计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
50(Suppl.): 216-233, 2013

大数据应用的现状与展望

张 引 陈 敏 廖小飞
(华中科技大学计算机科学与技术学院 武汉 430074)
(yinzhang.cs@gmail.com)

Big Data Applications: A Survey

Zhang Yin, Chen Min, and Liao Xiaofei
(*School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan* 430074)

Abstract Characteristics of big data can be summarized as four Vs, i.e. volume (great volume), variety (various modalities), velocity (rapid generation), and value (huge value but very low density). Big data application can provide useful values, judgments, suggestions, supports or decisions. In this paper, we introduce the methods, architecture and tools for big data analysis. We then examine six most correlated data analysis fields, including structured data analysis, text analysis, website analysis, multimedia analysis, network analysis, and mobile analysis. Finally, we summarize the research hotspots and possible research directions of big data. We also discuss potential development trends of big data.

Key words big data; data analysis; data mining; unstructured data; internet of things; social network data; crowdsourcing

摘 要 大数据具有规模大、种类多、生成速度快、价值巨大但密度低的特点.大数据应用就是利用数据分析的方法,从大数据中挖掘有效信息,为用户提供辅助决策,实现大数据价值的过程.主要介绍了大数据分析方法、分析模式以及常用的分析工具,将大数据应用归纳为6个关键领域——结构化数据分析、文本分析、Web分析、多媒体分析、社交网络分析和移动分析,并列举了6个大数据的典型应用.最后,从基础理论、关键技术、应用实践以及数据安全4个方面总结了大数据的研究现状,并对大数据应用未来的研究进行展望.

关键词 大数据;数据分析;数据挖掘;非结构化数据;物联网;社交网络数据;众包

中图法分类号 TP311

在过去的20年中,各个领域都出现了大规模的数据增长,包括医疗保健和科学传感器、用户生成数据、互联网和金融公司、供应链系统等.国际数据公司(IDC)报告称[1],2011年全球被创建和复制的数据总量为1.8 ZB(1 ZB≈10^{21} B),在短短5年间增长了近9倍,而且预计这一数字将每两年至少翻一番.大数据这一术语正是产生在全球数据爆炸增长的背景下,用来形容庞大的数据集合.与传统的数据集合相比,大数据通常包含大量的非结构化数据,且大数据需要更多的实时分析.此外,大数据还为挖掘隐藏的价值带来了新的机遇,同时给我们带来了新的挑战,即如何有效地组织管理这些数据.如今,工业界、研究界甚至政府部门都对大数据这一研究领域产生了巨大的兴趣.例如,我们经常在公共媒体领域听到

收稿日期:2013-11-26
基金项目:国家自然科学基金项目(61300224)

大数据这一话题,包括《经济学人》[2-3]、《纽约时报》[4]、《全国公共广播电台》[5-6].《自然》和《科学》杂志也分别开放了特殊专栏,来讨论大数据带来的挑战和重要性[7-8].政府机构最近也宣布了一项加快大数据进程的重大计划[9],各行各业也都在积极讨论大数据的吸引力[10].

随着网络的快速发展,索引和查询的内容也在迅速增加,大数据给搜索公司带来了巨大的挑战.谷歌创建了谷歌文件系统(GFS)[11]和 MapReduce 编程模型[12]来应对网络规模的数据管理和分析所带来的挑战.此外,用户生成数据、各种传感器和其他的数据源也助长了这种势不可挡的数据流,这就需要对计算架构和大规模数据处理机制进行一次根本的转变.2007 年 1 月,吉姆·格雷(Jim Gray)——数据库软件的先驱,将这种转变称为"第四范式"[13](表 1 所示为科学发现的 4 种范式).他还认为,应对这种范式的唯一方法就是开发新一代的计算工具,以对海量数据进行管理、可视化和分析.2011 年 6 月,EMC/IDC 发表了一篇题为"从混沌中提取价值"的研究报告[1],首次对大数据的概念和其潜在性进行了探讨.

表 1　科学发现的 4 种范式

科学范式	时间	方法
实证	一千多年前	描述自然现象
理论	过去数百年	使用模型、概括
计算	过去几十年	模拟复杂的现象
数据探索(eScience)	如今	使用工具采集数据,使用模拟器生成数据;使用软件处理数据;利用计算机存储信息;分析数据

1　相关研究

1.1　大数据的定义

目前,虽然大数据的重要性得到了大家的一致认同,但是关于大数据的定义却众说纷纭.大数据是一个抽象的概念,除去数据量庞大,大数据还有一些其他的特征,这些特征决定了大数据与"海量数据"和"非常大的数据"这些概念之间的不同.一般意义上,大数据是指无法在有限时间内用传统 IT 技术和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合.科技企业、研究学者、数据分析师和技术顾问们,由于各自的关注点不同,对于大数据有着不同的定义.通过以下定义,或许可以帮助我们更好地理解大数据在社会、经济和技术等方面的深刻内涵.

2010 年 Apache Hadoop 组织将大数据定义为,"普通的计算机软件无法在可接受的时间范围内捕捉、管理、处理的规模庞大的数据集".在此定义的基础上,2011 年 5 月,全球著名咨询机构麦肯锡公司发布了"大数据:下一个创新、竞争和生产力的前沿",在报告中对大数据的定义进行了扩充.大数据是指其大小超出了典型数据库软件的采集、存储、管理和分析等能力的数据集.该定义有两方面内涵:1)符合大数据标准的数据集大小是变化的,会随着时间推移、技术进步而增长;2)不同部门符合大数据标准的数据集大小会存在差别.目前,大数据的一般范围是从几个 TB 到数个 PB(数千 TB)[10].根据麦肯锡的定义可以看出,数据集的大小并不是大数据的唯一标准,数据规模不断增长,以及无法依靠传统的数据库技术进行管理,也是大数据的两个重要特征.

其实,早在 2001 年,就出现了关于大数据的定义.META 集团(现为 Gartner)的分析师道格·莱尼(Doug Laney)在研究报告中,将数据增长带来的挑战和机遇定义为三维式,即数量(Volume)、速度(Velocity)和种类(Variety)的增加[14].虽然这一描述最先并不是用来定义大数据的,但是 Gartner 和许多企业,其中包括 IBM[15]和微软[16],在此后的 10 年间仍然使用这个"3Vs"模型来描述大数据[17].数量,意味着生成和收集大量的数据,数据规模日趋庞大;速度,是指大数据的时效性,数据的采集和分析等过程必须迅速及时,从而最大化地利用大数据的商业价值;种类,表示数据的类型繁多,不仅包含传统的结构化数据,更多的则是音频、视频、网页、文本等半结构和非结构化数据.

但是,也有一些不同的意见,大数据及其研究领域极具影响力的领导者的国际数据公司(IDC)就是其中之一.2011 年,在该公司发布的报告中(由 EMC 主办)[1],大数据被定义为:"大数据技术描述了新一代的技术和架构体系,通过高速采集、发现或分析,提取各种各样的大量数据的经济价值."从这一定义来看,大数据的特点可以总结为 4 个 V,即 volume(体量浩大)、variety(模态繁多)、velocity(生成快速)和 value(价值巨大但密度很低),如图 1 所示.这种 4Vs 定义得到了广泛的认同,3Vs 是一种较为专业化的定义,而 4Vs 则指出大数据的意义和必要性,即挖掘蕴藏其中的巨大价值.这种定义指出大数据最为核心的问题,就是如何从规模巨大、种类繁多、生成快速的数据集中挖掘价值.正如 Facebook 的副总工程师杰伊·帕瑞克所言,"如果不利用所收

集的数据,那么你所拥有的只是一堆数据,而不是大数据"[18].

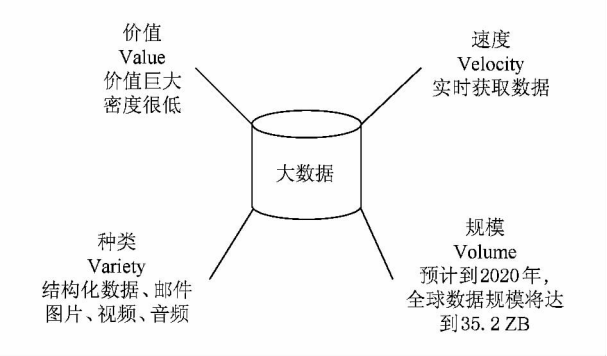

图 1 大数据的 4Vs 特点大数据分析

此外,美国国家标准和技术研究院(NIST)也对大数据做出了定义:"大数据是指其数据量、采集速度,或数据表示限制了使用传统关系型方法进行有效分析的能力,或需要使用重要的水平缩放技术来实现高效处理的数据."这是从学术角度对大数据的概括,除了 4Vs 定义所提及的概念,还特别指出需要高效的方法或技术对大数据进行分析处理.

就大数据究竟该如何定义,工业界和学术界已经进行了不少讨论[19-20].但是,大数据的关键并不在于如何定义,或如何去界定大数据,而应该是如何提取数据的价值,如何利用数据,如何将"一堆数据"变为"大数据".

我们认为大数据价值链可分为 4 个阶段:数据生成、数据采集、数据储存以及数据分析.数据分析是大数据价值链的最后也是最重要的阶段,是大数据价值的实现,是大数据应用的基础,其目的在于提取有用的值,提供论断建议或支持决策,通过对不同领域数据集的分析可能会产生不同级别的潜在价值[10].

1.2 传统数据分析方法

传统数据分析是指用适当的统计方法对收集来的大量第 1 手资料和第 2 手资料进行分析,把隐没在一大批看来杂乱无章的数据中的信息集中、萃取和提炼出来,找出所研究对象的内在规律,以求最大化地开发数据资料的功能,发挥数据的作用.数据分析对国家制定发展计划,对企业了解客户需求、把握市场动向都有巨大的指导作用.大数据分析可以视为对一种特殊数据的分析,因此很多传统的数据分析方法也可用于大数据分析.以下是可用于大数据分析的传统数据分析方法,这些方法源于统计学和计算机科学等多个学科.

1) 聚类分析[21].聚类分析是划分对象的统计学方法,指把具有某种相似特征的物体或者事物归为一类.聚类分析的目的在于辨别在某些特性上相似(但是预先未知)的事物,并按这些特性将样本划分成若干类(群),使在同一类内的事物具有高度的同质性,而不同类的事物则有高度的异质性.聚类分析是一种没有使用训练数据的无监督式学习.

2) 因子分析[22].因子分析的基本目的就是用少数几个因子去描述许多指标或因素之间的联系,即将相互比较密切的几个变量归在同一类中,每一类变量就成为一个因子(之所以称其为因子,是因为它是不可观测的,即不是具体的变量),以较少的几个因子反映原数据的大部分信息.

3) 相关分析[23].相关分析法是测定事物之间相关关系的规律性,并据以进行预测和控制的分析方法.社会经济形象之间存在着大量的相互联系、相互依赖、相互制约的数量关系.这种关系可分为两种类型.一类是函数关系,它反映着现象之间严格的依存关系,也称确定性的依存关系.在这种关系中,对于变量的每一个数值,都有一个或几个确定的值与之对应.另一类为相关关系,在这种关系中,变量之间存在着不确定、不严格的依存关系,对于变量的某个数值,可以有另一变量的若干数值与之相对应,这若干个数值围绕着它们的平均数呈现出有规律的波动.

4) 回归分析[24].回归分析是研究一个变量与其他若干变量之间相关关系的一种数学工具,它是在一组实验或观测数据的基础上,寻找被随机性掩盖了的变量之间的依存关系.通过回归分析,可以把变量间的复杂的、不确定的关系变得简单化、有规律化.

5) A/B 测试[10].也称为水桶测试,通过对比测试群体,确定哪种方案能提高目标变量的技术.大数据可以使大量的测试被执行和分析,保证这个群体有足够的规模来检测控制组和治疗组之间有意义的区别.

6) 数据挖掘[25].更为深入的数据分析就需要利用到数据挖掘技术,实现一些高级别的数据分析需求.数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的数据中,提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程.数据挖掘主要用于完成以下 6 种不同任务,同时也对应着不同的分析方法:分类(classification)、估值(estimation)、预言(prediction)、相关性分组或关联规则(affinity grouping or association rules)、聚集(clustering)、描述和可视化(description and

visualization).挖掘方法分大致分为为:机器学习方法、神经网络方法和数据库方法.机器学习可细分为:归纳学习方法、基于范例学习、遗传算法等.神经网络方法可细分为:前向神经网络、自组织神经网络等.数据库方法主要是多维数据分析或联机分析处理(on-line analytical processing, OLAP)方法,另外还有面向属性的归纳方法.

虽然这些传统的分析方法已经被应用于大数据领域,但是它们在处理规模较大的数据集合时,效率无法达到用户预期,且难以处理复杂的数据,如非结构化数据.因此,出现了许多专门针对大数据的集成、管理及分析的技术和方法.

1.3　大数据分析方法

随着大数据时代的到来,如何快速地从这些海量数据中抽取出关键的信息,为企业和个人带来价值,是各界关注的焦点.目前一些大数据具体处理方法主要有:

1) Bloom Filter:布隆过滤器,其实质是一个位数组和一系列 Hash 函数.布隆过滤器的原理是利用位数组存储数据的 Hash 值而不是数据本身,其本质是利用 Hash 函数对数据进行有损压缩存储的位图索引.其优点是具有较高的空间效率和查询速率,缺点是有一定的误识别率和删除困难.布隆过滤器适用于允许低误识别率的大数据场合.

2) Hashing:散列法,也叫做 Hash 法,其本质是将数据转化为长度更短的定长的数值或索引值的方法.这种方法的优点是具有快速的读写和查询速度,缺点是难以找到一个良好的 Hash 函数.

3) 索引:无论是在管理结构化数据的传统关系数据库,还是管理半结构化和非结构化数据的技术中,索引都是一个减少磁盘读写开销、提高增删改查速率的有效方法.索引的缺陷在于需要额外的开销存储索引文件,且需要根据数据的更新而动态维护.

4) Trie 树:又称为字典树,是 Hash 树的变种形式,多被用于快速检索,和词频统计.Trie 树的思想是利用字符串的公共前缀,最大限度地减少字符串的比较,提高查询效率.

5) 并行计算:相对于传统的串行计算,并行计算是指同时使用多个计算资源完成运算.其基本思想是将问题进行分解,由若干个独立的处理器完成各自的任务,以达到协同处理的目的.目前,比较典型的并行计算模型有 MPI(message passing interface), MapReduce, Dryad 等.

传统数据分析方法,大多数都是通过对原始数据集进行抽样或者过滤,然后对数据样本进行分析,寻找特征和规律,其最大的特点是通过复杂的算法从有限的样本空间中获取尽可能多的信息.随着计算能力和存储能力的提升,大数据分析方法与传统分析方法的最大区别在于分析的对象是全体数据,而不是数据样本,其最大的特点在于不追求算法的复杂性和精确性,而追求可以高效地对整个数据集的分析.总之,传统数据方法力求通过复杂算法从有限的数据集中获取信息,其更加追求准确性;大数据分析方法则是通过高效的算法、模式,对全体数据进行分析.

1.4　大数据分析模式

由于大数据来源广泛,种类繁多,结构多样且应用于众多不同领域,所以针对不同业务需求的大数据,应采用不同的分析模式.

1) 根据实时性,可分为实时分析和离线分析

实时分析,多用于电子商务、金融等领域.由于数据瞬息万变,因此需要及时的数据分析,在极短的时间能返回分析结果.目前,实时分析的主要模式是采用传统关系型数据库组成并行处理集群,并大多采用了内存计算平台.EMC 的 Greenplum[26]、SAP 的 HANA[27]等,都是进行实时分析的工具.

离线分析,往往用于对结果反馈时间要求不高的场合,比如机器学习、统计分析、推荐算法等.离线分析一般是通过数据采集工具将日志大数据导入专用的平台进行分析.在大数据环境下,为了降低数据格式转化的开销,提高数据采集的效率,很多互联网企业都采用基于 Hadoop 的离线分析模式.例如, Facebook 开源的 Scribe[28]、LinkedIn 开源的 Kafka[29]、淘宝开源的 Timetunnel[30]、Hadoop 的 Chukwa[31]等,均可以满足每秒数百 MB 的日志数据采集和传输需求.

2) 根据数据规模,可分为内存级、BI 级和海量级

内存级分析,是指数据总量不超过集群内存的最大值.目前的服务器集群的内存超过几百 GB,甚至达到 TB 级别都是很常见的,因此可以采用内存数据库技术,将热点数据常驻内存,从而达到提高分析效率的目的.内存级分析非常适用于实时分析业务.目前,MongoDB 是比较有代表性的内存级分析模式.随着固态硬盘(solid-state drive, SSD)的发展,内存级数据分析的能力和性能将会得到进一步的提升,其应用也越来越广泛.

BI 级分析,是指数据规模超出了内存级,但是又可以导入 BI 分析环境下进行分析,目前主流的 BI 产品都有支持 TB 级以上的数据分析方案.

海量级分析,是指数据规模已经完全超出 BI 产品以及传统关系型数据库的能力.目前,大多数的海量级分析都是采用 Hadoop 的 HDFS 分布式文件系统来存储数据,并使用 MapReduce 进行分析.海量级分析基本也都属于离线分析.

3) 根据算法复杂度的分类

根据业务数据和业务需求的不同,数据分析算法的时空复杂度也有巨大的差异性.例如,针对易并行问题,可以设计分布式算法,采用并行处理的模型进行分析.

1.5 大数据分析工具

目前,在众多可用于大数据分析的工具中,既有专业的也有非专业的工具,既有昂贵的商业软件也有免费的开源软件.根据 2012 年,KDNuggets 针对 798 名专业人员,做了一份"过去一年中在实际项目中所用到的大数据、数据挖掘、数据分析软件"[32]的调查结果,本文选取使用频率最高的前 5 名进行简单介绍:

1) R(30.7%)

R 是开源编程语言和软件环境,被设计用来进行数据挖掘/分析和可视化.在执行计算密集型任务时,在 R 环境中还可以调用 C,C++和 Fortran 编写的代码.此外,专业用户还可以通过 C 语言直接调用 R 对象.R 语言是 S 语言的一种实现.而 S 语言是由 AT&T 贝尔实验室开发的一种用来进行数据探索、统计分析、作图的解释型语言.最初 S 语言的实现版本主要是 S-PLUS.但 S-PLUS 是一个商业软件,相比之下开源的 R 语言更受欢迎.R 不仅在软件类中名列第一,在 2012 年 KDNuggets 的另一份调查"过去一年中在数据挖掘/分析中所使用的设计语言"中,R 语言击败了 SQL 和 Java,同样荣登榜首.在 R 语言盛行的大环境下,各大数据库厂商如 Teradata 和 Oracle,都发布了与 R 语言相关的产品.

2) Excel(29.8%)

Excel 是微软的 Office 办公软件的核心组件之一,提供了强大的数据处理、统计分析和辅助决策等功能.在安装 Excel 的时候,一些具有强大功能的分析数据的扩展插件也被集成了,但是这些插件需要用户的启用才能被使用,这其中就包含了分析工具库(Anlyasis ToolPak)和规划求解向导项(Solver Add-in)等插件.Excel 也是前 5 名中唯一的商业软件,其他软件都是开源的.

3) Rapid-I Rapidminer(26.7%)

Rapidminer 是用于数据挖掘、机器学习、预测分析的开源软件,在 2011 年 KDnuggets 的调查中,它比 R 的使用率还高,位于第一位.RapidMiner 提供的数据挖掘和机器学习程序包括:数据加载和转换(ETL)、数据预处理和可视化、建模、评估和部署.数据挖掘的流程是以 XML 文件加以描述,并通过一个图形用户界面显示出来.RapidMiner 是由 Java 编程语言编写的,其中还集成了 Weka 的学习器和评估方法,并可以与 R 语言进行协同工作.Rapidminer 中的功能均是通过连接各类算子(operataor)形成流程(process)来实现的,整个流程可以看做是工厂车间的生产线,输入原始数据,输入出模型结果.算子可以看作是执行某种具体功能的函数,不同算子有不同的输入输出特性.

4) KNMINE(21.8%)

KNIME(konstanz information miner)是一个用户友好、智能的、并有丰富功能的开源数据集成、数据处理、数据分析和数据勘探平台.它提供可视化的方式创建数据流或数据通道,可选择性地运行一些或全部的分析步骤,最终输出研究结果、模型以及可交互的视图.KNIME 由 Java 写成,其通过插件的方式来提供更多的功能.通过插件用户可以为文件、图片和时间序列加入处理模块,并可以集成到其他开源项目中,比如:R 语言,Weka.KNIME 是通过工作流来控制数据的集成、清洗、转换、过滤,再到统计、数据挖掘,最后是数据的可视化.整个开发都在可视化的环境下进行,通过简单的拖曳和设置就可以完成一个流程的开发.KNIME 被设计成一种模块化的、易于扩展的框架.它的处理单元和数据容器之间没有依赖性,这使得它们更加适应分布式环境及独立开发.另外,对 KNIME 进行扩展也是比较容易的事情.开发人员可以很轻松地扩展 KNIME 的各种类型的结点、视图等.

5) Weka/Pentaho(14.8%)

Weka 的全名是怀卡托智能分析环境(waikato environment for knowledge analysis),是一款免费的、非商业化的、基于 Java 环境下开源的机器学习以及数据挖掘软件.Weka 提供的功能有数据处理,特征选择、分类、回归、聚类、关联规则、可视化等.而 Pentaho 则是世界上最流行的开源商务智能软件.它是一个基于 Java 平台的商业智能(business intelligence, BI)套件,之所以说是套件是因为它包括一个 Web server 平台和几个工具软件:报表、分析、图表、数据集成、数据挖掘等,可以说包括了商务智能的各个方面.在 Pentaho 中集成了 Weka 的数

据处理算法,可以直接调用.

需要说明的是,虽然 KDNuggets 的调查是针对大数据,但是上述 5 种分析工具,并非全是针对大数据而设计的. 例如 excel,在大数据出现之前,就已经用于数据分析.

2　大数据应用

大数据应用,是利用大数据分析的结果,为用户提供辅助决策,发掘潜在价值的过程. 本节首先回顾各种数据源的应用演化,并研究由结构化数据分析、文本分析、网站分析、多媒体分析、网络分析和移动分析构成的 6 个关键分析领域,最后列举大数据的典型应用.

2.1　应用演化

数据驱动的应用程序被广泛地应用于各个领域,例如早在 20 世纪 90 年代商业智能就成为了一个在商界流行的术语,而 21 世纪早期就出现了基于海量数据挖掘处理的网站搜索. 下面列出了一些来自不同领域的很有潜力且影响较大的应用演变,并讨论其数据分析的特性.

1) 商业应用的演变

最早的商业数据通常为结构化数据,各公司从其旧有系统中收集这些数据并把它们存储到关系型数据库管理系统中. 这些系统中使用的分析技术在 20 世纪 90 年代非常流行,通常都很直观但也很简单,例如报表、仪表盘、条件查询、基于搜索的商业智能、联机事务处理、交互式可视化、记分卡、预测建模、数据挖掘等[33]. 自 21 世纪初以来,互联网和网站给各类组织机构提供了一个在线展示其业务并和客户直接互动的独特机遇. 大量的产品和客户信息,包括点击流数据日志、用户行为等,均可以从网站上获取. 这样通过采用各种文本和网站挖掘技术分析就可以实现产品布局优化、客户交易分析、产品的建议和市场结构分析.

2) 网络应用的演变

早期的网络主要提供电子邮件和网页服务,而文本分析、数据挖掘和网页分析技术也相应地用于挖掘电子邮件内容、构建搜索引擎等. 网络数据占据了全球数据量的大多数. 如今,Web 已经日渐成为相互关联的页面的世界,充满了各种不同类型的数据,例如文本、图像、视频、照片和交互内容等. 大量用于半结构化或非结构化数据的高级技术应运而生. 例如,图像分析技术可以从照片中提取有用的信息,从脸部识别个人等. 多媒体分析技术可以应用于商业、执法和军事应用中的自动化视频监控系统中.

2004 年后,在线社交媒体,例如论坛、网上群体、网络博客、社交网站,社交多媒体网站等,为用户创建、上传并分享内容提供了更为便捷的方式,社交数据开始爆发式的增长.

此外,网络应用所产生数据,不再仅源自互联网,移动网络和物联网也成为网络数据的重要来源. 据文献[34]显示,2011 年移动电话和平板电脑的数量第 1 次超过了笔记本电脑和 PC 的数量,移动电话和基于传感器的物联网正在开启新一轮网络应用的演变.

3) 科学应用的演变

许多领域的科研都在通过高通量传感器和仪器获取大量数据,从天体物理学和海洋学,到基因学和环境研究,无不如此. 美国国家科学基金会(NSF)最近公布了 BIGDATA 方案征集,以利于信息共享和数据分析. 一些学科已经开发了海量数据平台并取得了相应的收益. 例如,在生物学中,iPlant 正应用网络基础设施、物理计算资源、协作环境、虚拟机资源、可互操作的分析软件和数据服务来协助研究人员、教育工作者和学生建设所有的植物学科. iPlant 数据集形式变化多端,其中包括规范或参考数据、实验数据、模拟和模型数据、观测数据以及其他派生数据.

表 2 所示为具有代表性的大数据应用及其特征.

表 2　典型的大数据应用及其特征

应用	实例	用户数量	反应时间	数据规模	可靠性	准确性
科学计算	生物信息	小	慢	TB	适中	很高
金融	电子商务	大	非常快	GB	很高	很高
社交网络	Facebook	很大	快	PB	高	高
移动数据	移动电话	很大	快	TB	高	高
物联网	传感网	大	快	TB	高	高
Web 数据	新闻网站	很大	快	PB	高	高
多媒体	视频网站	很大	快	PB	高	适中

2.2　大数据分析的关键领域

根据数据的生成方式和结构特点不同,本文将数据分析划分为 6 个关键技术领域:

1) 结构化数据. 一直是传统数据分析的重要研究对象,目前主流的结构化数据管理工具,如关系型数据库等,都提供了数据分析功能.

2) 文本. 是常用的存储文字、传递信息的方式,也是最常见的非结构化数据.

3) Web 数据. Web 技术的发展,极大地丰富了获取和交换数据的方式,Web 数据高速的增长,使其成为大数据的主要来源.

4) 多媒体数据. 随着通讯技术的发展,图片、音频、视频等体积较大的数据,也可以被快速地传播,由于缺少文字信息,其分析方法与其他数据相比,具有显著的特点.

5) 社交网络数据. 从一定程度上反映了人类社会活动的特征,具有重要的价值.

6) 移动数据. 与传统的互联网数据不同,具有明显的地理位置信息、用户个体特征等其他信息.

结构化数据分析、文本分析、Web 分析、多媒体分析、社交网络分析和移动分析,这 6 个关键领域分类旨在强调数据的不同特性,其中的一些领域可能会利用类似的底层技术,或者存在交集,这样分类的目的在于理解和激发数据分析领域中的关键问题和技术.

2.2.1 结构化数据分析

商业和科研领域会产生大量的结构化数据,而这些结构化数据的管理和分析依赖于数据库、数据仓库、OLAP 和业务流程管理(business process management, BPM)[35]的成熟商业化技术. 得益于关系型数据库技术的发展,结构化数据的分析方法较为成熟,大部分都以数据挖掘和统计分析为基础.

2.2.2 文本分析

存储信息最常见的形式就是文本,例如电子邮件通信、公司文件到网站页面、社交媒体内容等. 因此,文本分析被认为比结构化数据挖掘更具有商业化潜力. 通常情况下,文本分析,也称为文本挖掘,指的是从非结构化文本中提取有用信息和知识的过程. 文本挖掘是一个跨学科领域,涉及到信息检索、机器学习、统计、计算语言学尤其是数据挖掘. 大部分文本挖掘系统都以文本表达和自然语言处理(NLP)为基础,重在后者.

文档介绍和查询处理是开发向量空间模型、布尔检索模型,概率检索模型[36]的基础,而这些模型又构成了搜索引擎的基础. 自 20 世纪 90 年代早期以来,搜索引擎已经演化成成熟的商业系统,通常包括快速分布式爬行、有效地倒排索引、基于 inlink 的网页排序和搜索日志分析.

NLP 技术可以提高关于期限的可用信息,这样计算机就可以分析、理解甚至产生文本. 下面是一些经常采用的方法:词法获取、词义消歧、词性标注、概率上下文无关文法[37]. 以 NLP 为基础,一些技术已经被开发出来并可以应用于文本挖掘,其中包括信息提取、主题模型、文本摘要、分类、聚类、答疑和意见挖掘. 信息提取是指自动地从文本中提取特定种类的结构化信息. 命名实体识别(NER)技术作为信息提取的一个子任务,旨在识别归属于预定类别(如人物、地点和组织等)的文本中的原子实体,近来已成功开发用于新的分析[38]和医学领域的应用[39]. 主题模型以“文档由主题组成,而主题是词汇的概率分布”这一观点建立. 主题模型是文档生成模型,规定了生成文档的概率程序.

现在已经有各种各样的概率主题模型用于分析文档的内容和词汇的意义[40]. 文本摘要是为了从单个或多个输入文本文件中生成一个缩减的摘要或摘录. 文本摘要的各种类型可以归结为具象性摘要和抽象性摘要[41]. 具象性摘要从源文档中选择重要的句子和段落等并把它们浓缩成较短的形式. 而抽象性摘要可以理解原文本并可以根据语言学方法用较少的词汇对原文本进行复述. 文本分类的目的在于通过将文档置入预定的主题集来识别文档的主题取向. 基于图表示和图挖掘的文本分类最近吸引了大家的研究兴趣[42]. 文本聚类用于给类似的文档分组,文档聚类通过预定的主题对文档进行分类. 在文本聚类中,文档可以出现在多个副主题当中. 通常采用数据挖掘领域的一些聚类算法来计算文档的相似性,但研究显示可以利用结构关系信息来增强聚类结果[43]. 答疑系统主要设计用于处理如何寻找给定问题的最佳答案. 它涉及问题分析、源检索、答案提取和回答演示[44]方面的不同技术. 答疑系统可以应用于许多领域,其中包括教育、网站、健康和国防. 意见挖掘与情感分析类似,是指提取、分类、理解和评估新闻、评论和用户生成的其他内容中表述的意见的计算技术. 它可以提供理解公众和客户对社会事件、政治运动、公司策略、营销活动和产品喜好的有利机会[45].

2.2.3 Web 分析

在过去的 10 年中,我们见证互联网信息的爆炸式增长,同时 Web 分析作为一个活跃的研究领域也已经出现. Web 分析旨在从 Web 文档和服务中自动检索、提取和评估信息用以发现知识. Web 分析建立在几个研究领域之上,包括数据库、信息检索、自然语言处理和文本挖掘等. 我们可以根据要挖掘的 Web 部分的不同将 Web 分析划分为 3 个相关领域:Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘[46].

Web 内容挖掘处理 Web 页面内容中有用信息

或知识的发现,Web 内容涉及多种类型的数据,例如文本、图像、音频、视频、代号、元数据以及超链接等.对图像、音频和视频挖掘的研究被称为多媒体分析,将在下一部分讨论.由于大部分 Web 内容数据为非结构化文本数据,大部分研究工作都是围绕文本和超文本内容展开.超文本挖掘涉及到具有超级链接的半结构化 HTML 页面的挖掘.

监督学习和分类在超文本挖掘中扮演重要角色,例如电子邮件、新闻组管理和维护 Web 目录[47]等.Web 内容挖掘可以采用两种方法进行:信息检索方法和数据库的方法.信息检索方法主要是协助或改善信息查找或根据推断或征求用户配置文件为用户过滤信息.数据库方法试图模拟并整合 Web 上的数据,这样就可以进行比基于关键词的搜索更为复杂的查询.

Web 结构挖掘涉及到发现 Web 链接结构相关的模型.这里的结构指的是网站中或网站间链接的示意图.模型是基于具有或没有链接描述的超链接的拓扑结构建立的.该模型揭示了不同网站间的相似性和相互关系,可以用来为网站页面分类.Page Rank[48]和 CLEVER[49]方法充分利用了该模型来查找相关网站页面.主题爬取[50]是另外一个利用该模型的成功案例.主题爬虫的目的在于有选择性地找出与预定的主题集相关的页面.主题爬虫会分析其爬行边界来寻找与爬取最有可能相关的链接并避免涉及 Web 的不相干区域,而不是收集和索引所有可访问的网页文件,来回答所有可能的即席查询.这样可以节约大量硬件和网络资源并帮助保持爬取更新.

Web 使用挖掘希望挖掘 Web 会话或行为产生的辅助数据,而 Web 内容挖掘和 Web 结构挖掘使用的是 Web 上的主要数据.Web 使用数据包括来自 Web 服务器访问日志、代理服务器日志、浏览器记录、用户配置文件、登记数据、用户会话或交易、缓存、用户查询、书签数据、鼠标点击和滚动以及用户和 Web 交互产生的任何其他数据.随着 Web 服务和 Web2.0 系统的成熟和普及,Web 使用数据正变得越来越多样化.Web 使用挖掘在个性化空间、电子商务、网络隐私/安全和其他一些新兴领域内扮演着关键角色.例如,协同推荐系统通过利用用户偏好的异同来使电子商务个性化.

2.2.4　多媒体分析

近来,多媒体数据(主要包括图像、音频和视频)正以惊人的速度增长,几乎无处不在.由于多媒体数据多种多样而且大多数都比单一的简单结构化数据和文本数据包含更丰富的信息,提取信息这一任务正面临多媒体数据语义差距的巨大挑战.多媒体分析的研究涵盖的学科种类非常多,从多媒体摘要、多媒体注解、多媒体索引和检索、多媒体的建议和多媒体事件检测等,此处仅举最近的几个研究重点.

音频摘要可以通过从原数据中简单地提取突出的词或句子或合成新的表述来实现.视频摘要可以理解最重要或更具代表性的视频内容序列,可以是静态的,也可以是动态的.静态视频摘要方法要利用一个关键帧序列或上下文敏感的关键帧来代表视频.这些方法都很简单,而且已经应用到商业应用(例如 Yahoo,Alta Visa 和 Google 等)中,但其可播放性很差.而动态视频摘要方法是使用一系列视频片段来表示视频,另外,还可以配置低级的视频功能并采取其他平滑措施使最终的摘要看起来更为自然.文献[51]提出了一个面向主题的多媒体摘要系统,该系统可以为一次观看完毕的视频生成基于短信息的重新计算.

多媒体注释指的是为图像和视频指派一组在句法和语义级别上描述其所含内容的标签.多媒体索引和检索指的是描述、存储并组织多媒体信息和协助人们方便、快捷地查找多媒体资源[52].

多媒体推荐的目的是要根据用户的喜好来推荐特定的多媒体内容.大多数现有的推荐系统分为两种:基于内容系统和基于协同过滤的系统.基于内容的方法识别用户或用户兴趣的一般特征并向用户推荐具有相似特征的其他内容,这些方法纯粹依赖于内容相似度测量,但大多受内容分析有限和过度规范困扰.基于协同过滤的方法识别具有相似兴趣的人群并根据小组成员的行为推荐内容[53].现在又引入了一种混合方法,融合了基于协同过滤和内容两种方法的长处来提高推荐的质量.

多媒体时间检测,是检测基于事件套件(event kit)的视频剪辑内某一事件的发生情况,而事件套件中含有一些有关概念和一些示例视频的文本描述.目前视频事件检测的研究仍处在初级阶段.事件检测的现有研究大多集中在体育或新闻事件以及监控录像中的奔跑或不寻常事件等之类的重复模式事件.作者在文献[54]中针对处理少数正例样本(positive training examples)的多媒体事件检测提出了一种新算法.

2.2.5　社交网络分析

网络分析从最初的计量分析[55]和社会学网络

分析[56]一直演化到21世纪初新兴的在线社交网络分析. 许多流行的在线社交网络,例如Twitter, Facebook和LinkedIn等近年来都日益普及. 这些在线社交网络通常都含有大量的链接和内容数据,其中链接数据主要为图形结构,表示两个实体之间的通信,而内容数据则包含有文本、图像以及其他网络多媒体数据. 这些网络的丰富内容给数据分析带来了前所未有的挑战,同时也带来了机遇. 按照以数据为中心的观点来看,社交网络上下文的研究方向可以分为两大类:基于链接的结构分析和基于内容的分析[57].

基于链接的结构分析研究一直着力于链接预测、社区发现、社交网络进化和社会影响分析以及其他一些领域. 社交网络可以作为图形实现可视化,图形中的定点对应于一个人,同时其中的边表示对应人士之间的某些关联. 由于社交网络是动态网络,不断会有新的顶点和边添加到图形中去. 链接预测希望能预测两个节点之间未来建立联系的可能性. 许多技术都可以用于链接预测,例如基于特征的分类、概率方法以及线性代数等. 基于特征的分类可以为定点对选择一组特征,然后再利用现有的链接信息来生产二元分类器以预测未来的链接情况[58]. 概率方法尝试为社交网络中的定点之间的连接概率建立模型[59]. 线性代数方法要根据降秩相似矩阵计算两个几点之间的相似性[60]. 社区指的是一个子图结构,该结构中子图中的定点上的边的密度更大,而子图间的定点上的变得密度较低. 人们提出并比较了许多针对社区检测的方法[61],大部分的方法都是基于拓扑并依赖于捕获社区结构概念的目标函数. Du等人利用现实生活中存在的重叠社区的性质提出了一种更为有效的大规模社交网络社区检测方法[62]. 针对社交网络的研究旨在寻找解释网络演化的法则和推导模型. 一些实证研究[63-65]发现近似偏见(proximity bias)、地域限制和其他一些因素在社交网络的演化过程中起着重要作用,同时还提出了一些生成方法[66]来协助网络和系统设计. 社交影响是指个人受网络中其他人的影响而改变自身行为. 社交影响的强弱[67]取决于人与人之间的关系、网络距离、时间效应、网络与个人的特点等许多因素. 营销、广告、推荐和其他许多应用都可以通过定性和定量测量个人对其他人的影响力[68]获取好处. 通常情况下,如果将社交网络之间的内容增殖考虑在内,基于链接的结构分析的性能都可以进一步改进.

得益于Web 2.0技术的革命性进展,社交网络中生成的内容呈爆炸性增长. 社交网络中基于内容的分析研究指的是社交媒体分析. 社交媒体内容包括文本、多媒体、定位和评论. 几乎所有的有关结构化分析、文本分析和多媒体分析的研究主题都可以解释为社交媒体分析,但社交媒体分析正面临着前所未有的挑战. 首先,我们需要在合理的时间期限内自动分析大量的而且不断增长的社交媒体数据. 其次,社交媒体数据中含有许多噪声数据. 例如博客圈中存在大量垃圾博客,Twitter中的trivial Tweets同样如此. 第三,社交网络是动态网络,常常在很短的时间内频繁变化和更新. 社交媒体紧贴于社交网络,因此社交媒体分析不可避免地要受社交网络分析的影响. 社交网络分析指的是社交网络上下文,尤其是社交和网络结构特征的文本分析和多媒体分析. 目前社交媒体分析的研究仍处在初级阶段. 社交网络文本分析的应用包括关键字搜索、分类、聚类和异构网络中的迁移学习. 关键字搜索试图同时使用内容和链接行为来进行搜索[69]. 这一应用背后隐藏的含义为含有类似关键字的文本文档通常都链接在一起. 在分类的过程中,假定社交网络中的节点都具有标签,然后再将这些加标的节点用于分类目的[70]. 在聚类过程中,研究人员尝试确定具有类似内容的节点集,并以此进行聚类[71]. 鉴于社交网络包含有大量的相互链接的不同种类对象的信息,例如文章、标签、图像和视频等,异构网络中的迁移学习旨在不同的链接之间迁移信息知识[72]. 社交网络中的多媒体数据集按照结构化的形式组织,纳入了丰富的信息内容,例如语义本体论、社会互动、社区媒体、地理地图以及多媒体意见. 社交网络中的结构化多媒体分析研究也被称为多媒体信息网络. 多媒体信息网络的链接结构主要为逻辑型结构,对多媒体网络中的多媒体来说至关重要. 多媒体信息网络中的逻辑连接结构可以分为4类:语义本体、社区媒体、个人照片相册和地理位置[57]. 我们可以根据逻辑连接结构进一步改善检索系统[73]、推荐系统结果[74]、协作标签系统[75]和其他一些应用的结果.

2.2.6 移动分析

随着移动计算的快速增长,世界上的移动终端(例如移动电话、传感器等)和应用也越来越多. 截止到2013年4月,安卓应用提供了超过650 000个应用,几乎涵盖了所有可以想见的种类. 截至2012年底,每个月的移动数据流量已经达到了885 PB[76]. 大量的数据和应用为移动分析开拓了广阔的研究领域,同时也带来了不少的挑战. 总体上来说,移动数

据的特征十分独特,例如移动感知、活动灵敏、嘈杂而且有大量冗余.近来不同的领域中均出现了新的移动分析研究来应对挑战.由于移动分析研究远未成熟,我们仅介绍一些最近的而且最具有代表性的分析应用.

随着移动电话用户数量的增长以及功能的改善,移动电话如今能够建立和维护社区,这些社区既可以区域进行划分,又可以文化兴趣进行划分,例如最近出现的微信.传统的互联网社区或社交网络社区缺乏成员间的在线互动,而且只有在成员在个人电脑前时社区才会活跃.而与此相反,移动电话可以支持随时随地的交互.移动社区被定义为一群具有相同爱好(即健康、安全、娱乐等)的人首先在网络上聚在一起,然后再亲自会面制定共同目标,商定措施以实现目标,再接着就开始实施其计划[77].

射频识别(radio frequency identification, RFID)技术使得传感器可以在没有光线的情况下远距离读取与标签相关的唯一产品识别码(EPC)[78].这些标签可以按照符合成本效益的方式识别、定位、跟踪和监控物理对象,因此 RFID 广泛应用于库存管理和物流行业.

近年来无线传感器、移动通信技术和流处理领域的进展使得人们可以建立体域网来实时监测个人身体健康状况.

2.3　大数据的典型应用

2.3.1　企业内部大数据应用

目前,大数据的主要来源和应用都是来自于企业内部,商业智能(business intelligence, BI)和 OLAP 可以说是大数据应用的前辈.企业内部大数据的应用,可以在多个方面提升企业的生产效率和竞争力.具体而言:市场方面,利用大数据关联分析,更准确地了解消费者的使用行为,挖掘新的商业模式;销售规划方面,通过大量数据的比较,优化商品价格;运营方面,提高运营效率和运营满意度,优化劳动力投入,准确预测人员配置要求,避免产能过剩,降低人员成本;供应链方面,利用大数据进行库存优化、物流优化、供应商协同等工作,可以缓和供需之间的矛盾、控制预算开支,提升服务.

在金融领域,企业内部大数据的应用得到了快速发展.例如,招商银行通过数据分析识别出招行信用卡价值客户经常出现在星巴克、DQ、麦当劳等场所后,通过"多倍积分累计""积分店面兑换"等活动吸引优质客户;通过构建客户流失预警模型,对流失率等级前 20%的客户发售高收益理财产品予以挽留,使得金卡和金葵花卡客户流失率分别降低了 15 个和 7 个百分点;通过对客户交易记录进行分析,有效识别出潜在的小微企业客户,并利用远程银行和云转介平台实施交叉销售,取得了良好成效.

当然最典型的应用还是在电子商务领域,每天有数以万计的交易在淘宝上进行,与此同时相应的交易时间、商品价格、购买数量会被记录,更重要的是,这些信息可以与买方和卖方的年龄、性别、地址、甚至兴趣爱好等个人特征信息相匹配.淘宝数据魔方是淘宝平台上的大数据应用方案,通过这一服务,商家可以了解淘宝平台上的行业宏观情况、自己品牌的市场状况、消费者行为情况等,并可以据此进行生产、库存决策,而与此同时,更多的消费者也能以更优惠的价格买到更心仪的宝贝.而阿里信用贷款则是阿里巴巴通过掌握的企业交易数据,借助大数据技术自动分析判定是否给予企业贷款,全程不会出现人工干预.据透露,截至目前阿里巴巴已经放贷 300 多亿元,坏账率约 0.3%左右,大大低于商业银行.

2.3.2　物联网大数据应用

物联网不仅是大数据的重要来源,还是大数据应用的主要市场.在物联网中,现实世界中的每个物体都可以是数据的生产者和消费者,由于物体种类繁多,物联网的应用也层出不穷.

在物联网大数据的应用上,物流企业应该有深刻的体会.UPS 快递为了使总部能在车辆出现晚点的时候跟踪到车辆的位置和预防引擎故障,它的货车上装有传感器、无线适配器和 GPS.同时,这些设备也方便了公司监督管理员工并优化行车线路.UPS 为货车定制的最佳行车路径是根据过去的行车经验总结而来的.2011 年,UPS 的驾驶员少跑了近 4 828 万公里的路程.

智慧城市,是一个基于物联网大数据应用的热点研究项目,图 2 所示为基于物联网大数据的智能城市规划.迈阿密戴德县,就是一个智慧城市的样板.佛罗里达州迈阿密戴德县与 IBM 的智慧城市项目合作,将 35 种关键县政工作和迈阿密市紧密联系起来,帮助政府领导在治理水资源、减少交通拥堵和提升公共安全方面制定决策时获得更好的信息支撑.IBM 使用云计算环境中的深度分析向戴德县提供智能仪表盘应用,帮助县政府各个部门实现协作化和可视化管理.智慧城市应用为戴德县带来多方面的收益,例如戴德县的公园管理部门今年因及时发现和修复跑冒滴漏的水管而节省了 100 万美元的水费.

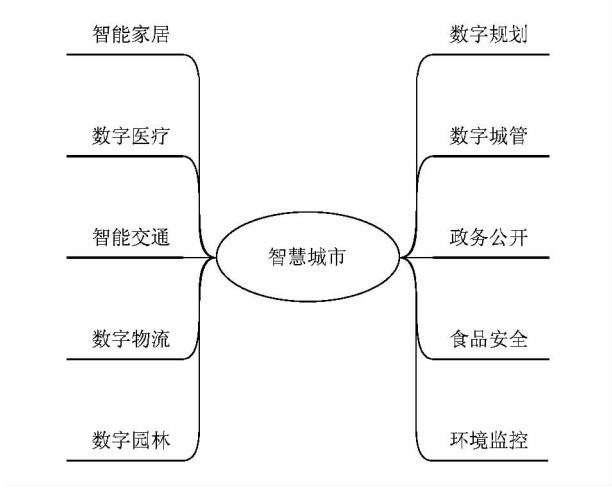

图 2 基于物联网的智能城市

2.3.3 面向在线社交网络大数据的应用

在线社交网络,是一种在信息网络上由社会个体集合及个体之间的连接关系构成的社会性结构.在线社交网络大数据主要来自即时消息、在线社交、微博和共享空间 4 类应用.由于在线社交网络大数据代表了人的各类活动,因此对于此类数据的分析得到了更多关注.在线社交网络大数据分析是从网络结构、群体互动和信息传播 3 个维度,通过基于数学、信息学、社会学、管理学等多个学科的融合理论和方法,为理解人类社会中存在的各种关系提供的一种可计算的分析方法.目前,在线社交网络大数据的应用包括网络舆情分析、网络情报搜集与分析、社会化营销、政府决策支持、在线教育等.

圣克鲁斯警察局是美国警界最早应用大数据进行预测分析的试点,通过分析社交网络,可以发现犯罪趋势和犯罪模式,甚至可以对重点区域的犯罪概率进行预测.

2013 年 4 月,美国计算搜索引擎 Wolfram Alpha,通过对 Facebook 中 100 多万美国用户社交数据进行分析,试图研究用户的社会行为规律.根据分析发现,大部分 Facebook 用户在 20 岁出头时开始恋爱,27 岁左右时订婚,30 岁左右结婚,而 30～60 岁之间,婚姻关系变化缓慢.这个研究结果与美国人口普查数据相比,几乎完全一致.

总得说来,在线社交网络大数据应用可以从以下 3 方面帮助我们了解人的行为,以及掌握社会和经济活动的变化规律:

1) 前期警告.通过检测用户使用电子设备及服务中出现的异常,在出现危机时可以更快速地应对.

2) 实时监控.通过对用户当前行为、情感和意愿等方面的监控,可以为政策和方案的制定提供准确的信息.

3) 实时反馈.在实时监控的基础上,可以针对某些社会活动获得群体的反馈信息.

2.3.4 医疗健康大数据应用

医疗健康数据是持续、高增长的复杂数据,蕴涵的信息价值也是丰富多样.对其进行有效的存储、处理、查询和分析,可以开发出其潜在价值.对于医疗大数据的应用,将会深远的影响人类的健康.

例如,安泰保险为了帮助改善代谢综合症的预测,从千名患者中选择 102 个完成实验.在一个独立的实验室工作内,通过患者的一系列代谢综合症的检测试验结果,在连续 3 年内,扫描 600 000 个化验结果和 18 万索赔事件.将最后的结果组成一个高度个性化的治疗方案,以评估患者的危险因素和重点治疗方案.这样,医生可以通过食用他汀类药物及减重 5 磅等建议而减少未来 10 年内 50%的发病率.或者通过你目前体内高于 20%的含糖量,而建议你降低体内甘油三酯总量.

西奈山医疗中心(Mount Sinai Meddical Center)是美国最大最古老的教学医院,也是重要的医学教育和生物医药研究中心.该医疗中心使用来自大数据创业公司 Ayasdi 的技术分析大肠杆菌的全部基因序列,包括超过 100 万个 DNA 变体,来了解为什么菌株会对抗生素产生抗药性.Ayasdi 的技术使用了一种全新的数学研究方法:拓扑数据分析(topological data analysis),来了解数据的特征.

微软的 HealthVault,是一个出色的医学大数据的应用,它是 2007 年发布的,目标是希望管理个人及家庭的医疗设备中的个人健康信息.现在已经可以通过移动智能设备录入上传健康信息,而且还可以第三方的机构导入个人病历记录,此外通过提供 SDK 以及开放的接口,支持与第三方应用的集成.

2.3.5 群智感知

随着技术的发展,智能手机和平板电脑等移动设备集成了越来越多的传感器,计算和感知能力也愈发强大.在移动设备被广泛使用的背景下,群智感知开始成为移动计算领域的应用热点.大量用户使用移动智能设备作为基本节点,通过蓝牙、无线网络和移动互联网等方式进行协作,分发感知任务分发,收集、利用感知数据,最终完成大规模的、复杂的社会感知任务.群智感知对参与者的要求很低,用户并不需要相关的专业知识或技能,只需拥有一台移动智能设备.

众包(crowdsourcing)是一种极具代表性的群智感知模式,是一种新型的解决问题的方式.众包以

用户为基础,以自由参与的方式分发任务.目前众包已经被运用于人力密集的应用,如语言翻译、语音识别、图像地理信息标记、定位与导航、城市道路交通感知、市场预测、意见挖掘等.众包的核心思想是将任务分而治之,通过参与者的协作来完成个体不可能或者说根本想不到要完成的任务.无需部署感知模块和雇佣专业人员,众包就可以将感知范围扩展至城市规模甚至更大.

其实,众包的应用早于大数据的兴起,宝洁、宝马、奥迪等许多公司都曾借助众包提升自身的研发和设计能力.而在大数据时代,空间众包服务(spatial crowdsourcing)成为了大家关注的热点.空间众包服务的工作框架如下:服务请求方要求获取与特定地点相关的资源,而愿意接受任务请求的参与者将到达指定地点,利用移动设备获取相关数据(视频、音频或图片),最后将这些数据发送给服务请求方.随着移动设备使用的高速增长以及移动设备提供的功能越来越复杂,可以预见空间众包将会变得比传统形式的众包服务更加流行,如 Amazon Turk[79]和 Crowdflower[80].

2.3.6　智能电网

智能电网,是指将现代信息技术融入传统能源网络构成新的电网,通过用户的用电习惯等信息,优化电能的生产、供给和消耗,是大数据在电力系统上的应用.智能电网可以解决以下几方面的问题:

1)电网规划

通过对智能电网中的数据进行分析,可以知道哪些地区的用电负荷和停电频率过高,甚至可以预测哪些线路可能出现故障.这些分析结果,可以有助于电网的升级、改造、维护等工作.例如,美国加州大学洛杉矶分校的研究者就根据大数据理论设计了一款"电力地图",将人口调查信息、电力企业提供的用户实时用电信息和地理、气象等信息全部集合在一起,制作了一款加州地图.该图以街区为单位,展示每个街区在当下时刻的用电量,甚至还可以将这个街区的用电量与该街区人的平均收入和建筑物类型等相比照,从而得出更为准确的社会各群体的用电习惯信息.这个地图为城市和电网规划提供了直观有效的负荷数预测依据,也可以按照图中显示的停电频率较高、过载较为严重的街区进行电网设施的优先改造.

2)发电与用电的互动

理想的电网,应该是发电与用电的平衡.但是,传统电网的建设是基于发—输—变—配—用的单向思维,无法根据用电量的需求调整发电量,造成电能的冗余浪费.为了实现用电与发电的互动,提高供电效率,研究者开发出了智能的用电设备——智能电表.德克萨斯电力公司(TXU Energy)已经广泛使用智能电表,并取得了巨大的成效.供电公司能每隔15 min就读一次用电数据,而不是过去的一月一次.这不仅仅节省了抄表的人工费用,而且由于能高频率快速采集分析用电数据,供电公司能根据用电高峰和低谷时段制定不同的电价,利用这种价格杠杆来平抑用电高峰和低谷的波动幅度,智能电表和大数据应用让分时动态定价成为可能,而且这对于TXU Energy和用户来说是一个双赢变化.

3)间歇式可再生能源的接入

目前许多新能源也被接入电网,但是风能和太阳能等新能源,其发电能力与气候条件密切相关,具有随机性和间歇性的特点,因此难以直接并入电网.如果通过对电网大数据的分析,则可对这些间歇式新能源进行有效调节,在其产生电能时,根据电网中的数据将其调配给电力紧缺地区,与传统的水火电能进行有效地互补.

3　研究现状与展望

3.1　研究现状

大数据应用面临着许多挑战,而目前的研究仍处于初期阶段,仍需要进行更多的研究工作来解决数据展示、数据储存以及数据分析的效率等问题.表3所示为目前大数据研究所取得的成果.

表3　大数据相关技术一览

类别		代表性例子
平台	本地	Hadoop,MapR,Cloudera,Hortonworks,InfoSphere BigInsights,ASTERIX
	云	AWS,Google Compute Engine,Azure
数据库	SQL	Greenplum,Aster Data,Vertica
	NoSQL	HBase,Cassandra,MongoDB,Redis
	NewSQL	Spanner,MegaStore,F1
数据仓库		Hive,HadoopDB,Hadapt
数据处理	批处理	MapReduce,Dryad
	流处理	Storm,S4,Kafka
查询语言		HiveQL,Pig Latin,DryadLINQ,MRQL,SCOPE
统计分析机器学习		Mahout,Weka,R
日志处理		Splunk,Loggly

3.1.1 基础理论研究

虽然大数据在学术和工业界是一个热点话题,但是有关它的一些科学问题并没有得到完整的解决:

1)大数据的基本问题.大数据的科学定义、大数据的结构模型、大数据的形式化表述、数据科学的理论体系等.现在有关大数据的讨论,并没有一个形式化、结构化的描述,无法严格界定并验证什么是大数据.

2)大数据的标准化工作.数据质量的评价体系、数据计算效率的评估标准等.很多大数据应用的解决方案都声称可以在各方面提升数据的处理分析能力,但是目前缺少一个统一的评估标准,以数学方法衡量大数据的计算效率,只能以实施大数据应用后的增益来评价其性能.这样无法横向比较各解决方案之间的优劣,甚至无法了解采用大数据方案前后的效率对比.此外,数据质量是数据预处理,数据的精简和筛选的一个重要依据,因此如何有效地评价数据质量也是一个急需的问题.

3)大数据计算模式的变革.外存模型、数据流模型、PRAM 模型、MR 模型等.大数据的出现,引发了算法设计的发展,算法已经从计算密集型转化为数据密集型.数据转移已经是大数据计算问题的主要瓶颈.因此,出现了很多专门针对大数据的新型计算模型,今后肯定还会出现更多的模型.

3.1.2 关键技术研究

大数据技术尚处于起步阶段,还有很多关键技术问题,如云计算、网格计算、流运算、并行计算、大数据体系结构、大数据的变成模型、支持大数据的软件系统等,需要深入研究.

1)大数据格式的转化

大数据由于其数据来源广泛且种类繁多,因此异构性、异质性一直是其大数据的特点,也是制约数据格式转化效率的关键.如果可以提高大数据格式转化的效率,那么大数据的应用可以获得更多的价值.

2)大数据的转移

大数据的转移,主要涉及大数据的生成、采集、传输、存储等数据在空间位置的变换.前面提到过,大数据环境下转移数据的开销很大,是大数据计算的瓶颈.但是,数据的转移是不可避免的,因此如何提高数据的转移速率,是提升大数据计算的关键.

3)大数据的实时性

针对不同的应用场合,大数据应用的实时性是一个核心问题.如何定义数据的生命周期,如何计算数据的"折旧率",如何建立实时应用和在线应用的计算模型,这些都会影响大数据分析反馈结果的价值.

4)大数据的处理

随着大数据的发展,对于大数据的处理也从简单的数据分析衍生出新的问题:数据再利用,大数据的一个典型特征就是价值大而密度低,随着数据规模的增大,对已有数据进行再次利用,会发掘出更多的价值;重组数据,将处于不同业务中的数据集进行重组,重组后的数据价值总和比单个数据集的总和还大;数据废气,就是数据采集过程中的错误数据,在大数据环境下并不是只有正确的数据才可以被利用,就像工业废气一样,错误的数据也可以被利用.

3.1.3 应用实践研究

尽管现在已经有很多成功的大数据应用案例,但是在实践过程中仍然存在着诸多问题亟待解决:

1)大数据管理.大数据的出现,对传统的数据管理带来了新的挑战.目前,针对大数据管理的研究主有适用于大数据的数据库和互联网应用、适用于新型硬件的存储模型和数据库、异构和多结构化数据的集成、移动和普适计算的数据管理、社交网络的数据管理、分布式数据管理等.

2)大数据搜索、挖掘与分析.数据处理一直是大数据领域的研究热点,如模型社交网络的搜索与挖掘、大数据搜索算法、分布式搜索、P2P 搜索、大数据的可视化分析、海量推荐系统和社交媒体系统、实时大数据挖掘、图像挖掘、文本挖掘、语义挖掘、多结构化数据挖掘、机器学习等.

3)大数据的集成和世系.上文提到过,将多个数据集进行综合利用所能获得的价值,远超过其个体价值的总和.因此,如何将不同的数据源整合在一起是亟待解决的问题.数据集成就是要将不同来源不同的数据集进行整合,其面临许多问题,如数据模式不同、冗余数据量大等问题.数据世系是用来描述数据的产生、并随时间推移而演化的过程信息[81].在大数据时代,数据世系的研究对象更多的是多个数据集,而不仅是单个数据集.因此,如何将来自不同数据集具有不同标准的数据世系信息整合,是值得研究的问题.

4)大数据应用.目前对大数据的应用还处于起步阶段,我们需要探索更多、更高效地利用大数据的模式.因此,科学、工程、医学、医疗、金融、商务、法律、教育、运输、零售、电信等特定领域的大数据应用,中小企业大数据应用,公共管理部门大数据应用,大数据服务,大数据人机交互等都具有较高的研究意义.

3.1.4　数据安全研究

信息技术中,安全和隐私一直是重点问题.大数据时代,随着数据的增多,数据面临更严峻的安全风险,传统的数据保护方法已经不适用于大数据,大数据安全面对挑战:

1)大数据的隐私

大数据时代,数据的隐私问题包括两个方面:一方面是个人隐私的保护,随着数据采集技术的发展,在用户无法察觉,个人的兴趣、习惯、身体特征等隐私信息可以被更容易地获取;另一方面,即使得到用户的许可,个人隐私数据在存放、传输和使用的过程中,也有被泄露的风险.大数据的分析能力导致看似简单的信息可能会被挖掘出其中的隐私,因此面对大数据时代的隐私保护将成为新的命题.

2)数据质量

数据质量影响着大数据的利用,低质量的数据不仅浪费了传输和存储资源,甚至无法被利用.制约数据质量的因素有很多,在生成、采集、传输和存储的过程中,都可能影响数据质量.数据质量具体表现在:准确性、完整性、冗余性、一致性.虽然有很多提升数据质量的措施,但是数据质量的问题是不可能完全根除的.因此,需要研究一种方法,可以对数据质量进行自动检测,并可以自行修复部分出现质量问题的数据.

3)大数据安全机制

大数据在数据规模和数据种类方面,给数据加密带来了挑战.以前针对中小规模的加密方法在性能上无法满足大数据的要求,需要研究高效的大数据密码学.针对不同结构的结构化、半结构化和非结构化数据,需要研究如何有效地进行安全管理、访问控制和安全通信.此外,在多租户的模式下,需要在保证效率的前提下,实现租户数据的隔离性、保密性、完整性、可用性、可控性和可追踪性.

4)信息安全领域的大数据应用

大数据不仅给信息安全带来了挑战,也为信息安全的发展带来注入了新的动力.例如,通过对入侵检测系统的日志文件进行大数据分析,可以发现潜在的安全漏洞隐患以及高级可持续性威胁(advanced persistent threat, APT).此外,病毒特征、漏洞特征和攻击特征等信息也更容易通过大数据分析而被掌握.

综上所述,大数据的安全问题已经获得了国内外研究学者的高度关注,然而,目前在在多源异构大数据的表示、度量和语义理解方法,建模理论和计算模型,能效优化的分布存储和处理的硬件及软件系统架构等方面相关的研究还并不多见,特别是在大数据安全方面,包括大数据的可信性问题、针对各应用领域的大数据备份与恢复技术、大数据完整性维护技术、大数据安全保密技术等还需进行进一步研究.

3.2　研究展望

大数据的出现,开启了一次重大的时代转型.在IT时代,以前技术(technology, T)才是大家关注的重点,是技术推动了数据的发展;如今数据的价值凸显,信息(information, I)的重要性日益提高,今后将是数据推动技术的进步.大数据不仅改变了社会经济生活,也在影响了每个人的生活和思维方式,而这样的改变才刚刚开始.

1)规模更大、种类更多、结构更复杂的数据

虽然目前以Hadoop为代表的技术取得了巨大的成功,但是随着大数据迅猛的发展速度,这些技术肯定也会落伍被淘汰.就如同Hadoop,它的理论基础早在2006年就已诞生.为了能更好地应对未来规模更大、种类更多、结构更复杂的数据,很多研究者已经开始关注此问题,其中最为著名的当属谷歌的全球级的分布式数据库Spanner[82],以及可容错可扩展的分布式关系型数据库F1[83].未来,大数据的存储技术将建立在分布式数据库的基础上,支持类似于关系型数据库的事务机制,可以通过类SQL语法高效地操作数据.

2)数据的资源化

既然大数据中蕴藏着巨大的价值,那么掌握大数据就掌握了资源.从大数据的价值链分析,其价值来自数据本身、技术和思维,而核心就是数据资源,离开了数据技术和思维是无法创造价值的.不同数据集的重组和整合,可以创造出更多的价值.今后,掌控大数据资源的企业,将数据使用权进行出租和转让就可以获得巨大的利益.

3)大数据促进科技的交叉融合

大数据不仅促进了云计算、物联网、计算中心、移动网络等技术的充分融合,还催生了许多学科的交叉融合.大数据的发展,既需要立足于信息科学,探索大数据的获取、存储、处理、挖掘和信息安全等创新技术与方法,也需要从管理的角度探讨大数据对于现代企业生产管理和商务运营决策等方面带来的变革与冲击.而在特定领域的大数据应用,更需要跨学科人才的参与.表4显示了与大数据相关的技术和学科预计的发展时间表.

表 4 大数据相关技术和学科预计的发展时间表

预计高峰期到来时间	技术萌芽期	期望膨胀期	泡沫谷底期	稳步爬升期	生产高峰期
<2 年高峰期到来			虚拟桌面	多媒体平板历理念管理	预测分析
2～5 年高峰期到来	硅阳极电池	大数据、无线充电、BYOD、社交分析、私有云计算、内存数据库、活动流	NFC、云计算、手势控制、内存分析、文本分析、移动 OTA 支付	IT 消费化、生物特征识别、语音识别	
5～10 年高峰期到来	自动内容识别、3D 扫描、自动驾驶、自然语言问答、语音翻译	众包、游戏化、HTML5、混合云计算、3D 打印、复合事件处理、APP 市场、增强实现	NFC 支付、互联网电视、语音挖掘、机器间通讯、家庭健康监控、虚拟世界	消费级车联网	
>10 年高峰期到来	人机技能增进、量子计算、3D 生物打印、全息显示、移动机器人、物联网		无线传感网络		

4) 大数据可视化

在许多人机交互场景中，都遵循所见即所得(what you see is what you get，WYSIWYG)的原则，例如文本和图像编辑器等. 在大数据应用中，混杂的数据本身是难以辅助决策的，只有将分析后的结果以友好的形式展现，才会被用户接受并加以利用. 报表、直方图、饼状图、回归曲线等经常被用于表现数据分析的结果，以后肯定会出现更多的新颖的表现形式，例如微软的“人立方”社交搜索引擎使用关系图来表现人际关系.

5) 面向数据

程序是数据结构和算法，而数据结构就是存储数据的. 在程序设计的发展历程中，也可以看出数据的地位越来越重要. 在逻辑比数据复杂的小规模数据时代，程序设计以面向过程为主；随着业务数据的复杂化，催生了面向对象的设计方法. 如今，业务数据的复杂度已经远远超过业务逻辑，程序也逐渐从算法密集型转向数据密集型. 可以预见，一定会出现面向数据的程序设计方法，如同面向对象一样，在软件工程、体系结构、模式设计等方面对 IT 技术的发展产生深远的影响.

6) 大数据引发思维变革

在大数据时代，数据的收集、获取和分析都更加快捷，这些海量的数据将对我们的思考方式产生深远的影响. 在文献[18]中，对大数据引发的思维变革进行了总结：

① 分析数据时，要尽可能地利用所有数据，而不只是分析少量的样本数据.

② 相比于精确的数据，我们更乐于接受纷繁复杂的数据.

③ 我们应该更为关注事物之间的相关关系，而不是探索因果关系.

④ 大数据的简单算法比小数据的复杂算法更为有效.

⑤ 大数据的分析结果将减少决策中的草率和主观因素，数据科学家将取代“专家”.

7) 以人为本的大数据

纵观人类社会的发展史，人的需求及意愿始终是推动科技进步的源动力. 在大数据时代，通过挖掘和分析处理，大数据可以为人的决策带来参考答案，但是并不能取代人的思考. 正是人的思维，才促使众多利用大数据的应用，而在大数据更像是人的大脑功能的延伸和扩展，而不是大脑的替代品. 随着物联网的兴起，移动感知技术的发展，数据采集技术的进步，人不仅是大数据的使用者和消费者，还是生产者和参与者. 基于大数据的社会关系感知、众包、社交网络大数据分析等与人的活动密切相关的应用，在未来会受到越来越多的关注，也必将引起社会活动的巨大变革.

4 总 结

大数据应用，是通过数据分析的方法从大数据中发掘潜在价值，具有重要的研究意义和实际价值. 本文重点介绍了大数据应用的相关概念、技术及方法，并根据数据的生成方式和结构特点的不同，创造性地将大数据应用划分为 6 大关键领域，并介绍了 6 个典型的应用. 最后，根据相关研究所存在的问题，本文从基础理论、关键技术、应用实践以及数据安全 4 个方面总结了大数据的研究现状，并从工程实践、交叉学科、方法论、人机交互等视角，对大数据应用的未来进行展望.

参 考 文 献

[1] Gantz J, Reinsel D. Extracting value from chaos. IDC iView, 2011: 1-12

[2] Cukier K. Data, data everywhere. The Economist, 2010, 394(8671): 3-16

[3] Drowning in numbers-digital data will flood the planet-and help us understand it better. The economist, 2011 [2013-07-13]. http://www.economist.com/blogs/dailychart/2011/11/bigdata-0

[4] Lohr S. The age of big data. New York Times, 2012, 11

[5] Noguchi Y. Following digital breadcrumbs to big data gold. National Public Radio, 2011 [2013-05-21]. http://www.npr.org/2011/11/29/142521910/thedigitalbreadcrumbs-that-lead-to-big-data

[6] Noguchi N. The search for analysts to make sense of big data. National Public Radio, 2011 [2013-05-11]. http://www.npr.org/2011/11/30/142893065/the-searchforanalysts-to-make-sense-of-big-data

[7] Big data. Nature, 2008 [2013-06-13]. http://www.nature.com/news/specials/bigdata/index.html

[8] Pecial online collection: Dealing with big data. Scinece, 2011 [2013-07-16]. http://www.sciencemag.org/site/special/data/

[9] Fact sheet: Big data across the federal government, 2012 [2013-08-07]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big data fact sheet 3292012.pdf

[10] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011: 1-137

[11] Ghemawat S, Gobioff H, Leung S T. The google file system //Proc of the 19th ACM Symp on Operating Systems Principles. New York: ACM, 2003: 29-43

[12] Dean J, Ghemawat S. Mapreduce: Simplified data processing on large clusters. Communications of the ACM, 2008, 51(1): 107-113

[13] Hey A J, Tansley S, Tolle K M, et al. The Fourth Paradigm: Dataintensive Scientific Discovery. New York: Microsoft Research, 2009

[14] Laney D. 3D data management: Controlling data volume, velocity and variety. META Group Research, 2001,6

[15] Zikopoulos P, Eaton C, et al. Understanding big data: Analytics for enterprise class hadoop and streaming data. New York: McGraw-Hill Osborne Media, 2011

[16] Meijer E. The world according to linq. Communications of the ACM, 2011, 54(10): 45-51

[17] Mark B. Gartner says solving big data challenge involves more than just managing volumes of data. Gartner Retrieved, 2011,13

[18] **维克托·迈尔-舍恩伯格, 肯尼思·库克耶. 大数据时代. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社**, 2013

[19] Team O R. Big Data Now: Current Perspectives from O'Reilly Radar. Sebastopol: O'Reilly Media, 2011

[20] Grobelnik M. Big data tutorial, 2012 [2013-08-17]. http://videolectures.net/eswc2012 grobelnik big data/

[21] **桂卫华, 刘晓颖. 基于人工智能方法的复杂过程故障诊断技术. 控制工程**, 2002, 9(4): 1-6

[22] **李德仁, 王树良, 李德毅, 等. 论空间数据挖掘和知识发现的理论与方法. 武汉大学学报: 信息科学版**, 2002, 27(3): 221-233

[23] **梅宏, 王千祥, 张路, 等. 软件分析技术进展. 计算机学报**, 2009, 32(9): 1697-1710

[24] **宋国杰, 唐世渭, 杨冬青, 等. 数据流中异常模式的提取与趋势监测. 计算机研究与发展**, 2004, 41(10): 1754-1759

[25] Hand D J. Principles of data mining. Drug Safety, 2007, 30(7): 621-622

[26] **孟小峰, 慈祥. 大数据管理: 概念、技术与挑战. 计算机研究与发展**, 2013, 50(1): 146-169

[27] **覃雄派, 王会举, 李芙蓉, 等. 数据管理技术的新格局. 软件学报**, 2013, 36(2): 175-197

[28] Thusoo A, Shao Z, Anthony S, et al. Data warehousing and analytics infrastructure at facebook //Proc of the 2010 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2010: 1013-1020

[29] Goodhope, Ken, et al. Building LinkedIn's Real-time Activity Data Pipeline. Data Engineering, 2012, 35(2): 33-45

[30] Ren Z, Xu X, Wan J, et al. Workload characterization on a production Hadoop cluster: A case study on Taobao // Workload Characterization (IISWC), 2012 IEEE Int Symp on IEEE. Piscataway, NJ: IEEE, 2012: 3-13

[31] Boulon J, Konwinski A, Qi R, et al. Chukwa, a large-scale monitoring system //Proc of CCA. 2008, 8

[32] What Analytics, Data mining, Big Data software you used in the past 12 months for a real project? 2012 [2013-07-11]. http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html

[33] Sallam J H R L, Richardson J, Hostmann B. Magic quadrant for business intelligence platforms. Stamford, CT: Gartner Group, 2011

[34] Economist T. Beyond the pc, Special Report on Personal TEchnology. 2011 [2013-07-16]. http://www.economist.com/node/21531109

[35] Agrawal P B D, Bertino E. Challenges and opportunities with big data. Washington, DC: The Computing Research Association, CRA Report, 2012

[36] Salton G. Automatic Text Processing, Reading. MA: Addison Wesley, 1989

[37] Manning C D, Schutze H. Foundations of Statistical Natural Language Processing. Cambridge, MA: The MIT Press., 1999

[38] Ritter A, Clark Mausam S, Etzioni O. Named entity recognition in tweets: An experimental study //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 1524-1534

[39] Li Y, Hu X, Lin H, et al. A framework for semisupervised feature generation and its applications in biomedical literature mining. IEEE/ACM Trans on Computational Biology and Bioinformatics, 2011, 8(2): 294-307

[40] Blei D M. Probabilistic topic models. Communications of the ACM, 2012, 55(4): 77-84

[41] Balinsky H, Balinsky A, Simske S J. Automatic text summarization and small-world networks //Proc of the 11th ACM Symp on Document Engineering. New York: ACM, 2011: 175-184

[42] Mishra M, Huan J, Bleik S, et al. Biomedical text categorization with concept graph representations using a controlled vocabulary //Proc of the 11th Int Workshop on Data Mining in Bioinformatics. New York: ACM, 2012: 26-32

[43] Hu J, Fang L, Cao Y, et al. Enhancing text clustering by leveraging wikipedia semantics //Proc of the 31st Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2008: 179-186

[44] Maybury M T. New Directions in Question Answering. Cambridge, MA: The MIT Press., 2004

[45] Pang B, Lee L. Opinion mining and sentiment analysis. Found. Trends in Information Retrieval, 2008, 2(1/2): 1-135

[46] Pal S, Talwar V, Mitra P. Web mining in soft computing framework: Relevance, state of the art and future directions. IEEE Trans on Neural Networks, 2002, 13(5): 1163-1177

[47] Chakrabarti S. Data mining for hypertext: A tutorial survey, SIGKDD Explorations Newsletter, 2000, 1(2): 1-11

[48] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998, 30(1): 107-117

[49] Konopnicki D, Shmueli O. W3qs: A query system for the worldwide Web //Proc of the 21st Int Conf on Very Large Data Bases. San Francisco, MA: Morgan Kaufmann, 1995: 54-65

[50] Chakrabarti S, van den Berg M, Dom B. Focused crawling: S new approach to topic-specific Web resource discovery, Computer Networks, 1999, 31(11-16): 1623-1640

[51] Ding D, Metze F, Rawat S, et al. Beyond audio and video retrieval: Towards multimedia summarization //Proc of the 2nd ACM Int Conf on Multimedia Retrieval, 2012: 2: 1-2: 8

[52] Lew M S, Sebe N, Djeraba C, et al. Content-based multimedia information retrieval: State of the art and challenges. ACM Trans on Multimedia Computing, Communications, and Applications, 2006, 2(1): 1-19

[53] Park Y J, Chang K N. Individual and group behavior-based customer profile model for personalized product recommendation. Expert Systems with Applications, 2009, 36(2): 1932-1939

[54] Ma Z, Yang Y, Cai Y, et al. Knowledge adaptation for ad hoc multimedia event detection with few exemplars //Proc of the 20th ACM Int Conf on Multimedia. New York: ACM, 2012: 469-478

[55] Hirsch J E. An index to quantify an individual's scientific research output. Proc of the National Academy of Sciences of the United States of America, 2005, 102(46): 16569

[56] Watts D J. Six Degrees: The Science of a Connected Age. New York: WW Norton & Company, 2004

[57] Aggarwal C C. An introduction to social network data analytics. Berlin: Springer, 2011

[58] Scellato S, Noulas A, Mascolo C. Exploiting place features in link prediction on location-based social networks //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2011: 1046-1054

[59] Ninagawa A, Eguchi K. Link prediction using probabilistic group models of network structure //Proc of the 2010 ACM Symp on Applied Computing. New York: ACM, 2010: 1115-1116

[60] Dunlavy D M, Kolda T G, Acar E. Temporal link prediction using matrix and tensor factorizations. ACM Trans on Knowledge Discovery from Data, 2011, 5(2): 10: 1-10: 27

[61] Leskovec J, Lang K J, Mahoney M. Empirical comparison of algorithms for network community detection //Proc of the 19th Int Conf on World Wide Web. New York: ACM, 2010: 631-640

[62] Du N, Wu B, Pei X, et al. Community detection in large-scale social networks //Proc of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. New York: ACM, 2007: 16-25

[63] Garg S, Gupta T, Carlsson N, et al. Evolution of an online social aggregation network: An empirical study //Proc of the 9th ACM SIGCOMM Conf on Internet Measurement. New York: ACM, 2009: 315-321

[64] Allamanis M, Scellato S, Mascolo C. Evolution of a locationbased online social network: Analysis and models // Proc of the 2012 ACM Conf on Internet Measurement. New York: ACM, 2012: 145-158

[65] Gong N Z, Xu W, Huang L, et al. Evolution of social-attribute networks: Measurements, modeling, and implications using google+ //Proc of the 2012 ACM Conf on Internet Measurement. New York: ACM, 2012: 131-144

[66] Zheleva E, Sharara H, Getoor L. Co-evolution of social and affiliation networks //Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 1007-1016

[67] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks //Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 807-816

[68] Li Y, Chen W, Wang Y, et al. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships //Proc of the 6th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2013: 657-666

[69] Lappas T, Liu K, Terzi E, Finding a team of experts in social networks //Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 467-476

[70] Zhang T, Popescul A, Dom B. Linear prediction models with graph regularization for web-page categorization //Proc of the 12th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2006: 821-826

[71] Zhou Y, Cheng H, Yu J X. Graph clustering based on structural/attribute similarities. Proc of the VLDB Endowment, 2009, 2(1): 718-729

[72] Dai W, Chen Y, Xue G R, et al. Translated learning: Transfer learning across different feature spaces //Proc of the Advances in Neural Information Processing Systems (NIPS). 2008: 353-360. doi:10.1.1.145.5832

[73] Rabbath M, Sandhaus P, Boll S. Multimedia retrieval in social networks for photo book creation //Proc of the 1st ACM Int Conf on Multimedia Retrieval. New York: ACM, 2011: 72: 1-72: 2

[74] Shridhar S, Lakhanpuria M, Charak A, et al. Snair: A framework for personalised recommendations based on social network analysis //Proc of the 5th Int Workshop on Location-Based Social Networks. New York: ACM, 2012: 55-61

[75] Maniu S, Cautis B. Taagle: Efficient, personalized search in collaborative tagging networks //Proc of the 2012 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012: 661-664

[76] Cisco. Cisco visual networking index: Global mobile data traffic forecast update, 2012c2017. 2013 [2013-08-15]. http://doc.mbalib.com/view/4d7d1101d4ba076636d19080594cc1b.html

[77] Rhee Y, Lee J. On modeling a model of mobile community: Designing user interfaces to support group interaction. Interactions, 2009, 16(6): 46-51

[78] Han J, Lee J G, Gonzalez H, et al. Mining massive rfid, trajectory, and traffic data sets //Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008

[79] Paolacci G, Chandler J, Ipeirotis P. Running experiments on amazon mechanical turk. Judgment and Decision Making, 2010, 5(5): 411-419

[80] Finin T, Murnane W, Karandikar A, et al. Annotating named entities in Twitter data with crowdsourcing //Proc of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Los Alamitos, CA: Association for Computational Linguistics, 2010: 80-88

[81] **高明, 金澈清, 王晓玲, 等. 数据世系管理技术研究综述. 计算机学报**, 2010, 33(3): 373-388

[82] Corbett J C, Dean J, Epstein M, et al. Spanner: Google's globally-distributed database. ACM Trans on Computer Systems, 2013, 31(3): 8

[83] Shute J, Oancea M, Ellner S, et al. F1: The fault-tolerant distributed RDBMS supporting google's ad business //Proc of the 2012 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012: 777-778

张　引　男,1986年生,博士,博士后,讲师,主要研究方向为系统分析与集成、半结构化数据管理、大数据、移动计算、云计算等(dr.yinzhang@gmail.com).

陈　敏　男,1980年生,博士,教授,博士生导师,主要研究方向为无线传感网、物联网、大数据等(minchen2012@hust.edu.cn).

廖小飞　男,1978年生,博士,教授,博士生导师,主要研究方向为虚拟化、对等计算、集群计算、流媒体服务等(xfliao@hust.edu.cn).