第 1 章　大数据概述

本章要点：

- 大数据的概念
- 大数据的结构类型
- 大数据的特征
- 大数据的关键技术
- 大数据的计算模式
- 大数据的应用
- 大数据的发展
- 大数据的意义

大数据是继移动互联网、物联网、云计算之后出现的新流行词语，已成为科技、企业、学术界关注的热点，被人们认为是继人力、资本之后的一种新的非物质生产要素，蕴含着极其巨大的价值。大数据将改变人类社会认识自然和宇宙的方式的广度及深度，大数据科学及相关工具使人类了解和利用自然变得更加全面、更加细化。

国际两大权威杂志 *Science* 和 *Nature* 专门出版了有关大数据的专刊，探讨大数据所带来的机遇以及挑战等问题。美国著名的未来学家阿尔文·托夫勒在《第三次浪潮》一书中，将大数据称为第三次浪潮的华彩乐章。大数据给人类社会的发展变化带来了机遇和挑战，人们必须要面对并迎接这些挑战，从中发现机会，抓住机会，利用机会，从而不断地适应和推动社会的发展进步。

1.1　大数据的概念

对大数据(big data)这个术语最早的引用可追溯到 Apache 的开源项目 Nutch，当时被用来描述在更新网络搜索引擎时需要同时批量处理或分析大量数据集。目前大数据已成为最流行的 IT 词汇，引领着各个应用领域的新一轮创新浪潮。Lisa Arthur 在《大数据营销》一书中将大数据表述为纷繁杂乱的、互动的应用程序、信息和流程。对于大数据概念的表述，不同的学者和机构给出的定义也不相同，目前比较权威的表述主要有麦肯锡、维基百科、IBM 公司、大数据研究机构高德纳(Gartner)、国际数据中心(IDC)以及美国国家标准技术研究院(NIST)等。多家权威机构针对大数据的数据体量大、数据类型繁多、价值密度低以及速度快等不同特征进行了不同的阐述。

尽管多家权威机构对大数据的定义并不相同。但综合多个表述，可以将大数据定义为：

大数据是指海量数据，是数据和大数据技术的综合体，既包括结构化、半结构化数据，还包括非结构化的数据。大数据具有种类繁多的信息价值，无法用目前的主流软件工具在一定的时间内去采取、分析、处理及管理海量的信息。大数据是信息产业持续高速增长的新引擎，在硬件与集成设备领域，大数据将对芯片、存储产业产生重要影响，还催生出一体化数据存储服务器、内存计算等市场。面向大数据市场的新技术、新产品、新服务会不断涌现。在软件与服务领域，大数据将引发数据快速处理分析技术、数据挖掘技术和软件产品的发展。大数据将成为许多行业提高核心竞争力的关键因素，各行各业的决策正在从业务驱动向数据驱动转变。在商业领域，对大数据的分析可以使零售商实时掌握市场动态并迅速做出应对，可以为商家制订更加精准有效的营销策略提供决策支持，可以帮助企业为消费者提供更加及时和个性化的服务。

1.1.1 数据与信息

从计算机科学的角度来看，数据是所有能输入到计算机并被计算机程序处理的符号的总称，是具有一定意义的数字、字母、符号和模拟量的统称。数据的形式具有多样化的特征，数据可以表现为连续的值，如声音和图像等；也可以是离散的，如符号和文字等。在计算机系统中，数据以二进制信息单元 0 和 1 的形式来表示。要保证数据的原始性和真实性，但很多数据通过后期加工才会变得有意义。信息是人们为了某种需求而对原始数据加工重组后所形成的有意义和有用途的数据。

数据与信息既有联系，又有所区别。数据是信息的表现形式和载体；信息是数据的内涵，它会加载于数据之上，并对数据作特定含义的解释。数据与信息是不可分离的，信息依赖数据来体现，数据则生动具体地表达出信息。数据是符号，具有物理性；信息是对数据进行加工处理之后所得到的对决策产生影响的内容，是具有逻辑性和观念性的。数据是信息的表达、载体，信息是数据的内涵，两者是形与质的关系。数据本身没有意义，数据只有对实体行为产生影响时才成为信息。在信息的基础上提炼和总结成具有普通指导意义的内容称为知识；知识可以进一步总结归纳成更普世的规律，可演化为更多的知识来指导客观实践，此时可称为智能。从数据到智能的过程，不仅仅是人们认识程度的提升，同时也是从部分到整体、从描述过去到预测未来的过程。数据、信息、知识和智能的演化过程如图 1.1 所示。

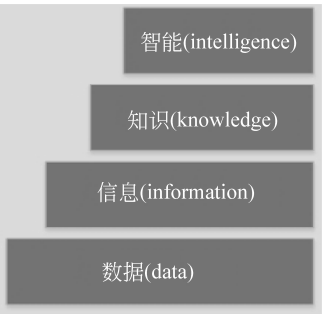

图 1.1　从数据到智能的阶梯

1. 数据

数据是客观对象的表示，是能够客观反映事实的数字和资料。《韦伯斯特大辞典》(*Merriam-Webster Dictionary*)将数据定义为：用于计算、分析或规划某种事物的事实或信息，并由计算机产生或存储。计算机对数据的处理，需要先对数据进行表示和编码，从而衍生出不同的数据类型。数据经过加工就会成为信息。根据数据所刻画的过程、状态和结构的特点，数据可以划分为不同的类型。数据是各种符号或原始事实，如数字、图片、声音、动

画以及视频等，数据要经过后期加工才有意义。数据体现的是一种过程、状态或结果的记录，这类记录经过数字化转化后可以被计算机存储和处理。

数据按照表现形式可以分为模拟数据和数字数据。模拟数据主要有声音和图像等；数字数据主要有文字和符号等，如在计算机系统中，数据以二进制信息单元 0 和 1 的形式来表示。

数据按照性质可以分为定位的数据、定量的数据、定时的数据和定性的数据，其中，定位的数据如个人运动轨迹的定位数据、店铺的定位数据等；定量的数据是以数量形式存在的属性，并能够反映事物数量特征的数据，如距离、质量、面积等；定时的数据是反映事物时间特性的数据，如时、分、秒等；定性的数据可以是表示事物属性的数据，如河流和道路等。

2. 信息

信息是对客观世界中各种事物的运动状态和变化的反映，是客观事物之间相互联系和相互作用的表征，是包含在数据之中的能够被人脑理解和进行思维推理的内容，表现的是客观事物运动状态和变化的实质内容。信息一词在我国古代称为消息，而在英文、法文、德文和西班牙文中均是 Information，在日文中为“情报”，在我国台湾则称为“资讯”。信息论的创始人克劳德·艾尔伍德·香农(Claude Elwood Shannon)对信息的定义给出了精确的表述，他认为信息是用来消除随机不确定性的东西。控制论的创始人诺伯特·维纳(Norbert Wiener)认为信息是人们在适应外部世界，并使这种适应反作用于外部世界的过程中，同外部世界进行互相交换的内容和名称。信息的特点为：没有大小和质量，容易复制。没有大小是指无论怎样小的空间，都可以存放大量的信息；无论怎样狭窄的通道，都能高速地传递大量的信息。没有质量是指信息没有重量，在处理时不需要能量。容易复制是指信息一旦产生，就很容易复制。

1.1.2　大数据的定义

关于大数据的定义，目前在学术界还未形成统一的标准化表述，比较被人们所接受的有以下几种表述。

大数据研究机构高德纳将大数据定义为：需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

国际数据中心将大数据定义为：大数据技术描述了一个技术和体系的新时代，被设计成用于从大规模、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。

麦肯锡将大数据定义为：大数据指的是其大小超出了传统软件工具的采集、存储、管理和分析等能力的数据集，具有海量的数据、快速的数据处理、多样的数据类型和价值密度低等多个特征。

维基百科将大数据定义为：大数据指的是所涉及资料量的规模庞大到无法通过目前主流的软件工具在合理时间内达到捕获、管理、处理并整理成为帮助企业经营决策更积极目的的资讯。

美国国家标准技术研究院(NIST)将大数据表述为：具有规模大、多样化、时效性和多变性等特性，需要具备可扩展性的计算架构来进行有效存储、处理和分析的大规模数据集。

综上所述，在大数据的定义中除了要关注它规模大、多样性化、时效性和多变性等特性

以外，还应关注它需要具备可扩展性的计算架构来进行有效存储、处理和分析。

大数据从本质上来讲，包含速度、数量和类型三个维度的问题。大数据的本质构建如图 1.2 所示，图中的速度主要是指对海量数据的处理速度，可实现海量数据的实时处理；数量是指数据量由 PB 级到 ZB 级；类型主要是指数据种类繁多，已打破了传统的仅仅为结构化数据的范畴，数据的处理还包括了海量半结构化和非结构化数据。

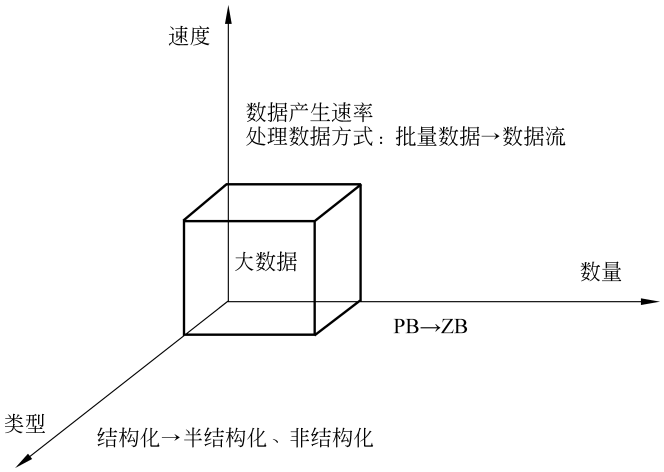

图 1.2　大数据的本质构建

1.2　大数据的结构类型

大数据需要特殊的技术，并利用特殊的数据结构来组织和访问巨大数量的数据，以便有效地处理跨多个服务器和离散数据存储的数据。按照数据是否有较强的结构模式，可将其划分为结构化数据、半结构化数据和非结构化数据，如表 1.1 所示。

表 1.1　结构化数据、半结构化数据、非结构化数据

数据类型	数据形成过程	数据特征	数据模型	不同点
结构化数据	先有结构，再有数据	由二维表结构来逻辑表达和实现，严格地遵循数据格式与长度规范，主要通过关系型数据库进行存储和管理	二维表（关系型）	多表现为行数据，存储在数据库里，可以用二维表结构来进行逻辑表达和实现
半结构化数据	先有数据，再有结构	数据结构描述具有复杂性、动态性的特点	树、图	介于完全结构化数据和完全无结构的数据之间
非结构化数据	先有数据，再有结构	数据结构不规则或不完整，没有预定义的数据模型，不方便用数据库二维逻辑表来表现		没有固定的数据结构，且不方便用数据库二维逻辑来表现，如存储在文本文件中的系统日志、文档、图形图像、音频和视频等数据

1. 结构化数据

结构化数据简单来说就是存储在结构化数据库里的数据，可以用二维表结构来逻辑表达并实现，如财务系统、企业 ERP、教育一卡通、政府行政审批等。

2. 半结构化数据

半结构化数据主要是指介于完全结构化数据（如关系型数据库和面向对象数据库中的数据）和完全无结构的数据（如声音、图像文件等）之间的数据。半结构化数据也具有一定的结构，但是不会像关系数据库中那样有严格的模式定义，其数据形成过程是先有数据，再有结构。半结构化数据使用标签来标识数据中的每个元素，通常数据组织成有层次的结构。常见的半结构化数据主要有 XML 文档和 JSON 数据，此外，还有 HTML 文档、电子邮件和教学资源库等。

3. 非结构化数据

非结构化数据主要是指没有固定的数据结构，且不方便用数据库二维逻辑来表现的数据。非结构化数据没有预定义的数据模型，因此，它覆盖的数据范围更加广泛，涵盖了各种文档。如存储在文本文件中的系统日志、文档、图形图像、音频和视频等数据，都属于非结构化数据。

1.3　大数据的特征

大数据的产生方式为主动生成数据，即利用大数据平台对需要分析事件的数据进行密度采样，从而精确获取事件的全局数据。大数据的数据源可以利用大数据技术，并通过分布式文件系统、分布式数据库等技术，对从多个数据源获取的数据进行整合处理。在数据处理方式上，大数据中较大的数据源、响应时间要求低的应用，可以采取批处理方式集中计算；而响应时间要求高的实时数据处理，采用流处理的方式进行实时计算，并通过对历史数据的分析进行数据预测。大数据可以依托云计算的分布式处理、分布式数据库、云存储和虚拟化技术对海量数据进行分析、处理。

大数据是一类反映物质世界和精神世界运动状态和状态变化的资源，具有功能多样性、决策有用性、应用协同性、可重复开采性以及安全风险性等特点。大数据的特征主要可以用 6 个 V 和 1 个 C 来概括，6 个 V 是指 volume（容量）、variety（种类）、velocity（速度）、value（价值）、veracity（真实性）、variability（可变性），1 个 C 是指 complexity（复杂性）。大数据的特征如图 1.3 所示。

1. 容量

容量主要是指数据量大，来源的渠道多。大数据通常是指 1PB（1PB＝1024TB）以上的数据。数据体量巨大是大数据的主要特征。根据著名的咨询机构 IDC 的估测，人类社会产生的数据一直都在以每年 50％的速度增长，也就是说，每两年就增加一倍，这被称为“大数

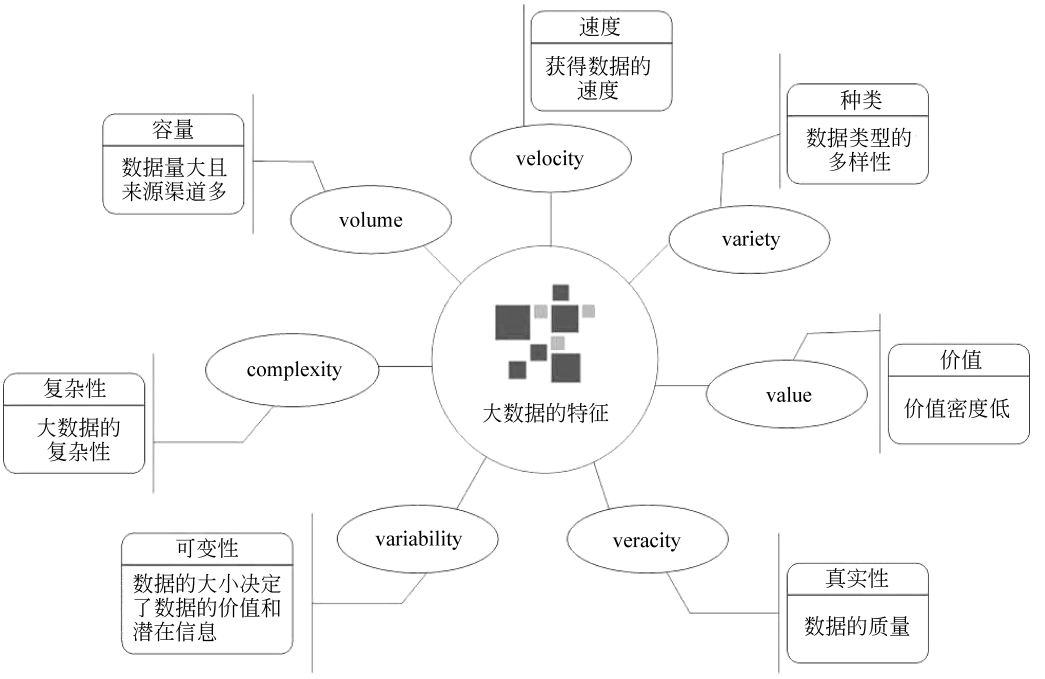

图 1.3　大数据的特征

据摩尔定律”。

2. 种类

种类是指数据类型的多样性，并表示所有的数据类型。除了结构化数据外，大数据还包括各类半结构化数据和非结构化数据，如电子邮件、办公处理文档、互联网上的文本数据、点击流量、文件记录、位置信息、传感器数据、音频和视频等。大数据的类型按照时效性还可分为在线实时数据和离线非实时数据；按照数据来源可分为个人数据、商业服务数据、社会公共数据、科学数据、物质世界数据、教育数据、医疗数据等；按照关联特性又可分为关联型数据和非关联型数据；按照数据类型则又可分为语音、图片、文字、动画、视频等类型。由于大数据来源的种类具有多样性和异构性的特点，因而会使后期海量数据的存储、分析、处理、查询及管理等变得更加困难，需要借助专业的大数据分析与软件处理工具才可解决。

3. 速度

速度是指获得数据的速度。大数据的计算处理速度是可用性和效益性的一个重要衡量指标，大数据的时效性要求对数据的处理能够做到实时和快速。要达到这一目标，要求使用的硬件平台能够及时更新换代，并将分布式计算、并行计算、软件工程以及人工智能等技术应用到其中。

4. 价值

价值是指价值密度低。互联网充斥着大量重复和虚假的信息，通常有价值的信息较为

分散，密度很低。大数据的价值具备稀疏性、多样性和不确定性等特点。许多数据采集和存储系统要求能够快速访问大数据的历史版本数据，要求备份数据的保存期限更长，但备份的时间不断缩短，甚至很多数据需要在线备份和实时对故障进行恢复等。大数据的安全维护对存储资源、计算资源、网络资源等提出了极高的性能要求。应合理利用大数据，并以低成本创造高价值。

5. 真实性

真实性是指数据的质量。数据真实性指数据中的内容与真实世界是紧密相关的，因此，研究大数据就是要从海量的网络数据中提取出能够解释和预测现实事件的过程。随着网络数据、社交数据、电信数据、医疗数据、金融数据、教育数据以及电商数据等新兴数据源的兴起，传统数据源的局限性被打破，多个行业和领域需要有效和真实的信息来确保数据的真实性和质量。数据的真实性和质量是获得真知和思路的最重要因素，是成功制订决策最坚实的基础。

6. 可变性

可变性是指数据的大小决定了所考虑数据的价值和潜在信息。可变性也指由于大数据具有多层结构，因此会呈现出多变的形式和类型。由于大数据的可变性、不规则以及模糊不清的特性，会导致无法用传统软件来对海量数据进行分析与处理。

7. 复杂性

复杂性是指大数据比较复杂，这是由于大数据的数据量较大且来源渠道较多，这是有别于传统数据的根本。大数据的复杂性主要表现在结构的复杂性、类型的复杂性和内在模式的复杂性等多个方面，从而使大数据的采集、分析与处理等变得困难。

1.4　大数据的关键技术

大数据时代，大数据技术在全世界范围内发展迅猛，全球学术界、工业界以及各国政府都给予了大数据高度的关注和重视，掀起了一场可与 20 世纪 90 年代的信息高速公路相提并论的发展热潮。大数据技术已被多国政府提升到国家重大发展战略的高度。大数据是未来的“新石油”，将大数据上升为国家发展战略，将会给未来的科技与经济发展带来重大影响。如美国政府在 2012 年由总统奥巴马签署了大数据研究发展创新计划(big data R&D initiative)，并投资两亿美元启动大数据技术和工具研发。英国、法国、德国以及日本等国家随后也纷纷推出了相应的大数据发展战略计划。大数据巨大的应用需求和隐含的深度价值极大地推动了大数据技术的快速发展，促进了大数据所涉及的各个技术层面和系统平台方面的长足发展。

大数据主要有 7 个关键技术，主要包括大数据采集技术、大数据预处理技术、大数据存储和管理技术、大数据处理技术、大数据分析和挖掘技术、大数据可视化技术、大数据安全和加密技术。

1. 大数据采集技术

大数据采集技术将分布在异构数据源或异构采集设备上的数据通过清洗、转换和集成技术，存储到分布式文件系统中，成为数据分析、挖掘和应用的基础。大数据采集技术是获取有效数据的重要途径，是大数据应用的重要支撑。大数据采集技术是在确定用户目标的基础上，针对该范围内所有结构化、半结构化和非结构化数据进行采集并处理。大数据采集技术与传统数据采集技术有很大的不同，传统数据采集的数据来源较为单一，数据量相对较小，但大数据采集的数据来源比较广泛且数据量巨大。传统数据采集的数据类型及结构简单；大数据采集的数据类型丰富，除了包括结构化数据，还包括半结构化数据和非结构化数据。传统数据采集中的数据处理使用关系型数据库和并行数据仓库，大数据采集中的数据处理使用分布式数据库。

2. 大数据预处理技术

大数据的多样性决定了通过多种渠道获取的数据种类和数据结构都非常复杂，这就给之后的数据分析和处理带来了极大的困难，通过大数据的预处理这一步骤，将这些结构复杂的数据转换为单一的或便于处理的结构，为后面的数据分析与处理打下良好的基础。

大数据预处理技术主要是指完成对已接收数据的辨析、抽取、清洗、填补、平滑、合并、规格化及检查一致性等操作。因获取的数据可能具有多种结构和类型，数据抽取的主要目的是将这些复杂的数据转化为单一的或者便于处理的结构，以达到快速分析处理的目的。要实现对巨量数据进行有效的分析，需要将来自前端的数据导入到一个集中的大型分布式数据库或分布式存储集群里，且能够在导入的基础上做一些简单的清洗和预处理。

大数据预处理的方法主要包括数据清洗、数据集成、数据变换和数据归约。数据清洗是在汇聚多个维度、多个来源、多种结构的数据之后，对数据进行抽取、转换和集成加载；数据集成是将大量不同类型的数据原封不动地保存在原地，而将处理过程适当地分配给这些数据；数据变换是将数据转换成适合挖掘的形式，并采用线性或非线性的数学变换方法，将多维数据压缩成较少维数的数据，消除它们在时间、空间、属性及精度等特征表现方面的差异；数据归约是从数据库或数据仓库中选取并建立使用者感兴趣的数据集合，然后从数据集合中过滤掉一些无关、有偏差或重复的数据。

3. 大数据存储和管理技术

大数据存储和管理技术的主要目标是用存储器把采集到的数据存储起来，建立相应的数据库，并进行管理和调用。大数据存储是数据处理架构中进行数据管理的高级单元，其功能是将按照特定的数据模型把组织起来的数据集合进行存储，并提供独立于应用数据的增加、删除、修改能力。

4. 大数据处理技术

大数据处理技术是对海量的数据进行处理，主要的处理模式有批处理模式和流处理模式。批处理模式是先存储后处理，如谷歌公司在 2004 年提出的 MapReduce 编程模型就是

典型的批处理模式。流处理模式与批处理模式不同，采用的是直接处理。流处理模式的基本理念是数据的价值会随着时间的流逝而不断减少，因此，要尽可能快地对最新的数据做出分析并给出结果。流处理模式将数据视为流，将源源不断的数据组成数据流，当新的数据到来时就立即处理并返回所需要的结构。需要采用流处理模式的大数据应用场景包括传感器网络、金融中的高频交易和网页点击数的实时统计等。

5. 大数据分析和挖掘技术

大数据时代，随着 5G 移动技术、在线学习、深度学习、人工智能、物联网、机器学习和云计算、移动计算、分布式计算、并行计算、批处理计算、边缘计算、流计算、图计算以及区块链等新技术的不断涌现，教育、科研、医疗、通信和电商等多个领域数据量的增加呈现出几何级数增长的态势，激增的数据背后隐藏着许多重要的信息，如何对其进行更加智能的分析，以便更好地利用这些数据挖掘出其背后隐藏的有价值的信息，是当前研究的热点问题。

(1) 大数据分析。大数据分析是大数据处理的核心，只有通过分析才能获取更多智能的、深入的和有价值的信息。大数据的分析方法在大数据领域比较重要，是决定最终信息是否有价值的关键，利用数据挖掘进行数据分析常用的方法主要有分类、回归分析、聚类、关联规则等。大数据分析的数据源除了传统的结构化数据，还包括半结构化和非结构化数据，针对不同的数据源可采用数据抽取，统计分析以及数据挖掘等多个步骤来进行分析与处理，以快速挖掘出有用信息，洞悉出数据价值。

(2) 大数据挖掘。主要从以下 6 个方面进行介绍。

① 数据挖掘。数据挖掘(data mining，DM)是数据库知识发现中的一个步骤，是指通过算法从大量的数据中搜索出隐藏于其中的信息的过程。数据挖掘又称为数据库中的知识发现(knowledge discover in database，KDD)，就是从大量的、不完全的、有噪声的、模糊的甚至随机的实际应用数据中，提取出隐含在其中的、人们事先不知道的但又潜在有用的信息和知识的过程。数据挖掘所挖掘的知识类型包括模型、规律、规则、模式和约束等。数据挖掘方法利用了来自多个领域的技术思想，如来自统计学的抽样、估计和假设检验，来自人工智能、模式识别和机器学习的搜索算法、建模技术及学习理论，来自包括最优化、进化计算信息论、信号处理、可视化和信息检索等方面的相关方法。

数据挖掘是一种决策支持过程，主要是基于人工智能、机器学习、模式识别、统计学、数据库和可视化等技术，自动分析企业的数据，做出归纳性的推理，从中挖掘出潜在的模式，为决策者调整市场策略及减少风险并做出正确的决策提供知识支持。数据挖掘的一般流程为：定义问题→数据准备→确定主题→读入数据并建立模型→挖掘操作→结果表达和解释。其中，定义问题是数据挖掘过程的第 1 个步骤，清晰地定义出业务的问题，认清数据挖掘的目的，是数据挖掘的重要一步。数据准备是数据挖掘的第 2 个步骤，可分为 3 个子步骤，分别为数据集成、数据选择以及数据预处理。数据集成将多文件或多数据库运行环境中的数据进行合并处理，解决语义模糊性，处理数据中的遗漏等；数据选择的目的是辨别出需要分析的数据集合，缩小处理范围，提高数据挖掘的质量；数据预处理是为了克服数据挖掘工具的局限性，提高数据质量，同时将数据转换成为一个适用于特定挖掘算法的分析模型。确定主题是数据挖掘过程的第 3 个步骤；读入数据并建立模型是继确定主题后的第 4 个步骤，主要是指在确定输入的数据之后，再用数据挖掘工具读入数据并从中构造出一个模型。

挖掘操作是数据挖掘的第 5 个步骤，是在前面的准备工作完成后，利用选好的数据挖掘工具在数据中查找。数据挖掘的最后一个步骤是结果表达和解释，即根据最终用户的决策目标对提取出的信息进行分析，把最有价值的信息区分出来，并通过决策支持工具交给决策者。

数据挖掘的任务主要有 4 种，分别为聚类分析、预测建模、关联分析和异常检测。聚类分析旨在发现紧密相关的观测值组群，使得与属于不同簇的观测值相比，属于同一簇的观测值相互之间尽可能类似，这也是将物理或抽象对象的集合分组成为由类似的对象组成的多个类的分析过程；聚类分析主要针对的数据类型包括区间标度变量、二元变量、标称变量、序数型变量、比例标度型变量以及由这些变量类型构成的复合类型。预测建模主要是以说明变量函数作为目标变量来建立模型。关联分析是用来发现描述数据中强关联特征的模式。异常检测是识别特征显著区别于其他数值的观测值。

数据挖掘的功能是指定数据挖掘任务的发现模式，可以将这些任务分为描述性的和预测性的。描述性数据挖掘可用于刻画目标数据的一般性质；预测性数据挖掘在当前数据上进行归纳，以便做出预测。常见的数据挖掘功能包括聚类、分类、关联分析、数据总结、偏差检测和预测等。其中，分类和预测可以作为预测性任务，其他的可以作为描述性任务。

数据挖掘运用的技术主要有统计学、机器学习、数据库与数据仓库、信息检索以及可视化。统计学主要研究数据的收集、分析、解释和标识；机器学习主要考查计算机如何基于数据学习；数据库与数据仓库主要是指数据挖掘能够利用可伸缩的数据库技术，以便获得在大型数据集上的高效率和可伸缩性；信息检索是指搜索文档或文档中信息的技术，其中，文档可以是结构化文本数据或非结构化多媒体数据，并且可能驻留在 Web 上；可视化是指数据的采集、提取和理解是人类感知和认识世界的基本途径。

② 大数据挖掘方法。大数据挖掘与传统数据挖掘有很大不同，大数据挖掘是在一定程度上降低了对传统数据挖掘模型以及算法的依赖，降低了因果关系对传统数据挖掘结果精度的影响，能够在最大程度上利用互联网上记录的用户行为数据进行分析。大数据挖掘方法主要有数据预处理技术、关联规则挖掘、分类、聚类分析、孤立点挖掘、数据演变分析、社会计算、知识计算、深度学习和特异群组挖掘等。其中，数据预处理技术能够有效提高数据挖掘的质量，进行异常数据清除，使其格式标准化；关联规则挖掘能够使项与项之间的关系在数据集中易于发现；分类是找出一组能够描述数据集合典型特征的模型，以便能够分类识别未知数据的归属或类别；聚类分析便于将观察到的内容分类编制成类分层结构，把类似的时间组合在一起；孤立点挖掘通常又称为孤立点数据分析，孤立点可以使用统计试验检测，是数据挖掘中的主要方法；数据演变分析是指对随时间变化的数据对象的变化规律和趋势进行建模描述；社会计算是由 Schuler 提出的，是大数据挖掘的新方法；知识计算是当前比较新的一种大数据挖掘方法；深度学习主要应用在计算机视觉、自然语言处理和生物信息学等方面，是当前研究的热点，是比较新的一种数据挖掘方法；特异群组挖掘是一种比较好的大数据挖掘方法，该挖掘方法可以应用在智能交通、生物医疗以及银行金融等多个领域。大数据挖掘方法如图 1.4 所示。

大数据时代，多源异构数据不断涌现，通过利用新的大数据挖掘方法（如特异群组挖掘和孤立点挖掘等），可以有效地实现数据挖掘，挖掘出数据背后隐藏的有用的价值信息。

③ 大数据挖掘类型。大数据挖掘类型主要有流数据挖掘、空间数据挖掘以及 Web 数据挖掘等多个类型。流数据挖掘是大数据挖掘类型中较为常见的一种类型，流数据挖掘主