在线学习行为数据的可视分析方法研究

Visual Analysis Methods on Online Learning Process Data

博 士 生　：贺欢
指导老师　：郑庆华 教授

学科专业　：计算机科学与技术
答辩时间　：2019年7月29日

西安交通大學
XI'AN JIAOTONG UNIVERSITY

内容概要

1. 研究背景与内容
2. 研究内容1：面向高维数据的学习参与度可视分析
3. 研究内容2：面向时序数据的学习时间管理可视分析
4. 研究内容3：面向大规模多属性数据的视频利用情况可视分析
5. 结论与展望

研究背景与内容

- 研究背景：在线学习行为数据

学习者使用教学功能进行在线学习的过程中产生的学习行为日志与相关信息。

随着专业、课程资源与教学功能的多样化，以及日益增长的学生数量，数据规模猛增。

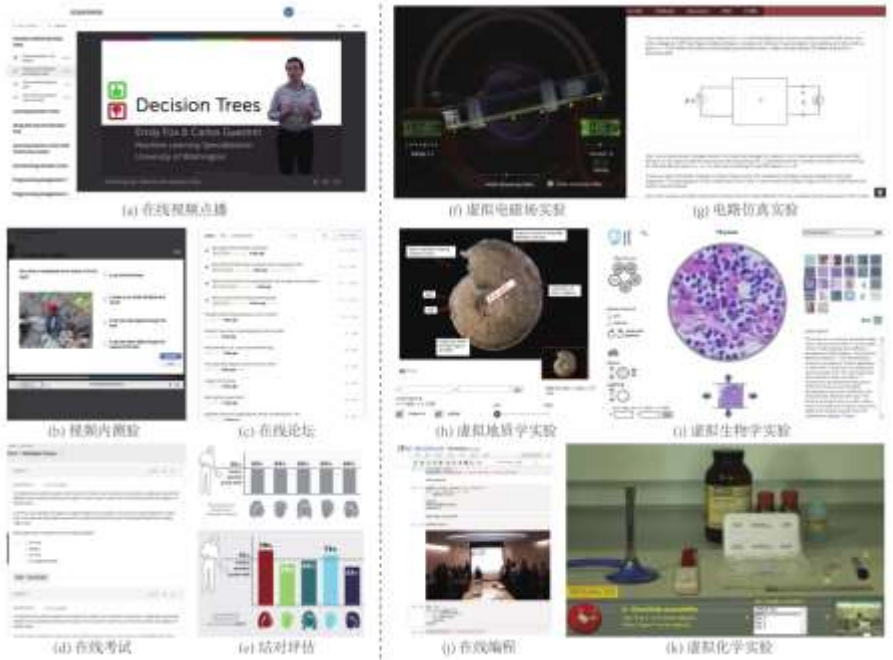
数据价值

- 学习者画像
- 成绩预测
- 智能导学
- 个性化学习
- 资源推荐

…

麻省理工学院 2012 年
电路与电子学
一学期15万学习者
2.3亿条学习行为数据

Nature 专刊讨论
在线学习行为数据的分析



多样化的在线教学过程

研究背景与内容

- 研究目的：旨在挖掘和呈现在线学习行为数据中潜在的特征与模式，
  探索学习参与度、时间管理特征与视频资源的利用情况。

- 在线学习行为数据的特点
  - 特征维度高
    - 多样化的数据来源与丰富的数据格式
    - 刻画在线学习行为特征的指标数量多
  - 时间周期长
    - 面向专项技能和专业学历的长周期多课程在线学习
    - 长期学习过程中的学习时间安排对进度
  - 大规模多属性相互关联
    - 学生的所在地区、入学机构、专业、课程表
    - 教学资源的课程编排、学科专业、难度类型

研究背景与内容

- 在线学习行为数据特点：**特征维度高**　学习周期长　大规模多属性

  教学多样性

  – 教学活动：课程视频、在线论坛、同伴互评、在线作业、测验考试、虚拟实验、阅读材料、资料检索等

  – 交互行为：播放、暂停、发帖、回复、关注、提交、转发、下载等

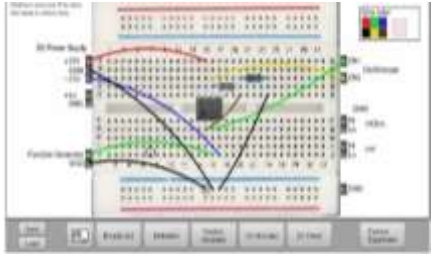
  – 评价指标：功能交互、资源利用、考核评估等

课程视频


在线讨论


阅读材料


测验考试




虚拟实验

研究难点1：如何刻画和分析高维度的在线学习行为数据中的学习过程特征？

研究背景与内容

- 在线学习行为数据特点：　特征维度高　**学习周期长**　大规模多属性

长周期的学习

  - 课程学习：
    - 单课程：2-18周
    - 专项课程：2-6个月
  - 学历教育：
    - 在线硕士教育：1-2年
    - 网络远程教育：2.5-5年

edX 机器学习课程 12周

Coursera 深度学习专项课程 3 个月

Coursera在线硕士项目12-40个月

网络远程教育项目2.5-5年

研究难点2：如何刻画和展示长时间的长期在线学习过程中时间特征？

研究背景与内容

- 在线学习行为数据特点：　　特征维度高　　学习周期长　　**大规模多属性**

大规模

- 随着多媒体技术的进步
  海量课程视频被创建以满足
  教学需求
- 网络技术的发展使世界各地的
  学生都能够参与学习

Coursera

edX

多属性

- 视频：专业、课程、类型等
- 学生：背景、生源地、
  学习环境、入学时间、
  学习计划等

网络远程教育

Udacity

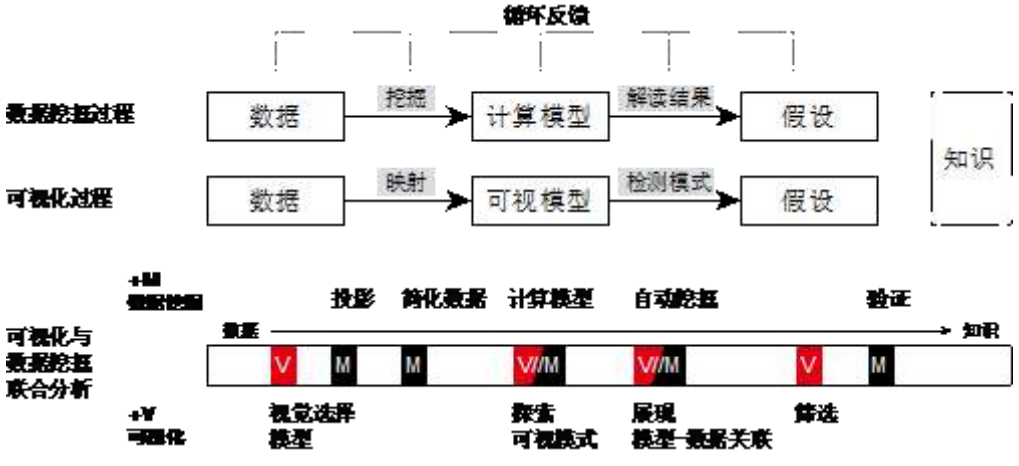
iMOOC

研究难点3：如何分析大规模多属性的不同属性、不同层面的视频利用模式?

研究背景与内容

- 研究方法：综合运用可视化技术、交互技术以及数据挖掘相关技术，以可视分析为主要手段探索在线学习行为数据

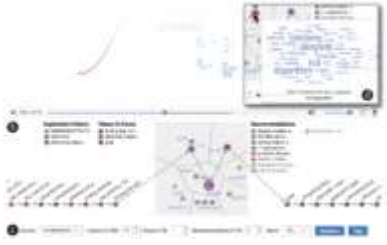
- 可视分析：交互视觉界面支持的分析推理科学 [Thomas, 2006]
  - 计算机科学、信息可视化、认知科学、图形学、社会科学等结合，提供交互式界面辅助用户，直观呈现分析过程与结果
  - 核心流程[Chen, 2013]：
    - 数据管理：汇聚数据
    - 数据映射视觉编码：可视化呈现
    - 交互探索分析：人类知识经验参与



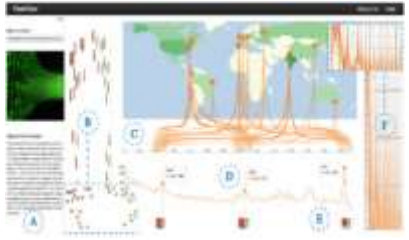
可视化过程与数据挖掘过程 [Bertini , 2009]

研究背景与内容

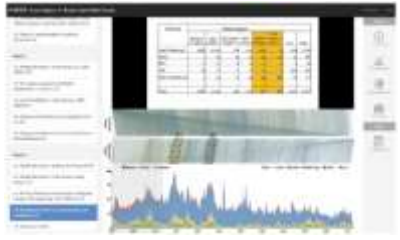
- 面向在线学习行为数据的可视分析研究
  - 视频观看行为
    - 视频内容推荐
    - 视频热点行为
  - 论坛交流行为
    - 学生群组
    - 活动分析
    - 话题演变
  - 其他教学活动
    - 题目推荐
    - 学习过程
    - 退课模式
  - 不足：
    - 只针对特定学习活动
    - 面向短期学习过程
    - 关联属性单一



MOOCex [Zhao, 2018]



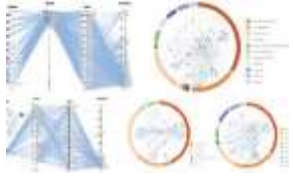
PeakVizor [Chen, 2016]



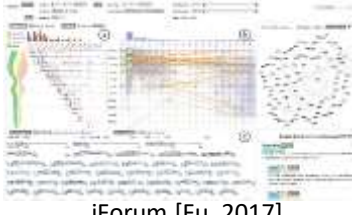
VisMOOC [Shi, 2015]

视频观看行为的可视分析

VisForum [Fu, 2018]



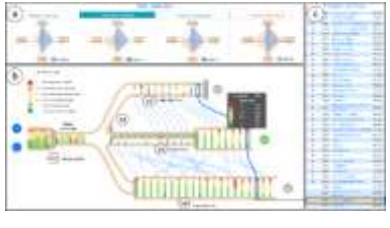
NetworkSeer [Wu, 2018]



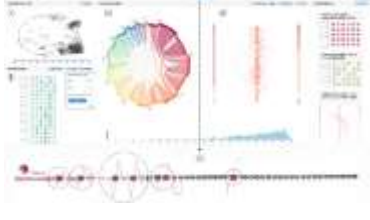
iForum [Fu, 2017]

ToPIN [Sung, 2017]

讨论行为的可视分析



PeerLens [Xia, 2019]



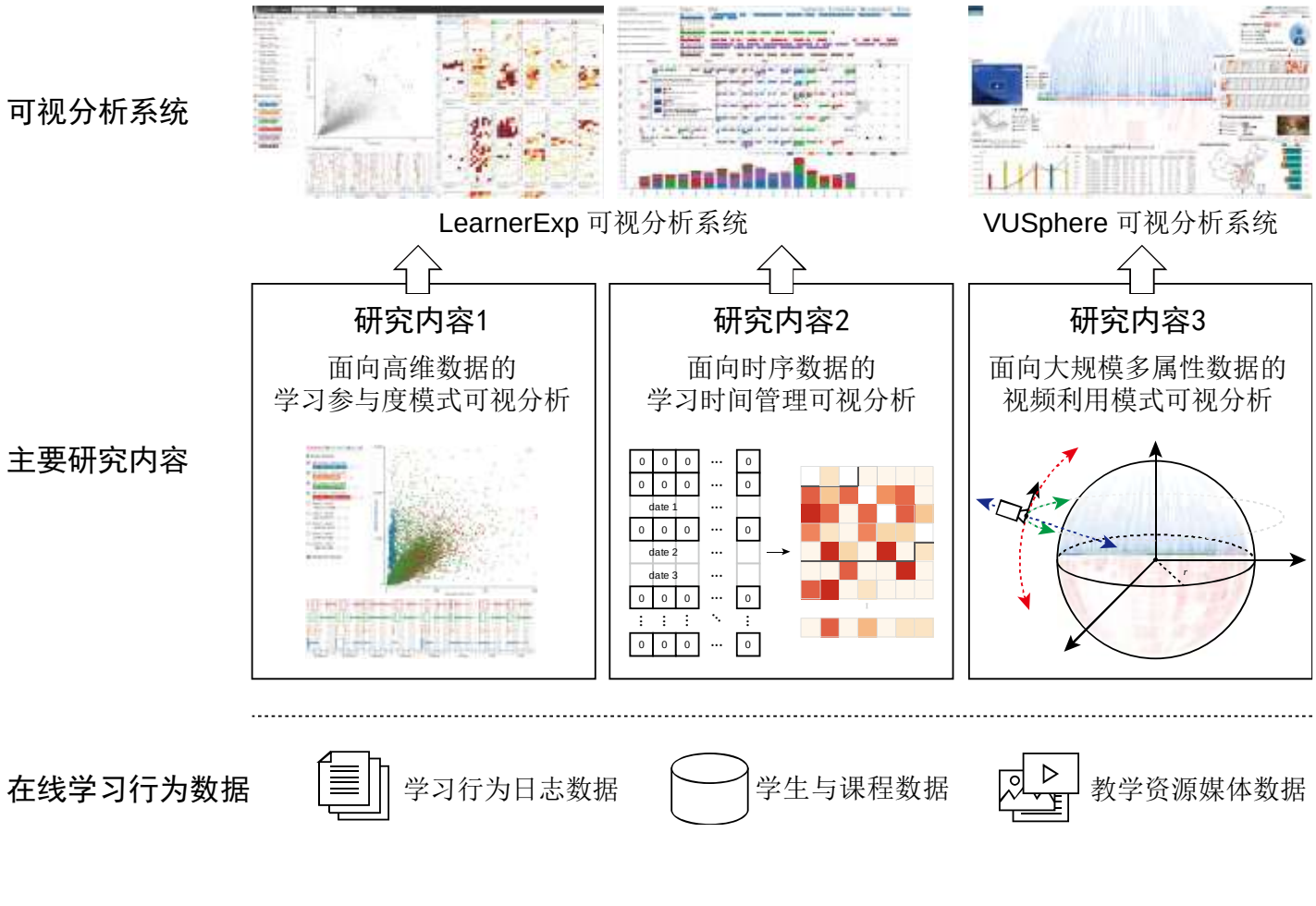
ViSeq [Chen, 2018]

DropoutSeer [Chen, 2016]

其他教学活动分析

研究背景与内容

- 研究框架



内容概要

1. 研究背景与内容
2. **研究内容1：面向高维数据的学习参与度可视分析**
3. 研究内容2：面向时序数据的学习时间管理可视分析
4. 研究内容3：面向大规模多属性数据的视频利用情况可视分析
5. 结论与展望

研究内容1：面向高维数据的学习参与度可视分析

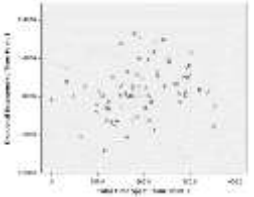
- 研究目的：旨在分析与在线学习行为数据中多样化的学习参与度分布，探索学生参与度模式及其影响因素

- 研究难点
  - 数据来源与格式异构多样
  - 高维数据模式难以探索和解释

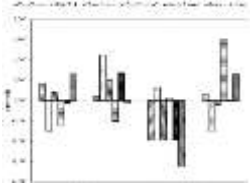
研究内容1：面向高维数据的学习参与度可视分析

- 研究现状
  - 在线学习行为数据管理方法
    - SCORM, xAPI, DataShop, 以及Moodle, edX, Coursera 等

不足：字段繁多、且设计复杂



[Henrie, 2018]

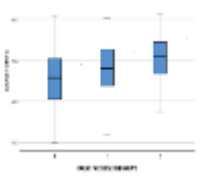


[Cerezo, 2016]

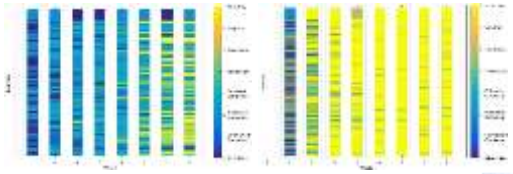
  - 高维度学习参与度分析方法：
    - 基于机器学习方法挖掘学生参与度的模式
      - 关联规则挖掘 [Howard, 2016]，聚类方法 [Cerezo, 2016; Li & Tsai, 2017; Kahan 2017]
    - 基于统计检验方法分析不同学生群组的参与度的影响因素
      - 相关性分析 [Henrie, 2018]
      - Kruskal-Wallis，Mann-Whitney U检验等 [Joksimović, 2015]

不足：学生参与度模式不直观，
不同群组的参与度模式特征难以理解

[Ilves, 2018]



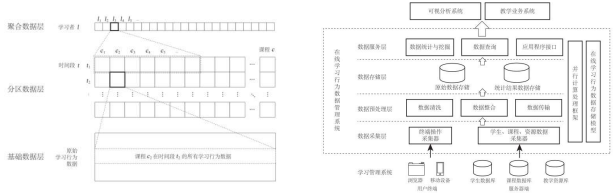
[Knobloch, 2018]

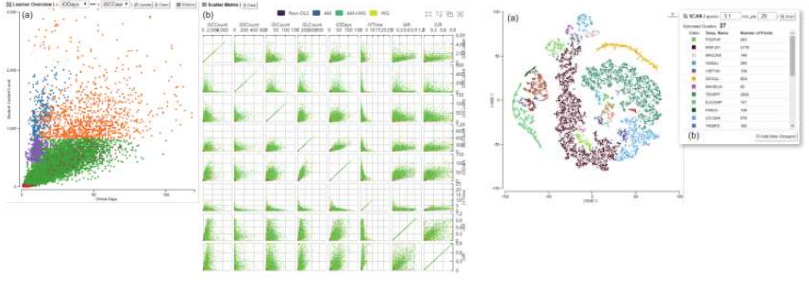


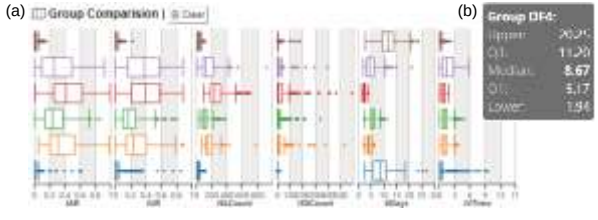
[Sunar, 2017]

研究内容1：面向高维数据的学习参与度可视分析

- 研究思路：结合降维技术与可视化交互呈现高维度的学习参与度模式
  - 在线学习行为数据管理
    - 层次化的存储结构与原始在学习行为数据存储方法
    - 在线学习行为数据管理系统框架
  - 可视化与交互设计
    - 基于散点图矩阵的学习参与度分布图
    - 基于t-SNE的学习参与度降维
    - 学生群组创建工具与学习参与度对比工具

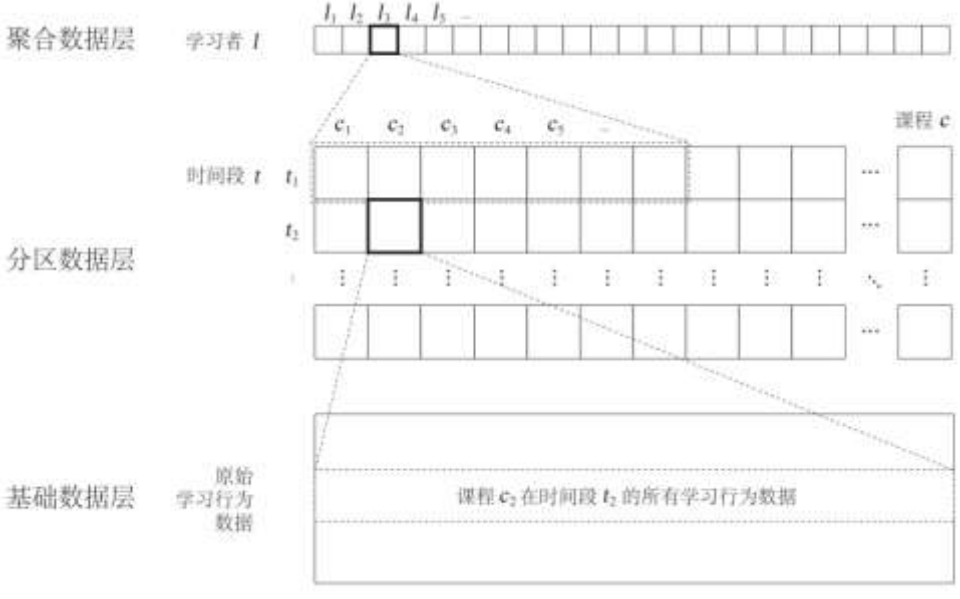




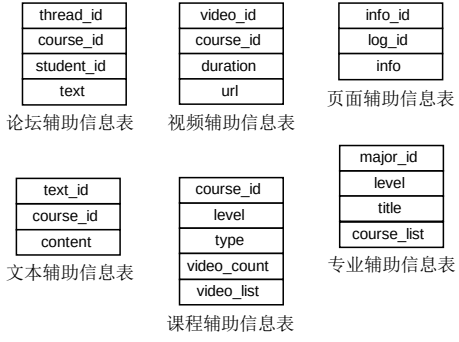


研究内容1：面向高维数据的学习参与度可视分析

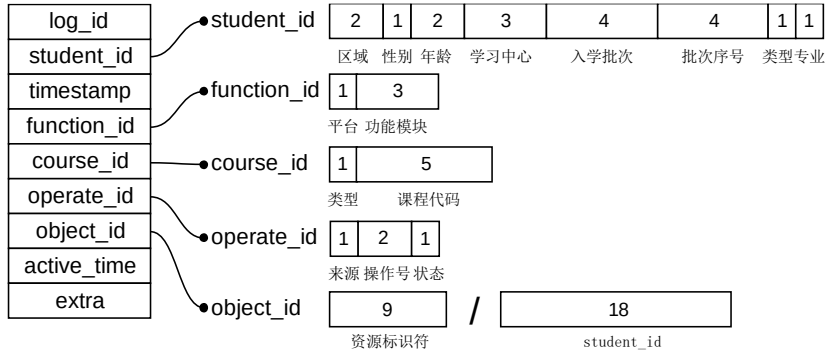
- 在线学习行为数据管理
  - 层次化的存储结构
  - 原始记录表



层次化的存储结构



辅助信息表

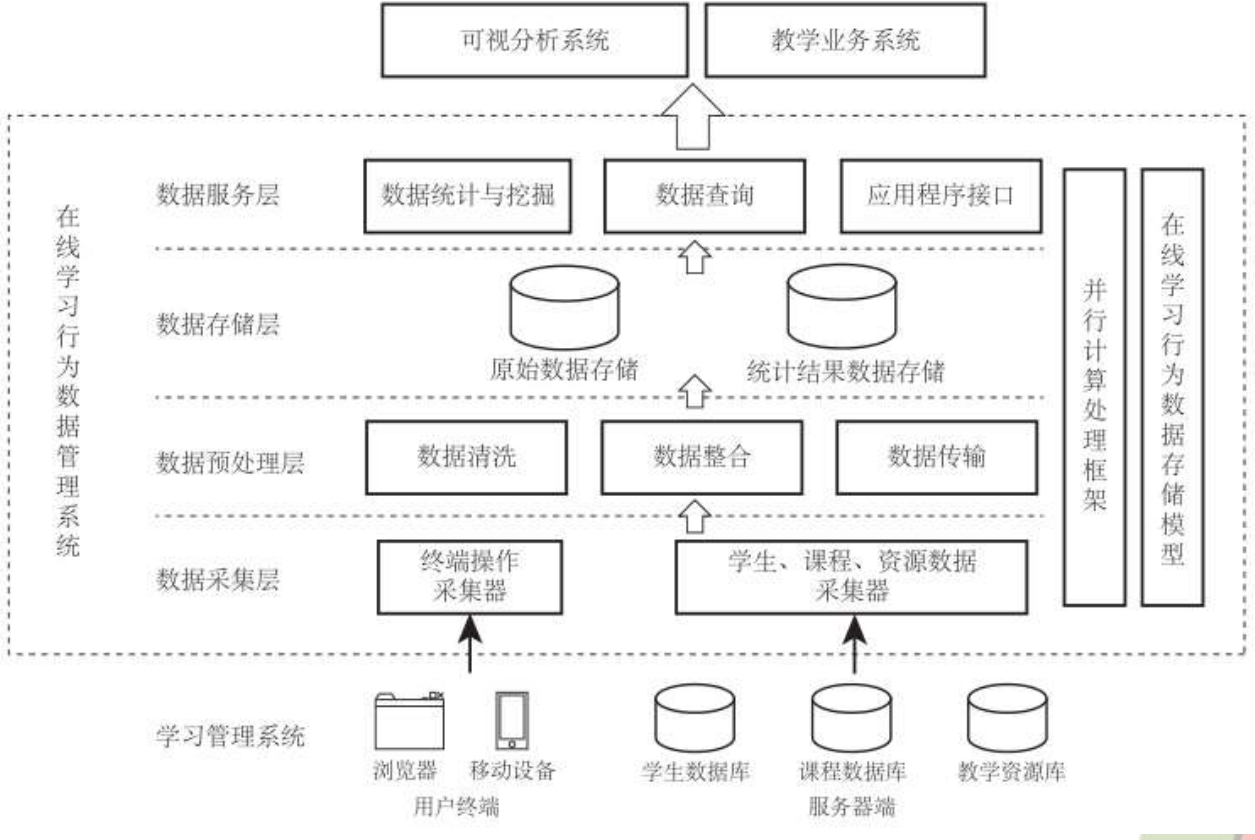


学生操作记录表　　操作记录字段编码

基础数据层的原始记录表

研究内容1：面向高维数据的学习参与度可视分析

- 在线学习行为数据管理：在线学习行为数据管理系统框架
  - 数据采集层
    - 终端采集
    - 服务端采集
  - 数据预处理层
    - 数据清洗
    - 数据整合
    - 数据传输
  - 数据存储层
    - 原始数据存储
    - 统计结果存储
  - 数据服务层
    - 统计
    - 查询
    - API



研究内容1：面向高维数据的学习参与度可视分析

- 在线学习的学生参与度 (Student Engagement) [Fredricks, 2004]

  行为参与度(Behavioral) + 情感参与度(Emotional) + 认知参与度 (Cognitive)

  - 衡量学生参与度的指标
    - 基于交互对象：学生与课程内容、学生、教师、系统的交互数量与时长 [Joksimović, 2015]
    - 基于任务活动：考试、交作业、查询资料、写作的时间与表现 [Bote-Lorenzo, 2017]
    - 基于学习资源：视频、讲义、作业、题目、讨论帖等 [Li & Tsai, 2017]

- 在线学习行为抽象
  - 学习者（Learner）
  - 操作（Operator）
  - 时间（Time）
  - 资源（Resource）
  - 属性集合（Attributes）

在线学习行为五元组

$$b_l = \langle l, o, r, t, \Omega \rangle$$

学生/学习者 l，操作 o，时间 t，资源 t，属性集合 Ω

- 在线学习过程 $P_l = [(l, o_1, r_1, t_1, \Omega_1), (l, o_1, r_1, t_1, \Omega_1), \dots (l, o_n, r_n, t_n, \Omega_n)] \quad t_1 < t_2 < \cdots < t_n$

- 学习参与度指标向量 $\boldsymbol{e}_l = [e_{l,1}, e_{l,2}, \cdots, e_{l,k}] \quad e_{l,i} = f_i(P_l)$

研究内容1：面向高维数据的学习参与度可视分析

- 学习参与度指标向量
  - 学习参与度基础指标：捕捉局部细粒度学习过程特征
    - 基于交互操作数量的学生参与度指标

$$e_{o^j} = |\{b_{l,i} | b_{l,i}.o = o^j \wedge b_{l,i} \in P_l\}|$$

    - 基于资源访问时长的学生参与度指标

$$e_{o^j} = \sum_{b_{l,i}.r = r^k \wedge b_{l,i} \in P_l} (b_{l,i+1}.t - b_{l,i}.t)$$

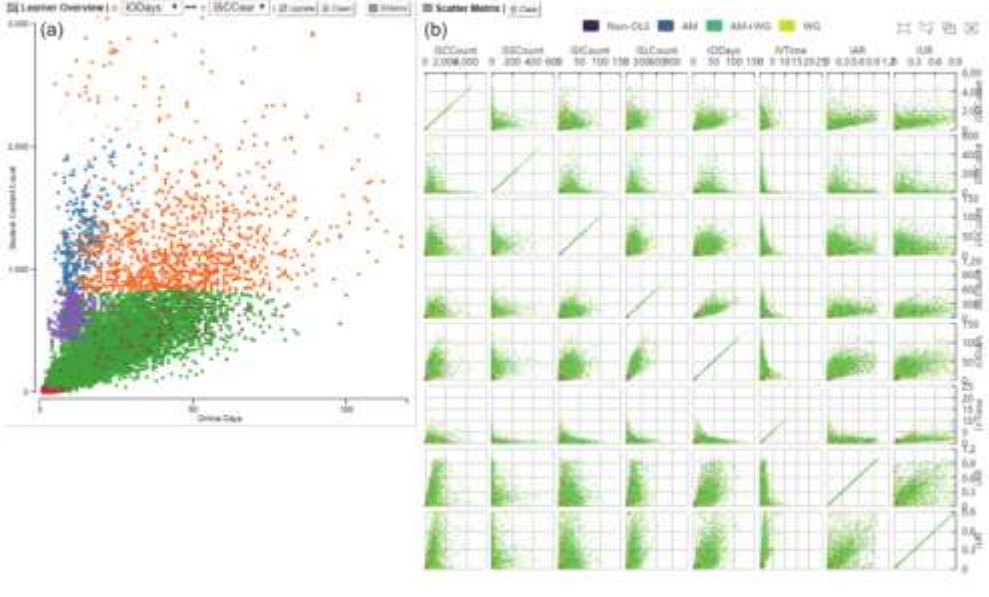
  - 学习参与度概况指标：解释学生学习参与度的整体表现

学习参与度类别	指标	简述
交互数量	学生-内容	学生对教学内容的交互操作次数，如播放、暂停等
	学生-学生	学生之间的看帖、回复、消息等操作数量
	学生-教师	学生与教师之间交流的次数
	学生-系统	学生使用学习管理系统其他功能的次数
学习时间	在线天数	实际在线学习的日期数量
	单日时长	一天内在线的总时长均值
资源利用	阅读数量	阅读不重复文本材料的总数量
	观看数量	观看不重复课程视频的总数量
	观看时长	观看课程视频的总时长

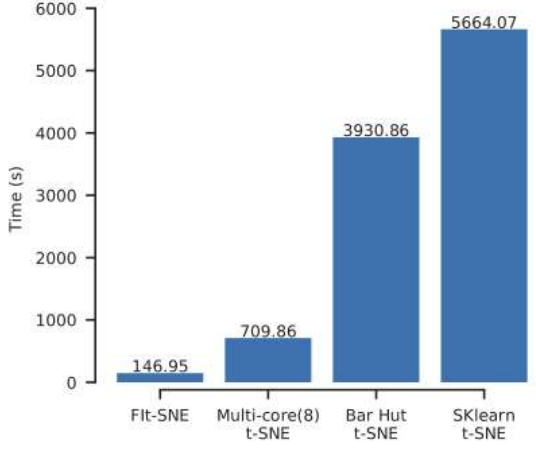
研究内容1：面向高维数据的学习参与度可视分析

- 可视化设计需求
  - R.1 体现学生学习参与度在各维度的分布模式
  - R.2 展示不同群组之间的学习参与度分布差异

- 可视化设计：学习参与度分布图
  - 基于散点图矩阵的高维数据分布展示



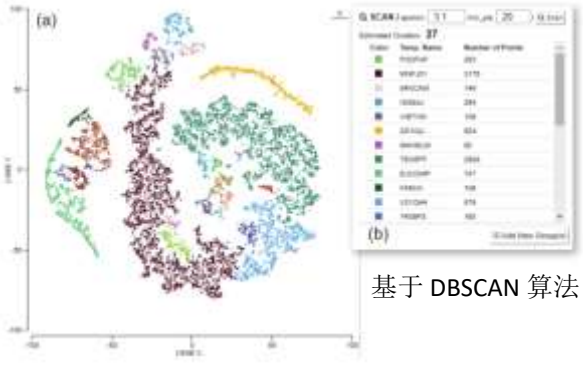
  - 基于 t-SNE 的学习参与度向量降维
    - FIt-SNE 实现版本的降维



各 t-SNE 实现版本的运行时间对比

研究内容1：面向高维数据的学习参与度可视分析

- 交互设计：自由探索不同属性的学生群组以及各群组的学习参与度分布（R.2）
  - 学生群组创建工具



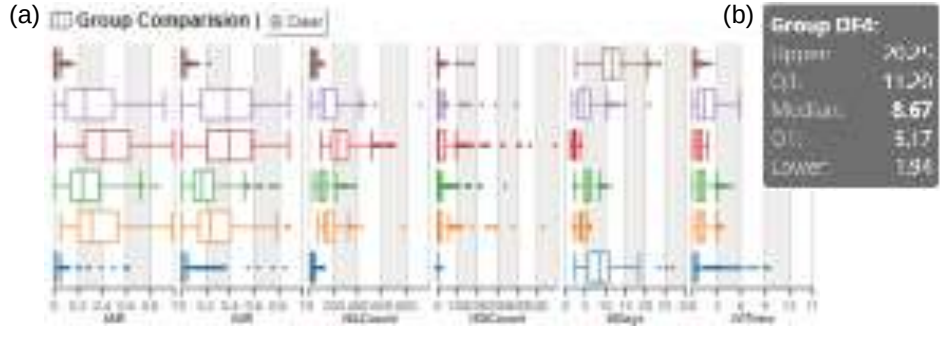
基于 DBSCAN 算法

条件选择器

SQL选择器

Student ID 选择器

  - 学生群组学习参与度对比工具
    - 群组→ 颜色编码
    - 学习参与度盒图
    - 交互详情



学生群组学习参与度对比工具

研究内容1：面向高维数据的学习参与度可视分析

- 案例分析：使用在线学习者支持服务对学生参与度的影响
  - 在线学习者支持服务 (Online Learner Support)：
    - 智能客服(Assistance Messager, AM) 和 导学材料(Weekly Guidance, WG)

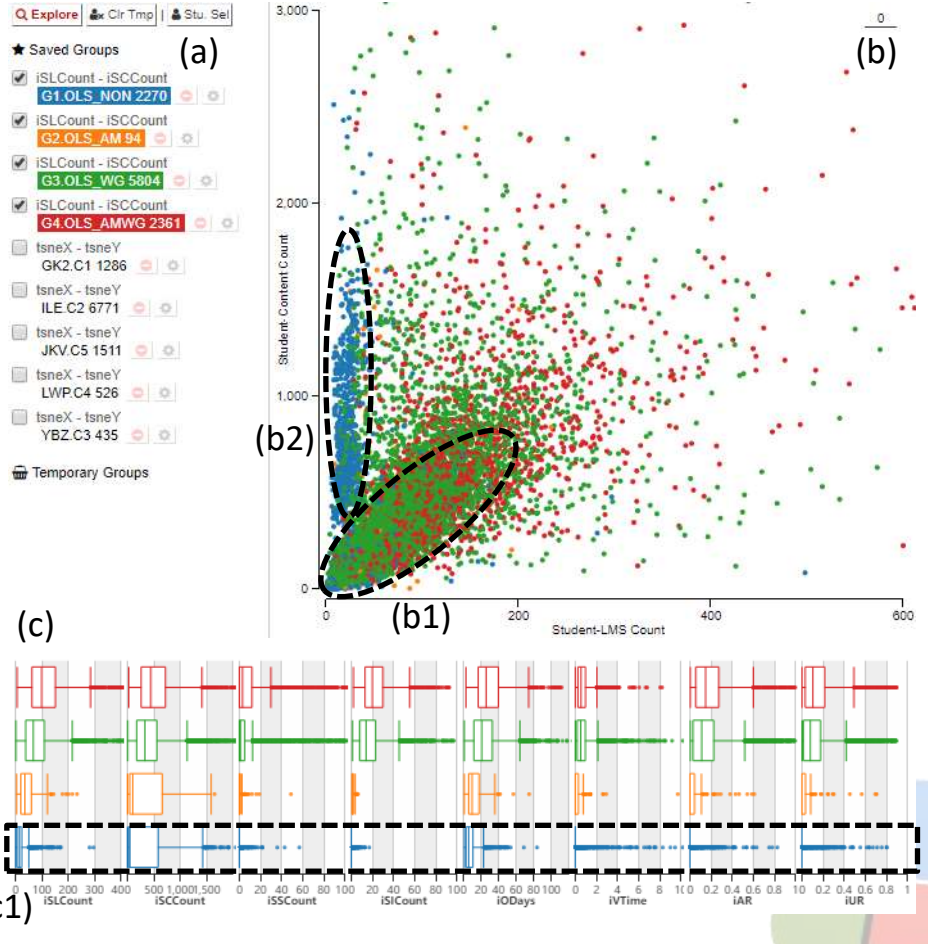
智能客服 (Assistance Messager, AM)

导学材料 (Weekly Guidance, WG)

  - 数据集
    - 本校网络远程教育学院2015年春季和秋季批次第一学期的学习日志
    - 学生：10,529
    - 日志：14,942,964

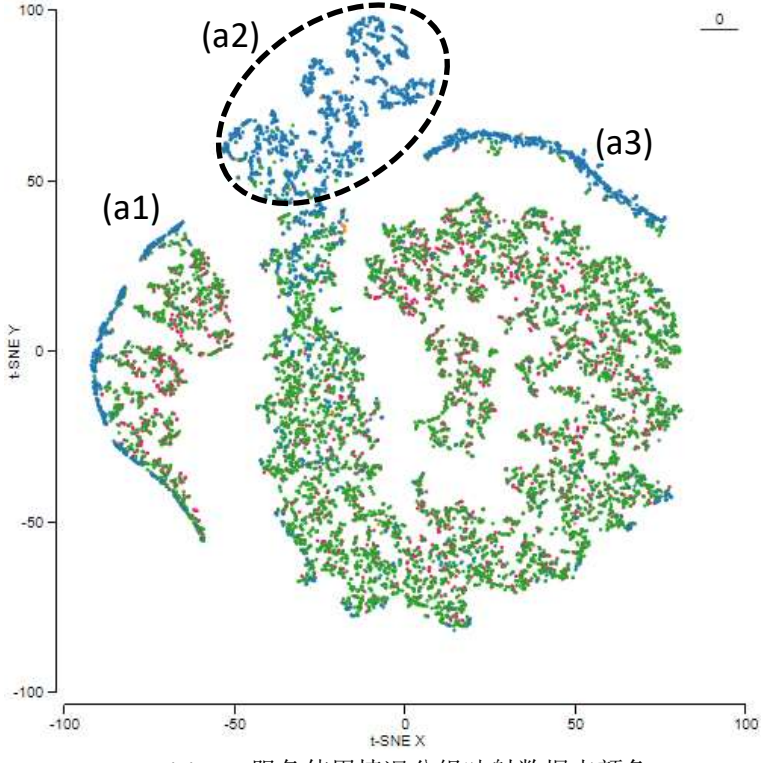
研究内容1：面向高维数据的学习参与度可视分析

- 案例分析1：学习参与度的分组分析
  - 按 OLS 服务使用情况分组
    - Non-OLS Group: 2,270
    - AM Group: 94
    - WG Group: 5,804
    - AM + WG Group: 2,361
  - 发现
    - 不使用 OLS 服务的学生整体学习参与度较低 (c1)
    - 学习参与度分布模式与使用 OLS 服务相关

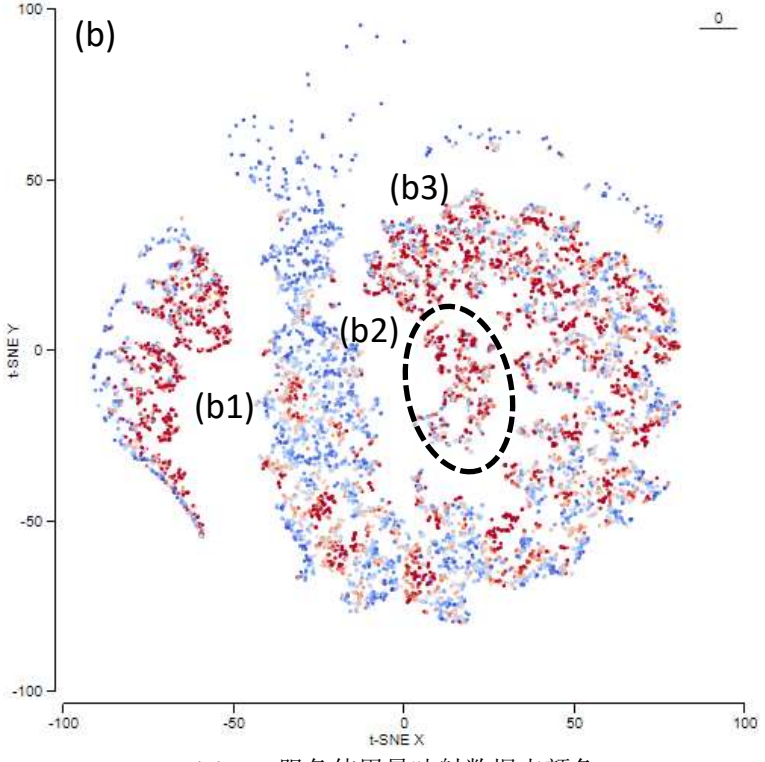


研究内容1：面向高维数据的学习参与度可视分析

- 案例分析2：使用 OLS 服务与学习参与度模式的关系



(a) OLS 服务使用情况分组映射数据点颜色



(b) OLS 服务使用量映射数据点颜色

学生参与度呈现不同模式，且与使用OLS服务相关

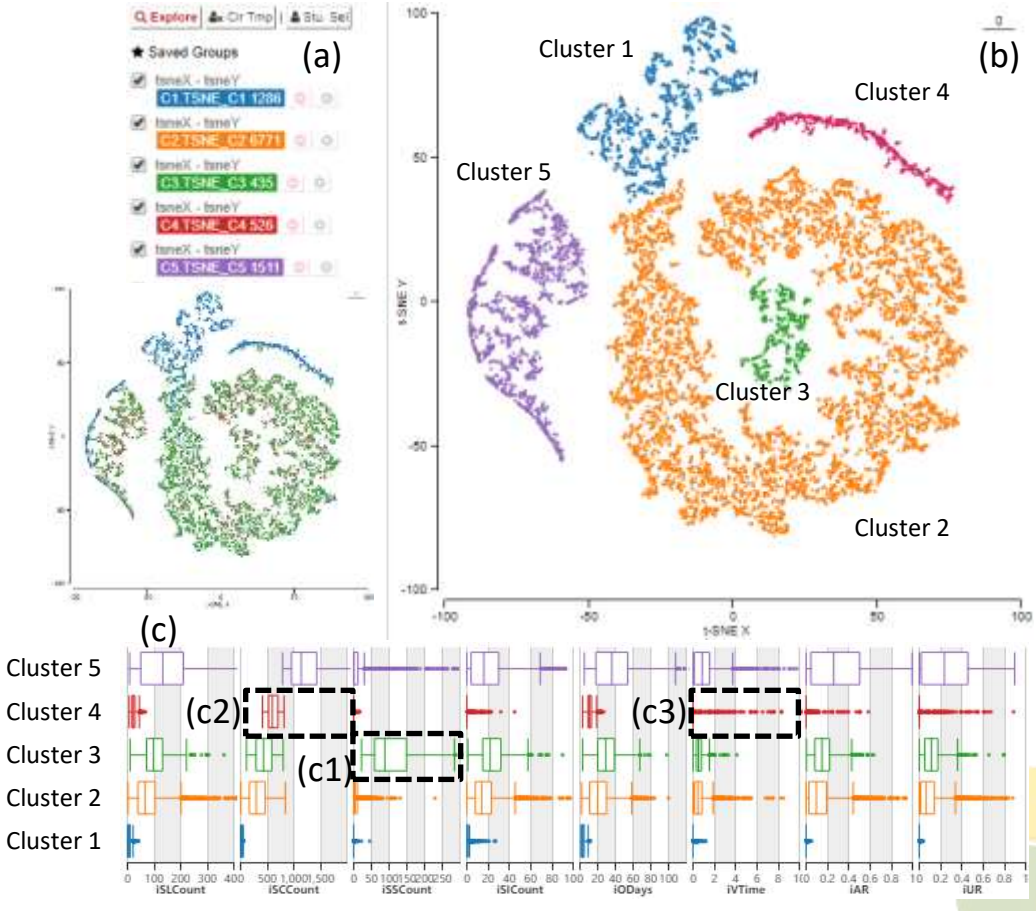
OLS服务使用量与学生参与度呈现正相关

研究内容1：面向高维数据的学习参与度可视分析

- 案例分析2：使用 OLS 服务与学习参与度模式的关系

集群1：低学习参与度
集群2：中等参与度
集群3：高论坛参与
集群4：高内容参与，低视频观看
集群5：高学习参与度

Non-OLS 主要分布在
集群1，集群4和集群5边缘
表现出3种模式



研究内容1：面向高维数据的学习参与度可视分析

- 小结
  - 设计了一个层次化的数据存储模型及数据管理框架
    并基于该框架实现了在线学习行为数据的采集、清洗、抽象和分层存储
  - 提出一种基于散点图矩阵与 t-SNE 数据降维技术的的高维度学习参与度可视化方法

- 成果
  - How Learner Support Services Affect Student Engagement in Online Learning Environment [J]. **IEEE Access**, 2019 (SCI:000466757700001)
  - Measuring Student's Utilization of Video Resources and its Effect on Academic Performance [C]. IEEE 18th International Conference on Advanced Learning Technologies (**ICALT 2018**, EI:20183605769536)
  - Big Log Analysis for e-Learning Ecosystem [C]. IEEE 11th International Conference on e-Business Engineering (**ICEBE 2014**, EI:20150300423652)

内容概要

1. 研究背景与内容
2. 研究内容1：面向高维数据的学习参与度可视分析
3. **研究内容2：面向时序数据的学习时间管理可视分析**
4. 研究内容3：面向大规模多属性数据的视频利用情况可视分析
5. 结论与展望

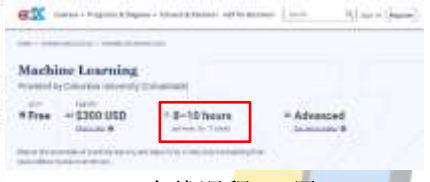
研究内容2：面向时序数据的学习时间管理可视分析

- 研究目的：从时间维度分析在线学习过程的时间管理模式，
  解决学习过程时序特征建模以及时间管理解释的问题。
- 研究现状：时间管理(Time Management)
  - 时间管理模式
    - 自律学习 (Self-regulated Learning) [Zhao, 2014; Milikic, 2018]
    - 拖延 (Procrastination) [You, 2015; Yamada, 2016; Park, 2018]
  - 时间管理的影响
    - 学生参与度 (Student Engagement) [Nawrot, 2014]
    - 学习成果 (Learning Outcome) [You, 2016; Kizilcec, 2017]
    - 退学 (Dropout) [You, 2015]
- 研究难点
  - 长时间在线学习过程的时序特征难以刻画
  - 长时间在线学习过程的时间管理风格难以呈现

Coursera 深度学习专项课程 3 个月

Coursera在线硕士项目12个月以上

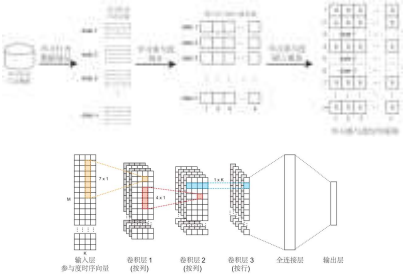

网络远程教育项目2.5-5年



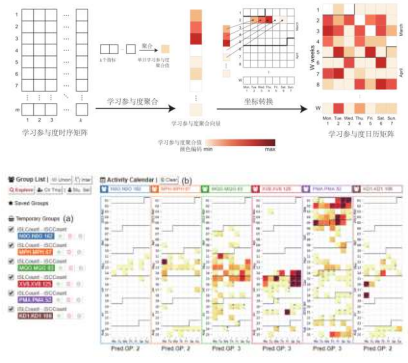
edX在线课程 12 周

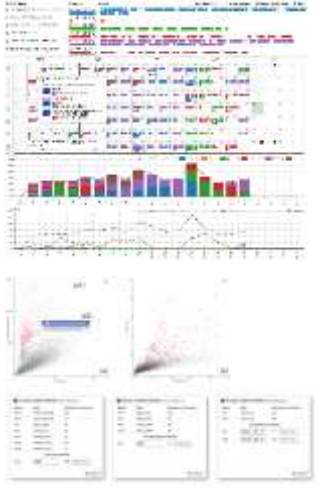
研究内容2：面向时序数据的学习时间管理可视分析

- 研究思路：通过可视化方法呈现时序数据的学习时间管理特点
  - 学习参与度时序特征建模
    - 学习参与度时序矩阵：刻画在线学习行为数据中包含的时间特征
    - 基于卷积神经网络的成绩预测模型：提取时间维度信息



  - 可视化与交互设计
    - 基于日历坐标系的学习参与度日历矩阵
    - 基于日历图的学习时间分布图
    - 交互式学生群组创建工具与学习时间分布图对比工具
    - 多课程视频观看可视化





研究内容2：面向时序数据的学习时间管理可视分析

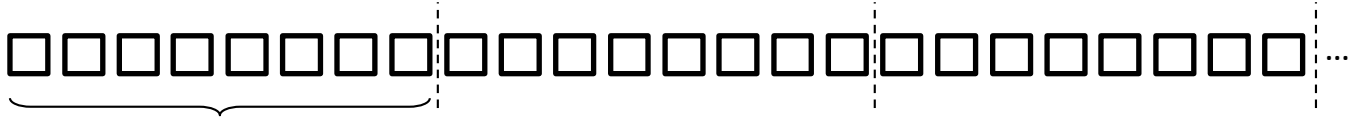
- 可视分析任务
  - T.1 分析学生在给定时间段内的学习时间安排
  - T.2 分析学生对多个课程的学习时间安排
  - T.3 分析不同时间管理风格与学习参与度的关系

- 设计需求
  - R.1 呈现学习参与度随时间的变化情况
  - R.2 展现多课程学习参与度随时间的变化情况
  - R.3 学生群体对比分析
  - R.4 交互探索

研究内容2：面向时序数据的学习时间管理可视分析

- 学习参与度时序矩阵：刻画在线学习行为数据中包含的时间特征

学习过程P_l



单位时间段长度 d

子学习过程ΔP_t^l $\quad \Delta P_t^l = \{(l, o_1, t_1, r_1, \Omega_1), (l, o_2, t_2, r_2, \Omega_2), \cdots, (l, o_j, t_j, r_j, \Omega_j)\}, t_0 \le t_j < t_0 + d$

$$e_{t,i}^l = f_i(\Delta P_t^l)$$

子学习参与度向量$\Delta \boldsymbol{e}_t^l$ $\quad \Delta \boldsymbol{e}_t^l = [e_{t,1}^l, e_{t,2}^l, \cdots, e_{t,k}^l]$

学习参与度时序矩阵$\boldsymbol{E}_l$

$$\boldsymbol{E}_l = \begin{bmatrix} e_{1,1}^l & e_{1,2}^l & \cdots & e_{1,k}^l \\ e_{2,1}^l & e_{2,2}^l & \cdots & e_{2,k}^l \\ \vdots & \vdots & \ddots & \vdots \\ e_{m,1}^l & e_{m,2}^l & \cdots & e_{m,k}^l \end{bmatrix}$$

		SCCount	SSCount	SICount	SLCount	STime	VTime	AR	UR
学期第一天 →	2018.02.18	32	0	1	11	44	40	.05	.04
	2018.02.19	0	0	0	0	0	0	.0	.0
	2018.02.20	0	0	0	0	0	0	.0	.0
	2018.02.21	0	0	0	0	0	0	.0	.0
	2018.02.22	11	2	0	5	21	12	.05	.01
	2018.02.23	0	0	0	0	0	0	.0	.0
	2018.02.24	1	17	0	4	12	10	.05	.01
	⋮								
	2018.06.30	0	0	0	0	0	0	.0	.0
学期最后一天 →	2018.07.01	91	32	2	13	244	232	.32	.34

SCCount	学生–内容交互数量
SSCount	学生–学生交互数量
SICount	学生–教师交互数量
SLCount	学生–系统交互数量
STime	在线时长(分钟)
VTime	观看视频时长(分钟)
AR	课程视频到课率
UR	课程视频利用率

单位时间段长度 d = 24小时

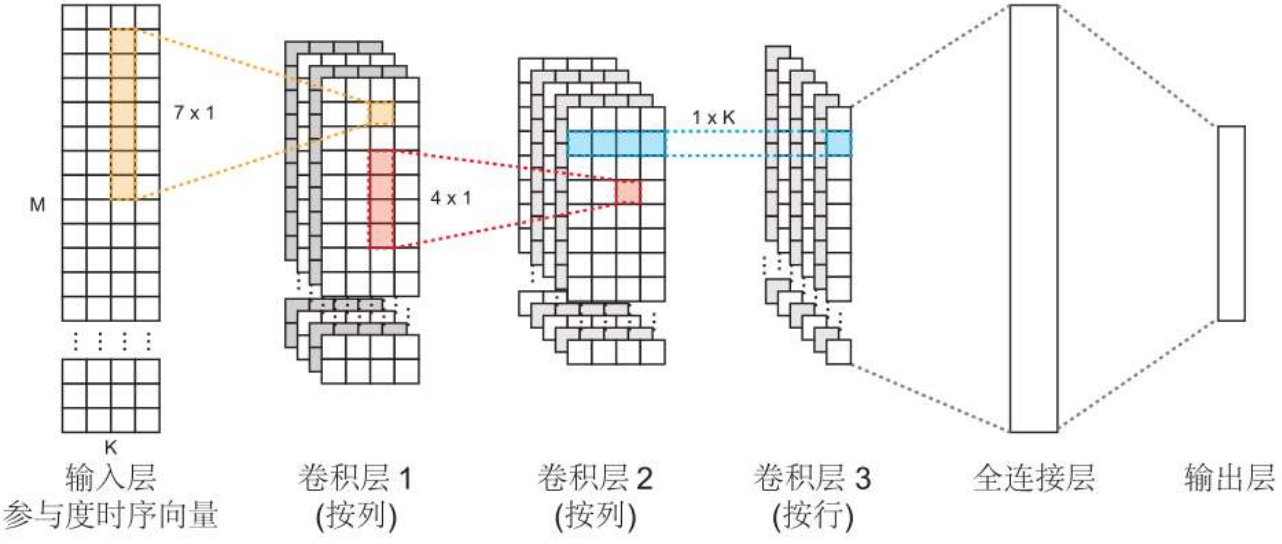
研究内容2：面向时序数据的学习时间管理可视分析

- 基于卷积神经网络的成绩预测模型
  - 考试成绩

    分数转换为类别

    0, 1, 2, 3, 4

    不及格，及格，中，良，优
  - 考试成绩预测模型
    - 基于卷积神经网络
    - 3 卷积层 + 1 全连接
    - 输出各类成绩概率
  - 数据集样本
    - 2015年学生 10,529
    - 英语课成绩及相关日志
    - 准确率 72.2%



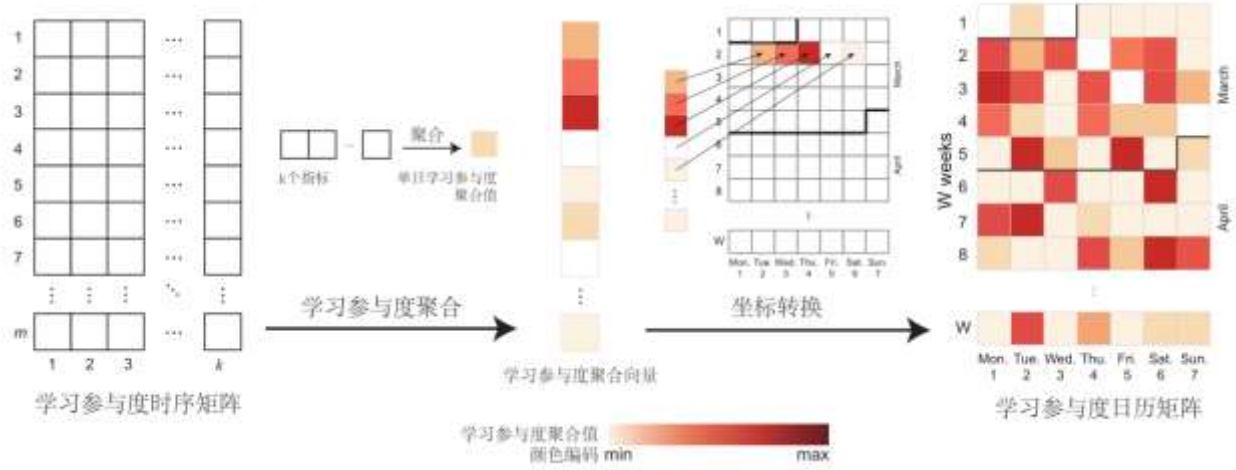
研究内容2：面向时序数据的学习时间管理可视分析

- 可视化设计：基于日历坐标系的学习参与度日历矩阵
  - 学习参与度聚合：$D \times K$ 维降至 $D \times 1$ 维
    - 将日活动聚合为单个数值反映单日总活动量

$$v_t^l = \alpha_1 g_1(e_{t,1}^l) + \alpha_2 g_2(e_{t,2}^l) + \cdots + \alpha_k g_k(e_{t,k}^l), \sum_{i=1}^{k} \alpha_i = 1$$

  - 坐标转换： $D \times 1$ 维转换至 $W \times 7$ 维
    - 按日期将聚合指标嵌入到日历坐标系矩阵

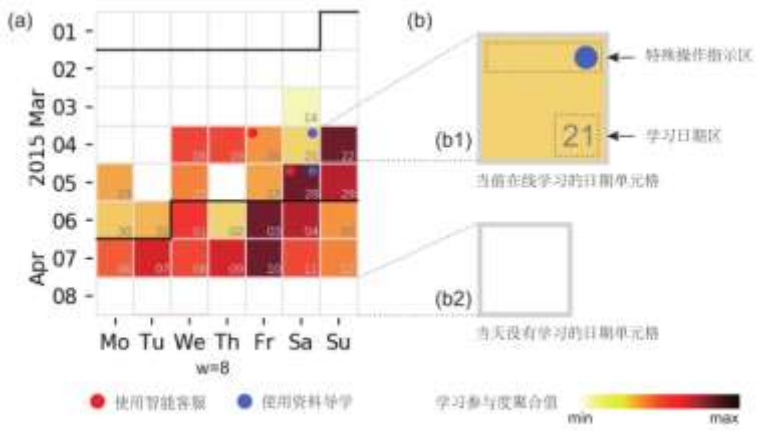
$$row = \lfloor \frac{i}{7} \rfloor, col = \begin{cases} (\lfloor \frac{23m}{9} \rfloor + (d+y-1) + 4 + \lfloor \frac{y}{4} \rfloor - \lfloor \frac{y}{100} \rfloor + \lfloor \frac{y}{400} \rfloor) \bmod 7 & \text{if } m < 3 \\ (\lfloor \frac{23m}{9} \rfloor + (d+y-2) + 4 + \lfloor \frac{y}{4} \rfloor - \lfloor \frac{y}{100} \rfloor + \lfloor \frac{y}{400} \rfloor) \bmod 7 & \text{if } m \geq 3 \end{cases}$$



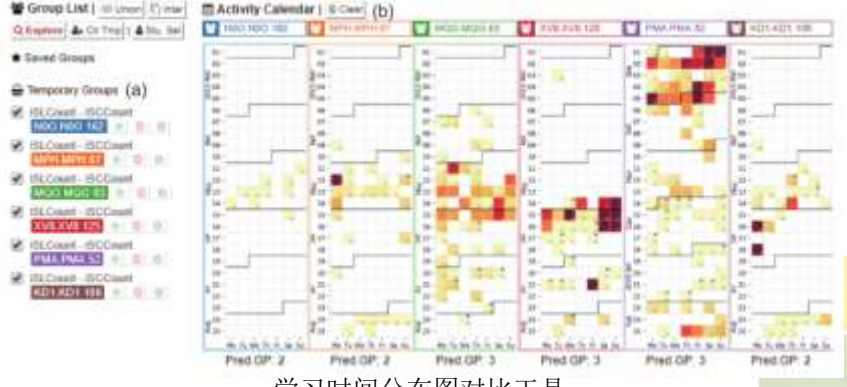
学习参与度时序矩阵转换为学习参与度日历矩阵

研究内容2：面向时序数据的学习时间管理可视分析

- 可视化设计：基于日历图的学习时间分布图
  - 日历图布局呈现学习参与度日历矩阵
    - 交互活动时序→ 日历位置
    - 交互活动→ 颜色编码
    - 学习参与度聚合值→ 颜色编码
  - 学习时间分布图对比工具
    - 同时对比多个学习时间分布图
    - 学生分组→ 颜色编码



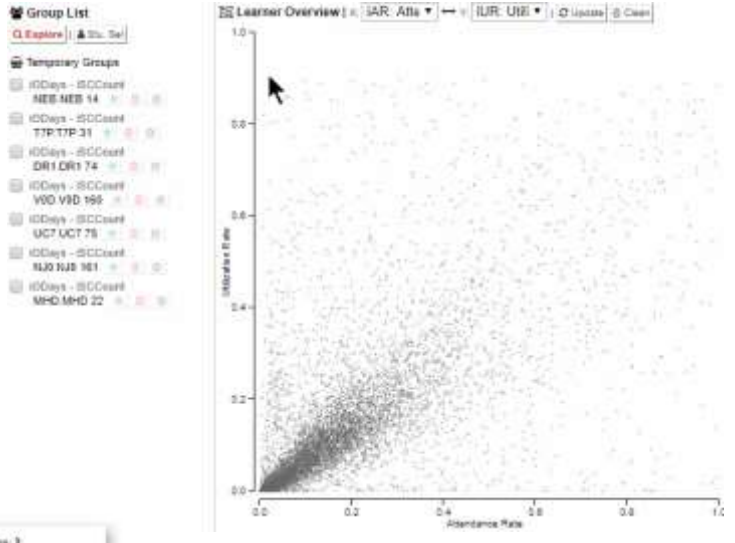
基于日历图的学习时间分布图



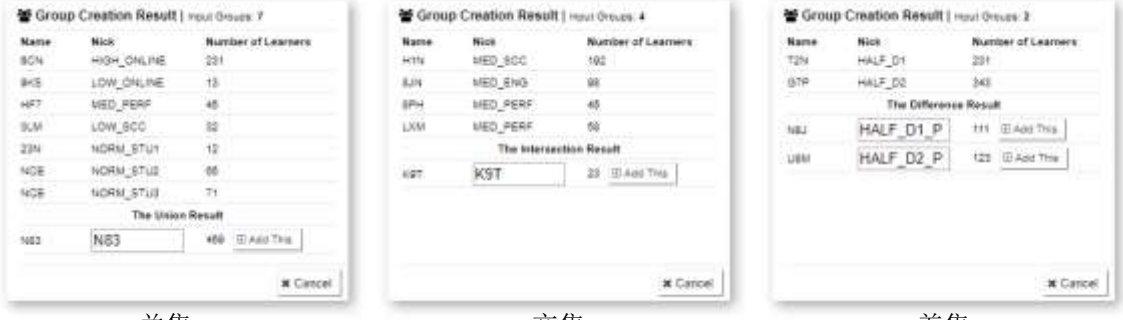
学习时间分布图对比工具

研究内容2：面向时序数据的学习时间管理可视分析

- 可视化交互设计：学生群组创建工具
  - 基于交互选择
    - 数据点分布稀疏不规则
      难以量化描述
    - 鼠标圈选任意封闭形状区域
  - 基于集合运算
    - 并集运算符，交集运算符，差集运算符



基于交互选择的学生群组创建工具



并集　　交集　　差集

基于集合运算的学生群组创建工具

研究内容2：面向时序数据的学习时间管理可视分析

- 可视化设计：
  多课程视频观看可视化
  - 可视化设计
    - 课程进度图
    - 观看时间分布图
    - 每周视频观看量统计图
  - 数据映射
    - 时序→ 日历位置
    - 课程→颜色编码
    - 到课率AR→ 单元格
    - 到课率UR→ 格内柱状图
  - 交互设计
    - 多视图关联

研究内容2：面向时序数据的学习时间管理可视分析

- 案例分析1：学习时间管理风格对比
  - 数据集：网络教育学院2015年第一学期英语课，学生10,529名，共 23,918 条

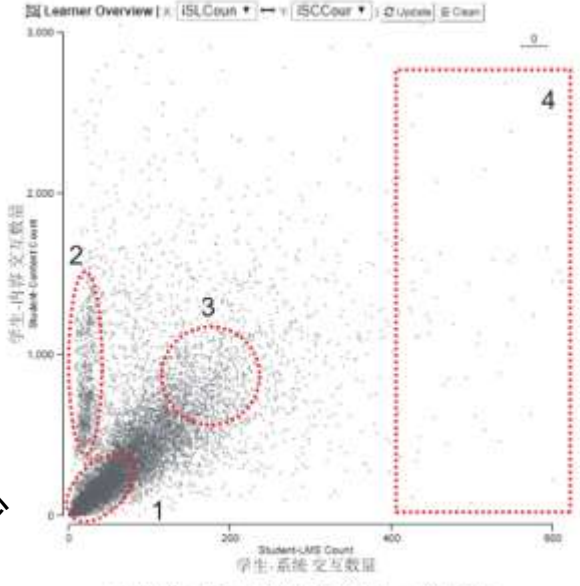
不同学习参与度模式的
时间管理风格存在差异

区域1：学习时间少

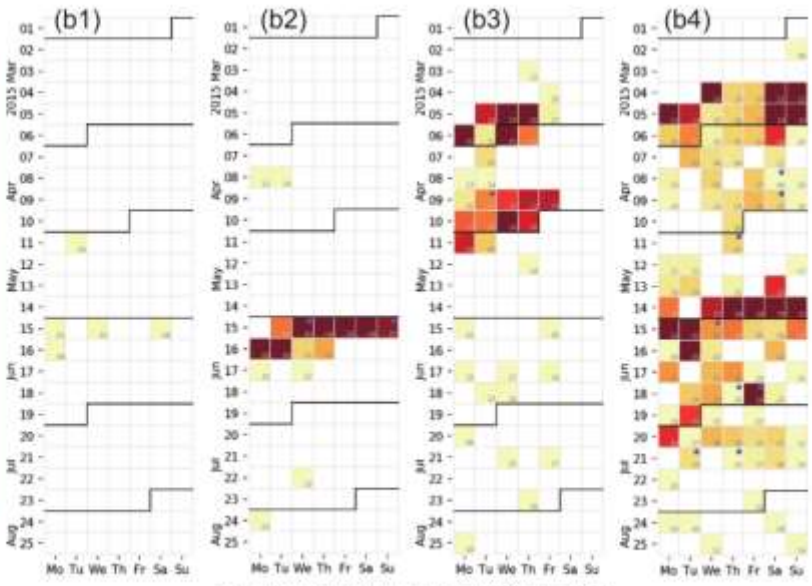
区域2：拖延学习。
考前连续密集学习

区域3：松懈疲劳。
前期较高，后期显著减少

区域4：自律学习。
整个学期持续学习



(a) 学习参与度分布图的 4 个区域



(b) 4个区域的典型学习时间分布图

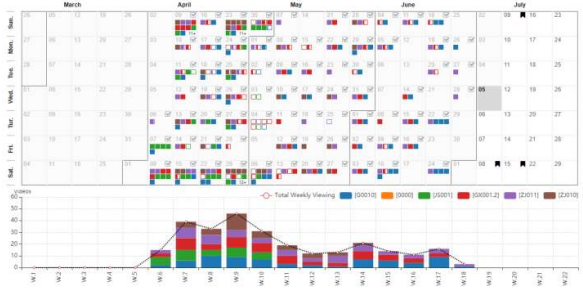
区域1：两项指标低，说明很少使用学习管理系统且很少访问教学资源
区域2：有限的上线学习次数中，大量访问课程视频、阅读材料等教学资源
区域3：使用学习管理系统的次数和访问教学资源的次数较多
区域4：频繁使用学习管理系统，教学资源访问量很高

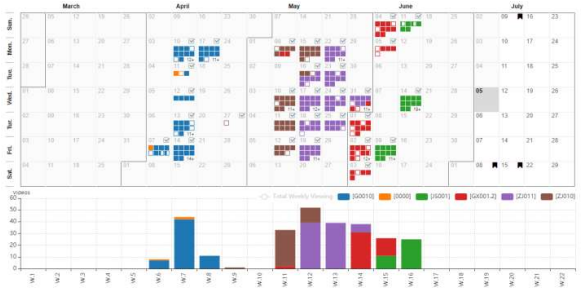
研究内容2：面向时序数据的学习时间管理可视分析

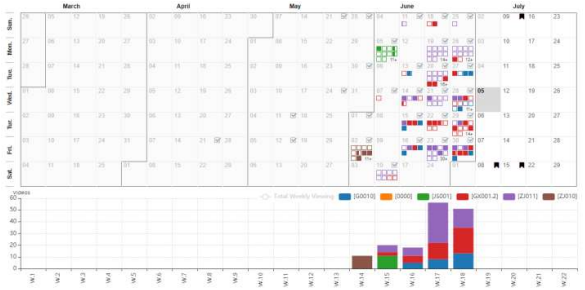
- 案例分析2：多课程学习时间管理风格
  - 数据集：2017年春季网络教育学院计算机专业第一学期的6门课程
    - 学生：164
    - 数据量：40,928

第一学期		
课程编号	课程名称	学分
GX001.2	英语(二)	3
G0010	毛泽东思想和中国特色社会主义理论体系概论	3
ZJ011	高等数学(专升本)	4
0000	远程教育学习学	1
ZJ010	程序设计基础	3
JS001	计算机及网络应用基础	2
合计	必修 16 学分	

  - 3种学习时间管理风格
    - 多课程同时规律学习， 单课程逐个规律学习，多课程混合拖延学习

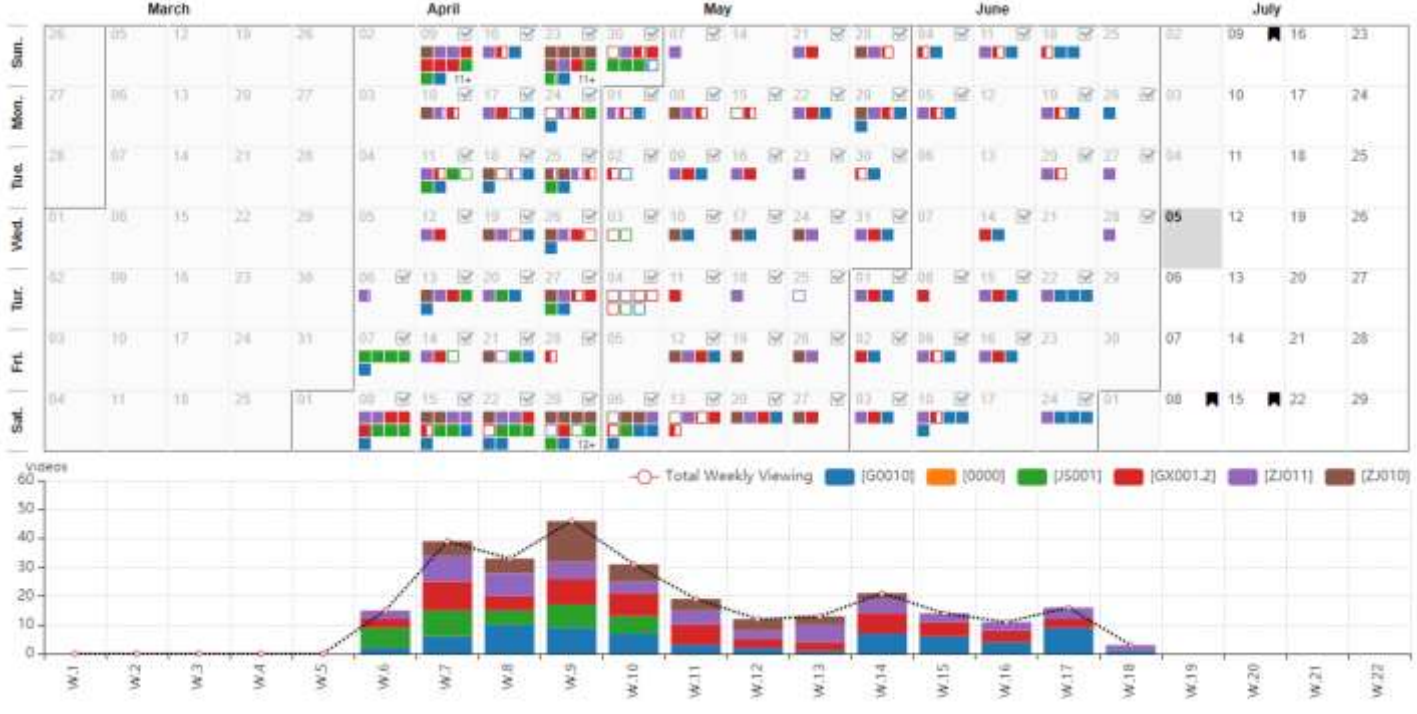






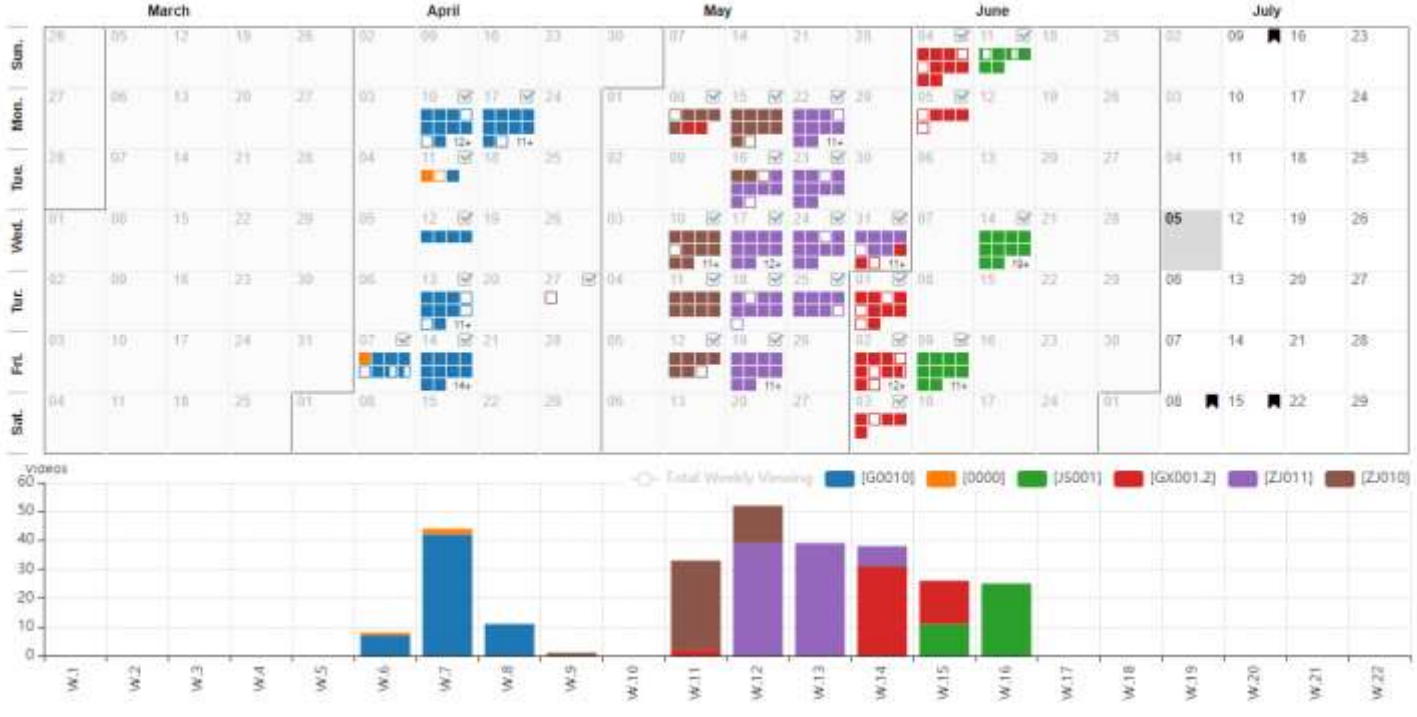
研究内容2：面向时序数据的学习时间管理可视分析

- 案例分析2：多课程学习时间管理风格
  - 多课程同时规律学习
    - 每周学习 4-5 门课程，并且每天完整观看 2-3 门课程的少量视频
    - 自律学习，并行管理多个课程的学习时间和学习进度



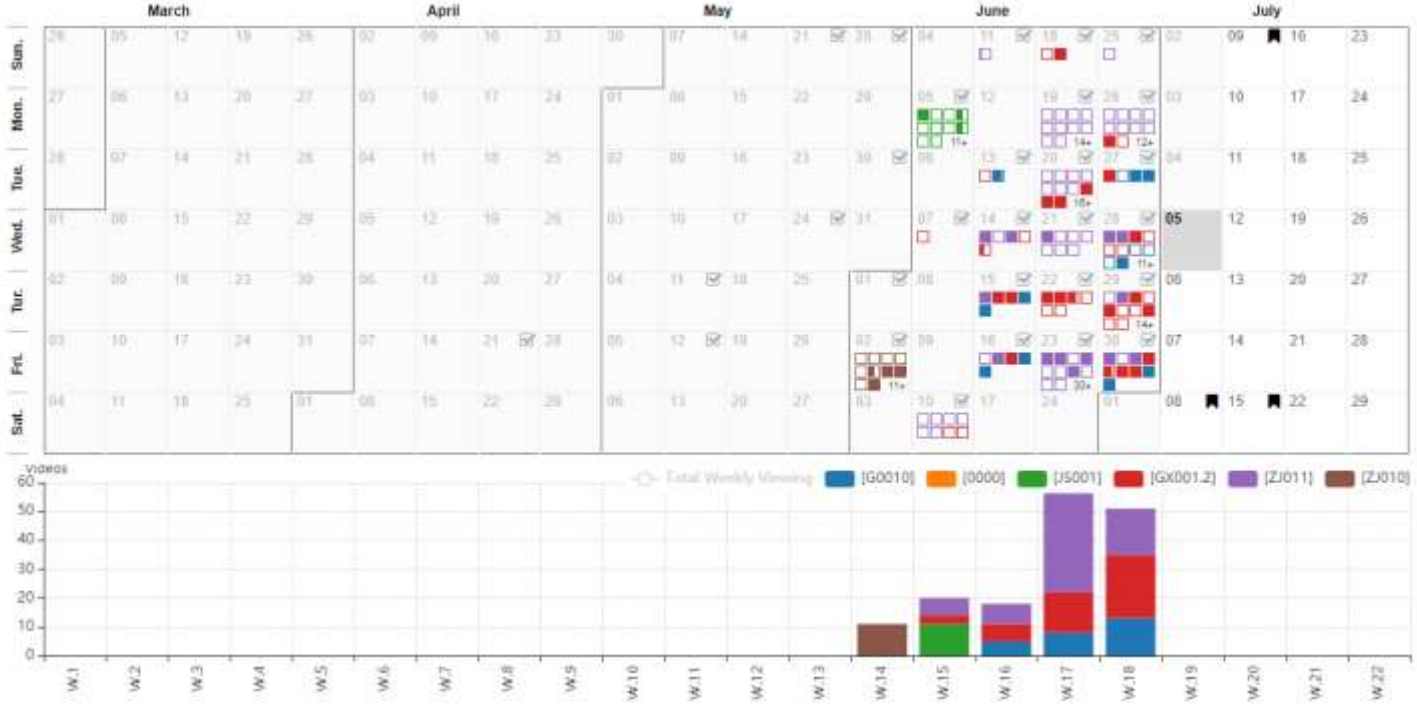
研究内容2：面向时序数据的学习时间管理可视分析

- 案例分析2：多课程学习时间管理风格
  - 单课程逐个规律学习
    - 每周学习 1-2 门课程，每天完整观看 1门课程的大量视频
    - 自律学习，串行的管理多个课程的学习事件，逐个课程的完成学习任务



研究内容2：面向时序数据的学习时间管理可视分析

- 案例分析2：多课程学习时间管理风格
  - 多课程混合拖延学习
    - 学期的前中期基本不观看视频，学期末开始连续数周突击观看，但不完整观看
    - 拖延学习，说明学生可能不擅长管理自己的多个课程的学习时间和进度



研究内容2：面向时序数据的学习时间管理可视分析

- 小结
  - 提出了一种刻画在线学习过程中学习参与度时序特征的抽象模型
  - 设计了一种基于日历坐标系的学习参与度时序特征表示方法与可视化方法
  - 设计了一种反映多课程同期学习时间安排的可视化方法

- 成果
  - VUC: Visualizing Daily Video Utilization to Promote Student Engagement in Online Distance Education [C]. ACM Global Computing Education Conference 2019 (**CompEd 2019**, EI: 20192106966416)
  - LearnerExp: Exploring and Explaining the Time Management of Online Learning Activity [C]. The Web Conference 2019 Demonstration Track (**WWW 2019**, **CCF A类会议Demo论文**, EI: 20192407027811)
  - Visual Analysis of the Time Management of Learning Multiple Courses in Online Learning Environment [C]. IEEE Conference on Visual Analytics Science and Technology 2019 Short Paper Track (**VAST 2019**, **CCF A类会议短论文**，已录用)

内容概要

1. 研究背景与内容
2. 研究内容1：面向高维数据的学习参与度可视分析
3. 研究内容2：面向时序数据的学习时间管理可视分析
4. **研究内容3：面向大规模多属性数据的视频利用情况可视分析**
5. 结论与展望

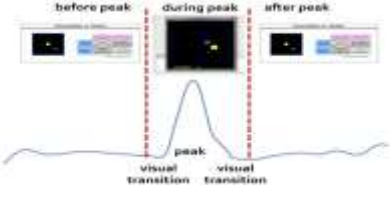
研究内容3：面向大规模多属性数据的视频利用情况可视分析

- 研究目的：从不同层面探索在线学习过程的资源利用模式，
  主要解决资源利用情况衡量以及多层面关联分析问题
- 研究现状
  - 基于数量或时长等指标的细粒度行为模式分析
    - 视频热点操作模式[Kim, 2014]、视频风格[Crook, 2017]
    - 成绩、留存率、退课率预测[Chen and Wu, 2015]
    - 不足：样本规模小，且指标仅适合单个课程
      缺少跨课程、多层面的模式分析
  - 分析结果呈现
    - 表格、折线图、柱状图等基本图表
    - 不足：视频利用模式与分析结果不直观

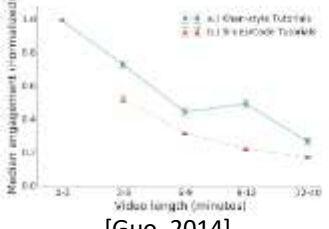
Coursera


网络远程教育

教学视频是在线学习的主要教学媒介



[Kim, 2014]



[Guo, 2014]


[Crook and Schofield, 2017]

- 研究难点
  - 如何衡量大规模视频资源的利用情况？
  - 如何对比分析不同属性层面的视频利用情况？

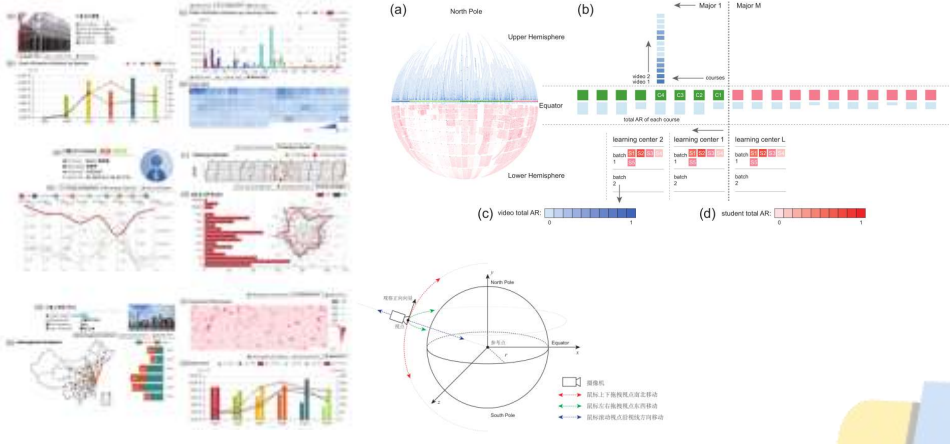
研究内容3：面向大规模多属性数据的视频利用情况可视分析

- 研究思路：基于关联协调视图呈现大规模多层面视频利用模式
  - 视频观看行为建模模型
    - 基于观看行为的视频到课率与利用率
    - 扩展到其他属性的视频到课率与利用率
  - 可视化与交互设计
    - 基于球面布局的概览视图
    - 基于混合图表的多属性详情视图与基于平行坐标系的多属性对比视图
    - 固定参考点导航工具
    - 多视图协调关联

$$ar_{s,v} = \begin{cases} 1 & viewed \\ 0 & not\ viewed \end{cases} \qquad ar_{s,c} = \frac{|W_{s,c}|}{|V_c|} \qquad ur_{s,v} = \frac{wt_{s,v}}{vt_v} \qquad ur_{s,c} = \frac{\sum_{v \in c \wedge v \in W_{T_{s,c}}} wt}{\sum_{v \in c \wedge v \in T_c} vt}$$

$$ars_{L^t,B^t,M^t,C^t} = \frac{1}{|S_{L^t,B^t,M^t,C^t}|} \sum_{c \in C^t \wedge s \in S_{L^t,B^t,M^t,C^t}} ar_{s,c}$$

$$urs_{L^t,B^t,M^t,C^t} = \frac{1}{|S_{L^t,B^t,M^t,C^t}|} \sum_{c \in C^t \wedge s \in S_{L^t,B^t,M^t,C^t}} ur_{s,c}$$



研究内容3：面向大规模多属性数据的视频利用情况可视分析

- 可视分析任务
  - T.1 分析全部视频的利用情况分布
  - T.2 分析每个课程的视频利用分布
  - T.3 分析每个学生的视频利用分布
  - T.4 分析每个学生群体的视频利用分布

- 设计需求
  - R.1 多尺度探索
  - R.2 多角度呈现
  - R.3 对比分析
  - R.4 交互探索

研究内容3：面向大规模多属性数据的视频利用情况可视分析

- 基于观看行为的视频到课率与利用率

到课率（Attendance Rate，AR）

衡量学生s是否观看了视频v

$$ar_{s,v} = \begin{cases} 1 & viewed \\ 0 & not\ viewed \end{cases} \qquad ar_{s,c} = \frac{|W_{s,c}|}{|V_c|}$$

利用率（Utilization Rate，UR）

衡量学生s观看视频v的程度

$$ur_{s,v} = \frac{wt_{s,v}}{vt_v} \qquad ur_{s,c} = \frac{\sum_{wt \in WT_{s,c}} wt}{\sum_{vt \in VT_c} vt}$$

– 扩展：学习中心L'、批次B'、专业M'、课程C'

$$ar_{S_{L',B',M',C'}} = \frac{1}{|S_{L',B',M',C'}|} \sum_{c \in C' \wedge s \in S_{L',B',M',C'}} ar_{s,c}$$

$$ur_{S_{L',B',M',C'}} = \frac{1}{|S_{L',B',M',C'}|} \sum_{c \in C' \wedge s \in S_{L',B',M',C'}} ur_{s,c}$$

– 存储：预计算全部组合，组合名存哈希表

研究内容3：面向大规模多属性数据的视频利用情况可视分析

- 可视化设计：基于球面布局的概览视图
  - 课程 c 坐标

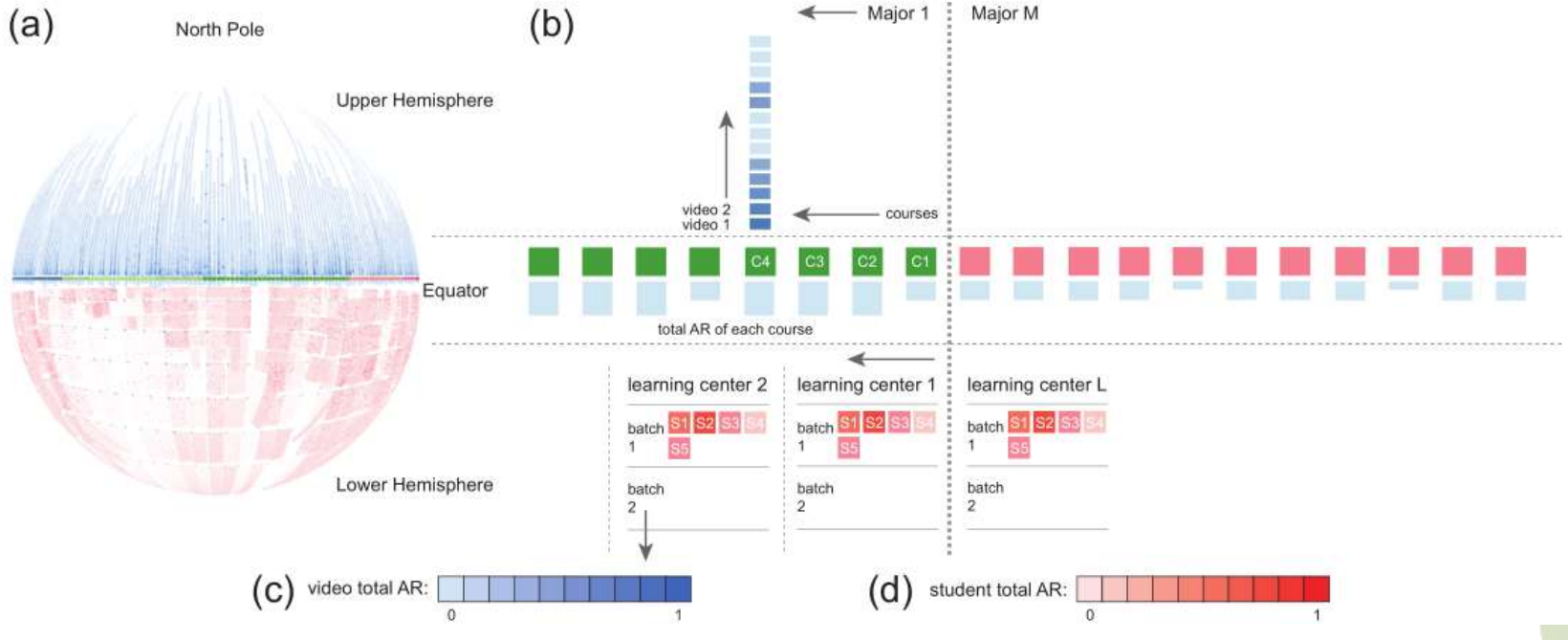
$$\varphi_c = 0$$
$$\theta_c = c\frac{2\pi}{C}$$

课程 c 视频 v 坐标

$$\varphi^v_{c,i} = i\frac{\pi}{2V_{max}} + \varphi_{base}$$
$$\theta^v_{c,i} = i\frac{2\pi}{|C|}$$

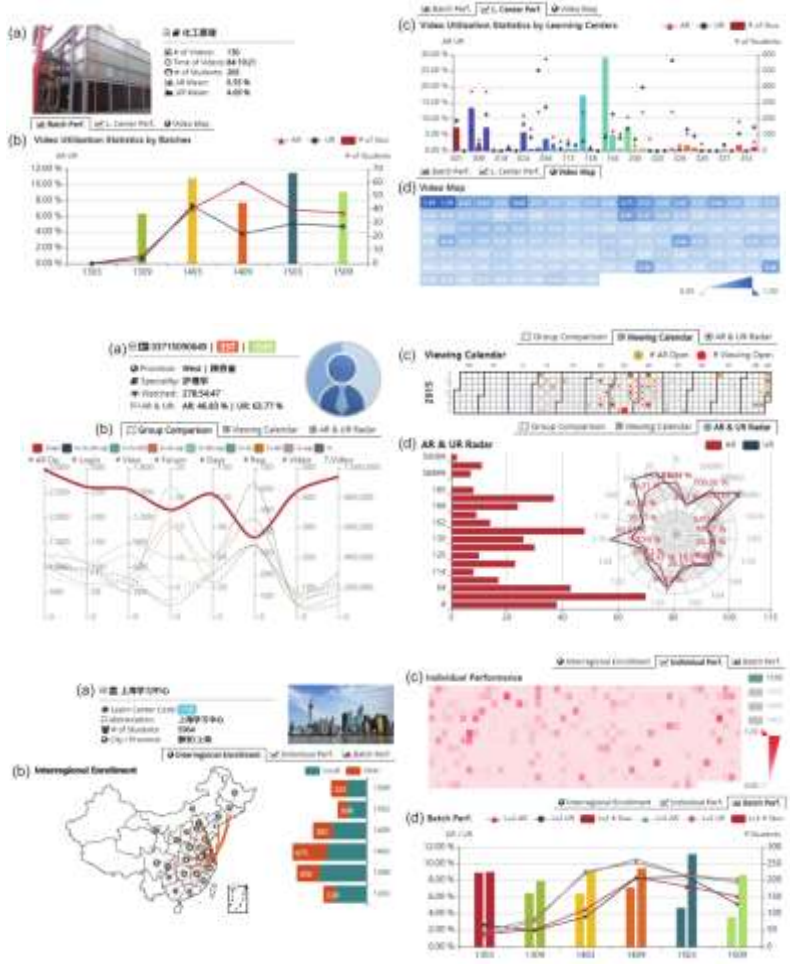
学习中心 l 批次 b 第 j 个学生 s 坐标

$$\varphi^s_{l,b,j} = b\frac{\alpha_B\pi}{2|B|} + \beta_B\lfloor\frac{j}{N_w}\rfloor\frac{\alpha_B\pi}{2|B|} + \varphi_{base}$$
$$\theta^s_{l,b,j} = l\frac{2\pi}{|L|} + \beta_L(j \bmod N_w)\frac{2\pi}{|L|}$$



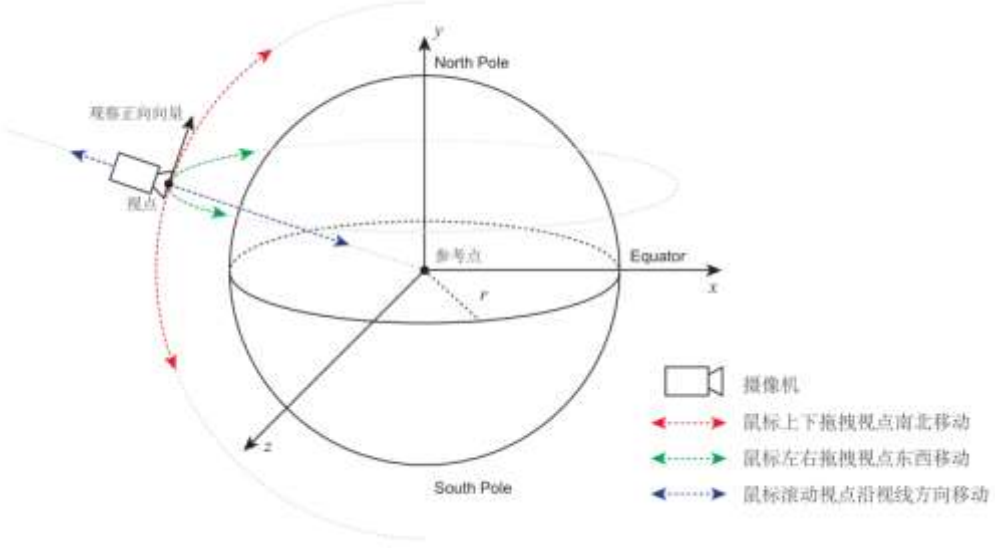
研究内容3：面向大规模多属性数据的视频利用情况可视分析

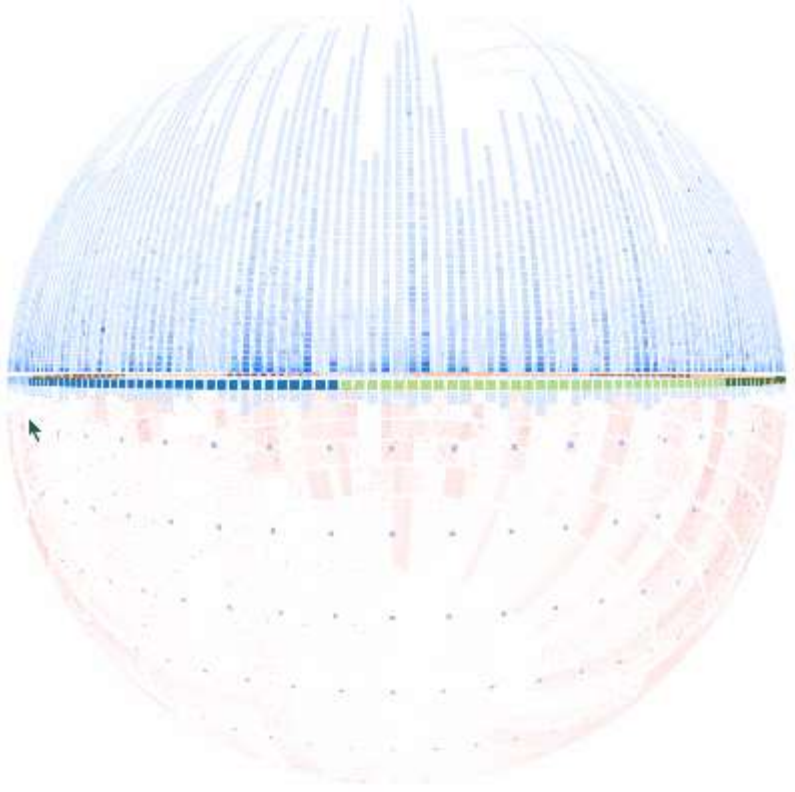
- 可视化设计：基于混合图表的多属性详情视图
  - 课程详情面板
    - 各入学批次视频利用柱状图
    - 各学习中心视频利用柱状图
    - 视频利用情况热力图
  - 学生详情面板
    - 每个课程的到课率雷达图
    - 对比平行坐标
    - 长期学习过程日历图
  - 学习中心详情面板
    - 招生区域地图
    - 学生视频利用率热力图
    - 各入学批次视频利用情况柱状图



研究内容3：面向大规模多属性数据的视频利用情况可视分析

- 交互设计：固定参考点导航工具
  - 在球面轨道上任意视角查看概览视图
    - 参考点固定球心
    - 鼠标操作映射轨道移动

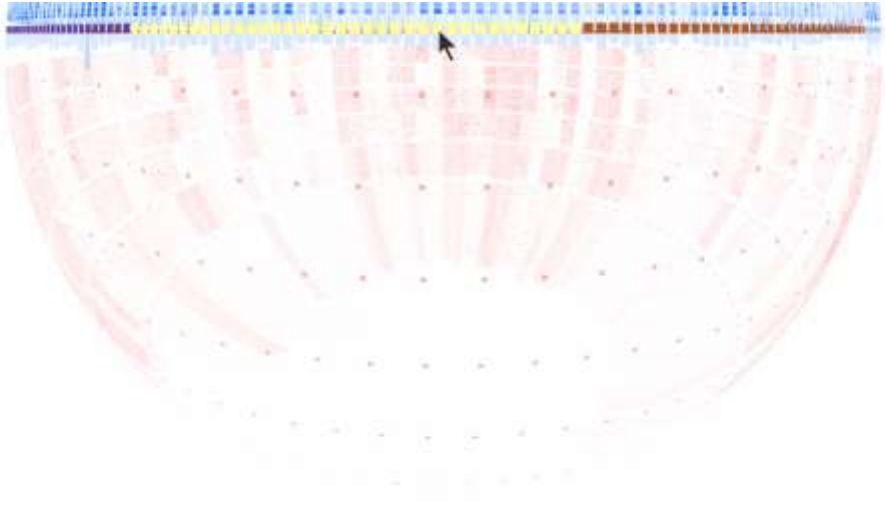




固定参考点导航工具示例

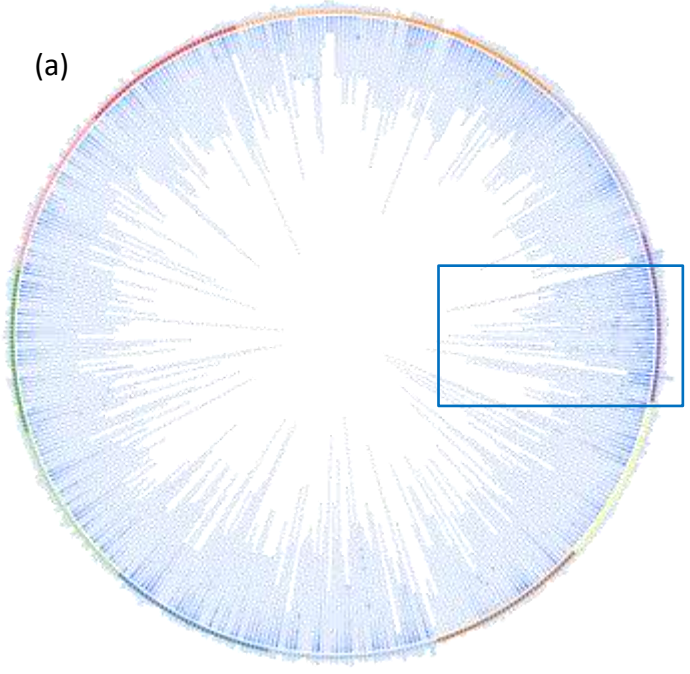
研究内容3：面向大规模多属性数据的视频利用情况可视分析

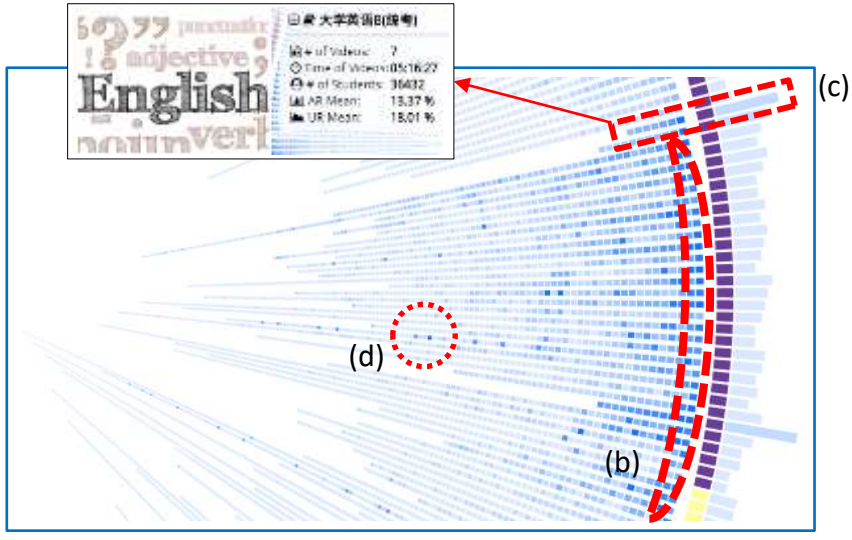
- 交互设计：多视图协调关联
  - 概览视图
    - 视频、课程、学习中心、学生符号
      关联详情视图内对应面板
  - 详情视图
    - 各面板的元素彼此关联



研究内容3：面向大规模多属性数据的视频利用情况可视分析

- 案例分析1：整体视频利用模式
  - 数据集：网络教育学院2013-2015数据
    学生：80,484，学习中心：91，课程：219
    视频：26,660，日志：100,359,271

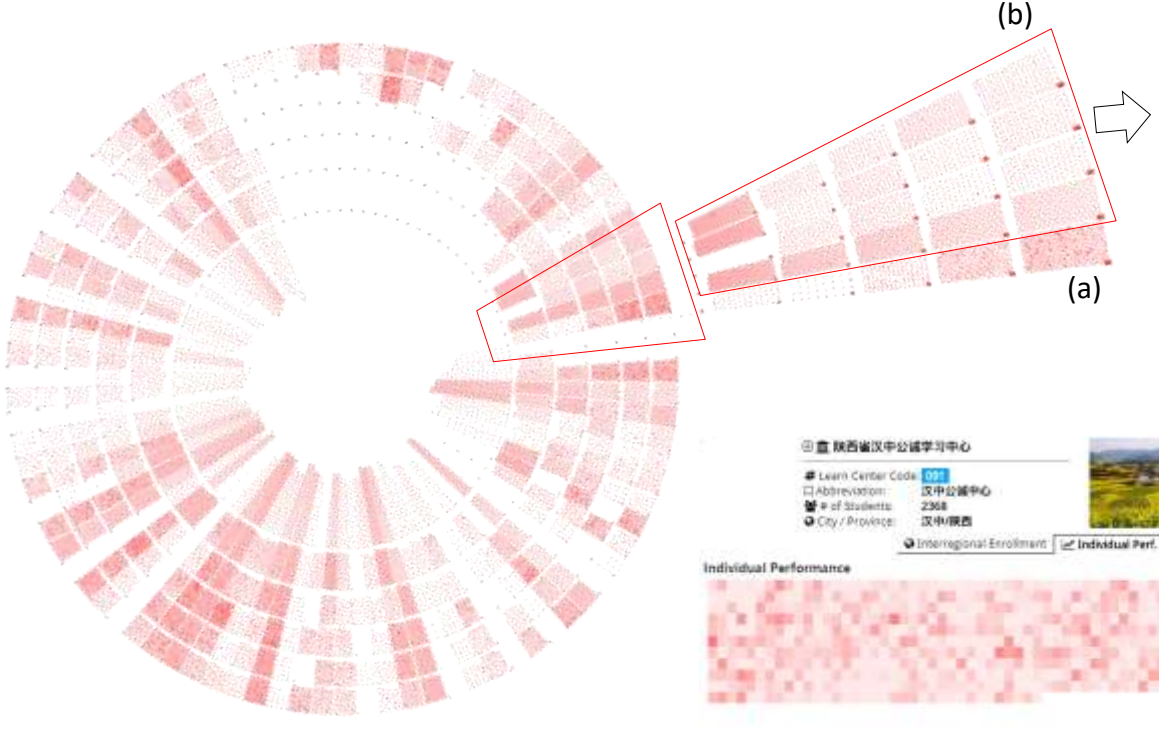




(a) 大部分视频利用程度低
(b) 前3节视频利用程度较高
(c) 短课程、学位相关课程利用程度较高
(d) 考试重点视频利用程度较高

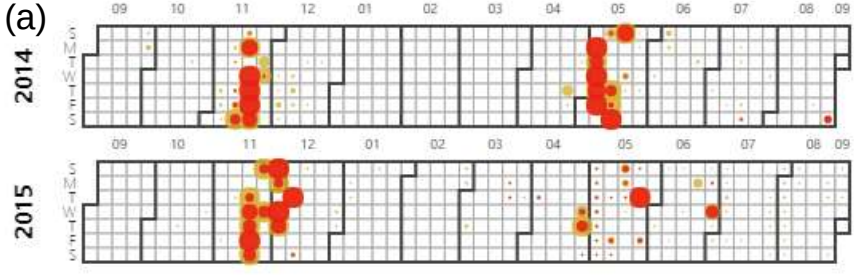
研究内容3：面向大规模多属性数据的视频利用情况可视分析

- 案例分析1：整体视频利用模式
  (a) 学习中心内学生视频利用率差异显著
  (b) 部分学习中心整体AR低且没有显著差异

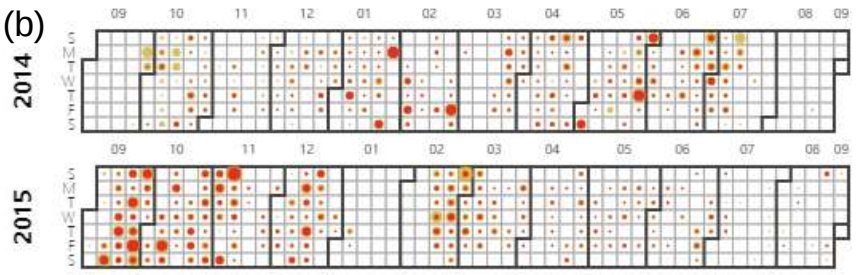


研究内容3：面向大规模多属性数据的视频利用情况可视分析

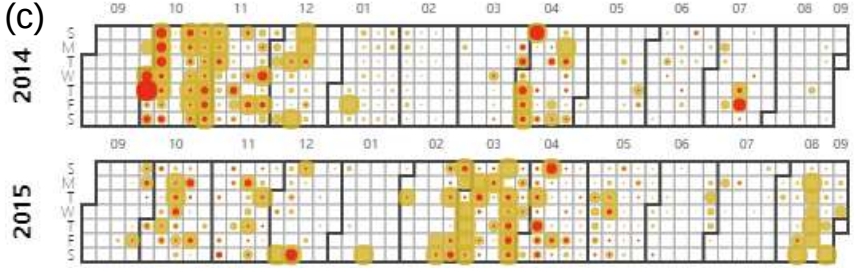
- 案例分析2：学生资源利用的时间模式
  - 存在明显不同的模式



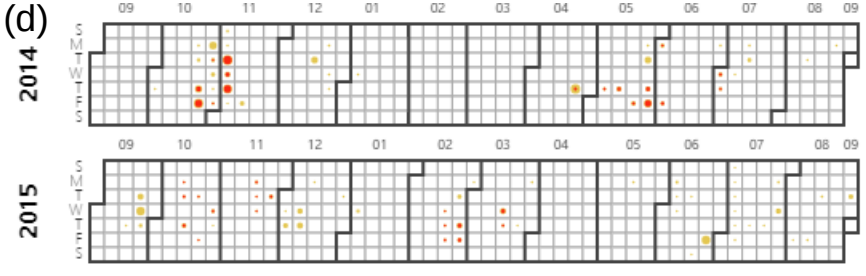
周期性拖延学习



长期规律学习



长期规律讨论学习



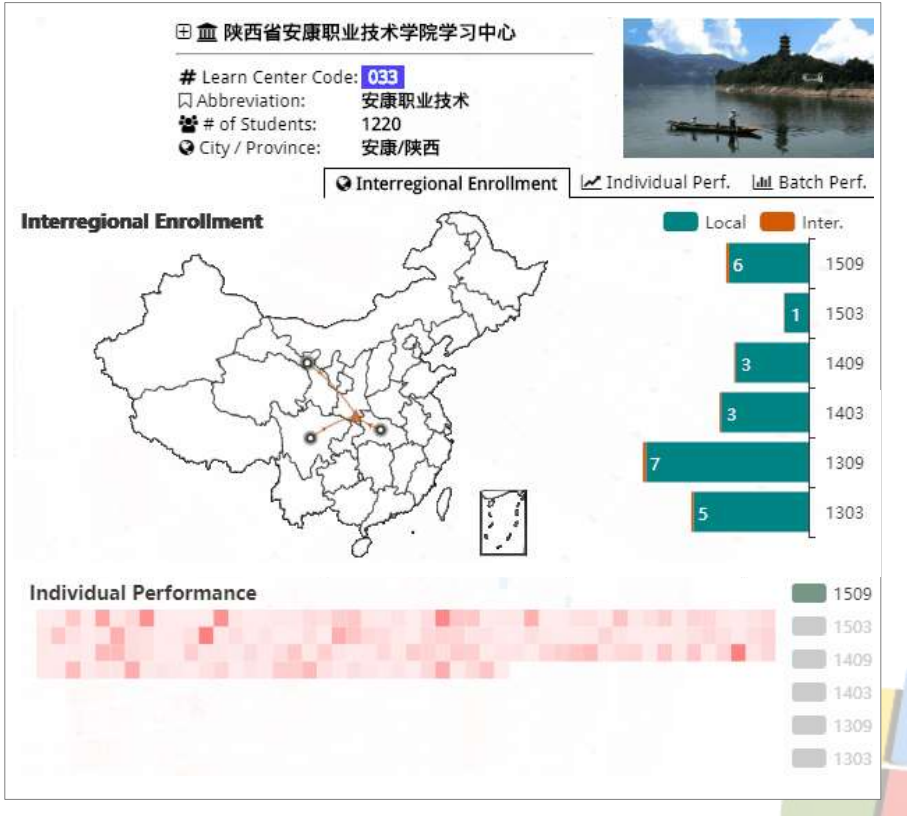
长期低参与度学习

研究内容3：面向大规模多属性数据的视频利用情况可视分析

- 案例分析3：生源分布与视频利用率模式

  生源地数量、外地学生数与视频利用情况相关

生源地数量多且外地生源人数多



生源地数量少且外地生源人数少

研究内容3：面向大规模多属性数据的视频利用情况可视分析

- 小结
  - 提出了通用的到课率和利用率指标，衡量学生对课程视频资源的利用情况
    并将其扩展到学生和视频的不同层面和属性
  - 提出了一种基于球面布局的多属性数据的可视化设计展示整体和局部视频利用分布
  - 设计了混合图表的多属性详情视图与对比视图展示多属性的视频利用分布

- 成果
  - VUSphere: Visual Analysis of Video Utilization in Online Distance Education [C].
    IEEE Conference on Visual Analytics Science and Technology 2018
    (**VAST 2018**, **CCF A类会议论文**)

攻读博士学位期间取得的研究成果

- 论文：已发表或录用 11 篇（第一作者 8 篇）

1. **Huan He**, Qinghua Zheng, and Bo Dong. VUSphere: Visual Analysis of Video Utilization in Online Distance Education [C]. IEEE Conference on Visual Analytics Science and Technology 2018 (**VAST 2018**), Berlin, Germany, Oct 21-26, 2018 (**CCF A 类会议论文**，已录用)
2. **Huan He**, Qinghua Zheng, Dehai Di and Bo Dong. How Learner Support Services Affect Student Engagement in Online Learning Environment [J]. **IEEE Access**, 2019, 7: 49961 - 49973 (SCI: 000466757700001)
3. **Huan He**, Bo Dong, Qinghua Zheng, Dehai Di and Yating Lin. Visual Analysis of the Time Management of Learning Multiple Courses in Online Learning Environment [C]. IEEE Conference on Visual Analytics Science and Technology 2019 Short Paper Track (**VAST 2019**), Vancouver, BC Canada, Oct 20-25, 2019 (**CCF A类会议短论文**，已录用)
4. **Huan He**, Qinghua Zheng, and Bo Dong. LearnerExp: Exploring and Explaining the Time Management of Online Learning Activity [C]. The Web Conference 2019 Demonstration Track (**WWW 2019**), San Francisco, CA USA, May 13-17, 2019 (**CCF A类会议 Demo 论文**，EI: 20192407027811)
5. **Huan He**, Qinghua Zheng, Bo Dong, and Guobin Li. VUC: Visualizing Daily Video Utilization to Promote Student Engagement in Online Distance Education [C]. ACM Global Computing Education Conference (**CompEd 2019**), Chengdu, China, May 17-19, 2019 (EI: 20192106966416)
6. **Huan He**, Qinghua Zheng, Bo Dong and Hongchao Yu. Measuring Student's Utilization of Video Resources and its Effect on Academic Performance [C]. Proceedings of IEEE 18th International Conference on Advanced Learning Technologies (**ICALT 2018**), Mumbai, India, Jul 9-13, 2018, pp. 196-198 (EI: 20183605769536)
7. **Huan He**, Qinghua Zheng, Rui Li, and Bo Dong. Using Face Recognition to Detect "Ghost Writer" Cheating in Examination [C]. The 12th International Conference on E-learning and Games (**Edutainment 2018**), Xi'an, China, June 28-30, 2018 (Best Paper Award Honorable Mention)
8. Qinghua Zheng, **Huan He**, Tian Ma, Ni Xue, Bing Li, Bo Dong. Big Log Analysis for e-Learning Ecosystem [C].Proceedings of IEEE 11th International Conference on e-Business Engineering (**ICEBE 2014**), Guangzhou, China, pp. 258-263, 2014 (EI: 20150300423652)
9. Hongchao Yu, **Huan He**, Qinghua Zheng, and Bo Dong. TaxVis: A Visual System for Detecting Tax Evasion Group [C]. The Web Conference 2019 Demonstration Track (**WWW 2019**), San Francisco, CA USA, May 13-17, 2019 (**CCF A类会议 Demo 论文**，EI:20192407027829)
10. Qinghua Zheng, Yating Lin, **Huan He**, Jianfei Ruan, and Bo Dong. ATTENet: Detecting and Explaining Suspicious Tax Evasion Group [C]. The 28th International Joint Conference on Artificial Intelligence Demonstration Track (**IJCAI 2019**), Macao, China, August 10-16, 2019 (**CCF A类会议 Demo 论文**，已录用)
11. Ni Xue, **Huan He**, Jun Liu, Qinghua Zheng, Tian Ma, Jianfei Ruan, and Bo Dong. Probabilistic Modeling Towards Understanding the Power Law Distribution of Video Viewing Behavior in Large-Scale e-Learning [C]. 2015 IEEE TrustCom/BigDataSE/ISPA, Helsinki, Finland, August, 2015 (CCF C 类会议，EI:20162102416538)

攻读博士学位期间取得的研究成果

- 已授权专利：学生一作 2 项，其余 4 项

1. 张未展, **贺欢**, 薛妮, 郑庆华, 董博. 一种面向MapReduce框架的地理归属信息查询方法, ZL201410328449.0 (2016-03-30 已授权)
2. 郑庆华, 董博, **贺欢**, 宋凯磊, 徐海鹏, 马天, 陈亚兴. 一种基于HBase的构建和检索增量索引的方法, ZL201310298976.7 (2016-03-30 已授权)
3. 董博, 薛妮, **贺欢**, 郑庆华, 马天. 一种基于DOM的网页关键内容抽取方法, ZL201410840805.7 (2016-03-30 已授权)
4. 郑庆华, 马天, 李冰, **贺欢**, 阮建飞, 张镇潮, 施建生, 王培勇, 钱运辉. 一种基于HBase的税收统计报表存储与计算的方法, ZL201410658492.3 (2016-06-08 已授权)
5. 刘均, 徐海鹏, 董博, 郑庆华, 马天, **贺欢**, 李冰. 基于重叠点识别的网络重叠社团检测方法, ZL201310272890.7 (2015-04-29 已授权)
6. 张未展, 张汉宁, 郑庆华, 董博, **贺欢**. 基于主从架构的MapReduce任务跨数据中心调度系统及方法, ZL201410344242.2 (2015-09-30 已授权)

- 参与课题项目

1. 全球汉语言文化传播服务平台及应用示范，国家科技支撑计划
2. 国家志愿者公益服务支撑平台关键技术研究与示范应用，国家科技支撑计划
3. 碎片化知识聚合方法研究，国家自然科学基金重点项目
4. 上海网达 GridFS 项目，横向课题

评审专家的问题

- 张向荣教授：

1. 第五章中提到的多属性数据中属性的内涵并不明确，需要补充说明。

回答：感谢张老师的意见，已在第五章5.1节和5.2节中增加了对视频和学生属性的说明。例如，视频的属性可能包括课程、学科、类型和时长等，学生的属性可能包括个人背景、专业、学习时间和学习环境等。

2. 可视化有效性评价主要体现在哪几个方面？用户体验如何？以及可视化的全面性如何？

回答：感谢张老师的意见，已在第二章2.1节、第三章3.5节、第四章4.5节和第五章5.2节中分别增加了可视化有效性评价的的主要方法、用户体验和反馈。

评审专家的问题

- 盲审专家提出的问题：

1. 文中提出了几种可视化分析方法，但这些方法所展示的效果与数据中所蕴含的学习行为的一致性讨论较少，建议增加这方面的分析

回答：感谢评审专家的意见，已在本文第三章2.3.1和2.3.2节、第四章4.5.1和4.5.2节、第五章5.5.1、5.5.2和5.5.3节分别对各可视化方法的案例分析增加了进一步说明，特别是增加了领域专家对于可视化呈现的效果与真实的在线学习行为之间的关系进行解释。

2. 第三章“面向高维数据的学习参与度可视分析”中建立了基于学生行为的参与度可视化模型，但是缺少跟以前模型性能的比较分析，同时也缺少对学生参与度的波动原因的分析

回答：感谢评审专家的意见，已在第三章3.3.4节、3.5.1节和3.5.2节中分别增加了对学习参与度模型中聚合数据层的读写时间复杂度的分析，同时也增加了对各群组之间学生参与度差异的讨论。

致谢

- 衷心感谢各位老师!