编者按

近年来,随着信息化技术与互联网的飞速发展,各行各业积累的数据资源急剧增长,大数据已成为国家层面的基础性战略资源,引起各国政府高度重视.例如,国务院于2015年8月印发了《促进大数据发展行动纲要》,美国政府于2016年5月发布了《联邦大数据研究与开发战略计划》,欧盟 Horizon 2020 计划把大数据提高到国家安全战略层面.发展大数据分析理论与技术方法,是实现我国从数据大国向数据强国转变的重要基础,将对国计民生的诸多方面产生深远影响.

大数据分析旨在从海量数据中提炼知识价值,涉及机器学习、可视分析、数据挖掘、知识处理等多个研究领域,而大数据的高维、多源、不确定、领域化等特性给现有分析体系带来严峻挑战.为及时反映国内大数据分析方面的最新发展,《中国科学: 信息科学》特组织本期“大数据分析专题”,邀请国家重点研发计划项目“大数据分析的基础理论和技术方法”团队成员介绍相关研究进展,并邀请国内同行专家评审,最终有7篇论文入选本期专题.

在大数据分析方面,机器学习与可视分析是实现原始数据至知识信息有效提炼的基础支撑,本期专题中有三篇文章是关于这方面的内容:

“元自步学习”(束俊, 孟德宇*, 徐宗本)针对现有自步学习算法存在的超参数选择问题,提出了一种基于元学习机理的自步学习算法.该算法基于元数据的指导,在同一优化框架中基于数据实现参数动态调节,并将所提元学习策略嵌入三种典型的自步学习实现.

“基于众包学习的交互式特征选择方法”(陈长建, 姜流, 雷娜, 刘世霞*)针对现有集成特征选择算法未考虑不同特征选择方法性能差异性的问题,提出了一种基于可视分析的交互式特征选择方法,通过可视分析系统为基于众包学习的集成特征选择算法提供丰富的排序方式,从而交互迭代地改善现有特征子集.

“一种保持语义关系的词向量复用方法”(李新春, 詹德川*)针对词向量复用面临的新词出现问题,提出了一种简单有效的语义关系保持词向量复用方法.该方法利用当前语料库构建词语之间的语义关系映射,并在预训练词向量基础上施加该映射获得新词的词向量.

数据挖掘与知识处理是实现领域知识至大数据核心价值有效转化的核心途径,本期专题中有两篇文章是关于这方面的内容:

“基于邻域视角的关联关系挖掘方法”(成红红, 钱宇华*, 胡治国, 梁吉业)对大数据背景下海量变量间的关联关系挖掘进行研究,尝试给出了大数据关联关系度量需满足的性质.在剖析两类传统的基于邻域视角的关联关系度量方法基础上,通过引入 k-NN 粒与多粒度融合提出了基于数据驱动的关联关系度量方法.

“一种高可读低冗余实体摘要的生成方法”(刘庆霞, 程龚*, 瞿裕忠)针对知识图谱的实体摘要问题进行研究,提出了一种具备高可读性和低冗余性的实体摘要生成方法.该方法通过综合考虑实体描述三元组的重要性,可读性以及摘要的冗余性,将实体摘要问题建模为二次背包问题进行求解.

引用格式: 周志华, 张敏灵. 大数据分析专题编者按. 中国科学: 信息科学, 2020, 50: 779-780, doi: 10.1360/SSI-2020-0155

数据驱动是大数据分析应用的基本特征, 面向视频、文本等大数据应用需求的研究具有重要意义, 本期专题中有两篇文章是关于这方面的内容:

“基于视觉 - 文本关系对齐的跨模态视频片段检索”(陈卓, 杜昊, 吴雨菲, 徐童 *, 陈恩红) 针对日益增长的视频片段精细检索需求, 对视觉 - 文本跨模态视频片段检索问题进行研究. 通过构建文本关系图与视觉关系图刻画查询文本与视频片段的语义关系, 并通过跨模态对齐的图卷积网络评估文本关系与视觉关系的相似度.

“基于长短时预测一致性的大规模视频语义识别算法”(王铮, 翁泽佳, 王锐, 陈静静, 姜育刚 *) 针对大规模视频语义识别面临的视频数量巨大且缺乏网络标签的问题, 提出了一种基于长短时预测一致性的视频语义识别算法. 在片段视频语义识别任务上, 文中算法性能在第三届 YouTube-8M 比赛中排名第二.

感谢《中国科学: 信息科学》编委会和编辑部所给予的指导和帮助, 感谢国家重点研发计划项目“大数据分析的基础理论和技术方法”(2018YFB1004300) 项目组的支持, 感谢审稿专家认真仔细的评审工作, 祝愿大数据分析领域优秀成果不断涌现.

特约编辑: 周志华 南京大学
张敏灵 东南大学