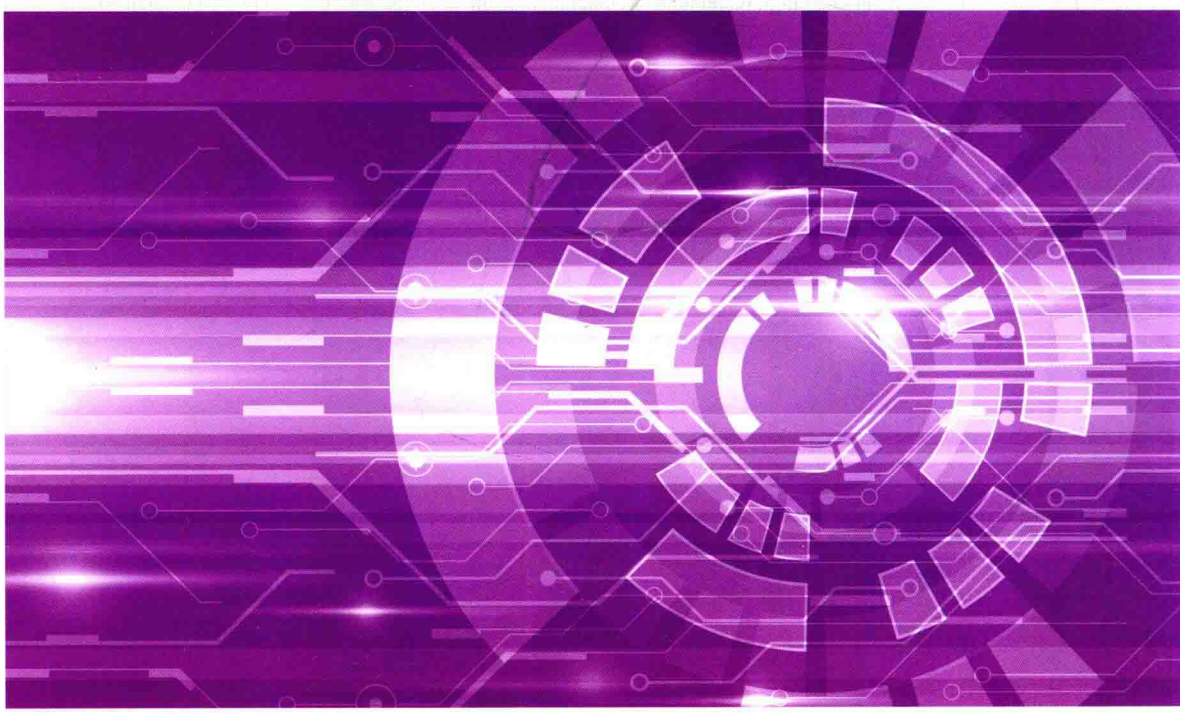
• 大数据应用人才培养系列教材 •

数据清洗

■ 总主编◎刘　鹏　张　燕　■ 主编◎李法平　■ 副主编◎陈潇潇



清华大学出版社

大数据应用人才培养系列教材

数据清洗

总主编　刘　鹏　张　燕
主　编　李法平
副主编　陈潇潇

清华大学出版社
北　京

内 容 简 介

数据清洗是大数据领域不可缺少的环节，用来发现并纠正数据中可能存在的错误，针对数据审查过程中发现的错误值、缺失值、异常值、可疑数据，选用适当方法进行“清理”，使“脏”数据变为“干净”数据。

本书共分为8章：第1章主要介绍数据清洗的概念、任务和流程，数据标准化概念及数据仓库技术等；第2章主要介绍Windows和类UNIX操作系统下的数据常规格式、数据编码及数据类型转换等；第3章介绍ETL概念、数据清洗的技术路线、ETL工具及ETL子系统等；第4章介绍Excel、Kettle、OpenRefine、DataWrangler和Hawk的安装及使用等；第5章介绍Kettle下文本文件抽取、Web数据抽取、数据库数据抽取及增量数据抽取等；第6章介绍数据清洗步骤、数据检验、数据错误处理、数据质量评估及数据加载；第7章介绍网页结构，利用网络爬虫技术进行数据采集，利用JavaScript技术进行行为日志数据采集等；第8章介绍RDBMS的数据清洗方法和数据脱敏处理技术等。

本书系统地讲解了数据清洗理论和实际应用，适用于高职高专院校和应用型本科的大数据课程教学，也适用于希望了解数据清洗的广大读者。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。侵权举报电话：010-62782989　13701121933

图书在版编目（CIP）数据

数据清洗/李法平主编. —北京：清华大学出版社，2018
（大数据应用人才培养系列教材）
ISBN 978-7-302-49327-3

I. ①数… II. ①李… III. ①数据处理-技术培训-教材 IV. ①TP274

中国版本图书馆CIP数据核字（2018）第004243号

责任编辑：贾小红
封面设计：刘　超
版式设计：刘艳庆
责任校对：赵丽杰
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　**邮　　编：**100084
社 总 机：010-62770175　　**邮　　购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张：**15.75　　**字　　数：**280千字
版　　次：2018年6月第1版　　**印　　次：**2018年6月第1次印刷
印　　数：1～2500
定　　价：58.00元

产品编号：075032-01

编写委员会

总主编　刘　鹏　张　燕

主　编　李法平

副主编　陈潇潇

编　委　付　雯　葛　斌　秦　毅　王海涛

文　华　徐佩锋　于　澄　岳宗辉

朱堂勋

总 序

短短几年间，大数据就以一日千里的发展速度，快速实现了从概念到落地，直接带动了相关产业的井喷式发展。数据采集、数据存储、数据挖掘、数据分析等大数据技术在越来越多的行业中得到应用，随之而来的就是大数据人才缺口问题的凸显。根据《人民日报》的报道，未来3～5年，中国需要180万数据人才，但目前只有约30万人，人才缺口达到150万之多。

大数据是一门实践性很强的学科，在其金字塔型的人才资源模型中，数据科学家居于塔尖位置，然而该领域对于经验丰富的数据科学家需求相对有限，反而是对大数据底层设计、数据清洗、数据挖掘及大数据安全等相关人才的需求急剧上升，可以说占据了大数据人才需求的80%以上。比如数据清洗、数据挖掘等相关职位，需要源源不断的大量专业人才。

迫切的人才需求直接催热了相应的大数据应用专业。2018年1月18日，教育部公布了“大数据技术与应用”专业备案和审批结果，已有270所高职院校申报开设“大数据技术与应用”专业，其中共有208所职业院校获批“大数据技术与应用”专业。随着大数据的深入发展，未来几年申请与获批该专业的职业院校数量仍将持续走高。同时，对于国家教育部正式设立的“数据科学与大数据技术”本科新专业，除已获批的35所大学之外，2017年申请院校也高达263所。

即使如此，就目前而言，在大数据人才培养和大数据课程建设方面，大部分专科院校仍然处于起步阶段，需要探索的问题还有很多。首先，大数据是个新生事物，懂大数据的老师少之又少，院校缺“人”；其次，院校尚未形成完善的大数据人才培养和课程体系，缺乏“机制”；再次，大数据实验需要为每位学生提供集群计算机，院校缺“机器”；最后，院校没有海量数据，开展大数据教学实验工作缺少“原材料”。

对于注重实操的大数据技术与应用专业专科建设而言，需要重点面向网络爬虫、大数据分析、大数据开发、大数据可视化、大数据运维工程师的工作岗位，帮助学生掌握大数据技术与应用专业必备知识，使其具备大数据采集、存储、清洗、分析、开发及系统维护的专

试读结束：需要全本请在线购买： www.ertongbook.com

业能力和技能，成为能够服务区域经济的发展型、创新型或复合型技术技能人才。无论是缺“人”、缺“机制”、缺“机器”，还是缺少“原材料”，最终都难以培养出合格的大数据人才。

其实，早在网格计算和云计算兴起时，我国科技工作者就曾遇到过类似的挑战，我有幸参与了这些问题的解决过程。为了解决网格计算问题，我在清华大学读博期间，于 2001 年创办了中国网格信息中转站网站，每天花几个小时收集和分享有价值的资料给学术界，此后我也多次筹办和主持全国性的网格计算学术会议，进行信息传递与知识分享。2002 年，我与其他专家合作的《网格计算》教材正式面世。

2008 年，当云计算开始萌芽之时，我创办了中国云计算网站（chinacloud.cn）（在各大搜索引擎“云计算”关键词中排名第一），2010 年出版了《云计算（第 1 版）》，2011 年出版了《云计算（第 2 版）》，2015 年出版了《云计算（第 3 版）》，每一版都花费了大量成本制作并免费分享对应的几十个教学 PPT。目前，这些 PPT 的下载总量达到了几百万次之多。同时，《云计算》一书也成为国内高校的优秀教材，在中国知网公布的高被引图书名单中，《云计算》在自动化和计算机领域排名全国第一。

除了资料分享，在 2010 年，我们在南京组织了全国高校云计算师资培训班，培养了国内第一批云计算老师，并通过与华为、中兴、360 等知名企业合作，输出云计算技术，培养云计算研发人才。这些工作获得了大家的认可与好评，此后我接连担任了工信部云计算研究中心专家、中国云计算专家委员会云存储组组长、中国大数据应用联盟人工智能专家委员会主任等。

近几年，面对日益突出的大数据发展难题，我们也正在尝试使用此前类似的办法去应对这些挑战。为了解决大数据技术资料缺乏和交流不够通透的问题，我们于 2013 年创办了中国大数据网站（thebigdata.cn），投入大量的人力进行日常维护，该网站目前已经在各大搜索引擎的“大数据”关键词排名中位居第一；为了解决大数据师资匮乏的问题，我们面向全国院校陆续举办多期大数据师资培训班，致力于解决“缺人”的问题。

2016 年年末至今，我们已在南京多次举办全国高校/高职/中职大数据免费培训班，基于《大数据》《大数据实验手册》以及云创大数据提供的大数据实验平台，帮助到场老师们跑通了 Hadoop、Spark 等多个大数据实验，使他们跨过了“从理论到实践，从知道到用过”的门槛。

其中，为了解决大数据实验难问题而开发的大数据实验平台，正在为越来越多的高校教学科研带去方便，帮助解决缺“机器”与缺“原材料”的问题。2016 年，我带领云创大数据（www.cstor.cn，股票代码：835305）的科研人员，应用 Docker 容器技术，成功开发了 BDRack 大数据实验一体机，它打破了虚拟化技术的性能瓶颈，可以为每一位参加实验的人员虚拟出 Hadoop 集群、Spark 集群、Storm 集群等，自带实验所需数据，并准备了详细的实验手册（包含 42 个大数据实验）、PPT 和实验过程视频，可以开展大数据管理、大数据挖掘等各类实验，并可进行精确营销、信用分析等多种实战演练。

目前，大数据实验平台已经在郑州大学、成都理工大学、金陵科技学院、天津农学院、西京学院、郑州升达经贸管理学院、信阳师范学院、镇江高等职业技术学校等多所院校部署应用，并广受校方好评。该平台也可以云服务的方式在线提供（大数据实验平台，https://bd.cstor.cn），实验更是增至 85 个，师生通过自学，可用一个月时间成为大数据实验动手的高手。此外，面对席卷而来的人工智能浪潮，我们团队推出的 AIRack 人工智能实验平台、DeepRack 深度学习一体机以及 dServer 人工智能服务器等系列应用，一举解决了人工智能实验环境搭建困难、缺乏实验指导与实验数据等问题，目前已经在清华大学、南京大学、南京农业大学、西安科技大学等高校投入使用。

在大数据教学中，本科院校的实践教学应更加系统性，偏向新技术的应用，且对工程实践能力要求更高。而高职、高专院校则更偏向于技术性和技能训练，理论以够用为主，学生将主要从事数据清洗和运维方面的工作。基于此，我们联合多家高职院校专家准备了《云计算导论》《大数据导论》《数据挖掘基础》《R 语言》《数据清洗》《大数据系统运维》《大数据实践》系列教材，帮助解决“机制”欠缺的问题。

此外，我们也将继续在中国大数据（thebigdata.cn）和中国云计算（chinacloud.cn）等网站免费提供配套 PPT 和其他资料。同时，持续开放大数据实验平台（https://bd.cstor.cn）、免费的物联网大数据托管平台万物云（wanwuyun.com）和环境大数据免费分享平台环境云（envicloud.cn），使资源与数据随手可得，让大数据学习变得更加轻松。

在此，特别感谢我的硕士导师谢希仁教授和博士导师李三立院士。谢希仁教授所著的《计算机网络》已经更新到第 7 版，与时俱进，日臻完美，时时提醒学生要以这样的标准来写书。李三立院士是留苏

博士，为我国计算机事业做出了杰出贡献，曾任国家攀登计划项目首席科学家。他的严谨治学带出了一大批杰出的学生。

本丛书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者致敬！书中难免会有不当之处，请读者不吝赐教。我的邮箱：gloud@126.com，微信公众号：刘鹏看未来（lpoutlook）。

刘　鹏

于南京大数据研究院

2018 年 5 月

前 言

随着信息技术的发展和科技的进步，人类步入了大数据时代。作为当前高科技时代的产物，大数据由大量结构化、半结构化和非结构化数据组成，它需要经过采集、清洗、存储、分析、建模、可视化等过程加工处理之后，才能真正产生价值。数据清洗是大数据技术不可缺少的环节，用来发现并纠正数据中可能存在的错误，针对数据审查过程中发现的错误值、缺失值、异常值、可疑数据，选用适当方法进行“清理”，把“脏”的数据变为“干净”的数据。

本书共分 8 章，下面分别对每章内容进行简单介绍。

第 1 章主要介绍数据清洗的概念、任务和流程，数据标准化概念及数据仓库技术等知识点。通过本章的学习，读者能够初步认识数据清洗、数据标准化及数据仓库。

第 2 章为数据格式及编码，主要介绍 Windows 和类 UNIX 操作系统下的数据常规格式，如文本格式、xls 及 xlsx 格式、JSON、XML、HTML 等，并针对数据的类型、数据编码及字符集进行了阐述，最后介绍格式间的相互转换，包括电子表格转换、数据库数据转换等。通过本章的学习，了解当前主流的数据格式、数据编码及格式间相互转换等知识。

第 3 章为数据清洗基本技术方法。本章从 ETL 技术出发，介绍 ETL 概念、数据清洗的技术路线、ETL 工具及 ETL 子系统等知识。通过本章的学习，进一步了解数据清洗的概念、技术路线及主要功能。

第 4 章为数据清洗常用工具及基本操作。介绍了 Microsoft Excel 数据清洗操作步骤、Kettle 安装使用及操作步骤、OpenRefine 的安装使用及操作步骤、DataWrangler 的安装使用及操作步骤、Hawk 网页数据采集的方法及操作实例。通过本章的学习，掌握当前市面主流的数据清洗工具的使用，为后面进行数据清洗做必要的准备工作。

第 5 章为数据抽取。本章以 Kettle 开源工具为载体，介绍文本文件抽取、Web 数据抽取、数据库数据抽取及增量数据抽取等知识。通过本章的学习，能够掌握借助 Kettle 实现文本文件抽取、网页文本抽取、数据库数据的导入导出、关系数据库到 NoSQL 的抽取转换及增量抽取等。

第 6 章为数据转换与加载。本章详细介绍数据清洗步骤、数据检验、错误处理、数据质量评估及数据装载等知识。通过本章的学习，掌握数据清洗具体方法和数据转换过程中的数据检验、错误处理等，以及数据加载和批量加载技术。

第 7 章为采集 Web 数据实例，介绍了网页结构、网络爬虫、行为日志数据采集等知识。通过本章的学习，了解网络爬虫技术采集 Web 数据的方法以及行为日志分析方法。

第 8 章为清洗 RDBMS 数据实例，介绍了 RDBMS 的数据清洗方法和数据脱敏处理技术，使读者进一步掌握关系型数据库清洗方法和敏感数据脱敏处理技巧。

本书的编写和整理工作由数据清洗教材编写组和南京云创大数据科技股份有限公司完成，主要参与人员有王海涛、于澄、岳宗辉、徐佩锋、秦毅、葛斌、文华、朱堂勋、陈潇潇、付雯等。全体成员在近一年的编写过程中付出了辛勤的汗水，在此由衷感谢。本书的问世也要感谢清华大学出版社王莉编辑给予的宝贵意见和支持。

尽管我们付出了最大的努力，但教材中难免存在不妥之处，欢迎各界专家和读者朋友提出宝贵意见，我们将不胜感谢。您在阅读本书时，如发现任何问题或不认同之处，可以通过电子邮件与我们联系。

请发送邮件至：DataCleaning@163.com。

李法平

2017 年 12 月

目 录

试读结束：需要全本请在线购买： www.ertongbook.com

第1章

数据清洗概述

在当今信息技术时代，大数据堪称是一项伟大技术，它改变了传统的数据收集、处理和应用模式，为众多领域的跨越式发展带来了新的机遇和挑战。大数据的战略价值不是追求掌握庞大的数据量，而在于对这些富有内涵的数据进行专业化处理，获取具有更强决策力、洞察力和流程优化能力的信息资产，进而指导科学决策和生产实践。人类在努力将数据转化为信息和知识的同时，也面临着海量数据中夹杂着“脏”数据的挑战。因此，对原始数据进行有效清洗并将其转化为易理解和易利用的目标数据，已成为人类进行大数据分析和应用过程中的关键一环。数据清洗（Data Cleaning）用来对数据进行审查和校验，进而删除重复信息，纠正存在的错误，并保持数据的一致性、精确性、完整性和有效性。由此可见，数据清洗在整个大数据分析过程中扮演着重要的角色。作为本书的引子，本章主要阐述数据清洗的基本概念和相关技术。

1.1 数据清洗简介

1.1.1 数据科学过程

现代社会的各个角落无不充斥着种类繁多、数量庞大的数据，这些数据不仅包括传统的结构型数据，还包括如网页、文本、图像、视频、语音之类的非结构型数据。大数据的兴起和研究热潮将数据科学推到风口浪尖。大数据不仅是一门技术，更代表了一种潮流和一个时代，而数

据科学则是一门新兴的以数据为研究中心的学科。作为一门学科，数据科学以数据的广泛性和多样性为基础，探寻数据研究的共性。例如，自然语言处理和生物大分子模型里都用到了隐式马氏过程和动态规划方法，其根本原因是它们处理的都是一维的随机信号。再如图像处理和统计学习中都用到正则化方法，因此用于图像处理的算法和用于压缩感知的算法有着许多共同之处。（参见文末参考文献[1]）

数据科学是一门关于数据的工程，它需要同时具备理论基础和工程经验，需要掌握各种工具的用法。数据科学主要包括两个方面，即用数据的方法来研究科学和用科学的方法来研究数据。前者包括生物信息学、天体信息学、数字地球等领域；后者包括统计学、机器学习、数据挖掘、数据库等领域。这些学科都是数据科学的重要组成部分，但只有把它们有机地放在一起，才能形成整个数据科学的全貌。数据科学的综合性也对数据科学家们提出了较高的技能要求，他们需要掌握的知识包括计算机、统计学、数据处理和数据可视化等。（参见文末参考文献[2]）

数据清洗是数据科学家完成数据分析和处理任务过程中必须面对的重要一环。具体来说，数据科学的一般处理过程包括如下几个步骤。

（1）问题陈述：明确需要解决的问题和任务。

（2）数据收集与存储：通过多种手段采集和存放来自众多数据源的数据。

（3）数据清洗：对数据进行针对性的整理和规范，以便于后面的分析和处理。

（4）数据分析和挖掘：运用特定模型和算法来寻求数据中隐含的知识和规律。

（5）数据呈现和可视化：以恰当的方式呈现数据分析和挖掘的结果。

（6）科学决策：根据数据分析和处理结果来决定问题的解决方案。

需要指出的是，上述数据科学过程的 6 个步骤并非全部需要，而且上述步骤的执行是一个反复迭代的过程。例如，在一个数据分析项目中可能需要不止一次地执行数据清洗和数据呈现操作。此外，数据分析和挖掘方法会影响数据清洗的手段和方式。

1.1.2 数据清洗定义

来自多样化数据源的数据内容并不完美，存在着许多“脏”数据，即数据不完整、存在错误和重复的数据、数据的不一致和冲突等缺陷。统计资料表明，“脏”数据大约占到总数据量的 5%，“脏”数据会对建立的数据处理和应用系统造成不良影响，扭曲从数据中获得的信息，影

响数据应用系统的运行效果，进一步影响数据挖掘效能，最终影响决策管理。为了减少这些“脏”数据对数据分析和挖掘结果的影响，必须采取各种有效的措施对采集的原始数据进行有效的预处理，这一预处理过程称为“数据清洗（Data Cleaning/Cleansing）”，即在数据集中发现不准确、不完整或不合理数据，并对这些数据进行修补或移除以提高数据质量的过程。

目前，对于数据清洗并没有统一的定义，其定义依赖于具体的应用领域。从广义上讲，数据清洗是将原始数据进行精简以去除冗余和消除不一致，并使剩余的数据转换成可接收的标准格式的过程；而狭义上的数据清洗特指在构建数据仓库和实现数据挖掘前对数据源进行处理，使数据实现准确性、完整性、一致性、唯一性和有效性以适应后续操作的过程。一般而言，凡是有助于提高信息系统数据质量的处理过程，都可认为是数据清洗。简单地说就是从数据源中清除错误数值和重复记录，即利用特定技术和手段，基于预定义的清洗规则从数据源中检测和消除错误数据、不完整数据和重复数据，从而提高信息系统的数据质量。

1.1.3　数据清洗任务

数据清洗就是对原始数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并使得数据保持精确性、完整性、一致性、有效性及唯一性，还可能涉及数据的分解和重组，最终将原始数据转换为满足数据质量或应用要求的数据。对于任何大数据项目而言，数据清洗过程都是必不可少的步骤。此外，格式检查、完整性检查、合理性检查和极限检查也在数据清洗过程中完成。数据清洗对保持数据的一致和更新起着重要的作用，因此被用于如银行、保险、零售、电信和交通等多个行业。（参见文末参考文献[3]）

当前，数据清洗主要有 3 个应用领域：数据仓库（Data Warehouse，DW）、数据库中知识的发现（Knowledge Discovery in Database，KDD）和数据质量管理（Data Quality Management，DQM）。

在数据仓库领域中，当对多个数据库合并时或多个数据源进行集成时需要使用数据清洗。例如，当同一个实体的记录在不同数据源中以不同格式表示或被错误表示的情况下，合并后的数据库中就会出现重复的记录。数据清洗就需要识别出重复的记录并消除它们，也就是所谓的数据合并/清除（Merge/Purge）问题。在数据仓库环境中，需要考虑数据仓库的集成性与面向主题的需要，包括数据的清洗及结构转换。

在数据库中的知识发现领域，数据清洗为 KDD 过程的首个步骤，

试读结束：需要全本请在线购买：www.ertongbook.com