大数据科学与工程的挑战与思考

马 帅 李建欣 胡春明
北京航空航天大学

关键词：互联网 大数据

引言

IBM前董事长兼首席执行官郭士纳（Louis Gerstner）认为，“计算模式每隔15年发生一次变革”：1965年前后出现大型机，1980年前后出现PC，1995年前后发生了互联网革命，2010年前后随着信息世界网络化、普适化、智能化，网络与传统技术交叉、融合，催生出云计算、物联网等新兴产业平台。虽然计算模式几经变迁，从单机到多机到协同等，围绕数据处理能力的研究应用一直都是IT发展的永恒主题。2007年美国总统科学技术顾问委员会（President's Council of Advisors on Science and Technology，PCAST）的报告以及英国e-Science计划前首席科学家托尼·海（Tony Hey）的著作《第四范式：数据密集型科学发现》（*The Fourth Paradigm: Data-intensive Scientific Discovery*）都揭示出数据分析已经成为继理论、实验和计算之后的第四种科学发现基础，成为产生经济价值的新源泉。它有助于分析社会学、市场预测以及医学等领域的规律和趋势，形成“真理尽在数据中”的效应。“数据科学”随之成为一个新兴的研究领域。

早在20世纪70年代，针对商业事务处理需求，数据处理的基础软件开始出现，而软件从过去的文件系统、操作系统、数据库系统，无一例外都是以更有效的数据处理为目标，来实现对数据更有效、更客观的分析和处理。然而，数据处理在科学计算、商业计算和社会计算等不同时期发生了本质变化。在科学计算时期，以科学数据的实时处理为主要目标，算法及算法复杂性是研究重点。在商业计算时期，以金融、电信等商业智能分析为主要目标，数据流程管理以及数据智能分析成为研究重点。如今，在社会计算时期，数据大规模、个性化和大众化特性显著，例如2011年Internet World统计互联网用户近20亿，社交网站Facebook活跃用户已突破8亿，其上市使风投获千倍回报，彻底改变了传统IT的应用模式。在社会计算时期，网络和应用升级推动数据量几何级数增长，数据变得愈为重要。继人力、资本之后，数据成为一种新的非物质生产要素，成为支撑科学研究和各类应用服务不可或缺的战略资源。社会计算进入了“大数据”时代。2012年3月29日，美国政府宣布了“大数据研究与发展计划”，初始启动经费2亿美元，其重要性堪比当年的“信息高速公路计划”，这标志着大数据已经上升到国家战略层面。

目前大家对大数据的基本特征还没有统一的认识和定义。一部分观点认为大数据只是对海量数据、数据规模等的描述和称谓，另一部分观点认为大数据就是指基于现有技术、方法和理论所无法处理的数据。然而，无论我们如何认识和描述大数据，科学、商业和社会中涌现的各类数据及其规模、处理能力，已经使得数据或信息从“匮乏”、“充足”进入到“无能为力”的时代。

本文以大数据科学与工程为切入点，以互联网网络化应用的大数据处理需求为核心，围绕大数据的三个关键问题，重点阐

述五个方面的研究：（1）海量异构数据模型理论与管理技术；（2）海量复杂数据智能分析理论与技术；（3）大数据分布式处理技术；（4）数据质量管理基础理论与技术；（5）大数据的安全与隐私保护。

大数据的三个关键问题

在“数据科学”领域，大数据管理及处理能力已经成为引领网络时代IT发展的关键。获取大量真实的运行数据并建立对其进行动态高效处理的能力，将成为产业竞争力的体现。在这样的背景下，社会计算引起的应用模式变革将深刻地影响或改变IT技术的研究理论和手段。在互联网数据为各领域应用带来新契机的同时，由于数据的异质异构、无结构及不可信等特征，互联网时代大数据的管理和分析研究需要解决可表示、可处理和可靠性三个关键问题。

可表示问题 当前互联网中的数据向着异质异构、无结构趋势发展。非结构化数据在互联网大数据中占有的比例大幅增加。美国弗雷斯特研究公司（Forrester）分析师在2010年《政府今天所面临的挑战》[46]报告中预计：“数据将会在今后的5年内增加8倍，其中非结构化数据在各组织机构的数据中所占份额超过70%到80%，并且这些非结构化数据的增长速度是结构化数据的10~50倍”。从数据管理的角度看，非结构化数据很难按照统一的模型进行分析处理，比结构化数据处理难得多。正是这些非结构化数据，使企业面对信息的快速增长猝不及防。因此，如何有效地表示这些非结构化数据成为首要问题。

可处理问题 如今数据规模急剧扩张，远远超越现有计算机处理能力。图灵奖获得者吉姆·格雷（Jim Gray）和IDC公司曾预测，全球数据量每18个月翻一番。目前全球数据的存储和处理能力已远落后于数据的增长幅度。例如，淘宝网每日新增的交易数据达10TB；eBay分析平台日处理数据量高达100PB，超过了美国纳斯达克交易所全天的数据处理量；沃尔玛是最早利用大数据分析并因此受益的企业之一，曾创造了“啤酒与尿布”的经典商业案例。现在沃尔玛每小时处理100万件交易，将有大约2.5PB的数据存入数据库，此数据量是美国国会图书馆的167倍；微软花了20年，耗费数百万美元完成的Office拼写检查功能，谷歌公司则利用大量统计数据直接分析实现。此外，在数据处理面临规模化挑战的同时，数据处理需求的多样化逐渐显现。相比支撑单业务类型的数据处理业务，公共数据处理平台需要处理的大数据涉及在线/离线、线性/非线性，流数据和图数据等多种复杂混合计算方式。例如，2011年Facebook首度公开其新数据处理分析平台PUMA，通过对数据多处理环节区分优化，相比之前单纯采用Hadoop和Hive进行处理的技术，数据分析周期从2天降到10秒之内，效率提高数万倍。因此，互联网数据规模的集聚使IT数据的处理能力成为保持企业核心竞争力的关键。大数据的高效处理已经成为一个核心问题，而数据处理在不同阶段形式不同。传统数学方法已无法适应不确定、动态大数据的分析，需要将计算科学与数学、物理等学科结合，建立一种新型数据科学方法，以便在数据多样性和不确定性前提下进行数据规律和统计特征的研究。

可靠性问题 由于互联网的开放性，使得大数据管理系统在数据输入时的质量确保和数据输出时的隐私保护面临考验。在传统数据库中假设数据是确定的，而互联网的数据采集和发布更灵活，容易将各种类型的不确定数据大量引入系统，造成数据中含有各种各样的错误和误差，体现为数据不正确、不精确、不完全、过时陈旧或者重复冗余。据高德纳公司（Gartner）统计，在全球财富1000强公司中有超过25%的公司关键数据不正确或不精确。在美国企业中有1%~30%的公司数据存在各类错误和误差，仅就医疗数据而言，有13.6%~81%的关键数据遗缺或陈旧。而数据是企业降低成本、损失和增加收入不可或缺的工具，例如英国BT公司

（British Telecom）因使用数据质量工具而创造的企业效益每年高达6亿英镑。同时，用户在享受数据价值的同时，也面临日益严重的安全威胁和隐私风险。趋势科技称2011年为数据泄露年，国内CSDN网站被曝600万用户的数据库信息数据保护不妥，导致用户密码泄露。据安全机构统计，此次隐私信息泄露涉及5000万互联网用户。而著名社会网络Facebook的Beacon广告系统可以追踪到5500万用户在其他网站的活动，严重威胁用户隐私信息。因此，大数据的可靠性已经成为一个重要问题。一方面通过数据清洗、去冗等技术提取有价值数据，实现数据质量高效管理；另一方面实现对数据的安全访问和隐私保护，两方面已成为大数据可靠性的关键需求。

因此，针对互联网大规模真实运行数据的高效处理和持续服务需求，以及出现的数据异质异构、无结构乃至不可信特征，数据的表示、处理和质量已经成为互联网环境中大数据管理和处理的三个重要问题。

海量异构数据模型与存储

在信息时代，图像、视频、音频、文本等异构数据每天都在以惊人的速度增长。不断膨胀的信息数据使系统资源消耗量日益增大，运行效率显著降低。海量异构数据资源规模巨大，新数据类型不断涌现，用户需求呈现出多样性。针对海量异构数据，如何构建一个模型来对其进行规范表达，如何基于该模型来实现数据融合，以及对其进行有效存储和高效查询是亟须解决的问题。

数据模型

现有的数据模型主要有关系模型、扩展关系模型、面向对象模型、E-R（Entity-Relation）模型以及分层式数据模型等。基于关系数据库，研究者提出用结构化的方法管理非结构化数据[1]，并采用关系模型表达非结构化数据的描述性信息[2]，但关系模型无法表达非结构化数据的复杂结构。扩展关系模型是在关系模型的二维表结构中，增加新的字段类型，表达非结构化数据信息。在多媒体数据库和空间数据库中，多采用面向对象模型。这种模型将具有相同静态结构、动态行为和约束条件的对象抽象为一类。各个类在继承关系下构成网络，使得整个面向对象的数据模型构成一个有向无环图。面向对象模型能够根据客观世界的本来面貌描述各种对象，能够表达对象间各种复杂关系。该模型存在的问题是缺乏坚实的理论基础，并且实现复杂。阿斯兰多根（Y. Alp Aslandogan）等人在SCORE系统中提出了用E-R方法表达图形数据的逻辑模型，西德特（Siadat）等人[3]和朱（Chu）等人[4]也提出了基于E-R方法的非结构化数据模型。在基于内容的多媒体数据检索系统中，马库斯（Marcus）等人[5]和阿玛托（Amato）等人[6]提出了基于语义描述、底层特征、原始数据的分层式数据模型。但是，这些模型不能很好地表达各类非结构化数据的各组成部分的关系以及各类数据之间的关系。

现有的非结构化数据管理技术，包括基于文本的信息检索系统、基于内容的信息检索系统和多媒体数据库系统，各自具有独立的数据表达方法与操作。在海量非结构化数据管理中，用户不仅希望使用基于文本和内容的信息检索，还需要进行数据分析、数据挖掘等一体化、智能化的数据处理。这就需要建立一种将非结构化数据的文本描述性信息与特征等信息整体表达，并且能够描述各种非结构化数据的统一数据模型。

数据存储

目前海量异构数据一般采用分布式存储技术。现有的分布式存储系统有美国麻省理工学院的CFS[7]、加州大学圣地亚哥分校的Total Recall[8]、谷歌文件系统（Google file system，GFS）[9]以及HDFS（Hadoop distributed file system）[10]。目前的存储架构仍不能解决数据的爆炸性增长带来的存储问题，静态的存储方案满足不了数据的动态演化所带来的挑战。因而在海量分布式存储和查询方面仍然需要进一步研究。

复杂数据智能分析技术

现在从海量的非结构化数据中归纳、过滤信息并依据这些信息进行快速、准确地决策已经成为用户最为迫切的需求。复杂数据的智能分析包括海量图数据的匹配分析和海量社交数据分析等。

图匹配查询

图的表达能力强，应用广泛，在社交网络、生物数据分析、推荐系统、复杂对象识别、软件代码剽窃检测等领域都起着重要的作用。图匹配的核心的关键问题是建立满足新型应用需求的图匹配理论和模型，并提供高效的匹配查询技术，以提高查询的效率和查询结果的准确性。大数据时代的图匹配理论和技术是目前国际上数据库领域的研究热点之一。

从查询语言的功能来看，图的查询语言可以分为两类。一类是Ad-hoc图查询语言，用以完成图中的某个单项查询任务。通常这类图的查询没有明确规定查询语言的语法，比如最短路径[11]、邻接查询[12]、可达性查询[13]、图同态及其扩展查询[14]、子图同构查询[15]、图模拟查询[16]及其扩展查询[17]等。另一类是通用图查询语言，可以完成多项查询任务，通常这类图的查询明确规定了查询语言的语法和表达能力，比如GraphQL[18]等。

通过拓展已有的图查询语言来设计新型的图查询语言是目前的一个研究热点，通过增强其表达能力，来适应新的应用需求[14]。新型语言的设计需要在图的表达能力和查询复杂性之间有一个权衡。另外，图表达和图划分是与大规模图的分布式查询密不可分的，前者可以提高单个计算节点的图存储能力，后者通过分布式多计算节点进一步提高整体的图存储能力和查询性能，但目前在这方面系统的研究工作还比较欠缺。

社会网络数据分析

社会网络分析包括社会网络结构分析[10]、信息传播方式分析[20]、社区群体结构分析（中心性分析、凝聚子群分析）[21]和用户间关系的预测[22]。随着微博客用户数量的激增以及人们获取信息方式的改变，最近针对社会媒体信息内容的研究也取得了一定的进展，其中包括重要舆情信息的发现[23]，以及从微博客数据中挖掘社会网络的结构，进而预测舆情在社会网络中的信息传播模式[24]。

融合社交网络数据的推荐系统也是目前的一个研究热点。为了给网络用户生成合适的推荐并保证推荐系统的性能，研究者提供了很多解决方案，比如基于协同过滤的推荐技术[25]、基于内容的推荐方法[26]和基于模型的推荐系统[27]。另外传统的数据挖掘技术，如聚类技术[28]和关联规则方法[29]也被应用到推荐系统中。目前这些传统的推荐算法最大的不足之处是没有考虑用户间的隐性相似度，仅通过内容或用户的评分历史等要素显示的相似度来计算用户间可能存在的共同兴趣或爱好。随着社会网络和社会媒体的发展，互联网用户更希望从在线好友或在线社区等社交网站中得到有关产品的评论。如何考虑信息在新兴的社会化媒体中的传播特性及社会网络拓扑对用户需求的影响，从而为用户提供更好的产品和内容推荐是一个亟待解决的问题。

大数据处理技术

由于海量数据的数据量和分布性的特点，使得传统的数据管理技术不适合处理海量数据。因此对海量数据的分布式并行处理技术提出了新的挑战，开始出现以MapReduce为代表的一系列研究工作。

数据并行处理

MapReduce[30]是2004年由谷歌公司提出的一个用来进行并行处理和生成大数据集的模型。Hadoop[31]是MapReduce的开源实现，是企业界、学术界共同关注的大数据处理技术。针对并行编程模型易用性，出现了多种大数据处理高级查询语言，如Facebook的Hive[32]、雅虎的Pig[33]、谷歌的Sawzall[34]等。这些高层查询语言通过解析器将查询语句解析

为一系列MapReduce作业，在分布式文件系统上执行[35]。与基本的MapReduce系统相比，高层查询语言更适于用户进行大规模数据的并行处理[36]。MapReduce及高级查询语言在应用中也暴露了在实时性和效率方面的不足，因此有很多研究针对它们进行优化。

MapReduce作为典型的离线计算框架，无法满足许多在线实时计算需求。目前在线计算主要基于两种模式研究大数据处理问题：一种基于关系型数据库，研究提高其扩展性，增加查询通量来满足大规模数据处理需求；另一种基于新兴的NoSQL数据库，通过提高其查询能力丰富查询功能来满足有大数据处理需求的应用。使用关系型数据库为底层存储引擎上层对主键空间进行切片划分，数据库全局采用统一的哈希方式将请求分发到不同的存储节点以达到可以水平扩展的要求，这种方案一般不能对上层提供原存储引擎的全部查询能力。Oracle NoSQL DB、MySQL Cluster、MyFOX即是典型系统，通过扩展NoSQL数据库的查询能力的方法来满足大规模数据处理需求的最典型的例子就是谷歌的BigTable及其一系例扩展系统。

如何处理分布式的海量复杂数据也是目前的研究热点。MapReduce的设计初衷是分析Web Graph，但处理图数据常常需要大量的迭代运算，而MapReduce不擅长处理这类复杂数据，已有的并行图算法库Parallel BGL或CGMgraph又没有提供容错功能。于是谷歌公司开发了Pregel[37]，可以在通用分布式服务器上处理PB级别图数据，与之对应的开源项目Giraph[38]也得到了学术界的关注。

增量处理技术

如何设计高效的增量算法，进行分布式大数据的动态更新也是目前的研究热点。谷歌公司已经采用增量索引过滤器（Percolator for incremental indexing）而不是MapReduce来分析频繁变化的数据集，使搜索结果返回速度越来越接近实时。Percolator通过只处理新增的、改动过的或删除的文档和使用二级指数来高效率地创建目录，并返回查询结果。Percolator将文档处理延迟缩短了99%[39]，其索引万维网新内容的速度比MapReduce快很多。

数据质量基础理论与关键技术

数据的价值涉及很多因素，数据质量是决定数据价值的关键因素之一。数据质量管理的研究旨在建立识别数据错误的理论和模型，提供自动发现和定位数据错误的方法，设计高效修复错误数据的关键技术，最终达到提高数据可用性的目的。数据质量管理研究与传统的数据管理研究具有本质区别。数据管理研究专注于管理数据的“量”，即如何快速集成、存储、查询大量数据。然而，由于这些系统中存储的数据具有各种各样的错误和误差，无论这些系统和引擎的速度多么快，处理的数据量多么大，都无法为用户提供正确的信息。数据质量的研究则侧重于管理数据的“质”，其目的是提供完整的理论体系和数据质量管理系统，自动发现和更正数据中的错误，保证数据和查询结果的正确性，从而提高数据的可用性。现实世界对数据质量管理的需求给数据质量的研究带来了诸多挑战。

统一的逻辑框架

当前学术界大多将上述课题视为单独的学术问题，如美国国家科学基金会和欧盟资助的3～5年的项目大多只关注于其中一个课题。然而在现实中，一个数据集可能同时包含各类错误，而且这些关键问题可能会相互影响。例如，部分数据可能不正确、不完全、过于陈旧或含有冗余。提高数据的正确性可帮助识别实体。反过来，有效识别实体能够帮助提高数据的精确度、完全性和时效性。这意味着我们不能仅关注解决某一个关键问题，而必须同时考虑如何保证数据的正确性、精确性、完全性、时效性和无冗余。因此，提出一个统一的逻辑框架，并在此框架下研究上述五个核心问题的相互影响，是数据质量理论和技术研究面临的巨大挑战。这也加大了数据规则描述定义、发掘、推理以及数据

错误检错和纠错算法的难度。

半结构化数据数据质量

现实中的数据不只限于传统的关系数据，更经常以半结构化形式出现，如XML（extensible markup language）或具有图结构的数据。当前XML已成为现今数据交换和合成的标准模式，而具有图结构的数据在生物、社会网络、交通网络等领域应用很广泛。如上所述，这些问题对于传统的关系数据而言已非易事，对半结构化而言，则更具挑战性。因此，我们需要针对半结构化数据进行研究。

分布式数据清洗

在实际应用中，数据往往被划分为若干片段并分布存储于不同的网络站点上。因此，我们不仅需要针对集中存储的数据质量进行研究，还需要对分布式存储的数据进行同样的研究。目前已有的研究主要针对集中存储的数据，几乎未涉及分布式存储数据的质量问题。研究分布数据的质量问题更困难。例如，检测分布数据中的错误需要将一些数据传输到其他站点，保障数据传输量最小化的数据错误检测问题在分布式环境下成为NP完全问题[40]，这就需要为分布式数据开发全新的错误检测算法和数据修复算法。因此，如何在分布网络环境下研究分布式存储数据的质量问题就成了一个极具挑战性的问题。

大数据安全与隐私保护

数据安全是互联网中大数据管理的重要组成部分。然而随着互联网规模不断扩大，数据和应用呈现出指数级增长趋势，给动态数据安全监控和隐私保护带来了极大的挑战。

文件的安全性

文件是数据处理和运行的核心，当前很多用户文件在第三方的运行平台中存储和进行处理，这些数据文件往往包含很多企业或个人的敏感信息，其安全性和隐私性自然成为一个需要重点关注的问题。目前，文件保护提供了对文件的访问控制和授权。例如Linux自带的文件访问控制机制，通过文件访问控制列表来限制程序对文件的操作。然而大部分文件保护机制都存在一定程度的安全问题。它们通常使用操作系统的功能来实现完整性验证机制，因此只依赖于操作系统本身的安全性。现代操作系统由于过于庞大，不可避免地存在安全漏洞，其本身的安全性都难以保证。基于主机的文件完整性保护方法将自身暴露在客户机操作系统内，隔离能力差，恶意代码可以轻易发现检测系统并设法绕过检测对系统进行攻击。例如Tripwire[41]，它本身是用户级应用程序，很容易被恶意软件篡改和绕过。

动态数据安全监控

对数据处理平台运行态数据（如内存数据、进程等）的安全监控与检测是保证数据安全性的重要环节，也是保证分布式计算系统健康运行的关键。在这方面的研究工作非常活跃。从操作系统层次看，包括内存、磁盘以及网络I/O数据的全面监控检测；从应用层次看，包括对进程、文件以及网络连接的安全监控。例如，IBM公司研发的IMA框架[42]。由于虚拟机技术具有隔离性、安全性等特点，也被用于验证和保护上层应用程序乃至操作系统的完整性。因此在存在规模化的分布式计算平台之中，如何有效通过动态数据的细粒度安全监控和分析，对大数据分布式处理平台可靠运行是一个重要需求。

数据的隐私保护

与私密性等传统安全属性不同，海量数据处理中的隐私性主要体现在如何在不暴露用户敏感信息的前提下进行数据挖掘。数据隐私性技术最早在统计数据库数据研究中得到关注，近年来逐渐成为数据管理、数据挖掘等领域领域的研究热点。保护隐私的数据挖掘（privacy preserving data mining）这一术语首先是由阿格拉瓦尔（Agrawal）和林德尔（Lindell）等人分别在文献[43]和文献[44]中提出的。大部分PPDP文献假设攻击者背景知识有限，攻击者可以把从外部表

得到的个人记录与发布数据表的记录、敏感属性或者数据表本身链接。

数据匿名性、关联性一直作为数据隐私分析的重要概念，可对数据隐私程度进行度量。例如，k-匿名性（k-anonymity）、l-多样性（l-Diversity）、t-Closeness 和FF-Anonymity 等概念和方法的针对不同需求相继出现。2006 年 Dwork 证明了在具有背景知识情况下，是不可能实现该定义要求的隐私的，并提出一种量化的隐私描述－称为差分隐私（Differential Privacy）[45]。差分隐私是针对概率攻击原理提出的，可以较准确度量数据发布前和后信息量的变化，但过于抽象很难用于设计隐私保护机制，而且在用于具有复杂关联性数据，2011年丹尼尔（Daniel）指出差分隐私会遇到失效问题。尽管如此，差分隐私概念是得到较为广泛的认可，多个研究者对其进行改进应用，例如，被用于空间数据发布的隐私度量，分布式数据查询的隐私泄露的分布式隐私模型，最近李凝晖（Ninghui Li）等则将差分隐私与k-匿名技术结合。

随着互联网技术和大数据应用的发展，隐私的研究问题和概念在不断发展。第一，随着分布式计算广泛应用，如何在实现多点高效协同工作的同时，保证在频繁的信息交互、数据传输过程中，不会给隐私信息、敏感数据带来威胁？如何在保护各独立站点私有隐私的同时，实现对整个分布式系统的隐私的共同保护？如何在确保隐私保护策略或算法有效的同时，对分布式查询、存储以及网络拓扑结构的负面影响尽量小？第二，现有隐私保护技术主要基于静态数据集，而现实世界中,数据库中的数据却是无时无刻不在变化,包括数据表现形式的改变、属性的增减、新数据的加入、旧数据的删除等。并且，数据库数据的这种变化，一般都不是完全随机、独立的。数据与数据之间，数据与数据变化之间都是相互关联的。因此，怎样在这种更加复杂的环境下同时实现对动态数据的利用和隐私保护将更具挑战。第三，在社会网络的隐私保护技术近年也逐渐得到关注，大部分现有隐私保护模型和算法都是针对传统的关系型数据，不能将其直接移植到这些应用中。原因在于：攻击者的背景知识更加复杂也更难模拟；不能通过简单地对比匿名前后的网络进行信息缺损判断，社会网络数据的发布信息缺损度量标准复杂；由于背景知识和度量标准的复杂性使得设计社会网络数据的匿名处理方法更具有难度和挑战性。

结语

本文介绍了互联网时代大数据的管理和分析研究需要解决的可表示、可处理和可靠性三个关键问题。以大数据科学与工程为切入点，以互联网网络化应用的大数据处理需求为核心，重点阐述五个方面的研究。

在“数据科学”领域，大数据管理及处理能力已经成为引领网络时代IT发展的核心。我们有充分的理由相信大数据管理和分析将成为与国计民生紧密相关的研究领域。■

马　帅

CCF会员。北京航空航天大学教授。主要研究方向为数据库理论与系统、图匹配和数据质量等。

mashuai@buaa.edu.cn

李建欣

CCF会员。北京航空航天大学副教授。主要研究方向为分布式系统、信息安全等。

lijx@buaa.edu.cn

胡春明

CCF会员、YOCSEF学术秘书。北京航空航天大学副教授。主要研究方向为分布式系统等。

hucm@buaa.edu.cn

参考文献

[1] Doan A, Naughton J F, Baid A, et al. The case for a structured approach to managing unstructured data. CIDR 2009

[2] Srivastava D, Velegrakis Y. Intentional associations between data and metadata. SIGMOD 2007

[3] Siadat M, Soltanian-Zadeh H, Fotoub I F, et al. Data modeling for content-based support environment (C-BASE): application on epilepsy data mining. ICDM 2007
[4] Chu E, Baid A, Chen T, et al. A relational approach to incrementally extracting and querying structure in unstructured data. VLDB 2007
[5] Marcus S, Subrahmanian V S. Foundations of multimedia database systems. JACM, 1996, 43: 474~523
[6] Amato G, Mainetto G, Savino P. An approach to a content-based retrieval of multimedia data. Multimed Tools Appl, 1998, 7: 9~36
[7] Dabek F, Kaashoek M,Karger D,et al.Wide-Area cooperative storage with CFS.SOSP 2001
[8] Bhagwan R,Tati K,Cheng Y,et al.Total recall:System support for automatedavailability management.In:Proc.of the 1st ACM/Usenix Symp.on NetworkedSystems Design and Implementation, 2004.
[9] Ghemawat,Gobioff,Leung,The Google file System.In Proc of ACM SOSP,2003
[10]The Apache Software Foundation,Apache Software Foundation. Storage Networking Industry Association and the OpenGrid ForumCloud Storage for Cloud Computing[EB/OL]
[11]G. Ramalingam and Thomas Reps, An incremental algorithm for a generalization of the shortest-path problem. Journal of Algorithms, 1996
[12]Hossein Maserrat and Jian Pei, Neighbor query friendly compression of social networks. KDD 2010
[13]Elmagarmid, Ahmed K. and Aggarwal, Charu C. and Wang, Haixun, Managing and Mining Graph Data, Advances in Database Systems, Springer 2010
[14]Wenfei, Jianzhong Li, Shuai Ma, Hongzhi Wang and Yinghui Wu, Graph Homomorphism Revisited for Graph Matching. PVLDB 2010
[15]Ullmann, J. R., An Algorithm for Subgraph Isomorphism. Journal of ACM, 1976
[16]M. R. Henzinger, T. Henzinger and P. Kopke, Computing simulations on finite and infinite graphs. FOCS 1995
[17]Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Yinghui Wu, Adding Regular Expressions to Graph Reachability and Pattern Queries. ICDE 2011
[18]H. He and A. K. Singh. Graphs-at-a-time: query language and access methods for graph databases. SIGMOD 2009
[19]J. Binder, A. Howes, and A. Sutcliffe The problem of conflicting social spheres: effects of network structure on experienced tension in social network sites. CHI 2009
[20]M. Cha, A. Mislove, and K. P. Gummadi.A measurement-driven analysis of information propagation in the flickr social network. WWW 2009
[21]M. J. Dombroski, K. M. Carley. NETEST: estimating a terrorist network's structure. Computational and Mathematical Organization Theory, 2002: 235~241
[22]V. Leroy , B. B. Cambazoglu , F. Bonchi, Cold start link prediction, KDD 2010
[23]J. Allan, A. Feng, and A. Bolivar, Flexible intrinsic evaluation of hierarchical clustering for TDT. CIKM 2003
[24]M. Jamali and H. Abolhassani. Different aspects of social network analysis. Web Intelligence 2006
[25]Z. Huang, D. Zeng, H. Chen. A comparison of collaborative filtering recommendation algorithms for e-commerce. IEEE Intelligent Systems, 2007(22):68~78
[26]A. Popescul, L. H. Ungar, D. M. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. UAI 2001
[27]J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. Madison, Wisconsin, USA, 1998.Morgan Kaufmann, 43~52
[28]N. J. Nnadi. Applying relevant set correlation clustering to multi-criteria recommender systems. RecSYS 2009
[29]W. Lin, S. A. Alvarez, and C. Ruiz. Efficient adaptive-support association rule mining for recommender systems. DMKD, 6(1), 2002:83~105
[30]J.Dean and S.Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. CACM,51(1), 2008: 107~113
[31]Apache. Hadoop: Open-source implementation of MapReduce. http://hadoop.apache.org
[32]Apache. Hive: http://hive.apache.org/
[33]Apache. Pig: http://pig.apache.org/
[34]R.Pike et al. Interperting the Data: Parallel Analysis with Sawzall. Scientific Programming Journal, 13(4), 2005: 227~298
[35]M. Isard and Y. Yu. Distributed data-parallel computing using a high-level programming language. SIGMOD 2009
[36]Biswapesh , Liang Lin, et al., Tenzing A SQL Implementation On The MapReduce Framework, VLDB 2011
[37]Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser and Grzegorz Czajkowski, Pregel: a system for large-scale graph processing. SIGMOD 2010
[38]Apache. Giraph http://giraph.apache.org/
[39]Daniel Peng and Frank Dabek, Large-scale Incremental Processing Using Distributed Transactions and Notifications, OSDI 2010

[40]Wenfei Fan, Floris Geerts, Shuai Ma, and Heiko Müller, Detecting inconsistencies in distributed data, ICDE 2010
[41]G. Kim and E. Spafford. The Design and Implementation of Tripwire: A File System Integrity Checker, CCS 1994
[42]R. Sailer, X. Zhang, T. Jaeger, and L. van Doorn. Design and implementation of a tcg-based integrity measurement architecture, USENIX Security Symposium. 2004
[43]Rakesh Agrawal , Ramakrishnan Srikant, Privacy-preserving data mining, SIGMOD 2000
[44]Y.Lindell and B.Pinkas. Privacy preserving data mining. CRYPTO 2000
[45]Cynthia Dwork. Differential Privacy, International Colloquium on Automata, Languages and Programming (ICALP), Lecture Notes in Computer Science, 2006, Volume 4052/2006, 1~12
[46]Today's Challenge in Government: What to do with Unstructured Information and Why Doing Nothing Isn't An Option, Noel Yuhanna, Principal Analyst, Forrester Research, Nov. 2010