软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(4):691−692 [doi: 10.13328/j.cnki.jos.004572]

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

大数据专题前言*

孟小峰[1], 高 宏[2]

[1](中国人民大学 信息学院,北京 100872)

[2](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

通讯作者: 孟小峰, E-mail: xfmeng@ruc.edu.cn

中文引用格式: 孟小峰,高宏.大数据专题前言.软件学报,2014,25(4):691−692. http://www.jos.org.cn/1000-9825/4572.htm

随着信息技术的不断发展,以及云计算、物联网、社交网络等新兴技术和服务的不断涌现和广泛应用,数据种类日益增多,数据规模急剧增长,大数据时代悄然来临.由于大数据对政府决策、商业规划和知识发现等所起的重大作用,大数据逐渐成为一种重要的战略性资源,受到政府、工业界及学术界的普遍关注.大数据的多样性(variety)、规模性(volume)和高速性(velocity)等特点,使得传统的数据存储、管理以及数据分析技术已经无法满足大数据的处理要求.

为了实现对大数据的高效存储管理和快速分析,云计算、内存计算、流计算等新兴技术不断涌现;同时,为了实现对不同类型数据的有效管理,产生了文档数据库、图数据库、列存储、键值数据库等不同的数据管理方法;同时,自然科学、社会科学等不同学科的研究人员开始探讨本领域在大数据场景下所面临的挑战和机遇,并逐步尝试利用大数据思维将不同学科进行交叉、不同领域的数据进行集成管理和分析,以期得到新的重大发现.

为了反映大数据在不同学科和领域的研究现状及最新研究成果,展示大数据面临的理论和技术上的新挑战,揭示大数据的研究热点及研究方向,《软件学报》和我们共同策划和组织了“大数据专题”.本期专题通过公开征文收到 50 余篇投稿,论文分别在多个方面阐述了大数据领域具有重要意义的研究成果.本专题的审稿严格按照期刊审稿要求进行,特约编辑先后邀请了 30 余位相关领域的专家参与评审,每篇论文邀请至少 2~3 位专家进行评审,历经初审、复审、终审等阶段,整个流程历经半年,最终从中遴选出 9 篇高质量的论文入选本专题.这 9 篇论文分别涵盖位置大数据、社交网络分析、大数据 OLAP 技术、Top-K 查询、大数据降维等研究内容,在一定程度上反映了当前国内各研究单位在大数据领域的主要研究方向.

大数据的表现形式多样,位置大数据作为其中重要的一类得到了广泛的关注.《位置大数据隐私保护研究综述》介绍了位置大数据的概念以及位置大数据的隐私威胁,对目前位置大数据隐私保护领域已有的研究成果进行了归纳.文章对各类位置隐私保护技术的基本原理、优缺点进行了对比、分析,并对目前该领域的前沿问题——基于隐私信息检索的位置隐私保护技术进行了重点阐述.最后,作者结合自身对位置大数据隐私保护技术的研究和理解提出了若干研究方向,如位置大数据与非位置大数据相结合的隐私保护、移动社交网络中的位置隐私保护和针对用户背景知识的位置大数据隐私保护.

为了获得更为准确的移动行为模式和区域局部特征,从而还原和生成满足关联应用分析的整体数据模型,《位置大数据的价值提取与协同挖掘方法》针对位置大数据存在的混杂性、复杂性和稀疏性等特点,分别提出了相应的处理方法,并提出了针对位置大数据的价值提取和协同挖掘方案,文章还从软件工程角度提出了位置大数据分析的整体框架.

对于多维信息的共享需求产生了 OLAP 技术,大数据时代,数据分析的实时性等要求使得传统的 OLAP 技术面临着严峻的挑战.《大数据分析的分布式 MOLAP 技术》提出了大数据环境中一种基于 Hadoop 分布式文件系统(HDFS)和 MapReduce 编程模型的分布式 MOLAP 技术,称为 DOLAP(distributed OLAP).实验结果表明,尽管数据装

* 收稿时间: 2014-01-28

载性能略显不足,但DOLAP的性能要优于基于HBase,Hive,HadoopDB,OLAP4Cloud等主流非关系数据库系统实现的 OLAP 性能.《一个基于三元组存储的列式 OLAP 查询执行引擎》则研究了新硬件平台下针对大规模数据的OLAP 查询的性能,设计了新的列存储 OLAP 查询执行引擎,基于三元组的物化策略有效地减少了内存列存储模型上表连接操作访问基表和中间数据结构的次数.

社交网络作为大数据的重要承载体,众多学者从不同方面对其展开了全面的研究.《基于节点分割的社交网络属性隐私保护》针对当前社交网络隐私属性匿名算法中存在的缺乏合理模型、属性分布特征扰动大、忽视社交结构和非敏感属性对敏感属性分布的影响等弱点,提出一种基于节点分割的隐私属性匿名算法.此外,文章量化了社交结构信息对属性分布的影响,根据属性相关程度进行节点的属性分割,能够很好地保持属性分布特征,保证数据可用性.《社会化媒体大数据多阶段整群抽样方法》针对传统的社交媒体数据抽样方法存在大型马尔可夫链难以并行化、样本局部性陷入、马尔可夫链燃烧预热等问题,提出了在线社会化媒体大数据整群多阶段抽样方法OSM-MSCS.该方法首先进行整群分解,将总体分解成若干小型凝聚子群,而后使用延迟拒绝(DR)方法,以并行化方式进行子群内部关系的抽样,最后使用 Gibbs 方法完成不同子群之间相干关系的筛选,从而获得整个样本序列.《基于图压缩的 k 可达查询处理》研究了基于图压缩的 k 可达查询处理,提出了一种支持 k 可达查询的图压缩算法 k-RPC 及无需解压缩的查询处理算法,k-RPC 算法在所有基于等价类的支持 k-reach 查询的图压缩算法中是最优的.由于 k-RPC 算法是基于严格的等价关系,进一步提出了线性时间的近似图压缩算法 k-GRPC.

Top-K 查询在搜索引擎、电子商务等领域有着广泛的应用.《一种云环境下的大数据 Top-K 查询方法》结合MapReduce 的特点,从数据划分、数据筛选等方面对云环境下的大数据 Top-K 查询问题进行了深入研究.实验结果表明,该方法具有良好的性能和扩展性.

《基于边界判别投影的数据降维》提出了一种新的有监督的线性降维算法——边界判别投影,即最小化同类样本间的最大距离,最大化异类样本间的最小距离,同时保持数据流形的几何形状.与经典的基于边界定义的算法相比,边界判别投影可以较好地保持数据流形的几何结构和判别结构等全局性.实验结果表明,该方法是一种有效的降维算法,可应用于大数据上的特征提取.

承蒙各位作者、审稿专家和编辑部等方面的全力支持,本专题得以顺利出版.目前大数据研究涉及领域十分广泛,这给审稿人及特约编辑的审稿、选稿带来巨大挑战.由于投稿数量大、主题广泛、时间安排紧张、专题容量有限等原因,本专题仅选择了部分有代表性的研究工作予以发表,无法全面体现大数据领域所有的最新研究工作.部分优秀稿件无法列入发表,敬请谅解.

我们要特别感谢《软件学报》编委会和编辑部,从专题的立项到征稿启示的发布,从审稿专家的邀请到评审意见的汇总,以及最后的定稿、修改、出版,他们都付出了辛勤的汗水.本专题的出版期望能给广大研究人员带来启发和帮助.在审稿过程中难免出现不尽如人意之处,希望各位作者和读者包容和谅解,希望同行不吝批评指正.最后,衷心感谢各位作者、审稿专家和编辑部的辛勤工作.

孟小峰(1964－),男,博士,教授,博士生导师,现为 CCF 会士,常务理事,中国计算机学会数据库专委秘书长.主要研究领域为 Web 数据管理,移动数据管理,XML 数据管理,云数据管理,新型存储数据库.先后获得电子部科技进步特等奖(1996 年),北京市科技进步二等奖(1998 年,2001 年),中国计算机学会"王选奖"一等奖(2009 年),北京市科学技术奖二等奖(2011 年),"第三届北京市高校名师奖"(2005 年).发表论文 200 多篇,获国家发明专利授权 13 项.
E-mail: xfmeng@ruc.edu.cn

高宏(1966－),女,博士,教授,博士生导师,IEEE 高级会员,CCF 高级会员,中国计算机学会传感器网络专业委员会委员,中国计算机学会数据库专业委员会副主任.主要研究领域为基于并行与压缩的海量数据计算,无线传感网数据管理,复杂图数据管理与计算,数据质量等.发表学术论文 200 余篇,获国家科技进步二等奖 1 项,省自然科学一等奖 1 项.
E-mail: honggao@hit.edu.cn