软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(7):1603−1604 [doi: 10.13328/j.cnki.jos.005042]

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

大数据可用性理论、方法和技术专题前言*

李建中[1], 杜小勇[2]

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

[2](中国人民大学 信息学院,北京 100872)

通信作者: 杜小勇, E-mail: duyong@ruc.edu.cn

中文引用格式: 李建中,杜小勇.大数据可用性理论、方法和技术专题前言.软件学报,2016,27(7):1603−1604. http://www.jos.org.cn/1000-9825/5042.htm

信息技术的快速发展,特别是信息获取技术、信息物理系统、互联网、物联网、社交网络等突飞猛进,引发了数据规模的爆炸式增长.能源、制造业、交通运输业、服务业、科教文化、医疗卫生等领域都积累了 TB 级、PB 级乃至 EB 级的大数据.这些大数据已经开始造福于人类,成为信息社会的重要财富.大数据蕴含着巨大的价值,对社会、经济、科学研究等各个方面都具有重要的战略意义,为人们更深入地感知、认识和预测物理世界提供了前所未有的丰富信息.由于大数据的迅速涌现及其蕴藏的巨大价值,已引起国内外学术界、工业界和政府部门的广泛关注.

伴随着数据的爆炸式增长,数据质量问题也随之而来,劣质数据的存在,极大地降低了数据可用性.事实表明,大数据在可用性方面存在着严重问题(以下简称数据可用性问题).国外权威机构的统计结果表明,美国企业信息系统中 1%~30%的数据存在各种错误和误差,美国医疗信息系统中 13.6%~81%的关键数据不完整或陈旧.国际著名科技咨询机构 Gartner 的调查结果显示,全球财富 1 000 强企业中超过 25%的企业信息系统中的数据不正确或不准确.可以预见,随着大数据应用的不断扩大,数据可用性问题将日趋严重,也必将导致源于数据的知识和决策的严重错误.

大数据可用性已经成为国内外学术界、产业界和用户普遍关注的热点问题,在国内外掀起了一个空前的研究热潮.本专题为“大数据可用性理论、方法和技术”,将突出目前大数据可用性研究中的热点技术,如大数据可用性基础理论、数据获取过程中的可用性问题、错误自动检测与修复、弱可用数据(指不完整数据、不一致数据、不精确数据、时效性错误数据等)上的知识发现与近似计算,不确定知识的演化与管理等等.

专题公开征文,共征得投稿 25 篇,内容涉及数据可用性的理论、方法和技术.特约编辑先后邀请了国内外在该领域的一些重要学者参与审稿工作,每篇投稿至少邀请 2 位专家进行初审,部分稿件还另外邀请专家进行复审.稿件评审历经约 6 个月的时间,经初审、复审等各个阶段,最终有 8 篇论文入选本专题,录用率 30%左右.此外,国家 973 项目“海量信息可用性基础理论与关键技术研究”首席专家、哈尔滨工业大学李建中教授为本专题撰写了一篇综述,对该项目所取得的一些重要成果进行了介绍.但总体而言,本专题的内容并没有覆盖数据可用性的全部方面,一方面说明这个领域的研究还属于起步阶段、成果不多,还需要吸引更多的人去研究,另一方面也说明这是大数据面临的重大挑战,难度大,需要更多的时间去积累.

李建中等人的《大数据可用性的研究进展》是一篇综述性文章,首先对数据可用性的概念进行了界定,之后归纳了研究挑战和研究问题,该文综述了数据可用性的研究成果,并给出了未来的研究方向.丁小欧等人的《数据质量多种性质的关联关系研究》探讨了数据可用性各要素之间的关联关系.针对影响数据可用性的 4 个重要性质:精确性、完整性、一致性、时效性整理出在数据集合上的操作方法,并逐一介绍其违反模式的定义,随后给出其具体关系证明,进而确定数据质量多维关联关系评估策略.

“不确定性建模”是数据可用性的一个基础性的问题.陈俞等人的《统计粗糙集》针对现有的模糊粗糙集方

* 收稿时间: 2016-05-20

法难以应用到大规模数据集上的问题,将随机抽样引入到经典的模糊粗糙集理论中,建立了一种统计粗糙集模型. 相比经典模糊粗糙集模型,该文提出的方法以随机抽样得到的小容量样本代替大规模全集,从而显著降低了计算量.而且,随着全集数量的增大,抽样样本数量并不会显著增大.

“真值发现”就是要感知数据缺陷的存在.李天义等人的《一种多源感知数据流上的连续真值发现技术》探讨了多源数据流中的真值发现问题,提出一种变频评估数据源可信度的策略,提高了每一时刻多源感知数据流真值发现的效率.

“数据修复”是数据可用性研究中的重要话题之一,现有的工作主要研究基于函数依赖的数据修复技术,即以函数依赖来描述数据一致性约束,通过变更数据库中部分元组的属性值(而非增加/删除元组)以使得整个数据库遵循函数依赖集合.本专题收录了 2 篇超越传统的基于函数依赖的数据修复方法的研究成果.金澈清等人的《基于函数依赖与条件约束的数据修复方法》探讨了开发新型数据清洗技术来提升数据质量的问题.该文考虑了其他的一些一致性约束描述,例如硬约束、数量约束、等值约束、非等值约束等,考虑以函数依赖与其他一致性约束共同表述数据库的一致性约束时的数据修复算法设计.徐耀丽等人的《基于可能世界模型的关系数据不一致性的修复》提出一种基于可能世界模型的不一致性修复方法.它首先构造可能的修复方案,然后从修复代价和属性值相关性两个方面量化各个候选修复方案的可信性程度,并最后找出最优的修复方案.实验结果验证了本文提出的修复方法,取得了比现有基于代价的修复方法更好的修复效果.

“劣质容忍”是指明知道数据存在缺陷,但是还需要在数据上进行查询和挖掘,并保证查询或者挖掘的结果可用.章志刚等人的《面向海量低质手机轨迹数据的重要位置发现》研究手机轨迹数据上的挖掘问题.手机轨迹数据具有规模庞大、位置精度低以及手机用户的多样性的特点,该文提出了一个通用框架以提高轨迹数据可用性.该框架包含一个基于状态的过滤模块,提高了数据的可用性,以及一个重要位置挖掘模块,提高了挖掘结果的准确性和精确度.除了上述工作以外,压缩和加密是数据存在的两种常见的形态,也可以看作是一种“劣质”存在.如果采用先解压或解密再进行检索的思路,势必会影响系统的效率.从近似检索的意义上看,如果能够直接在压缩或者加密数据上进行检索,也就是“劣质容忍”进行检索和挖掘,可以极大地提高数据的可用性.王佳英等人的《面向压缩生物基因数据的高效的查询方法》探讨了在压缩 DNA 数据上直接进行查询的高效方法,重点放在提高索引和查询的可伸缩性上,支持任意长度序列的精确和近似查询.黄冬梅等人的《基于 Henon 映射的加密遥感图像的安全检索方案》则探讨在加密遥感大数据上进行直接检索的方法.该文提出了一种基于 Henon 映射的遥感图像可搜索加密方案.可有效地提高检索密文遥感图像的安全性及准确性,且计算复杂度低、通信成本开销小.

本专题主要面向大数据领域的研究人员,包括数据库、数据挖掘、机器学习等,专题反映了我国学者在大数据可用性领域的最新研究进展.在此,我们要特别感谢《软件学报》编委会对专题工作的指导和帮助,感谢编辑部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,感谢专题国内外评审专家及时、耐心、细致的评审工作.此外,我们还要感谢向本专题踊跃投稿的作者对《软件学报》的信任.

最后,感谢专题的读者们,希望本专题能够对相关领域的研究工作有所促进.

李建中(1950－),男,黑龙江哈尔滨人,博士,教授,博士生导师,CCF 会士,现任 CCF 传感器网络专委会主任,主要研究领域为数据库系统实现技术,数据仓库,半结构化数据,传感器网络,压缩数据库技术,Web 数据集成,数据挖掘,计算生物学.

杜小勇(1963－),男,博士,中国人民大学信息学院教授,博士生导师,CCF 会士.现任 CCF 常务理事、数据库专委会主席,担任 APWEB2010,DASFAA2016 大会程序委员会共同主席等.主要研究领域为数据库与智能信息检索.