《海量数据处理》专辑 前言

周傲英

1965年，数据库领域第一个获得图灵奖的计算机科学家Charles Bachman发表了他的重要论文“Integrated Data Store”，向世人介绍了世界上第一个数据库系统IDS.这一事件标志着数据库的诞生.到现在，数据库概念已经耳熟能详，深入人心，数据库被认为是信息化社会的重要基础设施之一.随着信息技术的发展，数据的生成和收集技术迅猛发展，数据量呈爆炸式增长态势，人们在日常生产生活和科学研究中对数据的依赖越来越强.传统的数据管理技术和系统难以应对海量多源异构数据的维护、保存和使用中所面临的问题.海量数据处理极大地扩展了传统数据库研究的内涵和外延.从应用方面而言，海量数据处理已成为当前信息服务和科学发现的基石；从技术上而言，海量数据处理是传统数据管理体制的一次变革，是当前技术和应用发展的必然趋势.

在国际上，2011年2月11日发生的两件事可以用来说明海量数据在当前科学研究和信息服务中的重要性.这一天在美国出版的《科学》(Science)杂志刊登了一个名为数据处理(Dealing with Data)的专辑，《科学》还联合了《科学—信号传导》(Science：Signaling)、《科学—转化医学》(Science：Translational Medicine)和Science Careers推出相关专题，其主题是围绕目前科学研究数据的海量增加展开讨论，说明海量数据对科学研究的重要性.也在这同一天，在美国很受欢迎的智力竞答“危险边缘(Jeopardy)”电视节目中，IBM的“沃森”计算机以绝对优势战胜两名人类顶级选手，这使得继“深蓝”计算机1997年战胜人类国际象棋大师后再次引发关于机器能力的关注.和14年前的“深蓝”相比，“沃森”除具有超群的计算能力外，更拥有超大规模的数据以及数据处理能力.

对许多学科而言，海量数据意味着更严峻的挑战，更好地组织和使用这些数据会有助我们将巨大机遇变为现实.美国总统科技顾问委员会(PCAST)2010年12月提交给总统和国会的报告中明确提出“数据密集的科学和工程”(DISE)概念，并在随后几个月的国家科学局和国家科学基金的各种会议上进行了深入的讨论.与发达国家相比，我国在海量数据的收集、管理和应用方面亟待加强.

海量数据的产生和使用是与应用密切相关的，就当前的情况而言，基于Web的互联网应用、支持商务智能的大型数据中心建设以及当代科学研究所依托的科学研究数据管理是海量数据管理和计算的三个重要领域.就Web应用而言，传统的电子商务系统和搜索引擎应用以及正在兴起的社会网络和社会计算是典型的“以数据为中心”的应用.电子商务和搜索引擎厂商已经经过了早期的粗放式的仅仅依靠创新的商业模式就取得成功的发展阶段，他们的核心业务已经变成了商品推荐、客户关系管理、促销策略设计、广告关键字竞标、广告投放等.而这些业务完全依赖于海量的客户行为数据以及Web内容和结构数据的分析.社会网络和社会计算则是更加综合性的应用，交互性更强，数据的产生和来源也更多.其成功的商业模式必然需要精细的快速的数据处理和分析.除了商业应用外，社会网络和社会分析对于政府把握民众意愿、了解社会热点问题、改善管理、及时化解社会矛盾等具有重要的意义.在商务智能方面，随着技术的进步和理念的更新，大型数据中心的建设已被大型跨国跨地域企业、政府服务机构提上议事日程.针对科学研究，科学实验数据的共享以及跨地域的科学协作研究在互联网时代已经成为一个潮流.传感器网络等各种数字化科学数据采集手段的发展使得科学实验数据的产生更加便利、全面和及时.在互联网环境下对海量的科学数据和科学文献进行集成和分析并支持协同合作研究是我们面临的一个重要问题.

基于以上认识，本人在2011年2月15日召开的《计算机学报》主编会议上提出了结合数据库年会组织一个专辑的请示，获得了批准和授权.本专辑拟定的重点放在数据库研究新方向的介绍或重大研究进展的展示，以展示我国数据库界关于未来研究的新观点，引领研究方向，或反映相关研究的当前水平.

本专辑由两部分组成.第一部分是数据库年会的优秀论文，共有15篇.会议程序委员会推荐了20篇论文，后经于戈教授和其他4位《计算机学报》数据库领域编委按照学报正常要求重新评审，同意收录其中的15篇.第二部分是本次专辑特邀的论文11篇.通过向国内外知名数据库学者约稿，我们共收到15篇稿件，经过《计算机学报》数据库领域编委的严格评审以及作者的修改，本次专辑同意收录其中的

11 篇. 约稿的对象主要包括现任的数据库专委会负责人、《计算机学报》数据库领域编委、国家“千人计划”入选专家、国家杰出青年基金获得者等.

特邀的 11 篇论文可以分为三类:海量数据管理架构和处理算法、海量数据应用、海量数据处理关键技术.

在架构和算法方面,王珊教授等的论文分析了海量数据管理所面临的挑战,分析了现状,列举了他们的科研进展,并且展望了未来的研究方向. 论文体现了深厚的理论功底和很强的应用价值. 于戈教授等结合云计算的特点,介绍了在云计算环境下进行大规模图数据处理的关键问题. 内容全面,对读者了解并开展相关研究具有很好的指导价值. 李战怀教授等改进了 MapReduce 的处理框架,方便高效地处理需要迭代处理的图应用问题. 提出了将 BSP 模型嵌入到 MapReduce 模型的 Map 阶段或 Reduce 阶段中,从而降低了需要多次迭代处理的图处理应用的多次级联 Job 启动的开销问题,设计了两种消息传递机制的实现方法,以适应大消息量和小消息量的情况,讨论了容错问题,并在较大规模的数据上验证了提出的处理框架的有效性. 王国仁教授的论文针对 Skyline 查询处理中的大数据问题,提出了在 Map-Reduce 平台上的优化处理算法. 论文反映了当前在 Skyline 的海量数据查询处理的前沿研究,显示了论文工作在 Skyline 查询处理性能方面的优势.

在应用方面,周立柱教授等探讨了 Deep Web 的查询策略,提出了基于随机游走的算法对查询日志进行挖掘,从而对用户查询进行分析;由于海量 Deep Web 数据不可能通过 crawler 的方式下载后进行查询,他们提出了一种基于采样的方式来获取数据源,解决 Deep Web 的数据库选择问题. 计算广告是随着 Web 研究与应用的深入而发展起来的新的应用,目前尚不为人所熟知. 周傲英教授等就计算广告的来源与发展进行了全面的介绍,是一篇让读者从中受益的综述性论文. 孟小峰教授的论文就移动情景下的用户轨迹隐私研究进行了综述,反映了隐私研究方面的最新动向,分析了目前的轨迹隐私研究三类方法各自的优缺点,提出了轨迹隐私的性能指标与分析方法,说明了轨迹隐私的当前挑战. 唐常杰教授等的论文关注近年来出现的干预规则挖掘. 论文以四年的研究实践为背景,介绍了干预规则挖掘的研究源革和现状,给出了干预规则挖掘的任务分类. 从三个角度,即干预效果预测、干预方法发现和未知干预探测三方面,介绍了干预规则挖掘的研究问题、困难和成果. 展望了干预规则挖掘未来的研究方向.

在关键技术方面,樊文飞教授等的论文介绍了复杂数据上实体识别的概念,讨论了 XML 数据、图数据和复杂网络上的实体识别方法,给出了复杂数据上实体识别方面有待研究的问题,对数据质量等相关领域的研究人员具有指导价值. 林学民教授等侧重于研究计算机学科中涉及多个领域的相似度查询问题,他们的文章对数据库领域近十年来针对这个问题的这部分研究成果进行了总结性的讨论. 文章着眼于集合和字符串的相似性查询问题,根据技术的改进流程,对相关的工作做了较为深入、详细的讨论,并对未来的研究提出了三个有价值的方向. 周晓方教授等的论文主要探讨了在关系数据库中如何有效地管理查询结果的溯源信息,提出了以溯源树结构来存储和管理关系查询的溯源信息,并对溯源树的构造和优化进行了讨论,并且通过实验对溯源树的有效性进行了验证.

在本专辑出版之际,衷心感谢各位受邀和赐稿的学者,感谢《计算机学报》主编会议的信任和编辑部的协助. 特别感谢《计算机学报》数据库领域编委王珊教授、周立柱教授、李建中教授和于戈教授,正是他们及时严格的审稿和把关才使得本专辑得以顺利完成.

周傲英,《计算机学报》副主编. 现任华东师范大学教授、软件学院常务副院长、海量计算研究所所长. 国家自然科学基金杰出青年基金获得者,教育部长江学者特聘教授. 现任中国计算机学会数据库专业委员会副主任. 担任《VLDB Journal》、《WWW Journal》等学术期刊编委会成员. 目前的主要研究领域为数据密集型计算的数据管理、分布 P2P 数据存储和管理、中文 Web 基础设施与 Web 挖掘、不确定性数据管理与数据流分析、数据管理服务.