# 前　言

已故的 Jim Gray 在其《事务处理》一书中提到：6000 年以前，苏美尔人(Sumerians)就使用了数据记录的方法，已知最早的数据是写在土块上，上面记录着皇家税收、土地、谷物、牲畜、奴隶和黄金等情况.随着社会的进步和生产力的提高，类似土块的处理系统演变了数千年，经历了殷墟甲骨文、古埃及纸莎草纸、羊皮纸等.19 世纪后期，打孔卡片出现，用于 1890 年美国人口普查，用卡片取代土块，使得系统可以每秒查找或更新一个“土块”(卡片).可见，用数据记录社会由来已久，而数据的多少和系统的能力是与当时的社会结构的复杂程度和生产力水平密切相关的.

随着人类进入 21 世纪，尤其是互联网和移动互联网技术的发展，使得人与人之间的联系日益密切，社会结构日趋复杂，生产力水平得到极大提升，人类创造性活力得到充分释放，与之相适应的数据规模和处理系统发生了巨大改变，从而催涌了当下众人热议的大数据局面.

从历史观的角度看，数据($D$)和社会($S$)形成一定的对应关系，即：$D_1 \sim f(S_{\text{Sumerians}})$，…，$D_{\text{big}} \sim f(S_{\text{present}})$，…，$D_n \sim f(S_{\text{future}})$.从量的关系上，$D_1$，…，$D_{\text{big}}$，…，$D_n$ 可能存在大小关系，还可形成包含关系，但它们只是与当时的社会发展状况相对应：$D_{\text{big}}$ 不可能反映代表未来的 $D_n$，因为我们不知道未来会有什么新的社会结构(诸如当下社交网络一类的事物)出现，也不知道会有什么新的生产活动(诸如电商一类的事物)产生；同样 $D_1$ 也不需要具有 $D_{\text{big}}$ 的规模，当时人们并没有如此频繁的联系.近期，网络上热议美国加州大学伯克利分校 Michael I. Jordan 教授提出的“大数据的冬天即将到来”，如果我们能历史地认识 $D_{\text{big}}$ 的地位，没有把 $D_{\text{big}}$ 当 $D_n$，就不存在“冬天”与“春天”的问题.这是历史客观发展的事实.

基于以上分析，当下大数据的产生主要源于人类社会生活网络结构的复杂化、生产活动的数字化、科学研究的信息化等，其意义和价值在于如何帮助人们解释复杂的社会行为和结构，以及提高人们生产制造的能力，进而丰富人们发现自然规律的手段.本质上，大数据具有以下 3 方面的内涵，即：大数据的“深度”、大数据的“广度”、以及大数据的“密度”.所谓“深度”是指单一领域数据汇聚的规模，可以进一步理解为数据内容的“维度”.而数据的“广度”则是指多领域数据汇聚的规模，侧重体现在数据的关联、交叉和融合等方面.大数据的“密度”是指时空维上数据汇聚的规模，即数据积累的“厚度”以及数据产生的“速度”等.

面对不断涌现的大数据应用，数据库乃至数据管理技术面临新的挑战.传统的数据库技术侧重考虑数据的“深度”问题，主要解决数据的组织、存储、查询和简单分析等问题.其

后,数据管理技术在一定程度上考虑了数据的“广度”和“密度”问题,主要解决数据的集成、流处理、图结构等问题.这里提出的大数据管理是要综合考虑数据的“广度”、“深度”、“密度”等问题,主要解决数据的获取、抽取、集成、复杂分析、解释等技术难点.因此,与传统数据管理技术相比,大数据管理技术难度更高,处理数据的“战线”更长.

基于以上分析,借《计算机研究与发展》专题化改革的契机,我们适时提出组织本期“大数据管理”专题,将大数据管理主题定位在“大数据管理理论和方法”、“大数据管理系统和技术”以及“面向新型存储器件的大数据管理”等方面.其中大数据管理理论和方法包括大数据集成技术、大数据分析与查询技术、大数据可视化技术、大数据隐私管理等.大数据管理系统和技术主要包括大数据管理的编程语言、大数据管理的编译技术以及大数据管理的生态系统(分布式、众包、实时等)等.面向新型存储器件的大数据管理则主要包括新型体系结构、基于SSD的混合存储系统以及高效节能系统等.

在征文发出之后,本期“大数据管理专题”得到同行的广泛关注.通过专题公开征文以及约稿征得近百篇高质量的投稿,这些论文分别在多个研究方向上阐述了大数据管理领域具有重要意义的研究成果,展示这个领域近年来的热点及研究现状.本专题的审稿严格按照期刊审稿要求进行,特邀编委先后邀请了10多位大数据管理及相关领域的专家参与评审,历经初审、复审、专家会议终审等阶段,最终从中遴选出13篇论文入选本专题.包含5篇综述、8篇研究性论文,内容分别涵盖大数据管理的理论和方法、大数据管理的生态系统、面向新硬件的大数据管理技术以及网络大数据管理等内容,在一定程度上反映了当前国内学者在大数据管理领域的主要研究工作.

大数据管理理论和方法问题中,首当其中的是隐私管理问题.传统的针对小数据的被动式保护方式在大数据环境下已经不太适用.因此“大数据隐私管理”(孟小峰、张啸剑)一文分析了大数据时代的隐私特征以及大数据管理中存在的隐私风险和隐私管理关键技术,第一次提出了大数据隐私主动式管理建议框架.数据质量一直是数据管理的核心关注点,在大数据环境下,数据质量变得更为复杂和不确定,函数依赖发现是一种非常有效的修正数据质量的方法,但是其通常仅适用于数据规模较小的情况,针对这种情况,“分布式大数据函数依赖发现”(李卫榜、李战怀等)提出了一种分布式环境下大数据的函数依赖发现算法,为大数据环境下数据质量管理提供了一种有效的方法.大数据具有多源异构的特点,不同数据源之间的数据不一致是一个普遍存在的问题.“Web大数据环境下的不一致跨源数据发现研究”(余伟、李石君等)一文针对这个问题,通过建立统一数据抽取算法和对象数据模型,在MapReduce架构之下提出了不一致数据的自动发现算法.

大数据管理的生态系统主要包含众包、分布式、实时处理等几类,本专题针对这几类典型系统,分别选择国内相关代表性工作进行了介绍.面对大数据的“广度”问题,人机协作的群体计算是进行数据分析和复杂认知推理的有效途径,“大数据群体计算中用户主题

感知的任务分配”(张晓航、李国良等)针对大数据群体计算中任务分配问题,提出了一种基于用户主题感知的迭代式任务分配算法.面对大数据的“密度”问题,流数据分布式处理是解决问题的有效途径,“分布式流处理技术综述”(崔星灿、禹晓辉等)一文回顾分布式流处理技术产生的背景以及技术演进过程,指出了现有系统的解决方案的优势和不足,对当前比较流行的 S4,Storm,Spark Streaming 等几种代表性的分布式流处理系统对比,指出了该领域进一步的研究方向.“大数据分析与高速数据更新”(陈世敏)一文详细解读了大数据“速度”的重要性以及面临的主要挑战,指出优化高速数据更新的数据组织和数据分布方式,是保证提高数据分析运算的效率的重要方面.

大数据时代,传统的计算机系统架构已经很难满足处理需求 ,因此必须利用新的硬件技术来提升系统效能.从目前来趋势来看,新型处理器以及新型存储设备是主流研究方向,因此本专题组织了 3 篇文章,分别从不同角度对新硬件条件下的大数据管理技术进行了阐述.兼具内存读取速度快和磁盘非易失特性的相变存储器 PCM 的出现为大数据处理带来了新的契机,“基于 PCM 的大数据存储与管理研究综述”(吴章玲、金培权等)一文概述了相变存储器的发展现状,总结了当前的研究进展,指出了若干未来发展方向.在某些大数据应用中,GPU 有着甚于 CPU 的处理效率,“基于 GPU 加速的超精简型编码数据库系统”(骆歆远、陈刚等)一文实现了一种新型超精简型编码的数据库系统 HEGA-STORE,在规则挖掘和编码中使用 GPU 作为协处理器并行处理从而提高了执行效率和系统性能.降低能耗是大数据管理中亟待解决的问题,“一种异构集群中能量高效的大数据处理算法”(丁有伟、秦小麟等)一文针对异构集群下 I/O 密集型的大数据处理任务提出一种新的能量高效算法 MinBalance,相比于传统方法可以减少超过 60%的能量消耗.

从征文整体来看,目前大数据的研究主要还是集中在网络数据,有近一半的投稿集中在这方面,其主要原因可能在于网络数据本身所具备的一些特性:一方面数据规模较大,是典型的大数据场景;另一方面网络数据的获取相对较易,且复杂的网络关系之间蕴含着很多可研究的问题.网络数据之间总体呈现出图的结构特征,“大规模图数据匹配技术综述”(于静、刘燕兵等)对大规模图数据上的高效查询、匹配技术进行了全面的总结和归纳,对当前的主流方法进行了介绍,并通过实验对这些方法的效率进行了详细的对比分析.如果在简单的网络图基础上赋予其边正、负关系,则形成了符号网络,这种符号网络具有非常丰富的应用场景,“符号社会网络中正负关系预测算法研究综述”(蓝梦微、李翠平等)对符号社会网络中链接的正负预测问题进行了研究,对其研究现状和最新进展进行了全面的剖析,重点介绍了主流的预测算法,分析了当前符号社会网络预测问题面临的问题和挑战,并给出了未来的发展方向.网络数据的链接预测也是当前网络研究的一个热点,由于真实世界数据大都存在噪声,传统链接预测方法效果不是很好,“基于低秩和稀疏矩阵分解的多源融合链接预测算法”(刘冶、朱蔚恒等)针对这一问题提出了一种新的链接预测方

法,通过多数据源的融合,有效地利用主数据源和附加数据源的信息.微博等社交网络中事件的传播分析是一个非常重要的研究点,传统的分析方法仅能对单条博文进行影响传播分析,限制了这些方法的实际应用,“基于微博的事件传播分析”(朱湘、贾焰等)通过对热点博文构成的事件进行用户去重,构建事件传播拓扑图来过滤垃圾用户节点,最后在此基础上利用概率阅读模型进行传播分析,有效的弥补了传统方法的不足.

尽管网络大数据研究非常火热,也产生了一定的经济价值.但是结合前面的分析可以发现,网络大数据有广度但深度不足.因此用网络数据解释社会、分析现象存在一些值得商榷的地方.当前的网络大数据研究大多只是对传统意义下的社会现象或社会科学规律的再验证,没有真正地帮助社会管理者或人文社会科学家解决过他们关心的重大问题,这也许正是传统人文社会科学学者对全新的大数据技术无动于衷的主要原因.在网络大数据研究中,研究者一般不是从问题出发搜集数据,而是看数据能研究什么问题,数据本身又常常缺乏社会科学研究所需的属性,因此大数据并不能直接回答研究者真正感兴趣的问题,也很难获得清晰的因果关系,这是学术界需要审视的问题.

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.由于大数据管理研究领域主题十分广泛,对于特邀编委而言,面临着更加艰巨的选稿挑战.由于来稿数量大、时间安排紧张、专题容量有限等原因,部分优秀稿件无法列入发表,因此本专题无法全面体现大数据管理领域最新的所有研究工作.

我们要特别感谢《计算机研究与发展》编委会和编辑部,他们为本期专题编辑和出版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(中国人民大学信息学院)

2015 年 1 月