软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(3):1007−1009 [doi: 10.13328/j.cnki.jos.006796]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

大数据治理的理论与技术专题前言*

杜小勇[1,2], 杨晓春[3], 童咏昕[4]

[1](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)

[2](中国人民大学 信息学院, 北京 100872)

[3](东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

[4](北京航空航天大学 计算机学院, 北京 100191)

通信作者: 杜小勇, E-mail: duyong@ruc.edu.cn; 杨晓春, E-mail: yangxc@mail.neu.edu.cn;
童咏昕, E-mail: yxtong@buaa.edu.cn

中文引用格式: 杜小勇, 杨晓春, 童咏昕. 大数据治理的理论与技术专题前言. 软件学报, 2023, 34(3): 1007–1009. http://www.jos.org.cn/1000-9825/6796.htm

数字经济时代, 数据已成为新型生产要素, 大数据技术更是数据要素市场发展的核心科技引擎. 然而, 近年来大数据使用中普遍存在着 “重采集轻管理、重规模轻质量、重利用轻安全” 的现象. 科学而有效地进行大数据治理将有助于提升数据质量、降低管理成本、增强决策能力. 本专题旨在探究大数据治理所面临的核心技术挑战, 面向数据的全生命周期, 不仅研究劣质数据的清洗与修复等数据治理技术, 也讨论隐私安全与开放共享等内容, 还研究利用区块链、联邦学习、知识图谱、数据定价等新技术形成大数据治理的新理论与新方法, 同时关注大数据治理在各应用领域的最新成果.

本专题公开征文, 共收到投稿 40 篇. 论文均通过了形式审查, 内容涉及大数据治理的理论与技术. 特约编辑先后邀请了 40 多位专家参与审稿工作, 每篇投稿至少邀请 2 位专家进行评审. 稿件经初审、复审、NDBC 2022 会议宣读和终审 4 个阶段, 历时 3 个月, 最终有 15 篇论文入选专题. 根据主题, 这些论文可以分为 4 组.

(1) 大数据质量管理技术

《面向列语义识别的共现属性交互模型构建与优化》针对政务数据孤岛系统中元数据语义难以互联互通的问题, 提出了基于预测阶段和纠错阶段的两阶段模型. 在预测阶段, 提出了共现属性交互的 CAI 模型;在纠错阶段, 结合语义标签之间的共现性, 通过引入纠错机制优化模型预测结果.

《面向聚合查询的 Apache IoTDB 物理元数据管理》提出了一种面向聚合查询的 Apache IoTDB 物理元数据管理方案. 该方案按照数据文件的物理存储特性切分数据, 并结合同步计算和异步计算策略, 优化数据的写入性能与系统效率.

《基于多视角的多类型错误全面检测方法》提出了一种基于多视角的多类型错误全面检测模型 CEDM. 结合现有约束条件在属性、单元和元组层面进行多维度的统计分析, 构建基础检测规则, 进而基于语义关系从多个维度上更新扩展基础规则, 进而联合多个视角实现对多种类型错误的全面检测.

《兼顾行列的时序数据质量规则发现》提出了一种针对劣质时序数据治理的数据质量规则发现方法, 依据数据在行与列上依赖信息形成数据质量规则, 并对已有的数据质量规则体系进行表达力的扩展, 同时设计了时序数据质量规则挖掘方法, 实现了高效、准确地挖掘时序数据中隐藏的数据质量规则.

《预训练语言模型实体匹配的可解释性》提出了一系列面向预训练语言模型的实体匹配技术. 针对预训练语言模型的实体匹配技术效果不稳定、匹配结果不可解释的挑战, 采用数据集元特征属性相似度计算与预训练语言模型注意力机制相结合的方法增强低置信度预测结果, 提升实体匹配质量.

(2) 大数据联邦计算技术

* 收稿时间: 2022-10-28; jos 在线出版时间: 2022-10-28

《面向数据联邦的安全多方 θ-连接算法》提出了一种数据联邦的安全多方 θ-连接算法, 在不泄露各自原始数据的前提下, 结合安全多方计算等隐私计算技术设计了一系列优化策略, 显著减少连接查询所需安全计算代价, 从而较大幅度地提升查询效率.

《基于联邦学习的跨源数据错误检测方法》提出了一种基于联邦学习的跨源数据错误检测方法 FeLeDetect, 以在数据隐私保证的前提下利用跨源数据信息提高错误检测精度. 为了降低联邦训练的通信开销和人工标注成本, 设计了一系列优化方法. 从而使得在本地场景和集中场景下错误检测率均有较大幅度的提升.

《基于贡献度证明共识机制的去中心化联邦学习框架》设计了一种高效的去中心化联邦学习框架 EDFL. 通过融合基于贡献度证明的共识机制, 角色自适应激励算法和区块链分区存储策略, 令 EDFL 框架可以降低存储开销, 同时提升联邦学习的学习效率.

《联邦学习贡献评估综述》综述了联邦学习领域中多参与方对学习过程贡献数据的估值指标、贡献评估方案和相关优化技术, 并展望了联邦学习贡献评估当前面临的挑战和未来发展方向.

(3) 复杂动态环境的大数据治理技术

《跳跃滤波: 一种面向大数据治理的动态数据摘要设计》提出了一种面向大数据治理的动态数据摘要技术. 该方法可随数据基数线性增长实现数据处理分析常数级别的处理效率, 从而有效支撑要求苛刻的大数据处理分析任务.

《面向开放大数据环境的动态数据保护系统》提出了一个面向开放大数据环境的动态数据保护系统 BDMasker, 通过基于查询依赖模型的精准查询分析及查询改写技术, 能够实现动态脱敏全过程对业务场景零影响.

《面向大数据分析的分布式矩阵计算系统研究进展》综述了面向大数据治理应用的分布式矩阵计算系统的研究进展, 并从编程接口、编译优化、执行引擎、数据存储这 4 个层面分析了该领域所面临挑战并展望了潜在研究方向.

(4) 大数据治理的应用技术

《基于多粒度注意力网络的知识超图链接预测》提出了一种知识超图多元关系表示模型, 旨在增强知识图谱的数据质量, 进而基于多粒度神经网络对知识图谱缺失关系进行链接预测, 实现多维度、多元关系的整体性图谱补全.

《属性公平的异质信息网络上的社区搜索算法》提出了基于属性公平的异质信息网络上的极大 core 挖掘问题, 设计了 Adv-FkPcore 算法以避免挖掘阶段中子图判定的高计算复杂性挑战, 并结合点标记方法优化算法针对异质信息网络的遍历效率.

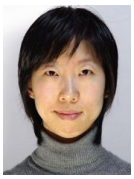
《基于宽容训练和隐私保护的快速监控视频检索模型》提出了一个面向大规模监控视频的安全、快速的视频检索模型. 针对云端算力大、监控摄像头算力规模小的特点, 设计宽容训练策略对其进行定制化知识蒸馏, 将蒸馏后的轻量级模型部署在监控摄像头内, 同时使用局部加密算法对图像敏感部分进行加密, 在极低资源消耗的情况下实现隐私保护.

本专题主要面向数据库、数据挖掘、大数据、机器学习、推荐系统等多领域的研究人员和工程人员, 反映了我国学者在大数据治理的理论与技术领域最新的研究进展. 感谢《软件学报》编委会和数据库专委会对专题工作的指导和帮助, 感谢专题全体评审专家及时、耐心、细致的评审工作, 感谢踊跃投稿的所有作者. 希望专题能够对大数据治理的理论与技术相关领域的研究工作有所促进.

附专题评审专家名单(按姓氏拼音首字母排序):

柴成亮 崔斌 陈璐 成雨蓉 丁琳琳 范举 冯恺宇 高军 谷峪 何震瀛 洪亮 李川 李佳佳 李艳辉 刘海龙 罗浩 宁博 潘巍 彭煜玮 秦建斌 任飞亮 邵鋆侠 唐博 童咏昕 魏哲巍 许嘉 徐建良 许建秋 杨晓春 游进国 袁冠 袁野 岳昆 张东祥 张昕 张小旺 张岩峰 张志威 赵翔 郑凯 郑臻哲 朱睿 朱扬勇 祝园园 邹兆年

杜小勇(1963－), 男, 博士, 中国人民大学教授, 博士生导师, 教育部数据工程与知识工程重点实验室主任, CCF 会士, CCF 大数据专委会主任委员. 主要研究领域为智能信息检索, 高性能数据库, 非结构化数据管理. 曾主持国家重点研发计划等项目 20 余项, 获得国家级与省部级科技奖励多项.



杨晓春(1973－), 女, 博士, 东北大学教授, 博士生导师, CCF 杰出会员, 入选国家级人才. 主要研究领域为大数据管理与知识工程, 大数据治理与质量管理, 数据隐私保护, 智能推荐等. 曾主持国家重点研发计划、自然科学基金优秀青年科学基金和重点项目等 20 余项, 获得国际会议最佳论文奖 4 项、全国会议最佳论文奖 5 项, 获得省部级奖励多项.

童咏昕(1982－), 男, 博士, 北京航空航天大学教授, 博士生导师, CCF 杰出会员, 国家优秀青年科学基金获得者. 主要研究领域为大数据, 数据库, 联邦学习, 隐私计算与群体智能等. 曾主持国家自然科学基金重点项目与国家重点研发计划课题等科研项目 10 余项, 获得中国电子学会自然科学一等奖等科技奖励多项.