软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(3):525−527 [doi: 10.13328/j.cnki.jos.005458]

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

基于图结构的大数据分析与管理技术专刊前言*

林学民[1,2], 杜小勇[3,4], 李翠平[3,4]

[1](The University of New South Wales, Sydney 1466)

[2](华东师范大学 软件学院,上海 200062)

[3](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)

[4](中国人民大学 信息学院,北京 100872)

通讯作者: 李翠平, E-mail: licuiping@ruc.edu.cn

中文引用格式: 林学民,杜小勇,李翠平.基于图结构的大数据分析与管理技术专刊前言.软件学报,2018,29(3):525−527. http://www.jos.org.cn/1000-9825/5458.htm

作为一种常用的描述事物之间复杂关联关系的普适结构,图结构被广泛用于表示万维网、社交网络、蛋白质交互网络、化学分子结构、软件代码剽窃检测、复杂对象识别、公路网络、模式识别、超大规模集成电路设计和生态系统中的食物链等.图结构在大数据价值发现中也发挥着重要的作用,如何对基于图结构的大数据进行有效的分析和管理,已成为学术界和工业界关注的新的热点.研究人员提出了很多新颖的图算法,如图生成器、图的可达性判定、相似子图查询、图的特性分析、图查询语言、图上的关键字查询、社交网络分析等,也出现了一些开源的图数据库系统.但总体而言,针对基于图结构的大数据的研究仍然处于起步阶段,还有很多需要研究的问题.

本专刊公开征文,共收到投稿 53 篇(包括第 34 届中国数据库学术会议(NDBC 2017)推荐的 22 篇高质量论文).其中 41 篇论文通过了形式审查,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 80 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 2017 会议宣读和终审 4 个阶段,历时 5 个月,最终有 23 篇论文入选本专刊.根据主题这些论文可以分为 4 组.

(一) 基于图结构的大数据并行计算模型、框架与系统.

《流式处理的异步图处理框架》结合累加迭代计算和单机并行处理技术,提出流式处理的异步计算模型 ASP.

《路径-维度 GraphOLAP 大规模多维网络并行分析框架》设计并提出了一种图立方体模型:路径-维度立方体,并提出了立方体的物化策略和基于 Spark 框架的相关算法.

《分布式图处理系统技术综述》总结了分布式图处理系统的 3 个优化目标,从计算粒度、任务调度、通信方式、负载划分这 4 个维度,对现有分布式图处理系统中的各类优化技术进行详细的综述.

《Coteries 轨迹模式挖掘及个性化旅游路线推荐》提出基于语义的距离敏感推荐策略(DRSS)和基于语义的从众性推荐策略(CRSS).

《基于距离度量的多样性图排序方法》提出一种描述节点间不相似性的距离度量,将多样性图排序问题建模为一个带权完全图的最大和 k-dispersion 优化问题.

《基于端到端分布式框架的符号网络预测方法》提出一种端到端的分布式框架,并将其应用于分布式环境下符号网络中的链接关系的正负预测问题.

《基于 MapReduce 的图结构聚类算法》关注图结构聚类(SCAN)算法的可扩展性问题,提出了基于 MapReduce 的海量图结构聚类算法 MRSCAN.

收稿时间: 2017-12-14

(二) 基于图结构的大数据索引和查询技术.

《路网环境下的最近邻查询技术》对路网环境下的最近邻查询技术进行综述,分别从最近邻查询采用的索引结构和查询处理过程对现有路网环境下的最近邻查询方法进行了分析和比较.

《动态图模式匹配技术综述》关注动态更新的图数据中进行高效的查询、匹配问题.从关键技术、代表性算法和性能评价方面对动态图匹配技术进行了综述.

《基于 SQL 的图相似性查询方法》研究基于编辑距离的图相似性查询处理问题.针对已有方法在过滤阶段自身存在优缺点和适用性的问题,提出一种面向关系型数据库的过滤框架.

《路网环境下兴趣点查询的隐私保护方法》针对在路网环境下,用户查询过程中位置隐私泄露的问题,提出了一种位置 k 匿名隐私保护方法,克服了传统 k-匿名不能抵御推断攻击的缺陷.

《基于疾病信息网络的表型相似基因搜索》利用疾病公开数据库构建了疾病信息网络并设计了基于此的相似基因搜索算法 gSim-Miner.

《路网感知的在线轨迹压缩方法》提出了一种路网感知的在线轨迹压缩方法,在综合考虑移动轨迹的特点和地图质量的基础上,针对轨迹压缩的需要,设计了一种距离有界的地图匹配算法.

(三) 基于图结构的大数据分析和挖掘技术以及深度学习方法.

《基于边采样的网络表示学习模型》提出一种能够编码节点间丰富关系信息的无监督网络表示学习模型 NEES.

《社交网络高效高精度去匿名化算法》提出了一种高效高精度的无种子去匿名化算法 RoleMatch,基于社交网络的拓扑结构识别个体身份.

《一种融合节点先验信息的图表示学习方法》提出了一种改进的图表示学习方法 GeVI.该方法将已知的节点特征看做先验知识,并基于 DeepWalk 思想,将图表示学习问题转化为词表示学习问题.

《基于循环神经网络的数据库查询开销预测》提出了一种基于循环神经网络的查询开销预测方法,该方法不仅能够预测出查询计划的执行时间,而且在查询执行前就能得到预测结果.

《全视角特征结合众包的跨社交网络用户识别》关注识别出不同社交网络上的同一用户,提出了基于全视角特征结合众包的跨社交网络用户识别方法(OCSA).

《基于树分解的空间众包最优任务分配算法》研究空间众包中最优任务分配问题,利用树分解技术将工人分割成独立的集合,并提出一种带启发式的深度优先搜索算法.

《多维图结构聚类的社交关系挖掘算法》提出了一种有效的子空间聚类算法 SCA,首次对多维度下子空间的图结构聚类进行研究,目的是探索如何通过图数据挖掘发现对象间真实的社交关系.

《基于社区的动态网络节点介数中心度更新算法》针对动态网络中节点介数中心度计算困难的问题,提出一种基于社区的节点介数中心度更新算法.

(四) 新型硬件下的图数据管理技术.

《应对倾斜数据流在线连接方法》基于二部图连接模型,提出应对倾斜数据流的在线连接方法.

《基于向量引用 Platform-Oblivious 内存连接优化技术》通过优化内存哈希表设计,消除哈希代价对内存连接算法的影响,从而更加准确地测量内存连接算法在不同硬件相关因素影响下的性能特征.

本专刊主要面向大数据、数据库、数据挖掘、机器学习、体系结构等多领域的研究人员和工程人员,反映了我国学者在基于图结构的大数据分析与管理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刊工作的指导和帮助,感谢专刊全体评审专家及时、耐心、细致的评审工作,感谢踊跃投稿的所有作者.希望本专刊能够对大数据相关领域的研究工作有所促进.

林学民(1961－),男,IEEE Fellow,新南威尔士大学杰出教授(Scientia Professor),中国国家“千人计划”学者.《IEEE Transactions on Knowledge and Data Engineering》主编.主要研究领域为数据库理论、算法与技术研究,时空数据和流数据的查询,图和文本的匹配查询和计算,不确定数据的概化查询等.在本领域国际学术会议与期刊上发表论文 270 余篇,其中,140 余篇发表在顶级会议和期刊上,16 篇优秀会议论文.

杜小勇(1963－),男,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,数据库专委会主任,《大数据》副主编,国家重点研发专项“云计算和大数据”专家组成员.长期从事数据库与大数据方面的教学与研究工作,先后承担核高基、973 等多项国家级课题,在本领域国际重要期刊和会议上发表高水平学术论文 100 余篇.

李翠平(1971－),女,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为数据仓库和数据挖掘,社会网络分析.