第3章
大数据分析

导学

本章主要介绍大数据分析的基础知识、大数据分析的主要技术及分析系统，以及医学大数据分析实证应用案例，使读者对大数据分析有个概括性的了解和掌握。

了解：大数据分析的基本思想、目前国内外大数据分析的主要状况、大数据分析的应用案例操作。

掌握：大数据分析的基本概念、大数据分析操作流程、大数据分析的研究方面、大数据分析使用的基本技术、大数据分析处理系统的类型及特点和作用。

大数据分析就是研究包含各种数据类型的大型数据集的过程。大数据技术可以发现隐藏的数据模式、未知数据的相关性、发展趋势和其他有用的商业信息。就医学大数据分析而言，其分析结果可以带来更有效的医疗诊治、更好的医疗服务、提高医疗效率、获得竞争优势和其他医疗与商业利益。

3.1 大数据分析简介

大数据具有价值密度低的特征，必须通过分析、处理对这些数据去伪存真，获得有用的数据及其相互关系，才能得到有价值的信息。大数据应用中的核心技术就是从大量数据中提取出我们所需要的信息并进行分析和处理，因此大数据分析是决定最终信息是否有价值的决定性因素。

大数据分析需要解决的问题主要包括如何通过构建数据库存储并处理这些大量、生成快速、模态繁多、异构的数据；如何将这些数据的结构标准化，从中提取出有用的信息；如何对大数据资源进行分配；如何实现大数据的安全、可靠传输等。

3.1.1 大数据的分类与存储方式

在大数据分析里，第一个问题是要明确分析的对象，即数据的概念。那么，什么是数据呢？数据是指所有能输入计算机并被计算机程序处理的符号的介质的总称，是用于输入电

子计算机进行处理，具有一定意义的数字、字母、符号和模拟量等的通称[1]。

单一的数据记录一般并不独立形成概念，为了产生有价值的、可靠的新知识，需要将不同记录的数据进行有效关联和组织，通过数据分析，把握体现数据共性和差异的关键线索，从而对在数据中的信息进行有序解读，实现对隐藏于数据中的线索和联系进行归纳与推理。从数据的复杂性来看，数据可分为显数据和隐数据。

显数据是指按照某种规律或理论通过测量能够得到的数据，用以描述观察到的现象和对概念做出量化描述。比如药品的大小、疾病的血项特征、CT 影像的特征等，显数据就是对参数的个体、部分或整体的观测。

对于无法直接测量的知识，则需要通过模型辅助推断，而推理建模的数据称为隐数据。隐数据的主要作用是揭示隐性知识成立的可靠依据，例如区分两类药品的关键要素、用于疾病诊断的基本症状等这类问题，其特点是概念构成多样化、内外影响机制不确定等，常常涉及不同因素或群体之间的相互影响作用关系的变化的揭示。

数据是个很广泛的概念，随着大数据时代的到来，数据量的激增越来越明显，各种各样的数据铺天盖地地砸下来。许多存储于数据库的大数据主要实现了事实的描述性功能，但其分析潜力没有得到深度开发；在复杂问题中，无论是已知概念的统计描述还是未知概念的统计推断常常同时被需要，显性数据和隐性数据都是不可或缺的。

1. 大数据的分类

计算机存储和处理的对象十分广泛，随着数据量的急剧增加，这些对象的处理将变得越来越复杂，根据复杂程度可将大数据的类型作如下分类。

1）按字段类型分类

按字段类型分类，大数据可分为文本类、数值类、时间类。

（1）文本类数据。文本类数据常用于描述性字段，如姓名、地址、交易摘要等数据，这类数据不是量化值，不能直接用于四则运算。在使用时，可先对该字段进行标准化处理（比如地址标准化），再进行字符匹配，也可直接模糊匹配。

（2）数值类数据。数值类数据用于描述量化属性或用于编码，如交易金额、额度、商品数量、积分数、客户评分等都属于量化属性，可直接用于四则运算，是日常计算指标的核心字段。

邮编、身份证号码、卡号之类的则属于编码，是对多个枚举值进行有规则编码，可进行四则运算，但无实质业务含义，不少编码都作为维度存在。

（3）时间类数据。时间类数据仅用于描述事件发生的时间。时间是一个非常重要的维度，在业务统计或分析中非常重要。

2）按描述事物的角度分类

按描述事物的角度分类，大数据可分为状态类数据、事件类数据、混合类数据这三种类型。

（1）状态类数据。状态类数据用数据来描述客观世界的实体，如心脏、肝脏、血液等对象；不同种类的对象拥有不同的特征，如血液的特征包括血型、红细胞和白细胞，心脏的特征包括心房和心室，这些数据可以随时间发生变化，每个时点的数据反映这个时点对象所处的状态，因此称之为状态类数据。

（2）事件类数据。事件类数据用于描述客观世界中对象之间是怎么互动的，我们把这

一次次互动或反应记录下来，这类数据称之为事件类数据。如患者到医院就医，这里出现三个对象，分别是患者、医院、药品，这三个对象之间发生了一次交易关系。

（3）混合类数据。混合类数据理论上也属于事件类数据范畴，两者的差别在于，混合类数据所描述的事件发生过程持续较长，记录数据时该事件还没有结束，还将发生变化。如药品临床反应，从注射药品到药品后期反应，整个过程需要持续很长一段时间，首次记录药品临床数据是在药品服用或注射后，血液白细胞、血液红细胞等各项功能的多次变化情况。

3）按数据处理的角度分类

按数据处理的角度分类，大数据可分为原始数据和衍生数据两种类型。

（1）原始数据。原始数据指来自上游系统的，没有做过任何加工的数据。虽然会从原始数据中产生大量衍生数据，但还是会保留一份未作任何修改的原始数据，一旦衍生数据发生问题，可以随时从原始数据重新计算。

（2）衍生数据。衍生数据是指通过对原始数据进行加工处理后产生的数据。衍生数据包括各种数据集市、汇总层、宽表、数据分析和挖掘结果等。从衍生目的上，可以简单分为两种情况：一种是为提高数据交付效率，数据集市、汇总层、宽表都属于这种情况；另一种是为解决业务问题，数据分析和挖掘结果就属于这种。

4）按数据粒度分类

按数据粒度分类可分为明细数据、汇总数据两种类型

（1）明细数据。通常从业务系统获取的原始数据，是粒度比较小的，包括大量业务细节。比如，就医信息表中包含每个患者的性别、年龄、姓名等数据，信息表中包含每笔就医的时间、症患、用药情况等数据。这种数据我们称之为明细数据。明细数据虽然包括了最为丰富的业务细节，但在分析和挖掘时，往往需要进行大量的计算，效率比较低。

（2）汇总数据。为了提高数据分析效率，需要对数据进行预加工，通常按时间维度、地区维度、产品维度等常用维度进行汇总。分析数据时，优先使用汇总数据，如果汇总数据满足不了需求则使用明细数据，以此提高数据使用效率。

5）按数据结构分类

按数据结构分类，大数据可分为结构化数据、半结构化数据、非结构化数据三种类型。

（1）结构化数据。通常是指用关系数据库方式记录的数据，数据按表和字段进行存储，字段之间相互独立。图 3-1 所示为结构化数据表示的二维表格。

2013年度第二学期(1)班第一次月考

学号	姓名	语文	数学	英语	自然	社会	总分	名次	百分比排名	等级
1	应祖	101	97	50	169	75	492			
2	慕如	117	103	89	166	84	559			
3	林海	118	121	90	146	89	564			
4	唐乐	111	108	94	180	79	572			
5	陈纳	108	114	81	160	93	556			
6	陈靖	116	95	92	163	87	553			
7	林瑶	107	88	80	157	81	513			
8	王青	125	108	96	180	87	596			
9	林辉	109	133	103	172	90	607			
10	周祖	95	100	57	149	85	486			
11	林强	120	89	105	152	90	556			
12	董益	116	111	84	161	88	560			
13	黄妙	111	78	88	161	89	527			
14	高翔	126	121	107	195	96	645			

图 3-1　结构化数据表示的二维表格

（2）半结构化数据。半结构化数据是指以自描述的文本方式记录的数据，由于自描述数据无须满足关系数据库那种非常严格的结构和关系，在使用过程中非常方便。很多网站和应用访问日志都采用这种格式，网页本身也是这种格式，如图 3-2 所示。

```
<!DOCTYPE html>
<html>
  <head>…</head>
  <body class="cms-area-hidden">
    <script type="text/javascript">…</script>
    <div class="child_fixed" id="search">…</div>
    <script>…</script>
    <div id="wrapper">…</div>
    <div id="loginbar" style="display: none;">…</div>
    <script>…</script>
    <script>…</script>
    <a id="backTop" style="display: inline;" href="javascript:;" data-title="返回顶部">…</a>
    <a id="report_link" href="javascript:;" data-title="反馈图片">…</a>
    <script>…</script>
    <script>…</script>
    <script></script>
```

图 3-2　半结构化数据表示的 DOM

（3）非结构化数据。非结构化数据通常是指语音、图片、视频等格式的数据，如图 3-3 所示。这类数据一般按照特定应用格式进行编码，数据量非常大，且不能简单地转换成结构化数据。

图 3-3　非结构化数据

从上述分析可见，结构化数据是传统数据的主体，而半结构化和非结构化数据是大数据的主体；在数据平台设计时，结构化数据用传统的关系数据库便可高效处理，而半结构化和非结构化数据必须用 Hadoop 等大数据平台；在数据分析和挖掘时，不少工具都要求输入结构化数据，因此必须把半结构化数据先转换成结构化数据。

2．大数据的存储方式

由于大数据具有数据量大、模态和种类繁多、异构的特征，用传统的存储产品很难对这些海量数据进行存储，需要运用资源云系统对大数据进行资料存储、应用服务和资源共享等。存储产品已不再是附属于服务器的辅助设备，而成为互联网中最主要的花费所在。

存储产品已不再是附属于服务器的辅助设备，而成为互联网中最主要的花费所在。海量存储技术已成为继计算机浪潮和互联网浪潮之后的第三次浪潮，磁盘阵列与网络存储成为先锋。

资源云系统是大规模数据存储及应用服务的中心，用户把大数据资源存储到云系统中，

当用户需要得到数据资源时可通过互联网获取，当不需要这些数据资源时，还可以删除、释放这些资源[2]。

资源云系统的功能主要包括虚拟存储技术、高性能 I/O、网格存储系统等。

1）虚拟存储技术

存储虚拟化的核心工作是物理存储设备到单一逻辑资源池的映射，通过虚拟化技术，为用户和应用程序提供了虚拟磁盘或虚拟卷，并且用户可以根据需求对它进行任意分割、合并、重新组合等操作，并分配给特定的主机或应用程序，为用户隐藏或屏蔽了具体的物理设备的各种物理特性，如图 3-4 所示。

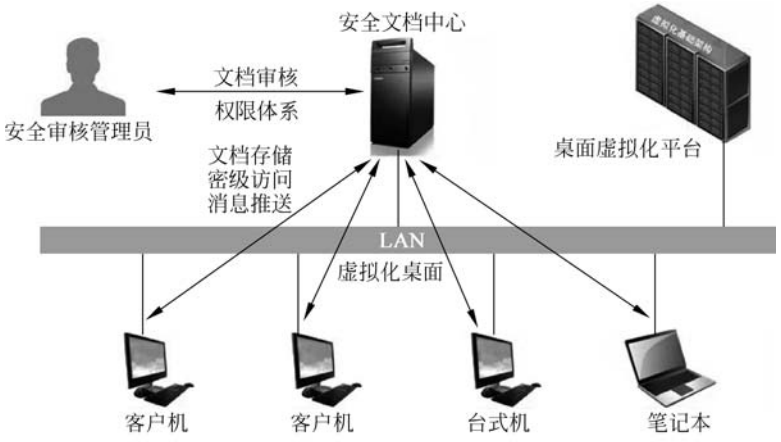

图 3-4 虚拟存储示意图

存储虚拟化可以提高存储利用率，降低成本，简化存储管理，而基于网络的虚拟存储技术已成为一种趋势，它的开放性、扩展性、管理性等方面的优势将在数据大集中、异地容灾等应用中充分体现出来。

2）高性能 I/O[3]

缓存系统在通信方面受制于传统以太网的高延迟，在存储方面受限于服务器内可部署的内存规模，亟须融合新一代高性能 I/O 技术来提升性能、扩展容量。高性能 I/O 设备的性价比正逐步提高，具备推广普及的条件。在高性能网络领域，随着高性能 I/O 技术的不断发展成熟及生产规模的扩大，未来高性能 I/O 的性价比将对商用数据中心产生巨大的吸引力。

集群由于其很高的性价比和良好的可扩展性，近年来在高性能计算集群（HPC）领域得到了广泛的应用。数据共享是集群系统中的一个基本需求，当前经常使用的是 Memcached[4] 存储。

基于日益流行的高性能远程直接内存访问通信协议，并针对不同的 Memcached 操作及消息大小设计不同的策略，降低了通信延迟；利用高性能 NVMe SSD 来扩展 Memcached 存储。

计算结点首先通过 NFS 协议从存储系统中获取数据，然后进行计算处理，最后将计算结果写入存储系统。在这个过程中，计算任务的开始和结束阶段数据读写的 I/O 负载非常

大，而在计算过程中几乎没有任何负载。

3）网格存储系统[5]

网格分布式存储系统的基本思想是利用网格存储技术，通过网络将网内零散的存储设备连接起来，汇集这些设备上的空闲存储空间，形成一个高可扩展、高可靠、高性能分布式存储系统。如果将各结点主机的存储空间构建成一个巨大的从属性系统功能，则系统可分为3层：应用层、服务层和资源层，如图3-5所示。

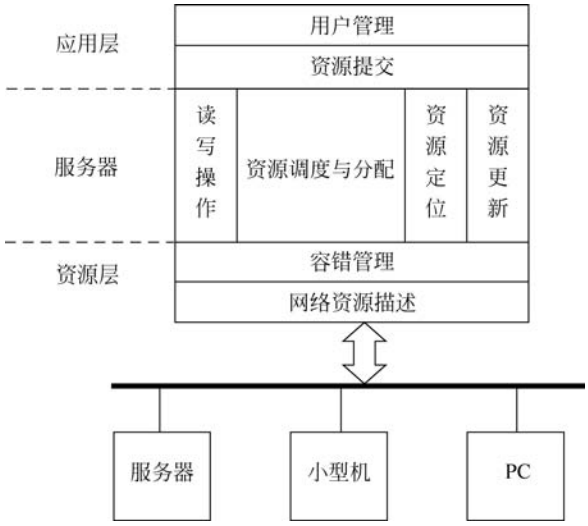

图3-5　网格存储系统层次结构

（1）应用层。用户通过用户界面直接与应用层交互。通过应用层提供的资源服务接口，用户看到的将是一个虚拟的海量存储空间，用户可以上传、下载、共享自己的资源，也可以访问由其他用户共享出来的资源。应用层主要包括用户管理模块和资源提交模块。用户管理模块负责对系统中的用户统一管理，用户按角色分类，各个角色的用户具有不同的操作权限，主要包括存储用户的申请注册、增加、修改等功能。

（2）服务层。服务层是用户使用网络资源的一个窗口，主要包括以下4个模块。

① 用户读写模块。用户登录到存储系统后，将所在结点名、主机IP地址、需要上传/下载的文件名称等基本信息报告给网格中心，并实现读写文件操作。

② 资源调度与分配模块。资源调度的目标是当用户通过接口提出任务请求时，尽可能高效及时地在分布式存储系统中找到合理的资源。

③ 资源定位模块。其主要功能是给定一个资源的描述，资源定位部件返回一个或者多个满足该描述的资源的位置。

④ 资源更新模块。当用户进行读写操作或当存储结点加入或离开时，需要对系统进行更新。文中具体表现在对全局目录索引及活动结点列表更新机制的研究方面。

（3）资源层。由地理分布的具有存储空间的主机即系统结点以及连接它们之间的底层网络构成。将社会中人们喜好的就近原则应用到系统设计中，系统将存储结点，同时根据地理位置划分为不同域。

3.1.2 大数据分析概述

大数据分析是指对规模巨大的数据进行分析，是一组能够高效存储和处理海量数据，并有效达成多种分析目标的工具及技术的集合。

下面通过美国利用大数据分析实现精准推送健康知识宣传的案例来初步认识大数据分析。案例的大数据分析基本过程如图 3-6 所示。

第 1 步：提出分析问题，精准定向投放健康诊疗知识材料。

将健康诊疗知识精准地送到需要人的手中，提升公共卫生宣传效果是有社会意义的问题。一般医疗宣传的做法是大量投放广告，需要大量人力物力，而且很难分清广告的作用。大数据技术可以对某个地区某些疾病的相关数据进行收集和分析，从而找到需要医学材料的人群。

第 2 步：大数据采集，获得居民的医院诊疗及医学网站上咨询的数据。

分析团队搜索采集数据，如这个地区居民的诊疗数据、相关的医学网站上的问诊数据，形成数据集，为数据分析做准备。

第 3 步：大数据分析，给出具体医学知识材料投放方案。

对采集的数据进行分析挖掘，为需要帮助的患者提供精准可靠的医学资料，哪个地区的患者对某种疾病知识有需求，相应医学知识就送到其电子邮箱和地区的报纸上，非常精准，节省人力物力。

第 4 步：结果可视化展示，将医学知识材料投放方案图形化。

根据数据分析结果，用图表等方式将解决方案展示出来。

第 5 步：效果评估，提升健康宣传工作效率。

与传统的医学知识宣传相比，通过大数据分析的创新方案，相关公共卫生宣传部门提高工作效率，大幅度地提高了健康宣传对象的精准度。

图 3-6 大数据分析基本过程

3.1.3 大数据分析的研究方向

大数据分析包括预测性分析、可视化分析、大数据挖掘分析、语义引擎分析、数据质量和数据管理分析 5 个主要方向。

1. 预测性分析

大数据分析最普遍的应用就是预测性分析，从大数据中挖掘出有价值的知识和规则，通过科学建模的手段呈现出结果，然后可以将新的数据代入模型，从而预测未来的情况。

例如，麻省理工学院的研究者创建了一个计算机预测模型用来分析心脏病患者丢弃的心电图数据。他们利用数据挖掘和机器学习在海量的数据中筛选，发现心电图中出现三类异常者一年内死于第二次心脏病发作的概率比未出现者高 1～2 倍。这种新方法能够预测出更多的、无法通过现有的风险筛查被探查出的高危病人，如图 3-7 所示。

2. 可视化分析

不管是对数据分析专家还是普通用户，对于大数据分析最基本的要求就是可视化分析，因为可视化分析能够直观地呈现大数据特点，同时能够非常容易地被用户所接受。可视化可以直观地展示数据，让数据自己说话，让观众看到结果。数据可视化是数据分析工具最基本的要求。图 3-8 所示是社区卫生服务站分布位置可视化。

图 3-7　心电图大数据分析

图 3-8　社区卫生服务站分布位置可视化

3. 大数据挖掘分析

可视化分析结果是给用户看的，而数据挖掘算法是给计算机看的，通过让机器学习算法，按人的指令工作，从而呈现给用户隐藏在数据之中的有价值的结果。大数据分析的理论核心就是数据挖掘算法，算法不仅要考虑数据的量，也要考虑处理的速度。目前在许多领域的研究都是在分布式计算框架上对现有的数据挖掘理论加以改进，进行并行化、分布式处理。

常用的数据挖掘方法有分类、预测、关联规则、聚类、决策树、描述和可视化、复杂数据类型挖掘(Text、Web、图形图像、视频、音频)等。有很多学者对大数据挖掘算法进行了研究。

例如，有文献提出了对适合慢性病分类的 C4.5 决策树算法进行改进，对基于 MapReduce 编程框架进行算法的并行化改造；也有文献提出对数据挖掘技术中的关联规则算法进行研究，并通过引入兴趣度对经典 Apriori 算法进行改进，提出了一种基于 MapReduce 的改进的 Apriori 医疗数据挖掘算法。

4. 语义引擎分析

数据的含义就是语义。语义技术是从词语所表达的语义层次上来认识和处理用户的检索请求。

语义引擎通过对网络中的资源对象进行语义上的标注，以及对用户的查询表达进行语义处理，使得自然语言具备语义上的逻辑关系，能够在网络环境下进行广泛有效的语义推理，从而更加准确、全面地实现用户的检索。大数据分析广泛应用于网络数据挖掘，可通过用户的搜索关键词来分析和判断用户的需求，从而实现更好的用户体验。

例如，一个语义搜索引擎试图通过上下文来解读搜索结果，它可以自动识别文本的概念

结构。如搜索“血型”,语义搜索引擎可能会获取包含“A 型血”“B 型血”和“O 型血”的文本信息。也就是说,语义搜索可以对关键词的相关词和类似词进行解读,从而扩大搜索信息的准确性和相关性。

5. 数据质量和数据管理分析

数据质量和数据管理是指为了满足信息利用的需要,对信息系统的各个信息采集点进行规范,包括建立模式化的操作规程、原始信息的校验、错误信息的反馈、矫正等一系列的过程。大数据分析离不开数据质量和数据管理。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的真实和有价值。

3.2 大数据分析处理系统

针对不同业务需求的大数据,应采用不同的分析处理系统。国内外的互联网企业都在对基于开源性面向典型应用的专用化系统进行开发。

3.2.1 批量数据及处理系统

1. 批量数据

批量数据通常是数据体量巨大,如数据从太字节(TB)级别跃升到拍字节(PB)级别,且是以静态的形式存储。这种批量数据往往是从应用中沉淀下来的数据,如医院长期存储的电子病历等。对这种数据的分析通常使用合理的算法,才能进行数据计算和价值发现。大数据的批量处理系统适用于先存储后计算,对实时性要求不高,但对数据的准确性和全面性要求较高的场景。

2. 批量数据分析处理系统

Hadoop 是典型的大数据批量处理架构,由 HDFS 负责静态数据的存储,并通过 MapReduce 实现计算逻辑、机器学习和数据挖掘算法。

3.2.2 流式数据及处理系统

1. 流式数据

流式数据是一个无穷的数据序列,序列中的每一个元素来源不同,格式复杂,序列往往包含时序特性。在大数据背景下,流式数据处理常见于服务器日志的实时采集,将拍字节级数据的处理时间缩短到秒级。数据流中往往含有错误元素、垃圾信息等,因此流式数据的处理系统要有很好的容错性,还要完成数据的动态清洗、格式处理等。例如,远程卫生环境监控是通过传感器和移动终端,对一个地区的环境卫生指标进行实时监控、远程查看、智能联动、远程控制,是通过流数据的方法系统地解决卫生环境问题。

2. 流式数据分析处理系统

流式数据分析处理系统有 Twitter 的 Storm,Facebook 的 Scribe,Linkedin 的 Samza 等。其中,Storm 是一套分布式、可靠、可容错的用于处理流式数据的系统。其流式处理作业被分发至不同类型的组件,每个组件负责一项简单的、特定的处理任务。

Storm 系统有其独特的特性。

(1) 简单的编程:类似于 MapReduce 的操作,降低了并行批处理与实时处理的复杂性。

（2）容错性：如果出现异常，Storm 将以一致的状态重新启动处理以恢复正确状态。

（3）水平扩展：其流式计算过程是在多个线程和服务器之间并行进行的。

（4）快速可靠的消息处理：Storm 利用 ZeroMQ 作为消息队列，极大地提高了消息传递的速度，任务失败时，它会负责从消息源重试消息。

3.2.3　交互式数据及处理系统

1. 交互式数据

交互式数据是操作人员与计算机以人机对话的方式产生的数据，操作人员提出请求，数据以对话的方式输入，计算机系统便提供相应的数据或提示信息，引导操作人员逐步完成所需的操作，直至获得最后处理结果。交互式数据处理灵活、直观、便于控制。采用这种方式，存储在系统中的数据文件能够被及时处理修改，同时处理结果可以立刻被使用。在互联网中的各种平台产生大量交互式数据，如搜索引擎、电子邮件、即时通讯工具、社交网络、微博以及电子商务等。

2. 交互式数据分析处理系统

交互式数据分析处理系统有 Berkeley 的 Spark 和 Google 的 Dremel 等。其中 Spark 是一个基于内存计算的可扩展的开源集群计算系统。

3.2.4　图数据及处理系统

1. 图数据

图数据是通过图形表达出来的信息含义，图自身的结构特点可以很好地表示事物之间的关系。图数据主要包括图中的结点以及连接结点的边。在图中，顶点和边实例化构成各种类型的图，如标签图、属性图、特征图以及语义图等，见图 3-9～图 3-11）。

图 3-9　标签图

图 3-10　医学特征图

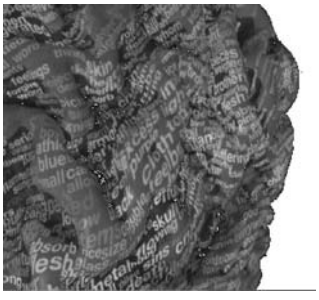

图 3-11　人脑语义地图

2. 图数据分析处理系统

典型的图数据分析处理系统有 Google 的 Pregel 系统、Neo4j 系统和微软的 Trinity 系统。Trinity 是 Microsoft 推出的一款建立在分布式云存储上的计算平台，可以提供高度并行查询处理、事务记录、一致性控制等功能。Trinity 主要使用内存存储，磁盘仅作为备份存储。

Trinity 有以下特点。

（1）数据模型是超图。超图中，一条边可以连接任意数目的图顶点，此模型中图的边称

为超边，超图比简单图的适用性更强，保留的信息更多。

（2）并发性。Trinity 可以配置在一台或上百台计算机上，提供了一个图分割机制。

（3）具有数据库的一些特点。Trinity 是一个基于内存的图数据库，有丰富的数据库特点。

（4）支持批处理。Trinity 支持大型在线查询和离线批处理，并且支持同步和不同步批处理计算。

3.3 大数据分析在医学领域的应用

大数据分析在医学领域有广泛的应用。本节以实证案例来介绍大数据分析的实际应用。

3.3.1 智能健康管理

1977 年，世界卫生组织对健康概念作出定义：不仅仅是没有疾病和身体虚弱，而是身体、心理和社会适应的完满状态。20 世纪 90 年代，健康的含义被注入了环境的因素：生理、心理、社会、环境四者的和谐统一。21 世纪，出现了健、康、智、乐、美、德六个字共同组成的全面的大健康概念。

当今社会已经迈入了工业 4.0 时代，整个医疗行业也进入了以信息化、大数据为导向的健康产业 4.0 时代。所谓健康产业 4.0 时代，是指利用大数据，将各种健康数据、各种生命体征的指标，集合在每个人的数据库和电子健康档案中；并且通过大数据的分析应用，推动覆盖全生命周期的预防、治疗、康复和健康管理的一体化健康服务。随着云计算平台、物联网、移动互联网等技术的快速发展，健康数据管理正逐渐成为现实。同时，新医改激活了长期进展缓慢的卫生信息化，引来了全国各地数字医院和区域医疗网络的建设浪潮，很多和医疗相关的 IT 新技术和新应用也随之进入医疗健康领域，智能健康管理的概念逐渐进入人们的视野。

智能健康管理整合了医疗与信息技术相关部门、企事业单位的资源。通过新型信息化技术、健康管理信息的获取、传输、处理和反馈技术，打造区域一体化协同医疗健康服务，建立高效率的健康监测、疾病防治服务体系、健康生活方式和健康风险评价体系，对区域化居民健康进行健康评价、制订健康计划、实施健康干预等过程，从而改善区域化居民健康状况，防治区域化居民常见和慢性疾病的发生和发展，提高区域化居民生命质量，降低医疗费用，最终实现全人全程全方位的智能健康管理。

2. 智能健康管理实例

实例 1：饮食控制 App

“每日三次”是一款由北京郁金香伙伴科技有限公司研发的饮食控制 App，旨在帮助用户形成更科学、健康的饮食习惯。这款 App 的设计主要基于计算机视觉识别和机器学习，利用图像识别技术，将用户拍照上传的菜品传输到模型中，自动识别其中的食物种类，判断菜品所含的热量、胆固醇、脂肪、升糖指数等指标，并根据用户的身体状况（如减肥、高尿酸、高血脂、脂肪肝、痛风等）进行饮食指导。App 的工作原理如图 3-12 所示。

图 3-12　饮食控制 App

实例 2：妙健康

“妙健康”是一家以人工智能和健康大数据为基础的健康科技公司。通过数字化精准健康管理平台专注为每位用户进行个性化健康管理，实现健康促进，降低疾病风险。平台设计如图 3-13 所示。

图 3-13　“妙健康”平台个性化健康管理

健康大数据的采集主要通过智能硬件来实现。通过不同的采集硬件，健康数据将存在于不同的智能硬件中，由此形成“数据孤岛”现象。针对这一问题，苹果生态打造了一套基于 iOS 的供应商智能穿戴等设备对数据进行收集；小米生态形成了一套丰富的自有产品线进行数据采集，但其他品牌智能设备无法接入。“妙健康”平台则致力于打造将各品类智能硬件设备融合的数据管理平台，由此实现智能健康管理的全流程，如图 3-14 所示。

图 3-14　智能硬件接入

实例 3：AI＋精神疾病管理

Avalon AI 是一家位于英国伦敦，专门研究阿尔茨海默病等神经退行性疾病预测的人工智能公司。阿尔茨海默病的病症主要包括记忆障碍、失语、失用、失认、视空间技能损害、执行功能障碍以及人格和行为改变等，病因未明。Avalon AI 公司通过深度学习核磁共振成像技术，首先对人的大脑制作 3D 磁共振图像，然后利用卷积神经网络将 3D 磁共振图像与已有的阿尔茨海默病研究中产生的失智大脑图像进行层层对比分析，最终建立病症的特征模型，从而实现对阿尔茨海默病的预测（包括是否发病、大脑损伤程度等）。磁共振影像案例如图 3-15 所示。

(a) 结构性影像　(b) 扩散性影像　(c) 功能性影像

图 3-15　磁共振影像案例

图 3-15(a)～(c)分别是结构性、扩散性、功能性的磁共振影像案例，通过深度学习方法并结合三种成像方式，可以使阿尔茨海默病漏诊概率降低 50％。

3.3.2　智能医学影像分析

1. 智能医学影像分析

医疗影像（X 射线、CT、MRI 等）数据占医疗数据中的 90％且呈增长趋势。目前大部分地区还是采用传统方法，即通过人力来分析医学影像数据，图像的复杂度、影像数据的规模等诸多因素都会降低临床诊断的精准度。因此，面对大数据时代下快速增长的影像数据，人工处理方式已经渐渐无法满足临床诊断的需求。

人工智能是一门包含计算机、数学等多种学科在内的新型交叉学科，数据资源、计算机学习能力、计算能力、算法模型等基础条件是人工智能发展的重要力量。近年来，有越来越多的人工智能方法通过改进对传统图像的处理方法，将其应用到医学图像中，形成智能医学影像分析。

人工智能等新技术的应用在提高影像医生的工作效率的同时，更能提高临床诊断的准确率。2017 年国务院正式印发《新一代人工智能发展规划》，规划提出要加快人工智能的创新应用，包含要加速实现智能影像识别、病理分型等目标。

2. 智能医学影像分析实例

实例 1：Airdoc 智能影像识别

人工智能企业 Airdoc 目前已掌握了世界领先的图像识别能力，在心血管、肿瘤、神内、五官等领域建立了多个精准的深度学习医学辅助诊断模型。在智能医学影像识别类辅助诊断系统领域享有一定的知名度，如 Airdoc DR 系统可帮助医生识别筛查糖尿病视网膜病变，如图 3-16 所示。

图 3-16　智能医学影像识别辅助诊断系统

实例 2：深睿医疗智能影像云

深睿医疗人工智能医学辅助诊断系统，运用国际前沿人工智能技术，使医学影像诊断达到国际先进水平，不仅在各系统疾病的精确诊断方面处于行业前列，更为医生进一步诊疗决策提供精准的临床建议。主要实现功能如下。

（1）影像云端存储及云端病灶筛查。云端 AI 算法的高速迭代及强大的计算设备，能够支持对病灶属性进行更完整的分析。

（2）医联体模式连通基层医院和上级医院。医生可以在不同终端、不同地点访问医学影像数据及 AI 算法 DE 处理结果，而且上级医院还可为基层医院提供更加精准的诊断建议，使得影像阅片更加灵活。

（3）内置 Dr. Wise 人工智能辅助诊断系统。病灶检出率高、假阳性率低、高效自动分割。

智能影像云的工作原理如图 3-17 所示。

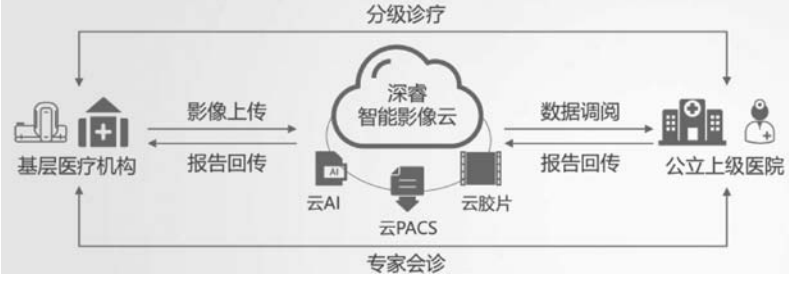

图 3-17　智能影像云

实例 3：数坤科技“加菲医生”影像诊断平台

数坤科技依托 AI 神经网络算法，打造了全球首个涵盖心脏、神经、肿瘤等多病种的 AI 影像诊断平台，并提供包括心脏病、脑卒中、癌症等危重症疾病的智能诊疗方案。依托强大的人工智能科研平台及云平台，数坤科技联合科研院校以及医疗机构，打造了人机协同的互联网智慧医疗模式，并开展基于大数据人工智能的跨区域、跨学科、多中心的临床科研工作，加速了医疗服务队伍专业化能力的培养，对用户实施精准的病患教育以及疾病监测，最终实现全民健康大目标。平台示意图如图 3-18 所示。

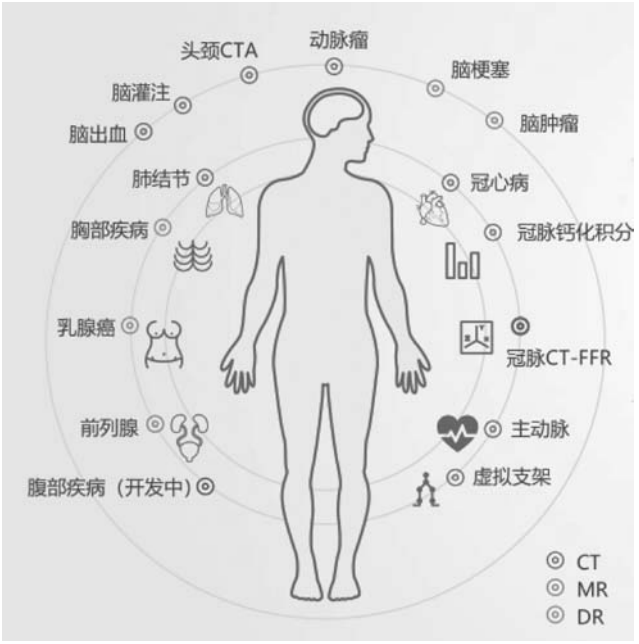

图 3-18　影像诊断示例

3.3.3　智能临床决策

1. 智能临床决策概述

临床决策支持系统(Clinical Decision Support System,CDSS)是基于人机交互的医疗信息技术应用系统,旨在为医生和其他卫生从业人员提供临床决策支持(CDS),并通过数据、模型等辅助完成临床决策。CDSS 的应用可降低因用药不当或操作不当造成的医疗事故的概率,减少对患者的不必要的伤害。CDSS 的根本目的是评估以及提高医疗质量,减少医疗差错。

CDSS 按系统结构分为两类：基于知识库的 CDSS 和基于非知识库的 CDSS。

基于知识库的 CDSS 的主要作用是满足用户的查询需求。这一类型的 CDSS 因为比较封闭并且缺乏机器的深度学习功能,所有信息的采集、编译、整理及规则均需要人工来完成,不仅维护成本高昂,还存在信息更新时效性不强的问题。

基于非知识库的 CDSS 一般采用人工智能的形式,人工神经网络具有机器学习的能力,可以在人机交互、不断训练的过程中自行总结和明确知识,并利用得到的知识为用户提供建议。随着医疗行业科技化、信息化程度的逐步提高,以及结合大数据技术的基础上,CDSS 的功能将拓展至更加广阔的空间,如医院/科室管理、科研协作平台搭建、结构化病历系统、患者交互及患者教育、医生继续教育、药物警戒、医疗控费等方向。

2. 智能临床决策实例

实例 1：深度学习应用于临床决策

深度学习是机器学习研究中的一个新领域,它的动机是建立以及模拟人脑的神经网络来进行分析学习,它通过模仿人脑的机制来解读数据,例如图像、声音和文本等。2016 年年

初，AlphaGo 击败了前世界第一的围棋选手李世石，使“深度学习”这个名词吸引了全球的关注目光。深度学习的概念源于对人工神经网络的研究，其目的是让计算机具有像人一样的智慧。深度学习利用层次化的架构学习，使得研究对象在不同层次上都得到表达，这种层次化的表达可以用来解决更加复杂抽象的问题。在层次化架构中，高层的概念往往是通过低层的概念来定义的，深度学习可以将人类难以理解的底层数据特征进行层层抽象，从而来提高数据学习的精度。让计算机建立类似人脑的神经网络进行机器学习，模仿人脑的机制来分析数据，从而实现对数据的有效表达、解释和学习，这种技术在人工智能上无疑是前景无限的。

近几年，深度学习在语音、图像、自然语言理解以及医疗诊疗等领域取得了一系列重大进展。在自然语言理解类辅助诊断系统领域，著名的 IBM Watson 机器人（如图 3-19 所示）经过 4 年多的训练，学习了 200 本肿瘤领域的教科书、290 种医学期刊和超过 1500 万份的文献后，开始被应用于临床，在肺癌、乳腺癌、直肠癌、结肠癌、胃癌和宫颈癌等领域向人类医生提出辅助建议。2015 年，Watson 用 10 分钟左右的时间就为一名 60 岁女性患者诊断出白血病，并向东京大学医科学研究所提出了适当的治疗方案。

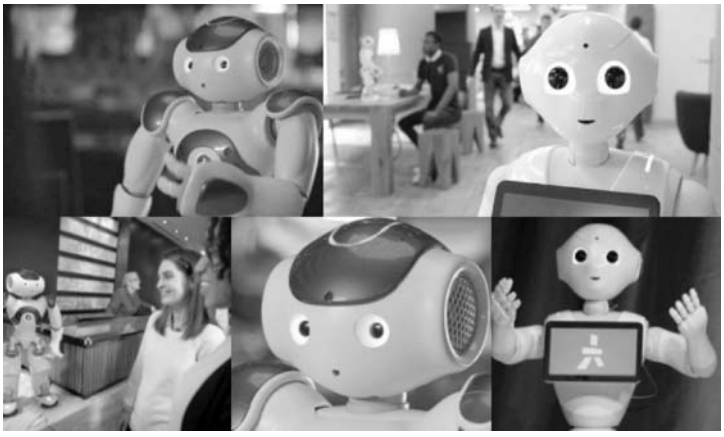

图 3-19　自然语言理解类辅助诊断机器人 Watson

实例 2：知识计算应用于临床决策

知识计算是从大数据中首先获得有价值的知识，并对其进行进一步深入的计算和分析的过程，也就是要对数据进行高度有效的分析。这就需要从大数据中先抽取出有价值的知识，并把它单独构建成可支持查询、分析与计算的知识库。知识库中的知识是显式的知识，通过利用显式的知识，人们可以进一步计算出隐式知识。支持知识计算的基础是构建知识库，这包括 3 部分：知识库的构建、多源知识的融合与知识库的更新。知识库的构建就是要构建几个基本的构成要素，包括抽取概念、实例、属性和关系。从构建方式上，可以分为手工构建和自动构建。多源知识的融合是为了解决知识的复用问题。知识库构建的代价是非常大的，为了避免从头开始，需要考虑知识的复用和共享，这就需要对多个来源的知识进行融合，即需要对概念、实例、属性和关系的冲突，重复冗余，不一致进行数据的清理工作，按融合方式可以分为手动融合和自动融合。

知识图谱泛指各种大型知识库，是把所有不同种类的信息连接在一起而得到的一个关

系网络，是机器大脑中的知识库。这个概念最早由 Google 提出，提供了从“关系”的角度去分析问题的能力。在国内，中文知识图谱的构建与知识计算也有大量的研究和开发应用。图 3-20 所示是心房颤动知识图谱；图 3-21 所示是心肌炎知识图谱。这些知识图谱必将使知识计算在医学领域发挥更大的作用。

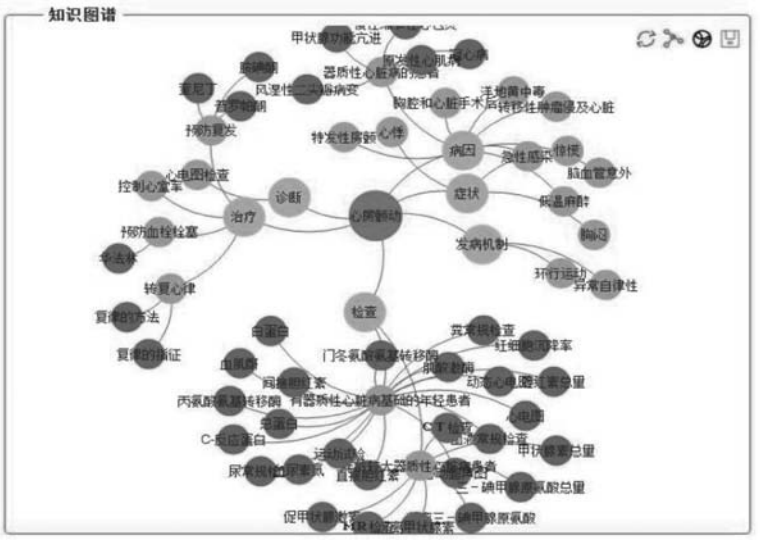

图 3-20　心房颤动知识图谱

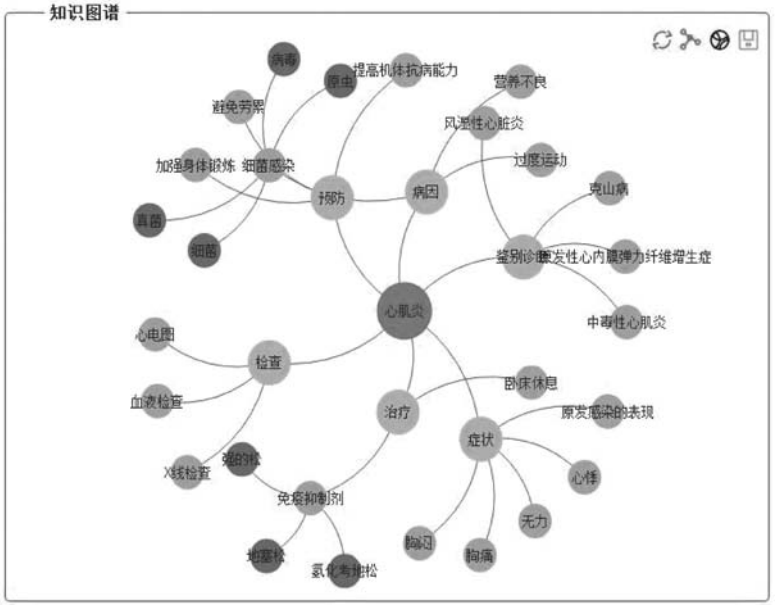

图 3-21　心肌炎知识图谱

实例 3：自然语言处理应用于临床决策

灵医智惠是由百度大脑技术驱动的 AI 医疗品牌。秉承“循证 AI 赋能基层医疗”愿景，基于灵医智惠技术平台能力，构造临床辅助决策系统、眼底影像分析系统、医疗大数据整体

解决方案、智能诊前助手、慢病管理平台等产品系列，服务院内院外全场景；广泛联合医院、医生、HIS 厂商、电子病历厂商、政府、监管等合作伙伴，通过共同推动基层医疗过程的标准化、规范化，提升基层医疗能力，降低医疗风险，控制医疗费用，服务“健康中国 2030”的国家战略。其构建的辅助问诊平台通过学习海量教材、临床指南、药典及三甲医院优质病历，基于百度自然语言处理、知识图谱等多种 AI 技术，构建符合基层医生使用习惯的辅助系统，支持分级诊疗落地，为基层医疗保驾护航。辅助问诊的工作原理如图 3-22 所示。

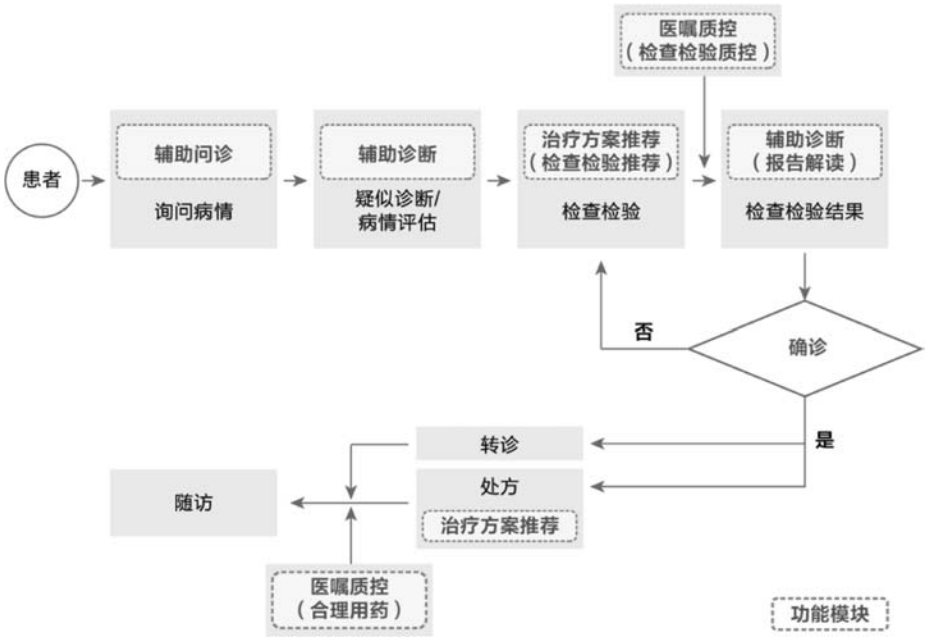

图 3-22　辅助问诊工作原理

本章小结

通过本章内容的学习，学生应该学会大数据分析的方法，掌握大数据分析的一般流程与主要技术，为医学大数据的分析应用奠定基础。大数据分析为处理结构化与非结构化的数据提供了新的途径，这些分析在具体应用方面还有很长的路要走，在未来的日子里将会看到更多的产品和应用系统在生活中出现。

【参考文献】

[1] 王星，等. 大数据分析：方法与应用[M]. 北京：清华大学出版社，2013.

[2] 张春丽，成彧. 大数据分析技术及其在医药领域中的应用[J]. 标记免疫分析与临床，2016(3)：327-329.

[3] 安仲奇，等. 基于高性能 I/O 技术的 Memcached 优化研究[J]. 计算机研究与发展，2018，20(7)：82-83.

[4] Memcached Organization. Memcached-Adistributed memory object caching system [EB/OL]. [2016-07-31]. http://memcached.org.

[5] 熊晓峰.基于网格的分布式存储系统的研究与设计[J].信息与电脑,2010,8:82-83.
[6] 张家亮.大数据分析在医疗领域中的应用[J].信息系统工程,2018,20(11):52.
[7] 郭清.智能健康管理[J].健康研究,2011,31(02):81-85.
[8] Airdoc 官网.
[9] 妙健康官网.
[10] 亿欧智库.
[11] 金征宇.前景与挑战:当医学影像遇见人工智能[J].协和医学杂志,2018(1):2-4.
[12] 王弈,李传富.人工智能方法在医学图像处理中的研究新进展[J].中国医学物理学杂志,2013,30(3):4138-4143.
[13] 深睿医疗官网.
[14] 数坤科技官网.
[15] 灵医智惠官网.

习题 3

一、填空题

1. 大数据分析处理系统有批量数据处理系统、流数据处理系统、交互数据处理系统和_____。

2. 大数据分析的基本方面有预测性分析、可视化分析、_____、语义引擎、数据质量和数据管理。

3. 大数据分析流程可以分解为:提出问题、_____、数据分析、结果可视化及结构评估等。

4. 深度学习和_____是大数据分析的基础。

5. 知识图谱泛指各种大型_____,是把所有不同种类的信息连接在一起而得到的一个关系网络。

6. 图数据中主要包括图中的结点以及连接结点的边。在图中,顶点和边实例化构成各种类型的图,如标签图、属性图、语义图以及_____等。

7. 大数据分析技术中的深度学习在语音、图像、_____等领域取得了一系列重大进展。人们对大数据的处理形式主要是对静态数据的批量处理,_____,以及对图数据的综合处理等。

8. _____是典型的大数据批量处理架构。

9. 交互式数据处理系统的典型代表是 Berkeley 的_____系统等。

10. 图数据处理有一些典型的系统,如微软的_____系统。

二、简答题

1. 简述大数据分析的概念及分析过程。

2. 简述深度学习的概念及应用。

3. 简述知识计算的概念及应用。

4. 简述批量数据的概念及特点。

5. 简述流式数据的概念及特点。

6. 简述大数据分析有哪些主要方面。