

面向工程能力培养的大数据教学研究

戴壮红

(湖南大学信息科学与工程学院, 湖南 长沙 410082)

✉dmh@hnu.edu.cn



摘要: 针对目前研究生专业选修课教学中采用教师讲课、学生写读书报告等教学形式, 缺少工程素养的系统性训练等问题, 在“大数据处理技术”课程中, 探究专业选修课的教学改革方法。在讲授大数据知识的同时, 通过设计与实现小组项目, 培养学生系统分析与设计能力, 以及团队协作和解决工程问题的能力。实践证明, 面向能力培养的教学方式效果良好, 学生高度认可。

关键词: 新工科; 大数据; 工程能力; 教学研究

中图分类号: TP392 **文献标识码:** A

Research on Big Data Teaching for Cultivation of Engineering Ability

DAI Muhong

(College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China)

✉dmh@hnu.edu.cn

Abstract: Currently, teacher lectures and reads reports are popular teaching methods in current postgraduate professional elective courses, which lacks of systematic training of engineering literacy. Therefore, it is necessary to explore a new teaching method for data processing technology course. The new teaching method aims to cultivate students' capabilities of system analysis and design, teamwork and engineering problem solving capabilities, through design and implementation of group projects. The capability-oriented teaching method has proved to be effective and highly recognized by students.

Keywords: new engineering; big data; engineering ability; teaching research

1 引言(Introduction)

新工科致力于培养德才兼备的具有深度学习、人工智能和大数据处理与分析等多元化技术的创新型人才, 使之具有较强的工程实践能力^[1]。随着大数据应用的兴起, 对大数据应用人才的需求急剧增长。大数据应用人才应具备数据科学相关的专业知识和工程能力。专业知识包括人文社会科学知识、数据科学基础知识和数据科学专业知识^[2]。大数据创新型人才的培养首先必须培养学生跨学科的数据思维, 同时还应培养他们具备工程应用的能力。“大数据处理技术”是湖南大学信息科学与工程学院面向硕士研究生开设的选修课程, 旨在系统地介绍大数据处理和应用开发相关知识与技能^[3]。“大数据处理技术”课程在传授研究生大数据专门知识的同时, 特别强调培养他们解决新工科工程问题的能力。

2 课程现状(Course status)

目前国内一些大学研究生选修课程的教学形式主要是采

用任课老师讲解部分内容, 博士生做几个技术讲座, 研究生写一篇读书报告或者是写一篇课程论文, 老师根据读书报告或者课程论文给出学生课程成绩。这样的教学方式存在着一些问题, 要学生在一个学期的课程学习中写出一篇有价值的论文实在是困难。如何提高研究生课程的授课质量? 如何实现从知识传授到能力培养的转变? 如何把学生的被动学习转变为自主学习, 提高学生的终身学习能力? 这些问题需要高等教育的管理人员和任课教师深入思考。

3 课程教学模式(Course teaching mode)

目前许多大学都开设大数据分析相关课程, 不同大学开设的大数据课程, 其知识的侧重点都不一样。但是, 基本上可以分成下面几类: 一是针对数据分析和数据挖掘的课程; 二是针对并行与分布式编程架构和模型的课程, 主要是对Hadoop MapReduce以及Spark生态系统的讲述; 三是面向非结构化数据处理与存储管理的NoSQL数据库和NewSQL数据

库技术以及面向云存储的课程^[4]。

通过查阅世界一流大学大数据相关课程,采用下面的知识点作为大数据应用技术课程的主要教学内容:大数据存储与管理技术、大数据编程模型、大数据分析和可视化技术,以及商业智能和数据湖等新的大数据应用技术^[5]。

课程学习是研究生面临的一个重要学习过程,专业课程学习可以使研究生掌握更广泛的专业知识,在未来的研究中发挥重要作用。研究生专业选修课教学应该根据课程的性质和学生的特点,采用不同于以往的教学方式和教学模式。教学模式不应该是单一的,需要结合课程特点采用多种教学方法。在“大数据处理技术”研究生课程教学中,采用授课、技术讲座、课程实验、讨论会、课程小组项目等多种方式进行。课程讲解大数据生态系统中的软件架构和相应的软件产品与处理技术,对学生难以理解的重要理论知识作出详细和深入阐述,使学生掌握核心的数据科学专业知识。针对每一个重要的知识环节,安排独立的实验,学生必须单独完成。通过实验,使学生掌握大数据生态系统的基础和核心技术,结合实际的操作和编程实验,巩固课堂教学内容,使学生将所学理论与实际技能相结合,能够应用主流的大数据开发技术和开发工具。此外,课程需要结合当前研究的热点问题,通过提出问题和分小组完成一个大数据相关的工程项目。通过运用软件工程的原理与方法,小组成员全程参与问题描述、需求分析、算法(系统)设计、编程实现到测试等每一个开发环节,培养学生系统分析与设计能力以及团队协作和解决工程问题的能力。

4 课程教学目标(Course teaching objectives)

大数据已经成为企业和公司使用信息增进用户体验和改变运营模式的极其重要的新技术。如何有效使用大数据?如何从大数据分析中获取知识和智慧?“大数据处理技术”课程将把最新的大数据处理技术介绍给学生,帮助学生掌握从海量的异构数据源中对大数据进行存储与管理、高效处理、有效应用与分析的技能。

4.1 理论教学目标

课程涵盖大数据技术的基本概念、大数据生态系统、大数据存储模型、大数据编程模型、大数据分析技术、大数据可视化技术和大数据主流应用技术。

4.2 工程能力培养目标

学生解决工程问题能力的培养通过课程实验以及设计课程中综合性的工程项目来达成。工程能力的培养问题不仅仅是个人的问题,工程问题的解决需要团队协作。“大数据处理技术”课程希望通过项目小组的形式,培养学生团队合作、有效沟通、项目管理和终身学习的能力^[6]。(1)团队合作,让学生分成讨论与项目小组,因为工程项目无法一人独立解决,需与他人合作。(2)有效沟通,讨论课让学生书面及口头报告,大家一起讨论,形成最终解决方案。(3)主动学习、终身学习,让学生主动学习,减少教师课堂讲授。学生须设计实验来验证、解决问题,须自己或团队搜集资料、寻求资

源、增进专业素养。(4)项目管理,让学生尝试执行项目,规划任务和时间。

5 课程教学实施(Course teaching implementation)

5.1 课时安排

“大数据处理技术”为研究生选修课,共48学时,课程通过课堂讲授、课程实验、文献阅读、小组讨论、课程小组项目和期末考试等环节来组织教学。教师用32课时讲解核心的大数据生态系统架构和关键技术,选修课程的学生2—4人组成项目小组,必须完成一个与大数据处理技术相关的工程项目。教师安排4个课时点评学生的文献阅读报告和小组拟进行的工程项目开题报告,对工作量、难易程度、实验数据来源、关键技术和算法等方面进行把关。每个项目小组就项目的开题报告、实验环境搭建、实现的关键技术与算法设计、编程实现、系统测试、问题描述与解决方案等内容在课堂上进行讨论。由各研究生小组成员讲解或演示,共6次,12课时,教师参与讨论与引导,学生遇到问题时给出解决方案的建议。

5.2 理论教学内容

“大数据处理技术”课程的主要内容分为四个部分:大数据处理与算法、大数据存储与管理、大数据分析与可视化、大数据应用技术。课程内容和主要知识点如表1所示。

表1 课程内容与知识点

Tab.1 Course content and knowledge points

课程内容	知识点	课时
大数据、数据科学基础	大数据与数据科学	2
	MapReduce并行计算技术, hadoop的安装与生态系统	2
大数据存储与管理	分布式文件系统HDFS的基本原理和使用方法	2
	分布式数据库HBase的基本原理和使用方法	2
	NoSQL数据库的概念和基本原理,典型的NoSQL和NewSQL产品的使用方法	2
	云数据库的概念与技术	2
	Hadoop MapReduce并行编程框架	2
大数据编程模型	MapReduce编程技术与算法设计	4
	Spark大数据处理架构	4
大数据分析 with 可视化	大数据数据仓库hive技术	2
	Rhadoop平台下的大数据分析与可视化技术	2
大数据应用	商业智能、数据湖、数据中台	6
小组项目	课程论文宣讲与点评	4
	小组项目汇报与讨论	12

5.2.1 大数据存储与管理

近十多年来,对存储和处理大规模数据集的需求不断增长,一直推动着数据存储技术和并行计算技术以及数据库技术快速发展。从本地存储到集群,从分布式存储到云数据库存储,数据存储方法得到了极大的增强^[7]。数据库系统已从传统的关系数据库系统进化到最新的基于NoSQL和支持数据库完整性约束的NewSQL数据库系统。大数据除结构化的数据

外, 主要涉及大量的非结构化数据, NoSQL和NewSQL是非结构数据存储的新方法。大数据存储建立在分布式文件系统之上, HDFS是使用最广泛的分布式文件系统。NoSQL数据库的类型很多, 有文档的、图的、列簇的和键值等^[8]。课程将讲解大数据存储模型和管理技术, 大多数NoSQL数据库系统都是开源的, 因此很容易纳入大数据课程的实验环节中。

5.2.2 大数据编程模型

并行分布式计算与处理模型是大数据应用系统架构的核心。大数据编程模型提供了编写大数据应用系统通用的接口, 为构建大数据应用程序系统提供了简单的途径。大数据通常涉及在集群上的分布式并行计算技术。MapReduce是主流的并行计算算法框架, 它可以支持各种数据密集型程序, 是事实上的大数据编程模型。大数据处理的核心是采用两个基本过程对象Mapper和Reducer, 它通过两种特殊的计算方式Map和Reduce归约到集群不同节点上进行处理^[9]。大数据处理另外一个广泛应用的算法框架是基于Spark平台中实现的弹性分布式数据集RDD(Resilient Distributed Dataset)。通过使用RDD, 用户不必担心底层数据的分布式特性, 只需要将具体的应用逻辑表达为一系列转换处理, 就可以实现管道化, 从而避免了中间结果的存储, 大大降低了数据复制、磁盘I/O和数据序列化的开销。

课程将通过比较具体的代码示例, 全面阐述大数据框架Hadoop和Spark中的不同编程模型。

5.2.3 大数据分析可视化

可视化已成为从大数据分析结果中提取有意义信息并带来价值的一项核心技术。有效地工具和技术, 可以轻松地可视化杂乱的海量大数据, 通过可视化大数据来改善企业的经营决策能力。为了实现强大的数据分析和出色的可视化, 在Hadoop上使用R语言将提供一个弹性的数据分析平台。R语言提供了最流行的开放源代码数据统计分析软件包, Hadoop强大的数据处理能力和R强大的可视化分析能力相结合的Rhadoop, 提供了很好的大数据分析平台。课程将介绍R语言和Tableau等可视化工具。

5.2.4 大数据应用技术

随着大数据技术的不断更新和迭代, 数据管理工具得到了飞速的发展, 相关概念层出不穷, 如从最初决策支持系统(DSS)到商业智能(BI)、数据仓库、数据湖、数据中台等。数据湖是企业所有数据的单一存储, 包括源系统数据的原始副本, 以及用于报告、可视化、分析和机器学习等任务的转换数据。数据湖可以包括各种类型的数据: 关系数据库的结构化数据、半结构化数据(CSV、日志、XML、JSON); 非结构化数据(电子邮件、文档、PDF)和二进制数据(图像、音频、视频)。Hadoop是最常用的部署数据湖的技术。数据中台是一个承接技术、引领业务、构建规范定义的智能数据处理平台。许多企业已经开始尝试使用大数据和云技术来构建数据湖和数据中台, 并支持数据驱动的智能决策。如今, 世界上顶级的公司和大型企业已经部署或正在部署数据湖和数据

台。课程将向学生介绍这些新的技术, 让学生了解企业大数据应用的真实现状。

5.3 课程实验设置

大数据处理技术课程安排了多次实验, 老师编写了实验指导书, 对大数据实验环境的构建、开发平台和开发工具都有基本的规定。对大数据主要技术都安排了至少一次的实验内容, 需要学生在规定的时间内、课外独立完成。所有实验为一个整体, 有其延续性, 如果不能完成当次实验将会影响下次实验。每次实验后, 必须认真填写实验报告, 在要求的时间提交实验报告的电子版, 超过规定时间, 该次实验成绩以零分计。

5.4 小组课程项目

大学的核心任务是培养人才, 大学教学的本质应该是以学生为中心, 以能力为本位的研究型教学, 以培养学生的自主学习、终身学习、研究性学习和解决工程问题的能力。

“大数据处理技术”课程教学体系的另外一个重要组成部分是学生按2—4个人组成一个课程讨论与项目小组, 项目小组的人数不能超过5人。学生自选具有一定难度和工作量的与大数据处理相关的研究或工程类题目, 鼓励结合导师的研究工作选择项目, 完成一个项目的开发。各小组进行查阅资料、文献阅读、进度分享、实验操作、成果展示等多个阶段。在查阅文献、资料的过程中, 训练学生的文献检索能力以及理解能力, 组内每位成员参与讨论, 增强沟通能力。

(1) 项目小组的组织与控制

项目小组成员要求有熟悉Java、Linux或Unix等技术的学生。小组长负责任务的划分, 组织、讨论详细设计, 组织和协调本小组的开发、文档整理等工作, 确保开发过程中及时完成各个里程碑任务。研究生课程没有配备助教, 但是“大数据处理技术”课程采用自愿报名的方式选择一位课代表, 负责协助主讲教师管理课程的相关事宜, 包括实验报告的归档, 记录每组讨论内容。项目的讨论课上, 要求所有小组成员都能有机会上台讲解, 而不只是项目组长的事情, 锻炼学生的表达能力。

(2) 项目小组需要提交的技术文档

课程需要填写、检查、整理和归档项目的过程报告。开题报告的目的是为了评估课程设计选题的内容和难度是否达到一定要求, 主要内容包括: 小组成员信息(人员、学号、联系信息)、各个成员初步的项目分工计划、项目题目、研究问题背景、主要技术难点、主要解决方法和算法设计思路等。小组项目总结报告主要包括的内容: 小组信息(人员、学号、联系信息); 小组具体的分工, 需要明确说明各成员在整个课题中分工负责完成的内容, 所做贡献的比例; 研究问题的背景、主要技术难点、解决方法和算法设计思想; 详细的设计说明, 包括详细算法设计、程序框架、功能模块、主要类的设计说明、程序运行和实验结果的说明及分析, 在完成系统的功能、性能、扩展性等方面存在的不足和可能的改进之处。

5.5 成绩评定方法

课程强调学生大数据素养和能力培养,课程的考评采用课程实验成绩、小组项目、考勤和期末考试综合评定的方法进行,期末成绩=个人课程实验25%+小组项目35%+期末考试30%+考勤10%。

6 结论(Conclusion)

“大数据处理技术”课程的授课,让学生熟悉了大数据管理和分析的基本概念和主流技术;能够胜任处理大数据应用程序所面临的挑战,并可提出可扩展性的解决方案。“大数据处理技术”课程采用学生小组项目的方式,在教师讲解专业知识的同时,学生通过一个小组,完成大数据应用开发项目工程能力的培养。通过论文宣讲、学生汇报讨论,教师依据自己的经验点评,学生的收获很大。这一教学形式很受欢迎,效果良好。

大数据技术涵盖的主题非常多,要在一个学期的课程中全面涵盖大数据处理和应用的技術是不现实的,还有许多重要的技术,比如大数据的数据采集与预处理、大数据流处理以及大数据安全与隐私保护等只能引导学生在课后自学。

参考文献(References)

- [1] 陈燕.新工科研究进展与前瞻[J].天津大学学报(社会科学版),2020,22(3):214-222.
- [2] 田绪红,邝颖杰.大数据应用人才的知识、能力、素质结构及其培养[J].计算机教育,2017(8):57-60.

(上接第59页)

5 结论(Conclusion)

本研究以“人工智能+教育”融合视域下的人才培养为目标,分析了目前高职人工智能专业教育存在的问题,提出了解决问题的重要挑战,给出了解决问题的建议,取得了较好的教学效果。未来工作致力于高职人工智能人才培养模式的优化、课程体系的持续完善等,以提升学院人工智能相关专业建设水准,为国家培养更多的优秀人才,服务于国家“十四五”建设。

参考文献(References)

- [1] 张海生.我国高校人工智能人才培养:问题与策略[J].高校教育管理,2020,14(02):37-43;96.
- [2] 肖卓宇,陈果,徐运标,等.新工科视角下以计算思维为导向的高职信息类学生程序设计能力培养研究[J].计算机时代,2020,38(6):106-108.
- [3] 王雪,何海燕,栗革,等.人工智能人才培养研究:回顾、比较与展望[J].高等工程教育研究,2020(01):42-51.
- [4] 吴朝晖.交叉会聚推动人工智能人才培养和科技创新[J].中国大学教学,2019(02):4-8.
- [5] 姚新.为什么人工智能人才培养这么难?[J].中国大学教学,2020(07):25-29.
- [6] 肖卓宇,徐运标,陈果,等.“人工智能+教育”融合的实施路径研究[J].计算机时代,2020(11):103-105;109.
- [7] 肖卓宇,宋艺.基于CDIO的大数据技术课程教学研究[J].长沙民政职业技术学院学报,2018,25(02):125-126.
- [8] 肖卓宇,黄俊,徐运标,等.新工科视角下以计算思维为导向的

- [3] DePratti, R., Dancik, et al. Challenges in designing an introductory course in big data programming[J]. Journal of Computing Sciences in College, 2015, 30(6):104-105.
- [4] Adam H. Villa. Motivating the development of an advanced database systems course[J]. Journal of Computing Sciences in College, 2016, 31(3):119-127.
- [5] Yuri Demchenko. Big Data Platforms and Tools for Data Analytics in the Data Science Engineering Curriculum[C]. International Conference on Cloud and Big Data Computing, 2019:60-64.
- [6] 刘爽,陈鹏,王巍,等.基于工程教育认证背景的民族院校大数据课程体系设计[J].计算机时代,2020(5):74-77.
- [7] Xiong,W., Hawley, D., Monismith, D. NoSQL in database education[J]. Journal of Computing Sciences in College, 2015, 30(5):151-152.
- [8] Grolinger, K., Higashino, W.A., Tiwari, A.. Data management in cloud environments: NoSQL and NewSQL data stores[J]. Journal of Cloud Computing: Advances, Systems and Applications, 2013, 22(2):1-24.
- [9] 刘萍.基于Hadoop与Spark的高校校园大数据平台研究.软件工程,2018,21(5):15-18.

作者简介:

戴壮红(1964-),男,硕士,教授.研究领域:数据科学.

高职Python程序设计公共课程教学改革研究[J].电脑知识与技术,2020,16(21):154-155.

- [9] 吕薇,季波,张怡凡,等.美国一流研究型高校人工智能人才培养的经验与启示——基于“以学生为中心”的视角[J].大学教育科学,2019(06):102-109;123.
- [10] 肖卓宇,叶嘉笑.基于CDIO的IT类应用型创新人才培养模式的研究[J].湖南邮电职业技术学院学报,2019,18(01):89-90;97.
- [11] 肖卓宇,何锴,陈果,等.带特征指标约束描述的设计模式分类挖掘[J].山东大学学报(工学版),2020,50(6):48-58.
- [12] Timms M J. Letting artificial intelligence in education out of the box: educational cobots and smart classrooms[J]. International Journal of Artificial Intelligence in Education, 2016, 26(2):701-712.

作者简介:

肖卓宇(1979-),男,硕士,教授.研究领域:软件演化,程序理解,机器学习.

陈果(1980-),女,硕士,讲师.研究领域:软件演化,软件工程.

郭杰(1976-),女,硕士,讲师.研究领域:软件演化,软件工程.

黄俊(1990-),男,硕士,助教.研究领域:软件工程,智能制造.

徐运标(1983-),男,硕士,讲师.研究领域:软件演化,软件工程.