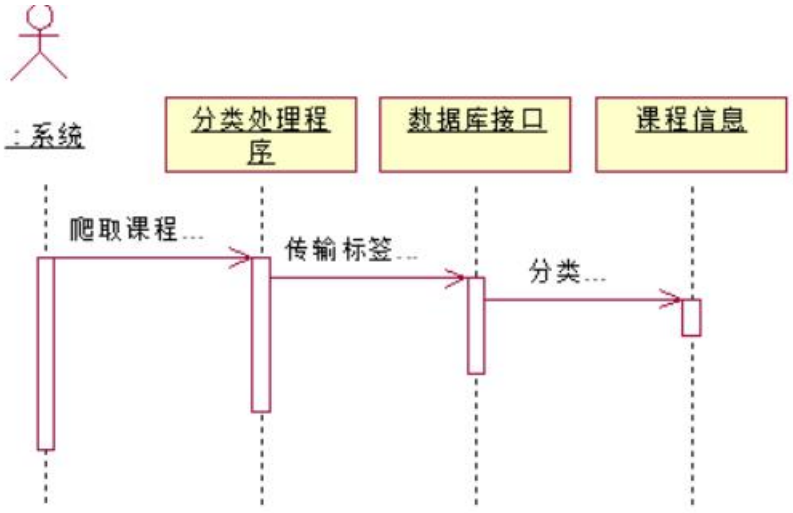


5.3 序列图

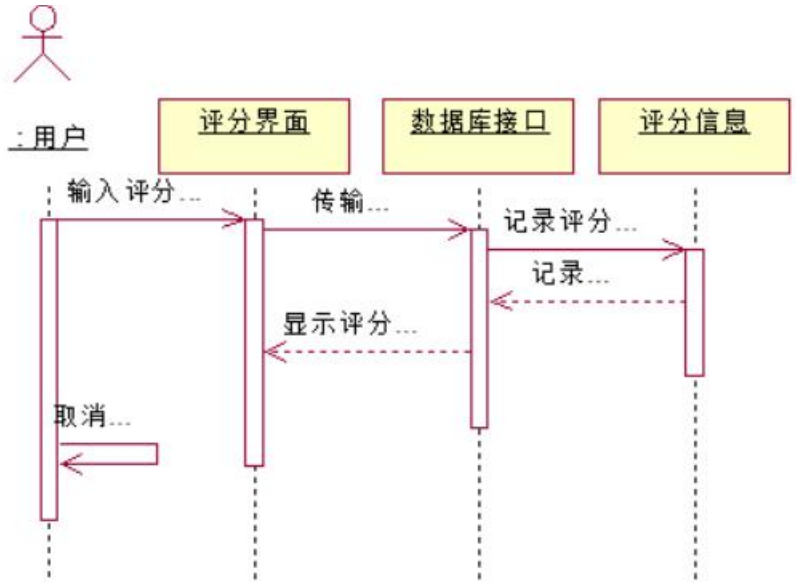
（a）登录


:用户
登录界面
数据库接口
用户信息
权限信息
输入用...
输入...
验证
点击...
非法...
登录信息...
获取用户...
返回用户...
获取用户...
返回用户...
登录...
取消...

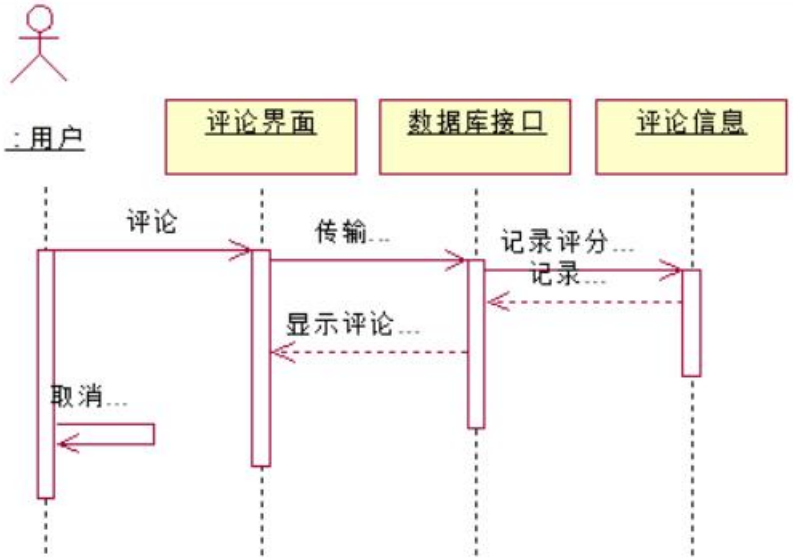
（b）分类


:系统
分类处理程序
数据库接口
课程信息
爬取课程...
传输标签...
分类...

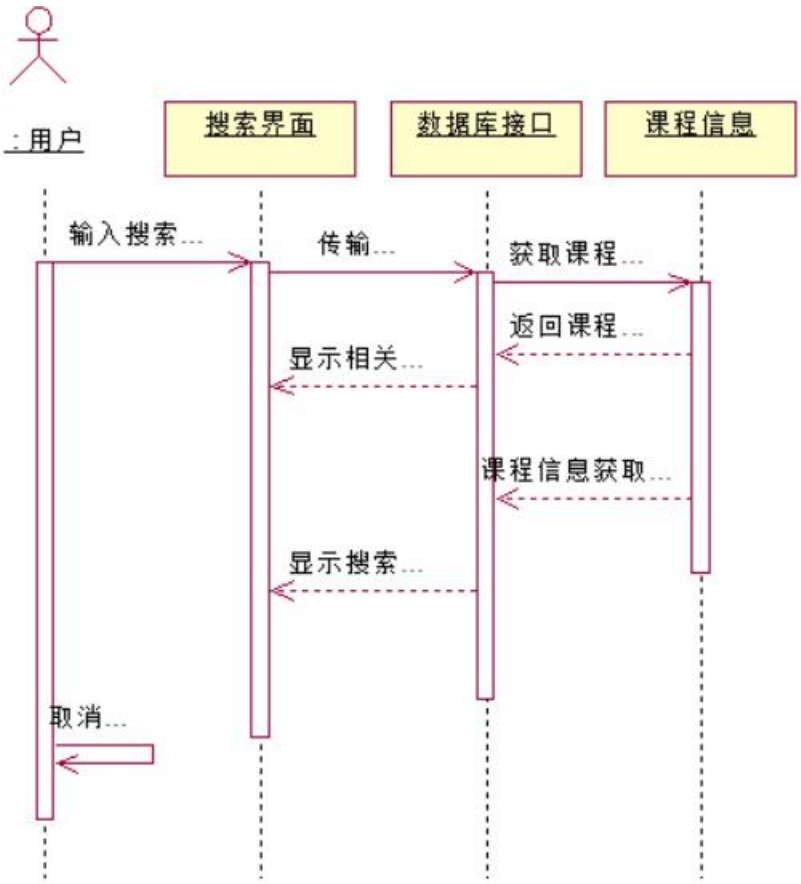
（c）评分


:用户
评分界面
数据库接口
评分信息
输入评分...
传输...
记录评分...
记录...
显示评分...
取消...

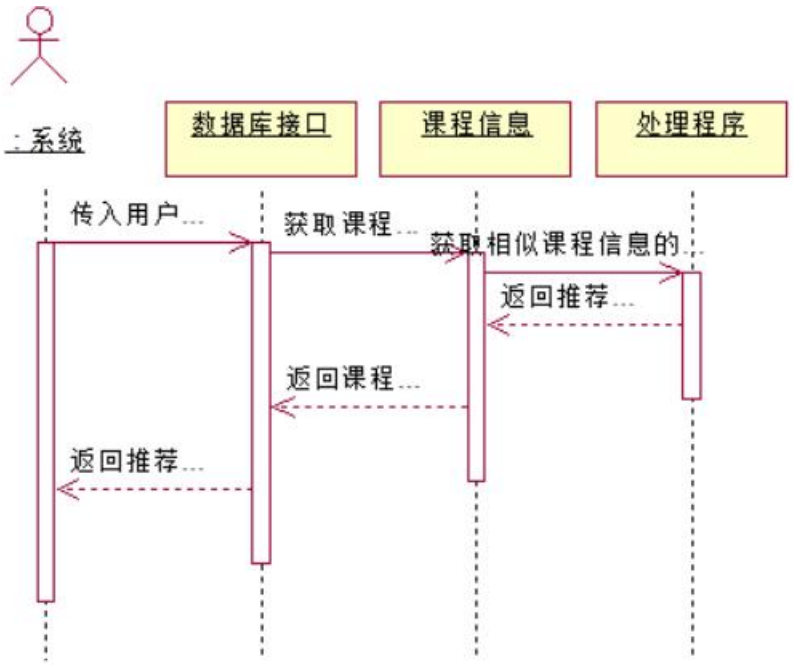
（d）评论


:用户
评论界面
数据库接口
评论信息
评论
传输...
记录评分...
记录...
显示评论...
取消...

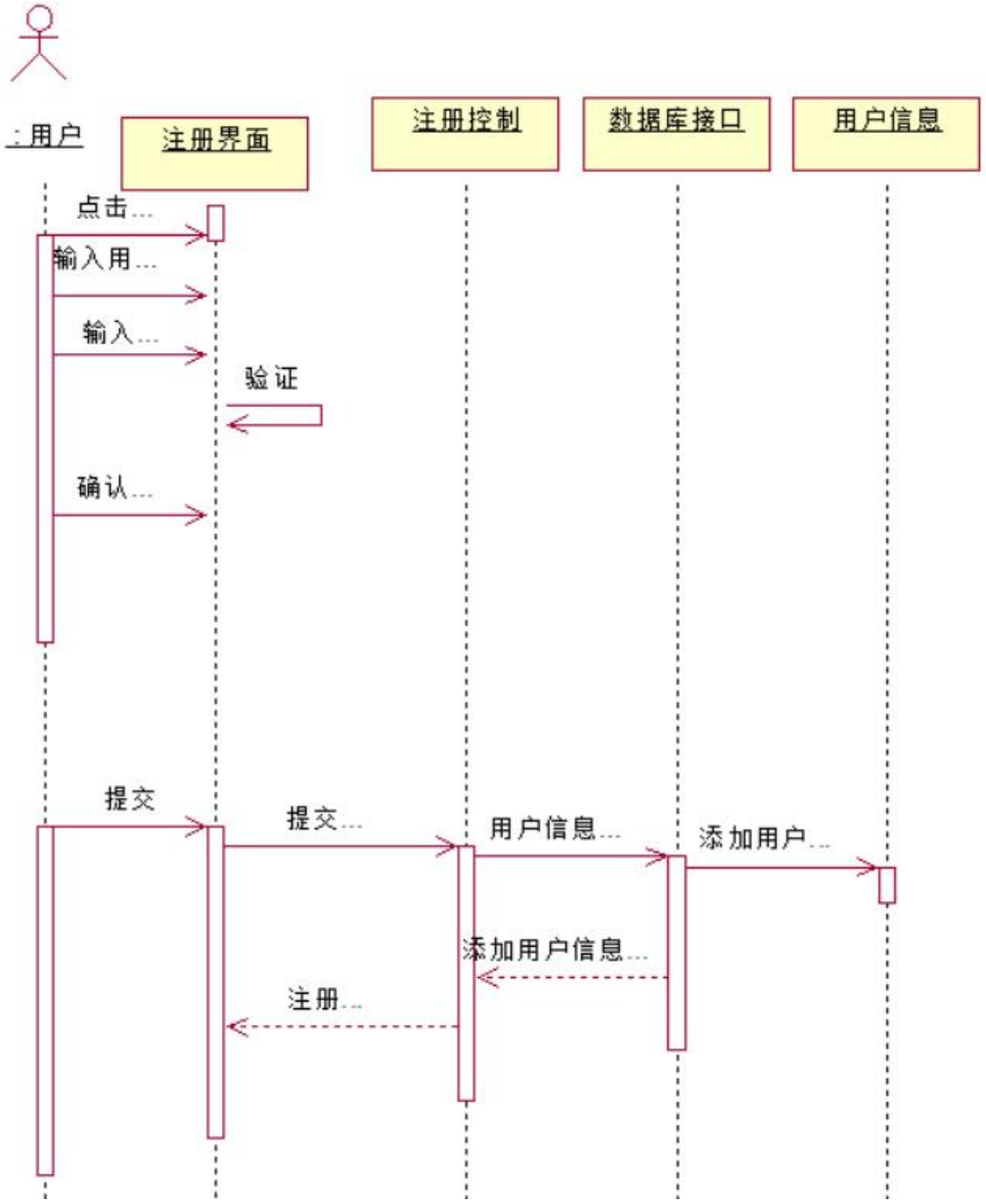
（e）搜索


:用户
搜索界面
数据库接口
课程信息
输入搜索...
传输...
获取课程...
返回课程...
显示相关...
课程信息获取...
显示搜索...
取消...

（f）推荐

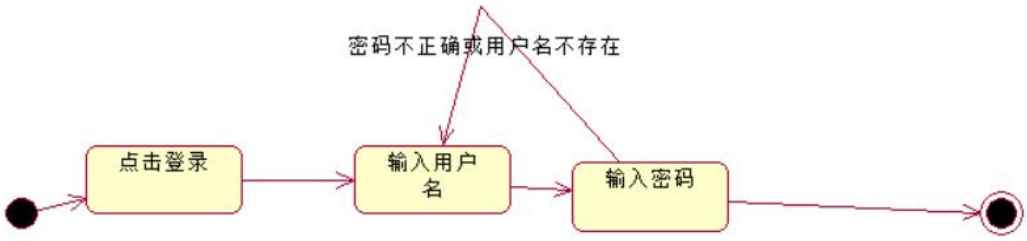

:系统
数据库接口
课程信息
处理程序
传入用户...
获取课程...
获取相似课程信息的...
返回推荐...
返回课程...
返回推荐...

（g）注册



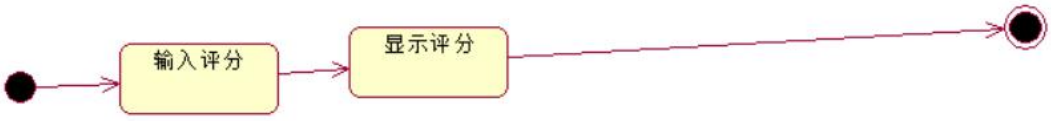
5.4 状态图

(a)登录登出



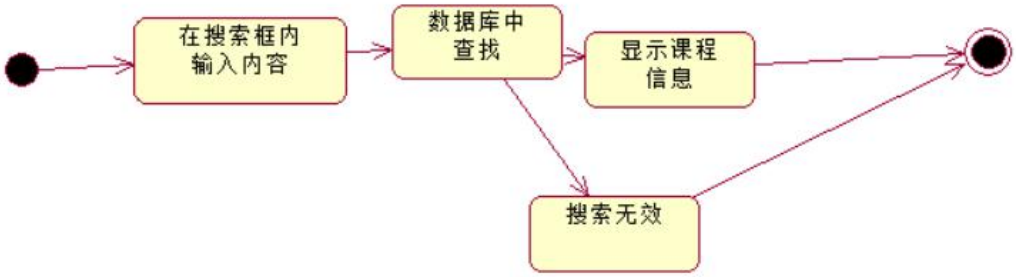
(b)分类

(c)评分



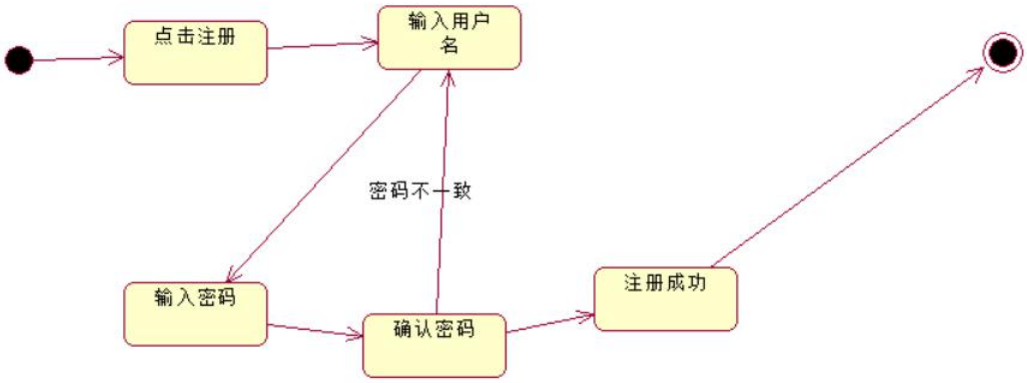
(d)评论

(e)搜索

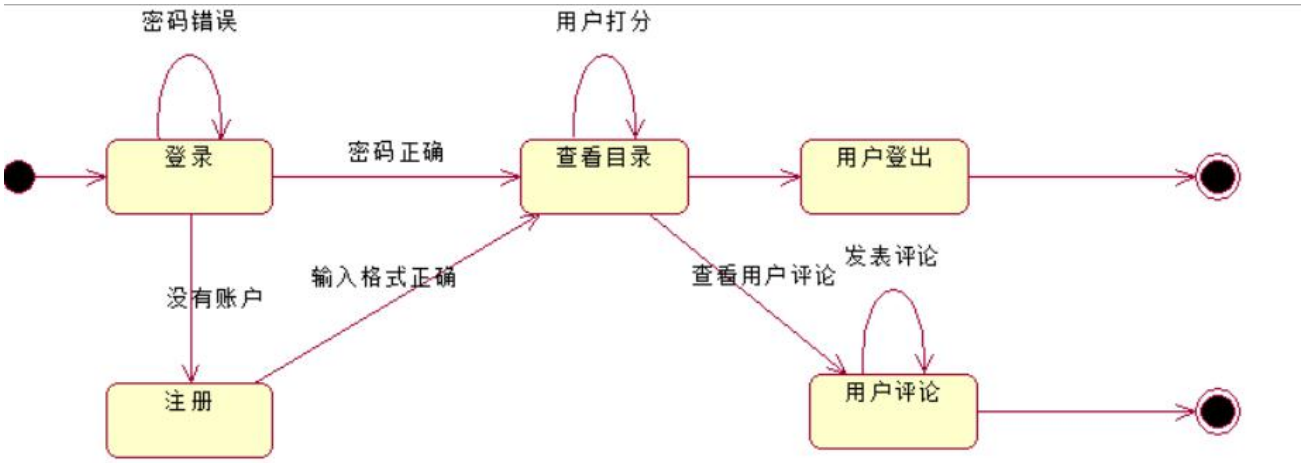


(f)推荐

(g)注册


点击注册
输入用户名
密码不一致
输入密码
确认密码
注册成功

（h）系统状态图


密码错误
用户打分
登录
密码正确
查看目录
用户登出
没有账户
输入格式正确
查看用户评论
发表评论
注册
用户评论