软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(9):1887−1888 [doi: 10.13328/j.cnki.jos.004652]

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

大数据分析专刊前言*

陈恩红[1], 于 剑[2]

[1](中国科学技术大学 计算机学院,安徽 合肥 230027)

[2](交通数据分析与挖掘北京市重点实验室(北京交通大学),北京 100044)

通讯作者: 于剑, E-mail: jianyu@bjtu.edu.cn

中文引用格式: 陈恩红,于剑.大数据分析专刊前言.软件学报,2014,25(9):1887−1888. http://www.jos.org.cn/1000-9825/4652.htm

自2008年《Nature》杂志发表大数据专辑以来,大数据得到越来越多的关注.2012年,美国和中国分别将大数据提升到国家战略高度.大数据技术是一个典型的跨领域研究方向,在数据的采集、存储、传输、管理、安全和分析等诸多方面均面临着挑战.在大数据分析方面,我国已经有国家自然科学基金重点项目、国家重点基础研究发展计划(973)在内的多个立项支持,并在学术界和工业界取得了一些有影响的研究与应用成果.然而,作为一个新兴的研究方向,大数据分析依然面临诸多挑战.本专刊收录的21篇论文反映了我国学者在大数据分析领域的部分近期研究成果.

本专刊收录了 3 篇综述性论文.

《大数据系统和分析技术综述》是特邀论文,分析了各种大数据处理系统以及相应的大数据分析技术,特别是批量数据处理系统、流量数据处理系统、交互式数据处理系统、图数据处理系统中的大数据分析技术,以及典型应用场景,总结出大数据处理系统的三大发展趋势:数据处理引擎专用化;数据处理平台多样化;数据计算实时化.随后,对系统支撑下的大数据分析技术和应用,包括深度学习、知识计算、社会计算与可视化等,进行了简要综述,总结了各种技术在大数据分析理解过程中的关键作用;最后梳理了大数据处理和分析面临的数据复杂性、计算复杂性和系统复杂性三个挑战,并逐一提出可能的应对之策.

《大数据可视分析综述》从可视分析领域所强调的认知、可视化、人机交互的综合视角出发,分析了支持大数据可视分析的基础理论,包括支持分析过程的认知理论、信息可视化理论、人机交互与用户界面理论.讨论了面向大数据主流应用的文本、网络(图)、时空、多维的可视化技术.同时,探讨了支持可视分析的人机交互技术,包括支持可视分析过程的界面隐喻与交互组件、多尺度/多焦点/多侧面交互技术、面向 Post-WIMP 的自然交互技术.最后,指出了大数据可视分析领域面临的瓶颈问题和技术挑战.

《图数据表示与压缩技术综述》分析了图数据压缩技术的研究现状,并将其分为 4 类:基于传统存储结构的压缩技术、网页图压缩技术、社交网络图压缩技术、面向特定查询的图压缩技术,详细分析了各类代表方法并比较其性能差异.最后,对未来大规模图数据压缩技术进行展望.

本专刊收录的另外 18 篇研究性论文主要涉及大数据环境下的个性化推荐、多标记学习、软件分类、相关性分析、海量信息融合、谱聚类、命名实体链接、人脸识别等方面的研究工作.

《基于大规模隐式反馈的个性化推荐》提出了潜在要素模型 IFRM 和并行化的隐式反馈推荐模型 p-IFRM,能够有效提高推荐质量并有良好的可扩展性.

《多标记分类和标记相关性的联合学习》提出了多标记分类和标记相关性的联合学习 JMLLC,从而增强了各个标记分类器的学习效果.

《基于代价敏感多标记学习的开源软件分类》将开源软件自动标注形式化为一个代价敏感多标记学习问题,并提出了一种代价敏感多标记学习方法 ML-CKNN.

* 收稿时间: 2014-07-10

《基于类属属性的多标记学习算法》验证了类属属性对多标记学习系统性能的影响,以及 LIFT 所采用的类属属性构造方法的有效性.

《时序数据曲线排齐的相关性分析方法》研究时序数据的相关性和协同性,提出了双序列的相关性判定方法和曲线排齐方法.

《有序判别典型相关分析》将有序类信息嵌入 CCA 进行扩展,提出了有序判别典型相关分析 OR-DisCCA.

《海量信息融合方法及其在状态评价中的应用》针对证据理论无法有效处理海量信息的融合问题,给出了一种结合聚类和凸函数证据理论的海量信息融合方法.

《基于自适应 Nyström 采样的大数据谱聚类算法》设计了自适应的 Nyström 采样方法,并提出一种适用于大数据的谱聚类算法.

《基于统计相关性与 *K*-means 的区分基因子集选择算法》针对高维小样本癌症基因数据集的有效区分基因子集选择难题,提出了基于统计相关性和 *K*-means 的混合基因选择算法.

《一种基于概率主题模型的命名实体链接方法》提出了语义层面对文档进行建模和实体消歧的思想,设计了一种基于概率主题模型的命名实体链接方法.

《关系抽取中基于本体的远监督样本扩充》提出了基于本体的远监督学习样本扩充方法,能够有效完成样本匮乏的关系抽取任务.

《基于图像分解的人脸特征表示》提出了基于图像分解的人脸特征表示方法 FRID.

《大数据环境下多决策表的区间值全局近似约简》针对电力大数据分类问题的特点,提出了基于依赖度和互信息的区间值约简算法,并针对大数据的分布式存储,提出了信息论下的区间值全局近似约简概念和方法.

《大数据下基于异步累积更新的高效 P-Rank 计算方法》将异步累积更新算法应用在了 P-Rank,能够有效提高计算收敛速度.

《支持向量学习的多参数同时调节》简化了传统模型选择的双层优化框架,提出了一种支持向量学习的多参数同时调节方法.

《一种求解强凸优化问题的最优随机算法》提出了一种能保证稀疏性的基于 COMID 的加权算法.

《基于空时极向 LBP 的极光序列事件检测》提出了用于检测全天空图像(ASI)序列中的弧状极光事件检测方法.

《基于学习的高分辨率掌纹细节点质量评价方法》提出了一种基于学习的高分辨率掌纹细节点质量评价方法,能够更好地区分真伪细节点,对细节点的质量做出更好的评价.

本专刊主要面向数据挖掘、机器学习、人工智能、模式识别、生物特征识别相关领域的研究人员.审稿过程历经 5 个月,有 30 余名相关领域的专家和学者参与审稿工作.全部预录用论文在中国数据挖掘会议(CCDM2014,金华)上交流.经过网审、会审等程序,最终收录以上 21 篇论文.在此,我们感谢踊跃投稿的相关领域学者,感谢辛勤工作的审稿专家和《软件学报》编辑部.

陈恩红(1968－),男,博士,教授,博士生导师,国家杰出青年基金获得者,IEEE 高级会员,CCF 理事.主要研究领域为机器学习,数据挖掘,社会网络,个性化推荐系统.
E-mail: cheneh@ustc.edu.cn

于剑(1969－),男,博士,教授,博士生导师,IEEE 会员,CCF 理事,主要研究领域为机器学习,计算智能,图像分析,数据挖掘.
E-mail: jianyu@bjtu.edu.cn