模块 7

大数据分析和挖掘

学习要点

- 数据仓库的特点。
- 大数据挖掘的概念及算法。
- 大数据分析方法及应用。

7.1 必 备 知 识

7.1.1 大数据分析概述

随着计算机技术及网络信息技术的融合并全面进入人们的生活，信息爆炸已经积累到了一个开始引发变革的程度。它不仅使世界出现了比以往更多的信息，而且其增长速度也在加快，进而创造出了“大数据”这个概念。如今，大数据的概念几乎应用到了所有人类智力与发展的领域中。

1. 大数据背景下的新型数据库

传统数据管理方法的局限性及大数据的现实条件促使新的数据库设计的出现，在新的数据库设计中，原本数据库模式中存在的记录和预设场域（常规数据的整齐排列）的规律被替代，同时最普遍的数据库查询语言，即 SQL 结构化查询语言也充分体现出其局限性。为适应近年信息发展的需要，非关系型数据库作为一种新型数据库出现，它不需要预先设定记

录结构，同时允许处理规模庞大、结构复杂的数据。但是，正是因为包容了数据结构的多样性，新型的数据库设计就需要相对更多的数据处理和存储资源。同时，考虑到对于数据处理和存储成本降低的需求，这种大的新型数据库往往并不是固定在某个地方的，它一般分散在多个硬盘和多台计算机上。为了确保其运行的稳定和速度，一个记录可能会分开存储在两三个地方。如果某个地方的记录更新了，其他地方的记录则只有同步更新才不会产生错误。传统的系统会一直等到所有地方的记录都更新，然而，当数据广泛地分布在多台服务器上且服务器每秒钟都会接受成千上万条搜索指令时，同步更新就比较不现实了。因此，数据结构的多样性成为一种解决的方法。在这种新型数据库设计中，非常典型的就是 Hadoop 处理平台。Hadoop 是与谷歌的 MapReduce 系统相对应的一种开源式分布系统的基础架构，它非常善于处理超大量的数据。它通过把大数据变成小模块然后分配给其他计算机进行分析，进而实现对超大量数据的处理。它在设计时充分考虑到硬件可能会瘫痪的可能性，所以在内部建立了数据的副本，同时它还假定数据量之大导致数据在处理之前不可能整齐排列。典型的数据分析需要经过“萃取、转移和下载”这样一个操作流程，但是 Hadoop 不拘泥于这样的方式。相反，它假定了数据量的巨大使得数据完全无法移动，所以人们必须在本地进行数据分析。但是，大数据自身的多样性和复杂性也决定了绝对不存在一套完全标准规范的解决方案。虽然 Hadoop 已经毫无疑问地获得了相当大的知名度，但是其也仅仅是适合大数据存储和管理的三种技术之一。其他两种技术则是 NoSQL 和大规模并行处理（MPP）数据存储。MPP 数据存储的例子包括 EMC 的 Greenplum、IBM 公司的 Netezza 和惠普的 Vertica。

此外，Hadoop 是一个软件框架，这意味着它包括若干专门设计的组件，是专门设计来解决大规模分布式数据存储、分析和检索任务的。不是所有的 Hadoop 组件都是必要的，对于一个大的数据解决方案，其中一些组件可取代其他技术，更好地配合用户的需求。一个例子是 MapReduce 的 Hadoop，其中包括 NFS 作为 HDFS 的替代，并提供了一个完整的随机存取、读/写文件系统。同时，Hadoop 的输出结果没有关系型数据库输出结果那么精确，它不适用于类似卫星发射、开具银行账户明细这类精确度要求很高的任务，但是对于精确度要求不高的任务，它的运行速度就明显优于其他系统。例如，实现商家对顾客分群并分别进行不同的营销活动，Hadoop 就表现得很优秀。

2. 数据仓库

数据仓库的出现和发展是计算机应用发展到一定阶段的必然产物。经过多年的计算机应用和市场积累，许多企业已保存了大量原始数据和各种业务数据，这些积累所得数据集合通常能够真实地反映企业主体和各种业务环境的经济动态。如何合理地集中对积累的数据进行集中的存储和管理，同时通过对数据进行有效的统计、分析和评估，为企业主体提供帮助成为一个急需解决的问题。

数据仓库的产生及发展大概经历了三个时期。20 世纪 70 年代出现并被广泛应用的关系型数据库技术为解决这一问题提供了强有力的工具。从 20 世纪 80 年代中期开始，随着市场竞争的加剧，商业信息系统用户已经不满足于用计算机仅仅去管理日复一日的事务数

据，他们更需要的是支持决策制定过程的信息。20世纪80年代中后期，出现了数据仓库思想的萌芽，为数据仓库概念的最终提出和发展打下了基础。20世纪90年代初期，数据仓库之父 W. H. Inmon 在其具有里程碑意义的著作《建立数据仓库》中提出数据仓库并不是简单的数据堆积，而是从数量巨大的事务型数据库中抽取并清理数据，将其转换为新的存储格式，即为获得决策目标把数据聚合在一种特殊的格式中。这也使得数据仓库的研究和应用得到了广泛的关注。数据仓库的提出对处在激烈竞争中的商业企业，也具有非常重要的现实意义。

W. H. Inmon 将数据仓库定义为“数据仓库是支持管理决策过程的、面向主题的、集成的、随时间而变的、持久的数据集合。”

数据仓库因其处理数据为获得决策目标的特殊性，与传统数据库相比具备以下一些自身的特点。

(1)数据仓库的数据是面向主题的。传统数据库往往是面向某种应用进行数据组织的，而数据仓库中的数据是面向主题进行数据组织的。此处主题具有两方面的含义：一是它指代一个抽象的概念，是较高层次上企业信息系统中的数据综合、归类并进行分析利用的抽象；二是主题在逻辑意义上指代与其相对应的企业中某一宏观分析领域所涉及的分析对象。数据仓库面向主题的数据组织方式就是在较高层次上对分析对象的数据的一个完整、一致的描述，能完整、统一地表示出各个分析对象所涉及的企业的各项数据，并表明数据之间的联系。所谓较高层次是指按照主题进行数据组织的方式相对于传统面向应用的数据组织具有更高的数据抽象级别。

(2)数据仓库的数据是集成的。从数据仓库的定义中可以看出，数据仓库中的数据是从原有的分散的数据库数据抽取来的。由于操作型数据与 DSS 分析型数据之间的差别很大，数据仓库的每一个主题所对应的源数据在原有的各分散数据库中有许多重复和不一致的地方，且来源于不同的联机系统的数据都和不同的应用逻辑捆绑在一起，同时，数据仓库中的综合数据无法从原有的数据库系统直接得到，数据进入数据仓库之前，必然要经过统一与综合，这一步是数据仓库建设中最关键、最复杂的一步。数据仓库中数据的统一主要包括统一源数据中所有矛盾之处，如字段的同名异义、异名同义、单位不统一、字长不一致等。然后，还要对数据进行综合和计算。数据仓库中的数据综合工作可以在从原有数据库抽取数据时生成，但更多的是在数据仓库内部生成的，即进入数据仓库以后进行综合生成的。

(3)数据仓库的数据在面向应用时是不可更新的。数据仓库的数据反映的是一段较长的时间内企业的历史数据，可以看成不同时点的数据库快照的集合，及基于这些快照进行统计、综合和重组的导出数据，它不是联机处理的数据。数据库中进行联机处理的数据需要经过集成后输入数据仓库中，当数据仓库存放的数据已经超过数据仓库中数据的存储期限时，这些数据将从当前的数据仓库中被删去。由于数据仓库的数据主要用于企业实现决策分析，因而一般执行的数据操作主要是数据查询，一般情况不进行修改操作。所以，数据仓库管理系统与传统数据库管理系统相比，在数据操作管理上有所简化。数据库管理系统中的许多技术难点，如完整性保护、并发控制等，在数据仓库的管理中几乎可以省去。但是，数据仓库的查询数据量的规模很大，所以数据仓库对数据查询提出的要求更高，它要求采用各种

复杂的索引技术；同时，由于数据仓库面向的是商业企业的高层管理者，他们会对数据查询的界面友好性和数据表示提出更高的要求。

(4)数据仓库的数据是随时间不断变化的。数据仓库中的数据不可更新是针对用户进行分析处理时来说的，但从数据集成输入数据仓库开始到最终被删除的整个数据生存周期中，所有的数据仓库数据都是不变的，而面向整个数据仓库的全部数据是随时间的变化而不断变化的。

①数据仓库随时间变化不断增加新的数据内容。数据仓库系统必须不断捕捉 OLTP 数据库中变化的数据，追加到数据仓库中去，也就是要不断地生成 OLTP 数据库的快照，经统一集成后增加到数据仓库中去；对于确实不再变化的数据库快照，若捕捉到新的变化数据，则只生成一个新的数据库快照增加进去，而不会对原有的数据库快照进行修改。

②数据仓库随时间变化不断删去旧的数据内容。数据仓库的数据也有存储期限，一旦超过了这一期限，过期数据就要被删除。只是数据仓库内的数据时限要远远长于操作型环境中的数据时限。在操作型环境中一般只保存 60～90 天的数据，而在数据仓库中则需要保存较长时限(5～10 年)的数据，以适应 DSS 进行趋势分析的要求。

③数据仓库中包含有大量的综合数据，这些综合数据中很多与时间有关，如数据经常按照时间段进行综合，或隔一定的时间进行抽样等。这些数据要随着时间的变化不断地进行重新综合。因此，数据仓库的数据特征都包含时间项，以标明数据的历史时期。

各种计算机技术的不断发展进步，如数据模型、数据库技术和应用开发技术也推动了数据仓库技术的发展，并使其在实际应用中逐渐发挥了重要的作用。企业运用数据仓库所产生的巨大效益也同时刺激了对数据仓库技术的需求，数据仓库市场也正以迅猛势头不断发展：一方面，数据仓库市场需求量迅速变大，每年约以 400%的速度扩张；另一方面，数据仓库产品日益成熟，致力于生产研究数据仓库工具的厂家也越来越多。

7.1.2 大数据挖掘概述

在大数据时代，数据的产生和收集是前提基础，从大数据中找出相关性，发现其内在价值是关键。数据挖掘正是在大数据中发现数据价值的关键。早期的数据挖掘技术大多停留在概念阶段，随着信息科技超乎想象的进展，诸多新的计算机分析工具问世，如关系型数据库、模糊计算理论、基因算法及类神经网络等，使得从数据中发掘宝藏成为一种系统性且可实行的程序。

1. 数据挖掘技术

数据挖掘或知识发现(KDD)泛指从大量数据中挖掘出隐含的、先前未知但潜在的有用信息和模式的一个工程化和系统化的过程。也就是说，大数据本身是一组表象信息，核心是要从中挖掘数据的价值。数据挖掘首先要对大数据进行收集和整合，然后对数据挖掘的结果进行验证和运用，同时也离不开相关人员的决策。数据挖掘的结果大多是相关关系，而不是因果关系，这些结果还可能有不确定性。具体而言，实际应用的需求是数据挖掘领域很多方法提出和发展的根源。从数据挖掘应用最开始的顾客交易数据分析、多媒体数据挖掘、隐

私保护数据挖掘到文本数据挖掘和 Web 挖掘，再到社交媒体挖掘，它的发展与进步都是由实际应用推动的。工程性和集合性决定了数据挖掘研究内容和方向的广泛性。其中，工程性使得整个研究过程里的不同步骤都属于数据挖掘的研究范畴。集合性使得数据挖掘有多种不同的功能，而如何将多种功能联系和结合起来，从一定程度上影响了数据挖掘研究方法的发展。

广义上一般将数据挖掘的理论技术分为传统技术与改良技术两个方向。传统技术中主要依赖统计分析，多运用统计学内所含序列统计、概率论、回归分析、类别数据分析等方法实现数据挖掘技术。针对数据挖掘中数据对象的变化性及规模的庞大性，在传统技术中往往采用高等统计学里的因素分析、判别分析及分群分析等方法。在改良技术方面，通常使用的有决策树理论、类神经网络及规则归纳法等。决策树是一种用树枝形状展现数据，一般指根据各变量的影响情形实现的预测模型，它根据目标变量所产生效应的不同而建构分类的规则，一般多运用在对客户数据的分析上。例如，针对有回函与未回函的邮寄对象找出影响其分类结果的变量组合，常用分类方法为分类回归树（CART）及卡方自动交互检测法（CHAID）两种。类神经网络是一种仿真人脑思考结构的数据分析模式，一般通过输入变量与数值实现自我学习，并根据学习经验所得的知识不断地调整参数以构建数据的型样（patterns）。类神经网络为非线性设计，与传统回归分析相比，其好处是在进行分析时无须限定模式，特别是当数据变量间存有交互效应时可自动侦测出来；缺点则在于其分析过程为一个黑盒子，因此通常无法以可读的模型格式展现，每个阶段的加权与转换也不明确。因此，类神经网络多用于数据高度非线性且带有相当程度的变量交感效应的情形。规则归纳法是知识发掘领域中最常用的格式，这是一种通过使用一连串的逻辑规则（If/Then）对数据进行细分的技术，问题的关键是在实际运用时如何界定规则，通常需先将数据中发生数太少的项目剔除，以避免产生无意义的逻辑规则。

实现正确的大数据挖掘通常需要完成以下几个方面的工作。

（1）理解业务与数据间的关系，获取数据挖掘及分析所在的数据背景。

（2）需要完成对业务和数据的整合并获取查询数据。

（3）将所获取的数据中明显有错误、不一致或不完整的数据去除，完成数据的清洗。

（4）在庞大的数据中选取部分数据作为样本先行试验。

（5）根据实验数据预先建立实验的数据模型，并在预建模型的前提下，进行数据挖掘的分析工作。

（6）运用数据挖掘的分析结果及进一步改进的模型对大数据进行测试与检验。

（7）找出假设并提出解释，同时持续应用于企业流程中。

由此可见，数据挖掘的整个过程中前期准备工作及规划过程占据重要的位置，事实上在整个数据挖掘的过程中数据前置作业阶段可能花费掉 80％的时间，其中包含数据的净化、格式转换及表格的连结。完整、真实地完成数据的格式转换是正确实现数据价值的前提。

2. 数据挖掘算法

数据挖掘的最终目标是获取数据的有效价值，其具体的实际应用功能可分为三大类（分

类区隔类、推算预测类和关联分析类）六分项（分类、聚类、回归分析、时间序列、关联规则、序列模式）：分类和聚类属于分类区隔类；回归分析和时间序列属于推算预测类；关联规则和序列模式则属于关联分析类。所以，数据挖掘最终完成的任务往往是对大数据的预测及相关性的分析。下面分别对数据挖掘的各项功能含义予以阐述。

（1）分类是根据数据及变量构造一个分类函数或分类模型（分类器），利用分类模型可将数据库中的数据项映射到给定类别中的某一个。一般分类器的构造都是通过对样本数据训练所得到的，样本训练集由一组数据库记录或元组构成，每个元组是一个由有关字段（属性或特征）值构成的特征向量。训练集中的每个训练样本带有一个类别标记。一个具体样本的形式可表示为：v1，v2，…，v*n*；c。其中，v1，v2，…，v*n* 表示字段值；c 表示类别。

分类常会被用来处理一些根据之前的经验已经分类好的数据来研究出数据的特征，然后根据这些特征对其他未经分类或新的数据进行预测。此类用来寻找特征的已分类数据事实上就来自现有的客户数据，或是对一个完整数据库做部分取样，再经由实际的运作来测试。在分类算法中比较典型的算法是 K 最近邻（KNN）算法和 Naive Bayes 算法。KNN 算法是向量空间中一种基于样本实例的分类算法，该算法是通过计算训练集合中文本与测试集合中文本的相似度，由大到小排序，选择与测试文本距离最近（相似度最高）的 k 个训练学习样本，选择 k 个样本中占多数的样本的所属类别，作为未知样本的类别。KNN 算法是一种懒惰学习算法，它是一种计算相对简单且分类速度较快的经典分类算法，对训练实例具有很强的依赖性。KNN 算法对存储空间的需求较好，对 k 的取值带有一定的经验性，但由于其具有较好的分类效果，通常也被广泛采用。例如，利用一个大型新闻数据源库的部分取样来建立一个不同新闻数据的分类模型，从而可以对不断采集获取的新的新闻数据利用模型进行分类预测。Naive Bayes 算法又称朴素贝叶斯算法，该算法是根据训练数据集合计算待分类文本属于各个类别的后验概率，将文本划分到所计算得到概率值最大的类别中。Naive Bayes 算法会将未知类别的文档划分到计算得到后验概率值最大的类别中。在实际应用中，对特征项间的独立性假设存在一定的不合理之处，但 Naive Bayes 算法因其表现出的较快的学习过程和稳定较好的分类效果，仍被广泛地应用于分类器中。

（2）聚类用于将数据分群，其目的是寻找不同群间的差异，同时发现群内成员的相似性。与分类不同的是，聚类在进行数据分析前并不确定使用何种方式或划分成何种类别，将数据库中的对象进行聚类是聚类的最基本操作。聚类分析的准则是使隶属于同一类的个体间的距离尽可能小，而不同类的个体间的距离尽可能大。根据划分条件的不同有多种聚类算法，典型的有 K-means 算法、K-medoids 算法、CLARANS 算法、BIRCH 算法等，这些算法适用于特定的问题及用户。以 K-means 算法为例，K-means 算法的基本思想是初始随机给定 K 个簇中心，按照最邻近原则把待分类样本点分到各个簇。然后，按平均法重新计算各个簇的质心（这个质心可以不是样本点），从而确定新的簇心。一直迭代，直到簇中心的移动距离小于某个给定的值。具体的 K-means 算法主要包含三个步骤：首先是为待聚类的点寻找聚类中心，然后计算每个点到聚类中心的距离，同时将各个点聚类到离该点最近的聚类中，最后计算每个聚类中全部点的坐标平均值，并将这个平均值作为新的聚类中心，迭代执行第一、二步，直到聚类中心不再进行大范围移动或聚类次数达到要求为止。

(3)回归分析是使用一系列的现有数值来预测一个连续数值的可能值。一般的回归分析法指利用数理统计方法建立因变量与自变量之间的回归关系函数表达式,回归分析中比较典型的算法就是线性回归法,线性分析的任务就在于根据 $x1,x2,\cdots,xp$ 线性回归和 Y 的观察值,去估计函数 f,寻求变量之间近似的函数关系。例如,预测房价,当前自变量(输入特征)是房子面积 x,因变量是房价 y,同时给定一批训练集数据。回归分析要做的是利用手上的训练集数据,得出 x 与 y 之间的函数 f 关系,并用 f 函数来预测任意面积 x 对应的房价 y。若将范围扩大亦可利用逻辑性回归来预测类别变量,特别在广泛运用现代分析技术,如类神经网络或决策树理论等分析工具,推估预测的模式已不再止于传统线性的局限,在预测的功能上大大增加了选择工具的弹性与应用范围的广度。

(4)时间序列与回归分析的功能类似,只是它是用现有的数值来预测未来的数值。两者最大的差异在于时间序列所分析的数值都与时间有关,时间序列预测的工具里会带有并可以处理有关时间的一些特性,如时间的周期性、阶层性、季节性及其他一些特别因素(过去与未来的关联性)。

(5)关联规则和序列模式同属于关联分析类。关联规则是寻找出在同一个事件中出现的不同项的相关性,即不同项中某事件同时出现的概率。例如,购买铁锤的顾客中70%的都购买了铁钉,发掘出此类关联规则对于商场管理人员很有价值,他们可以根据这些规则更好地进行规划,如把铁锤和铁钉这样的商品摆放在一起,能够促进销售。关联规则的基本思想:一是找到所有支持度大于最小支持度的频繁项集,即频集;二是使用第一步找到的频集产生期望的规则,其核心方法是基于频集理论的递推方法。1993 年,R. Agrawal 等人首次提出了挖掘顾客交易数据中项目集间的关联规则问题,其核心是基于两阶段频繁项集思想的递推算法。该关联规则在分类上属于单维、单层及布尔关联规则,典型的算法是 Aprior 算法。Aprior 算法将发现关联规则的过程分为两个步骤:第一步通过迭代,检索出事务数据库中的所有频繁项集,即支持度不低于用户设定的阈值的项集;第二步利用频繁项集构造出满足用户最小信任度的规则。其中,挖掘或识别出所有频繁项集是该算法的核心,占整个计算量的大部分。

序列模式与关联规则非常类似,但序列模式中事件的相关性往往特指事件之间时间上的相关性,如今天银行利率的调整与明天股市的变化等。

虽然数据挖掘的各项功能在方法和目标上各有不同,但它们都不是独立存在的,而是在数据挖掘中互相联系,发挥作用。

7.1.3 大数据对思维方式的影响

维克托·迈尔-舍恩伯格在《大数据时代:生活、工作与思维的大变革》一书中明确指出,大数据时代最大的转变就是思维方式的三种转变:全样而非抽样、效率而非精确、相关而非因果。

1. 全样而非抽样

过去,由于数据存储和处理能力的限制,在科学分析中,通常采用抽样的方法,即从全集

数据中抽取一部分样本数据，通过对样本数据的分析来推断全集数据的总体特征。通常，样本数据的规模要比全集数据小很多，因此，可以在可控的代价内实现数据分析的目的。现在，我们已经迎来了大数据时代，大数据技术的核心就是海量数据的存储和处理，分布式文件系统和分布式数据库技术提供了理论上近乎无限的数据存储能力，分布式并行编程框架MapReduce提供了强大的海量数据并行处理能力。因此，有了大数据技术的支持，科学分析完全可以直接针对全集数据而不是抽样数据，并且可以在短时间内迅速得到分析结果，速度之快，超乎我们的想象。例如，谷歌公司的Dremel可以在2～3秒内完成PB级别数据的查询。

2. 效率而非精确

过去，我们在科学分析中采用抽样分析方法，就必须追求分析方法的精确性，因为抽样分析只是针对部分样本的分析，其分析结果被应用到全集数据以后，误差会被放大，这就意味着，抽样分析的微小误差被放大到全集数据以后，可能会变成一个很大的误差。因此，为了保证误差被放大到全集数据时仍然处于可以接受的范围，就必须要确保抽样分析结果的精确性。正是由于这个原因，传统的数据分析方法往往更加注重提高算法的精确性，其次才是提高算法效率。现在，大数据时代采用全样分析而不是抽样分析，全样分析的结果就不存在误差被放大的问题。因此，追求高精确性已经不是其首要目标；相反，大数据时代具有“秒级响应”的特征，要求在几秒钟内就迅速给出针对海量数据的实时分析结果，否则就会丧失数据的价值，因此，数据分析的效率成为关注的核心。

3. 相关而非因果

过去，数据分析的目的，一方面是解释事物背后的发展机理，比如，一个大型超市在某个地区的连锁店在某个时期内净利润下降很多，这就需要IT部门对相关销售数据进行详细分析找出发生问题的原因；另一方面是预测未来可能发生的事件，比如，通过实时分析微博数据，当发现人们对雾霾的讨论明显增加时，就可以建议销售部门增加口罩的进货量，因为人们关注雾霾的一个直接结果是，大家会想到购买口罩来保护自己的身体健康。不管是哪个目的，其实都反映了一种“因果关系”。但是，在大数据时代，因果关系不再那么重要，人们转而追求“相关性”而非“因果性”。比如，我们在购物网站购物时，当我们购买了一个汽车防盗锁以后，购物网站还会自动提示你，与你购买相同此物品的其他用户还购买了汽车坐垫。也就是说，购物网站只会告诉你“购买汽车防盗锁”和“购买汽车坐垫”之间存在相关性，但是并不会告诉你为什么其他用户购买了汽车防盗锁以后还会购买汽车坐垫。

7.1.4 大数据分析师

1. 商业智能的概念

商业智能（简称BI），又称商务智能或商业智慧，其概念于1996年由Gartner Group提出。Gartner Group将商业智能定义为商业智能是描述了一系列的概念和方法，通过应用基于事实的支持决策系统来辅助商业决策的制定和实施。商业智能提供使企业迅速计算分析数据的技术和方法，包括收集、组织、管理和分析数据，并将这些数据转化为有用的信息，然

后分发到企业各处。不过，目前公认的商业智能的定义是指企业在收集、组织、管理和分析结构化与非结构化的数据和信息时，使用现代信息技术，使商务决策水平得以提升，商务知识和见解得以创造和增加，并且能够帮助企业完善商务流程，采取更有效的商务行动，提升各方面商务绩效，提高综合竞争力的智慧和能力。商业智能是一系列技术、方法和软件的总称，其最终目的是提高企业运营性能以及增加企业商业利润。对于商业智能这个概念的正确理解，应从四个层面展开。

(1)信息系统层面。它是商业智能系统(BI System)的物理基础，是一个面向特定应用领域的信息系统平台，一个独立的软件工具，具有非常强大的决策分析能力。

(2)数据分析层面。商业智能是一系列具有计算、分析功能的工具、算法或模型的总称。在数据分析层面，首先是获取数据，获取与所关心主题有关的高质量的数据或信息，然后自动或人工参与使用具有分析功能的算法、工具或模型，其间包括分析信息、得出结论、形成假设与验证假设等一系列的过程。

(3)知识发现层面。它与数据分析层面一样，也是一系列工具、算法或模型的总称。这一层面可以直接将信息转变成知识，或者是把数据转变成信息后，借助于大数据分析挖掘技术发现信息背后隐藏的东西，然后将其转变成知识。

(4)战略层面。这一层面主要是将知识或信息应用在改善运营能力和提高决策能力以及企业建模等上面。商业智能的战略层面是提高企业决策能力，是通过利用应用假设或经验以及一个或多个数据源的信息所形成的一组方法、概念和过程的集合。它通过获取、组织、管理和分析数据，将数据和信息提供给贯穿企业组织的各类人员，使得企业的决策能力得以提高。

2.大数据分析师的分类

数据分析师是一个统称，不同的公司对于数据分析师的需求不一样，有主要负责产品、运营的数据分析师(business partner)，有负责集团管理层的分析师(类似军师)，也有行业研究的分析师(比如一些行业咨询公司)。

因此对于不同阶段的需求，需要不同专长的人员，他们大体可以分为三类。

(1)业务分析人员。要求精通业务，能够解释业务对象，并根据各业务对象确定出用于数据定义和挖掘算法的业务需求。

(2)数据分析人员。精通数据分析技术，并熟练掌握统计学相关知识，有能力把业务需求转化为数据挖掘的各步操作，并为每步操作选择合适的技术。

(3)数据管理人员。精通数据管理技术，并从数据库或数据仓库中收集数据。

综上可见，数据挖掘是一个多学科专家合作的过程，也是一个在资金上和技术上高投入的过程。这一过程要反复进行，并在反复过程中不断地趋近事物的本质，不断地优化问题的解决方案。进行数据重组和细分、添加和拆分记录，选取数据样本、可视化数据、探索聚类，分析神经网络、决策树数理统计、时间序列结论、综合解释评价数据、知识数据取样、数据探索、数据调整和模型化评价。

7.2 扩展知识

大数据时代展现出数据的规模庞大性和数据存储管理的复杂性，但大数据所提供的仅仅是具备一定价值的数据集合，如何对大数据分析并提取出其所具有的价值是需要解决的关键问题。同时，随着计算技术和互联网技术的不断发展和进步，大数据分析技术已经逐渐延伸应用于社会的各个领域。

7.2.1 大数据分析法

大数据本身的数据规模大，数据结构复杂，但它仅仅是带有数据价值的一手材料，无法直接运用，如何对规模庞大、结构复杂的大数据进行分析，从而进一步获取更多有规律、智能且极具价值的信息也成为大数据研究中最重要的问题之一。尤其是随着信息技术的迅速发展，大数据的应用领域越来越广泛，使得大数据的复杂性日益增加，尤其是在数据结构多样性上，所以大数据的分析方法在大数据领域就显得尤为重要，可以说是决定最终信息是否有价值的决定性因素。下面将介绍一些当前主流的大数据分析的方法理论。

1. 大数据处理

大数据的应用逐渐由早期的数据抽样，追求绝对准确率及查找因果关系的时代转变为运用大规模数据寻找内在规律及相关性。面对纷繁复杂的大规模数据集合，首先需要完成对数据的处理，才能进一步实现对数据的分析和应用。

具体的大数据处理方法其实有很多，通常一个基本的大数据处理流程可以概括为数据的采集、数据的导入和预处理、数据的统计和分析、数据的挖掘四个步骤。

(1)数据的采集多指使用多个数据库来接收来自不同客户端(Web、App 或传感器形式等)的数据，同时用户还可以通过使用这些数据库来完成简单的查询和处理工作。例如，在电子商务中电商往往会采用一些如 MySQL 和 Oracle 的传统关系型数据库存储其每一笔事务数据。除此之外，一些 NoSQL 数据库(Redis 和 MongoDB)也常用于数据的采集。大数据的采集过程中面临的最主要的挑战就是并发数很高的问题，因为在同一时刻可能会有成千上万的用户正在进行访问和操作，比较典型的如火车票的售票网站及网上购物的淘宝网站，此类网站的并发访问量在其峰值时可能达到上百万，这往往需要在采集端部署数目庞大的数据库才能予以支撑，同时，还要深入思考和设计如何在这些数据库之间进行负载均衡和分片。

(2)数据的导入和预处理是将在数据采集端分布在多个不同数据库上的前端数据导入一个集中的大型分布式数据库或分布式存储集群中，进而在导入数据的基础上做一些简单的数据清洗和预处理工作。在数据导入阶段，为满足一些业务的实时计算要求，有些用户可能会运用来自 Twitter 的 Storm 对导入数据进行流式计算。在数据导入和预处理阶段面临的最大问题就是数据规模非常庞大，通常每秒钟的数据导入量可能会达到百兆，甚至千兆

级别。

(3)在数据的统计、分析阶段主要是利用导入数据后的分布式数据库或分布式计算集群来对存储在它们之中的海量数据进行一些普通的分析和分类汇总等,进一步来满足一些较为常见的分析需求。在数据分析和统计阶段,为满足一些实时性需求,通常会使用 EMC Greenplum 或 Oracle 的 Exadata,以及基于 MySQL 的列式存储 Infobright 等。针对批处理或基于半结构化数据的需求则可以选择使用 Hadoop。在数据统计与分析阶段关键的问题是统计与分析所涉及的数据量大,会极大地占用系统资源,特别是对 I/O 资源的占用非常突出。

(4)数据挖掘在前一节已经有所介绍。它与数据的统计和分析过程完全不同的是,数据挖掘通常不会预先设定主题,而是在现有数据基础上运用各种算法计算进而实现对效果的预测,从而进一步完成一些高级别数据分析的需求。在数据挖掘中较为典型的算法通常有基于 K-Means 的聚类算法、用于统计学习的 SVM 算法和用于分类的 Naive Bayes 算法,而在数据挖掘中主要使用的工具有 Hadoop 的 Mahout 等。数据挖掘中最为关键和复杂的问题在于挖掘算法的选择与设计以及挖掘算法自身的复杂性,而且在数据挖掘中计算所涉及的数据量和数据的计算量非常庞大,另外,一般的数据挖掘算法都以单线程为主,所以在此阶段往往会耗费大量的时间和精力。

2. 大数据分析的五个基本方面

1)可视化分析

可视化就是运用图形或图表的方式向数据分析的专家或普通的用户直观地展示数据,即运用一些数据分析的工具将数据转化为图的形式,达到“看图说话”的效果。在大数据分析中无论是对数据分析专家还是普通用户,数据可视化都是数据分析工具应具备的最基本的前提。当前一些常见的数据可视化工具有 Weka、R Project、NodeBox、Gephi、Google Chart API 等。运用这些数据可视化工具,大数据可视化可以通过多种方法来实现,如多角度展示数据、聚焦大量数据中的动态变化及筛选信息(包括动态问询筛选、星图展示和紧密耦合)等方法。当前大数据可视化面临的主要挑战:一是针对一些大数据分析工具(如 Hadoop,High Performance Computing and Communications 等),它们可以轻而易举地处理 ZB 级和 PB 级数据,但它们往往不能将这些数据可视化;二是如何实现大数据可视化的动态化,大数据的交互式可视化比静态数据工具能够更好地进行工作,也为大数据带来了无限前景。基于 Web 的可视化为大数据处理及时获取动态数据提供了有利的前提,同时也为交互式可视化提供了可能。

2)数据挖掘算法

大数据的可视化是为了方便数据分析专家和用户直观感受数据的存在,但大数据分析的目标是通过运用计算机深入数据内部进而发掘出数据的自身价值。所以,大数据分析理论核心是数据挖掘算法及如何改进传统的数据挖掘算法使其适用于不同的数据类型和格式,进而运用这些挖掘算法处理海量的数据并更加科学地呈现出数据本身具备的特点。外,还要求这些数据挖掘的算法能够更快速地处理大数据,大数据的价值具有时效性,也

求数据处理的速度足够快。

3)预测性分析能力

在大数据分析应用中非常重要的一点就是做预测性分析，一般是先从大数据中挖掘出数据的规律及特点，再科学地完成模型的建立，通过将新的数据带入之前建立的新模型，进而预测未来新的数据或未来可能产生的结果。如果说数据的可视化和数据挖掘能够让数据分析员和用户更好地理解和计算数据，那么预测性分析的作用则是让数据分析员可以根据数据可视化分析和数据挖掘计算所得结果对未来的相关事物做出一些预测性的判断。

4)语义引擎

大数据与传统数据集除在数据规模上有所区别外，还有一点不同就是其数据中包含大量的非结构化数据，这种数据自身的多样性为数据分析也带来了新的挑战。数据分析者往往需要一系列的工具去完成数据的解析、提取及进一步的分析。语义引擎需要被设计成能够从“文档”中智能提取信息。结合大数据分析在网络数据挖掘中的广泛应用，借助语义引擎，可从用户的搜索关键词、标签关键词或其他输入语义分析、判断用户需求，从而实现更好的用户体验和广告匹配。

5)数据质量和数据管理

大数据分析的前提是数据，而数据质量和数据管理也显得尤为重要。高质量的数据和有效的数据管理，无论是在学术研究还是在商业应用领域，都能够保证分析结果真实、有价值。专业的数据分析工具只有在高质量、管理规范、合理的大数据环境中才能提取出隐含的、准确的、有用的信息，否则即使使用先进的数据分析工具在大数据环境中也只能提取出毫无意义的信息。通过标准化的流程和工具对数据进行处理可以保证一个预先定义好的高质量的分析结果。

大数据时代的来临无疑使得整个社会中的各个行业，包括政府、商界、企业等都面临着如何处理来自各个行为、各个对象的接触、交易、互动时产生的海量数据，大数据分析流的结果可以说是企业发展的新机遇，但同时也具有一定的挑战性。它要求管理层或相关负责人员必须准确获取高质量的数据，并提炼出一套合适的模式进而将其转化为可靠的决策依据。

一般大数据分析的基础就是以上五个方面，当然随着大数据分析研究的不断深入，还会不断产生实用意义更强、更具特点、更加深入专业的大数据分析方法。

7.2.2 大数据分析的应用

大数据分析的应用已经渗透到了人们生活中的方方面面。最典型的案例就是营销推荐。你会发现如果自己在淘宝网站选购图书或连续使用百度搜索引擎定向查找特定内容一段时间后，淘宝网站会自动为用户定制推荐书目，百度搜索引擎会提供相关的内容推荐。事实上网站所具备的这种功能正是根据用户的消费数据及查找的语义分析获取信息的，从而预测用户可能使用的相关信息。数据流借助大数据分析从而拥有了预测性分析的能力，它能够通过数据的实时变化迅速建立预测判断，收集到有利于提升企业决策，提高企业赢利的数据，而这些数据的来源包括新闻消息、摄像头、卫星、网络爬虫、服务器、传统数据库等，甚至可能是 Hadoop 系统的数据，而大多数企业的工作都能够从数据流的处理方案中获得帮

助从而转换为利益点。大数据在企业商业智能营销、公共服务和用户个体服务三个领域都拥有巨大的应用潜力和商机，主要表现在以下几个方面。

1. 大数据分析使企业和商业能理解客户并预测客户服务需求

当前大数据分析在营销推荐中的应用是最广泛的，也是与人们生活联系最密切的。具体地讲就是提升消费者与企业之间的关系，使商品卖得更多、更快，同时消费者的采购更有效率。大数据应用在此领域中需要研究的重点是如何应用大数据更好地了解客户及他们的爱好和行为，从中发现商业机遇并取得商业价值。它所借助的最大的数据系统是发布在Web上的多方位的用户信息。与传统营销渠道完全不同的是，企业与客户之间的接触点从过去简单的电话和邮件地址，发展到网页、社交媒体账户、博客等。企业借助这些用户发布的个人信息及偏好的渠道跟踪客户，将他们的每一次点击、收藏、“顶”、分享、加好友、转发等行为纳入企业的销售行为中，并将其转化成客户价值。使用数字化营销手段，企业可以有效、及时地完成对用户的个性化和精准定位。

移动互联网技术及社交网络的便利性无形中将大数据分析与在线营销交织在一起，其应用可以分为以下两大类。

(1)从线上到线下。配备近场数据通信交换技术的智能手机和基于位置的签到为营销人员提供了有利的条件。他们将能跟踪商场人流，把在线零售的分析优化应用于线下。

(2)数据分析工具将更加容易使用(面向中小企业应用的大数据创业形势非常乐观)，中小企业可能不具备商务智能平台，但他们可以使用平板电脑和智能手机，移动联网客户智能分析将会改变企业使用营销工具的方式。大数据分析使企业能更加全面地了解客户。灵活运用所得用户不同偏好的信息，发现其中的规律，建立出数据模型并进行预测。例如，美国著名的零售商Target运用大数据的分析获得客户消费习惯相关的有价值信息，进而精准地预测到客户选择生育小孩的时间，从而完成持续的产品营销链。另外，通过大数据的分析，超市、汽车、保险等线下行业可通过线上大数据反馈的用户喜好信息，预测出受欢迎及适合特定人群的产品，进而有针对性地实现产品销售。

2. 大数据分析应用于社会公共服务

大数据分析被广泛应用的另一个重要的领域是社会公共服务。如今，数据挖掘已经能够预测疾病暴发、理解交通模型并改善安全执法情况。大数据是公共服务现代化的一种技术路径，为公共服务方式的变革与模式的创新提供了技术支撑。

以公共卫生与医疗为例，大数据的应用为公共卫生管理及时提供了流行病的症状、种类及传染途径信息，同时实时反映流行病爆发区域居民的身体健康状况和环境卫生状况信息，这为及时、有效地控制疾病提供了保障。以大数据为基础建立疾病监测与响应信息平台，可以做到在疾病暴发前做好相应的准备与措施，爆发中采取有效手段控制疾病蔓延，爆发后提供相应配套设施预防疾病再次发生。同时，对原先发生疾病信息的整合与分析，为预测疾病发生控制措施提供了借鉴。

大数据分析可以更好地优化配送业务流程。借助用户所在社交媒体、网络搜索及天气预报挖掘出有价值的数据，大数据分析得到的结果可以提供更优质的供应链及配送路线

在这两个方面，地理定位和无线电频率的识别追踪货物和送货车，利用实时交通数据制定更加优化的路线。

大数据还被应用于改善城市实时交通信息、利用社交网络和天气数据来优化最新的交通情况，利用大数据可以建立以高速公路监控和信息诱导系统、车速信息系统、优化交通系统、路口监测系统为主的综合信息平台。对大量的交通状况数据进行快速处理和实施分析，依靠所覆盖的网络，对交通状况进行全景式的观察，实现数据采集、信息发布及依靠数据化调控改善城市交通，信息技术与公共交通相连接，突破原有增量解决交通拥堵的框架从而提高道路利用率，实现减缓交通拥堵情况，运用大数据分析历史数据，获得道路交通流量及事件的固定模式可以预测未来一小时内的道路交通流量状况。

大数据分析在政府安全执法的过程中也有所应用。例如，美国国家安全局利用大数据打击恐怖主义，企业应用大数据技术防御网络攻击，警察应用大数据工具捕捉罪犯，信用卡公司应用大数据工具来检测欺诈性交易。

3. 大数据分析对用户个体的影响

移动互联网技术的进步、智能手机及 Twitter 等社交网络的普及让人类社会首次实现了公民的联网。大数据的规模及数据形式的多样性也因全民的参与面临更大的挑战。大数据分析的应用不再单一面向企业或政府，而是同样适用于参与其中的每个个体。用户个人信息往往保存在第三方手里。例如，个人用户在互联网上留存、在政府部门登记在案的各类信息，此类信息实际上也是互联网、政府和企业用于分析用户行为的基础。此外，随着个体可穿戴设备等新事物的产生与不断发展，个人信息的获取方式趋于便利化且形式更加多样化，用户数据信息的积累也在不断完善。例如，用户的身体数据、健康数据、地理位置信息、运动数据、社会关系数据、饮食数据等个体信息可通过可穿戴设备或植入芯片等感知技术被实时获取。未来此类应用针对个性化人群可以设计的应用场景是，个体用户可以将个人数据授权给第三方机构用于实现特定用途，如高血压患者可以将个人血压数据、身体机能数据、饮食数据等授权给健康管理机构使用，由他们监控和使用这些数据，进而为用户制定有效的健康维护方案。

大数据分析应用的计算能力在个体医疗卫生方面可以实现对整个 DNA 序列的解码，进而针对个体制定出最新的治疗方案，同时可以更好地去理解和预测疾病。典型的应用是，大数据技术目前已经在医院应用于监视早产婴儿和患病婴儿的情况，通过记录和分析婴儿的心跳，医生针对婴儿的身体可能会出现的不适症状做出预测。

7.3 实　训

大数据应用开启了一个大规模产生、分享和应用数据的时代，同时也给信息技术和商业带来了巨大的变化。大数据在社会核心领域的渗透速度有目共睹，然而研究表明，在大数据时代未被真正利用到的信息比例高达 99.4%，很大程度都归因于高价值的信息无法被采集。