软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(3) [doi: 10.13328/j.cnki.jos.005171]

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

大数据管理技术专刊前言*

马 帅[1,2], 崔 斌[3]

[1](软件开发环境国家重点实验室(北京航空航天大学),北京 100191)

[2](大数据与脑机智能高精尖创新中心(北京航空航天大学),北京 100191)

[3](北京大学 信息科学技术学院,北京 100871)

通讯作者: 马帅, E-mail: mashuai@buaa.edu.cn

中文引用格式: 马帅,崔斌.大数据管理技术专刊前言.软件学报,2017,28(3). http://www. jos.org.cn/1000-9825/5171.htm

大数据管理及处理能力已经成为引领网络时代 IT 发展的关键;获取大量运行数据并建立对其进行动态高效处理的能力,已经成为产业竞争力的体现.从 2012 年美国政府宣布“大数据研究计划(Big Data Initiative)”和我国发布的《“十二五”国家科技计划信息技术领域项目》将大数据研究列在首位以来,大数据管理分析技术得到了学术界和工业界空前的关注,并且 2016 年 5 月美国又进一步发布了《联邦大数据研究与开发战略计划》.然而,如何对大数据进行有效的管理分析仍然面临巨大的挑战.本专刊选题为“大数据管理技术”,将突出目前大数据的整个生命周期的管理研究中的热点技术,包括大数据系统和应用等诸多方面的研究.

专刊公开征文,共征得投稿 25 篇(其中包括 NDBC 2016 推荐的 5 篇高质量论文).这 25 篇文章通过特约编辑形式审查,有 23 篇论文进入到评审阶段.上述稿件研究内容涉及大数据系统和应用的方方面面,特约编辑先后邀请了 24 位数据库和大数据等相关领域的专家参与审稿工作,每篇投稿邀请 2 位专家进行评审.稿件评审历经 5 个月,经初审、复审、NDBC 2016 会议宣读和终审四个阶段,最终有 17 篇论文入选本专刊.

(一) 大数据分析处理系统和平台及其关键技术是大数据管理技术的研究重点之一.

《集群数据库系统的日志复制和故障恢复》观察到基于集群环境的数据库系统中,传统主备架构的日志复制在异常情况下对未决事务日志处理不佳,数据副本存在不一致的问题;分布式系统领域的一致性算法缺乏对事务一致性的处理,在选主时存在活锁、多主和频繁选主的问题,无法直接适用于事务日志复制,因此提出了一种集群环境下的事务日志复制策略和恢复机制,提供强弱两种读一致性,并且提出了一种轻量级的选主算法.

《一种基于众核架构 Phi 协处理器的内存 OLAP 外键连接算法》关注缓存相关的分区哈希连接和缓存不相关的无分区哈希连接的缓存友好型外键连接算法,以适应新兴主流高性能计算平台的 Xeon Phi 协处理器较小的 LLC 和高并发线程的特点,提出了将复杂的哈希表和 CPU 代价较高的哈希探测操作简化为通过映射外键值为代理键向量内存偏移地址的方法对代理向量直接访问的方法.

《一种面向 HDFS 的多层索引技术》观察到 SOH 系统在处理选择型查询或交互式查询时的性能缺陷,分析了 SOH 系统访问 HDFS 文件的过程,得出了其中影响数据加载时间的关键因素,从而提出一个通用的索引技术,以提高 SOH 系统查询处理的效率.

《MapReduce 大数据处理平台与算法研究进展》介绍了近年来基于 MapReduce 编程模型的大数据处理平台与算法,将大数据处理算法抽象为外存算法,并对外存算法的特征加以梳理,提出了普适的外存算法性能优化方法的研究思路和研究问题.

(二)分布式与云环境中大数据分发与任务调度是对大数据管理技术的有效支撑.

《面向多源大数据云端处理的成本最小化方法》观察到由于数据的动态性以及资源价格的波动性,将数据

* 收稿时间: 2016-11-27; jos 在线出版时间: 2016-11-29
CNKI 网络优先出版: 2016-11-29 13:35:11, http://www.cnki.net/kcms/detail/11.2560.TP.20161129.1335.013.html

迁移至哪些数据中心并提供合适的计算资源来处理成为处理多源数据的一大问题,提出了将该问题转换成联合随机优化问题,利用李雅普诺夫优化框架分解成两个独立的子问题进行求解的在线算法.

《分布式数据流上的高性能分发策略》观察到数据的倾斜分布及数据流本身实时、动态和数据规模不可预知等特性使得数据流分布并行处理系统存在持续且动态的负载不均衡现象,综合基于key的迁移和基于元组拆分的随机分发两种方法,提出对key按需拆分和尽量合并的方法.

《一种云环境中数据流的高效多目标调度方法》优化云数据流中的调度过程,将优化目标分别划分为用户指标和云系统指标,并将调度问题制定成为一个连续的合作博弈,提出一种快速收敛的高效 Muliti-Objective Game 调度算法,在优化用户指标的同时,实现系统指标的约束,并保证云资源的效率和公平度.

《基于最小费用最大流的大规模资源调度方法》观察到采用队列进行资源调度建模,仅能满足局部最优解,只能适应调度目标固定不变的场景,从而提出了一种基于最小费用最大流的大规模资源调度建模方法,将任务的资源需求和物理资源供给问题转换成最小费用最大流图的构造和求解问题.

《空间众包环境下的三类对象在线任务分配》观察到已有空间众包研究的假设过强:通常假设基于静态场景和均假设仅有两类众包参与对象,忽略了第三方众包工作地点对任务分配的影响,因此提出了一类新型空间众包环境下的三类对象在线任务分配问题,并采用在线学习方法进一步优化了随机阈值算法,设计自适应随机阈值算法,并证明该优化策略可逼近随机阈值算法使用不同阈值所能达到的最佳效果.

(三)网络大数据尤其是社会网络数据的分析与应用得到了广泛关注.

《复杂网络大数据中重叠社区检测算法》基于模块度聚类和图计算思想，应用新的节点和边的更新方法,利用平衡二叉树对模块度增量建立索引,基于模块度最优的思想设计一种新的重叠社区检测算法.相对于传统重叠节点检测算法,对每个节点分析的频率大大降低.

《基于深度稀疏自动编码器的社区发现算法》观察到在处理网络的高维矩阵时使用这些经典聚类方法得到的社区往往不够准确,因此提出一种基于深度稀疏自动编码器的社区发现算法CoDDA,从而提高使用经典方法处理高维邻接矩阵进行社区发现的准确性.

《动态信息网络中基于角色的结构演化与预测》使用“角色”来量化动态网络的结构,得到动态网络的角色模型,应用并改进多类标分类问题的“问题转换”思想,将动态网络的角色预测问题视为多目标回归问题,以历史网络数据作为训练数据构建模型,预测未来时刻网络可能的角色分布情况,提出基于多目标回归思想的动态网络角色预测方法.

《基于语义约束 LDA 的商品特征和情感词提取》根据中文商品评论文本的特点,从句法分析、词义理解和语境相关等多角度获取词语间的语义关系,然后将其作为约束知识嵌入到主题模型,提出语义关系约束的主题模型SRC-LDA,用来实现语义指导下LDA的细粒度主题词提取.

《基于社交关系的微博主题情感挖掘》考虑微博用户相互关联的事实,提出基于主题模型 LDA 和微博用户关系的主题情感模型,在主题模型中加入情感层与微博用户关系参数,利用微博用户关系与微博主题学习微博的情感极性.

《融合主题模型和协同过滤的多样化移动应用推荐》观察到已有推荐方法多注重推荐准确率,忽视多样性,改进了两个已有推荐方法,提出了将用户的主题模型和应用的主题模型与MF相结合的LDA_MF模型,以及应用的标签信息和用户行为数据同时考虑 LDA_CF 算法,并提出了融合 LDA_MF、LDA_CF 以及经典的基于物品的协同过滤模型的混合推荐算法.

《社交网络环境下基于信任的推荐算法》提出一种基于信任的推荐算法,该方法结合全局和局部信任,利用信任的传播性质对信任关系进行建模,综合相似度和信任度来构建用户间的偏好关系;将基于记忆的协同过滤思想和社交网络的信任关系融入概率矩阵分解模型,同时使用自适应权重动态决定各部分的影响程度,形成高效统一的可信推荐模型 Trust-PMF.

《透析计算:面向 OLGP 的 InfoNetCube 高效物化》基于“透析计算”思想的信息网络立方物化策略,通过主题图度量在信息维和拓扑维上反单调性运用,提出基于“透析计算”的空间剪枝算法,快速透析掉不可能命

中的子图度量、方体单元、方体乃至方体格.

本专刊主要面向数据库、数据挖掘和人工智能等相关领域的从事大数据的研究人员,反映了我国学者在大数据管理技术最新的研究进展.在此,我们要特别感谢《软件学报》编委会和数据库专委会对专刊工作的指导和帮助,感谢编辑部和数据库专委会各位老师从征稿启示发布、审稿专家邀请至评审意见汇总、论文定稿、修改及出版所付出的辛勤工作和汗水,感谢专刊评审专家及时、耐心、细致的评审工作.此外,我们还要感谢向本专刊踊跃投稿的作者对《软件学报》的信任.

最后,感谢专刊的评审专家、编辑和读者们,希望本专刊能够对大数据相关领域的研究工作有所促进.

马帅(1975－),男,山东潍坊人,博士,教授,博士生导师.现任中国计算机学会数据库专业委员会常务委员,大数据专家委员会委员.曾获 VLDB 最佳论文奖(2010 年),国家自然科学基金优秀青年科学基金(2013 年).主要研究领域为数据库系统和理论,大数据.

崔斌(1975－),男,浙江宁波人,博士,北京大学信息学院教授、博士生导师.现任中国计算机学会数据库专业委员会秘书长,担任 VLDBJ、TKDE、Information Systems 和软件学报等多个学术期刊编委.主要研究领域为数据库系统、数据管理和分析技术.