西安交通大學

博士学位论文

在线学习行为数据的可视分析方法研究

学位申请人：贺欢

指导教师：郑庆华教授

学科名称：计算机科学与技术

2019 年 7 月

Visual Analysis Methods on Online Learning Behavior Data

A dissertation submitted to

Xi'an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

By

Huan He

Supervisor: Prof. Qinghua Zheng

Computer Science and Technology

July 2019

博士学位论文答辩委员会

在线学习行为数据的可视分析方法研究

答辩人：贺欢

答辩委员会委员：

西北工业大学教授：周兴社 (主席)

西安电子科技大学教授：苗启广

西安交通大学教授：赵银亮

西安交通大学教授：张未展

西安交通大学教授：李辰

答辩时间：2019 年 7 月 29 日

答辩地点：西安交通大学彭康楼 218 会议室

摘 要

近年来，随着互联网技术的快速发展以及个人移动计算终端的广泛普及，在线学习的规模迅猛增长，并随之产生了海量的在线学习行为数据。然而，随着学习者数量以及在线教学手段的复合增长，海量的在线学习行为数据对探索分析学习行为规律和模式提出了多方面挑战，具体表现在：首先，互联网应用技术水平的提升使在线课程的教学功能更加丰富，多样化的教学功能使在线学习行为数据蕴含的特征维度更高，现有的教育数据分析工具和分析方法难以刻画和解释高维度的学习参与度模式；其次，长期的在线学习过程使得在线学习行为数据天然具有时序性，然而常用的聚合数值指标与基于时序图的可视化难以刻画和呈现学生的时间管理特征；最后，视频资源与学生数量的猛增使得在线学习过程中各对象属性的差异变得更加显著，用户需要在大规模在线学习行为数据中探索不同属性的视频利用规律，而现有的静态统计图表难以提供支持。针对上述挑战，本研究提出了基于可视分析方法的在线学习行为数据分析方法，帮助用户直观探索大规模在线学习行为数据中蕴含的学习参与度模式、时间管理风格以及视频资源利用规律，为改进教学服务、提高教学质量提供支持 。本文的主要研究内容及贡献概括为以下三点：

第一，提出了一种基于散点图矩阵与 t-SNE 数据降维技术的学习参与度可视分析方法。首先，设计了层次化的在线学习行为数据存储模型，并基于该模型设计了在线学习行为数据管理系统，解决了多源异构在线学习行为数据的采集、清洗、抽象和学习指标计算的问题；其次，提出了一种学习参与度指标模型用来刻画复杂多样的在线学习行为特征，并提出了基于散点图矩阵和 t-SNE 算法的学习参与度分布图，再结合基于 DBSCAN 算法的学生群组创建工具，分析高维度学习特征中不同学生群体的学习参与度模式；最后，为了验证所提方法的有效性，本文对网络教育学习者的在线学习行为数据进行研究，发现了多种学习参与度模式。

第二，在前一研究的基础上，进一步提出了基于日历坐标系的学习参与度时序特征可视分析方法。首先，提出一种刻画长期在线学习过程中时序信息的数据模型，将学习参与度随时间的变化过程表示为学习参与度时序矩阵；然后，设计了一个基于卷积神经网络的成绩区间预测模型，捕捉学习参与度时序矩阵中局部列向量包含的时间维度信息与行向量包含的学习参与度信息；随后，提出一种基于日历坐标系的学习参与度日历矩阵，并设计了基于日历图的学习时间分布图展示学习参与度时序矩阵中包含的时序特征，直观呈现学习参与度在日、周、月和学期尺度上的周期性变化；并且，针对在线多课程同期学习的特点，设计了多课程学习进度图，通过日历图与堆叠图编码多课程学习的时间安排信息，直观呈现学生对多个课程的时间管理特征；最后，基于前一研究内容与本研究内容设计了 LearnerExp 可视分析系统，并结合真实的在线学习行为数据验证了所提方法的有效性，并讨论了学生在长期学习过程中的时间管理风格。

本研究得到国家重点研发计划”教育大数据分析挖掘技术及其智慧教育示范应用”（2018YFB1004500），教育部创新团队（IRT17R86），国家自然科学基金（61721002, 61502379, 61532015）和中国工程科技知识中心项目资助

第三，提出一种基于球面布局的视频资源利用情况分布可视分析方法，并进一步将前两项研究拓展到大规模多属性的在线学习行为数据。首先，根据在线学习行为数据中视频观看行为的特点，提出了到课率和利用率指标衡量学生对课程视频资源的使用情况，并将其拓展到视频与学生的其他属性，解决了跨属性对比分析视频利用情况的指标一致性问题；然后，提出了一种基于球面布局的多属性数据可视化方法，采用球面坐标编码视频和学生的多个属性，从不同尺度同时展示整体与局部各属性的视频利用模式；同时，提出了固定参考点的导航工具与关联视图，实现任意视角和多层面的自由探索多属性数据的分布情况。最后，基于上述方法开发了 VUSphere 可视分析系统，并在大规模真实数据集上验证了所提可视化方法的有效性，通过不同侧面的案例分析发现了课程和学生的视频资源利用模式。

关 键 词：在线学习行为数据，可视分析，学习参与度，时间管理，视频利用模式

论文类型：应用基础

ABSTRACT

In recent years, with the rapid development of internet technologies and the popularity of personal mobile devices, the scale of online learning has grown exponentially and it has produced a huge amount of online learning behavior data. However, with the compound growth of learner population and teaching methods, the massive online learning behavior data poses some challenges to explore and analyze its hidden laws and patterns. More specifically, first, the improvement of internet technologies makes the teaching methods more abundant and makes the online learning behavior data contain higher dimensional characteristics. Existing educational data analysis tools and cluster analysis methods are difficult to characterize and explain high-dimensional learning engagement patterns. Second, the long-term learning process makes the online learning behavior data naturally have time-series information. However, the commonly used methods are difficult to describe the time management style in the data; Finally, the increase of teaching resources and the number of students makes the differences in attributes more significant. Static statistical charts are difficult to explore the multiple attribute distribution patterns of teaching resources and students in large-scale learning behavior data. To address the above challenges, this study proposes the following online learning behavior data analysis methods based on visual analytics, which is helpful to visualize the information of learning engagement pattern, time management style and resource utilization distribution in large-scale online learning behavior data, and to provide support for improving teaching services and teaching quality. The main contributions of this study are as follows :

First, this dissertation proposes a learning engagement visual analysis method based on scatter plot matrix and t-SNE data dimensionality reduction technology. In the first place, this dissertation designs a hierarchical online learning behavior data storage model, and designs storage system based on the model to solve the problems of collecting, clearning, abstracting and extracting the learning engagement indicators. Then, this dissertation proposes a learning engagement indicator model to describe the complex online learning features, and proposes a high-dimensional data analysis method based on scatter plot matrix and t-SNE algorithm and interactive tools based on DBSCAN algorithm to explore the characteristics of learning engagement patterns. Finally, in order to verify the effectiveness of the our proposed methods, the online learning behavior data of online education learners are studied and several learning engagement patterns are founded.

This research was partially supported by ”The Fundamental Theory and Applications of Big Data with Knowledge Engineering” under the National Key Research and Development Program of China with Grant No. 2018YFB1004500, the MOE Innovation Research Team No. IRT17R86, the National Science Foundation of China under Grant Nos. 61721002, 61502379, 61532015, and Project of China Knowledge Centre for Engineering Science and Technology.

Second, on the basis of previous study, this dissertation further proposes a visual analysis method for learning engagement time-series features based on calendar coordinate system. First, this dissertation proposes a data model to describe the time-series information in long-term online learning process, which abstracts the complex changes in learning engagement over time into a unified engagement time-series matrix. Then, this dissertation designs a grade point prediction model based on convolutional neural network, which captures the time information contained in the column vectors and the learning engagement information in the row vectors. Then, a calendar-based learning time distribution map is designed to show the time patterns of the learning process, which can represents the periodic changes of engagement in the day, week, month and semeseter. In addition, a multi-course learning progress chart is designed to explore the time patterns of watching multiple courses. Finally, with the real online learning behavior data, this dissertation designs the LearnerExp visual analysis system to verify the effectiveness of the visual design with real data. The time management style of students in the long-term learning process are also discussed.

Third, this dissertation proposes a visual analysis method for video utilization based on spherical layout, and further extend previous two studies to large-scale, multi-attribute online learning behavior data. First, this dissertation proposes the attendance rate and utilization rate to measure the students' utilization of video resources and extend these two indicators to other attributes of videos and students, which build a unified basis for video utilization comparison between different attributes. Then, based on the attendance rate and utilization rate, this dissertation proposes a visual analysis method based on spherical layout to show the overall and local video utilization patterns from different perspectives, which encodes the multiple attributes of video and student to show the overall and local video utilization patterns from different scales. Finally, this dissertation designs the VUSphere visual analysis system and demonstrates the effectiveness of the design on a real large-scale dataset by exploring the video resource utilization patterns of courses and students from various aspects.

KEY WORDS: Online Learning Behavior Data, Visual Analytics, Learning Engagement, Time Management, Video Utilization Pattern

TYPE OF DISSERTATION: Application Fundamentals

目 录

CONTENTS

1 绪论

本章首先介绍本文的研究背景、面临的挑战、研究的目的和意义。然后介绍国内外研究现状，主要研究内容、研究框架、以及本文的组织结构。

1.1 研究背景与意义

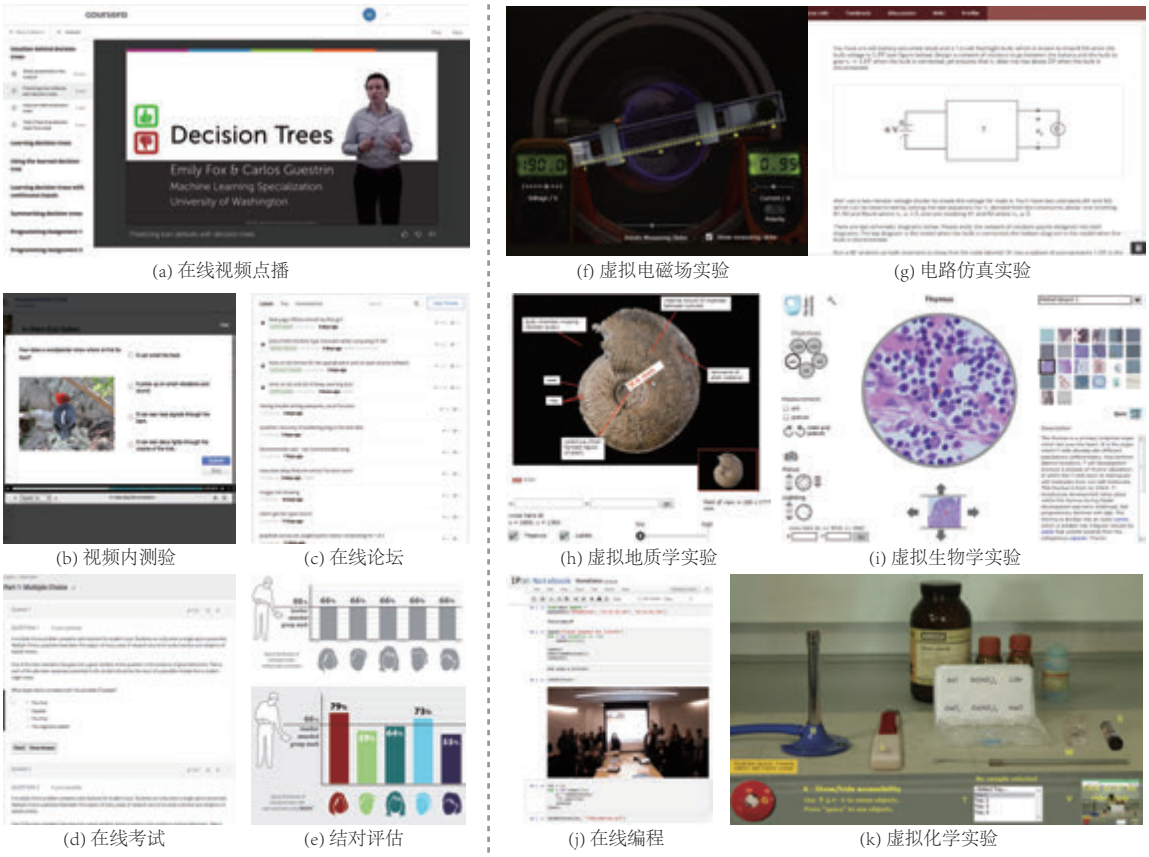
随着互联网技术快速发展以及个人移动计算终端设备的广泛普及，在线学习已经成为补充专业知识、拓展职业技能、提升学历水平、实现自我发展、以及实现终身教育的重要途径和手段[3]。在这个过程中，在线学习行为数据的规模呈现指数级增长，并与整个在线学习过程相互交织影响。一方面，在线视频、学生论坛、虚拟实验以及在线测验考试等多样化的在线教学活动不断涌现，海量的教育数据随之源源不断产生，其中以学习行为日志为代表的在线学习行为数据规模最为庞大。以麻省理工学院 2012 年发布的电路与电子学课程为例，仅在 2012 年春季一个学期就吸引了超过 15 万的全球学习者注册学习，并随之产生了超过 2.3 亿条学习行为数据[1]。在随后的 4 年里麻省理工学院联合哈佛大学又发布了 290 门课程，吸引超过 450 万的全球学习者注册学习，并产生超过 23 亿条学习行为数据[2]。另一方面，随着数据采集、存储与分析技术的不断提升，这些在线学习行为数据也进一步为理解在线学习特征、解决在线学习需求、生产高质量教学资源、以及营造高效在线学习环境等方面提供了有力的支撑。已有研究开始利用这些数据在多个方面展开研究。例如，2013 年《Nature》期刊的专题《Learning in Digital Age》讨论了 MOOC 在美国高校以及全球各地的应用情况，并提出需要使用大数据分析技术处理大规模的在线学习行为数据[3]。同期美国教育部发布研究报告《Expanding Evidence Approaches for Measuring Learning in Digital World》，提出要在未来的在线教育中利用大数据分析方法，辅助建立个性化的学习系统为学习者提供学习资源[4]。通过深入挖掘和分析这些在线学习行为数据，能够开发多种数据驱动的教学辅助功能，例如学习者画像[5]、成绩预测[6]、智能导学[7]、学习路径和资源推荐[8] 等，为提高在线教育的教学质量提供重要的数据与技术支撑。

大规模在线学习行为数据中蕴含着大量学习模式与教育规律，但是由于其飞速增长的规模与丰富多样的内涵，分析这些数据不仅需要海量的存储能力与强大的计算能力，还需要有效的数据分析方法对其展开探索。因此，如何挖掘和呈现潜藏其中的模式与规律，进而为教育者和学习者提供服务是急需解决的问题。以下将针对在线学习行为数据分析的挑战展开讨论。

1.1.1 在线学习行为数据分析的挑战

得益于人机交互技术的迅猛发展，在线学习中多样化的教学功能呈现出显著增加的趋势[9, 10]。来自不同学科的各种课程除了通过在线视频[11]、视频内练习[12]、在线论坛[13] 以及结对评估[14, 15] 等通用教学功能（图 1-1(a-e)）传授课程知识以外，还根据课

程自身学科特点与教学目标，设计了全新的在线教学功能。例如，在麻省理工学院开设的电路与电子学课程中，开发了基于 HTML5（Hyper Text Markup Language 5th version）技术的在线电路仿真模拟实验室，使学生可以在线设计、搭建、检验各种虚拟电器原件，学习相关的电路原理[16]（图 1-1(g)）。英国开放大学的电气工程（图 1-1(f)）、地球科学（图 1-1(h)）、生物学（图 1-1(i)）、以及化学（图 1-1(k)）等各类课程，分别针对课程知识特点设计了不同的在线仿真实验，帮助学生了解各课程的核心概念[17]。另外，主流 MOOC 平台上的编程语言类、统计类与机器学习类课程均采用在线编程环境 Jupyter Notebook 辅助学习者编写程序从而训练编程能力。目前该环境已经支持 Python，R，Mathematica，Maple 和 MATLAB 等编程语言，并支持调用服务器端的图形计算资源，开展深度学习、数据挖掘方面的训练[18–20]（图 1-1(j)）。



(a) 在线视频点播　(f) 虚拟电磁场实验　(g) 电路仿真实验

(b) 视频内测验　(c) 在线论坛　(h) 虚拟地质学实验　(i) 虚拟生物学实验

(d) 在线考试　(e) 结对评估　(j) 在线编程　(k) 虚拟化学实验

通用教学功能　课程定制的教学功能

图 1-1 多样化的在线教学过程

随着在线学习人数与在线教学功能数量的双重增长，在线学习行为数据在数据规模与格式上呈现出特征维度高、时间周期长、大规模多属性相互关联的特点，并且这些特点也对在线学习行为数据的分析提出了挑战：

1) **特征维度高**。一方面，由于上述教学功能的多样化与学习行为数据采集点的增加，在线学习行为数据格式从只包含交互过程记录的单一结构化数据，拓展到同时具有结构化数据，半结构化数据与非结构数据的复合数据。例如交互行为记录、讨论文本、实验记录图像、过程视频、作业程序代码等，因此，多样化的来源与丰富的数据

格式使得原始数据的维度不断增加；另一方面，刻画在线学习行为的特征指标数量也显著增加。每一个教学功能都可能伴随着数十个刻画学生学习过程特征指标，而每个课程又包含多个教学功能，这使得反映学生完整学习过程的特征指标维度进一步增长。探索高维度的学习过程特征有助于教师识别不同学习模式的差异，进而辅助不同学习模式的学生提高学习收益。但是，异构多样的在线学习行为数据难以管理，同时高维度的学习过程特征不容易直观理解，也难以识别其中各学生群体的学习模式差异和相关影响因素。

2) **时间周期长**。根据课程内容的不同，一门在线课程的学习时间约为 4-18 周。而随着越来越多的课程开始以在线学习的形式传播，学习时间更长的长周期多课程的在线学习正在兴起，如专项技能培训和专业学历教育等。以 Coursera 平台的计算机科学硕士项目为例，学生需要在 12-36 个月内获得多门在线课程的学分以获得硕士学位。在如此长期的在线学习过程中，学习时间的安排和学习进度的管理会直接影响课程知识掌握与最终的学习成效，不仅教师需要了解学生在长期学习过程中的时间管理情况，学生也需要掌握和调整自己的学习时间安排。尽管在线学习行为数据完整记录了学生长期学习过程中所有活动并包含时间字段记录，但是如何从这些数据的时间字段记录中刻画学习过程的时间特征，并识别其中蕴含的在线学习时间管理特点需要进一步深入研究。

3) **大规模多属性相互关联**。在线学习过程涉及的学生与教学资源双方都具有多个相互关联的属性，例如学生的所在地区、入学机构、专业、课程表，教学资源的课程编排、学科专业、难度类型等。这些相互关联的属性之间存在层次化交织又相互影响的关系，并直接反映在在线学习行为数据中，使得兼顾不同层面的数据探索分析变得困难。特别是当学生数量与教学资源规模都非常庞大时，难以同时在不同尺度、不同层面探索和分析学生和教学资源之间的相互关系与影响。

在线学习行为数据的这些特点为数据分析带来了三个个挑战：第一，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；第二，如何刻画长期的在线学习过程中的时间特征并向教师和学生展示；第三，如何分析大规模的不同尺度与层面下的学生对教学资源的利用模式。这些数据分析的挑战涉及学生、课程、资源以及管理等多种相互关联的因素，并且数据分析的结果需要被教师、管理人员以及学生理解，因此迫切需要一种直观的、交互的数据分析方法帮助用户充分挖掘在线学习行为数据中蕴含的价值。

1.1.2 在线学习过程数据的可视分析

根据分析技术的不同，在线学习行为数据的分析可以分为三类：基于教育学经验模型的方法，基于机器学习的方法，以及基于可视分析的方法。教育学经验模型方法侧重根据先验知识或者假设建立模型，然后依据模型收集数据进行分析。例如，在分析学生在线学习参与程度影响因素的研究中，常用的方法是通过设计包含多个关联问题

https://www.coursera.org/courses?query=master

的调查问卷，从受控样本中调研获得关于研究问题模型的数据，再基于结构方程模型（Structural Equation Model）或者线性回归分析等方法检验初始假设[4]。然而，在大规模的在线课程中，由于学生的匿名性，其个人背景差异极大，小样本采样可能存在样本偏差，难以确保问卷数据的真实性和可靠性；基于机器学习的方法能够使用大量样本数据训练计算模型，并让计算模型从数据中自动的提取知识[21]。但是，由于在线学习行为数据的格式多样化与高维度，潜在的学习模式往往并不确定，并且会随课程、学生、时间等因素发生变化，难以套用固定的计算模型从多样化的数据中直接挖掘模式信息。因此，基于机器学习的方法会针对一个具体问题设计专门的计算模型，从而自动高效的在数据中提取模式信息完成预定的分析任务。与前两种方法不同，可视分析方法降低了样本数据与模型之间耦合关系的紧密程度，用户可以通过可视化界面呈现的不同层面数据特征，结合交互技术提供的数据操作，运用教育专家的知识和经验，探索数据中潜在的模式与规律，为其他两类方法提供启发和引导。

可视分析是一种“交互式可视化界面促进的分析推理科学”[22, 23]，它通过分析推理技术、可视化表示技术、交互技术、数据挖掘技术等多种跨学科领域技术的紧密结合[24]，将领域专家的专业知识和业务经验通过交互操作与视觉呈现带入到数据分析的过程中[25]。目前可视分析方法已经应用在多个领域中，包括天气气象研究[26]、医学影像分析[27]、电子商务交易[28]、体育运动分析[29]、社交媒体分析[30] 以及城市交通管理[31]等。

可视分析的基础在于数据可视化，Bertini 和 Lalanne [32] 讨论了数据分析中数据挖掘过程与可视化过程的特点与差异，如图 1-2所示。从整体步骤上看，数据挖掘过程与可视化过程都是以数据为起点，以验证假设获得领域知识为目标和终点，但是在实现目标的过程和中间结果上存在差异。数据挖掘过程通过数据训练、建模等方法得到关于数据的计算模型，通过解读各种分析指标验证问题假设。而可视化过程则是通过映射视觉编码的方式，将数据从抽象的数学空间映射到人类可直观感知的视觉空间，进而利用领域专家和用户的业务知识与经验，通过探索分析视觉空间中呈现的视觉模式检验问题假设获得知识[33]。

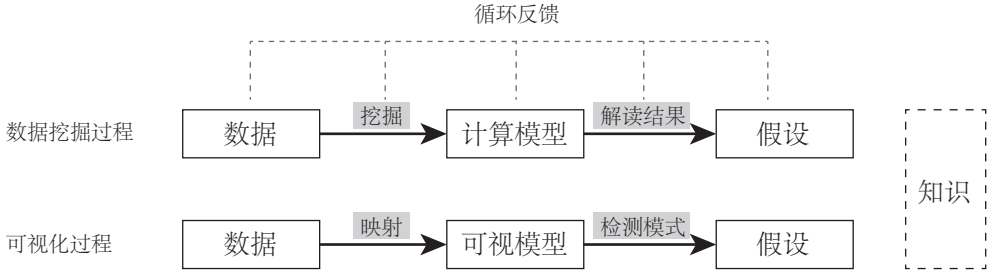


图 1-2 数据挖掘与可视化过程的区别[32]

随着现实世界数据规模和复杂度的迅猛增长，特别是大数据时代的到来，数据挖掘方法得到的抽象知识与人类可理解的知识之间存在显著差异，如何分析和解释这两

种知识之间的差异需要领域专家的专业经验与领域知识。因此，结合人脑智能与机器智能的可视分析成为一种自然的解决方法[34]。Keim 等[35] 认为可视分析的基本流程是将人类掌握的知识和机器自动分析的结果相结合，经过人类的视觉通道，形成数据在人脑和机器的双向转换，特别是将人类知识和个性化经验通过可视手段融入到整个数据分析和推理决策的过程中。并且文献 [35] 针对可视化技术的特点，提出了可视分析方法的基本流程，如图 1-3所示。

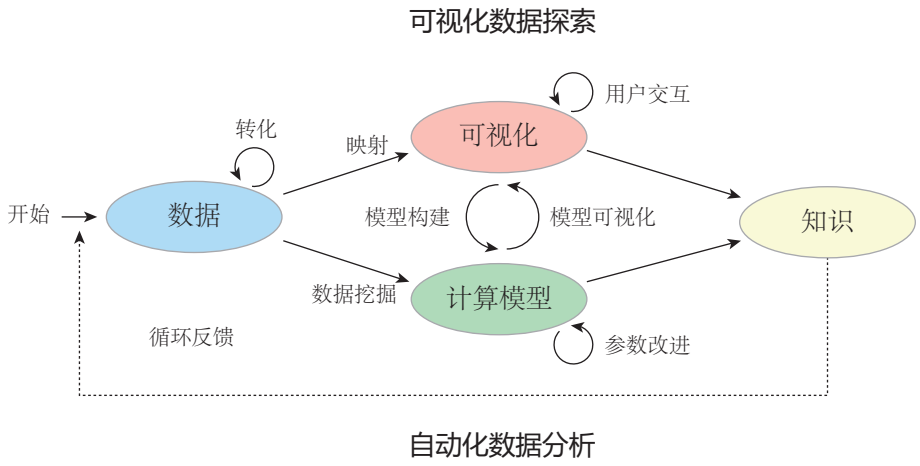


图 1-3 可视分析的过程[35]

在可视分析的基本流程中，起点仍是输入的数据，终点是提炼的知识。通过基于可视化方法的数据探索和基于数据挖掘方法自动化的数据分析这两条途径的交互迭代，实现从数据输入到知识提炼的过程。在两类途径的交织过程中，可视化交互探索中得到的视觉模式与数据模型得到的计算模式能够相互支持。在可视化探索的过程中，通过组合可视化方法的七个基本抽象操作[36]：概览（Overview），缩放（Zoom），过滤（Filter），详查（Details-on-Demand），关联（Relate），历史（History），提取（Extract），领域专家不仅可以自顶向下的从数据的整体分布出发，通过多种交互操作逐步缩小观察的范围，深入到数据分布的细节与规律中，也可以自底向上的在部分数据中继续探索性数据分析。领域专家借助直观的动态图表与多种交互式的分析工具[37]，总结数据在局部的规律模式及影响因素，进而推广到更大的范围[38]。在运用数据挖掘方法的过程中，领域专家既可以利用可视化探索计算结果从而指导计算模型的设计和改进，也可以调整计算模型的参数验证可视化结果。在这两条路径的迭代过程中，数据挖掘过程的计算模型与可视化模型之间彼此相互补充，相互促进，交互可视化方法产生的中间结果可以用于改进计算模型的参数和结构，而挖掘得到的模型和模式可以通过可视化加深用户理解。通过这个双向过程的循环往复，最后实现对复杂数据的探索分析得到领域知识[39]。

如前一节所述，由于在网络课程的学习过程中课程数量、学生规模以及教学功能的多重复合增长，使得在线学习行为数据的体量庞大且复杂多样，并且，为了从在线学习行为数据中挖掘学习行为的模式与规律仍面临着三个挑战。针对上述挑战，本文以

在线学习行为数据为分析对象，提出了以可视分析为主要手段的数据分析方法，通过综合运用可视化技术、交互技术以及数据挖掘相关技术，帮助教师、学生以及教学相关人员直观的探索高维度、长周期、大规模多属性关联的在线学习行为数据，发现其中潜在的规律与模式，为理解在线学习过程、提高学习效果、改进教学质量提供技术手段和数据支持。

1.1.3 研究目的和意义

当前主流的在线学习平台，如 edX，Coursera，以及学习管理系统 Moodle，Blackboard 等主要采用以数值统计和基本图表为基础的在线学习行为数据分析，难以应对日益增长的数据规模与数据分析需求。本文所做研究工作将可视分析方法引入在线学习行为数据分析，有助于丰富和完善在线学习行为数据分析的技术与方法，并为构建新型的数据驱动的个性化学习服务平台提供理论与关键技术支撑。

1.2 国内外研究现状

本文的研究对象是在线学习行为数据，针对前一节讨论的在线学习行为数据的特点，本节将分别从高维数据可视分析、时序数据可视分析、以及在线学习行为数据可视分析三个方面，总结与本文密切相关的重要研究。

首先，在高维数据可视分析方面，针对高维数据的可视分析通常包括数据转换（Data Transformation），视觉映射（Visual Mapping），以及视图转换（View Transformation）这三个步骤[40]。其中，数据转换步骤中的数据降维是与数据可视化紧密相关的一类任务，通常可以采用特征选择（Feature Selection）或者特征提取（Feature Extraction）这两类方法降低数据的维度 [41]。视觉映射是利用可视化与交互技术将数据转化为图形化的过程，包括位置、大小、图形、符号、颜色以及纹理等。视图转换是利用计算机图形技术将可视化设计与交互综合渲染呈现为一个或多个相关联的视图，从而将可视化设计最终呈现给用户。

高维数据可视分析的相关研究会根据数据的特点与分析任务，选择多个技术的组合实现分析目标。例如，通过概览 + 详情，鱼眼缩放、焦点 + 上下文等交互技术探索高维数据的整体分布情况和局部子空间分布情况，进而加深对高维数据特征模式的理解[42]。Zhao 等[43] 采用 t-SNE 方法将 NBA 运动数据投影在二维平面，并针对运动过程的主要指标设计了球员能力符号，再结合其他多种视图综合分析 NBA 的比赛统计数据轮廓。Lee 等[44] 在 MDS 算法的基础上，综合平行坐标系反映的数据结构信息衡量数据点之间的距离，辅助用户可视化的探索高维数据的分布。Wei 等[45] 采用 SOM 算法将电子商务平台中的用户点击流数据（Clickstream Data）降至二维平面，再通过聚类算法分析用户的购买模式，并通过其他关联视图展示低维分布中各集群在主要原始特征上的分布情况。Sucontphunt 等[46] 采用 PCA 方法将面部表情高维特征降维到三维空间的实例，通过对表情的颜色与位置编码帮助用户分析不同表情的差异与特点。另外，在教

育数据的分析中，特征选择方法是较为常用的一类方法[21]，例如 Li 等[47] 采用相关性分析方法从 14 个反映学习过程的教学资源使用情况指标中选择了 4 个指标作为分析对象，并进一步使用这 4 个指标使用 k-means 方法展开聚类分析。类似的，Cerezo 等[48] 从学习系统提供的 76 个指标中选择了 6 个指标用来分析学生的学习参与度模式。综上所述，现有数据降维方法能够较好的保留高维数据局部特征，并通过可视化方法呈现低维数据的分布规律。然而，教育数据分析对学习行为特征指标与分析结果的可解释性有较高要求，使用非线性数据降维技术得到的低维学习行为特征不具有实际的现实含义，难以被用户理解和使用。现有的高维数据可视化方法与在线学习行为数据分析需求之间仍存在差距。

其次，在时序数据可视分析方面，针对时序数据的可视分析通常会根据时序数据的时间属性特点，包括线性时间、周期性时间、以及时间点或时间间隔[49] 等，采用多种不同的可视化方法进行展示和分析。例如，对于线性时间模式，Heer 等[50] 提出多水平带图（N-band Horizon Graph）用来反映时序数据在水平线上的涨跌变化情况。Sun 等[30] 基于主题河流图设计了 EvoRiver 可视分析系统用于展示、分析和理解社交媒体上话题之间的随时间变化的竞争合作变化过程。对于周期性时间模式，Shen 和 Ma[51] 结合环状图与柱状图反映移动数据中用户在一周和一天的活动规律。Woodring 和 Shen[52] 基于小波变换方法将时序数据转换到不同时间尺度的曲线集合，并采用折线图矩阵的形式展现时序数据的不同周期模式。

通过分析时序数据的可视化研究工作可以发现，针对时序数据可以利用数据中时间属性的特点，选择合适的可视化设计展现数据中的时间属性，结合符号设计和其他图表展现数据中的时间模式和周期性规律。然而，现有针对在线学习行为数据中时序特征的可视分析仍然较少，并且现有的河流图、环状图等可视化方法难以呈现长期学习过程中的短期和长期时间管理特征。

最后，在线学习行为数据的可视分析研究目前相对较少，其研究对象主要是针对具体的在线学习行为，包括视频观看行为分析、讨论行为分析以及其他教学活动分析等。例如，Qu 和 Chen[53] 采用可视化方法对 MOOC 课程的学习者来源、完成率、课程领域等方面的数据展开分析。Zhao 等[54] 设计了一个可视化推荐系统 MOOCex，该系统能够基于视频、课程之间的知识点关联以及学生当前的学习状况向学生推荐课件视频。Chen 等[55] 提出了 ViSeq 可视分析系统从不同粒度和层次探索学习顺序。Fu 等[56] 提出了一种群组检测与排序算法，并开发了 VisForum 可视分析系统探索网络教育在线论坛中的学生群组，通过不同粒度的群组符号、用户符号以及集合符号，用户可以分析论坛里具有较高影响力的文章与成员。Meng 等[57] 针对在线编程学习过程中，测验题库的题目筛选问题开发了 PeerLens 可视分析系统分析编程练习题的学习路径。上述研究利用可视分析方法，从不同侧面细粒度的分析了在线学习行为数据中隐含的学习过程模式与特点，从而为深入理解特定的在线学习行为模式与规律提供了支持。但是，现有研究主要关注单个课程或少量课程在短期内的视频利用情况，而对大规模视频长期的、多层面的利用模式的分析较少。

1.3 论文的主要工作

论文的主要工作是面向在线学习行为数据，分别从高维学习参与度可视分析、时间管理可视分析、大规模多属性关联的视频利用模式可视分析这三个方面展开研究，论文的整体框架如图 1-4所示。

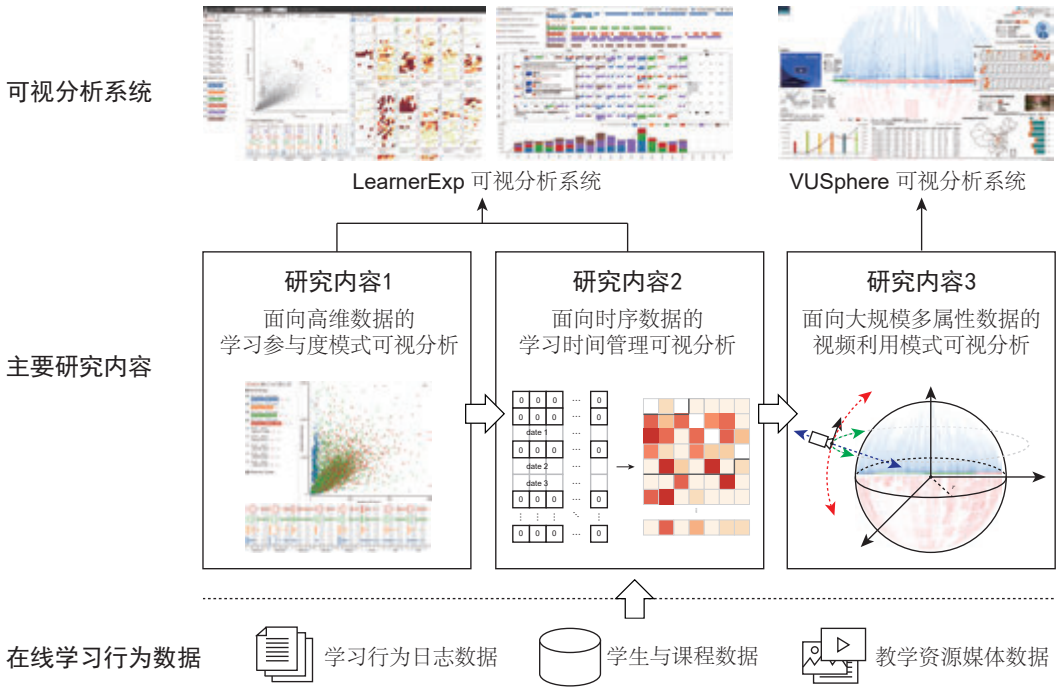


图 1-4 论文整体的研究框架图

1）**面向高维数据的学习参与度模式可视分析**。通过学习参与度指标可以刻画学生在线学习过程的学习模式，然而从在线学习行为数据中提取的学习参与度具有较高的特征维度，难以直观探索学习参与度分布模式差异以及影响因素。因此，学习参与度模式可视分析主要包括两个任务。第一，学习参与度高维度特征刻画。该任务主要实现从在线学习行为数据中提取反映学习过程特征的学习参与度指标模型，并将高维度的学习参与度特征映射为可呈现的视觉编码。第二，学习参与度模式解释。该任务主要用来发现不同的学习参与度模式关联的影响因素。本研究利用大数据存储计算平台与可视化技术，提出一种面向高维数据的学习参与度模式可视分析方法，首先，本文提出了层次化的在线学习行为数据的存储模型，并在此基础上设计了基于该模型设计了在线学习行为数据管理系统；其次，本文提出了基于散点图矩阵和 t-SNE 降维方法的可视化方法，将学习参与度模式直观展现为易于理解的散点图形式。最后，通过颜色编码与关联多视图查看不同因素与学习参与度模式之间的联系。利用上述可视化设计，本文在真实的在线学习行为数据集上展开研究，分析在线学生支持服务对于学生参与度的影响。

2）**面向时序数据的学习时间管理可视分析**。由于网络课程的学习具有长周期性以及时间安排灵活性，学生需要根据课程学习进度自行安排自己的学习时间。然而，不同的时间管理风格可能对学生的学习时间安排与学习进度产生影响，并进一步影响学习

成效。因此，本研究在前一研究中学习参与度指标模型的基础上，进一步提出了出学习参与度时序矩阵以保留学习过程的时序特征，并采用基于卷积神经网络的模型预测不同时间管理风格的考试成绩。随后，为了直观理解学习参与度时序矩阵中刻画的学习时序特征，本文设计了基于日历坐标系的学习参与度日历矩阵和基于日历图的学习时间分布图，用以展现长期学习过程中的时间信息与周期变化情况。最后，本文针对在线多课程同期学习的特点，设计了多课程学习进度图，从课程资源和时间两个维度细粒度呈现学习过程。综合上述可视化设计与交互技术，本文开发了 LearnerExp 可视化分析系统，探索长期学习过程中的时间管理特点。

3）**面向大规模多属性数据的视频利用模式可视分析**。在大规模在线学习过程中，学生数量与视频资源数量极大且相互之间多属性关联，难以从整体到局部自顶向下的探索两者之间不同尺度不同层面的规律。针对这一问题，本文在前两个研究成果的基础上，进一步将研究对象拓展到大规模长周期多个专业的在线学习行为数据。首先以在线学习过程中视频观看行为这一核心交互行为作为对象，抽象出多尺度的视频利用率指标用以衡量学生和视频资源双方在学习过程中的表现，反映不同属性之间的共性特征。然后，本文设计了基于球面布局的视频利用情况分布图，将学生与视频资源的主要属性映射为多种视觉编码，实现全面展示所有学生与视频资源的利用率分布规律，识别整体视频利用率模式和局部的异常模式。最后，本文基于多种数据图表设计了反映不同尺度局部数据的多个关联视图，通过固定参考点的导航工具等交互技术深入探索学生和视频资源的细粒度利用模式，实现自顶向下和自底向上的双向探索分析，从而直观的展现在线学习行为数据中多个属性的视频利用模式。综合上述可视化设计与交互技术，本文开发了 VUSphere 可视分析系统，探索大规模的学生与视频资源在多个属性的利用情况分布。

1.4 论文组织结构

本论文共分为六章，针对在线学习行为数据的可视分析问题，以数据特征和可视分析方法的研究思路组织全文和各章内容，具体章节安排如下：

第 1 章绪论给出了本研究课题产生的背景，介绍了在线学习行为数据的特点与数据分析工作面临的挑战，并简要阐述了国内外关于可视分析与在线学习行为数据分析方面的工作，分析了本课题所要解决的研究问题、主要贡献、研究思路与研究框架。

第 2 章相关工作着重介绍了高维数据可视分析、时序数据可视分析以及在线学习行为数据可视分析这三个方面的相关研究，以及上述领域已有工作和本研究的关系，为本文的相关研究提供理论基础和参考依据。

第 3 章研究了学习参与度高维数据的可视分析方法，主要提出了一个学习参与度指标模型以及基于散点图矩阵和 t-SNE 算法的学习参与度可视分析方法，并通过集成多种学生群组创建工具与协调关联的多个视图，分析在线学习行为数据的高维学习参与度分布模式与影响因素。

第 4 章研究了学习时间管理时序数据的可视分析方法，主要针对在线学习过程的长期性特点，提出了基于日历坐标系的学习参与度时序特征可视分析方法，包括刻画在线学习行为数据时序特征的学习参与度时序矩阵、基于日历图的学习时间分布图以及多课程视频观看进度图等，分析长期在线学习过程中的学习时间管理特点。

第 5 章研究了大规模多属性的视频利用率模式的可视分析方法，主要提出基于球面布局的视频资源利用情况分布可视分析方法，包括视频资源的多尺度利用率指标、基于球面布局的概览视图以及多属性详情视图等多个关联视图，实现从不同尺度与层面对关联属性的视频利用模式展开探索性分析。

第 6 章为结论与展望。本章对论文的各项研究进行了总结，并提出了论文的主要贡献，最后对未来的研究方向进行了展望。

2 相关工作

海量数据的探索分析需要综合利用统计学、机器学习、数据库等技术，使用计算机编程语言编写复杂的分析程序才能开展，因此要求分析人员必须具备相应的技术基础。然而，迫切需要分析业务数据的领域专家通常只具备领域内的专业知识和业务经验，他们不得不经过较长时间的计算机数据分析技术培训才能掌握相关编程语言和数据挖掘软件，难以直接对大规模业务数据开展分析。

为了解决海量数据分析需求与分析手段之间的鸿沟，可视分析应运而生。可视分析通过结合图形化的用户界面、交互式的操作体验以及针对领域知识建立的数据模型等多个方面要素，构建针对具体分析问题的可视分析系统，实现业务数据到视觉编码的映射并提供交互式的分析界面，让领域专家能够专注于分析任务本身，而不需要深入了解底层所采用的计算机相关技术细节。利用可视分析系统，领域专家能够直观的查看和操作数据，从而探索分析潜藏在海量数据中的规律和模式。相比于已经广泛应用的数据挖掘方法，可视分析方法通过运用可视化技术与交互技术，将人类经验和领域知识集成到数据挖掘的过程中，从而提高数据挖掘过程的表现，更好的发现和促进知识的产生[32]。这一过程如图 2-1所示：

图 2-1 可视化与数据挖掘过程的结合[32]

通过在数据建模、模式分析以及模型评价等各个环节运用可视化技术和交互技术，领域专家可以直观的理解数据，并将领域知识与专业经验运用在分析过程中，从而发现潜藏在数据中的规律和模式。由于可视化技术通常会针对用户的现实需求进行设计，所以需要通过严谨的用户评测实验评估可视化方法与结果，为可视化技术的应用提供有效性验证[49]，从而评测可视化技术是否能够更好的帮助用户解读数据。目前可视化有效性和用户体验方面的评测主要来自于人机交互（Human-Computer Interaction，HCI）领域的评测方法，常见的评测方法包括：用户实验（User Studies），专家评估（Expert Review），案例研究（Case Studies / Use Cases），指标评估（Metrics）和众包（Crowdsourcing）等[49]。根据研究对象和目标的不同，评测采用的具体方法会发生改变，但大体都遵循以下基本流程：定义评估研究目的和问题、选择评估方法、设计实验、采集数据、分析实验结果、验证研究假设并得到结论。在本文的三个研究内容中，可视化系统的目标用户包括网络教育技术支持人员与授课教师，他们具有计算机专业技术背景以及在线教学相关

的经验。另外，本文访谈的领域专家长期参与了本研究的全部过程，对研究问题与分析任务需求具有较为深入的理解。因此，本文采用了领域专家用户参与评估的案例分析方法对可视化设计的有效性与用户体验进行评测，全面分析可视化设计对各项分析任务的完成情况。

目前针对特定领域的可视分析方法主要遵循 Munzner [58] 提出的层次化嵌套模型，如图 2-2所示。

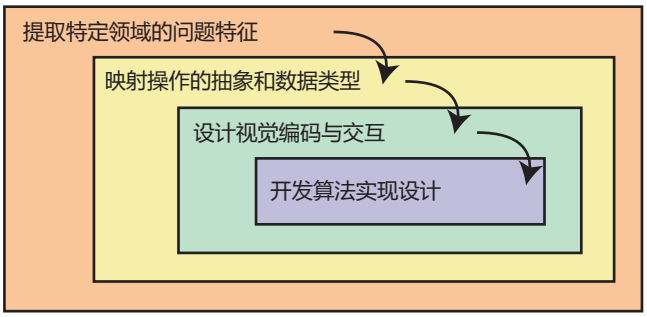


图 2-2 可视化设计的层次模型[58]

该层次化嵌套模型包括四个步骤：首先，可视分析通常与具体的领域问题相结合，针对特定问题进行可视化设计并构建可视分析系统，因此需要针对特定领域问题，通过访谈专家用户和目标用户，提取分析任务，分析问题的特征；其次，针对目标用户的特点和要求，在领域问题和分析任务的基础上，将数据和问题抽象为数据类型和操作；然后，根据数据类型的特点和操作的需求，设计相应的视觉编码与交互方法；最后，设计高效的算法实现上述的数据映射与交互方法。本文的研究工作同样遵循了上述分析模型，并且在整个数据分析与系统设计过程中紧密结合领域专家与目标用户，通过专家访谈与实地研究等方式验证可视化设计的有效性。通过这个模型可以发现，可视分析围绕着具体应用领域逐步展开深入，并且可视化与交互技术的设计与编码实现均与数据的特性紧密相关，因此，领域问题与数据的特征直接影响所选择的技术与实现。

针对这一特点，结合本研究涉及的领域问题与数据特点，本章综述了与本研究工作密切相关的国内外研究现状，分别从高维数据可视分析、时序数据可视分析以及在线学习行为数据可视分析这三个方面综述各领域的典型方法和研究思路，并分析已有工作和本文研究工作之间的关系。

2.1 高维数据可视分析

随着各领域数据产生能力的日益增强，数据的规模与形式与日俱增，人们迫切需要挖掘海量数据中隐藏的重要价值。其中，大量的模式信息隐藏在高维度数据的高维特征中，如何分析和挖掘高维数据成为数据科学以及其他各领域所关注的热点问题。其中，利用可视化技术提供的直观图形化展示，能够有效促进对高维数据特征的理解和

改进机器学习模型，从而深入探索高维数据的分布与模式。

通常针对高维数据的可视分析方法主要包括三个步骤，数据转换（Data Transformation），视觉映射（Visual Mapping），以及视图转换（View Transformation）[40]。首先，在数据转换的过程中，通常包括数据降维、回归分析、子空间聚类[59, 60] 等任务，其中数据降维是其他后续分析任务以及可视化的基础。当数据维度较高时，直接展示全部维度信息可能引起视觉混乱，造成数据理解困难。因此，需要利用数据降维方法在减少数据的维度的同时，保留数据的分布特征以满足分析和显示的需求。其次，在视觉映射的过程中，通常将数据的多个维度映射到不同的视觉编码通道，包括位置、大小、图形、符号、颜色以及纹理等，通过组合多个视觉编码通道和图表形式实现同时展示数据的多个维度。最后，视图转换是利用计算机图形技术将可视化设计与交互综合渲染呈现为一个或多个相关联的视图。通过用户与视图的交互操作，结合领域知识与视图中呈现的数据效果，实现对领域问题的分析和推理，从而理解数据的内涵和规律。以下将重点讨论本研究所关注的高维数据降维方法与视觉映射方法。

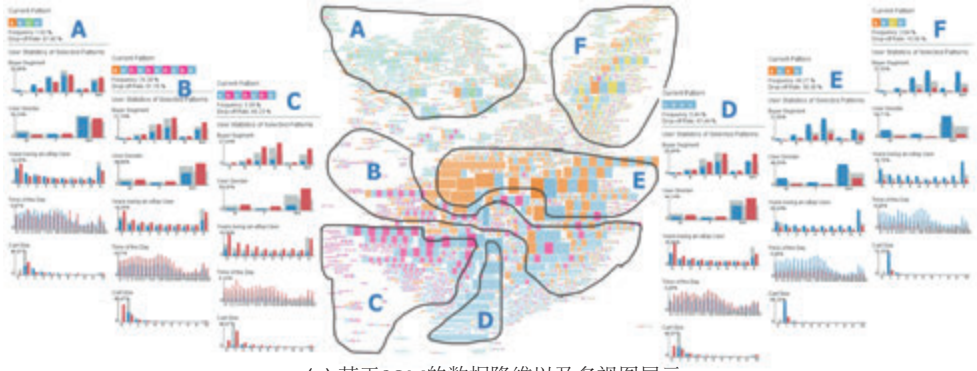
2.1.1 高维数据降维方法

针对维度较高的数据，可以通过特征选择（Feature Selection）或者特征提取（Feature Extraction）这两类方法降低数据的维度 [41, 61]。其中，特征选择是从数据的 M 个特征中按照一定标准或方法选择 $N(N < M)$ 个特征的过程。而特征提取是从初始特征集合通过线性或者非线性的方法，将原始数据从高维空间映射到低维空间中，使得数据的低维表示仍保留原始数据的部分特征。

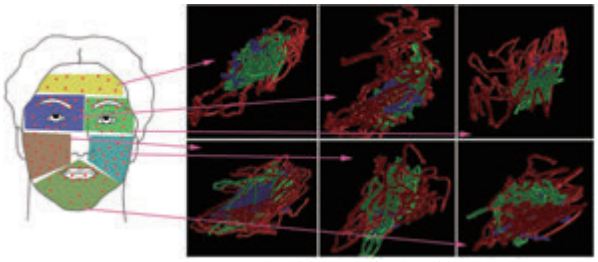
特征选择的方法主要包括 3 类：过滤（Filter）、包装（Wrapper）以及嵌入式方法（Embedd）。其中，基于过滤的方法通过分析特征的分布、相关性[62] 或者信息增益[63] 从特征集合中去除冗余的或者不相关的特征；包装方法根据目标函数的预测效果，逐次从特种集合中选择或者排除若干特征[64]。而嵌入式方法则是一些学习算法在回归或者分类的训练过程中伴随产生的维度降低。通常在教育领域的数据分析中，为了确保分析过程中指标与结果之间的相关性与因果关系，现有研究对指标的可解释性要求较高，因此，基于过滤和包装的特征方法在教育数据分析相关研究中使用较多[21]，例如，Joksimović 等[65] 采用回归分析方法选择与成绩相关的交互特征，从而了解不同类型的交互对于学习成绩的影响。Li 等[47] 采用相关性分析方法从 14 个反映学习过程的教学资源使用情况指标中选择了 4 个指标作为分析对象，并进一步使用这 4 个指标用 k-means 方法展开聚类分析。类似的，Cerezo 等[48] 从学习系统提供的 76 个指标中选择了 6 个代表性指标衡量学生的在线学习活动，并使用聚类方法分析学生的学习参与度模式。

特征提取的方法主要包括线性方法（Linear）以及非线性方法（Non-linear）2 类。其中，线性方法包括主成分分析 PCA（Principal Component Analysis）[66]，独立成分分析 ICA（Independent Component Analysis），线性判别分析 LDA（Linear Discriminant Analysis）以及多尺度分析法 MDS（Multidimensional Scaling）[44] 等。非线性方法包括局部线性嵌入（Locally-linear Embedding，LLE），等距映射（Isomap），自组织映射

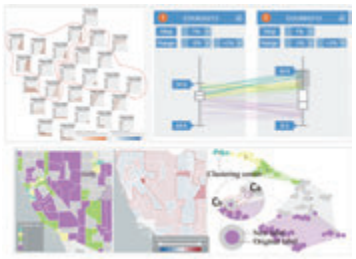
（Self-organizing Map，SOM），自动编码器（Autoencoder），核 PCA（Kernal PCA），以及 t-分布领域嵌入（t-Distributed Stochastic Neighbor Embedding，t-SNE）[67–70] 等。上述方法在可视分析中均有应用，例如，Tatu 等[71] 将高维数据划分为不同的子空间，并采用 MDS 方法将子空间降维投影在二维平面，通过用户交互探索不同子空间的特征进而理解数据的高维模式。Lee 等[44] 在 MDS 算法的基础上，综合平行坐标系反映的数据结构信息衡量数据点之间的距离，辅助用户可视化的探索高维数据的分布。



(a) 基于SOM的数据降维以及多视图展示



(b) 基于PCA的数据降维与特征解释



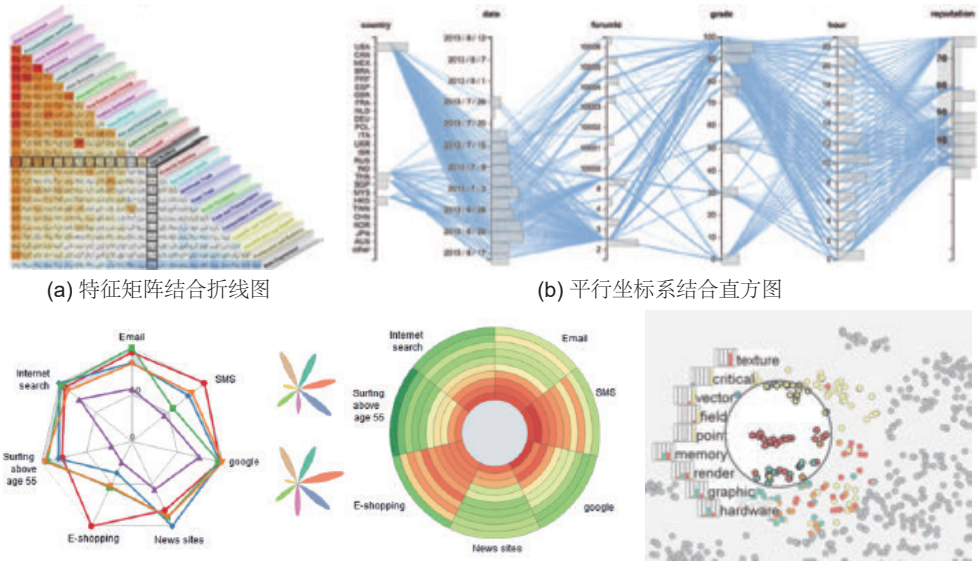
(c) 基于t-SNE的数据降维与多视图

图 2-3 采用数据降维的可视分析。(a)，(b) 和 (c) 分别来自文献 [45]，[46] 和 [72]

特征提取方法在各领域的可视分析中有十分广泛的应用，能够有效的解决维度爆炸带来的分析难题。然而，通过特征提取方法得到的低维特征虽然与原始特征相关并能够表现原始数据的分布，但不再具有原始特征维度的物理意义，难以向直接解释其含义。针对这一问题，现有研究通过运用可视化技术提出了多种方法提高降维后数据的可解释性，包括概览 + 局部统计[73]，降维投影 + 符号[43] 等。例如，如图 2-3(a) 所示，Wei 等[45] 采用 SOM 算法将电子商务平台中的用户点击流数据（Clickstream Data）降至二维平面，再通过聚类算法分析用户的购买模式，并通过其他关联视图展示低维分布中各集群在主要原始特征上的分布情况。图 2-3(b) 展示了 Sucontphunt 等[46] 采用 PCA 方法将面部表情高维特征降维到三维空间的实例，通过对表情的颜色与位置编码帮助用户分析不同表情的差异与特点。如图 2-3(c) 所示，Huang 等[72] 使用 t-SNE 算法将地理区域的高维统计数据映射到二维平面，通过运用颜色编码以及关联视图展示数据在地理以及其他维度的信息。另外，Zhao 等[43] 将篮球运动员的比赛统计指标通过 t-SNE 算法降维，并结合指标符号设计展示运动员的能力特征。

2.1.2 高维数据视觉映射方法

当高维数据的维度较低或者数据量较少时，可以采用多种可视化技术直接展示，包括运用多视觉通道编码[74]，散点图矩阵（Scatterplot Matrix），平行坐标系（Parallel Coordinate），星型图/雷达图（Radar），符号设计（Glyph）以及切尔诺夫脸谱图（Chernoff Face）[75] 等。在这些基本图表的基础上，有研究通过拓展图表设计、组合多种设计以及增加交互等方式进一步展现高维数据的统计特征或者分布规律。例如，Dang 等[76] 将散点图矩阵与折线图结合，展示不同指标随时间的变化情况（图 2-4(a)）。Wu 等[77] 将柱状图与平行坐标系结合，显示数据在每个维度上的分布，同时增加交互式的范围选择器限制输出的数据数量（图 2-4(b)）。Albo 等[78] 分析了星型图/雷达图等多种展示多维数据的图表在用户表现方面的差异，发现雷达图的有效性和用户表现最低，而星型图和圆环图在特定的任务上更好一些（图 2-4(c)）。Heimerl 等[73] 采用概览 + 局部统计交互设计，同时展示数据在二维平面上的分布与局部数据的统计特征（图 2-4(d)）。Palmas 等[79] 将边绑定技术[80]（Edge Bundling）应用在平行坐标系中，缓解高维数据连线密集难以识别的问题，便于用户查看数据在不同维度的分布。



(a) 特征矩阵结合折线图　(b) 平行坐标系结合直方图

(c) 雷达图/花瓣图/环状图　(d) 概览+局部结合图标设计

图 2-4 高维数据的可视化。(a)，(b)，(c) 和 (d) 分别来自文献 [76]，[77]，[78] 和 [73]

通过分析上述研究工作可以发现，针对高维数据可以采用多个关联协调视图，从不同侧面展现数据的分布特点，同时结合数据降维技术，通过概览 + 细节等交互技术分析降维后数据的整体与局部分布规律，为发现潜在的高维数据模式提供支持。

2.2 时序数据可视分析

时序数据广泛存在于自然科学研究、工程项目以及商业环境中，并且随着数据采集能力与数据分析需求的日益增长，因此面向时序数据的可视化方法也成为学术研究

的热点。针对时序数据的时间特征与各类分析任务，如时间特征提取、聚类与相关性分析和模式识别等，已有多个文献[81–83] 对时序数据的可视化方法做了系统性的综述，本文将结合时序数据的特征，讨论部分广泛使用的可视化方法。

(a) 人物年份图 (1765)　(b) 帝国历史疆域图 (1765)　(c) 贸易差额时序图 (1786)

图 2-5 早期反映时间维度信息的时序图。(a) 和 (b) 来自文献 [84]，(c) 来自文献 [85]

从宏观上看，时序数据是以时间轴排列的事件序列数据（Time-series Data），例如股票的价格波动数据，传感器监测数据以及日程安排数据等，同时还有类似新闻文本和蛋白质状态数据这样内在包含时间信息的数据。针对时序数据的可视化分析，按照表现形式通常可以分为两类：一类方法采用静态形式展示数据内容，通过多视图关联、交互等方法展示数据在时间维度的变化与规律[83]。此类图表可追溯到 Joseph Priestley 的人物年份图（图 2-5(a)）和帝国历史疆域图（图 2-5(b)）[84] 以及 William Playfair 的贸易差异时序图（图 2-5(c)）[85]，通过将时间信息映射作为图表中的一个坐标轴，展示其他信息随时间的变化情况。另一类方法采用动画形式展示数据随时间的变化，利用视觉感知连续画面变化的趋势与规律。主流的可视化研究观点认为，动画过程会影响用户的认知过程，需要根据具体问题谨慎的使用动画形式[86]，因此本文主要讨论静态形式的时序数据可视化方法。

时序数据的时间属性可以作为时间轴变量，从而将每个数据事件的各维度映射到数据轴或者其他视觉编码。针对时序数据中时间属性的性质差异，可以采用不同的可视化方法呈现数据中潜在的特征和规律。本节主要讨论线性时间、周期性时间、以及时间点这三种时序数据的可视化方法。

2.2.1 线性时间数据可视化

针对线性时间的时序数据，主流可视化方法通常将时间属性作为二维图表中的一个轴，通过折线图、柱状图、散点图、堆叠图等形式展现其他变量随时间的变化。在此基础上，通过结合多种图表并利用计算机交互技术，产生了众多的衍生变体与扩展。例如，Heer 等[50] 提出多水平带图（N-band horizon graph）用来反映时序数据在水平线上的涨跌变化情况。Liu 等[87] 扩展了堆叠图用于展示不同话题的摘要信息随时间的演化过程。Luo 等[88] 提出了 ProcessLine 将生产过程中的关键要素映射到符号与颜色，并结合鱼眼放大（Fisheye）等交互工具辅助监控生产过程中各生产单位质量控制。

此外，通过关联多视图和符号设计展示时间特征也是常用的可视分析手段。例如，

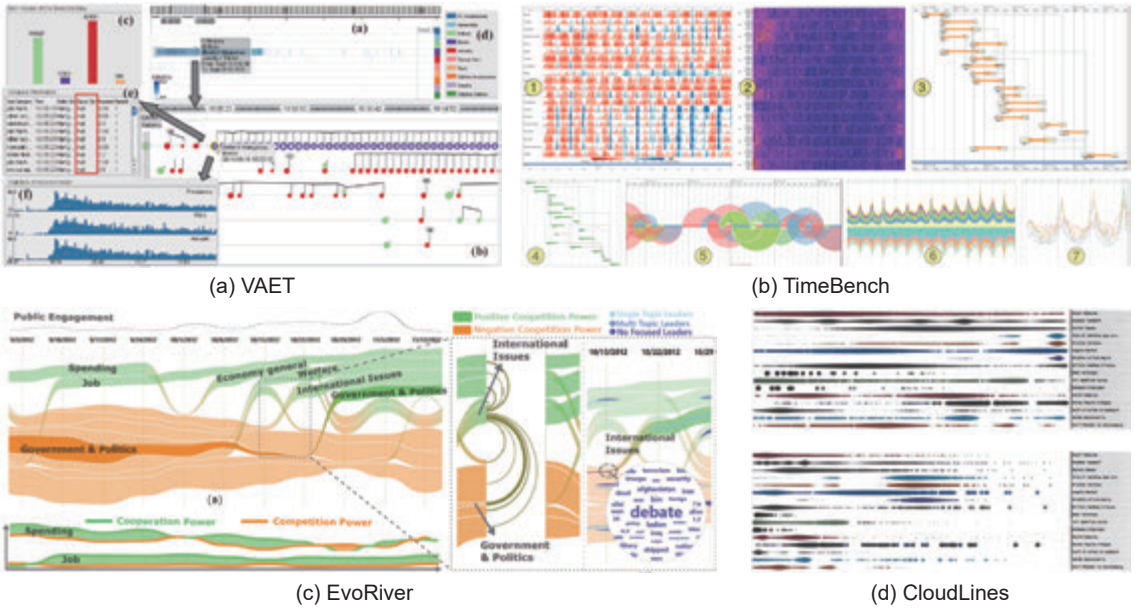


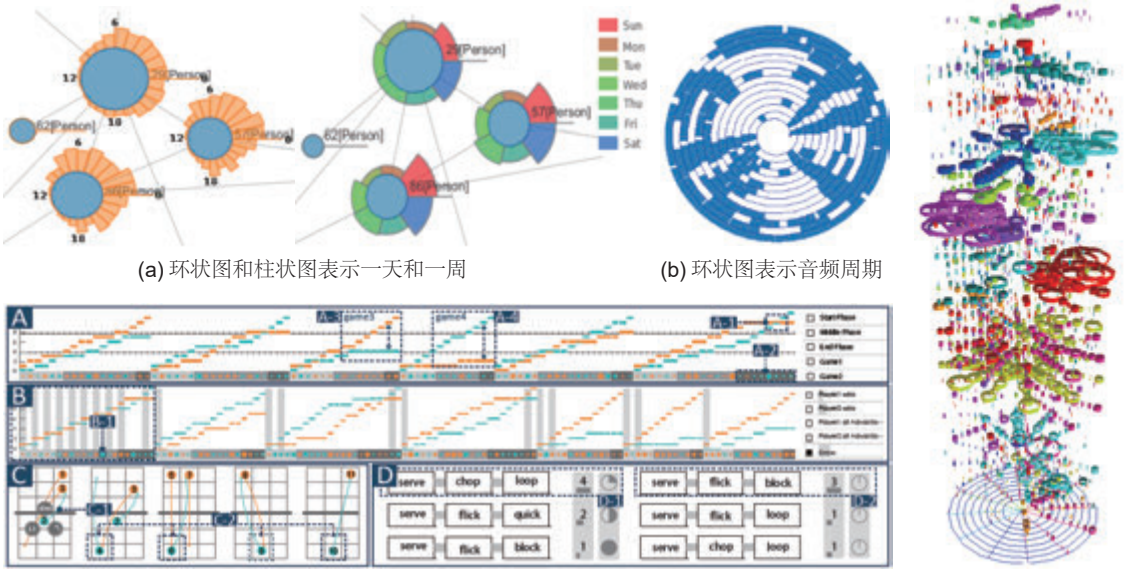
图 2-6 线性时间数据可视化。(a)，(b)，(c) 和 (d) 分别来自文献 [89]，[90]，[30] 和 [91]

如图 2-6(a) 所示，Xie 等[89] 设计了一种紧凑的视觉编码用来刻画电子交易过程中的时间变化，并基于该设计开发了 VAET 可视分析系统通过不同尺度的时间视图帮助用户分析大规模的交易数据。如图 2-6(b) 所示，Rind 等[90] 综合使用甘特图（Gantt chart）、弧状图（Arc Diagram）、热力图（Heatmap）以及堆叠图等多种图表开发了 TimeBench 可视分析系统，从不同侧面反映长期时序数据的特征变化。如图 2-6(c) 所示，Sun 等[30] 基于主题河流图设计了 EvoRiver 可视分析系统用于展示、分析和理解社交媒体上话题之间的随时间变化的竞争合作变化过程。如图 2-6(d) 所示，Krstajic 等[91] 设计了 CloudLines 图表将事件流数据映射为时间散点图，并结合变形透镜（Distortion Lens）交互工具帮助用户分析多个时间序列的视觉模式。另外，赵等[92] 提出了一个网络流量时序分析流程模型并基于该模型设计了一个多视图合作的网络流量时序数据可视分析，支持自顶向下的、从整体到个体的网络攻击分析。Dang 等[76] 将散点图矩阵与折线图结合，展现数据中多个维度随时间变化的趋势。Holtz 等[93] 集成地图视图展示时间关联的地理信息。

2.2.2 周期性时间数据可视化

对于长期的时序数据，展现和分析数据的周期模式也是常见的可视化任务[82]，可以利用螺旋图（Spiral，如图 2-7(d)）[94, 95]、环状图（如图 2-7(b)）[51]、子图聚类[96] 等方式实现。Woodring 和 Shen[52] 基于小波变换方法将时序数据转换到不同时间尺度的曲线集合，通过对这些集合进行聚类分析挖掘数据的周期模式，并利用多种过滤器和多个视图辅助用户选择不同时间尺度的集合深入分析。Shen 和 Ma[51] 结合环状图与柱状图反映移动数据中用户在一周和一天的活动规律（图 2-7(a)）。Wu 等[29] 针对一种展示乒乓球运动数据中的回合与击球策略，分析每一局比赛的每个回合中双方运动员的

击球和得分过程（图 2-7(c)）。Walker 等[97] 综合使用像素图（Pixel Plot）、河流图以及多种透镜交互工具设计了 ChronoLenses 可视化分析系统帮助用户多尺度的分析时间数据中的频率特征。金等[98] 综合使用折线图、堆叠图、平行坐标系等可视化设计反映不同传染病的长期趋势、季节消长规律以及在不同省份的空间分布规律。



(a) 环状图和柱状图表示一天和一周　(b) 环状图表示音频周期

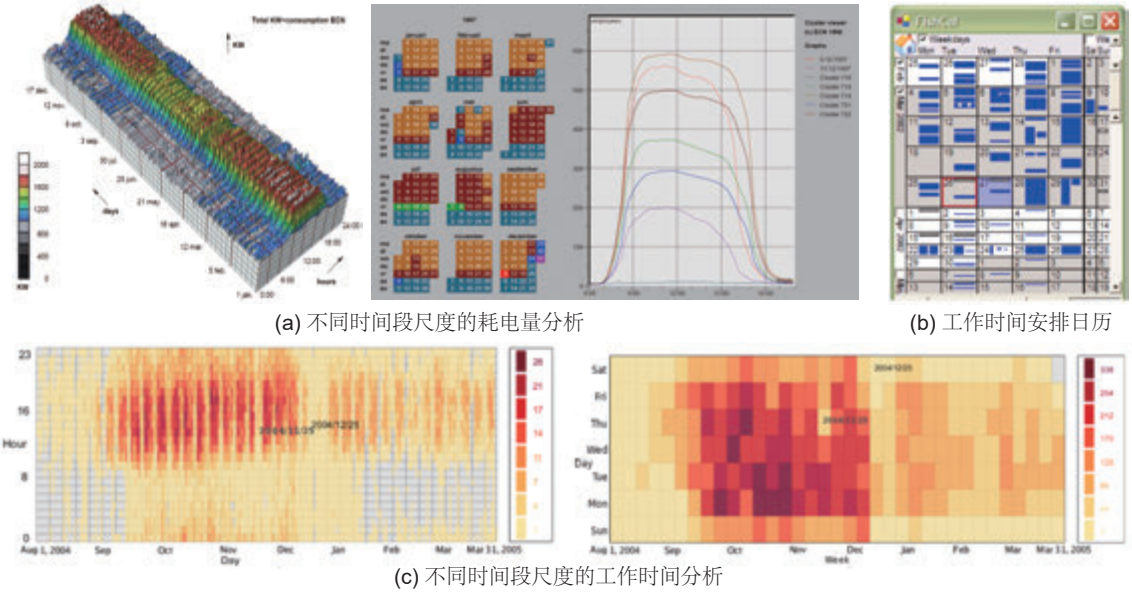
(c) 柱状图展示乒乓球每局比赛与每个回合　(d) 螺旋图

图 2-7 可视分析时序数据中的周期性规律。(a)，(b)，(c) 和 (d) 分别来自文献 [51]，[96]，[29] 和 [95]

2.2.3 时间点与时间间隔数据可视化

离散的时间点将数据中的事件描述为离散空间中的点，没有持续过程的概念。而时间间隔则表示了一段线性时间域，数据中的事件描述了一个持续的时间段，例如几小时、几天、几周或几个月。将时间点和时间间隔与日历中的时间单位对应，并分为年、月、周、日、小时等多个等级，能够符合人类对时间的认知习惯，从而促进此类数据的可视化。Wijk 等[99] 将小时、日期作为 x, y 轴，将耗电量作为 z 轴，同时展现每天和全年的耗电量周期，并采用日历图展示耗电量在每周的使用分布（图 2-8(a)）。Bederson 等[100] 将日历视图结合鱼眼交互技术应用在手持设备的小尺寸屏幕上，帮助用户安排分析日程活动（图 2-8(b)）。Shen 和 Ma[51] 基于日历和月历分析移动设备数据中体现的工作时间段，直观的展示出每天的工作时间从早上 11 点到晚上 7 点并在 12 月底有明显的休假时间（图 2-8(c)）

通过分析上述三种类型时序数据的可视化研究工作可以发现，针对时序数据首先可以针对数据中时间属性的特点，选择合适的可视化设计展现数据中的时间属性，然后，通过符号设计和其他图表展现数据中与分析任务相关的属性和维度，最后，通过利用结合关联协调视图，概览 + 细节等交互技术分析时序数据中的时间模式和周期性规律。



(a) 不同时间段尺度的耗电量分析　(b) 工作时间安排日历

(c) 不同时间段尺度的工作时间分析

图 2-8 利用日期属性分析数据的可视化设计。(a)，(b) 和 (c) 分别来自文献 [99]，[100] 和 [51]

2.3 在线学习行为数据可视分析

在传统的课堂教育的数据分析中，由于样本数量少且数据采集手段缺乏的限制，所以获得的教育数据集规模也相对较小，数据可视化通常仅作为分析结果的呈现手段，而交互技术在教育数据的探索分析中也很少使用。然而，随着网络教育的兴起特别是大规模在线开放课程（Massive Online Open Courses，MOOC）的热潮[101]，教育数据集的规模正在以前所未有的速度增长。其中以学习日志为代表的在线学习行为数据增速尤其明显，仅一门 MOOC 课程的数据就可能含有超过 10 万名学生的 1 亿条学习行为记录[3]。面对如此大规模的数据，已有的数据挖掘方法尽管能够自动化的提取模式信息，但是这些分析结果难以被教师和学生直观理解和进一步运用[102]。针对这一问题，已有研究开始运用可视化技术与交互技术分析教育数据[7, 21, 103]。通过利用直观的可视化界面和交互探索工具，教师可以将自己的教育授课经验与在线学习行为数据分析相结合，从不同角度探索和理解学生的学习模式，为改进教学手段提高教学质量提供参考[104]。

由于在线学习过程包含了多种各异的教学活动，使得在线学习行为数据具有明显的多样性，同时不同用户对数据分析的需求也各不相同，面向在线学习行为数据的可视分析需要针对教学活动中具体的学习行为数据展开。因此，本文将分别从学习行为数据分析任务，以及在线学习行为数据可视分析这两个方面讨论相关的研究工作。

2.3.1 在线学习行为数据分析任务

随着学习管理系统（Learning Management System，LMS）的广泛使用，学生从入学到毕业的完整在线学习过程能够被充分的记录下来，形成数字化的教育数据（Educational Data）为后续的教学评估与学术研究提供支持。对这类数据的分析主要涉及教育数据挖

掘（Educational Data Mining）和学习分析（Learning Analytics）这两个领域[105]。根据相关综述文章的总结，针对教育数据的分析按照研究目标可以分为学生/学习行为建模、成绩预测、自觉性增强、留存/退出预测、评估反馈服务、以及资源推荐这六大类[106, 107]。本研究针对的对象是在线学习行为数据，主要涉及上述六类分析任务中学生/学习行为建模相关的研究。

学生/学习行为建模的相关研究通过分析在线学习行为数据及其关联数据，从不同侧面刻画学生以及学习过程的全方面特征，因此通常是其他各类任务的基础。其研究内容包括，衡量学生的学习参与度[108]，识别了不同特征的学习者群组[109]，挖掘学习工具使用情况的频繁模式[110]，识别不同类型活动学习模式[48]，分析学生的行为参与度与情感参与度、认知参与度之间的关系[111]，分析课程各类教学资源的使用情况与考试成绩之间的关系[47]，分析多种教学技术和支持服务对学习参与度的影响[112]，识别周期性的学习时间模式[113] 等。

针对这些分析任务，现有研究采用统计学、机器学习、以及教育技术学的方法分析涉及的学生人口统计学数据、课程数据、学习行为数据以及成绩数据开展分析。在这些数据中，学习行为数据的规模和数据类型最大，特别是随着 MOOC 的普及，数据规模猛增。以麻省理工学院在 edX 平台上 2012 年发布的电路与电子学课程为例，仅 2012 年春季和秋季分别产生了超过 2.3 亿条和 1.3 亿条学习行为记录[1, 114]。为了处理如此规模的数据，Kalyan 等[114] 通过运用大数据处理平台和相关工具，设计了 MOOCdb 规范并基于分布式计算平台处理这些数据，从入学区域分布、课程完成率、以及课堂讨论等方面分析了 MOOC 学生的学习特点。类似的，Chuang 和 Ho[2] 分析了哈佛大学和麻省理工学院在 2012 年至 2016 年间的 MOOC 课程学习数据，讨论了学生分布、学习趋势、以及课程完成率等问题。

2.3.2 在线学习行为数据可视分析

在针对上述分析任务的研究中，通常会采用基于统计学或机器学习的方法分析数据，然后采用基本的图表和界面展示统计分析结果，例如使用柱状图呈现视频复杂度分布[115]，饼图、折线图以及地图展示学生的人数和地理分布[103]，热力图展示学生在论坛中活跃程度[13]，状态转换图（State Transition Chart）展示学生群体在不同学习状态之间的人数变化[53, 116]，以及使用一些徽章符号（Badge）表现学习进度[117]。

但是随着在线学习行为数据规模的迅猛增长，如何从规模庞大的数据中探索学习行为的规律正在成为研究的热点。在这方面，已有研究工作针对部分具体的学习活动，通过运用可视化技术与交互技术在大规模学习行为数据中分析潜在的学习行为模式。按照这些研究针对的学习行为，可将现有的研究分为三类，包括视频观看行为分析、讨论行为分析以及其他教学活动分析。

视频观看行为分析：视频观看行为是在线学习行为的主要组成，分析视频观看行为数据可以为管理视频资源、评估学习过程、衡量参与度等方面提供支持，例如分析视频时长对学生参与度的影响[11]，分析视频表现风格对学生成绩的影响[121]，分析学习模

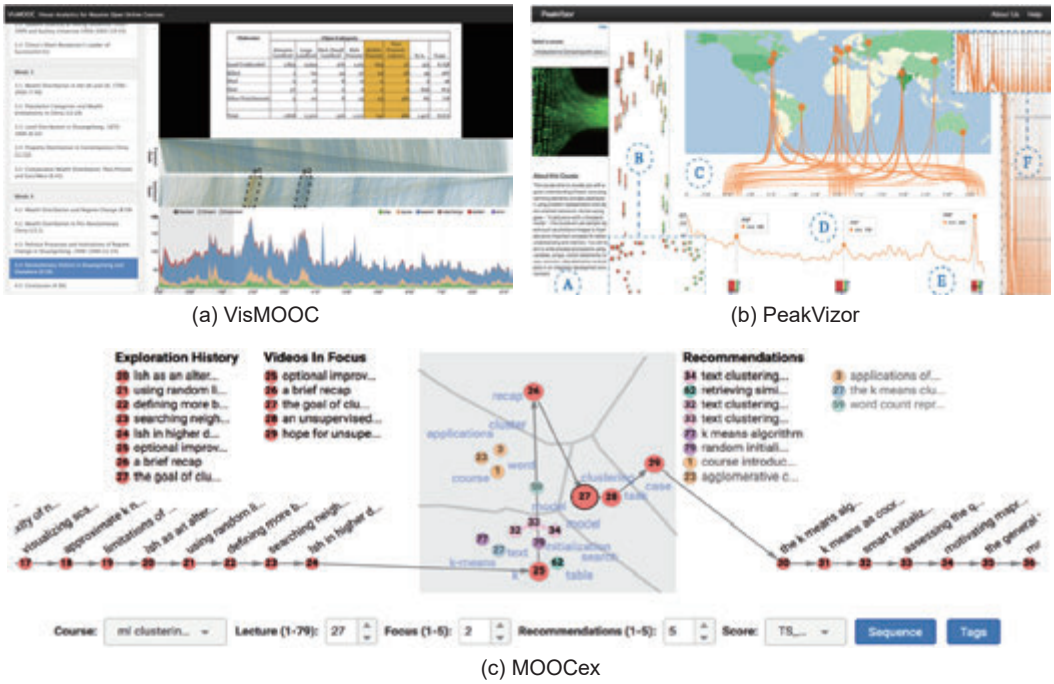


图 2-9 视频观看行为的可视分析。(a)，(b) 和 (c) 分别来自文献 [118]，[119] 和 [120]

式[47, 48] 等。在针对视频观看行为的可视分析方面，Shi 等[118] 设计了 VisMOOC 可视分析系统帮助教师分析课件视频内学生的操作过程。利用 VisMOOC 特别设计的事件图与查找活动图，教师可以直观的发现学生在观看视频时的频繁操作区间与模式（图 2-9(a)）。Chen 等[119] 设计了 PeakVizor 可视分析系统以帮助教师和教学专家查看点击流数据中视频片段上出现的峰值现象，并帮助他们分析其学习行为模式。PeakVizor 通过特别设计的符号展示了峰值的统计特征，并通过关联的多个视图展示峰值相关的时空信息与学习者信息（图 2-9(b)）。Zhao 和 Cooper 等[54, 120] 构建了 MOOCex 可视分析系统用来帮助 MOOC 学习者交互式的选择适合当前学习进度的视频。通过预先分析课程视频在语义空间的相似度以及课程之间的关联性，MOOCex 可以将适合学习者当前学习进度的课程视频以网络图的形式呈现，帮助学生提高学习效率（图 2-9(c)）。

讨论行为分析：在线论坛被认为是构建在线学习社区以及促进沟通交流的重要功能，利用学习管理系统提供的在线论坛功能，学生可以和其他学生以及教师之间通过发帖、浏览、回复等操作展开讨论，从而利用社区群体的支持解决个人的学习困难。在针对论坛内讨论行为分析的可视分析方面，Sung 等[122] 结合主题河流图设计了 ToPIN 可视化方法展示学生在不同时间段讨论的话题及其归属类别，帮助用户理解在线学习者的讨论的内容和学习状态（图 2-10(a)）。Wu 等[77] 结合网络图与平行坐标系设计了 NetworkSeer 可视化分析系统展示论坛中的交互统计情况，帮助用户分析学生使用论坛的原因以及讨论热点（图 2-10(b)）。Fu 等[56] 设计了 VisForum 可视化分析系统用来帮助用户分析 MOOC 论坛中的学生群组。通过多种创新的符号设计与多个关联视图，VisForum 能够识别论坛中具有重要影响力的论坛成员。（图 2-10(c)）。Fu 等[123] 设计了多种展现论坛讨论活动时序变化特征的图表与符号，并基于这些设计开发了 iForum

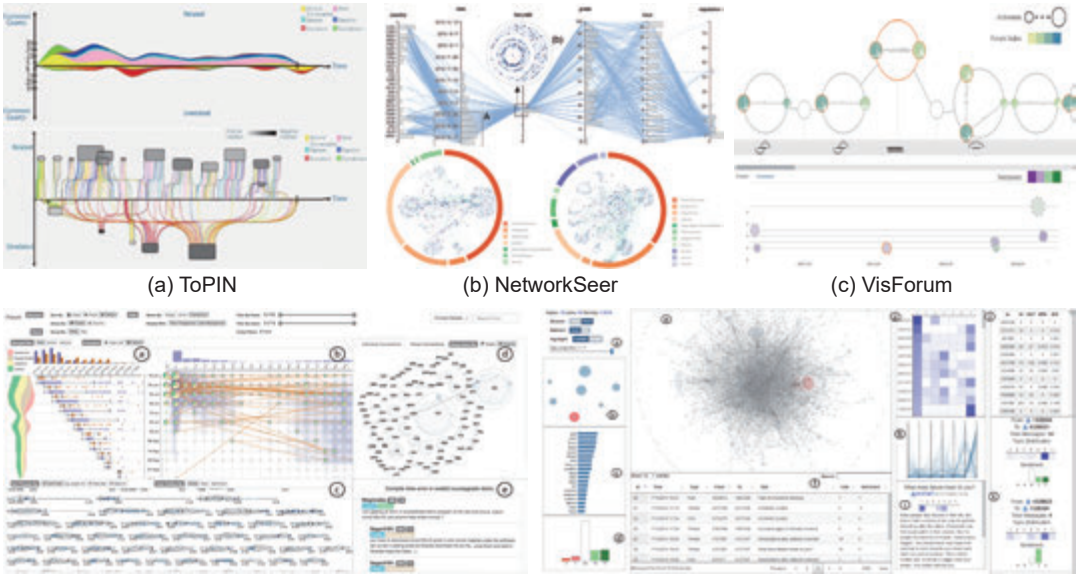


图 2-10 讨论行为的可视分析。(a)，(b)，(c)，(d) 和 (e) 分别来自文献 [122]，[77]，[56]，[123] 和 [124]

可视化分析系统，帮助用户分析论坛中每个讨论话题随时间的演变过程以及论坛整体的讨论趋势变化（图 2-10(d)）。Wong 和 Zhang[124] 集成多种图表与交互技术设计了 MessageLens 可视化分析系统，辅助用户了解论坛中学习者态度的变化与学习之间的沟通情况。（图 2-10(e)）。Atapattu 等[125] 设计了一种环状图用来分析学生在论坛中的讨论话题。

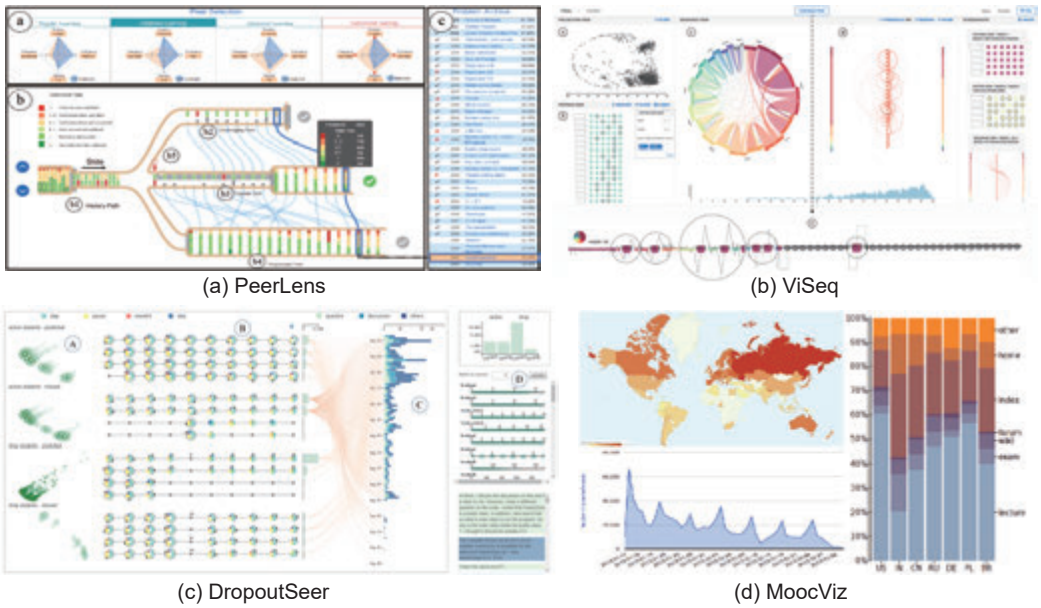


图 2-11 整体学习过程的可视分析。(a)，(b)，(c) 和 (d) 分别来自文献 [57]，[55]，[126] 和 [103]

其他教学活动分析：除了上述两个通用的教学功能以外，也有其他工作针对特定课程的学习功能或者整体教学过程展开研究，例如，Xia 等[57] 设计了一种学习路径图可视化方法 PeerLens 来帮助编程学习者选择适合自己的编程题练习路径，学习者可以

直观的查看自己的编程题历史与各种难度和类型的编程题路径，通过交互式的对比不同路径，从而选择适合自己的编程练习题目（图 2-11(a)）。Chen 等[55] 集成多种图表设计了 ViSeq 可视化分析系统用来帮助用户分析 MOOC 学习者学习顺序与成绩之间的关系，进而为实现识别不同学习行为背后的原因和发现不同学习模式的学生群组（图 2-11(b)）。Chen 等[126] 开发了 DropoutSeer 可视化分析系统并设计了反应学生学习活动的符号表示，帮助用户分析 MOOC 学习者退课不再学习的原因，并提供对可能退课学生的预测（图 2-11(c)）Dernoncourt 等[103] 开发了 MoocViz 分析平台帮助研究者探索 MOOC 学习行为数据中的统计情况，包括区域分布、活跃人数、活动数量等可视化展示界面（图 2-11(d)）。Pardos 等[127] 开发了 moocRP 分析平台帮助研究者处理多个 MOOC 平台的异构学习行为数据，并提供基于课程的人口统计分析、资源使用率分析等可视化展示界面。纪等人[128] 开发了 SPVAS 可视分析系统，通过结合热力图、平行坐标系、弧长链接图以及节点链接树等多种可视化技术与多视图协同交互技术，揭示学生成绩在年级和学期上的多维属性时序分布。

综合上述研究可以发现，利用可视分析方法能够从不同侧面，细粒度的分析在线学习行为数据中各种行为的模式与特点，为改进课程质量提高学习效果提供支持。通过利用多种可视化设计与多视图协同交互技术，从多个层面分析学习过程中的各种学习行为模式。但是，目前仍缺少对在线学习的高维度学习参与度、时间管理风格以及大规模视频利用模式的整体过程分析，因此，本文将借鉴已有研究工作的可视化设计与交互技术，开展针对上述问题的研究。

2.4 本章小结

本章着重介绍了包括高维数据可视分析、时序数据可视分析和在线学习行为数据可视分析在内的相关工作，为本文的相关研究提供理论基础和参考依据。

3 面向高维数据的学习参与度可视分析

3.1 引言

随着大规模开放在线课程（Massive Open Online Courses, MOOC）在全球范围的流行，多样化的学习功能与大规模的学习者催生出海量的在线学习行为数据，包括学生数据、课程资源数据、视频学习记录、论坛讨论记录、虚拟实验记录、测验考试记录、移动学习记录等。在线学习行为数据不仅有助于解决网络教学中的基本问题，如生源分析、课程完成率统计、资源评价、学习质量评估等，而且有助于分析学生的在线课程参与模式和影响因素。Kizilcec 等[129] 发表在《Science》上的论文研究了 211 个国家和地区的超过 180 万名学生的在线学习行为数据，发现在线课程的完成率不仅和学生的教育背景、英语能力、学生所在地区人类发展指数（Human Development Index）相关，还会受到学习氛围中社会认同感的影响，因此，通过在学习过程中提供心理学引导可以有效的提高发展中国家学生的课程参与度。

学生数量的猛增使得学生个体的学习行为差异性变得更加显著，例如，有的学生可能只通过观看课程视频学习，有的偏好阅读文本材料，而有的学生更偏好通过讨论交流学习。同时，技术手段的提升也使得课程教学手段更加丰富，有的课程以视频为主，有的是教材与视频并重，而有的课程则主要通过在线虚拟实验等实践环节学习。由于学生规模的庞大与课程学习活动的多样性，在线学习行为数据具有显著的异构性并且体量大、维度高。因此，如此大规模的高维度学习行为数据对学习行为数据的分析和挖掘提出了挑战。早期研究中采用的分析方法数据采集能力限制的制约，数据样本规模较小且限定于特定的学习行为，使得分析过程通常只包含有限的学生特征和行为特征[105]。然而，随着学习行为数据在规模与维度两方面的共同增长，现有研究也开始从中提取高维度的学习行为特征用于各项数据分析任务[107]。例如，Cerezo 等[48] 的研究发现，不同类型的学习行为数据能够从多个角度共同刻画学生的参与模式，全面的反映学生的在线学习过程。Jiang 等[6] 的研究也发现使用视频观看与作业提交两类学习活动能够反映出多种学习模式，从而为预测学习效果和评价教学质量提供依据。Bote 等[130] 从学习过程的数据中提取了多个特征用于衡量学生在网络课程学习情况，进而为预测学生的课程参与情况提供支持。

现有的在线学习行为数据分析方法主要分为基于教育学分析模型的方法和基于机器学习技术的方法[7, 21]。基于教育学分析模型的方法需要运用教育专家的经验和专业知识，根据认知理论建立数学模型，利用问卷与访谈数据分析各种因素对学习行为和学习成效的影响；基于机器学习技术的方法从在线学习行为数据中提取的学生的学习行为特征，利用分类或聚类算法挖掘学生的行为模式。然而，这两种方法都需要领域专家与数据分析专家运用长期的领域经验和专业知识才能对数据和模型进行判断和总结，耗费大量的人力和时间成本。特别是对于缺少数据挖掘经验的教师与助教，这些方法

更是难以直接运用。在实际工作中，学习行为分析是一个反复迭代的渐进式过程，现有的教学平台很少为教师设计交互式的分析工具帮助他们探索学生的学习过程。

在线学习行为模式分析存在以下两方面挑战：首先，在线学习方式的多样化，在线学习行为数据呈现明显的异构特性与多样性，因此需要有效的管理和抽象异构的学习过程数据，构建统一的学习行为数据模型和平台；其次，学习行为特征通常表现为高维度的学生参与度，用户需要直观的可视化界面与交互反馈探索不同维度的行为模式与影响因素。为了应对上述挑战，本章提出采用可视化技术分析学生参与度的方法，辅助教师和管理人员分析在线学习行为数据中课程与学生群组的参与度，并通过交互界面探索影响学生参与度模式的相关因素。为解决学习行为数据异构多样的问题，本章设计了一种在线学习行为数据层次化抽象模型和存储系统，实现多源学习行为数据的汇聚和表征。而针对高维数据难以直观探索的问题，本章提出一种综合散点图矩阵与 t 分布随机邻域嵌入（t-Distributed Stochastic Neighbor Embedding, t-SNE）数据降维技术的学习参与度分布图，并设计了多个交互视图与学生群组创建工具，辅助探索学生参与度的高维模式和不同学生群体的学习参与度分布。利用本章提出的在线学习数据管理系统与学习参与度可视化技术，本文采集了本校网络教育学院的在线学习行为数据集，并从中选择了一门课程的在线学习行为数据用以验证所提出的可视化分析技术的有效性。通过对该课程在线学习行为数据的探索分析，不仅发现了多种学习参与度模式，还进一步分析了在线学生支持服务与学习参与度模式之间的关系。

在后续的各节中，本文将在第 3.2节讨论在线学习行为数据的管理方法与系统，在第 3.3节说明如何抽象在线学习行为数据的学习参与度特征，在第 3.4节详细介绍学习参与度分布图的可视化设计与交互设计，在第 3.5节结合真实的学习行为数据讨论学生的学习参与度模式，并讨论在线学生支持服务对学习参与度的影响，最后在第 3.6节对本章工作进行分析和总结。

3.2 在线学习行为数据管理

本节针对在线学习行为数据多样化的特点，提出一个层次化的在线学习行为数据管理模型，将多样化的在线学习活动用统一的数据标准表示。然后，基于该数据管理模型，设计了一套学习行为数据管理框架，为后续的学习参与度等研究提供数据基础。

3.2.1 数据存储模型

在线学习行为数据的数据规模庞大，为了提高数据分析与查询访问的效率，在线学习行为数据的存储需要合理的结构，因此本文设计了一种层次化的数据存储模型分别保存原始学习过程数据、分区的分析数据以及统计结果数据。

如图 3-1所示，该数据存储模型包括三层，首先在基础数据层将在线学习行为数据按照统一的数据格式集中存储，然后在分区数据层按照课程与开课时间段分区，最后在统计数据层按学生分别存储统计结果数据。自底向上每层分别为全局统计、课程数

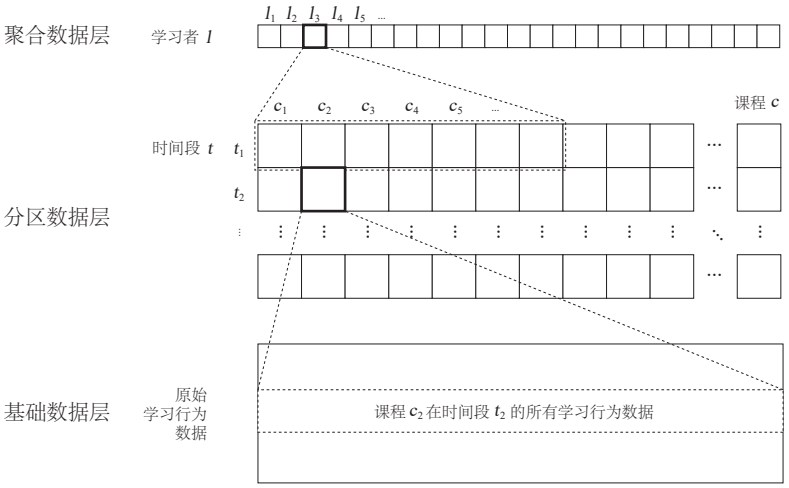


图 3-1 层次化的存储结构

据统计查询以及学生个人提供数据支持。各层的详细设计如下。

基础数据层：为了完整刻画多样化的在线学习行为的特征，需要设计数据模型规范用于描述学习行为的属性，保留学习行为的原始信息。在已有研究中已经提出了共享内容对象参考模型（Shareable Content Object Reference Model，SCORM），xAPI（Experience API）标准，DataShop[131]，以及斯坦福 CAROL 学习者数据模型[127] 等标准，此外 Moodle[132]、Blackboard，edX[133]，Coursera 等教育平台也提出了针对内部的数据标准。这些数据规范为了描述平台特有的学习行为特征而设计了复杂的结构与繁多的属性，不适合本文对于一般在线学习行为的描述。因此，本文借鉴 MOOCdb[114] 的基础数据设计，提出一种适合于大规模云存储环境的原始学习行为数据存储结构。如图 3-2所示，该存储结构以学习行为记录表为核心，包含全部的学习行为记录，并将部分信息内嵌在字段编码中，减少检索式的跨表数据访问。除学习行为以外的其他信息保存在各种辅助信息表中，通过各表的唯一标识符与学习行为记录关联。

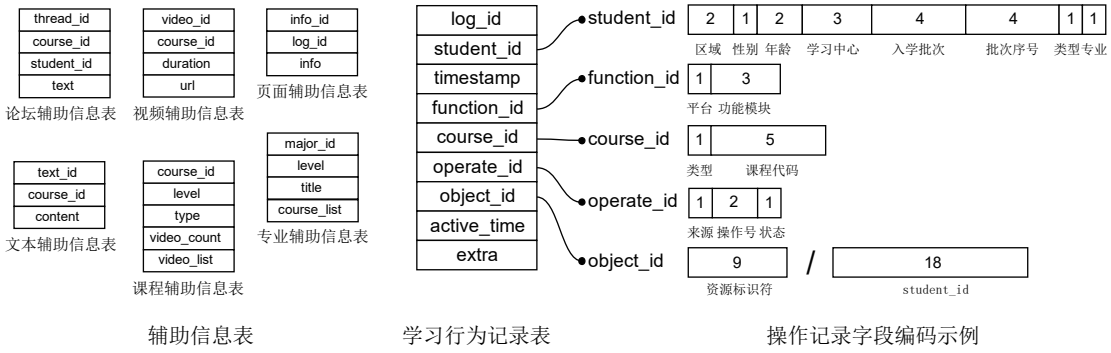


图 3-2 基础数据层的原始学习行为数据存储结构

分区数据层：在线学习行为数据的分析通常是针对特定的课程以及特定的时间段，分析单个课程的数据时不会涉及其他课程或其他时间段的数据。根据数据存储的一般原则，相关性较强的数据在物理存储时相邻有利于提高访问效率[134]，因此，在线学习

行为数据在存储上需要满足：1）同时间段的学习过程数据相邻存储；2）相同课程的学习过程数据相邻存储。根据上述需求，在该层中，学习过程数据先按课程分块、再按特定的时间段分段，从而提高分析时的处理效率。

统计聚合层：由于所有的统计分析指标是以学生为基本单位，并且数据的查询与展示也需要以学生为最小单位呈现，所以在该层中，所有数据按学生划分存储，包括学生在课程上的学习过程，以及按照分析指标计算的统计结果。

3.2.2 数据管理系统框架

基于上述数据存储模型，本文提出以下的在线学习行为数据管理框架，实现对在线学习行为数据的采集、处理、分析以及应用，其结构如图 3-3 所示。

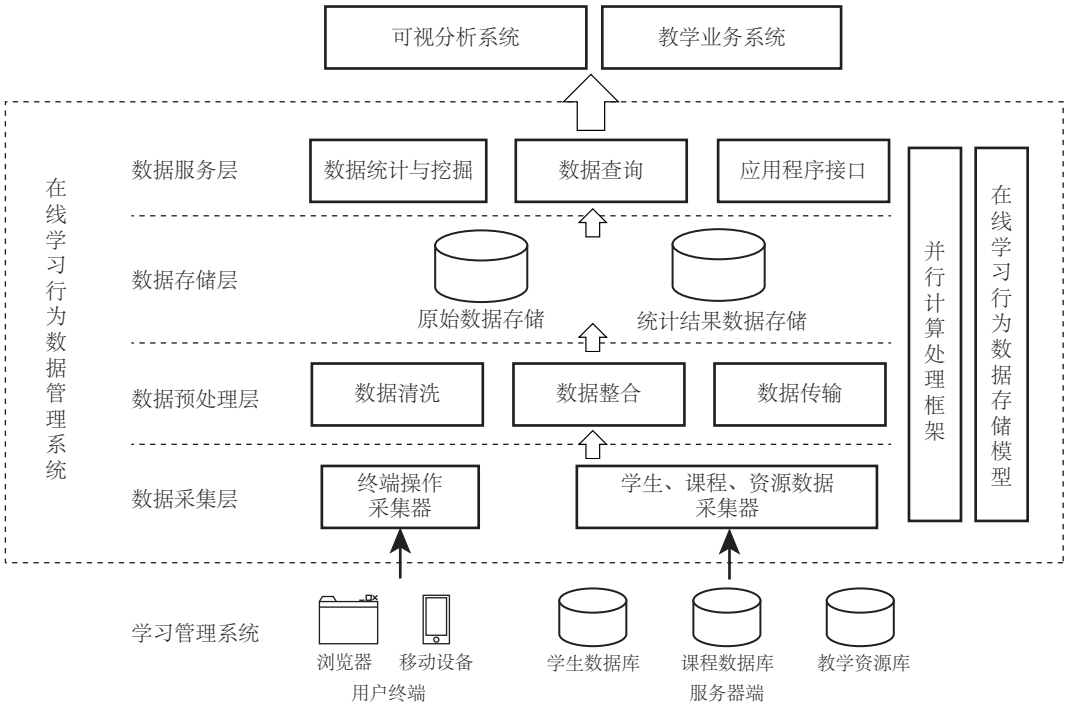


图 3-3 在线学习行为数据管理系统框架

该框架基于上述数据存储模型以及并行计算处理框架，包括四个主要层次：数据采集层、数据预处理层、数据存储层以及数据服务层。其中，数据采集层从学习管理系统中采集客户端的学生学习操作数据和服务器端的学生数据、课程数据和教学资源数据；数据预处理层按照预定义的方法从原始数据中删除无关的数据、清理异常编码，字段、并将所有异构的数据格式统一，按照前一节提出的数据存储模型将数据分层保存在数据存储层的原始数据存储中。数据采集与预处理可在数据分析之前定期增量执行；数据服务层通过结构化查询语言（Structured Query Language，SQL）脚本和其他脚本程序对分区数据进行统计分析，并将统计结果保存在数据存储层的统计结果存储中，同时，所有统计结果可以通过数据查询结构以及应用程序接口向外提供服务，为可视分析系统、教学管理系统、移动客户端等应用提供数据支持。上述框架中的各系统模块

可以采用多种主流的开源技术实现。

3.3 学习参与度指标模型

针对在线学习行为数据多源异构、教学形式繁多的问题，本节将介绍数据处理的过程。数据处理分为 3 个阶段，首先，汇聚各数据源采集到的学习行为相关数据并进行预处理，清理无关字段，修复错误字段和空字段等；然后，按照课程与时间段对数据进行分区存储；最后，以学生为单位统计各项学习参与度指标并建立索引。

3.3.1 数据预处理

原始的在线学习行为数据主要以学习日志数据为主，其中包含了大量冗余字段和记录，例如学习系统记录的会话编号、浏览器信息、客户端信息、多端时间戳等字段，以及与本次课程学习无关的缴费、学籍管理等记录。除了上述冗余字段与记录以外，原始学习日志数据中也包含了许多异常数据。例如，由于字符编码引起的异常编码文本数据、异常学习时间、不同版本客户端的数据格式不一致等问题。针对这些问题，数据预处理模块按以下步骤对学习日志数据进行了清理：

- 清理编码异常数据：由于多源数据库的异构性，各数据来源的编码格式均不相同，特别是涉及中文字符和特殊符号时，易产生由于编码不兼容引起的乱码造成数据丢失。因此，在数据预处理时需要先处理编码问题。首先，尝试将所有数据的编码转换为统一编码，如可变长的字符统一编码（Unicode Transformation Format），然后再针对乱码数据单独从数据源重新采集或删除。
- 删除无关数据：学习日志数据中包含了浏览器、操作系统、网络状态等与本研究关注的在线学习行为分析无关的字段，另外也包含了涉及学生个人隐私信息的一些记录，该步骤会将这些字段和记录删除。
- 处理异常学习时间的数据：在学习日志中记录了学生观看视频、阅读材料等活动的学习时间，若其学习时长超过合理范围或给定阈值（例如，无活动会话超时设置为 1 小时），将无法代表学生实际的学习时间，影响对后续学习参与度的分析。针对这些异常数据，将采用相邻日志记录的时间间隔长度作为替代，或者按照学习活动类型选取合适的时间间隔作为替代。例如，对于单次观看视频时长超过登录长的数据，以点击播放开始到结束播放操作的时间间隔作为替代。
- 统一数据格式：建立格式编码转换字典，采用统一的编码规则标记数据记录中的学生、操作、资源等字段，从而在整个系统中使用相同的规则访问数据。

经过数据预处理后，在线学习行为数据中学习日志数据的数量和大小显著下降，从而降低了对数据存储和数据分析的读写压力。

3.3.2 在线学习行为抽象

学习日志数据中记录的在线学习行为与学习系统提供的教学功能模块紧密相关，虽然各种教学手段呈现出显著的多样化特点，但是根据各种学习行为的共同要素，每一个在线学习行为都可以由为五个基本要素构成：学习者（Learner），操作（Operator），时间（Time），资源（Resource），以及属性集合（Attributes）。以论坛讨论行为数据和在线测验行为数据为例，论坛讨论行为数据的字段包括：{student_id, post_id, operate, content}，以及相关的文本数据：{post_id, last_update_timestamp, owner_id, content}。在线测验行为数据的字段包括：{student_id, quiz_id, operate, timestamp, answer}。其中，每条数据均包含学生唯一标识符 student_id 表明发起学习行为的主体，operate 属性表明具体的学习行为，timestamp 属性表明了发生学习行为的时间点，post_id 或者 quiz_id 表明该学习行为直接相关的教学资源，以及其他一些属性来标识该条数据的特征（例如发帖的内容 content 属性，提交的答案 answer 属性）。

根据这些学习行为数据的共性属性，本文提出以下的在线学习行为抽象模型。设 L 为学习者集合，O 为所有在线学习操作集合，R 为教学资源的集合，则可以将一个学习者 l 的一个在线学习行为 b_l 定义为一个五元组：

$$b_l = \langle l, o, r, t, \Omega \rangle \tag{3-1}$$

其中 l 表示学习者，o 表示在线操作，r 表示教学资源，t 表示学习行为发生的时间，Ω 表示其他各类属性的集合，且各元素满足 $l \in L \wedge o \in O \wedge r \in R$。

因此，对于给定的课程 c，任意学习者 l 在一段时间的全部在线学习行为构成的一个序列，称之为在线学习过程 P_l，定义如下：

$$P_l = [(l, o_1, r_1, t_1, \Omega_1), (l, o_2, r_2, t_2, \Omega_2), \cdots, (l, o_n, r_n, t_n, \Omega_n)] \tag{3-2}$$

式中，每个五元组按照时间顺序排列，满足 $t_1 < t_2 < \cdots < t_n$。

由公式 3-2可知，在线学习过程 P_l 中是由多个按时间顺序排列的学习行为 b_l 构成的一个学习行为序列。其中，每个学习行为 b_l 的操作 o 与学生使用的教学功能相关。在学习过程中，学生可能连续的使用不同的学习功能，例如一边查看论坛讨论一边观看视频，或者一边完成作业一边查阅资料。因此，相邻的两个学习行为 $b_{l,i}$ 和 $b_{l,i+1}$ 可能反映了不同的操作 o 和教学资源 r。另外，不同来源的数据在格式和内容上存在明显差异，例如视频观看数据包含了视频帧与时间信息，讨论数据包含了自然语言文本信息，而虚拟实验数据则包含了结构化的代码信息。这些差异化的信息保存在属性集合 Ω_i 中，通过在属性集合 Ω_i 中设置相应字段来标识数据的来源并保存具体内容，从而在统计计算时根据需求提供数据支持。

3.3.3 学习参与度指标

学生的学习参与度（Student Engagement）是教学过程中的重要因素，它对学生的学习满意度、学习过程、学习成绩、知识掌握程度、按期毕业、课堂留存以及休学退学等多个方面都有显著的影响[135]。教育学领域的研究通常将学习参与度定义为一个多维度的概念，包括行为参与度（Behavioral Engagement），情感参与度（Emotional Engagement）以及认知参与度（Cognitive Engagement）这三个方面[136]。其中，行为参与度主要指学生参加某个教学过程的行为，例如完成一项作业、到课堂听讲、或者回答题目等行为；情感参与度指学生对教师、其他学生以及相关人员的情绪反应；认知参与度则是指学生参与教学活动过程中对具体任务投入的思考。这三种参与度通常认为是相互影响的关联要素，并在学生的学习过程中同时存在[137]。然而，在学生参与度的这三个方面中，行为参与度是学生展现出学习状态的基础，在所有关于学生参与度的定义中均以行为参与度作为必要组成，并且学生的情感参与、认知参与也是通过具体的学习行为展现的[138]。因此，在本文中主要关注学生的行为参与度，下文中的学习参与度均指代行为参与度。

在网络课程的在线学习过程中，学习参与度是通过统计学生的学习行为数据得到的，得益于在线学习管理系统（Learning Management System，LMS）的使用，学生的在线操作行为可以被细粒度的、完整的记录并随时调取分析。基于前一节介绍的在线学习行为数据存储模型与数据管理系统采集的在线学习行为数据，可以从不同角度、不同层面刻画全部学生在整个在线学习过程的学习参与度。在已有针对在线学习过程的研究中，通常可以采用基于数量与时长的两类指标衡量学生参与度[136]。例如，Joksimović 等[65] 采用基于学生交互类型的 8 个统计指标衡量学生与学习管理系统的交互情况；Bote 等[130] 提出了 16 个指标用于衡量学生在网络课程每一章中的学习情况；Singh 等[139] 提出了 12 个反映不同层面学生参与度的指标并将其聚合为参与度评分（Engagement Score）用于综合衡量学生在网络学习过程中的参与情况。通过使用上述这些指标，可以将学生的在线学习过程刻画为多个指标构成的向量，进而使用统计检验和机器学习的方法展开进一步分析[140]。

根据上述研究，本文将学生 l 的学习参与度表示为一个 $1 \times k$ 维学习参与度指标向量 $\boldsymbol{e}_l$：

$$\boldsymbol{e}_l = [e_{l,1}, e_{l,2}, \cdots, e_{l,k}] \tag{3-3}$$

式中：$e_{l,i}$ 是学生 l 的第 i 个参与度指标的值。该值通过对学生 l 的学习过程 P_l 按照对应指标定义计算得到，即：

$$e_{l,i} = f_i(P_l) \tag{3-4}$$

式中：f_i 为第 i 个学习参与度指标的计算函数，根据指标的具体定义不同，函数 f_i 可能有多种形式。

对于学习参与度指标向量 $\boldsymbol{e}_l$，本文参考常见的基于交互类型的学习参与度指标分类方法[47, 65]，采用两类学习参与度指标从不同角度衡量学生的在线学习参与度：面向捕捉局部细粒度特征的学生参与度基础指标，以及面向解释学习过程特征的学生参与度概况指标。这两类指标共同构成学生 l 的学习参与度指标向量 $\boldsymbol{e}_l$。以下将分别说明上述两类学习参与度指标的具体定义与计算方法。

1) 学习参与度基础指标

首先，基础指标反映学生的基本的学生参与度，主要包括基于数量与基于时长的两类统计值反映学习过程中学习操作与资源使用情况的细节信息：

基于交互操作数量的学生参与度指标：此类指标反映了学生与 LMS 系统各项学习功能之间的交互活动数量，可以通过统计学习过程 P_l 中各个操作的数量得到，即对于操作集合 O 中的每一个操作 o^j，按以下公式计算：

$$e_{o^j} = |\{b_{l,i} | b_{l,i}.o = o^j \wedge b_{l,i} \in P_l\}| \tag{3-5}$$

基于资源访问时长的学生参与度指标：此类指标反映了学生对教学资源的使用时长，通过统计学习过程 P_l 中与资源集合 R 中的每个资源 r^k 被操作的时间距离下一个操作的时间差得到，按以下公式计算：

$$e_{o^j} = \sum_{b_{l,i}.r=r^k \wedge b_{l,i} \in P_l} (b_{l,i+1}.t - b_{l,i}.t) \tag{3-6}$$

以本文案例分析中使用的英语课程在线学习行为数据集为例，LMS 系统中为该课程提供了 76 个相关的在线学习行为操作以及 121 个短视频和 20 个阅读资料的教学资源。基于这些操作和资源，可以为每个学生生成一个 1×76 维的交互指标向量以及一个 1×141 维的资源指标向量，通过连接这两个指标向量可以得到一个 1×217 维的学习参与度基础指标向量。可以看出，由于网络课程的教学功能与教学资源在过程、内容、类型和数量方面存在明显差异，学习参与度基础指标向量的维度也是随之变化的。

2) 学习参与度概况指标

为了向领域专家和用户简要的解释学生学习参与度的整体表现，需要使用更加宏观的指标概括上述基础指标的总体特征。因此，本节基于相关文献中对学习参与度指标的研究[47, 109, 111] 以及领域专家的建议，采用以下 3 类经验指标从交互数量、学习时间、以及资源利用 3 个方面整体衡量一个学生在特定时间段内的学习参与度（例如，一个课程周期，一个学期、或者整个在读期间），如表 3-1所示。

其中，交互数量的 4 个指标涵盖了学生在线学习的主要行为类型，其数值大小反映了学生在线学习过程中的 4 类行为类型的活动强度[13, 65]；学习时间的 2 个指标从时间上概括了学生对在线学习的日程安排和投入情况[48, 139]；资源利用的 3 个指标概括了在线学习过程中最主要的两个教学资源，即课件视频和文本材料，的整体利用情况[11, 47, 130]。

表 3-1 学习参与度概况指标

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

根据表 3-1的定义，这些学习参与度指标可以通过结构化查询语言脚本以及其他程序脚本对学习日志数据分析得到。其中，交互数量的 4 个指标与前一节基于交互操作数量的学习参与度基础指标的计算方法类似，通过统计与特定交互类型相关的学习行为操作数量得到，可以在公式 3-5的基础上，将单个操作扩展到操作的集合，得到特定类型 s 的交互数量指标 e^s：

$$e^s = |\{b_{l,i}|b_{l,i}.o \in O^s \wedge b_{l,i} \in P_l \wedge O^s \subset O\}| \tag{3-7}$$

式中，O^s 代表特定交互类型 s 相关的学习行为操作集合，且该集合是全部操作的集合 O 的子集。

学习时间的 2 个指标中，在线天数指标可以通过计算学习过程 P_l 的全部学习行为的日期构成的集合的元素数量得到，其值 e^{dates} 如下：

$$\begin{aligned} D^{dates} &= \{date(b_{l,i}.t)|b_{l,i} \in P_l\} \\ e^{dates} &= |D^{dates}| \end{aligned} \tag{3-8}$$

式中，$date()$ 代表将学习行为时间戳转换为日期的函数，D^{dates} 是全部学习日期构成的集合。在此基础上，按学习日期集合统计每个学习日在线会话时长的平均值，得到单日时长指标值 e^{day}：

$$\begin{aligned} T^d &= \{b_{l,i}.t|date(b_{l,i}.t) = d \wedge b_{l,i} \in P_l\} \\ e^{day} &= \frac{1}{|D^{dates}|} \sum_{d \in D^{dates}} (max(T^d) - min(T^d)) \end{aligned} \tag{3-9}$$

式中，T^d 代表由学生在日期 d 当天内的全部学习行为的时间戳构成的集合。

资源利用的 3 个指标中，数量相关的前 2 个指标与基于资源访问时长的学习参与度基础指标的计算方法类似，通过按资源类型划分的集合统计得到，可以在公式 3-5的

基础上，得到资源利用数量指标 e^r：

$$e^r = |\{b_{l,i}.r | b_{l,i}.r \in R^s \wedge b_{l,i} \in P_l \wedge R^s \subset R\}) \tag{3-10}$$

式中，R^s 代表特定类型的教学资源集合，且该集合是全部教学资源的集合 R 的子集。另外，由于视频观看过程的操作复杂性，使用学习日志数据不易获得准确的观看时长指标值。为了提高指标准确度，本文借鉴文献 [65] 的计算方法，通过嵌入在客户端的 JavaScript 脚本程序监听文档对象模型（Document Object Model，DOM）的活动事件，实时监控学生的观看操作序列与视频活动状态得到。

3.3.4 基于哈希表的学习参与度数据存储

为了分析每个学生以及各学生群体的学习参与度情况，可视分析系统通常允许用户快速查看任意学生的学习参与度。因此，在线学习行为数据存储模型中设计的聚合数据层（图 3-1）需要提供对每个学生的学习参与度统计结果的索引访问。按照在线学习行为数据存储模型中的定义，文本为每个课程在每个特定时间段都关联一张选课学生的学生标识符列表，以及一张相关学生的学习参与度哈希表。在学习参与度哈希表中，每项记录以学生唯一标识符为键，以学生在该课程该时间段的学习参与度指标向量 $\boldsymbol{e}_l$ 为值。为了提高查询性能，该哈希表使用 B+ 树进行索引。

例如，对于课程 c 在时间段 t 的每个学生 l，按照上述的学习参与度基础指标与概况指标计算方法计算学习过程 P_l 对应的学习参与度指标向量 $\boldsymbol{e}_l$。随后，如图 3-4所示，将课程 c 在时间段 t 的全部学生构建学生标识符列表，并以学生标识符为键，学习参与度指标向量 $\boldsymbol{e}_l$ 为值构建课程 c 在时间段 t 的学习参与度哈希表。当探索指定课程与时间段的学习参与度时，可以通过该哈希表快速生成任意学生群组的指定特征维度列表。

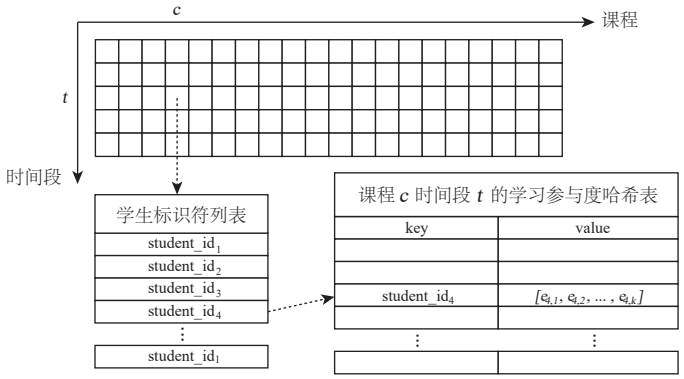


图 3-4 聚合数据层中按课程和时间段划分的学生标识符列表与学习参与度哈希表

基于哈希表的学习参与度数据存储在写入和读取两方面都具有较好的表现。首先，经过预处理后，对在线学习行为数据中的学习日志数据计算每个学生的学习参与度指标向量 $\boldsymbol{e}_l$，以学生标识符 l 为键插入学生标识符列表和学习参与度哈希表。由于键名的生成无需计算，单条记录写入列表和哈希表的时长约为常数时间，因此写入单条记录的时间复杂度为 $O(1)$，写入一批记录为线性时间复杂度 $O(n)$；其次，可视分析系统在读

取每个学生的学习参与度时，以学生标识符 l 为键查询学习参与度哈希表，因此，读取操作的时间复杂度为 $O(1)$ 常数时间。相较于常见的数据文件形式或者数据表形式，采用上述方式存储学生的学习参与度，能够实现更快的检索速度和更短的界面响应时间。

3.4 可视化设计与交互界面

上述学习参与度指标模型可以刻画在线学习过程数据中不同侧面的在线学习过程特征，但是使用该模型得到的学习参与度指标向量数据维度较高，同时由于学生数量较多，使得学习参与度数据量较大，进而对分析学习参与度模式提出了两个挑战：首先，大量的学习参与度数据的高维度特性使得其分布情况及其蕴含的模式规律不易直观理解，其次，在线学习过程涉及多种与学习相关的因素，这些因素对学习参与度模式的影响不易探索。为了解决这两个挑战，学习参与度的可视化需要满足 2 个方面设计需求：

R.1：体现学生学习参与度在各维度的分布模式。学习参与度的不同维度具有不同的分布，大规模高维的参与度中可能存在不同的学习模式，如何展示这些不同的参与度模式是可视化的重点。

R.2：展示不同群组之间的学习参与度分布差异。学习参与度数据可以按照学生的各种属性划分为不同的群组，查看群组学生的参与度分布以及对比不同群组之间的差异有助于理解不同因素对学习参与度模式的影响。

基于上述两个设计需求，本文设计了一种综合散点图矩阵与 t-SNE 数据降维技术的学习参与度分布图，并设计了学生群组创建工具和学生群组学习参与度对比工具，实现从整体上展示全部学生在各维度的分布情况，探索各学生群组的学习参与度差异。本节将对这些设计进行详细说明。

3.4.1 学习参与度分布图

为了体现高维度学习参与度的模式特征（R.1），需要直观的展示学习参与度在不同维度上的分布情况。常用的展示数据分布的可视化方法包括直方图、概率密度分布图、累计概率密度分布图、热力图、以及散点图等，这些方法能够直观的呈现一维或二维数据的分布特征，因此在教育数据的分析中经常使用并且易于理解。而对于高维数据可视化中常见的平行坐标系等方法，存在数据量较大时连线过于密集难以辨识的问题，不适合从中识别各学生群组之间的差异（R.2）。因此，为了便于用户以容易理解的可视化形式探索潜在的学习参与度模式，本节设计了结合散点图矩阵和 t-SNE 数据降维技术的学习参与度分布图显示高维度的学习参与度。以下将详细说明学习参与度分布图的可视化设计与使用的降维方法。

1) 可视化设计

由于散点图能够有效的展示每一个数据点在二维平面中的位置，并能够通过数据点密集区域识别出潜在的分布模式，本文采用散点图作为学习参与度分布图的基本图

表，并结合散点图矩阵展示学习参与度的其他维度分布（图 3-5）。

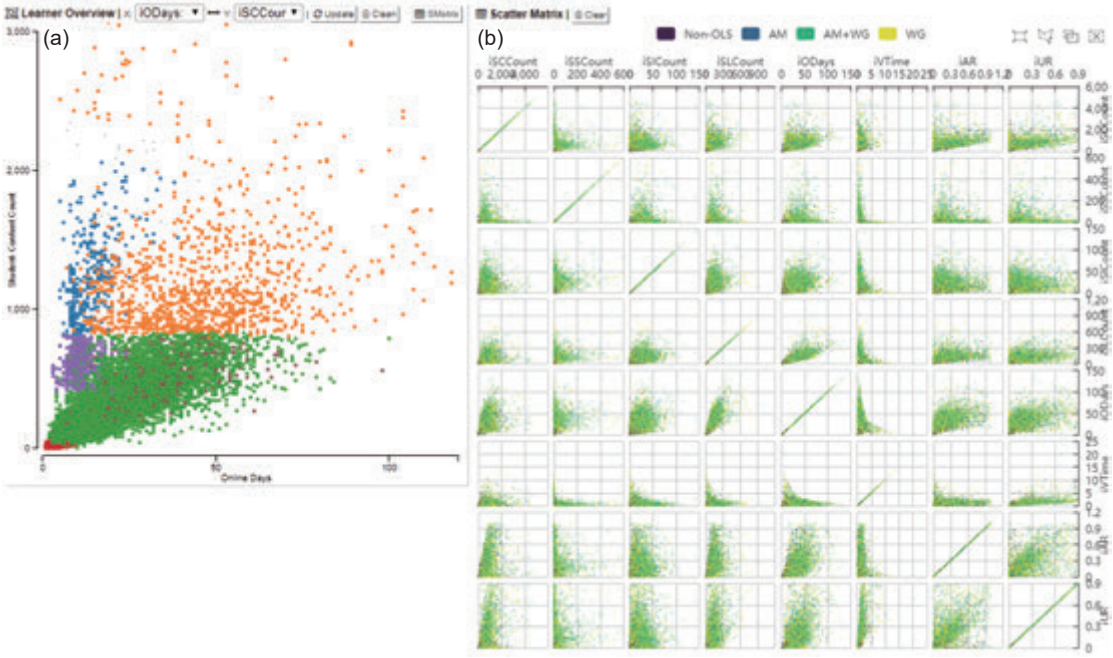


图 3-5 学习参与度分布图界面截图，其中 (a) 为学习参与度分布图，(b) 为学习参与度概况指标散点图矩阵

如图 3-5(a) 所示，学习参与度分布图以散点图为主要视图，该视图能够同时展示学习参与度中的两个维度的联合分布。其中，每个点代表一个学生数据点，横坐标和纵坐标分别对应两个维度。每个数据点的颜色默认为灰色，当选择学生群组后，各学生群组按分组信息映射为数据点颜色。该散点图视图与散点图矩阵关联，通过选择散点图矩阵中任意一个联合分布的缩略图（图 3-5(b)）可以更新散点图的坐标。

2) 基于 **t-SNE** 的学习参与度向量降维

尽管上述可视化设计能够帮助用户理解学习参与度概括指标的分布情况，但蕴含在高维学习参与度中的模式仍然无法呈现，需要降低数据维度以便探索分析。在现有学习参与度研究中，为了确保教育数据分析过程中指标与结果之间的相关性、因果关系，通常对指标的可解释性要求较高，因此基于特征选择的方法是较为常用的数据降维方法，而使用特征提取方法降低数据维度较少。本文利用可视化技术与交互技术呈现高维数据，能够在多个视图展示降维后数据的原有维度特征，可以缓解由于数据降维造成的可解释性问题，因此本节采用基于特征提取的数据降维方法将高维度的学习参与度映射到二维平面，协助用户在二维平面上观察数据模式。

如第 2 章所述，在众多基于特征提取的降维方法中，t-SNE 算法在降低数据维度的同时也在保留数据局部模式与整体分布方面具有较好的表现。t-SNE 通过构建一个高维数据点之间的概率分布，使得相似的数据点有更高的概率被选择；然后，在低维空间里构建一个 t 分布用于表达两个数据点之间的相似度，并使得这两个概率分布之

间尽可能相似。最终，通过仿射变换将高维数据点映射到低维概率分布上[67]。相比于其他降维算法，如主成分分析（Principal Component Analysis，PCA），多维尺度分析（Multidimensional Scaling，MDS），以及自组织映射（Self-organizing Map，SOM）等，t-SNE 能够更好的同时考虑全局与局部关系，适合将高维数据降至二维平面从而在视觉上直观的认识数据的结构特点。

t-SNE 算法在数据降维效果方面具有良好的表现，但在运行效率方面，标准 t-SNE 算法的时间复杂度为 $O(n^2)$[67]，广泛使用的 Barnes-Hut 实现版本[68] 的时间复杂度为 $O(nlog(n))$，因此在数据点规模较大或者维度较高时，生成低维映射的时间较长，无法及时展示。然而在探索性数据分析的过程中，用户可能会尝试不同学习特征特征维度的组合，并查看不同课程、不同时间段的学习参与度分布情况，降维算法的运行时间对于分析过程的用户体验存在直接影响。因此，本文调研选择了 4 种不同版本的 t-SNE 实现，并在 MNIST 数据集上对比了不同版本 t-SNE 的运行效率，包括 scikit-learn（SKlearn）[141], Barnes-Hut[68], Multicore t-SNE（本实验使用 8 核并行版本），以及 FIt-SNE（Fast Interpolation-based t-SNE）[70]。

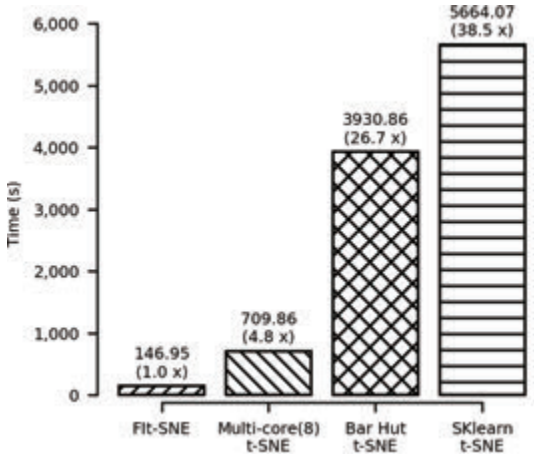


图 3-6 各 t-SNE 实现版本的运行时间对比

MNIST 数据集使用 60,000 个测试样本，并将原 28×28 维特征转换为 1×784 维特征。每个 t-SNE 版本的主要参数均为 n_components=2，perplexity=50。实验主机为 Intel i7-7700 4.2 GHz 处理器，8G 内存，Ubuntu 18.04 LTS 环境。各版本的运行时间表现结果如图 3-6所示。可以看出，FIt-SNE 版本的 t-SNE 算法运行时间显著小于其他版本，能够在可接受的时间范围内获得可以接受的数据二维表示。因此，本文选择使用 FIt-SNE 算法对高维度的学习参与度进行降维。

3.4.2 交互设计

为了支持用户自由探索不同属性的学生群组（R.2）以及各群组的学习参与度分布模式，本节设计了多种学生群组创建工具，允许用户采用基于人工规则和基于算法识

https://github.com/DmitryUlyanov/Multicore-TSNE

别这两种方法创建学生群组，并设计了学生群组学习参与度对比工具展现各学生群组在学习参与度概况指标方面的差异，为探索学习参与度的局部模式提供支持。

1) 学生群组创建工具

由于学习参与度高维度以及学生属性的多样性，学生群体可能以不同的方式被定义。例如，不同的年龄范围，生源地区，以及视觉上的数据点距离等。本节设计了 2 个创建学生群组的工具，分别是基于人工规则和基于算法识别的学生群组创建工具：

基于人工规则的学生群组创建工具：用户可以选择部分数据点作为群组进行分析，例如按照年龄区间、成绩分类、交互数量范围等划分群组，从而实现精确的筛选全体学生中的子集。针对目标用户群体的使用习惯，本节设计了 3 个基于人工规则的群组创建工具，分别是条件选择器（图 3-7(a)）、SQL 选择器（图 3-7(b)）、以及学生 ID 选择器（图 3-7(c)），分别提供自定义阈值规则列表、编写 SQL 条件子句、以及指定学生的编号 ID 列表这三种方式实现筛选数据点。

(a) 条件选择器　(b) SQL 选择器　(c) 学生ID选择器

图 3-7 基于规则的群组创建工具

基于算法识别的学生群组创建工具：使用 t-SNE 算法将学习参与度向量降至二维后，数据点在二维平面上可能会聚集形成多个连续形状的密集区域，这些不同密度的区域暗示了潜在的高维数据模式。利用这些区域之间的数据点密度差异，本文设计了一个基于 DBSCAN（Density-Based Spatial Clustering of Applications with Noise）算法的自动群组识别工具。通过调整 DBSCAN 算法的邻域样本数阈值 MinPts 和距离阈值 epsilon 这两个参数（图 3-8(b)），可以查看不同参数下集群的检测结果。在学习参与度分布图中，散点图的数据点将根据检测结果指定的集群被染色（图 3-8(a)）。当自动识别的集群符号分析需求时，可以将这些检测结果保存为学生群组。

2) 学生群组学习参与度对比工具

由于高纬度学习参与度在数据降维后的低维表示并不具有物理意义，且散点图呈现的各个密集区域之间的距离大小与各集群的相似度大小无关[67, 68]，使得对各集群之间差异的现实含义难以直观理解。针对该问题，本节设计了学生群组学习参与度对比

https://github.com/upphiminn/jDBSCAN

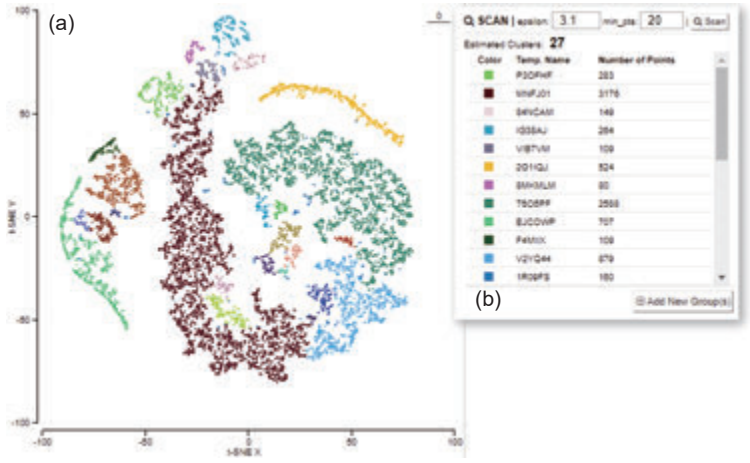


图 3-8 基于 DBSCAN 算法的学生群组创建工具界面截图，其中 (a) 为按学生分组染色的学习参与度分布图，(b) 为算法自动识别的学生分组列表

工具，辅助用户对比各学生群组之间在学习参与度概况指标上的分布情况，进而分析各集群的学习参与度模式差异（R.2）。

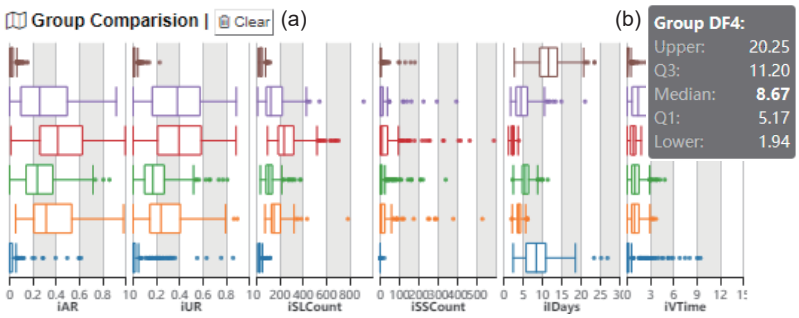


图 3-9 学生群组学习参与度对比工具界面截图，其中 (a) 为基于盒图展示了各学生群组在各学习参与度指标的分布，(b) 为鼠标悬停在盒图上时提示的分布详情信息

该对比工具包括 2 个部分，分别用于展示学生群组的学习参与度分布盒图（图 3-5(a)），以及学习参与度盒图详情（图 3-5(b)）。学生群组列表中显示了每个群组的主要特征，包括学习参与度维度、名称、编号、以及学生数量等。学习参与度盒图展示了各学生群组在主要学习参与度概要指标上的分布，每个盒图代表一项学习参与度概要指标，当鼠标停留在盒图符号时，会进一步显示该盒图对应的分布详情（图 3-5(b)）。为了在不同的视图中区分各个学生群组，首先将群组编号映射为颜色编码，各个视图中相同颜色的符号均代表同一学生群组。

3.5 案例分析

为了验证本章提出的可视化设计的有效性和用户体验，本节采用案例分析方法对基于上述可视化与交互设计开发的可视化分析系统进行评测。整个案例分析过程包括三个步骤：首先，向参与评测的本校网络教育学院专家用户介绍可视化分析系统的主

要功能；然后，由专家用户参考分析任务利用该系统自由探索真实数据集；最后，针对可视化系统的使用体验访谈专家用户并汇总他们的反馈意见。

案例分析中使用的数据集来自本校网络教育学院真实的在线学习行为数据集。该数据集包含 2015 年入学的学生在第一学期（2015 年 3 月至 7 月）的学习日志数据和相关其他类型数据。经过数据清洗后，共包含学生 10,529 名，学习日志记录 14,942,964 条。按照第 3.3节的方法预处理和抽象后，将学习行为数据保存在 HBase 集群，然后基于分布式内存计算框架 Spark 实现各项学习参与度指标的统计和分析，分析结果保存在 MongoDB 数据库中。基于上述数据集，本节以网络课程学习中学生支持服务与学习参与度的关系进行案例研究，包括分析学习参与度分布，以及学生支持服务使用情况对学习参与度模式的影响。

在线学生支持（Online Learner Support，OLS）服务是在线学习过程中重要环节，学生可以通过这个服务获得关于技术困难、课程学习、教学安排以及考试测验等方面的帮助。除了传统的线下答疑、热线电话以及辅导资料以外，本校网络教育学院利用网络平台技术，实现了在线的智能客服（Assistance Messenger，AM）和定期导学（Weekly Guidance，WG）两个在线支持服务。

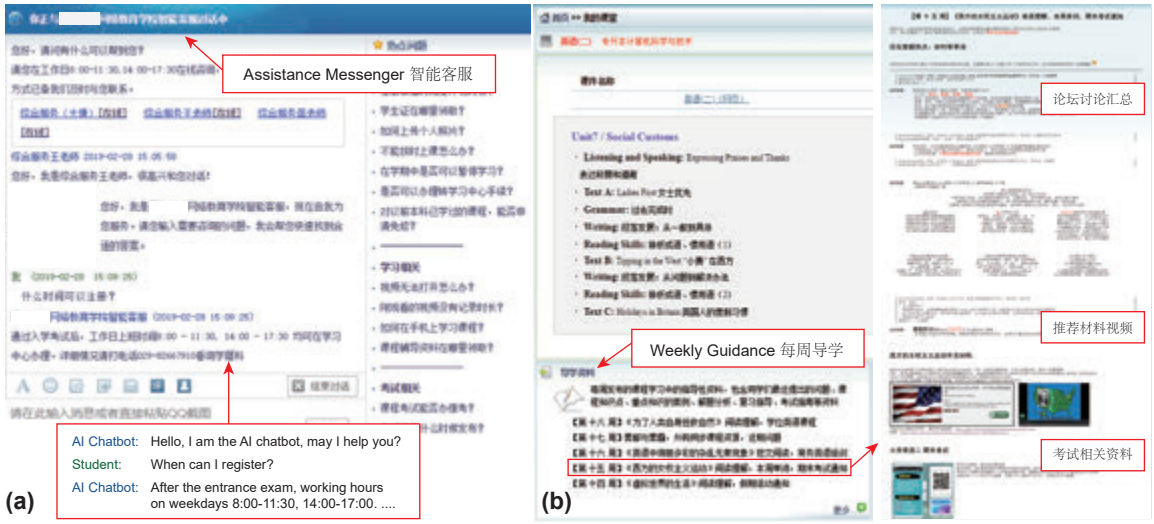


图 3-10 在线学生支持服务界面截图，其中 (a) 为智能客服（AM）用户界面，(b) 为定期导学服务（WG）的列表界面与详情缩略图，包含了论坛讨论、推荐材料和考试相关资料等

如图 3-10(a) 所示，智能客服是一个在线实时对话系统，学生可以随时通过会话式界面向教师、辅导员、技术人员发送文本、图片、语音等信息请求帮助，或通过人工智能对话机器人询问常见的问题。如图 3-10(b) 所示，定期导学服务是一个由教师定期发布的学习资料，内容包括近期学生论坛讨论问题、难点解读以及考试重点等，该服务可以帮助学生更好的学习课程知识，提高学习表现。教育研究认为，学生支持服务的使用对学生的满意度、参与度、学习成效等有直接影响 [4, 10]。因此，教师需要分析学生支持服务的使用情况，及其对学习参与度的影响。领域专家使用基于本章提出的可视化方法与交互工具开发的可视化系统探索该数据集，从不同角度分析学生的学习参与度分

布以及各学生群组的学习参与度模式。

3.5.1 学习参与度的分组分析

如图 3-11(a) 所示，采用本章提出的学生群组创建工具，将学生按是否使用 OLS 服务分成了 4 组并映射颜色编码：1）未使用任何服务组（Non-OLS Group）蓝色；2）仅使用智能客服组（AM Group）橙色；3）仅使用定期导学组（WG Group）绿色；以及 4）同时使用两个服务组（AM+WG Group）红色。使学习参与度分布图查看学习参与度概况指标的分布情况可以发现（图 3-11(b)），在“学生-系统”与“学生-内容”这两个维度的分布下，学生布存在两个显著较为密集的集群，分别位于图 3-11(b1) 区域以及图 3-11(b2) 区域。可以看出，(b1) 区域的学生数量庞大，并且以 WG 组和 AM+WG 组为主，而 (b2) 区域的学生较少，以 Non-OLS 组为主。

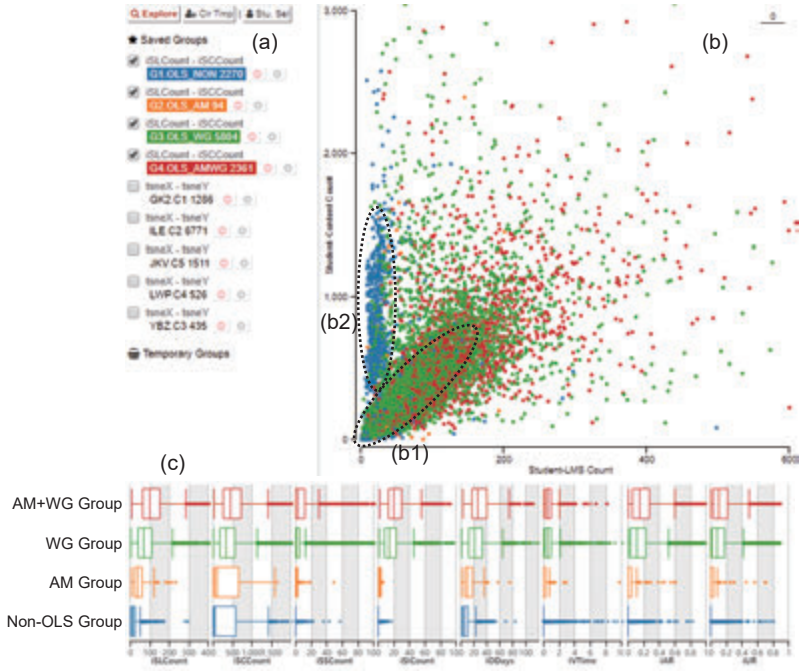


图 3-11 4 个学生群组的学习参与度分布图界面截图，其中 (a) 为选中的 4 个学生群组并以 4 种颜色编码，(b) 为 4 个学生群组按颜色编码的学习参与度分布图，(c) 为 4 个学生群组在主要学习参与度指标的分布情况

进一步使用学生群组学习参与度对比工具可以发现，这 4 个群组在主要的学习参与度概况指标上存在明显差异（图 3-11(c)）。其中，AM+WG 组则明显高于其他群组（红色），而 Non-OLS 组在各项指标均明显低于其他群组（蓝色）。

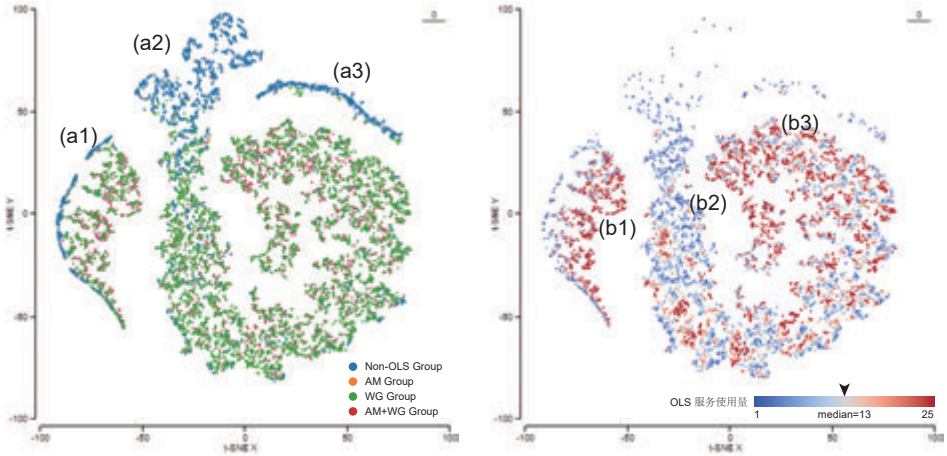
领域专家使用学习参与度分布图和基于人工规则的学生群组创建工具进一步查看了这 4 个学生群组在其他学习参与度指标上的分布情况，他们认为图 3-11(b2)（蓝色）区域呈现的集中分布与他们日常的教学经验一致，即一部分学生会在特定时间段密集学习，例如考试前突击学习或者提交课程作业前复习。利用本文提出的可视化设计，领域专家不仅从可视化效果中验证了他们的教学经验，更进一步发现了这部分学生在使

用 OLS 服务方面的学习行为特点。他们认为这个发现说明 OLS 服务的使用情况与学习参与度之间可能存在正相关关系，而且从图 3-11(b) 可以看出，不使用 OLS 服务学生的学习参与度呈现出不同的分布情况。

领域专家认为上述群组明显的学习参与度分布差异主要来自于“学生-内容”（iSCCount）指标的分布差异。Non-OLS 组学生在其他各项学习参与度指标明显较低的情况下，在“学生-内容”（iSCCount）指标上与其他各组的分布接近，说明该组学生可能在短时间内访问了大量的课程学习资源。这种分布差异表明学生的学习过程存在不同的模式，并且这些模式与他们使用 OLS 服务的情况有关。

3.5.2 使用学生支持服务与学习参与度模式的关系

为了进一步探索使用 OLS 服务与学生的高维度学习参与度模式之间的关系，领域专家选择学习参与度分布图的横纵坐标为 t-SNE 算法降维后的二维表示，并按是否使用 OLS 服务的分组颜色编码对数据点着色，得到如图 3-12(a) 所示的分布情况。



(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

图 3-12 学习参与度的降维分布，其中 (a1)、(a2) 和 (a3) 区域主要是 Non-OLS 组的学生，(b1)、(b2) 和 (b3) 区域反映 Non-OLS 组以外的其他群组学生使用 OLS 服务的数量

可以看出，降维后的数据呈现出多个明显不同的区域，并且区域分布与是否使用 OLS 服务相关。其中，Non-OLS 组（蓝色）主要集中在图 3-12(a) 的 (a1)，(a2)，以及 (a3) 这 3 个区域，而 WG 组（绿色）以及 AM+WG 组（红色）也呈现出不同密度区域分布，这些分布情况说明 Non-OLS 组以及其他组内部也可能有不同的学习参与度模式，并且各学习参与度模式之间的差异与是否使用 OLS 服务相关。进一步将 OLS 服务的使用数量映射为数据点颜色后，从图 3-12(b) 可以发现在 WG 组和 AM+WG 组分布密集的区域中，使用 OLS 服务数量的分布也有明显差别。其中，图 3-12(b1) 和 (b3) 区域的学生明显使用 OLS 服务的次数高于中位数，而 (b2) 区域学生使用 OLS 服务的数量明显较少。这个分布说明学习参与度的模式与使用 OLS 服务的使用量之间也存在关系，某些学习参与度模式的学生可能更倾向于使用 OLS 服务。

进一步使用基于 DBSCAN 算法的工具自动创建学生群组后，可以识别 5 个集群，

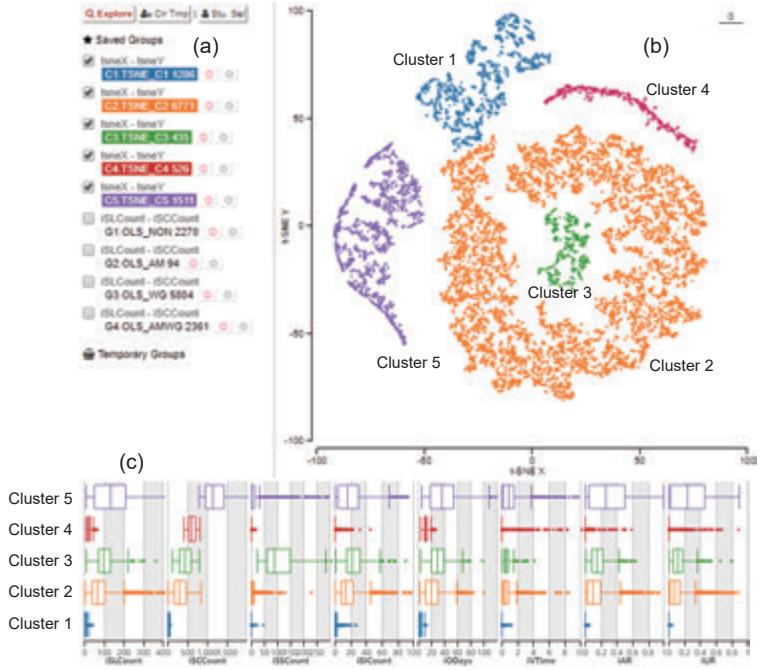


图 3-13 学习参与度的降维分布形成的局部集群，其中 (a) 为选中的 5 个学生集群并以 5 种颜色编码，(b) 为各集群按颜色编码的分布情况，(c) 为各集群在主要学习参与度指标的分布情况

如图 3-13所示。结合图 3-12(a) 并使用学习参与度对比工具查看各集群的学习参与度分布可以发现，Non-OLS 组所在的集群中，集群 1（蓝色）的各项指标均显著低于其他各集群，集群 4（红色）除了 iSCCount 指标（学生-内容交互数量）较高以外其他各项指标也低于其他集群，集群 5（紫色）除了 iSSCount 指标（学生-学生交互数量）以外所有指标明显高于其他所有群组。另外，集群 2（橙色）的学生数量最多且各项指标比较均衡，集群 3（绿色）的各项指标整体较高并且 iSSCount 指标（学生-学生交互数量）明显高于其他各集群。

这些可视化结果说明，不使用 OLS 服务学生的学习参与度呈现三种不同的特点：1）很少参与课程学习（集群 1）；2）少量浏览课程内容学习（集群 4）；以及 3）大量浏览课程内容但不观看视频（集群 5）。这些参与度模式反映的学习行为，用户可以进一步分析学生的学习动机与教育背景。使用支持服务的学生也表现出 3 种不同的参与度模式：1）少量参与课程学习（集群 2）；2）密集讨论型学习（集群 3）；以及 3）频繁学习（集群 5）。进一步将支持服务使用数量映射到学习参与度分布图的数据点的颜色，可以发现集群 2 和集群 4 的数据点整体呈蓝色，集群 3 和集群 5 整体呈红色，即具有较高学习参与度的学生可能具有较高的支持服务使用数量，说明支持服务使用数量也和学习参与度之间可能存在正相关性。

领域专家认为结合图 3-12和图 3-13的可视化结果符合他们日常观察的学生在线学习活动，首先，集群 3 整体表现出较高的学生-学生交互数量，说明集群 3 的学生可能更加愿意与 LMS 系统中的其他成员进行交流，而这种主动交流的学习行为与使用 OLS 服务之间显示出潜在的正相关，符合日常教学活动中具有积极主动交流特征的学生的

行为特点。其次，集群 5 表现出的频繁学习与使用 OLS 服务之间的正相关性也说明学习课程内容较多的同学可能在学习过程中遇到更多问题，因此使用 OLS 服务的数量也较高。同时，集群 1 显著较低的学习参与度和 OLS 服务使用量，说明当学生很少学习时，他们可能发现自己学习中的问题，进而没有使用 OLS 服务的需求。基于这些发现，领域专家建议可以将在线学生支持服务的使用情况作为分析学习动机、衡量学习参与度、以及评价学习满意度的参考依据，进而为改进教学管理、提高教学质量和学习体验提供支持。

3.6 本章小结

本章将高维数据的可视分析方法应用到在线学习过程数据的分析中，提出一种基于散点图矩阵与 t-SNE 数据降维技术的学习参与度可视化方法，设计了多种数据群组创建工具，并将其集成在协调关联的多个视图中，辅助探索学生参与度的高维模式和不同学生群体的学习参与度分布。

首先，本文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过程数据的采集、清洗、抽象和分层存储，同时从中提取了反映学生学习参与度的指标并建立索引。其次，本文针对高维学习参与度数据难以分析和解释的特点，提出了基于 t-SNE 算法的降维可视化分析方法，并设计多个关联视图与交互式的群组创建工具，通过对多个维度、学生群组等数据的可视化编码展示学习参与度模式特征与差异。为了验证方法的有效性，本文利用网络学习者的在线学习日志数据展开研究，发现了多种不同的学习参与度模式，并分析了在线学生支持服务使用情况与学习参与度之间的关联关系。

以本章研究内容为基础撰写的论文发表在国际期刊 IEEE Access，题为“How Learner Support Services Affect Student Engagement in Online Learning Environments”，国际会议 IEEE International Conference on Advanced Learning Technologies (ICALT 2018)，题为“Measuring Student’s Utilization of Video Resources and its Effect on Academic Performance”，以及国际会议 IEEE International Conference on e-Business Engineering (ICEBE 2014)，题为“Big Log Analysis for e-Learning Ecosystem”。

4 面向时序数据的学习时间管理可视分析

4.1 引言

知识的学习总是伴随着持续性的时间投入，小到一章课件或着一个实验、大至一门课程乃至一个专业，都需要学生投入大量时间参与多个教学环节才能完成。在线学习网络课程的过程也是如此，并且随着网络课程内容的系统化与教学活动的多样化，所需要的学习时间正在进一步增加。例如，常见的大规模开放在线课程（Massive Open Online Courses, MOOC）中，单门课程通常需要 4-18 周的学习时间[2, 142]，专项技能培训需要 6-12 个月[143, 144]，而在线学历教育则需要长达 2.5-5 年[65]。在如此长时间的在线学习过程中，学习时间的安排和学习进度的管理会直接影响课程知识掌握与最终的学习成效[145]。因此，不仅教师需要了解学生在长期学习过程中的时间管理情况进而提供支持帮助，学生也需要掌握和调整自己的学习时间安排保持学习进度。

得益于网络课程学习过程的异步性与教学资源的开放性，学生可以不受时空条件限制的随时随地展开学习。已有报告和研究发现，这种便利性在为学生提供充分机会管理自己的学习时间和学习进度同时，也对学生的自我认知、自律性、时间管理能力和风格的挑战[146, 147]。例如，You 等[148] 发现自律学习（Self-regulated Learning）、会话次数以及提交作业的时间与最终课程成绩之间有显著的相关性，按教学计划规律的观看视频并提交作业的学生可能取得更好的考试成绩。Broadbent 等[146] 对自律学习策略相关研究的综述发现，学生的时间管理策略，规律学习，批判性思维等因素与学生的学习收益正相关。Kizilcec 等[149] 对 MOOC 课程学习者的在线学习过程进行研究发现，学生的目标与学习策略有助于预测学生个人的课程成绩，并且具备更好自律学习技能的学生更有可能复习学过的课程内容，另外学生的个人背景信息可能与自律学习有关。Kim 等[150] 发现采用不同自律学习策略的学生在学习过程中呈现不同的模式，其中不自律学习的学生在时间管理，在线服务使用，学习频率等方面明显低于自律学习的学生。

尽管网络课程能够为学习者推荐有益的自律学习策略，但单纯的向学习生提供这些自律学习策略信息可能无法实际改进学生的学习时间安排，也难以改变学生的学习成绩[151]。由于在线学习过程中授课与学习这两个过程的时空分离性，教师无法像传统课堂中那样掌握学生的学习时间和进度，学生也无法理解自己的学习时间管理是否有效，教学双方对学生学习过程的时间管理特征都缺少统一的认识。现有针对学习行为时序数据的时间模式分析主要是基于教育学理论的方法，通过运用领域专家的专业知识和授课经验，利用聚类分析、相关性分析等方法研究时间管理相关的因素与模式，包括预测学习成效[149]、推荐自律学习策略[151]、提供学习辅助工具[152, 153]、实施教学干预[154]等。然而，在大规模的网络课程学习中，课程的复杂性与学生数量的庞大使得分析学生的学习时间安排更加困难，迫切需要利用技术手段分析学生的在线学习行为数据，分析数据背后的时间管理风格，为提高学习收益和教学质量提供支撑。

上一章介绍了在线学习行为数据的存储模型与管理系统，通过利用该系统可以得到包含时间戳字段在内的完整在线学习行为记录。得益于这些字段，在线学习行为数据具有天然的时序性，因此在线学习行为数据能够刻画学生在线学习过程时间维度信息。通过挖掘在线学习行为数据中时间维度信息，可以得到反映学生时间管理风格的学习时间安排情况。这不仅可以帮助教师与教育机构更好的编排教学大纲[11]、调整课程进度[122]、提供支持服务[113]、预测学习行为[149]，还可以帮助学生了解自身学习进度[153]、督促完成作业[154]、以及调整学习计划[152]。然而，由于在线学习行为数据中时序信息的复杂性，现有研究的分析结果通常将长期学习过程的时序信息压缩在有限的数值中，使得抽象的数值分析结果难以被教师和学生直观理解，且难以分析在压缩过程中损失的时间维度变化过程。

为了突破现有研究的局限性，本章在前一章学习参与度指标模型的基础上，进一步抽象出学习参与度时序矩阵以保留学习过程时序特征，并采用基于卷积神经网络的模型训练学习参与度时序矩阵数据，预测不同时间管理风格的考试成绩。随后，为了直观理解学习参与度时序矩阵中刻画的学习时序特征，本文设计了基于日历坐标系的学习参与度日历矩阵和基于日历图的学习时间分布图，展现长时间学习过程中的学习参与度的时间信息与周期变化情况。最后，针对在线多课程同期学习的特点，设计了多课程学习进度图，从课程资源和时间两个维度细粒度的分析长期学习过程中视频观看活动的时间安排。

在后续的各节中，本文将在第 4.2节讨论可视化设计的分析任务，在第 4.3节说明如何抽象在线学习行为数据中的时序特征，在第 4.4节详细介绍在线学习行为数据时序特征的可视化设计与交互设计，在第 4.5节结合真实的在线学习行为数据，利用上述设计分析学生的学习时间管理风格与多课程的学习时间安排情况，最后在第 4.6节对本章工作进行分析和总结。

4.2 在线学习时间管理模式分析任务

4.2.1 分析任务与设计需求

在传统课堂教育中，授课过程严格遵循既定的时间表，学生必须在统一的时间段同步学习课程内容，并且按照统一时间点完成实验、提交作业和参加考试。然而在网络课程的在线学习过程中，弹性的学习时间与开放的学习资源允许学生自己安排各项学习活动的时间。利用在线学习这种“随时随地”的便利性，学生可以按照自己的学习能力与学习进度，灵活的安排自己的学习计划。

但是，这种灵活性在赋予学生自主管理学习进度与安排学习内容的机会同时，也对学生的时间管理能力提出了挑战 [146, 147]。有些学生安排工作日时间学习，有些学生安排周末时间学习，而有些学生则可能集中安排一两周突击学习，这些不同的时间安排反映了学生的工作节奏和时间管理风格，并且可能对他们的学习成效产生影响[145]。因此，理解学生对学习时间的安排以及他们的时间管理风格，对于改善学习体验、改进教

学服务水平、提高学习成效具有显著意义。

为实现这一目标，通过访谈在线教育方面的领域专家，本章针对学生的学习时间管理的探索分析提出以下 3 个的分析任务：

- **T.1**：分析学生在给定时间段内的学习时间安排。在长期的学习过程中，多种因素会影响学生的在线学习时间，进而造成学生的学习时间安排的变化，直观呈现这种时间安排的变化对于分析和理解学生的学习过程有重要作用。
- **T.2**：分析学生对多个课程的学习时间安排。在网络课程的学习过程中，学生经常需要同时在线学习多门课程，由于没有像传统课堂中那样已经安排好的课程表，如何安排各个课程的学习时间与进度是学生需要自己根据个人情况调节的。分析同期学习多课程的时间安排有助于教师掌握学生的学习进度，也能帮助学生自己理解和规划学习时间。
- **T.3**：分析不同时间管理风格与学习参与度的关系。学生的时间管理风格会可能影响学生投入在线学习的时长以及学习时间的安排，而学习时间与时间安排会直接影响到学习参与度，进而影响最终的学习成效，研究这两者之间的关系有助于理解和预测学生的学习参与度变化并为实施教学干预提供参考数据。

为了完成上述分析任务，本章针对在线学习行为数据中学习过程的时序特点，概括了以下的设计需求指导可视化设计：

- **R.1**：呈现学习参与度随时间的变化情况。在长时间的学习过程中，学习时间的安排与学习参与度之间可能存在关联，呈现一个时间段内的学习参与度随时间的变化情况有助于理解这两者之间的关系（T.1，T.3)。为了探索学习参与度在时间维度上的变化，可视化应当直观的呈现这种变化过程。
- **R.2**：展现多课程学习参与度随时间的变化情况。在一段时间内多门课程的学习进度可能存在交织共存的状态，进而使多课程的学习参与度变化模式更加多样（T.2），可视化设计应当展示这种多课程交织状态的模式特点。
- **R.3**：学生群体对比分析。按照学习参与度的指标可以将学生划分为不同的学生群体，对比分析两个或者两个以上学生群体在各方面的差异有助于理解时间管理风格与其他因素的关系（T.2，T.3）。例如，个体学生与特定学习参与度模式群体学生、特定成绩区间学生之间的学习时间安排差异以及学习参与度差异。因此，可视化应当支持不同学生群体之间学习时间和学习参与度方面的对比。
- **R.4**：交互探索。由于学生在学习参与度、时间管理以及成绩等方面存在关联关系（T.1，T.2，T.3），可视分析系统应当提供对学生个体和群体作选择、过滤、关联等交互式探索。

4.2.2 系统架构

针对上述任务与可视化设计原则的分析，本文设计了 LearnerExp 可视分析系统，其系统架构如图 4-1所示。LearnerExp 可视分析系统整体分为 4 个部分：数据采集模块、数据存储模块、学习时间特征提取模块以及可视化模块。其中，数据采集模块负责从学

习管理系统中收集在线学习行为数据，并在预处理后保存在数据存储模块中；学习时间特征提取模块负责按照本文提出的学习参与度时序模型从在线学习行为数据中抽象出学习参与度的时序特征，并基于这些特征预测学习成绩；最后，可视化模块负责提供交互式的学习过程分析与展示，帮助教师和学生深入理解学生的学习时间管理特征。

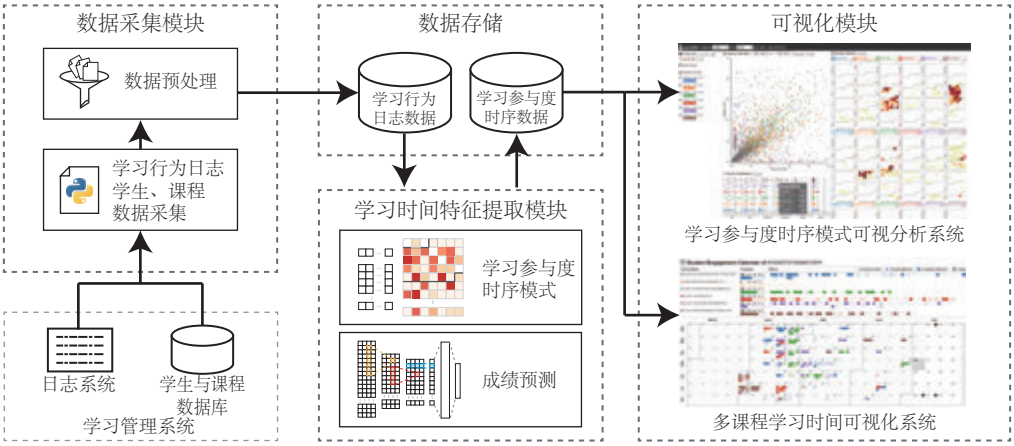


图 4-1 在线学习行为时间管理模式分析可视化查询平台架构

4.3 学习参与度时序模型

为了描述在线学习过程中的学生对学习时间的安排（T.1），需要刻画在线学习行为数据中包含的时间特征。前一章将学生 l 长时间的在线学习过程 P_l 抽象为学习向量 $\boldsymbol{e}_l$，虽然其中包含了关于学习时间的概况指标，如学习天数和在线时长等，从而反映学习过程总体投入的时间，但是在计算这些指标的标量数值时，已经从在线学习过程 P_l 中消除了过程性的时间信息与顺序信息。因此，无法进一步从这些指标的统计结果中理解学生是如何在长期的学习过程中安排每天的学习时间使用教学资源。

针对这个问题，本节对在线学习行为数据的时序特征进行抽象，提出一种基于日期的学习参与度时序矩阵，刻画在线学习过程的时序信息。另外，针对学习参与度时序矩阵中表达的时序信息与学习参与度信息，本节设计了基于卷积神经网络（Convolutional Neural Network，CNN）的预测模型从学习参与度时序矩阵中提取学习参与度的短期和长期时间特征，预测学生的学习成绩。

4.3.1 学习参与度时序矩阵

本文前一章的研究中，学生 l 的在线学习行为被定义为五元组 $b_l = \langle l, o, t, r, \Omega \rangle$，其中，$l$ 表示学习者，o 表示在线操作，r 表示教学资源，t 表示学习行为发生的时间，Ω 表示其他各类属性的集合。在这 5 个元素中，t 元素保存了学习行为的时间点信息。基于在线学习行为 b_l，学生 l 在时间段 T 期间的对课程 c 的学习过程 $P_i = [b_{l,1}, b_{l,2}, \cdots, b_{l,n}]$ 中，通过 t 的有序序列 $[t_1, t_2, \cdots, t_n]$ 保存了整个学习过程的学习时序信息。

按照上述学习过程 P_l 的定义，学生对学习时间的安排信息已经记录在 P_l 的时间维

度 t 中，需要进一步将学习过程 P_l 的时间维度抽象时序特征。在已有研究中，通常会以特定长度的时间段为单位，将时间序列划分为多个阶段，通过分析每个阶段的部分数据实现对时间特征的提取，例如 Mahzoon 等[113] 分别以周和学期为单位，刻画学生在短期学习过程以及长期学习过程中的周期性规律。Park 等[147] 以天和周为单位分析学生的拖延学习状态。本文借鉴上述时间分段的思路，提出一种从学习过程中提取学习参与度的时序特征的方法，其过程如下。

通过对总时间长度 D 的学习过程 P_l 在时间维度 t 按单位时间段长度 d 切分，可以得到学生在每个时间片段的子学习过程 ΔP_t^l，定义如下：

$$\Delta P_t^l = \{(l, o_1, t_1, r_1, \Omega_1), (l, o_2, t_2, r_2, \Omega_2), \cdots, (l, o_j, t_j, r_j, \Omega_j)\}, t_0 \leq t_j < t_0 + d \quad (4\text{-}1)$$

式中：t_0 是时间段的起始时间，且对于子学习过程 ΔP_t^l 中的任意时间 t_j，都满足 $t_0 \leq t_j < t_0 + d$ 的约束条件。

子学习过程 ΔP_t^l 是学习过程 P_l 的连续子序列，两者的形式一致。因此，每个子学习过程 ΔP_t^l 均可以通过前一章提出的学习参与度指标计算方法，转换为该时间段的子学习参与度向量 $\Delta \boldsymbol{e}_t^l$，定义如下：

$$\Delta \boldsymbol{e}_t^l = [e_{t,1}^l, e_{t,2}^l, \cdots, e_{t,k}^l] \quad (4\text{-}2)$$

式中：k 是学习参与度指标的总数量，$e_{t,i}^l$ 是学生 l 在时间段 t 至 $t+d$ 期间第 i 个学习参与度指标的值，通过第 i 个学习参与度指标的计算函数 f_i 得到 $e_{t,i}^l = f_i(\Delta P_t^l)$。

基于每个子学习过程 ΔP_t^l 的时间段学习参与度 $\Delta \boldsymbol{e}_t^l$，学生 l 在整个学习过程 P_l 中的学习参与度时序信息可以抽象为一个 $m \times k$ 维的学习参与度时序矩阵 $\boldsymbol{E}_l$，

$$\boldsymbol{E}_l = \begin{bmatrix} e_{1,1}^l & e_{1,2}^l & \cdots & e_{1,k}^l \\ e_{2,1}^l & e_{2,2}^l & \cdots & e_{2,k}^l \\ \vdots & \vdots & \ddots & \vdots \\ e_{m,1}^l & e_{m,2}^l & \cdots & e_{m,k}^l \end{bmatrix} \quad (4\text{-}3)$$

式中，每行为一个子学习过程 ΔP_t^l 的时间段学习参与度，$\boldsymbol{E}_l$ 的行是按照时间段切分后每个子学习过程的时间先后顺序排列，总行数为 $m = \lceil \frac{D}{d} \rceil$。

4.3.2 学习参与度时序矩阵计算

通过对齐每个 t_0 并调整单位时间段长度 d 的大小，即可调整每个子学习过程 ΔP_t^l 的时间跨度和学习参与度时序矩阵 $\boldsymbol{E}_l$ 的总行数 m。例如，对于为期 4 周 $D = 4\text{weeks}$ 的一个在线课程，当 t_0 为整点且 $d = 1\text{hour}$ 时，表示按照 1 小时划分学习过程，每个子学习过程向量 $\Delta \boldsymbol{e}_t^l$ 描述 1 小时时间长度的学习过程特征，学习参与度时序矩阵 $\boldsymbol{E}_l$ 共有 $m = 24 \times 7 \times 4 = 672$ 行；当 t_0 为 0 点且 $\Delta d = 24\text{hours}$ 即 1 天时，每个子学习过程向量 $\Delta \boldsymbol{e}_t^l$ 描述 1 天时间长度的学习过程特征，学习参与度时序矩阵 $\boldsymbol{E}_l$ 有 $m = 7 \times 4 = 28$ 行；而当 t_0 为周一 0 点，$\Delta d = 168\text{hours}$ 即 1 周时，$\Delta \boldsymbol{e}_t^l$ 则描述一周时间长度的学习过

程特征，学习参与度时序矩阵 $\boldsymbol{E}_l$ 有 $m=4$ 行。

由上述示例可见，不同的单位时间段长度 d 可以划分不同粒度的子学习过程，进而产生行数量 m 不同的学习参与度时序矩阵 $\boldsymbol{E}_l$，描述不同时长尺度的学习特征时间特征。较小的 d 可能使每个子学习过程 ΔP_t^l 中的学习行为 b_l 的数量偏低甚至没有学习行为，即产生了空子学习过程 $|\Delta P_i|=0$，从而无法捕捉持续的学习行为和学习参与度信息。例如，学生在半夜至凌晨期间，可能没有没有任何学习行为，从而产生数个空子学习过程，使得捕捉的时序信息较为稀疏。然而，较大的 d 可能使每个子学习过程 ΔP_t^l 中包含过多的学习行为 b_l，造成每个时间段学习参与度 $\Delta \boldsymbol{e}_t^l$ 压缩了过多时间信息。例如，当 d 设置为 1 个月时，整个学习过程可能只包含几个子学习过程，进而每个子学习过程的子学习参与度 $\Delta \boldsymbol{e}_t^l$ 会损失 1 个月内细粒度的学习时间维度信息。因此，需要选择合适大小的 d 划分学习过程以保留合适尺度的时序特征。

为选择合适的单位时间段长度 d，本文利用前一章介绍的数据集，本文分析了选取不同 d 时，空子学习过程数量与非空子学习过程比例的变化情况。为了使 d 的现实含义易于理解，d 的取值以 1 小时为步进单位，而当 d 大于 24 后，以 24 小时为步进单位，直到 $d=168$ 即 1 周。结果如图 4-2所示：

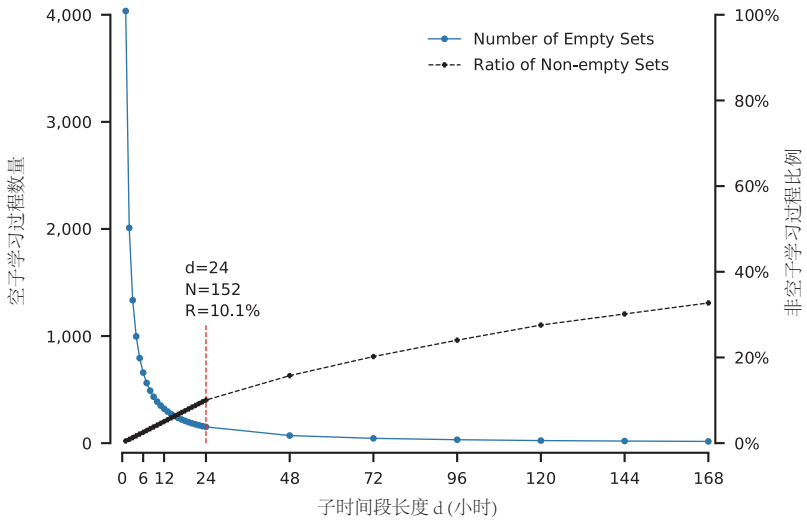


图 4-2 单位时间段长度 d 对子学习过程划分的影响

该数据集的学习过程总时长为 26 周，当单位时间段长度 $d<24$ 时，空子学习过程数量明显较高，随着 d 的增加，空子学习过程数量明显下降，而当 $d>24$ 后，空子学习过程数量虽然仍然保持下降趋势，但下降趋势明显放缓。另外，可以发现随着 d 的增加，非空子学习过程比例也显著增加。当 $d=24$ 时，非空子学习过程比例为 10.1%，即学生的平均学习天数约为 $26*7*0.1=18.2$ 天。而当 $d=168$ 时，非空子学习过程比例为 32.7%，说明在 26 周的学习过程中，平均有学习的周数约为 8 周。

综合上述结果可以发现，针对该数据集，选择以天数为单位的时间段长度 d 会产生相对较少的空子学习过程。但是当 $d>24$ 时，随着 d 的增加，非空子学习过程比例的增加趋势会放缓。并且由于单个子学习过程中学习行为数量的增加，会使得这些学

习行为的时间信息被压缩在子学习过程中从而造成时序信息损失。因此，相对较小的 d 能够保存较多的时序信息。另外，考虑到合适的 d 应当具有良好的可解释性并易于可视化，选择日常熟悉的时间段长度有助于用户理解时间特征，例如 Mahzoon 等[113] 提出基于以周为单位的分析模型分析一个学期的学习过程，Park 等[147] 以天为单位分析数周课程的学习参与度。因此，基于上述分析，本文采用 $d = 24\text{hours}$，即 1 天，作为时间段划分单位。

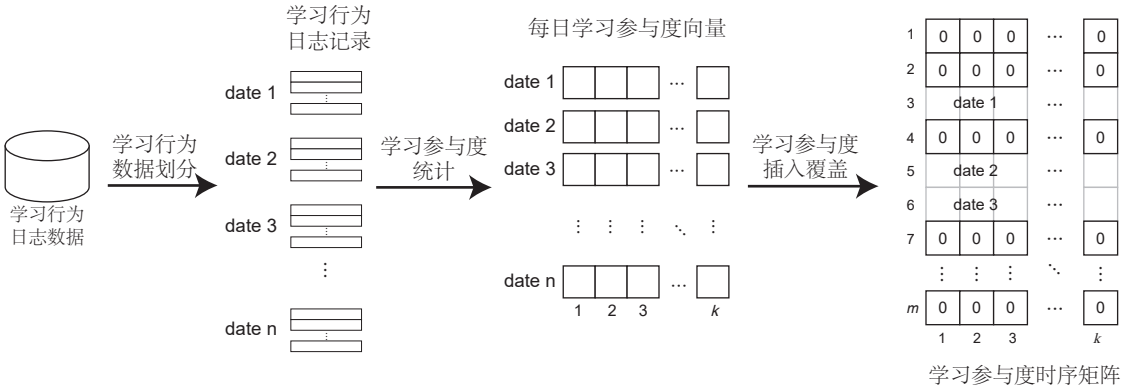


图 4-3 学习参与度时序矩阵的生成过程示意图

基于上述的学习参与度时序矩阵 $\boldsymbol{E}_l$ 的定义与单位时间段长度 $d = 24\text{hours}$，可以按图 4-3所示的步骤从在线学习行为数据中提取学习过程的时序特征。如图 4-3所示，首先，与前一章中处理学习行为日志数据不同，预处理后的学习行为数据不仅按照学生划分，还需要按照日期再进一步划分，如果学生在某个日期没有学习行为，则不划分该日期；然后，对每个学生按日期统计学习参与度指标统计，得到每天的子学习参与度向量；最后，先构建全空的学习参与度时序矩阵，再按照日期将每天的子学习参与度向量覆盖对应日期所在行，最终形成学生的学习参与度时序矩阵 $\boldsymbol{E}_l$。

通过学习参与度时序矩阵 $\boldsymbol{E}_l$，学生 l 在学习过程中的每一维学习参与度随时间变化的差异被保存在 m 维的列向量中。因此，通过学习参与度时序矩阵中刻画的学习过程的时序特征，能够反映学生在时间段 D 的学习时间管理风格。

4.3.3 基于卷积神经网络的成绩区间预测模型

为了充分利用学习参与度时序矩阵反映的学习时间管理特点，需要从学习参与度时序矩阵中提取时间维度信息，特别是短期的和长期的学习时间安排。通常针对时序数据的预测问题多采用基于整合移动平均自回归模型（Auto Regressive Integrated Moving Average，ARIMA）、隐马尔可夫模型（Hidden Markov Model，HMM）以及循环神经网络模型（Recurrent Neural Network，RNN）的方法，这些方法能够很好的实现对序列数据的回归预测分析。然而，在本研究中，教师和学生更加关心在长期学习过程后的考试成绩是否是不及格、及格、一般、良好或者优秀这五类之一，即可以将成绩区间预测

看作是一个分类问题。因此，本文借鉴 LeNet-5、WimNet[155] 等基于 CNN 的模型，基于 CNN 设计了一个成绩区间预测模型，利用 CNN 模型捕捉特征矩阵中局部特征的能力特点，提取学习参与度时序矩阵 $\boldsymbol{E}_l$ 中列向量包含的时间维度信息与行向量包含的学习参与度信息，预测学生最终的考试成绩区间。该模型如图 4-4所示。

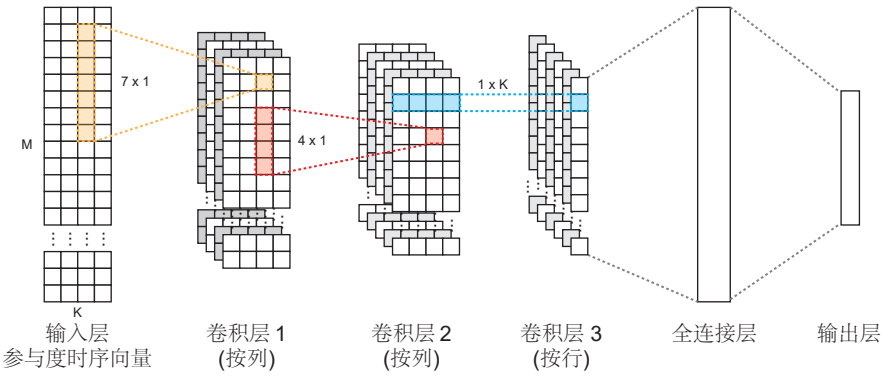


图 4-4 基于卷积神经网络的成绩预测模型结构

基于 CNN 的成绩预测模型的输入层为学习参与度时序矩阵，除此以外共包含 3 个卷积层、1 个全连接层与 1 个输出层。其中，第 1 个卷积层按列提时间维度的特征，卷积核大小为 $s \times 1$，提取 s 天的短期学习时间特征（例如，$s = 7$ 表示 1 周时间内的短期特征）；第 2 个卷积层以第 1 个卷积层的输出作为输入，同样按列提取时间维度的特征，卷积核大小维 $q \times 1$，提取 q 个短周期跨度的长期时间特征（例如，$q = 4$ 表示 4 个前一层短周期的长期特征，在 $s = 7$ 时即为 4 周）；第 3 个卷积层以第 2 个卷积层的输出作为输入，按行提取学习参与度的特征，卷积核大小为 $1 \times k$。同时，每个卷积层均采用 ReLU（Rectified Linear Unit）作为激活函数。利用 CNN 的平移不变性特点，前两个卷积层能够捕获整个学习过程中所有时间段的学习参与度随时间的变化，进而刻画学习过程的时序特征。

4.4 可视化设计与交互界面

为了分析学生在长期学习过程中学习时间的安排（T.1），需要将学生的学习参与度时序矩阵直观的呈现出来。而要分析不同学生群体的学习参与度与时间管理风格的差异（T.2，T.3），则需要为用户提供对比多组参与度时序向量的对比视图。因此，本文提出一种基于日历坐标系的学习参与度日历矩阵表示方法将转换学习参与度时序矩阵的表现形式，并基于此方法设计了基于日历图的时间管理可视化方法和多个关联视图。同时，在前一章学生群组创建工具的基础上，进一步设计了交互式的学生群组创建工具，帮助用户从不同层面探索学生的在线学习过程的时间管理风格。

4.4.1 基于日历坐标系的学习参与度日历矩阵

虽然学习参与度时序矩阵 $\boldsymbol{E}_l$ 通过 m 维列向量表达了在线学习过程的时序特征，但该表达方法仅适合作为计算模型的输入，而不适合直接可视化呈现。一方面，直接采用

常见的表格、热力图、像素图等可视化方法虽然能够显示全部数据，但由于学习参与度时序矩阵的两个维度都较高，过于密集的数据点不容易展示学习参与度随学习时间变化的周期性特征。另一方面，单独呈现一个子学习过程 ΔP_t^l 的时间段学习参与度 $\Delta \boldsymbol{e}_t^l$ 时可以使用星型图、平行坐标系或者前一章的多盒图方法，但叠加 m 维的时间维度后易造成图形符号重叠的问题，难以观察整体的变化趋势。

因此，为了让用户直观理解学生学习过程的时间安排，以及随时间变化的学习参与度，本文提出一种基于日历坐标系的学习参与度日历矩阵表示方法用以辅助学习参与度时序矩阵 $\boldsymbol{E}_l$ 的可视化，该过程如图 4-5所示。

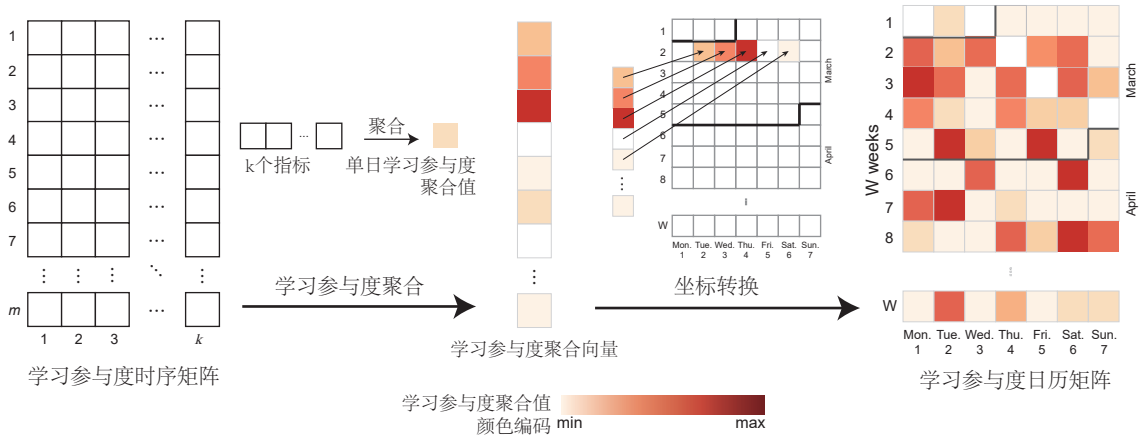


图 4-5 学习参与度时序矩阵转换为学习参与度日历矩阵的过程示意图

首先，针对学习参与度高维度的特点，前一章使用了 t-SNE 数据降维方法将高维度学习参与度降维至二维，但采用这个方法处理数据后会损失特征的物理意义，使降维后特征的数值大小与实际学习参与度的大小无关。然而，在分析学习过程时需要直观理解特征的含义，因此，为了体现学习参与度的变化情况（R.1），本节采用基于线性加权的方法处理学习参与度特征向量。便于学习参与度的视觉编码映射，促进用户对参与度整体变化情况的理解，我们按照以下公式将学习参与度时序矩阵 $\boldsymbol{E}_l$ 中每个 $1 \times k$ 维的子学习参与度向量 $\Delta \boldsymbol{e}_t^l$ 聚合为反映学习参与度程度大小的单一数值 v_t^l，定义如下：

$$v_t^l = \alpha_1 g_1(e_{t,1}^l) + \alpha_2 g_2(e_{t,2}^l) + \cdots + \alpha_k g_k(e_{t,k}^l), \sum_{i=1}^{k} \alpha_i = 1 \tag{4-4}$$

其中，$g_i()$ 代表各指标的归一化函数，由于各指标的数值范围差异较大，需要在聚合前先进行归一化处理。α_i 代表各指标的聚合系数，根据文献 [130] 介绍的学习参与度聚合方法，按照各类参与度在学习过程中的利用率预先设定。由于通过 $g_i()$ 归一化和聚合系数 α_i 的作用，学习参与度聚合值 $0 \le v_t^l \le 1$，并且 v_t^l 越大表示学习参与度越高。另外，$v_t^l = 0$ 时表示当天没有任何学习相关行为，而 $v_t^l = 1$ 时则表示当天的参与度已经达到或超过了领域专家预先设定的阈值。尽管聚合后损失了学习参与度在学习行为类型、资源类型方面的信息，但是由于线性方法产生的聚合值 v_t^l 具有可比性，通过聚合值的大小可以在一定程度上反映学习参与度的大小，从而为对比任意子学习参与度向

量 $\Delta \boldsymbol{e}_t^l$ 的变化提供了支持。

在上式基础上，可以进一步对每一天的子学习参与度向量进行聚合，构造学习参与度聚合向量 $\boldsymbol{v}^l$：

$$\begin{aligned}\boldsymbol{v}^l &= \begin{bmatrix} g_1(e_{1,1}^l) & g_2(e_{1,2}^l) & \cdots & g_k(e_{1,k}^l) \\ g_1(e_{2,1}^l) & g_2(e_{2,2}^l) & \cdots & g_k(e_{2,k}^l) \\ \vdots & \vdots & \ddots & \vdots \\ g_1(e_{m,1}^l) & g_2(e_{m,2}^l) & \cdots & g_k(e_{m,k}^l) \end{bmatrix} \times [\alpha_1, \alpha_2, \cdots, \alpha_k] \\ &= \left[v_1^l, v_2^l, \cdots, v_m^l\right]^T \end{aligned} \tag{4-5}$$

式中，利用聚合系数向量 $[\alpha_1, \alpha_2, \cdots, \alpha_k]$，将 $m \times k$ 维的参与度时序向量 $\boldsymbol{E}_l$ 压缩为 $m \times 1$ 维的学习参与度聚合向量 $\boldsymbol{v}^l$。

然后，由于稍后讨论的可视化方法采用了基于日历图的展示形式，因此需要将学习参与度聚合向量 $\boldsymbol{v}^l$ 转换为日历坐标系的学习参与度日历矩阵。首先，构造代表时间范围 D 的全部日期的 $w \times 7$ 维的零矩阵，随后将 $m \times 1$ 维的学习参与度聚合向量 $\boldsymbol{v}^l$ 按对应日期的周数与星期数逐项转换坐标至学习参与度日历矩阵 $\boldsymbol{A}_l$ 。在转换过程中，将 $\boldsymbol{v}^l$ 每个元素的下标 i 和对应的日期（格式为 $y-m-d$，分别对应年、月、日）按以下公式计算得到学习参与度日历矩阵 $\boldsymbol{A}_l$ 的行号和列号：

$$row = \lfloor \frac{i}{7} \rfloor$$

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

式中，row 为日期转换后在学习参与度日历矩阵中的行号，col 为列号。列号 col 的计算采用文献 [156] 中提出的方法。

转换完成后，得到构建 $w \times 7$ 维的学习参与度日历矩阵 $\boldsymbol{A}_l$，形式如下所示：

$$\boldsymbol{A}_l = \begin{bmatrix} v_{1,1}^l & v_{1,2}^l & v_{1,3}^l & v_{1,4}^l & v_{1,5}^l & v_{1,6}^l & v_{1,7}^l \\ v_{2,1}^l & v_{2,2}^l & v_{2,3}^l & v_{2,4}^l & v_{2,5}^l & v_{2,6}^l & v_{2,7}^l \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{w,1}^l & v_{w,2}^l & v_{w,3}^l & v_{w,4}^l & v_{w,5}^l & v_{w,6}^l & v_{w,7}^l \end{bmatrix} \tag{4-6}$$

式中，$\boldsymbol{A}_l$ 是日历坐标系的矩阵，其中每行代表每一周，每列自左向右分别为周一到周日，自上而下代表从时间段 D 的第一周到最后一周，w 为总周数。例如，对于一个学期 $w=26$，一个学年 $w=53$。由于构建 $\boldsymbol{A}_l$ 时，各元素初值设置为 0，表示当天没有任何学习行为。

4.4.2 基于日历图的学习时间分布图

日历图是展示时序数据的一种常用方式，通过将数据按照日期属性映射在日历上，帮助用户理解数据的时间趋势与周期性[49]。利用前一节提出的学习参与度日历矩阵，本

节提出一种基于日历图的学习时间分布图，将在线学习过程中的时间信息映射为日历坐标，并将每天的学习参与度映射为颜色编码，从而通过日历图的形式呈现学生在一段时间内的学习时间安排以及学习参与度随时间的变化情况，直观的理解和分析学生的时间管理风格（R.1）。学习时间分布图的设计如图 4-6所示。

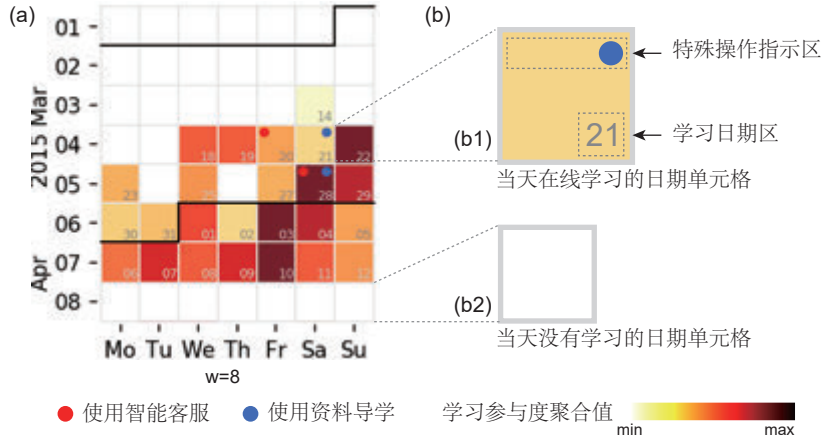


图 4-6 基于日历图的学习时间分布图，其中 (a) 为一个 8 周学习过程的学习时间分布图，(b) 为两个典型日期单元格的放大，(b1) 为当天在线学习的单元格，(b2) 为当天没有学习的单元格

图 4-6(a) 展示了一个 $w = 8$ 的典型学习事件分布图。该设计基于日历图的日期单元格展示学习参与度，每行从周一到周日共 7 天，纵向逐行展示学习过程，并用边界线划分不同的月份。如图 4-6(a) 所示，每个日期单元格按照对应日期是否在线学习而展示不同信息。对于日历时间范围内的任意一天 t，如果学生当天有在线学习，即学习参与度聚合值 $v_t^l > 0$，则按图 4-6(b1) 方式显示单元格，将学习参与度聚合值 v_t^l 按照图例所示的颜色编码映射为单元格填充颜色（颜色越深代表学习参与度聚合值越高），将当天的日信息标记在学习日期区。另外，如果当天有用户关注的特定操作，则在特殊操作指示区用对应符号标记。如果学生没有在线学习，即学习参与度聚合值 $v_t^l = 0$，则单元格填充颜色为白色（图 4-6(b2)）。例如，如图 4-6(a) 所示，该学生在第 4 周 3 月 21 日有一定的学习参与度并使用了定期导学服务，而在第 5 周的周 2 和周 4 没有学习，随后连续学习了 17 天直到 4 月 12 日。在这个 8 周的时间段中，该学生在 3 月 22 日周日、3 月 28 日周六，4 月 3 日周五、以及 4 月 10 日周五学习参与度明显较高，表明该学生在周末可能会更加投入学习。从上述示例可以看出，利用学习时间分布图可以发现学生在一个时间段内的学习时间安排，以及在不同时间点的学习参与度程度，从而了解学生的学习时间管理风格。

4.4.3 多课程视频观看可视化

上述的学习时间分布图展示了学生对一门课程的学习时间安排的时序信息，然而在网络课程的学习过程中，学生可能同期学习多门课程，他们需要在有限的时间段内安排多门课程的学习时间（T.2）。而网络课程在线学习的多种教学环节与学习行为中，观看课程视频是最主要的学习形式并且也是耗费时间最多的，因此，本文选择学生观

看课程视频的时间安排作为研究对象展开设计分析。为了展示学生在一段时间内多个课程的观看视频的时间安排（R.2），本节设计了三个关联视图从三个角度展示学生学习多个课程时观看视频的时间安排情况和观看进度，包括多课程学习进度图、观看时间分布图、以及观看量统计图。

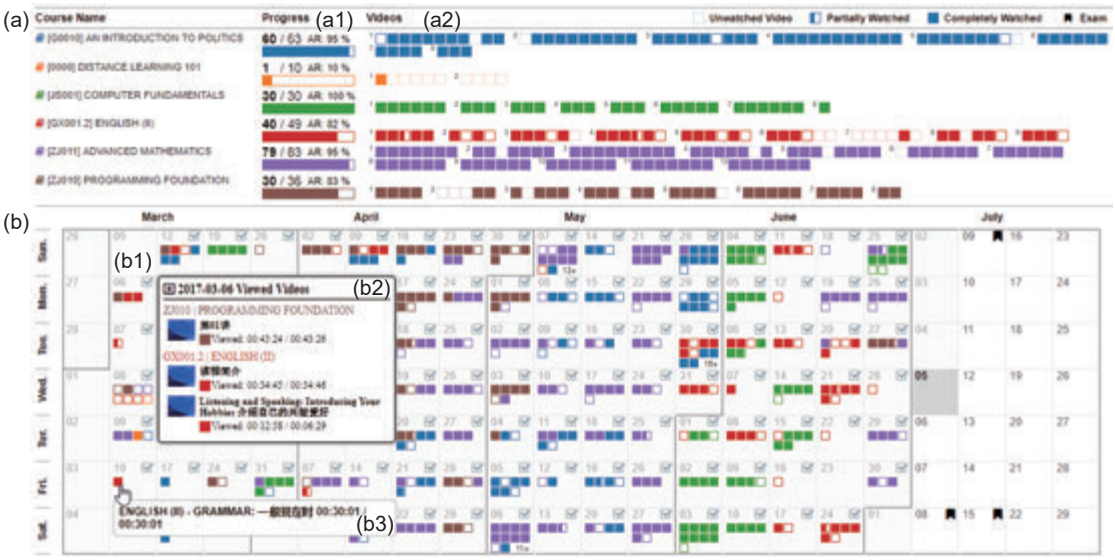


图 4-7 一个典型学生的多个课程的课程进度图与观看时间分布图，其中 (a) 为课程进度图，(a1) 和 (a2) 分别表示课程总观看进度与课程内各视频观看进度，(b) 为观看时间分布图，(b1) 代表 3 月 6 日的观看情况，(b2) 为 3 月 6 日的观看详情，(b3) 为单个视频的观看详情

课程进度图。如图 4-7(a) 所示，课程进度图以表格为基本形式将学生在特定时间段内的所有课程以及学习进度逐行展示。其中，每门课程映射为颜色编码，显示已经观看课程视频数量与总课程视频数量，并将两者的比值映射为矩形符号（图 4-7(a1)）。同时，课程中每一个视频的观看进度也会以视频单元格符号展示，所有视频单元格按照章节顺序从左到右依次罗列，并按照章顺序分隔（图 4-7(a2)）。其中，尚未观看的视频映射为点状线边框空白单元格，已观看的视频映射为实线边框单元格，并将将已观看时长占视频时长的比例映射为单元格内矩形的宽度比例。通过观察课程进度图中各种单元格的分布，可以直观的理解学生各个课程当前的学习进度。

观看时间分布图。为了进一步展示学生如何安排各个课程的视频观看时间，本文设计了基于日历图的观看时间分布图（图 4-7(b)）。与前一节学习时间分布图不同，该图为了充分利用宽屏幕的水平空间，将日历设计为横向分布，逐列展示学习过程的每一周。其中，已观看的视频映射为视频单元格并按照观看日期逐个排列在对应的日期单元格中，每个视频单元格对应的视觉编码规则与课程进度图一致。例如，如图 4-7(b1) 所示，学生在 3 月 6 日完整观看了 2 个课程的 3 个视频。点击该日期单元格，会进一步显示该日期观看每个视频的详情（图 4-7(b2)）。另外，当选择单个视频时也会显示观看详情（图 4-7(b3)）。利用观看时间分布图中呈现的每日观看视频的数量与课程类型，可以直观的了解学生在一段时间内如何安排各课程的视频观看。

(a) 每周视频观看量堆叠柱状图

(b) 每周视频观看量折线图

图 4-8 每周视频观看量统计图

每周视频观看量统计图。为了进一步理解学生观看各课程视频随时间的变化情况，本文设计了与观看时间分布图关联的每周观看量统计图（图 4-8），包括堆叠柱状图与折线图两种表现形式，并可以随时切换。该图两种形式的纵坐标均为视频观看数量，横坐标均为周数，并且横坐标的宽度直接对应观看时间分布图中每一列的宽度，有助于理解统计值与对应周的观看活动详情。堆叠柱状图展示了每周视频观看总数量的课程构成，折线图展示了单独每个课程的每周视频观看量，并且两种形式的课程颜色编码与课程进度图一致。通过观察该统计图，可以进一步直观理解学生对每个课程的时间安排情况。例如，如图 4-8(a) 所示，该学生在学期初每周学习 4-5 门课程，而在学期中后期每周只学习 2、3 门课程，在第 13 周只学习了一门课程。而图 4-8(b) 显示，绿色折线显示该学生学期初和学期末观看计算机基础的课程视频，褐色折线显示学生主要前期和中期观看程序设计基础的课程视频。

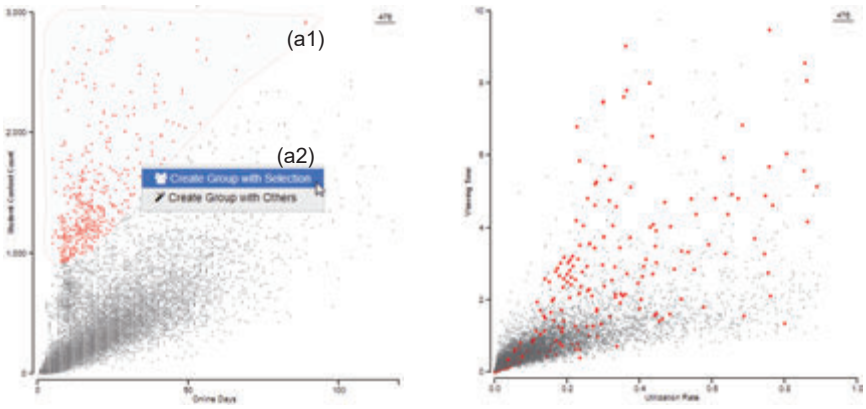
利用上述的课程进度图、观看事件分布图以及每周视频观看量统计图这三个相互关联的可视化设计，不仅可以帮助教师详细的了解每个学生对多门课程的时间安排，也可以帮助学生了解自身学习进度和为管理学习计划提供支持。

4.4.4 交互界面设计

由于学生在线学习过程的长期性与时间管理的复杂性，在上述可视化设计的基础上，需要从学生中选择部分具有特定学习参与度模式的学生构建学生群组，并对这些学生群组或者个体之间的时间管理风格做对比分析（T.3）。因此，为了支持交互的探索学生群组的时间管理风格（R.4），本节在前一章介绍的学生群组创建工具基础上，进一步设计了基于交互选择的和基于集合运算的学生群组创建工具，并构建了学习时间分布图对比工具（R.3）。

1) 学生群组创建工具

利用前一章设计的 2 种学生群组创建工具，能够按照学生的属性值或者分布区域精确的创建不同属性群组，然而在探索性数据分析时，需要更多交互式的手段创建学生群组（R.4）。因此，针对该问题，本节设计了以下两种学生群组创建工具。



(a) 鼠标圈选学生数据点　　(b) 圈选数据点在其他横纵坐标下的分布

图 4-9 基于交互选择的学生群组创建工具

基于交互选择的学生群组创建工具。在前一章设计的学生参与度分布图中，学生数据点的分布形状除了密集区域以外，还呈现出分布稀疏不规则的特点，难以通过精确的数值规则或者自动识别方式创建群组。针对这一问题，本节设计了一个交互式区域选择器（图 4-9(a)），可以使用鼠标圈选任意封闭形状区域以内的数据点从而构建学生群组。例如，在图 4-9(a1) 中，使用鼠标圈选一片区域，区域边界和区域内选中的数据点将被高亮为红色以识别区域范围（图 4-9(a1)），然后用户可以将区域以内或以外的数据点构建为一个学生群组（图 4-9(a2)）。同时，通过切换横纵坐标轴，可以查看选中的数据点在其他学习参与度维度下的分布情况（图 4-9(b)）。

(a) 并集运算符　　(b) 交集运算符　　(c) 差集运算符

图 4-10 基于集合运算的学生群组创建工具

基于集合运算的学生群组创建工具。由人工规则或者圈选方式创建的多个学生群组之间可能存在重叠或者互补，而重新录入规则或者重新圈选难以准确提取重叠部分或合并群组。针对这一问题，如图 4-10所示，本节设计了基于集合运算的学生群组创建

工具，利用并集、交集和差集这 3 种集合运算符，在已经创建的群组基础上通过集合运算生成新的学生群组。首先，用户在学生群组列表中选中需要操作的学生群组；然后，选择并集、交集或差集运算符的其中一个，系统会根据操作符生成相应的候选结果集合，例如差集运算符会产生两个待选结果集合；确认后即产生新的学生群组。

2) 学习时间分布图对比工具

不同的学习参与度模式的学生群体在学习时间管理风格上可能存在差异，为了直观对比这种学生群组间的时间管理风格差异（R.3），本节设计了学习时间分布图对比工具。首先，如图 4-11(a) 所示，在群组列表中选中需要对比的学生群组。然后，如图 4-11(b) 所示，被选中的各个学生群组逐列显示在对比视图中，每列均包含该群组的所有学生的学习分布图，并且每列学生群组的颜色编码与其他视图保持一致。同时，每个学生的课程成绩预测值也会展示在学习时间分布图下方。

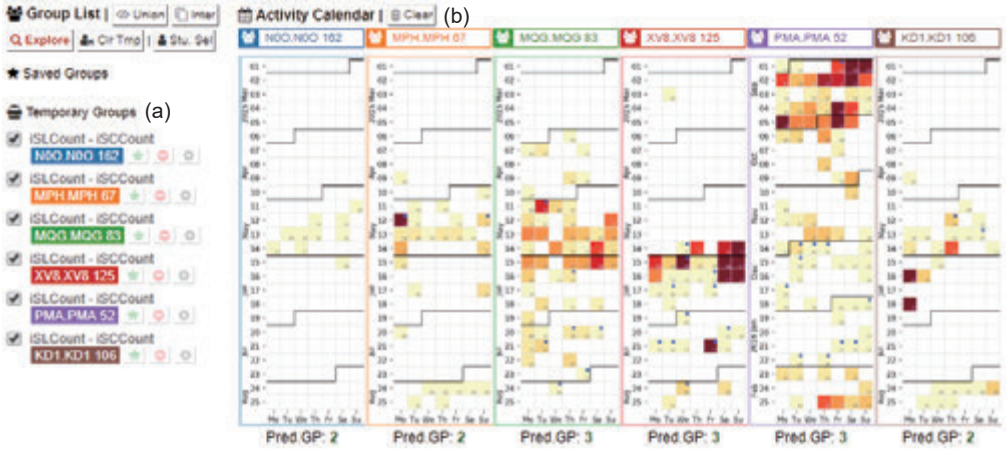


图 4-11 学习时间分布图对比工具，其中 (a) 为选中的待对比学生群组并按颜色编码，(b) 为各学生群组每个学生的学习时间分布图以及成绩区间预测结果

4.5 案例分析

为了验证本章提出的可视化设计的可用性，本文作者邀请本校网络远程教育学院的两名授课教师作为领域专家开展案例分析。两名领域专家均熟悉网络远程教育的学习过程，他们一致认为分析学生如何管理在线学习时间对于提高教学质量和改进学习支持服务有重要作用。然而，LMS 现有的统计功能仅提供了基本的学习时长、天数等宏观统计值，难以理解学生在长期学习过程中具体的学习时间安排。因此，他们目前只能根据课堂教学经验，推测学生的学习时间管理特点并提供的基本的建议。

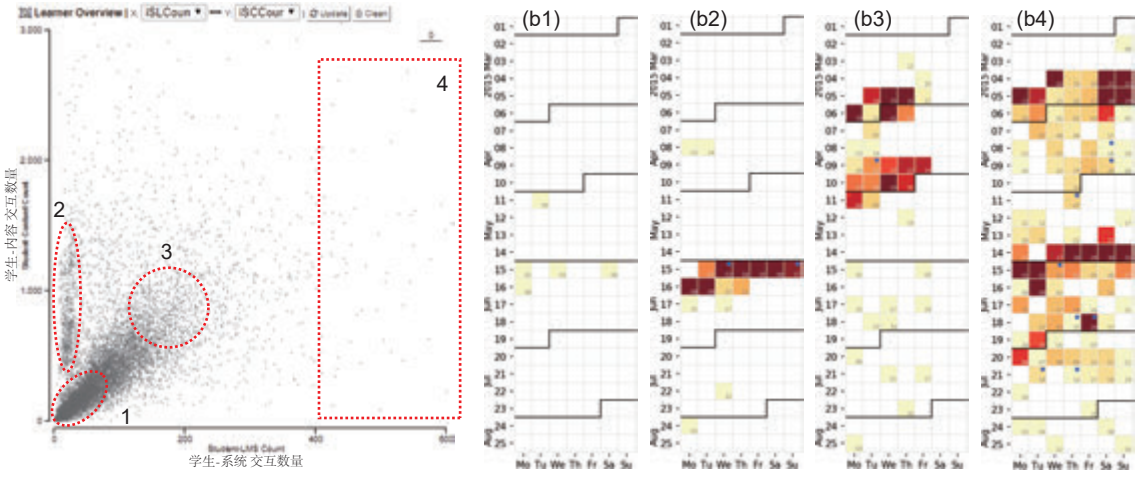
在开展案例分析前，本文作者首先向领域专家介绍了 LearnerExp 可视分析系统的设计和主要功能，然后，由领域专家探索真实的在线学习行为数据集。在此期间，本文作者会回答他们的问题并观察操作行为。最后，访谈他们对系统的使用体验并记录反馈意见。总体而言，两位领域专家都认为他们能够比较容易的理解系统的可视化界面

和交互操作，并使用这个系统识别不同学生的时间管理风格。此外，他们也发现了一些典型的时间管理风格模式，并建议将这些发现集成在 LMS 中，为开展教学干预、提供学习支持服务以及改进课程设计提供支持。

本节案例分析中利用了本校网络教育学院的真实在线学习行为数据，运用 4.3 节提出的学习参与度时序模型刻画在线学习行为数据中的时序特征，并通过 4.4 节提出的多种可视化方法展示学生在长期在线学习过程中的时间管理风格。本节分析的数据可以分为两个数据集，这两个数据集均反映了网络远程教育标准的在线学习过程。网络远程教育的在线学习参照全日制专业培养计划，每个专业的培养计划包括 4-6 个学期，每个学期通常包括 25 周学习 5-7 门课程。学生需要在每个学期完成课程学习、课程训练、测验与作业、并通过期末的课程考试获得学分。在学期开始后学生可以访问所有的教学资源，包括课件视频、讲义、阅读材料以及往年的课程讨论等。学生需要自己根据课程知识掌握情况制定学习计划，并根据工作生活情况安排时间学习。这两个数据集的原始数据记录和分析结果分别保存在 HBase 集群和 MongoDB 数据库中，基于分布式内存计算框架 Spark 实现统计和分析。领域专家使用本文设计的 LearnerExp 可视分析系统对这两个数据集展开分析，探索学生的学习时间安排以及多课程的学习时间管理风格。

4.5.1 学习时间管理风格对比

该案例分析使用的数据集包含 2015 年入学的学生在第一学期（2015 年 3 月至 7 月）的学习日志数据和相关其他类型数据。经过数据清洗后，共包含学生 10,529 名，日志记录 14,942,964 条。本文选择所有学生必修的英语公共基础课的在线学习行为数据作为分析对象。



(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

图 4-12 学习参与度分布模式与时间管理风格

如图 4-12(a) 所示，在前一章提出的学习参与度分布图上，以“学生-系统交互”与“学生-内容交互”这两个维度作为横坐标和纵坐标，学习参与度分布图中呈现出两个明

显的密集区域，即区域 1 和区域 2。另外根据教学经验，领域专家进一步圈选了区域 3 和区域 4。这 4 个学习参与度分布区域反映不同的学习参与度模式，其中，区域 1 的学生两项指标都较低，说明这些学生很少使用学习管理系统，并且也很少访问教学资源；区域 2 的学生在系统交互数量与区域 1 相似的情况下，教学资源交互数量明显较高，说明这些学生可能在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 的学生也分布相对较密集，并且其内容交互数量与区域 2 近似，同时系统交互数量高于区域 1 和区域 2，说明这些学生使用学习管理系统的次和访问教学资源的次数都较多。区域 4 的学生数量较少而且内容交互数量的分布范围广，但系统交互数量最高，说明这些学生频繁的使用学习管理系统，并且部分学生对学习资源的访问数量也很高。

进一步使用学习时间分布图查看这 4 个区域学生的学习时间安排，可以发现每个区域的学习时间安排各有特点，而在一个区域内的部分学生呈现出相似的时间管理风格。如图 4-12(b1) 所示，区域 1 学生的学习参与度低学习时间少，并且没有明显的规律性；区域 2 学生的学习参与度和学习时间明显高于区域 1，并且呈现在前期很少学习而学期中后期、期末考试前的 1-2 周连续密集学习的特点，反映出拖延学习的特点。图 4-12(b2) 展示了区域 2 学生的典型学习时间安排，该学生在第 15 周和第 16 周连续的高参与度学习，而在平时很少学习；区域 3 学生的学习参与度和学习时间在学期前期较高，但随后在学期中后期显著减少，呈现出松懈疲劳的特点。如图 4-12(b3) 所示，该学生在学期前期的第 5 周、第 6 周、第 9 周和第 10 周学习的天数较多，而随着时间推移，学期中后期学习天数明显减少；区域 4 学生的学习参与学习时间最高，并且在整个学期始终持续性的学习，呈现出自律学习的特点。如图 4-12(b4) 所示，该学生除了第 10 周和第 11 周的学习参与度较低以外，在整个学期的学习天数和学习参与度都较高。

从这些发现可以看出，学生的学习参与度模式与学习时间管理风格之间存在一定程度的相关性，并且通过本文提出的学习时间分布图可以有效的呈现不同的学习时间管理风格。领域专家认为上述可视化结果与他们在日常授课中观察到的学生长期学习行为一致，即学生的学习行为在整个学期中会呈现多种时间管理的风格，包括图 4-12所示的拖延、松懈以及自律等特点，并且图 4-12(a) 中呈现的这些学习风格的学生数量分布也符合实际教学过程中的感受。

4.5.2 多课程学习时间管理风格

该案例分析使用的数据集为 2017 年入学的计算机科学与技术专业学生在第一学期（2017 年 3 月至 7 月）的学习日志数据与学生的学籍管理数据。经数据清洗和筛选后，包含学生 164 名，日记记录 40,928 条。由于在网络远程教育的学习过程中，学生每个学期需要学习多门课程，而如何安排协调多门课程的学习时间可能会影响他们每个课程的学习参与度以及学习成效。利用本文设计的课程进度图、多课程观看时间分布图以及每周视频观看量统计图，可以直观的呈现长期学习过程中对多门课程的视频观看时间安排。领域专家使用 LearnerExp 可视分析系统对数据集展开探索分析后发现，与前一案例分析类似，大部分学生在所有课程的学习参与度都很低，没有呈现出明显的

学习管理风格。因此，领域专家进一步对学习天数高于第三四分位数的 41 名学生进行仔细观察，可以识别以下 3 种学习时间管理风格。

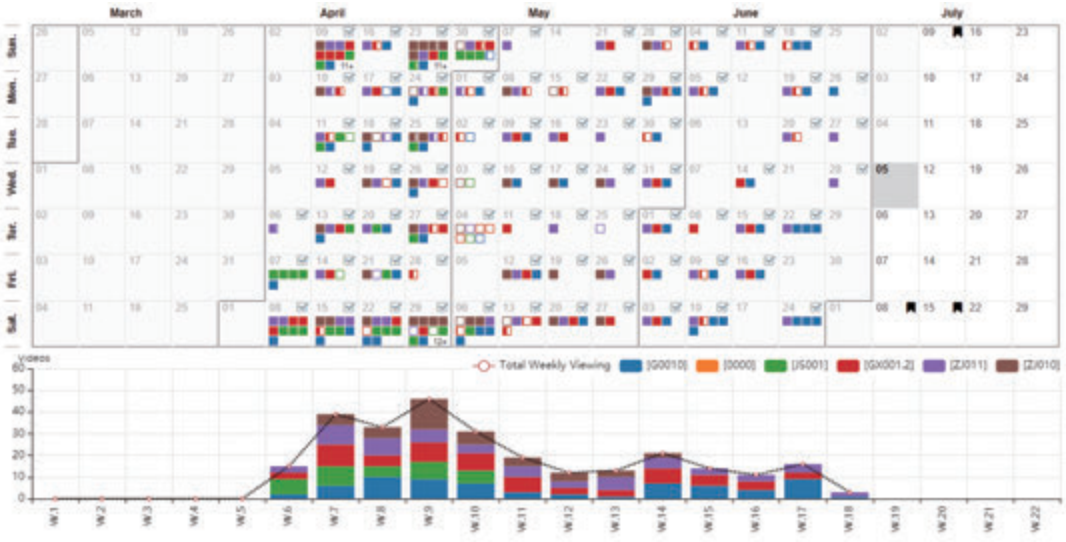


图 4-13 多课程同时规律学习的典型学习时间安排

首先，多课程同时规律学习。该风格的学习者数量较少，领域专家共标记了 9 人为此风格的学习者（5.5%）。图 4-13展示了该风格的一个典型实例，该学生自学期初开始，在整个学期保持每周学习 4-5 门课程，并且每天完整观看 2-3 门课程的少量视频，周六和周日可能会观看更多。随着时间推移，学期中后期每周仍学习多门课程，但视频观看量逐渐减少。这种时间安排说明该学生能够比较有效的并行管理自己多个课程的学习时间和学习进度，并且表现出显著的自律学习特点。

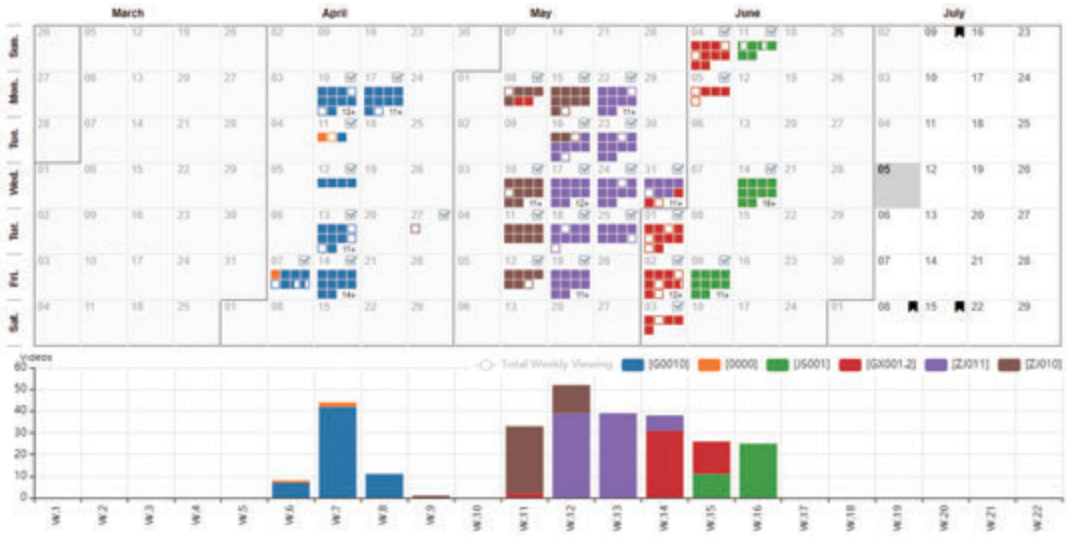


图 4-14 单课程逐个规律学习的典型学习时间安排

其次，单课程逐个规律学习。领域专家共标记了 11 人为此风格的学习者（6.7%）。如图 4-14所示的典型实例，该学生除了 4 月底至 5 月初一段时间以外，均能够保持比

较规律的每周学习。但与图 4-13不同，该学生平均每周只学习 1-2 门课程，周内每天基本只观看 1 门课程的大量个视频，并且周六和周日基本不学习。通过连续数日的完整视频观看学完一门课程之后，该生再开始下一门课程的学习。虽然按照这种时间管理风格取得的课程学习参与度与前一种类似，但过程差异明显，说明该学生在表现出显著的自律学习特点同时，更偏向于串行的管理自己多个课程的学习事件，逐个课程的完成学习任务。

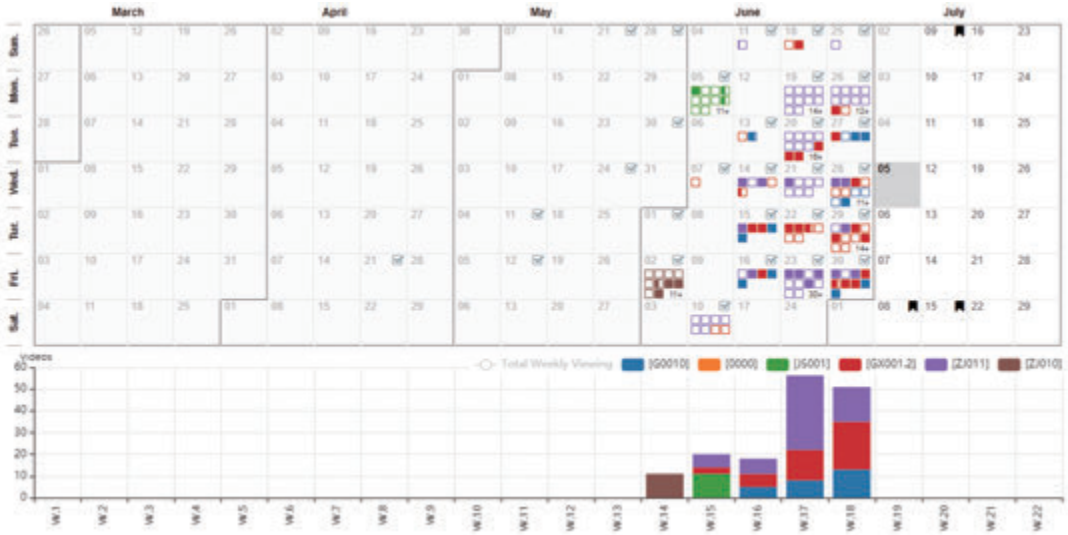


图 4-15 多课程混合拖延学习的典型学习时间安排

最后，多课程混合拖延学习。领域专家将其他 21 人标记为此风格的学习者（12.8%）如图 4-15所示，该学生对各课程的学习时间安排呈现了与前两种时间管理风格明显不同的特点。在总学习周数与天数方面，该学生的学习时间明显较少，且在学期的前期与中期基本没有在线观看视频，只从学期末开始连续数周突击学习。而在学习参与度方面，可以发现该学生每天虽然观看了大量课程视频，但大多没有完整观看，实际观看时长明显较低。在每天的课程学习安排方面，可以发现该学生没有明显特点的观看习惯，有时候只观看同一个课程的视频，有时候交错观看不同课程的视频。这种时间管理风格表现出了明显的拖延学习特点，并且也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。另外，由于这种时间管理风格特点使学习时间紧张，可能无法完整的使用课程资源，造成学习参与度较低。

领域专家认为上述可视化效果反映的多课程学习时间管理风格能够反映真实学习行为在时间方面的特点，例如，根据他们的教学经验，大部分参与网络远程教育的学生由于缺少在线学习课程以及独立管理多课程学习进度的经验，而在网络远程教育中由于课程教学资源数量较多且整个学期时间较长，因此容易产生拖延、松懈以及学习参与度低的现象。而网络远程教育的学生通常已经参加工作，部分学生可能已经在工作过程中培养了自我安排进度的能力，因此也体现在学习过程中能够规律的安排短期和长期的学习时间，呈现出自律的学习时间管理风格。通过总结上述多课程时间管理风格可以看出，本文提出的多课程学习时间可视化设计能够有效的呈现学生在面对多个

课程的学习压力时采取的间安排，为分析不同的学习时间管理风格提供支持。

4.6 本章小结

本章提出了一种基于日历坐标系的学习参与度时序特征可视分析方法，并基于该方法设计了 LearnerExp 可视分析系统，探索长期学习过程中学生的时间管理风格。

首先，为了刻画在线学习行为数据中反映的时序特征，本文将学习过程按照时间间隔长度划分为多个子学习过程，通过按时间顺序堆叠每个子学习过程的学习参与度向量构建学习参与度时序矩阵。然后，为了充分利用学习参与度时序矩阵中的时间维度信息，本文设计了一个基于卷积神经网络的成绩预测模型，捕捉学习参与度时序矩阵局部列向量中包含的时间维度信息与行向量中的学习参与度信息，预测学生最终的考试成绩。随后，为了直观展示从在线学习行为数据中提取的时序特征，本文将学习参与度时序矩阵映射到基于日历坐标系的学习参与度日历矩阵中，并进一步将该矩阵展示为基于日历图的学习时间分布图，帮助用户理解在线学习行为数据中反映的时序特征。同时，本文针对在线多课程同期学习的特点，设计了多课程学习进度图，支持从课程视频资源和时间两个方面细粒度分析学习过程中学生观看课程的时间安排特征。最后，为了验证数据模型与分析方法的有效性，本文基于上述可视化设计开发了 LearnerExp 可视分析系统对真实的在线学习日志数据进行分析，发现了多种学习时间管理风格及其与学习参与度模式之间的关系，并进一步识别和讨论了同期学习多门课程时的多种时间管理风格。

以本章研究内容为基础撰写的论文发表在国际会议 ACM Global Computing Education Conference (CompEd 2019)，题为“VUC: Visualizing Daily Video Utilization to Promote Student Engagement in Online Distance Education”，以及国际会议 IEEE Conference on Visual Analytics Science and Technology Short Paper Track(VAST 2019)，题为“Visual Analysis of the Time Management of Learning Multiple Courses in Online Learning Environment”；以本章研究内容为基础开发的可视分析系统，发表在国际会议 The Web Conference 2019 Demonstration Track (WWW 2019)，题为“LearnerExp: Exploring and Explaining the Time Management of Online Learning Activity”。

5 面向大规模多属性数据的视频利用情况可视分析

5.1 引言

教学视频是目前网络课程在线学习中最重要，也是最为广泛采用的教学媒介[121]。如图 5-1所示，包括大规模网络公开课程（MOOC）、开放课程课件（Open CourseWare, OCW）以及现代网络远程教育等在内的众多在线学习平台，均以不同形式与内容的教学视频为主要载体，实现课程知识的指导以及专业技能的传授，从而为全世界学习者带来丰富多样的学习体验。

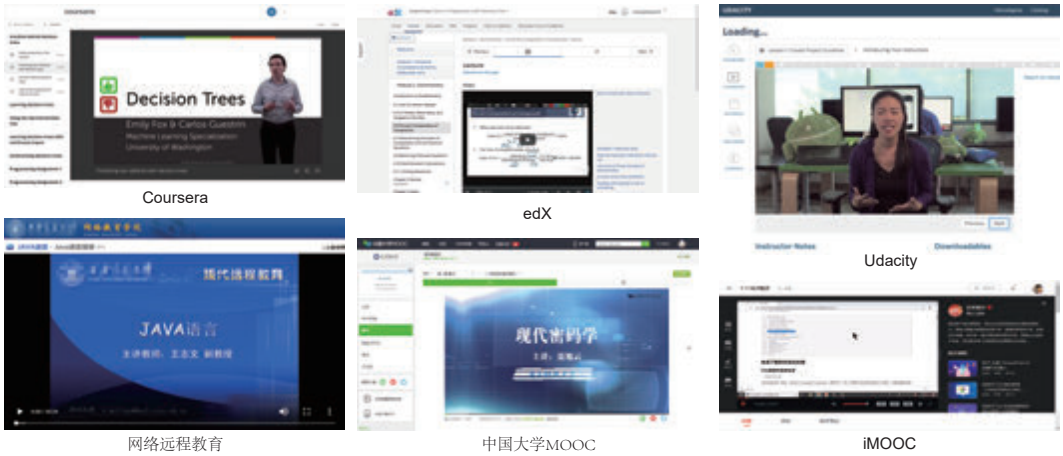


图 5-1 主流在线学习平台的视频课件资源

随着课程数量与学生规模的双重增长，课程与学生的多种属性对学习过程的影响也更加明显。在视频方面，其课程、学科、类型和时长等属性都可能对学习体验和学习效果产生直接影响[11, 121, 157]。例如，Chen 等[121] 讨论了不同的视频表现风格对学生成绩的影响，发现相比于只有语音和讲义画面或者分屏出现教师和讲义这两种视频形式，教师与讲义内容同时出现在画面中的视频形式更加能够提升学生的学习表现。Frans 等[115] 提出一个基于文本信息量的视频的复杂度模型，并利用这个模型分析视频复杂度对学生观看时间的影响，发现视频复杂度和学生停留观看时间成多项式关系，即低复杂度和高复杂度的视频有较长的停留时间，而中复杂度视频的停留时间较短。Dobrian 等[157] 分析了在线视频的传输质量对学生学习参与度的影响，发现视频缓冲时间占比较高会减少学生对课程视频的参与度。在学生方面，学生的个人背景、专业、学习时间和学习环境等属性也会对学习产生影响。例如，Kizilcec 等[149] 分析了学生的自律学习策略对在线学习行为的影响，发现学生的学习动机、家庭背景等属性与自律学习策略相关。Kim 等[12] 从学生观看长视频、重看视频、以及观看讲义视频的退出率较高，并识别了五种不同的学生观看模式解释这些高退出率现象。

上述研究说明，课程视频与学生的多种属性对在线观看视频的学习过程和学习成效具有显著的影响。因此，探索分析这些属性对视频资源的利用模式的影响具有重要

https://ocw.mit.edu/index.htm

的研究价值与现实意义，对于教师，掌握视频的利用模式不仅能够深入理解学生在线学习过程的特征，还能为改进教学手段和提高教学质量提供评价参考；对于教育机构，通过分析视频与学生在多个属性方面的分布与关系，能够为掌握视频资源的利用规律、提高视频制作水平、控制运营成本和改进学习体验等方面提供数据支持，从而更好的提供教学支持服务。

得益于学习管理系统（Learning Management System，LMS）在网络课程在线学习中的广泛使用，视频与学生的多种属性以及在线学习过程的所有操作都被完整的记录在在线学习行为数据中，形成了大规模多属性的在线视频观看行为数据，为分析教学视频的利用情况提供了基础支持。针对这种数据，现有研究从分析工具与分析方法两个方面展开了探索。在分析工具方面，现有的学习管理系统，如 Moodle 和 Blackboard 等，均为教师和管理人员提供了对教学视频利用情况的分析功能，包括访问次数统计、观看时长统计等[132]。然而，现有分析工具通常是针对一门课程在一个课程周期内的视频使用情况展开分析，且分析目标单一、范围有限，提供的界面主要以报表和基本统计图为主，不适用于高维度、多属性、周期长的大规模在线视频观看行为数据的分析；在视频观看行为数据分析方法方面，现有研究主要采用统计、机器学习等方法[21, 107]，分析视频观看行为的行为模式特征[5]、观看模式与学习效果的关系[158]、以及课程留存率[159]等。将可视分析方法应用在视频观看行为数据方面的研究相对较少，其主要关注细粒度的观看热点操作[118]、数量统计[103] 以及点击流分析（Clickstream）[119] 等，缺少综合分析大规模多课程以及多时间段视频观看行为数据的可视化设计。

现有分析工具和分析方法难以从大规模多属性关联的视频观看行为数据中提取有价值的宏观视频利用模式，无法满足教育机构和教师对数据探索分析的需求，其困难具体表现在：首先，由于大规模的在线学习过程中，视频观看行为不仅涉及视频本身的课程、类型和专业等属性，还包括了学生的个人背景、专业、学习时间等多个属性，这些属性之间存在关联，难以全面的分析不同属性的视频利用情况；其次，由于视频数量和学生数量都非常庞大，难以直观呈现大规模的视频利用情况的分布和潜在模式，帮助用户掌握整体和局部的视频利用情况。

为了解决上述困难，本章将可视分析方法应用于大规模多属性的视频观看行为数据，提出了一种探索不同属性层面的视频利用模式的可视分析方法，实现对视频观看行为数据的采集、预处理、组织、存储、计算以及交互可视化分析。首先，本文总结了视频利用模式的分析任务与设计需求，提出了一种通用的视频利用指标用来衡量不同对象、不同维度的视频利用情况；然后，在通用指标基础上，设计了前后端多层分离的可视分析系统架构；最后，设计了基于球面布局的多属性概览视图、基于混合图表的多属性详情视图以及基于平行坐标系的多属性对比视图这三个关联视图，分别从不同尺度和层面展示视频的利用情况分布.

在后续的各节中，本文将在第 5.2节讨论视频利用模式的分析任务与可视化设计需求，在第 5.3节说明如何从视频观看行为数据中抽象出反映视频利用情况的到课率与利用率指标，在第 5.4节详细介绍针对分析任务与设计需求提出的多个交互视图的可视化

设计与交互设计，在第5.5节结合大规模的在线学习行为数据介绍了三个案例，并讨论了从中发现的多种视频利用情况模式，最后在第5.6节对本章工作进行分析和总结。

5.2 视频利用模式分析任务

本章的可视化分析任务是以网络远程教育这种大规模在线学习过程为背景，由于其独特的运作模式和学习目标，本节将首先讨论关于网络远程教育的学习过程与其他在线学习平台的共性和差异，然后介绍本章的分析任务与设计需求，并总结针对分析任务提出的可视分析系统架构。

5.2.1 分析任务背景

网络远程教育的在线学习过程与MOOC等网络课程基本相同，包括使用学习管理系统观看课程视频、使用论坛与学生和教师讨论问题、在线测验与考试、提交作业、论文等，因此其在线学习行为数据在形式与规模方面同常见的MOOC课程基本相同。但是，由于社会功能定位与办学理念等方面的不同，网络远程教育的学习过程在以下方面存在显著差异：

1) 课程数量与学习周期差异。相比于普通MOOC课程或者专业培训只包含一门或数门课程，网络远程教育作为国家教育部认可的学历教育的组成部分，其定位是针对学科专业的学历教育，其培养模式参照全日制的学历教育制定。以网络远程教育中的专升本教育为例，按照各专业的教学大纲，每个专业平均具有约22门课程，包含1,348个课程视频，合计约760学时，涵盖公共基础课、专业基础课、专业课以及专业选修课等课程。并且，由于课程数量较多，网络远程教育的学制通常为2-5年，包含4个以上的学期，每个学期完成4-6门课程的学习。
2) 学籍管理差异。与随时随地入学的MOOC课程不同，网络远程教育每年固定有春秋两次入学时间，并且全国各地的学生需要在所在地区的学习中心报名入学，并由学习中心组织学生参加入学考试、线下辅导、学籍事务管理等工作。因此，学生的学习效果也会与归属的学习中心服务质量相关。

由上述差异的存在，使得对网络远程教育产生的在线学习行为数据的分析有所不同。首先，由于网络远程教育机构需要满足专业培养计划的课程体系设置要求，每个专业的课程建设都需要投入可观的人力和物力筹备大量的视频资源，不仅包括录像、剪辑、转码、更行等影视制作投入，还包括传输、存储、内容分发和流量监控的技术投入。因此，每个课程视频都有多个属性可能对学习过程产生影响，如课程归属、时长、分辨率、表现形式等。对于网络远程教育机构，掌握视频资源的利用情况在提高视频质量、控制运营成本、改进学习体验等方面都有促进作用，视频资源被学生、课程、专业以及其他相关群体的不同层面的利用情况是网络教育机构特别关注的对象。

其次，由于学籍管理模式的差异，学生的学习中心属性与学生的视频资源利用情况、学习过程和学习效果相关。因此，网络远程教育机构特别关注学生的学习中心属性

在整个学习过程中的影响。

5.2.2 视频利用模式分析任务与设计需求

本文与网络远程教育领域的专家合作，通过访谈、实地调研以及原型系统测试等方式，归纳了视频利用模式分析的主要任务：

- **T.1**：分析全部视频的利用情况分布。根据领域专家的教学经验，学生的视频利用分布具有显著差别，研究整体视频利用分布对于配置网站流量、提升观看体验等有非常重要的作用。
- **T.2**：分析每个课程的视频利用分布。不同专业、课程、章节的视频在内容上有显著差别，分析课程的视频利用分布有助于改进教学质量。
- **T.3**：分析每个学生的视频利用分布。从学生的视频利用情况在课程、时间等方面的分布可以了解学生个体的学习进度和在线学习时间管理的风格，从而为改进学习体验与促进学生支持服务提供参考。
- **T.4**：分析每个学生群体的视频利用分布。学生按照归属的学习中心可以划分为不同的学生群体，而学习中心在网络远程教育中具有重要的管理作用，视频利用分布有助于评价学习中心的招生与服务水平。

为了完成上述分析任务，本节针对视频观看行为数据规模大、属性多且相互关联的特点，概括了以下的设计原则指导可视化设计：

- **R.1**：多尺度探索。大规模的在线学习行为数据所具有的多个属性呈现出层次化特点，并且每个层次可能具有不同的模式，从全部视频的整体利用情况（T.1），到专业、课程、学生以及其他属性层面（T.2，T.3）。为了在不同属性层面上展开探索分析，可视化应当支持从全局到各尺度的数据展示。
- **R.2**：多角度呈现。每个属性层面都可能与其他属性相关联，产生不同侧面的分布特征（T.2，T.3，T.4）。例如课程的视频利用率既与章节编排顺序有关，也与学生群体相关，同时也可能与学期相关。因此，可视化应当从不同角度反映各属性层面的分布特点。
- **R.3**：对比分析。对于任意视频或者学生个体，应当能够对比分析两者或者两者以上对象的差异（T.2，T.3，T.4）。同时，对于不同属性层面的对象，也应该支持跨尺度的对比。例如个体学生与特定群体学生、全体学生之间的视频利用率差异。因此，可视化应当支持同类对象以及跨尺度对象之间的对比。
- **R.4**：交互探索。由于所有属性之间存在明显的关联关系，可视分析系统应当支持在不同属性之间往复选择，以及自上而下或者自下而上的双向导航。

5.2.3 系统整体框架

针对上述任务与可视化设计原则的分析，需要构建有效的数据处理与交互式数据分析平台。因此，本文设计了 VUSphere 可视分析系统以满足视频利用模式的探索分析。

其架构设计如图 5-2所示，总体包括四个部分：数据预处理、数据分析、数据存储、以及可视化。数据预处理、基于视频利用率的数据分析、数据存储、以及数据可视化分析。首先，数据预处理和数据分析在可视化之前离线完成。除了收集在线学习行为数据以外，数据预处理模块还从流媒体服务器采集了视频信息，包括视频时长与缩略图，分别用于计算视频利用率指标和可视分析系统展示。其次，为了提高可视化界面的数据读取效率，所有统计分析结果保存在 MongoDB 数据库中。最后，可视化模块提供多种视图展现分析结果，领域专家可以利用系统并结合自身经验对视频利用模式进行探索。

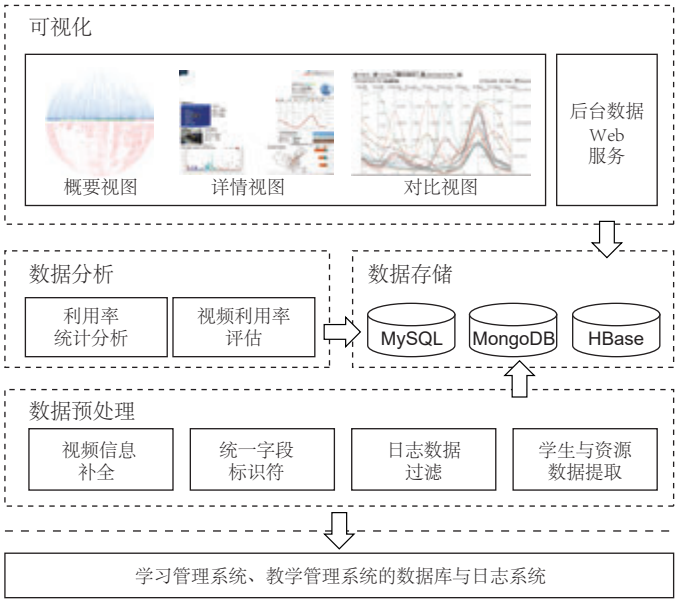


图 5-2 VUSphere 可视分析系统架构

5.3 基于观看行为数据的视频利用模型

5.3.1 到课率与利用率定义

视频观看行为数据在细粒度层面可以反映多种不同的观看模式，如完整观看、浏览、重点查看[12] 等，在这方面已有研究根据点击流数据（Clickstream）展开分析，包括视频内的操作热点[119]、观看注意力[11] 等。而在分析视频利用情况方面，通常会使用视频观看数量或者观看时长作为分析指标，为衡量学生的学习参与度提供支持[47, 65]。

这些数量和时长指标在分析单门课程或者相似时长的多门课程时是合适的，但是在大规模在线学习中，由于专业、课程以及学生群体的多样性，这些采用累加值的指标可能会引起误导。例如，假设 A 课程和 B 课程分别有 10 个和 40 个视频，当观察到 A 课程和 B 课程的视频都被观看了 20 次时，A 课程的视频可能已经被充分利用了，而 B 课程仍至少有一半视频没有被观看。为了解决上述问题，本文提出到课率（Attendance Rate，AR）和利用率（Utilization Rate，UR）作为现有指标的补充，从而为衡量视频的利用情况提供一致性的标准。

1) 基本到课率与利用率

到课率 AR 的基本定义如下，一名学生 s 对一个视频 v 的到课率 $ar_{s,v}$ 为：

$$ar_{s,v} = \begin{cases} 1 & \text{viewed} \\ 0 & \text{not viewed} \end{cases} \tag{5-1}$$

即，如果学生 s 已经观看过视频 v，则 $ar_{s,v} = 1$，反之 $ar_{s,v} = 0$。到课率与传统课堂中的“出勤”的概念类似，每个视频可以看作一节课，如果学生观看了这个视频，则类似于学生在这节课上“已出勤”。

利用率 UR 的基本定义如下，一名学生 s 对一个视频 v 的利用率 $ur_{s,v}$ 为：

$$ur_{s,v} = \frac{wt_{s,v}}{vt_v} \tag{5-2}$$

式中，$wt_{s,v}$ 为学生 s 观看视频 v 的累计时长，vt_v 为视频 v 的总时长。利用率 $ur_{s,v}$ 反映了学生观看视频的时间长短。学生观看视频的时间越长，则利用率 $ur_{s,v}$ 越大，反之越小。和到课率不同，由于 $wt_{s,v}$ 是累加计算的，所以重复观看同一个视频的时长会被累加，进而可能使利用率 $ur_{s,v} > 1$。

在单个学生和单个视频的到课率 $ar_{s,v}$ 和利用率 $ur_{s,v}$ 的基础上，可以进一步定义学生在课程上的到课率和利用率。学生的课程到课率定义如下，一名学生 s 对一个课程 c 的课程到课率 $ar_{s,c}$ 为：

$$ar_{s,c} = \frac{|W_{s,c}|}{V_c}, W_{s,c} = \{ar_{s,v} | ar_{s,v} = 1, v \in V_c\} \tag{5-3}$$

式中，V_c 是课程 c 的所有视频列表，$W_{s,c}$ 是学生在课程 c 中已经观看的不重复视频列表。根据这个定义，已经被观看过的视频只会被计数 1 次，重复观看已经观看的视频并不会增加课程的到课率 $ar_{s,c}$。例如，假设一个课程有 10 个视频，一个学生只观看了一个视频但是观看了 5 次，那么该学生对这门课程的课程到课率为 $ar_{s,c} = 1/10 = 0.1$，而不是 $ar_{s,c} = 5/10 = 0.5$。

学生的课程利用率定义如下，一名学生 s 对一个课程 c 的课程利用率 $ur_{s,c}$ 为：

$$ur_{s,c} = \frac{\sum_{wt \in WT_{s,c}} wt}{\sum_{vt \in VT_c} vt}, WT_{s,c} = \{wt_{s,v} | v \in V_c\} \tag{5-4}$$

式中，$WT_{s,c}$ 是学生在课程 c 中已经观看视频的观看时长列表，VT_c 是课程 c 的所有视频视频时长列表。和课程到课率不同，由于重复观看同一个视频的时长会被累加，所以可能使课程利用率 $ur_{s,c} > 1$。例如，假设一个课程有 10 个视频每个 1 小时，一个学生只观看了一个视频但是观看了 5 次，每次 0.5 小时，那么该学生对这门课的课程利用率为 $ur_{s,c} = (0.5 \times 5)/(10 \times 1) = 0.25$。因此，课程利用率也可能大于 1，说明学生重复观看了某些课程视频。

2) 多属性拓展的到课率与利用率

在上述学生的课程到课率和课程利用率的基础上，可以进一步拓展到学生与视频资源的其他多个属性，使视频利用情况可以在不同属性内部和属性之间展开对比分析。针对分析任务 T.2，T.3 和 T.4，本节将课程到课率和课程利用率拓展到学生的学习中心，入学批次，学习专业以及课程这些相互关联的属性上。

定义 L, B, M, C 分别为全部学习中心，全部入学批次，全部学习专业，全部课程的集合，则各属性的到课率与利用率的一般形式为：

$$ar_{S_{L',B',M',C'}} = \frac{1}{|S_{L',B',M',C'}|} \sum_{c \in C' \wedge s \in S_{L',B',M',C'}} ar_{s,c} \tag{5-5}$$

$$ur_{S_{L',B',M',C'}} = \frac{1}{|S_{L',B',M',C'}|} \sum_{c \in C' \wedge s \in S_{L',B',M',C'}} ur_{s,c} \tag{5-6}$$

式中，$S_{L',B',M',C'}$ 代表满足特定条件的学生集合，其中 L', B', M', C' 分别代表学习中心、入学批次，学习专业以及课程的条件集合，分别满足 $L' \subseteq L \wedge L' \neq \emptyset$，$B' \subseteq B \wedge B' \neq \emptyset$，$M' \subseteq M \wedge M' \neq \emptyset$，以及 $C' \subseteq C \wedge C' \neq \emptyset$；由于专业与课程存在关联关系，课程集合 C' 还满足 $\forall c \in C', c \in C_m \wedge m \in M'$，表示课程都属于学习专业集合 M' 的课程列表。

由于 L, B, M, C 任意集合的元素数量较多，L', B', M', C' 各属性的组合数均为 $C_N^1 + C_N^2 + \cdots + C_N^N = 2^N - 1$ 级别，所以学生集合 $S_{L',B',M',C'}$ 的取值空间极大。而在实际探索分析时会选择有明确物理含义的组合作为参考学生集合，例如将学生 s 在课程 c 的到课率与与同专业、同入学批次的全部学生的到课率作对比，则各条件集合为 $L' = L, B' = \{s.b\}, M' = \{s.m\}, C' = \{c\}$，即各属性只会选择全集或者元素数量为 1 的集合。因此，各属性的组合数只有 $C_N^1 + C_N^N = N + 1$ 个，全部参考学生集合的数量为：

$$N_S = (|L| + 1)(|B| + 1)(|C| + 1) + \sum_{m \in M} (|L| + 1)(|B| + 1)(|C_m| + 1) \tag{5-7}$$

式中，C_m 是专业 m 的课程列表。其中，前一项代表全部课程不分专业的全部属性集合，后一项代表每个专业每个课程。以本文案例分析的数据集为例，$|L| = 91, |B| = 6, |M| = 11, |C| = 219$，各专业含公共课的平均课程数量为 24 门，则全部参考学生集合的数量为 318,780 个。

5.3.2 到课率与利用率存储

同时，为了进一步提高可视化界面检索这些参考数据的效率，本文将所有的参考学生集合的到课率和利用率存储在哈希表中。每条记录以各属性元素值的组合为键，当查询任意学生或者课程时，即可通过关联属性查找所有相关的参考集合。哈希表所有记录的键均为 $l - b - m - c$ 的形式，即使用短横线连接将学习中心、入学批次、学习专业与课程这 4 个属性连接。当属性是单个元素时，使用属性的唯一标识符，而当属性是全

集时，使用 ALL 字符串占位代替。例如，对于属性组合为 081 号学习中心，全部入学批次，01 号计算机科学专业，全部课程的视频利用率记录，其键为 $081-ALL-01-ALL$。为了提高可视化界面展示相关信息的效率，全部学生集合的到课率 AR 和利用率 UR 都将被预先计算并保存在 VUSphere 可视分析系统的数据存储中。

5.4 可视化设计

根据第 5.2节讨论的分析任务与设计原则，视频观看行为的核心主体是视频和学生，因此，所有的可视化设计均应围绕这两个主体展开。并且，针对视频观看行为数据规模大、属性多的特点，应采用一些特别的可视化设计呈现多方面的视频利用率分布。首先，针对展示整体的视频利用情况分析（T.1，T.3），可视化设计应同时展现全体学生与全部视频的数据点。而针对课程、学生群体的分析和对比（T.2, T.4）则应从不同角度呈现统计结果与分布差异。数据点的分布可以采用散点图、热力图等方式呈现，统计结果则可以通过折线图、柱状图等方式呈现，分布差异可以通过雷达图、平行坐标系等方式呈现。其次，由于各项分析任务涉及的数据规模庞大，且数据属性有不同的特征并相互关联，因此针对各项分析任务，本节按照前述的设计原则设计了三个关联视图，通过综合运用多种图表从多个尺度与层面探索视频利用模式，其主要界面如图 5-3所示。

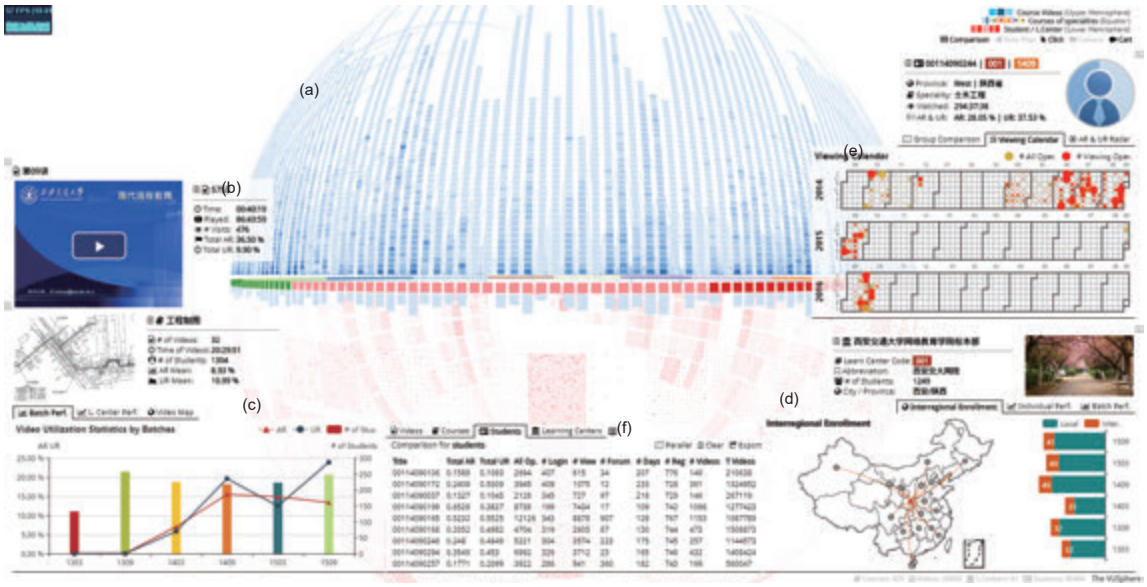


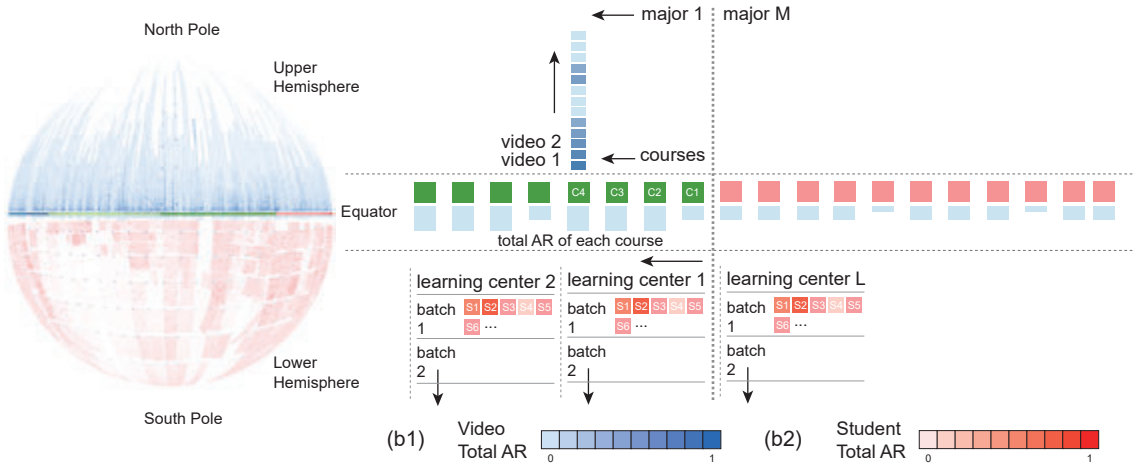
图 5-3 VUSphere 可视分析系统界面截图，其中 (a) 为基于球面布局的概览视图，(b) 为视频详情面板，(c) 为课程详情面板，(d) 为学习中心详情面板，(e) 为学生详情面板，(f) 为对比视图

本节将详细说明 VUSphere 可视分析系统三个关联视图的可视化设计，包括基于球面布局的多属性概览视图，基于混合图表的多属性详情视图，以及基于平行坐标系的多属性对比视图。

5.4.1 基于球面布局的多属性概览视图

为了分析整体的视频利用率模式（T.1），需要显示全部视频与全体学生的视频利用分布情况。然而，视频利用分布情况的数据存在以下问题：首先，由于视频与学生的数量庞大，直接以散点图、热力图等形式将所有数据点显示在平面上会造成视觉遮挡，数据点密集难辨等问题。其次，除了到课率与利用率是连续数值型数据，视频和学生的其他主要属性，如课程、专业等，都是离散的类别数据，需要将这些属性映射到有限的视觉编码。最后，在各属性上的数据量分布并不均匀，有些课程可能包含上百个视频，而有些只有几个视频。

因此，针对上述问题，本文提出一种基于球面布局的多属性概览视图，通过将数据点按多个属性映射为球体表面的球面坐标和其他视觉编码，展示大规模的学生与视频利用情况分布。其基本结构如图 5-4(a) 所示：



(a) 概览视图界面截图　　(b) 概览视图可视化设计展开图

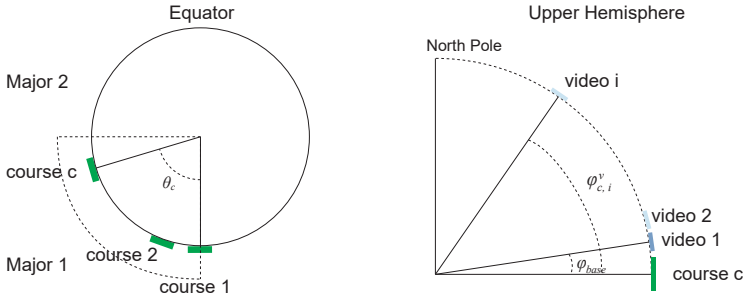
图 5-4 基于球面布局的多属性概览视图可视化设计

为了直观呈现全体数据点分布情况，本文设计了一个固定半径为 r 的三维球体并利用其表面展示所有数据点。尽管通常情况下采用三维空间展示数据可能会造成遮挡、透视变形等问题[86]，但是本文提出的设计是采用球体表面展示数据，而不使用球体内部的空间展示数据点，避免由在球体内部显示数据造成的视觉遮挡。并且，通过固定摄像机视角指向球心以及稍后说明的交互技术，可以保持视觉中心区域的图形元素具有相似的尺寸比例与可预期的视觉形变。另外，尽管球体边缘区域的图形元素存在表面形变，但是交互过程符合观察地球仪等球体表面信息的日常经验，因此不会造成观察困难和理解障碍。

图 5-4(b) 展示了按墨卡托投影将概览视图球面展开的二维平面示意图，共显示 3 类数据点：课程、视频、以及学生。其中，赤道带显示课程数据点，北半球显示视频数据点，下半球显示学生数据点。以下将分别说明这 3 类数据点的坐标设计。

课程：每个课程数据点映射为一个平面法线指向球心且尺寸相同的方块。所有课

https://en.wikipedia.org/wiki/Mercator_projection



(a) 专业和课程的布局设计　　(b) 课程视频的布局设计

图 5-5 课程、视频与学生的数据点球面坐标，其中 (a) 为 0 纬度赤道面，(b) 为上半球的经度线切面

程方块沿赤道带自西向东均匀分布在球面的赤道带，课程所归属专业映射为方块颜色。每个方块中心的纬度不变，经度按以下步骤计算：首先，所有课程按专业序号分组，其中公共课程划分到公共组，多专业共有的课程划分到序号较小的组；然后，同一个专业的课程按照教学顺序在组内排列；最后，将所有课程按照此规则排序。完成排序后按照总课程数量均匀分配每个课程的经度坐标。

如图 5-5(a) 所示，排序后第 c 个课程的球面坐标纬度 φ_c 和经度 θ_c 分别为：

$$\varphi_c = 0 \tag{5-8}$$

$$\theta_c = c\frac{2\pi}{|C|} \tag{5-9}$$

其中，C 代表全部课程集合。而课程方块均在赤道带上，所以纬度 φ_c 均为 0。

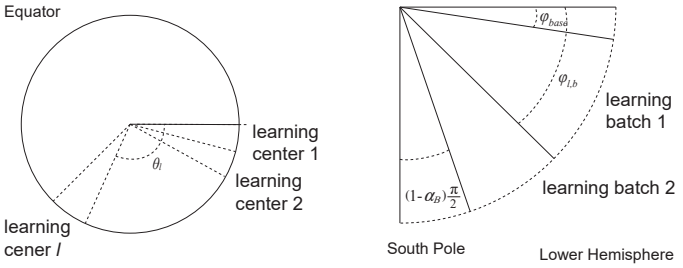
视频：每个视频数据点映射为一个平面法线指向球心、且尺寸相同的矩形，视频的到课率按照图 5-4(b1) 所示的颜色编码映射为矩形的颜色。所有视频矩形沿归属课程所在的经度线自赤道向北极点，按照视频在课程内的章节顺序排列。根据上述规则，如图 5-5(b) 所示，第 c 个课程的第 i 个视频 $v_{c,i}$ 映射到球面上的坐标纬度 $\varphi^v_{c,i}$ 和经度 $\theta^v_{c,i}$ 分别为：

$$\varphi^v_{c,i} = i\frac{\pi}{2V_{max}} + \varphi_{base} \tag{5-10}$$

$$\theta^v_{c,i} = c\frac{2\pi}{|C|} \tag{5-11}$$

其中 V_{max} 代表全部课程中一门课程中最大的视频数量，C 代表全部课程集合。由于在赤道带上已经有课程方块，为了避免符号重叠，计算视频纬度时有 φ_{base} 偏移量。由于课程的所有视频矩形与课程方块都在同一经度线上分布，所以视频矩形的经度坐标 $\theta^v_{c,i}$ 与课程的经度坐标 θ_c 相同。

学生：每个学生数据点映射为一个平面法线指向球心、且尺寸相同的方块，学生的到课率按照图 5-4(b2) 所示的颜色编码映射为方块颜色。利用学生的学习中心与入学批次属性，可以将全部学生分为多个群组，先将群组映射到球面的特定区域，然后在区域内显示每个学生方块。为了使区域布局中的每个区域的尺寸更加均匀，考虑到学习中



(a) 学习中心属性的布局设计　　(b) 入学批次属性的布局设计

图 5-6 课程、视频与学生的数据点球面坐标，其中 (a) 为 0 纬度赤道面，(b) 为下半球的经度线切面

心的数量（$|L| = 91$）显著大于入学批次的数量（$|B| = 6$），本文将将学习中心序号属性映射为经度，将入学批次属性映射为纬度。根据上述规则，如图 5-6(a) 和图 5-6(c) 所示，对于第 l 个学习中心在第 b 个入学批次，其在球面的学生区域基准纬度 $\varphi_{l,b}$ 与基准经度 $\theta_{l,b}$ 为：

$$\varphi_{l,b} = b\frac{\alpha_B \pi}{2|B|} + \varphi_{base} \tag{5-12}$$

$$\theta_{l,b} = l\frac{2\pi}{|L|} \tag{5-13}$$

其中，B 代表全部入学批次集合；L 代表全部学习中心集合；α_B 参数为显示入学批次的纬度宽度调整系数，该参数可以调节每个入学批次在球面上的纬度带宽度。由于在高纬度区域相同经度区间的球面面积比低纬度区域显著减少，为了避免数据点在靠近南极点附近的高纬度区域过于密集，需要避免生成过高的学生区域的基准纬度。例如，当 $\alpha_B = 0.8$ 时，高于 $0.8\pi/2$，即南纬 72° 以上的高纬度区域不会设置学生区域。

在学生区域基准纬度 $\varphi_{l,b}$ 与基准经度 $\theta_{l,b}$ 的基础上，可以进一步按照学生的注册时间顺序将学生数据点在区域内逐行依次排列。根据上述规则，对于第 l 个学习中心在第 b 个入学批次的第 j 名学生 s 映射在球面上的坐标纬度 $\varphi^s_{l,b,j}$ 和经度 $\theta^s_{l,b,j}$ 分别为：

$$\varphi^s_{l,b,j} = b\frac{\alpha_B \pi}{2|B|} + \beta_B \lfloor \frac{j}{N_w} \rfloor \frac{\alpha_B \pi}{2|B|} + \varphi_{base} \tag{5-14}$$

$$\theta^s_{l,b,j} = l\frac{2\pi}{|L|} + \beta_L (j \bmod N_w)\frac{2\pi}{|L|} \tag{5-15}$$

其中，N_w 代表在一个学生区域内同一纬度上的最大学生方块数量；β_B 参数代表一行学生方块占学生区域的纬度比例；β_L 参数代表一列学生方块占学生区域的经度比例，其数值为 $\frac{1}{N_w} - \varepsilon_L$，$\varepsilon_L$ 是两个学生区域之间经度间隙。为了避免渲染过程中发生学生方块溢出学生区域，上述参数均根据实际数据集统计得到，并满足约束条件 $\frac{1}{\beta_L}(\frac{1}{\beta_B} - 1) \leq max(\{S_{l,b}\}) \leq \frac{1}{\beta_L \beta_B}$，表示一个学生区域内可显示的数据点数量不小于某个学习中心 l 在入学批次 b 的最大学生数量，且不会多出超过 1 行。通过设置合适的 β_L 与 β_B 参数，可以在视觉上明显区分不同学习中心和入学批次的边界。

最后，在渲染时多属性概览视图的球体时，需要将球面坐标转换为直角坐标系，对

每个数据点的纬度 φ 和经度 θ，按以下公式转换：

$$x = r \cdot \sin\varphi \cdot \cos\theta$$
$$y = r \cdot \sin\varphi \cdot \cos\theta$$
$$z = r \cdot \cos\varphi$$

5.4.2 基于混合图表的多属性详情视图

为了呈现课程、学生以及学习中心的视频利用分布情况（T.2，T.3，T.4），本节设计了多属性详细视图从不同的角度展示每个课程、学生以及学生中心的详细统计（R.2）。该视图由 4 个面板构成，包括视频详情面板（图 5-3(b)），课程详情面板（图 5-3(c)），学生详情面板（图 5-3(e)），以及学习中心详情面板（图 5-3(d)）。这些面板分别位于概览视图的左右两侧，提供视频、课程、学生以及学习中心的基本信息与多方面的统计结果图。除了视频详情面板以外，其他各面板均包含多个标签页，从其他关联属性的角度展示统计结果，具体如下：

视频面板：如图 5-3(b) 所示，该面板提供单个视频的基本信息以及总体的到课率和利用率（T.1）。借鉴视频可视分析的相关研究，如 VisMOOC 等[118]，该面板内嵌了基于 HTML5 技术的视频播放器以便用户查看视频的具体内容。

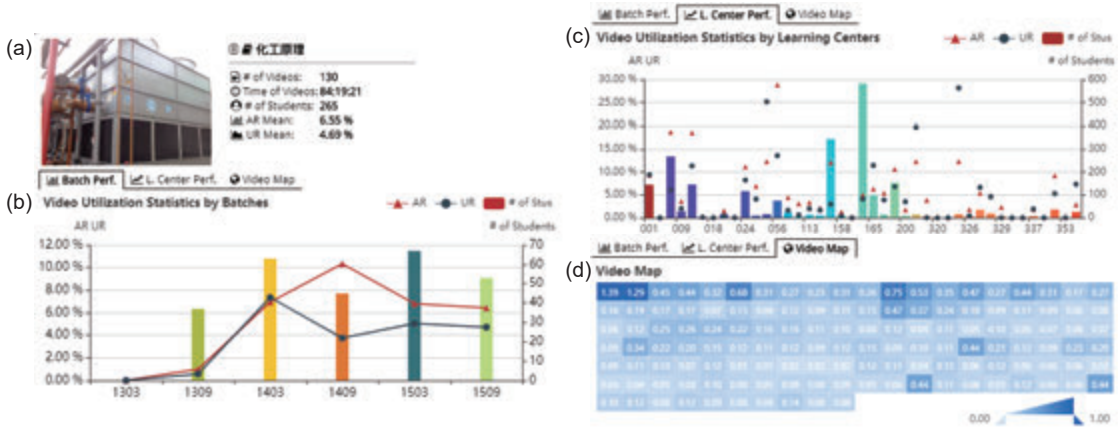


图 5-7 详细视图中的课程详情面板，其中 (a) 为课程基本信息，(b) 为基于柱状图和折线图的按入学批次统计 AR 和 UR 分布，(c) 为按学习中心统计的 AR 和 UR 分布，(d) 为课程各个视频的 AR 热力图

课程详情面板：该面板提供基本的课程信息以及与课程相关的入学批次，学习中心，课程视频这三个关联属性的视频利用情况统计（T.2）。如图 5-7所示，入学批次，学习中心这两个属性需要展现到课率、利用率、以及学生人数这三个方面按照批次和学习中心的统计结果，为了显示这方面的差异，本文采用柱状图、折线图和散点图分别展示。而对于课程中的所有视频，由于课程的视频数量差异大，柱状图或者折线图会出现过于稀疏或过于密集的问题，因此采用热力图以充分利用平面空间（图 5-7(c)），将每个视频的到课率 UR 映射为颜色编码，并显示其数值作为参考。

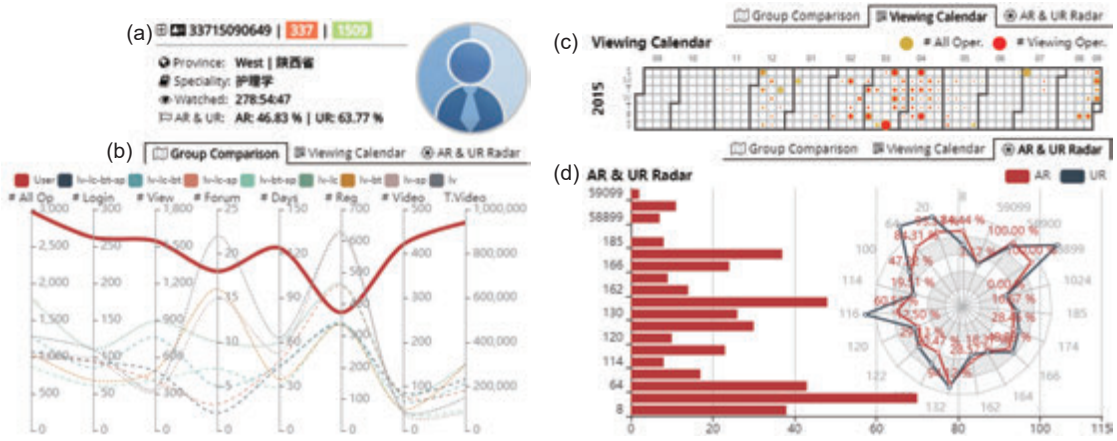


图 5-8 详细视图的学生详情面板，其中 (a) 为学生基本信息，(b) 为基于平行坐标系展示的该学生与相似属性学生群体的对比，(c) 为基于日历图的长期视频观看情况，(d) 为学生的每个课程的 AR 与 UR 情况

学生详情面板：为了全面的展示学生的视频利用情况（T.3），该面板提供与学生相关的全部课程、学习时间这两个关联属性的视频利用情况统计，并提供对比分析界面（R.4）。如图 5-8(d) 所示，为了展现学生对各个课程的视频利用情况，本文采用柱状图与雷达图展示每个课程的到课率与利用。为了显示单个学生学习参与度和整体之间的差异，本文将该学生与其他多个尺度学生群体的平均学习参与度共同展示在同一个平行坐标系中，从而为用户提供对比参照（图 5-8(b)）。为了展示学生在长期学习过程中观看视频的时间安排与视频利用情况，在前一章设计的基于日历图的学习时间分布图的基础上，本文设计了长期学习过程日历图，将图表的时间跨度从单个学期扩展到了整个在读期间。如图 5-8(c) 所示，每日视频观看数量与每日活动总数量分别映射为数据点的面积，数据类型映射为数据点颜色。另外，由于学生的视频观看数量与活动总数量存在显著差异，因此在映射时先将各数量对数转换再映射至有限的面积区间，避免面积过大造成视觉遮挡。

学习中心详情面板：该面板提供与学习中心相关的入学批次、全体学生、生源地域这三个关联属性的视频利用情况统计（T.4）。如图 5-9(a) 和 (b) 所示，针对入学批次，采用类似课程面板中的柱状图与折线图展现多方面的视频利用情况统计。针对各批次的学生，采用热力图将主要学生的专业到课率 AR 映射为颜色编码，并按照注册时间顺序排列。另外针对学生的生源地属性，本文采用地图展示学习中心各入学批次的招生地域差异（图 5-9(c)）。生源分布地图根据学生的工作地址和学习中心的招生地在地图上绘制学生来源分布，本地和外地的学生数量用不同颜色编码，并使用柱状图表现一个学习中心在本地和外地招生的数量差异。

5.4.3 基于平行坐标系的多属性对比视图

为了对比不同学生个体，视频以及各属性在不同角度的差异，本文设计了基于平行坐标系的多属性对比视图（R.3），从视频、学生、课程、以及学习中心这 4 个方面展示

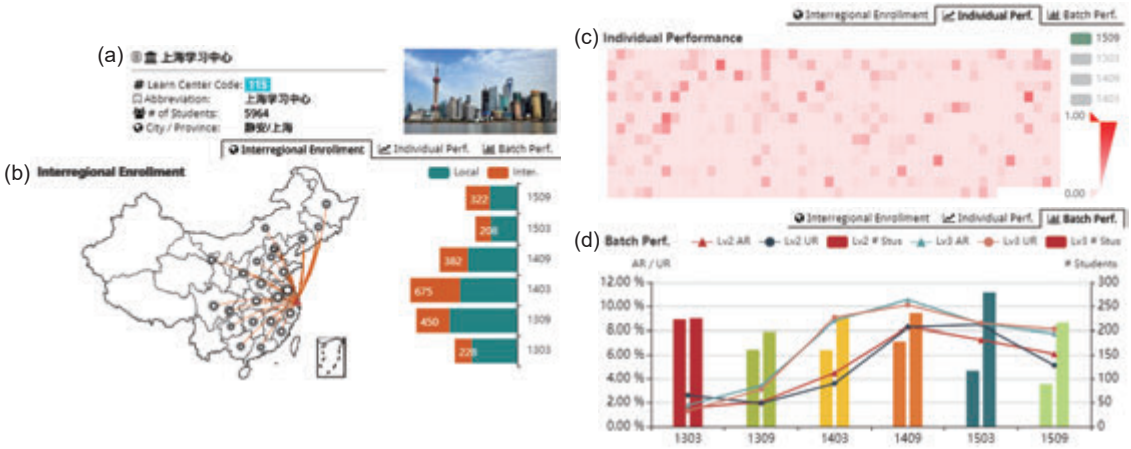
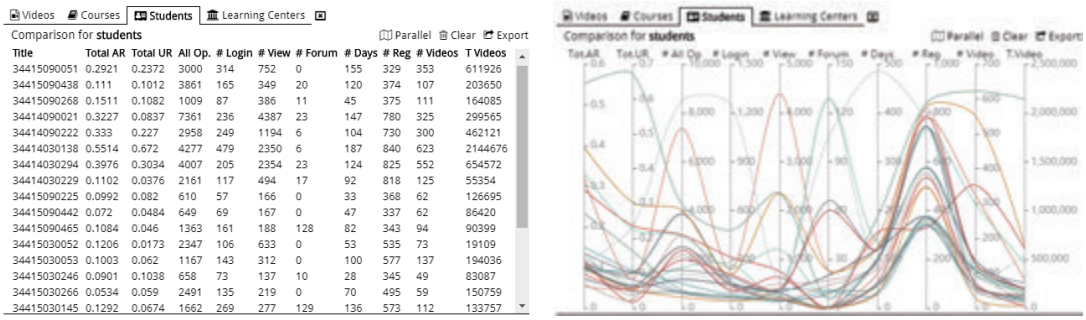


图 5-9 详细视图的学习中心详情面板，其中 (a) 为学习中心基本信息，(b) 为基于地图的学生生源分布情况，(c) 为基于热力图的每个学生的 AR 分布，(d) 为基于柱状图和折线图的按入学批次的 AR 和 UR 分布

他们在不同层面的差异（图 5-3(f)），对比视图共包括 4 个标签，分别对应上述 4 个方面。为了管理需要对比的数据点，本文设计了一个数据表格显示所有待对比展示的数据，其中每行代表一个数据点，每列为该数据点相关的特征值（图 5-10(a)）。如图 5-10(b) 所示，表格中的数据会按照列顺序和数值映射到平行坐标系的对应坐标轴中。



Title	Total AR	Total UR	All Op.	# Login	# View	# Forum	# Days	# Reg	# Videos	T Videos
34415090051	0.2921	0.2372	3000	314	752	0	155	329	353	611926
34415090438	0.111	0.1012	3861	165	349	20	120	374	107	203650
34415090268	0.1511	0.1082	1009	87	386	11	45	375	111	164085
34414090021	0.3227	0.0837	7361	236	4387	23	147	780	325	299565
34414090222	0.333	0.227	2958	249	1194	6	104	730	300	462121
34414030138	0.5514	0.672	4277	479	2350	6	187	840	623	2144676
34414030294	0.3976	0.3034	4007	205	2354	23	124	825	552	654572
34414030229	0.1102	0.0376	2161	117	494	17	92	818	125	55354
34415090225	0.0992	0.082	610	57	166	0	33	368	62	126695
34415090442	0.072	0.0484	649	69	167	0	47	337	62	86420
34415090465	0.1084	0.046	1363	161	188	128	82	343	94	90399
34415030052	0.1206	0.0173	2347	106	633	0	53	535	73	19109
34415030053	0.1003	0.062	1167	143	312	0	100	577	137	194036
34415030246	0.0901	0.1038	658	73	137	10	28	345	49	83087
34415030266	0.0534	0.059	2491	135	219	0	70	495	59	150759
34415030145	0.1292	0.0674	1662	269	277	129	136	573	112	133757

(a) 数据对比表格　　(b) 数据对比平行坐标系

图 5-10 基于平行坐标系的多属性对比视图

5.4.4 交互设计

从前述 4 项分析任务与总结的 4 项设计原则可以看出，需要从不同角度交互的查询视频利用分布情况（R.4）。因此，本文设计了多个交互工具允许用户使用多协作视图分析视频利用模式。

固定参考点导航工具。多属性概览视图（图 5-4(a)）基于三维球体表面布局，在三维空间中观察三维物体的常用工具是摄像机，通过设置摄像机的视点（eye point），参考点（at point），观察正向向量（view-up vector）以及视野、远近视截面等参数，实现在三维空间中观察对象[160]。常用的导航方式通过直线移动摄像机并调整参考点，提供了自由观察三维场景的视角。

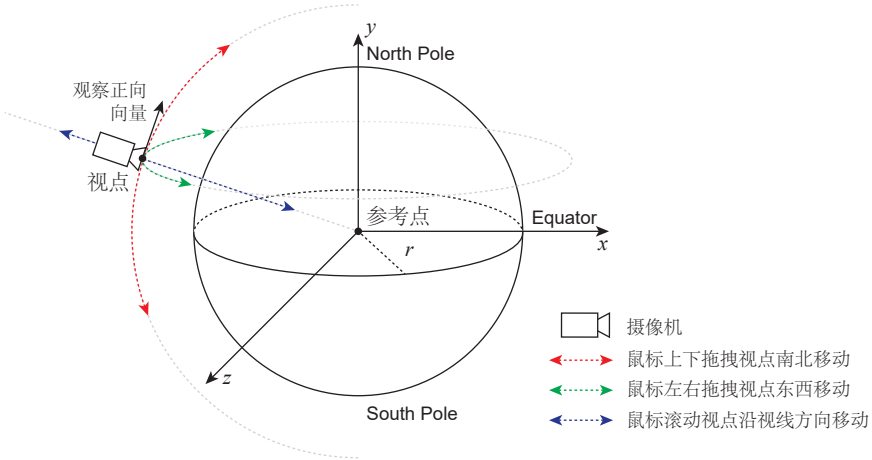


图 5-11 固定参考点导航工具

本文针对观察球面数据点的需求，设计了一种固定参考点导航工具，让领域专家可以在球面轨道上任意视角查看概览视图。如图 5-11所示，摄像机参考点固定在球心位置不改变，而。该导航工具将鼠标的两种操作与视点调整关联：1）鼠标拖拽操作改变视点在球面轨道上的位置。由于参考点固定，视点与参考点距离不变，拖拽操作将改变视点在球面轨道上的位置。上下拖拽时，沿经度线向南或者向北移动视点（红色虚线），左右拖拽时，沿纬度线向东或向西移动视点（绿色虚线）。并且，为了缓解鼠标拖拽操作准确度较低造成的摄像机视点抖动，垂直于拖拽方向的视点移动被消除；2）鼠标滚动操作沿当前视点与参考点的距离。鼠标前后滚动时，沿视点与参考点方向缩小或加大视点与参考点之间的距离（蓝色虚线）。领域专家使用该工具可以逐个查看球面的每个课程、每个学习中心以及每个学生的视频利用情况分布。

多视图协调关联。由于视频与学生的各属性之间存在关联，因此为了促进交互探索（R.4），概览视图、详情视图以及对比视图之间彼此动态关联。概览视图中的任意元素均关联至详情视图，点击课程方块、视频矩形、学生方块时，详情视图中的对应面板会显示对应信息；同时，查看详情视图中任意的视频与学生时，概览视图也会移动摄像机至相应视角；对比视图的数据表格和平行坐标系中的数据点也与其他两个视图动态关联，点击数据点时其他两个视图将更新内容为所点击数据点。另外，上述视图也支持键盘快捷键操作，用户可以关闭任意视图中的部分面板或数据以减少视觉干扰。

5.5 案例分析

为了评估 VUSphere 可视分析系统的有效性与用户体验，本文邀请了本校网络远程教育学院的两位工程师作为专家用户开展案例分析。这两位专家负责 LMS 系统的开发与视频资源的管理工作，他们全程参与了 VUSphere 可视分析系统的调研、设计与试用过程，因此对系统的基本功能较为熟悉，并且熟悉系统的主要分析任务。通过在真实的在线学习行为数据集上进行探索分析，两位专家发现了视频资源利用的多个模式，并

将这些模式按第 5.2节讨论的分析任务分为三类：总体视频利用模式、长期视频利用模式、以及生源分布与视频利用率模式。

为了进一步评估系统的有效性和所发现模式的现实意义，本文访谈了五位专家，其中三位来自本校网络教育学院，另外两位是来自本校的教育专家。首先，由于本校教育专家并未使用过该系统，本文作者介绍并演示了 VUSphere 可视分析系统的主要功能与界面；然后，专家们可以使用系统自由探索真实的在线学习行为数据集并向本文作者提问；最后，本文作者与专家们展开讨论，并收集整理了专家们的反馈意见。总体而言，所有专家都认可本章提出的可视分析方法的有效性，并认为通过利用该可视分析系统查看视频利用的分布情况能够为教学管理与学生服务提供有价值的参考信息。

本节利用本校网络教育学院 2013 年至 2015 年的在线学习行为数据集，验证本文提出的可视分析方法的有效性。该数据集包含 219 门课程的 26,660 个视频，91 个学习中心在 6 个批次的 80,484 名学生，以及这些学生在线学习产生的 100,359,271 条学习日志数据。从这些在线学习行为数据中提取视频观看行为数据后，按照本文第 5.3节提出的视频利用指标进行统计分析，得到所有视频、学生以及各属性的到课率与利用率，统计结果保存在 MongoDB 数据库中。

5.5.1 整体视频利用模式

图 5-12和图 5-13分别显示了对全部视频和学生的到课率进行可视化之后得到的结果。通过使用交互导航工具，领域专家从概览视图中发现了两个整体视频利用模式。

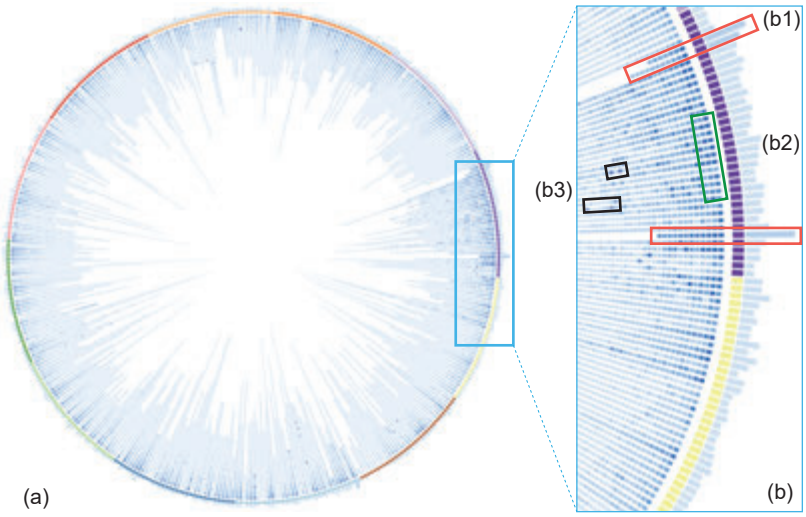


图 5-12 全部视频的整体视频利用模式，其中 (a) 为从南极点观察的概览视图截图，(b) 为蓝色矩形区域放大的细节，(b1)、(b2) 和 (b3) 是 3 种具有较高 AR 的视频代表

首先，使用固定参考点导航工具移动视点至南极点后观察概览视图，图 5-12(a) 显示了从南极点底部仰视概览视图得到的全部视频的利用率分布。全部视频矩形在视觉上投影在赤道平面上形成圆盘形状，每门课程的全部视频矩形自圆盘外边缘向圆心沿半径直线按章节顺序排列，每条半径线代表一门课程。观察圆盘的颜色分布可以发现，

圆盘整体颜色较浅，代表大部分视频的利用程度较低。进一步放大概览视图局部区域可以发现，部分视频的利用率明显高于其他视频，包括：1）视频数量较少的课程整体利用程度较高（图 5-12(b1)）；2）课程的前几节视频利用程度较高（图 5-12(b2)）；以及 3）课程中后部分的某些视频利用程度较高（图 5-12(b3)）。

领域专家表示这个可视化效果符合他们对视频点播日志统计的统计结果，即整体视频利用率较低，一少部分视频利用率相对较高。由于各个课程的视频数量较多，完整观看全部视频需要耗费较长的时间，而学生在平时需要工作，只能安排部分时间用于课程学习。受限于较少的学习时间，大部分学生无法观看所有视频，进而造成视频的整体利用率较低。而图 5-12(b) 展示的几种较高利用率的视频则反映了学生的在线学习行为，与领域专家的教学管理经验保持一致，他们解释这些较高的视频利用分布可能反映了学生的学习习惯，例如，学生对视频数量较少的课程对观看压力较低，因而能够完整观看，而对于视频数量较多的课程，学生可能观看了前几个视频后就由于观看压力较大而不再观看，或者选择观看与考试重点相关的视频。

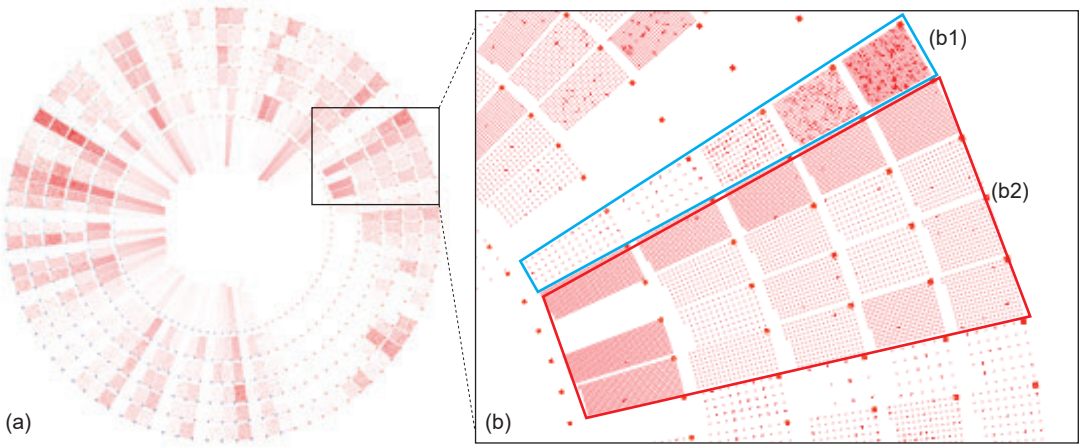


图 5-13 全部学生的整体视频利用模式，其中 (a) 为从北极点观察的概览视图截图，(b) 为黑色矩形区域放大的细节，(b1) 和 (b2) 是 2 种 AR 分布模式的学习中心代表

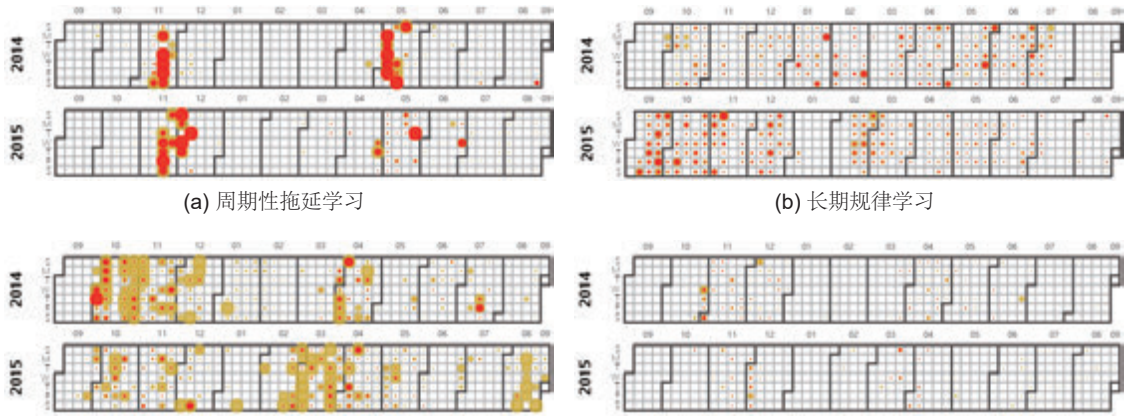
其次，图 5-13(a) 显示了将视点移动至北极点后观察得到的概览视图。全部学生矩形同样投影在赤道平面上形成圆盘，每个扇形区域代表一个学习中心，扇形区域内各扇段代表一个入学批次。可以发现学习中心的招生数量并不均匀，部分学习中心招生数量显著较多。另外，放大部分学习中心可以发现，各学习中心的视频利用率分布模式存在差异：有些学习中心的学生视频利用率明显不同（图 5-13(b1)），而有些非常一致（图 5-13(b2)）。

领域专家认为这种整体的按照学习中心统计的视频利用情况分布是以往统计工具中没有的，可视分析系统提供的可视化展示为他们带来了一种新的视角来分析学习中心的运营情况。领域专家表示，在通常情况下一个学习中心的全体学生的学习状态应当呈现明显的多样化分布，即有的学生视频利用率较高、有的较低；而过于整齐的低视频利用率可能反映了所在学习中心的管理与服务问题，例如虚假招生、跨省招生、缺少服务管理等。通过可视化呈现出的这种分布特点，网络远程教育学院可以从视频利用

的角度对学习中心的运营情况展开检测，从而发现潜在的异常，避免造成教学事故。

5.5.2 长期视频利用模式

网络远程教育的学习通常为 4-6 个学期，时间跨度为 2-3 年。本文前一章讨论分析了学生在第一学期的学习时间管理风格特点，而更加长期的整体视频学习时间模式仍不清楚。利用本章设计的学生详情面板中的长期学习过程日历图，可以查看学习时间管理风格在更长时间范围的变化情况。图 5-14显示了 4 个典型的前 2 年学习过程，分别呈现了 4 种长期学习时间管理模式。



(a) 周期性拖延学习　(b) 长期规律学习

(c) 长期规律讨论学习　(d) 长周期低参与度学习

图 5-14 长期的在线学习时间管理模式

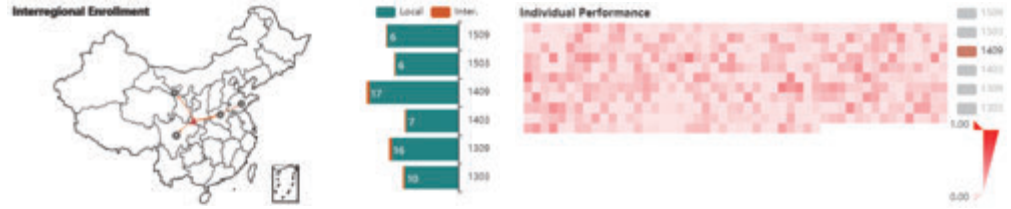
图 5-14(a) 显示了典型的拖延学习特征，该学生在学期前期不学习而在中后期集中 1-2 周内集中学习，短时间内观看大量视频；图 5-14(b) 显示了典型的规律学习特征，该学生在整个学期经常学习，两次学习日之间间隔天数很少；图 5-14(c) 显示了一种论坛讨论活跃的学习特征，该学生在每个学期很少观看视频，但是会有大量的看帖、发帖和回贴活动；图 5-14(d) 显示了典型的低学习参与度特征，该学生既很少在线学习，上线后也很少观看视频。进一步观察更多数据显示，这 4 种模式在较长时间范围内以学期为单位周期性出现，即学生的学习时间管理风格是比较稳定的，但是可能会随着周期数的增加而逐渐偏向松懈。

领域专家表示上述可视化效果与学生长期的学习过程一致，反映了学生在长期的学习过程中学习习惯的连贯性，通过这些可视化设计能够反映学生在单学期与多学期时间管理风格，从而为制定课程计划与教学进度安排提供参考信息。

5.5.3 生源分布与视频利用率模式

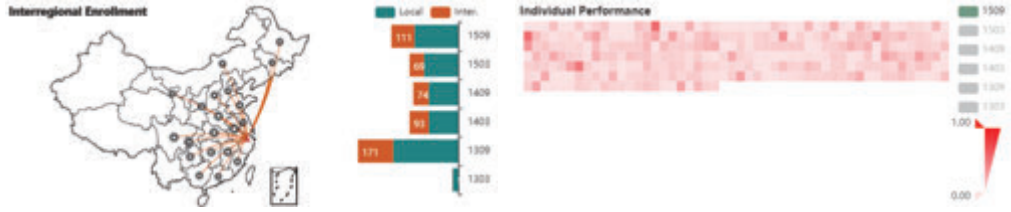
利用本章提出的详情视图中的学习中心详情面板，结合其他关联视图，可以查看不同学习中心的生源分布于视频利用率模式，从而为领域专家分析学习中心招生、生源区域以及学习参与度之间的关系提供参考。图 5-15 显示了 3 个具有代表性的学习中

心视频利用模式。



(a) 西部省份某学习中心的学生来源与到课率分布

(b) 中部省份某学习中心的学生来源与到课率分布



(c) 东部省份某学习中心的学生来源与到课率分布

图 5-15 不同区域的学习中心呈现差异化的招生和视频利用模式

图 5-15(a) 展示了西部省份一个学习中心，其外地生源远少于本地生源，同时学生的视频利用率差异显著；图 5-15(b) 展示了中部省份的一个学习中心，其外地生源远多于本地生源，但是学生的视频利用率很低且没有明显差异；图 5-15(c) 展示了一个东部省份的学习中心，虽然外地生源数量较多，但本地生源仍占多数，学生整体的视频利用率明显高于中部地区但略低于西部省份。

由于学习中心的数量较少，领域专家逐个查看了全部学习中心的到课率分布。从整体上，领域专家认为大部分的可视化效果符合学习中心招生和运营的真实情况，即存在部分学习中心由于招生方面存在问题造成所辖学生的学习参与度较低。而本文提出的可视化设计则直观的将学生的学习参与度与学习中心的招生生源地信息进行关联，进一步增加了对学习中心招生情况的评估手段。根据领域专家的分析，各学习中心的招生生源地原则上应为学习中心所在地区，但是由于不同区域的经济社会发展水平、工作机会、教育成本等方面的差异，使得各地区的人口流动性存在差异进而造成学习中心的招生来源地存在显著差异。例如，经济较发达、工作机会多的地区可能吸引经济欠发达地区的劳动力流入，同时学习相同专业知识的费用可能各不相同。受这些因素的共同影响，学习中心的学生生源地分布会产生差异。从图 5-15展示的可视化效果可以看出，某个学习中心过高的外地生源比例和整体偏低的到课率反映了这个学习中心可能存在招生问题。基于上述分析，领域专家认为直观呈现生源比例差异与学生视频利用情况可能为进一步评估学习中心的招生与运营中的问题提供了新的依据和参考。

5.6 本章小结

本章针对在线学习行为数据中包含的视频学习行为，提出了一种基于球面布局的多属性数据可视化设计，并基于该可视化设计开发了反映课程与学生视频利用情况分布的 VUSphere 可视分析系统，帮助领域专家从不同尺度探索和分析视频利用情况，及其在课程、专业、学习中心以及学生个人等多个方面的分布。

首先，本章针对分析任务概括了可视分析系统的设计原则，并根据在线学习行为数据中视频观看行为的特点，提出了到课率和利用率指标用来衡量学生对课程视频资源的利用情况，并将其拓展到视频与学生的其他属性，为跨属性的视频利用情况对比和统计提供了基础；然后，基于到课率与利用率，本章提出了一种基于球面布局的多属性展示方法，将视频观看行为的学生与视频按照其属性映射到三维球体表面坐标，从不同视角展示整体和局部的视频利用模式，同时配合协调关联的多个视图，展示各关联属性在不同侧面的视频利用率统计结果；最后，本章在大规模真实数据集上展示了该设计的有效性，通过分析三个不同尺度的案例探索了多方面的视频利用模式。从案例分析的结果看，通过基于球面布局的概览视图与固定参考点导航工具，能够有效的发现全部视频和学生的视频利用分布模式。并且通过结合详情视图，可以进一步探索学生的长期学习时间管理模式以及学习中心的生源分布模式。

以本章研究内容为基础撰写的论文发表在国际会议 IEEE Conference on Visual Analytics Science and Technology (VAST 2018，CCF A 类会议)，题为“VUSphere: Visual Analysis of Video Utilization in Online Distance Education”。

6 结论与展望

本章首先总结本文的研究内容和创新点，然后讨论在线学习行为数据可视分析在未来值得探索的改进方向和面临的全新挑战。

6.1 研究工作总结

本文首先针对在线学习行为数据的特点，分析了数据分析工作面临的挑战。然后分别从高维学习参与度可视分析、时间管理风格可视分析、大规模多属性关联的视频利用模式可视分析这三个方面展开研究，并开发了 LearnerExp 和 VUSphere 可视分析系统。主要的创新点如下：

1）针对在线学习行为数据的高维度特点，提出了一个基于散点图矩阵和 t-SNE 算法的学习参与度降维可视化分析方法，设计了多种数据群组创建工具，并将其集成在协调关联的多个视图中，让用户可以分析学习过程数据中高维学习参与度的分布与模式。首先，本文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过程数据的采集、清洗、抽象和分层存储，同时从中提取了反映学生学习参与度的指标并建立索引；其次，本文针对高维的学习参与度数据难以分析和解释的特点，提出了基于 t-SNE 算法的降维可视化分析方法，并设计多个关联视图与交互式的群组创建工具；最后，为了验证方法的有效性，本文利用该系统对网络学习者的在线学习日志数据进行研究，发现了多种不同的学习参与度模式，并分析了在线学生支持服务使用情况与学习参与度之间的关联关系。以该研究内容为基础撰写的论文发表在国际期刊 IEEE Access，题为“How Learner Support Services Affect Student Engagement in Online Learning Environments”，国际会议 IEEE International Conference on Advanced Learning Technologies (ICALT 2018)，题为“Measuring Student’s Utilization of Video Resources and its Effect on Academic Performance”，以及国际会议 IEEE International Conference on e-Business Engineering (ICEBE 2014)，题为“Big Log Analysis for e-Learning Ecosystem”。

2）在前一研究中学习参与度指标模型的基础上，进一步提出了学习参与度时序矩阵以保留学习过程的时序特征，并采用基于卷积神经网络的模型预测不同时间管理风格的考试成绩。随后，本文设计了基于日历坐标系的学习参与度日历矩阵和基于日历图的学习时间分布图，展现长时间学习过程中的时间信息与周期变化情况。最后，本文针对在线多课程同期学习的特点，设计了多课程学习进度图，直观呈现从课程资源和时间两个维度细粒度分析学习过程。综合上述可视化设计与交互技术，本文开发了 LearnerExp 可视化分析系统，探索长期学习过程中的时间管理风格。以该研究内容为基础撰写的论文发表在国际会议 ACM Global Computing Education Conference (CompEd 2019)，题为“VUC: Visualizing Daily Video Utilization to Promote Student Engagement in Online Distance Education”，以及国际会议 IEEE Conference on Visual Analytics Science and Technology Short Paper Track(VAST 2019)，题为“Visual Analysis of the Time Manage-

ment of Learning Multiple Courses in Online Learning Environment”；以本章研究内容为基础开发的系统，发表在国际会议 The Web Conference 2019 Demonstration Track (WWW 2019)，题为“LearnerExp: Exploring and Explaining the Time Management of Online Learning Activity”。

3）提出了多尺度的视频利用率指标用以衡量学生和视频资源双方在学习过程中的表现，并提出一种基于球面布局的多属性数据可视化设计同时基于该可视化设计开发了一个反映课程与学生视频利用情况分布的可视分析系统 VUSphere，帮助领域专家从不同尺度探索和分析视频利用情况及其在课程、专业、学习中心以及学生个人等多个方面的分布。首先，提出了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，并将其拓展到视频、学生的其他属性，为跨属性的视频利用情况对比和统计提供了基础。然后，提出了一种基于球面布局的视频利用情况展示，将视频与学生按照属性映射到三维球体表面，从不同视角展示整体和局部的视频利用模式，同时配合协调关联的多个视图，展示各关联属性在不同侧面的视频利用率统计结果。以该研究内容为基础撰写的论文发表在国际会议 IEEE Conference on Visual Analytics Science and Technology (VAST 2018，CCF A 类会议)，题为“VUSphere: Visual Analysis of Video Utilization in Online Distance Education”。

6.2 下一步工作展望

本文针对在线学习行为数据呈现的特征维度高、时间周期长、大规模多属性关联的特点，提出了多种可视分析方法探索数据中潜在的学习行为模式和规律。尽管本文在上述方面均取得了一定的成果，然而进一步深化在线学习行为数据的探索分析仍有许多工作和可改进的空间，主要包括以下几方面：

1）目前的学习参与度指标主要以行为交互相关的指标为主，没有对学习内容、作业、测验以及其他教学评估相关指标作研究。而这些内容蕴含了更多的信息，反映了学生对知识的深层次掌握情况，对于理解学生的学习状态和提升学习收益有重要作用。所以，未来将针对这些内容，利用可视分析方法在已有工作的基础上展开研究。

2）现有学习过程的时序模式主要来自于学习结束之后的统计分析，然而由于在线学习的长期性，可能需要在早期利用有限的时序数据进行分析和预测，提前实施教学干预，避免松懈、拖延等不利的时间管理风格对学习成效造成负面影响。因此，未来的研究需要在现有工作基础上，关注有限时序数据的时间管理风格识别。

3）目前 LearnerExp 与 VUSphere 可视分析系统只作为一个独立系统面向教学管理人员使用，在应用层面仍未与生产环境的其他系统集成，如教学管理系统、学习管理系统和学生服务系统等。这使得可视分析系统中发现的规律和模式，包括自动提取的模式与领域专家发现的模式，都无法直接应用于教学实践。因此，未来应进一步将可视分析系统以适当的形式与业务系统融合，并将利用可视分析系统得到的规律和数据应用到远程教育的不同层面。

致　谢

梧桐七繁七落，樱花七开七谢，迎新人辞旧友，如今也换己作别。回首这七年平凡又独特的时光，只觉得岁月匆匆，跨过而立，已近不惑，虽小有所得，但多是窥见日月苍穹，深觉所知甚少。向前行，虽不是波涛汹涌艰险难行，却也不是轻松自如一帆风顺，若孜身一人定然无法抵达这里。值此博士论文即将付梓之际，感谢在我博士旅途中引导我、帮助我、陪伴我的老师、朋友、同学和家人们，与你们的相遇相随都将成为我此生中最美好的回忆。

衷心感谢我的导师郑庆华教授。在这七年间，郑老师以高瞻远瞩的学术眼光为我指引了科学研究的方向与道路，以脚踏实地的务实作风为我提供了非常优越、团结和舒适的科研环境。郑老师不仅传授给我知识与科研方法，更重要的是通过一言一行以身作则的将交大人精勤、敦笃、果毅、忠恕的精神品格传承给我们，这不仅令我在学术研究和为人处事上受益匪浅，更是激励我不断向前的重要精神财富。值此博士毕业之际，特向郑老师致以我衷心的感谢和崇高的敬意。

非常感谢董博师兄一直以来的支持和关照。在这七年里，一起修改论文、指导学生、撰写材料、攻关项目、争取资源、出差北京、上海、杭州、柏林，一起讨论科研、工作以及生活等方方面面的问题。取得每一个工作成果的背后，都有董博师兄在各方面的帮助和指导，当我在课题研究和项目工作中遇到困难时，总是能帮我仔细分析，给我很多中肯的建议和指导。感谢实验室团队各位老师在我博士期间科研、工作、生活等方面给予的帮助，感谢田锋教授、张未展教授、李辰教授、刘均教授、陈丽莉老师、陈灵老师、钱步月老师、魏笔凡老师、朱海萍老师、罗敏楠老师、屈宇老师、郭强老师、刘峰老师、褚亮老师、王昱老师、王立军老师、李志兴老师。感谢众多校内校外合作单位专家的支持与指导，感谢韩博老师、李一鸣老师、锁志海老师、杨帆老师、刘红磊老师、张亚娟老师、杜丰老师、杨洁老师、张宁老师、谭薇老师、程洁老师以及杨源杰博士、夏梦博士、曾柯博士、胡文静博士、周杭谕女士、吉烨女士以及殷梦实女士。

感谢实验室同学们对我科研和生活上的所有帮助，感谢李睿师兄、杜海鹏师兄、骞雅楠师姐、张杰师姐、陈艳平师兄、赵辉师兄、阮建飞、马天、宋凯磊、薛妮、徐海鹏、钟阿敏、蔚文达、孟玮、李佳、李冰、张莎、于洪潮、林雅婷、李珍艳、陈浩、吴凡、乐佳、王萌、刘文强、宋凌云及其他帮助过我的同学。

衷心感谢我的父母一直以来无私的引导和关爱，让我在求学探索的道路上能够安心前行心无旁骛。感谢我的爱人，携手相伴，风雨同行。

最后，再次感谢所有帮助过我的人们，祝福你们身体健康阖家幸福。

参考文献

[1] Seaton D T, Bergner Y, Chuang I, et al. Who does what in a massive open online course?[J]. Communications of the ACM. April 2014, 57(4):58–65.

[2] Isaac C, Andrew H. HarvardX and MITx: Four Years of Open Online Courses – Fall 2012-Summer 2016[M/OL]. 2016. http://www.ssrn.com/abstract=2889436.

[3] Waldrop M M. Online learning: Campus 2.0[J]. Nature. March 2013, 495(7440):160.

[4] Means B, Anderson K. Expanding Evidence Approaches for Learning in a Digital World[M]. Washington DC, USA: Office of Educational Technology, US Department of Education, February 2013.

[5] Uchidiuno J, Hammer J, Yarzebinski E, et al. Characterizing ELL Students' Behavior During MOOC Videos Using Content Type[C]. Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale. New York, NY, USA: ACM, 2017: 185–188.

[6] 蒋卓轩, 张岩, 李晓明. 基于 MOOC 数据的学习行为分析与预测 [J]. 计算机研究与发展. 2015, 52(3):614–628.

[7] Baker R S J D, Yacef K. The State of Educational Data Mining in 2009: A Review and Future Visions[J]. Journal of Educational Data Mining. 2009, 1(1):3–16.

[8] Zhu H, Tian F, Wu K, et al. A multi-constraint learning path recommendation algorithm based on knowledge map[J]. Knowledge-Based Systems. March 2018, 143:102–114.

[9] Kellogg S. Online learning: How to make a MOOC[J]. Nature. July 2013, 499(7458):369–371.

[10] McAndrew P, Scanlon E. Open Learning at a Distance: Lessons for Struggling MOOCs[J]. Science. December 2013, 342(6165):1450–1451.

[11] Guo P J, Kim J, Rubin R. How video production affects student engagement: an empirical study of MOOC videos[C]. Proceedings of the First (2014) ACM Conference on Learning @ Scale. New York, NY, USA: ACM Press, 2014: 41–50.

[12] Kim J, Guo P J, Seaton D T, et al. Understanding in-video dropouts and interaction peaks in online lecture videos[C]. Proceedings of the First (2014) ACM Conference on Learning @ Scale. New York, NY, USA: ACM, 2014: 31–40.

[13] Sunar A S, White S, Abdullah N A, et al. How Learners' Interactions Sustain Engagement: A MOOC Case Study[J]. IEEE Transactions on Learning Technologies. October 2017, 10(4):475–487.

[14] Lu Y, Warren J, Jermaine C, et al. Grading the Graders: Motivating Peer Graders in a MOOC[C]. Proceedings of the 24th International Conference on World Wide Web. Republic and Canton of Geneva, Switzerland: ACM Press, 2015: 680–690.

[15] Staubitz T, Petrick D, Bauer M, et al. Improving the Peer Assessment Experience on MOOC Platforms[C]. Proceedings of the Third (2016) ACM Conference on Learning @ Scale. New York, NY, USA: ACM, 2016: 389–398.

[16] Breslow L, Pritchard D E, DeBoer J, et al. Studying Learning in the Worldwide Classroom Research into edX's First MOOC[J]. Research & Practice in Assessment. 2013, 8:13–25.

[17] Waldrop M M. Education online: The virtual lab[J]. Nature. July 2013, 499(7458):268.

[18] Shen H. Interactive notebooks: Sharing the code[J]. Nature. November 2014, 515(7525):151.

[19] O'Hara K, Blank D, Marshall J. Computational Notebooks for AI Education[C]. The 28th International Flairs Conference. New York, NY, USA: AIED, April 2015.

[20] Kery M B, Radensky M, Arya M, et al. The Story in the Notebook: Exploratory Data Science Using a Literate Programming Tool[C]. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, 2018: 174:1–174:11.
[21] Peña-Ayala A. Educational data mining: A survey and a data mining-based analysis of recent works[J]. Expert Systems with Applications. March 2014, 41(4, Part 1):1432–1462.
[22] Thomas a J. Visual Analytics[J]. IEEE Computer Graphics and Applications. September 2004, 24 (5):20–21.
[23] 任磊, 杜一, 马帅等. 大数据可视分析综述 [J]. 软件学报. 2014, 25(09):1909–1936.
[24] Thomas J J, Cook K A. A visual analytics agenda[J]. IEEE Computer Graphics and Applications. January 2006, 26(1):10–13.
[25] Wijk J J v. The value of visualization[C]. VIS 05. IEEE Visualization, 2005. New York, NY, USA: IEEE, October 2005: 79–86.
[26] 梅鸿辉, 陈海东, 肇昕等. 一种全球尺度三维大气数据可视化系统 [J]. 软件学报. 2016, 27(05): 1140–1150.
[27] 魏敏, 王松, 吴亚东. 医学图像可视化的视觉优化方法 [J]. 计算机辅助设计与图形学学报. 2019, 31(04):659–667.
[28] 贾若雨, 曾昂, 朱敏等. 面向在线交易日志的用户购买行为可视化分析 [J]. 软件学报. 2017, 28 (09):2450–2467.
[29] Wu Y, Lan J, Shu X, et al. iTTVis: Interactive Visualization of Table Tennis Data[J]. IEEE Transactions on Visualization and Computer Graphics. January 2018, 24(1):709–718.
[30] Sun G, Wu Y, Liu S, et al. EvoRiver: Visual Analysis of Topic Coopetition on Social Media[J]. IEEE Transactions on Visualization and Computer Graphics. December 2014, 20(12):1753–1762.
[31] 黄文达, 陶煜波, 屈珂等. 基于 OD 数据的群体行为可视分析 [J]. 计算机辅助设计与图形学学报. 2018, 30(06):1023–1033.
[32] Bertini E, Lalanne D. Surveying the Complementary Role of Automatic Data Analysis and Visualization in Knowledge Discovery[C]. Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration. New York, NY, USA: ACM, 2009: 12–20.
[33] Chabot C. Demystifying Visual Analytics[J]. IEEE Computer Graphics and Applications. March 2009, 29(2):84–87.
[34] Chang R, Ziemkiewicz C, Green T M, et al. Defining Insight for Visual Analytics[J]. IEEE Computer Graphics and Applications. March 2009, 29(2):14–17.
[35] Keim D, Kohlhammer J, Ellis G, et al. Mastering the Information Age Solving Problems with Visual Analytics[M]. Goslar, Germany: Eurographics Association, 2010.
[36] Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations[C]. Proceedings 1996 IEEE Symposium on Visual Languages. New York, NY, USA: IEEE, September 1996: 336–343.
[37] Yi J S, Kang Y a, Stasko J. Toward a Deeper Understanding of the Role of Interaction in Information Visualization[J]. IEEE Transactions on Visualization and Computer Graphics. November 2007, 13 (6):1224–1231.
[38] Fekete J D, Wijk J J v, Stasko J T, et al. The Value of Information Visualization[C]. IEEE Symposium on Information Visualization, 2008. InfoVis 2008. New York, NY, USA: IEEE, 2008: 1–18.

[39] Keim D, Andrienko G, Fekete J D, et al. Lecture Notes in Computer Science Visual Analytics: Definition, Process, and Challenges[M]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 154–175.

[40] Liu S, Maljovec D, Wang B, et al. Visualizing High-Dimensional Data: Advances in the Past Decade[J]. IEEE Transactions on Visualization and Computer Graphics. March 2017, 23(3):1249–1268.

[41] Sorzano C O S, Vargas J, Montano A P. A survey of dimensionality reduction techniques[J]. arXiv:1403.2877 [cs, q-bio, stat]. March 2014.

[42] Cockburn A, Karlson A, Bederson B B. A Review of Overview+Detail, Zooming, and Focus+Context Interfaces[J]. ACM Comput. Surv. January 2009, 41(1):2:1–2:31.

[43] Zhao X, Wu Y, Cui W, et al. SkyLens: Visual Analysis of Skyline on Multi-dimensional Data[J]. IEEE Transactions on Visualization and Computer Graphics. 2017, PP(99):1–10.

[44] Lee J H, McDonnell K T, Zelenyuk A, et al. A Structure-Based Distance Metric for High-Dimensional Space Exploration with Multidimensional Scaling[J]. IEEE Transactions on Visualization and Computer Graphics. March 2014, 20(3):351–364.

[45] Wei J, Shen Z, Sundaresan N, et al. Visual cluster exploration of web clickstream data[C]. 2012 IEEE Conference on Visual Analytics Science and Technology (VAST). New York, NY, USA: IEEE, October 2012: 3–12.

[46] Sucontphunt T, Yuan X, Li Q, et al. A Novel Visualization System for Expressive Facial Motion Data Exploration[C]. 2008 IEEE Pacific Visualization Symposium. New York, NY, USA: IEEE, March 2008: 103–109.

[47] Li L Y, Tsai C C. Accessing online learning material: Quantitative behavior patterns and their effects on motivation and learning performance[J]. Computers & Education. November 2017, 114:286–297.

[48] Cerezo R, Sánchez-Santillán M, Paule-Ruiz M P, et al. Students' LMS interaction patterns and their relationship with achievement: A case study in higher education[J]. Computers & Education. 2016, 96:42–54.

[49] 陈为, 沈则潜. 数据可视化 [M]. 北京, 中国: 电子工业出版社, 2013.

[50] Heer J, Kong N, Agrawala M. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations[C]. In Proc. ACM Human Factors in Computing Systems (CHI. New York, NY, USA: ACM, 2009: 1303–1312.

[51] Shen Z, Ma K. MobiVis: A Visualization System for Exploring Mobile Data[C]. 2008 IEEE Pacific Visualization Symposium. New York, NY, USA: IEEE, March 2008: 175–182.

[52] Woodring J, Shen H. Multiscale Time Activity Data Exploration via Temporal Clustering Visualization Spreadsheet[J]. IEEE Transactions on Visualization and Computer Graphics. January 2009, 15 (1):123–137.

[53] Qu H, Chen Q. Visual Analytics for MOOC Data[J]. IEEE Computer Graphics and Applications. November 2015, 35(6):69–75.

[54] Zhao J, Bhatt C, Cooper M, et al. Flexible Learning with Semantic Visual Exploration and Sequence-Based Recommendation of MOOC Videos[C]. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, 2018: 329:1–329:13.

[55] Chen Q, Yue X, Plantaz X, et al. ViSeq: Visual Analytics of Learning Sequence in Massive Open Online Courses[J]. IEEE Transactions on Visualization and Computer Graphics. 2018, 1–1.

[56] Fu S, Wang Y, Yang Y, et al. VisForum: A Visual Analysis System for Exploring User Groups in Online Forums[J]. ACM Trans. Interact. Intell. Syst. February 2018, 8(1):3:1–3:21.

[57] Meng X, Mingfei S, Huan W, et al. PeerLens: Peer-inspired Interactive Learning Path Planning in Online Question Pool[C]. ACM CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, 2019: 51–60.
[58] Munzner T. A Nested Model for Visualization Design and Validation[J]. IEEE Transactions on Visualization and Computer Graphics. November 2009, 15(6):921–928.
[59] 陈谊, 张聪. 一种基于维度投影的多维数据相关性可视分析方法 [J]. 计算机辅助设计与图形学学报. 2018, 30(04):592–601.
[60] 夏佳志, 张亚伟, 张健等. 一种基于子空间聚类的局部相关性可视分析方法 [J]. 计算机辅助设计与图形学学报. 2016, 28(11):1855–1862.
[61] Hohman F M, Kahng M, Pienta R, et al. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers[J]. IEEE Transactions on Visualization and Computer Graphics. 2018, 1–10.
[62] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research. 2003, 3(7-8):1157–1182.
[63] Lee C, Lee G G. Information gain and divergence-based feature selection for machine learning-based text categorization[J]. Information Processing & Management. January 2006, 42(1):155–165.
[64] Kohavi R, John G H. Wrappers for feature subset selection[J]. Artificial Intelligence. December 1997, 97(1):273–324.
[65] Joksimović S, Gašević D, Loughin T M, et al. Learning at distance: Effects of interaction traces on academic achievement[J]. Computers & Education. September 2015, 87:204–217.
[66] Jolliffe I T. Springer Series in Statistics Principal Component Analysis[M]. 2nd ed. New York: Springer-Verlag, 2002.
[67] Maaten L v d, Hinton G. Visualizing Data using t-SNE[J]. Journal of Machine Learning Research. 2008, 9(Nov):2579–2605.
[68] Maaten L v d. Accelerating t-SNE using Tree-Based Algorithms[J]. Journal of Machine Learning Research. 2014, 15:3221–3245.
[69] Rauber P E, Falcão A X, Telea A C. Visualizing Time-Dependent Data Using Dynamic t-SNE[C]. Eurographics Conference on Visualization (EuroVis) 2016. New York, NY, USA: IEEE, 2016.
[70] Linderman G C, Rachh M, Hoskins J G, et al. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data[J]. Nature Methods. March 2019, 16(3):243.
[71] Tatu A, Maaß F, Färber I, et al. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data[C]. 2012 IEEE Conference on Visual Analytics Science and Technology (VAST). New York, NY, USA: IEEE, October 2012: 63–72.
[72] Huang Z, Lu Y, Mack E, et al. Exploring the Sensitivity of Choropleths under Attribute Uncertainty[J]. IEEE Transactions on Visualization and Computer Graphics. 2019, 1–1.
[73] Heimerl F, John M, Han Q, et al. DocuCompass: Effective exploration of document landscapes[C]. 2016 IEEE Conference on Visual Analytics Science and Technology (VAST). New York, NY, USA: IEEE, October 2016: 11–20.
[74] Shen Z, Wei J, Sundaresan N, et al. Visual analysis of massive web session data[C]. IEEE Symposium on Large Data Analysis and Visualization (LDAV). New York, NY, USA: IEEE, October 2012: 65–72.
[75] Kabulov B T, Tashpulatova N B. Enhanced Chernoff faces[C]. 2010 4th International Conference on Application of Information and Communication Technologies. Berlin, Germany: Springer, October 2010: 1–4.

[76] Dang T N, Anand A, Wilkinson L. TimeSeer: Scagnostics for High-Dimensional Time Series[J]. IEEE Transactions on Visualization and Computer Graphics. March 2013, 19(3):470–483.
[77] Wu T, Yao Y, Duan Y, et al. NetworkSeer: Visual analysis for social network in MOOCs[C]. 2016 IEEE Pacific Visualization Symposium (PacificVis). New York, NY, USA: IEEE, April 2016: 194–198.
[78] Albo Y, Lanir J, Bak P, et al. Off the Radar: Comparative Evaluation of Radial Visualization Solutions for Composite Indicators[J]. IEEE Transactions on Visualization and Computer Graphics. January 2016, 22(1):569–578.
[79] Palmas G, Bachynskyi M, Oulasvirta A, et al. An Edge-Bundling Layout for Interactive Parallel Coordinates[C]. 2014 IEEE Pacific Visualization Symposium. New York, NY, USA: IEEE, March 2014: 57–64.
[80] Holten D. Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data[J]. IEEE Transactions on Visualization and Computer Graphics. September 2006, 12(5):741–748.
[81] Aigner W, Miksch S, Müller W, et al. Visualizing time-oriented data—A systematic view[J]. Computers & Graphics. June 2007, 31(3):401–409.
[82] Aigner W, Miksch S, Müller W, et al. Visual Methods for Analyzing Time-Oriented Data[J]. IEEE Transactions on Visualization and Computer Graphics. January 2008, 14(1):47–60.
[83] Sun G D, Wu Y C, Liang R H, et al. A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges[J]. Journal of Computer Science and Technology. September 2013, 28(5):852–867.
[84] Daniel Rosenberg. Joseph Priestley and the Graphic Invention of Modern Time[J]. Studies Inghteenth Century Culture. 2007, 36(1):55–103.
[85] Javed W, McDonnel B, Elmqvist N. Graphical Perception of Multiple Time Series[J]. IEEE Transactions on Visualization and Computer Graphics. November 2010, 16(6):927–934.
[86] Tamara Munzner. Visualization Analysis & Design[M]. Boca Raton: Taylor & Francis Group, 2014.
[87] Liu S, Zhou M X, Pan S, et al. TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis[J]. ACM Trans. Intell. Syst. Technol. February 2012, 3(2):25:1–25:28.
[88] Luo X, Wang H, Tian F, et al. ProcessLine: Visualizing time-series data in process industry[C]. 2009 IEEE Symposium on Visual Analytics Science and Technology. New York, NY, USA: IEEE, October 2009: 231–232.
[89] Xie C, Chen W, Huang X, et al. VAET: A Visual Analytics Approach for E-Transactions Time-Series[J]. IEEE Transactions on Visualization and Computer Graphics. December 2014, 20(12): 1743–1752.
[90] Rind A, Lammarsch T, Aigner W, et al. TimeBench: A Data Model and Software Library for Visual Analytics of Time-Oriented Data[J]. IEEE Transactions on Visualization and Computer Graphics. December 2013, 19(12):2247–2256.
[91] Krstajic M, Bertini E, Keim D. CloudLines: Compact Display of Event Episodes in Multiple Time-Series[J]. IEEE Transactions on Visualization and Computer Graphics. December 2011, 17(12): 2432–2439.
[92] 赵颖, 王权, 黄叶子等. 多视图合作的网络流量时序数据可视分析 [J]. 软件学报. 2016, 27(05): 1188–1198.
[93] Holtz S, Valle G, Howard J, et al. Visualization and pattern identification in large scale time series data[C]. 2011 IEEE Symposium on Large Data Analysis and Visualization. New York, NY, USA: IEEE, October 2011: 129–130.

[94] Carlis J V, Konstan J A. Interactive Visualization of Serial Periodic Data[C]. Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology. New York, NY, USA: ACM, 1998: 29–38.

[95] Weber M, Alexa M, Muller W. Visualizing time-series on spirals[C]. IEEE Symposium on Information Visualization, 2001. InfoVis 2001. New York, NY, USA: IEEE, October 2001: 7–13.

[96] Lin J, Keogh E, Lonardi S. Visualizing and Discovering Non-Trivial Patterns in Large Time Series Databases[J]. Information Visualization. June 2005, 4(2):61–82.

[97] Walker J, Borgo R, Jones M W. TimeNotes: A Study on Effective Chart Visualization and Interaction Techniques for Time-Series Data[J]. IEEE Transactions on Visualization and Computer Graphics. January 2016, 22(1):549–558.

[98] 金思辰, 陶煜波, 严宇宇等. 基于多维时空数据可视化的传染病模式分析 [J]. 计算机辅助设计与图形学学报. 2019, 31(02):241–255.

[99] Wijk J J V, Selow E R V. Cluster and calendar based visualization of time series data[C]. Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99). New York, NY, USA: IEEE, October 1999: 4–9.

[100] Bederson B B, Clamage A, Czerwinski M P, et al. DateLens: A Fisheye Calendar Interface for PDAs[J]. ACM Trans. Comput.-Hum. Interact. March 2004, 11(1):90–119.

[101] McNutt M. Bricks and MOOCs[J]. Science. October 2013, 342(6157):402–402.

[102] Reich J. Rebooting MOOC Research[J]. Science. January 2015, 347(6217):34–35.

[103] Dernoncourt F, Taylor C, Veeramachaneni K, et al. MoocViz: A Large Scale, Open Access, Collaborative, Data Analytics Platform for MOOCs[C]. NIPS 2013 Education Workshop. Cambridge, Massachusetts, USA: MIT Press, November 2013: 159–166.

[104] Vieira C, Parsons P, Byrd V. Visual learning analytics of educational data: A systematic literature review and research agenda[J]. Computers & Education. July 2018, 122:119–135.

[105] Romero C, Ventura S. Educational data mining: A survey from 1995 to 2005[J]. Expert Systems with Applications. July 2007, 33(1):135–146.

[106] Papamitsiou Z, Economides A A. Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence[J]. Journal of Educational Technology & Society. 2014, 17(4):49–64.

[107] Dutt A, Ismail M A, Herawan T. A Systematic Review on Educational Data Mining[J]. IEEE Access. 2017, 5:15991–16005.

[108] Phan T, McNeil S G, Robin B R. Students' Patterns of Engagement and Course Performance in a Massive Open Online Course[J]. Computers & Education. December 2015, 29–42.

[109] Kahan T, Soffer T, Nachmias R. Types of Participant Behavior in a Massive Open Online Course[J]. The International Review of Research in Open and Distributed Learning. September 2017, 18(6).

[110] Howard S K, Ma J, Yang J. Student rules: Exploring patterns of students' computer-efficacy and engagement with digital technologies in learning[J]. Computers & Education. 2016, 101:29–42.

[111] Henrie C R, Bodily R, Larsen R, et al. Exploring the potential of LMS log data as a proxy measure of student engagement[J]. Journal of Computing in Higher Education. August 2018, 30(2):344–362.

[112] Knobloch J, Kaltenbach J, Bruegge B. Increasing Student Engagement in Higher Education Using a Context-aware Q&A Teaching Framework[C]. Proceedings of the 40th International Conference on Software Engineering: Software Engineering Education and Training. New York, NY, USA: ACM, 2018: 136–145.

[113] Mahzoon M J, Maher M L, Eltayeby O, et al. A Sequence Data Model for Analyzing Temporal Patterns of Student Data[J]. Journal of Learning Analytics. April 2018, 5(1):55–74.

[114] Kalyan V, Franck D, Colin T, et al. MOOCdb: Developing Data Standards for MOOC Data Science[C]. 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings. New York, NY, USA: AIED, 2013: 17.

[115] Van der Sluis F, Ginn J, Van der Zee T. Explaining Student Behavior at Scale: The Influence of Video Complexity on Student Dwelling Time[C]. Proceedings of the Third (2016) ACM Conference on Learning @ Scale. New York, NY, USA: ACM, 2016: 51–60.

[116] Coffrin C, Corrin L, de Barba P, et al. Visualizing Patterns of Student Engagement and Performance in MOOCs[C]. Proceedings of the Fourth International Conference on Learning Analytics And Knowledge. New York, NY, USA: ACM, 2014: 83–92.

[117] Auvinen T, Hakulinen L, Malmi L. Increasing Students' Awareness of Their Behavior in Online Learning Environments with Visualizations and Achievement Badges[J]. IEEE Transactions on Learning Technologies. July 2015, 8(3):261–273.

[118] Shi C, Fu S, Chen Q, et al. VisMOOC: Visualizing video clickstream data from Massive Open Online Courses[C]. 2015 IEEE Pacific Visualization Symposium (PacificVis). New York, NY, USA: IEEE, April 2015: 159–166.

[119] Chen Q, Chen Y, Liu D, et al. PeakVizor: Visual Analytics of Peaks in Video Clickstreams from Massive Open Online Courses[J]. IEEE Transactions on Visualization and Computer Graphics. October 2016, 22(10):2315–2330.

[120] Cooper M, Zhao J, Bhatt C, et al. MOOCex: Exploring Educational Video via Recommendation[C]. Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. New York, NY, USA: ACM, 2018: 521–524.

[121] Chen C M, Wu C H. Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance[J]. Computers & Education. 2015, 80:108–121.

[122] Sung C Y, Huang X Y, Shen Y, et al. Exploring Online Learners' Interactive Dynamics by Visually Analyzing Their Time-anchored Comments[J]. Computer Graphics Forum. October 2017, 36(7): 145–155.

[123] Fu S, Zhao J, Cui W, et al. Visual Analysis of MOOC Forums with iForum[J]. IEEE Transactions on Visualization and Computer Graphics. January 2017, 23(1):201–210.

[124] Wong J S, Zhang X L. MessageLens: A Visual Analytics System to Support Multifaceted Exploration of MOOC Forum Discussions[J]. Visual Informatics. March 2018, 2(1):37–49.

[125] Atapattu T, Falkner K, Tarmazdi H. Topic-Wise Classification of MOOC Discussions: A Visual Analytics Approach[C]. Proceedings of the 9th International Educational Data Mining Society. New York, NY, USA: EDM, 2016.

[126] Chen Y, Chen Q, Zhao M, et al. DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction[C]. 2016 IEEE Conference on Visual Analytics Science and Technology (VAST). New York, NY, USA: IEEE, October 2016: 111–120.

[127] Pardos Z A, Kao K. moocRP: An Open-source Analytics Platform[C]. Proceedings of the Second (2015) ACM Conference on Learning @ Scale. New York, NY, USA: ACM Press, 2015: 103–110.

[128] 纪连恩, 高芳, 黄凯鸿等. 面向多主体的大学课程成绩相关性可视探索与分析 [J]. 计算机辅助设计与图形学学报. 2018, 30(01):44–56.

[129] Kizilcec R F, Saltarelli A J, Reich J, et al. Closing global achievement gaps in MOOCs[J]. Science. January 2017, 355(6322):251–252.

[130] Bote-Lorenzo M L, Gómez-Sánchez E. Predicting the Decrease of Engagement Indicators in a MOOC[C]. Proceedings of the Seventh International Learning Analytics & Knowledge Conference. New York, NY, USA: ACM, 2017: 143–147.

[131] Stamper J, Koedinger K, Baker R S J d, et al. PSLC DataShop: A Data Analysis Service for the Learning Science Community[C]// Aleven V, Kay J, Mostow J. Intelligent Tutoring Systems. Berlin, Germany: Springer Berlin Heidelberg, 2010: 455–455.

[132] Fenu G, Marras M, Meles M. A Learning Analytics Tool for Usability Assessment in Moodle Environments[J]. Journal of e-Learning and Knowledge Society. September 2017, 13(3).

[133] Lampietti V, Roy A, Barnes S. Managing and Analyzing Student Learning Data: A Python-based Solution for edX[C]. Proceedings of the Fifth Annual ACM Conference on Learning at Scale. New York, NY, USA: ACM, 2018: 48:1–48:2.

[134] John L. Hennessy, David A. Patterson. Computer Architecture: A Quantitative Approach 5th Edition[M]. Amsterdam: Morgan Kaufmann, September 2011.

[135] Hew K F. Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCS[J]. British Journal of Educational Technology. 2015, 320–341.

[136] Fredricks J A, Blumenfeld P C, Paris A H. School Engagement: Potential of the Concept, State of the Evidence[J]. Review of Educational Research. March 2004, 74(1):59–109.

[137] Bradford G, Wyatt S. Online learning and student satisfaction: Academic standing, ethnicity and their influence on facilitated learning, engagement, and information fluency[J]. The Internet and Higher Education. 2010, 13(3):108–114.

[138] Appleton J J, Christenson S L, Furlong M J. Student engagement with school: Critical conceptual and methodological issues of the construct[J]. Psychology in the Schools. May 2008, 45(5):369–386.

[139] Singh V, Padmanabhan B, de Vreede T, et al. A Content Engagement Score for Online Learning Platforms[C]. Proceedings of the Fifth Annual ACM Conference on Learning at Scale. New York, NY, USA: ACM, 2018: 25:1–25:4.

[140] Henrie C R, Halverson L R, Graham C R. Measuring student engagement in technology-mediated learning: A review[J]. Computers & Education. December 2015, 90:36–53.

[141] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research. October 2011, 12:2825–2830.

[142] MOOCs@Edinburgh Group. MOOCs @ Edinburgh 2013 –Report #1[M/OL]. May 2013. https://www.era.lib.ed.ac.uk/handle/1842/6683.

[143] Schulz M, Stamov Roßnagel C. Informal workplace learning: An exploration of age differences in learning competence[J]. Learning and Instruction. 2010, 20(5):383–399.

[144] Gu J, Churchill D, Lu J. Mobile Web 2.0 in the workplace: A case study of employees' informal learning[J]. British Journal of Educational Technology. 2014, 45(6):1049–1059.

[145] Dvorak T, Jia M. Online Work Habits and Academic Performance[J]. Journal of Learning Analytics. December 2016, 3(3):318–330.

[146] Broadbent J, Poon W L. Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review[J]. The Internet and Higher Education. 2015, 27:1–13.

[147] Park J, Yu R, Rodriguez F. Understanding Student Procrastination via Mixture Models[C]. Proceedings of the 11th International Conference on Educational Data Mining. New York, NY, USA: EDM, 2018: 1–11.

[148] You J W. Identifying significant indicators using LMS data to predict course achievement in online learning[J]. The Internet and Higher Education. April 2016, 29:23–30.

[149] Kizilcec R F, Pérez-Sanagustín M, Maldonado J J. Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses[J]. Computers & Education. January 2017, 104:18–33.

[150] Kim D, Yoon M, Jo I H, et al. Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea[J]. Computers & Education. December 2018, 127:233–251.

[151] Kizilcec R F, Pérez-Sanagustín M, Maldonado J J. Recommending Self-Regulated Learning Strategies Does Not Improve Performance in a MOOC[C]. Proceedings of the Third (2016) ACM Conference on Learning @ Scale. New York, NY, USA: ACM, 2016: 101–104.

[152] Mei J. Learning Management System Calendar Reminders and Effects on Time Management and Academic Performance[J]. International Research and Review. 2016, 6(1):29–45.

[153] Ilves K, Leinonen J, Hellas A. Supporting Self-Regulated Learning with Visualizations in Online Learning Environments[C]. Proceedings of the 49th ACM Technical Symposium on Computer Science Education. New York, NY, USA: ACM, 2018: 257–262.

[154] Davis D R, Abbitt J T. An Investigation of the Impact of an Intervention to Reduce Academic Procrastination Using Short Message Service (SMS) Technology[J]. Journal of Interactive Online Learning. 2013, 12(3):78–102.

[155] Kim S, Malik S, Lipka N, et al. WimNet: Vision Search for Web Logs[C]. Proceedings of the 26th International Conference on World Wide Web Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017: 803–804.

[156] Michael Keith, Tom Craver. The Ultimate Perpetual Calendar?[J]. Journal of Recreational Mathematics. 1990, 22(4):280–282.

[157] Dobrian F, Awan A, Joseph D, et al. Understanding the Impact of Video Quality on User Engagement[J]. Commun. ACM. March 2013, 56(3):91–99.

[158] MacHardy Z, Pardos Z A. Toward the Evaluation of Educational Videos using Bayesian Knowledge Tracing and Big Data[C]. Proceedings of the Second (2015) ACM Conference on Learning @ Scale. New York, NY, USA: ACM, 2015: 347–350.

[159] Gardner J, Brooks C. Dropout Model Evaluation in MOOCs[C]. Thirty-Second AAAI Conference on Artificial Intelligence. New York, NY, USA: AAAI, April 2018.

[160] Jos Dirksen. Learning Three.js: The JavaScript 3D Library for WebGL - Second Edition[M]. Birmingham, UK: Packt Publishing, 2015.

攻读学位期间取得的研究成果

本人在攻读博士学位期间取得以下研究成果。其中已录用或已发表的 11 篇论文中，第一作者或导师第一作者 8 篇。已授权国家发明专利第一学生作者 2 项，其余 4 项。

[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，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)

国家发明专利：

[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, 西安交通大学 (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, 西安交通大学 (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, 西安交通大学 (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, 西安交通大学 (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, 西安交通大学 (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, 西安交通大学 (2015-09-30 已授权)

学位论文独创性声明 (1)

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1) 交回学校授予的学位证书；
2) 学校可在相关媒体上对作者本人的行为进行通报；
3) 本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉；
4) 本人负责因论文成果不实产生的法律纠纷。

论文作者 (签名)：　　日期：　　年　　月　　日

学位论文独创性声明 (2)

本人声明：研究生_____ 所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1) 学校可在相关媒体上对本人的失察行为进行通报；
2) 本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉；
3) 本人接受学校按照有关规定做出的任何处理。

指导教师 (签名)：　　日期：　　年　　月　　日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者 (签名)：　　日期：　　年　　月　　日

指导教师 (签名)：　　日期：　　年　　月　　日