人在回路的数据准备技术研究进展

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

摘要
随着数据分析技术的迅猛发展，数据准备越来越成为一个瓶颈性问题。以真实的数据分析场景为背景，分析了数据准备的两大核心挑战：人力成本高与时间周期长。在此基础上，介绍了人在回路数据准备技术的研究进展。交互式数据准备技术面向终端用户，通过与用户的交互预测其意图，并通过有效的预测算法来节省数据准备的时间。基于众包的数据准备技术引入互联网上的海量用户作为众包工人扩展计算能力，从而支持数据准备的基本任务，并研究如何对众包做质量控制与成本优化。最后，对人在回路的数据准备做出总结并探讨未来的挑战性问题。

关键词
数据治理；数据准备；众包；交互机制

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2019046

Progress on human-in-the-loop data preparation

FAN Ju[1,2], CHEN Yueguo[1,2], DU Xiaoyong[1,2]
1. DEKE Lab & Information School, Renmnin University of China, Beijing 100872, China
2. School of Information, Renmnin University of China, Beijing 100872, China

Abstract
With the rapid development of data analytics, data preparation has become a major bottleneck. The two essential challenges for data preparation on cost and time were analyzed. To address the challenges, the research progress on human-in-the-loop data preparation was reviewed. Firstly, interactive data preparation was reviewed, which aimed to reduce the time for data preparation by predictively interacting with the end users. Then, crowdsourced data preparation was introduced, which utilize human's computational power from the crowd to support foundamental data preparation tasks, and developed algorithms for controlling result quality and reducing crowdsourcing cost. Finally, future research directions were summarized and discussed.

Key words
data governance, data preparation, crowdsourcing, interactive mechanism

1 引言

近年来，以机器学习特别是深度学习为代表的大数据驱动的分析技术取得了突飞猛进的进展，在很多重要的领域（如图像识别、自然语言处理与无人驾驶等）得到了成功的应用。然而，开发一个数据驱动的分析应用并非易事，图1给出了一个典型的流程示例。首先，需要通过数据准备步骤将原始数据转换为训练数据；进而，通过训练数据上的模型训练，得到预测模型；然后，通过预测模型在真实场景数据上进行预测，并记录预测日志；最后，通过预测日志不断监控模型预测效果，如果预测的偏差超过了一定限度，则需要重新回到数据准备阶段，准备新的数据以更新模型。

数据准备（data preparation）也被称为数据整理（data wrangling），是上述流程的第一步，其作用是将具体某个领域的原始数据转换成机器学习算法可以使用的训练数据[1]。包括数据的结构化、清洗与转换、集成、标注和开放共享等多个阶段。数据准备看似无足轻重，只完成了前期的“预处理”工作，但在绝大多数机器学习应用中，已越来越成为影响整个开发流程的瓶颈性问题。有调查研究表明[2]，很多机器学习或大数据分析应用80%以上的工作花在了数据准备上，而模型训练与预测的工作量往往相对小些。

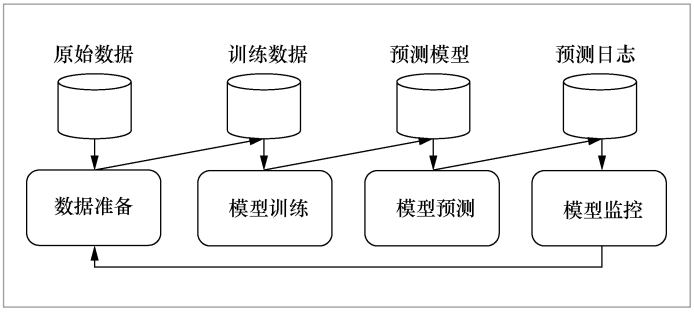


图1 一个典型机器学习应用的开发流程

传统的数据准备技术，如数据库领域的ETL，即抽取（extract）、转换（transform）、加载（load），多面向数据库管理员或IT专家。这类用户熟悉数据处理技术、了解数据底层模式，并有着丰富的编程经验。然而，随着大数据分析的普及，从数据中发现价值的主体已经变成了领域用户，如金融分析师或分析病人数据的医生。领域用户相对缺乏数据管理与处理的能力，但对数据背后的领域知识更为了解，对数据分析的趋势更为洞察。因此，如何赋予领域用户数据准备的能力变得十分迫切，也颇具挑战。具体来说，有以下2个问题。

一是人力成本大。数据准备工作难以由机器自动完成，需要大量的人力介入。这方面典型的例子是数据标注。众所周知，深度学习技术在显著提高数据分析能力的同时，也对数据提出了更高的要求：很多深度学习模型需要标注大量的数据，从而达到让人满意的效果与避免模型过拟合（over-fitting）。例如，深度学习在图像识别领域的突破性进展在很大程度上得益于ImageNet数据集[3]。截至2018年，该数据集包含了超过1 400万张的标注图片。不难想象，标注的过程需要投入大量的人力成本。因此，很多数据分析设想因为难以承担数据准备阶段的人力成本投入，而被迫最终放弃。

二是时间周期长。不同应用的数据通常千差万别，这给数据准备带来了很大的不确定性，导致其过程冗长。以应用深度神经网络（如卷积神经网络或循环神经网络）做文本分析为例，其数据准备包括分词、规范大小写、去除停用词、删除特殊符号、填充（padding）、单词嵌入（embedding）等步骤。不同任务需要对上

述步骤做不同的组合，而且每步都需要用户选择合适的参数（如单词嵌入的维度）。用户需要反复尝试才能确定更优的步骤与参数，这使得数据准备过程十分耗时。针对上述挑战，人在回路（human-in-the-loop）的技术路线近年来备受学术界和工业界的关注。本文将系统性地介绍人在回路的数据准备技术的研究进展。传统的数据准备，如上文介绍的ETL[4]，本身就是一个人在回路的过程。然而，面向大数据分析的数据准备使人在回路的方式产生了深刻的变革，这既体现在终端用户的技术背景上，也体现在人参与计算的方式上。因此，本文首先分析人在回路与数据准备的内在关联，并将现有人在回路的数据准备工作分类为交互式数据准备与众包数据准备。前者面向的是需要进行数据准备的终端用户，目的是通过尽可能少的交互预测用户的意图；而后者面向的是互联网上海量的众包工人，目的是借助他们的识别能力提升数据准备的效果。基于上述分类，本文深入地梳理了交互式数据准备和众包数据准备的最新研究进展。

2 人在回路的数据准备概览

人在回路是指计算问题的求解需要人的参与或引入人的参与能提升问题求解的效果[5]。传统的数据准备，如前文介绍的ETL，其终端用户通常是企业的IT专家，他们通过一定的领域描述语言（domain specific language，DSL）进行编码，完成数据从一个数据源到另一个数据源的转换工作流，例如，从业务数据库中提取用户的行为，转换为数据立方体（data cube），并存储到数据仓库中。

美国威斯康星大学Anhai Doan教授对人在回路系统中人的定义提出了2个刻画维度[5]，这里用来分析数据准备中参与人的变化。

一是终端用户的技术背景，由传统ETL系统中的IT专家变为没有或是仅具备有限编程能力的领域用户。这使得通过DSL编写数据准备工作流的方式变得越来越困难，亟待新型的交互方式，如通过更好的用户界面、可视化等方式，让用户更加便捷地表达数据准备的需求。

二是人的参与方式，由传统ETL系统中的一个IT专家或一个小规模的专家团队变为更丰富的参与方式：既包含传统的单一用户、小规模团队，也包含互联网上海量的群众用户，即众包工人。这给数据准备的计算模式带来了新的机会——由传统的单纯依赖机器变成人机协作的计算模型。例如，针对多源实体识别问题，传统的ETL系统多采用基于相似度函数的模糊匹配方法，往往难以根据实际数据选择恰当的相似度函数与阈值，而借助众包工人，通常能得到更好的结果。

人在回路方式的变化驱动了数据准备技术的革新。**图2**给出了相关研究的概览。人在回路的数据准备技术主要侧重服务领域用户，其目的是将相关应用场景的原始数据整理为数据分析或机器学习方法可以使用的数据。本文结合前面的分析从以下2个方面对现有的研究工作进行梳理。

首先，由于数据准备的终端用户逐渐以仅具备有限IT知识的领域用户为主，传统依靠IT专家编写DSL脚本的方式变得不再适用。因此，越来越多的研究工作向为终端用户提供交互式数据准备的方式转变，一方面提供更为优化的界面，通过迭代的方式与用户交互；另一方面更加依赖算法推断用户的数据准备需求，仅在必要的时候引入用户的参与。交互式数据准备

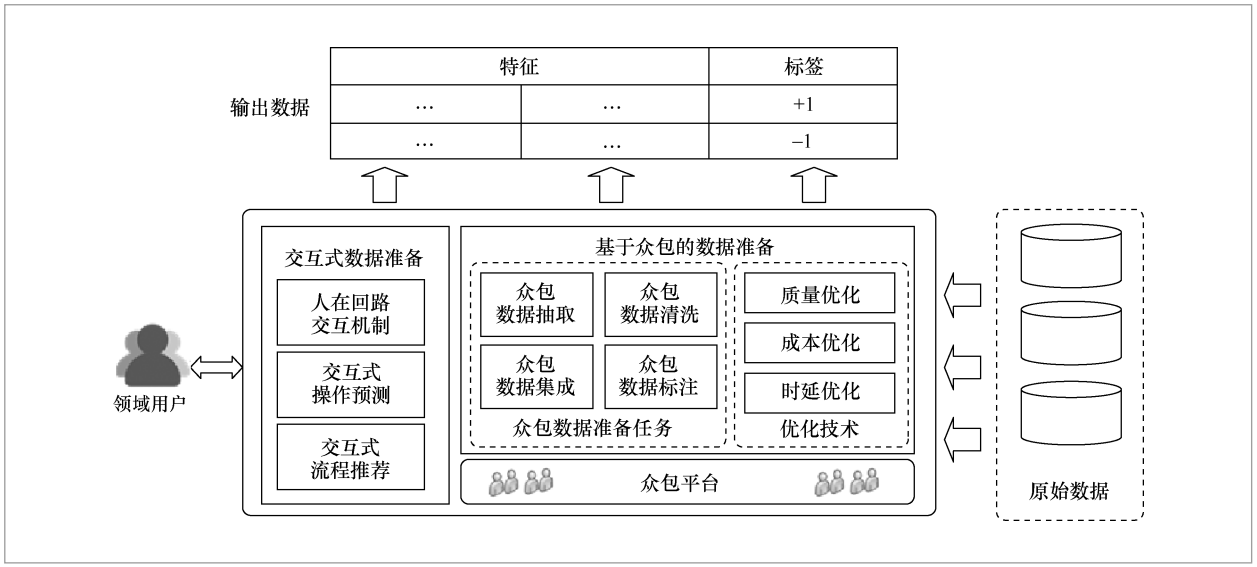


图 2 人在回路数据准备的研究概览

的难点一方面在于需要准确地预测适用于用户数据集与分析任务的数据准备步骤以及最优的参数，另一方面在于需要提升算法执行的效率，以保证交互的实时性。本文的第3节将从交互机制、操作预测与流程推荐这3个方面梳理交互式数据准备的研究进展。

其次，众包工人的参与给数据准备带来了新的机会——能够通过人机协作的方式解决传统单纯依靠机器难以解决的问题，提升数据准备的效果。因此，众包技术被广泛应用于数据准备过程中，如数据抽取、数据清洗、集成与标注。其基本思想是向互联网上的众包平台发布众包问题，吸引大量的互联网用户（称为众包工人）作答，以相对低廉的单价，将人的认知与处理能力引入数据准备的过程中。然而，由于数据集的规模越来越大，保证众包结果的质量、成本与时延变得越来越有挑战性。第4节将梳理众包数据准备的研究进展，包括众包数据准备的核心任务与众包优化。

不难看出，虽然都引入了人的参与，交互式数据准备与众包数据准备在谁来做（who）、为何做（why）和做什么（what）这3个基本问题上存在本质不同。具体地，交互式数据准备面向的是需要进行数据准备的终端用户（如上文提到的领域用户），目的是通过尽可能少的交互预测用户的意图，主要的工作是设计有效的交互机制和预测算法。与此相对，众包数据准备面向的是互联网上海量的众包工人，目的是借助他们的识别能力提升数据准备的效果，主要的工作是设计众包任务和进行质量控制。需要指出的是，很多相关工作也研究了众包的交互界面和交互机制[6-9]，但其目的是提升众包完成的质量、降低众包成本或时延，与本文的交互式数据准备是不同的概念。

此外，现有的文献也采用了类似的方式对人在回路的数据准备进行了分类。例如Data Tamer系统将参与人分为管理员与领域专家[10]，前者通过与系统的交互形成数据准备工作流，后者起到了类似众包的作用，完成一些挑战性强的任务，如实体识别。采用类似分类方式的还包括参考文献[11-14]

等。与这些文献不同，本文提供了更为全面的调研方法。

3 交互式数据准备技术

交互式数据准备的目标是通过与领域用户构建有效的交互机制，高效高质地完成数据转换、清洗、集成等相关操作，节省数据准备的时间。本文对领域用户做了进一步的细分，将现有的交互式数据准备工作按照交互机制分为以下2类。

- 基于菜单界面的数据准备。这类方法假设领域用户不具备编程能力，或编程能力极为有限，主张构建简单的界面与用户进行交互，将用户在界面上的操作转换为数据准备的脚本去执行。在此基础上，现有参考文献提出了一些预测性的研究，如自动补全、基于实例的脚本生成等。
- 基于交互式编程的数据准备。这类方法假设领域用户具备一定的编程能力，如掌握Python语言，能够使用交互式编程环境（如IPython）进行编程。然而，由于领域用户缺乏数据处理的经验以及数据准备过程冗长，系统通过交互式推荐的方式对领域用户进行指引，并辅助其确定参数设置。

3.1 基于菜单界面的数据准备

基于菜单界面的数据准备假设用户不具备编程能力，希望通过界面与领域用户反复交互，并将用户在界面上的操作“翻译”成数据准备操作脚本，以缩短数据准备的时间。**图3**给出了交互式数据准备系统Wrangler[15]的界面示意。用户可以直接在数据表格上进行操作（**图3**中右侧），如选择出州信息，并将其提取到一个新的数据列中。**图3**的左侧显示的是系统根据用户操作翻译出的脚本，以便用户直接在脚本上编辑。后续Wrangler系统得到了进一步的优化[16]，并产业化为商用数据整理工具Trifacta。类似的系统还包括美国谷歌公司的OpenRefine系统和麻省理工学院的Data Tamer系统[10]。

围绕这种用户交互机制，近年来涌现出了一系列的研究工作。这些工作的基本思路是预测用户意图，自动或半自动地生

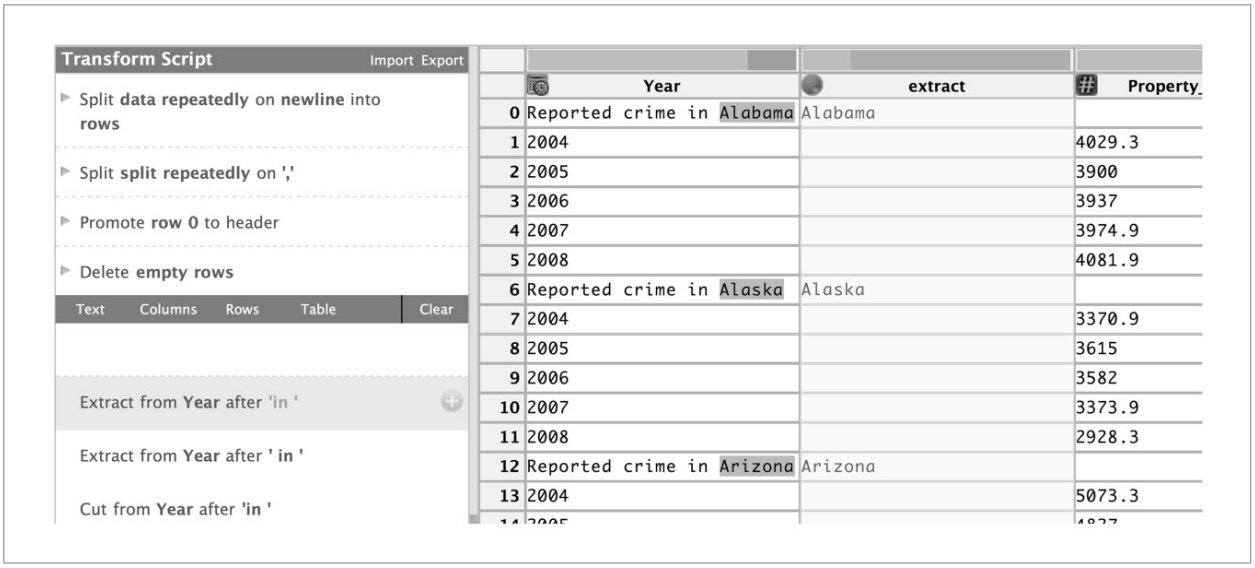


图 3 基于交互界面的数据准备 [15]

成数据准备脚本，可以将相关工作分为以下两大类。

（1）基于用户操作的预测

这方法直接针对用户在界面上的操作进行预测。这方面典型的工作有Trifacta、InformaticaRev与Talend。以**图3**中的州名提取为例，如果用户针对“Alabama”进行了高亮操作，潜在的提取脚本可能有多种，如仅提取“Alabama”这一单词，提取“Reported crime in”后面的第一个单词，或是提取字符串下标18~24之间的子串等。系统对这些候选提取脚本按照预测的相关性进行排序，用户可以从中选择最符合自己意图的脚本，并将其应用到数据集的其他部分，如从图中的第6行与第12行分别提取的单词Alaska与Arizona。同时，为了更好地与用户进行交互，也有参考文献研究数据准备脚本的可视化问题[17]。然而，基于用户操作的预测方法往往准确率不高，其原因是该方法只利用了“输入”，即用户选择哪些数据进行转换，而忽略了与“输出”（即用户希望达成的转换目标）进行对应。同时考虑“输入”和“输出”的方法就是下面介绍的基于输入输出实例的预测。

（2）基于输入输出实例的预测

这类方法让用户提供数据准备前后的输入输出实例，从而根据实例生成数据准备的操作脚本。这里以字符串转换为例，如将日期格式由“yyyymmdd”转换为“dd/mm/yy”，用户只需要提供一个或多个输入输出实例，如“20190101”“01/01/19”“19980316”“16/03/98”，系统就将自动生成数据转换的脚本。为了实现这一目标，现有参考文献提出了2种技术路线。

- 基于实例的程序合成方法。程序自动合成（program synthesis）或实例编程（program by example）是编程语言领域的经典研究问题[18]。一些研究工作将相关的技术引入数据转换[19-20]、数据清洗[21]与数据集成[22-23]中，通过实例自动合成一段程序。基于实例程序合成方法的基本想法是首先抽象数据准备的基本算子，进而搜索能够将输入转换成输出的算子序列，再通过一定的策略选择最优的算子序列。然而，基于实例的程序合成的方法难以解决复杂的数据准备问题[24]。
- 基于搜索引擎的方法。与程序合成不同，基于搜索引擎的方法将实例预测建模成一个搜索的问题。例如，一些方法使用搜索引擎找到与输入输出实例相关的万维网表格数据[25-26]（即Web Tables[27]），并从这些数据中提取其他转换实例。然而，由于数据的稀疏性和需求的定制性，该方法在相对复杂的场景下效果有限。因此又有方法提出对企业内部或互联网上（如Github和Stackoverflow）与数据准备有关的源代码进行索引[24]。在给定输入输出实例后，推荐相关的源代码以合成处理脚本。

3.2 基于交互式编程的数据准备

随着人工智能相关技术的不断普及，数据分析呈现了2个新的趋势。一是数据分析任务越来越“定制化”，导致数据准备的复杂程度越来越高。**图4**给出了一个使用深度神经网络——长短期记忆网络（LSTM）[28]进行文本分析的数据准备流程，先后经过数据集切分、分词、规范大小写、去除停用词、删除特殊符号、填充、单词嵌入等多个步骤，而且每步都需要选择合适的参数。二是随着Python语言与交互式编程环境的普及，越来越多的领域用户掌握了基础的编程能力。例如有调查研究表明，越来越多的社会科学领域（如教育、传媒、经济）的研究者开始使用计算和编程的方法进行科学研究。因此，阻碍

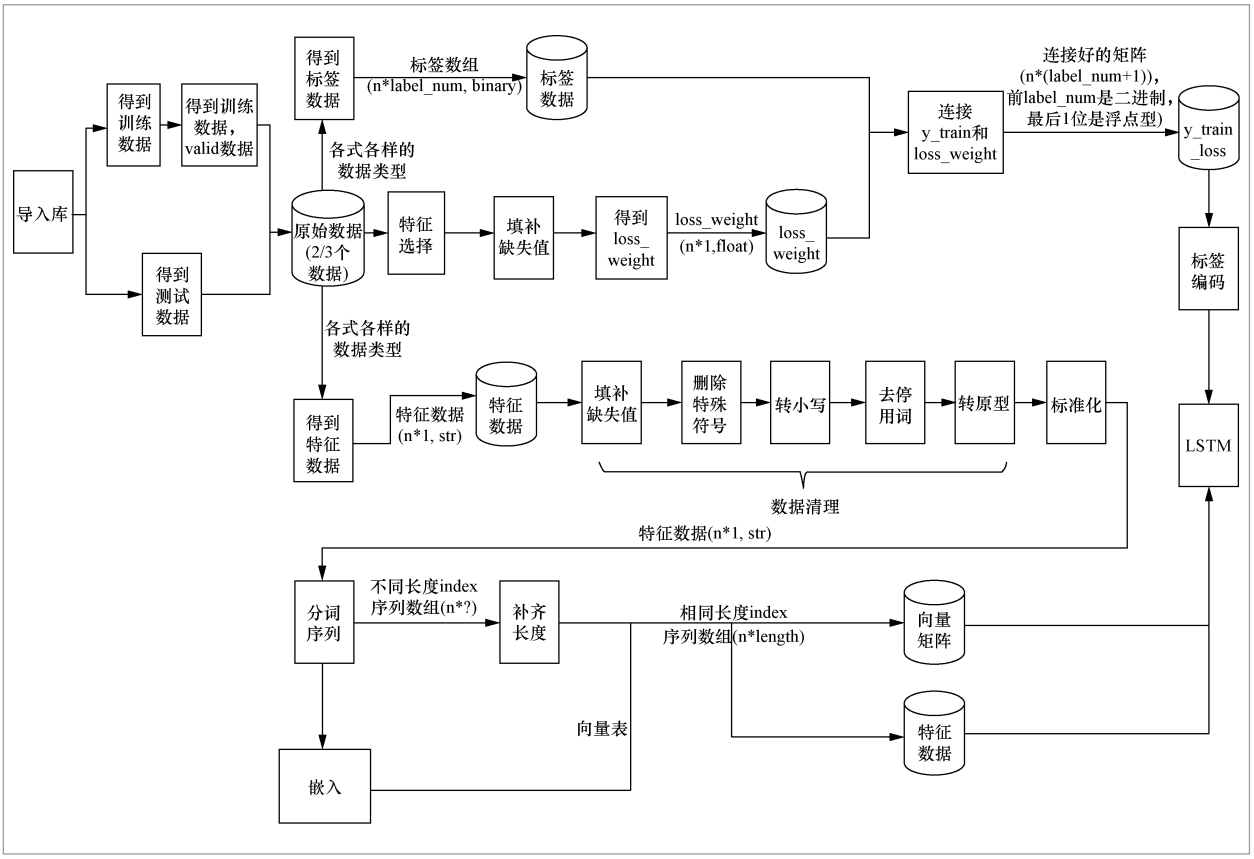


图 4 使用 LSTM 模型进行文本分析的一般流程

领域用户做数据准备的核心瓶颈因素已经不再是编程，而是数据处理的技能与数据准备的经验，如应该调用哪些函数完成特定功能以及如何选择最佳参数等。

在这一背景下，基于交互式编程的数据准备得到了越来越多的关注，它假设用户可以通过编写程序进行数据准备。因此，基于交互式编程的数据准备更侧重通过与用户的交互，为用户推荐相关的数据集、恰当的处理函数、优化的参数等。这方面的研究方兴未艾，本节将着重从以下两点进行介绍。

(1) 交互式编程界面

目前，领域用户或数据科学家越来越倾向于使用计算型笔记本（computational notebook）进行编程，例如Jupyter Notebook、Apache Zeppelin与RStudio Notebook等。这种编程方式的好处是多方面的：一是基于网络，用户本地可以不用部署开发环境；二是支持交互，用户执行计算步骤，可以实时地看到结果；三是文档驱动，用户可以将代码段和文字内容编排在一起，提升程序的解释性与复现性。因此，近期的一些研究工作将计算型笔记本的交互方式引入数据准备中。例如：Juneau[29-30]在Jupyter Notebook上进行了扩展，它首先在数据科学网站Kaggle上下载了大量的网友分享的数据分析可编程记事本，进而通过这些Notebook为新的用户数据准备过程推荐相关数据集、辅助提取特征或数据清洗与集成的工作流。

（2）交互式可视化与工作流生成

领域用户在使用交互式编程做数据准备的过程中，通常希望可以得到来自系统2个层面的帮助。第一是数据层面，用户希望可以从数据中观察到一些现象或趋势，从而有针对性地进行清洗、集成或扩充操作。例如：如果发现某个属性的数据分布不符合预期，用户可能会采取异常值检测等操作。近些年交互式可视化技术得到了迅猛的发展[31-32]，为用户便捷地观察数据提供了更好的帮助。第二是工作流层面，用户希望能够得到工作流的推荐——针对当前的数据集与任务，应该进行哪些步骤以及每个步骤中参数如何设定。在这方面，一些工作已经开始了尝试，例如Data Civilizer系统[33]整体考虑数据准备与数据分析，做出整体优化；Alpine Medow系统[34]提出了基于规则的优化方法，通过与用户交互生成数据准备与分析的工作流，能够显著地节省领域用户的时间。

3.3 未来挑战

尽管已经取得了一定的进展，交互式数据准备还有一系列挑战性问题亟待解决。

- 基于交互推荐的数据准备。推荐技术被广泛地应用在商品、电影等领域，取得了显著的成果。本文认为推荐技术也同样可以应用在数据准备中。互联网上有着丰富的数据准备代码（如Kaggle网站分享的notebooks或Github上开源的代码）与数据资源，这些都为基于交互推荐的数据准备创造了条件。
- 交互性能优化。在交互式数据准备中，性能是一个十分关键的因素，因此需要研究性能优化方法。一方面，需要对单一操作进行优化，例如使用采样技术，提升数据清洗或其他步骤的性能[35]；另一方面，需要对整体的工作流做性能优化，例如避免重复的数据迁移、提高数据处理的并行度等。
- 易用性交互方式研究。交互式数据准备直接面向领域用户，如何给他们提供易用的交互方式非常重要。例如，传统的数据库领域使用断言式的查询，这种方式是否有可能应用到数据准备中，能否融合菜单界面与交互式编程二者的优势，提供更加易用的交互方式。这些都是开放且需要深入研究的问题。

4 基于众包的数据准备技术

众包是一种将某个复杂的计算问题分解成大量简单任务（称为微任务）发布给互联网众包平台上的众包工人进行分布式解答的技术。需要指明的是，众包数据准备同样是一个迭代的过程，即通过多个轮次地向众包平台发布任务与收集众包工人返回的答案，而有效地支持数据准备的核心任务。例如，通过众包进行实体识别，可以发布一定的众包任务，并基于返回的结果按照一定的规则（如传递性）进行推理。不断迭代地重复上述过程，直到完成所有实体对的判别工作。此过程中的技术挑战在近些年得到了广泛的关注和深入的研究，包括众包质量控制、众包数据库系统与众包激励机制等，有相关的综述类文章可以参考[36]。首先在第4.1节介绍众包数据准备的核心任务，进而在第4.2节梳理众包的成本优化技术，最后在第4.3节探讨未来的研究挑战。

现有文献将众包技术应用于数据准备的核心任务上，包括数据抽取、清洗、集成与标注。为了更好地说明这些任务，**图5**给出了一个公司数据集准备场景下的示例。

4.1 众包数据准备的核心任务

(1)众包数据抽取

数据抽取是指从非结构化数据(如文本)中提取结构化的信息,如**图5**从HTML网页中抽取公司名及其营收属性。传统的数据抽取方法基于规则、词典或机器学习模型,在抽取精度上难以达到令人满意的效果。因此一些方法引入了人提取信息的能力,提出了众包数据抽取,取得了明显高于自动数据抽取方法的效果。相关的研究包括众包实体识别[37]、众包实体属性与实体关系抽取[38-39]以及众包类别体系构建[40]。这些方法首先通过自动的数据抽取方法生成候选的结果,然后让众包工人进行验证。

(2)众包数据清洗

数据清洗是指发现并纠正数据中潜在的错误,以确保数据的质量。如**图5**所示,真实世界的数据可能会有各种各样的错误,包括数据缺失(左上表格中沃尔玛的产业属性值缺失)、取值错误(左下表格中沃尔玛的营收仅为469.2美元)、记录重复(左下表格的第1条与第4条记录重复)等。众包数据清洗技术借助众包工人对上述错误进行检测与修复,提升数据清洗的效果,其基本思想是生成一些验证性问题(如沃尔玛的产业是否是零售)发布给众包工人进行验证。相关的技术手段包括借助知识图谱或Web资源生成最有收益的众包问题[41-42]、使用抽样技术生成有一定理论保证的众包问题[35,43]等。核心的挑战在于可能的验证性问题会很多,因此需要设计有效的问题选择策略,选择最有价值的问题进行众包,并通过数据之间的关联或相应的领域知识进行结果推理。

(3)众包数据集成

数据集成是指融合多个数据源以得到更加全面的数据,如在**图5**中将关系数据(左上表格)与网页抽取数据(左下表格)融合,从而汇总公司基础信息与营收数据。数据集成包含3类基本任务:一是模式匹配

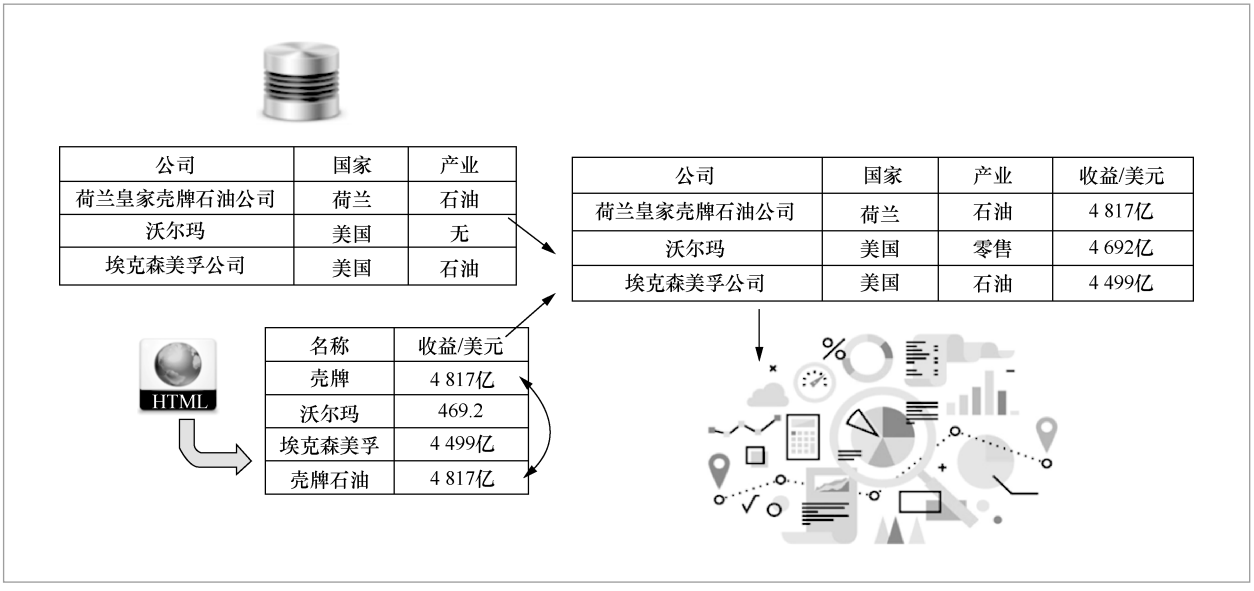


图 5 众包数据准备的基本任务示例

（schema matching）[27]，即将不同数据源中语义相同的数据列关联起来，如左上表格的公司列和左下表格的名称列；二是实体识别（entity resolution），即将不同数据源中表征不同的实体匹配起来[6-7]，例如左上表格中的荷兰皇家壳牌石油公司和左下表格中的壳牌/壳牌石油；三是属性融合，即对不同数据源中属性值可能存在的冲突现象进行消解[44]。由于不同数据源之间存在的异质性，自动的数据集成方法通常难以达到满意的精度。因此，近些年众包数据集成得到了广泛的研究，现有工作相对聚焦于挑战性最强的实体识别任务。主要解决的核心难点是成本优化，方法包括众包问题设计[6-7]、基于传递性或偏序关系的问题推理[13,45-47]与问题选择[14,48]等。

（4）众包数据标注

数据标注是指根据特定的数据分析任务为数据打上类别标签，从而支撑后续的模型训练。此外，前文介绍的数据抽取与数据集成任务也可以采用机器学习模型解决，数据标注在其中起关键作用。数据标注是公认的“脏活累活”，需要耗费大量的人力成本。众包的引入虽然降低了数据标注的单价，但当数据规模急剧扩张时，成本也是难以承受的。因此，现有的数据标注工作侧重研究如何降低众包成本。相关文献[49-50]提出了基于弱监督规则的方法，将领域用户的专家规则、领域知识图谱、众包等一系列标注数据源融合起来提高数据标注的覆盖率与准确率。然而，由于弱监督规则通常良莠不齐，一些方法提出使用众包对规则进行选择[51-52]。

4.2 众包数据准备优化技术

众包数据准备的优化目标包含3个方面：质量优化、时延优化与成本优化。首先，由于众包平台的开放性，众包工人提供的答案可能存在噪声，因此需要进行质量优化[53-55]；其次，由于众包平台的动态性，发布的任务不一定能够及时完成，因此需要进行时延优化[56]；最后，大多数众包平台的激励机制是付费（如亚马逊MTurk平台、CrowdFlower平台等），尽管每个任务的费用相对低廉，但问题规模变大时，成本也是难以承受的，因此，需要进行成本优化。

成本优化是众包数据准备的核心优化问题，现有文献提出了很多不同的优化策略。这些方法的基本思路均是通过迭代的方式进行众包，从而在质量、成本、时延3个方面进行权衡取舍。本节将其中具有代表性的工作分为以下3类，分别做出介绍。

（1）众包结果推理策略

众包结果推理策略通过一定的规则从已经众包的结果中推理出部分未众包任务的结果，从而降低众包问题的规模。在数据准备的很多任务中，一种典型的推理规则是传递性[45,47]。以**图5**中的实体识别任务为例：如果众包已经返回荷兰皇家壳牌石油公司和壳牌是同一实体、壳牌和壳牌石油是同一实体，那么无须问众包即可得到荷兰皇家壳牌石油公司和壳牌石油是同一实体。基于传递性的众包结果推理研究聚焦2个难点问题：一是计算最优的提问次序，以保证传递性最大程度的发挥；二是控制个别问题的错误通过传递性被传播与扩大，以保证众包结果的质量。除了传递性，一些研究也采用了偏序关系进行推理[13,46]，其基本思想是根据数据准备的场景建立众包任务之间的偏序关系（如任务A的结果为匹配，则任务B的结果也为匹配），将众包任务建模成一个有向无环图，进而利用图上的算法进行任务的选择与推理。此外，结合特定数据准备场景，如企业数据清洗与基于知识图谱的实体抽取，一些领域相关的规则也可以用来进行推理[37,41]。众包结果推理策略的优点在于

设计简单，实际效果也比较好；缺点是比较依赖推理规则的可靠性，如果推理规则不适用于当前数据集，往往会放大错误，产生低质的众包结果。

（2）众包任务选择策略

众包任务选择策略的基本思想是选择出最有“价值”的任务进行众包。这类策略中最典型的例子是基于主动学习（active learning）的方法[47,57]。与传统的监督学习不同，主动学习的目标是选择对训练模型最有价值的数据实例，其中对“价值”的度量有很多标准，比如当前模型预测的不确定性、模型更新的程度等。在一些数据准备任务中，使用主动学习的任务选择策略能够在众包成本受限的前提下得到很好的效果[47]。同时，相关的研究也提出了针对具体众包操作（如求最大值或条件过滤）的任务选择策略[58-60]。最近，一些文献提出了博弈众包的策略进行任务选择[51-52]，从而支持数据标注任务。该策略首先生成候选规则，然后考虑覆盖率和精度来选择高质量的规则。博弈众包策略雇佣两组众包工人：一组回答规则验证任务，以发挥规则生成器的作用，而另一组回答元组检查任务，扮演规则审查者的角色。让2个小组玩一个双人游戏，迭代地调用规则生成器和规则转换器，直到众包预算用完为止。实验表明，博弈众包策略能够生成高质量的数据标注规则，在保持质量的同时大幅降低众包成本。众包任务选择策略的优点在于人机协作，仅在必要的时候引入众包的参与，并在众包结果的基础上用机器学习模型进行推断，因此能够在成本和质量方面做较好的权衡。该策略的缺点在于其通常采用多轮众包方式，增大了时延。

（3）众包任务抽样策略

众包任务抽样策略的基本思想是让众包只解决一小部分数据样本上的问题，进而推理出整个数据集上的结论。任务抽样策略被成功地应用于数据清洗任务中[35]。以**图5**中左下表格的收益（revenue）属性为例。考虑数据准备的目的是统计出公司的平均营收，其中部分数据可能存在的错误（如第2行的469.2美元）会对结果造成很大影响。众包任务抽样策略首先获取全体数据的一个采样，然后只在采样数据上发布众包任务，并收集答案，最后推断出全体数据上的结果。以平均营收为例，现有抽样技术的一些性质可以从理论上保证推断出的结果在一定的误差范围内。然而，该策略的缺点在于其局限于特定的数据准备任务，如统计公司的最大营收，则采样的策略很难给出准确的推断结果。

4.3 未来挑战

尽管已经取得了一定的进展，众包数据准备还有一系列具有挑战性的问题亟待解决。

- 众包数据准备系统。现有的研究工作多侧重单点的理论研究，缺乏系统性。因此，亟待研究一个整合众包数据准备核心任务、提供系统性优化的系统。为了达成这一目标，需要设计并实现众包数据准备算子，构建数据准备工作流，提供整体优化机制，设计良好数据准备查询语言与交互方式等。
- 人机混合优化算法。众包数据准备成本优化的趋势是人机混合，即综合发挥众包的识别能力与机器的推理能力，在成本可控的前提下，提高数据准备的效果。尽管目前已经提出了一些人机混合优化的算法，但系统性、理论性的研究还不多，结合众包实际场景设计的算法还比较欠缺，需要进一步的深入。
- 众包隐私保护机制。众包隐私保护分为2个层面：第一，数据已成为政府、企

业和机构的重要资源，如何既保护数据隐私，又能借用众包解决数据准备的问题，是一个非常大的挑战；第二，众包工人也面临着隐私泄露的风险（如位置信息等），如何在保证答题质量的前提下，保护众包工人的隐私，也是研究的热点问题[61]。

5 结束语

数据准备已经成为数据治理的一项关键支撑技术。没有成功的数据准备，数据的可用性与数据分析的效果就会大打折扣。本文从人在回路的角度介绍了数据准备的研究进展，然而需要进一步研究的工作还有很多，除了前文提及的之外，这里再给出一些代表性的挑战问题。

● 隐私保护。数据已成为政府、企业和机构的重要资源，妥善处理隐私保护与开放共享之间的矛盾变得越来越迫切。这里需要权衡考虑2个维度：一是隐私保护的质量。例如能否应对典型的数据隐私攻击，如重识别攻击（reidentification attack）与成员攻击（membership attack），或是否有理论保证；二是生成数据的效用，主要考量使用生成数据训练分析模型能否在相同的测试数据上达到与真实数据训练的模型近似的性能。

● 系统构建。目前，数据准备的相关工作还比较分散，没有进行系统化的集成。在这方面，一些研究者做出了很有价值的尝试，如美国麻省理工学院的Michael Stonebraker教授等人[10]搭建Data Tamer系统；美国威斯康星大学麦迪逊分校的AnHai Doan教授等人[11]提出不必另起炉灶，应充分利用Python开源社区PyData。然而，整体来看，数据准备系统构建方面的研究和业界的尝试还很匮乏，这使数据准备难以形成整体的研究。

● 评测基准（benchmark）。评测基准是数据库领域经常采用的一种衡量系统或算法性能的方法，如评测联机事务处理（OLTP）型关系数据库的TPCC基准。然而，在数据准备领域缺乏系统性的评测基准，这带来两方面的问题：首先，不同的数据准备算法可能只对特定的数据有效，缺乏一般性；其次，在给定具体分析人物的情况下，用户不知道如何选择有效的数据准备方法。因此，需要建立合理有效的评测基准。

参考文献:

[1] 杜小勇，陈跃国，范举，等. 数据整理——大数据治理的关键技术[J]. 大数据，2019，5(3)：13-22.
DU X Y, CHEN Y G, FAN J, et al. Data wrangling: a key technique of data governance[J]. Big Data Research, 2019, 5(3): 13-22.

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

[3] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]// Computer Vision and Pattern Recognition (CVPR), June 20-25, 2009, Miami, USA. Piscataway: IEEE Press, 2009: 248-255.

[4] YANG Y, MENEGHETTI N, FEHLING R, et al. Lenses: an on-demand approach to ETL[J]. Proceedings of the VLDB Endowment, 2015, 8(12): 1578-1589.

[5] DOAN A H. Human-in-the-loop data analysis: a personal perspective[C]// The Workshop on Human-In-the-Loop Data Analytics (HILDA@SIGMOD 2018), Jun 10-15, 2018, Houston, USA. New York: ACM Press, 2019: 1-6.

[6] VERROIOS V, GARCIA-MOLINA H, PAPAKONSTANTINOU Y. Waldo: an adaptive human interface for crowd entity resolution[C]// International Conference

on Management of Data (SIGMOD), May 14–19, 2017, Chicago, USA. New York: ACM Press,2017: 1133–1148.

[7] WANG J, KRASKA T, FRANKLIN M J, et al.CrowdER: Crowdsourcing Entity Resolution[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1483–1494.

[8] BERNSTEIN M S, BRANDT J, MILLER R C, et al. Crowds in two seconds: enabling realtime crowd-powered interfaces[C]// Annual ACM Symposium on User Interface Software and Technology (UIST), October 16–19, 2011, Santa Barbara, USA. New York: ACM Press, 2011: 33–42.

[9] HAAS D, WANG J, WU E, et al. CLAMShell: speeding up crowds for low-latency data labeling[J]. Proceedings of the VLDB Endowment, 2015, 9(4): 372–383.

[10] STONEBRAKER M, BRUCKNER D, ILYAS I F, et al. Data curation at scale: the data tamer system[C]// Biennial Conference on Innovative Data Systems Research (CIDR),January 6–9, 2013, Asilomar, USA. [S.l:s.n.], 2013.

[11] 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.

[12] CHEN C, GOLSHAN B, HALEVEY A Y, et al. BigGorilla: an open-source ecosystem for data preparation and integration[J]. IEEE Data Engineering Bulletin, 2018, 41(2): 10–22.

[13] LI G.Human-in-the-loop data integration[J]. Proceedings of the VLDB Endowment, 2017, 10(12): 2006–2017.

[14] FAN J, LI G. Human-in-the-loop rule learning for data integration[J]. IEEE Data Engineering Bulletin, 2018, 41(2): 104–115.

[15] KANDEL S, PAEPCKE A, HELLERSTEIN J M, et al. Wrangler: interactive visual specification of data transformation scripts[C]// International Conference on Human Factors in Computing Systems (CHI), May 7–12, 2011, Vancouver, Canada. New York: ACM Press,2011: 3363–3372.

[16] HEER J, HELLERSTEIN J M, KANDEL S. Predictive interaction for data transformation[C]// Biennial Conference on Innovative Data Systems Research (CIDR), January 4–7, 2015, Asilomar, USA. [S.l:s.n.], 2013.

[17] KHAN M A, XU L, NANDI A, et al. Data tweening: incremental visualization of data transforms[J]. Proceedings of the VLDB Endowment, 2017, 10(6): 661–672.

[18] LIEBERMAN H. Your wish is my command: programming by example[M]. Morgan Kaufmann Publishers, 2001.

[19] JIN Z, ANDERSON M R, CAFARELLA M J, et al. Foofah: Transforming data by example[C]// International Conference on Management of Data (SIGMOD), May 14–19, 2017, Chicago, USA. New York: ACM Press, 2017: 683–698.

[20] BLINKFILL R S. Semi-supervised programming by example for syntactic string transformations[J]. Proceedings of the VLDB Endowment, 2016, 9(10): 816–827.

[21] SINGH R, MEDURI V V, ELMAGARMID A K, et al. Synthesizing entity matching rules by examples[J]. Proceedings of the VLDB Endowment, 2017, 11(2): 189–202.

[22] BONIFATI A, COMIGNANI U, COQUERY E, et al. Interactive mapping specification with exemplar tuples[C]// International Conference on Management of Data (SIGMOD),May 14–19, 2017, Chicago, USA. New York: ACM Press, 2017: 667–682.

[23] ZHU E, HE Y, CHAUDHURI S. Auto-join: joining tables by leveraging transformations[J]. Proceedings of the VLDB Endowment, 2017, 10(10): 1034–1045.

[24] HE Y, CHU X, GANJAM K, et al. Transform-data-by-example (TDE): an extensible search engine for data transformations[J]. Proceedings of the VLDB Endowment, 2018, 11(10): 1165–1177.

[25] MORCOS J, ABEDJAN Z, ILYAS I F, et al. DataXFormer: an interactive data transformation tool[C]// International

Conference on Management of Data (SIGMOD), May 31-June 4, 2015, Melbourne, Australia. New York: ACM Press,2015: 883-888.

[26] ABEDJAN Z, MORCOS J, ILYAS I F, et al. DataXFormer: a robust transformation discovery system[C]// IEEE International Conference on Data Engineering (ICDE), May 16-20, 2016, Helsinki, Finland. Piscataway: IEEE Press, 2016: 1134-1145.

[27] FAN J, LU M, OOI B C, et al. A hybrid machine-crowdsourcing system for matching web tables[C]//IEEE International Conference on Data Engineering (ICDE), March 31-April 4, 2014, Chicago, USA. Piscataway: IEEE Press, 2014: 976-987.

[28] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[29] ZHANG Y, IVES Z G. Juneau: data lake management for Jupyter[J]. Proceedings of the VLDB Endowment, 2019, 12(12): 1902-1905.

[30] IVES Z, ZHANG Y, HAN S, et al. Dataset relationship management[C]// Biennial Conference on Innovative Data Systems Research (CIDR), January 13-16, 2019, Asilomar, USA. [S.l:s.n.], 2019.

[31] VARTAK M, RAHMAN S, MADDEN S, et al. SEEDB: efficient data-driven visualization recommendations to support visual analytics[J]. Proceedings of the VLDB Endowment, 2015, 8(13): 2182-2193.

[32] LUO Y, QIN X, TANG N, et al. DeepEye: towards automatic data visualization[C]// IEEE International Conference on Data Engineering (ICDE), April 16-19, 2018, Paris, France. Piscataway: IEEE Press, 2018: 101-112.

[33] REZIG E K, CAO L, STONEBRAKER M, et al. Data civilizer 2.0: a holistic framework for data preparation and analytics[J]. Proceedings of the VLDB Endowment, 2019, 12(12): 1954-1957.

[34] SHANG Z,ZGRAGGEN E, BURATTI B, et al. Democratizing data science through interactive curation of ML pipelines[C]// International Conference on Management of Data (SIGMOD), June 30 - July 5, 2019, Amsterdam, The Netherlands. New York: ACM Press, 2019: 1171-1188.

[35] WANG J, KRISHNAN S, FRANKLIN M J, et al. A sample-and-clean framework for fast and accurate query processing on dirty data[C]// International Conference on Management of Data (SIGMOD), June 22-27, 2014, Salt Lake City, USA. New York: ACM Press, 2014: 469-480.

[36] LI G, WANG J, ZHENG Y, et al. Crowdsourced data management: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(9): 2296-2319.

[37] DEMARTINI G, DIFALLAH D E, CUDRÉ-MAUROUX P. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking[C]// International World Wide Web Conferences (WWW), April 16-20, 2012, Lyon, France. [S.l:s.n.], 2012: 469-478.

[38] KONDREDDI S K, TRIANTAFILLOU P, WEIKUM G. Combining information extraction and human computing for crowdsourced knowledge acquisition[C]// International Conference on Data Engineering (ICDE), March 31 - April 4, 2014, Chicago, USA. Piscataway: IEEE Press, 2014: 988-999.

[39] ABAD A, NABI M, MOSCHITTI A. Self-Crowdsourcing training for relation extraction[C]// Annual Meeting of the Association for Computational Linguistics (ACL), July 30 - August 4, 2017, Vancouver, Canada. [S.l:s.n.], 2017: 518-523.

[40] CHILTON L B, LITTLE G, EDGE D, et al. Cascade: crowdsourcing taxonomy creation[C]// International Conference on Human Factors in Computing Systems (CHI), April 27 - May 2, 2013, Paris, France. New York: ACM Press, 2013: 1999-2008.

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

[42] TONG Y, CAO C C, ZHANG C J, et al. CrowdCleaner: Data cleaning for multi-version data on the web via crowdsourcing[C]// IEEE International Conference on Data Engineering (ICDE), March 31 - April 4, 2014, IL, USA. [S.l:s.n.], 2014: 1182-1185.

[43] DOLATSHAH M, TEOH M, WANG J, et al. Cleaning crowdsourced labels using oracles for statistical classification[J]. Proceedings of the VLDB Endowment, 2018, 12(4): 376-389.

[44] GAO J, LI Q, ZHAO B, et al. Truth discovery and crowdsourcing aggregation: a unified perspective[J]. Proceedings of the VLDB Endowment, 2015, 8(12): 2048-2049.

[45] WANG J, LI G, KRASKA T, et al. Leveraging transitive relations for crowdsourced joins[C]// International Conference on Management of Data (SIGMOD), June 22-27, 2013, New York, USA. New York: ACM Press, 2013: 229-240.

[46] CHAI C, LI G, LI J, et al. Cost-effective crowdsourced entity resolution: a partial-order approach[C]// International Conference on Management of Data (SIGMOD), June 26 - July 1, 2016, San Francisco, USA. New York: ACM Press, 2016: 969-984.

[47] WANG S, XIAO X, LEE C. Crowd-based deduplication: an adaptive approach[C]// International Conference on Management of Data (SIGMOD), May 31-June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 1263-1277.

[48] DAS S, C P S G, DOAN A, et al. Falcon: scaling up hands-off crowdsourced entity matching to build cloud services[C]// International Conference on Management of Data (SIGMOD), May 14-19, 2017, Chicago, USA. New York: ACM Press, 2017: 1431-1446.

[49] RATNER A, BACH S H, EHRENBERG H R, et al. Snorkel: rapid training data creation with weak supervision[J]. Proceedings of the VLDB Endowment, 2017, 11(3): 269-282.

[50] RATNER A J, SA C D, WU S, et al. Data programming: creating large training sets, quickly[C]// Neural Information Processing Systems (NeurIPS), December 5-10, 2016, Barcelona, Spain. [S.l:s.n.], 2016: 3567-3575.

[51] YANG J, FAN J, WEI Z, et al. Cost-effective data annotation using game-based crowdsourcing[J]. Proceedings of the VLDB Endowment, 2018, 12(1): 57-70.

[52] LIU T, YANG J, FAN J, et al. CrowdGame: a game-based crowdsourcing system for cost-effective data labeling[C]// International Conference on Management of Data (SIGMOD), June 30 - July 5, 2019, Amsterdam, The Netherlands. New York: ACM Press, 2019: 1957-1960.

[53] LIU X, LU M, OOI B C, et al. CDAS: a crowdsourcing data analytics system[J]. Proceedings of the VLDB Endowment, 2012, 5(10): 1040-1051.

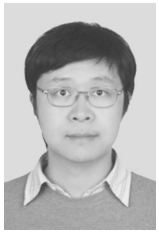
[54] FAN J, LI G, OOI B C, et al. iCrowd: an adaptive crowdsourcing framework[C]// International Conference on Management of Data (SIGMOD), May 31 - June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 1015-1030.

[55] ZHENG Y, WANG J, LI G, et al. QASCA: a quality-aware task assignment system for crowdsourcing applications[C]// International Conference on Management of Data (SIGMOD), May 31 - June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 1031-1046.

[56] HAAS D, WANG J, WU F, et al. CLAMShell: speeding up crowds for low-latency data labeling[J]. Proceedings of the

VLDB Endowment, 2015, 9(4): 372-383.
[57] MOZAFARI B, SARKAR P, FRANKLIN M J, et al. Scaling up crowd-sourcing to very large datasets: a case for active learning[J]. Proceedings of the VLDB Endowment, 2014, 8(2): 125-136.
[58] VERROIOS V, LOFGREN P, GARCIA-MOLINA H. tDP: an optimal-latency budget allocation strategy for crowdsourced MAXIMUM operations[C]// International Conference on Management of Data (SIGMOD), May 31 - June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 1047-1062.
[59] SARMA A D, PARAMESWARAN A G, GARCIA-MOLINA H, et al. Crowd-powered find algorithms[C]// IEEE International Conference on Data Engineering (ICDE), March 31 - April 4, 2014, Chicago, USA. Piscataway: IEEE Press, 2014: 964-975.
[60] BOIM R, GREENSHPAN O, MILO T, et al. Asking the right questions in crowd data sourcing[C]// IEEE International Conference on Data Engineering (ICDE), April 1-5, 2012, Washington, USA. Piscataway: IEEE Press, 2012: 1261-1264.
[61] TO H, SHAHABI C, XIONG L. Privacy-preserving online task assignment in spatial crowdsourcing with untrusted server[C]// IEEE International Conference on Data Engineering (ICDE), April 16-19, 2018, Paris, France. Piscataway: IEEE Press, 2018: 833-844.

作者简介



范举（1984- ），男，博士，中国人民大学数据工程与知识工程教育部重点实验室与信息学院副教授，中国计算机学会会员，数据库专业委员会委员。主要研究方向为数据库与大数据、众包数据管理、数据准备。

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信委员。主要研究方向为大数据分析系统和语义搜索。

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，数据库专业委员会主任，《大数据》期刊编委会副主任，*ACM Transactions on Data Science*编委。主要研究方向为数据库与大数据、智能信息检索、知识工程。

收稿日期： 2019-09-20
基金项目： 国家自然科学基金资助项目（No.61602488，No. 61632016，No. U1711261）
Foundation Items: The National Natural Science Foundation of China(No.61602488, No. 61632016, No. U1711261)