21 世纪高等院校
云计算和大数据人才培养规划教材

Ruijie Networks

CLOUD COMPUTING AND BIG DATA

大数据技术与应用基础

陈志德 曾燕清 李翔宇 ◎ 编著

- 内容新颖，可操作性强，层层深入，简明易懂
- 以实际行业应用案例讲解大数据处理的方法和计算工具的使用

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

21世纪高等院校
云计算和大数据人才培养规划教材

CLOUD COMPUTING AND BIG DATA

大数据技术与应用基础

陈志德 曾燕清 李翔宇 ◎ 编著

人民邮电出版社
北京

图书在版编目（CIP）数据

大数据技术与应用基础 / 陈志德，曾燕清，李翔宇编著. -- 北京 : 人民邮电出版社，2017.1
21世纪高等院校云计算和大数据人才培养规划教材
ISBN 978-7-115-44347-2

Ⅰ. ①大… Ⅱ. ①陈… ②曾… ③李… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第300291号

内 容 提 要

本书在介绍大数据发展背景、特点及主要技术的基础上，对大数据的数据采集、数据存储、常见计算模式进行了分析介绍。本书同时对各种典型系统工具进行了讲解，包括大数据查询分析计算典型工具（HBase、Hive）、批处理计算典型工具（MapReduce、Spark）、流式计算典型工具（Storm、Apex、Flink）、事件流典型工具（Druid）等。

本书提供了大量的实例和源代码供读者参考，指导读者快速、无障碍地了解和掌握常见大数据分析工具。本书适合作为计算机及相关专业的教学用书，也可以作为大数据初学者的自学教材和参考手册。

◆ 编　　著　陈志德　曾燕清　李翔宇
责任编辑　桑　珊
执行编辑　左仲海
责任印制　焦志炜

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京艺辉印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：13.75　　2017 年 1 月第 1 版
字数：266 千字　　2017 年 1 月北京第 1 次印刷

定价：39.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前 言

近几年，大数据技术迅猛发展，在各个领域都得到广泛关注，推动了新一轮技术发展浪潮。大数据技术的发展，已被列为国家重大发展战略。到2016年为止，大数据已经是第三次出现在政府工作报告中；而在过去的几年里，无论是聚焦大数据发展的《促进大数据发展行动纲要》，还是“十三五”规划中都深刻体现了政府对大数据产业和应用发展的重视。

大数据发展一般会经历扩散期、加速期、转型期、成熟期。目前国内发展还处于第二和第三阶段，数据与传统产业的融合还处于起步阶段，各行业对大数据分析和挖掘的应用还不理想。但随着市场竞争的加剧，各行业对大数据技术研究的热情越来越高，未来几年，各领域的数据分析都将大规模应用。本书在注重大数据时代应用环境前提下，考虑大数据处理分析需求多样、复杂的基本情况，从初学者角度出发，以轻量级理论、丰富的实例介绍大数据常用计算模式的各种系统和工具，注重大数据分析实践操作。本书主要特点如下。

1. 内容丰富多样，对比学习

考虑到当前大数据发展处于起步并逐步赶超先进的阶段，其应用领域丰富广泛，本书除了介绍典型开源大数据处理框架——Apache Hadoop框架之外，还介绍了批处理计算Spark、流式计算及典型工具（Storm、Apex、Flink）和事件流及典型工具（Druid）等，让读者了解不同类型工具系统的特点，并配以丰富简单易上手的实例，让读者能够切实体会和掌握各种类型工具的特点和应用。

2. 轻量级理论，重在培养动手实践能力

为了让读者能够快速掌握技能并保证理论能够适应实践要求，本书本着轻量级理论原则，给出丰富的实例、详实的实验操作步骤，使读者易于配置的实验环境，让读者能够快速上手，在做中学。

3. 有效结合实际应用

除了各章节给出的配套实例外，本书在最后还给出电商领域的大数据分析综合实例，以实际行业应用案例说明大数据处理和计算工具的使用，并进一步阐述大数据行业应用的重大意义。

为了方便读者学习和使用，本书中所有实验操作和实验代码均经过实际运行测试，可直接使用运行。

本书由陈志德、曾燕清、李翔宇共同完成，陈志德统编全稿。由于编者水平有限，书中不妥或错误之处在所难免，不当之处敬请读者批评指正，并将反馈意见发送到邮箱feedbackbigdata@163.com，以便我们及时修正完善。

编者

2016 年 10 月

目 录 CONTENTS

试读结束：需要全本请在线购买： www.ertongbook.com

PART 1

第1章 大数据概述

随着大数据技术的发展，大数据处理及其行业应用价值有目共睹。本章将从大数据发展、大数据的基本概念和特点、大数据的来源、大数据的主要技术层面及大数据的应用等方面简要介绍大数据的基础知识。

1.1 大数据的发展

近年来，随着计算机和信息技术的迅猛发展和普及应用，行业应用系统的规模迅速扩大，行业应用所产生的数据呈爆炸性增长。互联网（社交、搜索、电商）、移动互联网（微博、微信）、物联网（传感器、智慧地球）、车联网、GPS、医学影像、安全监控、金融（银行、股市、保险）、电信（通话、短信）都在疯狂地产生数据。Google 上每天需要处理 24PB 的数据；每个月网民在 Facebook 上要花费 7 000 亿分钟时间，被移动互联网使用者发送和接受的数据量高达 1.3EB；百度目前的总数据量已超过 1000PB，每天需要处理的网页数据达到 10～100PB；每天亚马逊上要产生 630 万笔订单；淘宝累计的交易数据量高达 100PB；Twitter 每天发布超过 2 亿条消息，新浪微博每天发帖量达到 8000 万条；每天会有 2.88 万小时的视频上传到 YouTube；中国移动一个省级公司的电话通联记录数据每月可达 0.5～1PB；一个省会城市公安局道路车辆监控数据 3 年可达 200 亿条、总量 120TB。根据国际数据公司（IDC）的检测，人类产生的数据量正呈指数级增长，大约每两年翻一番，这个速度在 2020 年之前会继续保持，意味着人类在最近两年产生的数据量相当于之前产生的全部数据量。根据 IDC 的测算，到 2020 年数字世界将产生 35 000EB 的数据。行业/企业大数据已远远超出了现有传统的计算技术和信息系统的处理能力，因此，寻求有效的大数据处理技术、方法和手段已经成为现实世界的迫切需求。

前些年人们把大规模数据称为“海量数据”，但大数据（Big Data）的概念早在 2008 年就已被提出。2008 年，《自然》杂志出版了一期专刊，专门讨论未来的大数据处理相关的一系列技术问题和挑战，其中就提出了“Big Data”的概念。

1.2 大数据的概念及特征

1.2.1 大数据的概念

关于大数据，难以有一个非常定量的定义。

麦肯锡对大数据的定义是：大数据指的是那些大小超过标准数据库工具软件能够收集、存储、管理和分析的数据集。

维基百科给出的大数据概念是：在信息技术中，“大数据”是指一些使用目前现有数据库管理工具或者传统数据处理应用很难处理的大型而复杂的数据集。其挑战包括采集、管理、存储、搜索、共享、分析和可视化。

“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。从数据的类别上看，“大数据”指的是无法使用传统流程或工具处理或分析的信息。它定义了那些超出正常处理范围和大小、迫使用户采用非传统处理方法的数据。Gartner 在阐述大数据概念时，提出如上论述。

复旦大学朱扬勇教授提出，大数据本质上是数据交叉、方法交叉、知识交叉、领域交叉、学科交叉，从而产生新的科学研究方法、新的管理决策方法、新的经济增长方式、新的社会发展方式等。

1.2.2 大数据的特征

大数据具备以下四个维度的特征（如图 1-1 所示）：

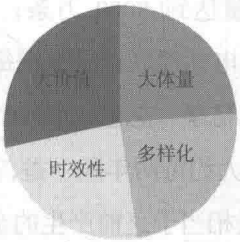

图 1-1 大数据的特征

（1）大体量（Volume）。数据量可从数百 TB 到数百 PB 甚至 EB 的规模。

（2）多样化（Variety）。大数据所处理的数据类型早已不是单一的文本数据或者结构化的数据库中的表，而是包括各种格式和形态的数据，数据结构类型复杂。

（3）时效性（Velocity）。很多大数据需要在一定时间限度下得到及时处理，处理数据的效率决定企业的生命。

（4）大价值（Value）。大数据包含很多深度的价值，通过强大的机器学习和高级分析对数据进行“提纯”，能够带来巨大商业价值。

1.3 大数据的产生及数据类型

1.3.1 大数据的产生

大量数据的产生是计算机技术和网络通信技术普及的必然结果，特别是近年来互联网、云计算、移动互联网、物联网及社交网络等新型信息技术的发展，使得数据产生来源更加丰富。

（1）企业内部及企业外延。企业原有内部系统（如 ERP、OA 等应用系统）所产生的存储在数据库中的数据，属于结构化数据，可直接进行处理使用，为公司决策提供依据。另外，企业内部也存在大量非结构化的内部交易数据，并且随着移动互联网、社交网络等的应用越来越广泛，信息化环境的变化促使企业越来越多的业务需要在互联网、移动互联网、社交网络等平台开展，使得企业外部数据迅速扩展。

（2）互联网及移动互联网。随着社交网络的发展，互联网进入新的时代，用户角色也发生了巨大的变化，从传统的数据使用者转变为随时随地的数据生产者，数据规模迅猛扩展。另外，移动互联网更进一步促进更多用户成为数据生产者。

（3）物联网。物联网技术的发展，使得视频、音频、RFID、M2M、物联网和传感器等产生大量数据，其数据规模更巨大。据 IDC 预测，到 2020 年，由 M2M 产生的数据将占到全世界数据总量的 42%。由此可见物联网产生的数据在整体数据来源中的比重之大。

1.3.2 数据类型

大数据除了数据量巨大外，另一个特点就是数据类型多。在海量数据中，仅有 20%属于结构化数据，其余均为非结构化数据。

按照数据结构，数据可以分为结构化数据、半结构化数据和无结构的非结构化数据。结构化数据存储在数据库中，逻辑结构清晰，易于使用。非结构化数据不方便用数据库二维表来表现，如文档、图片、XML、图像、音频、视频等。非结构化数据中有半结构化数据和无结构化的数据。

按照生产主体，数据可以分为企业应用产生的少量数据、用户产生的大量数据（社交、电商等）、机器产生的巨量数据（应用服务器日志、传感器数据、图像和视频、RFID 等）。

按照数据作用的方式，数据可以分为交易数据和交互数据。海量交易数据指企业内部的经营交易信息，主要包括联机交易数据和联机分析数据，是结构化的、可以通过关系数据库进行管理和访问的静态历史数据。海量交互数据由源于 Facebook、Twitter、微博及其他来源的社交媒体数据构成，包括呼叫详细记录（CDR）、设备和传感信息、GPS 和地理位置映射数据、通过管理文件传输协议传送的海量图像文件、Web 文本和点击流数据、科学信息、电子邮件等。

两类数据的有效融合将是大势所趋，大数据应用要有效集成两类数据，并实现数据的处理和分析。

1.4 大数据计算模式和系统

大数据计算模式，是指根据大数据的不同数据特征和计算特征，从多样性的大数据计算问题和需求中提炼并建立的各种高层抽象和模型。传统的并行计算方法主要从体系结构和编程语言的层面定义了一些较为底层的抽象和模型，但由于大数据处理问题具有很多高层的数据特征和计算特征，因此大数据处理需要更多地结合其数据特征和计算特征考虑更为高层的计算模式。根据大数据处理多样性的需求，出现了各种典型的大数据计算模式，并出现了与之相对应的大数据计算系统和工具。表 1-1 所列为大数据计算模式及典型系统和工具。

表 1-1 大数据计算模式及典型系统和工具

大数据计算模式	典型系统和工具
大数据查询分析计算	HBase，Hive，Cassandra，Premel，Impala，Shark，Hana，Redis 等
批处理计算	MapReduce，Spark 等
流式计算	Scribe，Flume，Storm，S4，Spark Steaming、Apex、Flink 等
迭代计算	HaLoop，iMapReduce，Twister，Spark 等
图计算	Pregel，Giraph，Trinity，PowerGraph，GraphX 等
内存计算	Dremel，Hana，Redis 等

1.5 大数据的主要技术层面和技术内容

从信息系统的角度来看，大数据处理是涉及软硬件系统各个层面的综合信息处理技术。从信息系统角度可以将大数据处理分为基础层、系统层、算法层以及应用层，表 1-2 所列是从信息处理系统角度所看到的大数据技术的主要技术层面和技术内容。

表 1-2 大数据主要技术层面和技术内容

应用层	大数据行业应用/服务层	电信/公安/商业/金融/遥感遥测/勘探/生物医药/教育/政府
		领域应用/服务需求和计算模型
	应用开发层	分析工具/开发环境和工具/行业应用系统开发

试读结束：需要全本请在线购买： www.ertongbook.com

续表

算法层	应用算法层	社交网络、排名与推荐、商业智能，自然语言处理，生物信息媒体分析检索，Web 挖掘与检索，大数据分析与可视化计算……
	基础算法层	并行化机器学习与数据挖掘算法
系统层	并行编程模型与计算框架层	并行计算模型与系统批处理计算，流式计算，图计算，迭代计算，内存计算，混合式计算，定制式计算……
	大数据存储管理	大数据查询（SQL，NoSQL，实时查询，线下分析） 大数据存储（DFS，Hbase，MemD，RDM） 大数据采集（系统日志采集、网络数据采集、其他数据采集）与数据预处理
基础层	并行构架和资源平台层	集群，众核，GPU，混合式架构（如集群+众核，集群+GPU）云计算资源与支撑平台

（1）基础层。基础层主要提供大数据分布存储和并行计算的硬件基础设施。目前大数据处理通用化的硬件设施是基于普通商用服务器的集群，在有特殊的数据处理需要时，这种通用化的集群也可以结合其他类型的并行计算设施一起工作。随着云计算技术的发展，也可以与云计算资源管理和平台结合。

（2）系统层。在系统软件层，需要考虑大数据的采集、大数据的存储管理和并行化计算系统软件几方面的问题。常见大数据数据采集方法主要有系统日志采集法、网络数据采集法和其他数据采集法。大数据处理首先面临的是如何解决大数据的存储管理问题。为了提供巨大的数据存储能力，通常做法是利用分布式存储技术和系统提供可扩展的大数据存储能力。首先需要有一个底层的分布式文件系统，但文件系统通常缺少结构化/半结构化数据的存储管理和访问能力，而且其编程接口对于很多应用来说过于底层。当数据规模增大或者要处理很多非结构化或半结构化数据时，传统数据库技术和系统将难以适用。因此，系统层还需要解决大数据的存储管理和查询问题，因此人们提出了一种 NoSQL 的数据管理查询模式。但最理想的状态还是能提供统一的数据管理查询方法，为此，人们进一步提出了 NewSQL 的概念和技术。解决了大数据的存储问题后，进一步面临的问题是如何能快速有效地完成大规模数据的计算。大数据的数据规模极大，为了提高大数据处理的效率，需要使用大数据并行计算模型和框架来支撑大数据的计算。目前，最主流的大数据并行计算框架是 Hadoop MapReduce 技术。同时，人们开始研究并提出其他的大数据计算模型和方法，如高实时、低延迟的流式计算，针对复杂数据关系的图计算，查询分析类计算，以及面向复杂数据分析挖掘的迭代和

交互计算，高实时、低延迟的内存计算。

（3）算法层。基于以上的基础层和系统层，为了完成大数据的并行化处理，进一步需要考虑的问题是，如何能对各种大数据处理所需要的分析挖掘算法进行并行化设计。

（4）应用层。基于上述三个层面，可以构建各种行业或领域的大数据应用系统。

1.6 大数据的典型应用

医疗大数据。医疗行业拥有大量的病例、病理报告、治愈方案、药物报告等，如果这些数据可以被整理和应用，将会极大地帮助医生和病人。如果未来基因技术发展成熟，还可以根据病人的基因序列特点进行分类，建立医疗行业的病人分类数据库。在医生诊断病人时可以参考病人的疾病特征、化验报告和检测报告，参考疾病数据库来快速帮助病人确诊，明确定位疾病。同时，这些数据也有利于医药行业开发出更加有效的药物和医疗器械。

生物大数据。自人类基因组计划完成以来，以美国为代表，世界主要发达国家纷纷启动了生命科学基础研究计划，如国际千人基因组计划、DNA 百科全书计划、英国十万人基因组计划等，这些计划引领生物数据呈爆炸式增长。目前，每年全球产生的生物数据总量已达 EB 级，生命科学领域正在爆发一次数据革命，生命科学某种程度上已经成为大数据科学。

金融大数据。大数据在金融行业的应用可以总结为精准营销、风险管控、决策支持、效率提升、产品设计五个方面。

零售大数据。未来考验零售企业的是挖掘消费者需求及高效整合供应链满足其需求的能力，因此信息科技水平的高低成为获得竞争优势的关键要素。

电商大数据。由于电商的数据较为集中，数据量足够大，数据种类较多，因此未来电商数据应用将会有更多的想象空间，包括预测流行趋势、消费趋势、地域消费特点、客户消费习惯、各种消费行为的相关度、消费热点、影响消费的重要因素等。

农牧大数据。大数据在农业应用主要是指依据未来商业需求的预测来进行农牧产品生产，降低菜贱伤农等的概率。同时，大数据的分析将会更加精确预测未来的天气气候，帮助农牧民做好自然灾害的预防工作；可以通过大数据帮助农民依据消费者消费习惯决定农作物生产的种类和数量，提高单位种植面积的产值；可以通过大数据分析来帮助牧民安排放牧范围，有效利用牧场；可以利用大数据帮助渔民安排休渔期、定位捕鱼范围等。

交通大数据。目前，交通的大数据应用主要在两个方面；一方面可以利用大数据传感器数据来了解车辆通行密度，合理进行道路规划（包括单行线路规划）；另一方面可以利用大量数据来实现即时信号灯调度，提高已有线路运行能力。科学地安排信号灯是一个复杂的系统工程，必须利用大数据计算平台才能计算出一个较为合理的方案。机场的航班起降依靠大数

据将会提高航班管理的效率，航空公司利用大数据可以提高上座率，降低运行成本。铁路利用大数据可以有效安排客运和货运列车，提高效率、降低成本。

教育大数据。毫无疑问，在不远的将来，无论是教育管理部门，还是校长、教师、学生和家长，都可以得到针对不同应用的个性化分析报告。通过大数据的分析来优化教育机制，也可以做出更科学的决策，这将带来潜在的教育革命。

体育大数据。大数据对于体育的改变可以说是方方面面。对运动员而言，可通过穿戴设备收集的数据更了解身体状况；对媒体评论员而言，通过大数据提供的数据可以更好地解说比赛、分析比赛。

环保大数据。借助于大数据技术，天气预报的准确性和实效性将会大大提高，预报的及时性将会大大提升，同时对于重大自然灾害，如龙卷风，通过大数据计算平台，将会更加精确地了解其运动轨迹和危害的等级，有利于帮助大众提高应对自然灾害的能力。

食品大数据。随着科学技术和生活水平的不断提高，食品添加剂及食品品种越来越多，传统手段难以满足当前复杂的食品监管需求，从不断出现的食品安全问题来看，食品监管成了食品安全的棘手问题。通过大数据管理将海量数据聚合在一起，将离散的数据需求聚合能形成数据长尾，从而满足传统中难以实现的需求。

政府调控和财政支出。政府利用大数据技术可以了解各地区的经济发展情况、各产业发展情况、消费支出和产品销售情况，依据数据分析结果，科学地制定宏观政策，平衡各产业发展，避免产能过剩，有效利用自然资源和社会资源，提高社会生产效率。大数据还可以帮助政府进行监控自然资源的管理，无论是国土资源还是水资源、矿产资源、能源等，大数据都可以通过各种传感器来提高其管理的精准度。同时，大数据技术也能帮助政府进行支出管理，透明合理的财政支出将有利于提高公信力和监督财政支出。

舆情监控大数据。国家正在将大数据技术用于舆情监控，其收集到的数据除了解民众诉求、降低群体事件之外，还可以用于犯罪管理。大量的社会行为正逐步走向互联网，人们更愿意借助互联网平台来表述自己的想法和宣泄情绪。国家可以通过社交媒体分享的图片和交流的信息来收集个体情绪信息，预防个体犯罪行为和反社会行为。

1.7 本章小结

本章对大数据及其分析框架整体流程进行了简单介绍，从大数据数据来源、数据采集、数据存储、数据计算、数据分析及应用等几方面介绍大数据的基本概念。

PART 2

第2章 数据获取

数据是大数据分析与应用的前提和基础，利用网络爬虫进行数据获取是非常高效的方法之一。在本书中，实验所用数据均采用编写的网络爬虫进行获取。本章将简单介绍 Scrapy 爬虫框架，并通过一个爬虫实例介绍运用该框架如何编写一个爬虫的全过程。

2.1 Scrapy 环境搭建

所需环境：

- Python 2.7
- lxml-3.5.0
- pyOpenSSL-0.13.1
- pywin32-219
- setuptools-0.7
- twisted-15.4.0
- zope.interface-4.1.3
- Scrapy-1.0

搭建时，应先搭建 Python 和其他几个环境，最后安装 Scrapy 环境。对依赖包的选择需要根据 Python 的位数（32 位或 64 位）进行，以避免兼容性问题，同时要注意环境变量的设置。

2.2 爬虫项目创建

单击“开始”菜单，输入“cmd”,进入计算机的命令行操作模式（Windows 7 操作系统的进入方式），运行图 2-1 所示的命令，进入到爬虫代码所需存放的目录，然后运行如下命令：

```
scrapy startproject SinanewsSpider
```

其中，SinanewsSpider 为所创建的爬虫项目的名称。此时在相应的目录下出现 SinanewsSpider 爬虫项目，如图 2-2 所示。

在项目路径下的 SinanewsSpider→SinanewsSpider→spiders 文件夹下，可以创建属于自己的

爬虫文件。如图 2-3 所示，我们建立一个自己的爬虫，文件名为 SinanewsSpider.py。爬虫代码则主要是在所建的爬虫文件中。

另外，在 SinanewsSpider→SinanewsSpider 路径下，文件 items.py、pipelines.py 以及 settings.py 都是后续需要使用到的文件，我们将在本章后续小节中依次进行介绍。

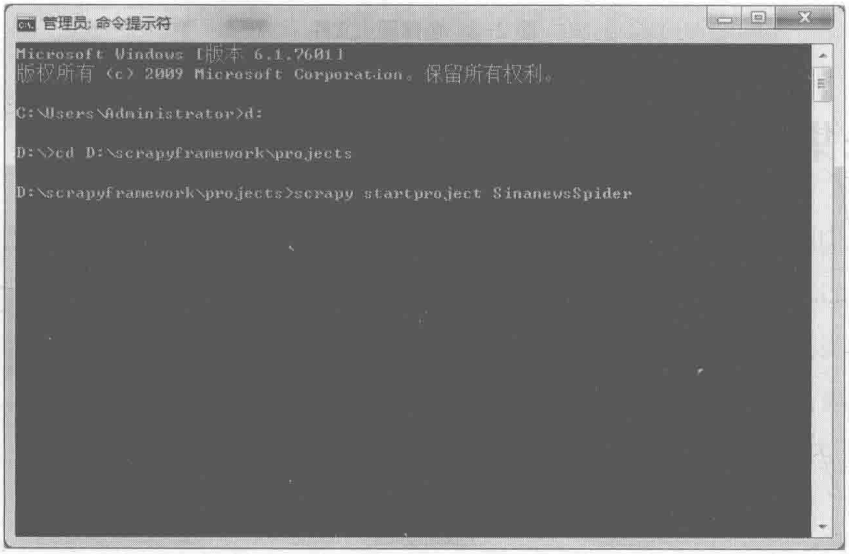

图 2-1　创建爬虫项目

图 2-2　爬虫项目

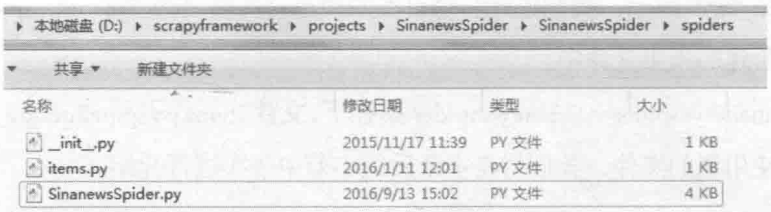

图 2-3 创建爬虫文件

2.3 采集目标数据项定义

这里我们以采集新浪本地新闻为例，介绍一个爬虫实例的实现过程。新闻的列表页地址为 http://roll.news.sina.com.cn/news/gnxw/gdxw1/index.shtml，如图 2-4 所示。我们的采集目标是从该列表页中获取所有列表新闻的链接地址，并访问各条新闻的详情页，爬取各条新闻的详细数据项，需要的数据项定义为：

（标题，内容，时间，图片链接地址，网页链接地址，发表时间）

图 2-4 新浪本地新闻列表页

并且，数据爬取完成后，我们希望存储到数据库中，这里我们以 MySQL 数据库为例。首先建立存储新闻的数据库表，详细如图 2-5 所示。

接下来，我们介绍在爬虫项目中的 items.py 文件。当我们确定好需要采集的目标数据之

试读结束：需要全本请在线购买： www.ertongbook.com