文章编号:1007-5321(2015)03-0001-12 **DOI**:10.13190/j.jbupt.2015.03.001

面向大数据的分析技术

高志鹏[1], 牛 琨[2], 刘 杰[3]
(1. 北京邮电大学 网络与交换技术国家重点实验室, 北京 100876; 2. 北京邮电大学 软件学院, 北京 100876;
3. 北京邮电大学 电子工程学院, 北京 100876)

摘要: 大数据分析作为整个大数据处理流程的核心,旨在从对大数据的分析中获取知识,其相关内容包括可视化分析、数据挖掘、预测分析、语义分析及数据质量管理. 从大数据时代背景出发,介绍大数据分析的基础理论,阐述大数据分析相关的前沿技术和处理工具,总结当前大数据分析所面临的机遇和挑战,并就大数据分析的发展方向和未来前景进行讨论.

关 键 词: 大数据; 可视化分析; 数据挖掘; 预测分析; 语义分析; 数据质量管理
中图分类号: TN911.22 **文献标志码**: A

Analytics Towards Big Data

GAO Zhi-peng[1], NIU Kun[2], LIU Jie[3]
(1. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China;
3. School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Big data analysis aiming at knowledge discovery from the analysis of big data, which involves visualization analysis, data mining, prediction analysis, semantic analysis and data quality management, plays an essential part in the procedures of big data processing. In this paper, it introduces the basic theories, cutting-edge technologies and processing tools for big data analysis, and summarizes encountered opportunities and challenges, along with discussions about the development trend and future prospects of big data analysis.

Key words: big data; visual analysis; data mining; predictive analysis; semantic analysis; data quality management

2011 年麦肯锡全球研究院发布的《大数据:下一个创新、竞争和生产力的前沿》研究报告中首次提及大数据(big data)概念,麦肯锡称:“数据,已经渗透到当今每个行业和业务职能领域,成为重要的生产因素. 人们对于海量数据的挖掘和运用,预示着新一波生产率增长和消费者盈余浪潮的到来. ”根据国际数据公司(IDC)监测,人类产生的数据量正在呈指数级增长,大约每两年翻一番,且此速度将在 2020 年前继续保持. 与此同时,互联网全球化、移动设备普及化、云计算存储低成本化、物质世界网络化,都在为“数据大爆发”储蓄能量,大数据已成为继云计算、物联网之后 IT 产业又一次颠覆性的技术变革. 大数据时代,人们能从数据中获得可转化为推动人类生活方式变革的有价值知识,它将成为下一个科技创新、市场竞争与生产力提高的前沿.

《Nature》于 2008 年出版了大数据专刊“Big

收稿日期: 2015-03-27
基金项目: 国家自然科学基金项目(61272515); 国家科技支撑计划项目(2015BAH03F02); 北京高等学校青年英才计划项目(YETP0474)
作者简介: 高志鹏(1980—), 男, 副教授, E-mail: gaozhipeng@bupt.edu.cn.

Data",专门讨论了巨量数据对互联网、经济、环境以及生物等各方面的影响与挑战.《Science》于2011年出版了如何应对数据洪流(data deluge)的专刊"Dealing with Data"[1]. 2013年,中国产生的数据总量超过0.8 ZB,是2012年的2倍,相当于2009年全球的数据总量. 预计到2020年,中国产生的数据总量将是2013年的10倍. 2013年11月19日,国家统计局与阿里巴巴、百度等11家企业签署大数据战略合作框架协议. 2014年3月,"大数据"首次出现在全国两会的《政府工作报告》中. 2014年5月,美国白宫发布了2014年全球"大数据"白皮书的研究报告《大数据:抓住机遇、守护价值》,鼓励使用数据以推动社会进步. 大数据已成为国家战略布局的重要组成部分,给各行各业带来根本性变革.

大数据时代的战略意义不仅在于掌握庞大的数据信息,还在于发现和理解信息内容及信息与信息之间的关系[2],而大数据分析就是大数据研究领域的核心内容之一,数据分析是决策过程中的决定性因素,也是大数据时代发挥数据价值的最关键环节. 从有数据分析的历史开始,根据需要对收集到的有限量的数据进行分析即为知识发现的关键环节. 由于各种条件的限制,如资源、计算能力等,人们基本处理的都是小量的、结构化的数据. 由于大数据的4 V特性,即体量巨大(volume)、类型繁多(variety)、时效性高(velocity)以及价值高密度低(value))[1],使大数据分析与传统的数据分析有较大差异,随之给数据分析带来诸多挑战.

笔者主要从大数据分析的基础理论出发,介绍大数据分析的关键技术和前沿工具,通过国内外对大数据分析技术的典型应用进一步说明大数据分析面临的挑战和机遇,最后对大数据分析的未来发展进行讨论.

1 大数据分析基础概念

1.1 大数据基础

IDC于2011年给出大数据定义[3],"大数据技术描述了一个技术和体系的新时代,被设计于从大规模多样化的数据中通过高速捕获、发现和分析技术提取数据的价值". 此定义体现了大数据的4V特征:体量巨大(volume)、类型繁多(variety)、时效性高(velocity)以及价值高密度低(value).

1) 体量巨大(volume)

大数据包括3种主要形式:结构化、半结构化和非结构化. 20世纪80年代末期,数字技术的盛行导致数据容量从GB上升到TB级别,超出了单个计算机系统的存储和处理能力,数据并行化和几种基于底层硬件架构的并行数据库被提出. 20世纪90年代末期,互联网时代带来了PB级别的半结构化和非结构化数据,相对于结构化数据而言,不方便用数据库二维逻辑表来表现的非结构化数据不仅规模巨大而且增长迅速,占据了总数据量的80% ~90%. 模式自由、快速可靠、高度可扩展的NoSQL数据库技术开始出现并被用来处理这些数据. 实现第4范式的唯一方法就是开发新一代的计算工具用于管理、可视化和分析数据[4].

2) 类型繁多(variety)

大数据的异构和多样性在于其有多样的形式(不仅包括传统的关系数据类型,也包括以网页、视频、音频、E-mail、文档等形式存在的未加工的、半结构化的和非结构化的数据[5]),且较多情况下难以预先确定模式[6],同时具有不连贯的语法或句意. 这些特征使得能在不同数据类型中进行交叉分析的技术成为大数据核心技术之一,包括语义分析技术、图文转换技术、模式识别技术、地理信息技术等.

3) 时效性高(velocity)

时效性高主要表现为数据流和大数据的移动性,要求对大数据进行实时分析而非批量式分析,数据的输入、处理与丢弃必须立竿见影而非事后见效,一般要在1 s时间给出分析结果,否则处理结果就是过时和无效的[7]. 实时处理的要求是区别大数据应用和传统数据仓库技术、BI技术的关键差别之一. 在实时处理的模式选择中,主要有3种思路:流处理模式、批处理模式以及二者的融合[6].

4) 价值高密度低(value)

大数据的价值具有稀缺性、不确定性和多样性,有价值的信息可能转瞬即逝,给对未来趋势和模式的可预测分析和深度复杂分析带来了挑战. 在大数据时代,对数据的接收和处理思想都需要转变,如何通过强大的机器算法更迅速地完成数据的价值"提纯"成为目前大数据背景下亟待解决的难题[1].

1.2 大数据分析

大数据分析是在强大的支撑平台上运行分析算法发现隐藏在大数据中潜在价值的过程[5]. 从异构数据源抽取和集成的数据构成了数据分析的原始数据,而大数据分析的核心问题是如何对这些数据进行有效表达、解释和学习[8].

大数据时代完成分析挖掘的理念需要 3 大转变[2].

1）要全体不要抽样

传统的统计学观点认为分析即为通过局部推断统计,从而了解其总体规律性. 但随机采样不仅严重依赖于采样的绝对随机性,需要被严密的安排和执行,而且不适合考查子类别的情况. 当人们想了解更深层次的细分领域时,随机采样的方法就会失效[2]. 大数据时代需要的是全数据模式来让人们正确地考查细节并进行分析[5].

2）要效率不要绝对精确

对“小数据”而言,最基本、最重要的要求就是减少错误. 但在大数据时代,精确性的优劣需要被重新审视,由于数据规模的宏大,确切数据的重要性降低,某个数据点对整个分析的不利影响可以被忽略. 大数据的混乱虽然提高了挖掘价值的难度,也增加了数据分析时设计衡量的方法以及指标的难度,但让人们对事物的认识变得更加完整真实[2].

3）要相关不要因果

小数据时代,相关关系分析和因果分析都不易执行且耗费巨大. 大数据时代往往通过应用相关关系分析海量数据,大数据处理和分析的终极目标即借助对数据的理解辅助人们在各类应用中做出合理的决策[8].

数据分析的主要目标包含推测或解释数据并确定如何使用数据、数据合法性校验、决策支持、错误原因识别、预测分析. 综上所述,大数据分析面临着来自数据量膨胀、数据深度分析需求的增长和数据类型多样化等挑战[9].

1.3　大数据处理流程

大数据处理包含 4 大重点问题:大数据的采集与管理、实时处理、分析挖掘和机器学习[10].

1）大数据采集:利用多种形态收集多源数据.

2）大数据的导入和预处理:将来自前端的数据导入至集中大型分布式数据库,或者分布式存储集群,并完成数据预处理过程.

3）分析沙漏:根据分析目标对大数据进行价值挖掘,主要在现有数据上进行基于各种算法的计算.

4）大数据展示:数据的展示需要借助可视化技术,常见的可视化技术有标签云(tag cloud)、历史流(history flow)、空间信息流(spatial information flow)等[6].

2　大数据分析技术

2.1　可视化分析

可视化分析的基础理论包括支持分析过程的认知理论、信息可视化理论以及人机交互与用户界面理论. 支持分析过程的认知理论重点研究从大数据中获取信息并形成知识的过程[1],信息搜索和获取的行为本质是意义构建行为,Pirolli 等[11-13] 的信息觅食理论为这种行为提供了理论基础. Card 等[14] 建立了意义建构循环模型,在信息觅食的基础上搜索并分析潜在的规律和模式并利用它分析解决问题的过程,形成一定决策. Green 等[15] 以信息发现活动为核心根据人和计算机各自的优势,对分析推理过程中各自的角色进行建模,提出了支持人机交互可视分析的用户认知模型.

信息可视化被 Card 等认为是从原始数据到可视化形式再到人的感知认知系统的可调节的一系列转换过程[14],或者理解为编码和解码 2 个映射过程[16],重点是能瞬间感知大量信息并在真实的基础上具有丰富的表达能力.

人机交互与用户界面理论则包括 3 个模型.

1）任务建模理论模型:支持并辅助用户认知过程,指导可视分析系统的用户界面设计与实现,追求具有多层次多粒度特征并且多领域相关[17-20].

2）交互模型:描述用户与系统为了协作完成任务目标,在互动过程中各自的角色与关系、承担的任务以及相互之间的消息反馈与影响. Keim 等[21] 对人、机两侧承担的最佳任务范畴进行了划分,同时 Pike 等[22] 根据任务的多层次特点,从高层与低层映射的维度建立了信息可视化与分析的交互模型.

3）用户界面模型:定义界面中的各种组成元素以及对于交互事件的响应方式,是任务模型和交互模型的最终实现. 对此,Puerta 等[23] 定义了完备的用户界面模型.

面向大数据主流应用的信息可视化技术,主要包括文本可视化、网络(图)可视化、时空数据可视化、多维数据可视化技术等.

1）文本可视化. 文本可视化旨在将文本中蕴含的语义特征直观展现,不仅有 DAViewer[24] 以树的形式进行的可视化、DocuBurst[25] 以放射状层次圆环的形式展示文本结构,还有 Hipp[26] 提供的将一维的文本信息投射到二维空间以便展示聚类关系的基于层次化点排布的投影方法. 此外,将动态变化

的文本中时间相关的模式与规律进行可视化展示,也是文本可视化的重要内容[1].

2) 网络(图)可视化. 网络可视化基于网络节点和连接的拓扑关系直观地展示网络中潜在的模式关系. 研究重点是解决在有限空间中可视化大规模网络和可视化网络的动态特征. Herman 等[27]综述了图可视化的基本方法和技术,基于节点和边的可视化方法(如 H-Tree)、空间填充法(如树图技术 Treemaps[28-29])和两者结合(如 TreeNetViz[30])的可视化方式直观表达了图节点之间的关系但受到规模限制.

3) 时空数据可视化. 时空数据可视化对时间与空间维度以及与之相关的信息对象属性建立可视化表征,并对与时间和空间密切相关的模式及规律进行展示,重点解决时空数据的高维性、实时性等特点[1]. 典型方法有将时间事件流与地图进行融合并使用边捆绑方法[31-32]或密度图技术[33]的流地图 Flow map[34],以三维方式直接展现时间、空间及事件的时空立方体(space-time cube)[35].

4) 多维数据可视化技术. 多维数据可视化技术的目标是探索多维数据项(基于传统关系数据库以及数据仓库的应用中具有多个维度属性的数据变量)的分布规律和模式,并揭示不同维度属性之间的隐含关系[1]. 散点图(scatter plot)[36]是最为常用的多维可视化方法,投影(projection)[37]尤其是平行坐标(parallel coordinates)[38]也被广泛使用.

2.2 数据挖掘

大数据分析核心即为挖掘,从技术角度看,数据挖掘就是从大量的、复杂的、不规则的、随机的、模糊的数据中获取隐含的、人们事先未发觉的、有潜在价值的信息和知识的过程[41]. 基本过程包括数据准备、数据挖掘、解释评估和知识运用[39].

数据准备是长期的、无规律的数据积累的结果,过程分为数据源的集成(数据对象整理、清洗等)、数据的选择(根据需求分类和提取数据集合)、数据预处理(消除数据中的非主体数据,检查数据的一致性和完整性)和数据转换(完成数据从数据源向目标数据仓库的转化过程)4 大部分[39].

数据挖掘是整个程序的关键过程,通过挖掘的目标要求选定合适的算法和数据挖掘模式,从海量数据中多次提取并转化为用户需要的知识,常见的算法有决策树、分类、神经网络等[41].

解释评估是根据一定的评估标准最终甄别并提取出有价值的模式知识. 数据挖掘发现的知识常见的有广义知识(实现方法如数据立方体、面向属性的归约等)、关联知识(发现方法如 Agrawal R 提出的 Apriori 算法)、分类知识(发现方法如 ID3 决策树方法)、预测型知识(发现方法如时间序列预测方法、神经网络和机器学习)和偏差型知识.

知识运用就是对挖掘的评估结果在现实决策中的运用,是数据挖掘价值的体现[41].

数据挖掘的分析方法包括聚类分析、分类和预测、关联分析等.

1) 聚类分析

聚类分析就是把大量的数据对象聚集成若干个簇的过程,并使得簇内对象尽量相似而簇间对象尽量相异. 现有的聚类算法大致分为划分方法(如 *K*-means、*K*-中心点算法(PAM 算法、CLARA 算法、CLARANS 算法等))、层次方法(如 BIRCH 方法、CURE 方法和 CHameleon 方法)、基于密度的方法(如 DBSCAN 算法、OPTICS 算法、DENCLIUE 算法)、基于网格的方法(如 CLIQLE 算法、STING 算法等)以及基于模型的方法(如 EM 算法)[40]. 能适用于大数据、处理不同类型数据、发现任意形状的簇、处理高维数据、具有处理噪声的能力和聚类结果可解释、易使用是聚类分析的目标.

2) 分类和预测

分类和数值预测是问题预测的 2 种主要类型. 分类是预测分类(离散、无序的)标号,而预测则是建立连续值函数模型. 分类是对已知的训练数据集表现出来的特性,获得每个类别的描述或属性来构造相应的分类器或者分类,是一种有监督的学习过程,根据训练数据集发现准确描述来划分类别[41]. 常见的分类算法主要有决策树、粗糙集、贝叶斯、遗传算法、神经网络(如 BP 和 RBF 网络)等,评估的要素为预测的准确度、计算复杂度、模型描述的简洁性、模型的可解释性和避免过度拟合.

3) 关联分析

关联分析就是利用事物之间存在的联系和相互之间的依赖性的规律,对这些事件进行的预测. 著名的关联规则发现方法有 Agrawal R 提出的挖掘布尔关联规则频繁项集的 Apriori 算法,此外还有 Han J 等提出的解决 Apriori 算法缺陷的不产生候选挖掘频繁项集的频繁模式树算法等.

近年来,大数据领域的数据挖掘方面的研究进展主要包括可扩展性、并行性、分布式算法等方面,

在大规模数据下，如何保证现有数据挖掘算法的时间和空间复杂度的应用成为研究热点. 李翠平等将经典的计算节点相似度的 SimRank 算法通过 GPU 的并行加速实现 20 倍加速比. 朱军等提出可扩展的最大化边界的话题模型来学习话题的概率分布崔鹏等提出了基于关系的异构哈希框架，做到同时优化同构和异构映射关系. 张岩等研发出 WorkiNet 工具，通过新词的及时发现有效帮助多个文本挖掘任务. Canny 和 Zhao 通过全新的算法设计方案提出了 BID 大数据处理框架. Michael I Jordan 等提出“bag of little bootstraps”方法解决传统分布式计算和并行计算中存在的问题. Karthik Raman 等[51]将大数据上的复杂分析任务分解为一系列的简单任务. 此外，社交网络分析和信息网络分析方面，Yang 等[8]提出时间序列聚类方法，从 Twitter 数据中挖掘热门话题发展趋势的规律.

2.3　预测分析

预测分析是利用统计、建模、数据挖掘工具对已有数据进行研究以完成预测. 传统预测分析技术与大数据预测分析技术有 2 点不同：首先，传统预测分析是基于关系数据仓库中的数据的，而关系数据库只对结构化数据进行批量处理；其次，传统预测分析需要特征设计，然后由特征通过假设和测试过程去驱动分析.

预测方法从技术上分为定性预测与定量预测. 定性预测是基于经验和判断对预测对象做定性分析，主要有集思广义法和德尔菲法，预测的准确程度主要取决于预测者的经验、理论以及掌握的情况和分析判断能力等，近年来人工智能也产生了如 Boosting、贝叶斯网络等一些定性预测算法. Yuan 等[42]、Laakso 等[43]、Karni 等[44]、Jin 等[45]分别对定性预测各种方法进行了探索与应用. 定量预测则是使用数学模型，根据已有的历史统计数据运用数学方法得到变量间的规律性联系，如统计分析、因果联系模拟、人工智能算法等. 常用的统计分析模型主要有指数平滑法、趋势外推法、移动平均法等，常用的因果联系模型主要有线性回归因果模型等. 定量预测的步骤主要包括分析数据识别数据模式或规律、通过数据模型进行描述和将数学模型在时间域上扩展完成预测. Togn[46]、Chen 等[47]、Zhang 等[48]对定量预测各种方法进行了研究和改进.

预测的过程主要考虑 3 个方面：计算复杂性、分类变量的因果关系以及预测模型的寻优. 选择一个恰当的预测算法需要考虑现有数据、预测形式、预测精度、实时性要求、可理解性和可操作性等因素.

预测分析是大数据技术的核心应用，但是预测分析的成功与否取决于数据质量、数据科学家（指能运用统计分析、机器学习、分布式处理等技术，从大量的数据中提炼出有价值的信息，以简单易懂的形式传达给决策者，其工作包括数据架构的搭建、数据模型的建立和数据分析）、预测分析软件（供数据科学家使用，用来评估数据科学家建立的数据模型和分析规则）3 个要点.

针对大数据数据量庞大的特点研发了专门的大数据处理平台，包括 2005 年由 Apache 发布的 Hadoop 分布式系统基本架构（其中 HDFS 提供海量的数据存储，MapReduce 提供海量的数据计算，YARN 框架负责作业调度和集群资源管理）、专门的数据分析机 Oracle Exadata 以及由数据库服务器与存储服务器组成的一体机硬件平台.

2.4　语义分析

由于非结构化数据与异构数据等的多样性带来了数据分析的新的挑战与困难，需要一系列的工具去解析、提取、分析数据. 语义引擎的设计需要其能从文档中智能提取信息，并能从大数据中挖掘出特点，通过科学建模和输入新的数据，从而预测未来的数据[14]. 语义分析即对信息所包含的语义的识别[49]. 语义分析技术是智能语义分析，包括 3 个方面，一是通过语义识别处理非结构化的社会性信息，二是通过支持大规模程序计算的自动分析应对持续快速增长的大数据，三是通过人工智能对信息进行及时处理，提高数据处理的时效性.

对于词语语义分析的研究主要是确定词语意义，衡量词与词之间的语义相似度，大体分为基于词语语义知识规则和基于统计的词语语义分析. 常见的基于词语语义分析的知识表示方法有语义场、语义网络（最早由 Quillian M R 提出）、概念图（由 Sowa J F 提出）和本体论，它们的本质都在强调对现实社会存在的概念及其之间的关系进行精确描述和建模. 现有的典型的语义知识库有 WordNet、FrameNet、MindNet、知网（HowNet）等. 对于基于规则的词语语义分析，最常用的知识规则库是语义词典，其将所有词语组织成树状层次结构，将词语在树结构图中的路径长度作为词语语义距离的度量方法. Google 于 2013 年公开的基于 Deep Learning 的学习工具 Word2vec 是将词语转换成向量的分析

工具[49].

文本语义分析就是识别文本的意义、主题、类别等语义信息的过程. 研究大体分为基于统计的文本语义分析和基于语义学的文本语义分析. 典型的大规模文本语义分析研究大多是基于统计的经验主义方法,其将文本看作一个个独立词语形成的无序词袋,利用词语的统计信息将文本表示为词语向量集合并据此分析隐含的主题等信息,代表方法有潜在语义分析、概率潜在语义分析和隐含狄利克雷分配. 而基于语义学理论的文本语义分析除了格语法、概念层次理论,还有框架语义学和本体语义学[49].

此外,情感分析是一个新兴的研究课题,它以应用为导向帮助用户快速获取和整理用户分享的评价信息,包括情感信息抽取、情感信息分类和情感信息的检索与归纳3部分任务. 目前的情感分析系统有Liu等研发的OpinionObserver系统和Wilson等研发的OpinonFinder系统等[50].

2.5 数据质量管理

大数据的4V特性使大数据存在数据质量问题,即不一致、不精确、不完整、过时或描述同一实体时数据出现冲突. 对大数据进行有效分析的前提是高质量的数据. 数据质量的评估维度:完整性(completeness)——度量丢失数据或不可用数据;规范性(conformity)——度量未按统一格式存储的数据;一致性(consistency)——度量值在信息含义上是冲突的数据;准确性(accuracy)——度量不正确信息、数据,或者超期数据;唯一性(uniqueness)——度量重复数据或者属性重复数据;关联性(integration)-度量缺失或未建立索引的关联数据.

数据质量提高技术涉及实例和模式2个层面. 数据清洗(data cleansing, data scrubbing)解决的是数据实例层面的问题,它主要研究以下内容[52-53].

1) 重复对象检测

数据集成时要判断是否不同数据源中数据是同一个实体[51-53],主要研究的2个方面:数据的重复记录检测和XML重复元素检测. 在数据重复记录方面,韩京宇等[54]、邱越峰等[55]提出了识别出相似重复记录的方法;Chaud等[56]对干净的参照表数据建立一个ETI(error tolerance index)索引,在线输入数据据此找到最匹配干净记录;Telcordia[57]具有模糊记录识别功能; Bhattacharya等[58]用迭代的方法解决了作者识别的问题. 在XML重复元素检测方面,Rohrs等[59]提出了XML文档中识别重复XML元素的方法;Pluempitiwiriyawe等[60]提出了将具有相似结构的XML元素进行合并的方法,XML数据清洗在处理数据结构多样性、发现父子元素间复杂关系以及计算复杂度方面有待探索.

2) 缺失数据处理

在统计领域,主要研究方法是单一填补法(single imputation)和多重填补法(multiple imputation),在表现数据集不确定性方面常常采用的是多重填补法. Zhao等[61]解决了含有空缺值的数据的相似记录的处理. Wu等[62]对多维数据集度量值的缺失数据进行填充,还利用约束来实现数据立方缺失值的填充,并利用条件表(conditional table)来填补关系数据库的数据表缺失的数据. Neal[63]用马尔可夫链蒙特卡罗法实现对缺失数据的填充.

3) 异常数据检测

在数据清洗时,对异常数据的处理主要采用数据审计的方法,先进行数据的概化以提取数据特征,接着进行数据挖掘发现数据异常. Hipp等[64]采用关联规则挖掘的方法发现标称型(nominal)数据异常. Lueebber等[65]采用C4.5决策树算法产生模拟数据, 检验数据挖掘算法发现异常数据的能力. 数据挖掘中的基于统计模型、基于距离、基于偏离等方法在检测异常数据时颇有成效,但在数据清洗领域,在运用数据挖掘算法之前,数据概化(data profiling)是必不可少的,其有助于探测性挖掘(exploratoty mining)[66],发现异常数据.

4) 逻辑错误检测

数据编辑修正(editing and imputation)关注如何运用自动化的方法除去信息系统中不符合业务的逻辑错误. 解决方法是根据某领域知识建立相应的领域规则体系来自动处理. Fellegi等[67]提出了形式化的数学模型(Fellegi-Holt模型),对某领域知识做显示约束规则(explicit rules),用数学方法求出规则的闭集,自动检测是否违反规则约束.

5) 不一致数据处理

不一致数据处理是为了解决由多个数据源在集成时出现数据重叠难以获取一致的理想数据的问题. Direct[68]将数据不一致划分为上下文独立冲突(context independent conflicts)和上下文依赖冲突(context dependent conflicts). 上下文独立冲突通常是由于外部原因造成的,需要人工干预. 上下文依赖冲突是由于各个数据源的数据系统设计和表达方式不同造成的,需要通过数据转换规则来解决.

Motro 等[69]对每个数据值从特征(feature)来评估，各个特征评估值的线性组合决定唯一正确的值.

此外，针对大数据的特点，传统的关系型数据库在数据管理性能方面已不能胜任. 无论是操作型还是分析型应用，为了能对大数据进行处理，并行处理是必由之路. 依赖大量的节点并行处理提高性能，MapReduce 技术部署的节点数量远远超过关系型数据库. NoSQL 采用与关系型数据库不同的数据模型，满足对大数据的读写要求.

操作型 NoSQL 技术可划分成基于 Key Value(键值对)存储模型、基于 Column Family(列分组)存储模型、基于文档模型和基于图模型的 4 类 NoSQL 数据库技术. 基于 Key Value 存储技术利用散列表维护 Key 值到具体数据(value) 的映射.

MapReduce 是面向分析型应用的 NoSQL 技术，包括 Map 阶段和 Reduce 阶段，Map 函数处理键值对，Reduce 函数合并相同 Key 值的中间键值对. MapReduce 技术因其高度的扩展性和容错性，引起了广泛研究：赋予 MapReduce 结构化存储模型(行存储和列存储)，以便 MapReduce 能有效处理结构化数据[70]；为 MapReduce 的数据处理提供索引支持[71]；对 MapReduce 框架的扩展，利用 MapReduce 框架处理流数据[72]；各种连接算法的优化[73-74]和查询算法[75]、调度算法的优化；MapReduce 计算框架的安全[76]与节能[77].

3　大数据分析工具

3.1　Weka

Weka 全名是怀卡托智能分析环境(Waikato environment for knowledge analysis)，是一个用于数据挖掘和知识发现的开源项目，能根据属性分类和集群大量数据，是现今最完备的数据挖掘工具之一. Weka 使数据挖掘的执行无须编程，集合了大量承担数据挖掘任务的机器学习算法，包括对数据进行预处理、分类、回归、聚类、关联规则以及在新的交互界面上的可视化，作为数据分析的强大工具，Weka 主要用于机器学习，它提供了在不同数据集上完成数据预处理及数据挖掘的多种算法，同时，在灵活性和可扩展性方面，它提供了文档全面的 Java 函数和类库，方便扩展.

3.2　R

R 在 1996 年由新西兰奥克兰大学的两位统计学教授——Ross Ihaka 和 Robert Gentleman 发明，是属于 GNU 系统的一个统计编程环境相对开放的软件，用来分析大数据集的统计组件包，拥有强大的社区和组件库；同时，R 是一个用于统计计算和统计制图的优秀工具，它提供的统计和制图技术包括线性和非线性建模，经典的统计测试，时间序列分析、分类、收集，等等.

R 是面向对象的统计编程语言，与 C ++ 等主流编程语言及数据库之间都有接口. R 独特的扩展插件可以提供免费的扩展，同时保证 R 语言引擎能运行在 Hadoop 集群之上. R 最主要的优点在于其灵活、开源、集思广益、更新速度快，并且使用的人越多，其数据资源就越丰富，这样会使得其呈现出来的能量呈几何级数增加.

3.3　Enterprise Miner

Enterprise Miner 作为 SAS 的功能模块，是一个集成的数据挖掘系统，可以使用、比较不同的算法模型对数据进行分析处理. SAS EM 按照“抽样—探索—转换—建模—评估”的工作过程进行数据挖掘，在每个步骤中可以按照数据挖掘项目的进展知道用户完成各种操作.

SAS EM 提供大量前沿的、用于深入分析的预测性和描述性建模算法，包括决策树、Bagging 和 Boosting、神经网络、基于记忆推理、分级聚类、线性和对数回归、关联规则、时间序列分析和 Web 路径分析等，可以与 SAS 数据仓库和 OLAP 集成，实现从提出数据、抓住数据到得到解答的“端到端”知识发现.

4　大数据分析面临的机遇和挑战

2014 年，互联网全球化、移动设备普及化、云计算存储低成本化、物质世界网络化迅速发展，大数据开始由概念走向应用，大数据技术已经在互联网、运营商、IT 服务提供商，以及众多传统企业中落地实践. 随着大数据时代的到来，“向数据要价值”使得几乎每个行业都面临着大数据问题，大数据分析面临的挑战主要体现在以下几方面.

4.1　大数据复杂性

大数据类型和模式多样，关联关系繁杂，质量良莠不齐. 大数据内在的复杂性使得数据的感知、表达、理解和计算等多个环节面临着巨大的挑战，导致了传统全量数据计算模式下时空维度上计算复杂度的激增，传统的数据分析与挖掘任务如检索、主题发现、语义和情感分析等变得异常困难. 如何形式化

或定量化地描述大数据复杂性的本质特征及其外在度量指标,进而研究数据复杂性的内在机理是个根本问题. 同时,大数据的4V特性使得传统的机器学习、信息检索、数据挖掘等计算方法不能有效支持大数据的处理、分析和计算,研究面向大数据的新型高效计算范式,改变人们对数据计算的本质看法,提供处理和分析大数据的基本方法,支持价值驱动的特定领域应用,是大数据计算的核心问题. 此外,以高效能为目标的大数据处理系统的系统架构设计、计算框架设计、处理方法设计和测试基准设计还有待研究[8].

4.2 大数据处理的实时性

随着时间的流逝数据中所蕴含的知识价值随之递减,实时处理成为大数据分析的典型需求. 大数据时代的数据实时处理面临着新挑战,主要体现在数据处理模式的选择及改进. 在实时处理的模式选择中主要有3种思路,即流处理模式、批处理模式以及2者的融合. 虽然已有的研究成果很多,但是仍未有一个通用的大数据实时处理框架. 各种工具实现实时处理的方法不一,支持的应用类型都相对有限,这导致实际应用中往往需要根据自己的业务需求和应用场景对现有的这些技术和工具进行改造才能满足要求[6].

4.3 大数据分析

挖掘大数据价值必须对大数据进行内容上的分析与计算,这就依赖于高精度的深度学习来对人类难以理解的底层数据特征进行层层抽象. 在面对大数据分析时,一方面半结构化和非结构化数据的存在,数据很难以类似结构化数据的方式构建出其内部的正式关系;另一方面很多数据以流的形式源源不断地到来,需要实时处理的数据很难有足够的时间去建立先验知识. 如何对数据庞大的结构化和半结构化数据进行高效率地深度分析、挖掘隐性知识,如语义、情感分析,即通过人工智能和机器学习技术的大数据分析技术还有待于进一步的探索[6]. 同时,借助知识计算将碎片化的多源数据整合成反映事物全貌的完整数据,从而增加数据挖掘的深度,但如何基于大数据实现新知识的感知,知识的增量式演化和自适应学习是其中的重大挑战. 对蕴含着一个体量庞大、关系异质、结构多尺度和动态演化的网络的社会媒体大数据的分析既需要有效的计算方法,更需要支持大规模网络结构的图数据存储和管理结构,以及高性能的图计算系统结构和算法. 大数据查询和分析的实用性和实效性对于人们能否及时获得决策信息非常重要,如何提出新的可视化方法帮助人们分析大规模、高维度、多来源、动态演化的信息,并辅助做出实时的决策,是大数据分析领域最大的挑战[8]. 总的来说,分为5方面:一是大数据的集成和接口问题;二是匹配心理映像的可视化表征设计与评估;三是最大限度发挥人、机各自优势的人机交互与最优化协作还需求解;四是使用户快捷方便地、自助式地实现大数据可视分析系统,满足自己的个性化需求的方法需要解决;五是超高维数据的降维、大规模并行处理方法与超级计算机的结合、可视化算法与人机交互技术的提升与拓展等系统可扩展性问题[1].

5 大数据分析未来趋势

当前的数据分析技术的研究可以分为6个重要方向:结构化数据分析、文本数据分析、多媒体数据分析、Web数据分析、网络数据分析和移动数据分析[9]. 2014中国大数据技术大会指出2015年大数据10个主要发展趋势:大数据与人工智能的融合;跨学科领域交叉的数据分析应用;数据科学带动多学科融合;深度学习成为大数据智能分析的核心技术;利用大数据构建大规模、有序化开放式的知识体系;大数据的安全持续令人担忧;开源继续成为大数据技术的主流;大数据与云计算、移动互联网等的综合应用;大数据提升政府治理能力,数据资源化、私有化、商品化成为持续的趋势;大数据技术课程体系建设和人才培养快速发展. 程学旗[78]将2015年大数据发展趋势预测总结为"融合、跨界、基础、突破".

1) 结合智能计算的大数据分析成为热点,包括大数据与神经计算、深度学习、语义计算以及人工智能其他相关技术结合. 得益于以云计算、大数据为代表的计算技术的快速发展,使得信息处理速度和质量大为提高,能快速、并行处理海量数据.

2) 跨学科领域交叉的数据融合分析与应用将成为今后大数据分析应用发展的重大趋势. 由于现有的大数据平台易用性差,而垂直应用行业的数据分析又涉及领域专家知识和领域建模,目前在大数据行业分析应用与通用的大数据技术之间存在很大的鸿沟,缺少相互的交叉融合. 因此,迫切需要进行跨学科和跨领域的大数据技术和应用研究,促进和推动大数据在典型和重大行业中的应用和落地,尤

其是与物联网、移动互联、云计算、社会计算等热点技术领域相互交叉融合[78].

3）大数据安全和隐私. 大数据时代，各网站均不同程度地开放其用户所产生的实时数据，一些监测数据的市场分析机构可通过人们在社交网站中写入的信息、智能手机显示的位置信息等多种数据组合进行分析挖掘. 然而，大数据时代的数据分析不能保证个人信息不被其他组织非法使用，用户隐私安全问题的解决迫在眉睫. 安全智能更加强调将过去分散的安全信息进行集成与关联，独立的分析方法和工具进行整合形成交互，最终实现智能化的安全分析与决策.

4）各种可视化技术和工具提升大数据分析. 进行分析之前，需要对数据进行探索式地考查. 在此过程中，可视化将发挥很大的作用. 对大数据进行分析以后，为了方便用户理解结果，也需要把结果展示出来. 尤其是可视化移动数据分析工具，能追踪用户行为，让应用开发者得以从用户角度评估自己的产品，通过观察用户与一款应用的互动方式，开发者将能理解用户为何执行某些特定行为，从而为自己完善和改进应用提供依据.

6　结束语

阐述了大数据分析的基础理论，概括了大数据分析的关键技术及典型工具，介绍了国内外对大数据分析技术的典型应用及相关研究，说明了大数据分析面临的主要挑战和机遇，最后对大数据分析的未来发展趋势进行了讨论.

参考文献：

[1] 任磊，杜一，马帅，等. 大数据可视分析综述[J]. 软件学报，2014，25(9)：1909-1936.
Ren Lei，Du Yi，Ma Shuai，et al. Visual analytics towards big data[J]. Journal of Software，2014，25(9)：1909-1936.

[2] 维克托·迈尔·舍恩伯格，肯尼斯·库克耶. 大数据时代[M]. 杭州：浙江人民出版社，2012.
Viktor M S，Kenneth C. Big data：a revolution that will transform how we live，work and think[M]. Hangzhou：Zhejiang People's Publishing House，2012.

[3] Gantz J，Reinsel D. 2011 digital universe study：extracting vaLue from chaos[M]. IDC Go-to-Market Services，2011：1-12.

[4] 李学龙，龚海刚. 大数据系统综述[J]. 中国科学：信息科学，2015，45(1)：1-44.
Li Xuelong，Gong Haigang. A survey on big data systems[J]. Science China Information Sciences，2015，45(1)：1-44.

[5] 陶雪娇，胡晓峰，刘洋. 大数据研究综述[J]. 系统仿真学报，2013，8(25)：144-145.
Tao Xuejiao，Hu Xiaofeng，Liu Yang. Overview of big data research[J]. Journal of System Simulation，2013，8(25)：144-145.

[6] 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，50(1)：146-169 .
Meng Xiaofeng，Ci Xiang. Big data management：concepts，techniques and challenges[J]. Journal of Computer Research and Development，2013，50(1)：146-169.

[7] 宫夏屹，李伯虎，柴旭东，等. 大数据平台技术综述[J]. 系统仿真学报，2014，26(3)：489-496.
Gong Xiayi，Li Bohu，Chai Xudong，et al. Survey on big data platform technology[J]. Journal of System Simulation，2014，26(3)：489-496.

[8] 程学旗，靳小龙，王元卓，等. 大数据系统和分析技术综述[J]. 软件学报，2014，25(9)：1889-1908.
Cheng Xueqi，Jin Xiaolong，Wang Yuanzhuo，et al. Survey on big data system and analytic technology[J]. Journal of Software，2014，25(9)：1889-1908.

[9] 覃雄派，王会举，杜小勇，等. 大数据分析——RDBMS 与 MapReduce 的竞争与共生[J]. 软件学报，2012，23(1)：32-45.
Qin Xiongpai，Wang Huiju，Du Xiaoyong，et al. Big data analysis—competition and symbiosis of RDBMS and MapReduce[J]. Journal of Software，2012，23(1)：32-45.

[10] 顾君忠. 大数据与大数据分析[J]. 软件产业与工程，2013(4)：17-21 .
Gu Junzhong. Big data and big data analysis[J]. Software Industry and Engineering，2013(4)：17-21.

[11] Pirolli P，Card S K. The sensemaking process and leverage points for analyst technology as identified though cognitive task analysis[C]// Maybury M. Proc. of the Int'l Conf. on Intelligence Analysis. MacLean：MITRE，2005：1-6.

[12] Pirolli P，Card S K. Information foraging in information access environments[C]// Katz I R. Proc. of the CHI. New York：ACM Press，1995：51-58.

[13] Pirolli P. Information foraging theory：adaptive interaction with information[M]. New York：Oxford University Press，2007：31-35.

[14] Card S K，Mackinlay J D，Shneiderman B. Readings in information visualization：vsing vision to think[M]. San

Francisco: Morgan-Kaufmann Publishers, 1999: 1-712.

[15] Green T M, William R, Brian F. Visual analytics for complex concepts using a human cognition model[C]// Grinsten G. Proc. of the VAST. Columbus: IEEE Press, 2008: 91-98.

[16] Wunsche B. A survey, classification and analysis of perceptual concepts and their application for the effective visualisation of complex information[C]// Chrucher N, Churcher C. Proc. of the APVIS. Darlinghurst: Australian Computer Society, 2004: 17-24.

[17] North C, Chang R, Endert A, et al. Analytic provenance: process + interaction + insight[C]// Tan D. Proc. of the CHI. New York: ACM Press, 2011: 33-36.

[18] Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations[C]// Gershon N. Proc. of the INFOVIS. San Francisco: IEEE Press, 1996: 336-343.

[19] Eades P, Huang Maolin. Navigating clustered graphs using force-directed methods[J]. Journal of Graph Algorithms and Applications, 2000, 4(3): 157-181.

[20] Brehmer M, Munzner T. A multi-level typology of abstract visualization tasks[J]. IEEE Trans. on Visualization and Computer Graphics, 2013, 19(12): 2376-2385.

[21] Keim D, Andrienko G, Fekete J, et al. Visual analytics: definition, process, and challenges[C]// Kerren A. Proc. of the Information Visualization. LNCS 4950. Berlin: Springer-Verlag, 2008: 154-175.

[22] Pike W A, Stasko S J, Chang R, et al. The science of interaction[J]. Information Visualization, 2009, 8(4): 263-274.

[23] Puerta A, Eisenstein J. Towards a general computational framework for model-based interface development systems[J]. Knowledge-Based Systems, 1999, 12(8): 433-442.

[24] Zhao Jian, Chevalier F, Collins C, et al. Facilitating discourse analysis with interactive visualization[J]. IEEE Trans. on Visualization and Computer Graphics, 2012, 18(12): 2639-2648.

[25] Collins C, Carpendale S, Penn G. Docuburst: visualizing document content using language structure[J]. Computer Graphics Forum, 2009, 28(3): 1039-1046.

[26] Paulovich F V, Minghim R. Hipp: a novel hierarchical point placement strategy and its application to the exploration of document collections[J]. IEEE Trans. on Visualization and Computer Graphics, 2008, 14(6): 1229-1236.

[27] Herman I, Melancon G, Marshall M S. Graph visualization and navigation in information visualization: a survey[J]. IEEE Trans. on Visualization and Computer Graphics, 2000, 6(1): 24-43.

[28] Shneiderman B. Tree visualization with tree-maps: 2-d spacing-filling approach[J]. ACM Trans. on Graphics, 1992, 11(1): 92-99.

[29] Zhang Xiu, Yuan Xiaoru. Treemap visualization[J]. Journal of Computer-Aided Design and Computer Graphics, 2012, 24(9): 1113-1124.

[30] Gou Liang, Zhang Xiaolong. Treenetviz: revealing patterns of networks over tree structures[J]. IEEE Trans. on Visualization and Computer Graphics, 2011, 17(12): 2449-2458.

[31] Phan D, Lin Xiao, Yeh R, et al. Flow map layout[C]// Andrews K. Proc. of the INFOVIS. Los Alamitos: IEEE Press, 2005: 219-224.

[32] Buchin K, Speckmann B, Verbeek K. Flow map layout via spiral trees[J]. IEEE Trans. on Visualization and Computer Graphics, 2011, 17(12): 2536-2544.

[33] Scheepens R, Willems N, Wetering V D H, et al. Composite density maps for multivariate trajectories[J]. IEEE Trans. on Visualization and Computer Graphics, 2011, 17(12): 2518-2527.

[34] Tobler W. Experiments in migration mapping by computer[J]. The American Cartographer, 1987, 14(2): 155-163.

[35] Peuquet D J, Kraak M J. Geobrowsing: creative thinking and knowledge discovery using geographic visualization[J]. Information Visualization, 2002, 1(1): 80-91.

[36] Ahlberg C, Shneiderman B. Visual information seeking: tight coupling of dynamic query filters with starfield displays[C]// Beth A, Susan D, Judith O. Proc. of the CHI. New York: ACM Press, 1994: 313-317.

[37] Jing Yang, Hubball D, Ward M S, et al. Value and relation display: interactive visual exploration of large data sets with hundreds of dimensions[J]. IEEE Trans. on Visualization and Computer Graphics, 2007, 13(3): 494-507.

[38] Inselberg A, Dimsdale B. Parallel coordinates: a tool for visualizing multi-dimensional geometry[C]// Kaufman A. Proc. of the Visualization. San Francisco: IEEE Press, 1990: 361-378.

[39] 马斌, 周平, 张建业, 等. 大数据时代的数据挖掘[J]. 中国科技信息, 2014(23): 117-118.

Ma Bin, Zhou Ping, Zhang Jianye, et al. Data mining in times of big data[J]. China Science and Technology Information, 2014(23): 117-118.

[40] 冯宏亮. 数据挖掘中若干关键算法的研究[D]. 西安：西安科技大学, 2010.
Feng Hongliang. The research on several key algorithms of data mining[D]. Xi'an: Xi'an University of Science and Technology, 2010.

[41] 李平荣. 大数据时代的数据挖掘技术与应用[J]. 重庆三峡学院学报, 2014, 30(151): 45-47.
Li Pingrong. Data mining technology and its applications in big data era[J]. Journal of Chongqing Three Gorges University, 2014, 30(151): 45-47.

[42] Yuan Soetsyr, Chen Yenchuan. Semantic ideation learning for agent-based e-brainstorming[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(2): 261-275.

[43] Kimmo L, Anita R, Hannu L. Delphi method analysis: the role of regulation in the mobile operator business in Finland[C]// Technology Management for Global Economic Growth (PICMET), 2010 Proceedings of PICMET'10: IEEE Press, 2010: 1-7.

[44] Edi Ka. Unknowable states and choice-based definitions of subjective probabilities [J]. Economics Letters, 2008, 99: 534-536.

[45] Zhi Jin, Chen Xiaohong, Didar Z. Performing projection in problem frames using scenarios[C]// Proc. 16th Asia-Pacific Software Engineering Conference: IEEE Press, 2009: 249-256.

[46] Togn H. On a threshold model pattern recognition and signal processing[C]// Amsterdam: Sijthoff and Noordhoff Press, 1978: 18-29.

[47] Chen Guanglei, Wang Zhaojun. The multivariate partially linear model with b-spline[J]. Chinese Journal of Applied Probability, 2010, 26(2): 138-150.

[48] Zhang G P. Time series forecasting using a hybrid ARIMA and neural network mode [J]. Neurocomputing, 2003, 50(5): 159-175.

[49] 秦春秀, 祝婷, 赵捧未, 等. 自然语言语义分析研究进展[J]. 图书情报工作, 2014, 58(22): 130-137.
Qin Xiuchun, Zhu Ting, Zhao Pengmo, et al. Research review on semantics analysis of natural language[J]. Library and Information Service, 2014, 58(22): 130-137.

[50] 赵妍妍, 秦兵, 刘挺. 文本情感分析综述[J]. 软件学报, 2010, 21(8): 1834-1848.
Zhao Yanyan, Qin Bing, Liu Ting, et al. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834-1848.

[51] 唐杰, 梅俏竹. 数据挖掘学科发展研究[A]. 2012—2013 控制科学与工程学科发展报. 北京：中国科学出版社, 2014.
Tang Jie, Mei Qiaozhu. Recent advances of data mining in China[A]. 2012—2013 Report on Advances in Control Science and Engineering. Beijing: China Science and Technology Press, 2014.

[52] 韩京宇, 徐立臻, 董逸生. 数据质量研究综述[J]. 计算机科学, 2008, 35(2): 1-5, 12.
Han Jingyu, Xu Lizhen, Dong Yisheng. An overview of data quality research[J]. Computer Science, 2008, 35(2): 1-5, 12.

[53] 王宏志. 大数据质量管理：问题与研究进展[J]. 科技导报, 2014, 32(34): 78-84.
Wang Hongzhi. Big data quality management: problems and progress [J]. Science and Technology Review, 2014, 32(34): 78-84.

[54] 韩京宇, 徐立臻, 董逸生. 一种大数据量的相似记录检测方法[J]. 计算机研究与发展, 2005, 42(12): 2206-2212.
Han Jingyu, Xu Lizhen, Dong Yisheng. An approach for detecting similar duplicate records of massive data [J]. Journal of Computer Research and Development, 2005, 42(12): 2206-2212.

[55] 邱越峰, 田增平. 一种高效的识别相似重复记录的方法[J]. 计算机学报, 2001, 24(1): 69-77.
Qiu Yuefeng, Tian Zengping. An efficient approach for detecting approximately duplicate database records[J]. Chinese Journal of Computers, 2001, 24(1): 69-77.

[56] Chaud H S, Ganjam K, Ganti V, et al. Robust and efficient fuzzy match for on line data cleaning[C]// Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. San Diego: ACM Press, 2003: 313-324.

[57] Grzymala B J, Hu Ming. A comparison of several approaches to missing attribute values in data mining[C]// Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing. Banff: Springer Berlin Heidalberg, 2001: 378-385.

[58] Bhattacharya I, Getoor L. Iterative record linkage for cleaning and integration[C]// Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. Maison de la Chimie: ACM Press, 2004: 11-18.

[59] Rohrs C E, Berry R A. A linear control approach to ex-

plicit rate feedback in ATM networks[C]// Proc. of IEEE in Focom'97. Kode: IEEE Press, 1997: 277-282.

[60] Pluempitiwiriyawej C, Hammer J. Element matching across data-oriented XML sources using a multi-strategy clustering model[J]. Data and Knowledge Engineering, 2004, 48(3): 297-333.

[61] Zhao Li, Yuan Sungsam, Yang Qixiao, et al. Dynamic similarity for fields with null values[C]// Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery. France: Springer Berlin Heidelberg, 2002: 161-169.

[62] Wu Xintao, Barbara D. Modeling and imputation of large incomplete multidimensional data sets[C]// Proceedings of the International Conference on Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg, 2002: 286-295.

[63] Neal R M. Probabilistic inference using Markov chain Monte Carlo methods[M]. CRG TR-93-1. Department of Computer Science, University of Toronto, 1993.

[64] Hipp J, Guntzer U, Grimmer U. Data quality mining: making a virtue of necessity[C]// In Proc. of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. New York: ACM Press, 2001: 52-57.

[65] Lueebber D, Grimmer U. Systematic development of data miningbased data quality tools[C]//29th VLDB. Berlin: Morgan Kaufmann, 2003: 548-559.

[66] Dasu T, Johnson T. Exploratory data mining and data cleaning[M]. Hoboken: John Wiley and Sons, Inc, 2003.

[67] Fellegi I P, Holt D. A systematic approach to automatice dit and imputation[J]. Americ Statistics of Association, 1976, 71(353): 17-35.

[68] Fan Weigui, Lu Hongjun, Madnick S E, et al. Discovering and reconciling value conflicts for numerical data integration[J]. Information Systems, 2001(26): 635-656.

[69] Motro A, Anokhin P, Acar A C. Utility-based resolution of datainconsistencies[C]// IQIS 2004: 35-43.

[70] Kaldewey T, Shekita E J, Tata S. Clydesdale: structured data processing on MapReduce[C]//Rundensteiner E A, Markl V, Manolescu I, et al. Proc. of the EDBT 2012. Berlin: ACM Press, 2012: 15.

[71] Jindal A, Quiane-Ruiz J A, Dittrich J. Trojan data layouts: right shoes for a running elephant[C]// Chase J S, Abbadi A E, Babu S, et al. Proc. of the SOCC. Cascais: ACM Press, 2011.

[72] Neumeyer L, Robbins B, Nair A, et al. S4: distributed stream computing platform[C]//Fan W, Hsu W, Webb G I, et al. Proc. of the ICDM Workshops 2010. Sydney: IEEE Computer Society, 2010: 170-177.

[73] Afrati F N, Ullman J D. Optimizing joins in a Map Reduce environment[C]// Manolescu I, Spaccapietra S, Teubner J, et al. Proc. of the EDBT 2010. Lausanne: ACM Internatinoal Conf. Proc. Series, 2010: 99-110.

[74] Okcan A, Riedewald M. Processing theta-joins using MapReduce[C]//Sellis T K, Miller R J, Kementsietsidis A, et al. Proc. of the SIGMOD 2011. Athens: ACM Press, 2011: 949-960.

[75] Nykiel T, Potamias M, Mishra C, et al. MRShare: sharing across multiple queries in MapReduce[J]. Proc. of the VLDB Endowment, 2010, 3(1-2): 494-505.

[76] Roy I, Ramadan H E, Setty S T V, et al. Airavat: security and privacy for MapReduce[C]//Castro M, Snoeren A C. Proc. of the NSDI 2010. San Jose: USENIX Association, 2010: 297-312.

[77] Willis L, Jignesh M P. Energy management for MapReduce clusters[J]. Proc. of the VLDB Endowment, 2010, 3(1-2): 129-139.

[78] 2015年度大数据发展的十大趋势[J]. 中国有线电视, 2015(1): 100-101.
Trends of big data development in 2015[J]. China Digital Cable TV, 2015(1): 100-101.