**面向移动互联网的自主学习系统研制及应用示范**

**技术报告**

|  |  |
| --- | --- |
| **项目名称：** | **面向移动互联网的自主学习系统研制及应用示范** |
| **项目编号：** | **14511109300** |
| **依托单位：** | **华东师范大学出版社有限公司** |
| **负责人：** | **吴永和** |
| **起止日期：** | **2014年6月30日~2016年6月30日** |

目录

[一、 需求分析及设计 1](#_Toc467056398)

[1.1 项目背景 1](#_Toc467056399)

[1.1.1 中小学学生个性化自主学习的需求及问题 1](#_Toc467056400)

[1.1.2 互联网+教育的大势 2](#_Toc467056401)

[1.1.3 互联网+教育的特征 3](#_Toc467056402)

[1.2 建设目标 6](#_Toc467056403)

[1.3 总体架构设计 6](#_Toc467056404)

[1.4 数据库逻辑结构设计 9](#_Toc467056405)

[1.5 系统部署拓扑图 9](#_Toc467056406)

[二、 系统实现 11](#_Toc467056407)

[2.1 教育资源管理系统的设计和实现 11](#_Toc467056408)

[2.1.1 功能概述 12](#_Toc467056409)

[2.1.2 个人信息管理 13](#_Toc467056410)

[2.1.3 机构管理 14](#_Toc467056411)

[2.1.4 用户管理 15](#_Toc467056412)

[2.1.5 系统配置管理 16](#_Toc467056413)

[2.1.6 图书资源管理 17](#_Toc467056414)

[2.1.7 二维码管理 19](#_Toc467056415)

[2.1.8 教育资源管理 20](#_Toc467056416)

[2.2 移动客户端系统设计和实现 21](#_Toc467056417)

[2.2.1 功能概述 22](#_Toc467056418)

[2.2.2 功能清单 23](#_Toc467056419)

[2.2.3 系统界面 25](#_Toc467056420)

[三、 关键技术 29](#_Toc467056421)

[3.1 教辅视频学习资源的推荐 29](#_Toc467056422)

[3.1.1 概述 29](#_Toc467056423)

[3.1.2 视频学习资源用户行为数据 30](#_Toc467056424)

[3.1.3 推荐模型的设计 31](#_Toc467056425)

[3.2 智慧型学习资源模型 35](#_Toc467056426)

[3.2.1 总体思想 36](#_Toc467056427)

[3.2.2 智能资源设计与聚合 37](#_Toc467056428)

[3.2.3 基于Tin Can API增强学习资源数据采集能力 39](#_Toc467056429)

[3.3 基于浏览器的实时沉浸式辅导技术 42](#_Toc467056430)

[3.3.1 关键技术 42](#_Toc467056431)

[3.3.2 实现案例 45](#_Toc467056432)

[3.4 用户行为数据采集、表示和存储 46](#_Toc467056433)

[3.4.1 用户行为的采集 47](#_Toc467056434)

[3.4.2 面向用户行为分析的元数据管理技术 47](#_Toc467056435)

[3.4.3 用户行为建模技术 48](#_Toc467056436)

[3.4.4 用户行为特征模版 49](#_Toc467056437)

[3.4.5 用户行为模型生成引擎 49](#_Toc467056438)

[3.5 用户行为特征模式抽取 51](#_Toc467056439)

[3.5.1 用户的兴趣模版 52](#_Toc467056440)

[3.5.2 用户模型生成引擎 54](#_Toc467056441)

[3.6 用户流失分析 55](#_Toc467056442)

[3.6.1 面向教育资源的冷热度预测算法 55](#_Toc467056443)

[3.6.2 基于用户细分的流失率预测算法 56](#_Toc467056444)

[四、 技术创新点 57](#_Toc467056445)

[4.1 多维度数据构建模型大幅提升分析系统的效率 57](#_Toc467056446)

[4.2 可定制的业务分析模型提升系统交互能力和实用性 58](#_Toc467056447)

[4.3 采用多算法协同方法提升系统业务分析的全面性和准确性 59](#_Toc467056448)

[五、 技术展望 59](#_Toc467056449)

1. 需求分析及设计

项目背景

### 中小学学生个性化自主学习的需求及问题

从炙手可热的互联网在线教育回归到细分的中小学生在线教育，我们看到了普遍存在的需求和亟待解决的问题：

1. **中小学学生个性化学习的强烈需求**

课堂教学的系统性、互动性具有无法取代的重要价值，但一个老师对应多个学生的客观现实也导致了学生很难在课堂上实现个性化学习，每个学习者的学习需求、活动方式、学习进度都可能不一样，因此，应主要依靠信息技术手段而不是教师主观判断来掌握和分析学习者的行为动态，从而提供自适应个性化的环境。课后的家庭学习时间成为了个性化学习的一个主要环境，围绕家庭作业和学习教辅学生要完成一天课程的巩固，在课程上每个学生掌握知识的情况都不相同，作业过程遇到的问题也非常具有个性化特征，以个性化问题为导向的一种自主学习方式是中小学课后个性化学习的一个基础需求。

1. **中小学学生自主学习的强烈需求**

目前中小学学生的课堂学习以统一的、填鸭式的方式为主，课后的作业是强迫式的、必修式的，造成了学生的被动接收，无法激发自我学习兴趣。但老师、家长却又都希望尽早培养孩子的自主学习能力，这就形成了一种现实的矛盾对立。其实中小学生的自主学习可以通过探究问题的环境营造，在学生课后习题过程遇到困难的时候给予及时的有针对性的讲解，让学生有一种自主发现的乐趣和成就，这类需求的满足同时还可以减轻家长辅导的压力和学生对家长辅导的一种依赖。

1. **目前数字出版对于精准学习资源服务的支持能力不够**

**难以做到精准出版。**目前的数字出版及相关的教育服务还没有突破“单向”出版的模式。基本是以传统出版，在线课程，在线题库等资源形式存在，并没有实现随时、按需、双向的教育精准服务。这种出版模式相比新型移动互联网在线教育服务反应慢、调整慢、几乎不可能获取用户行为偏好，从而制约了个性化学习资源的按需生成和发布。

**难以做到对学生的精准资源服务**。由于没有能够获得学习的学习数据，无法进行学习分析，因此难以获得学生个人“画像”，从而无法根据学生特征为其推荐适合学习者特征的的相关的学习资源。

### 互联网+教育的大势

2015年7月4日，国务院印发《[国务院关于积极推进“互联网+”行动的指导意见](http://baike.baidu.com/item/%E5%9B%BD%E5%8A%A1%E9%99%A2%E5%85%B3%E4%BA%8E%E7%A7%AF%E6%9E%81%E6%8E%A8%E8%BF%9B%E2%80%9C%E4%BA%92%E8%81%94%E7%BD%91%2B%E2%80%9D%E8%A1%8C%E5%8A%A8%E7%9A%84%E6%8C%87%E5%AF%BC%E6%84%8F%E8%A7%81)》指出，“互联网+”是把互联网的创新成果与经济社会各领域深度融合，推动技术进步、效率提升和组织变革，提升实体经济创新力和生产力，形成更广泛的以互联网为基础设施和创新要素的经济社会发展新形态。在全球新一轮科技革命和产业变革中，互联网与各领域的融合发展具有广阔前景和无限潜力，已成为不可阻挡的时代潮流，正对各国经济社会发展产生着战略性和全局性的影响。

意见提出探索新型教育服务供给方式。鼓励互联网企业与社会教育机构根据市场需求开发数字教育资源，提供网络化教育服务。鼓励学校利用数字教育资源及教育服务平台， 逐步探索网络化教育新模式，扩大优质教育资源覆盖面，促进教育公平。鼓励学校通过与互联网企业合作等方式，对接线上线下教育资源，探索基础教育、职业教育 等教育公共服务提供新方式。推动开展学历教育在线课程资源共享，推广大规模在线开放课程等网络学习模式，探索建立网络学习学分认定与学分转换等制度，加快 推动高等教育服务模式变革。

“互联网+教育”对教育改革与创新提出了新的挑战和基于。互联网+教育是可以将微课、慕课、翻转课堂、移动学习、泛在学习、学习者为中心的学习等各种不同纬度的学习概念整合的发展趋势。“互联网+教育”对于解决上述需求问题具有良好的实践性。

### 互联网+教育的特征

“互联网+教育”具有很多特征，就本项目而言，它具有突破地域局限，突破资源服务的困境，具有长尾效应，可实现学习服务和学习资源的精准投放的特征。

1. **突破地域局限**

互联网应用技术的快速发展，智能移动设备的功能快速更迭以及价格日益下降到大规模可接受水平，高速移动互联网的大规模覆盖，随时随地的接入网络进行在线学习成为可能，这些因素确保无论学生处在和种情况，都可能接入到在线学习环境中。因此突破学校场所和家庭场所的学习已经随处可见，如碎片化学习就是典型案例，学生在上学途中可在地铁、公交上开展基于微课的学习。

1. **突破资源服务的困境**

学习资源的首要存在价值就是需要能达到学习者学习的目标。“互联网+教育”以无处不在的网络、学习终端、以及在线学习环境作为载体，能实现各种学习资源快速流通。

1. **学习资源的精准投放**

学习者对学习资源的需求各有不同，而这种差异需要精准的服务来满足更高的要求。学习资源是支持学习者学习的资源，不同种类的资源因其独特性而能够较好地契合不同类型学习者的需求（周素萍，2012），度量学习资源必须要考虑学习者的能力和需求。学习者的能力、偏好和现有水平会影响其对学习资源的需求和选择。通过对真实学习者的调查，发现在学习过程中，那些具有相似学习偏好及学习水平的学生通常都对学习资源具有相似的需求（王志梅等，2006）。学习者的兴趣和动机也会影响其对网络资源的选择和使用（王广新等，2006）。另一方面，学习者的能力是动态演变的，当新的知识建立在已有知识基础上，学习就有了收益（Hattie, J.,2014）。学习资源的内容和组织形式等会影响学习者的学习兴趣、学习结果，进而改变学习者的能力水平；使用社会网络资源进行非正式学习已经影响了在职的和受教育人（F. J. García-Peñalvo,2012）。同时，网络学习资源的内在特性是促进深度学习（C.Cechinel,2011），而学习者的个性化学习、自适应学习和无缝学习已成为资源发展的要点，进而对资源的服务也要更加精准，满足不同学习者的学习需求，会在很大程度上促进教育社会化和终身化的进程（陈时见等，2003）。

在互联网+教育形态中，信息技术与教育学习深度融合，让我们能够对学习行为进行持续的跟踪和记录，从而形成庞大的实时的学习者状态数据集。新技术的深入应用带来了教育“大数据”的高速增长，挖掘这些教育数据潜在价值的迫切需求，使得学习分析应运而生（吴永和，2013）。地平线报告于2011—2014连续四年将学习分析列为教育发展的关键技术，于2016年将学习分析与自适应学习相结合，再次推动其发展，说明了学习分析技术在教育中的重要性。

随着科技的发展，移动学习平台、学习管理系统等广泛应用。移动学习鼓励学习者进行自主探究和协作学习，为学生积极贡献、分享、反馈个人学习经验提供平台（Boticki，2015），探究式学习最大限度地提高了学习者的智力参与(Savery, 2006)，学习从以教师为中心转向以学生为中心(Moyle, 2008)。在这些软件系统中存储着海量的学习者信息及学习过程数据（魏顺平，2013）。学习分析技术是测量、收集、分析和报告有关学生的学习行为以及学习环境的数据，用以理解和优化学习及其产生的环境的技术（顾小清，2012）。运用学习分析技术可以从量化的角度分析学习者的个性化行为，进而评估学习者的学习状态，预测学习结果（武法提，2016）；可以更好地理解个性化学习过程，进而较为精准地了解学习者的风格偏好，定位学习者的能力水平（魏雪峰，2013）。

建设目标

依托华东师范大学出版社在中小学教育辅导资源及品牌方面的优势，建立面向移动互联网的个性化自主学习系统，实现百万级的中小学生应用示范，并积极探索传统纸质图书与移动互联网结合的O2O商业模式。同时在大数据和学习分析技术支撑下，研究用户阅读行为分析，个性化推送等技术。形成“精品”数字出版和“精准”教育服务深度结合的移动学习新模式。促进教育信息化成果的普及，提高数字教育辅助出版精度、速度和适应性，逐步构建新型数字化出版与学习服务高度融合、相互促进的生态系统。

总体架构设计

本项目研究主要研究内容是研制一个“精品”数字出版和“精准”教育服务深度结合的个性化自主学习系统，其中包括：研制一套个性化自主学习系统，出版一批优质中小学教育辅助材料，开发一个简单好用的智能交互终端，建立试题讲解资源库等，具体分解如图 1研究内容架构图：

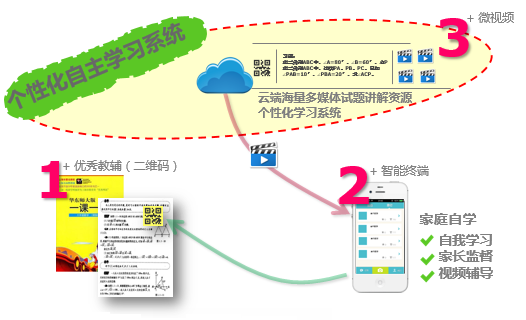


图 1研究内容架构图

为了实现个性化自主学习系统的云端数据承载，用户端的友好交互，我们将系统设计划分为三个层次：应用、业务、数据三个层次，设计架构如图 2 系统设计架构图：



图 2 系统设计架构图

系统的各个子系统及模块具备独立的可扩展性，并可根据需要进行模块化组装或定制裁剪。

**数据层**

按照资源被访问的开放性，分为云存储和私有存储。作为被用户频繁访问的微视频、广告资源，放在云端存储，可提高访问效率和减小业务层相关业务的访问压力；元数据、用户数据、系统数据、日志等内部使用的数据则存储在私有存储设备中。

**业务层**

按业务类型分为微视频服务平台、广告系统、分析系统3个系统，各自提供独立的服务。微视频服务平台，主要管理二维码书、微视频资源以及二维码等相互关系，作为用户使用的核心系统之一；同时提供对上述资源的维护管理服务，提供相关资源的新建、更新以及审批。

广告系统，主要管理广告内容、广告策略以及更新、访问服务等，作为一个独立系统存在。

分析系统根据另外两个系统的日志以及基础数据，对用户、微视频资源、知识点3个对象进行分析，并提供数据分析的呈现服务。

**应用层**

应用层主要包含两类：外部用户使用的APP和内部用户使用的管理工具。

外部用户使用的APP完成，用户通过二维码最终访问微视频以及相关的其他功能。也可以通过微信等其他第三方扫码工具进行扫描并完成知识点视频的关联阅读。

内部用户使用的管理工具主要负责，微视频数据更新、广告数据更新、分析的最终呈现。

数据库逻辑结构设计

面向移动互联网的个性化自主学习系统分为前台移动客户端和后台资源管理子系统，两个子系统的数据逻辑设计如下两图，分别包括了用户表，资源表，知识点，图书等数据视图和关系表，设计架构如图 3数据逻辑设计所示：

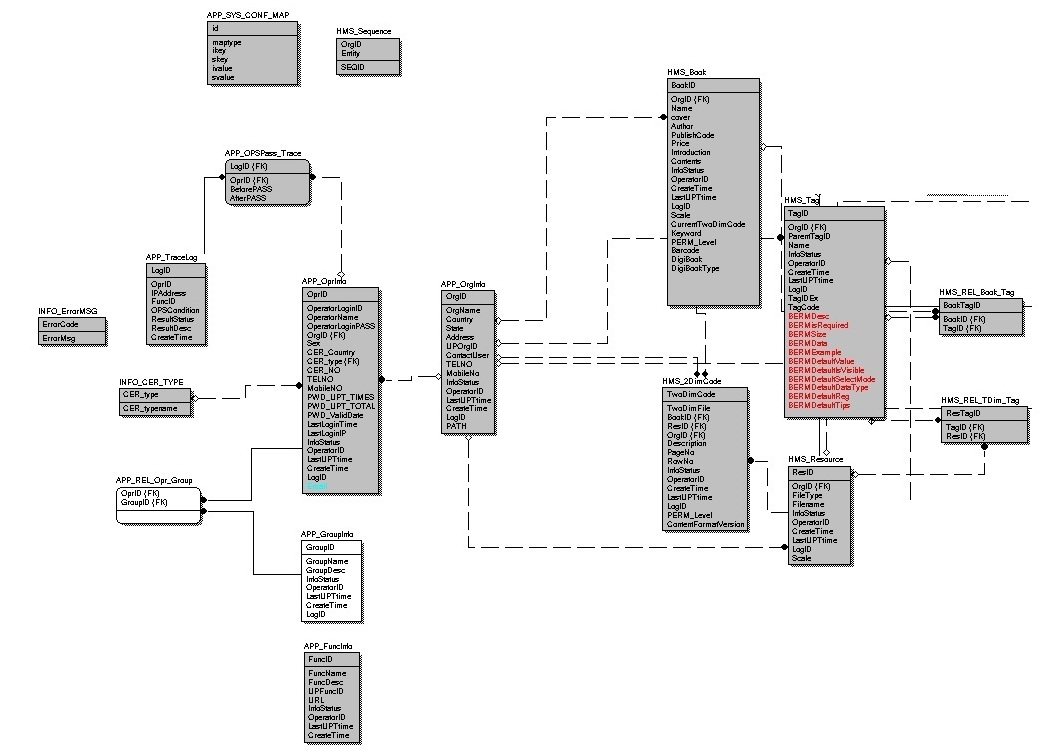


图 3数据逻辑设计

系统部署拓扑图

本系统总体部署拓扑图如图 4 系统部署拓扑图所示：



图 4 系统部署拓扑图

如上图所示，系统部署主要分为：

租用的公有云存储，主要存储终端用户频繁访问的微视频资源和广告资源；

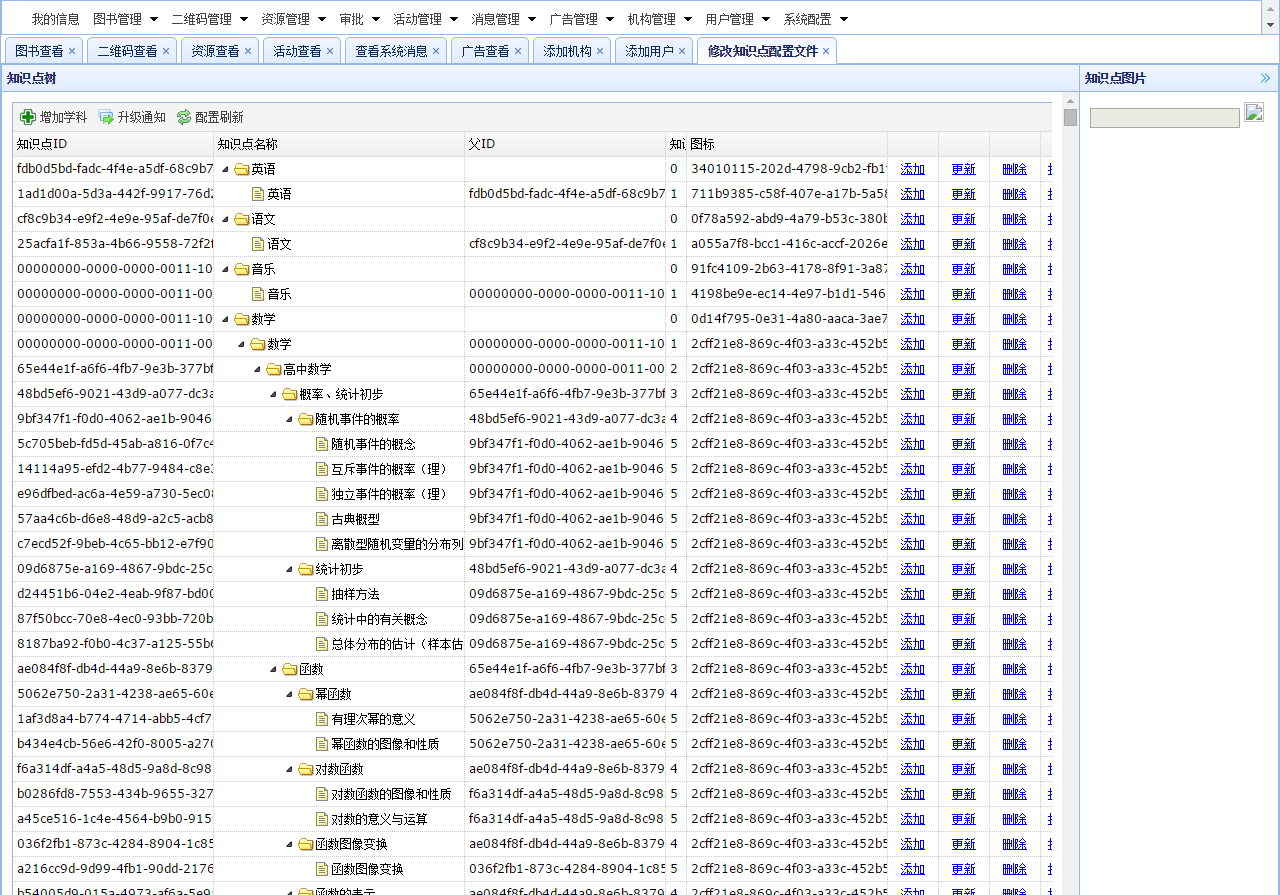
服务平台及私有存储，这个部分可以考虑租用云服务或自建，这个部分的服务都部署在同一局域网内。包括微视频服务平台、广告系统、分析系统以及私有存储。

服务平台采用Linux系统，J2EE构架，提供WebService；管理工具部分由部署在微视频服务平台服务器上的tomcat提供网页服务。

1. 系统实现

教育资源管理系统的设计和实现

教育资源采集系统是指，资源提供方对图书或二维码资源的管理。资源提供方可以通过本采集系统直接上传教育资源，这些资源将最终供二维码客户端用户搜索点播。样例界面如图 5 教育资源管理系统总界面：



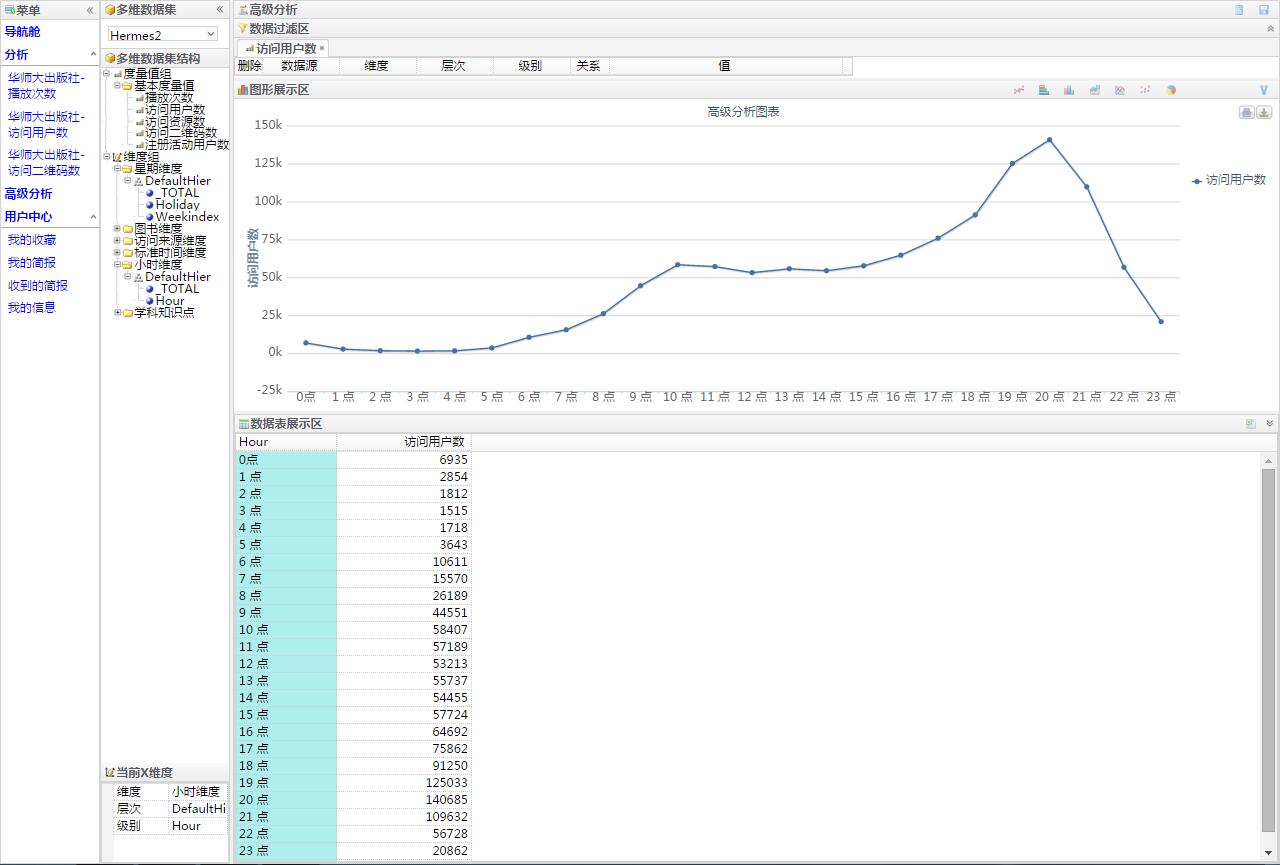


图 5 教育资源管理系统总界面

### 功能概述

**（1）资源提供用户管理**

* 机构管理：新增、删除、修改出版机构。

满足不同资源拥有者均可以上传、发布自己的教育资源。

* 用户管理：新增、删除、修改出版机构用户。

操作由拥有管理员权限的员进行，满足平台管理、数据运营的分离。

**（2）二维码资源管理**

图书或二维码资源所有者可对图书或二维码资源进行管理。这些资源将最终被会员用户在各个子系统上浏览或使用。系统可以完成以下功能：

* 配置文件管理：新建、删除、修改模板

为方便资源数据管理，可建立不同类型数据的数据结构模板，在建立数据时即保证了数据的规范，又减少了录入的工作量。

* 图书管理：新建、删除、修改图书

所有录入的图书，必须先建立书籍资源。书籍资源最终被体现在用户终端搜索结果、二维码与教育资源的关联绑定等。

* 教育资源管理：新建、删除、修改教育资源

以图书为分类建立教育资源库，上传资源均与二维码对应。

* 二维码管理：二维码资源绑定、对应资源更新、批量导出二维码列表

最终将图书、资源通过二维码绑定，使得用户拍摄某个二维码即可获得该图书对应的教育资源；同时为了适应图书不重印对应资源可更新的情况，提供对应资源更新。

### 个人信息管理

后台管理员成功登录后，可以自行修改个人信息与密码。如表格 1 个人信息管理表和图 6 用户信息管理所示：

表格 1 个人信息管理表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **功能** | **输入内容** | **输出内容** | **结果** | **备注** |
| 用户信息修改 | 个人信息 | 个人信息 | 更新个人信息 |  |
| 用户密码修改 | 用户名，密码,旧密码 | 更新成功 | 修改个人密码 |  |

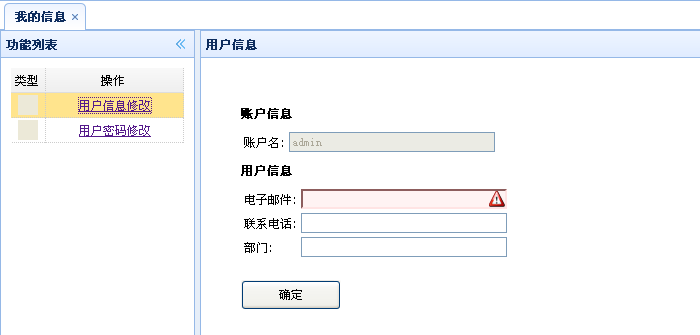


图 6 用户信息管理

### 机构管理

对机构进行添加、删除、修改的管理。以分类管理用户，如表格 2 机构管理表。

表格 2 机构管理表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **功能** | **输入内容** | **输出内容** | **结果** | **备注** |
| 添加机构 | 机构名称 | 新建成功 | 机构信息更新 |  |
| 删除机构 | 机构选择 | 删除成功 | 机构信息更新 |  |
| 修改机构 | 机构名称 | 更新成功 | 机构信息更新 |  |

### 用户管理

* 对用户的所属机构、密码、权限等个人信息、进行管理。
* 系统操作用户必须归属于机构。
* 自己不能操作自己的任意信息；只能在个人信息管理中操作。
* 系统功能组合在一起形成权限组，再将权限组赋予用户，使得用户具备了系统功能操作的权限。具备了相关权限，菜单中会出现相应的菜单项。

表格 3 用户管理表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **功能** | **输入内容** | **输出内容** | **结果** | **备注** |
| [添加用户](javascript:void(0)) | 用户信息 | 用户信息 | [添加用户](javascript:void(0)) |  |
| [更新用户信息](javascript:void(0)) | 用户信息 | 更新成功 | 用户信息更新 |  |
| [删除用户](javascript:void(0)) | 选择用户 | 删除成功 | 用户列表更新 |  |
| [重置用户密码](javascript:void(0)) | 选择用户 | 重置密码成功 | 修改用户密码 |  |
| [更新用户权限](javascript:void(0)) | 权限信息 | 更新成功 | 权限更新 |  |
| [添加权限组](javascript:void(0)) | 权限组信息 | 权限组信息 | 权限组更新 |  |
| [修改权限组](javascript:void(0)) | 权限组信息 | 更新成功 | 权限组更新 |  |
| [删除权限组](javascript:void(0)) | 权限组信息 | 删除成功 | 权限组更新 |  |

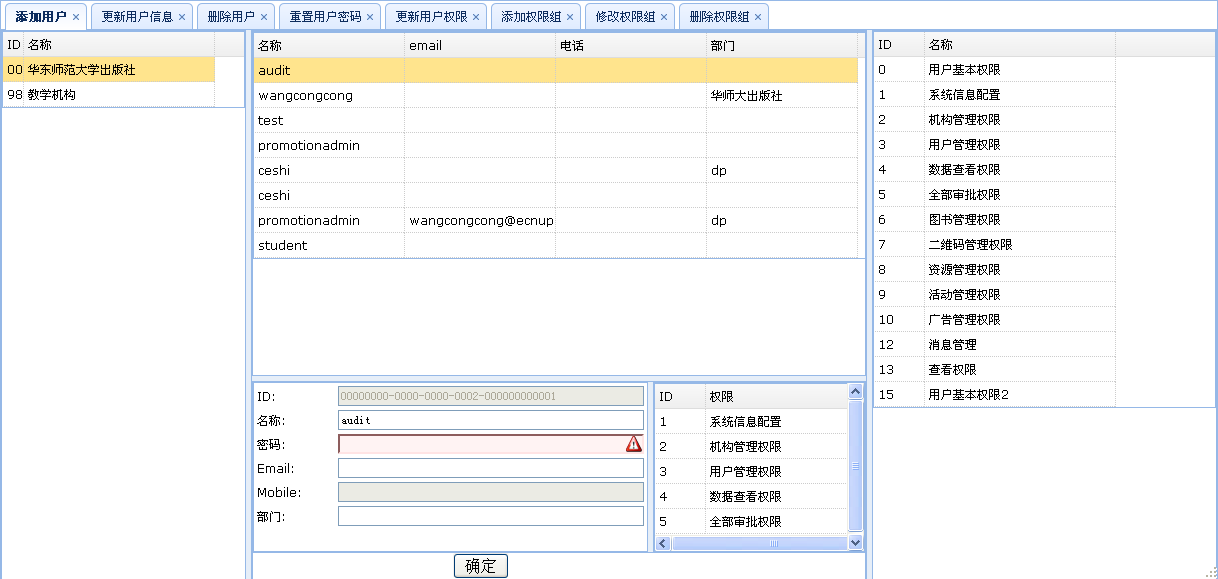


图 7 用户管理

### 系统配置管理

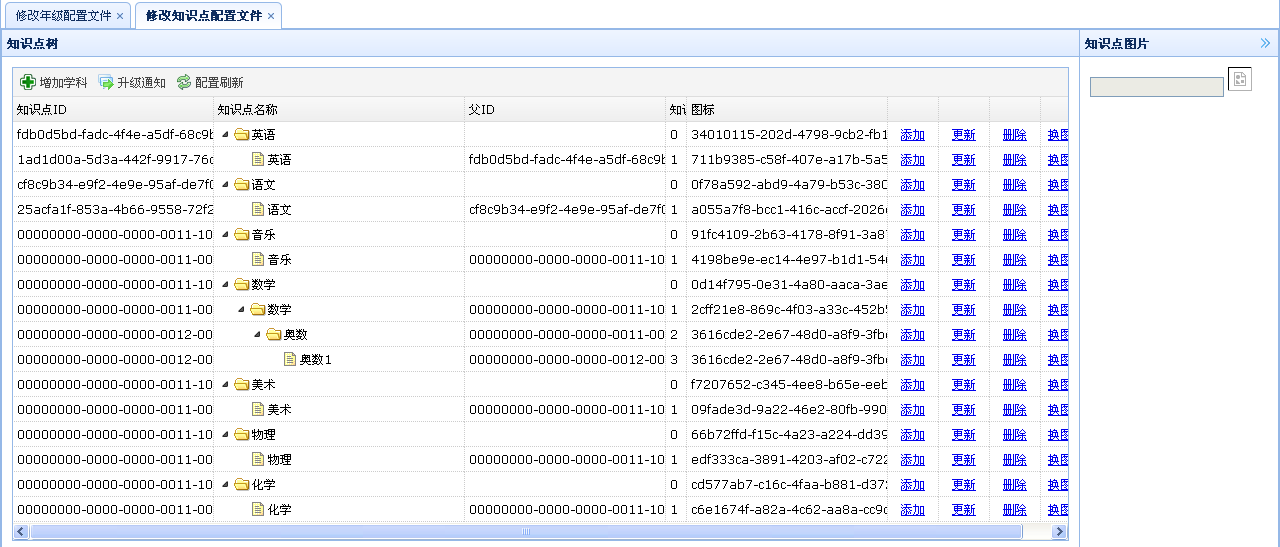
* 对年级配置和知识点进行管理。以在新建资源时配对相应信息。如表格 4 配置管理。
* 年级配置文件，包括ID，名称，阶段ID，阶段名称四项，其中：ID不可修改，如图 8 年纪配置管理。
* 知识点配置文件，包括ID，名称，父ID，深度、图片ID五项，其中：ID、父ID、深度、图片ID不可直接修改；名称可以修改。

表格 4 配置管理

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **功能** | **输入内容** | **输出内容** | **结果** | **备注** |
| [修改年级配置文件](javascript:void(0)) | 年级信息 | 修改成功 | 年级列表更新 |  |
| [修改知识点配置文件](javascript:void(0)) | 知识点信息 | 修改成功 | 知识点列表更新 |  |



图 8 年纪配置管理

图 9 知识点管理

### 图书资源管理

* 所有录入的图书，必须先建立书籍资源。
* 上传的二维码资源与图书的对应，使得用户可以检索到相关二维码资源。
* 图书信息项中，必填的数据为：书名：显示在app中。ISBN：在本系统中此ISBN必须是唯一的，且只用真实ISBN的后四位数字（必须唯一）。搜索关键词：第一个字符必须是英文逗号“,”；多个关键词，用英文逗号”,”分隔。年级：可多选。封面：可不上传图片。

表格 5 图书管理

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **功能** | **输入内容** | **输出内容** | **结果** | **备注** |
| [图书查看](javascript:void(0)) | 图书选择 | 图书信息 |  |  |
| [创建图书](javascript:void(0)) | 图书信息 | 新建成功 | 图书信息更新 |  |
| [修改图书信息](javascript:void(0)) | 图书信息 | 修改成功 | 图书信息更新 |  |
| [删除图书](javascript:void(0)) | 选择图书 | 删除成功 | 图书列表更新 |  |

### 二维码管理

* 绑定二维码与二维码资源文件；
* 对指定的图书进行电子资源与二维码对应关系编辑。将自动生成编号与二维码，电子资源对应关系；
* 导出二维码文件，系统会形成完成的二维码编号对应电子资源对应关系表；二维码图片包；
* 利用导出结果可以进行二维图书的排版，校验工作；

表格 6 二维码管理

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **功能** | **输入内容** | **输出内容** | **结果** | **备注** |
| [二维码查看](javascript:void(0)) | 二维码选择 | 二维码信息 |  |  |
| [创建二维码](javascript:void(0)) | 资源编码 | 新建成功 | 对应二维码生成 |  |
| [修改二维码](javascript:void(0)) | 二维码信息 | 更细成功 | 二维码信息更新 |  |
| [删除二维码](javascript:void(0)) | 选择二维码 | 删除成功 | 二维码列表更新 |  |
| [批量导入二维码](javascript:void(0)) | 图书信息表格 | 二维码列表 | 对应二维码生成 |  |
| [批量导出二维码](javascript:void(0)) | 图书选择 | 二维码打包图片 | 二维码导出 |  |

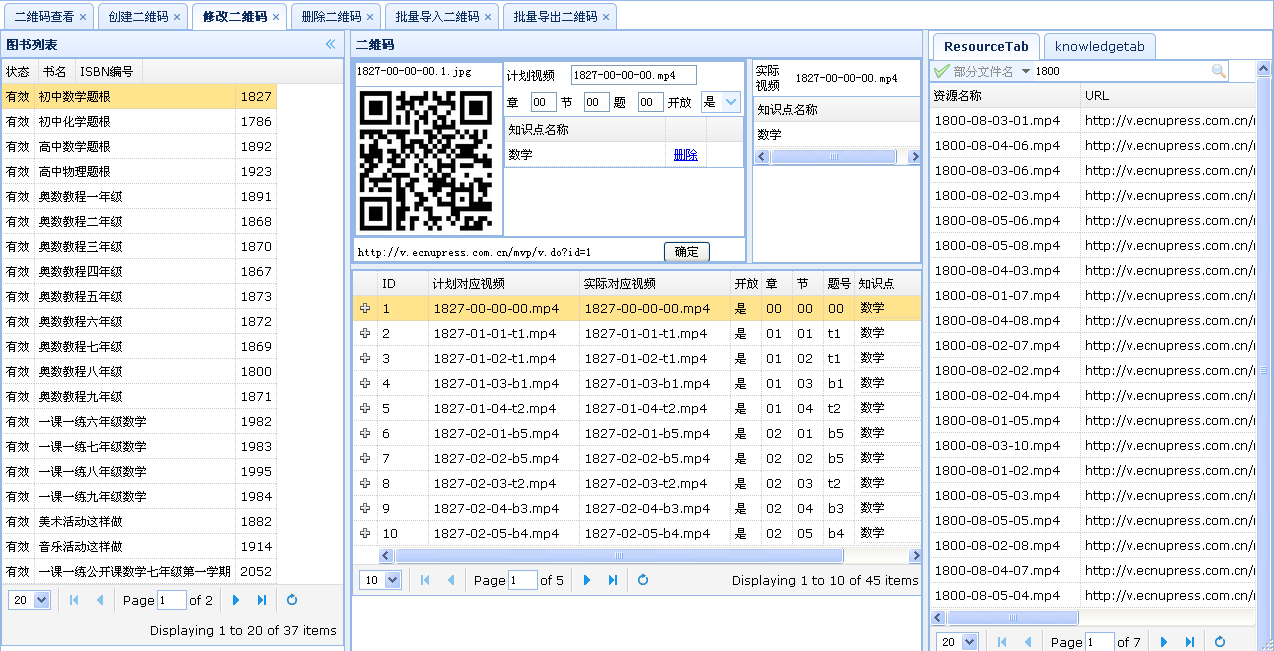


图 10 二维码管理界面图1

### 教育资源管理

* 提供：查看、新增、修改、删除、导入操作。
* 资源信息中，必填数据为：上传文件，文件名必须是ISBN-章-节-题.mp4名称： ISBN-章-节-题.mp4 必须符合此规则。关键词：首字母必须是英文逗号“，”，如果多个关键词，则通过英文逗号分隔。年级：可多选。知识点：从右侧拖拽，可多个，必须勾选一个作为主知识点。

表格 7 资源管理

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **功能** | **输入内容** | **输出内容** | **结果** | **备注** |
| [资源查看](javascript:void(0)) | 资源选择 | 资源信息 |  |  |
| [创建资源](javascript:void(0)) | 资源信息 | 新建成功 | 资源信息更新 |  |
| [修改资源信息](javascript:void(0)) | 资源信息 | 更细成功 | 资源信息更新 |  |
| [删除资源](javascript:void(0)) | 资源选择 | 删除成功 | 资源列表更新 |  |
| [批量导入资源](javascript:void(0)) | 资源列表信息 | 资源导入信息 | 资源批量创建 |  |

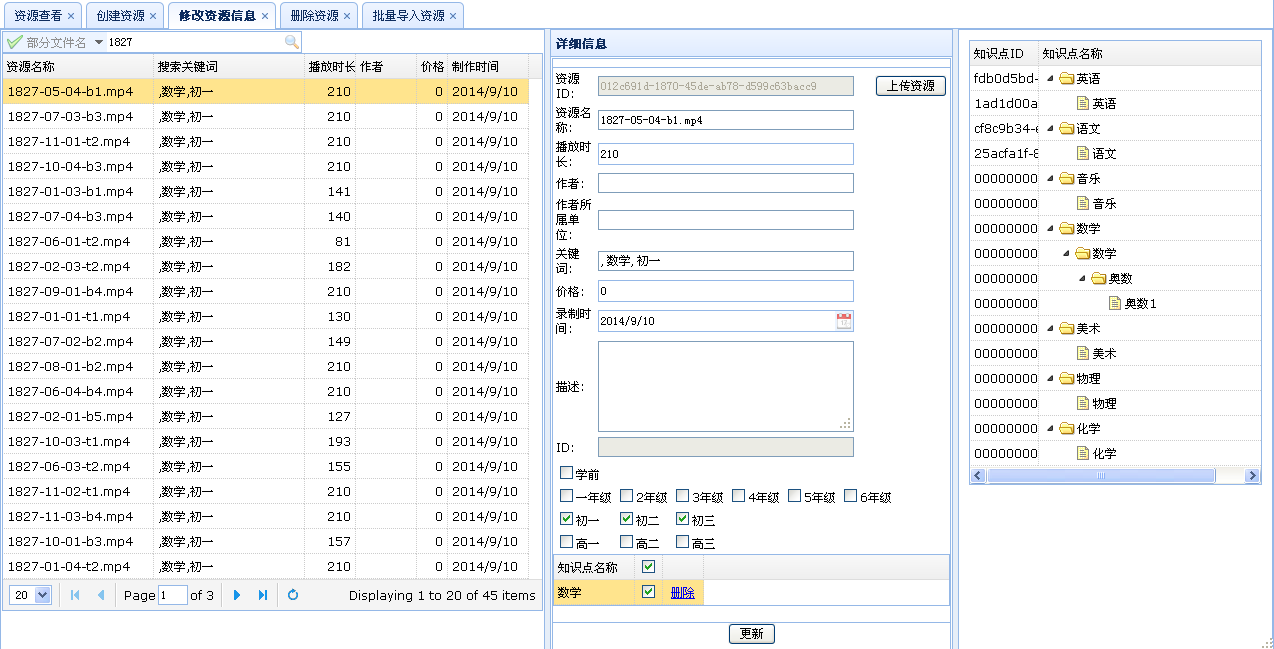


图 11 资源管理

移动客户端系统设计和实现

“华师微视”客户端用户可通过下载安装移动终端应用程序，完成注册，登录，通过拍摄纸质图书上的二维码获取电子资源，可进行点击观看资源或收藏资源等操作；用户基于应用程序所能实现的功能，输入内容与操作如下描述：

### 功能概述

**（1）用户功能**

* 用户管理：注册、修改用户信息

系统通过用户信息收集来满足后续用户个性化教育服务分析。

* 成就记录：可查看学习时长与历史观看记录，把握学习进度。

通过总时间长度和图表来显示学习进度，并显示在所有用户中的排名占比，以提高用户积极性。且保存历史记录以供用户查看资源观看的情况。

* 个人信息管理：个人信息修改，活动消息获取。

**（2）资源获取**

* 学习资源管理： 按热门、筛选、关键字搜索获取资源。
* 通过用户的观看习惯与筛选条件设置热门资源的显示条目，按照热门程度进行排序帮助用户获得最有效的资源。
* 二维码拍摄：拍摄二维码进行直接在线播放教育资源。
* 学生在阅读纸质书时，可通过拍摄书上的二维码来观看对应的教学资源。
* 视频相关信息：通过播放数、收藏数、点赞、分享数量显示资源热度。
* 用户再次观看的教学资源时，可点击收藏，将资源保存入个人收藏中，登录即可观看。对教学资源表示满意时可点赞，也可将视频内容分享到自己的各种朋友圈子内，进行交流学习。

### 功能清单

表格 8 系统功能清单表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **功能模块** | **功能** | **输入内容** | **输出内容** | **结果** | **备注** |
| 用户管理 | 注册 | 个人用户信息 | 个人用户信息 | 注册成功 | 建立会员用户信息 |
| 登录 | 登录 | 用户名，密码 | 登录结果 | 登录成功 | 登录验证 |
| 资源列表 | 热门资源 | 筛选条件 | 资源列表 | 资源按播放次数排列 |  |
| 资源搜索 | 关键字 | 资源列表 | 显示关键字对应结果 |  |
| 资源筛选 | 选择年级学科 | 资源列表 | 资源列表更新 |  |
| 成就记录 | 学习成就 | 播放统计 | 学习时长与用户排名 |  |  |
| 播放记录 | 播放统计 | 历史播放内容 |  |  |
| 二维码 | 二维码扫描 | 二维码信息 | 资源对应链接 | 后台识别二维码并返回对应资源，应用播放二维码资源 | 检索二维码资源 |
| 试题讲解 | 视频播放 | 二维码扫描或资源选择 | 资源对应链接 | 显示对应视频页面 | 可全屏播放 |
| 播放数统计 | 用户页面加载 | 加载次数统计 | 显示播放次数 |  |
| 收藏 | 点击收藏 | 已收藏 | 在我的收藏栏目下显示 |  |
| 分享 | 选择分享 | 分享成功 | 分享次数更新 |  |
| 点赞 | 点击赞图标 | 图标变色 | 点赞次数更新 |  |
| 相关资源 | 选择资源条目 | 对应资源链接 | 跳转至对应资源 |  |
| 学习测试 | 试题测验 | 答案 | 显示测试结果 |  |
| 我的信息 | 基本信息 | 修改用户信息 | 个人用户信息 | 更新用户信息 |  |
| 我的收藏 | 资源选择 | 对应资源链接 | 显示对应资源 |  |
| 我的成就 | 播放统计 | 学习时长与用户排名 |  |  |
| 我的消息 | 消息选择 | 显示对应消息 |  |  |
| 系统设置 | 修改密码 | 原密码，新密码，确认密码 | 修改结果 | 修改用户密码 |  |
| 版本升级 | 检测新版本 | 新版本可更新 | 选择后更新应用 |  |
| 关于 | 点击查看 | 版本信息 |  |  |
| 注销 | 点击退出 | 注销账号 | 回到登录页面 |  |

### 系统界面

**（1）用户登录和注册，如图 12 用户注册和登录图。**



图 12 用户注册和登录图

**（2）****教育资源浏览和筛选，如图 13 教育资源浏览和筛选。**

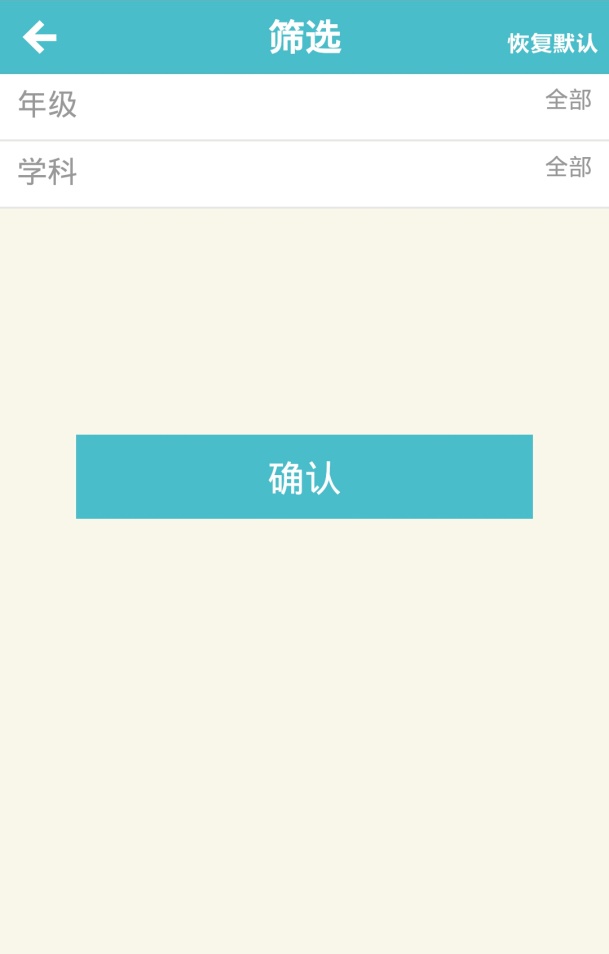


图 13 教育资源浏览和筛选

**（3）****个人成就管理，如图 14 个人成就管理。**



图 14 个人成就管理

**（4）****二维码拍摄和资源浏览，如图 15 二维码拍摄和资源浏览。**

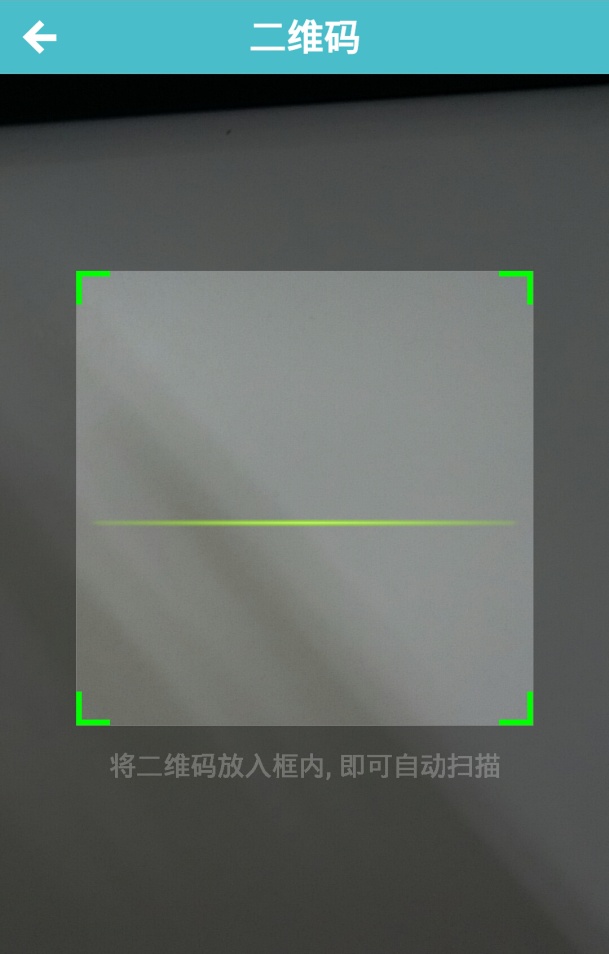


图 15 二维码拍摄和资源浏览

**（5）****视频资源观看界面，如图 16 视频资源观看界面。**

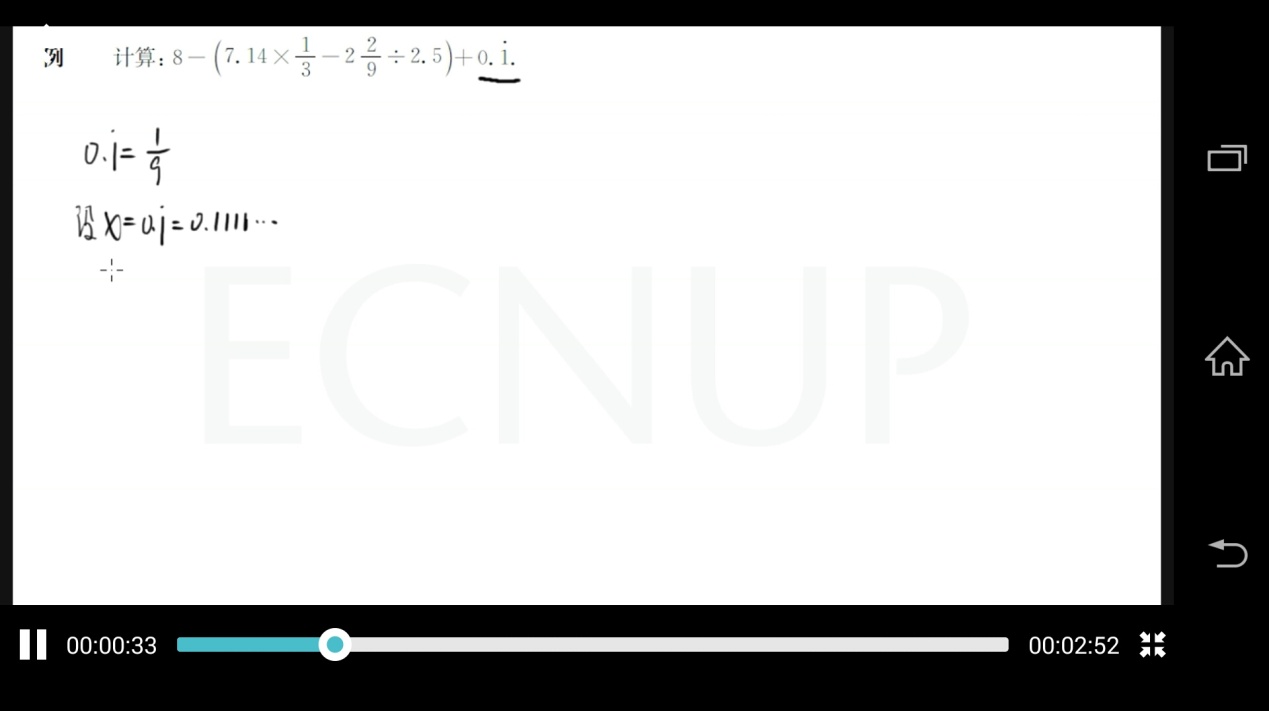


图 16 视频资源观看界面

**（6）****资源分享和设置管理，如图 17 资源分享和设置管理。**



图 17 资源分享和设置管理

1. 关键技术
   1. 教辅视频学习资源的推荐

### 概述

推荐系统的产生是信息时代一个必然产生的结果，当我们面对互联网上纷繁复杂的信息时，我们关注的是如何快速、准确的获得符合自己需求的信息。传统的搜索引擎在用户没有明确的需求或者潜在需要不得而知时信息筛选效果不佳。而一些购物网站的分类信息方法虽然可以让用户通过类别对想要的物品进行搜索，但面对海量商品信息时依然存在信息过载的问题，推荐系统为解决这类问题提供了全新的视角。推荐系统能够在信息过载的环境中帮助用户发现令他们感兴趣的信息，自动关联用户和物品，也能将信息推送给对它们感兴趣的用户。以做电子商务的亚马逊网站为例，它为用户做的个性化推荐核心就是比较用户的消费偏好，并于其他用户进行对比，借以预测用户可能感兴趣的商品分区的混合的机制，并将不同的推荐结果分不同的区显示给用户。另一个典型案例如基于图书分享的社交网络豆瓣网，它的核心也是基于社会化的协同过滤的个性化推荐，给用户提供最满意的电影音乐资源。

将推荐系统应用于教育领域是众多学者一直在研究的问题，对网上学习系统来讲，教育资源推荐系统是可以根据自身的知识背景、教育程度和学习进度,通过网络随时获取适合的教学资源,并依据个人要求完成对学习资源的组织提升学习效果。

### 视频学习资源用户行为数据

推荐系统中，用户行为特征是关键信息。不同的业务关注的行为数据有所不同。在线服务系统中，用户点击数据，登陆数据通常是共性关键数据。在学习系统中，学习者对于学习资源的关注是关键。本系统中，我们设计如表所示的用户行为模型。在二维码增强的教辅服务中，用户行为有表所示。我们在该系统中设计如表格 10 用户行为数据。

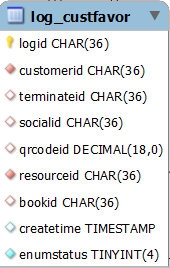
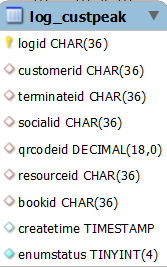
表格 9 用户行为数据

| **用户行为** | **类型** | **特征** | **作用** |
| --- | --- | --- | --- |
| 评分 | 显式 | 整数量化的偏好，可能的取值是 [0, 5]； | 通过学生对二维码视频内容的评分，可以精确的得到用户的偏好 |
| 点赞 | 显式 | 布尔量化的偏好，取值是 0 或 1 | 通过学生对二维码视频内容的点赞，可以较精确的得到用户的偏好 |
| 转发 | 显式 | 布尔量化的偏好，取值是 0 或 1 | 通过用户对二维码视频的投票，可以精确的得到用户的偏好。 如果是站内，同时可以推理得到被转发人的偏好（不精确） |
| 保存书签 | 显示 | 布尔量化的偏好，取值是 0 或 1 | 通过用户对二维码视频的是否保存为favorite，可以精确的得到用户的偏好。 |
| 评论 | 显示 | 一段文字，需要进行文本分析，得到偏好 | 通过分析用户对二维码视频的评论，可以得到学生的情感：喜欢还是讨厌 |
| 视频播放次数 | 隐式 | 受多种因素影响：包括讲解的是否深入浅出；知识点本身的难易程度；学术的接受能力等； | 学生播放次数一定程度上反映了学生的注意力和需求 |

### 推荐模型的设计

#### 数据分析与预处理

所使用的数据中二维码对视频自由共有5146个，注册使用用户共有4220位。我们获取了该项服务的半年累计数据，经过初步分析，得知在本批数据中，收集到了一定数量的用户点赞、播放和收藏数据，而没有收集到有效数量的评分数据和用户评论数据。因此选择用户点赞、用户播放、用户收藏、用户评论作为推荐实验数据。数据库中的相关实体如图 18 数据库实体：



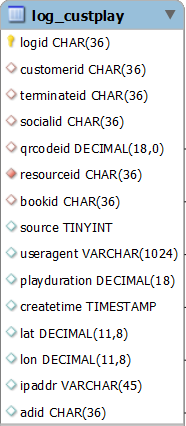
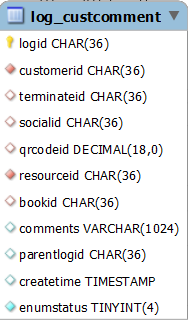


图 18 数据库实体

1）用户点赞数据

其中log\_custpeak记录用户关于某一个知识点对应的视频资源的点赞，点赞数值类型是布尔型（0,1），值1表示用点赞。此表中我们只关心字段customerid和qrcodid，数据库原始数据部分如下,标示筛选出来的点赞用户及相关视频资源对应的二维码id：

Customerid,qrcodid,enumstatus

'008f85de-52df-4ba3-abbe-21e8a3d43655', '1211', '1'

'00999195-4196-44d9-b4d2-f4863dadcede', '3253', '1'

据此获得如表所示的矩阵模型，如表格 10 矩阵模型表。

表格 10 矩阵模型

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | QrResource1 | QrResource1 | QrResource1 | … | QrResourceN |
| Customer1 | 0 | 1 | 0 | 0 | 1 |
| Customer1 | 1 | 1 | 1 | NA | 1 |
| … | 1 | NA | 0 | 1 | NA |
| CustomerN | 0 | 0 | NA | 0 | 1 |

1）用户收藏数据

log\_custfavor记录用户是否收藏某一个知识点对应的视频资源，用布尔型（0,1）标示，值1表示用点赞,数据效果同log\_cuspeak相似。同理可获得用户收藏矩阵模型，如表格 11 用户收藏矩阵模型。

表格 11 用户收藏矩阵模型

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | QrResource1 | QrResource1 | QrResource1 | … | QrResourceN |
| Customer1 | 1 | 1 | 0 | 1 | 1 |
| Customer1 | 0 | 0 | 1 | NA | 1 |
| … | 1 | NA | 0 | 1 | NA |
| CustomerN | NA | 0 | NA | 0 | 0 |

3）用户播放行为

log\_custplay记录用户播放视频资源的动作，记录了其播放时间相关信息。本表中可记录用户播放视频的时长，可通过计算统计用户播放同一个视频的次数。本次实验中，关注用户播放次数，数据形式如下，其中ount（play）有数据库中统计而出，标示用户customerid播放qrcodeid资源的次数。

Customerid,qrcodid,count(play)

'22ec4a89-bea4-4d47-9396-f9ad8d4480db', '1241', '3'

'22ec4a89-bea4-4d47-9396-f9ad8d4480db', '1244', '3'

'23036042-5780-4f8a-9432-da2b4bef21ce', '1351', '3'

'250a7770-7d84-4536-a279-90d2b649734d', '2458', '3'

同理获得用户播放行为矩阵模型，如表格 12 用户播放行为矩阵模型。

表格 12 用户播放行为矩阵模型

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | QrResource1 | QrResource1 | QrResource1 | … | QrResourceN |
| Customer1 | 0 | 3 | NA | 1 | 1 |
| Customer1 | 3 | 2 | 2 | NA | 2 |
| … | 4 | NA | 2 | 3 | NA |
| CustomerN | 1 | 5 | NA | 0 | 1 |

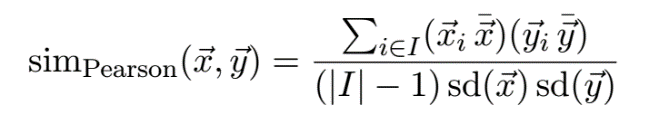
#### 多维度推荐模型

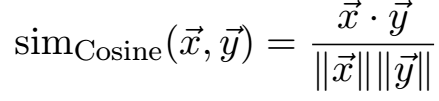
多维推荐模型如图 19 多维推荐模型图。

用户点赞模型和用户收藏模型属于（0-1）类型数据模型，使用Jaccard相似度计算方法。其公式如下：

其中x和y分别是用户和用户对于item的偏好为1的集合。将用户点赞模型相似度记为，用户收藏模型记为。

对于用户播放行为矩阵，将其当做用户评分表来设计。用户相似度可以采用常见的余弦相似度和皮尔森相似度计算公式来进行计算。





通过上述方法计算获得的行为相似度记为。最终相似度为

其中系数a，b，c是实验确定的加权因子。

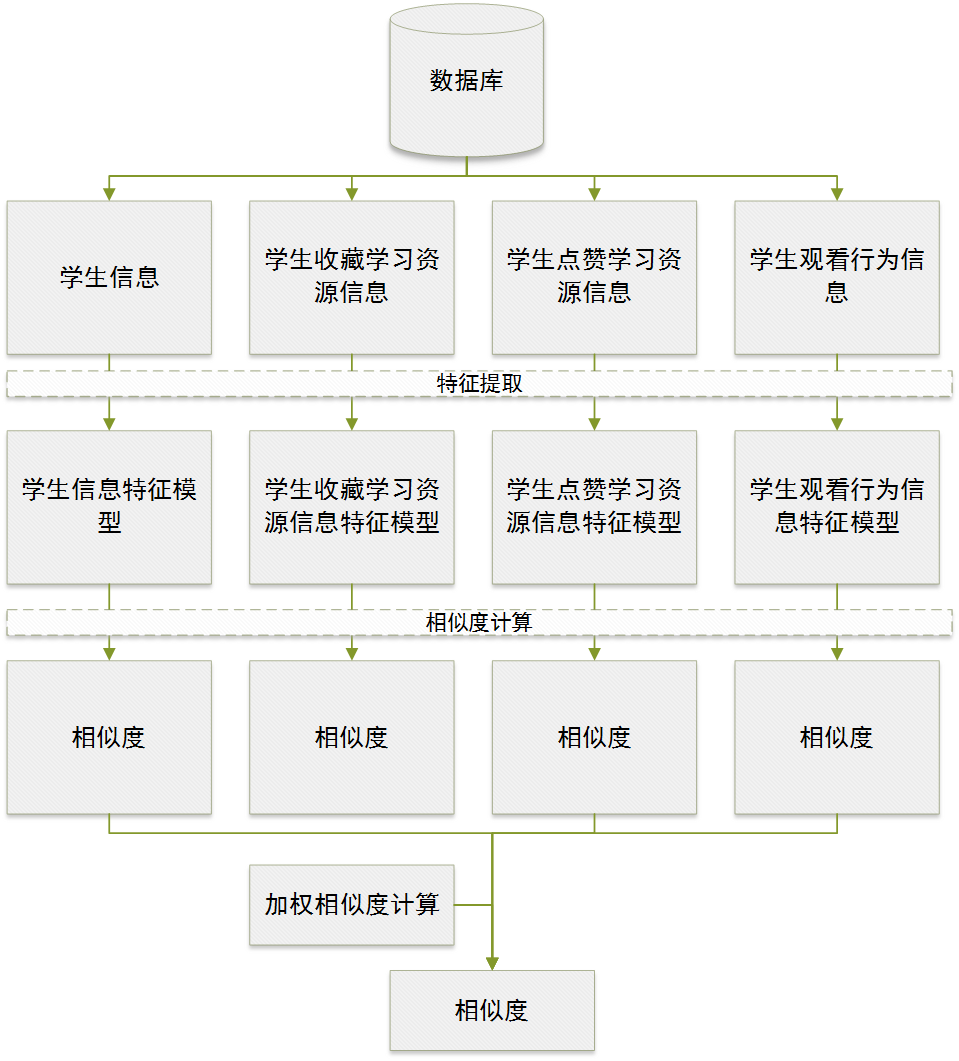


图 19 多维推荐模型图

智慧型学习资源模型

教育理论、教育技术和资源建设的观点来讲.AECT(美国教育与传播学会)对学习资源的定义为大家所公认。在AECT’77定义中，学习资源分为“设计的资源”和“利用的资源”。在AECT’94定义中，学习资源是支持学习的资源，包括教学材料、支持系统、学习环境。甚至可以包括能帮助个人有效学习和操作的任何因素[2](file:///C:\Users\meflyup\AppData\Local\gitbook\app-5.0.4\resources\app.asar\editor.html)。显然，这是一种极为宽泛的定义。

学习对象，如基于IEEE LOM，ADL SCORM标准制作的学习资源也属于学习资源的范畴。学习对象本质上可以看做是通过一些列的元数据描述来规范上述宽泛的学习资源类型，其设计思想来源于面向对象的程序设计，学习对象的分发依赖于资源打包系统以及LMS对于学习对象的解析支持功能。学习对象规范和技术提出的目的是解决学习资源的重用性、可操作性、可获取性、可聚合性等等，其设计思想来源于面向对象的程序设计，如通过IEEE LOM来规范实现学习对象的可发现性[3](file:///C:\Users\meflyup\AppData\Local\gitbook\app-5.0.4\resources\app.asar\editor.html)，通过IMS规范实现重用性，通过SCORM实现互操作性等[4](file:///C:\Users\meflyup\AppData\Local\gitbook\app-5.0.4\resources\app.asar\editor.html)。  
然而随着技术的进步和学习情境、学习需求的变化，学习对象难以适应时代的发展需求。如上所述，学习对象设计目标是学习资源的重用性、可操作性、可获取性、可聚合性，这一目标考虑的出发点本质上是如何经济的使用学习资源，也就是说它是面向投入产出这样的经济效益角度来出发设计的。对于面向学习者的学习体验上的设计考虑不足，比如情境感知、自适应、学习者学习行为数据的有效采集和管理、自发现和组装、学习环境独立性等等，也就是学习对象缺乏智能性。 本项目吸收利用学习对象，是开发智能型学习资源关键技术。

### 总体思想

我们吸收采纳学习对象的成果，同时放弃学习对象打包发布方式和对LMS的需求，采用web上基于URL寻址的学习资源和基于web浏览器的学习系统来作为智能型学习资源的关键技术，如图 20 智能型学习资源模型应用生态。

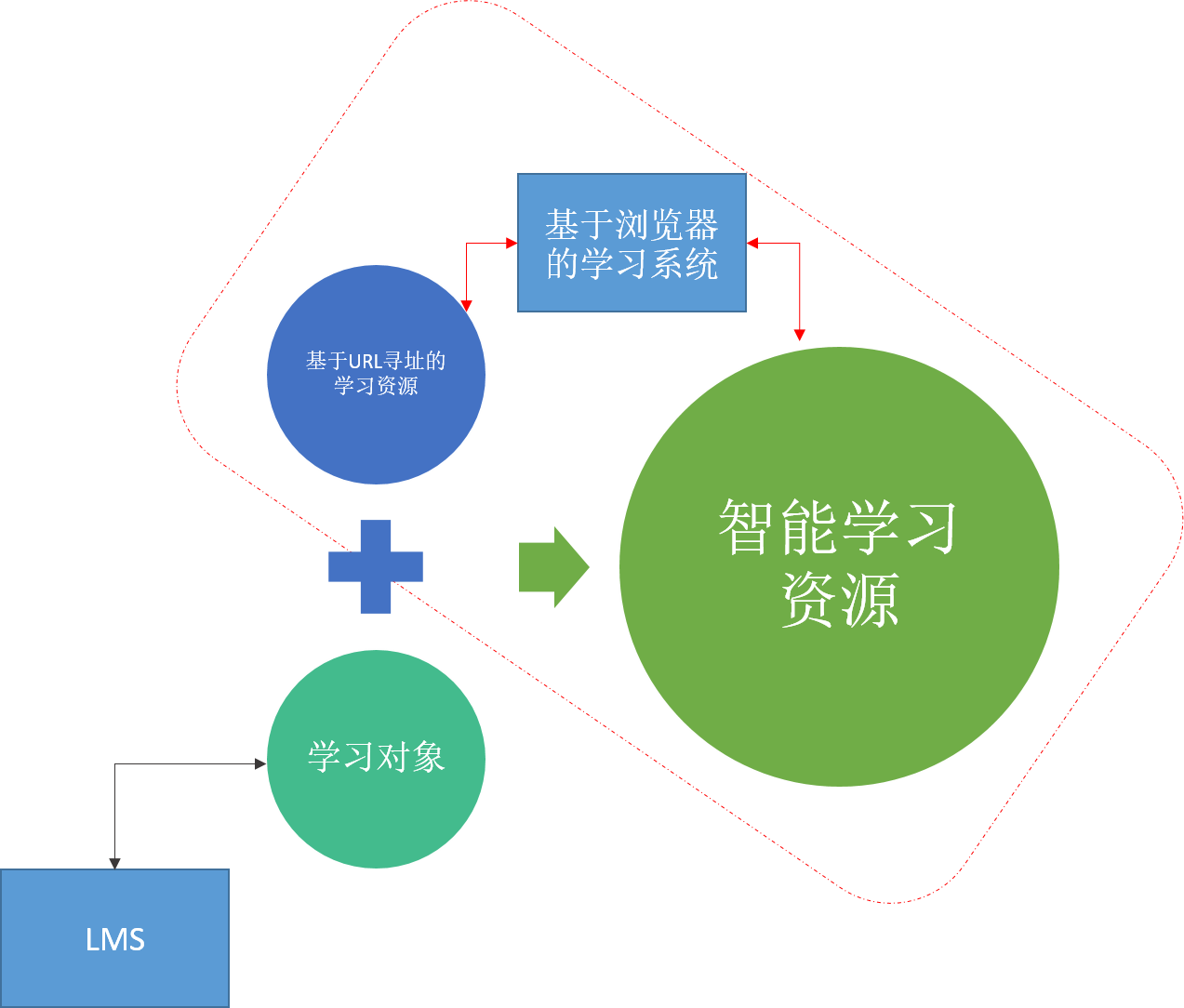


图 20 智能型学习资源模型应用生态

### 智能资源设计与聚合

学习资源开发者在一定的规则下开发基于Web技术的学习资源，这些学习资源中的网页在传统Web网页基础上嵌入了一种于HTML5标准中提出的新的事件机制：PostMessage。通过该机制绕开浏览器同源策略（SOP，Same Origin Policy）限制，该Web页面可以同其他不同来源的页面在同一网页进行通讯，从而达到聚合的目的。

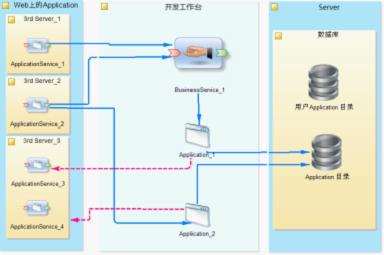


图 21 智能型学习资源聚合引擎及server 交互-a

图 21 智能型学习资源聚合引擎及server 交互-a和图 22 智能型学习资源聚合引擎及server 交互-b显示了整体聚合机制以及聚合引擎所在的位置。一个开发者通过平台搜索已注册的Web应用数据,搜索到applicationService1和applicationService2满足其需求。开发者通过Intent机制，将二者结合起来。需要强调的是，在这个过程中，开发者并不是将这两个应用的代码融合到一个Web页面中，开发者仅仅使用Intent声明了要使用那个applicationService，在随后聚合过程中，聚合引擎将可以把不同来源的应用聚合在用户界面之中。

开发者将applicationService1和applicationService2在开发阶段聚合成为新的应用application1；开发者也可以直接用applicationService2开发新的应用application2，随后他将把application1和application2发布到服务器上，并且可将其注册到平台的application目录上以供其他开发者共享使用。

图 22 智能型学习资源聚合引擎及server 交互-b表示的是在用户使用阶段，聚合引擎的工作机制。一个认证用户可以从application 目录中挑选满足需求的应用，这些应用条目将被自动记录到用户的application 目录中，并且在适当的时候保持同浏览器端用户离线application 目录的同步。依然以contact和sendsms为例，把contact作为applicationService1，把sendsms作为applicationService2。当用户启动application服务1时，启动过程从离线数据库获取了applicationService1的URL，随即聚合引擎依据该URL从3rd server上调出contact应用页面。此时聚合引擎没有进行任何关于sendsms机applicationService2的操作。当用户点击contact中的Mobile电话号码时，该过程通过代码Intent(Intent\_SMS，arg\_msg)在Intent analyzer协助下对sendsms及applicationService2进行定位、加载、赋值等一系列操作，最终显示给用户一个新的界面，该界面即为发送短信的界面，并且其上依据自动填充arg\_msg的相关的数据，此外，用户还可以根据需要修改该数据，最后将短信发送出去。

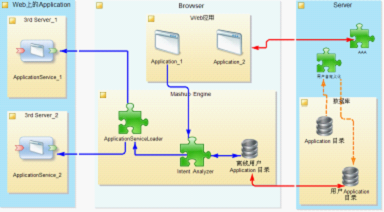


图 22 智能型学习资源聚合引擎及server 交互-b

### 基于Tin Can API增强学习资源数据采集能力

教育大数据和学习分析为教育信息化的发展带来了新的机遇，而对教育大数据应用的实现首先依赖于对来源广泛的教育数据的获取。在这些有待获取的海量的教育数据中，学习经历数据作为一种承载着学习者在学习过程中的学习行为、学习活动、学习进程和与之交互的学习环境等教育信息的数据，蕴含着丰富的学习分析价值。随着相关学习技术的迅速发展，学习经历的跟踪在技术层面已经得以实现，但学习经历的复杂性、多样性和数据劣构性等特点阻碍了对其广泛的应用。2014年，由美国“高级分布式学习”组织发布的Tin Can API(也称为Experience API), Tin Can API规范为破解这一难题提供了可行途径[[1]](#footnote-1)。

**1）技术生态**

设计Tin Can API的目的是为了存储和记录学习经验。它可以跟踪学习经验，包括传统的考试分数或完成进度记录，还包括学习者行动的记录，例如，读了一篇文章，或观看了一段视频。Tin Can API用于支持现有的SCORM及SCORM无法支持的运用，例如移动学习、未经过浏览器接触到的内容。

根据官方解释[[2]](#footnote-2)，Tin Can API的设计哲学是：当今社会，学习不再局限于LMS，而是无处不在，随时发生的。因此，必须有一个即能保持独立，又能根据需要而整合的记录系统，该系统能很好的适应大数据和云环境。为此Tin Can中设计了一个名为LRS（ Learning Record store）的关键系统，故简单来看，Tin Can API实际上是一个数据服务。通过该服务的客户端API接口，开发者将学习过程数据通过称为statement的数据形式发送给学习记录仓库,如图 23 LRS使用场景所示。

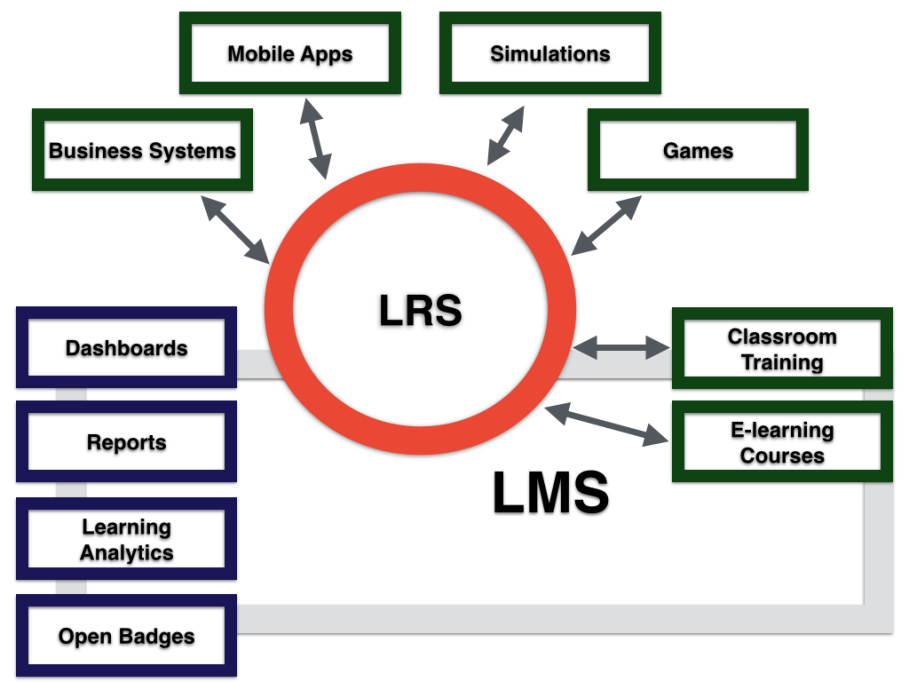


图 23 LRS使用场景

围绕这样的核心思想。Tin Can API已经形成了相对完备的技术生态。现今，已经推出了Tin Can API 客户端软件库，支持各种常见的编程语言和环境。

**2）运作机制**

Tin Can API技术利用“活动流”（Activity Stream）来描述学习经历。活动流规范是开源网络基金会（Open Web Foundation）下的一个项目，由 众 多 网 络 服 务 商 （ 包 括 Facebook、 Twitter、Google）所支持和维护，定义了基于 JSON 格式和XML 格 式 的 Atom 发 布 协 议 （Atom PublishingProtocol）。此外，该项目还初步定义了一些动词和活动，以辅助人们利用流（Stream）来对交互行为进行建模3。

第一，采集和记录行为。无论是在日常生活、工作中还是在学习中，人与人、人与内容之间都有交互行为。Tin Can API能够记录在任何地方发生的任何标识和事件。

第二，当一种活动需要被记录的时候，应用会以“主、谓、宾”的形式发出安全声明到学习记录库中。

第三，学习记录库记录所有的声明。不同记录库之间可以共享这些声明。一个学习记录库可以单独存在，也可以在LMS中分享。

使用Tin Can API的简化描述如图 24 Tin Can API 记录数据的简化原理所示：

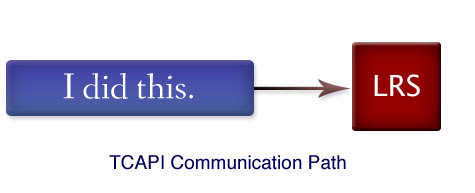


图 24 Tin Can API 记录数据的简化原理

基于浏览器的实时沉浸式辅导技术

### 关键技术

在web上的学习活动，特别是具有大量操作的学习活动，通常的解决思路是采用视频教程（tutorial）。然而，这样的方式有如下不足：1）制作视频tutoiral 费时费力，而且更新不及时。2）视频教学过程中，学生如果需要实战，就必须在视频和实际的环境中反复切换，极为不便。3）视频和环境分割，难以形成身临其境的感受。4）高质量视频资源带宽耗费大，对传播不利。因此，我们的研究考虑一种真实场景下的教程教辅技术。

1）基于互联网浏览器的学习环境

基于互联网浏览器的学习环境有别于传统的在线学习环境。该环境具有如下关键组件：

（a）内嵌为网页添加额外学习资源的脚本（Learning Resource Script）插件。该功能对用户透明。

（b）内嵌运行学习路径脚本程序（Learning Path Script，以下简称LPS）功能的插件。该功能提供真实目标网站的交互式学习体验。用户在向导脚本程序下课开展沉浸式互动学习。

技术概念设计效果如图 25 沉浸式辅导环境所示。

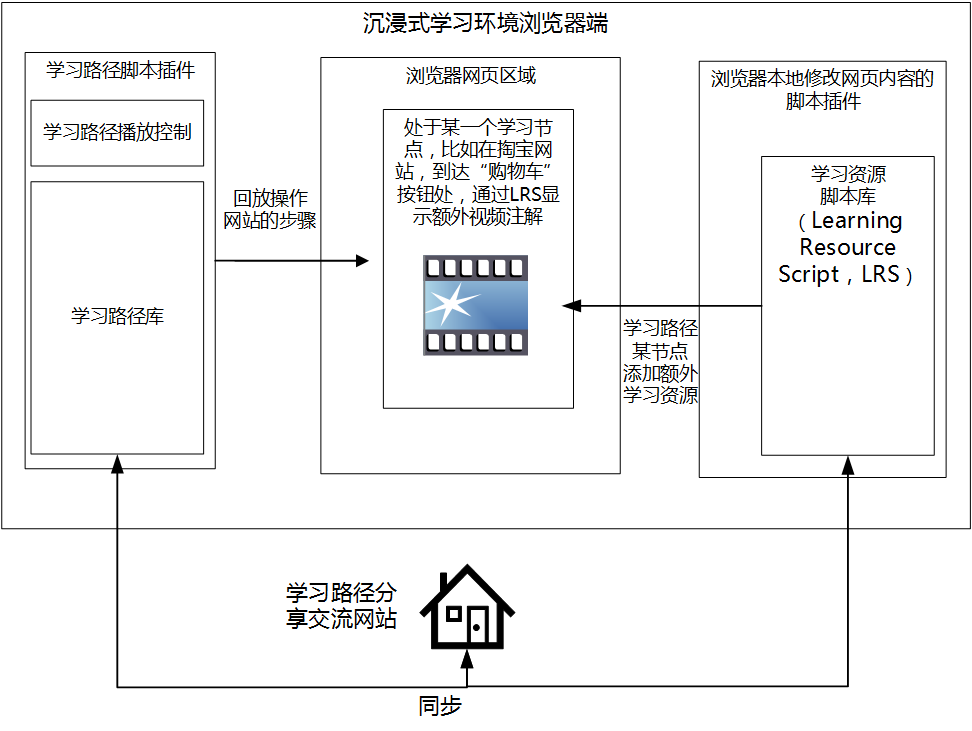


图 25 沉浸式辅导环境

要实现这样的学习环境，其关键技术在于：

(1)构建基于真实浏览器的LPS学习系统

本质上，这是一种在浏览器中的录制宏的技术。录制宏的作用是记录在网上的行为过程，然后重放。教师在目标网站上操作的同时开始录制宏，录制结束后可供学员回放。iMacros是一个能在主流浏览器IE、Firefox、Chrome运行的宏录制和播放扩展，被广泛应用web上的自动化任务，比如测试，数据抽取等。该技术为我们的开发提供了技术基础。

然而，尽管iMacros具有支持学习的潜力，但其开发之初不是为学习目的的，其设计目标是工程技术人员，因此将iMacros应用于LPS构建，是需要进行相关的适配：（1）扩展适合LPS制作的命令，（2）需改用户界面，使其更适合LPS制作和用户学习的交互式回放。

(2)在LPS的学习路径中添加辅助学习资源的技术

使用iMacros 扩展来录制和回放LPS只能对目标网站进行操作记录，不能在回放过程中修改目标网站源代码以支持添加学习资源。因此必须结合在本地修改浏览器中目标网站源代码能力。GreaseMonkey是一种广泛使用的开源本地脚本程序支持插件，使用它能够开发javascript脚本程序程序，用以在浏览器上重构目标网站网页的能力。项目将开发greamsMonkey上运行的javascript脚本程序程序，该脚本程序能够同LPS配合，LPS学习节点中的需要位置添加必要的学习资源。

(3)通常，用户需要自己在chrome 和firefox store上搜索、查找需要的插件并完成安装使用。在完成上述关键技术后，需要将上述插件预先打包到浏览器中，支持用户获得“开箱可用”的体验。

### 实现案例

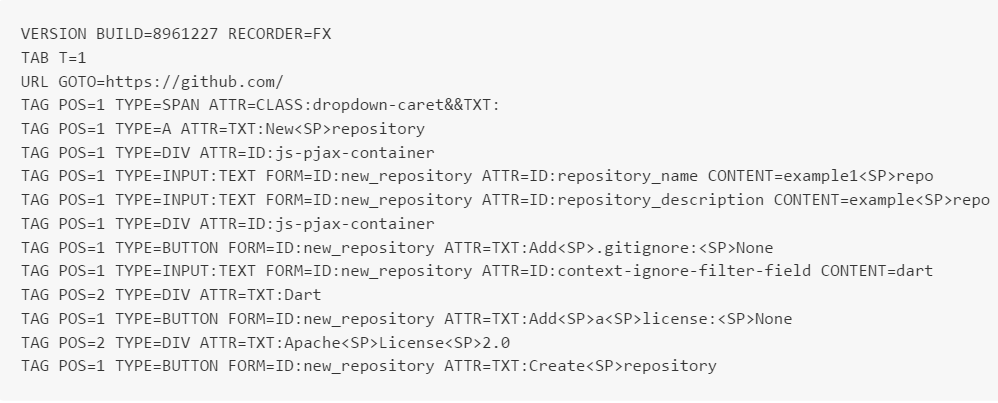
图 26 步骤辅导脚本



图 27 步骤辅导脚本增强脚本模板

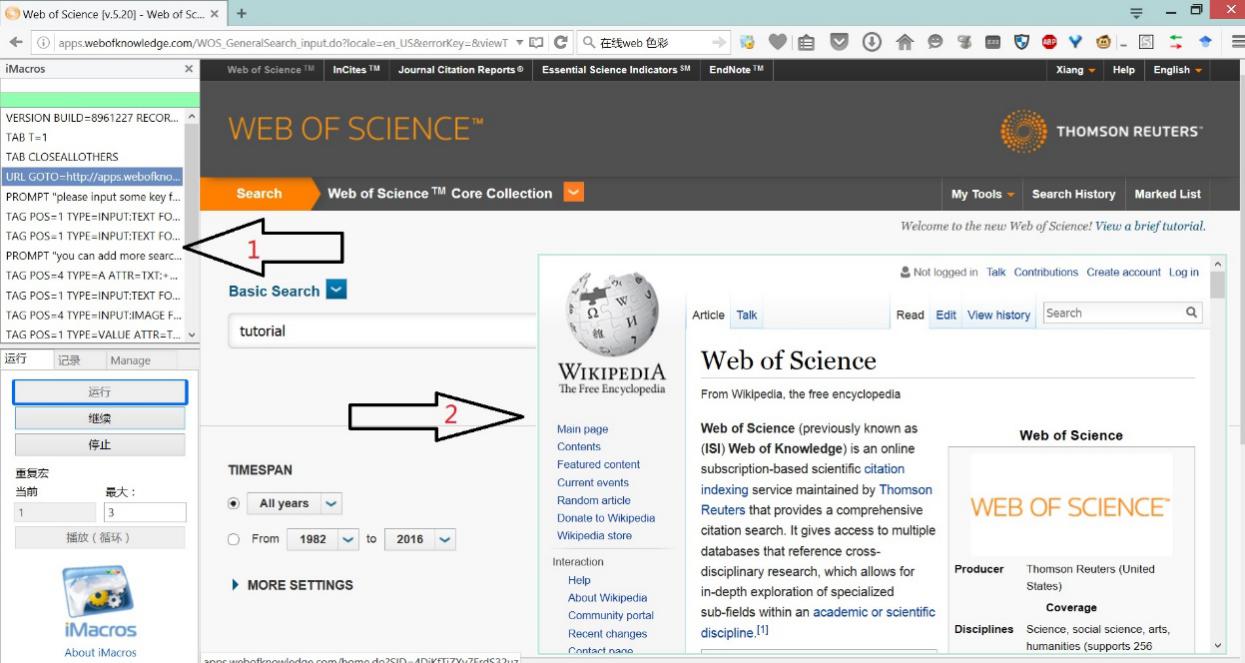


图 28 整体效果

其中1表示步骤引导；2表示在特定元素和时刻的增强解释

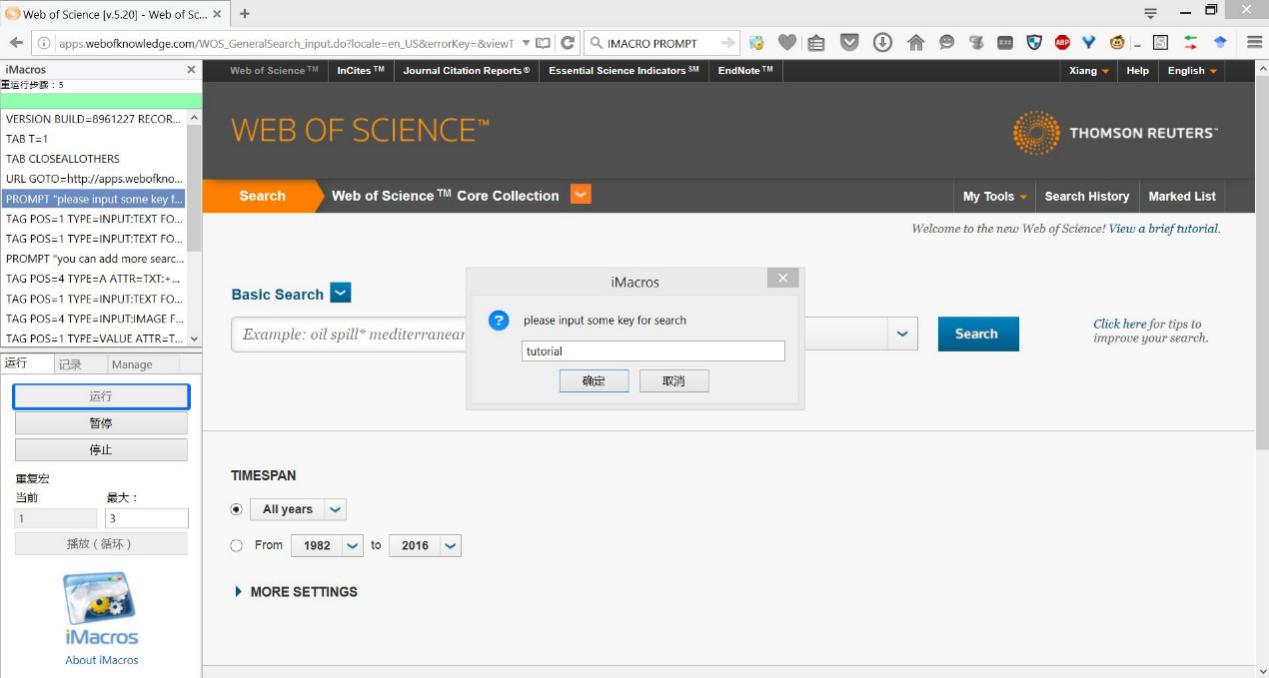


图 29 实现用户交互等待

用户行为数据采集、表示和存储

对于用户行为采集分析，主要需要研究的内容包含两个：用户行为的采集，以及用户行为的建模和存储，这两个相对独立的模块将分别采用如下技术路线实现：

### 用户行为的采集

主要依靠两种手段实现对用户行为和偏好的采集，一种手段是引导用户填写偏好设置，由用户主动将自己的喜好提交到服务端，并在服务端的数据库中记录下来，以备后用；另外一种手段是在用户无需感知的情况下，通过终端系统的配合，对用户使用服务的行为进行采集，由于在B/S或C/S架构的支撑平台上，所有的服务均面向用户提供服务，因此终端系统可通过探测用户正在使用的ID来获取其业务种类、服务内容，并将这些内容汇报给支撑平台，由平台实现用户行为的记录。

### 面向用户行为分析的元数据管理技术

用户行为分析技术中，采集用户数据的速度通常高于分析处理的速度。网络技术的社会化步调日益加快，表征用户行为的数据信息日益增大，用户特征数据信息的增加速度和分析速度的矛盾是用户行为分析中最为突出的问题。因此，需要找到更快、更加全面和更有效的网络数据管理和维护办法，以加快并能跟上用户行为分析的速度。在这种情况下，用户行为数据信息的内容、质量、状况等元数据信息变得更加重要，成为用户数据信息资源有效管理和应用的重要手段。

元数据是以数据高效利用和交换为目的的数据集说明性数据，主要包括对数据集、与数据集相关信息、数据集各数据项说明以及数据用户访问、检索、更新的方法。同时元数据也包括基于不同数据领域如何尽可能全面反映基本数据的信息。

为了实现网络行为分析技术中网络数据的快速有效交换，进行必要的数据映射或处理，研究制定网络行为数据规范和网络行为描述文件规范 。

### 用户行为建模技术

作为用户行为分析基础和核心，用户行为建模及模型的质量直接关系到行为分析的质量。只有当用户行为信息可以很好的被系统“理解”的时候，才可能实现理想的行为分析服务。

用户行为模型是根据网络应用业务来定制的一个模型。用户行为分析系统中的行为模型不是对用户行为个体的一般性描述，而是一种面向算法的、具有特定数据结构的、形式化的用户行为描述。为了实现用户行为分析，首先需要跟踪和学习用户行为，并在此基础上对用户行为进行建模。所以我们只有通过各种不同的途径充分收集、了解和跟踪用户行为的信息，通过网络业务建立用户行为模型和通过标签抽取规则建立网络标签抽取规则，再利用用户行为模型和标签抽取规则对用户行为信息进行处理，提取出标志着用户行为的标签，并用标签来表示不同用户行为的各种特征、取向，进而为下一步的趋势判断做好准备。因此需要设计一种合理的用户行为模型，来满足这种需求。本课题在以前的用户行为建模技术的基础上，加入了可定制性，使得用户行为模型更具有人性化、以及可扩展性。

为了实现系统的可扩展性和个性化定制需求，本项目将运用一种新的基于标签的用户行为建模技术对用户行为进行描述，并且可根据实际需求定制用户行为模板。

### 用户行为特征模版

用户行为特征模板将以标签的形式存在。标签是对被标签的物体的某个方面的属性进行的描述，往往采用简短的描述，被标签物体上的标签数量可以是无穷尽的。在本课题中，标签分为以下几种：

静态标签，静态标签的属性往往是不随时间和被标签物体的行为而变的。

动态标签，动态标签的属性是随着某特性而变的，可能是时间，也可能是某个触发条件，按照触发条件的不同，动态标签可以分为时间型动态标签和抑制/增强型动态标签。

链接型标签，链接型标签是一类特殊标签，本身并不具有任何内容，而是按需产生的，由其他标签推导而出的具有依附特性的标签，但其依附的主体发生变化时候，它也将发生变化，甚至消亡，他的作用是为了便于理解和系统处理。

标签作为一个实体，由标签集合和标签规则集组成的。其中标签集合由一组标签组成；标签规则集则由一组规则组成，每个规则通过一定的原则对动态标签进行更新，每个规则需包含其运行的触发条件。

### 用户行为模型生成引擎

用户行为标签库是根据系统的具体需要而生成，如果系统的需求不一样，就可以设定不同的用户行为标签库，即可定制。用户行为标签的结构是层次化的，其中叶子节点必定是标签，非叶子节点也可成为标签。

标签静态规则生成，基于用户行为抽取的静态数据。静态的属性可以从用户行为知识库中直接获得，比如ARP攻击，或者通过对数据的分析获得，比如蠕虫攻击。生成的标签存放在用户标签库中。

标签动态规则生成，动态生成规则基于用户行为的动态行为和信息。要生成某一动态用户特征，根据特定需求，由特定的规则方法来实现。譬如根据资源的标签与此特征的匹配算法进行映射，或者根据用户行为的特殊性推导出。生成的标签也存放在用户标签库中。

标签潜规则生成，基于用户行为特征模板的实时性，用户行为标签的生成并不是一次性的，是根据时间退化以及用户行为的改变不断更新。因此，需要一些潜规则来对标签进行约束和更新。如时间退化规则；抑制规则；链接规则。时间退化规则与需求相关，针对不同的标签，设定适合的退化周期；抑制规则考虑不同标签之间的抑制特性；链接规则可以生成新的由现有标签所组成的虚拟标签。潜规则也是可定制的，这些内容存放在用户行为规则库中。

本课题将分析用户行为历史数据库得到用户行为的静态信息及动态行为信息；根据需求模块定制的业务/用户需求构建用户行为模型，而需求模块配置的抽取规则构成用户行为标签的抽取规则。用户行为模型生成引擎综合这些信息生成用户行为数据库，用户行为标签库及用户行为规则库。该模块需要解决用户行为的自动标注，网络行为关联的标注主要来源于他访问过的资源的标签，及通过标注抽取规则从相关资源和网络资源上去挖掘。

用户行为特征模式抽取

个性化服务的形式是多种多样的，即可以是向用户推荐文档、商品或网页的个性化推荐服务，也可以是在用户检索信息的过程中提供个性化检索结果的个性化检索服务等等。但是有这些不同形式的个性化服务都首先需要建立对用户的描述，然后才能据此针对不同的用户提供不同的个性化服务。用户兴趣模型是个性化服务的基础和核心。

作为个性化服务的基础和核心，用户兴趣模型的质量直接关系到个性化服务的质量。只有当用户的兴趣、偏好和访问模式等用户信息可以很好的被系统“理解”的时候，才可能实现理想的个性化服务，我们的可定制用户行为建模也成功了一大半。根据用户信息构建用户兴趣模型，即用户兴趣建模，是个性化服务的核心和关键技术，也是个性化服务中最重要的一环。

用户兴趣模型是根据业务/用户需求来定制的一个用户兴趣模型。仅仅是对用户兴趣的准确描述还不能称为用户建模。个性化服务系统中的用户模型不是对用户个体的一般性描述，而是一种面向算法的、具有特定数据结构的、形式化的用户描述。为了实现播控平台上的个性化服务，首先需要跟踪和学习用户的兴趣和行为，并在此基础上对用户的行为兴趣进行建模。所以我们只有通过各种不同的途径充分收集、了解和跟踪用户的信息以及用户的行为，通过业务/用户需求建立用户兴趣模型和通过标签抽取规则建立用户标签抽取规则，再利用用户兴趣模型和用户标签抽取规则对用户信息进行处理，提取出标志着用户的标签，并用标签来表示不同用户的各种特征、兴趣，进而为下一步的推荐做好准备。因此需要设计一种合理的用户行为兴趣模型，来满足这种需求。本课题在以前的用户行为兴趣建模技术的基础上，加入了用户的可定制性，使得用户行为兴趣模型更具有人性化、以及可扩展性。

目前来说，常见的用户描述文件从内容上可以划分为基于兴趣的和基于行为的两种类型。基于兴趣的用户描述文件可以表示为加权矢量模型、类型层次结构模型、加权语义网模型、书签和目录结构等。基于行为的用户描述文件可以表示为用户访问模式。在相对不变的系统环境中，以上这些模板都是适用的。但是，随着播控平台的逐步发展、扩充，所提供的服务与涉及的领域也会越来越多，用户的兴趣模型也需要进行更新，添加新的内容和领域。有时，也需要根据业务或者用户需求筛选出具有某些特征的用户，而这些特征可能是现有的用户描述文件中未定义的。面对这样的情况，目前的技术就显得不够适用了。为了实现系统的可扩展性和个性化定制需求，本项目将运用一种新的基于标签的用户兴趣建模技术对用户的行为兴趣进行描述，并且可根据实际需求定制用户兴趣模板。

### 用户的兴趣模版

用户的兴趣模板将以标签的形式存在。标签是对被标签的物体的某个方面的属性进行的描述，往往采用简短的描述，被标签物体上的标签数量可以是无穷尽的。在本课题中，标签分为以下几种：

（1）静态标签

静态标签的属性往往是不随时间和被标签物体的行为而变的，这类标签往往与生俱来。如对于人来说，姓名，出生日期，性别，成人后的身高（一般意义认为不变），身份证号码，父母名称等。而对于信息内容而言，信息长度，信息创建时间，信息文本内容等，都属于静态标签的范畴。

（2）动态标签

动态标签的属性是随着某特性而变的，可能是时间，也可能是某个触发条件。动态标签是解决用户兴趣变迁的主要途径。按照触发条件的不同，动态标签可以分为：

a. 时间型动态标签：随时间影响，可能退化，可能增强，比如说喜好，爱打篮球，可能随着年龄的增长降低了，但也有可能增强；

b. 抑制/增强型动态标签：某些标签，会被其他标签所覆盖，也就是说这些标签会由于其他标签的产生而消亡或降低标签强度，如喜欢看动作片，会被在他之后的不喜欢动作片抑制或增强（后向抑制/增强），当然，当前设定的抑制/增强型标签也可能被前面已经设定的标签所抑制或增强（前向抑制/增强）。

（3）链接型标签

链接型标签是一类特殊标签，本身并不具有任何内容，而是按需产生的，由其他标签推导而出的具有依附特性的标签，但其依附的主体发生变化时候，它也将发生变化，甚至消亡，他的作用是为了便于理解和系统处理。

标签作为一个实体，由标签集合和标签规则集组成的。其中标签集合由一组标签组成；标签规则集则由一组规则组成，每个规则通过一定的原则对动态标签进行更新，每个规则需包含其运行的触发条件。

### 用户模型生成引擎

用户标签库是根据推荐系统的具体需要而生成，如果推荐的需求不一样，就可以设定不同的用户标签库，即可定制。用户标签的结构是层次化的，其中叶子节点必定是标签，非叶子节点也可成为标签。如“爱运动”这个大标签下有“篮球”、“足球”子标签。

（1）标签静态规则生成

静态生成规则基于用户静态数据。静态的属性可以从用户数据库中直接获得，比如性别，或者通过对数据的分析获得，比如年龄段。生成的标签存放在用户标签库中。

（2）标签动态规则生成

动态生成规则基于用户的动态行为和信息。要生成某一动态用户特征，根据特定需求，由特定的规则方法来实现。譬如根据资源的标签与此特征的匹配算法进行映射，或者根据用户的特定浏览行为推导出。生成的标签也存放在用户标签库中。

（3）标签潜规则生成

基于用户模板的实时性，用户标签的生成并不是一次性的，是根据时间退化以及用户行为的改变不断更新。因此，需要一些潜规则来对标签进行约束和更新。如时间退化规则；抑制规则；链接规则。时间退化规则与用户需求相关，针对不同的标签，设定适合的退化周期；抑制规则考虑不同标签之间的抑制特性；链接规则可以生成新的由现有标签所组成的虚拟标签。潜规则也是可定制的，这些内容存放在用户规则库中。

本课题将分析用户行为历史数据库得到用户的静态信息及动态行为信息；根据需求模块定制的业务/用户需求构建用户兴趣模型，而需求模块配置的抽取规则构成用户标签的抽取规则。用户模型生成引擎综合这些信息生成用户数据库，用户标签库及用户规则库。该模块需要解决用户的自动标注，用户关联的标注主要来源于他访问过的资源的标签，及通过标注抽取规则从相关资源和网络资源上去挖掘。

用户流失分析

将社会网络上的信息传播过程进行建模、学习和测试；将话题的传播趋势模型化为一组带有时间标注的序列，将当前已经发生的事件作为训练样本，通过对当前时刻之前的已发生的事实数据进行训练，从而生成分类和预测规则，以达到对当前时间之后的话题趋势进行预测分析的目标。

### 面向教育资源的冷热度预测算法

在教育资源平台上如何对新闻节目建立预测模型是一项关键任务，预测结果的好坏很大程度取决于模型是否能真正符合事物发展的规律，这里提出了结合教育资源平台上节目自身数据结构及节目发展过程，一种全新的基于用户-项目-时间的节目热度预测模型UIT-TPM(User-Item-Time Trend Prediction Model)。算法模型如图 30 基于用户-项目-时间的教育资源热度预测模型所示：



图 30 基于用户-项目-时间的教育资源热度预测模型

UIT-TMP模型按照节目发展趋势的过程主要分为如下两个过程：其一，为节目热度的前期预测；其二，为节目的后期趋势预测。

（1）节目热度前期趋势预测

该过程主要是依据对节目数据“降维”后的用户-项目关系和用户-时间关系，及部分节目-时间关系的分析。

（2）节目热度后期趋势预测

节目后期趋势预测过程，主要是基于前期节目的热度预测结果，在此基础上利用对节目-时间关系分析得到的成果，进行外界因素对节目热度影响的调整，从而得到节目的最终预测结果。

### 基于用户细分的流失率预测算法

伴着诸多因素的影响，用户的流失是一个不可忽视的问题。如果可以建立一种可靠的模型，分析用户的流失预测规律，从而可以在用户还没有流失的情况下，尽早的采取各种营销和挽留措施，使得用户的流失概率将到最低，从而可以降低损失。这里提出一种在现有流失率预测模型的基础上的基于用户细分的流失率预测模型。

这种基于用户细分的流失预测模型包括七大模块：客户数据模块、数据预处理模块、流失率预测模块、用户细分模块、结果反馈模块、流失预测模块以及结果输出模块。它是在现有用户流失预测模型的基础上，增加了用户细分模块和预测结果反馈模块。这种基于用户细分的流失预测模型具有如下几点创新点：首先在数据的预处理阶段，利用属性参量集代替所有的属性集合来进行流失预测，其次将用户细分算法应用到流失预测模型，可以根本不同的细分用户群进行流失预测，最后克服了现有流失预测模型预测结果只考虑单用户流失情况的。

1. 技术创新点

多维度数据构建模型大幅提升分析系统的效率

运用多维度数据构建技术，基于MOLAP多维度数据建模技术，设计实现了满足所有算法和统计分析要求的多Cube自动构建模型。在选择模型时不仅优先考察访问效率，还在尽量减少数据冗余，无论是在数据仓库还是多维数据库，都要在可能的范围内做到数据共用。同时，模型的设计还保证了足够的灵活性，以能应对系统中数据间引用关系不正确或是发生改动的状况。在操作功能方面，则是需要支持在不同维度以及时间等维度上进行便捷的分片钻取操作，为各种算法和业务统计分析提高高效的数据访问接口。

通过分析常见多维数据模型的优劣，选择灵活度较高、数据冗余度教的的雪花模型。这种模型在整体结构与星型模型类似，但是在一个维度上能够有多张表，这样可以更好地满足外键约束较多的的数据库；能够将一个维度上的各个属性拆分到不同的表中，并用外键连接，从而大大提高灵活性，对维度属性从属关系进行修改时的代价相对其他模型会少很多；还能增加表中数据重用的可能性，减少数据仓库中的数据冗余。

而且，还通过合理的任务调度，从底层的数据抽取、清洗和加载，到多维度数据Cube构建，一级任务完成自动激活下一级任务，形成完整的自动化链条。这样，不仅可以大幅提升分析系统的效率，还有效提升了系统智能化程度，便于部署、管理和维护。

可定制的业务分析模型提升系统交互能力和实用性

在实际的业务分析和个性化推荐算法中，不同的应用背景、不同的应用时段以及不同的应用目标，需要对所构建的模型作微调，例如模型参数的优化与配置，新业务分析定制，推荐算法的组合策略的优化与配置。这些都需要系统提供便捷的模型定制功能。

因此，通过对各种业务分析算法已有研究成果的总结分析，针对各种模型中正确率影响重要的可配置的参数做了系统的全面的分析，在实际的应用系统中设计实现了可定制的功能，以在需要的时候针对不同的模型进行分析测试，以选择最优模型。这使得系统的交互能力、实用性和可用性得到了有效的提升。

采用多算法协同方法提升系统业务分析的全面性和准确性

在与用户的交互过程中收集到大量用户的行为特征信息，为可定制的各种业务分析模型和个性化推荐算法提供了足够的数据源泉。通过分析各种业务分析的具体需求，采用多算法协同方式，通过协同技术凸显各算法的优点，降低各算法的不足，以提升系统业务分析能力的全面型和准确性。

例如，在个性化推荐中，通过对用户行为特征信息的分析与挖掘，针对不同用户的兴趣倾向和特征知识，按特定用户的个性化需求定制媒体内容和推荐方法，提高媒体内容的定向推送能力，加入了用户的可定制性，使得用户行为兴趣模型更具有人性化、以及可扩展性，改善了用户对数字媒体内容和服务商的满意度。

1. 技术展望

本课题的研究实施中，因时间有限等原因，还有一些方向未能进行深入的研究，主要归纳为以下几个方面，这些内容是后续工作中可以继续挖掘的重点。

要将大数据业务分析更广泛地应用于数字出版行业，可以从以下几个方面去探索：

**数据获取**

可以获得哪些数据。大数据时代的特点是数据来源广泛，可以是业内发布的数据，也可以是来自搜索引擎、社交媒体等的数据。有些数据看似关联不强（比如社交媒体数据），但往往能从中挖掘到用户的潜在需求。

**数据挖掘**

从数据中，我们想挖掘什么信息。各种点击率预测的模型，挖掘了搜索量等数据与点击率收入的关联；个性化推荐，则挖掘了观众对不同电影的偏好，以及其他的行为特点。挖掘什么信息，一方面取决于我们有哪些数据，另一方面也取决于什么样的信息可能有助于哪些业务分析。

**行业经验**

有什么行业经验是可以结合的。单纯地数据分析，可能会找到很多规律，但这些规律未必是有实际价值的。只有当数据结合行业经验，才更容易形成精准的行业模型，从而产生巨大的价值。

1. [] 顾小清, 郑隆威与简菁, 获取教育大数据:基于xAPI规范对学习经历数据的获取与共享. 现代远程教育研究, 2014(05): 第13-23页. [↑](#footnote-ref-1)
2. Ecosystem TinCan API[EB/OL],http://tincanapi.com/ecosystem/, 2014-10-10 [↑](#footnote-ref-2)