数据整理——大数据治理的关键技术

杜小勇[1,2]，陈跃国[1,2]，范举[1,2]，卢卫[1,2]
1. 中国人民大学信息学院，北京 100872；
2. 数据工程与知识工程教育部重点实验室（中国人民大学），北京 100872

摘要
数据是政府、企业和机构的重要资源。数据治理关注数据资源有效利用的众多方面，如数据资产确权、数据管理、数据开放共享、数据隐私保护等。从数据管理的角度，探讨了数据治理中的一项关键技术：数据整理。介绍了以数据拥有者和直接使用者（行业用户）为核心的数据整理的关键技术，包括数据结构化处理、数据质量评估及数据清洗、数据规范化、数据融合与摘取、数据整理的发布共享等。最后，针对加强数据整理方面的研究提出了一些思考。

关键词
数据整理；数据准备；数据治理；数据管理

中图分类号：TP315　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2019020

Data wrangling: a key technique of data governance

DU Xiaoyong[1,2], CHEN Yueguo[1,2], FAN Ju[1,2], LU Wei[1,2]
1. School of Information, Renmin University of China, Beijing 100872, China
2. Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing 100872, China

Abstract
Data is an important resource for governments, businesses and institutions. Data governance focuses on many aspects of effective use of data resources, such as data asset, data management, data sharing, and data privacy. A key technique-data wrangling-in data governance from the perspective of data management was explored. The key technologies of data wrangling based on data owners and direct users-industry users were emphasized, including data structure processing, data quality assessment and data cleaning, data normalization, data fusion and extraction, data publishing and sharing, etc. Finally, some thoughts on strengthening the research on data organization were put forward.

Key words
data wrangling, data preparation, data governance, data management

1 引言

大数据作为一种资源，在政府、大型企业和机构中发挥着越来越重要的作用。随着大数据应用的不断推进，与数据资源的价值提炼、保值和增值密切相关的大数据治理越来越引起人们的重视。大数据治理是一项复杂的工程，它需要在国家、行业、企业等多个层面上开展体系化的建设，技术上包含数据资产确权、数据管理、数据开放共享、数据隐私保护等诸多方面。这些技术面临的挑战多、难度大，很多方面还没有形成被广泛认可的系统化的解决方案。本文从数据管理这一关键环节出发，探讨其中的关键支撑技术：数据整理（data wrangling）。

数据整理也叫数据准备，是在挖掘提炼数据价值的过程中进行的前期的数据预处理工作。它看似不足轻重，实则非常重要。有调查研究表明[1]，很多大数据分析任务80%以上的工作花费在数据整理上，这给数据分析带来了巨大的人力成本。很多分析设想因为承担不起前期的数据整理工作而最终被放弃。更重要的是，由于缺少系统性和理论性的支撑，数据整理的质量千差万别，这给数据分析的结果带来了很大的不确定性，大大影响了大数据价值的挖掘与提炼。因此，人们很有必要重视数据整理的研究工作，它是整个数据治理环节中一项重要的基础性工作，但是这项工作在学术界和企业界并没有得到应有的重视。

2 数据整理概述

在数据仓库时代，数据预处理主要指的是抽取、转换和加载（ETL）过程。笔者探讨的数据整理和ETL过程有相似的地方，两者都将多源异构的数据集通过一系列处理和转换，变成想要的输出形式。但二者之间是存在较大差别的[2]，具体如下。

● 针对的用户不同。ETL服务于专业的数据工程师，而数据整理服务于企业内部所有的数据使用者，以对数据处理技术不熟悉的业务用户为主。这些用户虽然缺少数据管理与数据处理知识，但对业务非常熟悉，对数据背后的语义更清楚。他们是企业机构大数据价值发现的主力。如何针对这类业务型数据分析人员的需求和特点，提供高效的数据整理工具，是数据整理技术面临的一大挑战。

● 数据处理的目的不同。数据仓库中的ETL是为了建立数据仓库采用的相对固定的数据处理流水线。数据处理过程一旦建立，整个过程比较静态，很少再变化。数据整理是针对企业业务系统中的问题，动态构建的数据处理过程。它针对具体问题进行数据预处理，针对不同问题采用不同的数据整理过程，一些任务之间可以共享某些数据整理过程。

● 数据处理的对象不同。ETL处理的数据对象多为业务系统数据库中的结构化数据源，这些数据源有很规范的元数据。数据整理则面临更复杂、更多样化的数据源，直接应对大数据多样性（variety）的挑战。这种多源异构性在很多大数据应用中非常常见。数据整理技术通常需要帮助用户将其拥有的数据与外部的一些数据源进行关联和数据融合。融合过程中存在的大量数据质量问题（如数据项缺失、不一致、重复、错位、异常值等）给数据整理带来了巨大挑战。与ETL技术相比，这种变化是一种质的变化。

数据整理是为了使数据更好地服务于数据分析而对数据进行的审查和转换的过

程，它是整个数据分析流程中最占用精力的过程。从技术上讲，数据整理包含前期数据解析与结构化处理、数据质量评估与数据清洗、数据集成和提纯等过程。由于问题的复杂性，数据整理过程通常不是完全自动化的，而是需要用户介入的反复迭代和交互的过程。数据可视化、用户反馈与交互在整个过程中都发挥了重要作用。数据整理是由数据可视化领域的Jeffery Heer教授（华盛顿大学）和数据库领域的Joseph M. Hellerstein教授（加州大学伯克利分校）等人较早提出来并持续开展系列研究的[1-2]。他们还将研究成果进行了产业化，成功创立了以数据整理为主业的Trifacta公司。本文主要在上述两位教授及其合作者发表的一些成果[1-2]的基础上，对数据整理包含的一些核心要素进一步地阐述，以期引起人们对数据整理研究和应用的重视。

3 数据整理的核心技术

3.1 数据的结构化处理

很多数据模型和算法是构建在结构化数据基础上的，多源异构数据要更好地与其他数据集融合，结构化处理是必不可少的过程。数据结构化处理首先要对原始数据进行解析，提取出需要的信息，再进一步将其转换成结构化数据。很多非结构化数据、Web数据是以文本形式存在的，需要使用信息抽取技术识别文本中的实体、属性、关系等信息。也有很多数据采用的是结构化强一些的数据模型，如JSO格式，这类数据相对关系型数据更灵活，在结构化转换过程中也需要一些技术上的处理。结构化处理的主要输出形式是二维表或者图数据，它需要用户确定数据在转换过程中采用的规则。

3.2 数据质量评估与数据清洗

结构化处理主要是数据表达形式上的转换，数据结构化之后并不意味着能够直接使用。处理后的数据还要进行质量评估，如果发现数据中存在问题，则采取进一步的数据清洗措施。这个过程称作数据质量评估。一些简单的数据质量问题可以利用自动化的算法发现，因为数据质量问题的多样性和不可预测性，数据可视化技术成为数据质量评估的关键技术。借助可视化技术，对数据语义非常了解的业务人员更容易发现数据存在的质量问题（如缺失、不一致、异常等）。伴随着数据质量问题的发现，用户可以定义一些数据清洗规则，批量化地处理数据中存在的质量问题，提高数据清洗的效率。在数据库研究领域，也有人借助众包的思路提升数据清洗的效率[3]。这种做法也是基于用户在数据清洗过程中发挥的重要作用进行的。在数据清洗过程中，需要多轮次的人机交互，系统的交互界面和交互方式对于数据清洗算法的有效性尤为重要。

3.3 数据规范化

数据清洗还有一项重要的内容是数据规范化，这也是数据准备中常见的问题。规范化有简单的底层数据层面的，如数据类型转换、单位变换、格式表换等，也有较为复杂的数据项规范化处理，如电话号码、邮编、地址等。这类问题的主要成因是自然语言表达上的差异性会造成同一实体存在多种表达形式。比较典型的例子是地址，人们需要对其进行规范化处理，以提升数据的质量。地址的规范化面临的一个比较大的挑战就是粒度的选取，同一个地址可以用不同粒度进行表达。数据的规范化处理需要根据应用的需求特点，确定数据粒度

和表达方式。地址规范化处理背后的问题是实体链指问题，即把同一实体的不同表达形式（不同名字）映射到同一个实体名字上，消除实体表达的语义鸿沟，进而通过关联在数据集中不同地方出现的相同语义的实体，达到数据融合的目的。

此外，缺失值填充也是数据规范化处理过程中常见的问题。一种处理方式是利用缺失数据的上下文数据，采用数据插值的办法修复缺失数据；另一种处理方式是采用平均值或者缺省值的办法填充缺失数据，有时候也用这种办法替换系统发现的异常值。

3.4 数据融合与摘取

很多数据价值的发现源自于多源异构数据之间的关联和在关联数据基础之上进行的数据分析。将多个数据集（很可能来自于多个数据源）融合到一起，可使数据内容更丰富，更容易获得新的发现。然而，多源数据融合所需的数据整理过程面临的挑战是很大的。多源头的数据缺少统一的设计，这导致数据集成和数据融合的难度加大。传统的基于模式的数据集成方法很难发挥出大的作用，解决这一难题更多地要从数据项的层面关联数据。因此，实体链指操作在数据融合过程中就显得尤为重要。数据在实体层面的链指可以丰富实体的语义，建立跨数据项之间的关联。由于实体表达的模糊性，实体上下文信息对实体链指精度的影响非常大，有效利用实体上下文信息（如文本中的语境、表结构中同行属性值等）是实体链指的关键。

数据融合是数据集整合的过程，有些分析任务未必需要全部整合后的数据，可能仅需要一部分数据支撑分析任务。在这种情况下，需要从数据集中提取部分数据（如一些样本或者数据片段），降低数据量，供数据分析模型实现分析操作。这一过程称作数据摘取，它需要根据任务的特点摘取相关数据。

3.5 发布共享

企业中复杂的数据分析任务经常需要被共享，某些数据整理操作也会被重复使用，这意味着数据整理的操作也是企业机构的一种资源。企业需要将这些操作以脚本的形式物化出来，使其能够被检索、分享和重复利用。经过数据整理过程的数据，其世袭关系需要被记录下来，以确保用户能够追溯数据的来源，也便于利用索引技术检索需要的数据整理操作。企业内部对数据整理的共享对于企业内部知识管理、协同工作而言有很重要的意义。

4 以技术带动数据治理能力

通过以上分析可以看出，数据整理以提升数据分析的效率和质量为目的，在整个大数据分析流程中占有重要的地位。近些年来，尽管学术界在数据质量管理方面做了大量的研究性工作，但在实际应用中，很多数据整理的需求并没有得到很好的满足，还缺少数据整理方面的工具，尤其是系统化的数据整理工具[4]。对于工业界而言，数据整理工作更多地被看作数据分析人员应完成的工作，人们并没有从工具和系统的角度开发设计高效率的数据准备工具，这使得数据分析人员在执行数据整理任务时，执行了大量重复性的工作。因此，加强数据整理的研究和应用工作是很有必要的。

4.1 数据的结构化与规范化

信息抽取是指从非结构化的文本中识

别实体，并发现实体的属性、实体之间的关系，在互联网信息抽取、知识库构建等领域发挥着重要的作用。命名实体识别的目的是发现文档中的各种实体，如人物、地理位置、组织、日期、时间等。命名实体识别技术分为以下3类。

● 基于正则表达式的命名实体识别：把预先定义的正则表达式和文本进行匹配，把符合正则表达式的文本模式都定位出来。基于正则表达式的命名实体识别一般用于识别日期、时间、金额、电子邮件等规则的文本。

● 基于字典的命名实体识别：把文本和字典里的<短语，类别>对进行匹配，对匹配的短语进行实体标注，一般用于识别人名、地名。

● 基于机器学习模型的命名实体识别：预先对一部分文档进行实体标注，产生一系列的<短语，类别>对，利用这些文档进行机器学习模型的训练，然后用这个模型对没有遇到过的文档进行命名实体识别和标注。

指代消解是自然语言处理中和命名实体识别关联的一个重要问题。比如在对某位专家学者进行的一个访谈中，除了第一次提到其姓名、职务之外，之后提到这位专家，文本中可能使用“某博士”“某教授”“他”等代称，或者以其担任的职务相称，如“所长”等。如果访谈中还提及其他人物，并且也使用了类似的代称，那么把这些代称对应到正确的命名实体上就是指代消解。在自然语言处理中，经常遇到的一个问题是命名实体的歧义，比如重名问题。为了让计算机正确地分析自然语言书写的文本，命名实体的歧义需要被消除，也就是把具有歧义的命名实体唯一地标识出来。

关系抽取是信息抽取的一个重要的子任务，负责从文本中识别出实体之间的语义关系。它分为3类方法：有监督的学习方法，该方法包括基于特征向量的学习方法和基于核函数的学习方法；半监督的学习方法，该方法无需人工标注语料库，但是需要根据预定义好的关系类型人工构造出关系实例，将这个关系实例作为种子集合，然后利用Web或者大规模语料库信息的高度冗余性，充分挖掘关系描述模式，通过模式匹配，抽取新的实体关系实例；无监督的学习方法，该方法是一种自底向上的信息抽取策略，它假设拥有相同语义关系的实体对的上下文信息较为相似，其上下文集合代表该实体对的语义关系。较新的技术是使用向量（embedding，基于词或者实体）的方式将结构化和非结构化数据中提及的实体关联起来，利用向量间的相似性，实现以向量为中介的异构数据的结构化处理和关联。

4.2 数据集成

数据集成是伴随企业信息化建设的不断深入而形成的。例如，因业务的需要，企事业单位内部普遍构建了多个异构的信息系统（这些信息系统可以自主选择合适的操作系统，有独立的数据库和应用界面，完全是一个自治的系统），并积累了图片、Word、PDF、Excel、网页等大量非结构化文件。由于开发部门和开发时间的不同，这些信息系统中管理的数据源彼此独立、相互封闭，形成了“信息孤岛”，数据难以在系统之间形成快速有效的共享。数据管理与数据分析需要打破这些“信息孤岛”，实现不同“孤岛”信息系统的互联互通，进而施行精准的决策分析。例如，在电子政务领域中，很多地方的政府机关有多少个委、办、局，就有多少个信息系统，每个信息系统都由独立的信息中心进行维护。政府机关之间需要实现信息互联互通、资源共享，最终实

现政务服务的协同操作，从而使社会大众真正享受到一站式办公服务（例如杭州市政府工作报告中的“最多跑一次”改革）。事实上，许多互联网应用（包括机票、酒店、餐饮、租房、商品比价等服务）也是把来自不同数据源中的数据进行有效集成后，对外提供统一的访问服务的。

数据集成把一组自治、异构数据源中的数据进行逻辑或物理上的集中，并对外提供统一的访问接口，从而实现全面的数据共享。数据集成的核心任务是将互相关联的异构数据源集成到一起，使用户能够以透明的方式访问这些数据源。集成是指维护数据源整体上的数据一致性，提高信息共享利用的效率；透明的方式是指用户无需关心如何实现对异构数据源数据的访问，只关心以何种方式访问何种数据即可[5]。数据集成涉及的数据源通常是异构的，数据源可以是各类数据库，也可以是网页中包含的结构化信息（例如表格）、非结构化信息（网页内容），还可以是文件（例如结构化CSV文件、半结构化的XML文件、非结构化的文本文件）等。数据集成中涉及的数据源具有自治性，这些数据源可以在不通知集成系统的前提下改变自身的结构和数据。

数据源的异构性和自治性是数据集成系统面临的两个主要挑战。针对这两个挑战，数据集成通常采用如下两种解决方案。

(1) 数据仓库

人们把一组自治数据源中的数据加载并存储到一个物理数据库（称为数据仓库）中，然后在数据仓库上对集成后的数据进行后续的操作和分析。**图1**显示了基于数据仓库的数据集成系统架构。数据仓库技术涉及的技术包括ETL、元数据管理和数据仓库本身涉及的技术。ETL定期地从各个数据源中抽取（extract）、转换（transform）、加载（load）数据到数据仓库中。元数据管理涉及对数据源的描述、对数据仓库中数据的描述、数据仓库中数据与数据源中数据之间的语义映射。例如，针对关系数据库类型的数据源，语义映射维护数据源中的某个属性对应于数据仓库的某个属性，并指定如何把属性分配到不同的表中。此外，语义映射还要解决不同数据源间数据描述的不统一、语义冲突、数据的冗余等问题。

(2) 虚拟集成系统

在虚拟集成系统中，数据保存在原来的数据源中，只在查询时才需要访问。**图2**显示了一个典型的虚拟集成系统的架构，该类集成系统使用中间模式建立全局数据的逻辑视图，中间模式向下协调各数据源系统，向上为访问集成数据的应用提供统一数据模式和数据访问的通用接口。各数据源独立性强，虚拟集成系统则主要为异构数据源提供高层次的数据访问服务。元数据维护数据源的基本信息以及中间模式到数据源之间的语义映射等。虚拟集成系统接收到用户的查询请求后，根据元数据信息进行查询的重写，把对中间模式的查询转化为对数据源的查询。类似于数据库的查询处理，虚拟集成系统也会进行查询的优化，包括访问数据源的顺序、不同数据源之间的操作访问（例如两个数据源之间数据的连接算法）等。每个数据源都连有一个封装器，负责把上层用户的查询转发到数据源，并把数据源返回的结果转发给上层的应用。虚拟集成系统的关键问题是如何构造逻辑视图，并使得不同数据源的数据模式映射到这个中间模式上。

无论是基于数据仓库还是基于中间模式的数据集成系统，都需要完成实体与关联抽取、模式匹配（schema matching）、实体对齐（record linkage或entity resolution）和实体融合（data fusion）这4个步骤。面向结构化数据的实体与关联抽取技术比较直观，面向非结构化数据的实体与关联抽取可参考第4.1节。模式匹配主要用于发现并

映射两个或多个异构数据源之间的属性对应关系，在大规模数据背景下尤为重要。目前，基于朴素贝叶斯、stacking等机器学习算法的模式匹配得到了广泛的研究，并在某些特定领域得到了良好的应用。基于模式匹配，实体对齐的目标是根据匹配属性的记录特征，将数据源中指代同一实体的记录连接起来。实体对齐主要分为3个步骤：获取候选集、成对匹配、聚簇处理。广义地说，实体对齐的方法可以划分为无监督学习和有监督学习。随着人工智能技术的发展，基于决策树、Logistic回归、支持向量机（support vector machine，SVM）的机器学习方法以及基于词向量（word embedding）的深度学习方法被应用于实体对齐，以提高算法的性能。使用实体对齐可以把一组数据源中同一实体的不同记录连接起来，由于数据质量问题，这些记录在描述同一实体时可能存在数据冲突，例如同一个人的住址在不同数据源之间的描述可能是不一样的。因此，在数据集成的最终环节中，实体融合旨在消除不同数据源之间同一个实体属性值的冲突，将不同的数据信息进行综合，从而提取出统一、丰富、高精度的数据。实体融合的主要方法包括基于规则的无监督学习、结合标注数据的半监督学习等。虽然基于标注数据的半监督学习在精度、召回率等方面均获得了令人满意的效果，但是其最大的挑战在于带标签训练数据的获取往往需要耗费较大的人力和物力。如何利用主动学习获取训练数据以降低研究代价，是当前学术界和工业界研究的热点话题。

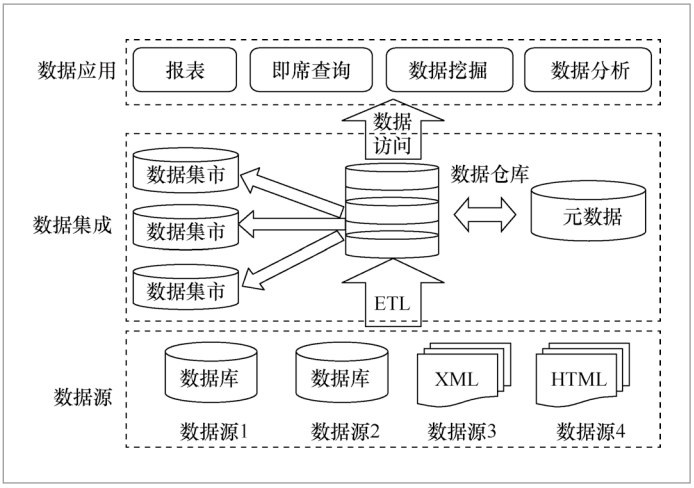

图1 基于数据仓库的数据集成系统架构

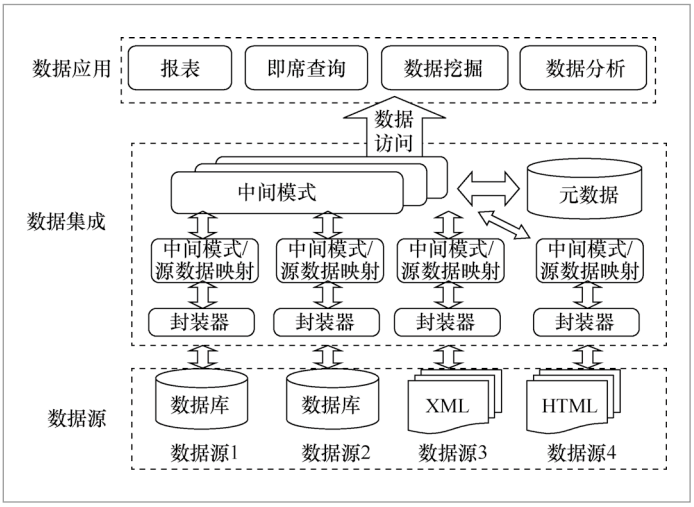

图2 基于中间模式的数据集成系统架构

4.3 数据清洗与数据质量评估

数据清洗是指从数据中检测并纠正可能的错误，以确保数据的质量并符合与领域相关的完整性约束。数据清洗是绝大多数数据驱动的任务的必要步骤。缺乏有效的数据清洗可能会使后续的数据分析产生垃圾进、垃圾出（garbage in, garbage out, GIGO）的不良后果。然而，由于数据越发显著的大规模、异质性、高噪音等特点，数据清洗也面临着极大的挑战，这也是近年来学术界和工业界的攻坚重点[6]。一般来说，数据清洗可以分为两个基本的任务：错误检测，即发现数据中潜在的错误、重复或缺失等；数据修复，即针对发现的错误，对数据进行修复。下面结合一个具体的实例分别进行介绍。

错误检测任务旨在发现影响数据质量的错误因素。一般将错误因素划分为4类，下面通过**图3**的示例进行说明。

(1) 异常值

异常值是指明显不符合属性语义的取值。例如，**图3**中t2的年龄为5岁，显然与其有工作单位这一事实是相悖的。然而，设计一种方法让计算机自动地、通用地检测出异常值是个挑战性很大的问题。现有的代表性解决方案包含以下几类。

● 基于统计的方法：首先使用一定的分布对数据进行建模，进而检测某个取值是否显著性地偏离正常值。例如，针对**图3**示例中年龄的例子，可以使用正态分布对数据建模，并计算均值与标准差。如果某个取值在k倍的标准差（如k=3）外，则认定其为异常值。更进一步地，由于均值对异常值比较敏感，很多方法使用中位数作为均值。

● 基于距离的方法：度量数据值之间的距离，将与大多数数据距离过远的值认定为异常值。

(2) 结构性错误

结构性错误是指数据不符合特定领域语义要求的完整性约束。例如**图3**示例中t1的工作单位是中国人民大学，其所在城市应该为北京，而非上海。检测结构性错误最直接的方法是从外部输入与领域相关的约束条件，如工作单位决定了所在城市。然而，这种方法往往耗时耗力，且很难达到通用性。因此，现有的大多数工作聚焦于从数据中发现潜在的约束条件，如条件函数依赖[7]、拒绝约束规则[8]等。近些年，也有些研究者考虑借助外部通用的知识图谱及互联网上公开可用的众包服务（crowdsourcing）[9]，其基本的思想是通过发现数据中与知识图谱或众包标注违背的部分，归纳出结构性错误。

	姓氏	名字	年龄/岁	工作单位	所在城市
t1	张	三	40	中国人民大学	上海
t2	李	四	5	上海交通大学	上海
t3	王	五	35	<缺失>	北京
t4	三	张	40	人大	北京

图 3 数据清洗中错误检测的示例

(3) 记录重复

记录重复在真实数据中十分普遍，其原因是多方面的，比如数据可能由不同的机构提供，或者数据整合自组织的内外部渠道。例如，**图3**中的t1和t4实际上指代同一个人，但由于数据存在结构性错误（如t1的城市）、缩写（如t4中的“人大”实为“中国人民大学”的缩写）、属性对应错误（t4中的姓氏与名字填反了）等问题，而被计算机认为是两条不同的记录。记录重复会对数据分析造成很大的影响。人们一般采取实体识别技术解决记录重复问题，其本质与上文提到的实体匹配是相同的。由于前文已经给出了详细的探讨，此处不再赘述。

(4) 数据缺失

数据缺失是指数据的部分属性不存在于数据库中，例如，**图3**示例中的t3缺失了工作单位信息。这会在两个层面给数据分析带来负面影响：一方面，数据缺失带来信息的损失；另一方面，不同数据源在数据缺失时使用的默认值不尽相同，如“NA”“NaN”“Null”等，这会进一步误导后续的分析过程。针对数据缺失，现有的方法是采用缺失值插补（data imputation）技术进行修复，其基本想法是使用合理的模型推断出缺失值。比较简单的办法是使用统一的全局值或其他记录在该属性的平均值进行插补，然而这些方法没有考虑具体的数据记录，在实际中难以得到良好的效果。更为有效的办法是采用最大可能性的数据值并进行推理，例如找出最相似记录的相应取值并进行插补，或通过建立贝叶斯或决

策树分类器，将缺失值插补建模成一个分类的问题。

数据修复任务是指根据检测出的错误对数据进行更新，以达到纠正错误的目的。与前文介绍的错误检测相比，数据修复的挑战性更大，因为通常缺乏对修复进行指导的信号。为了应对这一挑战，现有的方法往往采用外部知识或一些定量的统计指标[10]。最近，也有人提出一些新方法，即采用机器学习的手段融合多源信号，将数据修复建模成一个联合推理的问题[11]。

5 结束语

数据整理需要研究的工作还有很多。如何开展有针对性的研究工作，并系统化地集成各方面的相关研究工作，形成数据整理方面整体上的研究和应用影响力？威斯康辛大学麦迪逊分校的AnHai Doan教授等人[4]倡议，从事相关领域的研究学者应充分利用庞大的Python开源社区PyData，投入系统化的数据准备工具研制中，将研究成果更好地应用在实际场景中。这或许是一条较为可行的技术路线。

参考文献：

[1] HELLERSTEIN J M, HEER J, KANDEL S. Self-service data preparation: research to practice[J]. IEEE Data Engineering Bulletin, 2018, 41(2): 23-34.

[2] HEER J, HELLERSTEIN J M, KANDEL S. Data wrangling[M]//Encyclopedia of big data technologies 2019. [S.l.:s.n.], 2019.

[3] LI G L, ZHENG Y D, FAN J, et al. Crowdsourced data management: overview and challenges[C]//The 2017 ACM International Conference on Management of Data, May 14-19, 2017, Chicago, USA. New York: ACM Press, 2017: 1711-1716.

[4] DOAN A H, ARDALAN A, BALLARD J R, et al. Toward a system building agenda for data integration[J]. IEEE Data Engineering Bulletin, 2018, 41(2): 35-46.

[5] 宋晓宇, 王永会. 数据集成与应用集成[M]. 北京: 中国水利水电出版社, 2008.
SONG X Y, WANG Y H. Data integration and application integration[M]. Beijing: China Water and Power Press, 2008.

[6] ABEDJAN Z, CHU X, DENG D, et al. Detecting data errors: where are we and what needs to be done[J]. Proceedings of the VLDB Endowment, 2016, 9(12): 993-1004.

[7] BOHANNON P, FAN W F, GEERTS F, et al. Conditional functional dependencies for data cleaning[C]//2007 IEEE 23rd International Conference on Data Engineering, April 15-20, 2007, Istanbul, Turkey. Piscataway: IEEE Press, 2007: 746-755.

[8] CHU X, ILYAS I F, PAPOTTI P. Holistic data cleaning: putting violations into context[C]//2013 IEEE 29th International Conference on Data Engineering (ICDE), April 8-12, 2013, Brisbane, Australia. Piscataway: IEEE Press, 2013: 458-469.

[9] CHU X, MORCOS J, ILYAS I F, et al. KATARA: a data cleaning system powered by knowledge bases and crowdsourcing[C]//The 2015 ACM SIGMOD International Conference on Management of Data, May 31-June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 1247-1261.

[10] YAKOUT M, BERTI-ÉQUILLE L, ELMAGARMID A K. Don't be scared: use scalable automatic repairing with maximal likelihood and bounded changes[C]//The 2013 ACM SIGMOD International Conference on Management of Data, June 22-27, 2013, New York, USA. New York: ACM Press, 2013: 553-564.

[11] REKATSINAS T, CHU X, ILYAS I F, et al. HoloClean: holistic data repairs with probabilistic inference[J]. Proceedings of the VLDB Endowment, 2017, 10(11): 1190-1201.

作者简介

杜小勇 (1963-), 男, 博士, 中国人民大学信息学院二级教授、学术委员会主任、博士生导师, 中国人民大学校长助理, 数据工程与知识工程教育部重点实验室 (中国人民大学) 主任。兼任教育部科学技术委员会信息学部委员, 国家重点研发计划“云计算与大数据”专家组成员, 中国计算机学会常务理事、教育工作委员会主任、数据库专业委员会主任, 《大数据》期刊副主编, 全国信息技术标准化技术委员会大数据标准工作组副组长等。曾担任中国人民大学信息学院院长, 国家“863”计划数据库重大专项专家组组长, 国家“863”计划软件重大专项专家组成员等。先后获得国家科技进步奖二等奖, 北京市科技进步奖一等奖, 教育部科技进步奖一等奖, 中国计算机学会科学技术奖一等奖等奖项。

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, 数据库专业委员会委员, 大数据专家委员会通讯委员。主要研究方向为高性能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等国际重要会议和期刊上发表论文20余篇。

范举 (1984-), 男, 博士, 中国人民大学信息学院副教授、硕士生导师, 中国计算机学会会员, 数据库专业委员会委员, 主要研究方向为大数据分析、数据集成与众包计算。先后在SIGMOD、VLDB、ICDE、TKDE等国际重要会议和期刊上发表论文30余篇。担任国际重要会议SIGMOD 2020、VLDB 2018/2020的程序委员会委员。

卢卫 (1981-), 男, 博士, 中国人民大学信息学院副教授、硕士生导师, 中国人工智能学会智能服务专业委员会委员。近年来主要从事数据库基础理论、大数据系统研制等相关领域的研究, 先后在SIGMOD、VLDB、ICDE、SIGIR、AAAI、VLDB Journal、TKDE等国际重要会议和期刊上发表论文30余篇, 主持和参与多项国家自然科学基金项目。

收稿日期: 2019-01-18
基金项目: 国家自然科学基金资助项目 (No. U1711261)
Foundation Item: The National Natural Science Foundation of China(No. U1711261)