“数据科学”课程群与“数据科学导论”课程建设初探

覃雄派，陈跃国，李翠平，柴云鹏，徐君，文继荣，杜小勇
中国人民大学信息学院，北京 100872

摘要
为适应新时代数据科学和大数据人才培养的需求，中国人民大学信息学院近年来针对计算机课程体系进行改革，设计了以“问题求解”“系统平台”和“数据科学”为主线的三大课程群。在介绍这三大课程群设计理念的基础上，重点介绍“数据科学”课程群中的第一门课“数据科学导论”课程的目标和定位、内容安排、教学计划、考核方法以及实践环节设计。

关键词
人才培养；大数据；数据科学；课程体系

中图分类号：TP311.13 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018057

Toward construction of “data science” course group and “introduction to data science” course

QIN Xiongpai, CHEN Yueguo, LI Cuiping, CHAI Yunpeng, XU Jun, WEN Jirong, DU Xiaoyong
School of Information, Renmin University of China, Beijing 100872, China

Abstract
To adapt to data science and big data talent cultivation requirements of the new era, School of Information, Renmin University of China has reformed its course system in recent years. Three course groups as problem solving, computer system, and data science has been designed. Firstly, the behind idea of the three course groups was presented, then the focus was shifted to the introduction of the first course among the data science course group, i.e. “introduction to data science”, including objective and positioning of the course, course contents, teaching program, access method, and project design.

Key words
talent cultivation, big data, data science, course system

1 引言

数据科学是以计算机科学（特别是数据库、数据挖掘、机器学习等）和统计学为基础，融合其他学科的一门新兴的交叉学科。它研究数据的各种类型、状态、属性及其变化规律，研究如何对数据进行分析，从而揭示自然界和人类行为等现象背后的规律。

2016年，北京大学、对外经济贸易大学及中南大学3所高校申请并且成功获得教育部批准，建立“数据科学与大数据技术”本科专业。2017年，又有32所高校获批建立该专业，中国人民大学为其中之一。2018年3月，教育部公布第三批获批建立“数据科学与大数据技术”专业的院校。至此，共有280多所高校获批建立该专业。

如何设计“数据科学与大数据技术”专业的课程体系、建设各门课程，成为各个高校的教学主管领导和任课老师关心的问题。本文介绍中国人民大学信息学院在计算机专业教育课程体系方面进行的改革，并重点介绍其中“数据科学导论”课程的建设情况。

2 课程建设思路

2.1 计算机专业教育课程体系

随着计算机及互联网的飞速发展，当今社会已进入大数据时代，数据及其应用已经渗透到社会和生活的方方面面。2015年9月5日国务院发布了《促进大数据发展行动纲要》，系统部署了大数据发展工作，这意味着大数据发展正式上升为国家战略。数据的处理与分析是推动生产力增长的关键要素，国家急需大量能构建系统、会分析数据、懂领域业务的复合型人才。

在新的需求背景下，中国人民大学信息学院（以下简称信息学院）把计算机专业课程划分为“问题求解”“系统平台”和“数据科学”3条主线、“数学和自然科学”“人文社科教育”两条辅线，如图1所示。

“问题求解”类课程主要训练学生用计算机求解问题的思维方式、方法和实现手段（编程）。“系统平台”类课程主要介绍和训练支持问题求解的实际计算机软硬件系统的使用、设计与实现。“问题求解”和“系统平台”两条主线中的课程设置主要是为了使计算机专业的学生在两个方面打下扎实的基础。而“数据科学”类课程涉及大数据从获取、存储、管理、分析到领域应用等整个数据生命周期各个环节的内容，尤其是云计算、大数据、人工智能（基于大数据的智能）等新技术，将重点集中在这一主线中进行讲授和训练。“数学和自然科学”辅线课程将为计算机专业的学生打下扎实的数学和自然科学基础，“人文社科教育”类课程的主要目标是提高学生的人文素养，积累社会学科的基础知识和方法论，同时数据科学中的大数据和人工智能的课程也会与人文社会学科深度结合。

从纵向看，从大一到大四，4个年级的课程深度逐步加强。大一阶段，主要通过“问题求解”主线中的程序设计等课程，训练学生的计算机编程能力，培养计算思维。从大二开始，设置“系统平台”和“数据科学”类课程，系统地培养学生问题求解的方法和能力，训练学生如何抽象问题、建立模型、求解问题，并编程实现；通过系统平台方面的基础课让学生初步了解计算机软硬件系统，学习实际系统模块的设计与开发，同时也会通过导论类型的课程开启“数据科学”主线课程的学习。大三

阶段的学习会再增加一个难度，增强问题求解的理论基础和复杂问题求解的能力，并训练学生完整设计和开发复杂的计算机系统的能力；在“数据科学”方面，也对大数据从获取、存储、管理、挖掘到可视化等整个数据生命周期各个环节进行综合讲授和训练。在大四阶段的毕业设计中，学生针对导师提出的开放性问题进行探索，在调研、分析、研究、设计与开发等方面得到进一步的训练。“领域应用”板块通过大三和大四的选修课和毕业设计完成。

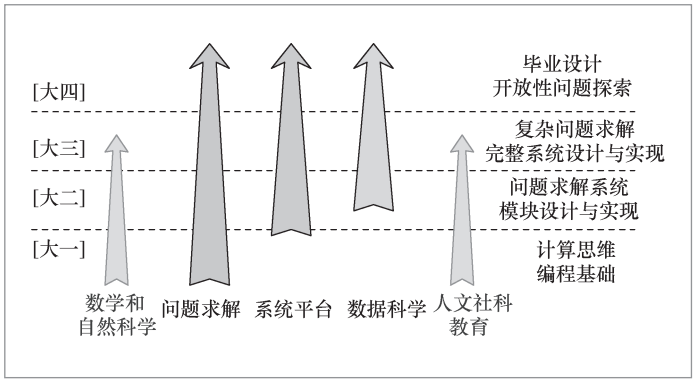

图1 计算机专业教育课程体系

2.2 “数据科学”课程群

传统计算机专业教育是以计算为中心的，与数据科学相关的课程设置比较薄弱，一般只包括关系数据库系统一门课程，甚至只是一门选修课。但是在数字经济环境下，云计算、物联网、人工智能都和大数据密切相关，需要转换思维，增加一系列重要的数据科学和与大数据相关的新课程。这些课程应涵盖数据从准备、存储、管理、分析到领域应用等整个数据生命周期各个环节的内容。

此外，在课程体系建设方面，需要加强实践课程的建设，尤其是设计和建设以数据管理和分析课程为核心的在线实验平台，加强学生的实践能力和动手能力；与企业合作建设新课程，给学生提供了解工业界的实际情况和最新技术的机会及充足的锻炼机会。

为了提高大数据人才的培养质量，信息学院专门为数据科学与大数据技术专业精心打造了一系列精品课程，试图在本科阶段给学生打下扎实的数学、大数据和计算机基础。数据科学与大数据技术是一个宽口径的专业，与很多行业有相关性，全面而扎实的基础可以提高本专业学生的适应能力。

在图2中，“数据科学导论”“大数据计算智能”“非结构化大数据分析”等课程是专门为大数据专业开设的新课程。信息学院精选了优秀教师组建课程建设团队，例如“数据科学导论”课程由杜小勇教授和文继荣教授牵头，陈跃国教授、徐君教授、覃雄派博士参与；“大数据计算智能”和“非结构化大数据分析”由李翠平教授和张静博士负责；“计算机系统基础”课程由柴云鹏副教授负责；“并行与分布式计算”由张峰博士负责。此外，信息学院的传统精品课程“数据库系统概论”由杜小勇教授和陈红教授联合授课。

2.3 “数据科学导论”课程

“数据科学导论”是一门入门课程，同时也是“数据科学”课程群统领式的课程，把学生引进数据科学的大门。它的目标有两个，一个是扩展宽广的视野和培养浓厚的兴趣，另一个是打下坚实的基础，有利于学生学习后续课程，有利于培养数据科学家。

一般来讲，该课程在大二下学期开设。在该时间点，学生已经学习了必备的数学知识和编程基础知识，为该课程的学习做好了准备。而这门课的学习，又为大三的专

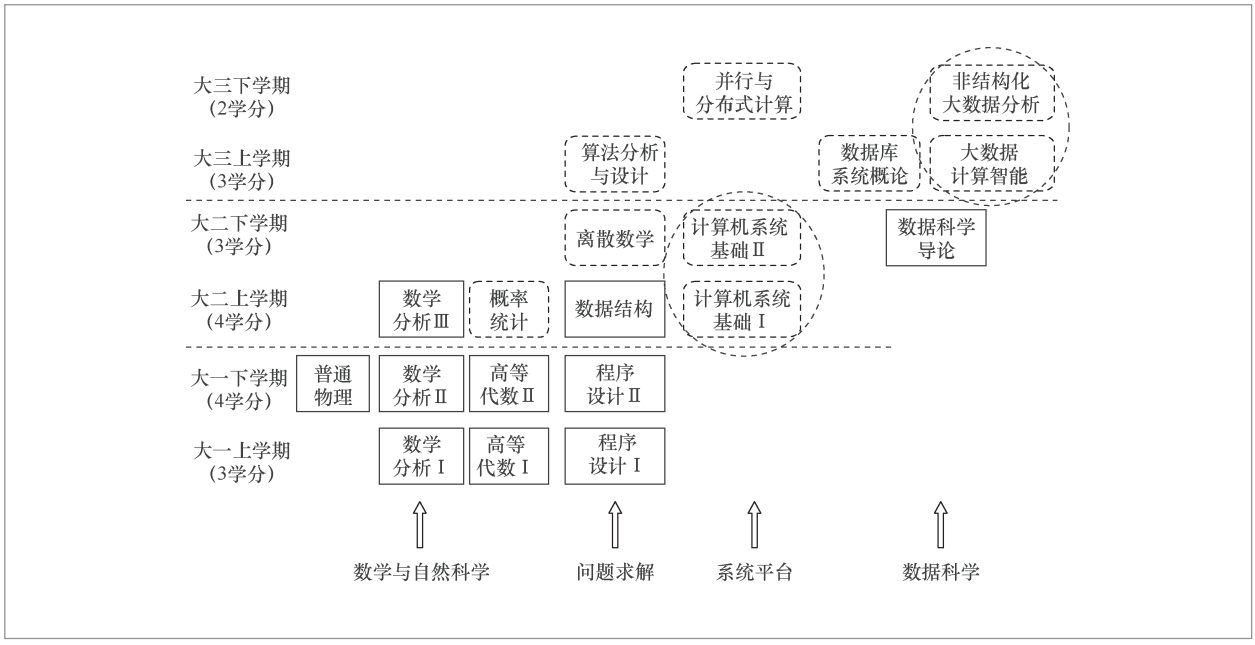

图 2 “数据科学”课程群

业课学习打下了必要的基础。

由此可见，“数据科学导论”课程的地位非常重要，它在学科基础课和专业课之间起到了承上启下的作用。由于它的入门和统领地位，在内容上它与后续的专业课程不可避免地会有部分重复。因此，这门课程在设计时，首先需要回答一个问题，即如何处理和后续课程的关系。换句话说，“数据科学导论”课程在广度和深度方面应该如何设计，才能既避免和后续课程的简单重复，同时又能够起到入门和引领的作用。

“数据科学导论”包含很多内容，而大部分内容可以扩展为一门课程。例如，在讲到文本分析的时候，“数据科学导论”课程里面只对文本分析的概念、技术和方法做了初步的介绍，让学生理解概念，会用工具，培养其数据思维。而在非结构化大数据分析课程里面，则深入介绍了文本的处理和分析方法。又如，在讲到人工神经网络的时候，“数据科学导论”课程只需把神经网络的基本原理、深度学习的基本概念介绍清楚，学生会使用相关工具即可，具体的神经网络模型很多，不胜枚举，可以选择一两种加以介绍，其他模型可以放在后续的机器学习或者深度学习课程中细细地展开分析。

3 “数据科学导论”课程内容安排及教学计划

3.1 国外若干著名大学的“数据科学导论”课程建设情况

在设计课程内容时，笔者团队首先调研了加州大学洛杉矶分校、斯坦福大学、华盛顿大学等几个国外著名大学的“数据科学导论”课程的内容安排，下面对其进行详细介绍。

加州大学洛杉矶分校的数据科学导论课程的主要内容包括数据管理、机器学

习、自然语言处理、统计分析、可视化等。该课程将自然语言处理进行了重点介绍，课程内容还包括大数据工具Hadoop、各种NoSQL数据库、Python编程语言等。课程最后安排了一系列的实践案例，要求学生在实际业务数据集上运用所学知识解决实际问题。该课程只需学生具有数学特别是概率统计的基础，会编程即可，没有其他先决条件，适用于本科生。

斯坦福大学的数据科学导论课程主要包含3个方面的内容，分别是数据可视化、机器学习、R编程语言。该课程同时包括一系列来自实际业务的案例分析讲座，包括天气预报、市场营销、生物学、股票市场、医药行业等，学生可以跟着教师的介绍，实际动手操作，一步步解决问题，这对于培养学生的动手能力是非常必要的。该课程在内容设计和教学设计上，照顾到没有计算机基础的学生，适用于本科生。

华盛顿大学的数据科学导论课程的主要内容包括实现（大规模）数据管理的SQL数据库（关系数据库管理系统）、NoSQL数据库技术及其选择和取舍（trade off）、数据挖掘和机器学习的基本技术、统计建模方法（如线性和非线性回归、数据可视化等）。该课程最后选择图数据分析作为特别的专题进行介绍。该课程虽然称为“数据科学导论”，但是其中的内容是以一系列相对独立的专题讲座的形式讲授的，前后两次课堂内容并不一定连续，适用于研究生或者高年级本科生。

纵观这些著名大学的数据科学导论课程可以发现，在内容安排方面有以下共同点：

- 数据管理不能缺失，数据科学导论不能只讲数据分析而不讲数据管理，把数据管理起来是分析的基础；
- 强调动手能力培养，选用的编程语言有R、Python等，问题来自实际应用；
- 选择某种数据类型进行深入介绍，比如图数据、文本数据等，不同学校在内容上各有取舍；
- 数据可视化内容必不可少。

3.2 信息学院“数据科学导论”课程建设情况和新规划

数据科学导论课程在中国人民大学开设已有5年时间。在探索的过程中，增长了课程建设的经验，同时也有一些教训。这门课程刚开始作为大三的选修课开设，随着数据科学和大数据人才需求的逐步旺盛，学生兴趣浓厚，选修这门课的学生逐渐增多。部分学生的学习主动性很高，积极参加中国计算机学会和各大互联网公司的大数据大赛，并且获得了好名次。

2017年，中国人民大学获批建设“数据科学与大数据技术”专业后，对该课程的定位为学科基础课。中国人民大学信息学院领导全方位督导这门课程的建设，对教学大纲、教学内容、教学环节、实践环节进行了重新梳理，为2017年开始招生的“数据科学与大数据技术”专业本科生开设这门课程做好了准备。

该课程作为选修课程讲授时，主要的教训有3点，分别是内容过于繁多、难度偏大、实践部分安排的时间不够。经过信息学院“数据科学导论”课程建设小组多次讨论，在教学目标方面，突出两个方面：一个是掌握数据分析的基本思路、基本理论与基本方法；另一个是实践应用能力培养，即面向实际应用中不同行业的需求，能够给出数据解决方案设计和最终实现。

对上述前期教学的3个经验教训进行梳理后，信息学院在内容方面删繁就简，突出结构化数据分析或者流数据处理、文本分析、图数据分析三大块；在难度方面，把该门课程定位为入门的课程，目的是培养

学生浓厚的兴趣，因此难度上要降低，对算法的数学基础和推导过程的把握由后续的各门专业选修课程来承担。本门课程仅要求掌握基本原理，加强实践环节，要求学生熟练掌握主流工具，完成数据管理和分析全流程的实践，切实培养学生的动手（编程）能力。在具体的教学中，设计了模块式教学法以及基于大作业（project）的教学法。通过3个大作业，把理论讲解和动手实践有机结合起来。

3.2.1 教学内容模块化

“数据科学导论”是一门入门性质的课程，内容安排上一定要注意取舍，达到广度和深度的平衡，既要避免泛泛而谈、没有重点，也要避免过于强调某个方面，但是广度不够。

从广度上来讲，要帮助学生构建起一个完整的知识体系。在理论部分，要包括数据科学的基本概念、技术和方法，数据的不同类型，数据处理的流程和数据的生命周期，数据处理的不同模式，数据分析的主要方法等。其中数据的类型是非常丰富的，包括文本、社交网络、社交媒体、时间序列、轨迹数据、音频/图像和视频等。在应用部分，要包括数据处理的基础设施、平台和各种工具以及一系列精心设计的案例。

在深度上，只能选择部分内容进行深入的介绍。在介绍完数据科学的基本概念、技术和方法以及主要的数据类型及其分析方法之后，可以选择某种具体的数据类型进行深入的介绍。同时在应用部分，要设计一些实际的案例，要求学生结合实际数据进行上机实践。

教学内容分为四大模块，分别是基础（base）模块、关系数据/流数据处理（relational & stream）模块、文本数据处理和分析（text）模块、图数据处理和分析（graph）模块。

基础模块为后续的3个实践模块做准备工作。它的内容包括以下几方面。

- 概念：数据科学概论，主要介绍数据科学的基本概念、大数据及其价值、数据处理的全生命周期，包括数据的采集和获取、数据预处理/清洗和集成、数据管理、数据分析、可视化和解释等。
- 方法：包括各种数据模型、数据处理的不同模式（批处理和流式处理）、通用的数据分析方法、数据可视化等。
- 平台和工具：包括分布式计算与大数据平台（Hadoop & Spark）以及Python语言。其中Python语言部分包括对Python语言基础以及Python的几个重要的库（数据预处理库pandas、机器学习库Scikit-Learn、可视化库Matplotlib）的介绍。

大数据具有4个关键的特点，分别是数据量大（volume）、数据类型多样（variety）、数据价值密度低（variety）以及速率快（velocity）。上述的教学内容完全体现了大数据的这些特点。分布式计算与大数据平台体现了数据量大，数据模型体现了数据类型多样，数据的清洗和集成体现了数据价值密度低，数据处理的不同模式体现了速率快。经过模块内部整理以后，基础模块的内容可以被划分成如下子模块，见**表1**。

后续的3个模块除了介绍理论之外，还要介绍平台、方法和应用。由于是导论课程，其主要目的是把学生引进数据科学的大门，培养其对数据的感觉和兴趣，基础理论不必讲得太深入（可以放在后续专业选修课程中深入介绍），而是偏向工具的使用和应用的开发，让学生迅速获得对数据价值的认识。特别需要注意的是，要准备好必要的数据集，以便学生可以利用一些工具对其进行处

理和分析。

关系数据/流数据处理模块提供两个选项，可以根据需要进行选择。第一个选项侧重于介绍大数据的实时流式处理，内容包括流式处理基础、流式处理工具和流式处理应用，可以选择Storm作为典型流式数据处理系统加以介绍，大作业部分可以针对传感器数据，实现实时汇总和可视化。第二个选项侧重于介绍关系模型、SQL语言以及分布式大数据多维分析（OLAP），工具部分可以选择对MySQL以及SQL on Hadoop系统加以介绍，做到会用即可，大作业部分可以针对销售数据进行多维分析以及可视化。

文本数据处理和分析模块依赖于分布式的大数据处理平台和分析工具，包括Hadoop/Spark、NLTK for Spark、Scikit Learn for Spark等，使用各种文本分析方法，实现对文本的分词、索引与检索、情感分析、实体识别、可视化等关键任务。大作业部分可以针对产品评论进行情感分析，并且从产品维度、时间维度进行汇总，最后进行可视化，让用户了解热门产品及客户对不同产品的态度变化等。

图数据处理和分析模块依赖于GraphX on Spark平台（工具）或者Neo4J数据库，对图数据进行创建、查询、路径分析、PageRank、社区检测等。大作业部分可以基于公开数据集进行重要节点排名、社区检测等，甚至可以扩展到金融领域，实现欺诈团伙的检测。

这3个模块都带有很强的实践性，教师在讲解平台、方法和应用的基础上，给学生布置大作业，要求学生熟练使用上述平台和工具，对数据集进行处理、分析和可视化，并且解释结果，体现数据的价值。

表1 数据科学导论课程内容的模块与子模块

子模块	内容
base1	数据科学概论、数据处理的全生命周期；从数据到大数据，再到大数据应用
base2	数据科学语言Python与数据可视化入门
base3	volume：分布式计算与大数据平台Hadoop/Spark
base4	variety：各种数据模型、通用的数据分析方法入门
base5	velocity & veracity：数据处理的不同模式（批处理和流式处理），数据清洗和数据质量保证、数据集成

3.2.2 教学计划

在时间安排上（共15个教学周，一个教学周2次课，每次课2课时，共30次课），基础模块被分配12次课，后续的3个模块分别被分配6次课，共18次课。

在基础模块，3位教师可以分别准备关于不同子模块的内容，3位教师的教学计划见**表2**（3位教师分别被命名为教师1、教师2、教师3）。

表2 3位教师3个教学班的教学计划（串行并行相结合）

上课次数	教学A班	教学B班	教学C班
2次课	教师1：base1模块概论	教师2：base1模块概论	教师3：base1模块概论
4次课	教师1：base2模块Python	教师2：base2模块Python	教师3：base2模块Python
2次课	教师1：base3模块volume	教师2：base4模块variety	教师3：base5模块velocity &veracity
2次课	教师3：base5模块velocity &veracity	教师1：base3模块volume	教师2：base4模块variety
2次课	教师2：base4模块variety	教师3：base5模块velocity &veracity	教师1：base3模块volume
6次课	教师1：relational/stream模块	教师2：text模块	教师3：graph模块
6次课	教师3：graph模块	教师1：relational/stream模块	教师2：text模块
6次课	教师2：text模块	教师3：graph模块	教师1：relational/stream模块

3.2.3 课程考核方法

采用上述教学计划，本课程的考核分为3个方面，包括平时的上机实践、大作业和期末闭卷考试。在平时的上机实践中，学生必须完成上机练习题目，并提交上机实验报告。大作业锻炼学生综合运用所学知识解决复杂问题的能力。期末考试考查学生对知识点的掌握和灵活运用能力。最终成绩按照15:15:15:15:40的比例，由平时成绩、大作业1、大作业2、大作业3和期末成绩构成。

3.3 课程参考教材及配套资源建设

自2013年以来，“数据科学导论”课程在中国人民大学信息学院计算机系开设已经有5年。笔者团队在授课的过程中，不断整理讲义。2016年开始，经过两年的精心编写和仔细校对，《数据科学概论》教材于2018年3月正式出版。

这本教材包括两个部分，分别是理论篇和应用篇。理论篇包括两个模块：一是数据科学基础，讲述数据科学的基本概念和原则；二是数据和数据上的计算，讲述不同的数据类型及其分析方法，数据类型包括结构化数据、非结构化数据、半结构化数据，分析方法包括统计学方法、数据挖掘和机器学习方法等。应用篇也包括两个模块：一是数据处理基础设施、平台和工具，讲述云平台、数据库、大数据平台、工具以及编程语言Python；二是数据科学案例和实践，讲述大数据应用的案例，并且面向金融领域的量化交易应用，从数据采集、模型训练、预测、评价到可视化等环节，带领读者完成数据分析处理的实践。在具体授课中，教师可以选择其他行业领域的实践案例进行介绍。

为了方便各个高校教师开设和讲授“数据科学导论”课程和提高学生的课程学习效果，笔者团队建设了课程配套网站，网站上提供了课程PPT、代码、数据、实验参考和技术博客等资源。配套资源网站还针对若干难点专题，给出了基于实例的详细和深入的剖析，帮助学生深入理解所学知识。这些难点包括最大期望（expectation maximization，EM）算法、反向传播算法、深度神经网络模型、奇异值分解（singular value decomposition，SVD）算法、概率潜在语义分析（probabilistic latent semantic analysis，PLSA）算法等。

4 结束语

本文介绍了中国人民大学信息学院面向大数据时代的课程体系改革，包括“问题求解”“系统平台”和“数据科学”三大课程群建设的构想和实践以及“数据科学”课程群的具体内涵。在此基础上，重点介绍了“数据科学导论”课程的建设情况，包括课程的目标和定位、课程内容安排、课程教学计划、考核方法、实践环节设计等。

参考文献:

[1] 覃雄派，王会举，杜小勇，等. 大数据分析-RDBMS与MapReduce的竞争与共生[J]. 软件学报，2012，23(1)：32-45.
QIN X P，WANG H J，DU X Y，et al. Big data analysis-competition and symbiosis of RDBMS and MapReduce[J]. Journal of Software，2012，23(1)：32-45.

[2] 王珊，王会举，覃雄派，等. 架构大数据：挑战、现状与展望[J]. 计算机学报，2011，34(10)：1741-1752.

WANG S, WANG H J, QIN X P, et al. Architecting big data: challenges, studies and forecasts[J]. Chinese Journal of Computers, 2011, 34(10): 1741-1752.

[3] 覃雄派, 陈跃国, 杜小勇, 等. “数据科学概论”课程设计[J]. 大数据, 2017, 3(6): 102-111.
QIN X P, CHEN Y G, DU X Y, et al. Course design of “introduction to data science”[J]. Big Data Research, 2017, 3(6): 102-111.

[4] 曹淑艳. “数据科学与大数据技术专业”课程体系与教学环节探讨[R]. 2017.
CAO S Y. Some thoughts on curriculum system and teaching design for the major of data science and big data technology[R]. 2017.

[5] 朝乐门, 邢春晓, 王雨晴. 数据科学与大数据技术专业特色课程研究[J]. 计算机科学, 2018, 45(3): 3-10.
CHAO L M, XING C X, WANG Y Q. Research on characteristic courses for the major of data science and big data technology[J]. Computer Science, 2018, 45(3): 3-10.

[6] 曹淑艳, 张莉, 黄浩. 大数据本科专业申报及认识[J]. 计算机教育, 2016(6): 30-33.
CAO S Y, ZHANG L, HUANG H. Application of undergraduate major of big data and some thoughts[J]. Computer Education, 2016(6): 30-33.

作者简介

覃雄派(1971-), 男, 博士, 中国人民大学信息学院讲师、硕士生导师, 目前主要从事高性能数据库、大数据分析、信息检索等方面的研究工作。主持1项国家自然科学基金面上项目, 参与多项“973”计划、“863”计划及国家自然科学基金项目, 在国内外期刊和会议上发表论文20余篇。

陈跃国(1978-), 男, 博士, 中国人民大学信息学院教授、博士生导师, 中国计算机学会高级会员, 数据库专家委员会委员, 大数据专家委员会通信委员, Frontiers of Computer Science青年编委。主要研究方向为大数据分析系统和语义搜索。主持国家自然科学基金重点项目1项, 广东省科技应用重大专项1项, 近年来在SIGMOD、SIGIR、ICDE、AAAI、TKDE、WWW等国际重要期刊和会议上发表论文20余篇。

作者简介

李翠平(1971-),女,中国人民大学信息学院教授、副院长、计算机系主任,中国计算机学会杰出会员,大数据专家委员会常务委员,数据库专家委员会委员。2003年于中国科学院计算技术研究所获得博士学位,2008年在美国伊利诺伊大学香槟分校(UIUC)计算机系访问交流。目前研究方向为数据仓库、数据挖掘、社会网络分析和社会媒体推荐等。主持和参与“973”计划、“863”计划、国家自然科学基金等10多项国家级和省部级项目,先后在SIGKDD、SIGMOD、VLDB、EDBT等重要国际会议和TKDE、KAIS、JCST等国内外期刊发表论文100多篇,出版著作2部,申请国家发明专利4项,获得软件著作权登记2项。2007年入选北京市委组织部优秀人才培养资助计划,2009年入选教育部新世纪优秀人才支持计划,2011年获得中国计算机学会“青年科学家奖”。

柴云鹏(1983-),男,博士,中国人民大学信息学院副教授、博士生导师,中国计算机学会信息存储技术专家委员会委员。担任TPDS、TC、TOS、JCST等期刊的评审专家。主要研究方向为云计算、存储系统、分布式系统等。

徐君(1979-),男,博士,中国人民大学教授,主要研究方向为用于互联网搜索的匹配和排序模型,在SIGIR、WWW、AAAI等国际学术会议和TKDE、TOIS、TIST等期刊发表论文50余篇,获专利授权10项;获CIKM'17 Best Full Paper Runner-up和AIRS'10最佳论文奖。担任JASIST编委和SIGIR、AAAI、WWW和ACML等国际会议的高级程序委员会委员。

文继荣(1972-),男,博士,中国人民大学信息学院教授、院长,国家“千人计划”特聘专家。1999年于中国科学院计算技术研究所获得博士学位,同年加入微软亚洲研究院,自2008年起担任高级研究员和互联网搜索与数据挖掘组主任。在微软亚洲研究院工作的14年中,获得50多项美国专利,其中一些成果已经被用于重要的微软产品中(如微软搜索引擎Bing)。领导的研究团队开发出了微软学术搜索、人立方、产品搜索等有影响力的互联网应用。在国际著名会议和期刊上发表了100多篇论文,担任过许多国际会议和研讨会的程序委员会委员和主席。

杜小勇(1963-),男,博士,中国人民大学信息学院教授、博士生导师,教育部数据工程与知识工程重点实验室主任,中国计算机学会会士,《大数据》期刊编委会副主任,ACM Transactions on Data Science编委。主要研究方向为智能信息检索、高性能数据库、知识工程。主持和参与多项国家核高基(核心电子器件、高端通用芯片及基础软件产品)、“973”计划、“863”计划、国家自然科学基金项目,近年来在SIGMOD、VLDB、AAAI、IEEE TKDE等国际重要期刊和会议上发表论文百余篇。

收稿日期:2018-08-20