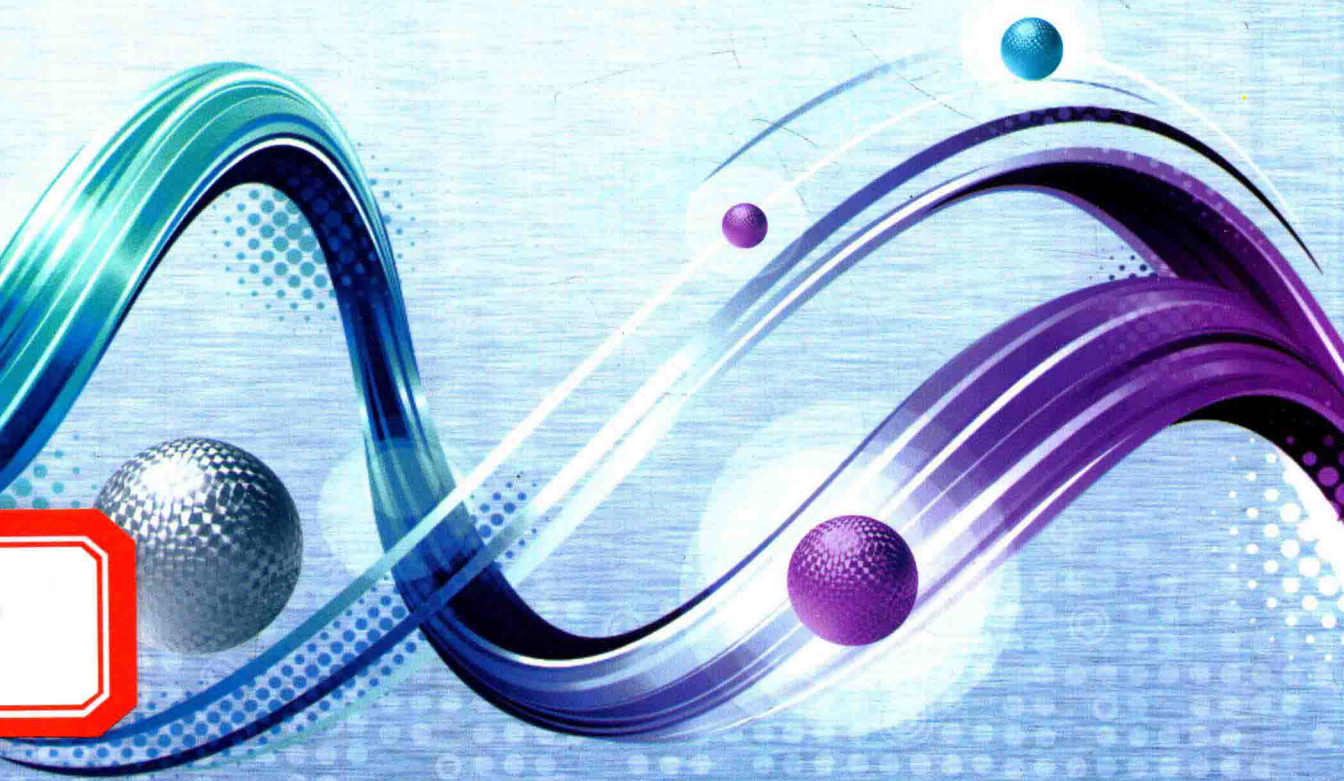
# 面向大数据的高维数据挖掘技术

王和勇　著



西安电子科技大学出版社
http://www.xduph.com

高职高专电子信息类“十三五”规划教材

# 面向大数据的高维数据挖掘技术

王和勇　著



西安电子科技大学出版社

## 内 容 简 介

本书从高维大数据的特性出发，指出了高维大数据的挖掘过程，介绍了大数据维数约简的目的和分类，对大数据特征选择和提取的线性和非线性方法进行了介绍并研究了相关的改进方法，给出了结合图的降维方法和稀疏大数据的维数约简方法。

本书可供从事大数据挖掘和商务智能研究的高校教师、研究生、科研院所的科研人员及有关工程技术人员使用。

**图书在版编目(CIP)数据**

**面向大数据的高维数据挖掘技术**/王和勇著. —西安：
西安电子科技大学出版社，2018.3
ISBN 978-7-5606-4218-5

① 面… Ⅱ. ① 王… Ⅲ. ① 数据收集—研究 Ⅳ. ① TP274

**中国版本图书馆 CIP 数据核字(2017)第 186217 号**

策　　划　胡华霖
责任编辑　马武装
出版发行　西安电子科技大学出版社(西安市太白南路 2 号)
电　　话　(029)88242885　88201467　　邮　　编　710071
网　　址　www.xduph.com　　电子邮箱　xdupfxb001@163.com
经　　销　新华书店
印刷单位　陕西华沐印刷科技有限责任公司
版　　次　2018 年 3 月第 1 版　2018 年 3 月第 1 次印刷
开　　本　787 毫米×1092 毫米　1/16　印张　8.375
字　　数　192 千字
印　　数　1～1000 册
定　　价　18.00 元
ISBN 978-7-5606-4218-5/TP

**XDUP　4510001-1**

# 前　言

随着移动互联网、物联网、社交网络等技术和应用的兴起，数据正以史无前例的规模汇集。与以往基于猜测或模型进行决策不同，如今，人们大多是根据数据本身进行决策。

越来越多的高维大数据出现在人们的生活中，在这些高维大数据中蕴含着许多无比重要的信息。在一定范围内，数据的分类效率会随着特征属性(即维度)的增长呈指数形式增长，但当数据的维度过高时，其中往往包含了过多无关项、冗余项、干扰项等，分类的效率反而会随着维数的增长而下降，当特征属性的数目达到一定数量时，就会产生“维数灾难”导致所谓的“数据爆炸”，从而不能达到我们最初预想的结果。因此，如何从高维大数据中剔除无关项、降低数据维数、提高分类效率，便成了处理数据的关键。

大数据的高维约简作为模式分类的预处理部分，发挥着重要的作用。首先通过大数据的高维约简，可去除不相关的和冗余的特征，滤除不必要的干扰，使模式语言更加精练，又使分类器的学习算法更具效率，达到更高的预测精度；其次通过维数约简，使用最少数目的特征来描述输入模式，可使输入的特征空间维度被最大程度地降低，避免发生“维数灾难”，同时降低了分类模型学习算法的空间复杂度，从而提高了训练速度，减少了学习时间；最后选取最显著相关的选定数目的输入特征，不但极大地降低了提供给分类模型的训练样本数，提高了学习算法的准确率，避免了发生“峰化”现象，也使分类模型的预测更有效，而且在满足样本数量存储量的需求的同时，也减少了存储数据的成本。

面向大数据的高维数据挖掘不只是需要统计学、人工智能的研究成果，也需要大数据管理平台提供有效的存储、索引、查询操作，其中源于高性能计算的并行操作技术在处理大数据方面是很重要的。因此在面向大数据的高维数据挖掘研究中常常涉及多个学科的多个领域知识。本书围绕面向大数据的高维数据挖掘这个主题，对高维大数据的数据特征选择、大数据特征提取的线性和非线性方法进行介绍并研究了相关的改进方法，给出了结合图形的降维方法和稀疏大数据的维数约简方法。各章的主要内容如下：

第 1 章介绍大数据的产生背景、机遇、挑战和应用的发展方向，并介绍了大数据挖掘过程和大数据维数。

第 2 章介绍大数据维数约简的目的、分类及有关定义。

第 3 章介绍特征选择的基本框架、相似性度量和特征选择分类及稳定性等方面的内容。

第 4 章介绍特征提取的概念、分类，主成分分析、线性判别分析、独立主成分分析、最大间距准则等线性特征提取方法，并给出了每种方法的改进方向。

第 5 章介绍核方法、流形学习等非线性特征提取方法并给出了每种方法的改进方向。

第 6 章介绍基于图的主成分分析、线性鉴别和边界费舍尔分析的图方法，并给出了图嵌入方法面临的挑战。

第7章介绍稀疏表示理论、岭回归、套索回归、稀疏保持映射、稀疏判别分析等稀疏大数据的维数约简方法。

本书是作者近年来从事大数据高维数据挖掘的研究成果的总结，在编写过程中，得到了美国迈阿密大学的meiling shyu教授、刘典婷博士、闫祎麟博士的指导和帮助，在此表示感谢。同时感谢数据挖掘组的蓝金炯、洪明和崔蓉同学，与他们就某些专业问题的讨论使我受益匪浅。

由于面向大数据的高维数据挖掘技术涉及的内容较多，本书不可能包含这个研究领域的全部内容，同时由于作者水平所限，书中不当与错误之处在所难免，敬请广大读者及相关专家批评指正。

作　者

2017.12

# 目 录

# 第 1 章　高维大数据

## 1.1　大数据介绍

### 1.1.1　大数据产生的背景

半个世纪以来，随着计算机技术融入社会生活，信息爆炸已经积累到了开始引发变革的程度[1]。不仅世界充斥着比以往更多的信息，而且信息增长速度也在加快。进入 21 世纪后，数据信息更迎来了大发展的时代，移动互联、社交网络、电子商务等极大地拓展了互联网的边界和应用范围，各种数据迅速膨胀并变大。互联网(社交、搜索、电商)、移动互联网(微博、微信)、物联网(传感器、智慧地球)、车联网、GPS、医学影像、安全监控、金融(银行、股市、保险)、电信(通话、短信)等都在疯狂产生着数据。地球上总共拥有的数据量增长迅速，2006 年，全球一共新产生了约 180EB 的数据；到 2011 年，这个数字达到了 1.8 ZB。有市场研究机构预测，到 2020 年，整个世界的数据总量将会增长四十多倍，达到 35.2 ZB (1 ZB≈$10^{10}$ TB)[2,3]。在这种情况下，信息爆炸性增长的学科中，如天文学和基因学，创造出了“大数据(Big Data)”这个概念。如今，这个概念几乎应用到了所有人类智力与发展的领域中。

### 1.1.2　大数据的重要性

随着移动互联网、物联网、社交网络等技术和应用的兴起。学术界和工业界都对大数据赋予大量的关注并展开了深刻的讨论。Nature 于 2008 年第一次推出 Big Data 专刊[4]。Science 在 2011 年 2 月推出专刊《Dealing with Data》[5,6]，主要围绕着科学研究中大数据的问题展开讨论，说明了大数据对于科学研究的重要性。麦肯锡研究院(McKinsey Global Institute，MGI)则于 2011 年 6 月发布名为《Big data：The next frontier for innovation，competition，and productivity》的研究报告，对大数据的影响、关键技术和应用领域等都进行了详尽的分析，并指出大数据将会是带动未来生产力发展和创新以及消费需求增长的风向标。2012 年以来，人们对大数据的关注度与日俱增。2012 年 3 月美国奥巴马政府发布了《大数据研究和发展倡议》(Big Data Research and Development Initiative)，投资 2 亿多美元，正式启动“大数据发展计划”。计划在科学研究、环境、生物医学等领域利用大数据技术进行突破。奥巴马政府的这一计划使大数据上升到国家战略。Gartner 在一年一度的技术成

熟度曲线(见图 1-1)报告中指出，大数据已进入膨胀期，并将在未来 2～5 年进入发展高峰期。由此可见，大数据是未来信息技术的重要发展方向之一。

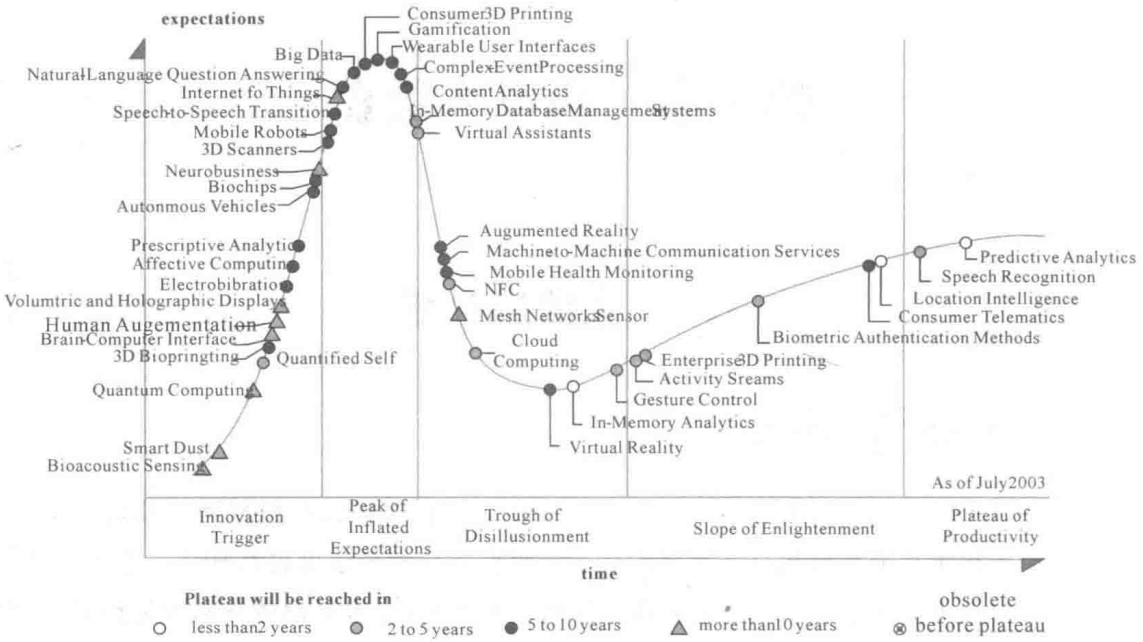



图 1-1　技术成熟度曲线

### 1.1.3　大数据的定义和特征

大数据这个名词的出现至少有 5 年以上的历史，然而至今业界对其也没有一个统一认同的完美定义。这好像是个不可思议的事情，因为表面看来“大数据”这个词汇已经直白得不能再简单了。这个“大”字只是个表象，而其内在蕴含着丰富的意义。让我们来看看众多权威机构和企业对大数据给予的不同定义[7]。

麦肯锡说：“大数据指的是所涉及的数据集规模已经超过了传统数据库软件获取、存储、管理和分析的能力。这是一个被故意设计成主观性的定义，并且是一个关于多大的数据集才能被认为是大数据的可变定义，即：并不定义大于一个特定 TB 数字的数据集才叫大数据。因为随着技术的不断发展，符合大数据标准的数据集容量也会增长，并且其定义随不同的行业也有变化，这依赖于在一个特定行业通常使用何种软件和数据集有多大。因此，大数据在今天不同行业中的范围可以从数十太(TB)字节到数拍(PB)字节。”

IBM 说：“可以用 3 个特征相结合来定义大数据：数量(Volume)、种类(Variety)和速度(Velocity)，或者就是简单的 3V 或 V3，即庞大容量、种类丰富和极快速度生成及处理的数据。”如图 1-2 所示。

数据量：如今存储的数据数量正在急剧增长，使我们深陷在数据之中。我们存储所有事物：环境数据、财务数据、医疗数据、监控数据等。有关数据量已从太字节(TB)级别转向拍字节(PB)级别，并且不可避免地会转向 ZB 级别。现在经常听到一些企业使用存储集群来保存数拍字节(PB)的数据。可供企业使用的数据量不断增长，而可处理、理解和分析的数据比例却不断下降。

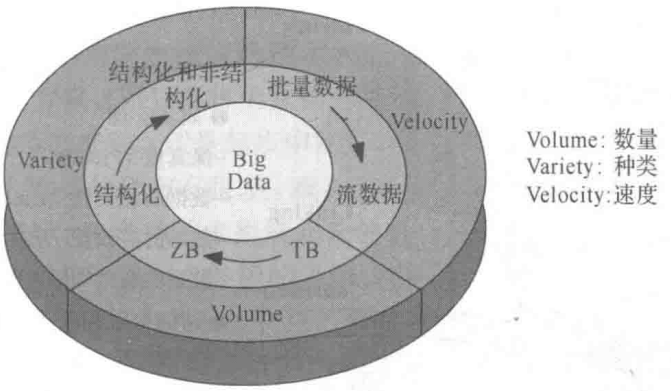



图 1-2　按数量、种类和速度来定义大数据

数据的多样性：与大数据现象有关的数据量为尝试处理它的数据中心带来了新的挑战：数据多样的种类。随着传感器、智能设备以及社交协作技术的激增，企业中的数据也变得更加复杂，因为它不仅包含传统的关系型数据，还包含来自网页、互联网日志文件(包括单击流数据)、搜索索引、社交媒体论坛、电子邮件、文档、主动和被动系统的传感器数据等原始、半结构化和非结构化数据。

数据的处理速度：就像我们收集和存储的数据量及种类发生了变化一样，生成和需要处理数据的速度也在变化。速度的概念不能限定为与数据存储相关的增长速率，应动态地将此定义应用到数据——数据流动的速度。有效处理大数据需要在数据变化的过程中对它的数量和种类进行分析，而不只是在它静止后进行分析。

最近，IBM 在以上 3 V 的基础上又归纳总结了第 4 个 V——Veracity(真实和准确)。“只有真实而准确的数据才能让对数据的管控和治理真正有意义。随着社交数据、企业数据、交易与应用数据等新数据源的兴起，传统数据源的局限性被打破，企业愈发需要有效的信息治理以确保其真实性及安全性。”

IDC 指出：“大数据是一个貌似不知道从哪里冒出来的大的动力。但是实际上，大数据并不是新生事物。然而，它确实正在走入主流，并得到重大关注，这是有原因的。廉价的存储、传感器和数据采集技术的快速发展，通过云和虚拟化存储设施增加的信息链路，以及创新软件和分析工具，正在驱动着大数据。大数据不是一个‘事物’，而是一个跨多个信息技术领域的动力和活动。大数据技术描述了新一代的技术和架构，其被设计用于：通过使用高速(Velocity)的采集、发现或分析，从超大容量(Volume)的多样(Variety)数据中经济地提取其价值(Value)。”

IDC 的定义除了揭示了大数据传统的 3 V 基本特征，即 Volume、Variety、Velocity，还增添了一个新特征：Value。

一个大数据实现的主要价值可以基于下面三个评价准则中的一个或多个进行评判：

- 它提供了更有用的信息吗？
- 它改进了信息的精确性吗？
- 它改进了响应的及时性吗？

Gartner 说：“实际上，大数据或者说‘极限信息’(Extreme Information)具有 12 个维度。”图 1-3 展示了极限信息管理的 3 个层次和 12 个象限。

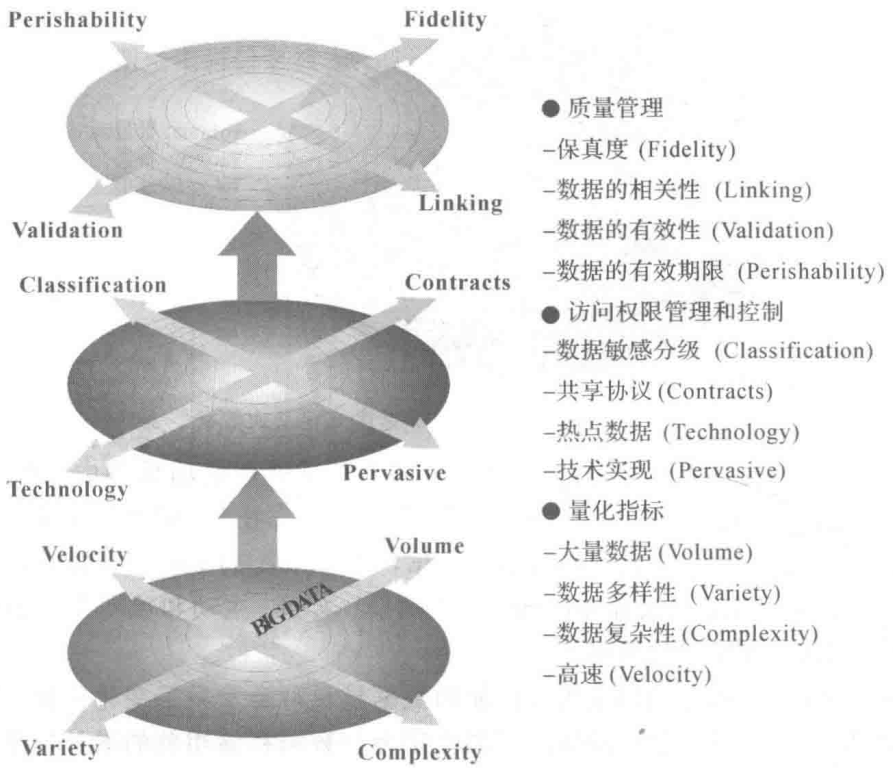



图 1-3 Gartner 极限信息管理的 3 个层次和 12 个象限

图 1-3 中的最下面一层“量化指标”指的是大数据的基本特征，即大数据量、多样性和高速，这也就是传统的 3 V 的概念。另外还加上了复杂性，包括空间维、时间维等多种数据复杂性。大数据解决方案应首先考虑以这些问题为出发点。然而，解决这 4 个方面的问题只是大数据解决方案的基础，用以支撑起大数据平台，在这之上还有很多问题需要解决。

第二层“访问权限管理和控制”有很多关于访问权限的问题。数据的敏感性是一个很基础的问题，但到现在为止，基于现有的技术和管理手段，还没有对数据的敏感性进行分析的优秀解决方案。

共享协议：数据将会以什么形式、什么格式和时间点，通过什么样的接口实现这些共享和数据的交换，这是大数据的重点问题之一。数据交换的所有方式都是以标准的协议来支持的，因为在大数据的时代，数据的来源本身是多样性的，数据的格式甚至是无法管理的，还有很多的数据是来自于企业的外部，来自于互联网的提供商。到底如何通过协议自动地将数据放到数据仓库里，这种情况下数据的共享协议是一个很关键的问题。

热点数据：在大数据时代，数据的管理与传统的方式有了非常明显的差别。传统的数据管理会把单独的时间点作为一个热点数据，但是在大数据时代，热点数据有可能是并行的多个。这些热点数据本身之间实际上是有可能有联系的。因为各种事件的相互触发，所以很有可能这些热点数据同时出现，而且是相互关联的，甚至于有可能是可以预测的。所以说在大数据时代，热点数据的管理也是一个重要的方面。

最上面一层“质量管理”在传统的数据管理里是非常重要的一个方面。这里面提到的有效性、有效期限，都有明确的技术工具来解决。但到现在为止，在这些方面，还是非常地依赖于传统的数据仓库的工具，没有专门针对大数据的工具和技术能够解决这些问题。所以

产生的结果是，现在产生的大数据的应用，一方面受制于用户接受的程度本身，另外一方面也受制于技术。现在，因为缺少相应的技术和工具的支持，很多用户仍然必须要依赖于传统的数据管理的解决方案，而只能将大数据的技术作为一个前台来做一些预处理。所以，大数据从 12 个象限的角度来说，还是起步中的起步，因为里面有一些非常基本的问题到现在还没有很好解决。大数据的形态有很多，现在仍然是非常雏形的阶段。数据的集成，尤其是跨行业、跨不同的部门、跨各种技术集成起来的机会还是非常少的。

EMC 指出："大数据并不是一个准确的术语，相反，它是对各种数据(其中大多数是非结构化的)永不休止的积聚的一种表征。它用以描述那些呈指数级增长，并且因太大、太原始或非结构化程度太高而无法使用关系数据库方法进行分析的数据集。不论是数 TB 的数据量还是数 PB 的数据量，数据的精确数量不如最终结果及数据如何使用来得重要。"

EMC 的大数据定义更强调大数据中的价值，特别是商业价值。大数据之所以流行，其主要的原因就是它能够给企业的核心业务带来直接的价值。具体的讲，大数据能够帮助企业做到以下 3 点：

- 发现新的收入增长点；
- 优化和完善现有的收入或利润空间；
- 获得超过其竞争对手的竞争优势。

上述定义中已经提到大数据有多种特征，其中最具代表性的是 3 个 V。除了上述业内主流的以大数据 3V 特征为基础的定义，还有使用 3S 或者 3I 描述大数据特征的定义。

3S 分别是 Size(大小)、Speed(速度)和 Structure(结构)。实际上，这个维度的特征与 3V 是异曲同工的，除了用词的不同，并没有太大的差别。

大数据的 3I 指的是：

(1) Ill-defined(定义不明确的)：多个主流的大数据定义都强调了数据的规模需要超过传统方法的处理能力。而随着技术的进步，数据分析的效率不断提高，符合大数据定义的数据规模也会相应地不断变大，因而并没有一个明确的标准。

(2) Intimidating(令人生畏的)：从管理大数据到使用正确的工具获取它的价值，利用大数据的过程充满了各种挑战。

(3) Immediate(即时的)：数据的价值会随着时间快速衰减。因此为了保证大数据的可控性，需要通过减少数据收集到获得数据使用之间的时间，使得大数据成为真正的即时大数据。这意味着能尽快地分析数据对获得竞争优势是至关重要的。

总而言之，大数据是个动态的定义，不同行业根据其应用的不同有着不同的理解，其衡量标准也在随着技术的进步而改变。

### 1.1.4　大数据的构成

大数据既是数据量的一个激增(从最开始的 ERP/CRM 数据，逐步扩大到增加互联网数据，再到物联网的传感器等相关信息数据)，同时也是数据复杂性的提升。大数据可以说是量积累到一定程度后形成的规模化质变。大数据的数据类型丰富多样，既有像原有的数据库数据等结构化信息，又有文本、视频等非结构化信息，而且数据的采集和处理速度要求也越来越快。

大数据包含了"海量数据"的含义，在内容上超越了海量数据，简而言之，大数据是"海

量数据”和复杂类型的数据。大数据包括交易和交互数据集在内的所有数据集，其规模或复杂程度超出了常用技术按照合理的成本和时限捕捉、管理及处理这些数据集的能力。

大数据由三类主要数据汇聚组成：

(1) 海量交易数据：在从 ERP 应用程序到数据仓库应用程序的在线交易处理(OLTP)与分析系统中，传统的关系数据以及非结构化和半结构化信息仍在继续增长。随着更多的数据和业务流程移向公共和私有云，这一局面变得更加复杂。内部的经营交易信息主要包括联机交易数据和联机分析数据，是结构化的、通过关系数据库进行管理和访问的静态历史数据。通过这些数据，我们能了解过去发生了什么。

(2) 海量交互数据：这一新生力量由源于 Facebook、Twitter、LinkedIn 及其他来源的社交媒体数据构成。它包括了呼叫详细记录(CDR)、设备和传感器信息、GPS 和地理定位映射数据、通过管理文件传输(Manage File Transfer)协议传送的海量图像文件、Web 文本和点击流数据、科学信息、电子邮件等等，这些数据可以告诉我们未来会发生什么。

(3) 海量数据处理：利用多种轻型数据库来接收发自客户端的数据，并将其导入到一个集中的大型分布式数据库或者分布式存储集群，然后利用分布式数据库对存储于其内的集中的海量数据进行普通的查询和分类汇总等，以此满足用户对大多数常见数据的分析需求，同时对基于前面的查询数据进行数据挖掘，能满足高级别的数据分析需求[8]。例如，YunTable 是在传统的分布式数据库和新的 NoSQL 技术的基础上发展而来的新一代分布式数据库，通过它能构建一个百台级别的分布式集群来管理 PB 级别的海量数据[9]。

### 1.1.5 大数据的机遇和挑战

#### 1. 大数据的机遇

在很多应用领域，数据正以史无前例的规模汇集，与以往基于猜测或模型进行决策不同，如今，人们大多是根据数据本身进行决策。大数据分析现在几乎遍及着社会生活的方方面面，包括移动服务、零售业、制造业、金融服务、生命科学和物质科学等[10]。

大数据给科学研究带来了变革。在天文学领域，天文学家的工作发生了大幅度转变。以前，天文学家的主要工作是进行太空拍照。如今，所有照片都已经存放在数据库中，天文学家的任务变为从数据库中发现有趣的物体或现象。在生物科学领域，目前已经建立起将科学数据存入公共数据存储集的良好传统。生物学家经常创建一些公共数据集供其他领域科学家使用。事实上，生物信息学领域有一个专门的学科致力于整理和分析这些数据。随着技术的进步，特别是新一代基因测序技术的问世，实验数据集越来越多，数据规模也在成倍增长。

大数据在给科学研究带来变革的同时，也为教育带来了变革。最近，一项在纽约市 35 所学校进行的不同教学方法的定量比较发现：使用数据指导教学是五种最有效的教学手段之一。

另外，通过连续监测、提前预防和个性化医疗，信息技术及大数据在降低医疗成本的同时可以提高医疗质量。麦肯锡估计，仅就美国而言，在医疗领域，信息技术的应用每年可节省上千亿美元。

类似地，大数据的价值还体现在多个方面，例如：城市规划(通过融合高清晰度的地理数据)、智能交通(通过分析与可视化现场的详细的道路网络数据)、环境建模(通过无处不在的传感器网络收集数据)、能源节省(通过发现使用模式)、智能材料(通过基因组计划发现新材料)、社会计算(由于获取数据的成本降低，该类方法越来越受欢迎)、金融风险分析(通过合同网络的综合分析寻找金融实体之间的依赖关系)、国土安全(通过分析潜在的恐怖分子的社交网络和金融交易有效保护国家安全)、计算机安全(通过分析日志信息和其他事件，掌握有关安全信息并可进行事件管理)等。

2010 年，企业和用户存储了超过 13 EB($13\times10^{18}$ B)的新数据，这比美国国会图书馆的数据多 5 万多倍。根据麦肯锡公司最近的一份报告，对于最终用户而言，全球个人定位数据的潜在价值估计为 7000 亿美元，可导致减少高达 50%的产品研发和组装费用。麦肯锡预测大数据将对就业产生很大影响，在美国，大约需要 14 万至 19 万名具有“深度分析”经验的分析师；此外，还需要 150 万懂得与数据打交道的管理人员。无独有偶，美国总统科技顾问委员会关于网络和信息技术研发的报告，将大数据定位为能够“促进优先发展”的“研究前沿”。另外，大数据还可以为很多商业提供如下服务：

(1) 精准广告投放。大数据最本质的应用就在于预测，即从海量数据中分析出一定的特征，进而预测未来可能会发生什么。大数据有数据量大、数据多样性等特征，实际是将各个维度的数据进行综合分析进而进行一定的预测。当不同的数据流被整合到大型数据库中后，预测的广度和精度都会大规模地提高。

(2) 医疗卫生体系更加精密。通过分析大量用户的搜索记录，比如“咳嗽”、“发烧”等特定词条，谷歌公司能准确预测美国冬季流感传播趋势。与官方机构相比，谷歌能提前一两周预测流感爆发，预测结果与官方数据的相关性高达 97%。

(3) 个性化教育可能真正实现。在传统教育模式下，分数就是一切，一个班上几十个人，使用同样的教材，同一个老师上课，课后布置同样的作业。然而，学生的素质是千差万别的，在这个模式下，不可能真正做到“因材施教”。

**2. 大数据的挑战**

大数据的潜在好处显而易见，在某些方面甚至已初见成效。但要充分发挥这一潜力仍需面对许多技术性挑战，这也是一个很大的问题。大数据分析的过程涉及多个不同阶段，每个阶段都存在挑战性。

面对大数据的汹涌来袭，传统的数据处理方式应对起来显得越来越困难，我们在很多时候就像面对一个金矿，却没有有效的工具和手段，只能望“数据”兴叹。传统分析技术面对大数据的困惑主要有：

(1) 由于分析手段限制，不能充分利用所有数据；

(2) 受限于分析能力而无法获取复杂问题的答案；

(3) 因为时限要求而不得不采用某项简单的建模技术；

(4) 因为没有足够时间运算，只好对模型精度进行妥协。

对于这些大数据的挑战，其实归纳起来只有两个目标：管理好大数据，从大数据的产生、存储、保护、归档到安全维护的各个角度，当数据量超出常规管理尺度后，对于管理维护的难度出现了跳跃式上升的态势；使用好大数据，这是我们管理的最终目标。大数据即

意味着大价值，数据与数据、数据与人、数据与事件(业务)之间都具有关联性。这些挑战既有流动性、关联性、智能的应用挑战，也有基于大数据深度挖掘的挑战。但是，这两个目标之间也不是分离的，而是相辅相成的关系，管理和维护的目的是使用，使用的基础是好的管理维护。在技术领域中，大数据对我们提出了以下一些挑战。

1) 对技术架构的挑战

对现有数据库管理技术的挑战。传统的数据库部署不能处理数个大字节(TB)级别的数据，也不能很好地支持高级别的数据分析。急速膨胀的数据体量即将超越传统数据库的管理能力。对技术架构的挑战包括如何构建全球级的分布式数据库(Globally-Distributed Database)，使之可以扩展到数百万台机器，数以百计的数据中心，上万亿的行数据。经典数据库技术并没有考虑数据的多类别(variety)，SQL(结构化数据查询语言)在设计的一开始也没有考虑非结构化数据。

2) 对实时性的技术挑战

一般而言，像数据仓库系统、商业智能应用，对处理时间的要求并不高。因此这类应用往往运行一两天获得结果依然是可行的。但实时处理的要求是区别大数据应用和传统数据仓库技术、商业智能技术的关键差别之一。

3) 对数据存储及软硬件的挑战

人们每天创建的数据量正呈爆炸式增长，但就数据保存来说，现有的技术改进不大，而数据丢失的可能性却不断增加。如此庞大的数据量首先在存储上就会是一个非常严重的问题，硬件的更新速度将是大数据发展的基石。

4) 对分析技术的挑战

传统意义上的数据分析主要针对结构化数据展开，并已经形成了一整套行之有效的分析体系。通过数据库来存储结构化数据，用数据挖掘的聚类、关联分析等技术梳理、分析、提炼、获取进一步层面的知识，这一系列的方法在处理一般结构化数据时极为高效，但在处理大数据的过程中，由于非结构化数据、半结构化数据量的极大增长，给传统的分析技术带来了巨大的冲击和挑战，主要体现在以下几个方面：

(1) 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往往也在衰减，因此很多领域对于数据的实时处理有需求。随着大数据时代的到来，更多应用场景的数据分析从离线转向了在线，开始出现实时处理的需求。大数据时代的数据实时处理面临着一些新的挑战，主要体现在数据处理模式的选择及改进上。在实时处理的模式选择中主要有三种思路，即流处理模式、批处理模式以及二者的融合。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。

(2) 动态变化环境中索引的设计。关系数据库中的索引能够加快查询速度，但是传统的数据管理中模式基本不会发生变化，因此在其上构建索引主要考虑的是索引创建、更新等的效率。大数据时代的数据模式随着数据量的不断变化可能会处于不断的变化之中，这就要求索引结构的设计简单、高效，能够在数据模式发生变化时很快地进行调整来适应。

在数据模式变更的假设前提下设计新的索引方案将是大数据时代的主要挑战之一。

(3) 先验知识的缺乏。传统分析主要针对结构化数据展开，这些数据在以关系模型进行存储的同时就隐含了这些数据内部关系等先验知识。比如我们知道所要分析的对象会有哪些属性，通过属性我们又能大致了解其可能的取值范围等。这些知识使得我们在数据分析之前就已经对数据有了一定的理解。而在面对大数据分析时，一方面是半结构化和非结构化数据的存在，这些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，这些需要实时处理的数据很难有足够的时间去建立先验知识。而无先验知识的数据更需要发现知识。

**3. 应对大数据挑战**

针对技术领域的挑战，科技工作者取得了很多研究成果。现有面向大数据的研究主要针对存储、处理、分析、可视化等某一方面的关键技术。在大数据存储方面，已有研究主要集中在各类 NoSQL 和分布式文件系统。随着互联网和云计算的不断发展，各种类型的应用层出不穷，对数据库技术提出了更多要求，主要体现在以下方面：

(1) 高并发读写需求。

(2) 海量数据的高效存储和访问需求。

(3) 高可扩展性和高可用性需求。

为了满足以上需求，出现了非关系型数据库(NoSQL)。典型的 NoSQL 数据库有 Redis、Memcached、Cassandra、MongoDB、Neo4j 等。NoSQL 虽然具有多方面优势，但是其最大的弱点就是不支持 SQL 查询，这为开发人员带来诸多不便。因此，为了同时满足高性能和支持 SQL 两方面的需求，一种全新的关系数据库产品 NewSQL 被设计出来，它或者通过将关系模型的优势与分布式体系结构结合，或者将关系数据库的性能提升到不必进行横向扩展的程度。典型的 NewSQL 有 VoltDB、Marklogic、Xeround、NuoDB 等；在大数据处理技术方面，最主流的平台是 Hadoop。Hadoop 由分布式文件系统 HDFS、并行计算框架 MapReduce 和非结构化数据 Hbase 组成，它们分别是 Google GFS、Google MapReduce 和 Google BigTable 的开源实现。HDFS 具有高容错性，因此适合部署在价格低廉的硬件上，同时它还适用于具有超大数据集的应用程序；在大数据分析方面，代表性研究有 Hive、Pig 等，Facebook 等公司在实时分析方面也进行了相关研究。

随着数据爆炸式的增长使得我们的时代进入到了真正的大数据时代，急需功能强大的工具和方法从这些海量数据中挖掘出数据内在的、深层次的、有价值的信息，把这些数据转化成为有组织的知识。如何有效利用这些海量数据中的信息来增强人们获取知识的能力，从而进一步加快生产力的发展，是目前及以后相当长时期内，全球科学与技术专家所面临的共同问题之一。数据挖掘(DM，Data Mining)、模式识别、机器学习、统计学习等就是一种有效的知识发现技术。

### 1.1.6　大数据应用的发展方向

美国政府在 2012 年 3 月 29 日宣布投资两亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家意志。美国奥巴马政府在白宫网站发布《大数据研究和发展倡议》，提出