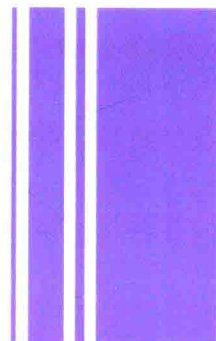
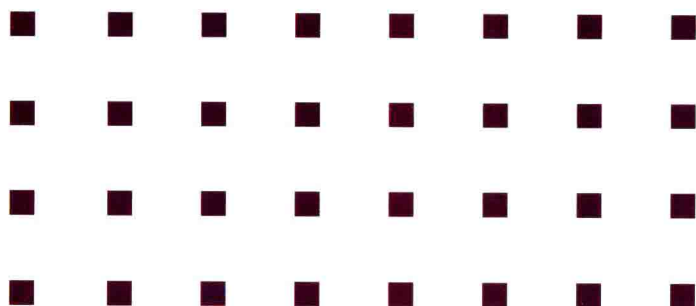
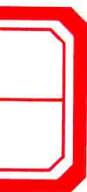


高等学校大数据系列教材



数据清洗基础 与实践

主 编 谢东亮 黄天春
副主编 廖 宁 陈怡然
刘振栋 王胜峰
主 审 周龙福



西安电子科技大学出版社
<http://www.xduph.com>

高等学校大数据系列教材

数据清洗基础与实践

主 编 谢东亮 黄天春

副主编 廖 宁 陈怡然

刘振栋 王胜峰

主 审 周龙福

西安电子科技大学出版社

内 容 简 介

数据清洗是大数据开发的基础环节,也是最重要的环节之一。数据清洗就是发现并纠正数据源中存在的错误,对错误值、异常值、缺失值等可疑数据按照一定的规则和方法,使用必要的工具进行清洗与整理,使其变得干净,达到可以进行数据分析的目的。

本书分为两篇,共九章,系统地讲解了数据清洗的理论与实践。第一篇为基础篇,其中第一章介绍数据清洗的基本概念、数据标准化以及数据仓库的概念等;第二章介绍数据格式类型与数据编码,从文件以及文本的格式、文档的归档与压缩、字符编码与数据转换方面做了介绍;第三章对最常用的文本与电子表格数据清洗做了介绍;第四章介绍了数据清洗的基本技术、方法与工具;第五章从文本、Web 文件、数据库数据的抽取方面做了介绍;第六章介绍了数据的转换与加载和数据的质量评估。第二篇为实战篇,其中第七章介绍了五种常用的数据清洗工具;第八章专门讲解了如何对 Web 的数据进行采集及预处理;第九章介绍了从 RDBMS 数据源中获取数据及进行数据的脱敏、清洗等操作。

本书可作为应用型本科及高职高专学生的大数据课程教材,也可供学习数据清洗的其他读者使用。

图书在版编目(CIP)数据

数据清洗基础与实践/谢东亮,黄天春主编. —西安:西安电子科技大学出版社,2019.9

ISBN 978-7-5606-5460-7

I. ① 数… II. ① 谢… ② 黄… III. ① 数据处理—高等学校—教材 IV. ① TP274

中国版本图书馆 CIP 数据核字(2019)第 190579 号

策划编辑 李惠萍

责任编辑 雷鸿俊 任倍萱

出版发行 西安电子科技大学出版社(西安市太白南路2号)

电 话 (029)88242885 88201467 邮 编 710071

网 址 www.xduph.com 电子邮箱 xduphxb001@163.com

经 销 新华书店

印刷单位 陕西天意印务有限责任公司

版 次 2019年8月第1版 2019年9月第1次印刷

开 本 787毫米×1092毫米 1/16 印 张 8.5

字 数 198千字

印 数 1~3000册

定 价 22.00元

ISBN 978-7-5606-5460-7 / TP

XDUP 5762001-1

如有印装问题可调换

前 言

大数据是当今信息化的产物，作为当前高科技时代的产物，大量的结构化、非结构化、半结构化的数据经过采集、清洗处理后才便于分析、建模等。这样的数据才能真正地产生价值。数据清洗技术是大数据必不可少的环节，通过数据清洗可以发现并纠正数据中可能存在的错误，并对发现的错误进行清理，使之变为干干净净的数据。

本书分为基础篇与实战篇两部分。在基础篇中主要讲述了数据清洗的流程以及在清洗过程中所使用的工具和数据清洗的方法，在实战篇中主要介绍了数据清洗工具的使用方法、基于 Web 的数据采集与清洗实战和基于 DBMS 的数据清洗实战。

基础篇包括第一章至第六章，其中：

第一章主要阐述数据清洗的基本概念，在数据清洗阶段的主要任务以及流程、数据的标准化以及为什么要进行数据清洗；通过一个简单的例子演示了现实世界中的脏数据，并阐述了各种清洗过程的优缺点以及清洗工作所带来的数据变化。

第二章主要阐述数据格式、编码、常见的数据类型以及字符集等，介绍格式间的相互转换。通过本章的学习，读者可了解当前流行的主要数据格式以及数据编码与格式间的相互转换等知识。

第三章描述如何从常见的电子表格和文本编辑器中尽可能多地发掘出数据清洗功能，主要介绍一些常见问题的简便处理方法，包括如何使用函数、搜索和替换、正则表达式来实现数据纠错和转换；最后利用已经掌握的技能，使用上述两种工具来完成一个与大学有关的数据清洗任务。

第四章从 ETL 技术入手，介绍 ETL 的相关概念、数据清洗遵从的基本步骤以及技术路线，介绍 ETL 常见的工具以及 ETL 子系统。通过本章的学习，读者能清晰地理解数据清洗的概念、技术路线及主要功能。

第五章以 Kettle 开源数据清洗工具为依托，阐述文本文件的抽取、Web 文件的获取、数据库文件的获取，以及全量与增量数据的抽取等。通过本章

的学习，读者能够借助 Kettle 实现文本的抽取、网页文件的获取以及数据库文件的导入与导出等。

第六章详细介绍数据的清洗步骤、数据检查、错误处理、数据质量评估和数据装载等知识。通过本章的学习，读者能够掌握数据清洗的具体方法以及数据转换过程中的数据检验和错误处理，同时掌握数据的批量加载等技能。

实战篇包括第七章至第九章，其中：

第七章主要阐述数据清洗使用的工具，从基础的 Microsoft Excel 工具进行讲解，阐述如何使用 Excel 来完成数据的分列校验和快速填充等操作以及使用 Kettle 进行数据的清洗操作。同时，对常用的清洗工具 OpenRefine 以及 DataWrangler 和 Hawk 等工具进行简单的介绍，使读者对常用的清洗工具有一个全面的认识。

第八章主要以常用的 Web 数据为基础，采用实际案例介绍如何对数据进行爬取采集以及清洗等预处理操作。

第九章主要以关键的关系型数据库为基础，采用实际案例介绍如何从关系型数据库中获取数据并进行准备、脱敏、格式转换等清洗工作，通过案例的实际讲解使读者掌握如何对关系型数据库的大量数据进行“清洗”操作。

在本书的编写过程中，北京雅丁信息科技有限公司提供了大量的支持，并给出了中肯的建议。黄天春老师除了编写第九章的内容之外还对全书进行了反复校验，才使本书得以成稿，在此表示衷心的感谢。同时，也要感谢西安电子科技大学出版社李惠萍编辑在本书出版过程中所给予的宝贵意见和大力支持。

尽管我们付出了很大的努力，但是书中可能仍存在不妥之处，欢迎读者朋友提出宝贵意见，我们将不胜感激。在阅读本书时，如果您发现有任何问题，可以通过邮件与我们联系，邮箱地址：donaldshieh@yeah.net。

谢东亮

2019 年 6 月

目 录

第一篇 基 础 篇

第一章 数据清洗概述.....	3
1.1 数据清洗简介	3
1.2 数据标准化	4
1.3 数据仓库	5
本章小结	7
第二章 数据格式类型与编码.....	8
2.1 文件及文本的格式.....	8
2.2 归档与压缩	10
2.3 数据类型	13
2.4 字符编码	16
2.5 数据转换	18
本章小结	18
第三章 电子表格与文本编辑器.....	19
3.1 电子表格中的数据清洗.....	19
3.2 文本编辑器里的数据清洗.....	24
本章小结	27
第四章 基本的技术与方法.....	28
4.1 ETL 基础知识	28
4.2 数据清洗的技术途径.....	31
4.3 常用的 ETL 工具	32
4.4 ETL 子系统介绍	34
本章小结	35
第五章 数据抽取	36
5.1 文本文件的数据抽取.....	36
5.2 Web 文件的数据抽取.....	39
5.3 关系型数据库数据抽取.....	41

5.4 增量数据抽取.....43

本章小结49

第六章 数据的转换与加载..... 50

6.1 数据清洗转换..... 50

6.2 数据质量评估..... 54

6.3 数据加载 56

本章小结 58

第二篇 实 战 篇

第七章 数据清洗工具介绍.....61

7.1 利用 Microsoft Excel 进行数据清洗.....61

7.2 使用 Kelltle 进行数据清洗 72

7.3 使用 OpenRefine 进行数据清洗 79

7.4 使用 DataWrangler 进行数据清洗..... 83

7.5 使用 Hawk 进行数据清洗..... 89

本章小结99

第八章 基于 Web 的数据采集实战 100

8.1 招聘信息采集..... 100

8.2 招聘信息数据预处理..... 104

本章小结 113

第九章 基于 RDBMS 的数据清洗实战 114

9.1 基础准备 114

9.2 编写数据清洗程序..... 117

9.3 编写脱敏程序..... 120

本章小结 130

第一篇 基础篇

基础篇主要从数据清洗应用背景入手，首先阐述了基本的数据清洗发展历程、背景以及数据的标准化过程、数据仓库的概念及其应用；其次，讲解了基本的数据格式与数据编码知识，包括数据的归档与压缩、数据的编码转换等；再次，讲解了使用常用的电子表格与文本编辑器进行数据清洗的方法以及数据清洗的基本技术与常用方法；最后，讲解了文本文件、Web 文件以及数据库文件的抽取与转换加载等基本知识。

第一章 数据清洗概述

随着物联网、云计算、数据技术的快速发展，迎来了大数据时代。大数据技术改变了传统的数据收集、处理与应用模式，为大量行业的跨越式发展带来了新的机遇。从字面上理解，数据清洗(Data Cleaning/Cleansing)就是把“脏”的数据进行“清洗”，也就是发现并纠正数据文件中可能出现的错误，包括检查数据一致性、处理无效值和缺失值等。通常在数据仓库中的数据都是面向某一主题的数据的集合，这些数据从多个业务系统中抽取而来而且包含历史数据，这样就避免不了有的数据是错误数据，有的数据相互之间有冲突，这些错误的或有冲突的数据显然是我们不想要的，称为“脏数据”。我们要按照一定的规则把“脏数据”洗掉，这就是数据清洗。数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉那些“脏数据”还是由业务单位修正之后再行抽取。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

数据科学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。数据科学也是一门关于数据工程的学科，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：用数据的方法来研究科学和用科学的方法来研究数据。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。

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、有缺失，存在错误和重复的数据，数据中有不一致和冲突等缺陷。数据清洗就是对数据进行审查和校验，发现不准确、不完整或不合理的数据，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，以提高数据的质量。

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性，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。

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。数据清洗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。

数据清洗对保持数据的一致性和更新数据起着重要的作用,因此被用于如银行、保险、零售、电信和交通等多个行业。数据清洗主要有三个应用领域: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)。

数据清洗对随后的数据分析非常重要,因为它能提高数据分析的准确性。但是数据清洗依赖复杂的关系模型,会带来额外的计算和延迟开销,必须在数据清洗模型的复杂性和分析结果的准确性之间进行平衡。

数据清洗通过分析“脏数据”的产生原因和存在形式,利用数据溯源的思想,从“脏数据”产生的源头开始分析数据,对数据流经环节进行考察,提取数据清洗的规则和策略,对原始数据集应用数据清洗规则和策略来发现“脏数据”,并通过特定的清洗算法来清洗“脏数据”,从而得到满足预期要求的数据。具体而言,数据清洗流程包含以下基本步骤:

- (1) 分析数据并定义清洗规则;
- (2) 搜寻并标识错误实例;
- (3) 纠正发现的错误;
- (4) 干净数据回流;
- (5) 数据清洗的评判。

数据清洗是一项十分繁重的工作,数据清洗在提高数据质量的同时要付出一定的代价,包括投入的时间、人力和物力成本。通常情况下,大数据集的清洗是一个系统性的工作,需要多方配合以及大量人员的参与,需要多种资源的支持。这些资源包括:

- (1) 数据清洗环境,指为进行数据清洗所提供的基本硬件设备和软件系统,特别是已得到广泛应用的开源软件和工具。
- (2) 终端窗口和命令行界面,比如 Mac OS X 上的 Terminal 程序或 Linux 上的 bash 程序。
- (3) 适合程序员使用的编辑器,如 Mac 上的 Text Wrangler, Linux 上的 vi 或 emacs,或 Windows 上的 Notepad++、Sublime 编辑器等。
- (4) Python 客户端程序,如 Enthought Canopy。另外,还需要足够的权限来安装一些程序包文件。
- (5) 电子表格程序,如 Microsoft Excel 和 Google Spreadsheets。它们可用于数据呈现和可视化,并且可以恰当的方式呈现数据分析和挖掘的结果。
- (6) 数据库软件,如 MySQL 数据库和 Microsoft Access 等。

1.2 数据标准化

数据标准化(Data Standardization/Normalization)是机构或组织对数据的定义、组织、分

类、记录、编码、监督和保护进行标准化的过程，它有利于数据的共享和管理，可以节省费用，提高数据使用效率和可用性。

数据标准化处理主要包括数据同趋化处理和无量纲化处理两个方面。数据同趋化处理主要解决不同性质数据问题，对不同性质指标直接加总不能正确反映不同作用力的综合结果，必须先考虑改变逆指标数据性质，使所有指标对测评方案的作用力同趋化，然后再加总才能得出正确结果。

数据无量纲化处理主要用于消除变量间的量纲关系，解决数据评价分析中数据的可比性。例如，多指标综合评价方法需要把描述评价对象不同方面的多个信息综合起来得到一个综合指标，由此对评价对象做整体评判，并进行横向或纵向比较。

数据标准化方法通常有下列几种：

(1) max-min 标准化：对原始数据进行线性变换。设 $\min A$ 和 $\max A$ 分别为属性 A 的最小值和最大值，将 A 的一个原始值 x 通过 max-min 标准化映射成在区间 $[0,1]$ 中的值 x' ，其计算公式为

$$x' = \frac{x - \min A}{\max A - \min A}$$

(2) z-score 标准化：基于原始数据的均值(mean)和标准差(standard deviation)进行数据的标准化，将 A 的原始值 x 标准化到 x' ，其计算公式为

$$x' = \frac{x - \text{mean}}{\text{standard deviation}}$$

(3) decimal scaling 标准化：通过移动数据的小数点位置来进行标准化。小数点移动多少位取决于属性 A 的取值中的最大绝对值。将属性 A 的原始值 x 标准化到 x' 的计算公式为

$$x' = \frac{x}{10^j}$$

其中， j 是满足条件的最小整数。

(4) 其他标准化方法：还有一些标准化方法的做法是将原始数据除以某一值，如将原始数据除以行或列的和，称总和标准化。如果原始数据除以每行或每列中的最大值，叫做最大值标准化。如果原始数据除以行或列的和平方的平方根，则称为模标准化(Normal Standardization)。

1.3 数据仓库

数据仓库是基于信息系统业务发展需要，基于传统数据库系统技术发展形成并逐步独立出来的一系列新的应用技术，目标是通过提供全面、大量的数据存储来有效支持高层决策分析。

W. H. Inmon 对数据仓库的定义是：数据仓库是决策支持系统和联机分析应用数据源的结构化数据环境，是一个面向主题的(Subject Oriented)、集成的(Integrated)、相对稳定的(Non-Volatile)、反映历史变化(Time Variant)的数据集合，用于支持经营管理中的决策制定过程。

数据库是面向事务设计的，而数据仓库是面向主题设计的。数据库设计是尽量避免冗余，一般采用符合范式的规则来设计；数据仓库在设计时有意引入冗余，采用反范式的方

式来设计。数据库是为捕获数据而设计的，数据仓库则是为分析数据而设计的。数据库一般存储的是在线交易数据，而数据仓库一般存储的是历史数据。

数据仓库不是一种提供战略信息的软件或硬件产品，而是一个便于用户找到战略信息和做出更好决策的计算环境，是一个以用户为中心的环境。数据仓库需要提供数据抽取、数据转换、数据装载和数据存储功能，并为用户提供交互接口。典型数据仓库的基本组成要素包括：源数据单元、数据准备单元、数据存储单元、信息传递单元、元数据单元和管理控制单元。

根据企业构建数据库的主要应用场景的不同，可以将数据仓库分为以下四种类型，每一种类型的数据仓库系统都有不同的技术指标与要求。

1. 传统数据仓库

企业把数据分成内部数据和外部数据，内部数据包括 OLTP(联机事务处理系统)交易系统和 OLAP(联机事务分析系统)分析系统的数据。企业首先需要将这些数据集中起来，经过转换放到这类数据库中，然后在数据库中对数据进行加工，建立各种主题模型，再提供报表分析业务。

2. 数据集市

数据集市一般是用于某一类功能需求的数据仓库的简单模式，往往由一些业务部门构建，也可以构建在企业数据仓库上。一般来说，数据集市的数据源较少，但对数据分析的延时有很高的要求，并需要和各种报表工具有很好的对接。

3. 关联发现数据仓库

在一些场景下，企业可能不知道数据的内联规则，而是需要通过数据挖掘的方式找出数据之间的关联关系、隐藏的联系和模式等，从而挖掘出数据的价值。很多行业的新业务都有这方面的需求，如金融行业的风险控制、反欺诈等业务。上下文无关联的数据仓库一般需要在架构设计上支持数据挖掘能力，并提供通用的算法接口来操作数据。

4. 实时处理数据仓库

随着业务的发展，企业客户需要对实时的数据做一些商业分析，譬如零售行业需要根据实时的销售数据来调整库存和生产计划。这类行业用户对数据的实时性要求很高，传统的离线批处理的方式不能满足需求，因此需要构建实时处理的数据仓库。数据可以通过各种方式完成采集，然后数据仓库可以在指定的时间限期内对数据进行处理和统计分析等，再将数据存入数据仓库以满足一些其他业务的需求。通常数据仓库有下列相关的技术：

1) 数据清洗

数据仓库需要从种类各异的多个数据源中导入大量数据，数据仓库的一个重要任务就通过数据清洗保证数据的一致性与正确性。

2) 数据粒度

数据仓库中存储的数据粒度将直接影响数据仓库中数据的存储量及查询质量，并进一步影响数据仓库能否满足最终用户的分析需求。设计数据仓库时要合理确定数据粒度。

3) 索引优化

不论是数据库还是数据仓库，索引查找是优化查询响应时间的重要方法，索引建立的

好坏直接影响数据访问效率。

4) 物化视图选择和维护

数据仓库中以物化视图(Materialized View)的形式存储大量来自多个异质数据源中的数据,数据仓库中采用物化视图进行快速查询和分析,能有效提高数据查询速度和响应时间。

5) 数据仓库的管理维护

为了减少数据更新量,数据仓库一般采用增量式更新策略。此外,数据仓库必须建立有效的安全策略和授权访问控制机制。最后,数据仓库必须提供稳定可靠的数据备份和恢复策略。

数据仓库不是一门纯粹的技术,确切地说,它更是一种架构和理念,核心在于对数据的整合集成,将企业原始数据进行集成、归类、分析,从而提供企业决策分析需要的目标数据。SQL Server、Sybase、DB2 和 Oracle 都是传统的关系数据库,但是只要经过合理的数据模型设计或参数设置,也可将其转变为很好的数据仓库实体。

目前,OLAP 已逐渐融合到数据仓库中,例如微软的 Analysis Service 和 DB2 的 OLAP Server,通过自身提供的专用接口可以加快多维数据的转换处理。当然,也有如 Essbase 这样纯粹的 OLAP 产品,实际上许多大型 OLAP 都采用 Essbase。

对于 ETL(提取、转换、加载)而言,广泛使用的 ETL 工具主要包括 Informatica PowerCenter、IBM 的 Datastage、SQL Server 搭配的 SSIS、Oracle 的 OWB 和 ODI 以及开源的 Kettle 等。

数据仓库可用的报表工具很多,专业性的报表工具有 Hyperion、B0、Cognos 和 Brio,这些产品价格相对昂贵;便宜的报表工具可选用微软的 ReportService。

本章小结

本章从数据清洗的概念开始,介绍了数据清洗的基本定义、数据清洗的过程与任务及数据清洗流程;其次讲解了数据标准化的概念和标准化的常用方法;最后对数据仓库进行了介绍,从数据仓库的定义到数据仓库的组成要素以及数据仓库的分类分别进行了详细介绍。

第二章 数据格式类型与编码

任何被存储的数据都存在格式(Data Format)，格式是数据的一种存在形式。数据的格式通常为数值、字符二进制等，一般用数据类型与数据长度来描述。因为计算机是统一自动化处理数据的，所以统一的格式是保证程序进行自动化处理的前提条件。编码(Code/Encode)是数据从一种形式转换成另外一种形式的过程，可用预先规定好的方法将文字、数据与其他信息编成固定格式的数字。编码的逆过程是解码(Decode)，是将数字码转换成文字、数字以及图像等信息。

本章将从文本及文本的格式以及文件的归档与压缩等基本知识入手，阐述数据的类型编码以及编码之间的转换等内容。

2.1 文件及文本的格式

文本是计算机保存数据的主要方式，其常见的文本格式有 txt、doc、zip、jpg 和 HTML 等几种。

Windows 操作系统下常见的文本格式有：txt、doc、xls 等。

类(UNIX)操作系统下常见的文本格式有：dmg、tar 等。

网络文本格式有：HTML、xml、php、jsp、css 等。

1. xls 及xlsx 文件格式

xlsx 是 Microsoft Office Excel 2007 或者更新版本保存的文件格式，即用新的基于 XML 的压缩文件格式取代了其之前专有的文件格式。此文件格式在传统的文件扩展名后面添加了字母 x(“.docx”取代“.doc”，“.xlsx”取代“.xls”)，使文件占用系统的空间更小。所有版本的 Microsoft Excel 都可以打开 xls 格式的文件。

2. JSON 文本格式

JSON 的全称是 JavaScript Object Notation，即 JavaScript 对象标记，是一种轻量级的数据传输格式，常用于网络信息的传输。JSON 基于 ECMAScript 规范，采用独立于编程语言的文本格式来存储和表示数据。

JSON 具有简洁和清晰的层次结构，是当下较为理想的一种数据传输语言。因为 JSON 易于阅读和编写，也易于机器解析和生成，所以能有效地提升网络传输效率，在现有的客户端和服务端数据交换传输中，JSON 的应用非常广泛。

3. HTML 和 XML 文本格式

HTML 的全称是 HyperText Markup Language，即超文本标记语言，这里的“超文本”指的是页面内可以包含图片、链接，甚至音乐、程序等非文字元素，HTML 是标准通用标

记语言下的一个应用。图 2-1 所示为一个典型的 HTML 文件内容。

```
<!DOCTYPE html>
<html lang="en" xmlns="http://www.w3.org/1999/html" xmlns="http://www.w3.org/1999/html">
  <head>
    <meta charset="UTF-8"/>
    <meta http-equiv="refresh" content="3"/>
    <meta name="keywords" content="HTML例子"/>
    <meta name="descriptions" content="HTML例子"/>
    <meta http-equiv="X-UA-Compatible" content="ie=IE9"/>
    <title>HTML例子</title>
    <link rel="icon" href="img/favicon.ico">
    <link rel="stylesheet" href="css/common.css">
  </head>
  <body>
    <ul>
      <li>hello</li>
      <li>html</li>
    </ul>
  </body>
</html>
```

图 2-1 典型 HTML 文件展示

4. XML

与 JSON 功能相同的另一种格式是 XML，其全称是 Extensible Markup Language，即可扩展标记语言，也是标准通用标记语言下的一个应用。XML 是各种应用程序之间进行数据传输最常用的工具。图 2-2 所示为一个典型的 xml 文件内容。

```
<w:lvl w:ilvl="4">
  <w:start w:val="1"/>
  <w:inf w:val="23"/>
  <w:lvlText w:val="" />
  <w:lvlJC w:val="left"/>
  <w:pPr>
    <w:ind w:left="2100" w:hanging="420"/>
  </w:pPr>
  <w:rPr>
    <w:rFonts w:ascii="Wingdings" w:h-ansi="Wingdings" w:hint="default"/>
  </w:rPr>
</w:lvl>
<w:lvl w:ilvl="5">
  <w:start w:val="1"/>
  <w:inf w:val="23"/>
  <w:lvlText w:val="" />
  <w:lvlJC w:val="left"/>
  <w:pPr>
    <w:ind w:left="2520" w:hanging="420"/>
  </w:pPr>
  <w:rPr>
    <w:rFonts w:ascii="Wingdings" w:h-ansi="Wingdings" w:hint="default"/>
  </w:rPr>
</w:lvl>
<w:lvl w:ilvl="6">
  <w:start w:val="1"/>
  <w:inf w:val="23"/>
  <w:lvlText w:val="" />
  <w:lvlJC w:val="left"/>
  <w:pPr>
    <w:ind w:left="2940" w:hanging="420"/>
  </w:pPr>
  <w:rPr>
    <w:rFonts w:ascii="Wingdings" w:h-ansi="Wingdings" w:hint="default"/>
  </w:rPr>
</w:lvl>
```

图 2-2 典型 XML 文件展示

通过上面的实例，对 XML 格式和 JSON 格式进行比较后发现：

- (1) JSON 和 XML 都是纯文本。
- (2) JSON 和 XML 都具有“自我描述性”。
- (3) XML 和 JSON 都可以通过 JavaScript 进行解析。
- (4) XML 有结束标签而 JSON 没有。
- (5) JSON 传输一般比 XML 更短、速度更快。

2.2 归档与压缩

现实中的许多数据、特别是分隔数据都是以压缩文件的形式存在的。那么对于一个数据科学家来说，常用的数据压缩格式有哪些呢？以及如何评估哪个压缩方法是最优的呢？下面我们就从文件的归档、文件的压缩，以及如何选择合适的压缩工具与程序等方面来回答这个问题。

1. 归档文件

归档文件就是一个对内部包含了许多文件的独立文件的归档。这些文件的内部包含文本文件或二进制文件，或者二者兼有。归档文件是使用一个特殊程序将给定的文件列表转换为一个文件而成的，当然它也可以反向操作被转换成多个文件。

通常，在进行数据科学工作的时候，碰到最多的归档文件就是磁带归档文件(tar)。这种文件是用 tar 程序创建的，以 .tar 为后缀名。其最初的设计目的是用于磁带归档技术。

tar 程序在类 UNIX 操作系统中都存在，我们也可以在苹果操作系统中的 terminal 程序中使用它。要想创建 tar 文件需要先设定几个参数，包括用户想压缩的文件路径、输出的文件名等(其中，c 表明想创建一个归档文件，v 表示程序在加载文件列表的时候打印出文件名，f 用于指定输出的文件名)。命令行代码如下：

```
tar cvf fileArch.tar bigFile.csv otherBigFile.csv
```

如果需要一个“untar”文件，只需要让 tar 程序指向用户想要展开的文件就可以了，命令行代码如下：

```
Tar xvf fileArch.tar
```

当我们知道了有一个以 .tar 为文件名后缀格式的归档文件包中包含多少个文件以及是什么文件时，在提取之前，我们需要掌握该压缩文件的一些基本信息，以及即将提取的文件是不是我们想要的，还有目前磁盘空间是否够用等情况。可用 tar 命名的选项 t 来以列表形式显示 tar 文件中的内容，命令行代码如下：

```
tar -tf fileArch.tar
```

当然，除了 tar 命令之外，还有许多别的归档程序，包括一些具备压缩功能的程序(例如，OS 系统内置的 zip 压缩软件和 Windows 上的各种 zip 以及 rar 工具)，它们在进行归档的同时还能进行文件的压缩处理。